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

Sanjoy Chowdhury*[1] Hanan Gani*[2] Nishit Anand[1]
Sayan Nag[3] Ruohan Gao[1] Mohamed Elhoseiny[4] Salman Khan[2†] Dinesh Manocha[1†]
[1]University of Maryland, College Park [2]MBZUAI [3]University of Toronto [4]KAUST
{sanjoyc, nishit, rhgao, dmanocha}@umd.edu {hanan.ghani, salman.khan}@mbzuai.ac.ae
sayan.nag@mail.utoronto.ca mohamed.elhoseiny@kaust.edu.sa

Abstract

Recent advancements in reasoning optimization have greatly enhanced the performance of large language models (LLMs). However, existing work fails to address the complexities of audio-visual scenarios, underscoring the need for further research. In this paper, we introduce AURELIA*, a novel actor-critic based audio-visual (AV) reasoning framework that distills structured, step-by-step reasoning into AVLLMs at test time, improving their ability to process complex multi-modal inputs without additional training or fine-tuning. To further advance AVLLM reasoning skills, we present AVReasonBench, a challenging benchmark comprising 4500 audio-visual questions, each paired with detailed step-by-step reasoning. Our benchmark spans six distinct tasks, including AV-GeoIQ, which evaluates AV reasoning combined with geographical and cultural knowledge. Evaluating 18 AVLLMs on AVReasonBench reveals significant limitations in their multi-modal reasoning capabilities. Using* AURELIA*, we achieve up to a 100% relative improvement, demonstrating its effectiveness. This performance gain highlights the potential of reasoning-enhanced data generation for advancing AVLLMs in real-world applications. Our code and data will be publicly released at: https://github.com/schowdhury671/aurelia.*

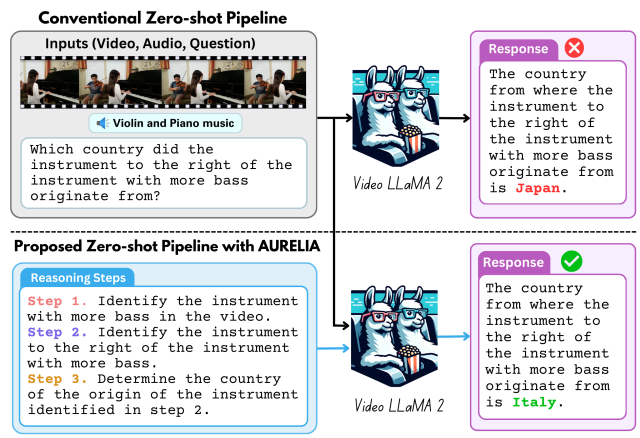


Figure 1. **Effect of injecting reasoning steps.** AURELIA enhances the ZS capabilities of audio-visual models (e.g., VideoLLaMA2). The conventional pipeline struggles in audio-visual comprehension, leading to incorrect responses. In contrast, AURELIA systematically breaks down the problem into intermediate reasoning steps, guiding the model toward more accurate and interpretable answer.

1. Introduction

Multi-agent AI systems powered by LLMs have excelled in structured reasoning tasks, including mathematical problem-solving [113, 125, 139, 143], coding assistance [158], and drug discovery [114]. These systems often employ systematic problem decomposition, as in chain-of-thought (CoT) reasoning [131]. More advanced approaches optimize reasoning through outcome reward models [144, 154], which refine solutions based on final results, and process reward models [70, 80, 152], which assess and improve intermediate steps.

Real-world reasoning extends beyond structured text-based tasks, often requiring multimodal integration, especially in audio-visual (AV) environments. Identifying a music performance's origin, for instance, involves both visual cues (e.g., attire, instruments) and audio cues (e.g., melody, language). AV reasoning is crucial for capturing abstract nuances that text or images alone cannot convey. Despite advancements in multimodal LLMs [17, 71, 110, 115, 116, 118, 126, 157], most benchmarks remain image-text focused, overlooking audio's role and its interplay with visual signals. AV reasoning presents unique challenges. Firstly, unlike static images, AV data unfolds over time, requiring models to track events, infer temporal relationships, and integrate multi-frame context. Secondly, audio often lacks direct textual mappings, making structured interpretation harder. For

*Equal contribution. †Equal advising.

example, a roaring crowd may signal excitement at a concert or unrest at a protest—context is essential for disambiguation. Current models often struggle with AV reasoning, relying on biases rather than deep cross-modal comprehension.

Moreover, current AVLLMs are susceptible to cultural, contextual, and perceptual biases embedded in their training data. As illustrated in Fig. 1, an AVLLM might incorrectly associate a musical instrument with Japan due to the presence of East Asian musicians and a Japanese track, even when the actual answer is Italy. This highlights the models' tendency to depend on dominant visual or auditory cues rather than true reasoning. While recent advances in test-time reasoning [56, 131, 162] have significantly improved text-based LLMs, these techniques remain largely unexplored for AV models.

To address these shortcomings, 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. Specifically, AURELIA employs an interactive LLM-based multi-agent framework that harnesses the reasoning capabilities of LLMs to iteratively generate high-quality reasoning data required for multimodal audio-video understanding. By leveraging the reasoning data, our approach distills structured reasoning into AVLLMs, enhancing their capabilities in multimodal audio-video commonsense reasoning, geographical understanding, music comprehension, and humor understanding.

To rigorously assess AVLLMs' reasoning capabilities, we further introduce **AVReasonBench**, a comprehensive benchmark comprising 4500 audio-visual questions, each paired with detailed step-by-step reasoning solutions generated through our pipeline. Our benchmark suite spans six distinct tasks, including the novel **AV-GeoIQ** task for geographical and cultural reasoning. Evaluating 18 existing AVLLMs on AVReasonBench reveals significant deficiencies in their ability to process dynamic audio-video content. However, incorporating AURELIA-generated reasoning solutions significantly enhances AVLLMs' performance, highlighting the impact of structured test-time reasoning. We summarize our contributions below:

- We present AURELIA, *a scalable and automated pipeline* for generating high-quality Audio-Visual reasoning data, serving as both an evaluation resource and to the best of our knowledge, the *first training-free reasoning distillation framework* for Audio Visual LLMs.
- Leveraging our proposed reasoning data generation pipeline, we introduce *AVReasonBench*, a comprehensive AV benchmark featuring 4500 audio-visual samples with detailed step-by-step reasoning solutions across six diverse tasks, encompassing multimodal commonsense reasoning, music comprehension, and humor detection. Additionally, as a part of our benchmark, we introduce a novel task *AV-GeoIQ* for geographical understanding and curate 1,000 AV-Compositional and 100 AV-Meme understanding samples through careful manual inspection.
- Leveraging our curated reasoning dataset, we demonstrate *up to 100% relative improvement* in AVLLM performance through zero-shot reasoning distillation, demonstrating the effectiveness of our approach in enhancing the reasoning capabilities of AV models.

2. Related Work

Reasoning in Multimodal LLMs. Researchers have been exploring ways to optimize CoT reasoning for MLLMs to handle increasingly complex tasks. The majority of studies focus on extracting graphical [29, 38, 51, 121], logical [30, 55, 129, 136, 161] or textual [3, 14, 138] information from images and using it to solve mathematical problems. LLaVA-CoT [138] investigates improved sampling and search algorithms to identify reasoning paths. Virgo [32] examines the organization of fine-tuning data and the transferability of text-based reasoning tasks to image-based reasoning. Recently, MAmmoTH-VL [44] developed a large-scale multimodal instruction-tuning dataset to enhance question-answering performance across various modalities. In a major departure from these previous bodies of work AURELIA specifically targets general video understanding scenarios, where various aspects of AV information are continuously referenced throughout the reasoning process.

Benchmarks for Audio-Visual LLMs. The rapid advancement of MLLMs [49, 74, 97, 104, 106, 117, 164] has driven the development of increasingly challenging video understanding benchmarks, shifting the focus from basic video description and perceptual abilities [10, 12, 22, 59, 83, 89, 109] to reasoning capabilities [24, 34, 36, 64, 66, 76]. Specifically, NExT-QA [135] emphasizes causal reasoning while Video-MME [36] features questions that necessitate integrating both audio and visual cues for effective reasoning. Our proposed AVReasonBench presents more challenging questions that demand deeper reasoning, extensive world knowledge, and a more seamless integration of AV information.

Reasoning Benchmarks While text-based benchmarks such as GSM8K [25] and MMLU [52] evaluate logical and commonsense reasoning, multimodal benchmarks remain underdeveloped. Recent efforts, such as MathVista [79] and VideoQA datasets [35, 61, 130, 137], attempt to introduce vision-based reasoning tasks, but they often emphasize perception over deeper reasoning. Moreover, existing benchmarks lack comprehensive challenges that require integration across multiple modalities, including audio, video, and world knowledge. Several works have proposed methods to assess reasoning quality in LLMs, e.g., logical consistency checks [40, 78, 120] and adversarial reasoning tasks [24, 84]. However, current benchmarks chiefly measure static performance rather than adaptive, context-dependent reasoning. Although multi-agent systems [46, 50, 88, 147] and collaborative rea-

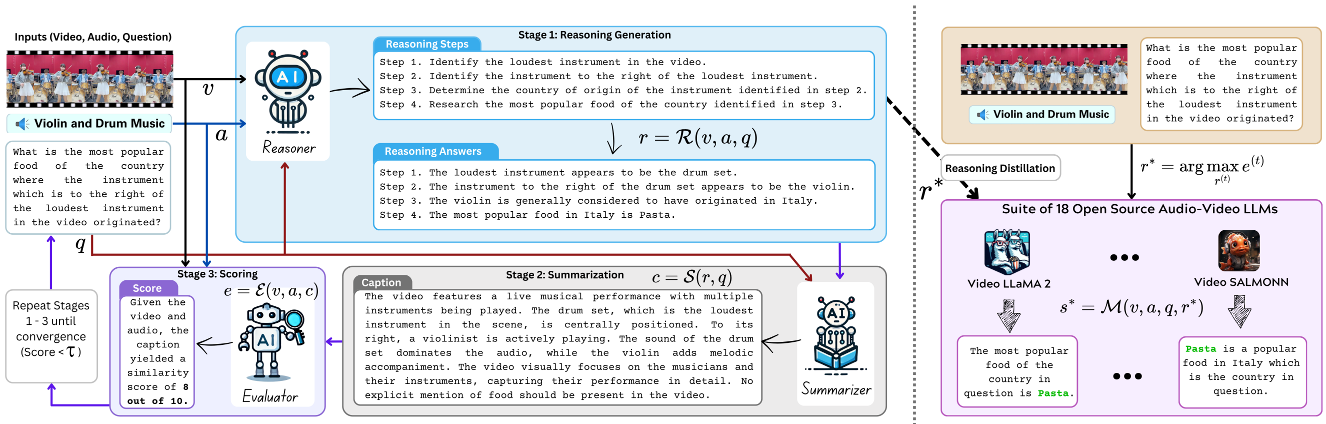


Figure 2. **Overview of AURELIA:** Our proposed AURELIA consists of a multi-agent interactive framework that functions in sync and generates reasoning steps that are then distilled inside the target model. The input set consisting of the audio, video, and question is first fed into the reasoning generator agent, which generates an initial set of reasoning steps that provide a structured pathway to reach the final answer. These reasoning steps are synthesized into a detailed caption by a Summarizer agent. The Evaluator agent then outputs a score that measures the relevance of the caption with the input audio and video. A feedback mechanism then provides supervision to the Reasoning generator based on the evaluation score, which adjusts its output to maximize the evaluation score. This actor-critique framework continues until the evaluation score exceeds a specific threshold or the number of iterations are exhausted.

soning frameworks [5, 112, 122] have shown promise in enhancing reasoning abilities, their evaluation remains fragmented across different domains.

Our work addresses these gaps by introducing a comprehensive reasoning benchmark AVReasonBench that evaluates multimodal reasoning skills in LLMs, integrating text, vision, audio, and external world knowledge. Unlike purely visual reasoning tasks, the AVR domain presents unique real-world challenges, such as temporal synchronization between audio and visual cues, ambiguity in auditory semantics, and the need for deeper cross-modal understanding.

3. Method

In this section, we will first provide an overview of audio-video multi-modal agents in Sec. 3.1, followed by a detailed description and working of AURELIA in Sec. 3.2.

3.1 Audio-Video Multi-Agent System

Our interactive audio-video multi-agent system is structured as a tuple $\langle \mathcal{R}, \mathcal{S}, \mathcal{E}, \mathcal{F} \rangle$, where multiple LLM-based agents collaboratively operate on the dataset comprising of video, audio and textual query, represented as $\langle \mathcal{V}, \mathcal{A}, \mathcal{Q} \rangle$, to enhance the performance of the target model $\mathcal{M}$. As shown in Fig. 2, the reasoning generator agent $\mathcal{R}$ processes the input video $v \in V$ and audio $a \in A$ and produces a sequence of reasoning steps r necessary for answering the given question $q \in Q$. Leveraging this information, the summarizer agent $\mathcal{S}$ extracts key cues and synthesizes them into a concise caption s that encapsulates the core content of both the video $v \in \mathcal{V}$ and the audio $a \in \mathcal{A}$. The relevance of the reasoning steps generated by $\mathcal{R}$ is assessed by the quality of the caption produced by $\mathcal{S}$. This assessment is conducted by the evaluation agent $\mathcal{E}$, a multi-modal model that takes $\{v, a, c\}$ as input and assigns a score quantifying the correctness and coherence of the reasoning steps. Based on this evaluation, a feedback mechanism ($\mathcal{F}$) iteratively refines the reasoning process by guiding $\mathcal{R}$ toward more effective reasoning paths. This interaction functions as an actor-critic framework, continuously optimizing until a satisfactory evaluation score is achieved. Ultimately, the refined reasoning steps, along with the original inputs $\langle v, a, q, r \rangle$, are fed into the target model $\mathcal{M}$. This process enhances the model's internal reasoning mechanism, leading to improved overall performance. We further present AURELIA mathematically in Algorithm 1.

3.2 AURELIA

Our proposed AURELIA enhances the performance of AVLLMS through a combination of multi-modal agents that interact with each other and generate a set of reasoning steps which distills the knowledge into the model in a training-free manner. Below, we describe the different components of AURELIA and their working in detail.

Reasoning Generator. The first component of AURELIA is a multi-modal reasoning generation agent, denoted as $\mathcal{R}$. Since our proposed method operates in a zero-shot setting, let $(x, y) \in \mathcal{D}^{test}$ represent samples from the test set, where each input x in $\mathcal{D}^{test}$ is a tuple $\langle v, a, q \rangle$, comprising a video v, an audio a, and a question q. The agent $\mathcal{R}$ processes this input tuple and produces three key outputs: a sequence of reasoning steps, a justification for these steps, and the final answer to the question. Formally,

$$r = \{r_1, r_2, r_3\} = \mathcal{R}(v, a, q), \tag{1}$$

where r_1 represents the reasoning steps, r_2 provides their justification, r_3 is the final answer to the question q.

Summarizer. The summarizer agent, denoted as $\mathcal{S}$, processes the reasoning information r generated in the previous stage along with the question q and synthesizes them into a caption c such that $c = \mathcal{S}(r, q)$. This caption provides a comprehensive summary of the video and its corresponding audio, encapsulating key details in a concise manner. The accuracy and relevance of the generated caption c depend on both the reliability of reasoning steps and the final answer produced by the reasoning generator agent. To ensure consistency and correctness, we introduce an evaluation agent $\mathcal{E}$ that assesses caption in relation to given audio and video.

Evaluator. The reliability of the reasoning steps and the generated answer directly impact the summarizer agent, which synthesizes the content into a detailed caption. Consequently, the quality of the caption is inherently tied to the correctness of the reasoning process. We hypothesize that an accurate caption aligns closely with well-formed reasoning steps, ultimately leading to a correct final answer.

To assess this alignment, we introduce a multi-modal evaluation agent $\mathcal{E}$ that serves as a judge. This agent receives the video v, audio a, and the corresponding caption c as input and assigns an evaluation score e based on their coherence. The score ranges from 1 to 10, where 1 indicates minimal alignment between the caption and the input data, while 10 signifies a perfect match.

$$e = \mathcal{E}(v, a, c), \tag{2}$$

where $e \in [1, 10]$ quantifies the relevance of the caption to the input signals and, by extension, evaluates the effectiveness of the reasoning steps in deriving the final answer.

Feedback Mechanism. Based on the evaluation score obtained in the previous step, we follow an Actor-Critic framework that facilitates iterative agent improvement through a feedback loop. In this case, the Actor is the Reasoning generating agent $\mathcal{R}$ which is evaluated by another agent $\mathcal{E}$ acting as a judge and based on the evaluation score, the Critic agent provides feedback to guide the Actor Agent in regenerating improved solutions. Let $\mathcal{F}$ be the feedback mechanism facilitating the interaction between the Actor and Critic, then the goal of the feedback mechanism is to maximize the evaluation score e such that e is above a certain threshold τ.

$$r^* = \arg\max_{r^{(t)}} \quad e^{(t)}, \quad \text{s.t} \quad e^{(t)} \geq \tau, \quad t \leq T. \tag{3}$$

If $e^{(t)} \geq \beta$ at any iteration t, the process terminates and returns the corresponding reasoning steps r^*. Otherwise, the system continues iterating, refining $r^{(t)}$ through $\mathcal{F}$ until T iterations are exhausted.

Reasoning Distillation. The optimal reasoning steps r^*, obtained through the multi-agent interaction process, serve as a structured sequence of logical inferences and contextual cues that can enhance the target model ($\mathcal{M}$) response. These steps encapsulate the essential knowledge relationships and transformations necessary to bridge the input modalities to derive an accurate and well-grounded solution. In other words, the knowledge inside the reasoning information is distilled in a training-free manner inside the target model $\mathcal{M}$ which now receives a refined and enriched input containing reasoning steps, in addition to the raw audio, video and the question, that highlight key features, intermediate conclusions, and decision pathways. By conditioning the target model $\mathcal{M}$ on the distilled reasoning steps r^*, we facilitate a more structured decision-making process, reducing ambiguity and improving model interpretability. The optimal solution s^* is formulates as,

$$s^* = \mathcal{M}(v, a, q, r^*) \tag{4}$$

💡 **AURELIA** is the first multi-agent framework capable of reasoning distillation in Audio Visual LLMs through an iterative actor-critique mechanism.
💡 **AURELIA** systematically mitigates visual and auditory biases by enforcing a structured reasoning process, leading to more objective and reliable cross-modal comprehension.
💡 **AURELIA** can scale and generalize to diverse audio-visual reasoning tasks due to zero-shot nature, where fine-tuning methods often fail due to training biases.

4. AVReasonBench: Audio-Visual Reasoning Benchmark

4.1 Why Designing AV Reasoning Tasks are Difficult?

Limitations in Forming Question-Answer Pairs for AV Setup. In vision-language tasks, the formation of question-answer pairs is relatively simple since objects have visible attributes (e.g., "What color is the book?"). However, in audio-visual reasoning, many objects do not make an inherent sound, making it harder to design meaningful QA pairs. For instance, "What does the book sound like?" lacks relevance unless an action (e.g., flipping pages) is involved. This necessitates carefully crafting interactions where both audio and visual cues contribute meaningfully.

Ambiguity in Audio-Visual Associations. Interpreting emotional tone in audio-visual tasks is challenging because the same visual cue, such as laughter, can convey different meanings depending on the accompanying audio. Cheerful music may indicate joy, while eerie background sounds might suggest nervousness or fear. Unlike vision-language tasks, where textual cues explicitly define emotions, AV models must infer meaning from the interplay of sound and visuals, requiring deeper multi-modal understanding. To encompass these scenarios we incorporate AV compositional understanding, meme understanding and dance matching tasks.

Cultural and Contextual Understanding. Object recognition and language understanding can often be generalised across cultures. If an image contains sushi, the model can easily label it as "sushi" using object detection and language mapping. However, AV tasks require deeper cultural and contextual awareness. For example, in music-dance matching, Flamenco music should pair with Flamenco dance rather than Hip-Hop. Similarly, laughter in a scene could indicate humour, but it could also indicate nervousness, depending on the visual cues. To address this gap we introduce AV-GeoIQ.

Audio-visual tasks pose additional challenges compared to only language or vision-language tasks due to the need for temporal synchronization [23], ambiguity resolution, noise handling, and cultural grounding. These challenges demand more sophisticated models that can process and align multimodal inputs dynamically over time, making AV reasoning a significantly harder problem than L/VL reasoning.

4.2 Task Overview

Audio-Visual Question Answering. Audio-visual question answering (AVQA) focuses on responding to questions that require both auditory and visual understanding. To construct our dataset, we gather question-answer pairs from AVSD [1] and MusicAVQA [60] enhancing them with detailed reasoning steps. We carefully curate samples which require strong audio-visual comprehension in terms of their interplay, association, dependency, etc.

Audio-Visual Captioning. This task involves generating detailed textual descriptions based on audio-visual inputs. Unlike image- or audio-only captioning methods, it demands robust multimodal understanding and advanced reasoning capabilities. We obtain samples from VALOR [11] for this task and augment them with reasoning annotations.

Audio-Visual Compositional Attribute Understanding. Inspired by [24], in this task we ensure each AV pair contains two separate events which are associated with two different attributes. For example, 'a cow is mooing' and a 'sheep is bleating'. Here the answer choices contain the same words but in a different sequence ('cow is bleating' and 'sheep is mooing'). An AVLLM must have a strong AV and linguistic understanding to comprehend the constituent modalities and semantically align them with the correct attributes.

AV-GeoIQ. We introduce AV-GeoIQ, a *novel* audio-visual reasoning task that integrates commonsense understanding with geographical and country-specific knowledge. This task challenges models to process and reason over multimodal inputs, requiring the alignment of audio cues, visual elements, and world knowledge. Unlike standard audio-visual question-answering tasks, AV-GeoIQ extends beyond perceptual understanding by incorporating reasoning over cultural and geographic attributes. For example, a question like *"What is the most famous drink of the country where the instrument to the left of the louder sounding instrument originates?"* necessitates multiple reasoning steps: identifying the loudest instrument, determining the relative position of another instrument, recognizing its country of origin, and retrieving cultural knowledge about that country's famous drinks—leading to the answer *Sangria*. Such questions require deep multimodal comprehension, contextual association, and factual world knowledge. AV-GeoIQ (Fig. 1) serves as a benchmark to evaluate the reasoning capabilities of AVLLMs in handling complex, real-world scenarios that go beyond direct perception.

Algorithm 1 AURELIA

1: **Input:** Data: D^{test}, Reasoning generator $\mathcal{R}$, Summarizer $\mathcal{S}$, Evaluator $\mathcal{E}$, Iterations T, Threshold τ
2: **Output:** Optimized Reasoning Steps r^*
3: Sample data: $\langle$Audio a, Video v, Question $q\rangle \subseteq \mathcal{D}^{test}$
4: Set iteration counter, $t = 1$
5: **while** $e \leq \tau$ **and** $t \leq T$ **do**
6: Generate Reasoning Steps, $r^{(t)} = \mathcal{R}(v, a, q)$
7: Generate Caption, $c^{(t)} = \mathcal{S}(r^{(t)}, q)$
8: Evaluate Generated Caption, $e^{(t)} = \mathcal{E}(v, a, c^{(t)})$
9: Feedback (Repeat Steps 6-8), $\mathcal{F}(v, a, q, e^{(t)})$
10: Update $t \leftarrow t + 1$
11: Select Optimal Reasoning, $r^* = \arg\max_{r^{(t)}} e^{(t)}$
12: **return** r^*

AV Meme Understanding. Inspired by AV-Odyssey Bench [42], we include AV-Meme a task that challenges models to interpret humour, sarcasm, and the context in multimodal memes by analyzing visual elements, audio cues, and text. Unlike traditional meme analysis, AV-Meme requires grasping subtle relationships between sound effects, expressions, and captions. For example, dramatic music over an ordinary event or mismatched audio-visual pairings create irony, demanding nuanced cultural awareness. This task serves as a benchmark for evaluating AVLLMs in recognizing implicit meanings and internet humour.

Dance and Music Matching. We also include Dance-Music Matching (DM-Match) [42], a task that evaluates a model's ability to align dance movements with appropriate musical styles by analyzing audio-visual correlations. Unlike standard motion or music classification, DM-Match requires understanding rhythm, tempo, and movement patterns to determine whether a given dance sequence matches the accompanying music. For instance, a ballet performance set to fast-paced electronic music may indicate a mismatch, while a tango paired with traditional tango music would be correct. This task serves as a benchmark for assessing AVLLMs in capturing temporal synchronization, genre compatibility, and expressive coherence between dance and music.

4.3 AVReasonBench Size

We carefully curate 1000 samples each from Music-AVQA, AVSD, and VALOR which are suitable for AV reasoning. For the AV compositional understanding task, we collect 1000 samples from the web through careful manual inspection. For AV-GeoIQ we again tailor-make 200 samples which

Models	AV-QA Music-AVQA	AV-QA AVSD	AV-Captioning	AV-Compositional	AV-GeoIQ	AV-Meme	DM-Match
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)	$62.6^{+6.82\%}$	$66.5^{+10.64\%}$	$78.8^{+8.09\%}$	$33.8^{+15.75\%}$	$39.0^{+52.94\%}$	$50.0^{+21.95\%}$	$35.0^{+27.27\%}$

Table 1. **Performance comparison of various models across multiple tasks in AVReasonBench**. The lower section highlights the performance improvement using AURELIA. The numbers in teal denotes relative gains over ZS results. Video-LLamA2 zero-shot is not reported because the publicly available model is already fine-tuned on the dataset. For ZS evaluation *A/B* represents best/mean of 3 runs evaluation. AV-Captioning values denote CIDEr scores.

Model	AV-Captioning BLEU@4 ↑	AV-Captioning METEOR ↑	AV-Captioning ROUGE ↑
Zero-shot			
AVLLM	10.2	18.1	34.6
OneLLM	11.3	19.7	36.1
AVicuna	10.6	19.1	35.4
CREMA	11.5	20.1	36.9
VITA	12.9	22.8	40.3
Zero-shot with AURELIA			
AVLLM	12.8	21.9	40.7
OneLLM	14.1	24.3	42.1
AVicuna	12.8	23.7	41.8
CREMA	13.8	24.9	43.3
VITA	14.5	26.0	46.4

Table 2. **Evaluation results of five models on the AV-Captioning**. The top section indicates ZS inference results of models. The bottom section indicates results after reasoning distillation with AURELIA. Clearly, the quality of the captions improves with our reasoning pipeline.

Subset	Modality	Category: *Knowledge*	*Film & Television*	*Sports Competition*	*Artistic Performance*	*Life Record*	*Multilingual*	Overall
Short	ZS	81.4	87.5	78.7	86.7	85.6	86.7	84.4
	+ AURELIA	85.6	91.3	81.2	88.0	88.9	89.4	**87.4**
Medium	ZS	80.2	83.9	72.1	84.3	76.8	100.0	82.8
	+ AURELIA	83.3	86.5	75.9	87.1	78.2	100.0	**85.16**
Long	ZS	81.1	73.2	72.6	63.3	66.7	83.3	73.3
	+ AURELIA	85.5	77.4	75.7	67.1	69.9	86.3	**76.98**
Overall	ZS	80.9	82.4	74.6	78.8	78.0	89.7	80.7
	+ AURELIA	83.4	85.3	77.8	81.0	82.3	92.6	**83.73**

Table 3. **Performance of VITA across Video-MME**. Table shows the performance of VITA on 6 major categories of Video-MME. The evaluation is done on audio-visual inputs.

require strong AV reasoning capabilities. We augment more videos to the original AV-meme set to make a total of 100 test samples while we adapt 200 samples of DM-Match to make the total size of our reasoning benchmark, AVReasonBench to 4500. We add further details in the supplementary.

4.4 Reasoning Data Generation

For each test sample comprising an audio, a video, and a question, we supplement the input with reasoning information at inference time before feeding it into the target model through a structured multi-agent pipeline. This ensures that model decisions are grounded in logical deductions rather than implicit associations, enhancing both accuracy and interpretability. For instance, in Fig. 2, the video showcases people playing musical instruments, accompanied by audio, and the question to identify the *most popular food of the country* through a complex audio-visual referral. To answer

Reason Gen.	Summ.	Eval.	AV-GeoIQ	AV-Comp	DM-Match
Gemini	Gemini	Gemini	36.5	30.2	33.0
Gemini	GPT-4o	Gemini	38.0	31.6	34.5

Table 4. **Effect of using a combination of agents.** Using a combination of different closed-source LLMs as agents proves beneficial compared to using a single type of LLM.

Iteration (T)	AV-Cap	AV-Meme	AV-GeoIQ	AV-Comp	DM-Match	Time
1	68.8	25.0	27.5	27.3	26.5	16.28
3	73.2	30.0	34.0	32.0	31.5	45.66
5	74.7	35.0	38.0	31.6	34.5	74.01

Table 5. **Effect of number of iterations.** The results improve as the number of feedback iterations increase. Time: time required to generate reasoning steps per sample

Threshold (τ)	AV-Cap	AV-Meme	AV-GeoIQ	AV-Comp	DM-Match	Time
4	69.6	26.5	28.5	27.9	28.5	23.90
6	72.2	30.0	32.5	29.7	32.0	47.15
8	74.7	35.0	38.0	31.6	34.5	61.28
10	74.8	35.0	38.0	31.4	34.5	65.81

Table 6. **Effect of Threshold Value.** A larger threshold for the evaluation score shows positive trend on the performance.

this, the model must first *identify the loudest instrument* via audio analysis followed by determining spatial relationships to *locate the musical instrument*. Once the instrument is located, the model must infer the instrument's origin, and finally retrieve the corresponding cuisine. This structured reasoning provided by our AURELIA enforces logical progression, reducing errors and hallucinations while enhancing interpretability. We defer more details to supplementary.

5. Experiments and Results

5.1 Baselines

We extensively evaluate VideoLLaMA [153], VideoLLaMA2 [18], Reka Core [119], Gemini 1.5 Pro [102], Unified-IO-2 [77], X-InstructBLIP [92], PandaGPT [106], OneLLM [48], AnyGPT [151], NExT-GPT [134], VITA [37], VideoSALMONN [111], ImagebindLLM [49], MacawLLM [81], CAT [142], AVicuna [117], CREMA [145]. AVLLM [104] on AVReasonBench.

5.2 Metrics

For AV-QA, AV-Comp, AV-GeoIQ, AV-Meme, and DM-Match, we report the Top-1 accuracy as the metric by extracting the model outputs using a choice extraction strategy outlined in the supplementary. We report the performance of AV captioning tasks on several established metrics, including BLUE@4 [94], METEOR [2], ROGUE [72], and CIDEr [123]. We employ GPT-based evaluation for AV-GeoIQ and AVSD which has open-ended answers.

5.3 Main Results

We extensively compare the performance of the baseline AVLLMs in Tab. 1 across all 6 AV tasks of our AVReasonBench benchmark. The experimental results reveal that closed-source models consistently outperform open-source ones in every reasoning task. Specifically, among the two closed-source models, we observe that Gemini 1.5 Pro surpasses Reka Core, likely due to its superior audio comprehension capabilities. This suggests that our AVReasonBench benchmark presents challenging scenarios that require strong audio-visual joint understanding. By leveraging the zero-shot reasoning distillation through AURELIA, we observe consistent boost in the performance of all the AVLLMs as seen from the experimental results with relative improvements up to 100% for X-InstructBLIP. Furthermore, for more challenging tasks such as AV-GeoIQ, AV-Meme, and DM-Match, we observe substantial improvements highlighting the importance of AURELIA's step by step reasoning distillation in deriving answers to complex AV queries.

We further note that recent approaches such as Unified-IO-2 XL and VITA demonstrate improved reasoning abilities over the other methods due to their stronger LLM backbone, which is capable of capturing finer multimodal information. Models with more robust audio encoders, such as AVicuna and Video-SALMONN, outperform alternatives like PandaGPT and Macaw-LLM. This highlights the critical role of the audio modality in leveraging the strengths of AVReasonBench.

Tab. 2 presents the AV-captioning results for five AVLLMs across three additional captioning metrics. As shown in the table, all models exhibit consistent improvements, highlighting the effectiveness of our reasoning-enhanced data in the dense captioning task.

Results on other benchmarks. Tab. 3 results demonstrate that our reasoning pipeline is generalizable across other benchmarks. We select VideoMME [36] as an alternative benchmark due to its tasks, which demand advanced reasoning abilities. Notably, the greatest improvements are observed in the long video *Knowledge* assessment categories, further emphasizing the generalizability of AURELIA.

5.4 Ablation Study

Combination of Agents. The multi-agent framework of AURELIA offers the flexibility to integrate various existing multi-modal LLMs as specialized agents. To assess the impact of different LLMs on reasoning generation, summarization, and evaluation, we conduct an analysis on three datasets across target model VideoLLaMA-2 (Tab. 4). Our findings indicate that leveraging a combination of models, specifically GPT-4o alongside Gemini yields superior performance compared to employing Gemini alone for all three agents roles as is evident from the higher accuracy scores in case of combination of agents. This suggests that while Gemini excels in processing multi-modal inputs such as video and audio, GPT-4o demonstrates stronger capabilities in textual comprehension and reasoning. The synergy between these models enhances the overall effectiveness of AURELIA, underscoring the advantages of a diversified agent selection.

Number of Generation Attempts. Our analysis reveals that the choice of T significantly influences overall performance. To evaluate this impact, we conduct an ablation study on five datasets across VideoLLaMA-2 model, as presented in Tab. 5. With just a single iteration, the obtained scores

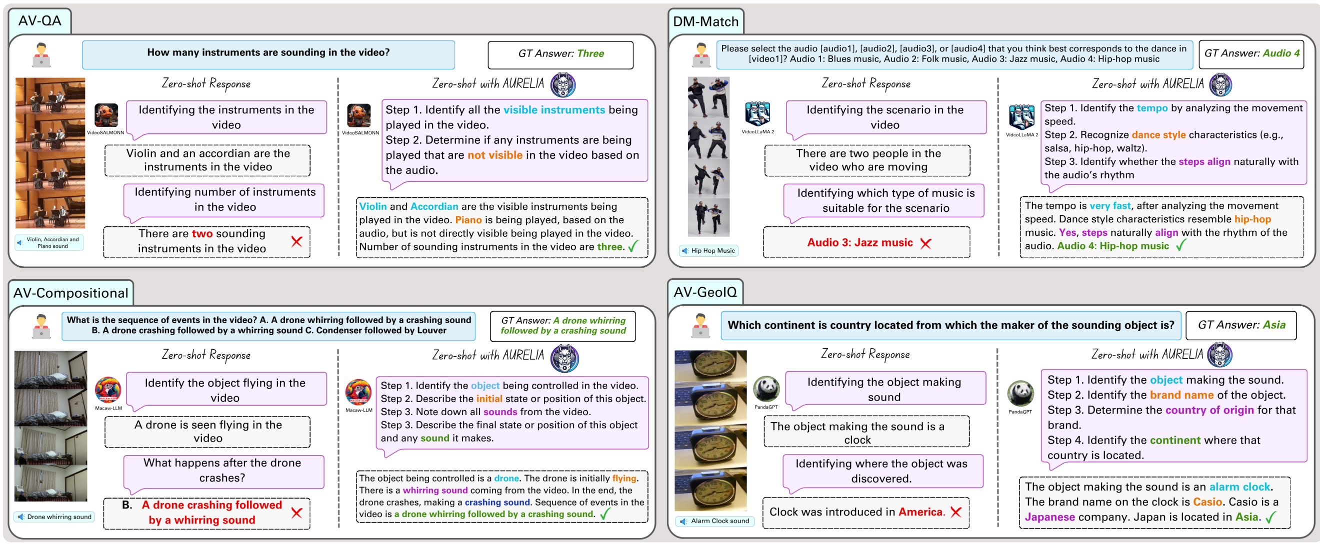


Figure 3. **Qualitative Visualizations.** Figure shows the qualitative visualizations of effect of AURELIA's reasoning distillation on the final answer across four tasks. Compared to vanilla zero-shot inference, AURELIA augments the target model with reasoning capabilities, leading to the improved answers.

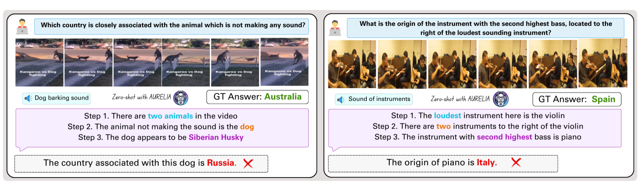


Figure 4. **Examples of Failure Cases. (Left)** AURELIA fails to comprehend audio, focus on single modality i.e. video, leading to incorrect reasoning chain. **(Right)** AURELIA fails to comprehend the dynamics of the video.

are notably low, whereas increasing the iterations to five yields substantial improvements across most datasets. This suggests that additional iterations allow AURELIA to progressively enhance its reasoning quality. However, considering computational efficiency and latency constraints, we cap the number of iterations at five for the final evaluation. AV-Cap values are CIDEr scores.

Threshold Value. Evaluation score (τ) quantifies the consistency of the reasoning steps with multimodal input. To empirically analyze the impact of the threshold (τ), we present results in Tab. 6 on five datasets across VideoLLaMA-2 model. As expected, a higher threshold value indicates stronger alignment, leading to superior model performance. However, we observe that a threshold of 8 yields performance comparable to the highest value, suggesting that setting the threshold at 8 or above ensures optimal reasoning quality. The increasing value of time required to generate the samples indicate to obtain improved reasoning steps we need more iterations. AV-Cap report CIDEr values.

5.5 Qualitative Results

To visualize the effect of AURELIA's reasoning distillation, refer to Fig. 3. We compare the performance of various AVLLMs on 4 tasks. We notice that in the absence of reasoning distillation, the target model faces difficulties in figuring out answers to the given queries. For example, in the AV-Captioning task, due to the step wise guidance to the AVLLM, the generated caption is dense and rick of contextual information compared to ZS response. Similarly, for AV-GeoIQ, powered by the sequence of prompts, the AVLLM is able to correctly respond to the query whereas, the response in ZS is wrong. Empirical studies reveal, with the addition of reasoning information, the decision making capability of model improves by structuring its response in accordance with the reasoning steps, thereby leading to correct answers. We add more qualitative results in the supplementary.

5.6 Failure Cases

Fig. 4 illustrates a few failure cases in our reasoning generation pipeline. In the first example, an error in interpreting the animal sounds leads to the assumption that the *dog* is silent. This assumption propagates through the reasoning steps, producing an incorrect response. In the second example, the pipeline fails to spot the *instrument with the second highest bass*, resulting in an erroneous conclusion. We believe that fine-grained AV comprehension and refining understanding of language instructions can help mitigate these issues.

6. Conclusion

In this work, we introduce AURELIA, a novel test-time framework designed to enhance the reasoning capabilities of AVLLMs through interactive multi-agent system which distills structured, step-by-step reasoning into AVLLMs without any training. To further advance the AVLLMs' reasoning abilities, we also present AVReasonBench, a comprehensive benchmark consisting of six diverse tasks including the novel AV-GeoIQ for geo-cultural knowledge reasoning. The samples in each task are paired with step-by-step reasoning data, generated using AURELIA, which facilitates both the evaluation and enhancement of existing AVLLMs. AURELIA serves as an essential step toward more robust, context-aware, and reasoning-driven multimodal AI, enabling future advancements in artificial audio-visual intelligence.

References

[1] Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K. Marks, Chiori Hori, Peter Anderson, Stefan Lee, and Devi Parikh. Audio-visual scene-aware dialog, 2019.

[2] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

[3] Zhenni Bi, Kai Han, Chuanjian Liu, Yehui Tang, and Yunhe Wang. Forest-of-thought: Scaling test-time compute for enhancing llm reasoning. *arXiv preprint arXiv:2412.09078*, 2024.

[4] Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*, 2023.

[5] Hanxiong Chen, Shaoyun Shi, Yunqi Li, and Yongfeng Zhang. Neural collaborative reasoning. In *Proceedings of the web conference 2021*, pages 1516–1527, 2021.

[6] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023.

[7] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

[8] Liqun Chen, Zhe Gan, Yu Cheng, Linjie Li, Lawrence Carin, and Jingjing Liu. Graph optimal transport for cross-domain alignment. In *International Conference on Machine Learning*, pages 1542–1553. PMLR, 2020.

[9] Pei Chen, Boran Han, and Shuai Zhang. Comm: Collaborative multi-agent, multi-reasoning-path prompting for complex problem solving. *arXiv preprint arXiv:2404.17729*, 2024.

[10] Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Weining Wang, Jinhui Tang, and Jing Liu. Valor: Vision-audio-language omni-perception pretraining model and dataset. *arXiv preprint arXiv:2304.08345*, 2023.

[11] Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Weining Wang, Jinhui Tang, and Jing Liu. Valor: Vision-audio-language omni-perception pretraining model and dataset. *arXiv preprint arXiv:2304.08345*, 2023.

[12] Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset. *Advances in Neural Information Processing Systems*, 36, 2024.

[13] Yongchao Chen, Jacob Arkin, Yang Zhang, Nicholas Roy, and Chuchu Fan. Scalable multi-robot collaboration with large language models: Centralized or decentralized systems? In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4311–4317. IEEE, 2024.

[14] Yongchao Chen, Harsh Jhamtani, Srinagesh Sharma, Chuchu Fan, and Chi Wang. Steering large language models between code execution and textual reasoning. *arXiv preprint arXiv:2410.03524*, 2024.

[15] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020.

[16] Zhipeng Chen, Kun Zhou, Wayne Xin Zhao, Junchen Wan, Fuzheng Zhang, Di Zhang, and Ji-Rong Wen. Improving large language models via fine-grained reinforcement learning with minimum editing constraint. *arXiv preprint arXiv:2401.06081*, 2024.

[17] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024.

[18] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024.

[19] Sanjoy Chowdhury, Aditya Patra, Subhrajyoti Dasgupta, and Ujjwal Bhattacharya. Audvisum: Self-supervised deep reinforcement learning for diverse audio-visual summary generation. In *BMVC*, page 315, 2021.

[20] Sanjoy Chowdhury, Sreyan Ghosh, Subhrajyoti Dasgupta, Anton Ratnarajah, Utkarsh Tyagi, and Dinesh Manocha. Adverb: Visually guided audio dereverberation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7884–7896, 2023.

[21] Sanjoy Chowdhury, Sayan Nag, and Dinesh Manocha. Apollo: Unified adapter and prompt learning for vision language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[22] Sanjoy Chowdhury, Sayan Nag, Subhrajyoti Dasgupta, Jun Chen, Mohamed Elhoseiny, Ruohan Gao, and Dinesh Manocha. Meerkat: Audio-visual large language model for grounding in space and time. In *European Conference on Computer Vision*, 2024.

[23] Sanjoy Chowdhury, Sayan Nag, KJ Joseph, Balaji Vasan Srinivasan, and Dinesh Manocha. Melfusion: Synthesizing music from image and language cues using diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26826–26835, 2024.

[24] Sanjoy Chowdhury, Sayan Nag, Subhrajyoti Dasgupta, Yaoting Wang, Mohamed Elhoseiny, Ruohan Gao, and Dinesh

Manocha. Avtrustbench: Assessing and enhancing reliability and robustness in audio-visual llms. *arXiv preprint arXiv:2501.02135*, 2025.

[25] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[26] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023.

[27] I De Zarzà, J De Curtò, Gemma Roig, Pietro Manzoni, and Carlos T Calafate. Emergent cooperation and strategy adaptation in multi-agent systems: An extended coevolutionary theory with llms. *Electronics*, 12(12):2722, 2023.

[28] Ahana Deb, Sayan Nag, Ayan Mahapatra, Soumitri Chattopadhyay, Aritra Marik, Pijush Kanti Gayen, Shankha Sanyal, Archi Banerjee, and Samir Karmakar. Beats: Bengali speech acts recognition using multimodal attention fusion. In *Proc. Interspeech 2023*, pages 3392–3396, 2023.

[29] Naihao Deng, Zhenjie Sun, Ruiqi He, Aman Sikka, Yulong Chen, Lin Ma, Yue Zhang, and Rada Mihalcea. Tables as texts or images: Evaluating the table reasoning ability of llms and mllms. *arXiv preprint arXiv:2402.12424*, 2024.

[30] Yuhao Dong, Zuyan Liu, Hai-Long Sun, Jingkang Yang, Winston Hu, Yongming Rao, and Ziwei Liu. Insight-v: Exploring long-chain visual reasoning with multimodal large language models. *arXiv preprint arXiv:2411.14432*, 2024.

[31] Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*, 2023.

[32] Yifan Du, Zikang Liu, Yifan Li, Wayne Xin Zhao, Yuqi Huo, Bingning Wang, Weipeng Chen, Zheng Liu, Zhongyuan Wang, and Ji-Rong Wen. Virgo: A preliminary exploration on reproducing o1-like mllm. *arXiv preprint arXiv:2501.01904*, 2025.

[33] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[34] Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. *Advances in Neural Information Processing Systems*, 37:89098–89124, 2025.

[35] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024.

[36] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024.

[37] Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Meng Zhao, Yifan Zhang, Shaoqi Dong, Xiong Wang, Di Yin, Long Ma, et al. Vita: Towards open-source interactive omni multimodal llm. *arXiv preprint arXiv:2408.05211*, 2024.

[38] Jiale Fu, Yaqing Wang, Simeng Han, Jiaming Fan, Chen Si, and Xu Yang. Graphic: A graph-based in-context example retrieval model for multi-step reasoning. *arXiv preprint arXiv:2410.02203*, 2024.

[39] Shengyi Gao, Zhe Chen, Guo Chen, Wenhai Wang, and Tong Lu. Avsegformer: Audio-visual segmentation with transformer. In *Proceedings of the AAAI conference on artificial intelligence*, pages 12155–12163, 2024.

[40] Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361, 2021.

[41] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023.

[42] Kaixiong Gong, Kaituo Feng, Bohao Li, Yibing Wang, Mofan Cheng, Shijia Yang, Jiaming Han, Benyou Wang, Yutong Bai, Zhuoran Yang, et al. Av-odyssey bench: Can your multimodal llms really understand audio-visual information? *arXiv preprint arXiv:2412.02611*, 2024.

[43] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[44] Jarvis Guo, Tuney Zheng, Yuelin Bai, Bo Li, Yubo Wang, King Zhu, Yizhi Li, Graham Neubig, Wenhu Chen, and Xiang Yue. Mammoth-vl: Eliciting multimodal reasoning with instruction tuning at scale. *arXiv preprint arXiv:2412.05237*, 2024.

[45] Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*, 2024.

[46] Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. Large language model based multi-agents: A survey of progress and challenges, 2024.

[47] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 976–980. IEEE, 2022.

[48] Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. Onellm: One framework to align all modalities with language. *arXiv preprint arXiv:2312.03700*, 2023.

[49] Jiaming Han, Renrui Zhang, Wenqi Shao, Peng Gao, Peng Xu, Han Xiao, Kaipeng Zhang, Chris Liu, Song Wen, Ziyu Guo, et al. Imagebind-llm: Multi-modality instruction tuning. *arXiv preprint arXiv:2309.03905*, 2023.

[50] Shanshan Han, Qifan Zhang, Yuhang Yao, Weizhao Jin, Zhaozhuo Xu, and Chaoyang He. Llm multi-agent systems: Challenges and open problems. *arXiv preprint arXiv:2402.03578*, 2024.

[51] Wei He, Zhiheng Xi, Wanxu Zhao, Xiaoran Fan, Yiwen Ding, Zifei Shan, Tao Gui, Qi Zhang, and Xuanjing Huang. Distill visual chart reasoning ability from llms to mllms. *arXiv preprint arXiv:2410.18798*, 2024.

[52] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021.

[53] Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. Large language models can self-improve. *arXiv preprint arXiv:2210.11610*, 2022.

[54] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19113–19122, 2023.

[55] Jihyung Kil, Zheda Mai, Justin Lee, Arpita Chowdhury, Zihe Wang, Kerrie Cheng, Lemeng Wang, Ye Liu, and Wei-Lun Chao. Mllm-compbench: A comparative reasoning benchmark for multimodal llms. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

[56] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

[57] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*, 2023.

[58] Nicholas Lee, Thanakul Wattanawong, Sehoon Kim, Karttikeya Mangalam, Sheng Shen, Gopala Anumanchipalli, Michael W Mahoney, Kurt Keutzer, and Amir Gholami. Llm2llm: Boosting llms with novel iterative data enhancement. *arXiv preprint arXiv:2403.15042*, 2024.

[59] Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19108–19118, 2022.

[60] Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19108–19118, 2022.

[61] Haopeng Li, Andong Deng, Qiuhong Ke, Jun Liu, Hossein Rahmani, Yulan Guo, Bernt Schiele, and Chen Chen. Sports-qa: A large-scale video question answering benchmark for complex and professional sports. *arXiv preprint arXiv:2401.01505*, 2024.

[62] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.

[63] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[64] 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. *arXiv preprint arXiv:2311.17005*, 2023.

[65] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022.

[66] Shicheng Li, Lei Li, Yi Liu, Shuhuai Ren, Yuanxin Liu, Rundong Gao, Xu Sun, and Lu Hou. Vitatecs: A diagnostic dataset for temporal concept understanding of video-language models. In *European Conference on Computer Vision*, pages 331–348. Springer, 2024.

[67] Siheng Li, Cheng Yang, Zesen Cheng, Lemao Liu, Mo Yu, Yujiu Yang, and Wai Lam. Large language models can self-improve in long-context reasoning. *arXiv preprint arXiv:2411.08147*, 2024.

[68] Xinyi Li, Sai Wang, Siqi Zeng, Yu Wu, and Yi Yang. A survey on llm-based multi-agent systems: workflow, infrastructure, and challenges. *Vicinagearth*, 1(1):9, 2024.

[69] Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*, 2023.

[70] Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2023.

[71] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023.

[72] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[73] Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Z Xu, Difei Gao, Rong-Cheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. *Advances in Neural Information Processing Systems*, 35:7575–7586, 2022.

[74] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[75] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023.

[76] 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? *arXiv preprint arXiv:2403.00476*, 2024.

[77] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. *arXiv preprint arXiv:2206.08916*, 2022.

[78] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.

[79] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.

[80] Liangchen Luo, Yinxiao Liu, Rosanne Liu, Samrat Phatale, Harsh Lara, Yunxuan Li, Lei Shu, Yun Zhu, Lei Meng, Jiao Sun, et al. Improve mathematical reasoning in language models by automated process supervision. *arXiv preprint arXiv:2406.06592*, 2, 2024.

[81] Chenyang Lyu, Minghao Wu, Longyue Wang, Xinting Huang, Bingshuai Liu, Zefeng Du, Shuming Shi, and Zhaopeng Tu. Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration. *arXiv preprint arXiv:2306.09093*, 2023.

[82] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023.

[83] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023.

[84] Atsuyuki Miyai, Jingkang Yang, Jingyang Zhang, Yifei Ming, Qing Yu, Go Irie, Yixuan Li, Hai Li, Ziwei Liu, and Kiyoharu Aizawa. Unsolvable problem detection: Evaluating trustworthiness of vision language models, 2024.

[85] Seungwhan Moon, Andrea Madotto, Zhaojiang Lin, Tushar Nagarajan, Matt Smith, Shashank Jain, Chun-Fu Yeh, Prakash Murugesan, Peyman Heidari, Yue Liu, et al. Anymal: An efficient and scalable any-modality augmented language model. *arXiv preprint arXiv:2309.16058*, 2023.

[86] Shehan Munasinghe, Rusiru Thushara, Muhammad Maaz, Hanoona Abdul Rasheed, Salman Khan, Mubarak Shah, and Fahad Khan. Pg-video-llava: Pixel grounding large video-language models. *arXiv preprint arXiv:2311.13435*, 2023.

[87] Sayan Nag, Koustava Goswami, and Srikrishna Karanam. Safari: Adaptive s equence tr a ns f ormer for we a kly supervised r eferring expression segmentat i on. In *European Conference on Computer Vision*, pages 485–503. Springer, 2024.

[88] Nathalia Nascimento, Paulo Alencar, and Donald Cowan. Self-adaptive large language model (llm)-based multiagent systems. In *2023 IEEE International Conference on Autonomic Computing and Self-Organizing Systems Companion (ACSOS-C)*, pages 104–109. IEEE, 2023.

[89] 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. *arXiv preprint arXiv:2311.16103*, 2023.

[90] OpenAI. Hello gpt-4, 2024.

[91] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022.

[92] Artemis Panagopoulou, Le Xue, Ning Yu, Junnan Li, Dongxu Li, Shafiq Joty, Ran Xu, Silvio Savarese, Caiming Xiong, and Juan Carlos Niebles. X-instructblip: A framework for aligning x-modal instruction-aware representations to llms and emergent cross-modal reasoning. *arXiv preprint arXiv:2311.18799*, 2023.

[93] Richard Yuanzhe Pang, Weizhe Yuan, Kyunghyun Cho, He He, Sainbayar Sukhbaatar, and Jason Weston. Iterative reasoning preference optimization. *arXiv preprint arXiv:2404.19733*, 2024.

[94] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[95] Xiangyu Peng, Congying Xia, Xinyi Yang, Caiming Xiong, Chien-Sheng Wu, and Chen Xing. Regenesis: Llms can grow into reasoning generalists via self-improvement. *arXiv preprint arXiv:2410.02108*, 2024.

[96] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023.

[97] Shraman Pramanick, Guangxing Han, Rui Hou, Sayan Nag, Ser-Nam Lim, Nicolas Ballas, Qifan Wang, Rama Chellappa, and Amjad Almahairi. Jack of all tasks, master of many: Designing general-purpose coarse-to-fine vision-language model. *arXiv preprint arXiv:2312.12423*, 2023.

[98] Shraman Pramanick, Li Jing, Sayan Nag, Jiachen Zhu, Hardik J Shah, Yann LeCun, and Rama Chellappa. Volta: Vision-language transformer with weakly-supervised local-feature alignment. *Transactions on Machine Learning Research*, 2023.

[99] Shraman Pramanick, Yale Song, Sayan Nag, Kevin Qinghong Lin, Hardik Shah, Mike Zheng Shou, Rama Chellappa, and Pengchuan Zhang. Egovlpv2: Egocentric video-language pre-training with fusion in the backbone. In

Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 5285–5297, 2023.

[100] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[101] Keshav Ramji, Young-Suk Lee, Ramón Fernandez Astudillo, Md Arafat Sultan, Tahira Naseem, Asim Munawar, Radu Florian, and Salim Roukos. Self-refinement of language models from external proxy metrics feedback. *arXiv preprint arXiv:2403.00827*, 2024.

[102] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

[103] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. *arXiv preprint arXiv:2312.02051*, 2023.

[104] Fangxun Shu, Lei Zhang, Hao Jiang, and Cihang Xie. Audio-visual llm for video understanding. *arXiv preprint arXiv:2312.06720*, 2023.

[105] Andries Smit, Paul Duckworth, Nathan Grinsztajn, Thomas D Barrett, and Arnu Pretorius. Should we be going mad? a look at multi-agent debate strategies for llms. *arXiv preprint arXiv:2311.17371*, 2023.

[106] Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. Pandagpt: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355*, 2023.

[107] Vighnesh Subramaniam, Yilun Du, Joshua B Tenenbaum, Antonio Torralba, Shuang Li, and Igor Mordatch. Multiagent finetuning: Self improvement with diverse reasoning chains. *arXiv preprint arXiv:2501.05707*, 2025.

[108] Chuanneng Sun, Songjun Huang, and Dario Pompili. Llm-based multi-agent reinforcement learning: Current and future directions. *arXiv preprint arXiv:2405.11106*, 2024.

[109] Guangzhi Sun, Potsawee Manakul, Adian Liusie, Kunat Pipatanakul, Chao Zhang, Phil Woodland, and Mark Gales. Crosscheckgpt: Universal hallucination ranking for multimodal foundation models. *arXiv preprint arXiv:2405.13684*, 2024.

[110] Guangzhi Sun, Wenyi Yu, Changli Tang, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, Yuxuan Wang, and Chao Zhang. video-salmonn: Speech-enhanced audio-visual large language models. *arXiv preprint arXiv:2406.15704*, 2024.

[111] Guangzhi Sun, Wenyi Yu, Changli Tang, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, Yuxuan Wang, and Chao Zhang. video-salmonn: Speech-enhanced audio-visual large language models. *arXiv preprint arXiv:2406.15704*, 2024.

[112] Zhongxiang Sun, Zihua Si, Xiaoxue Zang, Kai Zheng, Yang Song, Xiao Zhang, and Jun Xu. Large language models enhanced collaborative filtering. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 2178–2188, 2024.

[113] Zhiqing Sun, Longhui Yu, Yikang Shen, Weiyang Liu, Yiming Yang, Sean Welleck, and Chuang Gan. Easy-to-hard generalization: Scalable alignment beyond human supervision. *arXiv preprint arXiv:2403.09472*, 2024.

[114] Kyle Swanson, Wesley Wu, Nash L Bulaong, John E Pak, and James Zou. The virtual lab: Ai agents design new sars-cov-2 nanobodies with experimental validation. *bioRxiv*, pages 2024–11, 2024.

[115] Changli Tang, Yixuan Li, Yudong Yang, Jimin Zhuang, Guangzhi Sun, Wei Li, Zujun Ma, and Chao Zhang. Enhancing multimodal llm for detailed and accurate video captioning using multi-round preference optimization. *arXiv preprint arXiv:2410.06682*, 2024.

[116] Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. Extending large language models for speech and audio captioning. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11236–11240. IEEE, 2024.

[117] Yunlong Tang, Daiki Shimada, Jing Bi, and Chenliang Xu. Avicuna: Audio-visual llm with interleaver and context-boundary alignment for temporal referential dialogue. *arXiv e-prints*, pages arXiv–2403, 2024.

[118] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[119] Reka Team, Aitor Ormazabal, Che Zheng, Cyprien de Masson d'Autume, Dani Yogatama, Deyu Fu, Donovan Ong, Eric Chen, Eugenie Lamprecht, Hai Pham, et al. Reka core, flash, and edge: A series of powerful multimodal language models. *arXiv preprint arXiv:2404.12387*, 2024.

[120] Omkar Thawakar, Dinura Dissanayake, Ketan More, Ritesh Thawkar, Ahmed Heakl, Noor Ahsan, Yuhao Li, Mohammed Zumri, Jean Lahoud, Rao Muhammad Anwer, Hisham Cholakkal, Ivan Laptev, Mubarak Shah, Fahad Shahbaz Khan, and Salman Khan. Llamav-o1: Rethinking step-by-step visual reasoning in llms, 2025.

[121] Omkar Thawakar, Dinura Dissanayake, Ketan More, Ritesh Thawkar, Ahmed Heakl, Noor Ahsan, Yuhao Li, Mohammed Zumri, Jean Lahoud, Rao Muhammad Anwer, et al. Llamav-o1: Rethinking step-by-step visual reasoning in llms. *arXiv preprint arXiv:2501.06186*, 2025.

[122] Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O'Sullivan, and Hoang D Nguyen. Multi-agent collaboration mechanisms: A survey of llms. *arXiv preprint arXiv:2501.06322*, 2025.

[123] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.

[124] Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. Mixture-of-agents enhances large language model capabilities. *arXiv preprint arXiv:2406.04692*, 2024.

[125] Peiyi Wang, Lei Li, Zhihong Shao, RX Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd:

Verify and reinforce llms step-by-step without human annotations. *arXiv preprint arXiv:2312.08935*, 2023.

[126] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[127] Qineng Wang, Zihao Wang, Ying Su, Hanghang Tong, and Yangqiu Song. Rethinking the bounds of llm reasoning: Are multi-agent discussions the key? *arXiv preprint arXiv:2402.18272*, 2024.

[128] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023.

[129] Xiyao Wang, Yuhang Zhou, Xiaoyu Liu, Hongjin Lu, Yuancheng Xu, Feihong He, Jaehong Yoon, Taixi Lu, Gedas Bertasius, Mohit Bansal, et al. Mementos: A comprehensive benchmark for multimodal large language model reasoning over image sequences. *arXiv preprint arXiv:2401.10529*, 2024.

[130] Yan Wang, Yawen Zeng, Jingsheng Zheng, Xiaofen Xing, Jin Xu, and Xiangmin Xu. Videocot: A video chain-of-thought dataset with active annotation tool. *arXiv preprint arXiv:2407.05355*, 2024.

[131] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[132] Sean Welleck, Ximing Lu, Peter West, Faeze Brahman, Tianxiao Shen, Daniel Khashabi, and Yejin Choi. Generating sequences by learning to self-correct. *arXiv preprint arXiv:2211.00053*, 2022.

[133] Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. Wav2clip: Learning robust audio representations from clip. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4563–4567. IEEE, 2022.

[134] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. In *Forty-first International Conference on Machine Learning*, 2024.

[135] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786, 2021.

[136] Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts. *arXiv preprint arXiv:2407.04973*, 2024.

[137] Binzhu Xie, Sicheng Zhang, Zitang Zhou, Bo Li, Yuanhan Zhang, Jack Hessel, Jingkang Yang, and Ziwei Liu. Funqa: Towards surprising video comprehension. In *European Conference on Computer Vision*, pages 39–57. Springer, 2024.

[138] Guowei Xu, Peng Jin, Li Hao, Yibing Song, Lichao Sun, and Li Yuan. Llava-o1: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*, 2024.

[139] An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, et al. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024.

[140] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Faisal Ahmed, Zicheng Liu, Yumao Lu, and Lijuan Wang. Unitab: Unifying text and box outputs for grounded vision-language modeling. In *European Conference on Computer Vision*, pages 521–539. Springer, 2022.

[141] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.

[142] Qilang Ye, Zitong Yu, Rui Shao, Xinyu Xie, Philip Torr, and Xiaochun Cao. Cat: enhancing multimodal large language model to answer questions in dynamic audio-visual scenarios. *arXiv preprint arXiv:2403.04640*, 2024.

[143] Huaiyuan Ying, Shuo Zhang, Linyang Li, Zhejian Zhou, Yunfan Shao, Zhaoye Fei, Yichuan Ma, Jiawei Hong, Kuikun Liu, Ziyi Wang, et al. Internlm-math: Open math large language models toward verifiable reasoning. *arXiv preprint arXiv:2402.06332*, 2024.

[144] Fei Yu, Anningzhe Gao, and Benyou Wang. Ovm, outcome-supervised value models for planning in mathematical reasoning. *arXiv preprint arXiv:2311.09724*, 2023.

[145] Shoubin Yu, Jaehong Yoon, and Mohit Bansal. Crema: Generalizable and efficient video-language reasoning via multimodal modular fusion. *arXiv preprint arXiv:2402.05889*, 2024.

[146] Xiao Yu, Baolin Peng, Michel Galley, Jianfeng Gao, and Zhou Yu. Teaching language models to self-improve through interactive demonstrations. *arXiv preprint arXiv:2310.13522*, 2023.

[147] Yangyang Yu, Zhiyuan Yao, Haohang Li, Zhiyang Deng, Yuechen Jiang, Yupeng Cao, Zhi Chen, Jordan Suchow, Zhenyu Cui, Rong Liu, et al. Fincon: A synthesized llm multi-agent system with conceptual verbal reinforcement for enhanced financial decision making. *Advances in Neural Information Processing Systems*, 37:137010–137045, 2025.

[148] Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*, 2024.

[149] Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488, 2022.

[150] Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. Merlot reserve: Neural script knowledge through vision and language and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16375–16387, 2022.

[151] Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang,

Linyang Li, et al. Anygpt: Unified multimodal llm with discrete sequence modeling. *arXiv preprint arXiv:2402.12226*, 2024.

[152] Di Zhang, Jianbo Wu, Jingdi Lei, Tong Che, Jiatong Li, Tong Xie, Xiaoshui Huang, Shufei Zhang, Marco Pavone, Yuqiang Li, et al. Llama-berry: Pairwise optimization for o1-like olympiad-level mathematical reasoning. *arXiv preprint arXiv:2410.02884*, 2024.

[153] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023.

[154] Lunjun Zhang, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. Generative verifiers: Reward modeling as next-token prediction. *arXiv preprint arXiv:2408.15240*, 2024.

[155] Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Kai Chen, and Ping Luo. Gpt4roi: Instruction tuning large language model on region-of-interest. *arXiv preprint arXiv:2307.03601*, 2023.

[156] Yunxiang Zhang, Muhammad Khalifa, Lajanugen Logeswaran, Jaekyeom Kim, Moontae Lee, Honglak Lee, and Lu Wang. Small language models need strong verifiers to self-correct reasoning. *arXiv preprint arXiv:2404.17140*, 2024.

[157] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024.

[158] Yuxiang Zhang, Shangxi Wu, Yuqi Yang, Jiangming Shu, Jinlin Xiao, Chao Kong, and Jitao Sang. o1-coder: an o1 replication for coding. *arXiv preprint arXiv:2412.00154*, 2024.

[159] Yang Zhao, Zhijie Lin, Daquan Zhou, Zilong Huang, Jiashi Feng, and Bingyi Kang. Bubogpt: Enabling visual grounding in multi-modal llms. *arXiv preprint arXiv:2307.08581*, 2023.

[160] Zijia Zhao, Longteng Guo, Tongtian Yue, Sihan Chen, Shuai Shao, Xinxin Zhu, Zehuan Yuan, and Jing Liu. Chatbridge: Bridging modalities with large language model as a language catalyst. *arXiv preprint arXiv:2305.16103*, 2023.

[161] Haojie Zheng, Tianyang Xu, Hanchi Sun, Shu Pu, Ruoxi Chen, and Lichao Sun. Thinking before looking: Improving multimodal llm reasoning via mitigating visual hallucination. *arXiv preprint arXiv:2411.12591*, 2024.

[162] Tianyang Zhong, Zhengliang Liu, Yi Pan, Yutong Zhang, Yifan Zhou, Shizhe Liang, Zihao Wu, Yanjun Lyu, Peng Shu, Xiaowei Yu, et al. Evaluation of openai o1: Opportunities and challenges of agi. *arXiv preprint arXiv:2409.18486*, 2024.

[163] Jinxing Zhou, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, and Yiran Zhong. Audio–visual segmentation. In *European Conference on Computer Vision*, pages 386–403. Springer, 2022.

[164] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

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

Supplementary Material

We add the following details in this supplementary:

A. Supplementary Video

In our supplementary video, we provide several audio-visual examples for each task and compare the performance of different models before and after introducing the reasoning steps.

B. More related Works

Multi-Agent Systems with LLMs. Recent advancements in multi-agent systems [27, 45, 50, 68, 105, 108, 127] underscore the potential of large language models in tackling complex tasks. While some approaches [31] facilitate answer-sharing among agents for enhanced collaboration, Mixture-of-Agents [124] employs a hierarchical architecture where agents iteratively refine responses. Comm [9] proposed problem-solving through structured communication and role division while Multi-Persona [69] promotes varied agent behaviours by assigning unique personas. ChatEval [4] investigates various multi-agent debate strategies for effective interaction and response optimization while DMAS [13] examines token-efficient multi-agent planning frameworks to enhance coordination and task performance. Building on advancements in multi-agent systems, recent research has investigated fine-tuning independently specialized agents that collaborate to produce diverse reasoning chains [107]. In contrast to these approaches, our method emphasizes collaborative optimization via a shared experience library, allowing agents to collectively learn from and refine effective reasoning trajectories.

Self-improvement. Self-improving models [53, 95, 132, 146, 148, 156] have gained significant attention due to their potential to enhance reasoning abilities through iterative feedback and refinement. Various studies [58, 67, 93, 149] utilize bootstrapping methods by leveraging self-generated rationales, while other works [16, 43, 101, 148] introduce self-refinement mechanisms via reinforcement learning.

Multi-modal Learning. Conventional multi-modal methods incorporating vision-language [8, 15, 21, 41, 54, 62, 63, 65, 73, 87, 98–100, 140], audio-visual [19, 20, 23, 39, 41, 163], audio-language [28, 33, 47, 133, 150] have developed over the recent years with a focus to solve a variety of coarse-grained (question-answering, captioning, retrieval, etc.) and fine-grained (detection, segmentation, phrase grounding, etc.) understanding as well as generation tasks. However, these traditional models do not typically solve reasoning based tasks (with the exception of NLVR). With the advent of multi-modal LLMs [6, 7, 17, 22, 26, 37, 49, 57, 74, 77, 81, 82, 85, 86, 90–92, 96, 97, 103, 111, 117, 128, 134, 141–143, 151, 155, 159, 160, 164], although some recent efforts have been made to leverage reasoning capabilities of LLMs to solve complex visual question answering tasks, multi-step reasoning with complex questions in the audio-visual space remains underexplored.

C. GPT based evaluation

C.1. Choice Extraction

Choice extraction strategy. We utilize a two-step choice extraction strategy, detailed next. While humans can easily extract choices from free-form predictions, rule-based matching may struggle with this task. To address this, we develop a universal evaluation strategy applicable to all AVLLMs, regardless of their varying instruction-following capabilities.
Step 1. *Prediction matching:* We first apply heuristic matching to extract choice labels (e.g., 'A', 'B', 'C', 'D') from AVLLM predictions. If successful, the extracted label is used as the final prediction. If heuristic matching fails, we employ GPT-4 to extract the choice label instead.
Step 2. *GPT-4 processing:* Prior benchmarks [75] validate GPT-4's effectiveness as a choice extractor. If step 1 fails, we input the question, choices, and model prediction into GPT-4, instructing it to align the prediction with one of the provided choices and return the corresponding label. If no match is found, GPT-4 outputs 'No match found.'

We also employ the best-of-N (3) evaluation strategy to ensure a rigorous evaluation and effectively demonstrate the performance gap across various models.

Response matching. To apply the matching algorithm to the options, we follow these rules: If an option is represented solely by a letter (e.g., ‘A‘) or formatted as ‘A) <response>‘, ‘A. <response>‘, ‘A, <response>‘, or ‘(A) <response>‘, without embedding other choices within ‘<response>‘, it is interpreted as a prediction of option ‘A‘.

Where does heuristic matching fail? The heuristic matching strategy usually fails in the following scenarios: (i) when the AVLLM is unable to provide an answer and requests clarification, such as ‘Apologies, can you please clarify ...‘ or similar phrases, and (ii) when the AVLLM responds with multiple option choices (A, B, C, etc.). In such cases, we proceed to Step 2, which involves GPT-4 based choice extraction. A sample prompt for GPT-4 is provided below.

Choice extraction prompt for GPT-4

Can you help me match an answer with a set of options for a single correct answer type question? I will provide you with a question, a set of options, and a response from an agent. You are required to map the agent’s response to the most similar option from the set. You should respond with a single uppercase character in ‘A’, ‘B’, ‘C’, ‘D’, and ‘E’ depending on the choice you feel is the most appropriate match. If there are no similar options you might output ‘No match found’. Please refrain from being subjective while matching and do not use any external knowledge. Below are some examples:
Example 1:
Question: What color is the man’s shirt who is sitting left of the object making this sound?
Options: A. Green B. Red C. Yellow D. Black
Answer: The person sitting next to the record player is wearing a black color shirt
Your output: D
Example 2:
Question: What does the audio-visual event constitute?
Options: A. A dog barking at a cat B. A dog barking on being hit by a stick C. The dog is hungry D. The dog is chasing another dog
Answer: It is a wolf
Your output: No match found

Change in template for GPT-4 evaluation. Next, to identify the model’s prediction, we utilize GPT-4, following the approach in MMBench [75]. We prompt GPT-4 with a template that includes the question, options, and the corresponding AVLLM prediction. Additionally, we incorporate task-specific options to help GPT-4 recognize the model’s predictions.

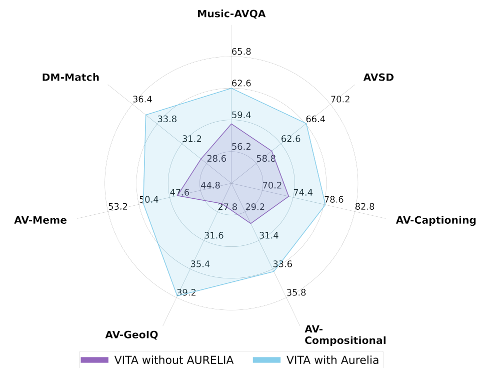


Figure 5. **Performance comparison across tasks.** The distillation of reasoning information in the VITA model via AURELIA enhances its performance across all the tasks.

C.2. Open-ended Answer Evaluation

To evaluate open-ended question answers with given ground truth answers using GPT, we design a prompt that instructs the model to assess the accuracy and relevance of the model’s answer in comparison to the ground truth. The prompt might be structured as: "Given the question, the model’s answer, and the ground truth answer, determine whether the model’s answer is correct or incorrect. If the model’s answer is factually accurate and appropriately aligns with the ground truth, even if expressed differently (e.g., ‘plane’ vs. ‘aeroplane’), output ‘Correct’. If the answer is incorrect or significantly deviates from the ground truth, output ‘Incorrect’." This ensures that GPT understands that synonymous or contextually equivalent terms (such as ‘plane’ for ‘aeroplane’) should be considered correct. Additionally, the evaluation will focus on factual accuracy and contextual alignment, and it will mark answers as ‘Correct’ if they are deemed effectively equivalent to the ground truth, despite minor wording differences.

D. Examples of Prompts

We use a combination of closed-source LLMs as specialized agents in AURELIA. To enable these LLM agents to interact with the input and with each other, we prompt them with appropriate instructions. We list these instruction prompts in Table 7.

E. Radar plot

The radar plot Fig. 5 illustrates the performance of the best performing open-source model VITA [37] on all 7 datasets before and after reasoning distillation is performed. We note that, upon ZS finetuning leveraging AURELIA the performance on each task is improved significantly with the maximum performance gain of 12.6% observed in the AV-captioning task. This underlines the efficacy of our pro-

Task	Instruct Prompt
Reasoning generation	Given the video and the audio and the question: `question` Task 1: generate detailed reasoning steps for solving the given question without revealing the answer. Task 2: provide detailed answers to each of these above reasoning steps generated in Task 1. Task 3: provide a final answer for the question. Your output should be in the form of a dictionary which looks like: `Task_1`: Task 1 answers, `Task_2`: Task 2 answers, `Task_3`: Task 3 answers.
Summarization	Given the reasoning steps, the answer to the reasoning steps, and the final response for the question, generate (come up with / guess) a detailed caption which is able to define the contents of the video and the audio. In the questions and the answers there may be things that might be outside the video and the audio context and needs world knowledge. You have to keep this in mind while generating the caption and you have to discard these information from the caption."
Evaluation	Given video and audio inputs, can you rate the following caption between 1 to 10 (1 being the lowest) based on its similarity with the corresponding inputs. Strictly output the numerical score only. Caption: `summary`
Feedback	The reasoning steps you previously generated: `{reasoning_steps}` to answer the question: `{question}` were evaluated and received a score of `{score}` out of 10. This score suggests that the reasoning steps may not be fully appropriate for answering the question correctly. Now, given the video, audio, and the question, carefully generate the correct reasoning steps to answer the question: `{question}` while strictly adhering to the following response format: Task 1: generate detailed reasoning steps for solving the given question without revealing the answer. Task 2: provide detailed answers to each of these above reasoning steps generated in Task 1. Task 3: provide a final answer for the question. Your output should be in the form of a dictionary which looks like: `Task_1`: Task 1 answers, `Task_2`: Task 2 answers, `Task_3`: Task 3 answers.

Table 7. **Details of Instruct Prompts.** Table presents the instruction prompts utilized by different agents in various stages of AURELIA.

posed reasoning data generation pipeline. AV-captioning often requires the model to draw intricate conclusions by critically analysing the audio-visual associations over multimodal temporal signals. A steady improvement in all the tasks underline the rich contextual understanding our reasoning augmented data can inject into a model.

F. Details on Reasoning Data Generation

To facilitate such reasoning generation, our framework, AURELIA, employs a multi-agent system that iteratively refines reasoning steps. A Reasoning Generator Agent first produces step-by-step deductions and explanations. The Summarization Agent then distills these steps into a structured caption without direct access to video or audio, ensuring reasoning quality is independent of raw inputs. A Multi-Modal Evaluator Agent assigns a similarity score based on how well the reasoning aligns with the original content, and a Feedback Agent iteratively refines the reasoning process to improve coherence and accuracy. Once the reasoning achieves an optimal evaluation score, it is integrated into the input before being fed into the target model. This explicit reasoning injection significantly enhances the model's ability to derive accurate, interpretable answers while minimizing errors and hallucinations. The process begins with the Reasoning Generator Agent, which analyzes the input set and produces step-by-step reasoning alongside an explanation for each step. Following this, the Summarization Agent interacts with the reasoning steps and generates a detailed caption crafted solely from the reasoning steps without any direct knowledge of the video or audio. This ensures that the caption's quality and accuracy are entirely dependent on the correctness of the generated reasoning. Next, a Multi-Modal Evaluator Agent assesses the alignment between the generated caption and the original video-audio content, assigning a similarity score between 1 and 10. A score of 1 indicates no alignment, while a 10 signifies perfect correspondence. Based on this evaluation, a Feedback Agent iteratively refines the reasoning steps by guiding the Reasoning Generator Agent to enhance its output by generating more coherent reasoning steps, aiming to maximize the evaluation score. This iterative loop continues until the reasoning quality surpasses a predefined threshold. Once the evaluation score pertaining to the reasoning steps reaches an optimal level, the reasoning information obtained at that step is integrated with the original audio, video, and question before being fed into the target model. By incorporating structured reasoning through distillation, AURELIA significantly improves the model's reasoning and overall performance.

G. AVReasonBench Statistics

G.1. Data Distribution

Tab. 8 reports different tasks along with various question categories associated with them. For example, QA pairs for AV-GeoIQ are collected from diverse categories of scenarios that require geographical and cultural knowledge combined with strong audio-visual reasoning. Similarly, samples for other tasks are also collected from diverse domains that span various categories. Fig. 6 reports data distribution for AV-GeoIQ and AV-Compositional understanding.

H. Breakdown results

In this section, we report the performance at a more granular level on AVReasonBench. We identify samples belonging to certain categories and consider only them for evaluation.

H.1. Performance on musical videos

We report the performance on musical videos category in Tab. 9. The samples under consideration require the AVLLMs to comprehend fine grained audio visual interactions followed by reasoning them with general knowledge/geo-cultural understanding. Experimental results demonstrate – best performance is achieved by VITA powered by its strong multimodal understanding. On an average, AV-compositional understanding task achieves most gains due to the reasoning supplement.

H.2. Performance on commonsense reasoning videos

Tab. 10 reports similar breakdown on commonsense reasoning examples. VITA outperforms other opensource models to achieve significantly improved performance upon treated with reasoning enhanced data generated by AURELIA. Highest performance gains are observed in AV-GeoIQ confirming the requirement of strong practical understanding of AV scenarios for this task.

I. Results on other benchmarks

We compare the performance of Video-SALMONN and Unified-IO-2 on VideoMME and report them in Tab. 11 and Tab. 12. As can be clearly seen, our synthetic reasoning data augmentation pipeline is generalizable to other benchmarks. Employing reasoning enhanced annotations generated by AURELIA boosts the performance in all the models. Instilling strong reasoning capabilities improves the average performance significantly.

J. Qualitative Results

Fig. 7 - Fig. 12 demonstrate several qualitative examples for each task. For AV-GeoIQ we design questions which require the model to reason at multiple levels and go through a series of derived steps to be able to come up with the correct response. As seen from these examples, injecting reasoning annotations into the AVLLMs significantly improves the performance in various audio-visual scenarios which require critical multimodal comprehension. Similar improvements can be observed for other tasks as well. AURELIA equips the models with a series of critical reasoning sequences which enables better decision making through step by step reasoning. Powered by reasoning annotated data significant improvements can be observed in AV-compositional understanding, AV-Meme understanding and AV-Dance matching tasks.

K. Key Observations

This section highlights key insights into the performance of AVLLMs when injected with reasoning data generated by AURELIA.

Open-ended evaluations. We observe that AVLLMs injected with the reasoning data generated by AURELIA, in addition to being effective on AV samples under close ended MCQ setting, are also effective in case of open-ended answers. The former evaluation has a predefined set of options out of which only one option is correct while latter is relatively harder to answer as it is not bounded by word vocabulary. We find that employing our reasoning augmented

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

Table 8. **Task Statistics.** Table shows detailed task statistics in AVReasonBench.

Models	**AV-QA**		**AV-Captioning**	**AV-Compositional**	**AV-GeoIQ**	**AV-Meme**	**DM-Match**
	Music-AVQA	**AVSD**					
Open-Source Models in ZS							
NExT-GPT	53.5	52.1	62.5	27.7	25.3	17.5	26.2
Unified-IO-2 XL	53.6	52.6	76.7	29.4	23.4	23.1	28.3
Bay-CAT	55.7	54.2	68.2	26.5	22.8	23.3	28.7
Video-SALMONN	57.6	58.8	73.4	25.5	23.0	23.0	24.5
VITA	59.2	62.3	74.6	27.4	26.6	46.4	28.8
Open-Source Models with AURELIA							
NExT-GPT	56.8	55.3	66.5	30.1	29.2	22.0	30.5
Unified-IO-2 XL	56.3	57.7	79.6	32.6	28.5	27.2	33.0
Bay-CAT	57.6	59.1	73.2	29.6	27.0	26.0	32.5
Video-SALMONN	61.8	62.6	76.8	29.1	28.6	28.0	29.0
VITA	61.4	65.3	78.3	32.5	30.7	49.2	33.9

Table 9. **Breakdown results on musical videos.** Performance comparison of various models before and after applying AURELIA.

data also improves the open-ended evaluation of existing AVLLMs.

Emphasis on one modality. It is observed that existing AVLLMs occasionally prioritizes one modality over the other, introducing biases in its decision-making process. Since AURELIA works on the synergy of AV input through the interaction of multiple agents, in such cases, our approach can mitigate the bias induced due to the model's focus on one modality by providing additional cues about the other modality through reasoning steps. However, we also notice occasionally (such as in Fig. 4 (left) of main paper), reasoning distillation becomes less effective in such extreme cases, as the model remains biased towards the dominant modality, neglecting the valuable information from the other. In this specific example, the AVLLM incorrectly assumes the dog is silent, even when audio information is present. We hypothesize that the error in such cases can propagate through the reasoning stages due to model being biased in initial step itself, ultimately resulting in a flawed conclusion.

Suboptimal Comprehension. AURELIA systematically distills the reasoning information in the AVLMMs to advance their AV comprehension capability. Leveraging strong multi-agent LLMs, AURELIA has an advanced comprehension of intricate AV relationships, which can help mitigate the weak reasoning comprehension in AVLLMs. Even though based

Models	AV-QA		AV-Captioning	AV-Compositional	AV-GeoIQ	AV-Meme	DM-Match
	Music-AVQA	AVSD					
Open-Source Models in ZS							
NExT-GPT	51.2	50.3	59.6	25.7	22.7	16.9	24.7
Unified-IO-2 XL	50.4	51.7	73.2	28.0	22.2	22.0	25.3
Bay-CAT	51.7	52.2	66.4	24.9	20.3	21.1	25.2
Video-SALMONN	53.7	52.2	70.1	22.7	21.3	20.2	21.9
VITA	55.7	59.7	71.2	24.0	22.3	43.5	26.5
Open-Source Models with AURELIA							
NExT-GPT	55.2	54.8	63.1	29.6	26.7	21.0	28.3
Unified-IO-2 XL	54.3	55.2	76.8	32.1	27.4	26.3	29.5
Bay-CAT	55.6	56.1	70.2	28.6	25.8	25.4	29.5
Video-SALMONN	58.8	57.6	74.8	27.1	26.6	25.0	26.2
VITA	60.4	64.7	74.7	29.5	27.7	48.1	31.2

Table 10. **Breakdown results on commonsense reasoning videos**. Table shows the performance comparison of various models before and after applying AURELIA specifically on commensense reasoning related videos.

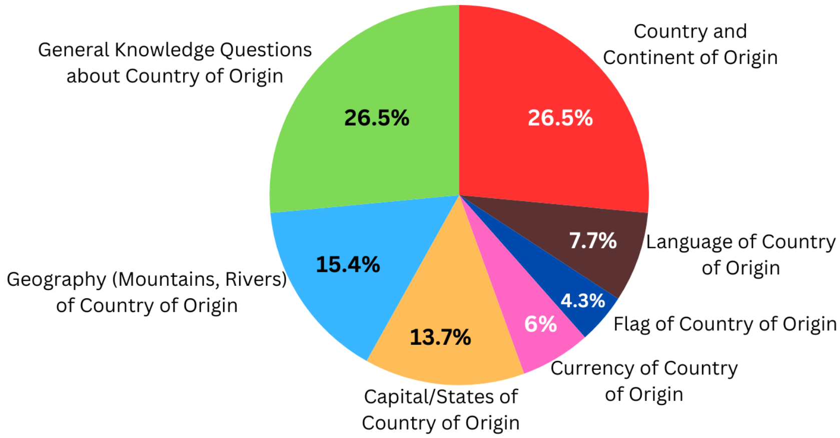
Subset	Modality	Category						
		Knowledge	*Film & Television*	*Sports Competition*	*Artistic Performance*	*Life Record*	*Multilingual*	Overall
Short	ZS	78.6	84.2	75.1	82.9	82.0	83.6	81.2
	+ AURELIA	82.1	88.3	78.4	85.7	85.2	86.4	**84.8**
Medium	ZS	77.3	80.7	69.0	80.6	72.6	96.8	78.5
	+ AURELIA	80.1	83.7	72.8	84.7	75.8	97.1	**82.7**
Long	ZS	78.6	70.8	69.4	60.3	63.0	80.9	70.2
	+ AURELIA	82.8	74.7	72.1	64.4	66.6	83.8	**73.7**
Overall	ZS	77.5	79.6	71.7	75.8	75.9	85.7	77.1
	+ AURELIA	80.5	82.7	74.9	78.4	78.7	89.0	**80.0**

Table 11. **Performance of Video SALMONN across Video-MME**. The evaluation is done on audio-visual inputs.

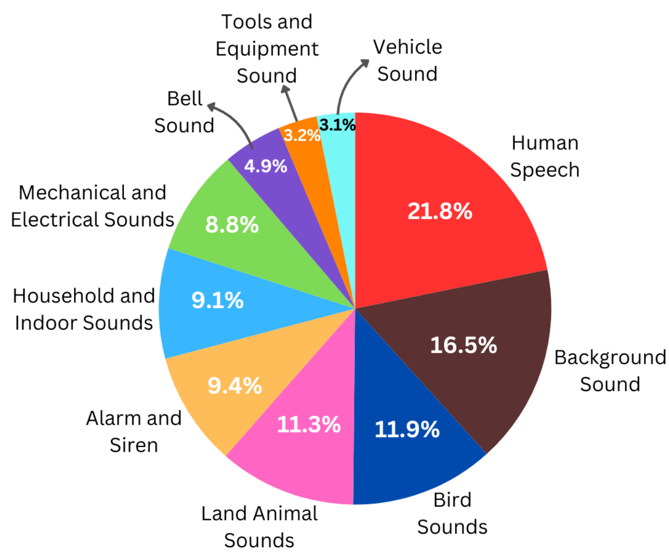
Subset	Modality	Category						
		Knowledge	*Film & Television*	*Sports Competition*	*Artistic Performance*	*Life Record*	*Multilingual*	Overall
Short	ZS	76.2	82.8	73.9	80.7	79.9	81.9	79.0
	+ AURELIA	78.9	84.1	74.9	82.9	81.0	83.1	**81.1**
Medium	ZS	75.6	78.9	67.9	78.1	70.7	94.4	76.4
	+ AURELIA	78.9	81.8	70.4	82.8	73.8	95.3	**80.5**
Long	ZS	76.7	68.8	67.3	58.5	61.9	78.0	68.2
	+ AURELIA	80.8	72.6	70.4	62.6	64.6	81.7	**71.6**
Overall	ZS	75.7	77.7	68.9	73.4	73.8	82.8	75.8
	+ AURELIA	79.5	80.4	72.6	76.5	76,8	87.4	**77.6**

Table 12. **Performance of Unified-IO-2 across Video-MME**. The evaluation is done on audio-visual inputs.

on strong closed-source LLMs, AURELIA can also incur errors sometimes in AV comprehension. Since AURELIA relies on a synergy of multi-modal agents, making any misunderstanding of audio-video input could be detrimental to the entire reasoning pipeline. Fig. 4 (right) of main paper illustrates such a case, where AURELIA struggles to grasp



(a) Distribution of AV-GeoIQ task.



(b) Distribution of AV-Compositional Understanding task.

Figure 6. **Distribution of AV-GeoIQ and AV-Compositional Understanding tasks.** **(a)** The pie chart shows the distribution of samples from our proposed AV-GeoIQ task. The collected samples exhibit diverse geographical and cultural characteristics. **(b)** The pie chart shows the distribution of samples from the AV-Compositional Understanding task. As seen from the pie chart, the data samples are collected from a diverse range of practical audio visual scenarios.

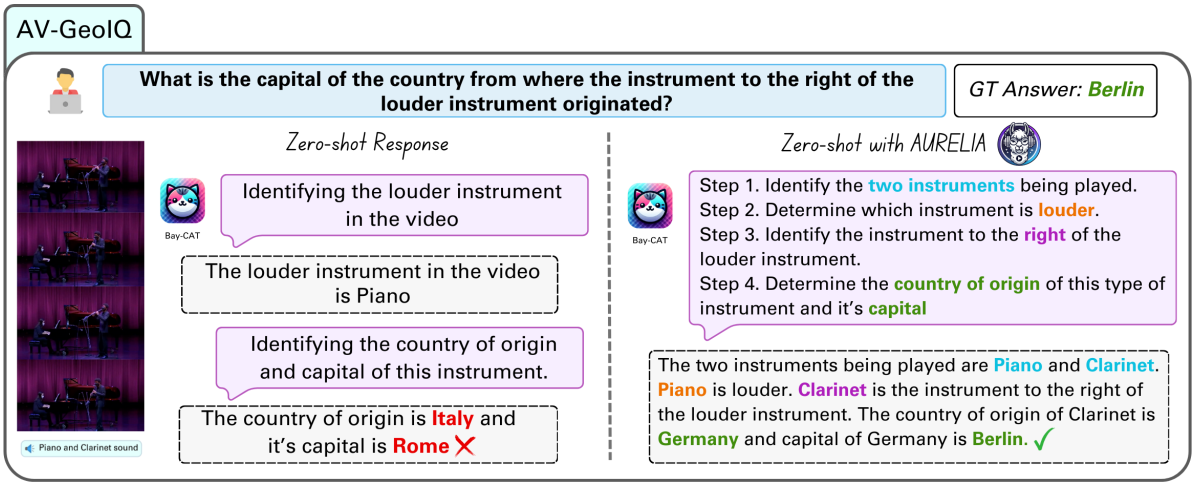
the interplay between video and audio.

L. Future Work

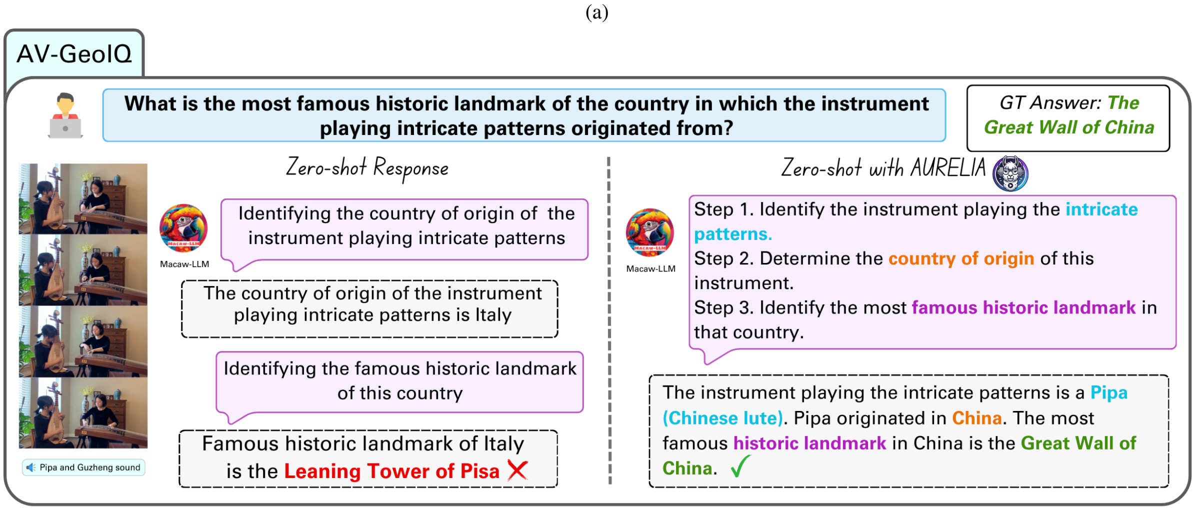
Currently, the multi-agent framework of AURELIA leverages a combination of closed-source LLMs as agents. A promising future direction would be to replace these proprietary models with open-source alternatives, enhancing accessibility and transparency. Additionally, another avenue for improvement lies in integrating reasoning directly into the training or instruction-tuning phase, rather than generating it dynamically at inference time. This would enable AVLLM to inherently develop step-by-step reasoning capabilities, allowing it to derive answers more naturally and effectively.

M. Societal Impact

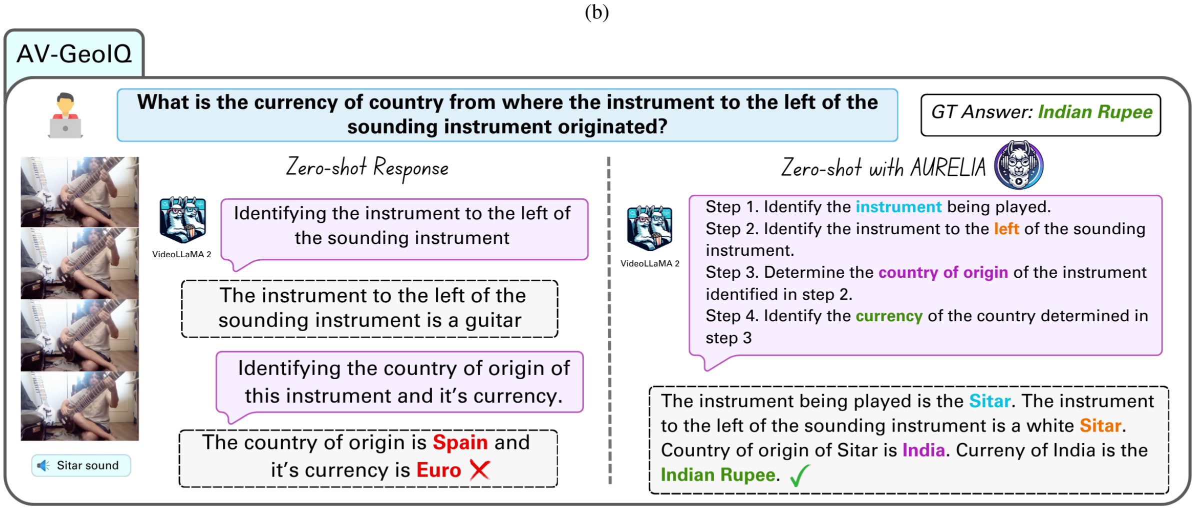
In this work, we perform an extensive analysis of reasoning capabilities of existing AVLLMs. Our study reveals that models lack sufficient audio-visual comprehension skills and most often fail to address scenarios that require commonsense reasoning. We believe our work can be useful to the community, and our findings can reveal the potential threats associated with deploying these models in real-time or accuracy-critical setups. We employ existing public datasets and in some cases, collect samples to curate the benchmark. We don't use any personal/human subject data without consent during data preparation and experiments.



(a)

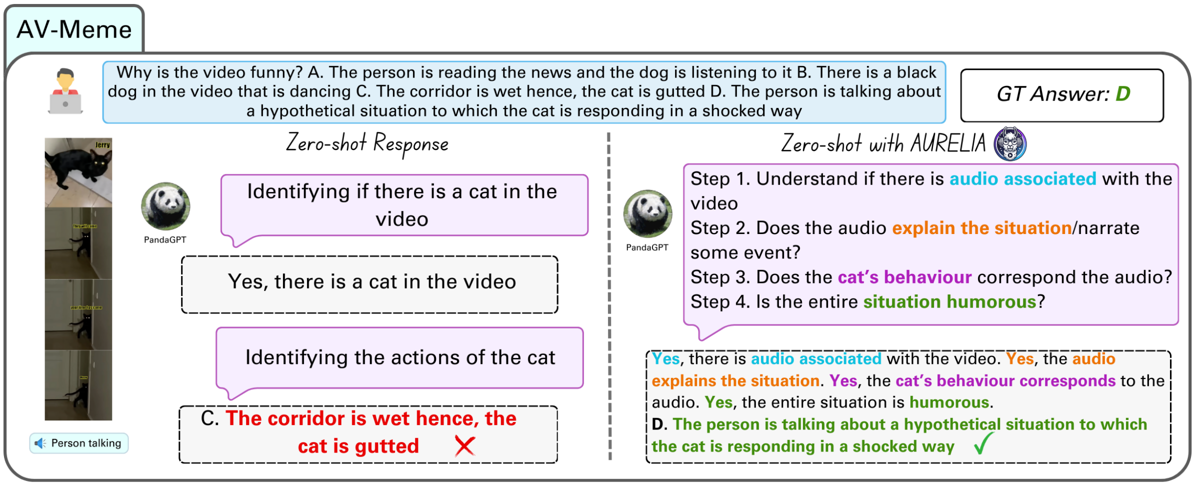


(b)

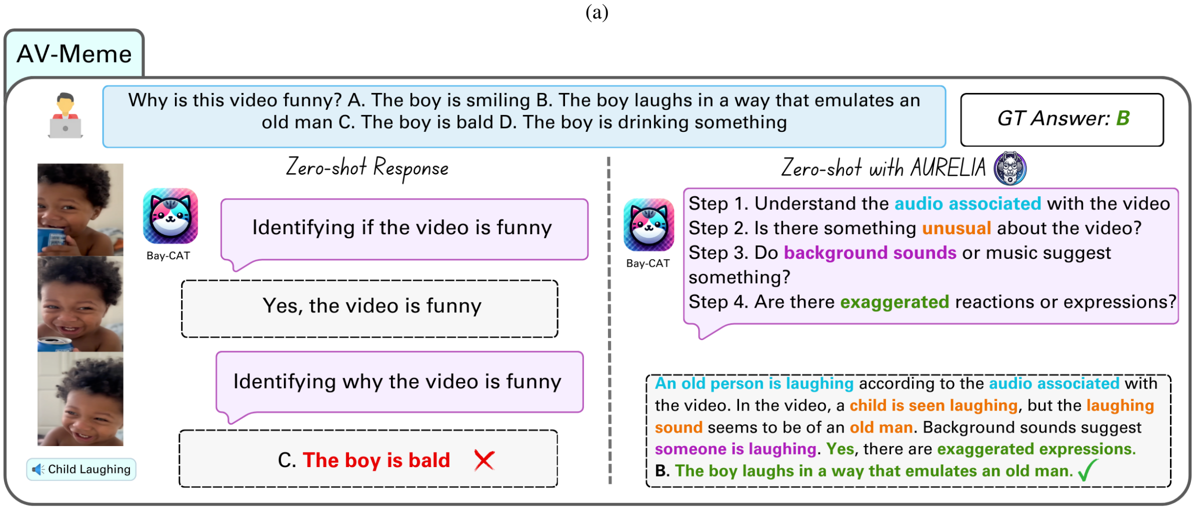


(c)

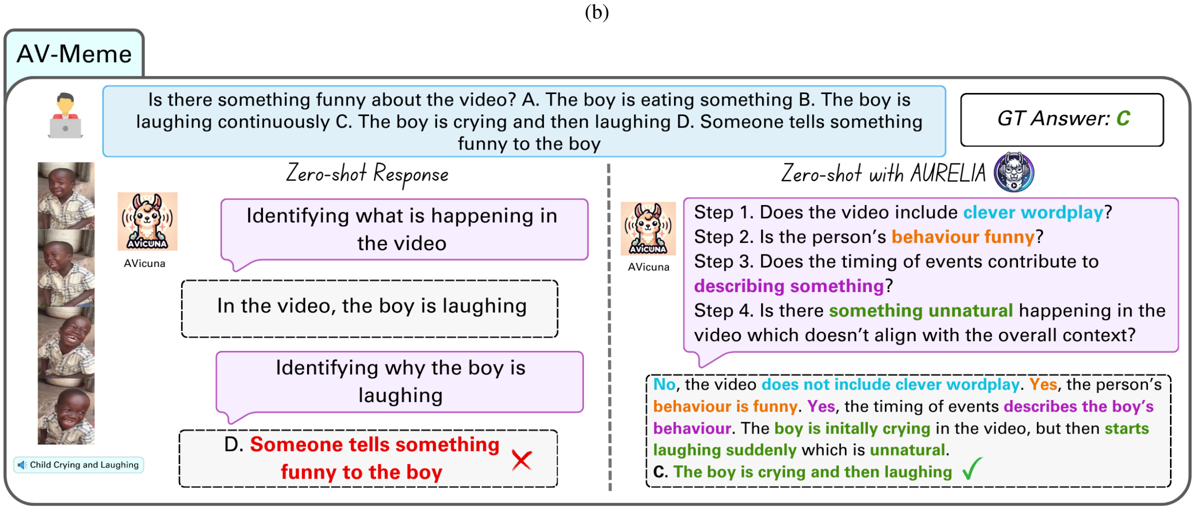
Figure 7. Qualitative visualization of AURELIA's reasoning distillation across AV-GeoIQ task.



(a)

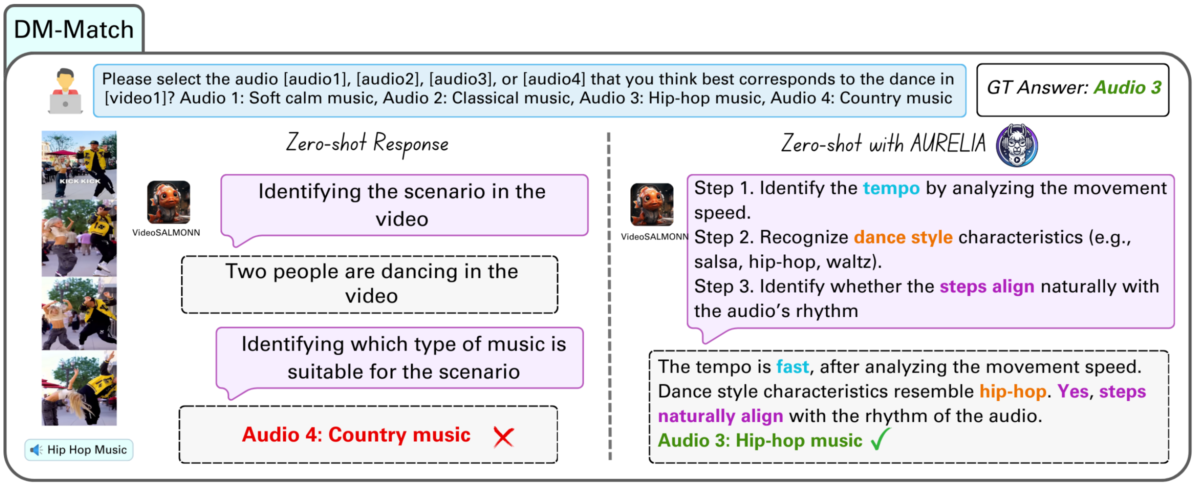


(b)

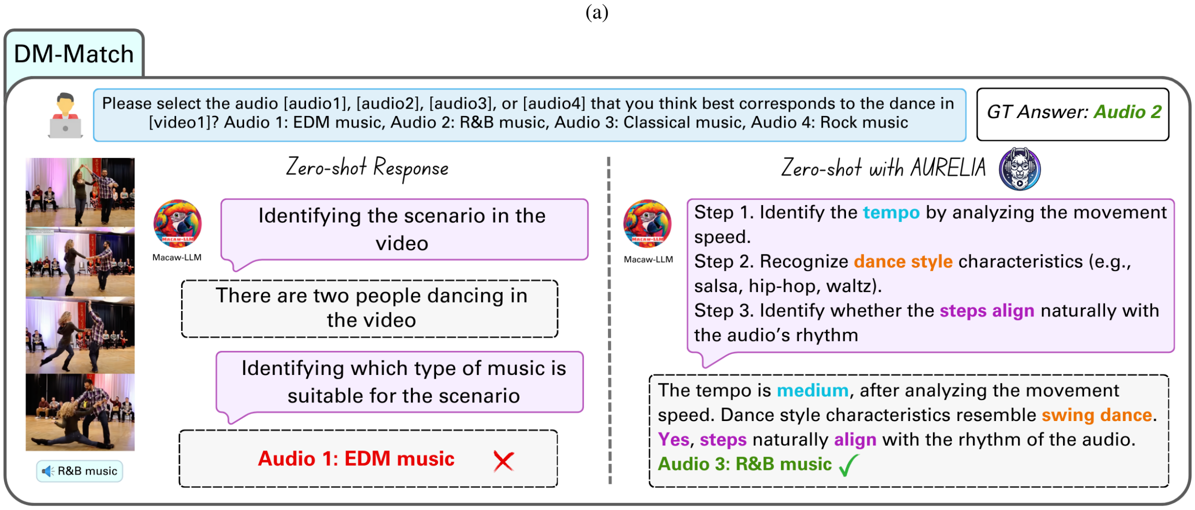


(c)

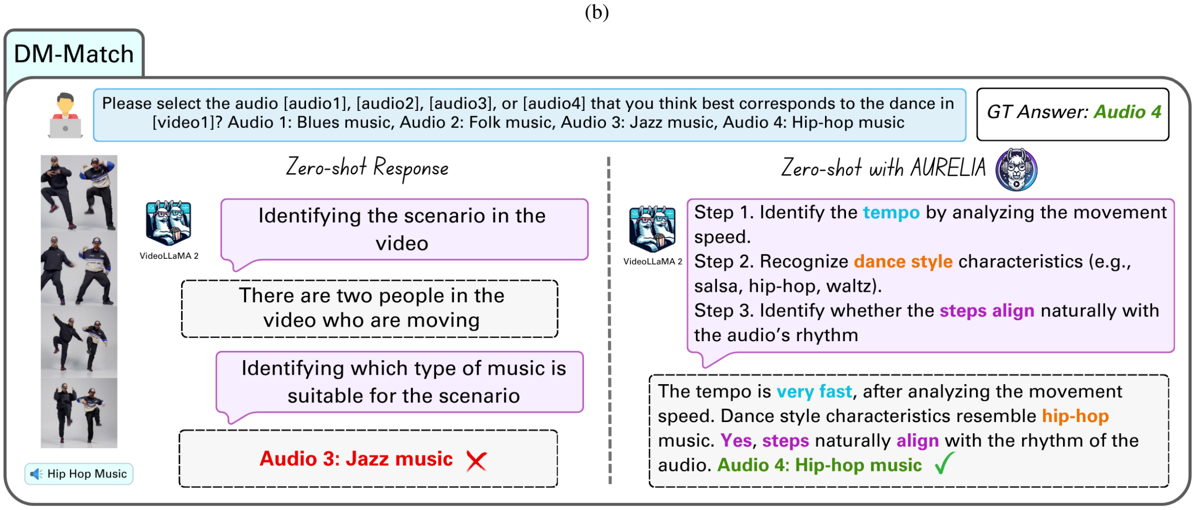
Figure 8. Qualitative visualization of AURELIA's reasoning distillation across AV-Meme task.



(a)

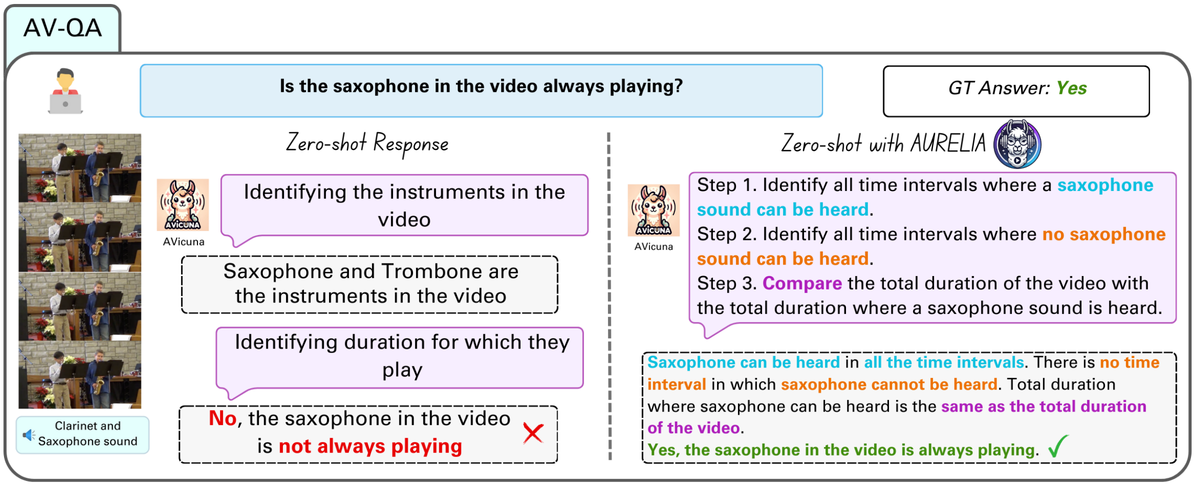


(b)

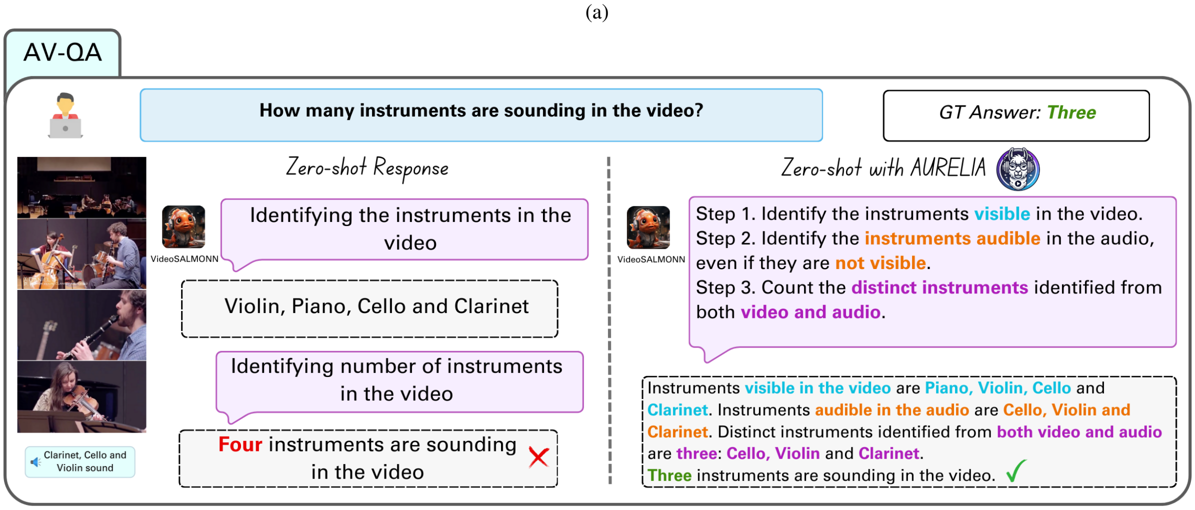


(c)

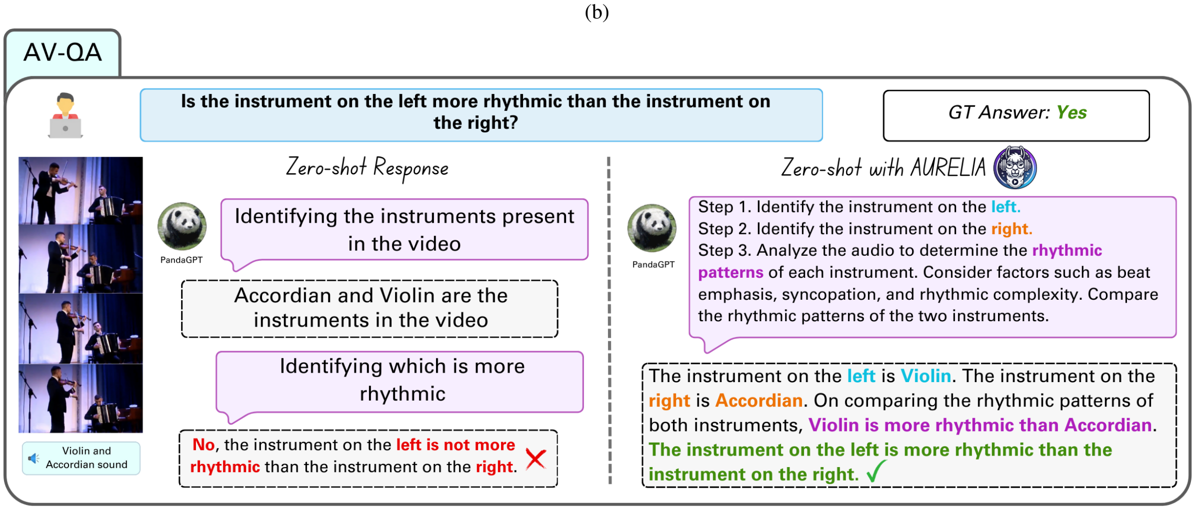
Figure 9. Qualitative visualization of AURELIA's reasoning distillation across DM-Match task.



(a)

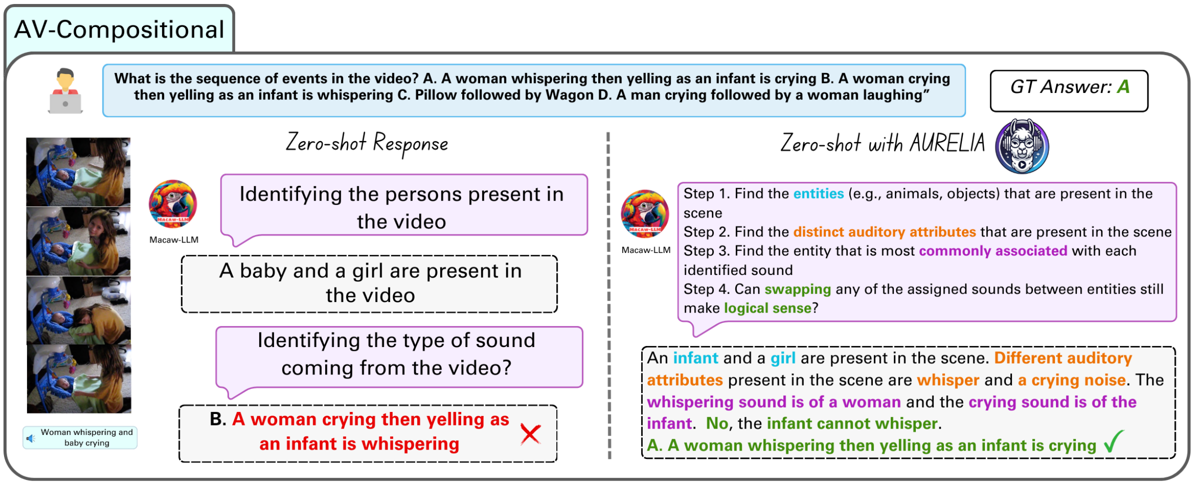


(b)

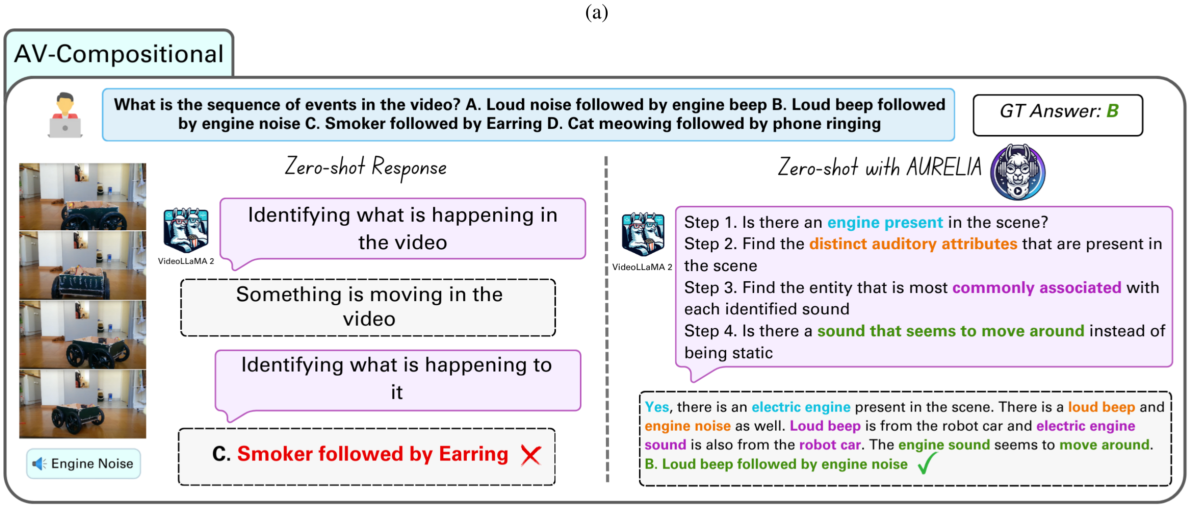


(c)

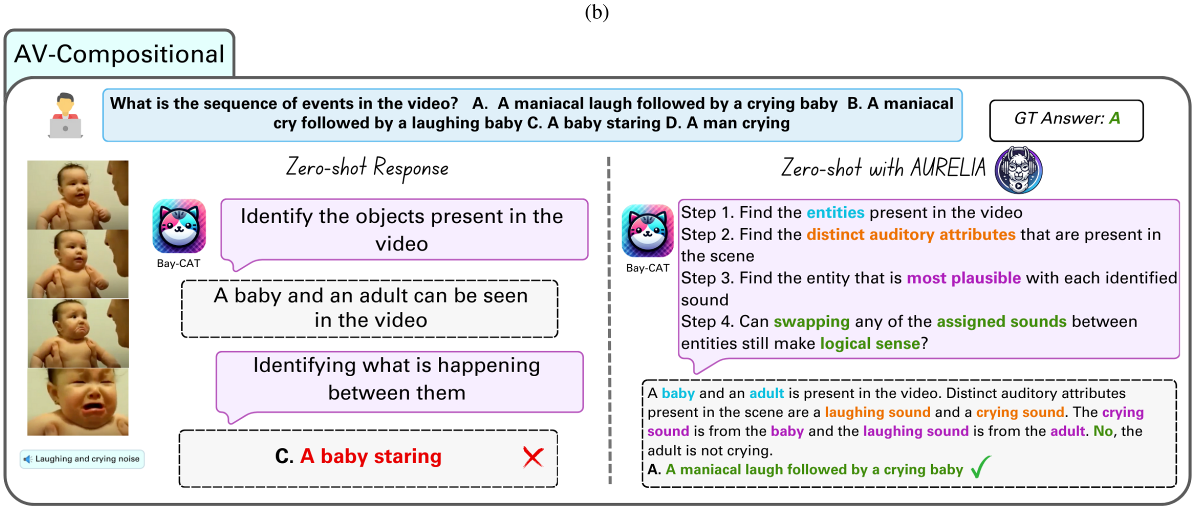
Figure 10. Qualitative visualization of AURELIA's reasoning distillation across AV-QA task.



(a)

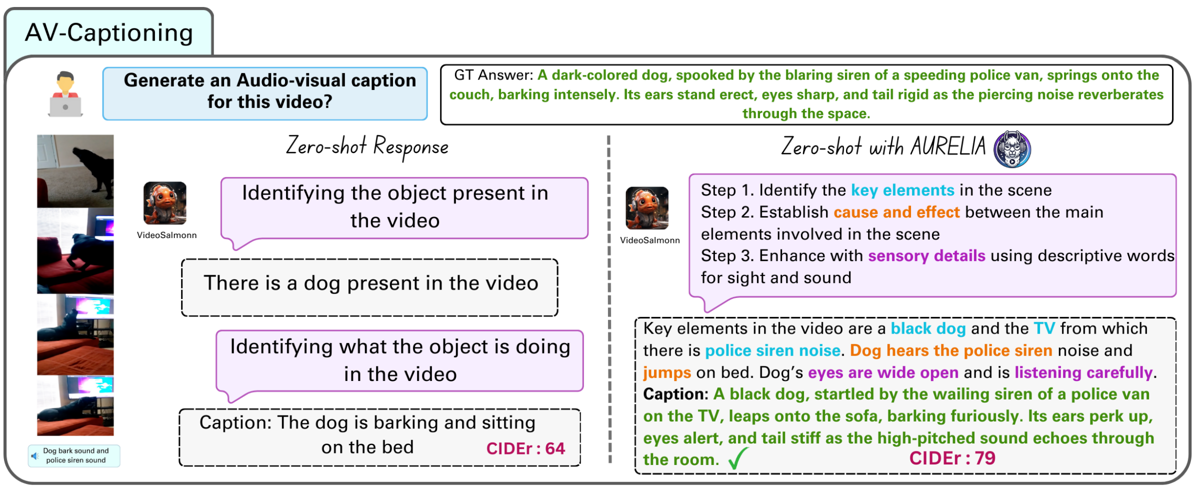


(b)

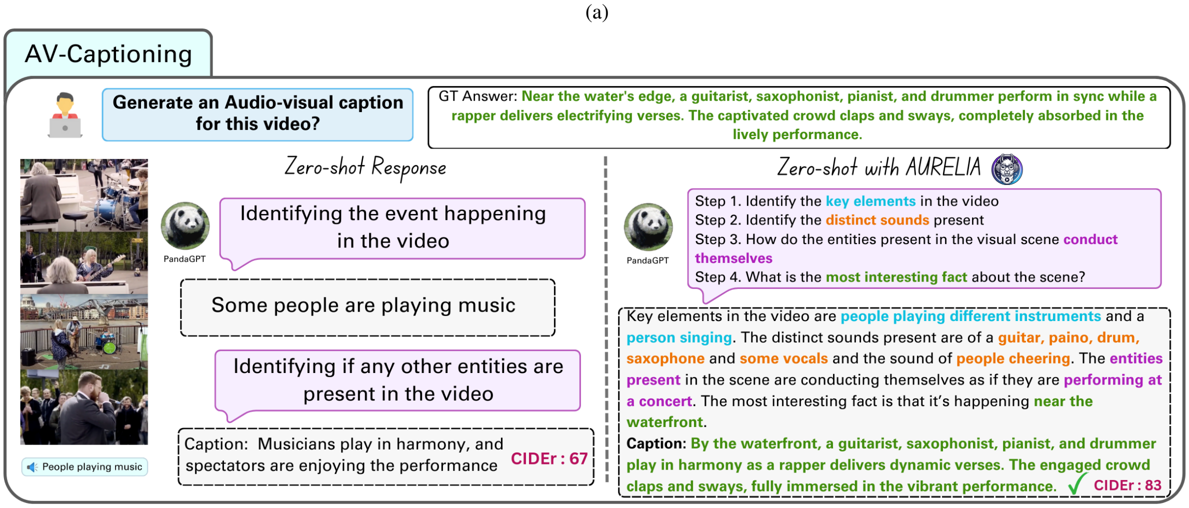


(c)

Figure 11. Qualitative visualization of AURELIA's reasoning distillation across AV-Compositional task.



(a)



(b)

Figure 12. Qualitative visualization of AURELIA's reasoning distillation across AV-Captioning task.