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

Supplementary Material

We add the following details in this supplementary:

1. 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.

2. More related Works

Multi-Agent Systems with LLMs. Recent advancements in multi-agent systems [16, 24, 27, 36, 59, 61, 65] underscore the potential of large language models in tackling complex tasks. While some approaches [18] facilitate answer-sharing among agents for enhanced collaboration, Mixture-of-Agents [64] employs a hierarchical architecture where agents iteratively refine responses. Comm [5] proposed problem-solving through structured communication and role division while Multi-Persona [37] promotes varied agent behaviours by assigning unique personas. ChatEval [1] investigates various multi-agent debate strategies for effective interaction and response optimization while DMAS [6] 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 [60]. 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 [28, 51, 67, 74, 75, 80] have gained significant attention due to their potential to enhance reasoning abilities through iterative feedback and refinement. Various studies [31, 35, 50, 76] utilize bootstrapping methods by leveraging self-generated rationales, while other works [8, 23, 57, 75] introduce self-refinement mechanisms via reinforcement learning.

Multi-modal Learning. Conventional multi-modal methods incorporating vision-language [4, 7, 12, 22, 29, 32–34, 38, 46, 54–56, 70], audio-visual [10, 11, 14, 21, 22, 83], audio-language [17, 19, 25, 68, 77] 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 [2, 3, 9, 13, 15, 20, 26, 30, 39, 41–45, 47–49, 52, 53, 58, 62, 63, 66, 69, 71–73, 78, 79, 81, 82, 84], 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.

3. GPT based evaluation

3.1. 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 [40] 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 [40]. 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.

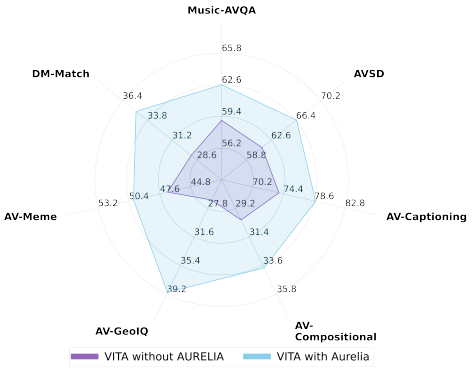


Figure 1. **Performance comparison across tasks.** The distillation of reasoning information in the VITA model via AURELIA enhances its performance across all the tasks.

3.2. 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.

4. 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 1.

5. Radar plot

The radar plot Fig. 1 illustrates the performance of the best performing open-source model VITA [20] 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 1. **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.

6. 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 de-

rive 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.

7. AVReasonBench Statistics

7.1. Data Distribution

Tab. 2 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. 2 reports data distribution for AV-GeoIQ and AV-Compositional understanding.

8. 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.

8.1. Performance on musical videos

We report the performance on musical videos category in Tab. 3. 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.

8.2. Performance on commonsense reasoning videos

Tab. 4 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.

9. Results on other benchmarks

We compare the performance of Video-SALMONN and Unified-IO-2 on VideoMME and report them in Tab. 5 and Tab. 6. 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.

10. Qualitative Results

Fig. 3 - Fig. 8 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.

11. 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 2. **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 3. **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 4. **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.

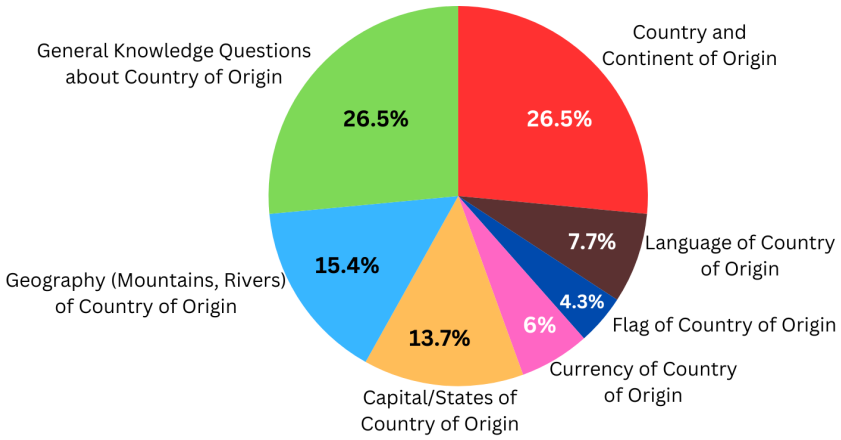
Subset	Modality	Category						
		Knowledge	*Film & Television*	*Sports Competition*	*Artistic Performance*	*Life Record*	*Multilingual*	Overall
Short	ZS	78.6	84.2	75.1	82.9	82.0	83.6	81.2
	+ AURELIA	82.1	88.3	78.4	85.7	85.2	86.4	**84.8**
Medium	ZS	77.3	80.7	69.0	80.6	72.6	96.8	78.5
	+ AURELIA	80.1	83.7	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 5. **Performance of Video SALMONN across Video-MME**. The evaluation is done on audio-visual inputs.

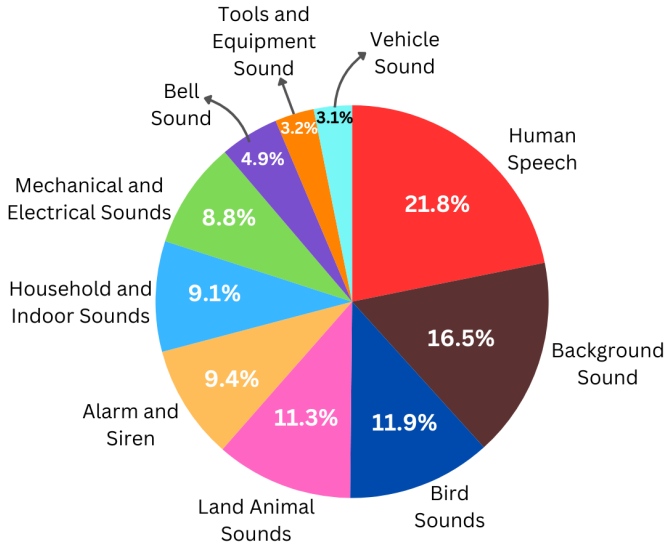
Subset	Modality	Category						
		Knowledge	*Film & Television*	*Sports Competition*	*Artistic Performance*	*Life Record*	*Multilingual*	Overall
Short	ZS	76.2	82.8	73.9	80.7	79.9	81.9	79.0
	+ AURELIA	78.9	84.1	74.9	82.9	81.0	83.1	**81.1**
Medium	ZS	75.6	78.9	67.9	78.1	70.7	94.4	76.4
	+ AURELIA	78.9	81.8	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 6. **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 2. **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.

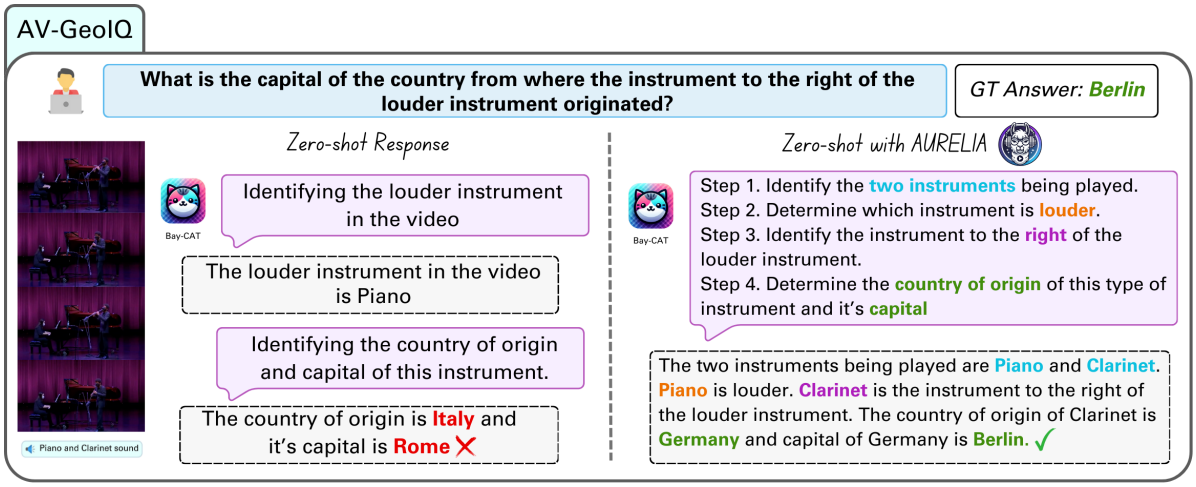
the interplay between video and audio.

12. Future Work

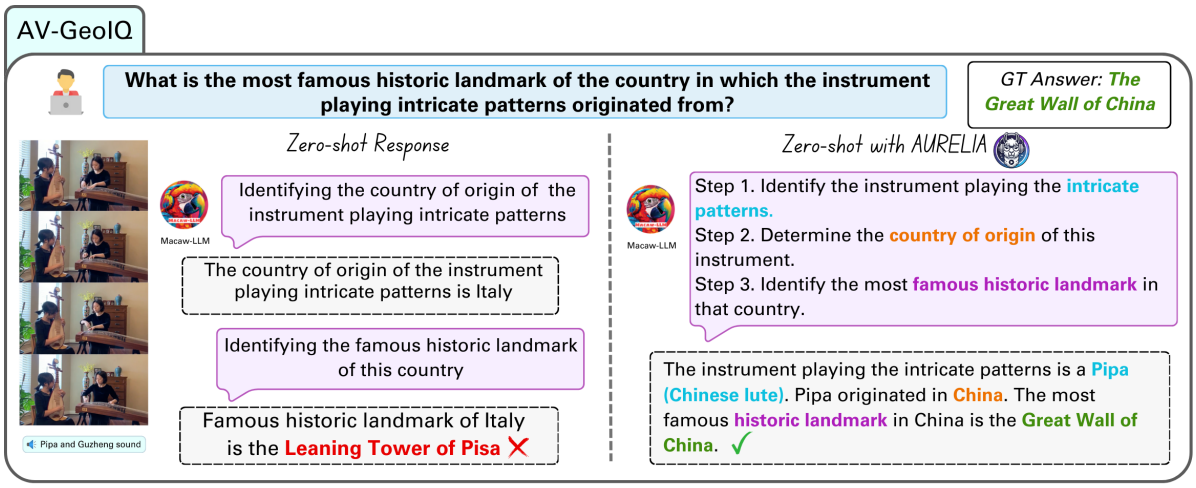
Currently, the multi-agent framework of AURELIA leverages a combination of closed-source LLMs as agents. A promising future direction would be to replace these proprietary models with open-source alternatives, enhancing accessibility and transparency. Additionally, another avenue for improvement lies 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.

13. Societal Impact

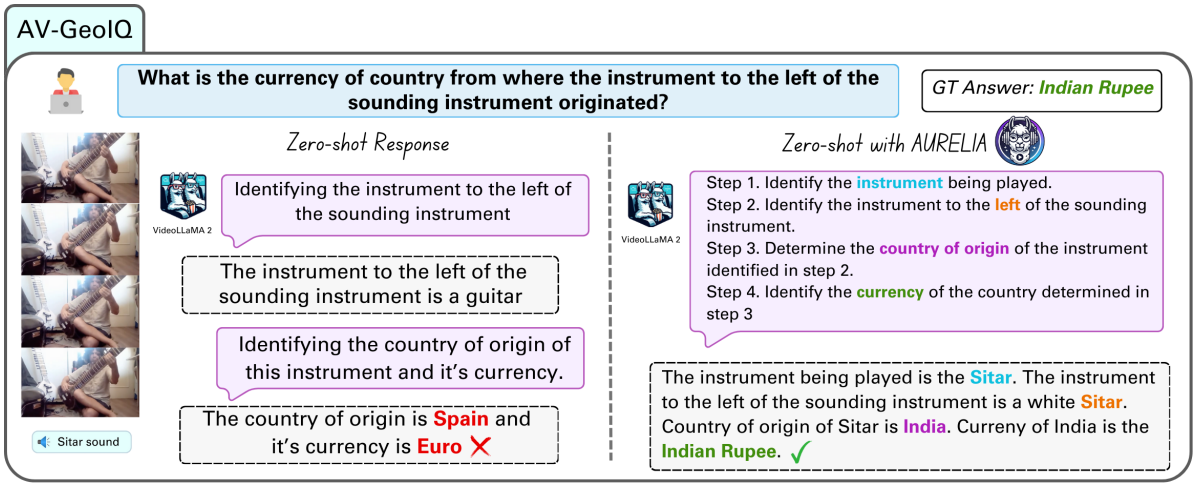
In this work, we perform an extensive analysis of reasoning capabilities of existing AVLLMs. Our study reveals that models lack sufficient audio-visual comprehension skills and most often fail to address scenarios that require commonsense reasoning. We believe our work can be useful to the community, 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)

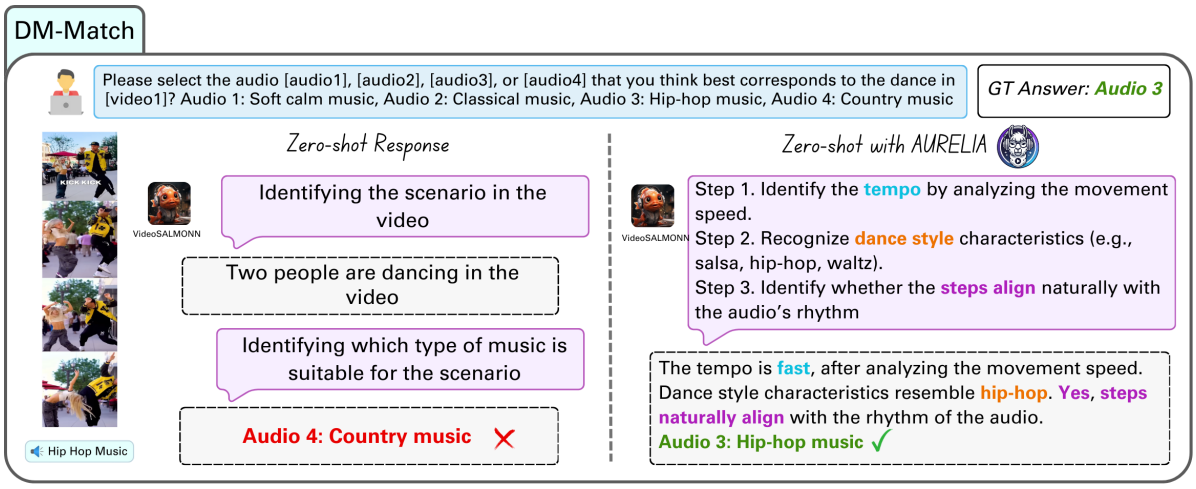
Figure 3. Qualitative visualization of AURELIA's reasoning distillation across AV-GeoIQ task.

(a)

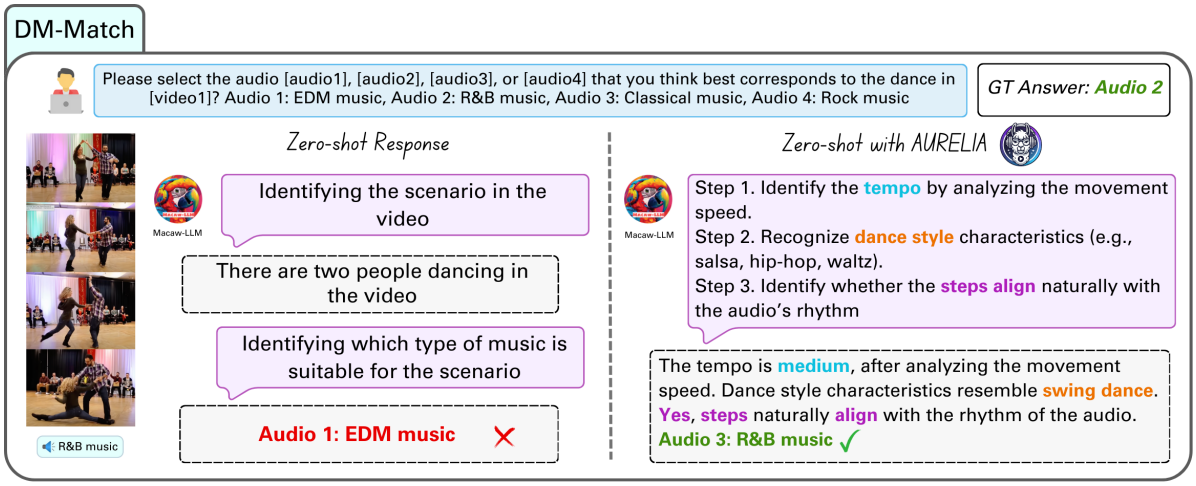
(b)

(c)

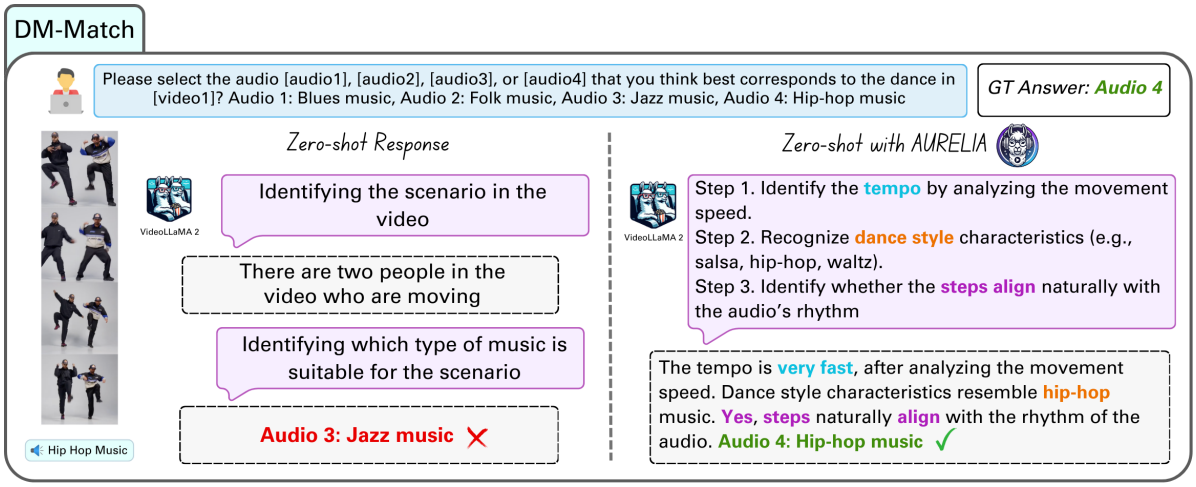
Figure 4. Qualitative visualization of AURELIA's reasoning distillation across AV-Meme task.



(a)

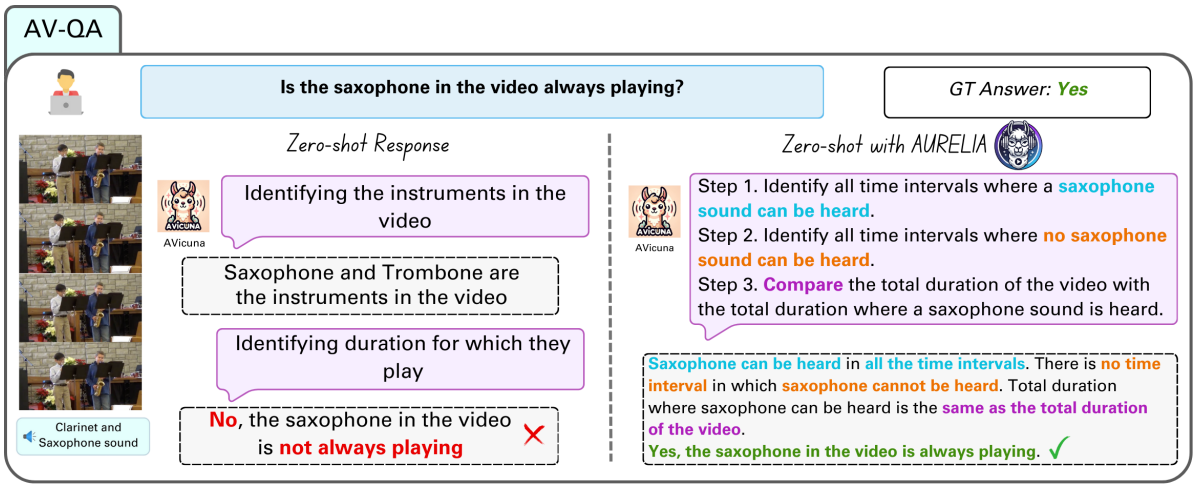


(b)

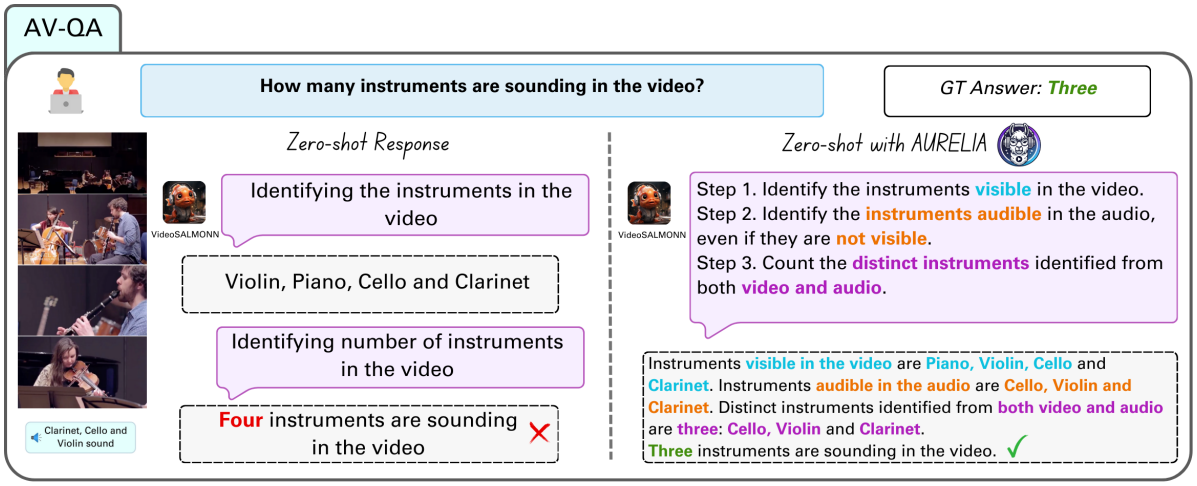


(c)

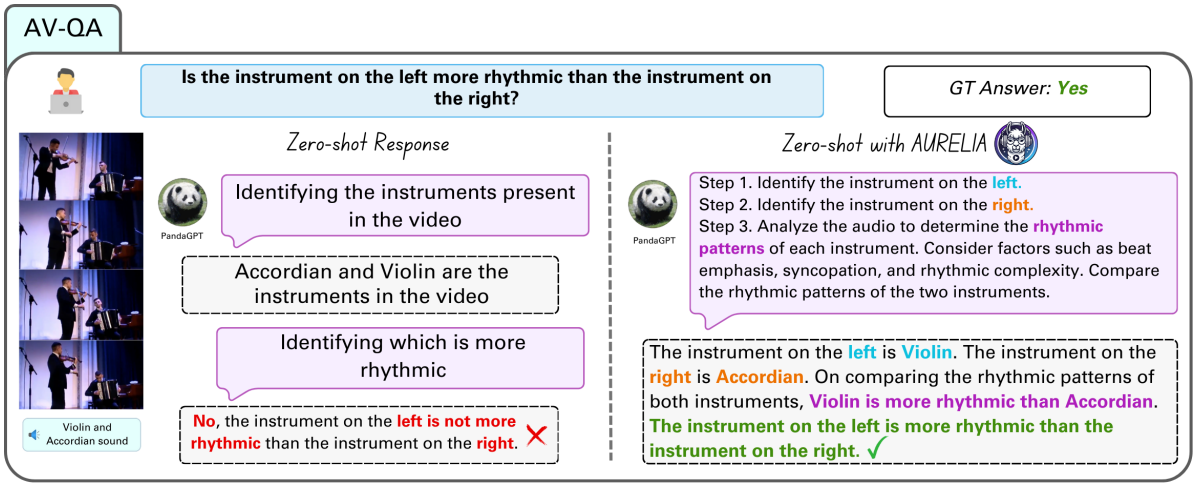
Figure 5. Qualitative visualization of AURELIA's reasoning distillation across DM-Match task.



(a)

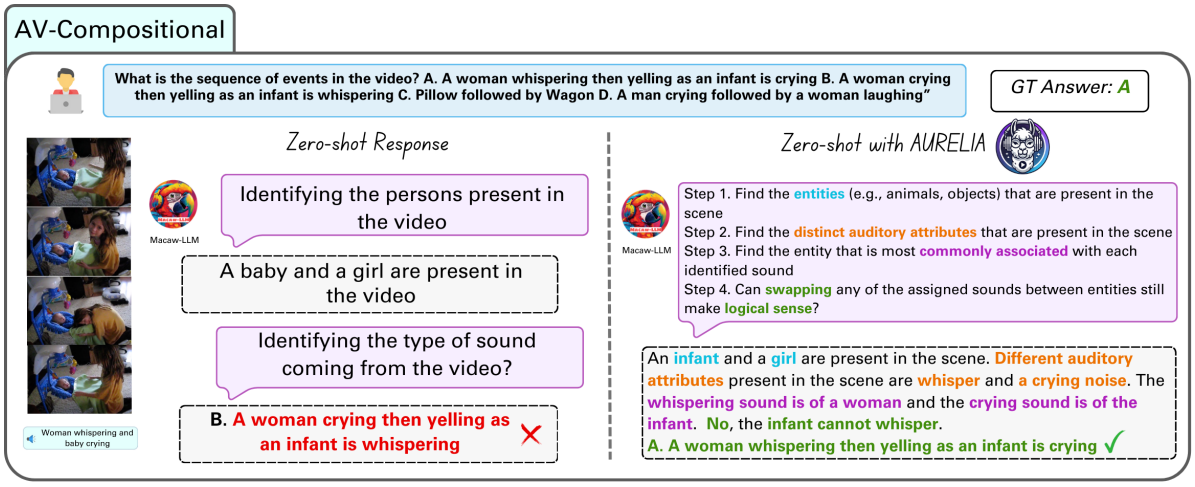


(b)

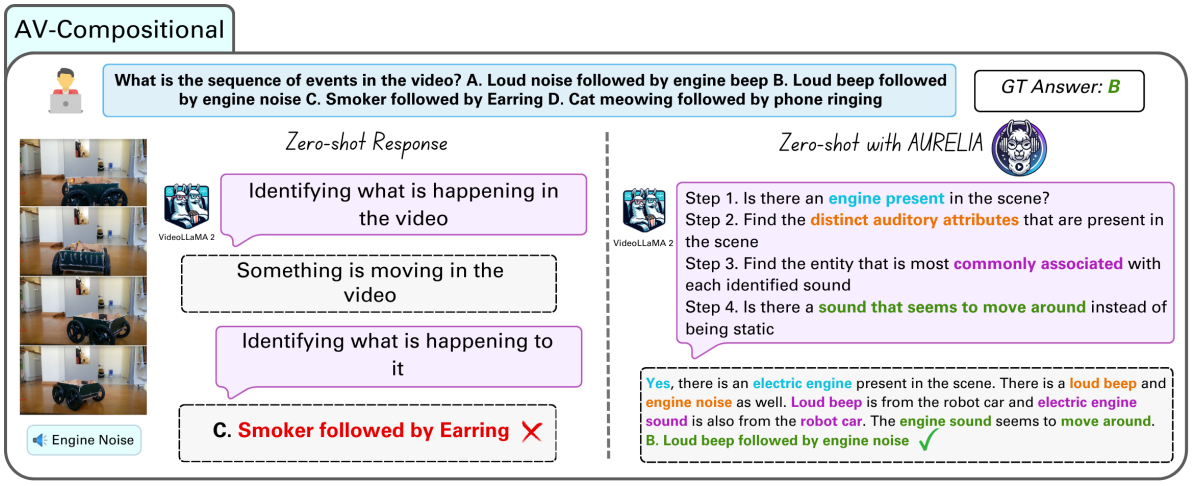


(c)

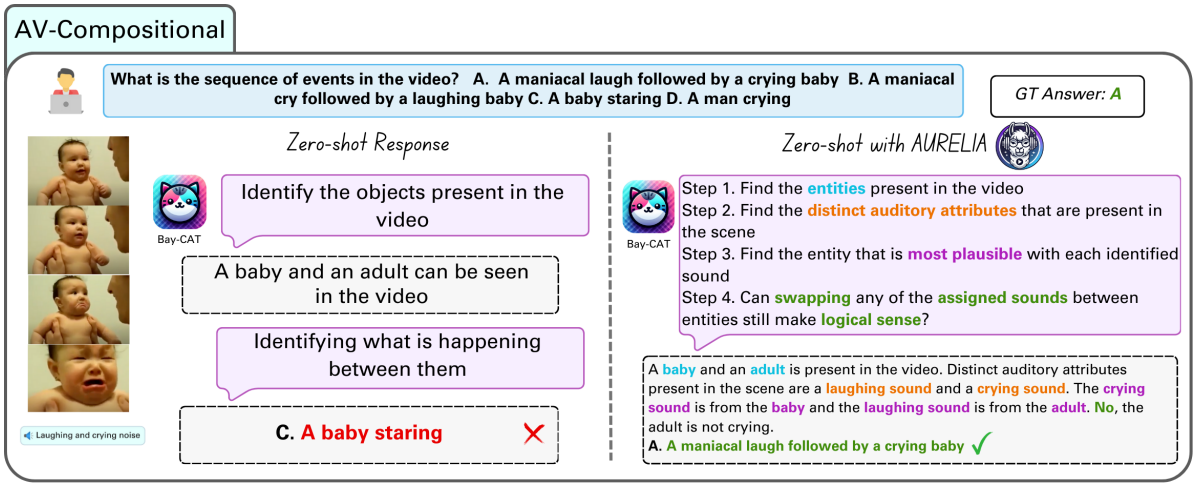
Figure 6. Qualitative visualization of AURELIA's reasoning distillation across AV-QA task.



(a)

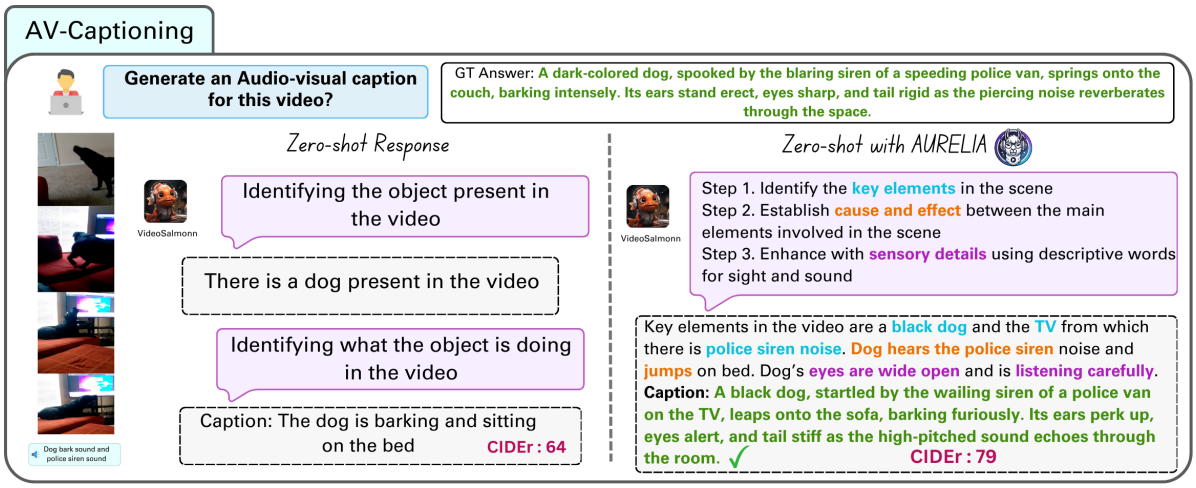


(b)

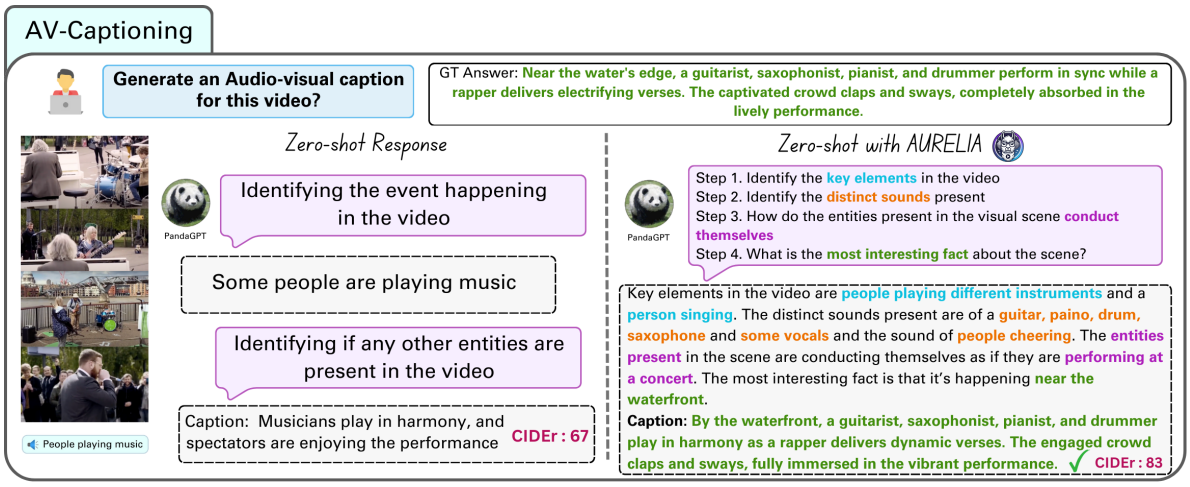


(c)

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



(a)



(b)

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

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

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

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

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

[48] 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.

[49] 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.

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

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] 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.

[56] 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.

[57] 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.

[58] 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.

[59] 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.

[60] 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.

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

[62] 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.

[63] 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.

[64] 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.

[65] 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.

[66] 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.

[67] 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.

[68] 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.

[69] 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.

[70] 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.

[71] 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.

[72] 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.

[73] 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.

[74] 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.

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

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

[77] 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.

[78] 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.

[79] 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.

[80] 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.

[81] 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.

[82] 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.

[83] 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.

[84] 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.