

AV-Reasoner: Improving and Benchmarking Clue-Grounded Audio-Visual Counting for MLLMs

Supplementary Material

1. Benchmark Construction

1.1. Reference-Query Modality Definition

To support comprehensive evaluation of multimodal counting capabilities, our benchmark defines five distinct modal settings based on the combination of reference and query modalities. Each setting reflects a different reasoning requirement, depending on which modality is used to locate the counting target and query interval:

- **Visual-only:** The model is required to both locate and count using only visual input. This setting targets scenarios where the counting question and answer are entirely grounded in visual content.
- **Audio-only:** The model must rely solely on audio cues to identify the relevant segments and determine the count. This setting emphasizes sound-based reasoning, such as counting distinct audio events or speaker turns.
- **Visual-reference, Audio-query:** The model uses visual input to locate the relevant temporal segment and then performs counting based on audio information within that scope (e.g., “How many people spoke in the scene showing the conference table?”).
- **Audio-reference, Visual-query:** The model leverages audio to identify the query interval and subsequently counts visually observable targets in the corresponding segments (e.g., “How many people are visible when the sound of clapping occurs?”).
- **Joint Audio-Visual:** Both audio and visual modalities are required to solve the task effectively. The model must integrate information across modalities to interpret the question, localize relevant segments, and produce an accurate count. This setting reflects complex scenarios where neither modality alone is sufficient, or where combining both provides more reliable counting signals.

1.2. Counting Target Definition

We define three types of counting targets in our benchmark to evaluate model capabilities from multiple perspectives:

- **Event:** A temporally localized activity or incident in the video. Event counting requires the model to recognize distinct occurrences over time.
- **Object:** A visually identifiable entity in the scene. Object counting involves detecting and enumerating instances. The model must avoid double-counting due to camera motion, scene transitions, or repeated appearances of the same entity.
- **Attribute:** A clustering of objects based on shared visual

or semantic properties. Rather than counting individual instances, the model must identify groupings according to the query attribute (e.g., “How many different clothing colors are worn by people?”).

1.3. More Dataset Statistics

As shown in Fig. 1, the videos selected from CG-Bench cover a diverse range of topics, including Life Record, Sports, Instruction, and TV Show, among more than ten categories. All videos are longer than 10 minutes, with the majority ranging from 20 to 30 minutes. This design encourages models to perform counting over long temporal contexts, which is essential for evaluating their capability in long-range temporal grounding and accumulation-based reasoning. This addresses a key limitation of existing counting benchmarks, which often rely on short clips that fail to capture the complexity of long-range dependencies.

1.4. Prompt for Generating Initial Question Proposals

```
You are given a video, and your task is to generate count-based audio-visual reasoning questions that can be answered by analyzing a clearly defined segment of the video.
```

```
### Definition of Count-Based Audio-Visual Reasoning Question
Each question should involve counting something that is either audibly heard, visually seen, or both, during a localized event in the video. The question must be specific, grounded in real scenes, and the answer must be objectively verifiable within a given time span.
```

```
### Modalities and Reasoning Types
- A2V (Audio to Visual): The question is triggered by an audio cue (e.g., a sound, noise, or dialogue) which helps the model identify the relevant time segment in the video. Once the time segment is located using the audio cue, the question asks about visual information (e.g., objects, people, or actions) within that segment. For example, "When the dog starts barking, how many people are visible in the background?" The audio cue identifies the time segment (e.g., when the dog barks), and the question asks about what is seen visually during that segment.
```

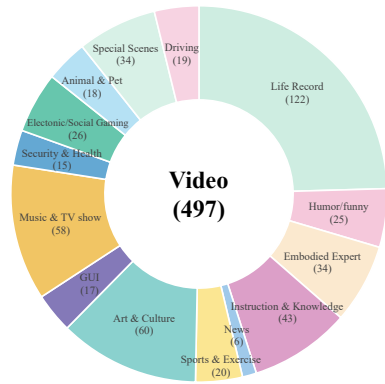
```
- V2A (Visual to Audio): The question is triggered by a visual cue (e.g., someone entering the frame, a person performing an action) which helps the model identify the relevant time segment in the video. The question then asks about auditory events (e.g., how many sounds or spoken words are heard) that occur during that time segment. For example, "When the firefighter enters the building, how many sirens can be heard in the background?" Here, the visual cue (the firefighter entering) locates the time segment, and the question asks about sounds during that segment.
```

```
- AV (Audio + Visual): The question requires both audio and visual cues to accurately count objects/events. Both cues help locate the relevant time segment in the video, and then the question asks for a count based on the interaction between both modalities. For example, "When the presenter gestures to the audience, how many people respond verbally?" Both audio (the people speaking) and visual (the presenter's gesture) cues are needed to define the time segment for counting.
```

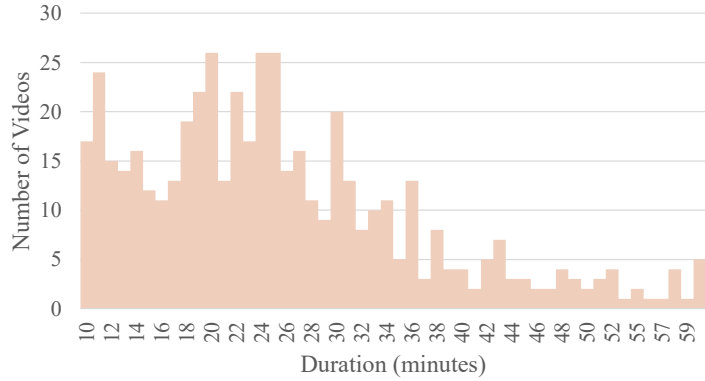
```
- A (Audio Only): The question is based solely on audio cues to identify a relevant time segment, and the task is to count auditory events within that segment. For example, "How many door slams can be heard in the scene?" The audio cue (the door slams) defines the time segment, and the model counts the auditory events within that period.
```

```
- V (Visual Only): The question is based solely on visual cues, with no need for audio. The task involves counting visible objects or actions in the defined segment. For example, "How many people are wearing blue shirts in the crowd at the park?" This is based purely on visual observation.
```

```
### Constraints
1. Each question must correspond to a specific and bounded video segment.
2. The answer must be a count (e.g., number of visual objects, number of auditory events, number of multimodal occurrences, etc.).
3. The answer must be clearly determined and not ambiguous within the video span.
```



(a) Statistics of Video Content Categories.



(b) Statistics of Video Duration.

Figure 1. Statistics of video content categories and duration.

- The difficulty of the question should be specified based on perceptual complexity (e.g., occlusion, background noise, overlapping motion/sound).
- Include a **diverse mix** of A2V, V2A, AV, A, and V questions in the output.
- Avoid redundant or trivial questions.
- The question should not include specific timepoints.

Output Format

Return your result as a **JSON array**, where each entry is a dictionary with the following fields:

- "question": A clear and specific counting question.
- "type": One of "A2V", "V2A", "AV", "A", or "V".
- "start_time": Start of relevant video segment ("MM:SS").
- "end_time": End of relevant video segment ("MM:SS").
- "counting_result": The correct count answer.

Example Output

Enclose the JSON block within `<json>`/`</json>` tags.

```
<json>
[
  {
    "question": "When the dog starts barking, how many people are visible in the background?",
    "type": "A2V",
    "start_time": "01:15",
    "end_time": "01:45",
    "counting_result": 3
  },
  {
    "question": "When the firefighter enters the building, how many sirens can be heard in the background?",
    "type": "V2A",
    "start_time": "02:00",
    "end_time": "02:30",
    "counting_result": 2
  },
  {
    "question": "When the presenter gestures to the audience, how many people respond verbally?",
    "type": "AV",
    "start_time": "03:00",
    "end_time": "03:40",
    "counting_result": 4
  },
  {
    "question": "How many door slams can be heard in the scene?",
    "type": "A",
    "start_time": "05:15",
    "end_time": "05:45",
    "counting_result": 3
  },
  {
    "question": "How many people are wearing different colors of clothes in the park?",
    "type": "V",
    "start_time": "02:00",
    "end_time": "02:30",
    "counting_result": 4
  }
]
</json>
```

1.5. Evaluation Prompts

1.5.1. Black-Box Evaluation

Watch the video and answer the question '(Question Here)' with a number. Just output the number itself, don't output anything else.

1.5.2. White-Box Evaluation (Event)

Watch the video and provide your answer to the question '(Question Here)', including the start and end timestamps for each event. Format your answer in JSON, enclosed in `<answer>` and `</answer>` tags. The output should look like this: `<answer>[{"start_time": "...", "end_time": "..."}]</answer>`. Ensure each timestamp is in seconds (e.g., 'xx.xx').

1.5.3. White-Box Evaluation (Object)

According to the given video frames, answer the question '(Question Here)', including the bounding box for the query object in the first frame where it appears. For subsequent frames where the object appears, do not provide the bounding box again. Format your answer in JSON, enclosed within `<answer>` and `</answer>` tags. The output should look like this: `<answer>[{"Frame1": [{"x_min", "y_min", "x_max", "y_max"}], "Frame2": [...]}</answer>`. In the output, each frame should either contain the bounding box of the object (if it appears for the first time in that frame) or an empty list '[]' (if the object does not appear or it has already been labeled in a previous frame). Ensure that bounding boxes are listed as [x_min, y_min, x_max, y_max].

1.5.4. White-Box Evaluation (Attribute)

According to the given video frames, answer the question '(Question Here)', clustering the objects based on the question. For each unique cluster, assign a unique label and return the bounding box for each object in the first frame where it appears. For subsequent frames where the object appears, do not output anything. Format your answer in JSON, enclosed within `<answer>` and `</answer>` tags. The output should look like this: `<answer>[{"Frame 1": [{"label": "Label 1", "x_min", "y_min", "x_max", "y_max"}], "Frame 2": [...]}</answer>`. In the output, each frame should either contain the bounding box and label for the object (if it appears for the first time in that frame) or an empty list '[]' (if the object has already been labeled or does not appear in that frame). The label should correspond to a unique object cluster according to the question.

1.6. More Evaluations

1.6.1. Model Performance across Different Counting Targets

Fig. 2 highlights a clear trend across both close-source and open-source MLLMs: models consistently perform better on object counting tasks compared to event and attribute counting. Most models demonstrate a noticeable advantage when counting concrete, visually grounded entities like objects. This performance gap is especially prominent in open-source models, where accuracy on event and attribute targets often drops substantially.

The relative ease of object counting can be attributed to the more direct visual correspondence between input and

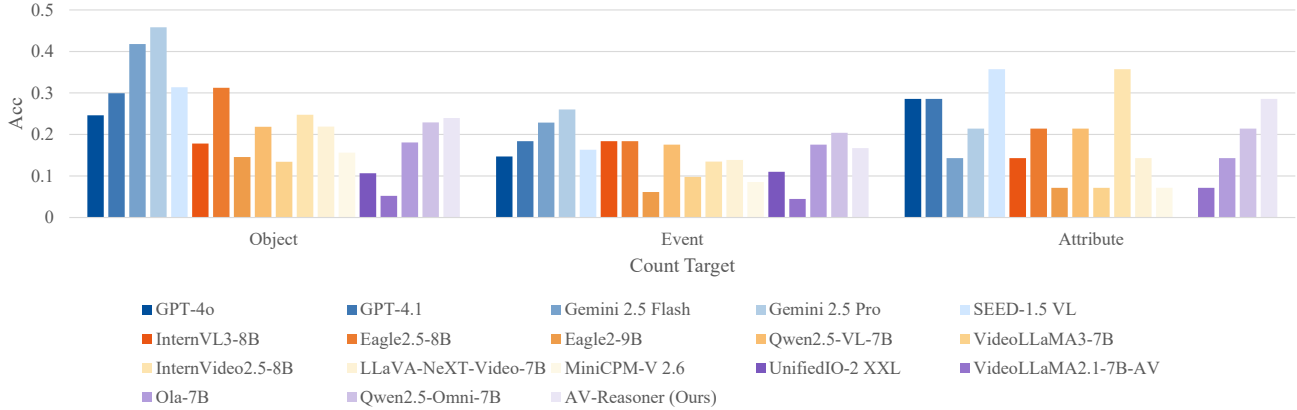


Figure 2. Model accuracy across different count targets.

target. Objects are typically well-defined spatially, consistently annotated in vision-language pretraining data, and often associated with discrete visual regions. In contrast, events may unfold over time and require temporal reasoning, while attributes tend to be abstract, context-dependent, or even implicit, making them harder to detect and quantify reliably.

These findings suggest that current MLLMs are more adept at processing perceptually salient elements. Addressing the challenges posed by more abstract or temporally extended targets like events and attributes may require stronger temporal modeling, better multimodal alignment, or targeted supervision in future model designs.

1.6.2. Model Performance across Different Counting Numbers

As shown in the Fig. 3, model performance clearly declines as the count range increases. Open-source models exhibit relatively high accuracy when the count is ≤ 5 , but their performance becomes more erratic and less reliable when the count exceeds 5.

1.6.3. Model Performance across Different Query Modalities

As shown in Fig. 4, model performance varies significantly across different query modalities. When the query modality is audio (A), most models—especially omni-MLLMs like Qwen2.5-Omni-7B and Ola-7B achieve relatively high accuracy and low error, indicating strong alignment with audio inputs. However, performance drops noticeably in cross-modal settings such as A2V and V2A, where both MAE and RMSE increase substantially. Notably, V2A and AV emerges as the most challenging configuration across all metrics. Overall, omni-MLLMs (purple) show clear advantages over VLMs (yellow) in audio-involved queries, but this advantage diminishes in purely visual or AV scenarios.

1.6.4. Model’s White-box Evaluation Performance across Different Counting Targets

As shown in Fig. 5, in the white-box evaluation, most models demonstrate relatively strong reasoning counting abilities in the event counting, but perform poorly in the object and attribute counting. Notably, only MiniCPM-V 2.6 [1] achieves a score in the attribute counting task. This suggests that current models are relatively proficient in temporal grounding, where understanding the sequence and timing of events is crucial, but perform poorly in spatial grounding, which requires precise localization and differentiation of objects or attributes.

2. Experiment Details

2.1. Experimental Setups

All experiments are conducted using the TRL framework on 8 NVIDIA A100 GPUs (each with 80GB memory). The base model used is Ola-Omni-7B. Training was performed with bf16 enabled to improve memory efficiency and computational performance.

We adopted the GRPO algorithm for fine-tuning. During the rollout phase, the generation temperature was set to 1.0, the maximum number of generated tokens is 256 or 1024 according to different tasks, and each sample was expanded into 8 rollout trajectories. The KL divergence penalty coefficient β was set to 0.1 to ensure controlled deviation from the initial policy.

2.2. Detailed Training Settings

Tab. 1 summarizes the training configurations across different stages. The curriculum-based RL process is divided into three subtasks: QA, grounding, and counting, each trained with its corresponding dataset. To mitigate forgetting and maintain performance across tasks, we adopt a stage review mechanism (SRM), which involves mixing a portion of pre-

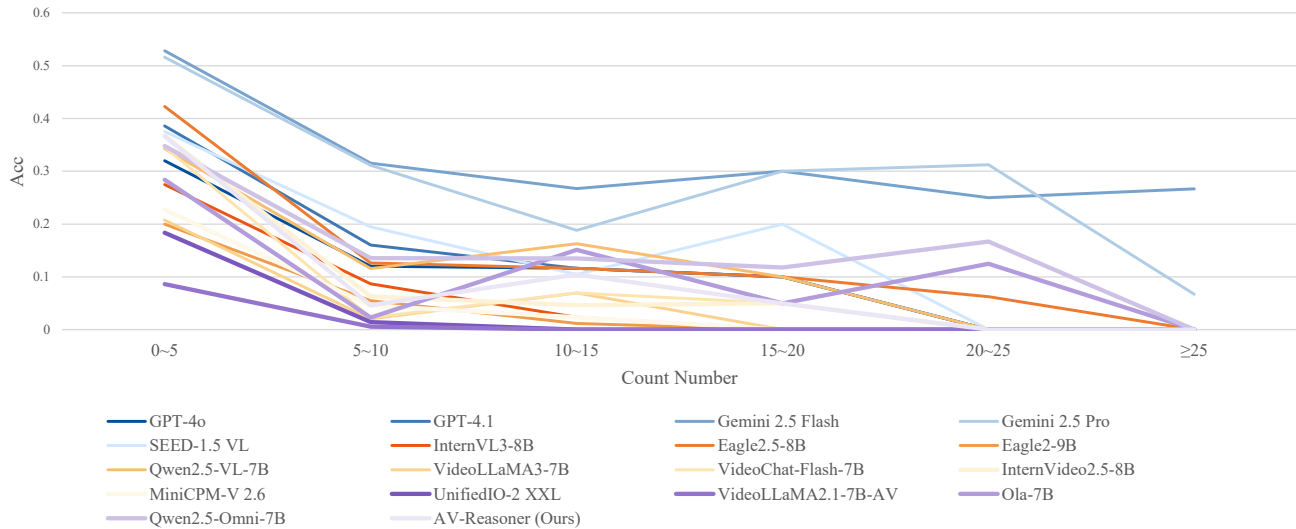


Figure 3. Model accuracy across different count ranges.

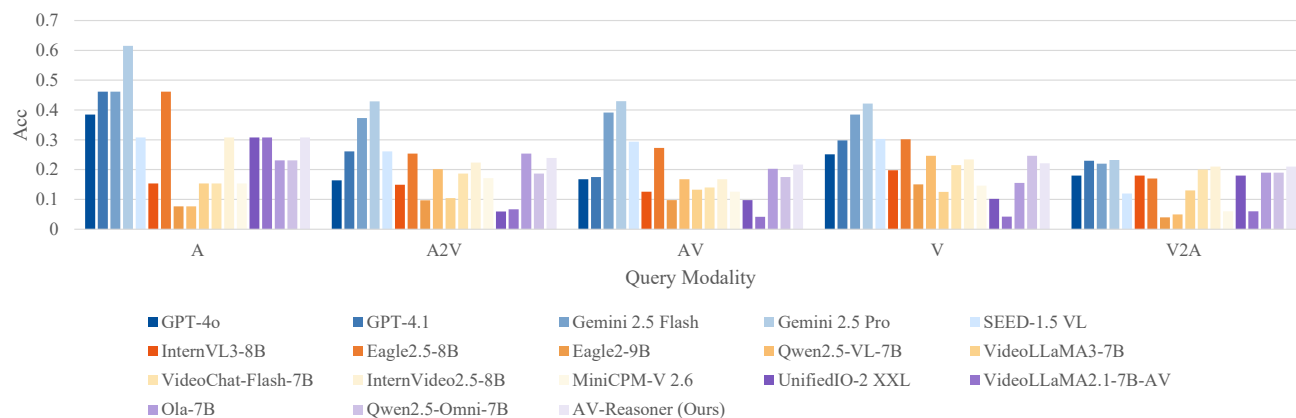


Figure 4. Model accuracy across different query modalities.

viously seen samples during training. The number of such samples is shown in parentheses.

3. Evaluation on More Benchmarks

The performance of our model is also evaluated on the AVQA and AVE benchmarks. The test results are shown in the Tab. 2.

4. Qualitative Results

Fig. 6 visualizes AV-Reasoner’s outputs on tasks including MCQ, QA, AVTG, ARIG, and Counting. The model not only provides correct answers but also generates coherent reasoning. For example, in AVTG, it identifies the sounds and determines the primary one, and in ARIG, it explains how the sounding object is located based on shape and audi-

tory cues.

References

- [1] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024. 3
- [2] Jiashuo Yu, Ying Cheng, Rui-Wei Zhao, Rui Feng, and Yuejie Zhang. Mm-pyramid: Multimodal pyramid attentional network for audio-visual event localization and video parsing. In *Proceedings of the 30th ACM international conference on multimedia*, pages 6241–6249, 2022. 6
- [3] Sanjoy Chowdhury, Sayan Nag, Subhrajyoti Dasgupta, Jun Chen, Mohamed Elhoseiny, Ruohan Gao, and Dinesh Manocha. Meerkat: Audio-visual large language model for grounding in

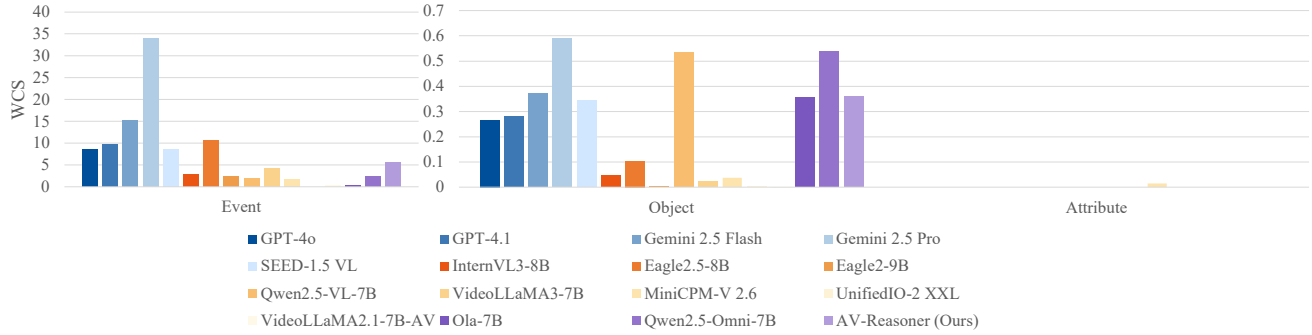


Figure 5. Model’s White-box Evaluation Performance across Different Counting Targets.

Table 1. Our detailed training settings. Data indicated in parentheses refers to that used in the SRM.

		Cold-start SFT	Curriculum-based RL (QA)	Curriculum-based RL (Grounding)	Curriculum-based RL (Counting)	Full-task RL
Data	Dataset	AVTG+ARIG+Counting	AVQA	AVTG+ARIG (AVQA)	Counting (AVTG+ARIG+AVQA)	AVQA+AVTG+ARIG+Counting
	#Samples	78K	72K	72K (14K)	6K (1K)	10K
Training	Thinking	✗			✓	
	Max New Tokens	-	256		1024	

space and time. In *European Conference on Computer Vision*, pages 52–70. Springer, 2024. 6

- [4] Zhuoming Liu, Yiquan Li, Khoi Duc Nguyen, Yiwu Zhong, and Yin Li. Pave: Patching and adapting video large language models, 2025. URL <https://arxiv.org/abs/2503.19794>. 6
- [5] Henghui Du, Guangyao Li, Chang Zhou, Chunjie Zhang, Alan Zhao, and Di Hu. Crab: A unified audio-visual scene understanding model with explicit cooperation. *arXiv preprint arXiv:2503.13068*, 2025. 6

Table 2. Performance comparison across AVQA and AVE Benchmarks.

Benchmark	MM-Pyramid [2]	MEERKAT [3]	PAVE [4]	Crab [5]	AV-Reasoner (Ours)	AV-Reasoner-Thinking (Ours)
AVQA Acc (%)	-	87.17	93.80	-	93.02	93.17
AVE Acc (%)	77.80	-	-	80.15	82.86	81.26

MCQ


 What is the main sound source of the video?
A. Car horn B. Civil defense alarm C. small bell D. Chimes




 <think> The video shows a civil defense siren, which is a type of alarm used for emergencies. The siren is clearly visible and is the main source of sound in the video. The other options, such as car horn, small bell, and chimes, are not present in the video. Therefore, the main sound source is the civil defense siren.</think>
<answer>B</answer>

QA

 Are there drum and saxophone sound?

 <think> The saxophone sound is present in the background, and the drum sound is present in the foreground. Both are part of the music being played. The saxophone is a common instrument in jazz music, and the drum is a fundamental component of any rhythm section. The presence of both instruments in the music is evident from the sound.</think> <answer>yes</answer>

AVTG

 From the 34.22-second video, 64 frames are sampled at these timestamps: Please describe the events and their time ranges from the video. Output in the following JSON Format: ...




 <think> The video shows a baby crying and sobbing, with occasional coughing. The baby's distress is evident throughout the video. The woman speaking is also noted, but it is not the main focus of the video. The baby's crying and sobbing are the primary events in the video.</think> <answer> [{"start": "0.00 seconds", "end": "34.01 seconds", "event": "baby crying", "type": "audio visual"}, {"start": "0.00 seconds", "end": "34.01 seconds", "event": "baby sobbing", "type": "audio visual"}, {"start": "31.01 seconds", "end": "34.01 seconds", "event": "woman speaking", "type": "audio visual"}]</answer>

ARIG

 Based on the given image in size of 1280x720 and its corresponding audio, please recognize the category of object making sound in the image, and then find out the bounding box coordinates of the object that makes the sound at the 2nd second of the audio.




 <think> The objects making sound are the two large, round, woven instruments that the man and boy are playing. These instruments are commonly used in Indian music and are known for their distinctive sound and appearance.</think> <answer> [{"coordinates": [152, 446], [1096, 720]}, {"name": "squirrel"}]</answer>

Counting

 Please answer the question 'during the earlier mentioned ball 's first flight , what number of stationary objects are there ?' with a number.



 <think> The ball is in motion, so it is not considered stationary. The other objects are stationary. Counting the stationary objects: cube, sphere, cylinder, cone, and another cone. Total count is 5.</think> <answer>5</answer>

Figure 6. Example Outputs on Various Tasks. AV-Reasoner not only outputs the correct answer, but also give the thinking process.