R-AVST: Empowering Video-LLMs with Fine-Grained Spatio-Temporal Reasoning in Complex Audio-Visual Scenarios

Lu Zhu[1, 2*], **Tiantian Geng**[1, 3*], **Yangye Chen**[1], **Teng Wang**[1, 4], **Ping Lu**[5], **Feng Zheng**[1, 2†]

[1]Southern University of Science and Technology
[2]Spatialtemporal AI
[3]University of Birmingham
[4]The University of Hong Kong
[5]ZTE Corporation
{zllllllau, gengtiantian97}@gmail.com, zhengf@sustech.edu.cn

Abstract

Recently, rapid advancements have been made in multimodal large language models (MLLMs), especially in video understanding tasks. However, current research focuses on simple video scenarios, failing to reflect the complex and diverse nature of real-world audio-visual events in videos. To bridge this gap, we firstly introduce R-AVST, a dataset for audio-visual reasoning featuring fine-grained spatio-temporal annotations. In constructing this, we design a pipeline consisting of LLM-based key object extraction, automatic spatial annotation and manual quality inspection, resulting in over 5K untrimmed videos with 27K objects across 100 types of audio-visual events. Building on this dataset, we define three core tasks for spatio-temporal reasoning in audio-visual scenes and generate more than 8K high-quality, evenly distributed question-answer pairs to effectively benchmark model performance. To further enhance reasoning, we propose AVST-Zero, a reinforcement learning-based model that avoids intermediate supervision, directly optimizing behavior via carefully designed multi-dimensional rewards. Extensive experiments validate the effectiveness of our R-AVST in advancing audio-visual spatio-temporal reasoning, upon which AVST-Zero demonstrates competitive performance compared to existing models. To the best of our knowledge, R-AVST is the first dataset designed for real-world audio-visual spatio-temporal reasoning, and AVST-Zero offers a novel perspective for tackling future challenges in this domain.

Introduction

The rapid advancement of multimodal large language models (MLLMs) recently has demonstrated their effectiveness in video understanding by integrating information from various modalities (Wang et al. 2024b; Zhang et al. 2024; Tang et al. 2025a; Geng et al. 2025). However, this rapid progress raises a question: *Do current models and datasets adequately account for the compositional audio-visual nature of real-world video scenes, and leverage it to enhance spatio-temporal understanding of videos?* This issue is particularly important, as many videos originate from real-world audio-visual events, which play a vital role in practical applications, such as human-computer interaction, autonomous driving, and so on, where audio and visual modalities are both crucial for temporal and spatial perception.

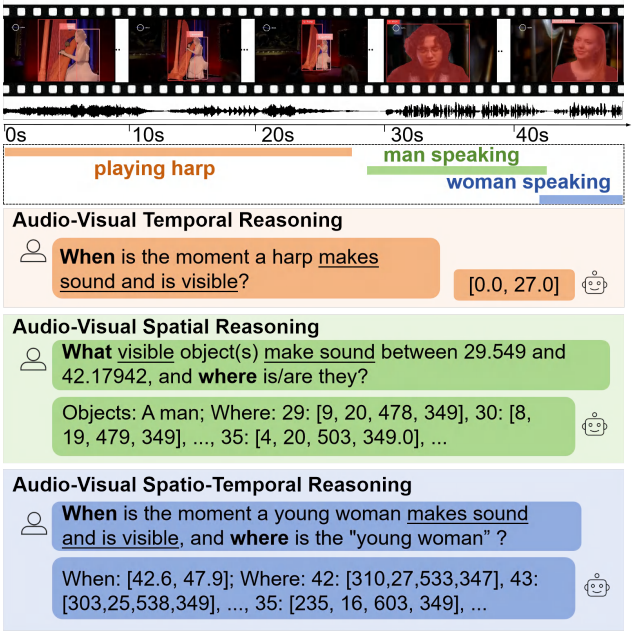

Figure 1: Unlike previous datasets, R-AVST focuses on spatio-temporal reasoning in complex audio-visual scenes of untrimmed videos, offering fine-grained temporal boundary and spatial localization annotations. This example shows three core tasks designed to evaluate reasoning over sounding objects, time, and space.

In response to this issue, existing datasets such as AVE (Tian et al. 2018), UnAV-100 (Geng et al. 2023), and PU-VALOR (Tang et al. 2025b), incorporate audio-visual information but primarily focus on temporal understanding while overlooking the spatial properties of audible objects. In contrast, spatio-temporal grounding datasets such as Vid-STG (Zhang et al. 2020), HC-STVG (Tang et al. 2021),

*Equal contribution.
†Corresponding author.

and V-STaR (Cheng et al. 2025) offer both temporal and spatial annotations, but they do not adequately capture the rich audio-visual dynamics of real-world scenes and typically involve only a limited range of object types. Therefore, proposing a fine-grained spatio-temporal dataset that jointly integrates audio and visual information is both valuable and urgent, as it can enhance models' reasoning capabilities in real-world scenarios.

At the model level, models such as LLaVA-ST (Li et al. 2025a), GroundingGPT (Li et al. 2024) and Grounded-VideoLLM (Wang et al. 2024a) have gradually extended their spatio-temporal modeling capabilities. However, these models require extensive high-quality labeled data and lack sufficient exploration capacity. With the promising performance of Deepseek-R1 (Guo et al. 2025) in rule-based reinforcement learning (RL), some studies have begun exploring the use of GRPO algorithms (Shao et al. 2024) to enhance the reasoning capabilities of MLLMs. Models like VideoChat-R1 (Li et al. 2025b), Video-R1 (Feng et al. 2025) and Omni-R1 (Zhong et al. 2025) have emerged in this direction, yet they offer limited reward designs for spatio-temporal reasoning and lack dedicated tasks tailored for complex audio-visual scenarios.

To address the above limitations, we propose R-AVST, the first video dataset with fine-grained spatio-temporal annotations in complex audio-visual scenarios, as illustrated in Fig. 1. In audio-visual scenarios, both auditory and visual attributes of objects can capture human perception and attention. Motivated by this, we use GPT-4o-mini (Hurst et al. 2024) to extract and analyze audio-visual event captions to label objects' attributes. This allows models to determine whether an object is audible, visible, or both, and subsequently supports automatic spatial annotation for grounding object locations. To further assess models' reasoning capabilities and align with human inference needs in complex audio-visual scenes, we define three targeted reasoning tasks, including temporal localization of sounding-visible objects, spatial localization of sounding-visible or silent-visible objects within a given duration, and spatio-temporal localization of sounding-visible objects. Based on these tasks, we automatically generate corresponding question-answer pairs (QAs) to enable large-scale evaluation. In total, R-AVST comprises 5,237 videos with 27,253 objects both sounding-visible and silent-visible and 8,166 QAs, covering over 100 types of audio-visual events, such as human speech, musical performances, and animal sounds.

Building upon R-AVST, we further focus on addressing the challenge of fine-grained spatio-temporal reasoning capabilities of models in complex audio-visual scenarios. To address the lack of advanced reasoning capabilities and the dependence on large-scale high-quality annotated data in models such as LLaVA-ST (Li et al. 2025a), we fine-tune our model using the data-efficient GRPO (Shao et al. 2024) method from DeepSeek-R1 (Guo et al. 2025), given the rule-based characteristics of our tasks. Meanwhile, to overcome the absence of task-specific objectives and reward designs for complex audio-visual spatio-temporal reasoning in models like VideoChat-R1 (Li et al. 2025b), we develop a multi-dimensional reward system tailored to our tasks. This system includes format, object, temporal, and spatial reward, which collectively enable effective policy gradient updates. Our experiments demonstrate that AVST-Zero achieves competitive performance on the three core tasks. It surpasses most Video-LLMs and sets a new perspective in the audio-visual spatio-temporal reasoning tasks.

Our contributions can be summarized as follows:

- We introduce R-AVST, the first video dataset encompassing a wide range of complex audio-visual events and featuring fine-grained spatio-temporal annotations, specifically designed to facilitate multimodal reasoning and evaluation in realistic scenarios of videos.
- Aiming to systematically evaluate models' spatio-temporal reasoning capabilities and to align more closely with human retrieval demands in complex audio-visual contexts, we introduce three specialized tasks: Audio-Visual Temporal, Spatial, and Spatio-Temporal Reasoning, alongside automatically constructed QAs based on LLM-generated labels.
- We construct AVST-Zero, a Video-LLM fine-tuned in fully GRPO, trained on R-AVST to enhance its performance on audio-visual spatio-temporal reasoning tasks. Experimental results demonstrate that AVST-Zero achieves competitive performance across all three core tasks, validating its effectiveness.

Related Work

Spatio-Temporal Understanding in Video-LLMs

Spatio-temporal understanding is crucial for extracting key information from videos (Goodge et al. 2025; Yuan et al. 2024; Zou et al. 2023; Geng et al. 2024). With the rise of MLLMs, general-purpose Video-LLMs like InternVL-2.5 (Chen et al. 2024b), Qwen2.5-VL (Bai et al. 2025), VideoLLaMA3 (Zhang et al. 2025), and GroundingGPT (Li et al. 2024) have made notable strides, along with specialized models such as LLaVA-ST (Li et al. 2025a), VideoMolmo (Ahmad et al. 2025), and Meerkat (Chowdhury et al. 2024) that focus on spatio-temporal reasoning. These models benefit from improvements like stronger encoders and audio integration, yet still struggle with complex audio-visual spatio-temporal tasks. On the dataset side, benchmarks like HC-STVG (Tang et al. 2021) and Vid-STG (Zhang et al. 2020) provide spatio-temporal grounding annotations, but feature limited object diversity. While BOSTVG (Yao et al. 2025) introduces a multi-object setting but uses short videos. V-STaR (Cheng et al. 2025) targets long-video scenarios, yet lacks realistic audio-visual content. AV-UIE (Du et al. 2025) centers on audio-visual understanding but mainly targets image-audio pairs, overlooking the construction of video-audio data.

Reinforcement Learning Enhancement in LLMs

Reinforcement Learning (RL) effectively enhances LLM reasoning with limited supervision. GPT-4 (Achiam et al. 2023) uses PPO with reward and value functions, while ChatGLM3-DPO (GLM et al. 2024) applies DPO through pairwise comparisons. DeepSeek-R1 (Guo et al. 2025) introduces GRPO (Shao et al. 2024), using rewards within

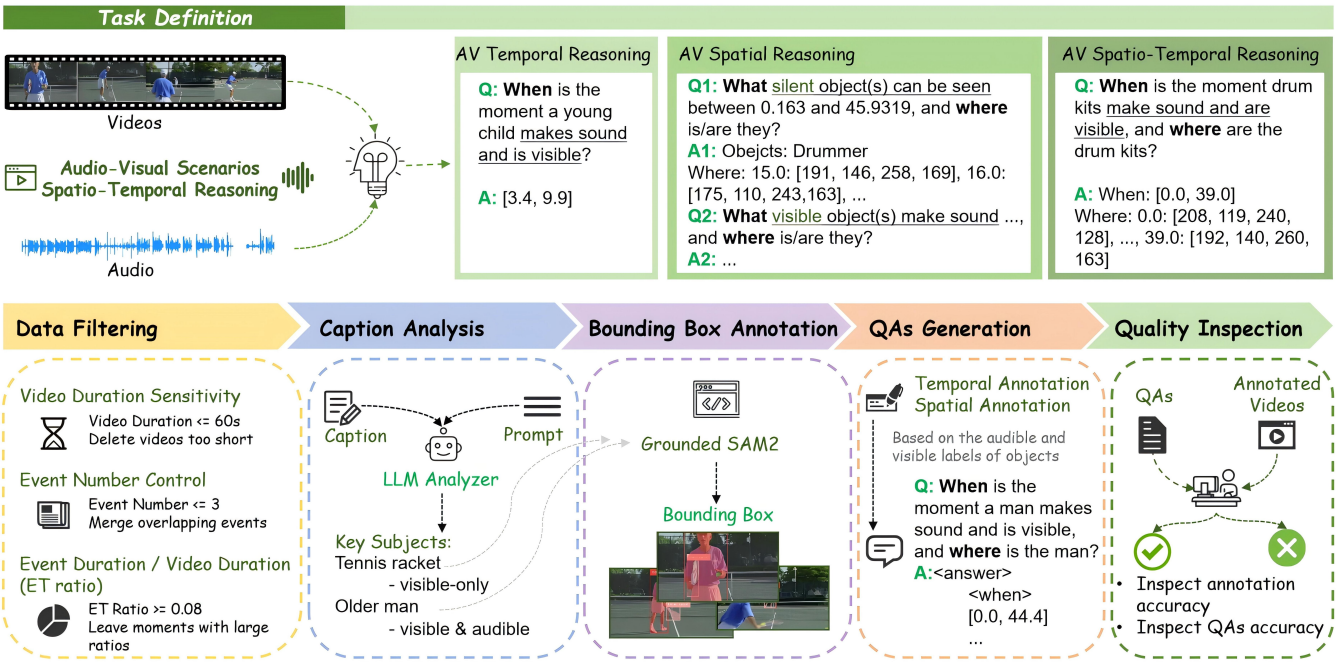

Figure 2: Data generation pipeline of R-AVST. The tasks are explicitly designed to capture both spatial and temporal aspects of complex audio-visual scenes. The dataset construction follows a five-step process, yielding fine-grained spatio-temporal annotations and task-oriented QAs.

groups to improve reasoning without value functions. GRPO has since been adopted in Video-LLMs as a method after supervised fine-tuning (SFT). VideoChat-R1 (Li et al. 2025b) improves spatio-temporal perception, Video-R1 (Feng et al. 2025) enhances temporal modeling, and R1-Omni (Zhao, Wei, and Bo 2025) incorporates audio for emotion understanding. Recent work (Guo et al. 2025) explores fully RL-based training, as in Omni-R1 (Zhong et al. 2025), which proposes an end-to-end GRPO framework. However, RL models specifically targeting spatio-temporal reasoning in audio-visual scenarios remain underexplored.

R-AVST Dataset

Task Definition

Existing datasets mainly focus on visual-based temporal and spatial reasoning, often neglecting the audio modality. In complex real-world audio-visual contexts, auditory cues are crucial for accurately localizing objects in both time and space. However, the absence of fine-grained audio-visual scene annotations limits comprehensive evaluation, as shown in Tab. 1. To fill this gap, we propose R-AVST dataset and three reasoning tasks designed for spatio-temporal reasoning in complex audio-visual scenarios, as Fig. 2 shows.

Audio-Visual Temporal Reasoning How to ground the temporal period when an object makes sound and is visible helps us better extract key information from audio-visual video scenarios. In this dimension, we design the Audio-Visual Temporal Reasoning task to infer about the time when the object appears and makes sound.

Audio-Visual Spatial Reasoning To better capture the spatial relationships among objects in audio-visual events, we introduce a novel Audio-Visual Spatial Reasoning task that considers both sound-emitting and non-sound-emitting objects. Given a specific temporal segment of an audio-visual event, Video-LLMs are supposed to accurately ground the target objects within the scene.

Audio-Visual Spatio-Temporal Reasoning In real-world scenarios, human perception of the audio-visual events is usually a process of simultaneously obtaining temporal and spatial information. Therefore, Audio-Visual Spatio-Temporal Reasoning task aims to evaluate the joint understanding ability of the model for temporal and spatial information in audio-visual scenes, making it closer to the real perception mechanism of human beings. Specifically, given that an object is known to have both visible and audible state, the task objective is to identify the temporal period when the object appears and further locate its spatial position within these periods.

Dataset Construction

Data Collection and Filtering We collect videos from UnAV-100 (Geng et al. 2023), an audio-visual dataset of untrimmed videos covering diverse domains. Starting from raw YouTube videos and the corresponding event captions, we employ a filtering strategy to select high-quality samples, aiming to ensure comprehensive coverage of complex audio-visual scenarios. As detailed in Fig. 2, the filtering involves three steps to balance video duration, event count, and event coverage ratio. First, videos are grouped

Figure 3: Statistics of R-AVST dataset. (a) Duration distribution of different event categories in descending order, where colors represent their corresponding coarse-grained categories. (b) Duration distribution of all videos. (c) Distribution of the number of audio-visual events per video. (d) Word cloud of event categories.

into short (0-20s), medium (20-40s), and long (40-60s) durations, excluding extremely short clips and merging overlapping events. Second, we limit videos to at most three audio-visual events, ensuring a balanced distribution of 1, 2, and 3-event videos to facilitate clearer scene-level evaluation. Third, videos with an event-to-total (ET) duration ratio below 0.08 are removed, retaining those where meaningful events occupy a substantial portion. Finally, we curate a dataset of 5,237 high-quality videos, encompassing a wide variety of complex audio-visual scenarios.

Caption Analysis Audio-visual spatio-temporal reasoning tasks typically focus on objects, mostly expressed as nouns in captions. We use GPT-4o-mini (Hurst et al. 2024) as an Analyzer LLM to extract noun-based objects from captions, as shown in Fig. 2. To capture the multi-modal nature, tailored prompts guide the model to annotate each object's auditory and visual attributes in a standardized format. We query the analyzer by emphasizing the definition of audibility in object-sound relations to improve analysis accuracy. For example, in the caption "A group of people are sailing on silver sailboats", "a group of people" is labeled "visible&audible" while "silver sailboats" is "visible-only". Finally, we identify 27,253 objects in audio-visual event captions in total, with 50.88% labeled in "visible&audible".

Bounding Box Annotation Based on caption analysis and temporal annotations, we perform spatial annotation on video frames within audio-visual event segments. To reduce the high annotation cost of large-scale videos, we leverage the automatic tool, Grounded-SAM2 (Ravi et al. 2024), for fine-grained, frame-by-frame object annotation. The pipeline extracts frames corresponding to each event and constructs textual prompts based on object information derived from caption analysis, which are then fed into Grounded-SAM2 (Ravi et al. 2024).

Automatic QAs Generation To align with the three reasoning tasks we propose for audio-visual scenes, R-AVST defines three corresponding question types: **when** (temporal), **where** (spatial), and **what** (object). Below are question detailed settings for each task:

- Audio-Visual Temporal Reasoning: With certain visible and audible objects, the question is formulated as: *When is the moment [objects] make sound and are visible?*
- Audio-Visual Spatial Reasoning: Given a time interval, the questions are formulated as: (1) For sounding-visible objects: *What objects make sound between [start_time] and [end_time], and where are they?* (2) For silent-visible objects: *What silent objects can be seen between [start_time] and [end_time], and where are they?*
- Audio-Visual Spatio-Temporal Reasoning: With certain visible and audible objects, the question is formulated as: *When is the moment [objects] make sound and are visible, and where are they?*

Dataset	#Vid	#Cls	Len	#Obj	Mod	TB	SA
Charades-STA	10,009	157	30s	-	V	✓	✗
AVE	4,143	28	10s	-	VA	✓	✗
UnAV-100	10,790	100	42.1s	-	VA	✓	✗
PU-VALOR	114,000	-	10s	-	VA	✓	✗
LongVALE	8,411	-	235s	-	VA	✓	✗
VidSTG	6,924	79	28.01s	1	V	✓	✓
HCSTVG-v1	5,660	1	20s	1	V	✓	✓
HCSTVG-v2	16,544	1	20s	1	V	✓	✓
BOSTVG	10,018	23	36.5s	2.4	V	✓	✓
V-STaR	2,094	-	110.23s	-	V	✓	✓
AVSBench-V1	5,356	23	5s	-	VA	✗	✓
AVSBench-V2	12,356	70	7.64s	-	VA	✗	✓
VPO	22,019	21	10s	-	VA	✗	✓
LU-AVS	7,200	88	41.97s	-	VA	✗	✓
R-AVST (Ours)	5,237	100	42.17s	5.2	VA	✓	✓

Table 1: Comparison of R-AVST with previous related datasets. #Cls: event category number; Len: average video duration; #Obj: average video object number; Mod: modality of event captions; V: visual events; VA: audio-visual events; TB: temporal boundary; SA: spatial annotation.

The answer follows a unified format, with different tags depending on the question type. Programs extract object audio-visual labels and generate corresponding QAs. The training set contains 2,663 temporal, 2,666 spatial, and 1,204 spatio-temporal questions, while the test set includes 663, 664, and 306 questions of each type, respectively.

Quality Control To ensure the high quality of the R-AVST dataset, we conduct manual verification of the spatio-temporal annotations and QAs at this stage. Videos with incorrect spatio-temporal annotations or inaccurate QAs are removed. This process enhances the robustness and accuracy of the R-AVST dataset, providing a more reliable basis for further evaluating of Video-LLMs.

Dataset Analysis

Overview Overall, we introduce R-AVST, the first dataset specifically designed to evaluate the spatio-temporal reasoning capabilities of Video-LLMs in complex audio-visual scenarios. As illustrated in Fig. 3, the dataset comprises 5,237 videos of 100 categories for over 220,833 seconds, with an average duration of 42.17 seconds per video. The training/test split has 4,171/1,066 videos with 6,533/1,633 QAs, respectively. These videos encompass a wide range of audio-visual events, with a relatively balanced distribution across videos containing 1, 2, or 3 events. Importantly, each video is accompanied by fine-grained spatio-temporal annotations, providing strong support for advancing research in reasoning within complex audio-visual video contexts.

Comparisons with Existing Datasets As shown in Tab. 1, we compare R-AVST with existing relevant datasets (Gao et al. 2017; Tian et al. 2018; Tang et al. 2025b; Geng et al. 2023, 2025; Zhang et al. 2020; Tang et al. 2021; Yao et al. 2025; Cheng et al. 2025; Zhou et al. 2022, 2025; Liu et al. 2024; Chen et al. 2024a). Existing datasets primarily focus on general temporal or visual grounding scenarios and lack fine-grained modeling of audio-visual scenarios. On the other hand, while datasets such as AVSBench (Zhou et al. 2022, 2025), VPO (Chen et al. 2023), and LU-AVS (Liu et al. 2024) focus on spatial localization of audible objects, they often neglect the temporal aspect. In contrast, R-AVST covers a wider range of events and objects, and is tailored for fine-grained spatio-temporal reasoning in real-world audio-visual scenes, offering a more comprehensive benchmark for complex video reasoning tasks.

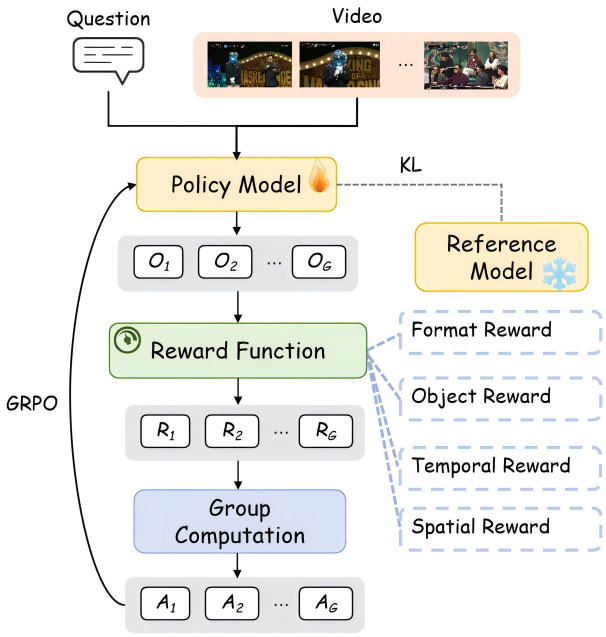

Figure 4: Model architecture of our AVST-Zero model. The multi-dimensional reward design allows AVST-Zero to perform exceptionally well in spatio-temporal reasoning tasks.

AVST-Zero

To further enhance the spatio-temporal reasoning capabilities of Video-LLMs, we fine-tune the model with GRPO (Shao et al. 2024), along with task-oriented rewards designed to improve performance on the audio-visual spatio-temporal reasoning domain.

Group Relative Policy Optimization

GRPO is a reinforcement learning (RL) method and can be used in multimodal large language models (MLLMs) to guide task alignment through reward functions. It differs from PPO in that it reduces reliance on a critic model by directly comparing groups of generated responses. Specifically, for each question q, GRPO (Shao et al. 2024) samples a set of outputs $\{o_1, o_2, \ldots, o_e\}$ from the old policy $\pi_{\theta_{\text{old}}}$, and then optimizes the policy π_θ by maximizing the objec-

tive defined as:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}\Bigg[q \sim P(Q),\, \{o_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(O \mid q)\Bigg] \frac{1}{G}\sum_{i=1}^{G}\Bigg(\min\Bigg(\frac{\pi_\theta(o_i \mid q)}{\pi_{\theta_{\text{old}}}(o_i \mid q)}A_i, \text{clip}\Bigg(\frac{\pi_\theta(o_i \mid q)}{\pi_{\theta_{\text{old}}}(o_i \mid q)}, 1-\varepsilon, 1+\varepsilon\Bigg)A_i\Bigg) - \beta\,\mathbb{D}_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}})\Bigg), \tag{1}$$

where A_i denotes the relative advantage of the i-th sample within the group of generated responses, estimated directly from rule-based rewards $\{r_1, r_2, \ldots, r_G\}$. The term $\mathbb{D}_{KL}\left(\pi_\theta \| \pi_{ref}\right)$ denotes the KL divergence, which measures the degree to which the optimized policy model π_θ deviates from the reference model π_{ref}. The hyper-parameters ε and β control the clipping threshold of the advantages and the penalty intensity of the KL-regularization term, respectively.

Rewards Design

To align with the specific characteristics of our tasks, we introduce four distinct reward components: format, object, temporal, and spatial. Each is designed to capture and reinforce a particular dimension of the tasks.

Format Reward The reward R_{format} assesses the format consistency of outputs by checking whether the required tag pairs (`<answer>`, `<object>`, `<when>`, `<where>`) are correctly included and matched, based on task requirements.

Object Reward We use Word2Vec (Mikolov et al. 2013) to quantify the semantic similarity between the predicted and ground truth object names, which helps mitigate misclassification caused by lexical variations. The similarity $\text{sim}\,(V_{\text{pred}}, V_{\text{gt}})$ is defined as:

$$\text{sim}\,(V_{\text{pred}}, V_{\text{gt}}) = \frac{V_{\text{pred}} \cdot V_{\text{gt}}}{\|V_{\text{pred}}\|\,\|V_{\text{gt}}\|}, \tag{2}$$

where V_{pred} and V_{gt} denote the word embeddings of the predicted and ground truth object names, respectively. Based on this, the object reward is defined as follow, where τ is the similarity threshold.

$$R_{\text{object}} = \begin{cases} 1, & \text{if}\,\text{sim}\,(V_{\text{pred}}, V_{\text{gt}}) \geq \tau. \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

Temporal Reward To enhance the model's temporal reasoning ability, this reward evaluates the accuracy of predicted audio-visual event segments by measuring the overlap between the predicted interval I_{pred} and the ground truth interval I_{gt}. The temporal reward computation is defined as their intersection-over-union (IoU) ratio:

$$R_{\text{temporal}} = \frac{|I_{\text{pred}} \cap I_{\text{gt}}|}{|I_{\text{pred}} \cup I_{\text{gt}}|}. \tag{4}$$

Spatial Reward To enhance fine-grained spatial reasoning, we define a spatial reward as the average 2D IoU between predicted and ground truth bounding boxes over their overlapping temporal interval. For each time point $t \in [T_{\text{start}}, T_{\text{end}}]$, the IoU is computed as follow if the object prediction is correct:

$$\text{IoU}(t) = \frac{\text{Area}(B_{\text{pred}}(t) \cap B_{\text{gt}}(t))}{\text{Area}(B_{\text{pred}}(t) \cup B_{\text{gt}}(t))}. \tag{5}$$

Therefore, the spatial reward R_{spatial} can be computed as the mean IoU over all N time points:

$$R_{\text{spatial}} = \frac{1}{N}\sum_{t=T_{\text{start}}}^{T_{\text{end}}} \text{IoU}(t). \tag{6}$$

Final Reward The total reward is computed by the weighted sum of the rewards in different parts as:

$$R = \lambda_{\text{f}} R_{\text{format}} + \lambda_{\text{t}} R_{\text{temporal}} + \lambda_{\text{o}} R_{\text{object}} + \lambda_{\text{s}} R_{\text{spatial}}. \tag{7}$$

For all tasks, $\lambda_{\text{f}} = 1$, while other parameters (λ_{t}, λ_{o}, λ_{s}) are set depending on the task type.

Experiments

Experimental Settings

In experiments, we base Qwen2.5-VL 7B (Bai et al. 2025) and Qwen2.5-Omni 7B (Xu et al. 2025) to fine-tune two model variants on our R-AVST dataset, namely AVST-Zero and AVST-Zero-Omni, respectively. Training is conducted on four NVIDIA RTX A6000 GPUs for a single epoch, with a batch size of 1 on each device. The group generation number is set to 6.

Quantitative Results

As shown in Tab. 2 and Tab. 3, we compare our models with existing advanced models (Bai et al. 2025; Xu et al. 2025; Cheng et al. 2024; Li et al. 2024; Chen et al. 2024b; Li et al. 2025b; Feng et al. 2025). The evaluation is conducted on the R-AVST test set. In Tab. 2, for the audio-visual spatial reasoning task, AVST-Zero performs similarly to other models in m_vIoU but slightly outperforms them in AP@0.3 with 3.12%. For the audio-visual spatio-temporal reasoning task, AVST-Zero shows significantly better performance, achieving 46.04% m_tIoU and 8.59% m_vIoU. As shown in Tab. 3, AVST-Zero leads in average temporal perception for the audio-visual temporal reasoning task with an m_tIoU of 47.96%. We also observe that AVST-Zero-Omni achieves higher prediction accuracy than AVST-Zero in the object and spatial dimensions, while performing worse in the temporal dimension. This is attributed to the base model's strong audio–visual joint perception but relatively weak temporal perception capabilities. These results show that our AVST-Zero variants perform competitively across all three tasks, confirming their effectiveness.

Ablation Study

We conduct ablation experiments on a test subset with a balanced ratio 1:1:1 of the three task types to comprehensively

Method	Audio-Visual Spatial Reasoning				Audio-Visual Spatio-Temporal Reasoning						
	Object	Spatial			Temporal				Spatial		
	Accuracy	m_vIoU	AP@0.3	AP@0.5	m_tIoU	R1@0.3	R1@0.5	R1@0.7	m_vIoU	AP@0.3	AP@0.5
Qwen2.5-VL(7B)	1.91	2.31	0.90	0.15	34.55	45.10	29.74	15.69	1.37	1.14	0.65
Qwen2.5-Omni(7B)	14.04	1.96	2.43	1.24	33.44	43.46	19.28	9.80	2.85	2.99	1.04
Video-LLaMA3(7B)	15.07	1.27	0.16	0.15	37.43	50.65	34.64	22.22	1.69	1.63	0.00
GroundingGPT(7B)	0.55	0.16	0.00	0.00	13.65	15.05	6.02	1.67	5.59	3.68	0.00
InternVL2.5(8B)	15.97	0.74	0.66	0.00	21.46	26.47	13.07	7.52	2.87	2.80	0.00
Video-R1(7B)	13.02	1.19	0.95	0.40	22.05	26.47	10.13	5.23	0.15	0.11	0.00
VieoChat-R1(7B)	15.54	1.99	3.11	0.36	41.81	60.78	44.77	**25.49**	2.15	3.21	0.60
AVST-Zero (7B)	14.34	2.27	3.12	0.87	**46.04**	**67.32**	**46.08**	23.53	8.59	10.38	3.83
AVST-Zero-Omni (7B)	**19.48**	**3.87**	**4.47**	**2.17**	35.97	50.00	20.92	10.46	**17.74**	**22.90**	**12.26**

Table 2: Comparison of different Video-LLMs for audio-visual spatial and spatio-temporal reasoning tasks on R-AVST test set.

Method	Audio-Visual Temporal Reasoning			
	m_tIoU	R1@0.3	R1@0.5	R1@0.7
Qwen2.5-VL(7B)	36.05	46.40	34.38	16.22
Qwen2.5-Omni(7B)	30.70	37.09	18.92	9.01
Video-LLaMA3(7B)	37.17	50.30	35.29	22.67
GroundingGPT(7B)	10.77	11.06	5.38	1.84
InternVL2.5(8B)	18.37	22.07	10.21	5.26
Video-R1(7B)	22.48	22.82	12.16	6.01
VieoChat-R1(7B)	43.17	60.81	46.70	**25.68**
AVST-Zero (7B)	**47.96**	**71.13**	**51.43**	23.91
AVST-Zero-Omni (7B)	34.79	51.05	21.32	11.56

Table 3: Comparison of different Video-LLMs for audio-visual temporal reasoning tasks on R-AVST test set.

Model	AVTR	AVSR		AVSTR	
	m_tIoU	Obj Acc	m_vIoU	m_tIoU	m_vIoU
SFT	42.84	9.52	3.42	38.40	4.26
AVST-Zero	**48.17**	20.72	**4.62**	**46.93**	**10.87**
AVST-Zero (w/o temporal reward)	46.67	**23.95**	4.54	45.82	8.31
AVST-Zero (w/o spatial reward)	47.03	23.17	3.28	44.29	9.23

Table 4: Ablation study on supervised fine-tuning (SFT) and the reward components. AVTR, AVSR, and AVSTR denote the Audio-Visual Temporal, Spatial, and Spatio-Temporal Reasoning tasks, respectively.

assess the impact of each reward. As shown in Tab. 4, removing the temporal reward reduces temporal accuracy from 48.17% to 46.67%, while removing the spatial reward significantly lowers m_vIoU, from 4.62% to 3.28% (spatial reasoning) and from 10.87% to 9.23% (spatio-temporal reasoning). Meanwhile, we also observe that the interdependence of the spatio-temporal dimension leads to cross-effects between different reward modules. Moreover, compared with simple SFT, directly applying RL yields more substantial benefits across all three tasks, suggesting that RL is better suited to the tasks' fine-grained nature.

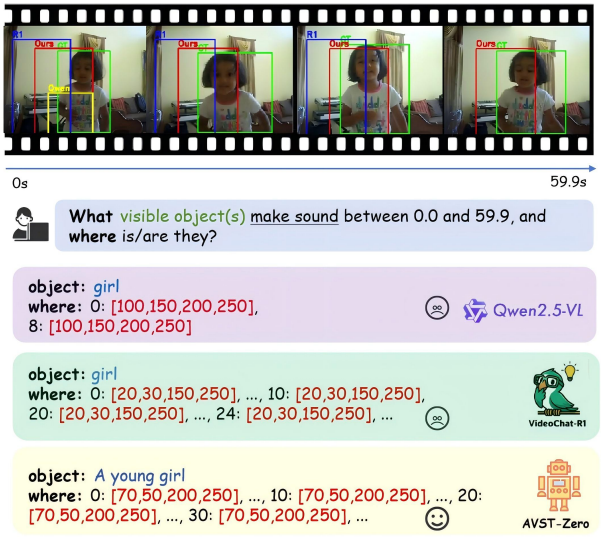

Figure 5: Qualitative results. For the bounding boxes in the video: green denotes the ground truth, blue comes from VideoChat-R1, yellow from Qwen2.5-VL, and red from our AVST-Zero.

Qualitative Results

As shown in Fig. 5, Qwen2.5-VL (Bai et al. 2025) predicts sparse and inaccurate object locations. VideoChat-R1 (Li et al. 2025b) correctly identifies the girl, but our model yields results closer to the ground truth, with more accurate object recognition and spatial localization.

Conclusion

We introduce R-AVST, a video dataset with fine-grained spatio-temporal annotations for complex audio-visual scenarios. Based on this dataset, we define three specialized reasoning tasks with automatically generated QAs. To support these tasks, we develop AVST-Zero, trained with GRPO and task-specific reward functions to improve spatio-temporal reasoning. Experimental results show that R-AVST advances research in audio-visual scenes, with our model demonstrating its effectiveness on these tasks.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China under Grant 2024YFE0203100, and in part by the ZTE Industry-University-Institute Cooperation Funds under Grant No.IA20240906004.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ahmad, G. S.; Heakl, A.; Gani, H.; Shaker, A.; Shen, Z.; Khan, F. S.; and Khan, S. 2025. VideoMolmo: Spatio-Temporal Grounding Meets Pointing. *arXiv preprint arXiv:2506.05336*.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Chen, Y.; Liu, Y.; Wang, H.; Liu, F.; Wang, C.; and Carneiro, G. 2023. A closer look at audio-visual semantic segmentation.

Chen, Y.; Liu, Y.; Wang, H.; Liu, F.; Wang, C.; Frazer, H.; and Carneiro, G. 2024a. Unraveling instance associations: A closer look for audio-visual segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26497–26507.

Chen, Z.; Wang, W.; Cao, Y.; Liu, Y.; Gao, Z.; Cui, E.; Zhu, J.; Ye, S.; Tian, H.; Liu, Z.; et al. 2024b. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Cheng, Z.; Hu, J.; Liu, Z.; Si, C.; Li, W.; and Gong, S. 2025. V-star: Benchmarking video-llms on video spatio-temporal reasoning. *arXiv preprint arXiv:2503.11495*.

Cheng, Z.; Leng, S.; Zhang, H.; Xin, Y.; Li, X.; Chen, G.; Zhu, Y.; Zhang, W.; Luo, Z.; Zhao, D.; et al. 2024. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*.

Chowdhury, S.; Nag, S.; Dasgupta, S.; Chen, J.; Elhoseiny, M.; Gao, R.; and Manocha, D. 2024. Meerkat: Audio-visual large language model for grounding in space and time. In *European Conference on Computer Vision*, 52–70. Springer.

Du, H.; Li, G.; Zhou, C.; Zhang, C.; Zhao, A.; and Hu, D. 2025. Crab: A unified audio-visual scene understanding model with explicit cooperation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 18804–18814.

Feng, K.; Gong, K.; Li, B.; Guo, Z.; Wang, Y.; Peng, T.; Wu, J.; Zhang, X.; Wang, B.; and Yue, X. 2025. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*.

Gao, J.; Sun, C.; Yang, Z.; and Nevatia, R. 2017. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, 5267–5275.

Geng, T.; Wang, T.; Duan, J.; Cong, R.; and Zheng, F. 2023. Dense-localizing audio-visual events in untrimmed videos: A large-scale benchmark and baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22942–22951.

Geng, T.; Wang, T.; Zhang, Y.; Duan, J.; Guan, W.; and Zheng, F. 2024. Uniav: Unified audio-visual perception for multi-task video localization. *CoRR*.

Geng, T.; Zhang, J.; Wang, Q.; Wang, T.; Duan, J.; and Zheng, F. 2025. Longvale: Vision-audio-language-event benchmark towards time-aware omni-modal perception of long videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 18959–18969.

GLM, T.; Zeng, A.; Xu, B.; Wang, B.; Zhang, C.; Yin, D.; Zhang, D.; Rojas, D.; Feng, G.; Zhao, H.; et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Goodge, A.; Ng, W. S.; Hooi, B.; and Ng, S. K. 2025. Spatio-temporal foundation models: Vision, challenges, and opportunities. *arXiv preprint arXiv:2501.09045*.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Li, H.; Chen, J.; Wei, Z.; Huang, S.; Hui, T.; Gao, J.; Wei, X.; and Liu, S. 2025a. Llava-st: A multimodal large language model for fine-grained spatial-temporal understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 8592–8603.

Li, X.; Yan, Z.; Meng, D.; Dong, L.; Zeng, X.; He, Y.; Wang, Y.; Qiao, Y.; Wang, Y.; and Wang, L. 2025b. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *arXiv preprint arXiv:2504.06958*.

Li, Z.; Xu, Q.; Zhang, D.; Song, H.; Cai, Y.; Qi, Q.; Zhou, R.; Pan, J.; Li, Z.; Tu, V.; et al. 2024. Groundinggpt: Language enhanced multi-modal grounding model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 6657–6678.

Liu, C.; Li, P. P.; Yu, Q.; Sheng, H.; Wang, D.; Li, L.; and Yu, X. 2024. Benchmarking audio visual segmentation for long-untrimmed videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22712–22722.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Ravi, N.; Gabeur, V.; Hu, Y.-T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rädle, R.; Rolland, C.; Gustafson, L.; et al. 2024. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024. Deepseekmath:

Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Tang, Y.; Bi, J.; Xu, S.; Song, L.; Liang, S.; Wang, T.; Zhang, D.; An, J.; Lin, J.; Zhu, R.; et al. 2025a. Video understanding with large language models: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*.

Tang, Y.; Shimada, D.; Bi, J.; Feng, M.; Hua, H.; and Xu, C. 2025b. Empowering llms with pseudo-untrimmed videos for audio-visual temporal understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 7293–7301.

Tang, Z.; Liao, Y.; Liu, S.; Li, G.; Jin, X.; Jiang, H.; Yu, Q.; and Xu, D. 2021. Human-centric spatio-temporal video grounding with visual transformers. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(12): 8238–8249.

Tian, Y.; Shi, J.; Li, B.; Duan, Z.; and Xu, C. 2018. Audio-visual event localization in unconstrained videos. In *Proceedings of the European conference on computer vision (ECCV)*, 247–263.

Wang, H.; Xu, Z.; Cheng, Y.; Diao, S.; Zhou, Y.; Cao, Y.; Wang, Q.; Ge, W.; and Huang, L. 2024a. Grounded-videollm: Sharpening fine-grained temporal grounding in video large language models. *arXiv preprint arXiv:2410.03290*.

Wang, J.; Jiang, H.; Liu, Y.; Ma, C.; Zhang, X.; Pan, Y.; Liu, M.; Gu, P.; Xia, S.; Li, W.; et al. 2024b. A comprehensive review of multimodal large language models: Performance and challenges across different tasks. *arXiv preprint arXiv:2408.01319*.

Xu, J.; Guo, Z.; He, J.; Hu, H.; He, T.; Bai, S.; Chen, K.; Wang, J.; Fan, Y.; Dang, K.; et al. 2025. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Yao, J.; Deng, X.; Gu, X.; Dai, M.; Fan, B.; Zhang, Z.; Huang, Y.; Fan, H.; and Zhang, L. 2025. OmniSTVG: Toward Spatio-Temporal Omni-Object Video Grounding. *arXiv preprint arXiv:2503.10500*.

Yuan, L.; Hui, C.; Wu, Y.; Liao, R.; Jiang, F.; and Gao, Y. 2024. Video Enhancement Network Based on CNN and Transformer. *ZTE Communications*, 22(4): 78.

Zhang, B.; Li, K.; Cheng, Z.; Hu, Z.; Yuan, Y.; Chen, G.; Leng, S.; Jiang, Y.; Zhang, H.; Li, X.; et al. 2025. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*.

Zhang, D.; Yu, Y.; Dong, J.; Li, C.; Su, D.; Chu, C.; and Yu, D. 2024. Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601*.

Zhang, Z.; Zhao, Z.; Zhao, Y.; Wang, Q.; Liu, H.; and Gao, L. 2020. Where does it exist: Spatio-temporal video grounding for multi-form sentences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10668–10677.

Zhao, J.; Wei, X.; and Bo, L. 2025. R1-omni: Explainable omni-multimodal emotion recognition with reinforcement learning. *arXiv preprint arXiv:2503.05379*.

Zhong, H.; Zhu, M.; Du, Z.; Huang, Z.; Zhao, C.; Liu, M.; Wang, W.; Chen, H.; and Shen, C. 2025. Omni-R1: Reinforcement Learning for Omnimodal Reasoning via Two-System Collaboration. *arXiv preprint arXiv:2505.20256*.

Zhou, J.; Shen, X.; Wang, J.; Zhang, J.; Sun, W.; Zhang, J.; Birchfield, S.; Guo, D.; Kong, L.; Wang, M.; et al. 2025. Audio-visual segmentation with semantics. *International Journal of Computer Vision*, 133(4): 1644–1664.

Zhou, J.; Wang, J.; Zhang, J.; Sun, W.; Zhang, J.; Birchfield, S.; Guo, D.; Kong, L.; Wang, M.; and Zhong, Y. 2022. Audio–visual segmentation. In *European Conference on Computer Vision*, 386–403. Springer.

Zou, W.; Gu, C.; Fan, J.; Huang, C.; and Bai, Y. 2023. Beyond Video Quality: Evaluation of Spatial Presence in 360-Degree Videos. *ZTE Communications*, 21(4): 91.