Omni-R1: Reinforcement Learning for Omnimodal Reasoning via Two-System Collaboration

Hao Zhong[1,2,*] **Muzhi Zhu**[1,2,*] **Zongze Du**[1,*] **Zheng Huang**[1,2] **Canyu Zhao**[1] **Mingyu Liu**[1] **Wen Wang**[1,2] **Hao Chen**[1] **Chunhua Shen**[1,2,†]

[1] Zhejiang University, China [2] Ant Group

Abstract

Long-horizon video–audio reasoning and fine-grained pixel understanding impose conflicting requirements on omnimodal models: dense temporal coverage demands many low-resolution frames, whereas precise grounding calls for high-resolution inputs. We tackle this trade-off with a *two-system architecture*: a **Global Reasoning System** selects informative keyframes and rewrites the task at low spatial cost, while a **Detail Understanding System** performs pixel-level grounding on the selected high-resolution snippets. Because "optimal" keyframe selection and reformulation are ambiguous and hard to supervise, we formulate them as a reinforcement-learning (RL) problem and present **Omni-R1**, an end-to-end RL framework built on Group Relative Policy Optimization. Omni-R1 trains the Global Reasoning System through hierarchical rewards obtained via online collaboration with the Detail Understanding System, requiring only one epoch of RL on small task splits. Experiments on two challenging benchmarks, Referring Audio-Visual Segmentation (RefAVS) and Reasoning Video Object Segmentation (REVOS), show that Omni-R1 not only surpasses strong supervised baselines but also outperforms specialized state-of-the-art models, while substantially improving out-of-domain generalization and mitigating multimodal hallucination.

Our results demonstrate the first successful application of RL to large-scale omnimodal reasoning and highlight a scalable path toward universally foundation models. Our code is released at: https://github.com/aim-uofa/Omni-R1.

1 Introduction

Enabling models to simultaneously perceive, understand, and reason over omnimodal inputs—such as text, video, and audio—in complex real-world scenarios remains a longstanding goal in artificial intelligence [1, 2, 3, 4]. Recent advances in omnimodal pretraining and instruction fine-tuning have led to the emergence of omnimodal models, bringing us closer to this objective [5, 6, 7]. Despite recent progress, current omnimodal models exhibit notable limitations in two key areas: (1) long-horizon reasoning over complex temporal sequences in video and audio, and (2) fine-grained spatial understanding at the pixel level. A fundamental challenge underlying these two problems is the inherent trade-off between temporal coverage and spatial resolution . Long-horizon reasoning [8, 9, 10] over video and audio typically requires high frame rates to capture global temporal context, which significantly increases memory and computational overhead—often forcing models to operate on low-resolution frames. Conversely, fine-grained pixel understanding [11, 12, 13, 14, 15] demands high-resolution inputs to preserve visual details, which in turn limits the number of frames that can

[*] Equal contribution.

[†] C. Shen is the corresponding author. This work was in part supported by The Pioneer R&D Program of Zhejiang (Grant No. 2025C01011).

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

Figure 1: Overview of the proposed **Omni-R1** system for collaborative video understanding. Left: Performance comparison across multiple benchmarks shows Omni-R1 significantly outperforms existing omni-modal and video-reinforced MLLMs on both segmentation-centric and reasoning-centric tasks. Top-right: Omni-R1 employs a two-stage collaborative framework, integrating a detail understanding system (for precise visual QA) and a global reasoning system (for temporal grounding and high-resolution key frame identification). Bottom: A qualitative example highlights Omni-R1's precise spatial-temporal segmentation and reasoning in identifying object-centric actions, outperforming prior expert models (e.g., Sa2VA) in complex scenarios.

be processed. This trade-off creates a tension between global context modeling and local detail preservation, making it difficult for existing models to excel at both simultaneously.

A natural way to address this trade-off is to decompose the problem into two stages. Accordingly, we frame our solution as a two-system architecture:

- System 1 (**Global Reasoning System**) performs coarse-grained, global reasoning over long video sequences at low spatial resolution—acting as a fast, context-aware selector that identifies critical temporal segments.
- System 2 (**Detail Understanding System**), in contrast, conducts detailed, high-resolution analysis over a small number of keyframes, focusing on precise grounding and fine-grained understanding.

To illustrate how these systems interact, consider a task where the goal is to segment the last person to disappear (*or* make a sound) in a scene. **System1** first processes the full video (*with* audio) sequence

to determine, through low-resolution multimodal abstraction, which person is the last to leave visually or to emit sound. It then selects a few key segments where this individual appears or speaks. Since **System 2** operates only on short segments with high-resolution input and lacks access to long-range temporal or auditory context, **System 1** needs to reformulate the original reference task—initially requiring long-horizon multimodal reasoning—into a simpler, localized problem. This reformulated task focuses on attributes, identity cues, and object permanence within the selected key segments, making it solvable using only fine-grained visual information. **System 2** then takes these key segments and performs fine-grained visual grounding directly on the high-resolution input, bypassing the need for global reasoning. This two-system design enables scalable and efficient multimodal reasoning by eliminating the need to process entire videos at high resolution, and effectively addresses the dual challenge of long-horizon reasoning and fine-grained visual understanding.

It is worth noting that current multimodal models already perform well as **Detail Understanding Systems** in tasks such as visual grounding [12, 16], OCR [2, 17, 18] and fine-grained image understanding [19, 20] on high-resolution inputs. Given this progress, the bottleneck in our two-system framework lies primarily in the capabilities of **System 1**. In this work, we therefore focus on improving **Global Reasoning System**, particularly its ability to select informative keyframes and reformulate the task. However, defining what constitutes an "optimal" keyframe selection or task reformulation is inherently ambiguous and task-dependent, making it impractical to rely on manually curated SFT data. To address this, we propose **Omni-R1**, an end-to-end reinforcement learning framework tailored for omnimodal reasoning. Built upon the Group Relative Policy Optimization (GRPO) [21, 22] algorithm, our method simulates online collaboration between **System 1** and **System 2**, applying policy gradient updates guided by a hierarchical reward framework to progressively train System 1 to select keyframes and reformulate tasks in long-horizon, omnimodal settings.

From a reinforcement learning (RL) perspective, although it has proven effective in enhancing reasoning within large language models [21, 22, 23], RL remains underexplored in omnimodal settings. One major challenge lies in the lack of effective multimodal reasoning data [24], along with uncertainty about whether language-based RL techniques can generalize across modalities. While Omni-R1 bridges this gap by reformulating long-horizon multimodal understanding as a collaborative process between two systems. In our design, the Global Reasoning System functions as an RL agent that selects keyframes and reformulates tasks for the Detail Understanding System to complete. Such an approach provides a scalable path toward improving temporal reasoning and summarization in omnimodal models, while also opening new opportunities for applying RL beyond purely linguistic tasks.

To validate the effectiveness of Omni-R1, we benchmark it on two especially demanding tasks, namely Referring Audio-Visual Segmentation (RefAVS [25]) and Reasoning Video Object Segmentation (REVOS [26]), both of which require temporal reasoning over video(audio) streams and fine-grained pixel understanding. Training Omni-R1 for just one epoch on the small datasets already lifts performance well beyond our baseline model and even surpasses the strongest, highly specialized state-of-the-art models on each benchmark. Even more striking, reinforcement learning improves out-of-domain generalization, whereas conventional supervised fine-tuning often weakens it. Omni-R1 achieves higher scores in both pure video-understanding and omnimodal understanding settings, outperforming recent RL methods tailored specifically to video-reasoning tasks. Finally, we conduct a comprehensive suite of diagnostic studies—including ablations over key architectural and training choices and an analysis of RL's impact on multimodal hallucination—which together highlight the versatility and reliability of our approach. We hope that Omni-R1 offers a new direction for applying reinforcement learning to future all-modality foundation models.

Our primary contributions are summarized as follows:

- We present a scalable **Global Reasoning**, and **Detail Understanding** two-system architecture that separates long-horizon video–audio reasoning from fine-grained pixel-level grounding, effectively resolving the temporal–spatial trade-off that constrains existing omnimodal models.
- We introduce an end-to-end reinforcement-learning framework **Omni-R1**, built on Group Relative Policy Optimization that trains **System 1** via hierarchical rewards and simulated collaboration with **System 2** to select keyframes and reformulate tasks in long-horizon omnimodal settings.

- With one epoch RL training, **Omni-R1** surpasses strong supervised baselines and specialized SOTA methods on RefAVS and REVOS, while markedly improving out-of-domain generalization including video understanding and omnimodal understanding.

2 Related Work

2.1 Omni-modal Large Models

The advent of Large Language Models (LLMs) has revolutionized artificial intelligence, showcasing unprecedented capabilities in understanding, generating, and reasoning with textual data [27, 28, 29, 30]. Building upon this foundation, Multimodal Large Language Models (MLLMs) have emerged, integrating multiple data modalities—such as vision, language, and audio—to achieve a more holistic understanding of complex tasks [12, 31, 32, 33, 34].

To differentiate from vision-language models (VLMs), multimodal large language models (MLLMs) incorporating the audio modality, such as Qwen2.5-Omni [35], are termed *omni-modal models*, abbreviated as *omni*. MiniCPM-o 2.6 [2] extends its vision-language foundation [2] with audio processing capabilities, allowing it to operate across more modalities. Baichuan-Omni-1.5 [1], trained and inferred in a fully end-to-end manner, surpasses GPT-4o-mini on the full-modality leaderboard OmniBench [36]. The recent development of omni-modal models further extends this integration, encompassing visual, linguistic, and auditory modalities to approach a comprehensive multimodal understanding [7].

2.2 MLLM with RL

Despite the remarkable progress enabled by supervised learning and instruction tuning, key challenges persist in aligning MLLMs with human preferences, mitigating harmful outputs, and enhancing their performance on complex reasoning tasks. Reinforcement Learning (RL), particularly Reinforcement Learning from Human Feedback (RLHF) [23], has proven effective in addressing these issues within unimodal LLMs, contributing to the success of models like ChatGPT [27]. A notable advancement in this domain is the introduction of DeepSeek-R1 [21], which employs Group Relative Policy Optimization (GRPO) to enhance reasoning capabilities. GRPO innovatively replaces traditional critic models with a group-based reward normalization approach, reducing computational costs while maintaining performance [22]. This technique has demonstrated that pure RL can effectively develop strong reasoning abilities without reliance on supervised data.

While reinforcement learning techniques have been widely explored in LLMs, their application to MLLMs is still at an early stage. Most recent efforts [37, 38, 39, 40, 41, 42] have primarily focused on vision and language modalities, with little attention paid to more comprehensive multimodal integration. Notably, a concurrent work, R1-Omni [43], is the first to include audio in addition to vision and language; however, its focus is limited to a single motion recognition task.

In contrast, our work targets more general long-horizon understanding tasks and conducts a more comprehensive and systematic investigation of omni-modal reinforcement learning. Building on recent advances, we propose **Omni-R1**, *an omni-modal framework that unifies vision, language, and audio processing under an end-to-end RL optimization pipeline*. Our two-system design, which separates temporal reasoning from spatial perception, enables enhanced long-horizon understanding and fine-grained attention, allowing Omni-R1 to better address complex multimodal tasks requiring both structured perception and dynamic decision-making.

3 Omni-R1

3.1 Task and System Formulation

We consider a long-horizon multimodal understanding task, where the model receives a video sequence $V = \{v_1, v_2, \ldots, v_T\}$ and a synchronized audio stream $A = \{a_1, a_2, \ldots, a_T\}$, along with an instruction or query q. The goal is to produce a task-specific output y (e.g., a localized segment, a textual response, or a grounding prediction) that reflects both global temporal reasoning and fine-grained visual understanding. To better facilitate global temporal reasoning, we transform the raw instruction q into a high-level instruction $q_{\text{global}} = \mathcal{T}(q)$ via a template-based rewriting function $\mathcal{T}$.

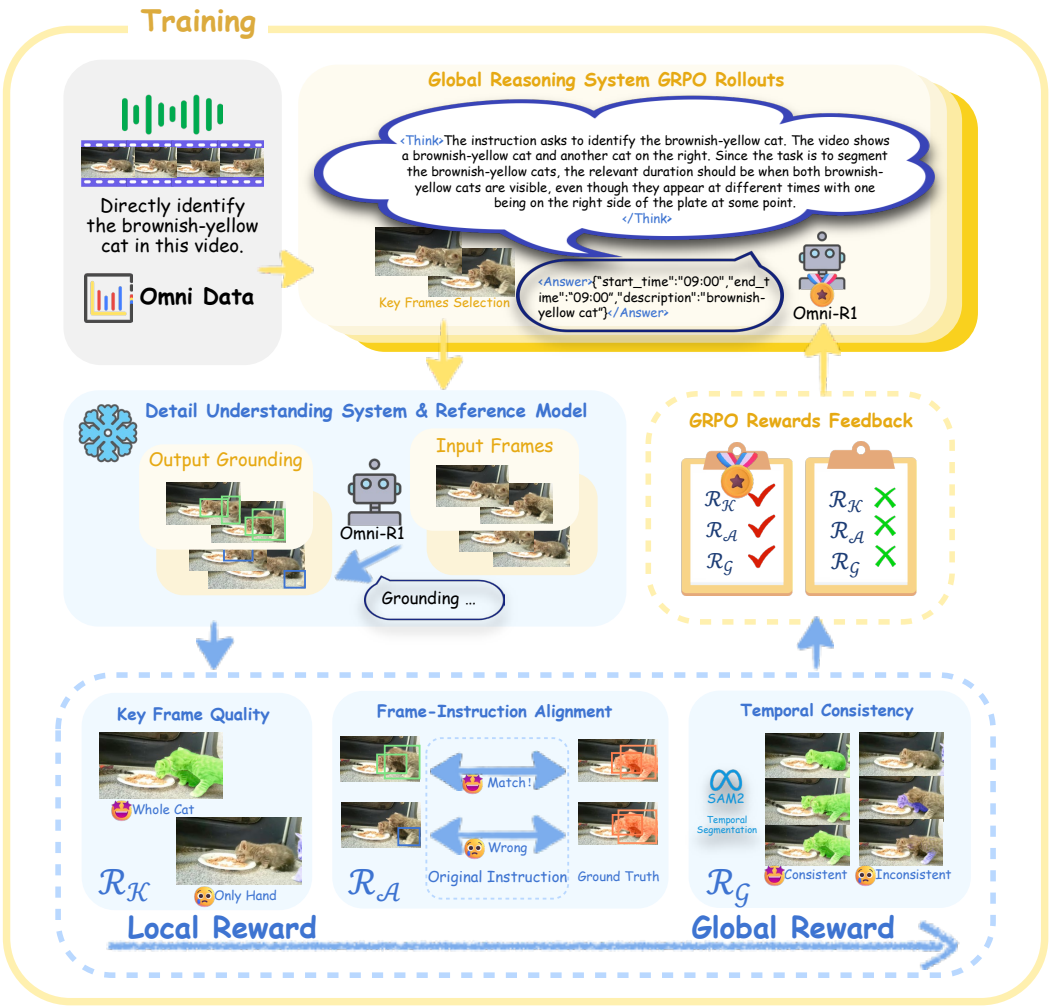


Figure 2: Exclusively trained as System 1 on video segmentation tasks in an End-to-End RL pipeline, Omni-R1 improved general understanding capabilities.

Stage 1: Global Reasoning System. We reduce the spatial resolution of the video as commonly adopted to obtain a low-resolution stream $\tilde{V} = \{\tilde{v}_1, \ldots, \tilde{v}_T\}$ suitable for efficient global processing. Given $(\tilde{V}, A, q_{\text{global}})$, System 1 produces a set of K selected segments(frames) $\mathcal{S} = \{s_1, s_2, \ldots, s_K\}$ and a corresponding set of local queries $\{q_{\text{local}}^{(i)}\}_{i=1}^{K}$, intended to simplify the reasoning objective for System 2:

$$\mathcal{S}, \{q_{\text{local}}^{(i)}\}_{i=1}^{K} = \pi^{(\text{S1})}(\tilde{V}, A, q_{\text{global}})$$

Stage 2: Detail Understanding System. System 2 then receives the high-resolution frames $V_{\mathcal{S}} = \{v_{s_1}, \ldots, v_{s_K}\}$ corresponding to the segments selected by System 1. Given $(V_{\mathcal{S}}, \{q_{\text{local}}^{(i)}\}_{i=1}^{K})$, it performs fine-grained multimodal reasoning and produces the final output:

$$y = \pi^{(\text{S2})}(V_{\mathcal{S}}, \{q_{\text{local}}^{(i)}\}_{i=1}^{K})$$

For tasks such as RefAVS and RVOS, one possible instantiation of System 2 is as a combination of a per-frame grounding model $\mathcal{F}_{\text{grounding}}$ and a frozen video segmentation model $\mathcal{F}_{\text{seg}}$ (e.g., SAM2 [44]). Given the selected high-resolution frames $V_{\mathcal{S}} = \{v_{s_1}, \ldots, v_{s_K}\}$ and the corresponding local instructions $\{q_{\text{local}}^{(i)}\}_{i=1}^{K}$, the grounding model is applied independently to each pair $(v_{s_i}, q_{\text{local}}^{(i)})$ to predict a set of bounding boxes $\mathcal{B}_{s_i} = \{b_1^{(i)}, \ldots, b_{N_i}^{(i)}\}$, where each $b_j^{(i)} \in \mathbb{R}^4$ denotes a box in

(x_1, y_1, x_2, y_2) format. These predicted boxes are then passed to the segmentation model to produce pixel-level instance masks and propagate them temporally across the entire video:

$$\hat{\mathcal{M}} = \mathcal{F}_{\text{seg}}(V, V_{\mathcal{S}}, \{\mathcal{B}_{s_i}\}_{i=1}^{K})$$

where the final output $\hat{\mathcal{M}} = \{\hat{m}_1, \ldots, \hat{m}_T\}$ is a sequence of temporally aligned masks. The corresponding ground-truth mask sequence is denoted as $\mathcal{M}^* = \{m_1^*, \ldots, m_T^*\}$, where each m_t^* is the binary segmentation mask for frame v_t. For more details on the segmentation model $\mathcal{F}_{\text{seg}}$, please refer to Section B in Appendix.

3.2 End-to-End Reinforcement Learning via GRPO

We now turn our focus to optimizing System 1 $\pi^{(\text{S1})}$, with the goal of improving both the selection of key segments $\mathcal{S}$ and the formulation of task-specific local instructions $\{q_{\text{local}}^{(i)}\}$, in order to better support System 2 in performing fine-grained understanding. However, $\mathcal{S}$, $\{q_{\text{local}}^{(i)}\}$, and the System 2 (i.e., $\pi^{(\text{S2})}$) are strongly coupled, making it difficult to directly define what constitutes an optimal pair $(\mathcal{S}, \{q_{\text{local}}^{(i)}\})$ for downstream performance. As a result, constructing high-quality supervised fine-tuning (SFT) data for $\pi^{(\text{S1})}$ is infeasible.

Instead, we propose to optimize $\pi^{(\text{S1})}$ via reinforcement learning by designing a reward function $R(\mathcal{S}, \{q_{\text{local}}^{(i)}\}, \pi^{(\text{S2})})$ that evaluates the utility of System 1's outputs in enabling System 2 to succeed. Under this framework, $\pi^{(\text{S1})}$ is trained to explore and generate candidate outputs, and receives feedback from the environment through this reward. Specifically, we adopt a GRPO-based policy optimization scheme. At each iteration, we sample N responses from the current policy $\pi^{(\text{S1})}$ and compute the corresponding rewards r_n using the reward function $R(\cdot)$. We then normalize the rewards to estimate the advantage of each sample:

$$A_n = \frac{r_n - \text{mean}(\{r_1, \ldots, r_N\})}{\text{std}(\{r_1, \ldots, r_N\})} \tag{1}$$

Based on the computed advantages $\{A_n\}$, we perform PPO-style policy gradient updates to improve $\pi^{(\text{S1})}$.

3.3 Hierarchical Reward Design for System 1

Designing an effective reward function that accurately reflects the quality of the action pair $(\mathcal{S}, \{q_{\text{local}}^{(i)}\})$ and provides a meaningful training signal for System 1 ($\pi^{(\text{S1})}$) is critical to the success of our framework. In this section, we describe our hierarchical reward formulation tailored for the Referring Video Object Segmentation (RVOS) task, which aims to guide System 1 to progressively learn to select informative keyframes and generate useful local instructions.

Due to the strong coupling among $\mathcal{S}$, $\{q_{\text{local}}^{(i)}\}$, and System 2 ($\pi^{(\text{S2})}$), relying solely on the final task objective (e.g., segmentation mIoU) as the reward leads to unstable and inefficient training. This is because such reward signals are sparse, non-decomposable, and difficult to attribute back to specific decisions made by $\pi^{(\text{S1})}$. To address this, we propose a set of hierarchical reward functions, organized from weakly coupled to strongly coupled, and from local to global. These rewards are designed to incrementally shape the learning of System 1, starting from simpler supervision signals and gradually incorporating a more task-specific structure. We define three types of rewards:

Key Frame Quality Reward ($R_{\mathcal{K}}$): This reward evaluates the quality of the selected keyframes $\mathcal{S}$, independently of the instructions or the performance of subsequent segmentation.

It provides early learning signals to encourage the selection of visually salient or semantically diverse frames.

We define the Key Frame Quality Reward as a weighted combination of three factors:

$$R_{\mathcal{K}} = \lambda_1 R_{\text{diversity}} + \lambda_2 R_{\text{num}} + \lambda_3 R_{\text{saliency}}$$

The first term, *Temporal Diversity Reward* $R_{\text{diversity}}$, encourages selected frames to spread over the video timeline, rather than being clustered within a short segment. This promotes broader temporal coverage and helps the model focus on long-range dynamics.

The second term, *Frame Count Regularization* R_{num}, regularizes the number of selected frames K to stay near a predefined target K_0, penalizes selections that include either too few or too many frames.

The third term, *Object-Centric Saliency Reward* R_{saliency}, rewards keyframes that contain a large visible portion of the target object. This is based on the hypothesis that selecting such frames provides stronger visual anchors, which can facilitate more accurate and stable object tracking and segmentation throughout the video. It is calculated as the normalized average GT mask area:

$$R_{\text{saliency}} = \frac{1}{K} \sum_{i=1}^{K} \frac{\text{area}(m^*_{s_i})}{\max_t \text{area}(m^*_t)}$$

Together, these components guide System 1 to select keyframes that are temporally diverse, reasonably sparse, and visually informative. Formal definitions of the reward are provided in Appendix Section B.

Frame-Instruction Alignment Reward ($R_{\mathcal{A}}$) measures how well each local instruction $q^{(i)}_{\text{local}}$ aligns with its corresponding keyframe v_{s_i}. This reward evaluates whether the instruction provides sufficient grounding cues to locate the correct object in the frame. As it operates independently per frame-instruction pair, it does not depend on the segmentation model $\mathcal{F}_{\text{seg}}$, and thus ignores temporal consistency. Concretely, given a frame v_{s_i} and its corresponding instruction $q^{(i)}_{\text{local}}$, we apply the grounding model $\mathcal{F}_{\text{grounding}}$ to predict a set of bounding boxes $\mathcal{B}_{s_i} = \{b^{(i)}_1, \ldots, b^{(i)}_{N_i}\}$. We compare these predictions against the ground-truth target boxes $\mathcal{B}^*_{s_i}$ defined for that frame. Since a single instruction may refer to multiple target objects, both $\mathcal{B}_{s_i}$ and $\mathcal{B}^*_{s_i}$ can contain multiple instances.The reward is computed as the negative Hungarian matching loss commonly used in object detection [45]:

$$R_{\mathcal{A}} = \frac{1}{K} \sum_{i=1}^{K} \left(1 - \mathcal{L}_{\text{Hungarian}}(\mathcal{B}_{s_i}, \mathcal{B}^*_{s_i})\right) \tag{2}$$

This loss is minimized when the predicted boxes exactly match the ground-truth targets.

Global Temporal Consistency Reward ($R_{\mathcal{G}}$) is the most strongly coupled and task-specific reward in our framework, directly reflecting the final objective of long-term video object segmentation. Unlike previous rewards, which evaluate the selected keyframes or instructions in isolation, $R_{\mathcal{G}}$ jointly considers how the selected keyframes $\mathcal{S}$ and local instructions $\{q^{(i)}_{\text{local}}\}$ influence the performance of System 2 throughout the video. This reward is designed to capture both the spatial accuracy and the temporal consistency of the predicted instance masks. In particular, it encourages System 1 to select frames that are critical for robust tracking—such as those appearing after significant object deformations, occlusions, or disappearances—so that the segmentation model (e.g., SAM2) can re-anchor to the target effectively. Formally, given a candidate keyframe set $\mathcal{S}$ and corresponding instructions $\{q^{(i)}_{\text{local}}\}$, we feed them into System 2 ($\pi^{(\text{S2})}$) to obtain a full sequence of predicted masks $\hat{\mathcal{M}} = \{\hat{m}_1, \ldots, \hat{m}_T\}$.

The reward is computed as the average frame-wise Intersection over Union (IoU) with the ground-truth masks $\mathcal{M}^* = \{m^*_1, \ldots, m^*_T\}$:

$$R_{\mathcal{G}} = \frac{1}{T} \sum_{t=1}^{T} \text{IoU}(\hat{m}_t, m^*_t) \tag{3}$$

Finally, we combine the above three components to form the overall reward used for training System 1. The total reward is a weighted sum of the three terms:

$$R = \alpha_{\mathcal{K}} R_{\mathcal{K}} + \alpha_{\mathcal{A}} R_{\mathcal{A}} + \alpha_{\mathcal{G}} R_{\mathcal{G}} \tag{4}$$

where $\alpha_{\mathcal{K},\mathcal{A},\mathcal{R}}$ are the weighting coefficients that control the importance of each reward component.

Table 1: Performance comparison across models grouped by Seen and Unseen sets in Ref-AVSBench [25]. Some metrics curated from [25]. $\mathcal{J}\&\mathcal{F}$ represents the average of ($\mathcal{J}$) score and ($\mathcal{F}$) score. † indicates the results are tested on the masks predicted by SAM2 according to model's grounding output.

Model	Seen			Unseen		
	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
AVSBench [49] + `text`	37.2	23.2	51.1	43.5	32.4	54.7
AVSegFormer [50] + `text`	40.2	33.5	47.0	43.1	36.1	50.1
GAVS [51] + `text`	39.4	28.9	49.8	39.8	29.8	49.7
ReferFormer [52] + `audio`	40.7	31.3	50.1	39.6	30.4	48.8
R2VOS [53] + `audio`	33.0	25.0	41.0	38.9	27.9	49.8
EEMC [25]	42.8	34.2	51.3	57.2	49.5	64.8
Qwen2.5-Omni-7B†	31.6	27.7	35.5	62.3	59.0	65.7
Qwen2.5-Omni-7B†(SFT)	39.1	35.4	42.8	66.2	63.1	69.3
Omni-R1-7B†	**47.2**	**43.0**	**51.4**	**74.2**	**71.3**	**77.0**
Δ	**+8.1**	**+7.6**	**+8.6**	**+8.0**	**+8.2**	**+7.7**

4 Experiments

4.1 Experiments Setting

System 1 and System 2 We adopt **Qwen2.5-Omni-7B** [7] as our base model, which serves as **System 1** responsible for high-level reasoning. To construct a lightweight and stable **System 2** during training, we use a frozen copy of the same pretrained Qwen2.5-Omni model, which also functions as a reference policy model for guiding optimization. For evaluation, unless otherwise stated, **Omni-R1** is serving as both System 1 and System 2 for resource efficiency. However, due to the modular design and decoupled functionality of the two systems, System 2 can be flexibly replaced with a stronger perception module in a **zero-shot** manner.

Training Paradigm We train System 1 on 1,600 samples randomly selected from the RefAVS [25] dataset and 2,600 videos from the ReVOS [26] and MeViS [46] datasets for 1 epoch. To further enhance the model's fine-grained understanding capabilities as system 2, we additionally train the model on 2,000 images from refCOCOg [47] for one epoch in the style of SegZero [48]. Unless otherwise specified, all experiments are conducted using a policy KL divergence hyperparameter of $\beta = 0.04$, a group size of 8, and an initial learning rate of 1×10^{-6} under the AdamW optimizer with a weight decay of 0.01. We adopt `sam2-hiera-large` as our SAM2 [44] version throughout the experiments.

4.2 Referring Video Segmentation

Referring Audio-Visual Segmentation Ref-AVS [25] is specifically designed for audio-visual segmentation tasks, offering a diverse and well-annotated collection of samples that require integrated reasoning across both modalities. The dataset comprises 2,908 audio-equipped video clips in the training set, covering 5,366 annotated objects across 39 semantic categories.

We evaluated the performance of our collaborative system on Ref-AVSBench [25] with other Referring AVS methods. Omni-R1 outperforms previous SOTA EMMC [25] by **+4.4%** on $\mathcal{J}\&\mathcal{F}$ in seen set and **+17.0%** on unseen set.

Reasoning Video Object Segmentation ReVOS [26] is a VOS dataset that emphasizes reasoning about temporal behaviors through implicit object descriptions, comprising 35,074 pairs of instruction-mask sequences derived from 1,042 diverse videos. In contrast to traditional referring video segmentation datasets, ReVOS includes text instructions that necessitate a sophisticated understanding of both video content and general world knowledge.

For our evaluation, we exclusively employed Sa2VA as **System 2** to investigate the full reasoning capabilities of Omni-R1 as **System 1**.

Table 2: Reasoning Video Object Segmentation performance comparison across different methods, the metric is $\mathcal{J}\&\mathcal{F}$ score(%). ‡ means the results are evaluated where Omni-R1-7B serves as System 1 and Sa2VA as System 2(1B and 4B).

Model	ReVOS				
	Referring	*Reasoning*	*Single*	*Multi*	***Overall***
LISA-13B [11]	-	-	-	-	41.6
TrackGPT-13B [54]	-	-	-	-	45.0
VISA-13B [26]	-	-	-	-	50.9
Sa2VA-8B [55]	-	-	-	-	57.6
Sa2VA-26B [55]	-	-	-	-	58.4
Qwen2.5-Omni-7B†	46.3	26.9	38.6	37.4	36.6
Omni-R1-7B†	52.5	36.9	45.0	46.6	44.7
Omni-R1-8B‡	61.6	50.7	56.6	47.3	56.2
Omni-R1-11B‡	**64.1**	**53.7**	**59.2**	**51.0**	**58.9**

Our **System 1** exhibits strong performance on video object segmentation tasks under both basic and reasoning-intensive conditions. When deployed as both systems (†), Omni-R1-7B significantly outperforms the base model on ReVOS, achieving a **+8.1%** improvement over Qwen2.5-Omni-7B. This result underscores its enhanced temporal reasoning and fine-grained recaption capabilities.

Furthermore, the collaborative system (‡) Omni-R1-11B achieves a score of **58.9%** on ReVOS, surpassing much larger segmentation-specialized models such as Sa2VA-26B [55]. Notably, it achieves the best performance across all categories, including the reasoning subset in ReVOS (53.7%), underscoring the effectiveness of our disentangled system architecture and reinforcement learning-based training paradigm.

4.3 General Omni-Modal Understanding

In this section, we focus on the impressive progress of Omni-R1 on multi-modal tasks, in comparison to its base model Qwen2.5-Omni-7B and other leading multi-modal models.

Omni-R1 shows stable improvements over its base model Qwen2.5-Omni. Omni-R1 achieves an average improvement of **+2.0%**, **+2.7%** and **+3.7%** over baseline on OmniBench [36], VideoMME [59]

Table 3: Performance comparison across models on general understanding QA benchmarks Omnibench, VideoMME, and MVBench.

Method	Omnibench	VideoMME		MVBench
	Avg	General	Short	General
Vision-Language Models				
Qwen2.5-VL-7B(CoT)	-	56.1	71.3	57.4
LLaVA-OneVision-7B [56]	-	58.2	-	56.7
Kangeroo-8B [57]	-	56.0	-	61.1
VideoChat-R1 [40]	-	-	72.2	67.9
Video-R1 [39]	-	59.3	-	63.9
Sa2VA-26B [55]	-	52.6	-	-
Omni-Modal Language Models				
VITA-1.5-7B [58]	33.4	57.3	-	55.5
MiniCPM-o 2.6-7B [2]	40.5	63.4	-	58.6
Baichuan-Omni-1.5-7B [1]	42.9	60.1	-	63.7
Qwen2.5-Omni-7B	47.3	58.3	69.8	66.1
Omni-R1-7B	**49.3**	**60.7**	**73.0**	**70.3**
Δ	**+2.0**	**+2.4**	**+3.2**	**+4.2**

and MVBench [60] respectively, surpassing all other open-source omni-models. Specifically, in video understanding tasks, Omni-R1 gains more progress on short videos (less than 2 min) than long videos. This could be attributed to our VOS training videos, where almost all videos are less than 2 minutes, with MeViS even being less than 30 seconds.

System 1's Strength in General Understanding Tasks Omni-R1 demonstrates significant improvements, achieving outstanding general performance on the omni-modal benchmark OmniBench, where it outperforms all other 7B models in the open-source space. With a score of **73.0** in the short subset of VideoMME, Omni-R1 surpasses VideoChat-R1 [40], which was exclusively fine-tuned for Video QA tasks through RL. Additionally, Omni-R1 achieves the highest score on MVBench, outperforming all other omni-modal models by a large margin.

The strong performance of Omni-R1 across both in-domain and general tasks showcases the effectiveness of our reinforcement learning approach. Leveraging System 1's multi-modal reasoning, the model excels in task-specific scenarios and generalizes effectively to unseen tasks, demonstrating its robustness and adaptability in real-world environments.

5 Conclusion

We present Omni-R1, a novel reinforcement learning framework that addresses a key limitation in omnimodal models: the trade-off between long-horizon temporal reasoning and fine-grained spatial understanding. By decoupling these objectives into a two-system architecture comprising a Global Reasoning System and a Detail Understanding System Omni-R1 enables scalable and efficient processing of complex video–audio–text inputs.

Through task reformulation and keyframe selection trained via Group Relative Policy Optimization, our approach significantly improves performance on challenging benchmarks like RefAVS and ReVOS, while also enhancing out-of-domain generalization. Our diagnostic studies further confirm the robustness and versatility of the framework. We hope that this work opens new avenues for integrating reinforcement learning into next-generation omnimodal foundation models.

References

[1] Yadong Li, Jun Liu, Tao Zhang, Song Chen, Tianpeng Li, Zehuan Li, Lijun Liu, Lingfeng Ming, Guosheng Dong, Da Pan, et al. Baichuan-omni-1.5 technical report. *arXiv preprint arXiv:2501.15368*, 2025. 1, 4, 9

[2] 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. 1, 3, 4, 9

[3] Chiori Hori, Takaaki Hori, Teng-Yok Lee, Ziming Zhang, Bret Harsham, John R Hershey, Tim K Marks, and Kazuhiko Sumi. Attention-based multimodal fusion for video description. In *Proceedings of the IEEE international conference on computer vision*, pages 4193–4202, 2017. 1

[4] 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. 1

[5] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 1

[6] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 1

[7] Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025. 1, 4, 8

[8] Heqing Zou, Tianze Luo, Guiyang Xie, Fengmao Lv, Guangcong Wang, Junyang Chen, Zhuochen Wang, Hansheng Zhang, Huaijian Zhang, et al. From seconds to hours: Reviewing multimodal large language models on comprehensive long video understanding. *arXiv preprint arXiv:2409.18938*, 2024. 1

[9] Yang Shi, Jiaheng Liu, Yushuo Guan, Zhenhua Wu, Yuanxing Zhang, Zihao Wang, Weihong Lin, Jingyun Hua, Zekun Wang, Xinlong Chen, et al. Mavors: Multi-granularity video representation for multimodal large language model. *arXiv preprint arXiv:2504.10068*, 2025. 1

[10] Zheng Huang, Mingyu Liu, Xiaoyi Lin, Muzhi Zhu, Canyu Zhao, Zongze Du, Xiaoman Li, Yiduo Jia, Hao Zhong, Hao Chen, et al. Notvla: Narrowing of dense action trajectories for generalizable robot manipulation. *arXiv preprint arXiv:2510.03895*, 2025. 1

[11] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589, 2024. 1, 9

[12] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1, 3, 4

[13] Muzhi Zhu, Hengtao Li, Hao Chen, Chengxiang Fan, Weian Mao, Chenchen Jing, Yifan Liu, and Chunhua Shen. Segprompt: Boosting open-world segmentation via category-level prompt learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 999–1008, 2023. 1

[14] Muzhi Zhu, Chengxiang Fan, Hao Chen, Yang Liu, Weian Mao, Xiaogang Xu, and Chunhua Shen. Generative active learning for long-tailed instance segmentation. *arXiv preprint arXiv:2406.02435*, 2024. 1

[15] Muzhi Zhu, Yang Liu, Zekai Luo, Chenchen Jing, Hao Chen, Guangkai Xu, Xinlong Wang, and Chunhua Shen. Unleashing the potential of the diffusion model in few-shot semantic segmentation. *Advances in Neural Information Processing Systems*, 37:42672–42695, 2024. 1

[16] 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. 3

[17] Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*, 2023. 3

[18] Yuliang Liu, Biao Yang, Qiang Liu, Zhang Li, Zhiyin Ma, Shuo Zhang, and Xiang Bai. Textmonkey: An ocr-free large multimodal model for understanding document. *arXiv preprint arXiv:2403.04473*, 2024. 3

[19] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13009–13018, 2024. 3

[20] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4818–4829, 2024. 3

[21] 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. 3, 4

[22] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 3, 4

[23] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. 3, 4

[24] Zehan Wang, Ziang Zhang, Hang Zhang, Luping Liu, Rongjie Huang, Xize Cheng, Hengshuang Zhao, and Zhou Zhao. Omnibind: Large-scale omni multimodal representation via binding spaces. *arXiv preprint arXiv:2407.11895*, 2024. 3

[25] Yaoting Wang, Peiwen Sun, Dongzhan Zhou, Guangyao Li, Honggang Zhang, and Di Hu. Ref-avs: Refer and segment objects in audio-visual scenes. In *European Conference on Computer Vision*, pages 196–213. Springer, 2024. 3, 8

[26] Cilin Yan, Haochen Wang, Shilin Yan, Xiaolong Jiang, Yao Hu, Guoliang Kang, Weidi Xie, and Efstratios Gavves. Visa: Reasoning video object segmentation via large language models. *arXiv preprint arXiv:2407.11325*, 2024. 3, 8, 9

[27] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 4

[28] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 4

[29] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024. 4

[30] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 4

[31] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 4

[32] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 4

[33] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 4

[34] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. Deepseek-vl: Towards real-world vision-language understanding, 2024. 4

[35] Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025. 4

[36] Yizhi Li, Ge Zhang, Yinghao Ma, Ruibin Yuan, Kang Zhu, Hangyu Guo, Yiming Liang, Jiaheng Liu, Jian Yang, Siwei Wu, Xingwei Qu, Jinjie Shi, Xinyue Zhang, Zhenzhu Yang, Xiangzhou Wang, Zhaoxiang Zhang, Zachary Liu, Emmanouil Benetos, Wenhao Huang, and Chenghua Lin. Omnibench: Towards the future of universal omni-language models, 2024. 4, 9

[37] Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*, 2025. 4

[38] Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, et al. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*, 2025. 4

[39] Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*, 2025. 4, 9

[40] Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *arXiv preprint arXiv:2504.06958*, 2025. 4, 9, 10

[41] Muzhi Zhu, Yuzhuo Tian, Hao Chen, Chunluan Zhou, Qingpei Guo, Yang Liu, Ming Yang, and Chunhua Shen. Segagent: Exploring pixel understanding capabilities in mllms by imitating human annotator trajectories. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3686–3696, 2025. 4

[42] Muzhi Zhu, Hao Zhong, Canyu Zhao, Zongze Du, Zheng Huang, Mingyu Liu, Hao Chen, Cheng Zou, Jingdong Chen, Ming Yang, et al. Active-o3: Empowering multimodal large language models with active perception via grpo. *arXiv preprint arXiv:2505.21457*, 2025. 4

[43] Jiaxing Zhao, Xihan Wei, and Liefeng Bo. R1-omni: Explainable omni-multimodal emotion recognition with reinforcement learning, 2025. 4

[44] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 5, 8

[45] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 7, 24

[46] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. MeViS: A large-scale benchmark for video segmentation with motion expressions. In *ICCV*, 2023. 8

[47] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. 2016. 8

[48] Yuqi Liu, Bohao Peng, Zhisheng Zhong, Zihao Yue, Fanbin Lu, Bei Yu, and Jiaya Jia. Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement. *arXiv preprint arXiv:2503.06520*, 2025. 8

[49] Jinxing Zhou, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, and Yiran Zhong. Avsbench: A pixel-level audio- visual segmentation benchmark. 8

[50] 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*, volume 38, pages 12155–12163, 2024. 8

[51] Yaoting Wang, Weisong Liu, Guangyao Li, Jian Ding, Di Hu, and Xi Li. Prompting segmentation with sound is generalizable audio-visual source localizer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5669–5677, 2024. 8

[52] Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4974–4984, 2022. 8

[53] Xiang Li, Jinglu Wang, Xiaohao Xu, Xiao Li, Bhiksha Raj, and Yan Lu. Robust referring video object segmentation with cyclic structural consensus. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22236–22245, 2023. 8

[54] Nicholas Stroh. Trackgpt–a generative pre-trained transformer for cross-domain entity trajectory forecasting. *arXiv preprint arXiv:2402.00066*, 2024. 9

[55] Haobo Yuan, Xiangtai Li, Tao Zhang, Zilong Huang, Shilin Xu, Shunping Ji, Yunhai Tong, Lu Qi, Jiashi Feng, and Ming-Hsuan Yang. Sa2va: Marrying sam2 with llava for dense grounded understanding of images and videos. *arXiv preprint arXiv:2501.04001*, 2025. 9, 28

[56] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 9

[57] Jiajun Liu, Yibing Wang, Hanghang Ma, Xiaoping Wu, Xiaoqi Ma, xiaoming Wei, Jianbin Jiao, Enhua Wu, and Jie Hu. Kangaroo: A powerful video-language model supporting long-context video input. *arXiv preprint arXiv:2408.15542*, 2024. 9

[58] Chaoyou Fu, Haojia Lin, Xiong Wang, Yi-Fan Zhang, Yunhang Shen, Xiaoyu Liu, Yangze Li, Zuwei Long, Heting Gao, Ke Li, et al. Vita-1.5: Towards gpt-4o level real-time vision and speech interaction. *arXiv preprint arXiv:2501.01957*, 2025. 9

[59] Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024. 9

[60] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024. 10

[61] David F. Crouse. On implementing 2d rectangular assignment algorithms. *IEEE Transactions on Aerospace and Electronic Systems*, 52(4):1679–1696, 2016. 24

[62] Yuqi Liu, Tianyuan Qu, Zhisheng Zhong, Bohao Peng, Shu Liu, Bei Yu, and Jiaya Jia. Visionreasoner: Unified visual perception and reasoning via reinforcement learning. *arXiv preprint arXiv:2505.12081*, 2025. 26

[63] Yunyang Xiong, Chong Zhou, Xiaoyu Xiang, Lemeng Wu, Chenchen Zhu, Zechun Liu, Saksham Suri, Balakrishnan Varadarajan, Ramya Akula, Forrest Iandola, et al. Efficient track anything. *arXiv preprint arXiv:2411.18933*, 2024. 27

[64] Kim Sung-Bin, Oh Hyun-Bin, JungMok Lee, Arda Senocak, Joon Son Chung, and Tae-Hyun Oh. Avhbench: A cross-modal hallucination benchmark for audio-visual large language models. *arXiv preprint arXiv:2410.18325*, 2024. 27

[65] Kaining Ying, Henghui Ding, Guangquan Jie, and Yu-Gang Jiang. Towards omnimodal expressions and reasoning in referring audio-visual segmentation. *arXiv preprint arXiv:2507.22886*, 2025. 30

[66] Henghui Ding, Song Tang, Shuting He, Chang Liu, Zuxuan Wu, and Yu-Gang Jiang. Multimodal referring segmentation: A survey. *arXiv preprint arXiv:2508.00265*, 2025. 30

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: See abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See supplementary material.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: We don't have theoretical assumptions.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide a detailed exposition of the sources for all training data and the specific configurations of our training parameters. We believe our work is fully reproducible. See Section 4 and Appendix A.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We plan to release the model and code later.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: See supplementary material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We follow the mainstream evaluation methods, which do not report error bars.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See supplementary material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: There is no societal impact of the work performed.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We don't have such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All papers and data are properly credited and cited.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: We don't release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: We don't involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: We don't involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: The usage of LLMs is described in detail. See Section 3 and 4.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.

A Appendix Overview

This appendix provides additional details on the experimental setup, model architecture along with training pipeline, and supplementary results that support the findings presented in the main paper.

- **Implementation Details:** this section details additional aspects of our method: our two-system architecture (including user instructions for its MLLM components and prompts for the downstream SAM2 model), our reward design, and the differences between training and inference procedures.
- **Ablation Studies:** this section provides ablation studies of System 1 regarding architectural selection, reward components, and dataset selection.
- **Visualization Results:** this section provides more visualization results, including examples in comparison with other methods and failure case analysis.
- **More Analysis:** this section provides an analysis on the hallucination issue and the influence of resolution on general video understanding tasks.
- **Limitations and Future Work:** this section discusses the limitations of our method and potential future work.

B Implementation Details

User Instructions on Two Systems. To enable MLLMs to perform keyframe selection and referred object captioning, we designed the prompt as shown in the figure 3. We formulate keyframes as time duration segments and assign spatial description text to each duration. Additionally, we observed that during training, the model could be influenced by the timestamp patterns seen in the prompt examples. Therefore, to increase the diversity of keyframe distributions during training and prevent the model from overfitting to specific timestamps, we incorporated randomized timestamps into the prompt, encouraging the model to focus on learning keyframe selection and caption rather than simply copying training timestamps. For AVS tasks, we designed a similar prompt (see Figure 4) to guide the model in analyzing the audio content and identifying the corresponding visual grounding description. The prompt emphasizes the need to avoid temporal expressions and instead focus on visual cues.

The intermediate results are then interpreted as frames and paired descriptions before being fed into System 2. The prompt for System 2 follows the official grounding prompt of Qwen2.5-VL, where the output is a list of bounding boxes and their corresponding labels in JSON format.

Prompt Design for SAM2 as Downstream Segmenter. Once the keyframe grounding results are obtained from System 2, we assign a unique identifier to each detection result using a tuple format: `(roll_out_idx, frame_idx, pred_obj_idx, bbox)`. This ensures that any detection within a single GRPO group can be uniquely referenced.

Since all detections within a group share the same input context, we optimize inference efficiency by processing all detection results in a single forward pass. Specifically, we feed the entire video segment into SAM2, and for each detection tuple, we assign a unique object ID. These object IDs are used as conditioning inputs to SAM2 to obtain their respective segmentation masks.

Specifically, we maintain a mapping dictionary between detection tuples and assigned object IDs $\mathcal{P} : \texttt{tuple} \rightarrow \texttt{obj\_id}$, which enables us to reverse-map SAM2's outputs $\{\hat{\mathcal{M}}_{\texttt{obj\_id}} | \texttt{obj\_id} \in \mathcal{P}(\texttt{tuple})\}$ back to the original detection structure $\hat{\mathcal{M}}_{\texttt{tuple}}$. The segmentation results are then matched to the corresponding grounded predictions and used for reward evaluation.

Reward Design of $R_{\mathcal{K}}$ To enable the model to learn from diverse keyframe selections, we design an evaluation reward function, $R_{\mathcal{K}}$, which assesses both the diversity and quality of the chosen keyframes. This function is formulated as:

$$R_{\mathcal{K}} = \lambda_1 R_{\text{diversity}} + \lambda_2 R_{\text{num}} + \lambda_3 R_{\text{saliency}}$$

where the last component, $R_{saliency}$, has been clarified in the main paper. Here, we detail the first two reward components.

Firstly, to discourage the model from selecting temporally adjacent keyframes, which can lead to redundant System 2 inference, we introduce a distribution reward function, $R_{diversity}$. This component evaluates the distributional diversity of keyframes by calculating the temporal intervals between them. Specifically, all keyframes are sorted in chronological order. We then compute the temporal interval $t_{i+1} - t_i$ between each pair of consecutive keyframes. The final $R_{diversity}$ value is subsequently determined based on the collection of all such inter-frame temporal intervals.

The diversity reward $\mathcal{R}_{\text{diversity}}(\mathcal{S})$ can be defined as the sum of an overlap punishment term and a distribution reward term:

$$\mathcal{R}_{\text{diversity}}(\mathcal{S}) = \textit{overlap\_punish} \cdot |\mathcal{I}| + \textit{dist\_reward} \cdot |\mathcal{D}|$$

where:

- $\mathcal{S}$ is the set of selected items. Let $\mathcal{S}_{\text{sorted}} = (s_1, s_2, \ldots, s_M)$ be the sequence of $M = |\mathcal{S}|$ items sorted according to the relevant criteria (e.g., timestamps).
- $\mathcal{I}$ is the set of indices i for which an "overlap" condition is met between s_i and s_{i+1}. Specifically, assuming $\text{idx}(s_j)$ gives an identifier for item s_j:

  $$\mathcal{I} = \{i \in \{1, \ldots, M-1\} \mid \text{idx}(s_i) = \text{idx}(s_{i+1})\}$$

  $|\mathcal{I}|$ is the number of such identified overlaps (e.g., pairs of consecutive items with identical identifiers).
- $\mathcal{D}$ is the set of indices of items in $\mathcal{S}_{\text{sorted}}$ that are not considered the start of an overlap as defined by $\mathcal{I}$:

  $$\mathcal{D} = \{j \in \{1, \ldots, M\} \mid j \notin \mathcal{I}\}$$

  Therefore, $|\mathcal{D}| = M - |\mathcal{I}|$.
- *overlap_punish* is the coefficient for the punishment. For this term to act as a punishment, *overlap_punish* should typically be a negative value (e.g., -0.2), or if it's a positive value, it should be subtracted from the reward.
- *dist_reward* is the coefficient for the reward given to items not initiating an overlap.

The formula $\mathcal{R}_{\text{diversity}}(\mathcal{S})$ can also be written as:

$$\mathcal{R}_{\text{diversity}}(\mathcal{S}) = (\textit{overlap\_punish} - \textit{dist\_reward}) \cdot |\mathcal{I}| + \textit{dist\_reward} \cdot M$$

Reward Design of $R_{\mathcal{A}}$ The specific formulation of $R_{\mathcal{A}}$ is as follows:

$$R_{\mathcal{A}} = \frac{1}{K} \sum_{i=1}^{K} \left(1 - \mathcal{L}_{\text{Hungarian}}(\mathcal{B}_{s_i}, \mathcal{B}^*_{s_i})\right)$$

where $\mathcal{B}_{s_i}$ denotes the set of predicted bounding boxes at the s_i-th frame, and $\mathcal{B}^*_{s_i}$ represents the corresponding set of ground truth bounding boxes. The function $\mathcal{L}_{\text{Hungarian}}$ refers to the Hungarian matching loss [45, 61], and K is the total number of selected keyframes.

The Hungarian matching loss $\mathcal{L}_{\text{Hungarian}}$ is computed based on the Intersection-over-Union (IoU) between predicted and ground truth bounding boxes. Specifically, a cost matrix $\mathcal{M}$ is first constructed using the IoU values between each pair of predicted and ground truth boxes. Then, the Hungarian algorithm is applied to the negative matrix $-\mathcal{M}$ to obtain the optimal one-to-one matching that minimizes the total negative cost, which corresponds to maximizing the overall IoU-based matching accuracy.

Reward Design of $R_{\mathcal{G}}$ For $R_{\mathcal{G}}$, we adopt a simple aggregated IoU as the reward function. Specifically, for each detected object, we accumulate the predicted segmentation masks across all objects to construct a per-frame mask set $\hat{\mathcal{M}}_t$. Then, we compute the Intersection-over-Union (IoU) between the predicted masks and the corresponding ground truth masks $\mathcal{M}^*_t$ on each frame. The final reward is obtained by averaging the IoU values across all frames.

$$R_{\mathcal{G}} = \frac{1}{T} \sum_{t=1}^{T} \text{IoU}(\hat{m}_t, m^*_t)$$

Prompt for OMNI-R1 as System 1

- Given a `[frames]` seconds video and a reference instruction: `[ref_prompt]` that may involve temporal behavior, identify the exact object(s) `[ref_prompt]` in the video that matches the description.
- Select about 4 most relevant moments that contain the referred object(s) with the best view.
- Then, simplify the identified object into a short and clear visual grounding description that can be used for single-image reference at each moment.
- Avoid temporal phrases and comparison phrases like "walking", "moving", "approaching", "bigger" or "smaller", but instead describe visible visual cues like clothing, pose, position, or grouping.
- Try to select moments that are **temporally well-distributed across the video**, rather than clustered in the same part of the timeline. Avoid selecting multiple timestamps that are adjacent or overlapping; instead, prefer clearly distinct moments that each offer unique visual information. It is better to choose the most relevant and highly representative moments **spanning the entire video**, rather than picking all from the beginning.
- Explain your reasoning in `<think></think>` and output the final result in `<answer></answer>`. Your final answer should be a JSON object in the following format:

```
<think> your analysis about the video and reference instruction </think>
<answer>
{
"start_time": "00:[start]",
"end_time": "00:[end]",
"description": "direct description of referred object(s) at this moment"
}
</answer>
```

Figure 3: Keyframe selection and recaptioning prompt for System 1.

Training and Inference Strategy For video clips, we first feed them into System 1 at a relatively low resolution of a per-frame pixel $128\times28\times28$, which allows us to process longer video segments during training and inference. Then System 2 predicts detection results at a higher resolution of $900\times28\times28$. **For VOS tasks**, we adopt a random uniform sampling strategy during training, selecting between 8 and 24 frames per video to enhance temporal diversity and robustness. All SAM2-based segmentation and reward evaluations are then applied to these resampled clips at their original input resolution. **For RefAVS tasks**, we observed severe cross-modal hallucination issues during preliminary experiments, particularly when reasoning jointly over full-length audio and multi-frame video inputs. To mitigate this, we introduce a simplified variant, RefAID (Referring Audio-Image Detection), which reduces the AVS problem to object detection using only the first video frame and the corresponding full audio query. In this setting, no SAM2 segmentation is used; training is driven solely by detection-based rewards.

During inference, we resample a fixed maximum of 24 frames per video for VOS tasks. Unlike training, segmentation and evaluation are conducted over the full video sequence using SAM2 to align with standard benchmark protocols. For RefAVS, we adopt the same resampling and evaluation procedure as in VOS, ensuring consistency across task settings.

C Ablation Studies

We conduct an ablation study to investigate the effect of progressively designed reward components $R_{\mathcal{K}}$ (keyframe coverage), $R_{\mathcal{A}}$ (alignment via Hungarian matching), and $R_{\mathcal{G}}$ (global grounding IoU) on the overall performance. To ensure a fair comparison, all models are trained for one epoch on the ReVOS and MeVIS datasets and are constrained to select exactly four keyframes unless otherwise noted.

Table 4 presents the results across four evaluation subsets of ReVOS: referring, reasoning, single-object, and multi-object. When using the baseline System 2, $R_{\mathcal{K}} + R_{\mathcal{G}}$ achieves the highest overall score (39.9%), outperforming the full combination $R_{\mathcal{K}} + R_{\mathcal{A}} + R_{\mathcal{G}}$ (38.4%). Interestingly, this trend reverses when employing Sa2VA as System 2, where $R_{\mathcal{K}} + R_{\mathcal{A}} + R_{\mathcal{G}}$ attains the best overall performance (54.0%). This indicates that while $R_{\mathcal{A}}$ may degrade performance with the original

Prompt for OMNI-R1 as System 1

- Given a `[audio_duration]` audio and a reference instruction: `[ref_prompt]`, which involves temporal and audio-related behavior, first analyze the objects in the image that are producing sound, including both human voices and instrument sounds.
- Based on the audio content, identify the exact object `[ref_prompt]` in the image that matches the audio.
- Then, simplify the identified object into a short and clear visual grounding description that can be unambiguously recognized in a single image without relying on audio.
- Avoid using temporal expressions such as "playing" or "moving"; instead, describe visible visual cues such as clothing, pose, position, or grouping.
- Explain your reasoning in `<think></think>` and output the final result in `<answer></answer>`.

Figure 4: Audio analyzing and recaptioning prompt for System 1.

System 2 due to the locality noises introduced with it, it substantially enhances generalization to alternative System 2 architectures.

We hypothesize that this is because $R_{\mathcal{A}}$ is a weakly coupled reward, focusing on the alignment between visual features and language prompts at the single-frame level. It provides better transferability across different System 2 implementations. In contrast, $R_{\mathcal{G}}$ tends to overfit to the specific structure and behavior of the System 2 used during training, limiting generalization.

These findings validate the effectiveness of $R_{\mathcal{G}}$ as a grounding-aware reward but also highlight the limitations of outcome-based $R_{\mathcal{G}}$ under two system reinforcement learning settings.

To assess the impact of architectural choices within our pipeline, we perform ablation studies by replacing key components, specifically the grounding and segmentation models, and analyzing their effects on both training and inference performance.

Grounding Model We substitute the original Qwen2.5-Omni-7B as System 2 with VisionReasoner [62], a unified visual perception model enhanced through GRPO optimization across grounding, counting, and segmentation tasks. To maintain experimental validity, we keep all training hyperparameters consistent across configurations and test two replacement strategies: (1) replacing the model during training only, and (2) replacing it during both training and inference phases.

As seen in 5, simply replacing the grounding model during training leads to noticeable gains, suggesting that current approach has further improvements under a more accurate System 2. A full switch to VisionReasoner during both training and inference yields even greater improvements across most sub-tasks.

Table 4: Ablation study on reward functions $R_{\mathcal{K}}$, $R_{\mathcal{A}}$ and $R_{\mathcal{G}}$ for System 1. We also evaluate Sa2VA as System 2 to investigate keyframe performance. The first model is trained with additional 2,000 samples from grounding dataset refcocog.

Method	**ReVOS**				
	Referring	*Reasoning*	*Single*	*Multi*	***Overall***
Omni-R1 + *refcocog*	52.5	36.9	45.0	46.6	44.7
$R_{\mathcal{K}}$+$R_{\mathcal{A}}$+$R_{\mathcal{G}}$	44.2	32.5	38.2	41.9	38.4
$R_{\mathcal{K}}$+$R_{\mathcal{G}}$	45.5	34.2	39.6	42.6	39.9
$R_{\mathcal{K}}$+$R_{\mathcal{A}}$	43.1	29.5	36.8	37.5	36.3
$R_{\mathcal{K}}$	44.1	26.2	36.6	34.1	35.2
$R_{\mathcal{G}}$	47.3	28.8	38.9	40.4	38.1
$R_{\mathcal{G}}$ + *Sa2VA*	58.1	43.5	50.6	53.5	50.8
$R_{\mathcal{K}}$ + $R_{\mathcal{G}}$ + *Sa2VA*	60.4	41.8	51.7	51.0	51.1
$R_{\mathcal{K}}$+$R_{\mathcal{A}}$+$R_{\mathcal{G}}$ + *Sa2VA*	61.8	46.1	54.6	50.7	54.0

Table 5: Ablation study of different grounding models as System 2.

Method	ReVOS				
	Referring	*Reasoning*	*Single*	*Multi*	***Overall***
Original	52.5	36.9	45.0	46.6	44.7
Replace during training	52.2	39.4	45.4	49.2	45.8
Replace also at inference	55.5	41.7	49.1	44.7	48.6

Segmentation Model In addition, we replace the segmentation model used during inference in 6. Specifically, we switched from SAM2 to TAM [63], a more efficient segmentation tracker, while keeping VisionReasoner as the grounding model (training only).

Table 6: Ablation study of different grounding models as System 2.

Segmentation Model	ReVOS				
	Referring	*Reasoning*	*Single*	*Multi*	***Overall***
SAM2 (baseline)	52.2	39.4	45.4	49.2	45.8
TAM	50.6	39.1	44.0	51.3	44.9

D Visualization Results

Mask Quality Since our method utilizes the SAM2 model for segmentation, without fine-tuning mask decoder, the final mask output is more stable than those methods that rely on additional training on segmentation mask decoder. As can be seen in Figure 5a, in this simple example, our method is able to segment the target object with a mask consistent with the ground truth, while Sa2VA predicts the right target but generates a mask with holes and noise.

Temporal Reasoning Our System 1 leverages temporal reasoning to improve segmentation accuracy. As can be seen in Figure 5b, in this example, one has to **watch the whole video and analyze the video context to make a correct prediction** about the next bottle to be picked up. Our method is able to select the bottle that is about to be picked up, while Sa2VA fails to do so and segments the bottle that is already picked up. A similar case is shown in Figure 6a, where one has to leverage world knowledge to understanding the target object and our method selects the right object while Sa2VA fails to do so. Both cases show that our method is able to leverage temporal reasoning to improve the segmentation accuracy.

Detail Understanding Our System 1 leverages detail reasoning to improve segmentation accuracy. Figure 5c shows a scenario where detail reasoning is needed to figure which wineglass is likely will be finished first. Our System 1 is **already able to select the right wineglass but still makes a loose description for System 2 to analyze detail information**, while Sa2VA fails to understand the instruction and segments all the wineglasses. This shows that our System 1 is able to delay detail reasoning for System 2 for detail understanding to improve the segmentation accuracy.

E More Analysis

Hallucination Analysis During the training of our RefAVS task, we identified a significant hallucination problem, which we attribute to the complexity of multi-modal video and audio inputs. To systematically evaluate this issue, we conducted targeted assessments on audio-related hallucinations using the JUDGE subset of AVHBench [64], the first comprehensive benchmark designed to evaluate the perception and comprehension abilities of audio-visual large language models (LLMs).

As shown in Table 8, our base model (Qwen2.5-Omni-7B) achieves an accuracy of 58.5% on the JUDGE subset. Training on 1600 AVS samples leads to a modest improvement (60.8%), which is further enhanced to 61.5% by applying the GRPO KL loss with a reduced coefficient ($\beta = 0.001$). Notably, increasing the AVS training samples to 10400 does not yield better results, suggesting potential overfitting or task imbalance.

Table 7: Video Object Segmentation performance on MeVIS across different methods, the metric is $\mathcal{J}\&\mathcal{F}$ score(%). ‡ means the results are evaluated where Omni-R1-7B serves as System 1 and Sa2VA-1B as System 2

Model	MeVIS *val_u*
Sa2VA-1B [55]	53.4
Sa2VA-4B [55]	55.4
Qwen2.5-Omni-7B†	33.6
Omni-R1-8B‡	**55.4**

Table 8: Performance on AVHBench (JUDGE subset, total 5302 samples). In the table, AVS tasks are trained on RefAVS dataset and VOS tasks are trained on ReVOS and MeViS datasets. The default GRPO KL loss weight $\beta = 0.04$.

Method	AVHBench JUDGE	
	Correct Answers	Accuracy
Base Model	3100	58.5%
AVS 1600 samples	3222	60.8%
AVS 1600 samples with $\beta = 0.001$	3261	61.5%
AVS 10400 samples	3120	58.9%
VOS	3500	66.0%
AVS and VOS	3811	**71.9%**

On the other hand, training with VOS tasks alone significantly boosts accuracy to 66.0%, and the best performance (71.9%) is obtained by jointly training on both AVS and VOS tasks. This represents a substantial improvement of 13.4% over the base model, demonstrating that multi-task training not only enhances audio-visual grounding but also mitigates hallucination issues more effectively.

These results confirm the effectiveness of leveraging task diversity and balanced reward optimization in improving the robustness of multimodal reasoning.

Table 9: Performance comparison across different resolutions and the use of a thinking prompt on VideoMME and MVBench. Resolutions are set to either `128×28×28` (default) or `256×28×28` (high). The thinking prompt provides an additional reasoning cue. The reported metric is the average of $\mathcal{J}$ and $\mathcal{F}$ scores (%).

Model	Resolution	Thinking	VideoMME		MVBench
			General	Short	Avg
Qwen2.5-Omni	128×28×28	No	58.3	69.8	66.1
Qwen2.5-Omni	256×28×28	No	58.7	69.9	67.0
Qwen2.5-Omni	128×28×28	Yes	59.3	70.1	68.1
Qwen2.5-Omni	256×28×28	Yes	59.8	70.9	68.3
Omni-R1-AVS	128×28×28	No	59.0	71.9	68.3
Omni-R1-AVS	256×28×28	No	59.4	71.9	68.7
Omni-R1-AVS	128×28×28	Yes	59.9	72.1	69.4
Omni-R1-AVS	256×28×28	Yes	60.0	72.1	69.5
Omni-R1-VOS	128×28×28	No	59.7	72.3	68.9
Omni-R1-VOS	256×28×28	No	59.6	72.5	68.9
Omni-R1-VOS	128×28×28	Yes	59.8	72.5	69.8
Omni-R1-VOS	256×28×28	Yes	60.1	72.8	69.9
Omni-R1-VOS-AVS	128×28×28	No	60.1	72.5	69.1
Omni-R1-VOS-AVS	128×28×28	Yes	60.7	73.0	70.3

Video Resolution Influence on General Video Understanding Tasks To evaluate the influence of input resolution and prompting strategy on general video understanding, we compare model performance across different configurations on the VideoMME and MVBench benchmarks, as summarized in Table 9. All models are evaluated under two resolution settings: the default resolution of $128\times28\times28$ and a higher resolution of $256\times28\times28$ (denoted with *), with and without the proposed *thinking* prompting strategy.

We observe that increasing the input resolution consistently leads to performance gains across all models. For instance, Qwen2.5-Omni improves from 66.1% to 67.0% on MVBench when evaluated at higher resolution. Similarly, our Omni-R1-AVS model benefits from the resolution increase, achieving a performance gain from 68.3% to 68.7%. These improvements suggest that higher spatial resolution enhances the model's ability to capture fine-grained visual details, particularly beneficial for multi-object reasoning and scene comprehension.

In addition to resolution, the *thinking* prompt designed to guide the model toward structured multi-step reasoning further boosts performance across all tested models. Omni-R1-AVS with *thinking* achieves 69.4% on MVBench, outperforming its baseline by 1.1%. The combination of both higher resolution and *thinking* yields the best results overall, with Omni-R1-VOS-AVS + *thinking** reaching 60.7% on VideoMME (general) and 70.3% on MVBench. This indicates that resolution and prompting act as complementary strategies: resolution improves visual precision, while prompting enhances reasoning capability.

However, the performance gain obtained by increasing video resolution is marginal, suggesting that in general understanding benchmarks, resolution plays a limited role, and temporal understanding is more critical than fine-grained spatial details. This observation is consistent with our findings in the main paper, where the dual-system design significantly enhances the model's temporal reasoning capabilities and yields notable improvements on reasoning-intensive VOS tasks.

F Limitations and Future Work

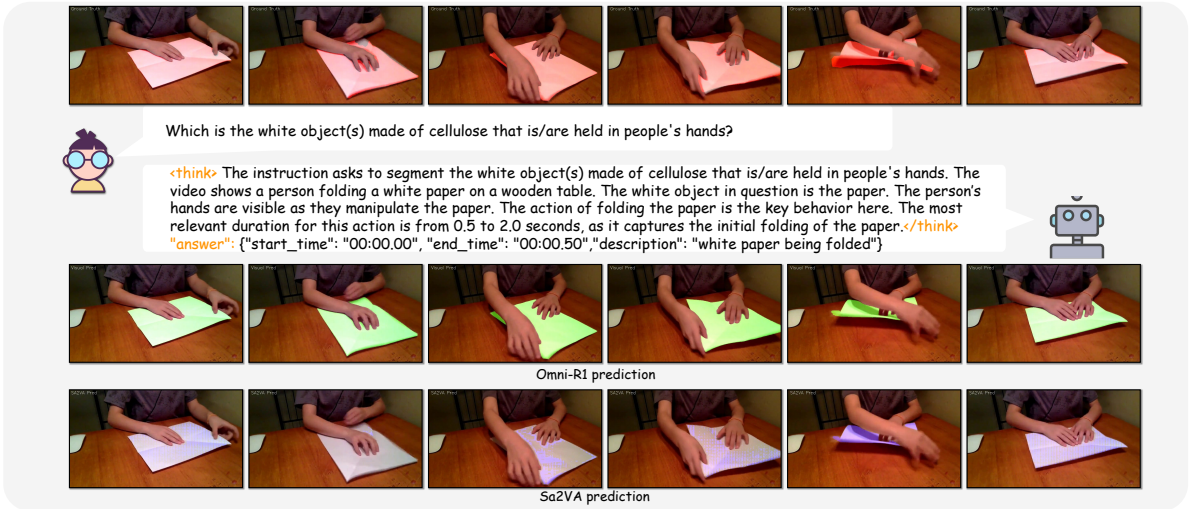
Limitations Although our dual-system design significantly enhances the temporal reasoning capability of **System 1**, the complete functional decoupling between System 1 and System 2 introduces certain limitations. In particular, **System 2** lacks temporal context, which may affect consistency in temporally coherent tasks. This consideration partially motivates our selection of **VOS** as a primary training task: while VOS emphasizes temporal consistency, it also provides dense per-frame annotations that allow us to design stable training strategies to mitigate the context gap—such as frame-wise Hungarian matching loss and aggregated mask-based rewards. During inference, the missing temporal cues in System 2 are partially recovered through SAM2's mask-based processing.

However, when extending to tasks requiring finer temporal sensitivity such as detecting and describing localized anomalous behaviors within a specific time span, our current architecture faces new challenges. While **System 1** can still progressively narrow down relevant temporal segments, the need for dynamic temporal granularity exposes the limitations of a fully decoupled, one-way reasoning architecture.

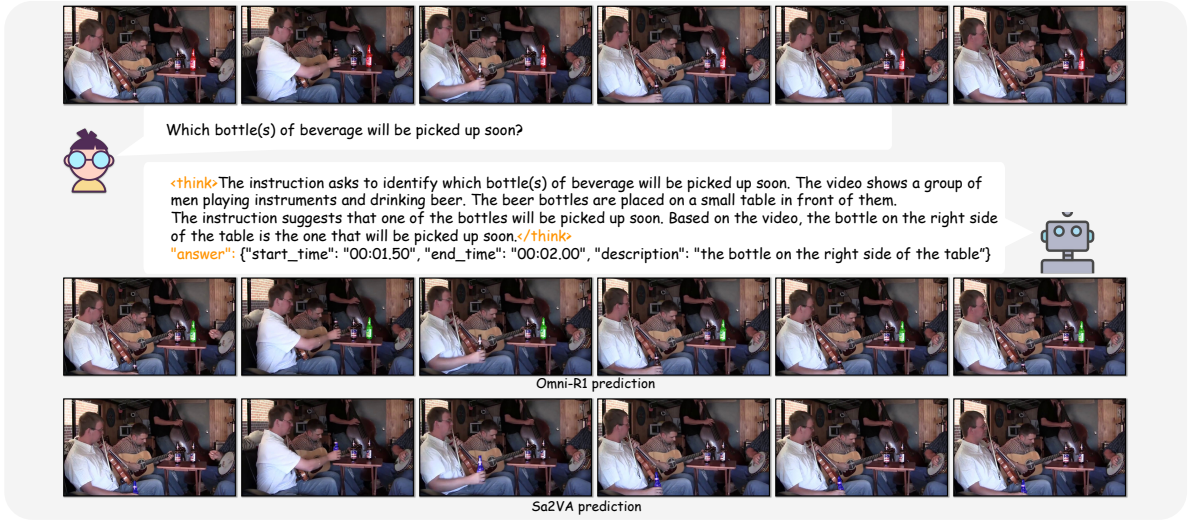
Discussion on Future Work Our coarse-to-fine reasoning pathway aligns closely with human cognitive intuition, yet we recognize the crucial importance of enabling a bidirectional flow of information between global and local levels. Our current design can be viewed as a context-constrained tree search structure, where only the root node (**System 1**) has access to full contextual information, while downstream nodes (**System 2**) operate solely on partial, local input. This diverges from typical hierarchical systems, which often allow child nodes to access aggregated information from their parent nodes.

While the VOS task structure inherently compensates for this limitation by providing dense temporal supervision, to further enhance the flexibility of our two-system framework, future work should explore more interactive architectures that facilitate explicit information exchange between **System 1** and **System 2**. Supporting backtracking within the reasoning tree would empower **System 1** to refine its global reasoning using local insights from **System 2**. Conversely, it would enable **System 2** to operate with broader contextual awareness provided by **System 1**. We firmly believe that transforming this pipeline into a bidirectional, cooperative reasoning structure holds immense

potential for advancing multi-modal temporal understanding abilities on more flexible and complex tasks [65, 66].



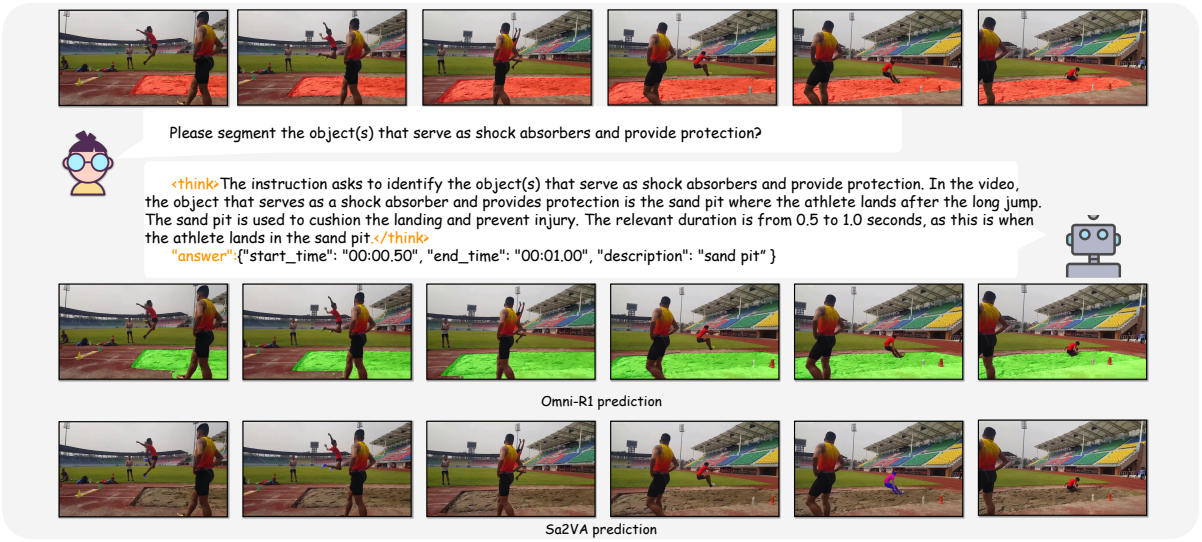
(a) In this simple example, both our method and Sa2VA are able to select the right target object, but our method is able to segment the target object with a mask consistent with the ground truth, while Sa2VA generates a mask with holes and noise.



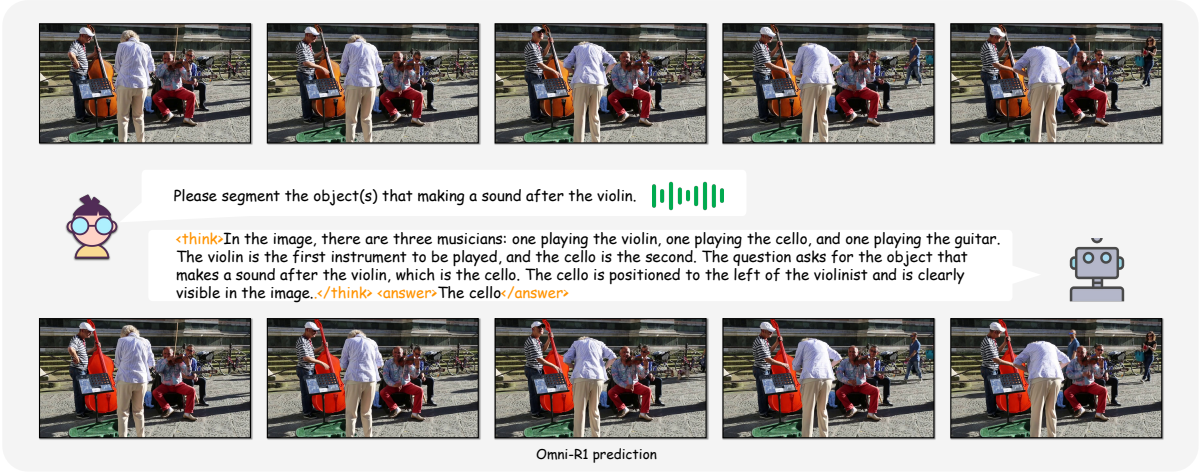
(b) The target object can only be predicted after one watches the whole video and makes a prediction according to the video context.

(c) Though our method is able to select the right wineglass at System 1, it still lets System 2 analyze the detailed information.

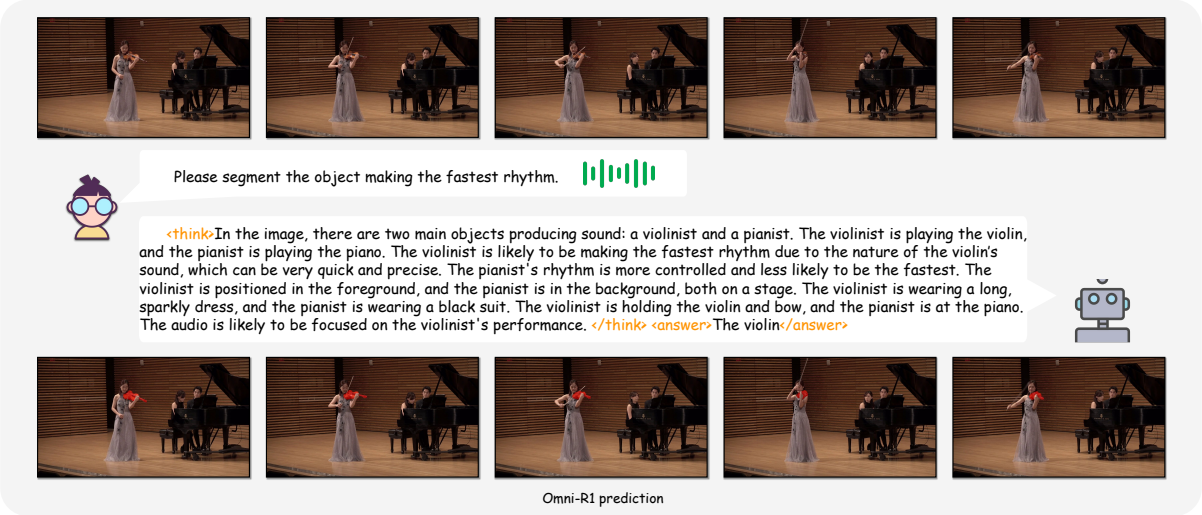
Figure 5: Visualization results of our method on three representative VOS cases. Each subfigure illustrates a different reasoning pattern.



(a) The case shows the video understanding abilities of our method. Our method analyses the function of the sand pit and is able to segment, while Sa2VA fails to do so.



(b) The case shows the video temporal understanding abilities of our method. Our method analyses the audio events and finds out the temporal behavior in visual context.



(c) The case shows the audio and video understanding abilities of our method. Our method analyses the music in the video and finds out the object making the fastest rhythm.

Figure 6: More visualization results of our method on representative VOS and AVS cases.