AURORA: Augmented Understanding via Structured Reasoning and Reinforcement Learning for Reference Audio-Visual Segmentation

Ziyang Luo[1], **Nian Liu**[1*], **Fahad Shahbaz Khan**[2], **Junwei Han**[1,3]

[1]Northwestern Polytechnical University
[2]Mohamed bin Zayed University of Artificial Intelligence
[3]Chongqing University of Posts and Telecommunications

Abstract

Reference Audio-Visual Segmentation (Ref-AVS) tasks challenge models to precisely locate sounding objects by integrating visual, auditory, and textual cues. Existing methods often lack genuine semantic understanding, tending to memorize fixed reasoning patterns. Furthermore, jointly training for reasoning and segmentation can compromise pixel-level precision. To address these issues, we introduce AURORA, a novel framework designed to enhance genuine reasoning and language comprehension in reference audio-visual segmentation. We employ a structured Chain-of-Thought (CoT) prompting mechanism to guide the model through a step-by-step reasoning process and introduce a novel segmentation feature distillation loss to effectively integrate these reasoning abilities without sacrificing segmentation performance. To further cultivate the model's genuine reasoning capabilities, we devise a further two-stage training strategy: first, a "corrective reflective-style training" stage utilizes self-correction to enhance the quality of reasoning paths, followed by reinforcement learning via Group Reward Policy Optimization (GRPO) to bolster robustness in challenging scenarios. Experiments demonstrate that AURORA achieves state-of-the-art performance on Ref-AVS benchmarks and generalizes effectively to unreferenced segmentation.

Code — https://github.com/Sssssuperior/AURORA

1 Introduction

Humans perceive and interact with a dynamic world through the seamless integration of multiple sensory modalities. While vision is dominant, auditory and textual cues are often essential for accurately identifying and understanding target objects, motivating the need for systems capable of similar multimodal integration and segmentation (Yao et al. 2025; Zhang et al. 2025). Recent advancements in multimodal large language models, such as Qwen-Omni (Xu et al. 2025), and VideoLLaMA2 (Cheng et al. 2024), have demonstrated significant progress in audio-visual comprehension. However, tasks like Reference Audio-Visual Segmentation (Ref-AVS) (Wang et al. 2024b) still remain challenging, requiring models to precisely segment specific sounding objects by integrating textual references, audio cues, and visual information.

While previous works have explored fusion methods (Wang et al. 2024b, 2025; Radman and Laaksonen 2025) to align modalities, often employing an auxiliary text encoder with self-attention fusion, it remains unclear whether these models truly grasp the meaning of textual references or merely exhibit language bias. The "black box" nature of these approaches hinders genuine semantic understanding. To address these interpretability concerns, recent approaches have begun incorporating Large Language Models (LLMs) to enhance reasoning capabilities in Ref-AVS tasks, as demonstrated by Crab (Du et al. 2025). However, their existing implementations often rely on supervised fine-tuning with pre-defined simple reasoning templates. This approach tends to lead the LLM towards "memorizing" these fixed patterns, which could result in outputs that reflect post-hoc rationalization rather than genuine and systematic reasoning. Moreover, the design of jointly training complex language reasoning with precise visual segmentation introduces new challenges, which risks compromising the pixel-level precision of the original segmentation models as pointed out by (Liu et al. 2025).

In this paper, we propose AURORA, a novel method that enhances both language comprehension and reasoning capabilities for audio-visual segmentation. Specifically, we leverage VideoLLaMA2 (Cheng et al. 2024) as our MLLM component to work in tandem with SAM (Kirillov et al. 2023) for segmentation, enabling reasoning-guided segmentation. To prevent models from merely rationalizing the final output, we introduce a structured Chain-of-Thought (CoT) (Wei et al. 2022) prompting mechanism via open-source Qwen-Omni (Xu et al. 2025) to cost-effectively generate diverse reasoning paths for training. It involves distinct analytical steps focusing sequentially on visual, audio, and textual reference cues, followed by a final integration of these analyses to derive the answer. To mitigate the potential conflict between reasoning and segmentation during joint training, we introduce a segmentation feature distillation loss. This loss distills knowledge from a segmentation-only model into the joint model, enabling it to acquire CoT reasoning capabilities without compromising segmentation performance. To endow the model with genuine reasoning ability, we further adopt a two-stage training strategy after supervised fine-

*Corresponding author: Nian Liu (liunian228@gmail.com)

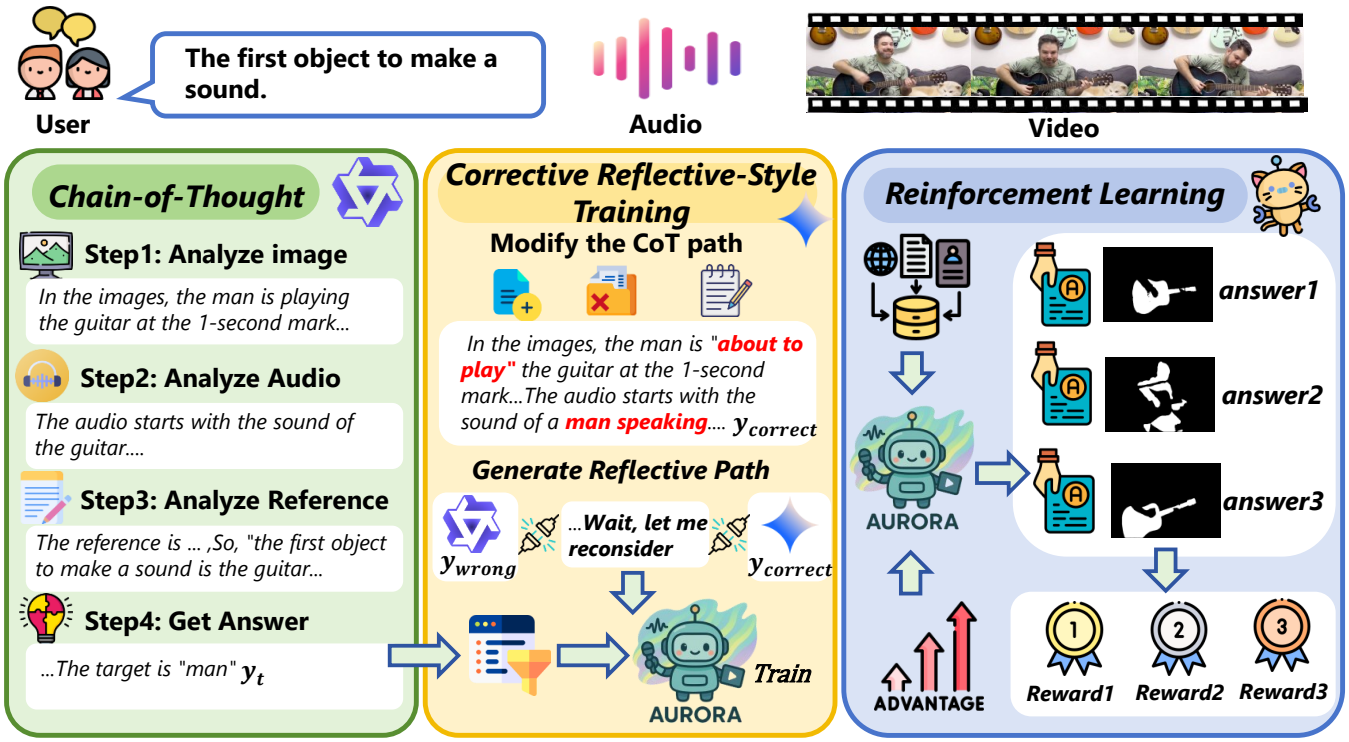

Figure 1: Overall training pipeline of our proposed model. The training pipeline consists of three stages: (1) Supervised Fine-Tuning with CoT paths (y_t) generated by Qwen-Omni (Xu et al. 2025). (2) Corrective Reflective-Style Training, in which we construct a "reflective-style" path by combining the reasoning output from the SFT-trained model, a self-correction trigger, and the corrected path ($y_{correct}$) from Gemini. (3)Reinforcement Learning via GRPO to further refine the model's reasoning.

tuning (SFT) with CoT, combining reflective learning and reinforcement learning via Group Reward Policy Optimization (GRPO) (Guo et al. 2025). In the corrective reflective-style stage, we transform low-quality reasoning paths from SFT-trained model into "reflective reasoning paths" by juxtaposing the initial flawed reasoning with a self-correction prompt followed by a Gemini-assisted high-quality revision. In the GRPO stage, we design a hybrid reward function incorporating format reward, IoU reward, and class reward to guide the model toward more robust and multimodally grounded reasoning. Our main contributions can be summarized as follows: (1) We propose AURORA, a reasoning-enhanced Ref-AVS framework that introduces structured Chain-of-Thought prompting with segmentation feature distillation loss for genuine multimodal reasoning and precise segmentation. (2) A two-stage refinement—reflective corrective learning and GRPO-based reinforcement—further mitigates foundation model biases and strengthens reasoning robustness. (3) AURORA achieves state-of-the-art performance on Ref-AVS benchmarks and generalizes well to unreferenced segmentation.

2 Related Work

2.1 Audio-Visual Segmentation

Audio-Visual Segmentation (AVS) aims to generate pixel-level masks for sounding objects and has attracted increasing research attention (Zhou et al. 2023). While some prior works have incorporated text modality into AVS, they primarily focus on class alignment without fully exploiting the semantic richness of textual information (Bhosale et al. 2023, 2024; Luo et al. 2025a). Furthermore, traditional AVS methods (Zhou et al. 2023; Li et al. 2023a; Gao et al. 2024; Li et al. 2024; Yang et al. 2024; Ling et al. 2024; Luo et al. 2025b) lack explicit guidance mechanisms, making it challenging to identify specific objects of interest within complex audio-visual scenes. To address this limitation, Ref-AVS (Wang et al. 2024b) introduces a reference-based framework that provides continuous segmentation guidance through audio-visual-text fusion via a multi-modal cue aggregation module. Building upon this approach, SAM2-LOVE (Wang et al. 2025) further enhances the framework by integrating SAM2 with multimodal fusion, token propagation, and accumulation strategies. TSAM (Radman and Laaksonen 2025) enhances SAM with temporal modeling capabilities for spatio-temporal learning across video frames and replaces interactive prompting with data-driven prompts, thereby extending SAM's functionality to dynamic video content. However, these methods treat text as an auxiliary modality and may fail to fully understand reference semantics.

Our work distinguishes itself by proposing a novel segmentation-centric framework integrating SAM (Kirillov et al. 2023) with VideoLLaMA2 (Cheng et al. 2024) for

reference-guided segmentation, employing CoT (Wei et al. 2022) reasoning to incrementally decompose references and a segmentation feature distillation loss to preserve segmentation accuracy in SFT training.

2.2 Multimodal Language Models in Audio-Visual Scenes

In audio-visual understanding, predominant Multimodal Language Models (MLLMs) such as Qwen-Omni (Xu et al. 2025), and VideoLLaMA2 (Cheng et al. 2024) integrate audio, visual, and textual modalities to achieve comprehensive scene understanding. MEERKAT (Chowdhury et al. 2024) advances this field by constructing a fine-grained, large-scale audio-visual instruction-tuning dataset that endows models with temporal and spatial grounding capabilities. More specifically for reference AVS, CRAB (Du et al. 2025) attempts to unify multimodal understanding and segmentation within an LLM framework. However, their reliance on supervised fine-tuning with pre-defined reasoning templates may lead to superficial pattern memorization rather than fostering genuine systematic reasoning or deep semantic comprehension, limiting their robustness in complex scenarios. To address these limitations, following SFT, we introduce an additional two-stage training process. The first stage, corrective reflective-style training, encourages the model to refine its prior knowledge and calibrate its foundational perceptual abilities. The second stage, GRPO, facilitates robust self-improvement by optimizing the model's reasoning pathways based on targeted reward functions.

3 Methodology

In this work, we introduce AURORA, a novel reasoning-enhanced framework for reference audio-visual segmentation that addresses the fundamental challenge of achieving genuine multimodal reasoning and maintaining precise segmentation performance. We implement this framework by integrating SAM (Kirillov et al. 2023) with VideoLLaMA2 (Cheng et al. 2024). Our approach employs a three-stage training procedure. In the first stage, we incorporate CoT reasoning using Qwen-Omni (Xu et al. 2025) for SFT with segmentation feature distillation loss. In the second stage, we enhance the model's foundational accuracy through a novel error-correction stage using structured "reflective examples" that address common perceptual and knowledge errors. The third stage employs GRPO (Guo et al. 2025) to jointly enhance both reasoning capabilities and segmentation performance. Figure 1 illustrates the overall architecture of our proposed framework.

3.1 Stage1: SFT Training with CoT

CoT To fully utilize GRPO, the initial procedure involves teaching the model proper reasoning techniques and establishing the basic format. Without this foundation, the model can hardly perform effective self-improvement, leading to suboptimal optimization directions. Therefore, in the first training stage, we implement a cold start using SFT. However, simple reasoning like basic scene description or superficial audio analysis in isolation is insufficient, as the model may engage in post-hoc rationalization of the final results, which could introduce complications in the subsequent GRPO procedure. To address this issue, we introduce CoT reasoning, which decomposes the reasoning process into step-by-step components rather than post-hoc explanations.

Specifically, we first generate CoT answers using the open-source Qwen-Omni (Xu et al. 2025) model to conserve computational costs compared to proprietary models, and establish the initial format through carefully designed prompts. Through prompt engineering, we decompose the reference reasoning into four distinct steps as shown below.

CHAIN OF THOUGHT REASONING FOR REF-AVS

Step 1 Video Description: *Extract important visual information from the video, including key actions, objects, and their spatial relationships.*
Step 2 Audio Description: *Extract and analyze audio content, identifying sound classes, and temporal characteristics.*
Step 3 Reference Analysis: *Analyze the relationship between multimodal information and the given reference, identifying relevant connections and correspondences.*
Step 4 Final Answer: *Generate the final segmentation decision as "the target is {class_name}" based on comprehensive analysis.*

The SFT objective aims to maximize the likelihood of a target CoT reasoning sequence $y = (y_1, ..., y_T)$ conditioned on the input query q, which consists of the instruction $\boldsymbol{x}_{inst}$, reference $\boldsymbol{x}_r$, video $\boldsymbol{v}$, and audio $\boldsymbol{a}$. The loss is defined as:

$$\mathcal{L}_{SFT} = -\mathbb{E}_{(q,y)\sim\mathcal{D}} \sum_{t=1}^{T} \log \pi_\theta(y_t|q, y_{<t}), \qquad (1)$$

where D denotes the fine-tuning dataset, and π_θ is the MLLM parameterized by θ. Since Ref-AVS is a binary segmentation task, we follow the LISA framework (Lai et al. 2024) and append the token "`It is [SEG]`" after the CoT reasoning. The hidden state feature of the `[SEG]` token is then used as a visual prompt for the SAM decoder.

Segmentation Feature Distillation Loss While CoT reasoning has demonstrated effectiveness in language tasks, it presents a challenge for multi-modal segmentation tasks. We hypothesize that fine-tuning with a joint objective may compromise the pixel precision of the segmentation models (Liu et al. 2025), particularly when the language-based reasoning loss dominates the optimization dynamics. This, in turn, can degrade the representation of the `[SEG]` token, which is critical for triggering high-quality mask generation, thereby impairing segmentation performance. To address this issue and effectively decouple the optimization of reasoning and segmentation capabilities, we introduce a segmentation feature distillation loss as shown in the gray block in Figure 2.

To obtain an ideal feature representation for distillation, we train a specialist teacher model. Unlike our main models, the teacher's training is deliberately confined to a pure segmentation task using only simple prompts (`It is [SEG]`). This specialized training regime, involving more extensive training steps, allows it to develop a

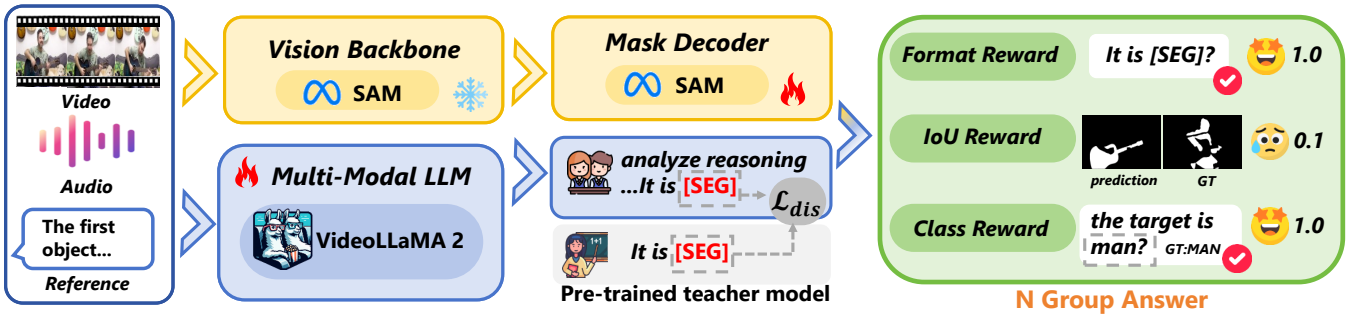

Figure 2: Overall framework of our proposed model. Our model integrates SAM and VideoLLaMA2. The gray block represents the Segmentation Feature Distillation Loss during the SFT stage, and the green block denotes the GRPO process with triplet rewards in Stage 3.

highly-optimized and uncompromised `[SEG]` feature embedding (f_t), even though it entirely lacks complex reasoning capabilities. Subsequently, we train a student model with the CoT objective, while compelling its `[SEG]` feature embedding f_s, to mimic the teacher's representation by minimizing the Mean Squared Error (MSE) between them:

$$\mathcal{L}_{dis} = \text{MSE}(f_t, f_s). \tag{2}$$

Please note that we use the SAM decoder from the pre-trained teacher model and freeze its parameters during the student's training. This strategy prevents potentially conflicting gradients from the complex reasoning task from altering the core segmentation module, ensuring that the student model acquires CoT reasoning skills without compromising the segmentation fidelity inherited from the teacher.

3.2 Stage2: Corrective Reflective-Style Training

While SFT with CoT data enables models to generate reasoned outputs and offers a cost-effective starting point, we identify a critical limitation: models often struggle with foundational perceptual and knowledge-based errors, such as misinterpreting audio cues (e.g., confusing the volume of a violin and a cello; ***detailed analysis is provided in the Supplementary Material***). Simply fine-tuning on correct CoT examples may not efficiently rectify these deep-seated biases. To address this, we introduce a corrective reflective-style training scheme. These structured examples explicitly present a common error followed by a detailed correction, effectively creating a "reflective" learning signal. Rather than teaching abstract self-reflection, this method's primary function is to refine the model's prior knowledge and calibrate its foundational perceptual abilities.

To endow our model with a more robust reasoning capability, we require high-quality examples of ideal thinking processes. We leverage a more powerful MLLM, Gemini, to serve as an expert annotator, whose role is to generate gold-standard reasoning paths by correcting flawed CoT samples. Specifically, we first prompt Gemini to critique the reasoning generated by our Stage 1 model. This process is focused on challenging samples from the training set, which we identify as those with both an IoU score below 0.6 and incorrect reasoning. This strict criterion ensures that corrective training targets only clear-cut failures where both the reasoning process and segmentation outcome are flawed. Adhering to the minimal modification principle (Yu et al. 2024), Gemini is instructed to rectify any identified flaws—such as incorrect conclusions or superficial analysis of audio information—by making the edits (e.g., modifying, adding, or deleting words). We denote the initial, potentially flawed CoT from the Stage 1 model as y_{wrong} and the corrected version from Gemini as $y_{correct}$. We then construct a "reflective path" as follows:

$$y_{reflective} = y_{wrong} \oplus x_{trigger} \oplus y_{correct}, \tag{3}$$

where $x_{trigger}$ is a reflective trigger phrase randomly sampled from a predefined collection (e.g., "Wait, let me re-evaluate..."). This path explicitly models the cognitive process of identifying a flaw and subsequently correcting it. Finally, we perform a second stage of fine-tuning on the model weights from our SFT stage. In this stage, we replace the standard target sequences y_t with our newly constructed $y_{reflective}$ paths, thereby directly injecting the desired "reflective" behavior into the model.

3.3 Stage3: Reinforcement Learning Training

Following the SFT stage, we incorporate reinforcement learning (RL) to further enhance the reasoning capability of the MLLM. Specifically, we adopt the GRPO (Guo et al. 2025) framework to enable self-refinement based on the relative quality of generated outputs within a group. However, GRPO was originally designed for text generation and does not directly align with the segmentation nature of Ref-AVS. To address this, we redesign the reward function tailored for segmentation as shown in the green block of Figure 2.

GRPO Given an input query q, GRPO samples a group of G candidate responses $o = \{o_1, ...o_G\}$ from the current policy π_θ. A rule-based reward model evaluates each response to produce scalar rewards $\{R_1, ..., R_G\}$. To assess the relative quality within the group, each reward is standardized to compute a normalized advantage for the i-th response:

$$\hat{A}_i = \frac{R_i - \text{mean}(\{R_i\}_{i=1}^G)}{\text{std}(\{R_i\}_{i=1}^G)}. \tag{4}$$

The optimization objective balances improving response quality with maintaining proximity to the previous policy

Method	Seen			Unseen			Mix (S+U)		
	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$	$\mathcal{J}\&\mathcal{F}\uparrow$	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$	$\mathcal{J}\&\mathcal{F}\uparrow$	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$	$\mathcal{J}\&\mathcal{F}\uparrow$
Audio-based Segmentation Methods									
AVSBench (Zhou et al. 2022)	23.2	51.1	37.2	32.4	54.7	43.5	27.8	52.9	40.3
AVSegFormer (Gao et al. 2024)	33.5	47.0	40.2	36.1	50.1	43.1	34.8	48.6	41.7
GAVS (Wang et al. 2024a)	28.9	49.8	39.4	29.8	49.7	39.8	29.4	49.8	39.6
Visual-based Segmentation Methods									
ReferFormer (Wu et al. 2022)	31.3	50.1	40.7	30.4	48.8	39.6	30.9	49.5	40.2
R2VOS (Li et al. 2023b)	25.0	41.0	33.0	27.9	49.8	38.9	26.5	45.4	35.9
Multi-modal Methods									
EEMC (Wang et al. 2024b)	34.2	51.3	42.8	49.5	64.8	57.2	41.9	58.1	50.0
SAM-LAVS (Wang et al. 2025)	43.5	51.9	47.7	66.5	72.3	69.4	55.0	62.1	58.5
TSAM (Radman and Laaksonen 2025)	43.4	56.8	50.1	54.6	66.4	60.5	49.0	61.6	55.3
Foundation-based Methods									
CRAB (Du et al. 2025)	40.5	58.0	49.3	45.6	63.0	54.3	43.1	60.5	46.2
AURORA(Ours)	**63.2**	**72.8**	**68.0**	**69.7**	**76.4**	**73.0**	**66.5**	**74.6**	**70.1**

Table 1: Performance comparison across different methods in Seen, Unseen, and Mix (S+U) settings of Ref-AVS benchmark. The mix indicates the average value of seen and unseen splits. "↑" indicates higher is better.

via a clipped objective. The final GRPO loss is defined as:

$$\mathcal{L}_{\text{GRPO}}(\theta) = \mathbb{E}_{(q,a)\sim\mathcal{D},\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}} \Bigg[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|} \min\Big(r_{i,t}(\theta)\hat{A}_{i,t}, \text{clip}\left(r_{i,t}(\theta), 1-\varepsilon, 1+\varepsilon\right)\hat{A}_{i,t}\Big) - \beta D_{\text{KL}}(\pi_\theta||\pi_{\text{ref}})\Bigg], \tag{5}$$

where $r_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\text{old}}(o_{i,t}|q,o_{i,<t})}$. Following (Yu et al. 2025), we set the KL divergence coefficient β to zero, effectively removing this term for simplicity.

Reward Design To effectively guide AURORA, we design a hybrid reward function consisting of Format Reward R_{format}, IoU Reward R_{IoU}, and Class Reward R_{class}. The coefficients for these reward components are all set to 1.

• **Format Reward** The format reward evaluates whether the output reasoning adheres to the required structural format. In tasks like mathematical problem-solving, reasoning is often followed by a variable-content textual answer enclosed in specific placeholders, such as `<answer></answer>`. While for our segmentation task, the goal is different. Although the model is still expected to generate a textual reasoning chain, the process must end with a fixed trigger phrase that directly leads to the generation of the segmentation mask. Therefore, we implement a simpler format constraint: if the final sentence of the reasoning process is `"It is [SEG]"`, the format reward is assigned a value of 1; otherwise, it receives a value of 0. This approach not only aligns conceptually with the segmentation task but also avoids introducing extraneous tokens absent from the vocabulary of our base model.

• **IoU Reward** Text-based rewards primarily optimize for linguistic coherence and semantic correctness, but they cannot directly assess the accuracy of the generated segmentation mask. Therefore, we introduce an IoU-based reward to measure segmentation mask quality. Specifically, we calculate the IoU between the predicted segmentation mask and the ground truth mask, which ranges from 0 to 1, to generate the reward signal, addressing a critical limitation that cannot be resolved through text generation rewards alone.

• **Class Reward** Simply introducing the IoU reward cannot directly supervise the correctness of reasoning; therefore, we introduce a class reward to evaluate the reasoning procedure. Since our CoT reasoning includes step 4, which outputs `"the target is {class_name}"`, we extract the `{class_name}` and compare it with the true class. If the `{class_name}` is consistent with the true class, we set the reward to 1; otherwise, we set it to 0. For simplicity, we do not consider synonyms in this evaluation.

4 Experiment

4.1 Experimental Settings

Dataset We evaluated our method on the Ref-AVS benchmark dataset (Wang et al. 2024b), which contains 4,000 videos with manual pixel-level annotations and expressions. The dataset is divided into a training set (2,908 videos), a validation set (276 videos), and a test set (818 videos). The test set is further split into three subsets: seen split, unseen split, and null split.

Implementation Details Our training unfolds in three stages. First, we perform an initial SFT for 720 steps using CoT reasoning generated by Qwen-Omni (Xu et al. 2025).

Method	Seen			Unseen		
	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
w/o MLLM	39.2	48.7	44.0	34.0	46.3	40.2
w/o CoT	54.1	64.0	59.1	64.2	71.9	68.1
CoT	**61.4**	**70.6**	**66.0**	**67.1**	**73.4**	**70.3**

Table 2: Ablation study of SFT training with CoT.

Method	Seen			Unseen		
	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
w/o $\mathcal{L}_{dis}$	60.2	70.2	65.2	65.7	72.4	69.1
with $\mathcal{L}_{dis}$	**61.4**	**70.6**	**66.0**	**67.1**	**73.4**	**70.3**

Table 3: Ablation study of segmentation feature distillation loss.

In the second stage, we conduct a 300-step corrective fine-tuning phase. For this, we curate 1,505 reflective-style examples by using Gemini 2.5-Pro to refine incorrect outputs from the first stage. These are then mixed with 3,500 standard CoT samples to prevent catastrophic forgetting. The final stage consists of 500 steps of GRPO (Guo et al. 2025) training, where we generate 3 candidate responses per input for preference alignment. We use a batch size of 2 with 1 sample per device and gradient accumulation steps of 4. All experiments are conducted on two A40 GPUs.

Evaluation Metrics In line with the evaluation protocol of Ref-AVS (Wang et al. 2024b), we employ the Jaccard index ($\mathcal{J}$), F-score ($\mathcal{F}$), and their average ($\mathcal{J}\&\mathcal{F}$) as the primary evaluation metrics.

4.2 Main Results

We compare AURORA with top SOTA methods on the Ref-AVS benchmark, including three audio-based (Zhou et al. 2022; Gao et al. 2024; Wang et al. 2024a), two visual-based (Wu et al. 2022; Li et al. 2023b), three multi-modal (Wang et al. 2024b, 2025; Radman and Laaksonen 2025), and one foundation-based method (Du et al. 2025). As shown in Table 1, AURORA achieves state-of-the-art performance across key metrics. On the seen test split, our model surpasses the second-best performing method by substantial margins. We attribute this significant leap in performance to our framework's core design: unlike prior methods that rely on simple modality fusion (Wang et al. 2024b, 2025; Radman and Laaksonen 2025) or train segmentation models from scratch (Du et al. 2025), AURORA leverages the power of a large language model for deep semantic reasoning and integrates it with a pre-trained foundation model for segmentation. More importantly, AURORA demonstrates even greater advantages on the challenging unseen test split. This consistent improvement on unseen categories indicates that AURORA possesses enhanced generalization capabilities, enabling it to effectively locate and segment novel objects not present in the training data. Figure 3 provides visual comparisons among the top-performing models.

Method	Seen			Unseen		
	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
SFT+GRPO	62.0	71.5	66.8	69.1	76.1	72.6
SFT+CT+GRPO	62.7	72.3	67.5	67.7	74.4	71.1
SFT+CRT+GRPO	**63.2**	**72.8**	**68.0**	**69.7**	**76.4**	**73.0**

Table 4: Ablation study of corrective reflective-style training.

Method	Seen			Unseen		
	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
with Reflective	61.8	71.1	66.5	67.2	73.9	70.6
R_{format} + R_{IoU}	62.4	72.2	67.3	68.3	75.4	71.9
R_{format} + R_{IoU} +R_{class}	**63.2**	**72.8**	**68.0**	**69.7**	**76.4**	**73.0**

Table 5: Ablation study of different designs in GRPO stage.

4.3 Ablation Study

Ablation study of SFT training with CoT To evaluate our method, we first compare it against a non-MLLM baseline where an ImageBind (Girdhar et al. 2023) text encoder processes simple prompts for SAM (Wang et al. 2025). As shown in Table 2 (w/o MLLM), our full model achieves superior performance, demonstrating the benefits of using a powerful MLLM backbone. To validate the contribution of CoT, we conduct an ablation study by training a variant of our model without the CoT component (w/o CoT). The resulting performance drop confirms that CoT is crucial for effective multimodal reasoning.

Effectiveness of Segmentation Feature Distillation Loss While our joint training approach is effective, it can create tension between the segmentation and reasoning tasks as they compete for shared parameters. Thus, we employ a feature distillation loss to instill robust segmentation capabilities by transferring knowledge from a specialist segmentation model. The results in Table 3 confirm that this loss maintains segmentation performance without harming reasoning, as evidenced by the results on the unseen split.

Effectiveness of Corrective Reflective-Style Training We evaluate our corrective reflective-style training by its final performance after our two-stage pipeline, as shown in Table 4. Our pipeline begins with an SFT stage to learn the $y_{reflective}$ format, followed by a GRPO stage to optimize this capability (SFT+CRT+GRPO). First, to demonstrate the value of incorporating reflective data, we compare our full pipeline against a baseline trained without any reflective examples (SFT+GRPO). The substantial performance gains of our method validate the overall effectiveness of our approach. Moreover, we test if performing SFT with only the correct answer $y_{correct}$ is enough (SFT+CT+GRPO). Our full reflective-style training still performs better. We believe this is because the reflective path explicitly shows the model both the mistake and the correct answer. This is more effective for rectifying the model's inherent biases than simply

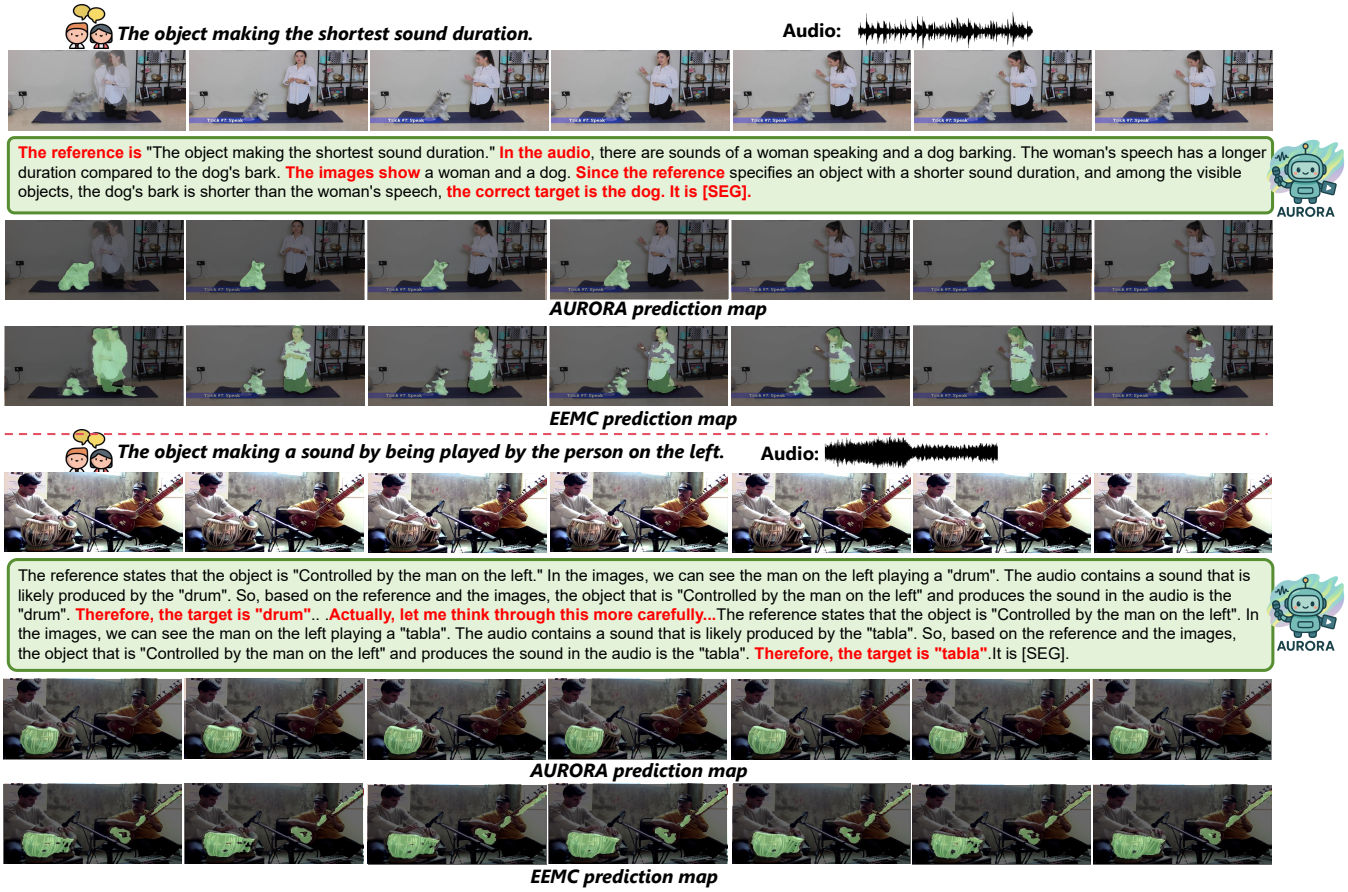

Figure 3: The visualization results of the referred objects in the Ref-AVS compared with EEMC (Wang et al. 2024b). Note that although the reasoning steps may appear in different orders to enhance diversity and support GRPO exploration, all outputs consistently contain the four key reasoning steps shown in Figure 1.

training on the correct answer alone.

Ablation study of GRPO We evaluate the effectiveness of our GRPO stage in Table 5, using the Stage 2 reflective model as our baseline (with Reflective). Our GRPO training first incorporates an R_{IoU} reward, which yields substantial improvements over the reflective model. To enhance performance on difficult samples, we introduce an additional R_{class} reward, which achieves further gains. These improvements confirm that our reward design effectively enhances segmentation by refining the model's reasoning abilities. For all GRPO experiments, a simple R_{format} reward was used as a training aid to ensure the model consistently generates the required `[SEG]` token, but it does not directly contribute to the final performance metrics.

4.4 Cross-Task Generalization Analysis

To evaluate the transferability of our model's learned representations, we assessed its performance on the AVS-Bench dataset (Zhou et al. 2022), which focuses on generic audio-visual segmentation without reference guidance. This task represents a significant departure from our training paradigm, as it requires identifying salient sounding objects rather than following specific textual references. To adapt our model, we guided the model with the generic prompt, "The sounding object," and exclusively fine-tuned our GRPO-trained model for 3 epochs. Crucially, we did not generate new CoT samples or reflective paths for the fine-tuning process. Our adapted model achieves competitive performance on the $\mathcal{J}$ metric, reaching 77.3 compared to 73.3 of the fully fine-tuned Crab model (Du et al. 2025), while maintaining comparable $\mathcal{F}$ scores (86.7 vs. 86.8).

5 Conclusion

We propose AURORA, a framework that endows Ref-AVS with authentic reasoning. First, we introduce a structured CoT prompting mechanism during SFT to build a strong foundation for reasoning. To mitigate the conflict between reasoning and segmentation during joint training, we introduce a feature distillation loss that preserves pixel-level precision. To elevate the model's capabilities from simple rationalization to authentic introspection, we further develop a two-stage refinement strategy combining reflective learning and GRPO-based reinforcement learning. AURORA achieves state-of-the-art performance on Ref-AVS benchmarks and generalizes well to the AVS task.

Acknowledgments

This work was supported in part by the Noncommunicable Chronic Diseases-National Science and Technology Major Project (Grant 2023ZD0500903, 2023ZD0500900), the National Natural Science Foundation of China under Grant 62136007, 62036011, U20B2065, 6202781, 62036005, 62293543.

References

Bhosale, S.; Yang, H.; Kanojia, D.; Deng, J.; and Zhu, X. 2024. Unsupervised Audio-Visual Segmentation with Modality Alignment. *arXiv preprint arXiv:2403.14203*.

Bhosale, S.; Yang, H.; Kanojia, D.; and Zhu, X. 2023. Leveraging foundation models for unsupervised audio-visual segmentation. *arXiv preprint arXiv:2309.06728*.

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

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

Du, H.; Li, G.; Zhou, C.; Zhang, C.; Zhao, A.; and Hu, D. 2025. Crab: A Unified Audio-Visual Scene Understanding Model with Explicit Cooperation. *arXiv preprint arXiv:2503.13068*.

Gao, S.; Chen, Z.; Chen, G.; Wang, W.; and Lu, T. 2024. Avsegformer: Audio-visual segmentation with transformer. In *AAAI*, volume 38, 12155–12163.

Girdhar, R.; El-Nouby, A.; Liu, Z.; Singh, M.; Alwala, K. V.; Joulin, A.; and Misra, I. 2023. Imagebind: One embedding space to bind them all. In *CVPR*, 15180–15190.

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

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *ICCV*, 4015–4026.

Lai, X.; Tian, Z.; Chen, Y.; Li, Y.; Yuan, Y.; Liu, S.; and Jia, J. 2024. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9579–9589.

Li, K.; Yang, Z.; Chen, L.; Yang, Y.; and Xiao, J. 2023a. Catr: Combinatorial-dependence audio-queried transformer for audio-visual video segmentation. In *ACMMM*, 1485–1494.

Li, X.; Wang, J.; Xu, X.; Li, X.; Raj, B.; and Lu, Y. 2023b. Robust referring video object segmentation with cyclic structural consensus. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22236–22245.

Li, X.; Wang, J.; Xu, X.; Peng, X.; Singh, R.; Lu, Y.; and Raj, B. 2024. QDFormer: Towards Robust Audiovisual Segmentation in Complex Environments with Quantization-based Semantic Decomposition. In *CVPR*, 3402–3413.

Ling, Y.; Li, Y.; Gan, Z.; Zhang, J.; Chi, M.; and Wang, Y. 2024. TransAVS: End-to-End Audio-Visual Segmentation with Transformer. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7845–7849. IEEE.

Liu, Y.; Peng, B.; Zhong, Z.; Yue, Z.; Lu, F.; Yu, B.; and Jia, J. 2025. Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement. *arXiv preprint arXiv:2503.06520*.

Luo, Z.; Liu, N.; Yang, X.; Khan, S.; Anwer, R. M.; Cholakkal, H.; Khan, F. S.; and Han, J. 2025a. TAViS: Text-bridged Audio-Visual Segmentation with Foundation Models. *arXiv preprint arXiv:2506.11436*.

Luo, Z.; Liu, N.; Yang, X.; Khan, S.; Anwer, R. M.; Cholakkal, H.; Khan, F. S.; and Han, J. 2025b. TAViS: Text-bridged Audio-Visual Segmentation with Foundation Models. In *ICCV*, 24014–24023.

Radman, A.; and Laaksonen, J. 2025. TSAM: Temporal SAM Augmented with Multimodal Prompts for Referring Audio-Visual Segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 23947–23956.

Wang, Y.; Liu, W.; Li, G.; Ding, J.; Hu, D.; and Li, X. 2024a. Prompting segmentation with sound is generalizable audio-visual source localizer. In *AAAI*, volume 38, 5669–5677.

Wang, Y.; Sun, P.; Zhou, D.; Li, G.; Zhang, H.; and Hu, D. 2024b. Ref-avs: Refer and segment objects in audio-visual scenes. In *European Conference on Computer Vision*, 196–213. Springer.

Wang, Y.; Xu, H.; Liu, Y.; Li, J.; and Tang, Y. 2025. SAM2-LOVE: Segment Anything Model 2 in Language-aided Audio-Visual Scenes. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 28932–28941.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Wu, J.; Jiang, Y.; Sun, P.; Yuan, Z.; and Luo, P. 2022. Language as queries for referring video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4974–4984.

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

Yang, Q.; Nie, X.; Li, T.; Gao, P.; Guo, Y.; Zhen, C.; Yan, P.; and Xiang, S. 2024. Cooperation Does Matter: Exploring Multi-Order Bilateral Relations for Audio-Visual Segmentation. In *CVPR*, 27134–27143.

Yao, J.; Guo, G.; Zheng, Z.; Xie, Q.; Han, L.; Zhang, D.; and Han, J. 2025. Prompting Vision-Language Model for Nuclei Instance Segmentation and Classification. *IEEE Transactions on Medical Imaging*.

Yu, Q.; Zhang, Z.; Zhu, R.; Yuan, Y.; Zuo, X.; Yue, Y.; Fan, T.; Liu, G.; Liu, L.; Liu, X.; et al. 2025. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*.

Yu, T.; Yao, Y.; Zhang, H.; He, T.; Han, Y.; Cui, G.; Hu, J.; Liu, Z.; Zheng, H.-T.; Sun, M.; et al. 2024. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13807–13816.

Zhang, D.; Li, H.; He, D.; Liu, N.; Cheng, L.; Wang, J.; and Han, J. 2025. Unsupervised Pre-training with Language-Vision Prompts for Low-Data Instance Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhou, J.; Shen, X.; Wang, J.; Zhang, J.; Sun, W.; Zhang, J.; Birchfield, S.; Guo, D.; Kong, L.; Wang, M.; et al. 2023. Audio-visual segmentation with semantics. *arXiv preprint arXiv:2301.13190*.

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