

Learning to Reason Before Masking for Referring Audio-Visual Segmentation

Zhixuan Wu¹, Ye Ni², Yiming Chen³, Wenyan Gu¹, Bo Cheng^{1*}

¹State Key Laboratory of Networking and Switching Technology,
Beijing University of Posts and Telecommunications,

²Southeast University, ³Huaqiao University

wzxmogu@bupt.edu.cn, niye@seu.edu.cn,

chenyiming@stu.hqu.edu.cn, {guwenyan, chengbo}@bupt.edu.cn

Abstract

Existing referring Audio-Visual Segmentation (Ref-AVS) benchmarks primarily focus on explicit expressions whose referents can be resolved through direct visual, acoustic, or linguistic correspondence. Consequently, they provide limited coverage of implicit references that require multi-step reasoning over multimodal evidence. To bridge this gap, we introduce **Reasoning Ref-AVS**, a new task that requires models to infer the referred object from audio-visual-language cues before generating its segmentation mask. We further construct **ReAVS**, a comprehensive dataset with manually annotated referring expressions, multimodal cues, object masks, and reasoning chains, covering diverse reasoning patterns. Evaluations of representative Ref-AVS models on ReAVS reveal substantial performance degradation when direct cross-modal correspondence is insufficient, exposing a clear gap between multimodal recognition and referent reasoning. Motivated by these challenges, we propose **Re2Seg**, an omni-modal framework that performs object-centric evidence encoding, multi-hop reasoning, and reasoning-conditioned mask decoding. Experiments demonstrate that Re2Seg achieves stronger segmentation performance under reasoning-intensive references and provides interpretable object-inference traces. These results highlight the importance of learning to reason before masking for bridging the gap between multimodal recognition and referring segmentation.

1 Introduction

Referring segmentation (Ding et al., 2025a; He et al., 2025) aims to localize and segment target objects from user-specified expressions, serving as a fundamental task for language-guided object grounding. Recent advances have extended this paradigm from static images to more complex

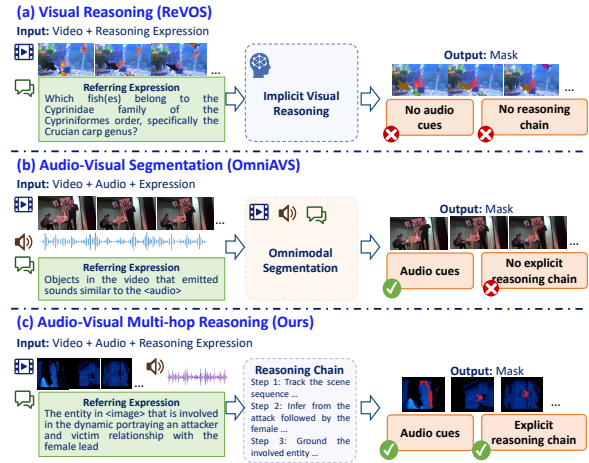


Figure 1: Comparison of Reasoning Ref-AVS with related referring segmentation tasks.

video scenarios. As shown in Fig. 1(a), video segmentation (Zheng et al., 2025; Xu et al., 2026; Yan et al., 2024a), such as ReVOS (Yan et al., 2024a), requires models to identify and segment target objects from visual-language evidence, but does not account for acoustic cues. This limitation becomes particularly pronounced when the referred object can only be identified by integrating visual, linguistic, and auditory information. Referring audio-visual segmentation (Wang et al., 2024; Ying et al., 2025; Zhou et al., 2026) addresses this limitation by incorporating audio signals into the grounding process. However, as illustrated in Fig. 1(b), existing approaches such as OmniAVS (Ying et al., 2025) primarily focus on direct audio-visual-language grounding, where the object can often be identified from explicit categories, attributes, actions, or sound-source correspondences. In contrast, complex referring expressions require models to combine multiple pieces of heterogeneous evidence and perform intermediate reasoning before determining the referred object. Existing methods largely treat this process as an implicit function of end-to-end

*Corresponding author.

prediction, without explicitly modeling or supervising the reasoning steps that connect multimodal evidence to the target object.

However, in real-world settings, referring expressions are often implicit and event-dependent, requiring models to infer the target from temporally accumulated audio-visual evidence (Liang et al., 2025; Liu et al., 2026). This exposes a limitation of existing systems, which may rely on shallow cue matching and segment salient objects without resolving the underlying referential intent. Addressing this limitation requires moving beyond perception-level grounding and evaluating whether models can reason over intermediate audio-visual evidence before mask prediction. As shown in Fig. 1(c), we therefore formulate **Reasoning Ref-AVS**, which shifts Ref-AVS from direct object localization toward reasoning-based referent inference. Existing benchmarks (Radman and Laaksonen, 2025; Seo et al., 2020a; Khoreva et al., 2018) typically provide only referring expressions and segmentation masks, without explicit annotations of reasoning chains, reasoning types, reasoning hops, or modality dependencies. This lack of diagnostic supervision makes it difficult to determine whether failures stem from object inference, temporal grounding, acoustic evidence modeling, or mask prediction.

To address these challenges, we introduce **ReAVS**, a fine-grained benchmark for Reasoning Ref-AVS. ReAVS evaluates whether models can infer and segment referred objects from implicit and evidence-dependent expressions rather than relying on direct textual, visual, or acoustic matching. Each sample contains an audio-visual video, a referring expression, object masks, and structured reasoning annotations covering reasoning types, reasoning hops, temporal dependencies, audio dependencies, and multimodal reference cues. These annotations enable diagnostic evaluation beyond mask quality and allow systematic analysis of model behavior under different reasoning conditions. We further propose **Re2Seg**, a reasoning-aware framework for audio-visual object segmentation. Rather than directly mapping multimodal inputs to masks, Re2Seg first performs entity-aware evidence encoding and chain-supervised object inference, followed by reasoning-conditioned mask decoding. By introducing reasoning chains as intermediate supervision, Re2Seg explicitly connects audio-visual-language reasoning with pixel-level object segmentation and provides a unified framework for multimodal evidence integration and mask

prediction. Overall, this work makes the following main contributions:

- **Task.** We formulate **Reasoning Ref-AVS**, a reasoning-oriented referring audio-visual segmentation task that requires models to infer the referred object from multi-step multimodal evidence before predicting its mask.
- **Benchmark.** We introduce **ReAVS**, a fine-grained benchmark with audio-visual videos, referring expressions, object masks, and structured reasoning annotations that cover diverse reasoning types, reasoning hops, temporal dependencies, audio dependencies, and multimodal reference cues.
- **Method.** We propose **Re2Seg**, a framework that integrates an omni-modal evidence encoder, MLLM-based multi-hop reasoning, and reasoning-conditioned mask decoding. Experiments on ReAVS and existing benchmarks demonstrate its effectiveness in referring scenarios that require multimodal reasoning.

2 Related Work

Audio-Visual Segmentation. Audio-visual segmentation (AVS) aims to localize and segment sound-related objects by exploiting the correspondence between visual content and acoustic signals, providing a fundamental step toward fine-grained audio-visual scene understanding (Zhou et al., 2022; Gao et al., 2024). Recent studies (Wang et al., 2024) further extend AVS to referring audio-visual segmentation (Ref-AVS), where models are required to segment the object specified by a referring expression with audio-visual cues. Existing Ref-AVS methods (Radman and Laaksonen, 2025; Wang et al., 2025) either learn unified audio-video-language representations for mask decoding, or employ MLLMs (Ying et al., 2025; Zhou et al., 2026; Zhao et al., 2026b) to infer the referred object before grounding and segmentation. However, they mostly rely on explicit audio-visual-language matching, leaving implicit references that require multi-hop object inference insufficiently explored. More details in Appendix. A.

Multi-Hop Reasoning. Multi-hop reasoning has been widely explored in question answering and vision-language understanding, where models are required to aggregate multiple pieces of evidence

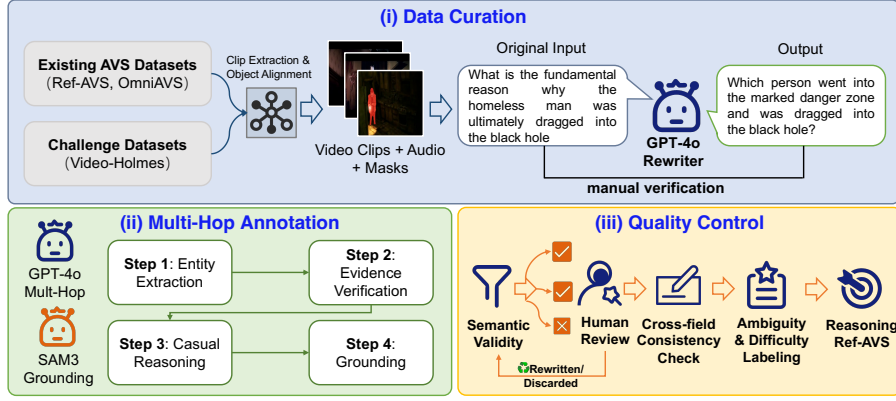


Figure 2: **Overview of the ReAVS benchmark.** (i) Data Curation, which collects audio-visual videos and converts explicit descriptions into implicit expressions; (ii) Multi-Hop Annotation, which constructs reasoning chains and semantic metadata with GPT-4o and SAM3 assistance; (iii) Quality Control, which verifies object consistency, evidence faithfulness, and annotation quality through human review.

and perform compositional inference over entities, relations, events, or commonsense knowledge (Trivedi et al., 2022; Schwenk et al., 2022). Recent chain-of-thought methods (Wei et al., 2022) further improve reasoning transparency by decomposing complex queries into intermediate steps, and have been extended to multimodal reasoning, visual grounding, and segmentation (Xu et al., 2025a; Jiang et al., 2025). Nevertheless, existing studies mainly address text-only QA, image-level reasoning, or box-level object grounding, leaving multi-hop reasoning for audio-visual mask prediction largely under-explored.

3 Reasoning Ref-AVS Benchmark

3.1 Task Setting

Given a video sequence $\mathcal{V} = \{v_t\}_{t=1}^T$ with synchronized audio $\mathcal{A}$ and a reasoning-intensive referring expression $\mathcal{Q}$, Reasoning Ref-AVS aims to segment the object referred to by $\mathcal{Q}$ across the video. The model outputs a sequence of binary masks $\hat{\mathcal{M}} = \{\hat{M}_t\}_{t=1}^T$, where $\hat{M}_t \in \{0, 1\}^{H \times W}$ denotes the predicted object region at frame t . We formulate the task as a two-stage process:

$$o^* = f_{\theta}(\mathcal{V}, \mathcal{A}, \mathcal{Q}), \quad \hat{\mathcal{M}} = g_{\theta}(\mathcal{V}, \mathcal{A}, \mathcal{Q}, o^*), \quad (1)$$

where f_{θ} infers the referred object from multimodal evidence, and g_{θ} predicts its mask sequence conditioned on the inferred object.

3.2 Benchmark Construction

As shown in Fig. 2, Reasoning Ref-AVS is constructed through a three-step pipeline to obtain samples that require evidence-dependent object inference rather than direct multimodal cue matching.

Data Curation. We initially collect a pool of candidate videos, which are manually reviewed and filtered to ensure that each video provides clear and unambiguous evidence for reliable reasoning-based reference construction. This process yields a carefully curated collection of 2,261 video clips covering 4,046 annotated objects and 49,762 reasoning-intensive referring expressions with corresponding explanations. As shown in Fig. 2(i), the videos are collected from two complementary domains: existing audio-visual segmentation datasets processed following Ref-AVS (Wang et al., 2024) and OmniAVS (Ying et al., 2025), and Video-Holmes (Cheng et al., 2025), which contains videos involving temporal, causal, and commonsense reasoning. We manually filter videos with severe occlusion, low visual quality, ambiguous objects, weak audio-visual correspondence, or insufficient reasoning evidence. For each retained video, GPT-4o (Hurst et al., 2024) is prompted to rewrite the original question or description into an implicit referring expression, while the original explanation is retained as evidence-grounded supervision. All generated expressions are manually reviewed to ensure that the referred object cannot be resolved by direct category, appearance, location, action, or sound-source matching alone. Detailed are provided in Appendix B.3 and B.4.

Multi-Hop Annotation. We define a reasoning hop as an evidence-dependent transition necessary for resolving the referred object. Each hop introduces a new evidence unit or inference relation, such as entity-action association, role inference, temporal ordering, or causal reasoning. Pure mask localization after object inference is not counted

Dataset	Content		Referring Cue			Reasoning Annotation					Statistics			
	Video	Audio	Text	Audio	Image	Avail.	Multi-hop	Chain	Expl.	Temp.	Video	Object	Expr.	Expl.
AVSBench (Zhou et al., 2022)	●	●	○	○	○	○	○	○	○	●	5,356	—	—	—
Ref-AVS (Wang et al., 2024)	●	●	●	○	○	○	○	○	○	●	4,002	6,888	20,261	—
OmniAVS (Ying et al., 2025)	●	●	●	●	●	●	●	○	●	●	2,104	4,277	61,095	34,841
ReVOS (Yan et al., 2024a)	●	○	●	○	○	●	●	○	○	●	1,042	5,535	35,074	—
R ² -AVSBench (Zhou et al., 2026)	●	●	●	○	○	●	●	○	○	●	400	—	400	—
ReAVS	●	●	●	●	●	●	●	●	●	●	2,261	4,046	49,762	49,762

Table 1: **Comparison of ReAVS with existing audio-visual segmentation datasets.** Avail., Expr., Expl., and Temp. denote reasoning availability, referring expressions, explanations, and temporal annotations, respectively. ●, ●, and ○ indicate full support, limited or implicit support, and no support.

Annotation Field	Agreement Metric	Score
Reasoning Type	Cohen’s κ	0.78
Reasoning-Hop Count	Weighted Cohen’s κ	0.71
Temporal Dependency	Cohen’s κ	0.86
Audio Dependency	Cohen’s κ	0.88
Difficulty Level	Weighted Cohen’s κ	0.78
Evidence Level	Weighted Cohen’s κ	0.77
Final Referent Identity	Exact agreement	0.91

Table 2: Inter-annotator agreement analysis.

as a hop. As shown in Fig. 2(ii), we prompt GPT-4o (Hurst et al., 2024) to extract key reasoning elements from each audio-visual clip and its referring expression: (1) the entities or events involved in the expression; (2) the intermediate visual, audio, or semantic evidence required for resolving the reference; (3) the relation, causal, temporal, or commonsense inference connecting the evidence to the object; and (4) the final groundable object or person corresponding to the segmentation mask. Annotators then verify and organize them into step-wise reasoning chains covering entity/event extraction, evidence verification, relation reasoning, and object grounding. Reasoning-relevant entities are linked to spatial masks using existing annotations or SAM3 (Carion et al., 2025), and temporal evidence is localized with core-frame captions and Video-Holmes annotations.

Quality Control. We remove invalid samples whose objects are absent, too small, severely occluded, or not observable within the relevant temporal window. We further discard expressions solvable by direct matching, and verify consistency among the expression, explanation, reasoning chain, object text, object ID, and mask. The final reasoning step must align with the annotated object mask. Ambiguous samples with multiple plausible objects are revised or removed. Each sample is independently reviewed by two annotators, with disagreements resolved by an expert annotator.

3.3 Benchmark Statistics and Analysis

As shown in Fig. 3 and Tab. 1, ReAVS provides a fine-grained benchmark for evaluating reasoning-intensive referring audio-visual segmentation. It contains 2,261 videos, 4,046 objects, 49,762 referring expressions, and 49,762 explanations, offering dense supervision for both object segmentation and reasoning evaluation. More details in Appendix B.

Fig. 3(a) shows that Reasoning Ref-AVS covers diverse reasoning types, including spatial (Spa.), temporal (Temp.), interaction (Inter.), behavioral (Behav.), social (Social), causal (Causal), appearance (App.), functional (Func.), intentional (Intent.), identity (Id.) reasoning. Functional and behavioral reasoning account for relatively large proportions, suggesting that many expressions require understanding object roles, actions, and high-level semantics rather than relying on direct appearance matching. The benchmark also emphasizes multi-step inference: 79.6% of the samples require multi-hop reasoning, with 39.8% involving three or more hops. Fig. 3(b) reflects realistic referring scenarios where users may specify objects through natural language, speech, acoustic evidence, visual references, or mixed multimodal cues. This reflects realistic referring scenarios in which object inference may depend on heterogeneous evidence sources. Fig. 3(d) further reports the difficulty-evidence distribution. Most samples fall into medium difficulty with partial evidence, indicating that models often need to integrate incomplete audio-visual clues rather than match a single explicit cue.

As summarized in Tab. 1, ReAVS provides more comprehensive reasoning-oriented annotations than existing audio-visual segmentation benchmarks, including multimodal referring cues, multi-hop labels, reasoning chains, explanations, and temporal annotations. These properties make ReAVS a suitable testbed for studying how multimodal models resolve implicit references, integrate

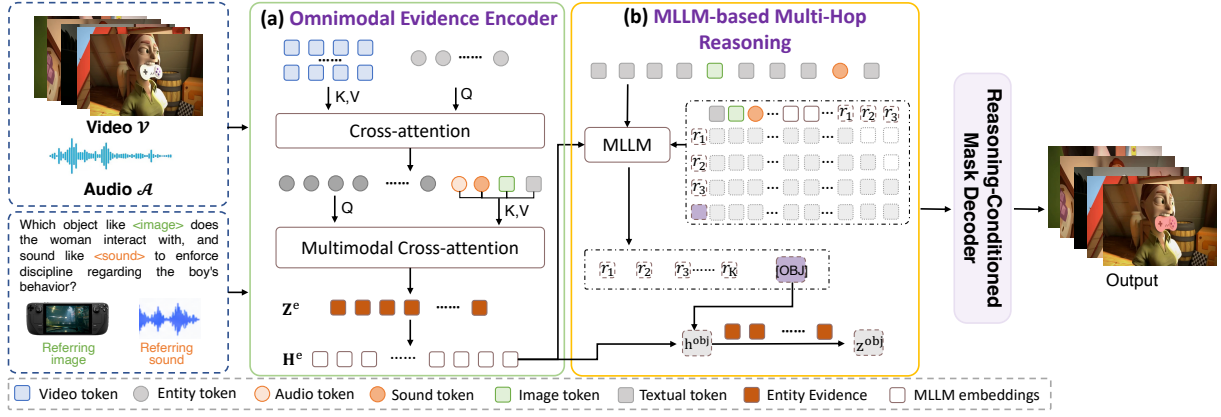


Figure 4: **Overview of Re2Seg.** Re2Seg constructs candidate-entity evidence, performs entity-conditioned continuous reasoning with an MLLM, and re-grounds the resulting [OBJ] state into the entity-evidence space to obtain a reasoning-aware object representation for frame-wise segmentation.

4.3 MLLM-based Multi-Hop Reasoning

We introduce an **MLLM-based Multi-Hop Reasoning** module to infer the referred object entity before mask decoding. Given the entity-evidence tokens $\mathbf{Z}^e$ from the Omnimodal Evidence Encoder, each token represents a candidate object enriched with visual appearance, temporal context, acoustic cues, and query semantics. The reasoner compares and integrates these heterogeneous evidence cues to determine which candidate entity best satisfies the referring expression. This process explicitly decomposes referent inference into a sequence of intermediate reasoning steps, allowing the model to resolve implicit references that cannot be determined from a single modality or isolated cue.

To enable language-based reasoning over entity-level evidence, we project $\mathbf{Z}^e$ into the MLLM embedding space $\mathbf{H}^e = \text{Proj}_e(\mathbf{Z}^e)$, and insert the projected embeddings as candidate entity tokens into the MLLM input. Conditioned on these object tokens and the referring expression $\mathcal{Q}$, the MLLM decomposes object inference into intermediate reasoning steps, including evidence localization, candidate comparison, cross-modal verification, relation reasoning, and object grounding.

During training, annotated reasoning chains provide intermediate supervision for object inference. We optimize the MLLM to generate the reasoning chain $\mathcal{R}$ followed by a special object token [OBJ]:

$$\mathcal{L}_{\text{reason}} = -\log p(\mathcal{R}, [\text{OBJ}] \mid \mathbf{H}^e, \mathcal{Q}). \quad (4)$$

[OBJ] token is appended after the reasoning chain, allowing its contextualized hidden state to aggregate information from the preceding reasoning

steps and multimodal entity evidence. This supervision encourages the model to associate reasoning steps with entity-level audio-visual evidence, rather than directly mapping the referring expression to a segmentation query. After decoding, we extract the hidden state of [OBJ] and use it to attend over the entity-evidence tokens. This attention-based matching softly selects the entity evidence most relevant to the inferred referent, yielding a reasoning-aware object representation $\mathbf{z}^{obj}$. The resulting $\mathbf{z}^{obj}$ is passed to the subsequent mask decoder for reasoning-conditioned mask prediction.

4.4 Reasoning-Conditioned Mask Decoding

The MLLM-based reasoning module outputs a reasoning-aware object representation $\mathbf{z}^{obj}$, which summarizes the inferred referent from the generated [OBJ] token and entity-level evidence. We use $\mathbf{z}^{obj}$ as the object query for mask decoding, so that pixel-level segmentation is conditioned on the object inferred through multi-hop reasoning. Given the video sequence $\mathcal{V} = \{I_t\}_{t=1}^T$, the decoder predicts a sequence of masks $\hat{\mathcal{M}} = \{\hat{M}_t\}_{t=1}^T$, where $\hat{M}_t \in [0, 1]^{H \times W}$ denotes the foreground probability map for frame I_t . Specifically, the decoder takes $\mathbf{z}^{obj}$ and multi-scale visual features as input. Through cross-attention, the object query interacts with frame-level visual features to produce reasoning-conditioned mask embeddings. The same object representation is shared across frames, while frame-specific visual features adapt the prediction to temporal appearance changes. This shared object representation provides a temporally consistent semantic query, while frame-specific visual features enable per-frame localization.

Method	Overall Performance					Fine-grained Reasoning Performance ($\mathcal{J}\&\mathcal{F}$)											
	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	OC	$\mathcal{S}$	Spa.	Temp.	Inter.	Behav.	Social	Causal	App.	Func.	Intent.	Id.		
<i>Specialized audio-visual segmentation methods</i>																	
AVSBench (Zhou et al., 2022)	24.3	26.1	25.2	–	–	30.4	21.1	23.8	26.7	21.4	18.5	32.2	23.5	20.8	33.1		
AVSegFormer (Gao et al., 2024)	28.5	30.7	29.6	–	–	34.8	24.6	28.2	32.5	26.9	22.8	35.4	27.9	25.9	37.3		
<i>General-purpose audio-capable MLLMs</i>																	
Qwen2-Audio-7B + SAM3 (Chu et al., 2023)	44.2	48.9	46.6	45.2	3.10	48.2	45.5	47.1	48.0	44.5	42.1	52.3	46.0	43.2	49.5		
Qwen2.5-Omni-7B + SAM3 (Xu et al., 2025b)	47.8	54.2	51.0	50.9	3.45	52.1	50.2	51.8	52.5	48.9	46.5	56.8	50.4	47.7	53.4		
<i>MLLM-based reasoning segmentation methods</i>																	
LISA-7B (Lai et al., 2024)	32.1	35.4	33.8	26.4	1.85	38.2	28.5	33.1	35.4	31.6	27.1	39.8	31.7	30.9	41.2		
VISA-7B (Yan et al., 2024b)	36.4	41.2	38.8	35.8	2.21	41.5	36.2	40.9	39.1	35.4	32.8	44.1	36.5	34.5	46.6		
OISA-1B (Ying et al., 2025)	47.2	53.6	50.4	40.2	2.45	51.4	49.8	51.6	51.2	49.5	47.3	56.0	48.7	45.2	53.3		
Ours (Re2Seg)	51.5	56.2	53.9	54.5	3.78	53.5	53.9	54.2	53.5	52.4	51.7	58.1	52.9	53.0	55.9		

Table 3: **Overall and fine-grained performance comparison on the ReAVS benchmark.** The left block reports segmentation quality and reasoning-chain metrics, while the right block reports $\mathcal{J}\&\mathcal{F}$ across ten reasoning types.

For each frame I_t , the decoder projects the corresponding mask embedding onto dense visual features to obtain the predicted mask:

$$\hat{M}_t = \sigma(\mathbf{P}_t \cdot \mathbf{m}_t), \quad (5)$$

where $\mathbf{P}_t$ denotes pixel-level visual features, $\mathbf{m}_t$ is the reasoning-conditioned mask embedding, and $\sigma(\cdot)$ is the sigmoid function.

For training, let $\mathcal{M} = \{M_t\}_{t=1}^T$ denote the ground-truth mask sequence, where $M_t \in \{0, 1\}^{H \times W}$. Following prior segmentation methods (Yan et al., 2024b), we supervise mask prediction with binary cross-entropy loss and Dice loss. The full objective is:

$$\mathcal{L} = \lambda_{\text{bce}} \mathcal{L}_{\text{bce}} + \lambda_{\text{dice}} \mathcal{L}_{\text{dice}} + \lambda_{\text{reason}} \mathcal{L}_{\text{reason}}, \quad (6)$$

where λ_{bce} , λ_{dice} , and λ_{reason} are balancing coefficients.

4.5 Inference

During inference, Re2Seg only requires the video $\mathcal{V}$, audio $\mathcal{A}$, and referring expression $\mathcal{Q}$. The Omnimodal Evidence Encoder first produces entity-evidence representations $\mathbf{Z}^e$, which are projected into the MLLM embedding space as candidate entity tokens. Conditioned on these tokens and $\mathcal{Q}$, the MLLM autoregressively generates a reasoning chain followed by the special object token [OBJ]. We then extract the hidden state of [OBJ] and align it with the entity-evidence representations to obtain the reasoning-aware object representation $\mathbf{z}^{obj}$. This representation summarizes the object inferred by the model through multi-hop audio-visual-language reasoning. Finally, the mask decoder uses $\mathbf{z}^{obj}$ as the object query and predicts the mask sequence $\hat{\mathcal{M}} = \{\hat{M}_t\}_{t=1}^T$. The

generated reasoning chain can be retained for interpretability, but mask prediction depends only on the model-inferred object representation rather than any ground-truth reasoning annotation.

5 Experiments

5.1 Experimental Setup

Datasets. ReAVS benchmark contains 2,261 videos and 49,762 reasoning referring samples with object masks. We split the dataset into training and test sets with no video overlap across splits. ReAVS covers diverse query modalities, including text, speech, image, sound, and their combinations, and supports evaluation across reasoning types, reasoning hops, query modalities, and difficulty levels defined by the required evidence. Structured reasoning chains and key evidence annotations are initially generated with the assistance of GPT-4o and subsequently verified through consistency checks and manual quality control. More benchmark details are provided in Appendix B.

Evaluation Metrics. Following previous work, we report $\mathcal{J}\&\mathcal{F}$ as the primary mask-quality metric, where $\mathcal{J}$ denotes region similarity measured by mask Intersection-over-Union (IoU), and $\mathcal{F}$ measures contour accuracy via bipartite matching. To evaluate the cognitive reasoning process behind the segmentation, we report Object Consistency (OC), which measures whether the final object inferred from the generated reasoning chain matches the ground-truth object category and ID. We further evaluate the quality of intermediate reasoning steps using an LLM-as-Judge protocol and report the reasoning score $\mathcal{S}$. Implementation details and judging prompts are provided in Appendix B.3.

Implementation Details. To comprehensively

Method	Ref-AVS			OmniAVS			MeViS			ReVOS		
	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
AVSBench (Zhou et al., 2022)	27.8	52.9	40.3	24.5	28.3	26.4	18.2	21.4	19.8	—	—	—
AVSegFormer (Gao et al., 2024)	34.8	48.6	41.7	28.1	32.5	30.3	21.5	24.9	23.2	—	—	—
LISA-7B (Lai et al., 2024)	44.1	50.3	47.2	29.1	37.5	33.3	35.1	39.4	37.2	39.1	42.7	40.9
VISA-7B (Yan et al., 2024b)	48.2	54.8	51.5	34.6	40.4	37.5	40.7	46.3	43.5	43.9	48.2	46.1
OISA-1B (Ying et al., 2025)	54.5	61.4	58.0	37.9	43.5	40.7	40.9	46.7	43.8	44.5	48.9	46.7
Ours (Re2Seg)	56.0	62.2	59.1	39.2	43.6	41.4	41.5	47.7	44.6	45.4	49.2	47.3

Table 4: **Performance comparison on the testing sets of Ref-AVS, OmniAVS, MeViS, and ReVOS benchmarks.** We report $\mathcal{J}$, $\mathcal{F}$, and $\mathcal{J}\&\mathcal{F}$ for each benchmark.

evaluate model capabilities on reasoning-guided referring audio-visual segmentation, we test all methods on the proposed ReAVS benchmark under the same evaluation protocol. Our Re2Seg is built upon OISA-1B (Ying et al., 2025), with the multi-modal perception and mask prediction components initialized from pretrained OISA-1B weights. We train our model on ReAVS. The learning rate is set to 2×10^{-5} with a warmup ratio of 0.03, and the model is trained for one epoch with a global batch size of 16. Training and evaluation are performed on 4 NVIDIA A100 GPUs.

5.2 Results on Benchmark

ReAVS Benchmark Results. Table 3 reports the overall and fine-grained results on ReAVS. Traditional audio-visual segmentation methods (Zhou et al., 2022; Gao et al., 2024) are evaluated only with mask-quality metrics, as they do not produce explicit reasoning chains. General-purpose audio-capable MLLMs achieve considerably stronger performance, with Qwen2.5-Omni-7B with SAM3 (Xu et al., 2025b; Carion et al., 2025) obtaining the best baseline results on the overall metrics, i.e., 51.0 $\mathcal{J}\&\mathcal{F}$, 50.9 OC, and 3.45 $\mathcal{S}$, demonstrating the benefit of large-scale omnimodal pretraining. Nevertheless, these models remain limited in explicitly identifying and grounding the referred object. Re2Seg achieves the best performance across all overall metrics and all ten reasoning categories. Compared with the strongest overall baseline, Qwen2.5-Omni-7B with SAM3, Re2Seg improves $\mathcal{J}\&\mathcal{F}$ by 2.9, OC by 3.6, and $\mathcal{S}$ by 0.33. Compared with the MLLM-based segmentation baseline OISA-1B (Ying et al., 2025), the corresponding gains are 3.5, 14.3, and 1.33, respectively. These results validate the effectiveness of entity-aware evidence modeling and chain-supervised object inference for reasoning-intensive

Variant	Segmentation	Reasoning	
	$\mathcal{J}\&\mathcal{F}_\Delta$	OC	$\mathcal{S}$
Full Model	53.9	54.5	3.78
w/o Entity Tokens	49.8 _{-4.1}	41.5	3.42
w/o $\mathcal{L}_{\text{reason}}$	49.5 _{-4.4}	40.6	2.53
w/o [OBJ] Token	52.1 _{-1.8}	53.7	3.75
w/o Reasoning-Conditioned Decoding	50.8 _{-3.1}	54.1	3.75

Table 5: **Ablation study on the main components of Re2Seg.** The subscript Δ denotes the performance drop in $\mathcal{J}\&\mathcal{F}$ compared with the full model.

audio-visual segmentation.

Results on Related Tasks. As shown in Tab. 4, Re2Seg achieves 59.1%, 41.4%, 44.6%, and 47.3% $\mathcal{J}\&\mathcal{F}$ on Ref-AVS, OmniAVS, MeViS, and ReVOS, respectively. Compared with the strongest baseline OISA-1B, our method yields consistent gains on all benchmarks, demonstrating that the proposed reasoning-aware entity modeling improves cross-dataset generalization rather than only fitting the proposed setting.

5.3 Ablation Studies

We conduct ablation studies and fine-grained evaluations using the structured annotations in ReAVS. Unless otherwise specified, we report $\mathcal{J}\&\mathcal{F}$ for each subset.

Component Ablation. Table 5 analyzes the contribution of each component. The full model achieves the best performance across both segmentation and reasoning metrics, with 53.9 $\mathcal{J}\&\mathcal{F}$, 54.5 OC, and 3.78 reasoning score. Removing entity tokens causes 4.1 drop in $\mathcal{J}\&\mathcal{F}$ and reduces OC from 54.5 to 41.5, showing the importance of entity-level evidence representations. Removing $\mathcal{L}_{\text{reason}}$ leads to the largest segmentation degradation, reducing $\mathcal{J}\&\mathcal{F}$ from 53.9 to 49.5 and the reasoning score from 3.78 to 2.53, indicating that reasoning-chain supervision is critical for object inference. Remov-

ing the [OBJ] token produces a smaller drop, while removing reasoning-conditioned decoding mainly hurts mask quality with little effect on OC and $\mathcal{S}$. These results suggest that entity-aware evidence modeling, reasoning supervision, and reasoning-conditioned decoding are complementary.

Reasoning Type. We first evaluate models under different reasoning types, including spatial, temporal, interaction, behavioral, social, causal, appearance, functional, intentional, and identity reasoning. Compared with appearance and spatial reasoning, functional, behavioral, interaction, social, and causal reasoning require models to infer object roles, actions, relations, temporal events, or causal consequences from audio-visual evidence. As shown in Tab. 3, our method obtains larger gains on these high-level semantic categories, indicating that explicit evidence modeling and object-level reasoning are particularly helpful for resolving implicit references.

Reasoning Hops. Fig. 3(c) reports performance under different reasoning-hop levels. All methods degrade as the number of hops increases, confirming that multi-hop Ref-AVS is substantially more challenging than direct object grounding. While OISA-1B already provides a strong omni-modal baseline, our model achieves consistent improvements across all hop levels, with larger gains on complex samples. In particular, it outperforms OISA-1B by 5.3 on 3-hop samples and 7.7 on >3 -hop samples. Removing $\mathcal{L}_{\text{reason}}$ notably weakens performance on high-hop cases, dropping from 49.8 to 44.2 on 3-hop and from 44.5 to 36.1 on >3 -hop. These results show that reasoning-chain supervision mainly benefits complex multi-step object inference, rather than merely improving simple object matching.

5.4 Qualitative Results

As shown in Fig. 5, the example requires identifying the object involved in the male protagonist’s emotional outburst after his girlfriend rejects the message. OmniAVS predicts the phone screen, suggesting that it is distracted by the visually salient message cue rather than the subsequent event. TGS-Agent outputs “stuffed bear” with a plausible rationale, but its prediction lacks temporal grounding, it treats the stuffed bear as the referred object throughout the whole video, although the object becomes referentially valid only after the rejection-triggered outburst. In contrast, Re2Seg decomposes the query into explicit reasoning steps: identifying the rejection event, locating the ensuing



Figure 5: Qualitative example of the Reasoning Ref-AVS task.

emotional outburst, and grounding the object violently interacted with during that event. This enables our model to segment the stuffed bear under the correct causal-temporal context, showing that reasoning-chain supervision helps capture when and why an object is referred to. More visualization in Appendix E.

6 Conclusion

We introduce **ReAVS**, a reasoning-oriented benchmark for referring audio-visual segmentation that requires models to identify the referred object from implicit and multi-hop audio-visual-language evidence before predicting its mask. ReAVS combines videos, synchronized audio, referring expressions, object masks, and reasoning-chain annotations, providing a systematic testbed for evaluating reasoning in audio-visual segmentation. We further propose **Re2Seg**, an omni-modal framework that integrates an omni-modal evidence encoder, MLLM-based multi-hop reasoning, and reasoning-conditioned mask decoding, explicitly connecting multimodal evidence with pixel-level prediction. Experiments show that Re2Seg outperforms existing baselines and produces interpretable reasoning traces, highlighting the importance of explicit object reasoning for pixel-level audio-visual segmentation.

Limitations

Although ReAVS provides a fine-grained benchmark for reasoning referring audio-visual segmen-

tation, it still has several limitations. First, the dataset is constructed from curated audio-visual videos, and its object categories and interaction patterns may not cover all real-world scenarios. Second, reasoning annotations such as reasoning types and reasoning hops inevitably involve a degree of subjectivity, although we mitigate this issue through structured annotation rules and quality control. Finally, the current framework mainly focuses on offline video segmentation, while more efficient online reasoning and segmentation remains an important direction for future work.

Acknowledgments

We would like to thank the anonymous reviewers, senior area chairs, area chairs, publication chairs for their thoughtful comments and support on this work. This work was supported by the National Natural Science Foundation of China (Grant No.62372058, U22A2026).

References

- John Atkinson and Diego Palma. 2025. An llm-based hybrid approach for enhanced automated essay scoring. *Scientific Reports*, 15(1):14551.
- Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, and 1 others. 2025. Llm instead of human judges? a large scale empirical study across 20 nlp evaluation tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 238–255.
- Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, and 1 others. 2025. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*.
- Jierun Chen, Fangyun Wei, Jinjing Zhao, Sizhe Song, Bohuai Wu, Zhuoxuan Peng, S-H Gary Chan, and Hongyang Zhang. 2025. Revisiting referring expression comprehension evaluation in the era of large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 513–524.
- Junhao Cheng, Yuying Ge, Teng Wang, Yixiao Ge, Jing Liao, and Ying Shan. 2025. Video-holmes: Can mllm think like holmes for complex video reasoning? *arXiv preprint arXiv:2505.21374*.
- Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.
- Henghui Ding, Chang Liu, Shuting He, Kaining Ying, Xudong Jiang, Chen Change Loy, and Yu-Gang Jiang. 2025a. Mevis: A multi-modal dataset for referring motion expression video segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(12):11400–11416.
- Henghui Ding, Song Tang, Shuting He, Chang Liu, Zuxuan Wu, and Yu-Gang Jiang. 2025b. Multimodal referring segmentation: A survey. *arXiv preprint arXiv:2508.00265*.
- Tianyuan Du, Haopeng Li, Zhen Fan, Jiarui Zhang, Panwang Pan, and Yang Zhang. 2026. Sam-veteran: An mllm-based human-like sam agent for reasoning segmentation. In *The Fourteenth International Conference on Learning Representations (ICLR)*.
- Shengyi Gao, Zhe Chen, Guo Chen, Wenhai Wang, and Tong Lu. 2024. Avsegformer: Audio-visual segmentation with transformer. In *Proceedings of the AAAI conference on artificial intelligence (AAAI)*, volume 38, pages 12155–12163.
- Shuting He, Guangquan Jie, Changshuo Wang, Yun Zhou, Shuming Hu, Guanbin Li, and Henghui Ding. 2025. Refersplat: Referring segmentation in 3d gaussian splatting. In *International Conference on Machine Learning (ICML)*, pages 22456–22467.
- Jiaxin Huang, Runnan Chen, Ziwen Li, Zhengqing Gao, Xiao He, Yandong Guo, Mingming Gong, and Tongliang Liu. 2026. Mllm-for3d: Adapting multimodal large language model for 3d reasoning segmentation. *Advances in Neural Information Processing Systems (NeurIPS)*, 38:54961–54984.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.
- Qing Jiang, Xingyu Chen, Zhaoyang Zeng, Junzhi Yu, and Lei Zhang. 2025. Rex-thinker: Grounded object referring via chain-of-thought reasoning. *arXiv preprint arXiv:2506.04034*.
- Anna Khoreva, Anna Rohrbach, and Bernt Schiele. 2018. Video object segmentation with language referring expressions. In *Asian conference on computer vision (ACCV)*, pages 123–141.
- Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. 2024. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9579–9589.
- Tianming Liang, Kun-Yu Lin, Chaolei Tan, Jianguo Zhang, Wei-Shi Zheng, and Jian-Fang Hu. 2025. Referdino: Referring video object segmentation with

- visual grounding foundations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20009–20019.
- Ye Liu, Zongyang Ma, Junfu Pu, Zhongang Qi, Yang Wu, Ying Shan, and Chang Chen. 2026. Unipixel: Unified object referring and segmentation for pixel-level visual reasoning. *Advances in Neural Information Processing Systems (NeurIPS)*, 38:126078–126108.
- Yuqi Liu, Bohao Peng, Zhisheng Zhong, Zihao Yue, Fanbin Lu, Bei Yu, and Jiaya Jia. 2025. Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement. *arXiv preprint arXiv:2503.06520*.
- Ye Ni, Ruiyu Liang, Xiaoshuai Hao, Jiaming Cheng, Qingyun Wang, Chengwei Huang, Cairong Zou, Wei Zhou, Weiping Ding, and Björn W Schuller. 2025. Affine modulation-based audiogram fusion network for joint noise reduction and hearing loss compensation. *Information Fusion*, page 103726.
- Ye Ni, Cong Pang, Chengwei Huang, and Cairong Zou. 2024. Msa-dpcrn: A multi-scale asymmetric dual-path convolution recurrent network with attentional feature fusion for acoustic echo cancellation. In *Proc. Interspeech 2024*, pages 162–166.
- Feiyu Pan, Hao Fang, Fangkai Li, Yanyu Xu, Yawei Li, Luca Benini, and Xiankai Lu. 2025. Semantic and sequential alignment for referring video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19067–19076.
- Abduljalil Radman and Jorma Laaksonen. 2025. Tsam: Temporal sam augmented with multimodal prompts for referring audio-visual segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23947–23956.
- Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision (ECCV)*, pages 146–162.
- Seonguk Seo, Joon-Young Lee, and Bohyung Han. 2020a. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In *European Conference on Computer Vision (ECCV)*, pages 208–223.
- Seonguk Seo, Joon-Young Lee, and Bohyung Han. 2020b. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In *European Conference on Computer Vision (ECCV)*, pages 208–223.
- Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Musique: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.
- Xiaolong Wang, Lixiang Ru, Ziyuan Huang, Kaixiang Ji, Dandan Zheng, Jingdong Chen, and Jun Zhou. 2026. Argenseg: Image segmentation with autoregressive image generation model. *Advances in Neural Information Processing Systems (NeurIPS)*, 38:28312–28336.
- Yaoting Wang, Peiwen Sun, Dongzhan Zhou, Guangyao Li, Honggang Zhang, and Di Hu. 2024. Ref-avs: Refer and segment objects in audio-visual scenes. In *European Conference on Computer Vision (ECCV)*, pages 196–213.
- Yuji Wang, Haoran Xu, Yong Liu, Jiaze Li, and Yansong Tang. 2025. Sam2-love: Segment anything model 2 in language-aided audio-visual scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28932–28941.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:24824–24837.
- Dongming Wu, Tiancai Wang, Yuang Zhang, Xiangyu Zhang, and Jianbing Shen. 2023. Onlinerefer: A simple online baseline for referring video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2761–2770.
- Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. 2022. Language as queries for referring video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4974–4984.
- Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. 2025a. Llava-cot: Let vision language models reason step-by-step. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2087–2098.
- Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfu Zhu, and 1 others. 2025b. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*.
- Zishan Xu, Yifu Guo, Yuquan Lu, Fengyu Yang, Junxin Li, and Lihua Cai. 2026. Videoseg-r1: reasoning video object segmentation via reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence (AAAI)*, volume 40, pages 11496–11504.
- Cilin Yan, Haochen Wang, Shilin Yan, Xiaolong Jiang, Yao Hu, Guoliang Kang, Weidi Xie, and Efstratios Gavves. 2024a. Visa: Reasoning video object segmentation via large language models. In *European Conference on Computer Vision (ECCV)*, pages 98–115.
- Cilin Yan, Haochen Wang, Shilin Yan, Xiaolong Jiang, Yao Hu, Guoliang Kang, Weidi Xie, and Efstratios

Gavves. 2024b. Visa: Reasoning video object segmentation via large language models. In *European Conference on Computer Vision (ECCV)*, pages 98–115.

Kaining Ying, Henghui Ding, Guangquan Jie, and Yu-Gang Jiang. 2025. Towards omnimodal expressions and reasoning in referring audio-visual segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22575–22585.

Henghao Zhao, Ge-Peng Ji, Rui Yan, Huan Xiong, and Zechao Li. 2026a. Videoexpert: Augmented llm for temporal-sensitive video understanding. *IEEE Transactions on Circuits and Systems for Video Technology*, 36(5):5893–5908.

Yuan Zhao, Zhenqi Jia, and Yongqiang Zhang. 2026b. Mar3: Multi-agent recognition, reasoning, and reflection for reference audio-visual segmentation. *arXiv preprint arXiv:2603.27706*.

Rongkun Zheng, Lu Qi, Xi Chen, Yi Wang, Kun Wang, and Hengshuang Zhao. 2025. Villa: Video reasoning segmentation with large language model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 23667–23677.

Jinxing Zhou, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, and Yiran Zhong. 2022. Audio-visual segmentation. In *European Conference on Computer Vision (ECCV)*, pages 386–403.

Jinxing Zhou, Yanghao Zhou, Mingfei Han, Tong Wang, Xiaojun Chang, Hisham Cholakkal, and Rao Muhammad Anwer. 2026. Think before you segment: An object-aware reasoning agent for referring audio-visual segmentation. In *Proceedings of the AAAI conference on artificial intelligence (AAAI)*, volume 40, pages 13665–13673.

A Related Work

A.1 Referring Audio-Visual Segmentation

Referring audio-visual segmentation (Ref-AVS) (Ding et al., 2025b; Radman and Laaksonen, 2025) belongs to the broader family of multimodal referring segmentation, which aims to segment objects in visual scenes according to referring expressions from different modalities, such as text, speech, sound, or their combinations. Beyond visual understanding, prior work (Ni et al., 2025, 2024) has extensively studied acoustic modeling and audio enhancement for improving the quality and robustness of audio representations. Compared with conventional audio-visual segmentation that mainly localizes sounding objects (Zhou et al., 2022; Gao et al., 2024), Ref-AVS (Wang et al., 2024) further requires models to identify the object

specified by a referring expression and generate its pixel-level masks in audible videos. Existing methods (Radman and Laaksonen, 2025; Wang et al., 2025) mainly follow two paradigms. The first line learns unified audio-video-language representations and uses them to prompt segmentation decoders, such as SAM or SAM2, for mask prediction. The second line introduces MLLMs or agent-based reasoning to infer the referred object before grounding and segmentation, improving semantic understanding for complex multimodal expressions (Ying et al., 2025; Zhou et al., 2026; Zhao et al., 2026b). Recent benchmarks (Ying et al., 2025; Zhou et al., 2026) further extend Ref-AVS toward omnimodal expressions and reasoning-intensive references, indicating that audio-visual segmentation increasingly requires high-level multimodal understanding beyond sound-source localization. However, most existing methods still focus on explicit references or relatively shallow object reasoning, where the referred object can often be determined through direct audio-visual-language matching. In contrast, our Reasoning Ref-AVS task emphasizes implicit references that require multi-hop inference over audio, visual, and linguistic evidence before mask prediction.

A.2 Referring Video Object Segmentation

Referring video object segmentation (Ref-VOS) (Seo et al., 2020b; Wu et al., 2023; Pan et al., 2025) aims to segment the object specified by a natural-language expression across video frames. Compared with image-level referring segmentation (Chen et al., 2025; Wang et al., 2026), Ref-VOS requires models to maintain temporal consistency while resolving object attributes, actions, and inter-object relations. Early methods (Wu et al., 2022) typically learn cross-modal fusion between language and video features, using attention or memory mechanisms to associate referring expressions with object regions over time. Recent works (Lai et al., 2024; Yan et al., 2024b; Carion et al., 2025; Yan et al., 2024a) further leverage powerful visual foundation models and MLLMs to improve open-vocabulary object understanding, temporal grounding, and mask decoding. However, existing Ref-VOS methods mainly rely on visual-language cues and do not explicitly model audio evidence, making them less suitable for audio-visual scenarios where the referred object is determined by sound events or

cross-modal interactions. In contrast, Reasoning Ref-AVS requires models to infer the referred object from audio, visual, and linguistic evidence before generating temporally consistent masks.

A.3 Reasoning Segmentation with MLLMs

Recent advances in multimodal large language models (MLLMs) (Huang et al., 2026; Du et al., 2026) have promoted reasoning-oriented segmentation, where models are expected to understand complex instructions and produce grounded masks. Representative works (Lai et al., 2024; Yan et al., 2024b) such as LISA and VISA introduce instruction-following segmentation by connecting MLLMs with mask decoders, while subsequent studies further incorporate chain-of-thought reasoning or reinforcement learning to improve interpretability and grounding reliability (Liu et al., 2025; Jiang et al., 2025). In video scenarios, reasoning-based methods extend this paradigm to temporally grounded object segmentation, enabling models to handle complex visual-language references across frames (Yan et al., 2024a; Zhao et al., 2026a). Despite their progress, these methods mainly rely on visual-language inputs and do not explicitly reason over audio-visual evidence. Our RAVS-Reasoner differs by performing object-centric multi-hop reasoning over audio, visual, and linguistic cues before mask decoding, which is essential for resolving implicit references in Reasoning Ref-AVS.

B ReAVS Benchmark Details

B.1 Candidate-Level Annotation Audit

The main text describes the annotation protocol and quality-control criteria. Here, we provide a candidate-level audit of the annotation outcomes to quantify how these procedures affect benchmark construction. **(1) Candidate disposition.** Table 7 summarizes the disposition of all 94,806 initial candidate expressions. Among them, 49,762 candidates are retained, corresponding to an overall retention rate of 52.49%. Only 26,450 candidates, or 27.90% of the initial pool, satisfy all criteria without modification. The remaining 23,312 retained candidates require manual revision. Consequently, 46.85% of the final benchmark instances are manually corrected before inclusion. Considering both revised and rejected candidates, 72.10% of the initial proposals require substantive human intervention rather than direct acceptance. **(2) Rejection**

profile. Of the 45,044 rejected candidates, insufficient multimodal evidence constitutes the largest category, accounting for 28.54% of all rejections. Directly solvable expressions account for a further 25.36%. Together, these two categories comprise 53.89% of the rejected pool, indicating that the primary filtering pressure concerns whether an expression is both evidence-grounded and genuinely reasoning-intensive. Ambiguous referents account for 19.20% of rejections, while inadequate object visibility and annotation inconsistency account for 14.28% and 12.63%, respectively. These statistics show that candidate removal is driven primarily by task validity, answerability, and groundability rather than surface-level linguistic quality.

B.2 Reasoning Types

Tab. 6 presents the reasoning-type taxonomy used in ReAVS. The taxonomy is designed to characterize the main inference pattern required to identify the referred object from audio-visual-language evidence. It covers low-level perceptual cues, such as appearance and spatial relations, as well as higher-level event and semantic reasoning, including temporal, causal, social, functional, behavioral, intentional, and identity reasoning. Each sample is assigned one or more reasoning types according to the evidence required by its referring expression. This categorization provides a diagnostic reference for analyzing whether a model fails due to weak visual grounding, insufficient acoustic understanding, missing temporal context, or inadequate high-level object reasoning.

B.3 Prompt for Expression Generation

Fig. 7 shows the structured prompt used for automated generation of reasoning-intensive referring expressions. By providing the original multiple-choice video QA sample, including the question, options, answer, answer text, and explanation, the prompt converts each sample into an object-centric grounding instance with a unique segmentable object. It further defines dynamic annotation fields, including object category, reasoning type, reasoning hops, temporal dependency, multi-frame dependency, and review flags. This systematic design ensures that the generated expressions are aligned with the ReAVS reasoning taxonomy while preserving the evidence implied by the original explanation and maintaining logical consistency across implicit, temporal, causal, and multi-hop referring scenarios.

Type	Definition	Example
Appearance	Identifying the referred object based on observable visual properties, such as color, shape, size, texture, clothing, facial attributes, or other appearance cues.	“Which person visually matches the man standing with the gun based on appearance?”
Spatial	Resolving the referred object according to its position, layout, orientation, or spatial relation with other objects, persons, or scene regions.	“Which person is standing closest to the man in the plaid shirt?”
Temporal	Inferring the referred object from event order, temporal change, duration, or the time period during which the object becomes relevant.	“Which object is heard ringing at the beginning and then changes its state over time?”
Interaction	Identifying the referred object through direct physical or communicative interaction with another object, person, animal, or scene element.	“Which person is interacting with the woman with a ponytail during their first encounter?”
Behavioral	Recognizing the referred object from actions, expressions, gestures, motion patterns, or other observable behaviors.	“Which person exhibits frustration by scratching the iron fence with a branch at 00:32?”
Social	Inferring the referred object from interpersonal relations, social roles, group dynamics, or changes in social relationships.	“Which persons exemplify the transition from rivalry to sisterhood due to interpersonal changes?”
Causal	Resolving the referred object by linking an event, action, or state to its cause, consequence, or triggered response.	“Which entity is responsible for extending the green light time to assist the old man?”
Functional	Identifying the referred object according to its function, affordance, intended use, or role in accomplishing an activity.	“Which object is used to send the message that is later rejected?”
Intentional	Inferring the referred object from an agent’s goal, intention, plan, desire, or purpose implied by audio-visual evidence.	“Which person is inferred to be expressing love through their actions at timestamp 0:58?”
Identity	Resolving the referred object through identity-level cues, including personal identity, narrative role, category identity, or cross-frame identity association.	“Who is the person inferred to be both the investigator and the thief connected to the victim and the stolen ring?”

Table 6: Definition and examples of the ten reasoning types in ReAVS.

Quality-control outcome	Count	Percentage
Initial candidate expressions	94,806	100.00%
Accepted without revision	26,450	27.90%
Accepted after manual revision	23,312	24.59%
Final retained samples	49,762	52.49%
Rejected: directly solvable expression	11,421	12.05%
Rejected: insufficient multimodal evidence	12,854	13.56%
Rejected: ambiguous referent	8,648	9.12%
Rejected: absent/small/occluded object	6,431	6.78%
Rejected: cross-field or mask inconsistency	5,690	6.00%
Total rejected samples	45,044	47.51%

Table 7: Quality-control statistics for the annotation pipeline.

B.4 Prompt for Reasoning Chain

We use GPT-4o to generate structured reasoning chains for each reasoning-intensive referring expression in **ReAVS**. The prompt asks the model to decompose object inference into a small number of evidence-grounded steps and to end with an explicit object-grounding step linked to the annotated object ID. To ensure consistency, we define reasoning hops as evidence-dependent transitions before final mask grounding: 1-hop samples require a single inference step before object grounding, 2-hop

samples require two dependent inference steps, and 3-hop samples require three dependent inference steps. The complete prompt is shown below, as shown in Fig. 8.

C Additional Method Details

C.1 Preliminaries

Reasoning Ref-AVS. Reasoning Ref-AVS aims to segment the object specified by a reasoning-intensive audio-visual-language expression. Given a video $\mathcal{V}$, synchronized audio $\mathcal{A}$, and referring expression $\mathcal{Q}$, the model predicts a mask sequence $\hat{\mathcal{M}}$ for the referred object. Different from conventional Ref-AVS, the referred object may not be directly determined by category, appearance, action, location, or sound-source matching. Instead, the model needs to infer the object through intermediate audio-visual-language evidence before mask prediction.

Training Instance. Each training sample contains the video, audio, expression, ground-truth mask sequence, and reasoning-chain annotation:

$$x = (\mathcal{V}, \mathcal{A}, \mathcal{Q}, \mathcal{M}, \mathcal{R}), \quad (7)$$

where $\mathcal{M}$ denotes the ground-truth mask sequence and $\mathcal{R}$ denotes the annotated reasoning chain. The reasoning chain is used only for training supervision. During inference, the model receives only $\mathcal{V}$, $\mathcal{A}$, and $\mathcal{Q}$.

Reasoning Chain. A reasoning chain is represented as an ordered sequence of intermediate inference steps:

$$\mathcal{R} = \{r_k\}_{k=1}^K. \quad (8)$$

Each step describes a necessary clue for resolving the referred object, such as visual evidence, acoustic evidence, temporal evidence, object interaction, or cross-modal verification. The final step is always object grounding, which determines the object representation used for mask prediction.

C.2 Reasoning Hop Definition

We define reasoning hops according to the number of non-trivial inference transitions required before object grounding. This definition focuses on reasoning dependency rather than sentence length.

A 1-hop expression can be resolved by one direct clue:

$$\mathcal{Q} \rightarrow o, \quad (9)$$

where o denotes the referred object. Typical cases include explicit category, appearance, action, position, or direct sound-source references.

A 2-hop expression requires one intermediate clue:

$$\mathcal{Q} \rightarrow c_1 \rightarrow o, \quad (10)$$

where c_1 may denote an action, event, relation, speaker, sound cue, or contextual clue. For example, the model may first identify the person described by an action and then ground the object manipulated by that person.

A 3-hop expression requires two intermediate clues:

$$\mathcal{Q} \rightarrow c_1 \rightarrow c_2 \rightarrow o. \quad (11)$$

For example, the model may first infer an event, then identify the event participant, and finally ground the object associated with that participant. Expressions requiring more intermediate clues are labeled as > 3 -hop, where only decisive reasoning steps are retained to avoid redundant annotations.

C.3 Reasoning Step Representation

To make reasoning-chain supervision structured and consistent, each reasoning step is represented

as:

$$r_k = (s_k, \tau_k, c_k), \quad (12)$$

where s_k is the step index, τ_k is the reasoning type, and c_k is the textual content of the reasoning step. The last step must be `target_grounding`. Its object phrase and object id are required to match the final referred object annotation.

C.4 Token-level Reasoning Supervision

The main paper defines reasoning supervision over the generated reasoning chain and the special object summary token [OBJ]. We provide the token-level form. Let the generated supervision sequence be:

$$\mathcal{V} = [\mathcal{R}; [\text{OBJ}]] = \{y_\ell\}_{\ell=1}^L. \quad (13)$$

This formulation provides supervision not only for the final object summary token, but also for the intermediate reasoning trajectory.

C.5 Object Summary Token

The special token [OBJ] bridges language-based reasoning and mask decoding. After the MLLM generates the reasoning chain, the hidden state of [OBJ] summarizes the inferred referred object. Instead of using the full reasoning text for segmentation, the model uses this compact object-level representation to retrieve the corresponding entity evidence and condition the mask decoder.

Let $\mathbf{h}^{obj}$ denote the hidden state of [OBJ] and $\mathbf{Z}^e = \{\mathbf{z}_i^e\}_{i=1}^N$ denote the entity-evidence representations. The object-aware attention weight over the i -th entity token is:

$$\alpha_i = \frac{\exp(\text{sim}(\mathbf{h}^{obj}, \mathbf{z}_i^e))}{\sum_{j=1}^N \exp(\text{sim}(\mathbf{h}^{obj}, \mathbf{z}_j^e))}, \quad (14)$$

where $\text{sim}(\cdot, \cdot)$ denotes a learnable similarity function. The reasoning-aware object representation is computed as:

$$\mathbf{z}^{obj} = \sum_{i=1}^N \alpha_i \mathbf{z}_i^e. \quad (15)$$

This representation is used as the object query for mask decoding.

C.6 Mask Supervision

Following the main paper, mask prediction is supervised by binary cross-entropy loss and Dice loss.

Method	Speech	Speech&Image	Speech&Sound	Speech&Sound&Image	Text	Text&Image	Text&Sound	Text&Sound&Image
AVSBench	27.8	25.4	20.2	18.4	38.2	28.2	22.8	20.6
AVSegFormer	30.1	27.8	23.5	24.2	39.6	33.2	30.2	28.4
LISA-7B	34.1	31.5	30.6	28.2	43.5	35.8	33.2	33.5
VISA-7B	39.4	36.6	32.8	29.4	48.7	45.1	40.4	38.2
OISA-1B	51.1	50.2	44.5	42.2	55.6	54.6	53.8	51.4
Ours w/o $\mathcal{L}_{\text{reason}}$	52.3	48.1	43.4	39.1	54.8	54.5	52.9	50.6
Ours	54.4	51.8	48.2	43.6	60.1	58.8	57.6	56.8

Table 8: Performance comparison under different expression modality groups. We report $\mathcal{J}\&\mathcal{F}$ (%) for each group, where bold text indicates the best performance.

Method	Params(B) ↓	Latency (s/video) ↓	Peak Memory (GB) ↓	$\mathcal{J}\&\mathcal{F}$ ↑
OISA-1B	1.61	6.58	16.2	50.4
Re2Seg	1.63	7.31	17.3	53.9
w/o Reasoning-Conditioned Decoding	1.63	6.95	16.8	50.8

Table 9: Efficiency comparison of different methods.

Method	$\mathcal{J}\&\mathcal{F}$	OC	$\mathcal{S}$
OISA-1B	50.4	40.2	2.45
OISA-1B + $\mathcal{L}_{\text{reason}}$	51.1	41.8	3.32
Re2Seg w/o $\mathcal{L}_{\text{reason}}$	49.5	40.6	2.53
Re2Seg	53.9	54.5	3.78

Table 10: Analysis of architecture and reasoning-chain supervision.

The binary cross-entropy term provides pixel-wise foreground-background supervision:

$$\mathcal{L}_{\text{bce}} = -\frac{1}{T|\Omega|} \sum_{t=1}^T \sum_{p \in \Omega} \left[M_t(p) \log \hat{M}_t(p) + (1 - M_t(p)) \log(1 - \hat{M}_t(p)) \right],$$

where Ω denotes the pixel set. The Dice loss is:

$$\mathcal{L}_{\text{dice}} = 1 - \frac{2\Sigma_{TM} + \epsilon}{\Sigma_{\hat{M}} + \Sigma_M + \epsilon}, \quad (16)$$

where $\Sigma_{TM} = \sum_{t=1}^T \sum_{p \in \Omega} \hat{M}_t(p) M_t(p)$, $\Sigma_{\hat{M}} = \sum_{t=1}^T \sum_{p \in \Omega} \hat{M}_t(p)$, $\Sigma_M = \sum_{t=1}^T \sum_{p \in \Omega} M_t(p)$, ϵ is a smoothing constant. The BCE term stabilizes pixel-level classification, while the Dice term alleviates foreground-background imbalance and directly optimizes region overlap.

C.7 Evaluation Metrics

We report region similarity $\mathcal{J}$ and contour accuracy $\mathcal{F}$. For a predicted mask $\hat{M}$ and ground-truth mask M , the region similarity is:

$$\mathcal{J} = \frac{|\hat{M} \cap M|}{|\hat{M} \cup M|}. \quad (17)$$

The contour score $\mathcal{F}$ is computed from boundary precision and recall:

$$\mathcal{F} = \frac{2P_{\text{bd}}R_{\text{bd}}}{P_{\text{bd}} + R_{\text{bd}}}, \quad (18)$$

where P_{bd} and R_{bd} denote boundary precision and boundary recall. The overall score is:

$$\mathcal{J}\&\mathcal{F} = \frac{\mathcal{J} + \mathcal{F}}{2}. \quad (19)$$

D Ablation Studies

D.1 Expression Modality

Table 8 reports performance across different expression modality groups. Compared with text-only expressions, multimodal expressions involving speech, sound, and image evidence are substantially more challenging, since the referred object cannot be resolved by direct linguistic matching alone. The performance gap becomes more evident for speech&sound&image and text&sound&image groups, where models must align semantic descriptions, acoustic cues, and visual entities jointly. RAVS-Reasoner consistently improves over the baselines across these groups, indicating that explicit evidence-aware reasoning helps integrate heterogeneous cues for robust object inference.

D.2 Efficiency Analysis

To evaluate the efficiency overhead introduced by Re2Seg, we compare its parameter count, inference latency, and peak memory usage against the baseline OISA-1B, as shown in Table 9, Re2Seg increases the parameter count by only 1.2% and peak memory by 6.8%, while improving $\mathcal{J}\&\mathcal{F}$ by



Figure 6: Visualization of Re2Seg.

3.5. The end-to-end latency increases from 6.58 to 7.31 s/video (+11.1%), indicating a reasonable accuracy–efficiency trade-off. Moreover, compared with the variant without reasoning-conditioned decoding, the full model adds only 0.36 s/video and 0.5 GB memory, while improving $\mathcal{J}\&\mathcal{F}$ by 3.1.

D.3 Disentangling architecture and reasoning supervision

Methods using different inputs or supervision are included as contextual references. To isolate the source of improvement, we compare Re2Seg with OISA-1B, from which it is initialized. As shown in Table 10, adding $\mathcal{L}_{\text{reason}}$ to OISA-1B improves $\mathcal{J}\&\mathcal{F}$, OC, and $\mathcal{S}$ by only 0.7, 1.6, and 0.87, respectively. Under the same supervision, Re2Seg further improves them by 2.8, 12.7, and 0.46. The difference-in-differences gains are 3.7, 12.3, and 0.38, indicating that the improvement mainly arises from the interaction between reasoning supervision and the proposed architecture.

E Visualization

Figure 6 presents a representative case requiring implicit object inference. Given the expression “The object like  whose failure to operate suggests that escape is more difficult than imagined,” the referred object cannot be identified by direct category matching. Instead, the model needs to associate the semantic clue of “failure to operate” with the functional role of an escape-related mechanism. With reasoning-chain supervision, RAVS-Reasoner correctly infers the referred object as the

door valve, since a malfunctioning valve prevents the door from being opened and therefore makes escape difficult. In contrast, removing $\mathcal{L}_{\text{reason}}$ leads the model to predict the more visually salient but semantically coarse object, i.e., the door. This comparison shows that $\mathcal{L}_{\text{reason}}$ encourages the model to resolve the functional evidence behind the expression, rather than relying on shallow visual or lexical cues. As a result, the full model produces a more precise object-level grounding under implicit reasoning scenarios.

Prompt for Expression Generation

System Prompt

You are an expert annotation assistant for constructing ReAVS, a benchmark for Reasoning Ref-AVS. Your task is to convert a multiple-choice video question-answering sample into an object-centric reasoning-grounding sample.

Given the provided {question}, {options}, {answer}, {answer_text}, and {explanation}, generate a reasoning-intensive referring expression whose answer is exactly one visually segmentable object in the video. The expression should require audio-visual-language reasoning before mask prediction, rather than direct matching of explicit textual, visual, or acoustic cues.

You must not output, modify, or infer any source identity fields, such as video ID, question ID, dataset name, or original split. The caller will preserve these fields separately. You should only generate the fields listed in the output schema.

Reasoning Type

- appearance: reasoning based on visual attributes such as color, shape, clothing, facial attributes, or other observable appearance cues.
- spatial: reasoning based on position, layout, orientation, or spatial relations with other objects or scene regions.
- temporal: reasoning based on event order, duration, temporal change, or the moment when the object becomes relevant.
- interaction: reasoning based on direct physical or communicative interaction with another object, person, animal, or scene element.
- behavioral: reasoning based on actions, gestures, expressions, motion patterns, or observable behaviors.
- social: reasoning based on interpersonal relations, social roles, group dynamics, or changes in social relationships.
- causal: reasoning based on cause, consequence, trigger, or response.
- functional: reasoning based on object function, affordance, use, or role in accomplishing an activity.
- intentional: reasoning based on an agent's intention, goal, plan, desire, or purpose.
- identity: reasoning based on personal identity, narrative role, category identity, or cross-frame identity association.

Reasoning Hop

A reasoning hop is a necessary evidence-dependent transition for identifying the referred object. Each hop should introduce a new evidence unit or inference relation, such as linking an event to its consequence, an action to an intention, a sound to an object, or a person to a social role. Pure mask localization after object inference must not be counted as an additional hop.

Detailed Analysis Guidelines

1. Identify the correct answer from {answer} and {answer_text}.
2. Determine whether the answer corresponds to a unique visible and segmentable object in the video.
3. Use {explanation} to infer the evidence needed to identify the object.
4. Rewrite the original QA sample into a concise referring expression expr.
5. Ensure that the answer to expr is exactly the referred object.
6. Preserve the reasoning implied by {answer_text} and {explanation}.
7. Make reasoning necessary. The expression should not be solvable only through object category, appearance, position, simple visible action, or direct sound-source matching.
8. Explicitly encode temporal, causal, behavioral, social, functional, intentional, or identity reasoning when supported by the explanation.
9. Assign the most appropriate object_category, reasoning_type, and reasoning_hops.
10. Determine whether the object requires temporal context or multi-frame reasoning.
11. If the sample is ambiguous, weakly grounded, not visually segmentable, or not reasoning-dependent, set needs_review=true.

Expression Writing Guidelines

Prefer natural question forms such as:

- Which person is inferred to ...?
- Which object is implied to ...?
- Which entity is responsible for ...?
- Which object becomes relevant after ...?
- Which person shows signs of ... because ...?

Avoid trivial rewrites such as:

- the man near the fence
- the person cleaning the windshield
- the dog on the left
- the object making sound

The expression must be concise, specific, and end with a question mark.

Task-Specific Constraints

- There must be exactly one ground-truth referred object.
- The referred object must be visually segmentable in the video.
- The referred object cannot be an abstract concept, event, emotion, relation, or narrative conclusion.
- The expression must preserve the reasoning evidence from the original QA sample.
- The expression must not directly reveal the object in a way that removes the need for reasoning.
- Set is_direct_referrable=true only if the object can be identified by direct category, appearance, position, simple action, or sound-source matching alone.
- Set requires_temporal_context=true if event order, duration, state change, or a specific temporal window is needed.
- Set requires_multi_frame_reasoning=true if resolving the object requires evidence from multiple frames.

Figure 7: Prompt for GPT-4o to generate reasoning referring expressions for **ReAVS**. Terms such as {question}, {answer_text}, and {explanation} denote input variables from the original video QA sample. Terms such as [reasoning_type], and [reasoning_hops] denote constrained output slots.

Prompt for Reasoning Chain Generation

System Prompt

You are an expert dataset annotation assistant for Reasoning Ref-AVS. Your task is to generate a minimal but structured reasoning_chain field for each sample.

The reasoning chain should explain how to infer the referred object from the given expression and explanation before mask prediction. Each step must be concise, factual, and grounded in the provided input. The chain should not introduce unsupported entities, events, names, sounds, actions, or attributes.

Input Fields

- expr: the reasoning-intensive referring expression.
- explanation: the human explanation or rationale.
- object_text: the final referred object phrase.
- object_category: the category of the final referred object.
- obj_id: the ground-truth object ID or IDs.
- reasoning_type: coarse reasoning type labels.
- reasoning_hops: the annotated number of reasoning hops.

Output Field

Generate only the following field:

- reasoning_chain

Do not modify or regenerate expr, explanation, object_text, object_category, obj_id, reasoning_type, or reasoning_hops.

Hop Definitions

A reasoning hop is a necessary evidence-dependent transition used to identify the referred object. Each hop should introduce either a new evidence unit or a new inference relation. Pure mask localization after object inference is not counted as a reasoning hop.

- 1-hop reasoning: one evidence-dependent transition before final grounding.
- 2-hop reasoning: two dependent transitions before final grounding; the second step must rely on the first.
- 3-hop reasoning: three dependent transitions before final grounding, such as temporal ordering, causal consequence, and object role inference.
- >3-hop reasoning: generate three or four compact reasoning steps covering the most essential dependencies.

The number of reasoning steps before target_grounding should match reasoning_hops when possible. If exact matching would make the chain redundant or unsupported, use the smallest faithful chain and set validity.ambiguous_without_context=true when appropriate.

Detailed Analysis Guidelines

1. Read expr and identify what object must be inferred.
2. Use explanation as the only source of supporting reasoning evidence.
3. Determine the minimal sequence of evidence-dependent transitions needed to infer object_text.
4. Align the number of reasoning steps before target_grounding with reasoning_hops when possible.
5. Each intermediate step should contain operation, description, intermediate_answer, and reasoning_type.
6. The final step must use operation="target_grounding".
7. final_target in the final step must exactly equal object_text.
8. obj_id in the final step must exactly equal the input obj_id.
9. Do not include mask prediction, segmentation quality, or pixel-level localization as reasoning steps.
10. Do not introduce unsupported entities, events, attributes, sounds, or causal relations.

Task-Specific Constraints

- Generate only the reasoning_chain field.
- The reasoning chain should usually contain 2 to 5 total steps, including target_grounding.
- The final step must always be target_grounding.
- The chain must be faithful to expr and explanation.
- The chain must be minimal and should not split one simple inference into redundant steps.
- If the sample is ambiguous but still has a plausible object, generate the best chain and set validity.ambiguous_without_context=true.
- If the object is not concrete or segmentable, set validity.has_groundable_target=false.
- Set validity.is_multi_hop=true when reasoning_hops is greater than 1.

Output Format

```
{
  "reasoning_chain": {
    "steps": [
      {
        "step_id": 1,
        "operation": "...",
        "description": "...",
        "intermediate_answer": "...",
        "reasoning_type": ["..."]
      },
      {
        "step_id": 2,
        "operation": "target_grounding",
        "description": "...",
        "final_target": "...",
        "obj_id": []
      }
    ],
    "validity": {
      "is_multi_hop": true,
      "has_groundable_target": true,
      "ambiguous_without_context": false
    }
  }
}
```

Figure 8: Prompt for GPT-4o to generate structured reasoning chains for **ReAVS**. Terms such as {expr}, {explanation}, {object_text}, and {obj_id} denote input variables from each annotated sample. Terms such as [operation], [reasoning_type], and [target_grounding] denote constrained output slots. The prompt defines 1-hop, 2-hop, 3-hop, and >3-hop reasoning according to the number of evidence-dependent transitions before final object grounding.