

Think Before You Segment: An Object-aware Reasoning Agent for Referring Audio-Visual Segmentation

Jinxing Zhou¹, Yanghao Zhou^{2*}, Mingfei Han¹, Tong Wang¹, Xiaojun Chang^{1,3},
Hisham Cholakkal^{1*}, Rao Muhammad Anwer^{1*}

¹Mohamed Bin Zayed University of Artificial Intelligence

²National University of Singapore

³University of Science and Technology of China

Abstract

Referring Audio-Visual Segmentation (Ref-AVS) aims to segment target objects in audible videos based on given reference expressions. Prior works typically rely on learning latent embeddings via multimodal fusion to prompt a tunable SAM/SAM2 decoder for segmentation, which requires strong pixel-level supervision and lacks interpretability. From a novel perspective of explicit reference understanding, we propose TGS-Agent, which decomposes the task into a Think-Ground-Segment process, mimicking the human reasoning procedure by first identifying the referred object through multimodal analysis, followed by coarse-grained grounding and precise segmentation. To this end, we first propose Ref-Thinker, a multimodal language model capable of reasoning over textual, visual, and auditory cues. We construct an instruction-tuning dataset with explicit object-aware think-answer chains for Ref-Thinker fine-tuning. The object description inferred by Ref-Thinker is used as an explicit prompt for Grounding-DINO and SAM2, which perform grounding and segmentation without relying on pixel-level supervision. Additionally, we introduce R²-AVSBench, a new benchmark with linguistically diverse and reasoning-intensive references for better evaluating model generalization. Our approach achieves state-of-the-art results on both standard Ref-AVSBench and proposed R²-AVSBench.

1 Introduction

In recent years, the trend in artificial intelligence research has been shifting toward omni-modal understanding, where models are expected to jointly perceive and reason across multiple modalities. Among various topics, the Referring Audio-Visual Segmentation (Ref-AVS) task aims to segment target objects in audible video scenes based on given natural language expressions (a.k.a *reference*). As shown in Fig. 1(a), the reference may contain either single-modal or multi-modal cues, requiring the model to automatically analyze and integrate relevant cues from audio, visual, and textual modalities. For instance, the reference “*The object making a sound on the left of the guitar*” demands integrating textual semantics, spatial visual cues, and temporal audio patterns to segment the correct object ‘piano’ across video frames.

Research on the Ref-AVS task remains an active and evolving area. The pioneering work, EEMC (Wang et al. 2024b),

*Corresponding author

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

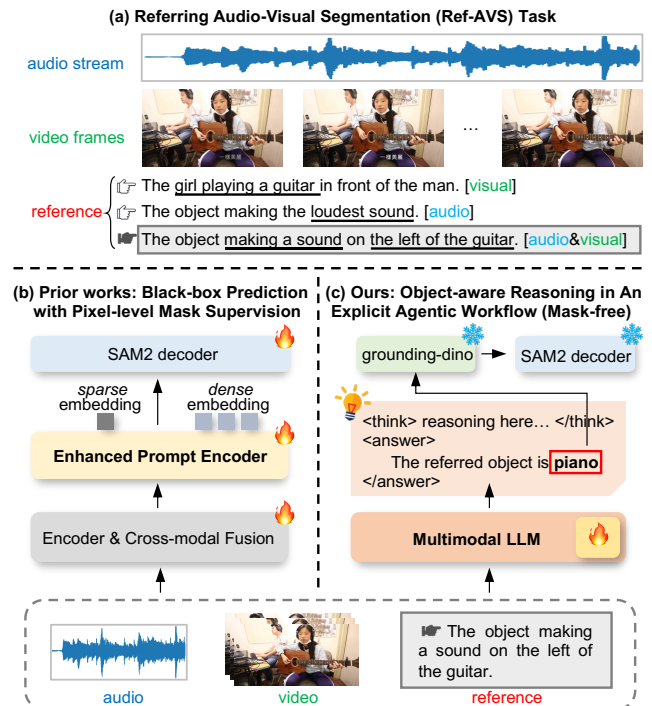


Figure 1: (a) Illustration of Ref-AVS task. (b) Prior works focus on enhancing sparse and dense prompt embeddings via implicit transformer fusion, requiring strong pixel-level supervision for model training. (c) Our method first performs explicit reasoning over the reference using a MLLM to identify the referred object, and then generates its bounding box and segmentation mask within an agentic workflow.

employs multiple transformer blocks to model interactions among audio, visual, and textual modalities. The resulting integrated multimodal features are then used as cues to prompt a segmentation decoder, Mask2Former (Cheng et al. 2021), to generate binary masks of the referred object. Recent approaches leverage more powerful segmentation models such as SAM (Kirillov et al. 2023) and SAM2 (Ravi et al. 2024), extending them to handle multimodal signals. As shown in Fig. 1(b), a common focus of these methods is to enhance the prompt encoder of SAM or SAM2 (*i.e.*, the sparse and dense

prompt embeddings) by modeling more informative multimodal cues. For instance, TSAM (Radman and Laaksonen 2025) utilizes cross-attention mechanism to fuse audio-visual-text modalities. Another recent method, SAM2-LOVE (Wang et al. 2025), proposes to summarize multimodal features into a learnable `[seg]` token, which serves as an improved sparse prompt for SAM2. The mask decoder of SAM/SAM2 is also fine-tuned to better adapt to the Ref-AVS task. Although these methods achieve strong performance, they rely on pixel-level ground truth masks for supervision, and their segmentation processes function as black boxes, lacking interpretability. In contrast, our work explores a ground truth *mask-free* and *more explainable* approach.

Let’s reconsider how humans naturally approach the Ref-AVS task. Given a reference, e.g., “*The object making a sound on the left of the guitar*” shown in Fig. 1(a), we first analyze this reference text, observe video frames to locate the *guitar*, shift our attention to its *left region*, and listen to the audio to identify the object which *is making a sound*. Following this chain of reasoning, we can clearly determine the referred object, i.e., the *piano*. Once the *piano* is identified, we can ground its position and precisely segment the corresponding pixels across video frames. We refer to this procedure as a ‘**Think-Ground-Segment (TGS)**’ decision process. This critical step is often overlooked by prior methods: before performing fine-grained segmentation, humans can first explicitly identify the target object by thinking about multimodal (audio, visual, textual) inputs.

Motivated by this observation, we propose **TGS-Agent**, an agentic framework with an explicit, explainable, object-aware reasoning chain for the Ref-AVS task (illustrated in Fig. 1(c)). Thanks to the advancements in foundation models for object detection and segmentation, the *Ground* and *Segment* steps can be effectively handled. In our framework, we adopt Grounding-DINO (Liu et al. 2024b) and SAM2 (Ravi et al. 2024) as the tools for these two stages, respectively. Given an appropriate textual prompt (related to target objects in Ref-AVS setting), Grounding-DINO generates its corresponding bounding box (object text \rightarrow bbox). Subsequently, SAM2 takes the bbox as prompt and accurately produces the corresponding pixel mask (bbox \rightarrow mask). Therefore, a critical challenge is how to address the *Think* step: identify the referred object according to multimodal signals (reference \rightarrow object text). Since the output of the *Think* step is an open-ended textual description related to the referred object, we introduce a multimodal large language model (MLLM), referred to as the **Ref-Thinker**, to infer the target object from the multimodal context (i.e., the reference alongside its audio and visual counterparts). To enhance the reasoning and instruction-following capabilities of Ref-Thinker, we utilize Gemini-1.5-Pro with carefully designed prompts to construct an instruction-tuning set, containing explicit think-answer reasoning chains. More details will be introduced in Sec. 3.2. As a result, the fine-tuned MLLM is able to reason over the audio, visual, and reference inputs and generate reliable descriptions of the referred object. We explore the impact of using simplified (category only) and fine-grained (category with more details like its attribute or spatial location) description texts (See more discussion in Sec. 5.4).

We evaluate TGS-Agent on the existing Ref-AVSBench dataset (Wang et al. 2024b). In addition, we propose **R²-AVSBench**, a new evaluation set designed to be more reasoning-intensive. Specifically, we observe that references in a portion of the Ref-AVSBench test set tend to be straightforward. For example, the target object name directly appears in the reference (e.g., ‘*The clarinet being played by a man*’). Moreover, Ref-AVSBench references are constructed using fixed templates, limiting their linguistic diversity. To improve this, we propose R²-AVSBench, which features references with greater lexical and structural diversity, and which require deeper reasoning to interpret. We prompt Gemini-1.5-Pro to achieve the reference transformation. Further details will be provided in Sec. 4. Our R²-AVSBench can serve as a more effective benchmark for evaluating a model’s generalizability across diverse reference types.

In summary, our main contributions are: (1) We propose TGS-Agent, decoupling the Ref-AVS task into a ‘Think-Ground-Segment’ agentic workflow. TGS-Agent provides a new paradigm that is mask supervision-free and also more explainable. (2) We propose Ref-Thinker, a MLLM with enhanced object-aware reasoning ability, which generates explicit descriptions of the referred object by analyzing multimodal signals. An instruction tuning set is constructed to support Ref-Thinker training. (3) We propose R²-AVSBench, a new evaluation benchmark featuring more challenging and reasoning-intensive references. It serves as an additional testbed to evaluate model performance in cross-reference scenarios. (4) Our method achieves state-of-the-art results on both Ref-AVSBench and R²-AVSBench, significantly outperforming existing methods.

2 Related Work

Audio-Visual Scene Understanding aims to analyze the rich and dynamic audio-visual signals present in real-world environments, which may facilitate compelling applications in areas such as video generation (Google 2025; Mao et al. 2025a, 2024), conversational assistant (Xu et al. 2025; Kryklyvets et al. 2025), and virtual reality. In academia, researchers have investigated various fundamental tasks, such as audio-visual event parsing (Tian et al. 2018; Tian, Li, and Xu 2020; Liu et al. 2025; Zhou et al. 2021; Zhou, Guo, and Wang 2022; Zhou et al. 2023a, 2024a,b; Zhao et al. 2025; Zhou et al. 2025e,a; Yu et al. 2025; Zhou et al. 2025c), sound source localization (Chen et al. 2021; Park, Senocak, and Chung 2024; Um et al. 2025), and audio-visual caption or question answering (Shen et al. 2023; Li et al. 2022; Yun et al. 2021; Yang et al. 2022; Li et al. 2024, 2025). These tasks involve audiovisual comprehension across diverse inputs, ranging from basic audio-image pairs and short video clips to complex, untrimmed long-form audible videos. The studied referring audio-visual segmentation task is also part of this field, requiring the understanding of audio, visual, and textual modalities. **Audio-Visual Segmentation** aims to generate the pixel-level segmentation maps of the sounding objects present in audible videos (Zhou et al. 2022, 2023b; Guo et al. 2025). The baseline AVS model (Zhou et al. 2022) adopts a

convolution-based decoder. Subsequent studies explore various transformer-based (Gao et al. 2024; Li et al. 2023b; Zhou et al. 2025d; Guo, Huang, and Zhou 2024; Ma et al. 2024; Yang et al. 2024; Zhou et al. 2025b), diffusion-based (Mao et al. 2025b), and SAM-based methods (Mo and Tian 2023; Liu et al. 2024a; Wang et al. 2024a; Bhosale et al. 2025; Luo et al. 2025). AL-Ref-SAM2(Huang et al. 2025) is the most relevant to our work. AL-Ref-SAM2 performs sample-specific inference at test time using GPT-4, requiring carefully crafted prompts to select pivot frames and describe candidate bounding boxes for segmentation. In contrast, our approach fine-tunes an open-source LLM with enhanced reasoning capabilities over the reference expression, enabling it to accurately identify the true target object across video frames and simplify segmentation mask prediction.

Referring Audio-Visual Segmentation aims to generate binary masks of the object referred to by a natural language expression involving either or both auditory and visual cues. The baseline method, EEMC (Wang et al. 2024b), integrates audio, visual, and textual features through a series of transformers to form a unified multimodal cue, which then prompts a segmentation decoder, Mask2Former (Cheng et al. 2021), via cross-attention to accurately identify and delineate the object of interest. The recent method TSAM (Radman and Laaksonen 2025) enhances SAM (Kirillov et al. 2023) encoder with a temporal modeling branch and designs data-driven multimodal sparse and dense prompts for SAM’s decoder. In particular, reference texts are used to guide the selection of relevant audio cues, which then interact with visual cues to form *sparse* prompts. The *dense* prompts are also generated through cross-attention between visual-audio and visual-text features. SAM2-LOVE (Wang et al. 2025) leverages SAM2 (Ravi et al. 2024) for Ref-AVS task. It integrates textual, audio, and visual representations into a learnable segmentation token $[seg]$. To enhance spatial-temporal consistency across video frames, SAM2-LOVE employs token propagation and token accumulation strategies to strengthen $[seg]$, which serves as a more effective *sparse* prompt for SAM2’s decoder. Unlike prior methods, we utilize a reasoning-enhanced MLLM to *explicitly* identify the referred object. Our method is more explainable and does not require pixel-level supervision.

3 Our Method

In this section, we first introduce the proposed TGS-Agent framework for the Ref-AVS task (Sec. 3.1). Next, we present Ref-Thinker, the core component of TGS-Agent, which is a reasoning-enhanced MLLM designed to explicitly generate textual descriptions of the referred object. We describe its architecture and training strategy in Sec.3.2.

3.1 TGS-Agent: An Object-aware Ref-AVS Agent

As discussed in the Introduction, our TGS-Agent decomposes the Ref-AVS task into multiple object-aware processing steps. The overall system flow is shown in Fig. 2. Below, we elaborate on each stage along with its corresponding agentic tool. **Think.** As a core innovation of our method, this first step aims to clearly answer what the referred object is. To achieve

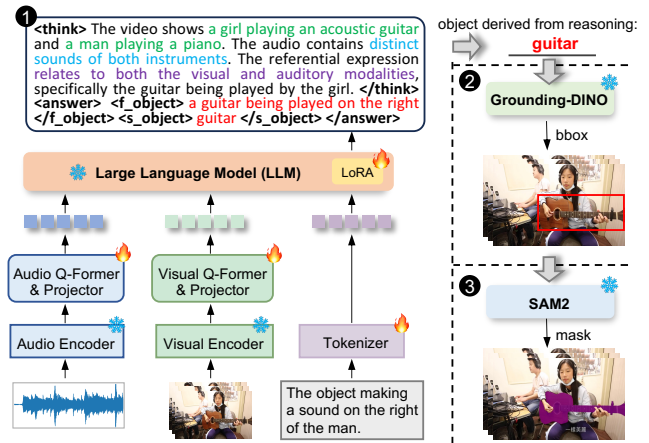


Figure 2: Workflow of our Think-Ground-Segment Agent.

this, we propose Ref-Thinker, a reasoning-enhanced MLLM, which can digest multimodal inputs (*i.e.*, audio, visual, and reference text) and output a description related to the referred object with an explicit reasoning chain. Specifically, the output text follows the format below:

== Object-aware Reasoning Chain of Ref-Thinker ==

```
<think>
The referential expression is “xxx”.
The video shows xxx (video analysis here).
The audio contains xxx (audio analysis here).
The reference related to xxx (modality analysis here).
</think>
<answer>
<f_object>
A fine-grained description of the referred object, including detailed attributes such as color, shape, or spatial location (e.g., a guitar being played on the right).
</f_object>
<s_object>
A simplified description of the referred object, specifying only its category name (e.g., guitar).
</s_object>
</answer>
```

In the thinking process, Ref-Thinker confirms the reference expression, analyzes multimodal contents, and decides which modality should be focused. Based on these analyses, Ref-Thinker provides answers with explicit information about the referred object. Notably, we explore two types of object descriptions: *fine-grained* (‘f_object’) and *simplified* (‘s_object’). The former one includes more details, such as the object’s appearance, attributes, or spatial relations (while remaining concise, typically fewer than 10 words), whereas the latter simplified description contains only the object category. We make this design considering that both the phases and class names are frequently used as prompts in object detection and segmentation. Moreover, the fine-grained description may be more beneficial for video scenes that contain different object instances that have a similar appearance or

the same category (as shown in Fig. 4).

The above process can be formalized as:

$$\mathbf{Think}(A, V, R, P) \rightarrow T, \quad (1)$$

where A , V , and R denote the audio stream, video frames, and reference, respectively. P is the user prompt text used to guide the Ref-Thinker (detailed in Sec. 3.2). T represents the generated reasoning text description. For convenience, we denote the fine-grained and simplified description of the referred object in T as T_f and T_s , respectively.

Ground. After obtaining the explicit object description T_f or T_s , the second step is to generate a bounding box (bbox) for the target object in each video frame. Regardless of whether the given reference is grounded in audio, visual, or both modalities, as illustrated in Fig. 1(a), it ultimately corresponds to a specific visible object (when it is present in the frame). We adopt Grounding-DINO(Liu et al. 2024b) as the tool for bounding box generation. Grounding-DINO is an advanced object detection model that combines the Transformer-based detector DINO(Zhang et al. 2022) with large-scale grounded pre-training, enabling it to detect arbitrary objects based on human language inputs, such as category names or referring expressions. Therefore, our object descriptions (T_f/T_s) are closely aligned with the model’s input format and intended usage. We denote this process as:

$$\mathbf{Ground}(T_f/T_s, V) \rightarrow B, \quad (2)$$

where $V = \{v_i\}_{i=1}^N$ denotes the N video frames, and $B = \{(x_1, y_1), (x_2, y_2)\}^N$ is corresponding bounding box set, with (x_1, y_1) and (x_2, y_2) indicating the coordinates of the top-left and bottom-right corners, respectively. Notably, two key hyper-parameters are involved in the bbox generation: τ_{bbox} controls the minimum confidence score for a bounding box to be considered valid, while τ_{text} sets the minimum confidence score for the text to be matched.

Segment. After obtaining the bounding box B of the referred object in each video frame, we use this explicit and strong guidance as a sparse prompt to drive the powerful segmentation foundation model, SAM2 (Ravi et al. 2024), to generate the corresponding segmentation masks. In particular, if the bounding box does not exist (*i.e.*, no matching object is found in the video frame), the segmentation mask defaults to all background pixels. We formalize this process as:

$$\mathbf{Segment}(B, V) \rightarrow M. \quad (3)$$

$M = \{m_i\}_{i=1}^N$ is the set of binary masks for N video frames. Notably, the prior SOTA method (Wang et al. 2025) also employs SAM2 but requires fine-tuning its mask decoder. In contrast, our method leverages the frozen SAM2 and achieves superior performance.

In this way, our TGS-Agent completes the transformation from audiovisual streams + reference text \rightarrow object description \rightarrow bbox \rightarrow mask, demonstrating an object-aware, reliable, and explainable decision process for Ref-AVS task.

3.2 Ref-Thinker: A Reasoning-enhanced MLLM

Architecture. As shown on the left side of Fig. 2, our Ref-Thinker is implemented as an audio-visual LLM. Given temporal audio and visual streams, modality-specific encoders

are used to extract segment-level features. The resulting audio and visual features are then compressed into a smaller number of token embeddings using independent Q-Formers (Li et al. 2023a). A projector module, implemented as two MLP layers, is further applied to align the audio/visual feature with the textual features processed by the LLM. The textual input consists of two parts: a task-relevant user prompt containing the specific reference expression, and the corresponding reasoning chain as the expected output. We provide the template of the user prompt (P) below:

==== User Prompt Template for Ref-Thinker ====
 This is a video:<video_start><video><video_end>.
 This is an audio:<audio_start><audio><audio_end>.
 Given the referential expression xxx, analyze the video and audio, and then generate a reasoning chain (<think>) and final answer (<answer>). Your output must follow this format: (reasoning chain introduced in Sec. 3.1).

Audio and visual feature embeddings are used to replace the placeholders <audio> and <video> tokens in the user prompt. The resulting textual input is then tokenized and passed to the LLM backbone for text generation.

Training Recipe. We train our Ref-Thinker in two phases. During the *pretraining* phase, the LLM is frozen, and domain-specific caption datasets(Kim et al. 2019; Lin et al. 2023) are independently used to train the audio and visual Q-Formers along with their corresponding projectors. Next, in the *instruction tuning* phase, we apply the LoRA (Hu et al. 2022) technique for parameter-efficient tuning of the LLM. This phase not only aims to improve the model’s instruction following ability, but, more importantly, enables the MLLM to better learn an object-aware reasoning chain, as introduced in Sec. 3.1. To support this, we construct an instruction tuning set. Specifically, given a reference expression and a video from the training set of the official Ref-AVSBench dataset (Wang et al. 2024b), we leverage Gemini-1.5-Pro (Google 2024) with carefully designed prompts to analyze the video and generate an object-aware reasoning chain that strictly follows the think-answer format shown in Sec. 3.1. Detailed prompt is provided in our Appendix A. Supervised with this high-quality instruction tuning set, our Ref-Thinker can effectively think about the reference (and audiovisual signals) and generate accurate object descriptions with an explicit reasoning chain. For both training phases, the autoregressive cross-entropy loss is used for optimization.

4 R²-AVSBench

We propose R²-AVSBench as an additional Reasoning-enhanced evaluation set for the Ref-AVS task, motivated by two key factors: 1) We observed that some reference expressions in the standard Ref-AVSBench (Wang et al. 2024b) test set are relatively simple. For example, ‘*The couch sit by a woman*’ (referred object: couch) and ‘*The object making a sound by using a saw*’ (referred object: man holding the saw). In these cases, the object name *couch* or key cues *saw* are explicitly mentioned in the reference, which may lead

Method	Venue	Task	Seen (S) \uparrow			Unseen (U) \uparrow			Mix (S+U) \uparrow			Null \downarrow
			\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{S}
AVSBench (Zhou et al. 2022)	ECCV'22	AVS	23.2	51.1	37.2	32.4	54.7	43.5	27.8	52.9	40.3	0.208
AVSegFormer (Gao et al. 2024)	AAAI'24		34.5	47.0	40.2	36.1	50.1	43.1	34.8	48.6	41.7	0.171
GAVS (Wang et al. 2024a)	AAAI'24		28.9	49.8	39.4	29.8	49.7	39.8	29.4	49.8	39.6	0.190
SAMA (Liu et al. 2024a)	AAAI'24		28.9	49.8	39.4	29.8	49.7	39.8	29.4	49.8	39.6	0.190
ReferFormer (Wu et al. 2022)	CVPR'22	Ref-VOS	31.3	50.1	40.7	30.4	48.8	39.6	30.9	49.5	40.2	0.176
R2VOS (Li et al. 2023c)	ICCV'23		25.0	41.0	33.0	27.9	49.8	38.9	26.5	45.4	35.9	0.183
EEMC (Wang et al. 2024b)	ECCV'24	Ref-AVS	34.2	51.3	42.8	49.5	64.8	57.2	41.9	58.1	50.0	0.007
Grounded-SAM2 (Ren et al. 2024)	ArXiv'24		28.5	39.9	34.2	59.8	68.1	63.9	44.2	54.0	49.1	0.277
Crab (Du et al. 2025)	CVPR'25		40.5	-	-	45.6	-	-	43.1	-	-	-
TSAM (Radman and Laaksonen 2025)	CVPR'25		43.4	<u>56.8</u>	<u>50.1</u>	54.6	66.4	60.5	49.0	61.6	55.3	<u>0.017</u>
SAM2-LOVE (Wang et al. 2025)	CVPR'25		<u>43.5</u>	51.9	47.7	<u>66.5</u>	<u>72.3</u>	<u>69.4</u>	<u>55.0</u>	<u>62.1</u>	<u>58.5</u>	0.23
TGS-Agent (ours)	-	Ref-AVS	49.5	60.4	54.9	73.2	80.6	76.9	61.3	70.5	65.9	0.035

Table 1: Comparison with prior methods on Ref-AVSBench test set. $\mathcal{J}\&\mathcal{F}$ is the average value of \mathcal{J} and \mathcal{F} .

Ref. Source	Method	Seen (S) \uparrow			Unseen (U) \uparrow			Mix (S+U) \uparrow		
		\mathcal{J}	\mathcal{F}	$\mathcal{J}\&\mathcal{F}$	\mathcal{J}	\mathcal{F}	$\mathcal{J}\&\mathcal{F}$	\mathcal{J}	\mathcal{F}	$\mathcal{J}\&\mathcal{F}$
Ref-AVSBench	Crab (Du et al. 2025)	20.7	33.8	27.2	39.4	54.9	47.2	30.1	44.4	37.2
	EEMC (Wang et al. 2024b)	26.4	40.9	33.7	39.7	55.9	47.8	33.1	48.4	40.7
	Grounded-SAM2 (Ren et al. 2024)	41.1	50.8	45.9	68.1	76.2	72.2	54.6	63.5	59.0
	TGS-Agent (ours)	48.5	58.4	53.4	71.8	80.4	76.1	60.1	69.4	64.8
R ² -AVSBench	Crab (Du et al. 2025)	19.6	32.1	25.9	36.9	52.0	44.4	28.2	42.1	35.2
	EEMC (Wang et al. 2024b)	22.6	39.9	31.2	38.7	56.9	47.8	30.6	48.4	39.5
	Grounded-SAM2 (Ren et al. 2024)	21.4	30.1	25.7	44.9	53.4	49.1	33.2	41.7	37.4
	TGS-Agent (ours)	42.7	52.3	47.5	68.3	77.0	72.7	55.5	64.7	60.1

Table 2: Evaluation on R²-AVSBench. We also report results on the same set of videos using original Ref-AVSBench references.

2) Crab (Du et al. 2025) is also an MLLM, designed for unified audio-visual task learning, including Ref-AVS. Despite using the same LLM backbone, our method significantly outperforms Crab. These improvements are attributed to our more effective agentic framework design and the instruction tuning strategy enhanced by explicit object-aware reasoning.

5.3 Evaluation on R²-AVSBench

We evaluate our method on the proposed R²-AVSBench. We primarily compare against those open-sourced methods: Crab (Du et al. 2025), EEMC (Wang et al. 2024b), and Grounded-SAM2 (Ren et al. 2024). Notably, all these pre-trained models are directly evaluated on R²-AVSBench. As shown in Table 2 (bottom four rows), our method consistently outperforms other competitors by a large margin. The upper part of the table reports results obtained on the same test videos but using the original Ref-AVSBench references. We can observe a consistent performance drop for all models when evaluated with the transformed references from R²-AVSBench, indicating the increased complexity. In particular, Grounded-SAM2 suffers a substantial drop of around 20% on both *Seen* and *Unseen* sets, highlighting its sensitivity to reference variation. In contrast, benefiting from the object-aware reasoning mechanism, our method can accurately identify the referred object from multimodal cues, thereby maintaining superior performance and demonstrating stronger generalizability across diverse and challenging reference types.

Setup	Seen (S) \uparrow			Unseen (U) \uparrow			Null \downarrow
	\mathcal{J}	\mathcal{F}	$\mathcal{J}\&\mathcal{F}$	\mathcal{J}	\mathcal{F}	$\mathcal{J}\&\mathcal{F}$	\mathcal{S}
Reference (R)	28.5	39.9	34.2	59.8	68.1	63.9	0.277
F-object (T_f)	43.9	54.1	49.0	69.9	77.4	73.7	0.043
S-object (T_s)	49.5	60.4	54.9	73.2	80.6	76.9	0.035

Table 3: Ablation on object text used as prompts.

5.4 Ablation Study and Qualitative Results

In this section, the experiments are conducted on the Ref-AVSBench dataset unless otherwise specified.

Impact of Object Description Types. Our method designs Ref-Thinker to generate two types of object descriptions, *i.e.*, a fine-grained version T_f containing more details (f_object), and a simplified version T_s consisting only of the object name (s_object). These are used as prompts in the *Ground* phase for bounding box generation. As shown in Table 3, compared to using original reference expressions, both T_f and T_s significantly improve the performance, highlighting the benefits of explicit object-aware reasoning. Interestingly, the simplified version (*i.e.*, object category) yields better performance. We hypothesize that this is related to the Grounding-DINO detector’s preference, which may be more friendly to short and specific prompts. However, in some cases, the fine-grained object phase can be more effective, especially when there are

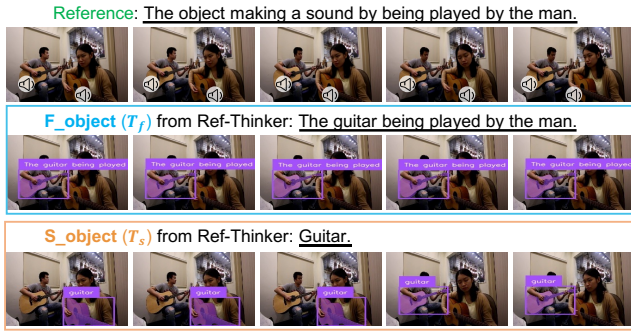


Figure 4: Visualization results of using different object descriptions as detection prompts. The fine-grained phase T_f may be more helpful in video scenes containing multiple instances of the same category (i.e., *guitar* in this example).

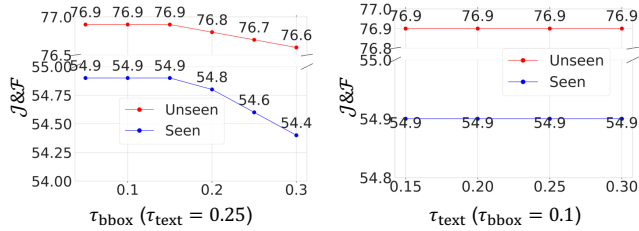


Figure 5: Parameter study on the τ_{bbox} and τ_{text} .

multiple similar/same object instances. As shown in Fig. 4, two guitars are played by different individuals. Using only the simplified object class *guitar* (T_s) as a prompt, the object detector is confused when recognizing the correct guitar at instance level. In contrast, the fine-grained phase (T_f) provides richer cues, specifying that the referred guitar is *played by the man*, which helps to improve the detection and segmentation results. Nevertheless, even using T_f , our method surpasses the prior state-of-the-art SAM2-LOVE (Wang et al. 2025) (Table 1), demonstrating the superiority of our method.

Impact of Threshold Parameters. The *Ground* phase of our TGS-Agent involves two hyperparameters, namely the bounding box threshold τ_{bbox} and the text similarity threshold τ_{text} , which are used to control the selection of object bounding boxes. We investigate their influence and present the results in Fig. 5. The model performance shows slight variation across different values of τ_{bbox} , while remaining stable with respect to τ_{text} . We set $\tau_{bbox} = 0.1$ and $\tau_{text} = 0.25$ as the default configuration in our experiments.

Qualitative Analysis. Fig. 6 presents some qualitative results. In the video example, a man is playing a bassoon while a woman is playing a piano. The referred object is the *piano*. Prior method EEMC (Wang et al. 2024b) identifies the piano but only segments part of it. Grounded-SAM2, which directly uses the reference expression for prediction, incorrectly segments another object, e.g., *the woman*, mentioned in the reference text. In contrast, our TGS-Agent method accurately identifies the referred object and achieves satisfactory segmentation. This improvement stems from our Ref-Thinker, which thoroughly analyzes the reference alongside the au-

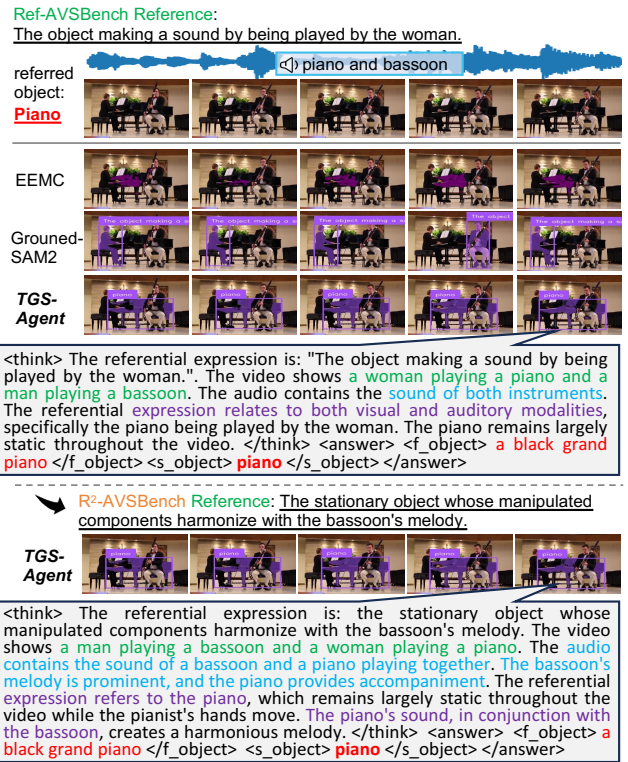


Figure 6: Qualitative example of the Ref-AVS task.

dio and visual content, and then explicitly outputs the target object for downstream tool invocation. The corresponding reasoning process is visualized in the figure. In addition, we provide results of our model on more complex reference from our R^2 -AVSBench. As illustrated, our Ref-Thinker can adapt its reasoning to new reference and still successfully identify the referred object. We provide more qualitative results and some failure case studies in Appendix D&E.

6 Conclusion

We propose TGS-Agent, addressing the Ref-AVS task through an agentic *Think-Ground-Segment* paradigm. At its core is Ref-Thinker, a reasoning-enhanced MLLM that transforms multimodal inputs (reference text and audio-visual streams) into explicit descriptions of the referred object, which better prompts subsequent grounding and segmentation processes. To further benchmark model robustness, we introduce R^2 -AVSBench, a new evaluation set featuring linguistically diverse and reasoning-intensive references. R^2 -AVSBench can be used for assessing a model's cross-reference generalization and may benefit the broader community as a challenging testbed. Our TGS-Agent operates without requiring pixel-level supervision, offering greater explainability. Despite its simplicity, it significantly outperforms prior models. Overall, our work highlights the importance of *explicit reference understanding* in Ref-AVS. Future directions include integrating our Ref-Thinker with SAM2 tuning-based methods to further enhance the performance.

Acknowledgments

We would like to express our sincere gratitude to the anonymous reviewers for their invaluable comments and insightful suggestions. This work is partially supported by the National Natural Science Foundation of China (NSFC) under Grant No. 62573399 and U25A20530, the *Google Research Gift Award 2023*, *Meta Regional Research Grant 2025*, and the *NVIDIA Academic Grant 2025*.

References

- Bhosale, S.; Yang, H.; Kanojia, D.; Deng, J.; and Zhu, X. 2025. Unsupervised audio-visual segmentation with modality alignment. In *AAAI*, volume 39, 15567–15575.
- Chen, H.; Xie, W.; Afouras, T.; Nagrani, A.; Vedaldi, A.; and Zisserman, A. 2021. Localizing visual sounds the hard way. In *CVPR*, 16867–16876.
- Chen, S.; Wu, Y.; Wang, C.; Liu, S.; Tompkins, D.; Chen, Z.; and Wei, F. 2022. Beats: Audio pre-training with acoustic tokenizers. *arXiv preprint arXiv:2212.09058*.
- Cheng, B.; Choudhuri, A.; Misra, I.; Kirillov, A.; Girdhar, R.; and Schwing, A. G. 2021. Mask2former for video instance segmentation. *arXiv preprint arXiv:2112.10764*.
- Du, H.; Li, G.; Zhou, C.; Zhang, C.; Zhao, A.; and Hu, D. 2025. Crab: A unified audio-visual scene understanding model with explicit cooperation. In *CVPR*, 18804–18814.
- Gao, S.; Chen, Z.; Chen, G.; Wang, W.; and Lu, T. 2024. Avsegformer: Audio-visual segmentation with transformer. In *AAAI*, volume 38, 12155–12163.
- Google. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. https://storage.googleapis.com/deepmind-media/gemini/gemini_v1_5_report.pdf.
- Google. 2025. Veo: a text-to-video generation system. <https://storage.googleapis.com/deepmind-media/veo/Veo-3-Tech-Report.pdf>.
- Guo, C.; Huang, H.; and Zhou, Y. 2024. Enhance audio-visual segmentation with hierarchical encoder and audio guidance. *Neurocomputing*, 594: 127885.
- Guo, R.; Ying, X.; Chen, Y.; Niu, D.; Li, G.; Qu, L.; Qi, Y.; Zhou, J.; Xing, B.; Yue, W.; et al. 2025. Audio-visual instance segmentation. In *CVPR*, 13550–13560.
- Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. In *ICLR*.
- Huang, S.; Ling, R.; Li, H.; Hui, T.; Tang, Z.; Wei, X.; Han, J.; and Liu, S. 2025. Unleashing the temporal-spatial reasoning capacity of gpt for training-free audio and language referenced video object segmentation. In *AAAI*, volume 39, 3715–3723.
- Kim, C. D.; Kim, B.; Lee, H.; and Kim, G. 2019. Audiocaps: Generating captions for audios in the wild. In *ACL*, 119–132.
- 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.
- Kryklyvets, Y.; Kurpath, M. I.; Mullappilly, S. S.; Zhou, J.; Khan, F. S.; Anwer, R. M.; Khan, S.; and Cholakkal, H. 2025. MAViS: A Multimodal Conversational Assistant For Avian Species. In *EMNLP*, 28601–28627.
- Li, G.; Wei, Y.; Tian, Y.; Xu, C.; Wen, J.-R.; and Hu, D. 2022. Learning to answer questions in dynamic audio-visual scenarios. In *CVPR*, 19108–19118.
- Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 19730–19742. PMLR.
- Li, K.; Yang, Z.; Chen, L.; Yang, Y.; and Xiao, J. 2023b. Catr: Combinatorial-dependence audio-queried transformer for audio-visual video segmentation. In *ACM MM*, 1485–1494.
- Li, X.; Wang, J.; Xu, X.; Li, X.; Raj, B.; and Lu, Y. 2023c. Robust referring video object segmentation with cyclic structural consensus. In *ICCV*, 22236–22245.
- Li, Z.; Guo, D.; Zhou, J.; Zhang, J.; and Wang, M. 2024. Object-aware adaptive-positivity learning for audio-visual question answering. In *AAAI*, 3306–3314.
- Li, Z.; Zhou, J.; Zhang, J.; Tang, S.; Li, K.; and Guo, D. 2025. Patch-level sounding object tracking for audio-visual question answering. In *AAAI*, 5075–5083.
- Lin, B.; Ye, Y.; Zhu, B.; Cui, J.; Ning, M.; Jin, P.; and Yuan, L. 2023. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*.
- Liu, J.; Wang, Y.; Ju, C.; Ma, C.; Zhang, Y.; and Xie, W. 2024a. Annotation-free audio-visual segmentation. In *WACV*, 5604–5614.
- Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Jiang, Q.; Li, C.; Yang, J.; Su, H.; et al. 2024b. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *ECCV*, 38–55. Springer.
- Liu, X.; Xia, N.; Zhou, J.; Li, Z.; and Guo, D. 2025. Towards energy-efficient audio-visual classification via multimodal interactive spiking neural network. *ACM TOMM*, 1–24.
- Luo, Z.; Liu, N.; Yang, X.; Khan, S.; Anwer, R. M.; Cholakkal, H.; Khan, F. S.; and Han, J. 2025. TAViS: Text-bridged Audio-Visual Segmentation with Foundation Models. In *ICCV*, 1–10.
- Ma, J.; Sun, P.; Wang, Y.; and Hu, D. 2024. Stepping stones: a progressive training strategy for audio-visual semantic segmentation. In *ECCV*, 311–327. Springer.
- Mao, Y.; Qin, Z.; Zhou, J.; Deng, H.; Shen, X.; Fan, B.; Zhang, J.; Zhong, Y.; and Dai, Y. 2025a. Autoregressive Image Generation with Linear Complexity: A Spatial-Aware Decay Perspective. *arXiv preprint arXiv:2507.01652*.
- Mao, Y.; Shen, X.; Zhang, J.; Qin, Z.; Zhou, J.; Xiang, M.; Zhong, Y.; and Dai, Y. 2024. TAVGBench: Benchmarking text to audible-video generation. In *ACM MM*, 6607–6616.
- Mao, Y.; Zhang, J.; Xiang, M.; Lv, Y.; Li, D.; Zhong, Y.; and Dai, Y. 2025b. Contrastive conditional latent diffusion for audio-visual segmentation. *IEEE TIP*.

- Mo, S.; and Tian, Y. 2023. AV-SAM: Segment anything model meets audio-visual localization and segmentation. *arXiv preprint arXiv:2305.01836*.
- Park, S.; Senocak, A.; and Chung, J. S. 2024. Can clip help sound source localization? In *WACV*, 5711–5720.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, 8748–8763. PmLR.
- Radman, A.; and Laaksonen, J. 2025. TSAM: Temporal SAM Augmented with Multimodal Prompts for Referring Audio-Visual Segmentation. In *CVPR*, 23947–23956.
- Ravi, N.; Gabeur, V.; Hu, Y.-T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rädle, R.; Rolland, C.; Gustafson, L.; Mintun, E.; Pan, J.; Alwala, K. V.; Carion, N.; Wu, C.-Y.; Girshick, R.; Dollár, P.; and Feichtenhofer, C. 2024. SAM 2: Segment Anything in Images and Videos. *arXiv preprint arXiv:2408.00714*.
- Ren, T.; Liu, S.; Zeng, A.; Lin, J.; Li, K.; Cao, H.; Chen, J.; Huang, X.; Chen, Y.; Yan, F.; et al. 2024. Grounded SAM 2: Ground and Track Anything in Videos. <https://github.com/IDEA-Research/Grounded-SAM-2>.
- Shen, X.; Li, D.; Zhou, J.; Qin, Z.; He, B.; Han, X.; Li, A.; Dai, Y.; Kong, L.; Wang, M.; et al. 2023. Fine-grained audible video description. In *CVPR*, 10585–10596.
- Tian, Y.; Li, D.; and Xu, C. 2020. Unified multisensory perception: Weakly-supervised audio-visual video parsing. In *ECCV*, 436–454.
- Tian, Y.; Shi, J.; Li, B.; Duan, Z.; and Xu, C. 2018. Audio-visual event localization in unconstrained videos. In *ECCV*, 247–263.
- Um, S. J.; Kim, D.; Lee, S.; and Kim, J. U. 2025. Object-aware Sound Source Localization via Audio-Visual Scene Understanding. In *CVPR*, 8342–8351.
- 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 *ECCV*, 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 *CVPR*, 28932–28941.
- Wu, J.; Jiang, Y.; Sun, P.; Yuan, Z.; and Luo, P. 2022. Language as queries for referring video object segmentation. In *CVPR*, 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, P.; Wang, X.; Duan, X.; Chen, H.; Hou, R.; Jin, C.; and Zhu, W. 2022. Avqa: A dataset for audio-visual question answering on videos. In *ACM MM*, 3480–3491.
- 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.
- Yu, X.; Fang, Y.; Jin, X.; Zhao, Y.; and Wei, Y. 2025. PreFM: Online Audio-Visual Event Parsing via Predictive Future Modeling. *arXiv preprint arXiv:2505.23155*.
- Yun, H.; Yu, Y.; Yang, W.; Lee, K.; and Kim, G. 2021. Pano-avqa: Grounded audio-visual question answering on 360deg videos. In *ICCV*, 2031–2041.
- Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L. M.; and Shum, H.-Y. 2022. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*.
- Zhao, P.; Zhou, J.; Zhao, Y.; Guo, D.; and Chen, Y. 2025. Multimodal class-aware semantic enhancement network for audio-visual video parsing. In *AAAI*, 10448–10456.
- Zhou, J.; Guo, D.; Guo, R.; Mao, Y.; Hu, J.; Zhong, Y.; Chang, X.; and Wang, M. 2025a. Towards open-vocabulary audio-visual event localization. In *CVPR*, 8362–8371.
- Zhou, J.; Guo, D.; Mao, Y.; Zhong, Y.; Chang, X.; and Wang, M. 2024a. Label-anticipated event disentanglement for audio-visual video parsing. In *ECCV*, 35–51.
- Zhou, J.; Guo, D.; and Wang, M. 2022. Contrastive positive sample propagation along the audio-visual event line. *IEEE TPAMI*, 45(6): 7239–7257.
- Zhou, J.; Guo, D.; Zhong, Y.; and Wang, M. 2023a. Improving audio-visual video parsing with pseudo visual labels. *arXiv preprint arXiv:2303.02344*.
- Zhou, J.; Guo, D.; Zhong, Y.; and Wang, M. 2024b. Advancing weakly-supervised audio-visual video parsing via segment-wise pseudo labeling. *International Journal of Computer Vision*, 132(11): 5308–5329.
- Zhou, J.; Li, Z.; Yu, Y.; Zhou, Y.; Guo, R.; Li, G.; Mao, Y.; Han, M.; Chang, X.; and Wang, M. 2025b. Mettle: Meta-Token Learning for Memory-Efficient Audio-Visual Adaptation. *arXiv preprint arXiv:2506.23271*.
- Zhou, J.; Shen, X.; Wang, J.; Zhang, J.; Sun, W.; Zhang, J.; Birchfield, S.; Guo, D.; Kong, L.; Wang, M.; et al. 2023b. 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.
- Zhou, J.; Zheng, L.; Zhong, Y.; Hao, S.; and Wang, M. 2021. Positive sample propagation along the audio-visual event line. In *CVPR*, 8436–8444.
- Zhou, J.; Zhou, Z.; Zhou, Y.; Mao, Y.; Duan, Z.; and Guo, D. 2025c. Clasp: Cross-modal salient anchor-based semantic propagation for weakly-supervised dense audio-visual event localization. *arXiv preprint arXiv:2508.04566*.
- Zhou, Y.-H.; Huang, H.; Guo, C.; Tu, R.-C.; Xiao, Z.; Wang, B.; and Mao, X.-L. 2025d. ALOHA: Adapting Local Spatio-Temporal Context to Enhance the Audio-Visual Semantic Segmentation. *ACM TOMM*.
- Zhou, Z.; Zhou, J.; Qian, W.; Tang, S.; Chang, X.; and Guo, D. 2025e. Dense audio-visual event localization under cross-modal consistency and multi-temporal granularity collaboration. In *AAAI*, 10905–10913.