REF-AVIS: REFERRING AUDIO-VISUAL INSTANCE SEGMENTATION IN LONG VIDEOS

Anonymous authors
Paper under double-blind review

ABSTRACT

In real-world audio-visual scenes, people use language to specify particular objects among multiple candidates. Existing benchmarks, however, provide limited support for independently segmenting and tracking these objects in complex, long videos. We introduce Ref-AVIS, a benchmark for referring audio-visual instance segmentation comprising 1,187 videos across 26 sound-source categories, with an average duration of 59.88 seconds. It covers scenes with multiple objects of the same or different categories and provides diverse referring expressions, independent masks, and persistent identities. We further develop a baseline that incorporates video-level referring understanding into instance segmentation through competitive instance grounding, audio-guided adaptation, and audio-visual discrimination. These designs enable independent segmentation and consistent tracking of multiple referred objects in dynamic audio-visual scenes. Extensive quantitative and qualitative experiments show that accurate foreground segmentation does not necessarily translate into reliable instance tracking, while demonstrating our baseline's improvements in both segmentation and tracking on Ref-AVIS. Together, Ref-AVIS and our baseline support the development and evaluation of multimodal models for language-guided instance segmentation and tracking in long videos. Our dataset and baseline code will be made publicly available upon acceptance.

1 INTRODUCTION

In the real world, people use language to specify objects in dynamic scenes where visual and acoustic information are closely intertwined in a continuous flow. Inspired by this ability, research on referring audio-visual segmentation explores how referring expressions can guide object segmentation using visual and acoustic information. Such language-guided interaction has practical value for applications including human–computer interaction, augmented reality, and intelligent multimedia processing.

Referring audio-visual segmentation (Ref-AVS) has introduced natural-language guidance into audio-visual segmentation, allowing objects to be specified through referring expressions that describe visual and acoustic cues (Wang et al., 2024). In parallel, audio-visual instance segmentation (AVIS) studies the segmentation and tracking of individual sounding objects across video frames (Guo et al., 2025). Its instance-level annotations distinguish objects of the same category and preserve their identities over time. Integrating referring expressions with this instance-level formulation would allow multiple objects to be selected by a single expression and tracked individually, supporting finer control over which objects are segmented in dynamic audio-visual scenes.

Associating language and sound with visual objects, while maintaining these associations over time, is central to these tasks. EEMC addresses referring segmentation by enriching referring expressions with visual and acoustic features and incorporating temporal changes into the segmentation process. OISA, the OmniAVS baseline, uses temporally interleaved audio-visual inputs in a multimodal large language model (MLLM) to interpret multimodal referring expressions (Ying et al., 2025). However, these methods offer limited support for distinguishing multiple referred objects as independent instances and maintaining their identities over time. At the instance level, AVISM uses window-based attention over object representations to associate frame-level segmentations across a video, but does not support target selection through referring expressions.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



Figure 1: **Comparison of Ref-AVIS with Ref-AVS, OmniAVS, and AVIS.** Top: Ref-AVIS supports independent segmentation and tracking of multiple referred instances in long videos. Bottom: Different referring expressions specify different target selections within the same scene.

Despite these advances, existing benchmarks provide limited support for evaluating referring instance segmentation in complex, long videos. Ref-AVS and OmniAVS primarily contain clips of around 10 seconds, restricting the temporal span over which target selection and tracking are evaluated. Moreover, Ref-AVS permits at most one target per referring expression, limiting the evaluation of multi-target instance segmentation and tracking. OmniAVS supports multiple targets but does not explicitly assess the identity of each selected object across frames. OmniAVS also suffers from a mismatch between the temporal conditions in referring expressions and its annotations (e.g., a person described as "currently coughing" remains annotated after the coughing stops). To support instance-level tracking over longer sequences, AVISeg provides videos with per-instance masks and consistent identities, but lacks referring expressions for selecting particular objects. These limitations leave a need for a benchmark that combines target selection through referring expressions with individual object segmentation and sustained tracking over longer sequences.

To address these limitations, we introduce Ref-AVIS, a task and dataset for referring audio-visual instance segmentation in long videos. The dataset contains 1,187 videos covering 26 sound-source categories, with 71,033 annotated frames over 19.74 hours. Multiple instances of the same or different categories appear in 87.95% of the videos. Ref-AVIS also provides diverse referring expressions for each video, allowing different objects to be specified within the same scene. This offers varied guidance for instance selection, especially when an expression refers to only a subset of objects from the same category. The benchmark therefore evaluates both whether models select the requested objects from multi-object scenes and whether they segment and track each selected object individually. Each referred instance is segmented only in frames where its state is consistent with the referring expression. In addition, with videos averaging 59.88 seconds, these requirements extend over longer sequences as object activity and scene content change.

Alongside this benchmark, we develop a baseline that integrates multimodal referring understanding with audio-visual segmentation through persistent instance representations. The framework incorporates three key designs: (1) **Competitive instance grounding**, which translates multimodal referring understanding into independent instance representations by adapting a pretrained MLLM and competitively aggregating visual features across frames. Each representation provides a shared identity reference for the corresponding object's mask predictions throughout the video, supporting both independent segmentation and consistent tracking. (2) **Audio-guided instance adaptation**, which integrates continuous acoustic information with the video-level instance representations to form segmentation prompts for SAM2 (Ravi et al., 2024). This design complements persistent

Table 1: Comparison of task characteristics and dataset statistics. Long videos denote an average duration of at least 20 seconds; state-aware masks are annotated only in frames where the target matches the referring expression, rather than in every frame where it remains visible.

Dataset	Modalities	Task characteristics				Statistics			
		Instance tracking	Multi-target references	State-aware masks	Long videos	Videos	Average length (s)	Total hours	Text references
Ref-AVS	A+V+T	✗	✗	✗	✗	3,993	10.01	11.10	20,434
AVISeg	A+V	✓	—	—	✓	926	61.44	15.80	—
OmniAVS	A+V+T+S	✗	✓	✗	✗	2,104	10.41	6.07	25,474
Ref-AVIS	A+V+T	✓	✓	✓	✓	1,187	59.88	19.74	20,170

target information with the changing auditory context, enabling segmentation to account for the current activity of each referred object. (3) **Instance-level audio-visual discrimination**, which relates each candidate region's appearance and concurrent audio to the representation of the referred instance. This assesses whether the candidate instance matches the referring expression under current audio-visual conditions, while SAM2's mask-quality estimate indicates how well the predicted mask matches the object. These assessments help suppress visually plausible masks that no longer match the referring expression. Together, these designs enable independent segmentation and consistent tracking of multiple referred objects under changing audio-visual conditions.

Our main contributions are as follows:

- **Benchmark.** We introduce Ref-AVIS, a benchmark for developing and evaluating multimodal models on long videos containing multiple objects of the same or different categories. It provides a unified testbed for advancing language-guided audio-visual understanding through independent instance segmentation and long-term tracking.
- **Method.** We develop a baseline that adapts multimodal language and segmentation models to combine instance selection, identity preservation, and segmentation under changing audio-visual conditions, providing a reference architecture for language-guided audio-visual scene understanding.
- **Experiments.** Extensive quantitative and qualitative experiments across Ref-AVIS, Ref-AVS, and AVIS demonstrate the value of our benchmark for evaluating multimodal instance understanding and the effectiveness of our baseline.

2 Dataset

2.1 Data Collection and Annotation

Data Collection and Composition. Ref-AVIS contains 1,187 videos spanning 26 sound-source categories, with 71,033 annotated frames and 20,170 English referring expressions over 19.74 hours. It includes 862 videos from AVISeg and 325 additional videos collected from YouTube, with the additional collection emphasizing same-category multi-instance scenes. Figure 2(b) summarizes the category composition. We divide the videos into 832 training, 119 validation, and 236 test clips, keeping clips derived from the same original video within a single split.

Referring Expression Design. We organize referring expressions into four target-selection relations: 1–1 specifies an individual instance; 1–n specifies multiple instances, including subsets of the available candidates; 1–all exhaustively selects the instances within the scope of the expression; and 1–null has no matching instance. Each video is paired with multiple expressions specifying different target selections. The instances selected by 1–all expressions may belong to the same category or to different categories. For 1–null expressions, either the specified category is absent from the video, or objects of that category are present but none matches the full referring expression. Figure 2(e) summarizes the four relations, and Figure 2(d) presents the subtype distributions of 1–all and 1–null.

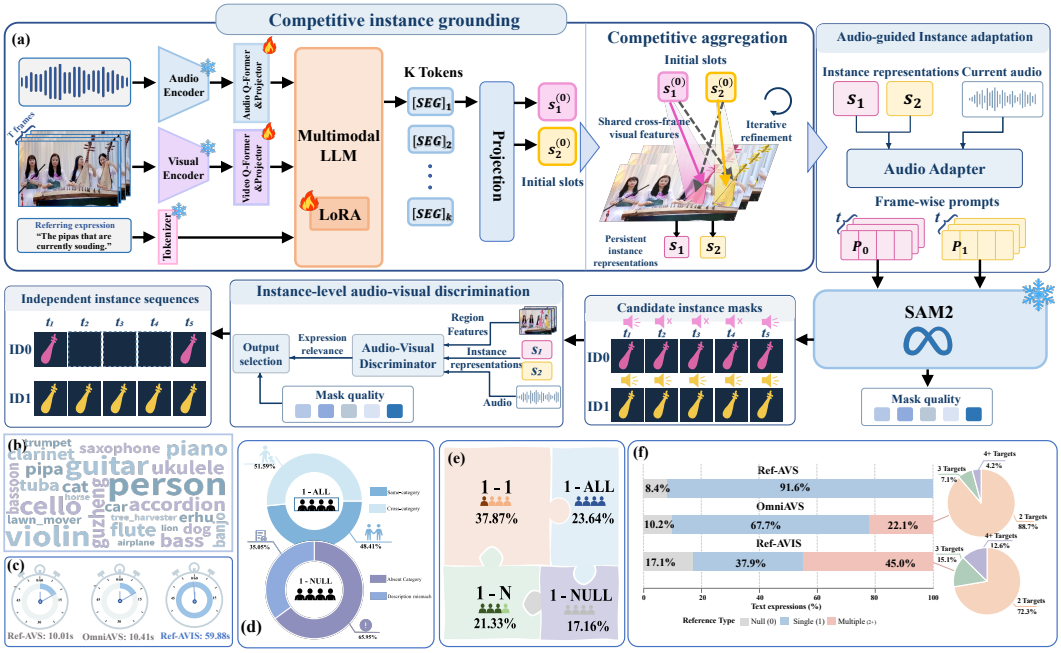


Figure 2: **Overview of RAVISeg and Ref-AVIS.** (a) RAVISeg for independent instance segmentation and tracking. (b) Sound-source categories. (c) Average video durations. (d) Subtype distributions of 1–all and 1–null expressions. (e) Distribution of the four referring relations. (f) Target-count distributions across benchmarks; pie charts show the breakdown within multi-target expressions.

Annotation Procedure. We reuse AVISeg's instance annotations and annotate the additional videos using SAM-assisted segmentation followed by manual refinement. Annotators inspect complete audio-visual clips, construct referring expressions, and associate each expression with its target instance IDs. Expressions may rely on visual cues, acoustic cues, or their combination. The temporal extent of the masks follows the conditions specified by each expression. When sounding is explicitly required, masks are restricted to frames in which the target is sounding and visibly segmentable, while its identity remains consistent across frames. Human review checks expression–instance correspondences, temporal validity, and the consistency of instance identities.

2.2 Dataset Statistics and Analysis

Multi-instance Scenes and Annotations. Multiple annotated instances occur in 87.95% of Ref-AVIS videos, including objects from the same or different categories. As illustrated in Figure 1, each selected object has independent masks and a consistent identity across frames. While Ref-AVS permits at most one target per expression and OmniAVS evaluates merged target masks, Ref-AVIS supports evaluating the segmentation and tracking of multiple selected instances individually.

Multi-target Referring Expressions. Beyond the presence of multiple objects in a scene, Ref-AVIS places greater emphasis on expressions that select multiple instances. Figure 2(f) compares the number of target instances per expression across Ref-AVS, OmniAVS, and Ref-AVIS. Multi-target expressions account for 45.0% of Ref-AVIS annotations, compared with 22.1% of OmniAVS text expressions, whereas Ref-AVS permits at most one target per expression. These proportions are computed over all text expressions, including null references. This places greater emphasis on identifying all objects specified by an expression without including other candidates.

Long-video Coverage. Videos in Ref-AVIS average 59.88 seconds, compared with 10.01 seconds for Ref-AVS and 10.41 seconds for OmniAVS, as shown in Figure 2(c). This extends referring instance segmentation and tracking to sequences nearly six times as long as those in the two referring benchmarks. These longer sequences test whether models can continue tracking the same referred objects as their activity and the scene change.

3 BASELINE MODEL

To address Ref-AVIS, we introduce RAVISeg (Figure 2(a)), a baseline that connects multimodal referring understanding with instance segmentation through persistent video-level representations. Competitive instance grounding constructs instance slots, an audio adapter converts them into frame-specific SAM2 prompts, and an instance-conditioned discriminator assesses whether the resulting predictions remain relevant to the expression under changing audio-visual conditions.

3.1 COMPETITIVE INSTANCE GROUNDING

The MLLM receives encoded video and audio embeddings v, a together with a referring expression r. Inspired by LISA (Lai et al., 2024), we fine-tune the MLLM with LoRA (Hu et al., 2021) to generate a variable number K of `[SEG]` tokens following predefined templates. Each token's hidden state h_k represents an instance and initializes a slot as $s_k^{(0)} = P(h_k)$, where P is a two-layer MLP.

To reduce ambiguity among these representations, Competitive Aggregation refines the slots using spatial features $\{v_n\}_{n=1}^{N}$ extracted from T frames by the SAM2 image encoder and performs iterative slot attention (Locatello et al., 2020). Here, $N = TH_fW_f$, with $H_f \times W_f$ denoting the feature resolution. At each refinement round, learned projections of layer-normalized slots and visual features produce queries q_k, keys κ_n, and values $\hat{v}_n$ in a shared d-dimensional space. The keys and values are shared across slots, whereas each query is computed from its corresponding slot state. Each round therefore assigns a common pool of visual evidence according to the current instance representations. Given $b_{nk} = q_k^\top \kappa_n / \sqrt{d}$, the assignments and visual updates are

$$\alpha_{nk} = \frac{\exp(b_{nk})}{\sum_{j=1}^{K} \exp(b_{nj})}, \qquad u_k = \frac{\sum_{n=1}^{N} (\alpha_{nk} + \epsilon)\hat{v}_n}{\sum_{n=1}^{N} (\alpha_{nk} + \epsilon)}.$$

Normalization across slots induces competition for each visual feature, while normalization across features yields a weighted visual update for each slot. We use $\epsilon = 10^{-8}$ for numerical stability. A gated recurrent unit (GRU) (Chung et al., 2014) integrates this update with the previous slot state, and a residual two-layer MLP further refines the representation:

$$\tilde{s}_k = \mathrm{GRU}(u_k, s_k^{(\ell-1)}), \qquad s_k^{(\ell)} = \tilde{s}_k + \mathrm{MLP}(\mathrm{LN}(\tilde{s}_k)) \,.$$

Here, LN denotes layer normalization. Recurrence proceeds across L refinement rounds over the same cross-frame feature set, rather than chronological frames. The GRU state carries accumulated instance information while incorporating newly assigned visual evidence. The final representations $s_k = s_k^{(L)}$ serve as shared instance references for subsequent frame-level predictions.

3.2 AUDIO-GUIDED INSTANCE ADAPTATION

To incorporate the current acoustic context, we average-pool BEATs tokens from the audio segment aligned with frame t to obtain e_t. The projection P_a, comprising a linear layer followed by layer normalization and GELU, maps this vector into the slot space as $a_t = P_a(e_t)$. The adapter constructs a segmentation prompt for each instance:

$$p_{k,t} = s_k + f_{\mathrm{aud}}([s_k; a_t; s_k \odot a_t]),$$

where $[\cdot;\cdot]$ denotes concatenation, $\odot$ denotes element-wise multiplication, and f_{aud} is a two-layer MLP. The multiplicative term allows the shared frame-level audio to condition each prompt through its corresponding slot. The residual connection preserves the video-level instance representation while allowing the prompt to adapt to changes in acoustic activity.

For each frame, the SAM2 mask decoder takes $p_{k,t}$ as a sparse prompt together with the image features and produces low-resolution mask logits $\tilde{m}_{k,t}$. Bilinear upsampling yields logits $m_{k,t}$ at the original frame resolution. Predictions associated with the same slot are assembled in temporal order into a candidate instance trajectory. Audio conditioning thus modifies the prompt before mask decoding, whereas the subsequent discriminator evaluates the predicted region.

3.3 Instance-level Audio-visual Discrimination

We assess each predicted region using its appearance, the persistent instance slot, and concurrent audio. Let $v_t(u)$ denote the SAM2 image feature at location u on grid Ω. Bilinearly resizing $\tilde{m}_{k,t}$ to this grid and applying sigmoid σ yields weights $w_{k,t}(u)$. The regional appearance is

$$z_{k,t} = \frac{\sum_{u\in\Omega} w_{k,t}(u) v_t(u)}{\max\left(\sum_{u\in\Omega} w_{k,t}(u), \epsilon\right)},$$

where $\epsilon = 10^{-6}$ ensures numerical stability. We then compute frame-level relevance as

$$z'_{k,t} = P_v(z_{k,t}) + s_k, \qquad e'_t = P_e(e_t), \qquad \rho_{k,t} = \sigma\left(f_{\text{disc}}([z'_{k,t}; e'_t; z'_{k,t} \odot e'_t])\right).$$

The projections P_v and P_e are separate two-layer MLPs with layer normalization and GELU between their linear layers. They align both modalities to the slot space. The two-layer MLP f_{disc} uses GELU and produces a scalar logit, which is converted into $\rho_{k,t}$ by the sigmoid function. The slot supplies the referring semantics, while the multiplicative interaction combines instance-conditioned appearance with audio. This relevance estimate complements SAM2's mask-quality prediction by assessing whether a region satisfies the expression under the current audio-visual conditions. A visually plausible mask can therefore receive a low relevance score when its appearance and acoustic context no longer match the expression.

At inference, we retain trajectories with peak relevance $E_k = \max_t \rho_{k,t} \geq 0.5$ and suppress runs of at least three consecutive frames with relevance below 0.1. We discard trajectories without nonempty retained masks. This temporal rule tolerates brief score fluctuations while suppressing sustained low-relevance intervals. Surviving trajectories are ranked by combining peak relevance with relevance-weighted SAM2 mask quality. A classifier predicts categories from the slots and regional features averaged over all frames before filtering; class-wise mask NMS then removes duplicates. A lower-scoring trajectory is suppressed when its mean mask IoU with a retained trajectory of the same category exceeds 0.5, computed over frames where both masks are nonempty.

4 Experiments

To evaluate Ref-AVIS and RAVISeg, we design the experiments to address following questions (Figure 3):

- **RQ1**: How do different methods perform on Ref-AVIS?
- **RQ2**: What does each mechanism contribute to the results?
- **RQ3**: Can our model adapt to other audio-visual segmentation tasks?
- **RQ4**: Is our model robust to reference rephrasing?
- **RQ5**: How does performance vary across categories?

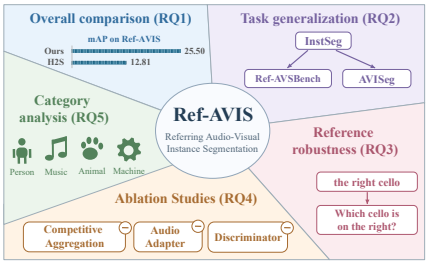


Figure 3: Overview of experiments.

4.1 Implementation Details

We employ Llama-2-7B-Chat (Touvron et al., 2023) as the base model, CLIP ViT-L/14 (Radford et al., 2021) and BEATs (Chen et al., 2022) as the visual and audio encoders, respectively, and SAM2.1 Hiera-L for mask prediction. We employ LoRA to adjust the attention and MLP projection layers within the language model, with rank $r = 8$, $\alpha = 16$, and a dropout rate of 0.05. We freeze the pretrained language model weights, CLIP, BEATs, and the SAM2 image and prompt encoders, while training the LoRA parameters, parameters associated with newly added tokens, multimodal projection modules, task-specific modules, and the SAM2 mask decoder. Each training sample contains 20 consecutive frames, with input resolutions of 224×224 for CLIP and 1024×1024 for SAM2. The instance representation dimension is set to 256, and competitive instance grounding performs 2 iterations. We train the model for 6,000 steps on four NVIDIA RTX A6000 GPUs using AdamW and BF16 mixed precision. The learning rate is warmed up to 1×10^{-4} and then decayed linearly. Per-GPU batch size is 1, with 8 gradient accumulation steps.

4.2 Overall Performance

RQ1: How do different methods perform on Ref-AVIS? To highlight the contributions of our dataset and method, we compare methods developed for the following tasks: (1) Ref-AVS, using

visual and acoustic cues to segment objects specified by referring expressions; (2) OmniAVS, supporting multi-object segmentation under diverse multimodal referring conditions; and (3) AVIS, segmenting and tracking individual sounding instances in videos. We evaluate all methods using mIoU for pixel-level region overlap, F-score for pixel-level precision and recall, mAP (Yang et al., 2019) for instance mask trajectories, HOTA (Luiten et al., 2021) for detection and temporal association, and FSLA (Guo et al., 2025) for the correctness of sounding-object counts, categories, and localization within each frame, providing a balanced assessment of referring segmentation, instance tracking, and sound localization performance. *We additionally evaluate methods for the Ref-VOS (Khoreva et al., 2018) and VIS (Yang et al., 2019) tasks; see the supplementary materials.*

Table 2: Performance comparison on Ref-AVIS. Red: best; blue: second best.

Task	Method	mIoU ↑	F-score ↑	mAP ↑	HOTA ↑	FSLA ↑
Ref-AVS	EEMC (Wang et al., 2024)	15.5	19.7	2.06	12.95	12.64
	Crab (Du et al., 2025)	23.7	31.7	1.70	20.55	18.18
	TGS-Agent (Zhou et al., 2026)	33.2	41.5	4.70	31.96	21.31
	AURORA (Luo et al., 2026)	33.8	44.0	5.54	30.30	23.88
OmniAVS	OISA-1B (Ying et al., 2025)	43.9	50.0	3.75	31.32	24.24
AVIS	AVISM (Guo et al., 2025)	37.2	42.5	12.01	32.27	21.05
	H2S (Liu et al., 2026)	45.7	52.2	12.81	34.60	15.80
	SAMOT (Peng et al., 2026)	48.2	54.8	24.70	38.13	20.64
Ref-AVIS	**Ours**	57.8	62.7	25.50	54.99	45.63

Table 3: Performance on Ref-AVSBench.

Method	**Seen** ↑			**Unseen** ↑			**Mix** ↑		
	J	F	$J\&F$	J	F	$J\&F$	J	F	$J\&F$
EEMC (Wang et al., 2024)	34.2	51.3	42.8	49.5	64.8	57.2	41.9	58.1	50.0
TSAM (Radman & Laaksonen, 2025)	43.4	56.8	50.1	54.6	66.4	60.5	49.0	61.6	55.3
SAM2-LOVE (Wang et al., 2025)	43.5	51.9	47.7	66.5	72.3	69.4	55.0	62.1	58.6
Crab (Du et al., 2025)	40.5	58.0	49.3	45.6	63.0	54.3	43.1	60.5	51.8
OISA-1B (Ying et al., 2025)	51.7	58.7	55.2	58.3	65.1	61.7	55.0	61.9	58.5
TGS-Agent (Zhou et al., 2026)	49.5	60.4	55.0	73.2	80.6	76.9	61.4	70.5	65.9
AURORA (Luo et al., 2026)	63.2	72.8	68.0	69.7	76.4	73.1	66.5	74.6	70.5
Ours	64.6	74.7	69.7	71.2	79.1	75.1	67.9	76.9	72.4

Table 4: Performance on AVISeg.

Method	FSLA	HOTA	mAP	FSLA_n	FSLA_s	FSLA_m
AVISM (Guo et al., 2025)	42.78	61.73	40.57	32.22	29.83	52.40
ACVIS (Seo et al., 2026)	42.87	62.09	42.14	21.16	30.62	52.34
H2S (Liu et al., 2026)	45.96	63.32	43.21	30.83	34.97	54.93
SAMOT (Peng et al., 2026)	48.18	67.27	47.62	36.73	32.87	60.68
Ours	59.35	72.42	49.74	89.47	51.64	60.30

As shown in Table 2, our method achieves the best results across all five metrics. Specifically, it surpasses the strongest competing result in this table for each metric by 9.6, 7.9, 0.80, 16.86, and 21.39 points in mIoU, F-score, mAP, HOTA, and FSLA, respectively. These consistent improvements demonstrate the effectiveness of our method in localizing referred instances, predicting their masks, and tracking instances across frames.

Ref-AVS is originally formulated for single-target referring segmentation, while OmniAVS supports multiple targets but without requiring separate instance identities and categories. Their performance on Ref-AVIS reveals a gap between region-level segmentation and instance-level prediction. For example, OISA-1B achieves 43.9 mIoU and 50.0 F-score. However, its mAP is only 3.75. TGS-Agent and AURORA also obtain relatively low mAP scores of 4.70 and 5.54, respectively. These results show that reasonable foreground overlap does not necessarily translate into accurate instance mask trajectories, highlighting the additional requirements exposed by Ref-AVIS.

AVIS methods are originally designed without language-based target selection. To adapt them to Ref-AVIS, we inject referring expressions through a frozen CLIP text encoder. Among AVIS methods, SAMOT achieves the highest baseline mIoU, F-score, mAP, and HOTA of 48.2, 54.8, 24.70,

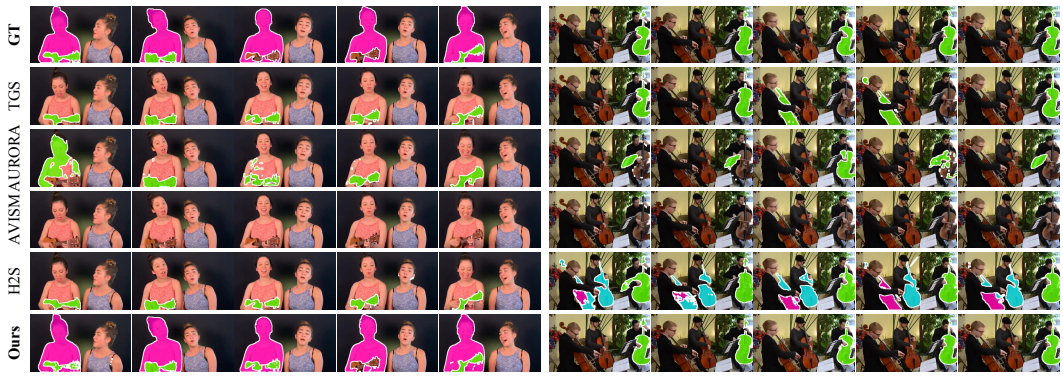


Figure 4: **Qualitative comparisons on Ref-AVIS.** Left reference: "the person holding the ukulele and the instrument itself, producing sound". Right reference: "The cello being played on the right".

and 38.13. These results demonstrate its competitive performance after adaptation, while all three AVIS baselines remain below our method across all five metrics. The results suggest that AVIS methods cannot fully bridge the performance gap on Ref-AVIS through simple text injection alone.

To complement the quantitative results, we present qualitative comparisons between Ours and methods of Ref-AVS and AVIS in challenging Ref-AVIS scenarios, as shown in Figure 4. The displayed frames are selected samples to highlight the challenging multi-instance scenes with 1-1 and 1-n referring expressions.

In the left example of Figure 4, EEMC incorrectly segments both people while missing the ukulele. TGS-Agent and AURORA segment only the person or the ukulele, failing to recover both targets specified by the expression. This inconsistency between the description and the predicted masks is consistent with the single-target bias inherited from Ref-AVS pretraining. Due to limitations in text comprehension, the displayed AVIS methods are also unable to identify the correct target. Our method correctly selects the person on the left and the ukulele, while suppressing the ukulele mask during its non-sounding intervals.

In the right example, TGS-Agent initially localizes the correct cello on the right but subsequently drifts to the cello on the left, failing to maintain consistent target tracking. AURORA switches between the music stand and the right cello, with severely fragmented masks indicating poor segmentation quality. The displayed AVIS methods also fail to follow the referring expression: H2S segments all instruments, whereas AVISM produces no foreground masks in the displayed frames. Our method consistently segments the referred cello and maintains its identity in the frames.

4.3 Ablation Studies

RQ2: What does each mechanism contribute to the results? To evaluate the contribution of the three core modules in our model, we individually remove the competitive aggregation, the audio adapter, and the audio-visual discriminator from the full model. As shown in Table 6, removing any module degrades all metrics, verifying that each component contributes to the performance of Ref-AVIS. In particular, removing competitive aggregation leads to the largest drops in mAP and HOTA, by 2.85 and 2.68 points, respectively, supporting its effectiveness in distinguishing instances and maintaining their identities across frames. Removing the audio adapter reduces mIoU and F-score by 2.3 and 2.2 points, indicating that frame-wise acoustic conditioning benefits sounding instance localization, as shown in Figure 5. Removing the audio-visual discriminator causes the largest drop in FSLA, by 4.16 points, highlighting its role in suppressing predictions that mismatch the referring expression under current audio-visual conditions. With all three modules enabled, the model achieves the best performance, supporting their respective contributions to instance grounding, audio-guided segmentation, and instance-level discrimination.

4.4 Generalization and Robustness Performance

RQ3: Can our model adapt to other audio-visual segmentation tasks? To evaluate the task generalization capability of our method, we conduct evaluations after fine-tuning the model on Ref-AVSBench and AVISeg. During inference, we disable the discriminator on Ref-AVSBench and fix the referring expression to "all sounding objects" on AVISeg. As shown in Tables 3 and 4, our

Table 5: Robustness to reference rephrasing on a 600-reference subset.

Reference	mIoU	F-score	mAP	HOTA	FSLA
Original	56.6	61.9	25.75	54.71	44.80
Rephrased	56.7	61.9	25.82	54.42	44.56

Table 6: Ablation studies. C.A.: competitive aggregation; Discr.: discriminator.

Setting	mIoU	F-score	mAP	HOTA	FSLA
w/o C.A.	57.0	62.0	22.65	52.31	42.32
w/o Adapter	55.5	60.5	24.95	54.11	44.25
w/o Discr.	54.7	59.6	22.67	53.32	41.47
Ours	57.8	62.7	25.50	54.99	45.63

Table 7: Performance across four broad category subsets.

Scenario	mIoU	F-score	mAP	HOTA	FSLA
Person	45.8	48.7	12.07	36.46	33.30
Music	61.2	67.2	28.43	51.50	47.38
Animal	52.4	54.1	8.08	26.43	43.45
Machine	62.2	64.9	26.81	59.52	57.12

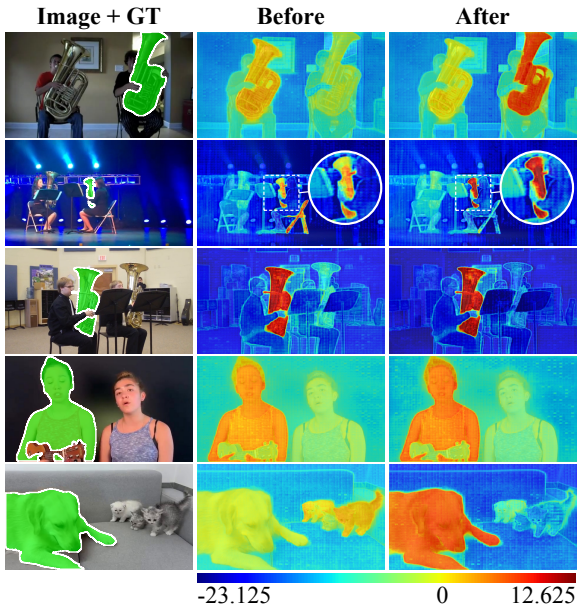


Figure 5: Heatmap of instance representations extracted before and after the audio adapter on five selected examples (GT in green).

method achieves competitive results, with $J\&F$ scores of 69.7, 75.1, and 72.4 on the Seen, Unseen, and Mix splits of Ref-AVSBench, respectively, and 49.74 mAP, 72.42 HOTA, and 59.35 FSLA on AVISeg. These results highlight its adaptability to different audio-visual segmentation tasks.

RQ4: Is our model robust to reference rephrasing? To evaluate robustness to reference wording, we randomly select 600 referring expressions and rephrase them as natural-language questions. We evaluate the model without additional training using both the original and rephrased references. As shown in Table 5, the rephrased references yield 56.7 mIoU, 61.9 F-score, 25.82 mAP, 54.42 HOTA, and 44.56 FSLA, with all metric changes remaining within 0.30 points of the original results. These results demonstrate robustness to reference rephrasing.

4.5 Further Analysis

RQ5: How does performance vary across different categories? To further investigate the challenges of Ref-AVIS, we evaluate our method across four broad categories. As shown in Table 7, Person and Animal obtain lower segmentation and tracking scores than Music and Machine. For example, their mAP scores are 12.07 and 8.08, compared with 28.43 and 26.81 for Music and Machine, respectively. This gap reflects the challenges introduced by non-rigid motion and pose changes, which complicate mask boundary prediction and temporal association. Interactions between visually similar instances may further hinder consistent identity tracking. These results suggest that segmenting and tracking dynamically deforming instances remains a challenging issue.

5 Conclusion

In this work, we introduced Ref-AVIS, a benchmark for language-guided instance segmentation and tracking in long audio-visual videos. It combines diverse referring expressions with independent masks and persistent identities to support the study of target selection and tracking in multi-instance scenes. We also developed a baseline that integrates referring understanding with segmentation under changing audio-visual conditions. Experiments demonstrate improved segmentation and tracking on Ref-AVIS and competitive performance on related tasks after fine-tuning. The results further highlight that accurate foreground segmentation alone does not imply reliable instance tracking, motivating continued research on language-guided audio-visual understanding.

AI USE STATEMENT

In this work, we used generative AI tools to assist with translation. We have not used generative AI tools to generate synthetic data sets, help develop theoretical models or conceptual frameworks, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, clean and reformat the dataset, support qualitative and thematic data analysis, and interpret results, and the rest of the required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for draft parts of a research paper and edit a research paper to improve readability. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This study adheres to the ICLR Code of Ethics. This study did not involve any human subjects or animal experiments. All datasets used (including Ref-AVSBench and AVISeg) were obtained in accordance with the relevant usage guidelines to ensure that privacy was not compromised. We have taken measures throughout the research process to avoid any biased or discriminatory results. No personally identifiable information was used in this study, nor were any experiments conducted that could raise privacy or security concerns. We are committed to maintaining transparency and integrity throughout the entire research process.

REPRODUCIBILITY STATEMENT

We have made every effort to ensure that the results presented in this paper are reproducible. Once the paper is accepted, all code and datasets will be made publicly available. Regarding the datasets, we describe the data collection and annotation processes in detail in Section 2.1; for the InstSeg model, the training steps, model configuration, and hardware information are detailed in Section 4.1. Additionally, the public datasets used in this paper (such as Ref-AVSBench and AVISeg) are publicly available, thereby ensuring the consistency and reproducibility of the evaluation results.

REFERENCES

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *International conference on learning representations*, volume 2026, pp. 138846–138923, 2026.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, and Furu Wei. Beats: Audio pre-training with acoustic tokenizers. *arXiv preprint arXiv:2212.09058*, 2022.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

Henghui Du, Guangyao Li, Chang Zhou, Chunjie Zhang, Alan Zhao, and Di Hu. Crab: A unified audio-visual scene understanding model with explicit cooperation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18804–18814. IEEE, 2025.

Shengyi Gao, Zhe Chen, Guo Chen, Wenhai Wang, and Tong Lu. Avsegformer: Audio-visual segmentation with transformer, 2023.

Ruohao Guo, Xianghua Ying, Yaru Chen, Dantong Niu, Guangyao Li, Liao Qu, Yanyu Qi, Jinxing Zhou, Bowei Xing, Wenzhen Yue, et al. Audio-visual instance segmentation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13550–13560. IEEE, 2025.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Anna Khoreva, Anna Rohrbach, and Bernt Schiele. Video object segmentation with language referring expressions. In *Asian conference on computer vision*, pp. 123–141. Springer, 2018.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9579–9589. IEEE, 2024.

Kexin Li, Zongxin Yang, Lei Chen, Yi Yang, and Jun Xiao. Catr: Combinatorial-dependence audio-queried transformer for audio-visual video segmentation. In *Proceedings of the 31st ACM international conference on multimedia*, pp. 1485–1494, 2023.

Tianming Liang, Kun-Yu Lin, Chaolei Tan, Jianguo Zhang, Wei-Shi Zheng, and Jian-Fang Hu. Referdino: Referring video object segmentation with visual grounding foundations. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20009–20019. IEEE, 2025.

Leiye Liu, Miao Zhang, Jiahong Jiang, Jingjing Li, Jialong Zhong, Kai Peng, Tingwei Liu, Wei Ji, Yongri Piao, and Huchuan Lu. Hear to see: Discerning stateful listening for audio-visual instance segmentation. *arXiv preprint arXiv:2608.03264*, 2026.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *Advances in neural information processing systems*, 33:11525–11538, 2020.

Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. Hota: A higher order metric for evaluating multi-object tracking. *International journal of computer vision*, 129(2):548–578, 2021.

Ziyang Luo, Nian Liu, Fahad Shahbaz Khan, and Junwei Han. Aurora: Augmented understanding via structured reasoning and reinforcement learning for reference audio-visual segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 7716–7724, 2026.

Kai Peng, Yunzhe Shen, Miao Zhang, Leiye Liu, Wei Ji, Jingjing Li, Yongri Piao, and Huchuan Lu. Samot: State-aware step modulation and optimal transport matching for audio-visual instance segmentation. *arXiv preprint arXiv:2608.08147*, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Abduljalil Radman and Jorma Laaksonen. Tsam: Temporal sam augmented with multimodal prompts for referring audio-visual segmentation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23947–23956. IEEE, 2025.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. URL https://arxiv.org/abs/2408.00714.

Jinbae Seo, Hyeongjun Kwon, Kwonyoung Kim, Jiyoung Lee, and Kwanghoon Sohn. Learning what to hear: Boosting sound-source association for robust audiovisual instance segmentation. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 16032–16036. IEEE, 2026.

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

Yaoting Wang, Peiwen Sun, Dongzhan Zhou, Guangyao Li, Honggang Zhang, and Di Hu. Ref-avs: Refer and segment objects in audio-visual scenes. In *European conference on computer vision*, pp. 196–213. Springer, 2024.

Yuji Wang, Haoran Xu, Yong Liu, Jiaze Li, and Yansong Tang. Sam2-love: Segment anything model 2 in language-aided audio-visual scenes. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28932–28932. IEEE, 2025.

Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4964–4974. IEEE, 2022.

Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5187–5196. IEEE, 2019.

Qi Yang, Xing Nie, Tong Li, Pengfei Gao, Ying Guo, Cheng Zhen, Pengfei Yan, and Shiming Xiang. Cooperation does matter: Exploring multi-order bilateral relations for audio-visual segmentation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27124–27133. IEEE, 2024.

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

Tao Zhang, Xingye Tian, Yu Wu, Shunping Ji, Xuebo Wang, Yuan Zhang, and Pengfei Wan. Dvis: Decoupled video instance segmentation framework. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1282–1291. IEEE, 2023.

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

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

Yingjian Zhu, Ying Wang, Yuyang Hong, Ruohao Guo, Kun Ding, Xin Gu, Bin Fan, and Shiming Xiang. Seavis: Sound-enhanced association for online audio-visual instance segmentation. *arXiv preprint arXiv:2603.01431*, 2026.

A Related Work

A.1 Audio-Visual Segmentation and Tracking

AVS (Zhou et al., 2022) aims to locate and segment sounding objects in a video. The pioneering work AVSBench establishes the foundational benchmark with audio-visual object segmentation for single sound sources (AVS-SS) and for multiple sound sources (AVS-MS), which focus on producing binary masks corresponding to active audio emitters. Building upon this, audio-visual semantic segmentation (AVSS) further enriches the task by introducing semantic labeling, requiring the model to predict category-specific segmentation maps aligned with sound cues. Extending AVS to the instance level, audio-visual instance segmentation (AVIS) requires separate masks for individual sounding objects and consistent identities throughout the video. AVISeg provides a benchmark for audio-visual instance segmentation (AVIS), which integrates detection, instance-level mask generation, and temporal association across frames.

For AVS, existing methods mainly improve sound-guided mask prediction through cross-modal interaction, query-based decoding, and temporal modeling. TPAVI (Zhou et al., 2022) injects audio semantics into visual features, while AVSegFormer (Gao et al., 2023) and CATR (Li et al., 2023) use audio-conditioned queries to guide segmentation; COMBO (Yang et al., 2024) further models bidirectional audio-visual relations and inter-frame consistency. In contrast, AVIS methods additionally associate object predictions over time to preserve instance identities. AVISM (Guo et al., 2025) constructs video-level tracks from frame-level object queries, ACVIS (Seo et al., 2026) improves sound-source discrimination before association, and SeaVIS (Zhu et al., 2026) performs online tracking by matching current instances with historical tracklets using representations that encode both appearance and sounding activity.

A.2 Multimodal Referring Segmentation

While AVS and AVIS focus on segmenting sounding objects, referring audio-visual segmentation (Ref-AVS) identifies and segments target objects using visual and acoustic cues described in referring expressions. Ref-AVS Bench establishes the first benchmark for this task with 10-second audio-visual clips, textual referring expressions, and pixel-level target masks. Its test set covers seen, unseen, and null-reference settings, and each non-null expression uniquely identifies one target object. Building on the textual referring expressions used in Ref-AVS, OmniAVS broadens the referring inputs to multiple modalities, supporting eight expression types that combine text or speech with optional sound and image cues. It further introduces expressions that require understanding sound content and reasoning with world knowledge, while allowing a single expression to refer to zero, one, or multiple target objects.

Existing Ref-AVS methods can be broadly grouped into feature-fusion and prompt-based approaches and MLLM-based approaches. EEMC combines textual, auditory, and visual features with temporal memory, while TSAM and SAM2-LOVE convert multimodal cues into prompts or tokens for pretrained segmentation models. MLLM-based methods, including SimToken, AURORA, and TGS-Agent, connect expression understanding with grounding and segmentation through semantic tokens or explicit reasoning. For OmniAVS, OISA further handles heterogeneous referring inputs by jointly encoding text, speech, sound, image cues, and temporally interleaved audio-visual content, and uses a segmentation token with online query propagation for mask prediction.

Existing benchmarks therefore provide complementary forms of supervision. AVISeg offers instance-level masks and persistent identities in long videos but does not provide referring expressions for selecting particular objects, whereas Ref-AVS Bench and OmniAVS support expression-guided target selection without explicitly evaluating the identity consistency of each selected object over time. To bridge this gap, we introduce Ref-AVIS, a long-video benchmark designed to jointly evaluate expression-guided target selection, instance-level segmentation, and identity-preserving tracking. Each expression may refer to one or multiple objects, which are annotated as separate instances with persistent identities, while their masks are provided only in frames where their states or activities match the conditions specified by the expression.

B MLLM INPUT PROMPT AND OUTPUT FORMAT

We provide the input prompt and response template used by RAVISeg on Ref-AVIS. Each input consists of a system message and a user message containing the video, audio, and referring expression. The messages are serialized with the chat template of Llama-2-7B-Chat. The video and audio placeholders below indicate the positions at which the corresponding encoded multimodal features enter the language model; {`referring expression`} is replaced with the input expression.

System message.

```
You are a helpful assistant.
```

User message.

```
This is a video:  <video_start><video><video_end>
This is an audio:  <audio_start><audio><audio_end>
Task rule:  segment the referred instances on their
annotated active frames; this dataset annotates sounding
targets.  Report the number of referred instances and one
[SEG] per instance, without category names.  If there are
none, report zero.
Can you segment {referring expression} in this video?
```

Assistant response. The response reports an instance count K and contains one `[SEG]` token for each candidate instance. For $K = 0$, no segmentation token is emitted. The following examples reproduce the fixed response template for zero, one, and two instances:

```
There are 0 referred objects.
There are 1 referred object: [SEG].
There are 2 referred objects: [SEG] [SEG].
```

The same template is used for larger instance counts. Each `[SEG]` hidden state initializes a separate instance slot, which is subsequently refined and used for mask prediction. Category names, instance IDs, and mask coordinates are not included in the textual response. Categories are predicted by the separate classification head, and temporal identity is maintained through the persistent instance slots. The generated count specifies the number of candidates before the discriminator and subsequent filtering; the final number of retained trajectories may therefore be smaller.

Table 8: Performance on 1–1 references. Red: best; blue: second best.

Task	Method	mIoU ↑	F-score ↑	mAP ↑	HOTA ↑	FSLA ↑
Ref-VOS	ReferFormer (Wu et al., 2022)	23.9	30.7	5.46	28.17	24.00
	ReferDINO (Liang et al., 2025)	52.0	56.7	22.98	52.09	52.50
VIS	DVIS (Zhang et al., 2023)	37.1	40.9	9.94	32.42	30.48
Ref-AVS	EEMC	18.3	22.3	3.23	16.27	17.86
	TGS-Agent	35.7	38.8	7.61	36.45	36.05
	AURORA	43.9	50.7	13.33	42.75	44.28
OmniAVS	OISA-1B	41.9	46.7	9.06	42.10	41.92
AVIS	AVISM	33.5	34.8	9.65	26.40	26.90
	H2S	41.2	45.4	12.15	33.13	20.01
Ref-AVIS	**Ours**	56.4	59.8	29.67	56.38	56.79

Table 9: Performance on 1–n references. Red: best; blue: second best.

Task	Method	mIoU ↑	F-score ↑	mAP ↑	HOTA ↑	FSLA ↑
Ref-VOS	ReferFormer	22.1	32.0	3.04	19.99	9.29
	ReferDINO	34.9	45.5	6.78	30.88	10.72
VIS	DVIS	34.4	41.4	11.28	30.95	23.02
Ref-AVS	EEMC	17.1	21.0	1.38	8.76	12.00
	TGS-Agent	29.1	39.7	3.43	26.03	8.87
	AURORA	26.3	38.4	3.56	21.12	7.19
OmniAVS	OISA-1B	41.0	47.9	1.70	21.42	11.23
AVIS	AVISM	38.5	45.5	10.22	27.43	21.28
	H2S	45.2	52.4	12.96	33.30	14.92
Ref-AVIS	**Ours**	50.5	56.8	20.82	47.73	32.38

Table 10: Performance on 1–all references. Red: best; blue: second best.

Task	Method	mIoU ↑	F-score ↑	mAP ↑	HOTA ↑	FSLA ↑
Ref-VOS	ReferFormer	18.9	30.4	3.51	20.97	4.92
	ReferDINO	41.6	55.5	6.64	32.22	7.29
VIS	DVIS	32.5	42.5	10.40	33.16	16.95
Ref-AVS	EEMC	9.1	13.9	1.46	10.61	4.57
	TGS-Agent	32.7	47.6	4.84	28.61	8.25
	AURORA	23.7	37.9	1.92	20.31	5.34
OmniAVS	OISA-1B	49.8	57.4	1.47	23.16	6.82
AVIS	AVISM	42.3	52.5	15.35	36.11	11.13
	H2S	53.8	63.2	14.44	33.88	9.64
Ref-AVIS	**Ours**	66.7	73.1	25.14	54.04	39.29

Table 11: Performance on 1–null references. Red: best; blue: second best.

Task	Method	Empty acc. ↑	FG frames ↓	FG pixels ↓
Ref-VOS	ReferFormer	0.00	95.31	5.27
	ReferDINO	3.70	83.32	4.03
VIS	DVIS	80.05	20.03	0.96
Ref-AVS	EEMC	26.68	48.49	3.64
	TGS-Agent	1.72	95.35	7.45
	AURORA	0.00	99.54	4.91
OmniAVS	OISA-1B	34.87	63.61	4.06
AVIS	AVISM	81.11	18.50	0.78
	H2S	77.54	21.89	0.87
Ref-AVIS	**Ours**	70.67	27.10	1.54

C Performance across Referring Types

To examine how target-selection requirements affect performance, we report results separately for the four referring types in Ref-AVIS. A 1–1 expression selects an individual instance; a 1–n expression selects multiple instances, including subsets of the available candidates; a 1–all expression selects all instances within its specified scope; and a 1–null expression has no matching target. The comparison includes Ref-VOS and VIS methods in addition to the Ref-AVS, OmniAVS, and AVIS methods.

Tables 8–10 report mIoU, F-score, mAP, HOTA, and FSLA for the three nonempty referring types. For 1–null, Table 11 reports the percentage of expressions with no predicted instances (Empty acc.), the percentage of evaluated frames containing predicted foreground (FG frames), and the percentage

of evaluated pixels predicted as foreground (FG pixels). All values are reported as percentages. Higher empty prediction accuracy and lower foreground rates indicate better rejection of unmatched references.

Single-target selection and multi-target tracking. Our method achieves the best results across all five metrics in each of the three nonempty referring types among the compared methods. ReferDINO is the strongest competitor on 1–1, achieving 22.98 mAP and 52.09 HOTA, compared with our 29.67 and 56.38. Its mAP decreases to 6.78 on 1–n and 6.64 on 1–all, while our method achieves 20.82 and 25.14, respectively. The HOTA advantage over the strongest competitor also increases from 4.29 points on 1–1 to 14.43 points on 1–n and 17.93 points on 1–all. These results highlight the additional demands of maintaining separate trajectories for multiple referred instances.

Region overlap and instance-level correctness. The multi-target results further distinguish foreground segmentation from instance-level prediction. On 1–all, OISA-1B obtains 49.8 mIoU and 57.4 F-score, but only 1.47 mAP and 6.82 FSLA. AVISM achieves a lower mIoU of 42.3 but a substantially higher mAP of 15.35. Likewise, H2S achieves 53.8 mIoU but 9.64 FSLA, whereas DVIS reaches 16.95 FSLA with 32.5 mIoU. This difference shows that accurately covering the foreground does not necessarily ensure correct instance separation, category prediction, or temporal association. Our method achieves 66.7 mIoU, 25.14 mAP, and 39.29 FSLA on this referring type, supporting both region-level and instance-level prediction.

Selective and exhaustive multi-target references. Our method obtains lower scores on 1–n than on 1–all, with 50.5 versus 66.7 mIoU and 47.73 versus 54.04 HOTA. This is consistent with the additional challenge of selecting a specified subset while excluding other candidate instances. However, the two groups contain different expressions and scenes, so these aggregate results do not isolate the effect of target selection alone. The gap identifies selective multi-target referring as an important setting for further analysis.

Rejection of unmatched references. On 1–null, our method achieves 70.67% empty prediction accuracy, exceeding the compared Ref-VOS, Ref-AVS, and OmniAVS methods but remaining below AVISM, DVIS, and H2S. AVISM achieves the highest accuracy of 81.11%, together with the lowest foreground frame and pixel rates of 18.50% and 0.78%, respectively. Our corresponding rates are 27.10% and 1.54%, indicating that false foreground predictions remain a limitation. These results complement the gains on nonempty references: selecting and tracking valid targets and rejecting unmatched expressions are distinct requirements, and improvements in one do not automatically ensure improvements in the other.

Table 12: Comparison with SAM3 tracking on Ref-AVIS.

Setting	mIoU ↑	F-score ↑	mAP ↑	HOTA ↑	FSLA ↑
SAM3 tracking	43.0	50.9	20.32	44.57	38.63
Ours	**57.8**	**62.7**	**25.50**	**54.99**	**45.63**

Table 13: Sensitivity to discriminator thresholds on Ref-AVIS. The default setting is highlighted.

τ_{track}	τ_{frame}	mIoU ↑	F-score ↑	mAP ↑	HOTA ↑	FSLA ↑
0.5	0.05	57.1	62.1	25.21	54.80	44.63
0.5	0.10	57.8	62.7	25.50	54.99	45.63
0.5	0.20	58.5	63.5	25.59	55.07	46.71
0.5	0.30	58.9	63.8	25.54	55.03	47.32
0.5	0.50	59.4	64.4	25.44	54.75	48.16
0.2	0.10	57.7	62.6	25.50	54.96	45.49
0.3	0.10	57.7	62.6	25.50	54.96	45.49
0.7	0.10	58.1	63.1	25.50	55.08	46.12
0.9	0.10	58.6	63.6	25.57	55.07	46.84

D Additional Experiments

Comparison with SAM3 tracking. Table 12 compares our method with SAM3 (Carion et al., 2026) tracking on Ref-AVIS. Our method achieves higher scores across all five metrics, improving mIoU, mAP, and HOTA by 14.8, 5.18, and 10.42 points, respectively. The gains cover both foreground overlap and instance-level prediction, showing the advantage of our task-specific audio-visual pipeline under the evaluated settings.

Sensitivity to discriminator thresholds. We vary the trajectory threshold τ_{track} from 0.2 to 0.9 and the frame threshold τ_{frame} from 0.05 to 0.5, changing one at a time while fixing the other at its default value. All configurations reuse the same cached model outputs without retraining. We retain the minimum run length of three consecutive low-relevance frames, with mask binarization and mask NMS thresholds fixed at 0.5. As shown in Table 13, mAP remains within 25.21–25.59 and HOTA within 54.75–55.08 across all nine configurations. FSLA varies more, from 44.63 to 48.16, reflecting the effect of filtering on frame-level sound localization. These results suggest that instance segmentation and temporal association are stable over the tested threshold ranges, while frame-level localization remains more sensitive to the filtering strength. We retain the default pair $(\tau_{\text{track}}, \tau_{\text{frame}}) = (0.5, 0.1)$ for the other experiments.

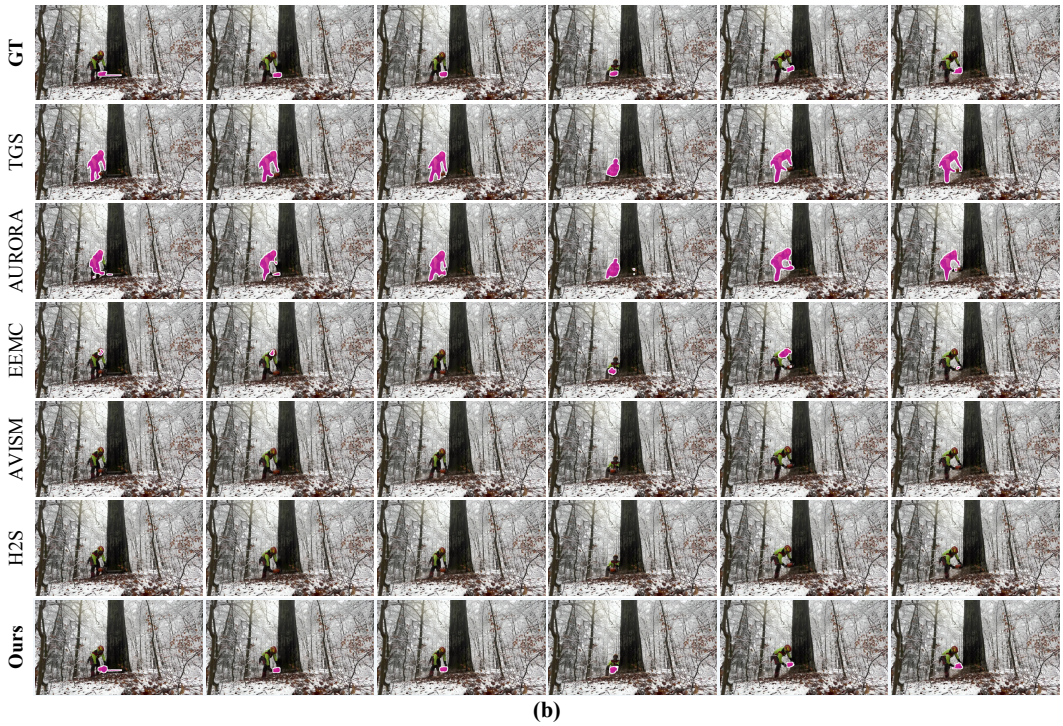
E Additional Qualitative Comparisons

Following the visualization style of Figure 4, Figures 6–8 present six additional videos covering single-target selection, multi-target selection, and exhaustive referring. Each video shows six uniformly sampled frames, including the first and last frames, with identical frame positions for all methods. Rows compare GT, TGS-Agent, AURORA, EEMC, AVISM, H2S, and Ours.

The single-target examples illustrate incorrect target selection and degraded mask quality, while the multi-target examples show the need to recover all specified instances without including surrounding objects. Our method better recovers the referred targets in these selected examples, but errors remain visible. In Figure 8(e), our method retains foreground in the last displayed frame despite the empty GT. In Figure 8(f), it misses one of the two referred dogs in the first displayed frame. These examples complement the aggregate results by showing both successful predictions and remaining errors under changing target activity.



(a)



(b)

Figure 6: **Additional single-target comparisons.** (a) "the sounding instrument being played on the right". (b) "The tree harvester making a sound by the tree". Both examples use 1–1 references.



(c)



(d)

Figure 7: **Additional multi-target comparisons.** (c) "The two instruments at the row's ends that are being played and producing sound". (d) "The two pipas producing sound on the right side". Both examples use 1–n references.



(e)



(f)

Figure 8: **Additional all-target comparisons.** (e) "By the brick wall, all the sounding persons and instruments in the performance are included." (f) "Every dog and man making sound near the fence and gate." Both examples use 1–all references.