Ref-AVS: Refer and Segment Objects in Audio-Visual Scenes

Yaoting Wang[1†], Peiwen Sun[2†], Dongzhan Zhou[3†], Guangyao Li[1], Honggang Zhang[2], and Di Hu[✉1,4]

[1] Gaoling School of Artificial Intelligence, Renmin University of China, China
yaoting.wang@outlook.com
{guangyaoli,dihu}@ruc.edu.cn
[2] Beijing University of Posts and Telecommunications, Beijing, China
{sunpeiwen,zhhg}@bupt.edu.cn
[3] Shanghai Artificial Intelligence Laboratory, Shanghai, China
zhoudongzhan@pjlab.org.cn
[4] Engineering Research Center of Next-Generation Search and Recommendation

Abstract. Traditional reference segmentation tasks have predominantly focused on silent visual scenes, neglecting the integral role of multimodal perception and interaction in human experiences. In this work, we introduce a novel task called Reference Audio-Visual Segmentation (Ref-AVS), which seeks to segment objects within the visual domain based on expressions containing multimodal cues. Such expressions are articulated in natural language forms but are enriched with multimodal cues, including audio and visual descriptions. To facilitate this research, we construct the first Ref-AVS benchmark, which provides pixel-level annotations for objects described in corresponding multimodal-cue expressions. To tackle the Ref-AVS task, we propose a new method that adequately utilizes multimodal cues to offer precise segmentation guidance. Finally, we conduct quantitative and qualitative experiments on three test subsets to compare our approach with existing methods from related tasks. The results demonstrate the effectiveness of our method, highlighting its capability to precisely segment objects using multimodal-cue expressions. Dataset is available at https://gewu-lab.github.io/Ref-AVS.

Keywords: Referring Audio-Visual Segmentation · Audio-Visual Segmentation · Multimodal Learning

1 Introduction

In the real world, visual scenes are often accompanied by diverse *multimodal information*[1], including audio and text modalities. This additional multimodal

† Equal contribution.
✉ Corresponding author.
[1] To clarify, we refer to the sound and vision in the scene as multimodal information, while the attribute description of audio and vision (*e.g.*, louder and on the right) in the reference expression is named multimodal cues.

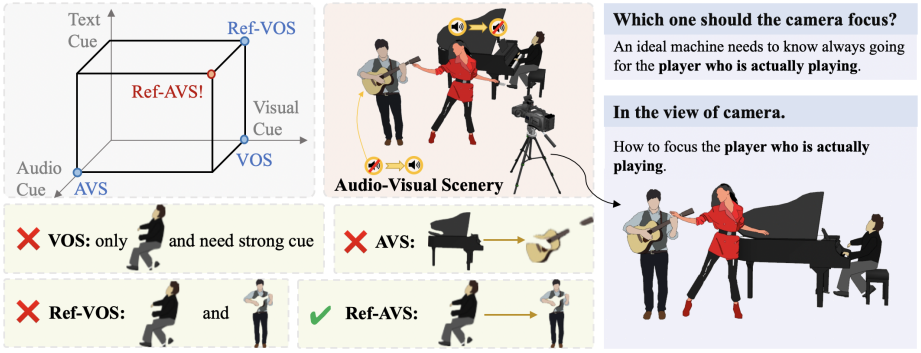

Fig. 1: Comparison of the Ref-AVS task with other related tasks. Ref-AVS challenges machines to locate objects of interest in the visual space using multimodal cues, just like humans do in the real world.

information offers valuable spatiotemporal and semantic cues, aiding individuals in locating objects of interest. Similarly, the ability to locate objects based on multimodal cues plays a crucial role in the functioning of machines. For example, stage cameras should be able to locate the musician playing instruments, while movie cameras should locate the superhero in front of the yelling villain. In the above scenarios, using a single modality for segmentation alone cannot achieve the desired behavior.

To date, researchers have studied the reference segmentation problem from somewhat limited scenarios. As shown in the coordinate system in Fig. 1, there are currently three mainstream developments in segmentation methods for different modalities. Firstly, from the perspective of visual cues, Video Object Segmentation (VOS) [9, 30–32, 40] emerges with annotated first frame mask as the reference, guiding the segmentation of specific objects in subsequent video frames. While VOS can achieve arbitrary object segmentation in videos, it heavily relies on precise annotation of the first frame, which can be challenging and time-consuming in practice. Secondly, from the perspective of text cues, Referring Video Object Segmentation (R-VOS) [8, 11, 13, 20] explores the segmentation ability by using attribute descriptions as guidance. As R-VOS has successfully replaced the mask annotation used in VOS with natural language, it provides a more accessible and user-friendly reference form but the capability to locate objects in more natural dynamic audio-visual scenes is still limited. In contrast, from the perspective of audio cues, Audio-Visual Segmentation (AVS) [26, 37, 44–47] leverages audio as a guide to segment sounding objects. This approach effectively addresses the challenge of locating objects in dynamic audio-visual scenes. However, AVS still faces limitations in its ability to focus on general objects that do not generate sound and cannot effectively locate objects of specific interest.

In simple terms, current work falls short of enabling machines to locate objects of interest in natural dynamic audio-visual scenes. For instance, as illus-

trated in Fig. 1, how can machines accurately locate a person who is genuinely playing a musical instrument over time? This requires machines to infer which instrument is making a sound and who is playing that instrument. Though Yan *et al*. [41] utilize either audio or language modality for multimodal reference segmentation, their method does not integrate multimodal information simultaneously nor support the aforementioned scenario. Therefore, introducing a task to explore the possibility of locating interested objects in natural dynamic audio-visual scenes holds significant practical potential.

In this paper, we propose a pixel-level segmentation task called ***Ref****erring* ***A****udio-* ***V****isual* ***S****egmentation* (Ref-AVS), which requires the network to densely predict whether each pixel corresponds to the given multimodal-cue expression, including dynamic audio-visual information. Top-left of Fig. 1 highlights the distinctions between Ref-AVS and previous tasks. The Ref-AVS task poses a greater challenge as it requires the network to accurately locate and segment objects in a more complex and stereoscopic modality space. To this end, a computation model with comprehensive multimodal understanding capabilities is necessary. To foster research advancements in this field, we introduce the *Ref-AVS Bench*, the pioneering benchmark that focuses on locating and segmenting objects of interest using referring multimodal-cue expressions. Considering the complexity of real-world audio-visual scenes, we collect about 4,000 audible video clips from YouTube of which more than 60% are multi-source sound scenarios. More than 20,000 reference expressions are collected and verified by experts to ensure accuracy and reliability, which adopt multimodal cues to describe objects in diverse and dynamic audio-visual scenes. Moreover, a special unseen test set is considered to evaluate the model's generalization ability in the growing demand for zero-shot scenarios.

Our contributions can be summarized as follows:

- We propose Ref-AVS as a challenging scene understanding task that segments objects of interest with multimodal-cue expressions and provide the corresponding Ref-AVS benchmark for performance training and validation.
- We design an end-to-end framework for Ref-AVS that efficiently processes the multimodal cues with a crossmodal transformer, serving as a feasible research framework for future development.
- We conduct extensive experiments to demonstrate the advantages of considering multimodal cues for visual segmentation, which also indicate the superiority of our approach in all subsets.

2 Related Work

2.1 Referring Video Object Segmentation (R-VOS)

The main objective of R-VOS is to perform object segmentation in streaming videos based on given natural language expressions. Initially proposed in 2018, A2D-Sentence [13] aims to segment actors in video content based on

descriptions of their actions. Subsequently, Refer-DAVIS17 [20] and JHMDB-Sentences [11] use extra language as references to replace previous VOS settings. To facilitate large-scale R-VOS training, Seo et al. [35] develop the Refer-YouTube-VOS dataset upon the YouTube-VOS-2019 dataset [40] and Henghui et al. [8] build the MeViS for complex motion expression. Overall, the primary R-VOS methods have evolved from semi-supervised VOS and incorporate attention mechanisms to enhance modalities interaction. For example, MAttNet [43] decomposes the expression information into language- and word-level attention, while URVOS [35] introduces crossmodal attention and memory attention. ReferFormer [39] and MTTR [1] utilize the crossmodal transformer to connect language expression with image regions. These datasets typically offer expressions that encompass the action or appearance attributes of the target object. However, these expressions primarily focus on visual information, which fails to meet the increasing multimodal demands in today's diverse audio-visual scenes.

2.2 Audio-Visual Segmentation (AVS)

The multimodal segmentation in audio-visual scenes, which is of great concern to us, has received some attention in AVS [12, 37, 46]. AVS aims to obtain finer pixel-level masks corresponding to sound-emitting objects in the visual space. In contrast, previous studies on audio-visual localization [17, 36] use heat maps for coarse localization in an unsupervised learning manner. The existing works on AVS can be broadly categorized into fusion-based methods [12, 18, 22, 24, 25, 46] and prompt-based methods [26, 28, 29, 37]. The former primarily focuses on localizing sounding objects by fusing audio and visual features, while the latter emphasizes constructing effective audio prompts for the visual foundation model. However, AVS simply segments all sound-emitting objects in the visual space, without the flexibility to combine the multimodal cues to segment specific objects of interest.

2.3 Language-aided Audio-Visual Scene Understanding

Currently, there is a limited number of public tasks offering datasets for audio-visual scene comprehension accompanied by language assistance. Two notable examples of such datasets are Music-AVQA [15] and AVQA [42], both encompassing audio, visual, and language information. These datasets contain rich audio-visual components and focus on annotations for textual questions and answers, emphasizing temporal (`Before`/`After`), spatial cues (`Left`/`Right`), *etc.*

Researchers [4, 19, 21] acknowledge the significance of the provided question, as it guides the feature extraction process for both audio and visual signals. Consequently, they have made efforts to identify pertinent segments by evaluating the semantic similarity between the question and temporal segments. Furthermore, [21] has been dedicated to locating crucial areas by leveraging semantic similarity between questions and visual tokens. Clearly, existing explorations have significantly advanced the research on language-aided audio-visual scene understanding. Nevertheless, the works that rely on the aforementioned datasets

are still unable to offer pixel-level scene understanding. This encourages us to begin exploring fine-grained segmentation of dynamic audio-visual scenes in the real world by constructing appropriate datasets.

3 Refer-AVS Dataset

3.1 Object Category

To ensure a diverse range of referred objects, we have carefully selected a wide variety of audible objects in 48 categories as well as a smaller selection of static, unsoundable objects in 3 categories. For the objects that can produce sound, we have chosen 20 categories of musical instruments, 8 of animals, 15 of machines, and 5 of humans. In the special case of humans, considering their diverse appearances, voices, and actions, we have employed the concept of morphology and classified them into 5 categories, based on age and gender.

3.2 Video Selection

During the process of video collection, we employed the techniques presented in [3, 46] to ensure the alignment of audio and visual snippets with the intended semantics. All videos were sourced from *YouTube* under the *Creative Commons license*, and each video was trimmed to a duration of 10 seconds. Throughout the manual collection process, we deliberately avoided videos falling into several categories (detailed in the appendix): 1) Videos with a large number of instances of the same semantics; 2) Videos characterized by extensive editing and camera switching; 3) Non-realistic videos containing synthetic artifacts.

To raise the alignment with real-world distributions, we carefully select videos that contribute to the diversification of scenes within our dataset. Specifically, we focus on selecting videos that involve interactions between multiple objects, such as musical instruments, people, vehicles, *etc*. The rich combination of categories indicates that our dataset is not limited to a narrow set of scenarios but rather encompasses a broad spectrum of real-life scenes where such objects are likely to naturally appear together. Refer to the supplementary materials for details.

In addition to diversity, we also carefully filter the videos to ensure that the dataset includes scenes with greater complexity and a larger number of objects. Specifically, 56% of the total videos contain two or more objects, while 13% of the total videos contain three or more objects.

3.3 Expressions

The diversity of expressions is one of the core factors in the assembly of the Ref-AVS dataset. The expressions consist of three dimensions of information, namely audio, vision, and time. The auditory dimension incorporates characteristics such as volume and rhythm while the visual dimension encompasses attributes like the appearance and spatial configuration of objects. We also leverage temporal

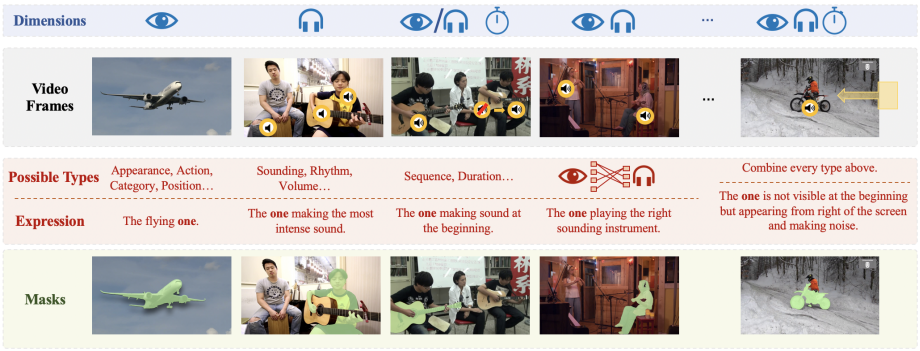

Fig. 2: The illustration of our Ref-AVS benchmark. Note,for , the sound volume increases successively from silent to loud. Our benchmark is meticulously designed to encompass multimodal expressions from multiple dimensions. By combining various types of modality expressions, we achieve a dataset that exhibits great diversity.

cues to generate references with sequential hints, *e.g.*, *"the one making sound first"* or *"the one appearing later"*. By integrating auditory, visual, and temporal information, we craft a rich array of expressions that not only accurately reflect multimodal scenarios but also meet users' specific needs for precise references. Fig. 2 illustrates examples of the combination of different modalities.

The accuracy of expressions is also a core concern. We follow three rules to generate high-quality expressions: 1) Uniqueness: The object referred to by an expression must be unique and an expression cannot refer to multiple objects simultaneously. 2) Necessity: Complex expressions can be used for reference, but each adjective in the sentence should narrow down the scope of target objects, avoiding unnecessary and redundant descriptions of the object. 3) Clarity: Certain expression templates involve subjective factors, such as *"the __ with a louder sound"*. The use of such expressions should only occur when the situation is clear enough to avoid ambiguous references.

Besides diversity and accuracy, we have graded the difficulty of expressions based on the number of cues involved, with the proportions of easy, medium, and hard samples in the overall dataset being 20%, 60%, and 20%, respectively. This gradation could be a great benefit to future research like curriculum learning. Please refer to the supplementary material for more details.

3.4 Segmentation Masks

We divide each 10-second video into ten equal 1-second snippets and the objective of annotation is to acquire the first frame's mask for each snippet. For these sampled frames, the ground truth labels are binary masks indicating the target object, according to expressions and multimodal information.

To obtain such a mask, initially, we manually select the pivotal frames for each 10-second video in which the target object is present. These pivotal frames

Table 1: Datasets comparison with related tasks.

Task	Dataset	Modality	Video	Frame	Object	Expression	Object/Video	Pub.
AVL	Flickr-SoundNet [34]	A+V	5,000	5,000	-	-	-	CVPR'2018
	VGG-SS [2]	A+V	5,158	5,158	-	-	-	CVPR'2021
AVS	AVS [46]	A+V	12,356	82,972	13,500∼	-	1.09∼	ECCV'2022
	VPO [44]	A+V	-	34,874	40,000∼	-	1.14 (Object/Frame)	CVPR'2024
R-VOS	J-HMDB Sentences [11]	T+V	928	928	928	928	1.28	CVPR'2018
	A2D Sentences [13]	T+V	3,782	11,936	4,825	6,656	1	CVPR'2018
	Refer-DAVIS$_{16}$ [20]	T+V	50	3,455	50	100	1	ACCV'2018
	Refer-DAVIS$_{17}$ [20]	T+V	90	13,543	205	1,544	2.27	ACCV'2018
	Refer-Youtube-VOS [35]	T+V	3,975	11,936	7,451	27,899	1.86	ECCV'2020
	MeViS [8]	T+V	2,006	443,000	8,171	28,570	4.28	ICCV'2023
Ref-AVS	**Ref-AVS Bench**	A+T+V	4,002	40,020	6,888	20,261	1.72	ECCV'2024

may occur at the beginning, middle, or end of the video, depending on when the target object is most clearly visible. Subsequently, we utilize Grounding SAM [33] to segment and label the pivotal frames, which will be manually checked and corrected subsequently. This process allows us to generate masks and labels for multiple target objects within the pivotal frames. Once the masks for the pivotal frames are established, we apply a tracking algorithm [6] to track the target object across the preceding and subsequent frames and obtain the ultimate mask and label for the target object in a span of 10 frames.

3.5 Dataset Statistics

In addition, we compare Refer-AVS with other popular audio-visual benchmarks in the Tab. 1. The Flickr-SoundNet [34] and VGG-SS [2] benchmarks are labeled at a patch level through bounding boxes and possess around 5,000 frame-level annotations. Compared with our dataset with pixel-level annotations, these benchmarks have a significantly lower quantity of annotations. Regarding the AVS dataset [44–46], the videos in our dataset contain a higher average number of objects about 1.72 objects per video, implying that our scenarios are more challenging with multiple sound sources and multiple semantics. Within such scenarios, our Ref-AVS benchmark is particularly valuable as it demonstrates the ability to effectively focus on objects of real interest. Our dataset also has more uniform video durations and a more refined video selection process than AVS. When it comes to datasets for R-VOS tasks, we maintain a consistent advantage in the number of videos compared to other datasets [8,11,13,20,35]. Moreover, we possess a considerable volume of data encompassing a large number of objects, expressions, and complex scenes.

Overall, our Ref-AVS dataset encompasses a substantial collection of 20,000 expressions and pixel-level annotations spread across 4,000 videos, totaling more than 11 hours. As a result, we believe that this benchmark adequately fulfills the requirements for facilitating research on the Ref-AVS task, while also presenting a significant level of challenge in this domain. We aim to continuously expand the dataset for broader community needs, like training larger foundation models.

Table 2: Dataset split of Ref-AVS Bench.

Subset	all	train	val	test	test (seen)	test (unseen)	test (null)
Category†	52	39	39	52	39	13	1
Video	4,002	2,908	276	818	292	269	257
Object	6,888	5,366	518	1,004	478	269	257
Expression	20,261	14,117	1,349	4,770	2,288	1,454	1,028

† The categories here contains "background".

3.6 Dataset Split

As shown in Tab. 2, the complete dataset is divided into three sets, *i.e.*, a training set of 2908 videos, a validation set of 276 videos, and a test set of 818 videos. The videos in the test set and their corresponding annotations undergo a meticulous review and re-annotation process conducted by experienced workers. In order to thoroughly evaluate the models' performance in the Ref-AVS task, the test set is further divided into three distinct subsets, each serving a specific purpose.

Seen: The first test subset, referred to as the "seen test set", comprises object categories that have already appeared in the training set. The purpose of establishing this subset is to evaluate the model's fundamental performance and assess how well it generalizes to familiar object categories.

Unseen: To address the growing demand for the generalization capabilities of various models in an open-world scenario, an additional test set was created specifically to assess the model's ability to generalize to unseen audio-visual scenes. This "unseen test set" consists of object categories that did not appear in the training set, although their super-categories (*e.g.*, animal, vehicle) may have been present in the training set. The purpose of this test set is to evaluate the model's capacity to generalize to novel object categories while leveraging its understanding of broader super-categories.

Null: A "null reference problem" arises when an expression refers to an object that either does not exist or is not visible in the given context. A model that accurately understands expression guidance should not segment any object in scenarios of a null reference [38]. Based on this consideration, we have developed a "null subset" to test the model's robustness against such challenges. This subset comprises object categories that were present during training but are paired with expressions that do not correlate, ensuring all objects in the video frames are irrelevant to the given reference expression. Therefore, the true masks for this subset are empty, and the model should refrain from segmenting any object.

4 Expression Enhancing with Multimodal Cues

4.1 Overall Architecture

Ref-AVS is designed to locate objects of interest in dynamic audio-visual scenes by utilizing multimodal cues. To solve this challenging task, we propose the

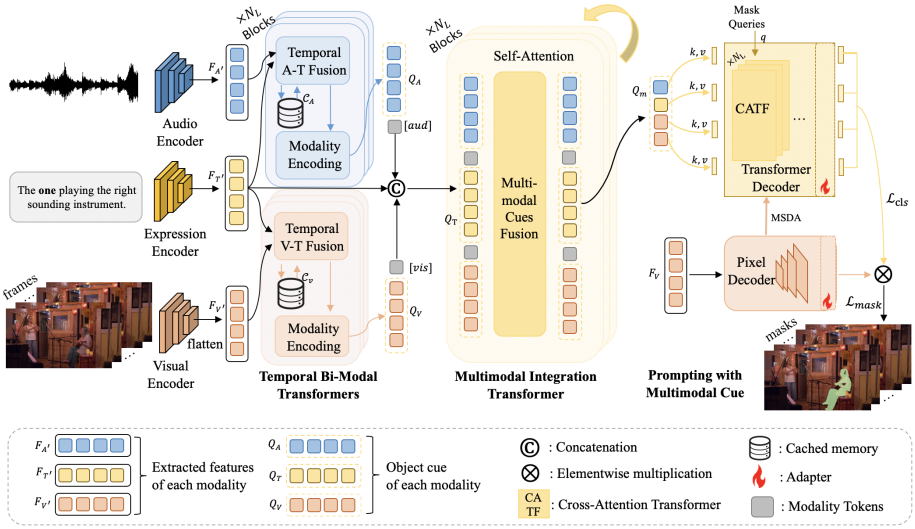

Fig. 3: The overall architecture of our method EEMC. We utilize cached memory to preserve temporal information, enabling the model to capture temporal mutations. Modality encoding provides a more clear context for the multimodal cue features Q_m, using the modality identify tokens $[aud]$ and $[vis]$. Lastly, we achieve efficient prompting of the visual foundation model by employing cross-attention, where mask queries q serve as the target and multimodal cue features Q_m act as the source.

Expression Enhancing with Multimodal Cues (EEMC) method, which focuses on integrating multimodal information in the dynamic audio-visual scenes into reference expressions with the corresponding multimodal cues. This approach enables the formation of a comprehensive multimodal reference feature. Additionally, we employ an attention mechanism to utilize the multimodal reference cues as a prompt for the visual foundation model, thereby facilitating the final segmentation process.

4.2 Multimodal Representations

Audio: Similar to the video processing, we divide the audio input into clips at 1-second intervals. Audio representations $F_A \in \mathbb{R}^{t \times d_A}$ is encoded from VGGish [14,16], where t represents the duration of the audio in seconds and matches the number of video frames. The audio representations are extracted offline and the audio encoder is not fine-tuned.
Visual: We sample t frames from the video input at 1-second intervals and extract visual representations $F_V \in \mathbb{R}^{t \times d_V \times H \times W}$ using a pre-trained Swin-base model [10]. The visual encoder is not fine-tuned.
Expression: We use RoBERTa [7, 27] as our text encoder to extract textual expression features $F_T \in \mathbb{R}^{L \times d_T}$, where L denotes the length of the expression. The expression representations are obtained off-the-shelf without fine-tuning.

4.3 Temporal Bi-modal Transformer

Temporal A-T & V-T Fusion: This module is introduced to retrieve the expression-related information of each modality. Firstly, for the convenience of the following multimodal fusion, we prepare the projected audio feature $F_{A'} \in \mathbb{R}^{t\times 1\times d_V}$, downsampled and flattened visual feature $F_{V'} \in \mathbb{R}^{t\times(\frac{W}{8}\times\frac{H}{8})\times d_V}$, and projected text feature $F_{T'} \in \mathbb{R}^{t\times L\times d_V}$ which is expanded along the temporal dimension. Subsequently, the features are combined to yield the expression-related multimodal features $\hat{F}_A$, $\hat{F}_V$ and modality-aware expression features $\hat{F}_{T_A}$, $\hat{F}_{T_V}$, with the same dimension as $F_{A'}$, $F_{V'}$, $F_{T'}$, respectively:

$$\hat{F}_M, \hat{F}_{T_M} = MF(Concat(F_{M'}; F_{T'})), M \in \{A, V\} \tag{1}$$

where $Concat(\cdot)$ denotes the concatenate operation, and $MF(\cdot)$ denotes the modality fusion function, which is employed as self-attention.

Cached Memory: Although the aforementioned fusion method can combine cross-modal information, it does not give much emphasis to strong temporal dynamics. To address this issue, we introduce a straightforward cached memory $\mathcal{C}_A \in \mathbb{R}^{t\times 2\times d_T}, \mathcal{C}_V \in \mathbb{R}^{t\times(W\times H+1)\times d_T}$ to explicitly capture the temporal mutations as they occur. The cache memory is required to store the mean modality features in the temporal domain from the beginning to the current moment. Then, the multimodal cues feature Q_A and Q_V output from temporal-aware transformers can be calculated as

$$Q_M^\tau = (\beta+1)\hat{F}_M^\tau - \beta\mathcal{C}_M^\tau, M \in \{A, V\}, \tag{2}$$

$$\mathcal{C}_M^\tau = \begin{cases} 0, & \text{if } \tau = 1, \\ \frac{1}{\tau-1}\sum_{i=0}^{\tau-1}\hat{F}_M^i, & \text{if } \tau > 1, \end{cases} \tag{3}$$

while τ is the time stamp of the specific feature at the temporal dimension and β is a hyper-parameter to control the sensitivity to temporal dynamics. When the current audio or visual features do not change significantly compared to the mean of past features (*i.e.*, $\hat{F}_M^\tau = \mathcal{C}_M^\tau$), Q_M^τ remains nearly unchanged. However, in cases where the current change is drastic (*i.e.*, $\hat{F}_M^\tau \neq \mathcal{C}_M^\tau$), the cached memory can amplify the difference in the current feature, leading to an output with salient features. It is important to note that text cue feature Q_T inherently possesses a highly abstracted semantic nature, which means manipulation may not be necessary at this stage. Therefore, at this stage, we combine and average the original textual features with the text features enhanced by multimodal information, resulting in a comprehensive enhanced text feature $Q_T = mean(F_{T'} + \hat{F}_{T_A} + \hat{F}_{T_V})$. Finally, we obtain the processed cue feature of each modality Q_A, Q_V and Q_T.

Modality Encoding: The modality encoding approach involves incorporating modality identify tokens $[aud] \in \mathbb{R}^{1\times d_V}$ for audio modality and $[vis] \in \mathbb{R}^{1\times d_V}$ for visual modality into the following multimodal cues integration process. Since only two tokens are required to split the sequence into three segments, we only need to manipulate the audio and visual cue features.

4.4 Prompting with Multimodal Cues

In this Prompting with Multimodal Cues (PMC) phrase, when we obtain the final multimodal cues features, we can concatenate the multimodal cues together and then apply self-attention to obtain comprehensive multimodal cues $Q_m \in \mathbb{R}^{t\times(H\times W+L+3)\times d_V}$ through the Multimodal Integration Transformer:

$$Q_m = MF(Concat([Q_A; [aud]; Q_V; [vis]; Q_T])), \tag{4}$$

where MF is the multimodal fusion function serving as the self-attention for multimodal cues interaction. Then we utilize these comprehensive multinational cues to prompt the learnable mask queries $q \in \mathbb{R}^{N\times d_V}$ in the transformer-based segmentation decoder with cross-attention, where N is the number of mask queries:

$$q_Q = CATF(query = q, key = Q_m, value = Q_m), \tag{5}$$

where $q_Q \in \mathbb{R}^{N\times d_V}$ is the updated mask queries. $CATF(\cdot)$ is the cross-attention transformer with q and Q_m as the information target and source, respectively.

5 Experiments

5.1 Implementation Details

Mask2Former [5] serves as our visual foundation model to provide the commonly used transformer-based segmentation decoder. By default, in this work, all input video frames are scaled to 384×384. The shape of the visual features is $[H = 64, W = 64, d_V = 256]$, and we employ 8-fold downsampling to reduce the computation costs. The audio features with $d_A = 128$ are extracted from the mono-channel waveform. The text features have a shape of $[L = 25, d_T = 768]$. For convenience, we map all modality feature dimensions to d_V. Hyper-parameter β is set to 1 by default. Transformer layers N_L of the Temporal Bi-modal Transformer, Multimodal Integration Transformer and the cross-attention transformer CATF are set to 4 by default. The number of mask queries N_q is fixed to 100.

5.2 Evaluation Metrics

To conduct a comprehensive evaluation of our Ref-AVS method, we employ the Jaccard index ($\mathcal{J}$) and F-score ($\mathcal{F}$) as performance metrics. Additionally, we introduce $\mathcal{S}$ in the null reference test set to assess the efficacy of expression guidance in the model. $\mathcal{S}$ denotes the square root of the ratio between the predicted mask area and the background area, a higher value indicates a larger proportion of the mask relative to the background area, suggesting a lack of precise expression guidance provided to the model.

Table 3: Performance on Ref-AVS Bench. We use three test subsets to evaluate the comprehensive ability of Ref-AVS methods. **Mix** is the average performance of **Seen** and **Unseen** test set. We also use **Null** test set to evaluate the robustness of multimodal-cue expression guidance.

		Seen		Unseen		Mix(S+U)		Null	Pub.
Method	Task	$\mathcal{J}(\%)$	$\mathcal{F}$	$\mathcal{J}(\%)$	$\mathcal{F}$	$\mathcal{J}(\%)$	$\mathcal{F}$	$\mathcal{S}$ (↓)	
AVSBench [46]	AVS	21.22	0.457	27.67	0.509	24.45	0.483	-	ECCV'2022
+*text*		23.20	0.511	32.36	0.547	27.78	0.529	0.208	
AVSegFormer [12]	AVS	29.16	0.423	32.71	0.451	30.94	0.437	-	AAAI'2024
+*text*		33.47	0.470	36.05	0.501	34.76	0.486	0.171	
GAVS [37]	AVS	24.85	0.455	27.86	0.460	26.36	0.457	-	AAAI'2024
+*text*		28.93	0.498	29.82	0.497	29.38	0.498	0.190	
ReferFormer [39]	Ref-VOS	27.46	0.477	24.94	0.504	26.2	0.491	-	CVPR'2022
+*audio*		31.31	0.501	30.40	0.488	30.86	0.495	0.176	
R2VOS [23]	Ref-VOS	21.34	0.309	23.91	0.431	22.63	0.370	-	ICCV'2023
+*audio*		25.01	0.410	27.93	0.498	26.47	0.454	0.183	
EEMC	Ref-AVS	**34.20**	**0.513**	**49.54**	**0.648**	**41.87**	**0.581**	**0.007**	ECCV'2024

5.3 Quantitative Results

We compare our Ref-AVS method with existing approaches in the relevant field using our Ref-AVS benchmark. The results from the seen test set demonstrate the superior performance of our Ref-AVS method on this dataset, outperforming other methods by a significant margin. To ensure fairness, we equip the other methods with the same inputs from both the audio and visual modalities as ours. However, even with this additional information, these methods still fail to achieve satisfactory results. The result indicates that simple modal fusion is inadequate for addressing the challenges of understanding multimodal cues within the Ref-AVS task. Instead of directly incorporating audiovisual information, our approach selects textual representations as carriers of multimodal information because they contain rich semantics and cues that are contextually relevant to the current audiovisual environment.

We conduct tests on the unseen test set and the null reference test set to explore the generalization and multimodal cues-following ability. Our method still maintains a leading position in the unseen test set because we leverage text with highly abstract semantic capabilities as a multimodal information carrier, instead of directly fusing audio and visual information. Therefore, the obtained multimodal cues can provide more robust semantic guidance. On the null test set, we obtain leading results among all other methods, indicating that our method can perceive multimodal cues to a reasonably accurate degree. Compared to other transformer-based and query-based frameworks, AVSBench, as a classic AVS method, obtains inferior results. This may be attributed to the model simply fuses multimodal cues and visual features in a progressive fashion, resulting in relatively weak guidance over the visual space.

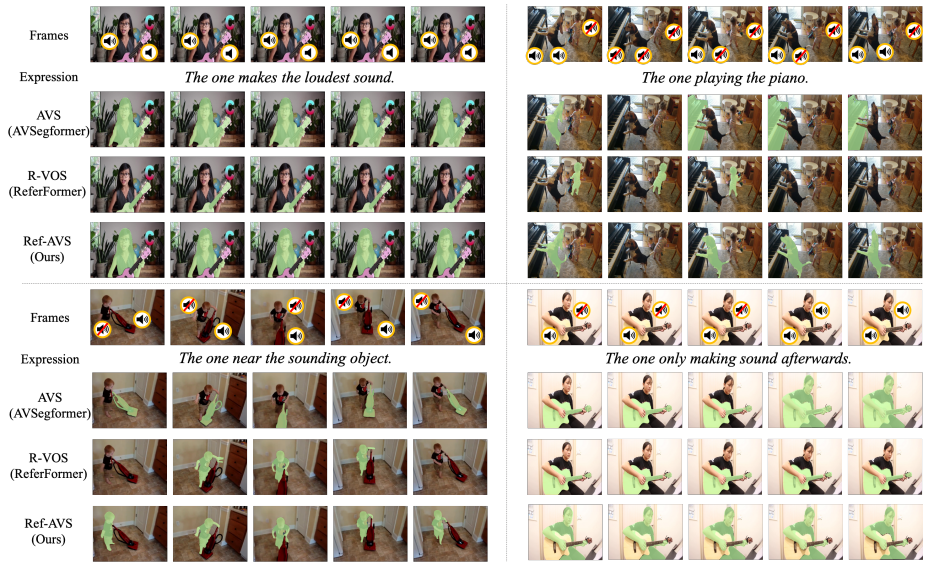

Fig. 4: Segmentation masks of the referred objects in the Ref-AVS Bench. Note, for 🔇🔉🔊, the sound volume increases successively from silent to loud. We compare our methodology with AVSegformer and ReferFormer. More cases can be found in the supplementary material.

5.4 Qualitative Results

We visualize the segmentation masks on the test set of the Ref-AVS Bench and compare them with AVSegFormer and ReferFormer, as shown in Fig. 4. From these qualitative results, we observe that both ReferFormer in the Ref-VOS task and AVSegFormer in the AVS task fail to accurately segment the object described in the expression. Specifically, AVSegformer struggles to fully comprehend the expression and tends to directly generate the sound source. This issue is exemplified in the bottom-left sample, where AVSegformer erroneously segments the vacuum cleaner instead of the boy. On the other hand, Ref-VOS may not adequately understand the audio-visual scene and thus mistakenly identify the toddler as the piano player, as shown in the top-right sample. In contrast, our Ref-AVS method demonstrates a superior capability to simultaneously process multi-modal expressions and scenes, enabling it to accurately interpret the user instruction and segment the intended object of interest.

5.5 Ablation Study

We conduct ablation studies to investigate the impact of two modalities (audio and text) information on the Ref-AVS task, as well as the effectiveness of the proposed method. As shown in the Tab. 4, we observe that in setting ②, removing the text information results in a significant performance degradation of 11.44% in

Table 4: Ablation study on modalities and method modules.

Setting	Seen		Unseen		Mix		Null
	$\mathcal{J}(\%)$	$\mathcal{F}$	$\mathcal{J}(\%)$	$\mathcal{F}$	$\mathcal{J}(\%)$	$\mathcal{F}$	$\mathcal{S}(\downarrow)$
① EEMC	**34.20**	**0.513**	**49.54**	0.648	**41.87**	**0.581**	**0.007**
② (-) Text	23.67	0.468	37.19	0.641	30.43	0.555	0.011
③ (-) Audio	27.27	0.492	43.64	0.656	35.46	0.574	0.010
④ (-) Cached Memory	25.30	0.494	36.73	**0.660**	31.05	0.577	0.009
⑤ (-) Modality Encoding	27.20	0.470	44.97	0.637	36.09	0.554	0.009

$\mathcal{J}$ and 2.60% in $\mathcal{F}$. This degradation is much higher compared to setting ③, where the removal of the audio information results in a combined mixed drop of 6.41% in $\mathcal{J}$ and 0.70% in $\mathcal{F}$. This phenomenon primarily results from the clarity and directness of textual information as a reference source. In contrast, when relying solely on audio information, the model often overlooks the referential content and instead focuses on identifying objects visually associated with audible behaviors. We also conduct experiments to examine the function of each module in this framework. The cached memory is designed to track significant changes in the temporal domain. Meanwhile, the purpose of modality encoding is to extract more distinct and thorough features from multimodal cues, enhancing modal perception. Results from settings ④ and ⑤ in Tab. 4 demonstrate the effectiveness of these modules.

6 Conclusion

In this work, we introduce a novel task, Ref-AVS, which requires the machine to segment objects of interest in dynamic audio-visual scenes based on expressions that incorporate multimodal cues and temporal information. To support this field, we develop the first benchmark, Ref-AVS Bench, for performance training and evaluation. Our method, EEMC, establishes a robust baseline with an advanced multimodal cues fusion module. We compare our method with several existing approaches on the Ref-AVS Bench dataset and demonstrate its promising performance in accurately locating objects using multimodal reference expressions. Essentially, Ref-AVS leverages audio, visual, and textual data to enrich real-world scene comprehension. Our task has potential in applications such as attention focusing and object editing in immersive audio-visual scenarios like extended reality, thereby enhancing user interaction experiences. In the future, we plan to further expand the scale of the current dataset to meet the growing data demands in the era of large language models.

Acknowledgements

This research was supported by National Natural Science Foundation of China (NO.62106272), and Public Computing Cloud, Renmin University of China. This research was partially supported by Shanghai Artificial Intelligence Laboratory.

References

1. Botach, A., Zheltonozhskii, E., Baskin, C.: End-to-end referring video object segmentation with multimodal transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4985–4995 (2022)
2. Chen, H., Xie, W., Afouras, T., Nagrani, A., Vedaldi, A., Zisserman, A.: Localizing visual sounds the hard way. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16867–16876 (2021)
3. Chen, H., Xie, W., Vedaldi, A., Zisserman, A.: Vggsound: A large-scale audio-visual dataset. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 721–725. IEEE (2020)
4. Chen, Z., Wang, L., Wang, P., Gao, P.: Question-aware global-local video understanding network for audio-visual question answering. IEEE Transactions on Circuits and Systems for Video Technology (2023)
5. Cheng, B., Choudhuri, A., Misra, I., Kirillov, A., Girdhar, R., Schwing, A.G.: Mask2former for video instance segmentation (2021)
6. Cheng, H.K., Oh, S.W., Price, B., Schwing, A., Lee, J.Y.: Tracking anything with decoupled video segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1316–1326 (2023)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
8. Ding, H., Liu, C., He, S., Jiang, X., Loy, C.C.: Mevis: A large-scale benchmark for video segmentation with motion expressions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2694–2703 (October 2023)
9. Ding, H., Liu, C., He, S., Jiang, X., Torr, P.H., Bai, S.: Mose: A new dataset for video object segmentation in complex scenes. arXiv preprint arXiv:2302.01872 (2023)
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
11. Dutta, U.K., Harandi, M., Sekhar, C.C.: Unsupervised deep metric learning via orthogonality based probabilistic loss. IEEE Transactions on Artificial Intelligence **1**(1), 74–84 (2020)
12. Gao, S., Chen, Z., Chen, G., Wang, W., Lu, T.: Avsegformer: Audio-visual segmentation with transformer (2023)
13. Gavrilyuk, K., Ghodrati, A., Li, Z., Snoek, C.G.: Actor and action video segmentation from a sentence. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5958–5966 (2018)
14. Gemmeke, J.F., Ellis, D.P., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M.: Audio set: An ontology and human-labeled dataset for audio events. In: 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 776–780. IEEE (2017)
15. Guangyao li, Yake Wei, Y.T.C.X.J.R.W.D.H.: Learning to answer questions in dynamic audio-visual scenarios. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
16. Hershey, S., Chaudhuri, S., Ellis, D.P., Gemmeke, J.F., Jansen, A., Moore, R.C., Plakal, M., Platt, D., Saurous, R.A., Seybold, B., et al.: Cnn architectures for

large-scale audio classification. In: 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 131–135. IEEE (2017)
17. Hu, D., Qian, R., Jiang, M., Tan, X., Wen, S., Ding, E., Lin, W., Dou, D.: Discriminative sounding objects localization via self-supervised audiovisual matching. Advances in Neural Information Processing Systems **33**, 10077–10087 (2020)
18. Huang, S., Li, H., Wang, Y., Zhu, H., Dai, J., Han, J., Rong, W., Liu, S.: Discovering sounding objects by audio queries for audio visual segmentation (2023)
19. Jiang, Y., Yin, J.: Target-aware spatio-temporal reasoning via answering questions in dynamics audio-visual scenarios. arXiv preprint arXiv:2305.12397 (2023)
20. Khoreva, A., Rohrbach, A., Schiele, B.: Video object segmentation with language referring expressions. In: Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part IV 14. pp. 123–141. Springer (2019)
21. Li, G., Hou, W., Hu, D.: Progressive spatio-temporal perception for audio-visual question answering. In: Proceedings of the 31st ACM International Conference on Multimedia. p. 7808–7816. MM '23, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3581783.3612293, https://doi.org/10.1145/3581783.3612293
22. Li, K., Yang, Z., Chen, L., Yang, Y., Xun, J.: Catr: Combinatorial-dependence audio-queried transformer for audio-visual video segmentation. arXiv preprint arXiv:2309.09709 (2023)
23. Li, X., Wang, J., Xu, X., Li, X., Raj, B., Lu, Y.: Robust referring video object segmentation with cyclic structural consensus. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22236–22245 (2023)
24. Ling, Y., Li, Y., Gan, Z., Zhang, J., Chi, M., Wang, Y.: Hear to segment: Unmixing the audio to guide the semantic segmentation (2023)
25. Liu, J., Ju, C., Ma, C., Wang, Y., Wang, Y., Zhang, Y.: Audio-aware query-enhanced transformer for audio-visual segmentation (2023)
26. Liu, J., Wang, Y., Ju, C., Zhang, Y., Xie, W.: Annotation-free audio-visual segmentation. arXiv preprint arXiv:2305.11019 (2023)
27. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
28. Ma, J., Sun, P., Wang, Y., Hu, D.: Stepping stones: A progressive training strategy for audio-visual semantic segmentation. IEEE European Conference on Computer Vision (ECCV) (2024)
29. Mo, S., Tian, Y.: Av-sam: Segment anything model meets audio-visual localization and segmentation. arXiv preprint arXiv:2305.01836 (2023)
30. Perazzi, F., Khoreva, A., Benenson, R., Schiele, B., Sorkine-Hornung, A.: Learning video object segmentation from static images. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2663–2672 (2017)
31. Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 724–732 (2016)
32. Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., Van Gool, L.: The 2017 davis challenge on video object segmentation. arXiv preprint arXiv:1704.00675 (2017)
33. Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., Zeng, Z., Zhang, H., Li, F., Yang, J., Li, H., Jiang, Q., Zhang, L.: Grounded sam: Assembling open-world models for diverse visual tasks (2024)

34. Senocak, A., Oh, T.H., Kim, J., Yang, M.H., So Kweon, I.: Learning to localize sound source in visual scenes. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
35. Seo, S., Lee, J.Y., Han, B.: Urvos: Unified referring video object segmentation network with a large-scale benchmark. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16. pp. 208–223. Springer (2020)
36. Tian, Y., Shi, J., Li, B., Duan, Z., Xu, C.: Audio-visual event localization in unconstrained videos. In: Proceedings of the European Conference on Computer Vision (ECCV) (September 2018)
37. Wang, Y., Liu, W., Li, G., Ding, J., Hu, D., Li, X.: Prompting segmentation with sound is generalizable audio-visual source localizer. arXiv preprint arXiv:2309.07929 (2023)
38. Wang, Y., Sun, P., Li, Y., Zhang, H., Hu, D.: Can textual semantics mitigate sounding object segmentation preference? IEEE European Conference on Computer Vision (ECCV) (2024)
39. Wu, J., Jiang, Y., Sun, P., Yuan, Z., Luo, P.: Language as queries for referring video object segmentation (2022)
40. Xu, N., Yang, L., Fan, Y., Yue, D., Liang, Y., Yang, J., Huang, T.: Youtube-vos: A large-scale video object segmentation benchmark. arXiv preprint arXiv:1809.03327 (2018)
41. Yan, S., Zhang, R., Guo, Z., Chen, W., Zhang, W., Li, H., Qiao, Y., He, Z., Gao, P.: Referred by multi-modality: A unified temporal transformer for video object segmentation (2023)
42. Yang, P., Wang, X., Duan, X., Chen, H., Hou, R., Jin, C., Zhu, W.: Avqa: A dataset for audio-visual question answering on videos. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 3480–3491 (2022)
43. Yu, L., Lin, Z., Shen, X., Yang, J., Lu, X., Bansal, M., Berg, T.L.: Mattnet: Modular attention network for referring expression comprehension. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1307–1315 (2018)
44. Yuanhong, C., Yuyuan, L., Hu, W., Fengbei, L., Chong, W., Gustavo, C.: A closer look at audio-visual semantic segmentation. arXiv preprint arXiv:2304.02970 (2023)
45. Zhou, J., Shen, X., Wang, J., Zhang, J., Sun, W., Zhang, J., Birchfield, S., Guo, D., Kong, L., Wang, M., Zhong, Y.: Audio-visual segmentation with semantics. arXiv preprint arXiv:2301.13190 (2023)
46. Zhou, J., Wang, J., Zhang, J., Sun, W., Zhang, J., Birchfield, S., Guo, D., Kong, L., Wang, M., Zhong, Y.: Audio-visual segmentation. In: European Conference on Computer Vision (2022)
47. Zhou, X., Zhou, D., Hu, D., Zhou, H., Ouyang, W.: Exploiting visual context semantics for sound source localization. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5199–5208 (2023)