This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

TSAM: Temporal SAM Augmented with Multimodal Prompts for Referring Audio-Visual Segmentation

Abduljalil Radman & Jorma Laaksonen
Department of Computer Science, Aalto University, Finland
{abduljalil.saif, jorma.laaksonen}@aalto.fi

Abstract

Referring audio-visual segmentation (Ref-AVS) aims to segment objects within audio-visual scenes using multimodal cues embedded in text expressions. While the Segment Anything Model (SAM) has revolutionized visual segmentation, its applicability to Ref-AVS, where multimodal cues act as novel prompts, remains unexplored. SAM's limitation to single-frame segmentation also hinders its ability to capture essential temporal context needed for multi-frame audio-visual segmentation. To address this gap, we propose TSAM, a novel extension of SAM designed to leverage multimodal cues for precise segmentation in dynamic audio-visual scenes. TSAM enhances SAM's image encoder with a temporal modeling branch, enabling spatio-temporal learning and deep multimodal fusion across video frames, while retaining SAM's pre-trained knowledge. Additionally, TSAM replaces SAM's user-interactive prompting mechanism with sparse and dense data-driven prompts, enabling more effective integration of audio-visual inputs and reference text expressions. Extensive experiments on the Ref-AVS dataset demonstrate TSAM's superiority over state-of-the-art methods. The results illustrate its effectiveness in segmenting objects in dynamic audio-visual scenes using text-based multimodal cues and its strong generalization to unseen objects. Project webpage: https://abdurad.github.io/TSAM/.

1. Introduction

In real-world scenarios, people often rely on multimodal information, such as audio-visual and visual-text, to gather spatiotemporal and semantic insights for distinguishing objects of interest [10, 39]. This multimodal approach is equally valuable for machines, enabling applications like stage cameras to track a performing musician or movie cameras to focus on a superhero confronting a yelling villain. Relying on a single sensory input, such as vision, is often insufficient for accurate object segmentation in complex, dynamic audio-visual scenes [16, 36]. In the literature, researchers have addressed this need for multimodal integration in object segmentation through three main approaches, each focused on a specific modality cue.

First, video object segmentation (VOS) [6, 30] relies on a mask annotation in the initial frame to guide segmentation of objects across subsequent frames. While VOS allows for flexible segmentation, it is heavily dependent on labor-intensive frame-level annotations, making it challenging to scale. **Second**, referring video object segmentation (Ref-VOS) [5, 9, 13] replaces the need for mask annotations with text expressions to identify targeted objects. This approach introduces a more intuitive, text-based method for segmentation, but struggles with tracking and segmenting objects in more dynamic audio-visual environments. **Third**, audio-visual segmentation (AVS) [21, 24, 33, 35, 39] leverages audio cues to segment sounding objects in video, effectively handling complex audio-visual scenarios, but remaining limited in scope as it does not generalize to inaudible objects. Each of these approaches, VOS, Ref-VOS, and AVS, contributes valuable segmentation capabilities within its modality focus, but remains challenged in more naturalistic audio-visual scenarios [36].

Recently, Wang *et al*. [36] introduced referring audio-visual segmentation (Ref-AVS), a task that segments objects in video by interpreting text expressions enriched with both audio and visual cues. They also presented Expression Enhancing with Multimodal Cues (EEMC), a simple method that integrates reference text expressions with dynamic audio-visual cues to segment targeted objects in scenes. By unifying these multimodal cues, EEMC enables robust segmentation in dynamic scenes, overcoming limitations of prior approaches that focused on narrower combinations of modalities.

We propose that enhancing Ref-AVS can be achieved by leveraging prior knowledge from large-scale visual foundation models, such as the Segment Anything Model (SAM) [15], for segmentation. SAM is a flexible segmentation model that utilizes an image encoder to generate high-dimensional image embeddings. It incorpo-

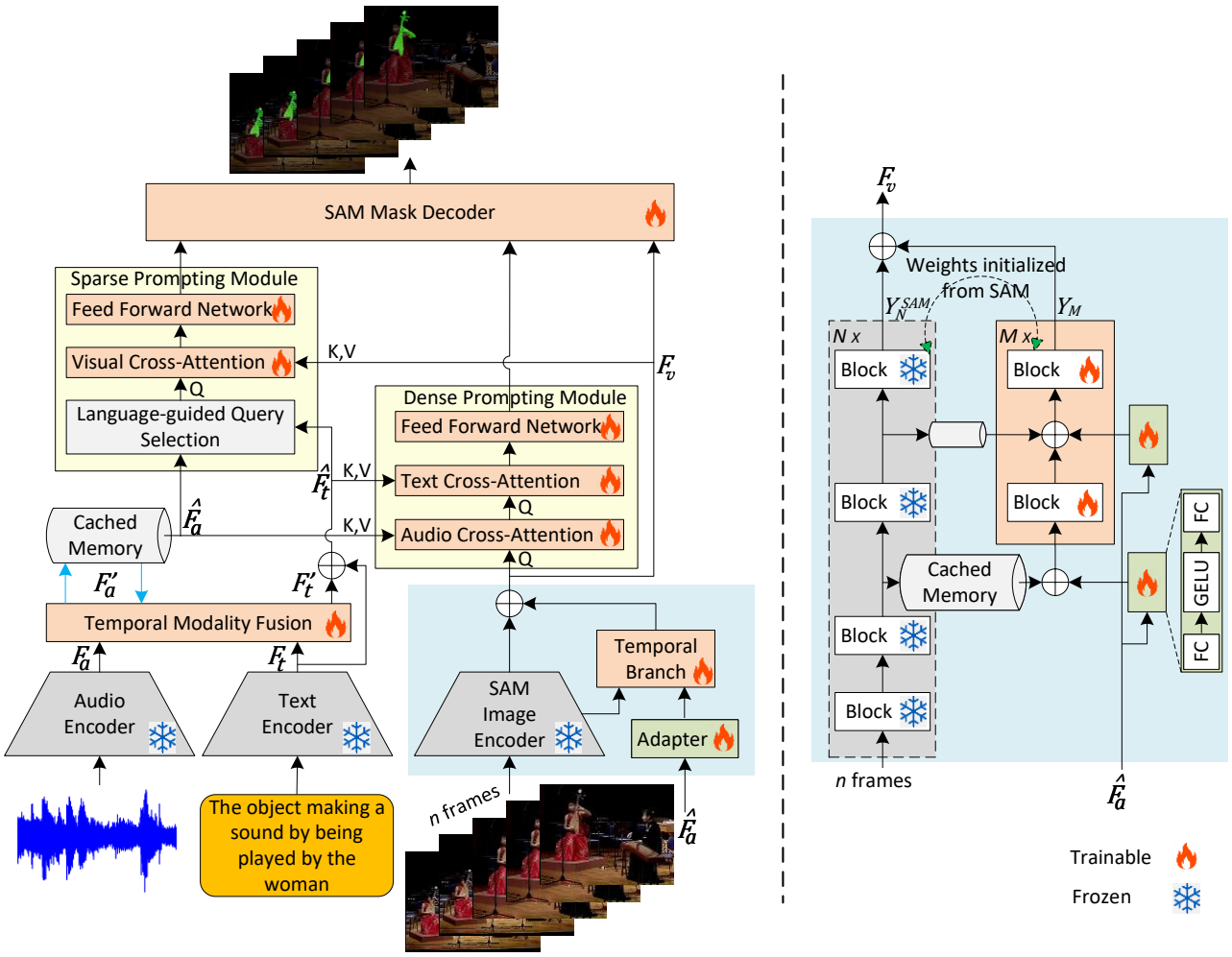

Figure 1. Overview of TSAM. Left: TSAM integrates multimodal cues from diverse inputs, leveraging the prompting modules to guide segmentation. Right: The temporal branch fuses multimodal cues and learns temporal context for segmentation in audio-visual scenes. The adapter input, $\hat{F}_a$, is the output from the cached memory.

rates user-provided visual prompts, which can be sparse (*e.g.*, points, boxes) or dense (*e.g.*, masks), processed by a dedicated prompt encoder. These embeddings are then passed through a lightweight, transformer-based mask decoder to generate accurate object masks based on the provided prompts. Although SAM has been applied to tasks such as AVS [24, 29, 33, 35], and Ref-VOS [19], its potential for segmenting objects in the visual domain using multimodal cues embedded in text expressions remains unexplored. This makes its application to the Ref-AVS task both novel and challenging as it requires the integration of diverse multimodal cues to guide segmentation, overcoming the limitations of SAM's traditional reliance on explicit user-interactive prompts.

The main challenge in adapting SAM for the Ref-AVS task, which requires nuanced multimodal understanding without direct user guidance, arises from SAM's lack of temporal awareness in its image encoder, limiting its ability to capture complex interactions across audio, visual, and text modalities. Additionally, SAM relies on user-provided visual prompts for explicit spatial information of targeted objects, while its capacity to process text and audio modalities remains limited.

This paper introduces TSAM (Fig. 1), a novel SAM-based architecture for the Ref-AVS task, designed to enhance segmentation by leveraging multimodal cues from audio, visual, and text inputs. TSAM incorporates an auxiliary temporal modeling branch in parallel with SAM's image encoder, preserving SAM's zero-shot capabilities while adapting it for temporally varying video content. More-

over, TSAM uses text cues to guide the selection of relevant audio cues, which are then used as sparse prompts, enriched with key visual cues, to query SAM's mask decoder. Since multimodal prompts lack explicit spatial information for the target objects, TSAM incorporates a lightweight adapter module into the temporal branch to enhance SAM's mask decoder prompts, facilitating deeper multimodal interactions during training. Through a cross-attention mechanism, TSAM synthesizes data-driven multimodal reference cues as dense prompts for SAM. Extensive experiments show that TSAM outperforms existing state-of-the-art methods in segmenting target objects in the Ref-AVS task. The main contributions of this work are:

1. We propose TSAM, an end-to-end architecture built on the Segment Anything Model (SAM), which redefines traditional visual prompts as data-driven multimodal prompts for object segmentation in Ref-AVS tasks.
2. We introduce a temporal modeling branch that enhances SAM's image encoder with temporal awareness, enabling the capture of intricate spatio-temporal interactions across multimodal cues.
3. Through comprehensive experiments, we demonstrate that the proposed TSAM achieves superior performance compared to state-of-the-art methods in Ref-AVS. Furthermore, extensive ablation studies provide a detailed evaluation of TSAM's architecture, emphasizing the pivotal role of multimodal cue integration and temporal modeling in enhancing segmentation performance.

2. Related work

2.1. Referring video object segmentation (Ref-VOS)

Referring video object segmentation (Ref-VOS) involves segmenting target objects within video frames based on textual descriptions. Gavrilyuk *et al*. [9] introduced an encoder-decoder model based on I3D [2], that generates dynamic convolutional filters from text and applies them to visual representations at different resolutions to produce segmentation masks for target objects. URVOS [32] enhanced object tracking by performing segmentation on a key frame and then propagating the mask across the video. This approach combines cross-modal attention to prevent drift and memory attention to maintain temporal consistency across frames. MTTR [1] formulated Ref-VOS as a sequence prediction task, utilizing multimodal transformers to process video and text jointly. ReferFormer [37] used Video-Swin (V-Swin) [26] as the visual backbone, linking textual expressions directly to relevant frame regions. In this approach, language-conditioned object queries serve as inputs, helping the model identify the referred objects and apply dynamic kernels as convolution filters to generate segmentation masks. Although Ref-VOS enhances user accessibility by replacing mask annotations with natural language expressions, it still struggles to accurately identify objects in dynamic audio-visual scenes [36].

R2-VOS [18] extended the Ref-VOS task to handle unpaired visual-text inputs, jointly modeling video segmentation and text reconstruction. By introducing a structural text-to-text cycle constraint, R2-VOS improves multimodal alignment across diverse text inputs, overcoming the limitations of prior point-wise methods. RefSAM [19] utilized text prompts with SAM [15], using a lightweight cross-modal MLP to project the referring expression's text embedding into prompt embeddings. RefSAM enables multimodal reference segmentation by integrating textual and visual modalities, but it does not fully exploit the simultaneous interaction of multimodal cues or support the temporal dynamics required for video segmentation tasks.

2.2. Audio-visual segmentation (AVS)

Audio-visual segmentation (AVS), introduced by Zhou *et al*. [39], is a challenging task that seeks to achieve pixel-level segmentation of sound-emitting objects in video, guided by the audio modality. AVSBench [39] used a temporal pixel-wise audio-visual interaction module to inject audio semantics as guidance for the visual segmentation process. Similarly, AVSegFormer [8], a transformer-based model built on the PVT-v2 backbone [34], used cross-attention to fuse spatial-temporal visual cues with audio cues. By decoding these combined audio-visual cues and employing audio queries, the model enhanced its ability to focus on sounding objects within the audio-visual scene. Other transformer-based models proposed in [14, 17, 23] integrated audio and visual cues across both temporal and spatial dimensions for the AVS task. Alternative approaches [20, 22] disentangled the audio input into audio queries, which interacts with the visual input to decode segmentation masks. These approaches emphasize implicit modality fusion strategies, training models to fit discrete samples in the dataset. However, their performance often falls short due to limitations in dataset size and diversity.

Alternatively, a Explicit Conditional Multimodal Variational Auto-Encoder (ECMVAE) [28] was proposed to extract semantically correlated cues from audio-visual inputs for segmenting audible objects in video frames. In another approach, the AVS Bidirectional Generation (AVS-BG) framework [11] was introduced, correlating visual cues of objects with their corresponding audio cues through a visual-to-audio projection and implicit volumetric motion estimation. Recent advancements in AVS [3, 27, 38] also integrated audio-visual contrastive learning methods to select informative samples and improve training efficiency and model performance. However, these methods often fail to fully leverage available information due to limited prior segmentation knowledge in their backbone architectures, hin-

dering their ability to effectively mine informative samples.

On the other hand, SAM-based methods showed improvements in the AVS task. AV-SAM [29] was an early attempt to apply SAM to AVS, utilizing pixel-wise audio-visual fusion to prompt SAM for sounding object segmentation. In [24], two SAM-based approaches for AVS were introduced: Audio-Prompted SAM (AP-SAM) and SAM Adapter (SAMA-AVS). AP-SAM encoded audio as a prompt for SAM's mask decoder, while SAMA-AVS incorporated an adapter module to inject the audio information during the training process, but did not fully leverage SAM's promptable nature. Recently, a Generalizable Audio-Visual Segmentation (GAVS) model [35] was introduced, targeting zero-shot and few-shot scenarios. GAVS proposed a semantic-aware audio prompt to bridge the semantic gap between visual and auditory modalities through contrastive learning. All of these methods take a single image and its corresponding audio as input, fusing multimodal inputs in the spatial domain, which limits their ability to capture temporal relationships frame by frame.

In a straightforward approach, Seon *et al*. [33] extended SAM for AVS by integrating a Spatio-Temporal, Bidirectional Audio-Visual Attention (ST-BAVA) module between SAM's image encoder and mask decoder. While this approach aims to model temporal cues for audible object segmentation, it oversimplifies the interaction between audio and visual modalities, treating them in a relatively shallow manner. The model relies on basic feature fusion instead of a more meaningful interaction, limiting its ability to fully leverage the potential of multimodal cues for AVS tasks.

2.3. Referring audio-visual segmentation (Ref-AVS)

Referring audio-visual segmentation (Ref-AVS) is a novel pixel-level segmentation task in which the model predicts, for each pixel, whether it corresponds to an object specified by multimodal cues in a text expression. Wang *et al*. [36] introduced the Ref-AVS task and developed a dataset with mask annotations for segmenting objects in video frames guided by multimodal expressions, creating the first benchmark for supervised Ref-AVS. They also proposed the Expression Enhancing with Multimodal Cues (EEMC) model, based on the Mask2Former (M2F) [4] segmentation decoder, which adaptively combines audio, visual, and text cues to capture spatio-temporal correspondences between target objects and text expressions.

Unlike existing SAM-based models for tasks like AVS and Ref-VOS [19, 24, 29, 33, 35], which either fail to address deep temporal multimodal alignment [19, 33, 35] or fail to leverage SAM's flexible prompt-based nature [24], our proposed TSAM, shown in Fig. 1, uniquely supports temporal modeling for SAM. It addresses the challenge of segmenting temporally varying objects and enabling a more cohesive and robust integration of multimodal cues for object segmentation in dynamic audio-visual scenes. Additionally, TSAM introduces a dual-pathway framework with sparse and dense prompting modules that replace SAM's visual prompts with data-driven multimodal prompts, fostering richer and more dynamic prompts.

3. TSAM for Ref-AVS

In this section, we introduce the proposed TSAM method (Fig. 1) for the Ref-AVS task. We begin by presenting the problem formulation, followed by an overview of the foundational components of SAM. Next, we describe the temporal modeling branch and the integration of multimodal prompts in TSAM. Finally, we present the training objective that guides TSAM's training.

3.1. Problem formulation

In the Ref-AVS task, the input includes a sequence of video frames, corresponding audio, and a reference text expression specifying the objects to be segmented in the video. The goal is to segment target objects within each video frame, using multimodal cues embedded in the text expression. The segmentation output is represented as binary masks, with each mask highlighting the target objects within an individual video frame.

The visual input includes n frames sampled at 1-second intervals, where each frame $\in \mathbb{R}^{3\times1024\times1024}$ represents RGB values at a resolution of 1024×1024 pixels. The audio input, temporally aligned with the video frames, is encoded offline using VGGish [12], resulting in representations $F_a \in \mathbb{R}^{n\times128}$. Similarly, the text expression is pre-extracted using RoBERTa [25] to generate a text representation $F_t \in \mathbb{R}^{l\times768}$, where l denotes the length of the expression.

To enable effective multimodal prompting within SAM, TSAM first projects both audio and text representations into the SAM visual embedding space with dimensionality d_{emb}. This results in audio cues $F_a \in \mathbb{R}^{n\times1\times d_{emb}}$ and text cues $F_t \in \mathbb{R}^{n\times l\times d_{emb}}$, with the text cues F_t temporally expanded along n. These projected cues are then fused across the temporal dimension in TSAM's temporal modality fusion layer (TMFL), as shown in Fig. 1, producing expression-related multimodal cues as follows:

$$F'_a, F'_t = \mathrm{SA}(\mathrm{Concat}(F_a, F_t)), \tag{1}$$

where $\mathrm{Concat}(\cdot)$ denotes concatenation, and $\mathrm{SA}(\cdot)$ applies temporal self-attention to enable modality fusion over the temporal domain. Following the approach in [36], TSAM integrates a basic cached memory (CM), which stores $\hat{F}_a$, an accumulated summary of F'_a over time. This memory captures the mean modality cues across the temporal domain and maintains a running summary of the evolving audio context. Concurrently, the attention-enhanced text cues

F_t' are added to the text cues F_t to create the unified multimodal text cues $\hat{F}_t \in \mathbb{R}^{n\times l\times d_{emb}}$, resulting in a cohesive and enriched representation of audio and text cues.

3.2. SAM as a foundational baseline

SAM [15] is the first large-scale pretrained foundational model for guided image segmentation. Trained on a massive dataset consisting of 1 billion masks across 11 million images, SAM has demonstrated remarkable effectiveness and versatility in image segmentation tasks. The core components of SAM are outlined here, with further details available in [15]. SAM's architecture is composed of three key modules:

(a) **Image encoder.** A pre-trained Vision Transformer (ViT) [7], minimally adapted to process high-resolution images, extracts spatial embeddings.

(b) **Prompt encoder.** SAM's prompt encoder handles various user prompts, including sparse (points, boxes, text) and dense (masks) formats, providing regional information about the target objects for segmentation.

(c) **Mask decoder.** The mask decoder consists of two layers. First, the dense prompt embedding is integrated with the image embedding. Then, cross-attention is applied between the image embedding and the sparse prompt embeddings to update both representations. In each attention layer, the original prompt token is re-integrated into both the query and key tokens. Finally, the dot product of the upsampled image embedding and the MLP-mapped sparse prompt embeddings produces the final segmentation mask.

3.3. Temporal modeling branch

The SAM image encoder does not account for the temporal dynamics inherent in video frames, limiting its capacity to handle the complex interplay between audio and visual cues in tasks such as object segmentation. To address this, the proposed TSAM enhances SAM's architecture by incorporating a dedicated temporal modeling branch at the final M blocks of the SAM image encoder, as shown in Fig. 1. While the SAM image encoder remains frozen during training, the newly introduced temporal blocks are the only trainable. These temporal blocks mirror the structure of the corresponding SAM encoder blocks, initialized with the same pre-trained weights.

Each video frame is processed independently through the SAM image encoder. At each temporal modeling block, TSAM integrates a straightforward cached memory and adapter module, which tracks and captures the evolving temporal and multimodal cues. Specifically, the m-th ($m = 1, 2, \ldots, M$) temporal block receives input x_m, combining the output y_{m-1} of the previous temporal block, the output $y^{\text{SAM}}_{(N-M)+(m-1)}$ of the previous SAM encoder block, and the textual enriched audio cues $\hat{F}_a$ via the adapter module:

$$x_m = y_{m-1} + \text{CM}_m(y^{\text{SAM}}_{(N-M)+(m-1)}) + \text{AM}_m(\hat{F}_a), \quad (2)$$

where $\text{CM}_m(\cdot)$ denotes the cached memory applied to the SAM block output, and $\text{AM}_m(\cdot)$ represents the adapter module applied to the audio cues $\hat{F}_a$. The adapter module, a bottleneck block shown in Fig. 1, integrates audio cues that have been refined by text cues, facilitating deeper interaction among the audio, visual, and text modalities within the temporal modeling branch.

Finally, the overall video cues, $F_v = \{f_v^i\}_{i=1}^n$, where each f_v^i represents a cue for frame i, are derived by integrating the output from the temporal modeling branch Y_M with the output from SAM's image encoder Y_N^{SAM}:

$$F_v = Y_M + Y_N^{SAM}. \quad (3)$$

Each video frame visual cues $f_v^i \in \mathbb{R}^{d_{emb}\times 64\times 64}$, with 64×64 representing the spatial dimensions.

3.4. Prompting SAM with multimodal cues

TSAM introduces a dual-pathway framework with dedicated sparse and dense prompting modules (SPM and DPM, respectively) to enhance SAM's segmentation capabilities. The SPM learns global context by leveraging sparse queries derived from multimodal cues, guiding TSAM toward high-level object localization. Meanwhile, the DPM operates at a finer granularity, providing dense prompts aligned with the frame's spatial features, highlighting its reliance on the outputs of SAM's image encoder, as shown in Fig. 1.

Sparse prompting module (SPM). To effectively guide segmentation within TSAM using the text expression, SPM first employs a language-guided query selection mechanism to select the k most relevant audio cues, associated with the input text expression. The top k query indices, denoted as A_k, are selected as:

$$A_k = \text{Top}_k\left(\text{maxf}(\hat{F}_a\hat{F}_t^\top)\right), \quad (4)$$

where Top_k picks the top k indices, maxf computes the maximum along the cues' dimension, and $(\cdot)^\top$ indicates matrix transposition. The selected audio cues are then fused with the visual cues through a cross-attention mechanism and transformed using a feed forward network-mapped sparse prompt embedding. This process generates contextually enriched prompts for SAM's mask decoder.

Dense prompting module (DPM). The DPM in TSAM generates spatially comprehensive embeddings by integrating visual, audio, and textual cues through cross-attention mechanisms, as illustrated in Fig. 1. First, an audio cross-attention layer aligns the audio cues with the visual cues, facilitating effective fusion of both modalities. Next, a text cross-attention layer incorporates the text cues, further refining the combined audio-visual cues to focus on the target object for segmentation. Finally, a feed forward network refines the multimodal embeddings, preparing them for input into SAM's mask decoder.

3.5. Training objective

We employ binary cross-entropy $\mathcal{L}_{\text{BCE}}$ and intersection over union $\mathcal{L}_{\text{IoU}}$ losses to measure the discrepancy between TSAM's predicted mask and the ground truth mask during training. The total loss $\mathcal{L}_{\text{total}}$ for model optimization is defined as a weighted sum of $\mathcal{L}_{\text{BCE}}$ and $\mathcal{L}_{\text{IoU}}$ losses, as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{BCE}} + \lambda \cdot \mathcal{L}_{\text{IoU}}, \tag{5}$$

where λ is set to 1.0 to balance the two loss components.

4. Experiments

4.1. Dataset

We evaluated TSAM using the Ref-AVS dataset [36], which contains a diverse collection of 20,000 text expressions and pixel-level annotations across 4,000 10-second videos. The dataset features a wide range of audible objects, including 20 categories of musical instruments, 8 animals, 15 machines, and 5 humans, as well as a smaller set of 3 categories of static, inaudible objects.

The dataset is divided into three sets: a training set of 2,908 videos, a validation set of 276 videos, and a test set of 818 videos. The test set is further divided into three subsets: 292 **Seen** videos, which includes object categories present in the training set; 269 **Unseen** videos, which consists of object categories not seen during training but belonging to super-categories represented in the training data, assessing the model's ability to generalize to new objects; and 257 **Null** videos, where the text expression refers to an object that either does not exist or is not visible in the given context, with empty true masks, requiring the model to avoid segmenting any object.

4.2. Evaluation metrics

Following [36], we used the Jaccard index $\mathcal{J}$ and F-score $\mathcal{F}$ as performance metrics. For the Null test set, we employed:

$$\mathcal{S} = \sqrt{\frac{\text{predicted mask area}}{\text{background area}}}, \tag{6}$$

a metric to assess expression guidance in the model. A higher $\mathcal{S}$ value indicates a larger mask relative to the background, suggesting less accurate expression guidance.

4.3. Implementation details

We initialized TSAM for the Ref-AVS task using the pre-trained ViT-B variant of SAM ($N = 12$, $d_{emb} = 256$). The temporal branch consists of $M = 4$ blocks, and the selected audio queries is $k = 5$. Training was performed with PyTorch on eight AMD GPUs in a distributed setup. The AdamW optimizer was used, with an initial learning rate of $1 \cdot 10^{-4}$ and a batch size of 1. The model was trained for 15 epochs with periodic evaluations, and the best-performing model on the validation subset was selected for testing.

Method	Task	Visual Backbone	Seen		Unseen		Null
			$\mathcal{J}(\%)$	$\mathcal{F}$	$\mathcal{J}(\%)$	$\mathcal{F}$	$\mathcal{S}(\downarrow)$
AVSBench [39]	AVS†	PVT-v2	23.20	0.511	32.36	0.547	0.208
AVSegFormer [8]	AVS†	PVT-v2	33.47	0.470	36.05	0.501	0.171
GAVS [35]	AVS†	SAM	28.93	0.498	29.82	0.497	0.190
SAMA [24]	AVS*	SAM	<u>39.22</u>	<u>0.562</u>	47.50	0.566	0.130
ReferFormer [37]	Ref-VOS‡	V-Swin	31.31	0.501	30.40	0.488	0.176
R2-VOS [18]	Ref-VOS‡	V-Swin	25.01	0.410	27.93	0.498	0.183
EEMC [36]	Ref-AVS	M2F	34.20	0.513	<u>49.54</u>	<u>0.648</u>	**0.007**
TSAM (Ours)	Ref-AVS	SAM	**43.43**	**0.568**	**54.58**	**0.664**	<u>0.017</u>

Table 1. Performance comparison on the Ref-AVS dataset. † denotes text integration in AVS from [36], ‡ indicates audio integration in Ref-VOS from [36], and * marks our implementation with text added to AVS.

4.4. Results and comparison

We compared the proposed TSAM method with the EEMC method [36] and existing methods from the AVS and Ref-VOS domains, including AVSBench [39], AVSegFormer [8], GAVS [35], SAMA [24], ReferFormer [37], and R2-VOS [18], using the Ref-AVS dataset [36]. For consistency and fairness, we reported the published results for all methods in [36] and re-evaluated the SAMA method by integrating text with both audio and visual inputs to align with our multimodal setup.

The results, presented in Tab. 1, demonstrate that even when augmented with text (for AVS methods) or audio (for Ref-VOS methods), these methods do not yield satisfactory performance in the Ref-AVS task. Interestingly, methods that utilize prior segmentation visual backbones, such as SAM and M2F (Mask2Former), tend to outperform those based on visual backbones like PVT-v2 and V-Swin.

Although GAVS and SAMA are SAM-based methods, they perform worse than EEMC. SAMA fails to leverage SAM's flexible, promptable nature, integrating additional modality cues superficially by layering them on SAM's frozen image encoder. This underutilizes the potential of multimodal cues, limiting SAMA's segmentation capabilities. GAVS, on the other hand, prompts SAM with multimodal cues but does not address the necessary alignment and interplay among these modalities, leading to a lack of cohesive multimodal fusion. These results confirm that simple modality fusion alone is insufficient for handling the complexity of multimodal integration in the Ref-AVS task.

In contrast, our TSAM leverages language-guided audio queries as sparse prompts, integrates multimodal dense prompts via cross-attention, and incorporates a temporal modeling branch and adapter module to enhance SAM's ability to perform accurate, contextually guided segmentation of dynamic objects in complex audio-visual scenes. Compared to the best-performing method, EEMC, on the Ref-AVS dataset, our TSAM achieves superior results: a $\mathcal{J} = 9.23\%$ and $\mathcal{F} = 5.5\%$ improvements on the Seen test set, and a $\mathcal{J} = 5.04\%$ and $\mathcal{F} = 1.6\%$ improvements

on the Unseen test set. Although TSAM falls slightly behind EEMC on the Null test set by $\mathcal{S} = 1\%$, likely due to SAM's limitation of always producing a segmentation mask even when no target object is present, TSAM maintains a strong advantage on Unseen test set. This is primarily due to TSAM's ability to generalize to unseen object categories, preserving SAM's pretrained knowledge while ensuring robust segmentation in novel audio-visual scenes, thanks to the added temporal modeling branch.

4.5. Qualitative results

We further evaluated our TSAM method qualitatively on both the Seen and Unseen test sets, as shown in Fig. 2 and Fig. 3, respectively. Fig. 2 demonstrates TSAM's consistent ability to produce high-quality masks for targeted objects, even when they are inaudible, as seen in the middle example. This highlights TSAM's proficiency in processing complex multimodal instructions through its prompting modules, enabling accurate segmentation based on textual cues. On the Unseen test set, Fig. 3, TSAM showcases a remarkable capacity to generalize to novel objects, effectively aligning audio-visual and textual inputs in novel scenes to accurately segment the targeted object of interest. This underscores TSAM's ability to generalize to new contexts, interpreting diverse multimodal instructions and segmenting unseen objects. TSAM's temporal branch enhances understanding of novel objects in dynamic scenes, while SAM's pre-trained knowledge is preserved for core segmentation capabilities. Additional qualitative results and comparisons are provided in the supplementary materials.

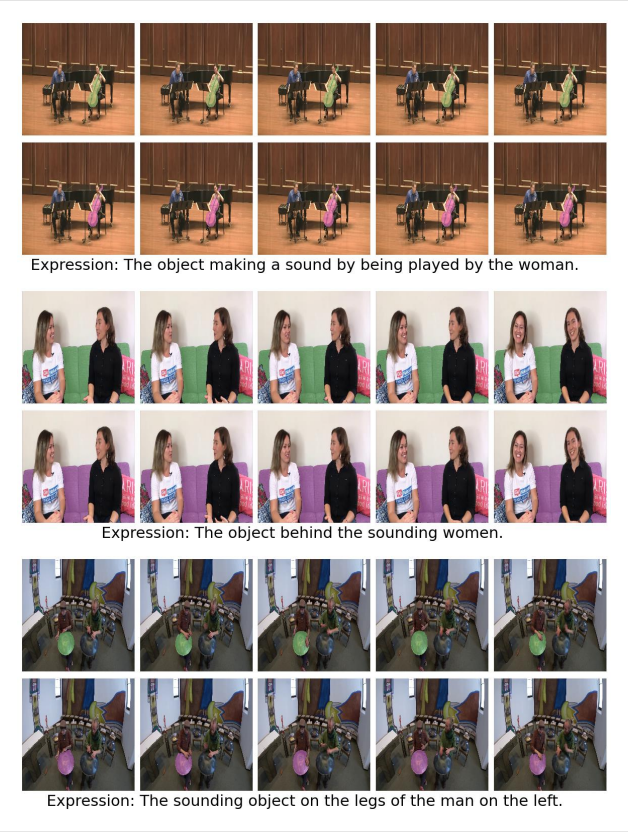

Figure 2. Segmentation masks for referred objects in the Ref-AVS dataset (Seen test set). Top row: ground truth with target object in green. Bottom row: our results with target object in violet.

4.6. Ablation study

We analyzed the impact of omitting key components of TSAM, including the temporal branch (TB), temporal modality fusion layer (TMFL), dense prompting module (DPM), sparse prompting module (SPM), cached memory (CM), and adapter module (AM), as shown in Tab. 2. CM(a+v) denotes cached memory for both audio and visual modalities, while CM(v) denotes cached memory on the visual modality only, see Fig. 1. Additionally, Tab. 2 presents the effect of the intersection over union ($\mathcal{L}_{\text{IoU}}$) loss, defined in Eq. (5), on TSAM's training.

As shown in Tab. 2, we observe a significant performance degradation in setting ② when the temporal branch (TB) is removed, with a drop of 7.24 %pt in $\mathcal{J}$ (from 49.01% to 41.77%) and 3.4 %pt in $\mathcal{F}$ (from 0.616 to 0.582). This performance loss is significantly higher than that in other settings, emphasizing the critical role of the temporal modeling branch in TSAM's ability to generalize across video frames and maintain coherence in dynamic scenes, which is essential for object segmentation in complex Ref-AVS tasks. Settings ④ and ⑤ demonstrate the importance of prompting SAM with multimodal cues through SPM and DPM. SPM and DPM play complementary roles in TSAM, and their joint contribution is pivotal for its overall performance. Omitting these prompts in setting ⑥ causes a notable decrease in segmentation performance, reinforcing our claim that the rich semantic cues from multimodal prompts are essential for improving segmentation performance.

Settings ⑦ and ⑧ evaluate the effect of cached memory on both audio and visual (CM(a+v)) and visual-only

Setting	Seen		Unseen		Mix(S+U)		Null
	$\mathcal{J}(\%)$	$\mathcal{F}$	$\mathcal{J}(\%)$	$\mathcal{F}$	$\mathcal{J}(\%)$	$\mathcal{F}$	$\mathcal{S}(\downarrow)$
① TSAM	**43.43**	0.568	**54.58**	**0.664**	**49.01**	**0.616**	0.017
② – TB	33.05	0.507	50.48	0.657	41.77	0.582	0.505
③ – TMFL	40.35	0.579	45.54	0.627	42.95	0.603	0.018
④ – DPM	42.72	0.580	49.10	0.647	45.91	0.614	0.018
⑤ – SPM	43.04	0.580	49.75	0.652	46.40	**0.616**	0.018
⑥ – SPM+DPM	42.60	0.602	40.58	0.604	41.59	0.603	0.018
⑦ – CM(a+v)	42.07	0.544	49.11	0.659	45.59	0.602	0.018
⑧ – CM(v)	42.75	0.549	51.18	0.660	46.97	0.605	0.018
⑨ – AM	43.13	**0.600**	40.79	0.617	41.96	0.609	0.017
⑩ – $\mathcal{L}_{\text{IoU}}$	38.29	0.564	42.15	0.631	40.22	0.598	**0.008**

Table 2. Ablation study on TSAM components and $\mathcal{L}_{IoU}$ loss. 'Mix' refers to the average performance on the Seen (S) and Unseen (U) test sets.

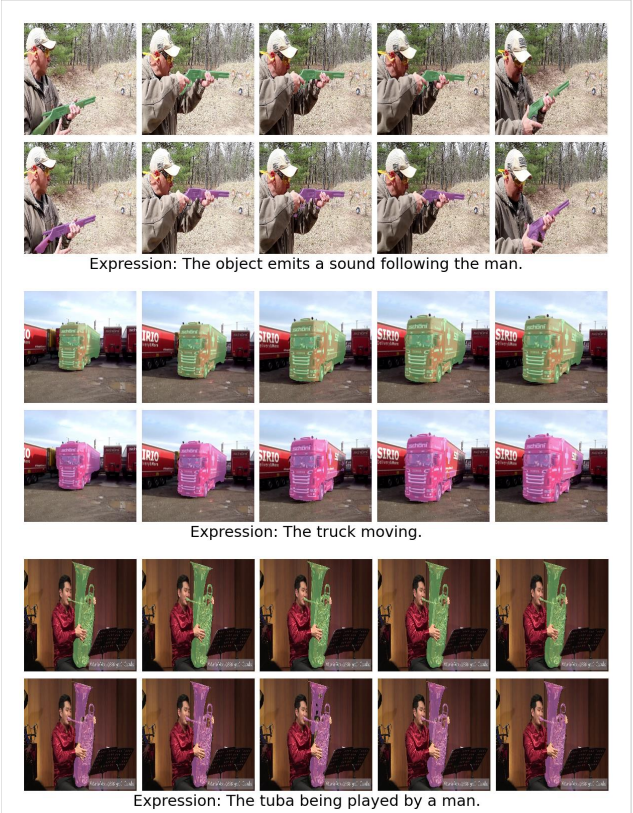

Figure 3. Segmentation masks for referred objects in the Ref-AVS dataset (Unseen test set). Top row: ground truth with target object in green. Bottom row: our results with target object in violet.

(CM(v)) modalities, demonstrate a performance degradation in TSAM when cached memory is not applied across modalities (CM(a+v)) and when only the visual modality's memory (CM(v)) is omitted. The latter specifically emphasizes the importance of maintaining shared memory across video frames for effective temporal alignment and multimodal integration, as evidenced by the results.

The results from setting ⑨ highlight the adapter module's importance. Its omission causes a significant performance drop, underscoring its crucial role in TSAM. We also observe that the components of TSAM, particularly in settings ③ – ⑨, have a substantial impact on performance, especially on the Unseen test set. This set challenges the model to segment novel objects, demonstrating TSAM's ability to generalize beyond seen categories, an essential feature for robust performance in open-world scenarios.

Finally, setting ⑩ highlights the distinct contribution of $\mathcal{L}_{\text{IoU}}$ to the overall loss function $\mathcal{L}_{\text{total}}$. Notably, the model performs best on both Seen and Unseen test sets when $\mathcal{L}_{\text{IoU}}$ is included in the total loss. Omitting $\mathcal{L}_{\text{IoU}}$ results in the highest performance on the Null test set. Further ablation studies of λ (Eq. (5)) are in the supplementary materials.

TSAM	Seen		Unseen		Mix(S+U)		Null
	$\mathcal{J}(\%)$	$\mathcal{F}$	$\mathcal{J}(\%)$	$\mathcal{F}$	$\mathcal{J}(\%)$	$\mathcal{F}$	$\mathcal{S}(\downarrow)$
$k=3$	**43.58**	**0.579**	50.43	0.655	47.01	**0.617**	0.018
$k=5$	43.43	0.568	**54.58**	**0.664**	**49.01**	0.616	**0.017**
$k=7$	43.24	0.573	46.95	0.630	45.10	0.601	0.018
$M=2$	42.04	0.566	53.51	0.651	47.76	0.609	0.023
$M=4$	**43.43**	0.568	**54.58**	**0.664**	**49.01**	**0.616**	**0.017**
$M=6$	43.27	**0.575**	49.86	0.640	46.57	0.608	0.020

Table 3. Effect of audio queries k and temporal branch depth M on TSAM's performance.

We further evaluated the impact of varying the number of audio queries k, defined in Eq. (4), and the temporal branch depth M, see Sec. 3.3, on TSAM's performance, as shown in Tab. 3. Among the configurations $k = 3$, $k = 5$, and $k = 7$, we observe that $k = 5$ performs best. This balance optimizes the relevance of audio cues without overloading the model with irrelevant queries to SAM's mask decoder.

Similarly, we find that $M = 4$ yields optimal results for the temporal branch. A shallower setup ($M = 2$) likely lacks the temporal depth required for effective generalization and multimodal integration, while a deeper branch ($M = 6$) may introduce excessive complexity, impairing the model's generalization to unseen objects.

5. Conclusions

This paper presented TSAM, a novel extension of the Segment Anything Model (SAM) tailored for the referring audio-visual segmentation (Ref-AVS) task. TSAM enhances SAM's segmentation capabilities by effectively integrating multimodal cues through data-driven dense and sparse prompting modules while modeling temporal dependencies via a dedicated temporal branch. This approach preserves SAM's foundational knowledge while improving performance, particularly with novel objects. Extensive experiments on the Ref-AVS dataset demonstrate TSAM's strong generalization ability, showcasing its effectiveness in segmenting objects from textual cues and its robustness in unseen scenarios. These results highlight TSAM's potential to advance cross-modal understanding in multimodal tasks.

Limitations and future work. Although TSAM outperforms existing Ref-AVS methods, a direct comparison with the recent SAM2 [31] remains to be explored. SAM2, designed for video segmentation with a memory module for temporal visual representations, lacks multimodal support. Future work will adapt SAM2 for the Ref-AVS task.

Acknowledgments. This work was supported by the Research Council of Finland through the Finnish Center for Artificial Intelligence flagship (grant 345552) and the USSEE project (grant 345791). We also acknowledge CSC–IT Center for Science, Finland, for awarding access to the LUMI supercomputer, owned by the EuroHPC Joint Undertaking.

References

[1] Adam Botach, Evgenii Zheltonozhskii, and Chaim Baskin. End-to-end referring video object segmentation with multimodal transformers. In *CVPR*, pages 4985–4995, 2022. 3

[2] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the Kinetics dataset. In *CVPR*, pages 6299–6308, 2017. 3

[3] Yuanhong Chen, Yuyuan Liu, Hu Wang, Fengbei Liu, Chong Wang, Helen Frazer, and Gustavo Carneiro. Unraveling instance associations: A closer look for audio-visual segmentation. In *CVPR*, pages 26497–26507, 2024. 3

[4] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, pages 1290–1299, 2022. 4

[5] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. MeViS: A large-scale benchmark for video segmentation with motion expressions. In *ICCV*, pages 2694–2703, 2023. 1

[6] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, Philip HS Torr, and Song Bai. MOSE: A new dataset for video object segmentation in complex scenes. In *ICCV*, pages 20224–20234, 2023. 1

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 5

[8] Shengyi Gao, Zhe Chen, Guo Chen, Wenhai Wang, and Tong Lu. Avsegformer: Audio-visual segmentation with transformer. In *AAAI*, pages 12155–12163, 2024. 3, 6

[9] Kirill Gavrilyuk, Amir Ghodrati, Zhenyang Li, and Cees GM Snoek. Actor and action video segmentation from a sentence. In *CVPR*, pages 5958–5966, 2018. 1, 3

[10] Yuxin Guo, Siyang Sun, Shuailei Ma, Kecheng Zheng, Xiaoyi Bao, Shijie Ma, Wei Zou, and Yun Zheng. CrossMAE: Cross-modality masked autoencoders for region-aware audio-visual pre-training. In *CVPR*, pages 26721–26731, 2024. 1

[11] Dawei Hao, Yuxin Mao, Bowen He, Xiaodong Han, Yuchao Dai, and Yiran Zhong. Improving audio-visual segmentation with bidirectional generation. In *AAAI*, pages 2067–2075, 2024. 3

[12] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. CNN architectures for large-scale audio classification. In *ICASSP*, pages 131–135, 2017. 4

[13] Xiao Hu, Basavaraj Hampiholi, Heiko Neumann, and Jochen Lang. Temporal context enhanced referring video object segmentation. In *WACV*, pages 5574–5583, 2024. 1

[14] Shaofei Huang, Han Li, Yuqing Wang, Hongji Zhu, Jiao Dai, Jizhong Han, Wenge Rong, and Si Liu. Discovering sounding objects by audio queries for audio visual segmentation. In *IJCAI*, pages 875–883, 2023. 3

[15] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, pages 4015–4026, 2023. 1, 3, 5

[16] Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. In *CVPR*, pages 19108–19118, 2022. 1

[17] Kexin Li, Zongxin Yang, Lei Chen, Yi Yang, and Jun Xiao. Catr: Combinatorial-dependence audio-queried transformer for audio-visual video segmentation. In *ACM MM*, pages 1485–1494, 2023. 3

[18] Xiang Li, Jinglu Wang, Xiaohao Xu, Xiao Li, Bhiksha Raj, and Yan Lu. Robust referring video object segmentation with cyclic structural consensus. In *ICCV*, pages 22236–22245, 2023. 3, 6

[19] Yonglin Li, Jing Zhang, Xiao Teng, Long Lan, and Xinwang Liu. RefSAM: Efficiently adapting segmenting anything model for referring video object segmentation. *arXiv preprint arxiv.org/abs/2307.00997*, 2024. 2, 3, 4

[20] Yuhang Ling, Yuxi Li, Zhenye Gan, Jiangning Zhang, Mingmin Chi, and Yabiao Wang. TransAVS: End-to-end audio-visual segmentation with transformer. In *ICASSP*, pages 7845–7849, 2024. 3

[21] Chen Liu, Peike Patrick Li, Xingqun Qi, Hu Zhang, Lincheng Li, Dadong Wang, and Xin Yu. Audio-visual segmentation by exploring cross-modal mutual semantics. In *ACM MM*, pages 7590–7598, 2023. 1

[22] Chen Liu, Peike Li, Hu Zhang, Lincheng Li, Zi Huang, Dadong Wang, and Xin Yu. BAVS: Bootstrapping audio-visual segmentation by integrating foundation knowledge. *IEEE TMM*, 26:10015–10028, 2024. 3

[23] Jinxiang Liu, Chen Ju, Chaofan Ma, Yanfeng Wang, Yu Wang, and Ya Zhang. Audio-aware query-enhanced transformer for audio-visual segmentation. *arXiv preprint arXiv:2307.13236*, 2023. 3

[24] Jinxiang Liu, Yu Wang, Chen Ju, Chaofan Ma, Ya Zhang, and Weidi Xie. Annotation-free audio-visual segmentation. In *WACV*, pages 5604–5614, 2024. 1, 2, 4, 6

[25] Y Liu, M Ott, N Goyal, J Du, M Joshi, D Chen, O Levy, M Lewis, L Zettlemoyer, and V Stoyanov. RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 4

[26] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video Swin Transformer. In *CVPR*, pages 3202–3211, 2022. 3

[27] Yuxin Mao, Jing Zhang, Mochu Xiang, Yunqiu Lv, Yiran Zhong, and Yuchao Dai. Contrastive conditional latent diffusion for audio-visual segmentation. *arXiv preprint arXiv:2307.16579*, 2023. 3

[28] Yuxin Mao, Jing Zhang, Mochu Xiang, Yiran Zhong, and Yuchao Dai. Multimodal variational auto-encoder based audio-visual segmentation. In *ICCV*, pages 954–965, 2023. 3

[29] Shentong Mo and Yapeng Tian. AV-SAM: Segment anything model meets audio-visual localization and segmentation. *arXiv preprint arXiv:2305.01836*, 2023. 2, 4

[30] Federico Perazzi, Anna Khoreva, Rodrigo Benenson, Bernt Schiele, and Alexander Sorkine-Hornung. Learning video object segmentation from static images. In *CVPR*, pages 2663–2672, 2017. 1

[31] 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 Dollar, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *ICLR*, 2025. 8

[32] Seonguk Seo, Joon-Young Lee, and Bohyung Han. URVOS: Unified referring video object segmentation network with a large-scale benchmark. In *ECCV*, pages 208–223, 2020. 3

[33] Juhyeong Seon, Woobin Im, Sebin Lee, Jumin Lee, and Sung-Eui Yoon. Extending segment anything model into auditory and temporal dimensions for audio-visual segmentation. In *ICIP*, pages 2480–2486, 2024. 1, 2, 4

[34] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. PVT v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):415–424, 2022. 3

[35] Yaoting Wang, Weisong Liu, Guangyao Li, Jian Ding, Di Hu, and Xi Li. Prompting segmentation with sound is generalizable audio-visual source localizer. In *AAAI*, pages 5669–5677, 2024. 1, 2, 4, 6

[36] Yaoting Wang, Peiwen Sun, Dongzhan Zhou, Guangyao Li, Honggang Zhang, and Di Hu. Ref-AVS: Refer and segment objects in audio-visual scenes. In *ECCV*, pages 196–213, 2024. 1, 3, 4, 6

[37] Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. In *CVPR*, pages 4974–4984, 2022. 3, 6

[38] Jiarui Yu, Haoran Li, Yanbin Hao, Jinmeng Wu, Tong Xu, Shuo Wang, and Xiangnan He. How can contrastive pre-training benefit audio-visual segmentation? A study from supervised and zero-shot perspectives. In *BMVC*, pages 367–374, 2023. 3

[39] 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 *ECCV*, pages 386–403, 2022. 1, 3, 6