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.

SAM2-LOVE: Segment Anything Model 2 in Language-aided Audio-Visual Scenes

Yuji Wang[1]*, Haoran Xu[2]*, Yong Liu[1], Jiaze Li[2], Yansong Tang[1]†
[1]Tsinghua Shenzhen International Graduate School, Tsinghua University, [2]ZJU
yuji-wan24@mails.tsinghua.edu.cn, †tang.yansong@sz.tsinghua.edu.cn

Abstract

Reference Audio-Visual Segmentation (Ref-AVS) aims to provide a pixel-wise scene understanding in Language-aided Audio-Visual Scenes (LAVS). This task requires the model to continuously segment objects referred to by text and audio from a video. Previous dual-modality methods always fail due to the lack of a third modality and the existing triple-modality method struggles with spatio-temporal consistency, leading to the target shift of different frames. In this work, we introduce a novel framework, termed ***SAM2-LOVE****, which integrates textual, audio, and visual representations into a learnable token to prompt and align SAM2 for achieving Ref-AVS in the LAVS. Technically, our approach includes a multimodal fusion module aimed at improving multimodal understanding of SAM2, as well as token propagation and accumulation strategies designed to enhance spatio-temporal consistency without forgetting historical information. We conducted extensive experiments to demonstrate that SAM2-LOVE outperforms the SOTA by 8.5% in $\mathcal{J}\&\mathcal{F}$ on the Ref-AVS benchmark and showcase the simplicity and effectiveness of the components. Our code will be available here.*

1. Introduction

In the process of human perception of the world, visual signals play a primary role, often accompanied by other multimodal signals such as audio and text. These additional multimodal cues are not only essential to locating and separating objects and areas of interest but also to understanding the changing behaviors of the objects. The dynamic scene that integrates visual elements with diverse multimodal cues including audio and text is regarded as the Language-aided Audio-Visual Scene (LAVS). To mimic the effect from humans, prior works such as Music-AVQA [10] and AVQA [39] have primarily focused on question-answering tasks. Nevertheless, these approaches lack a fine-grained understanding of LVAS at the pixel level. As shown in Figure 2, Reference Audio-Visual Segmentation (Ref-AVS) task [31] makes a good example to improve the pixel-wise

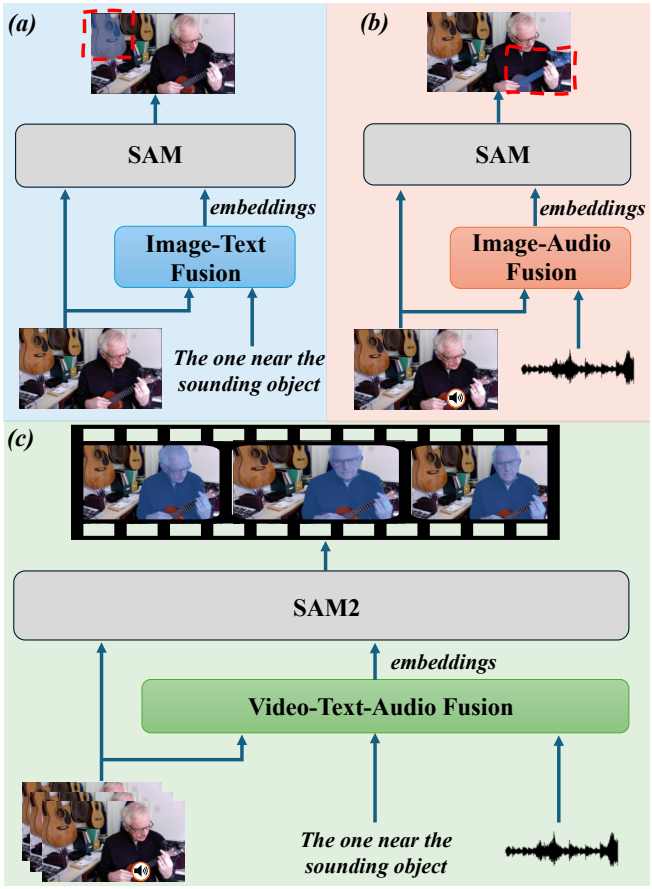


Figure 1. The comparison with dual-modality SAM pipelines. (a) The text-visual SAM struggles to distinguish the sounding object in the silent scenes. (b) The audio-visual SAM fails to understand the dynamic control from the text input. (c) Our model fuses three modalities' information and expands to the video due to SAM2.

*Equal contribution
†Corresponding author

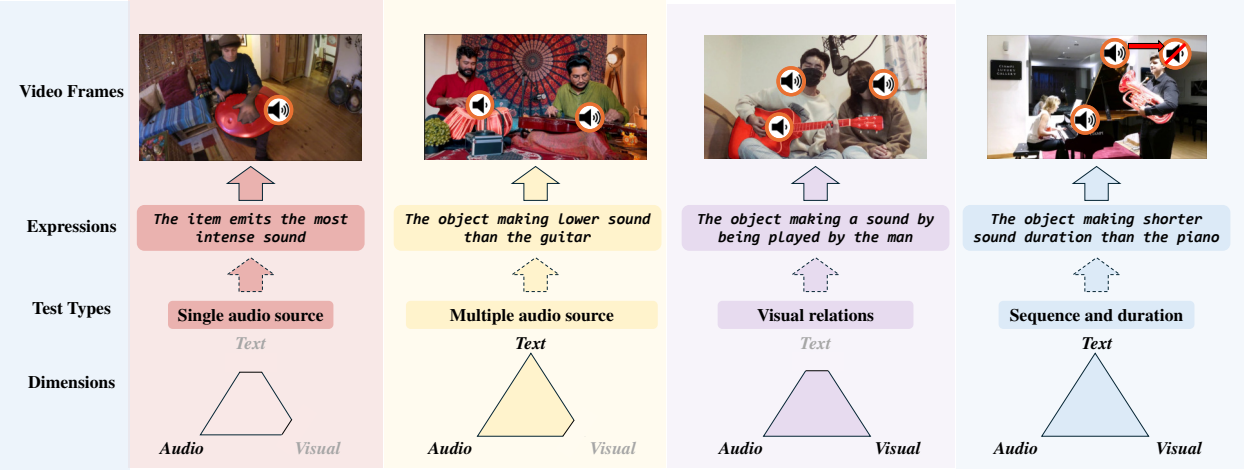


Figure 2. Illustration of the Ref-AVS task under varying modality cues. The expression can be treated as a kind of control signal to condition the referred object on audio. Different expressions can test the models' pixel-wise scene understanding capabilities in the LAVS with different dimensions. The absence angle of a triangle means that modality is not the main test dimension.

scene understanding of the model in the LAVS. The previous duality-referring segmentation works struggle with the Ref-AVS task where the interested object can only be located under triple modalities, shown in Figure 1. In neither Figure 1 (a) text-visual EVF-SAM [43] nor Figure 1 (b) audio-visual GAVS [30], the truly referred target is accurately segmented due to the absence of additional help from the other third modality. Additionally, they are constrained in the image level, which can not be directly adapted to the dynamic nature of video content. Therefore, methods for effectively fusing information from the three modalities for segmentation have garnered widespread attention.

Prior work, EEMC [31], simultaneously models three modalities' representations to achieve excellent results on the Ref-AVS benchmark. However, it showcases the relatively insufficient spatio-temporal consistency. Specifically, despite the existence of memory caches, the model fails to continuously track the target's position and shape, leading to a shift in the segmented area over time. Recently, SAM2 [22] has been proposed to generalize the segmenting anything capabilities of the image to the video, which demonstrates a powerful video tracking capability in interframe consistency. SAM2 follows the previous promptable visual segmentation of SAM [8] that requires a strong manual annotation to the interested target of the starting frame (not only limited to the first frame) and then propagates to the whole video based on the crucial frame. Nevertheless, the lack of dependency on other modalities (text or audio) constraints its further application in the LAVS.

Thus, in this paper, we propose a novel framework, namely **SAM2-LOVE**, which means **S**egment **A**nything **M**odel **2** in **L**anguage-aided audi**O**-**V**isual sc**E**nes. To enhance SAM2's understanding of multimodal information, we design the multimodal fusion module to compress all the modalities' information into a learnable token that aligns the SAM2 for segmenting the referring target. Additionally, we incorporate token propagation and accumulation strategies to strengthen spatio-temporal understanding of the video while avoiding forgetting historical frames. Our main contributions can be summarized as follows:

- We propose a novel framework, **SAM2-LOVE** that firstly leverages SAM2 to achieve pixel-wise understanding in the LAVS by designing a multimodal fusion module.
- We develop creative token propagation and accumulation strategies to improve spatio-temporal comprehension of the promtable token.
- Extensive experiments on Ref-AVS dataset demonstrate the superiority of our method, with ablation studies highlighting the simplicity and effectiveness of its modules.

2. Related Works

Segmentation via Language or Audio. Unlike semantic segmentation [1, 17, 29, 32], visual segmentation via language starts with referring expression segmentation task [7, 18, 21, 33, 34, 36, 40, 41] which performs target object segmentation based on given natural language expressions. X-Decoder [47] bridges vision and language, unifying multiple tasks within a single model. Moreover, GRES [13] defines a new task called generalized reference expression segmentation. With the development of multimodal large language models (MLLMs) such as LLaVA [14, 15], LISA [9] proposes the concept of reasoning segmentation with combination of pixel-level understanding and MLLMs. GSVA [37] and PixelLM [23] propose multi-target and null-target reasoning segmentation. On the other

hand, audio-visual segmentation (AVS) is also constantly improving. Audio-Visual Segmentation [45, 46] proposes the AVS problem and AVSBench dataset. Numerous subsequent methods [2, 4, 27] have been developed and refined based on the proposed benchmark.

Segment Anything Model. SAM [8] is an interactive segmentation model based on various types of prompts (points, boxes, coarse masks). It shows significant generalization ability in open-world segmentation. Several works strive to reduce the massive computation cost of SAM, such as EfficientSAM [38], FastSAM [44], and MobileSAM [42]. The advent of SAM2 [22] extends SAM's functionality to include video segmentation. While SAM and SAM2 perform effectively in visual segmentation tasks using box, point, or mask prompts, they currently do not possess language understanding capabilities, rendering the direct use of text prompts for referring segmentation infeasible. EVF-SAM [43] extends SAM's capabilities with text-prompted segmentation, while GAVS [30] equips the SAM with image-audio fusion to achieve AVS.

Language-aided Audio-Visual Scene. There is presently a scarcity of publicly available tasks that provide datasets for audio-visual scene comprehension with language-based support. The complexity of language-aided audio-visual scene understanding arises from its reliance on three modalities (vision, audio, and text) necessitating massive matched data and their corresponding labels, posing significant challenges for data collection and model design. Music-AVQA [10] and AVQA [39] are two of the classic audio-visual question-answering (QA) datasets. However, the above work focuses on multimodal QA and spatio-temporal reasoning of audio-visual scenes involving various relationships between objects or activities. Ref-AVS [30, 31] begins pixel-level understanding of dynamic audio-visual scenes.

In summary, no methods exist for integrating the foundational visual segmentation model SAM2 into language-based audio-visual scenes. Therefore, to achieve pixel-level understanding in the LAVS, **we introduce SAM2-LOVE**.

3. Method

3.1. Preliminary of SAM2

SAM2 can be treated as a generalization of SAM to segment anything in images and videos. It also supports visual prompts (point, box, and mask) to indicate the objects to be segmented in an individual frame (image as a single frame). Then, it generalizes across the whole video based on the prompted frame to achieve the video segmentation.

As shown in Figure 3 (Inference), this model begins with a hierarchical image encoder that encodes each frame into multilevel embeddings. Unlike SAM which directly feeds these features to the mask decoder for predictions, SAM2 incorporates a series of transformer blocks (i.e. memory attention) to condition the current image feature on memories of previous and prompted frames stored in a memory bank that is maintained by a FIFO (First In First Out) queue. Each memory entry is composed of two kinds of memories: (1) spatial feature maps that are generated through an element-wise summation of the downsampled and encoded mask features from the memory encoder and the unconditioned frame embedding from the image encoder, and (2) the object pointers that are obtained from the output tokens within the mask decoder contain high-level semantic information of the interested target. Therefore, the memory attention operation in the transformer makes the current frame's feature cross-attend both spatial fine-grained memory and abstract temporal object pointers. Second, the mask decoder of SAM2 largely follows the design of SAM, which is a "two-way" transformer that takes the prompt embeddings from the prompt encoder and frame embeddings as inputs. The difference lies in the additional occlusion score that aims to judge the object's presence in the current frame.

Thanks to the strong and generalized capability of SAM2 across images and videos, we can first compress all the information contained in three modalities into a single token and then use it to prompt SAM2 on a single frame. Finally, starting from this crucial frame, we can propagate through video to obtain the desired masks of all frames.

3.2. SAM2-LOVE

The overall pipeline of our model is shown in Figure 3, which integrates multimodal information in the LAVS into a representative token and uses it to provide a visual clue for SAM2 to segment the reference targets.

Multimodal Representations. Following the previous work [31], for each modality, we use three popular encoders to separately extract their representations, which were not aligned before in their pretraining stages. VGGish [5, 6] encodes an audio clip into audio embeddings $F_A \in R^{N \times d_A}$, where N denotes the duration of the audio in seconds and corresponds to the total number of frames in the video. The sampled N frames are encoded into $F_V \in R^{N \times d_V \times h \times w}$ by a ViT [3] pre-trained on ImageNet. For the text expressions, we utilize DistilRoBERTa [16, 24] to obtain the textual representations $F_T \in R^{P \times d_T}$, where P is the number of textual tokens. To unify and align all embeddings into the same semantic space, we use three MLPs to project them into a shared dimension respectively, shown in Equation (1).

$$\widehat{F}_Z = Linear(\mathrm{ReLU}(Linear(F_Z))), Z \in \{A, V, T\} \quad (1)$$

Multimodal Fusion. As shown in Figure 3, a multimodal fusion module is proposed to compress all three modalities' information into a single token and then leverage its embedding to represent the fused features. We define this token as a learnable segmentation token, $[seg]$, designed to prompt SAM2 as a clue to retrieve the targeted

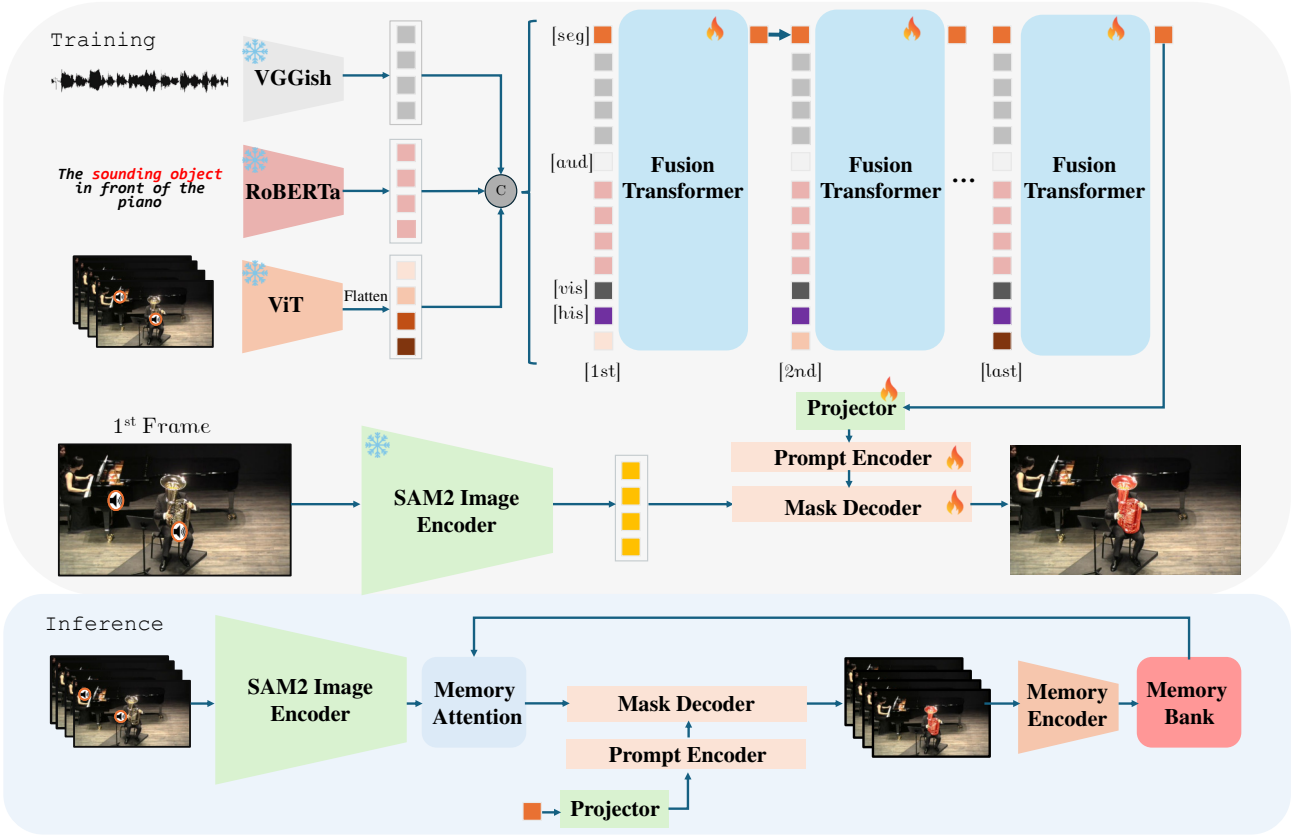


Figure 3. The pipline of SAM2-LOVE. We harness a fusion transformer to compress the tripe modality representations into a learnable token $[seg]$. For the input, $[aud]$ and $[vis]$ are added to indicate their corresponding modality sequences, and $[his]$ is to provide additional information about former frames. In the training stage, we use the first frame to align the SAM2 with the compressed and fused representation of $[seg]$. During inference, we harness the strong zero-shot capabilities of SAM2 to track the referred target at the pixel level.

object. The fusion module is a naive transformer encoder architecture with L stacked bi-directional transformer block that contains a multi-head self-attention (MSA), Layernorm (LN), and MLP, formulated below:

$$\begin{aligned} z'_l &= MSA(LN(z_{l-1})) + z_{l-1}, \quad l = 1...L \\ z_l &= MLP(LN(z'_l)) + z'_l, \quad l = 1...L \end{aligned} \tag{2}$$

where z_{l-1} is the input sequence of l-th layer.

Token Propagation. The behavior of the $[seg]$ token is a self-propagation process within the whole video, which adaptively captures the inner-frame spatial features and the strong continuity among consecutive frames to achieve spatio-temporal consistency. To be more specific, after obtaining the three projected multimodal representations $\widehat{F}_A$, $\widehat{F}_T$, and $\widehat{F}_V$, we first prepend the $[seg]$ token to the head of the sequence of modalities' embeddings to obtain the multimodal sequence F_M^i for the i-th frame, shown as follows:

$$F_M^i = Concat([[seg]; \widehat{F}_A; [aud]; \widehat{F}_T; [vis]; \widehat{F}_V^i]), \tag{3}$$

where $Concat$ denotes the concatenation operation in the feature dimension axis, two fixed tokens, $[aud]$ and $[vis]$ indicate the end and beginning of their corresponding sequences, respectively. This concatenated sequence will be fed into the fusion transformer for updating the status of the $[seg]$ token, allowing the token to capture the visual information of the current frame along with the integral audio and text representations. Since the transformer is the bi-directional encoder that enables all the tokens to interact with each other, we can harness the single token to serve as the multimodal representation of the whole sequence. We take the first element from the output of the fusion transformer $Q_i = Transformer(F_M^i)$, as the updated $[seg]$ token $[seg] = Q_i^0$, shown in Figure 4. Therefore, the token not only contains the current frame information but also the compressed information of the previous frames. Our proposed token propagation strategy allows it to model the video from both the spatial and temporal aspects and aggregate the information within the text and audio.

Token Accumulation. Although the token propagation strategy can effectively obtain the multimodal characteristics and be used to prompt the SAM2, it still faces the issue

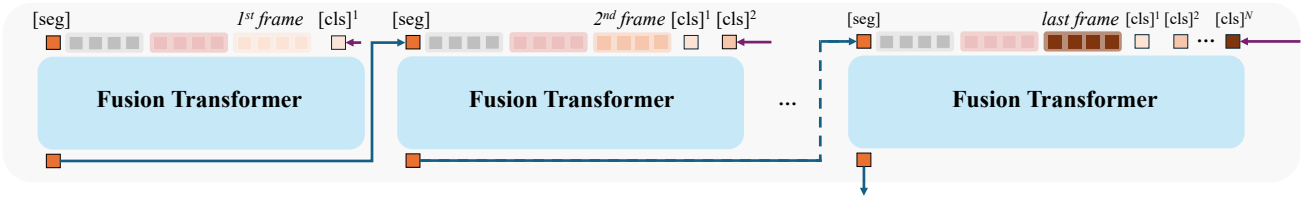


Figure 4. The token propagation and token accumulation strategy in the fusion transformer. For the token propagation, the output $[seg]$ of the current step serves as the input for the next step, which interacts with different frame features but the same audio and text features. To offer additional historical information, the token accumulation strategy appends the global representations of all the previous frames to update the maintained history token $[his]$.

that the information of the first several frames can be diluted and forgotten with the increasing propagation times. To address the issue, we propose the $[cls]$ token accumulation strategy, in which we keep a history of token sequences, $[his]$ to store the global and abstract representation of historical frames, shown in Figure 4. It is achieved that when propagating to the i-th frame, we provide additional global information of all the previous frames by appending and concatenating the $[his]$ tokens into F_M^i. The updation of the history token $[his]$ can be formulated by Equation (4).

$$[his]^i = Concat([[cls]^0; [cls]^1; ...; [cls]^i]) \quad (4)$$

where $[cls]^i$ is the global representation of i-th frame from ViT. These historical tokens can be referred to by the $[seg]$ of the current stage, promising the information of all the previous frames is not lost.

Unified Perspective of Token Behaviors. We interpret the two token operations described above from a unified knowledge transfer perspective. Token propagation of the $[seg]$ token is a direct flow, which expects the knowledge from different frames to be gradually connected through the propagation process. This is the **forward** knowledge transfer. In contrast, token accumulation is the replay of $[his]$ tokens, a reversed flow to prevent the $[seg]$ token from forgetting the knowledge of historical frames as it propagates forward. This is the **backward** knowledge transfer. We propose that these two token operations are functionally opposed yet conceptually unified.

3.3. Training

Trainable Modules. The encoders used for multimodal representations for audio, video, and text are fully frozen, perceiving the pre-trained knowledge. However, the multimodal fusion transformer and $[seg]$ token are fully trainable during our training process, allowing them to learn how to generate the fused and aligned embeddings tailored for prompting SAM2. For SAM2, we keep the image encoder and memory attention modules frozen while training the prompt encoder and mask decoder. Additionally, all the MLPs for projecting are trainable.

Training Strategy. Our training strategy initially activates SAM2's multimodal prompting capability at the image level and then leverages its extensive pre-trained knowledge for zero-shot video segmentation. In this setting, the fusion transformer and $[seg]$ token are trained with full access to the entire video sequence, while SAM2 is exposed only to the first frame during training. Specifically, we consolidate multimodal information (video, audio, and text) through the fusion module, which compresses these features into tokens that serve as prompts for SAM2 on the first frame, shown in Figure 3 (Training). Then, a supervised loss is applied to this frame to train SAM2's referring segmentation abilities. During testing, both the fusion module and SAM2 receive the entire video, allowing for comprehensive segmentation across all frames. This approach enables SAM2 to effectively utilize multimodal prompts on the initial frame and then generalize to the complete video sequence through its zero-shot Video Object Segmentation (VOS) capabilities, shown in Figure 3 (Inference).

Loss Function. The model is trained end-to-end under the supervised mask loss $\mathcal{L}_{mask}$, to encourage the SAM2 to predict the correct localized masks. To compute this loss, we combine the per-pixel binary cross-entropy (BCE) loss and DICE loss with their corresponding weights λ_{bce} and λ_{dice}, which can be formulated by:

$$\mathcal{L}_{mask} = \lambda_{bce}\textbf{BCE}(\hat{M}, M) + \lambda_{dice}\textbf{DICE}(\hat{M}, M) \quad (5)$$

where $\hat{M}$ is a single mask prediction for the first frame and M is the label of the first frame for supervision.

4. Experiments

4.1. Experimental Settings

Dataset. We evaluate our method on the Ref-AVS benchmark [31], a substantial collection of 4,000 videos with manual pixel-level annotations and 20,000 expressions reflecting the scenarios in audio, vision, and time dimensions. Additionally, the collected videos contain abundant and diverse visual elements. Each video encompasses a high average number of objects about 1.72, a wide variety of audible objects in 48 categories, and 3 kinds of static objects.

Method	Seen (%)			Unseen (%)			Mix (S+U)			NULL	VENUE
	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	s	
AVSBench [45]	21.2	45.7	33.5	27.7	50.9	39.3	24.5	48.3	36.4	-	ECCV'2022
+text	23.2	51.1	37.2	32.4	54.7	43.5	27.8	52.9	40.3	0.208	
AVSegFormer [4]	29.2	42.3	35.7	32.7	45.1	38.9	30.9	43.7	37.3	-	AAAI'2024
+text	33.5	47.0	40.2	36.1	50.1	43.1	34.8	48.6	41.7	0.171	
GAVS [28]	24.8	45.5	35.2	27.8	46.0	36.9	26.3	45.7	36.0	-	AAAI'2024
+text	28.9	49.8	39.4	29.8	49.7	39.8	29.4	49.8	39.6	0.190	
ReferFormer [35]	27.4	47.7	37.5	24.9	50.4	37.6	26.2	49.1	37.6	-	CVPR'2022
+audio	31.3	50.1	40.7	30.4	48.8	39.6	30.9	49.5	40.2	0.176	
R2VOS [11]	21.3	30.9	26.1	23.9	43.1	33.5	22.6	37.0	29.8	-	ICCV'2023
+audio	25.0	41.0	33.0	27.9	49.8	38.9	26.5	45.4	35.9	0.183	
EEMC [31]	34.2	51.3	42.8	49.5	64.8	57.2	41.9	58.1	50.0	**0.007**	ECCV'2024
SAM2-LOVE (ours)	**43.5**	**51.9**	**47.7**	**66.5**	**72.3**	**69.4**	**55.0**	**62.1**	**58.5**	0.23	CVPR2025
vs. prev. SOTA	**+9.3**	**+0.6**	**+4.9**	**+17.0**	**+7.5**	**+12.2**	**+13.1**	**+4.0**	**+8.5**	-	-

Table 1. Performance comparison across different methods in Seen, Unseen, Mix (S+U), and NULL settings of Ref-AVS benchmark. The mix indicates the average value of seen and unseen splits.

Therefore, the tested scenarios are more challenging with multiple sound sources and multiple semantics. The dataset is divided into three splits, a training set of 2908 videos, a validation set of 276 videos, and a test set of 818 videos. The test set is further divided into three distinct subsets: 1) *seen splits* that are composed of the same 39 categories in the training set. 2) *unseen splits* with 13 additional categories that do not appear in the training stage. 3) *null splits* to test the model when an expression refers to an object that either does not exist or is not visible in the given context.

Baselines. We compare our method with six challenging baselines: AVSBench [45], AVSegFormer [4], GAVS [28], ReferFormer [35], R2VOS [11] and EEMC [31]. AVSBench, AVSegFormer, and GAVS are designed for Audio-Visual Segmentation (AVS) tasks while ReferFormer and R2VOS claim for Referring Video Object Segmentation(Ref-VOS) tasks. We also report the results achieved by offering the third their absent modalities, which is consistent with EEMC. EEMC is specially designed to achieve fine-grained semantic understanding in the LAVS and also proposes the Ref-AVS task.

Implementation Details. For the training recipe, we adopt the DeepSpeed training engine with ZeRO-2 to train our model on 2 NVIDIA A100 GPUs. The AdamW [19] optimizer is harnessed with the learning rate and weight decay set to 1e-4 and 0, respectively. We also adopt WarmupDecayLR as the learning rate scheduler, where the warmup iterations are set to 100. The weights of the BCE loss λ_{bce} and Dice loss λ_{dice} are set to 1.0 and 1.0, respectively. The batch size is set to 8 and the gradient accumulation step is set to 2. For the network architecture, the fusion transformer is the stacked 6 transformer blocks with bi-directional self-attention. The hidden dimension of the transformer is 1024 and the number of $[seg]$ tokens is set to 1. SAM2 is initialized with the weight of SAM2-Hiera-large.

Metrics. Following [20, 31], we assess the performance of the Ref-AVS benchmark using the Jaccard index ($\mathcal{J}$), the F-score ($\mathcal{F}$), and their average value ($\mathcal{J}\&\mathcal{F}$) as key evaluation metrics.

4.2. Main Results

Table 1 shows the comparative results of our method on Ref-AVS benchmarks. Our SAM2-LOVE achieves state-of-the-art (SOTA) performance and significantly outperforms the second-best method EEMC on both seen and unseen test splits. The results indicate that our model has highly generalized capabilities that can be directly used for out-of-domain scenarios. We attribute two reasons leading to these huge performance improvements. Firstly, the fusion transformer can effectively step-by-step compress the multimodal cues into the learnable tokens through the $[seg]$ token propagation strategy and $[cls]$ token accumulation. The generated token contains high semantics as well as positional information, which can accurately prompt SAM2 to segment the inserted targets. Secondly, SAM2 showcases strong zero-shot VOS capabilities, accurately tracking the referred object despite the circumstance where it suddenly disappears and occurs again or similar objects have interacted together.

4.3. Quality Assessment

Figure 5 shows the visualization comparative results with the previous SOTA EEMC. In the first case, we can observe that the SAM2 shows strong consistency within different frames, which can track the referred object with high



Figure 5. The visualization results of the referred objects in the Ref-AVS. Notably, we use the legends of the trumpet to represent the different sound volumes of the audio objects, from silent to increasing loudness. More cases can be found in the supplementary.

CLIP-T CLIP-V	RoBERTa ViT	$[cls]$ Acc	Seen (%)			Unseen (%)		
			$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
✓		✓	42.5	51.2	46.8	65.1	71.4	68.2
	✓	✓	**43.5**	**51.9**	**47.7**	**66.5**	**72.3**	**69.4**
	✓		42.6	50.9	46.7	66.3	71.6	69.0

Table 2. Ablations on backbone design and our proposed $[cls]$ token accumulation strategy.

Number of transformer blocks	Seen (%)			Unseen (%)		
	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
1 layer	41.5	49.4	45.4	65.2	70.9	68.1
6 layers	**42.6**	**50.9**	**46.7**	66.3	71.6	69.0
12 layers	42.0	50.7	46.4	**67.7**	**73.3**	**70.5**

Table 3. Ablations on the number of fusion transformer blocks.

mask quality. Additionally, it also demonstrates that the $[seg]$ token can accurately locate the small referring target, while EEMC fails in the first several frames. The second case shows that our SAM2-LOVE is more robust regarding changes in audible objects. EEMC may not distinguish the difference or malfunction in understanding the hidden meaning in the expressions (i.e. "shorter duration" instead of the obvious "the man").

4.4. Ablation Study

Effect of Backbone. Table 2 shows the ablation study of backbones. Compared to CLIP vision and text encoder, RoBERTa and ViT achieve better performance, which can be the difference from the pre-trained paradigm. To align with visual elements, the CLIP textual encoder focuses more on the text with specific semantics such as cat and person. However, in the LAVS scenes, the massive expressions tend to provide a kind of control signal without referred category names, e.g., " the one making the sound." and " the one making the most intense sound." Therefore, the RoBERTa may demonstrate greater strength in understanding instructions due to expertise in natural language understanding during training.

Effect of $[cls]$ Token Accumulation. Since we harness the propagated token to prompt the first frame for SAM2, the attention for the first several frames may weaken during propagation, causing the token to depend only on the last frame. Therefore, in the fusion transformer, we attach the visual global $[cls]$ tokens of previous frames to the current frame to update the history token, avoiding the $[seg]$ token forgetting the information of former frames. Compared to the last row and second row in Table 2, we can observe that this strategy can effectively improve performance for free without additional design or resource consumption.

Effect of Fusion Transformer Design. The design of

Number of $[seg]$ tokens	Seen (%)			Unseen (%)		
	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
1	**42.6**	**50.9**	**46.7**	66.3	71.6	69.0
4	41.3	50.8	46.0	65.2	71.0	68.1
8	40.7	49.8	45.2	**67.0**	**72.2**	**69.6**

Table 4. Ablations on the numbers of $[seg]$ tokens.

Fusion Strategy	Seen (%)			Unseen (%)		
	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
learnable token	**42.6**	**50.9**	**46.7**	**66.3**	**71.6**	**69.0**
mean	41.6	50.1	45.8	65.9	71.6	68.7

Table 5. Ablations on the different fusion strategies for generating the prompt token for SAM2.

our fusion transformer is a naive transformer architecture without additional operations such as ROPE [26] or attention enhancement [25]. Table 3 shows the different performances under different numbers of transformer blocks. Increasing the model's volume can lead to gains in the unseen test set and comparable performance in the seen test set. The single layer can also achieve satisfying results, demonstrating that the simple multimodal fusion still makes a huge impact. However, considering the additional resource consumption and balance of efficiency, we use 6 layers of transformer instead of 12.

Effect of the Number of $[seg]$ Tokens. We only use a single token to promote SAM2 for localization of the correct target, but the number of tokens entering the propagation process can be multiple. As shown in Table 4, increasing the number of tokens can improve significant performances in the unseen split, while this sacrifices the performances of the seen category. We think that it fits the insight from LIBERO [12]. The additional tokens can be regarded as kinds of registers that store additional global information, and this prior knowledge would facilitate learning on new tasks (Unseen tasks). On the other hand, utilizing a single token implicitly encodes global information into that token, which can enhance performance on prior tasks (Seen tasks) but could negatively impact performance on new tasks (Unseen tasks). Despite the representation learned by a single token exhibiting a stronger inductive bias, it is still useful and we set it by default.

Effect of Fusion Strategy. The token used to prompt the SAM2 can be generated by different strategies. Besides using a learnable token, we can directly take the mean of whole the sequence that is generated by concatenating all the frames' embeddings with those of audio and text. As shown in Table 5, our learnable token can behave better since it can aggregate information from different frames gradually. In contrast, the direct mean operation may lose a large amount of inter-frame information.

Mask decoder	Prompt encoder	Seen (%)		
		$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
*	*	34.8	44.8	39.8
✓	*	42.2	50.7	46.5
✓	✓	**42.5**	**50.9**	**46.7**

Table 6. Ablations on the different training modules in SAM2. "$*$" denotes the frozen module while "✓" makes the module trainable.

Effect of Training Modules in SAM2. To preserve the strong VOS capabilities of SAM2, we freeze the memory attention modules but train the prompt encoder and mask decoder to further adapt the SAM2 to the multimodal clue reference, which can achieve the best results shown in Table 6. Without training the mask decoder, it will cause a huge performance drop due to the low capability to fit the multimodal representation of the $[seg]$ token.

4.5. Limitations

In this section, we discuss the limitations of our method. As shown in Table 1 NULL split, our method has a relatively low capability to reject the users' requirements for a null reference. It is the inherent limitation of the prompting paradigm for localization that SAM2 attends to segment something or some irrelevant regions in the image without any meaning. SAM2-LOVE lags behind the EEMC in the null test with 0.2 in S score, achieving 0.23 similar results with the *AVSBench plus text* method. The S denotes the square root of the ratio of the number of foreground pixels over the background pixels. This can be our future work on how to solve the null reference problem.

5. Conclusion

This paper introduces SAM2-LOVE, a novel framework for segmenting reference objects in Language-aided Audio-Visual Scenes, overcoming the limitations of previous dual-modality SAM approaches in handling scenarios that integrate three modalities. We propose a multimodal fusion module that compresses information from all modalities into a single token, which then prompts SAM2 for video segmentation. This compressed $[seg]$ token propagates forward through the video to capture both the spatial and temporal features, while also referring to the accumulated historical tokens backward preserving information from previous frames. Extensive experiments demonstrate the superiority of our model on the Ref-AVS benchmark and highlight the efficiency of our design. Future work will focus on addressing scenarios where the three modality references are misaligned, such as the null reference problem.

Acknowledgement

This work was supported in part by National Natural Science Foundation of China (Grant No. 62206153), and in part by Guangdong Natural Science Funds for Distinguished Young Scholar (No. 2025B1515020012).

References

[1] Sule Bai, Yong Liu, Yifei Han, Haoji Zhang, and Yansong Tang. Self-calibrated clip for training-free open-vocabulary segmentation. *arXiv preprint arXiv:2411.15869*, 2024. 2

[2] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26497–26507, 2024. 3

[3] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[4] Shengyi Gao, Zhe Chen, Guo Chen, Wenhai Wang, and Tong Lu. Avsegformer: Audio-visual segmentation with transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12155–12163, 2024. 3, 6

[5] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing*, pages 776–780. IEEE, 2017. 3

[6] 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 *2017 IEEE international conference on acoustics, speech and signal processing*, pages 131–135. IEEE, 2017. 3

[7] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing*, pages 787–798, 2014. 2

[8] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2, 3

[9] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589, 2024. 2

[10] Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19108–19118, 2022. 1, 3

[11] Xiang Li, Jinglu Wang, Xiaohao Xu, Xiao Li, Bhiksha Raj, and Yan Lu. Robust referring video object segmentation with cyclic structural consensus. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22236–22245, 2023. 6

[12] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36, 2024. 8

[13] Chang Liu, Henghui Ding, and Xudong Jiang. GRES: Generalized referring expression segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2

[14] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 2

[15] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 2

[16] Yinhan Liu. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364, 2019. 3

[17] Yong Liu, Sule Bai, Guanbin Li, Yitong Wang, and Yansong Tang. Open-vocabulary segmentation with semantic-assisted calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3491–3500, 2024. 2

[18] Yong Liu, Cairong Zhang, Yitong Wang, Jiahao Wang, Yujiu Yang, and Yansong Tang. Universal segmentation at arbitrary granularity with language instruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3459–3469, 2024. 2

[19] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 6

[20] Zhuoyan Luo, Yicheng Xiao, Yong Liu, Shuyan Li, Yitong Wang, Yansong Tang, Xiu Li, and Yujiu Yang. Soc: Semantic-assisted object cluster for referring video object segmentation. *Advances in Neural Information Processing Systems*, 36, 2024. 6

[21] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling context between objects for referring expression understanding. In *IEEE European Conference on Computer Vision*, pages 792–807. Springer, 2016. 2

[22] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 2, 3

[23] Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. Pixellm: Pixel reasoning with large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26374–26383, 2024. 2

[24] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller,

faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. 3

[25] Zhuoran Shen, Mingyuan Zhang, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Efficient attention: Attention with linear complexities. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3531–3539, 2021. 8

[26] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 8

[27] Luoyi Sun, Xuenan Xu, Mengyue Wu, and Weidi Xie. Auto-acd: A large-scale dataset for audio-language representation learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 5025–5034, 2024. 3

[28] Yaoting Wang, Weisong Liu, Guangyao Li, Jian Ding, Di Hu, and Xi Li. Prompting segmentation with sound is generalizable audio-visual source localizer. *arXiv preprint arXiv:2309.07929*, 2023. 6

[29] Yuji Wang, Ruojun Zhao, and Zijun Sun. Efficient remote sensing transformer for coastline detection with sentinel-2 satellite imagery. In *IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium*, pages 5439–5442. IEEE, 2023. 2

[30] Yaoting Wang, Weisong Liu, Guangyao Li, Jian Ding, Di Hu, and Xi Li. Prompting segmentation with sound is generalizable audio-visual source localizer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5669–5677, 2024. 2, 3

[31] Yaoting Wang, Peiwen Sun, Dongzhan Zhou, Guangyao Li, Honggang Zhang, and Di Hu. Ref-avs: Refer and segment objects in audio-visual scenes. *IEEE European Conference on Computer Vision*, 2024. 1, 2, 3, 5, 6

[32] Yuji Wang, Ruojun Zhao, Shicai Wei, Jingchen Ni, Meng Wu, Yang Luo, and Chunbo Luo. Convolution meets transformer: Efficient hybrid transformer for semantic segmentation with very high resolution imagery. In *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*, pages 9688–9691. IEEE, 2024. 2

[33] Yuji Wang, Jingchen Ni, Yong Liu, Chun Yuan, and Yansong Tang. Iterprime: Zero-shot referring image segmentation with iterative grad-cam refinement and primary word emphasis. *arXiv preprint arXiv:2503.00936*, 2025. 2

[34] Dongming Wu, Tiancai Wang, Yuang Zhang, Xiangyu Zhang, and Jianbing Shen. Onlinerefer: A simple online baseline for referring video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2761–2770, 2023. 2

[35] Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4974–4984, 2022. 6

[36] Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4974–4984, 2022. 2

[37] Zhuofan Xia, Dongchen Han, Yizeng Han, Xuran Pan, Shiji Song, and Gao Huang. Gsva: Generalized segmentation via multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3858–3869, 2024. 2

[38] Yunyang Xiong, Bala Varadarajan, Lemeng Wu, Xiaoyu Xiang, Fanyi Xiao, Chenchen Zhu, Xiaoliang Dai, Dilin Wang, Fei Sun, Forrest Iandola, et al. Efficientsam: Leveraged masked image pretraining for efficient segment anything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16111–16121, 2024. 3

[39] Pinci Yang, Xin Wang, Xuguang Duan, Hong Chen, Runze Hou, Cong Jin, and Wenwu Zhu. Avqa: A dataset for audio-visual question answering on videos. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3480–3491, 2022. 1, 3

[40] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2

[41] Zhao Yang, Jiaqi Wang, Xubing Ye, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Language-aware vision transformer for referring segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2

[42] Chaoning Zhang, Dongshen Han, Yu Qiao, Jung Uk Kim, Sung-Ho Bae, Seungkyu Lee, and Choong Seon Hong. Faster segment anything: Towards lightweight sam for mobile applications. *arXiv preprint arXiv:2306.14289*, 2023. 3

[43] Yuxuan Zhang, Tianheng Cheng, Rui Hu, Lei Liu, Heng Liu, Longjin Ran, Xiaoxin Chen, Wenyu Liu, and Xinggang Wang. Evf-sam: Early vision-language fusion for text-prompted segment anything model. *arXiv preprint arXiv:2406.20076*, 2024. 2, 3

[44] Xu Zhao, Wenchao Ding, Yongqi An, Yinglong Du, Tao Yu, Min Li, Ming Tang, and Jinqiao Wang. Fast segment anything. *arXiv preprint arXiv:2306.12156*, 2023. 3

[45] Jinxing Zhou, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, and Yiran Zhong. Audio-visual segmentation. In *European Conference on Computer Vision*, 2022. 3, 6

[46] Jinxing Zhou, Xuyang Shen, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, and Yiran Zhong. Audio-visual segmentation with semantics. *arXiv preprint arXiv:2301.13190*, 2023. 3

[47] Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, et al. Generalized decoding for pixel, image, and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15116–15127, 2023. 2