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

Supplementary Material

6. Additional qualitative results

Fig. 4 presents additional segmentation results from TSAM on the Ref-AVS test set, emphasizing moving objects to showcase its temporal modeling capabilities. In the top example, TSAM demonstrates its strong performance on the Unseen test set, accurately segmenting the target object "*car*" despite the absence of audio cues in the expression. In the middle and bottom examples, the same target object "*boat*" appears in different scenes from the Seen test set. The middle example incorporates visual cues ("*entity moving*") in the expression, while the bottom example relies solely on audio cues ("*sounding object*"). Overall, these examples validate TSAM's ability to segment objects in dynamic audio-visual scenes, demonstrating its strength in processing target objects referred to in textual expressions, regardless of the presence of visual or audio cues.

Figure 4. TSAM's segmentation of moving objects, highlighting temporal dynamics. Top row: ground truth with target object in green. Bottom row: TSAM results with target object in violet.

In contrast, Fig. 5 illustrates cases where TSAM fails to segment target objects referred to in the expression. In the top example, the scene contains two clarinets, with the target object specified as "*clarinet in front of the man*". However, another clarinet is also positioned in front of the man, but with a woman, and both clarinets produce the same sound. This similarity in sound hinders TSAM from distinguishing the target clarinet, leading to the segmentation of the wrong clarinet. In the middle example, the failure arises from the implicit reference to "*sheep*" in the phrase "*between the two donkeys,*" which challenges TSAM's ability to align the textual cues with the visual cues. Additionally, the accompanying sound corresponds to donkeys, further complicating the alignment between the text and audio. In the bottom example, the failure occurs due to misannotation of the target object as "*wolf*" instead of "*dog*", while the accompanying sound is consistent with a dog. It is important to highlight that the middle and bottom examples are from the Unseen test set, emphasizing the challenges TSAM encounters with novel, ambiguous references and mismatches between sound and target objects.

7. Training objective

To improve mask quality during TSAM training, we utilized binary cross-entropy $\mathcal{L}_{\text{BCE}}$ and intersection over union $\mathcal{L}_{\text{IoU}}$ losses, as defined in Eq. (5) in the main paper. The parameter λ is set to 1.0 to balance these losses. The effect of omitting $\mathcal{L}_{\text{IoU}}$ is analyzed in Tab. 2 in the main paper, while the impact of varying λ is explored in Figs. 6 and 7. Fig. 6 and Fig. 7 illustrate the impact of varying λ on segmentation performance in terms of the $\mathcal{J}$ and $\mathcal{F}$ metrics, respectively, across the Seen and Unseen test sets of the Ref-AVS dataset.

Fig. 6 illustrates that increasing λ improves performance on the Unseen test set, while its effect on the Seen test set remains minimal. This trend suggests that incorporating the $\mathcal{L}_{\text{IoU}}$ loss helps refine segmentation masks, particularly for novel, challenging objects in unseen scenarios. Particularly, the Unseen test set shows a consistent upward trend as λ increases from 0.4 to 1.0, highlighting that a stronger emphasis on $\mathcal{L}_{\text{IoU}}$ enhances the TSAM's ability to generalize, producing more accurate and spatially coherent masks for previously unseen objects.

Fig. 7 reinforces this observation, showing comparable trends in the $\mathcal{F}$ metric, further validating TSAM's ability to enhance segmentation quality by leveraging multimodal cues and optimizing spatial consistency. The results demonstrate that TSAM maintains robust performance on the Seen

test set while effectively adapting to challenging unseen objects through a well-calibrated balance of loss components.

Figure 5. TSAM's failure segmentation results. Top row: ground truth with target object in green. Bottom row: TSAM results with target object in violet.

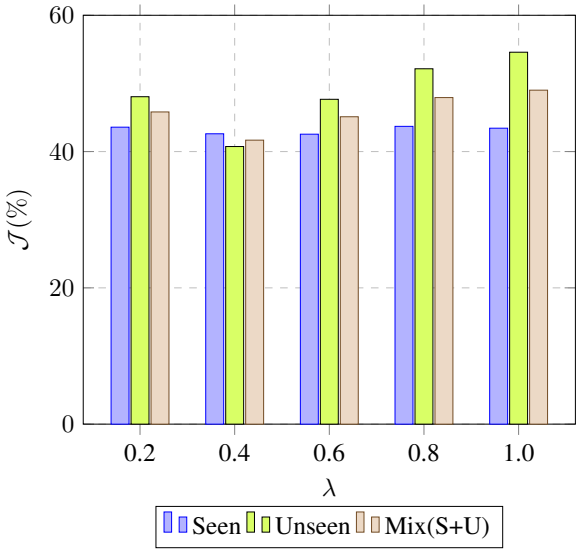


Figure 6. Segmentation performance ($\mathcal{J}$) with varying λ.

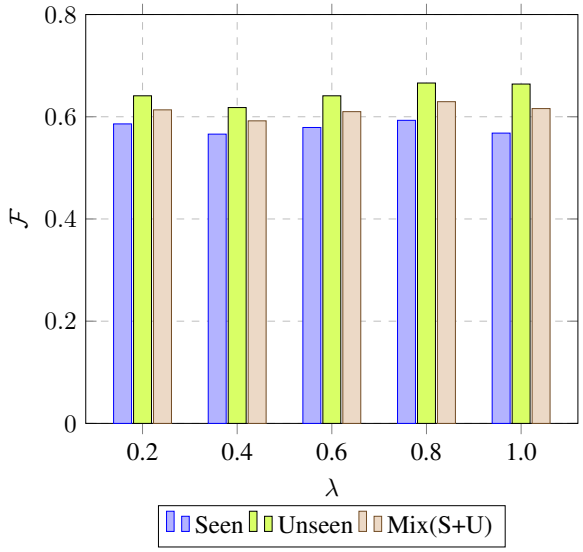
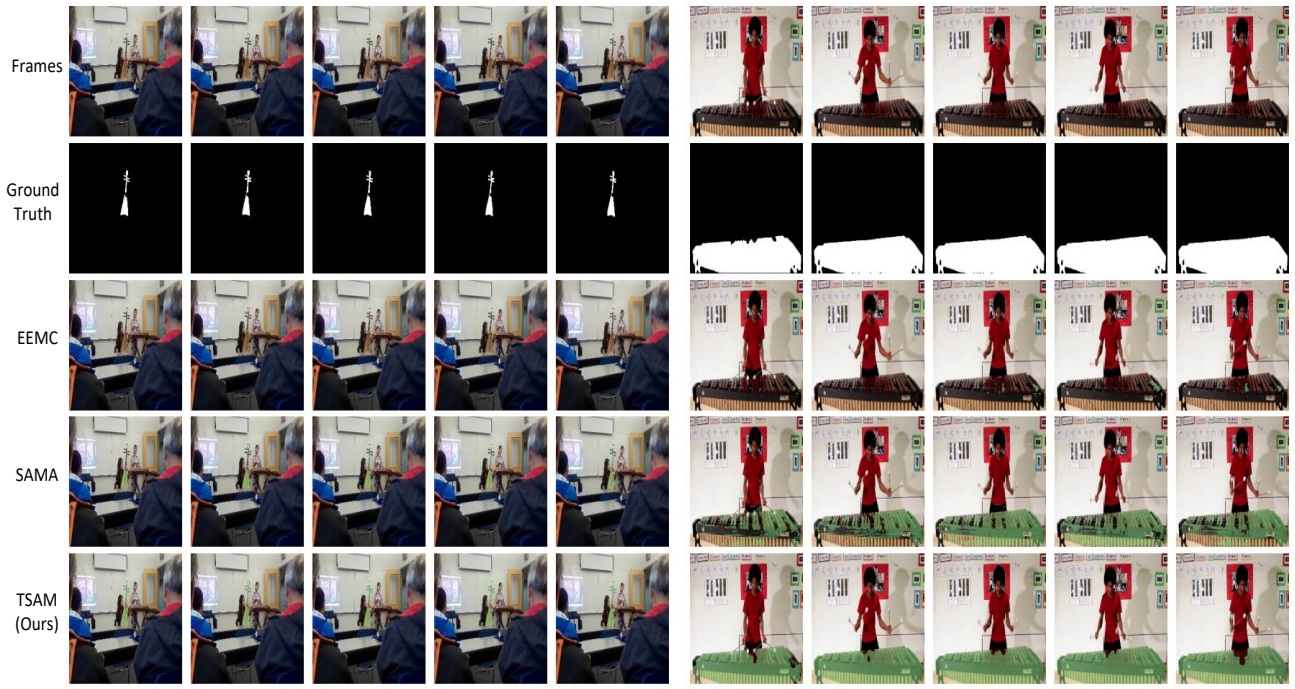


Figure 7. Segmentation performance ($\mathcal{F}$) with varying λ.

8. Qualitative comparison

Fig. 8 shows a qualitative comparison of TSAM with state-of-the-art methods, EEMC [36] and SAMA [24], on the Ref-AVS dataset. The visualizations highlight the limitations of EEMC, which consistently struggles to accurately segment the object specified in the expression. The SAMA method demonstrates some improvement, particularly when the expression explicitly describes a sounding object, as shown in Fig. 8 (a). On the contrary, SAMA struggles to fully understand nuanced expressions and often defaults for segmenting all sounding objects in the scene, as shown in Fig. 8 (b). This limitation is further demonstrated in Fig. 8 (c). In the left example, the SAMA method incorrectly segments a sounding object, "*baby*", instead of the target object, "*couch*", while in the right example, it completely fails to detect the target object, "*table*".

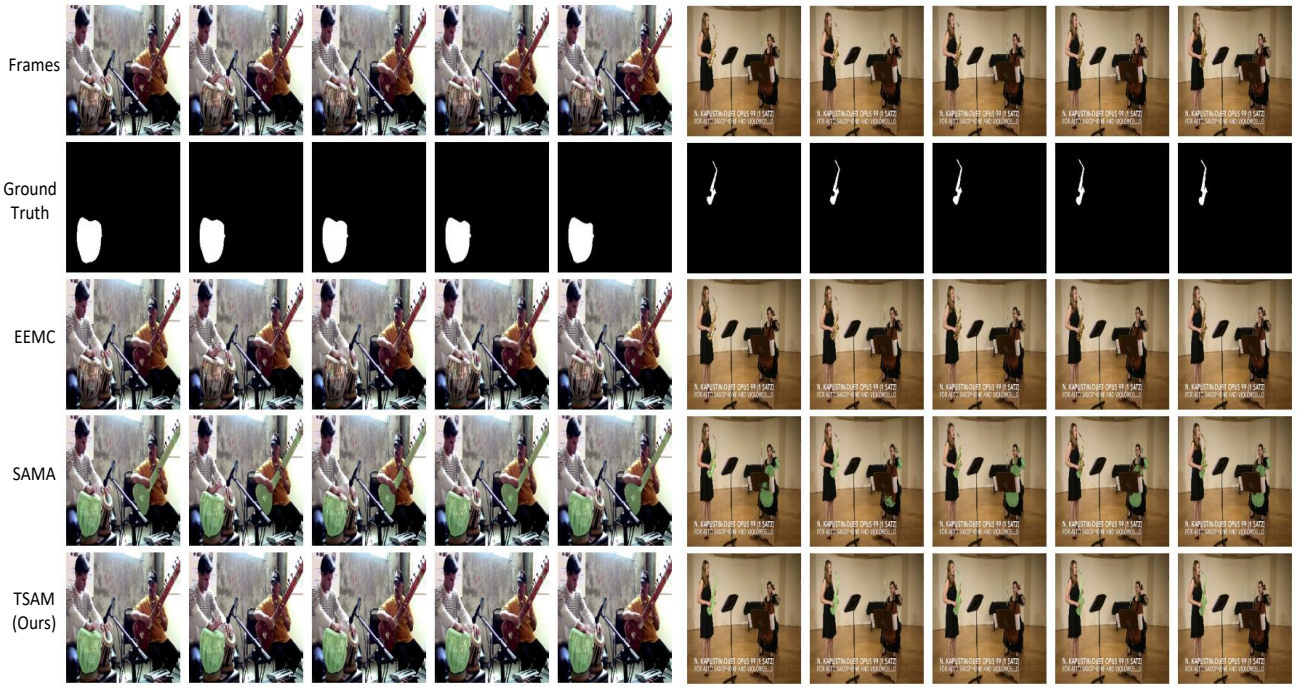
On the other hand, TSAM demonstrates a superior ability to process multimodal cues and dynamic audio-visual scenes, allowing it to accurately interpret textual expressions and segment the target objects. For example, Fig. 8 (a) (left) highlights TSAM's effectiveness in segmenting target objects described in the expression, even when they are small or distant from the camera. Fig. 8 (b) (right) further showcases TSAM's capability to isolate the target object described in the expression while ignoring irrelevant sounding objects in the scene. Additionally, Fig. 8 (c) underscores TSAM's robustness in handling more complex cases, such as segmenting non-sounding objects like "*couch*" (left) and "*table*" (right), even when "*table*" is not explicitly referenced in the expression. These results highlight TSAM's ability to leverage multimodal cues and seamlessly align with target objects in dynamic audio-visual scenes.



Expression: The object making a sound by being played by the woman in black.

Expression: The sounding object.

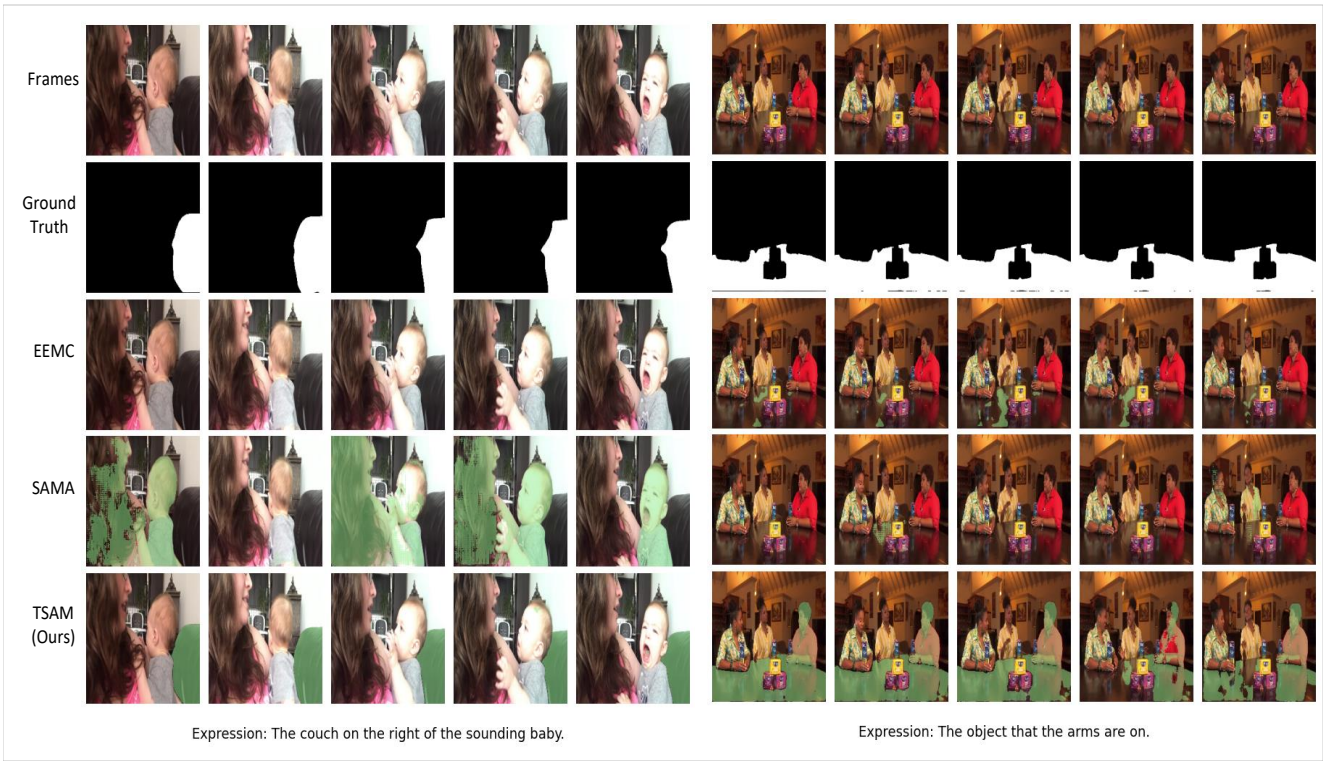
(a)



Expression: The sounding object far from the sitar.

Expression: Controlled by the woman on the left_the object sounded.

(b)



(c)

Figure 8. Qualitative comparison of segmentation results on the Ref-AVS test set between TSAM and the state-of-the-art methods, EEMC [36] and SAMA [24].