Supplementary Material: TAViS: Text-bridged Audio-Visual Segmentation with Foundation Models

Ziyang Luo[1] Nian Liu[2,*] Xuguang Yang[1] Salman Khan[2] Rao Muhammad Anwer[2]
Hisham Cholakkal[2] Fahad Shahbaz Khan[2] Junwei Han[1,3]
[1]Northwestern Polytechnical University [2]Mohamed bin Zayed University of Artificial Intelligence
[3] Institute of Artificial Intelligence, Hefei Comprehensive National Science Center

1. Additional Implementation Details

In the ImageBind-guided query decomposition, we configure all cross-attention layers to 4. Due to computational limitations, the number of queries is restricted to 5. For the audio-text dual prompt, the audio embedding is first transformed to predict 6 tokens, which are then padded to 77-token sequences, consistent with the original ImageBind implementation. These padded sequences are subsequently processed by the ImageBind encoder.

2. More Ablation Study

Due to space constraints, additional ablation studies on our design are provided in the supplementary material. All ablation studies are conducted using 224×224 image resolution on the S4 and MS3 datasets, following the methodology outlined in the main text.

2.1. Impact of Text Templates Design

Text serves as a vital bridge between visual and audio modalities. To explore its impact, we conduct extensive ablation studies on different text format designs. Inspired by CLIP [4], we utilize modality-specific text templates: "An image of [cls]" for visual inputs and "A sound of [cls]" for audio inputs. While these templates explicitly encode modality-specific information, potentially aiding the model in capturing relationships within each modality, they risk disrupting the unified feature space crucial to our approach. As shown in Table 1, modality-specific templates result in lower performance compared to universal text templates, emphasizing the importance of maintaining a shared representation space across all modalities.

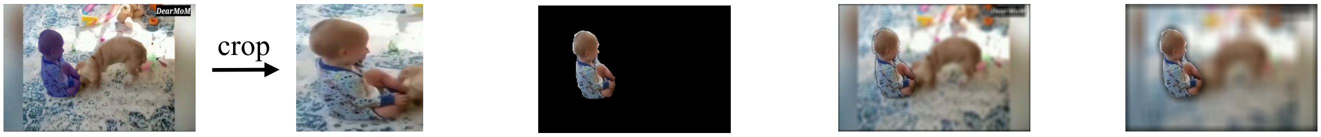
Settings	S4		MS3	
	$\mathcal{M}_\mathcal{J}$	$\mathcal{M}_\mathcal{F}$	$\mathcal{M}_\mathcal{J}$	$\mathcal{M}_\mathcal{F}$
separate text	84.1	0.910	64.3	0.722
universal text	84.8	0.912	68.2	0.759

Table 1. **Ablation studies of our TAViS for text template designs.** "Separate text" indicates that we use "An image of [cls]" and "A sound of [cls]" as text templates for the visual and audio modalities, respectively. For "universal text," we use the format "A [cls]" for both visual and audio modalities.

2.2. Effect of Region Preprocessing Strategies

Previous studies in open-vocabulary semantic segmentation [2, 5] have explored various image preprocessing strategies when feeding images into the encoder for class prediction. Following this line of research, we investigate the most appropriate approach for our model. Specifically, we experiment with three strategies: (1) cropping and resizing the predicted region to the initial size; (2) masking the background in black while preserving the foreground; and (3) blurring the background as described in our main methodology.

As shown in Table 2, blurring the background yields the best performance, as it highlights the foreground information while preserving contextual environmental cues. In contrast, cropping the foreground or setting the background to black compromises ImageBind's capability by eliminating important environmental context. Furthermore, we analyze the impact of different kernel sizes in Gaussian blur. As demonstrated in Table 2, a kernel size that is too small allows the encoder to still



(a) crop the foreground (b) black background (c) kernal=25 (d) kernal=75

Figure 1. **Illustration of our region preprocessing strategies.**

recognize background objects, which interferes with the model's ability to align modality-specific information for the target objects. All example designs are presented in Figure 1.

Settings	S4		MS3	
	$\mathcal{M}_{\mathcal{J}}$	$\mathcal{M}_{\mathcal{F}}$	$\mathcal{M}_{\mathcal{J}}$	$\mathcal{M}_{\mathcal{F}}$
crop	84.4	0.907	64.5	0.711
background black	83.4	0.905	63.8	0.706
background blur	84.8	0.912	68.2	0.759
kernal=25	84.0	0.905	66.5	0.733
kernal =75	84.8	0.912	68.2	0.759

Table 2. **Ablation studies of our TAViS for region preprocessing strategies.**

2.3. IBQD Design

To assess the effectiveness of our IBQD module, which processes audio *cls* tokens and features, we performed experiments comparing it against a baseline model that uses only $\boldsymbol{t}_W$ as prompts and for subsequent alignment. As shown in Table 3, our proposed design largely preserves a well-aligned feature space. However, since the IBQD module is trained between the audio encoder and the projection layer, it introduces a potential risk of overfitting to training classes, potentially hindering zero-shot generalization. To demonstrate the importance of our full two-stage mechanism, we evaluated a variant where the second cross-attention step is omitted, using the $\boldsymbol{t}'_a$ queries immediately . This simplified approach resulted in a performance drop, highlighting the necessity of the query refinement stage.

Settings	S4		MS3	
	$\mathcal{M}_{\mathcal{J}}$	$\mathcal{M}_{\mathcal{F}}$	$\mathcal{M}_{\mathcal{J}}$	$\mathcal{M}_{\mathcal{F}}$
based on $\boldsymbol{t}_W$	84.4	0.913	63.3	0.708
only add t_a	83.6	0.905	66.2	0.739
IBQD	84.8	0.912	68.2	0.759

Table 3. **Ablation studies of our TAViS for IBQD design.**

2.4. Segmentation Loss Design.

The AVS task primarily involves two types of predictions: binary masks for S4 and MS3 datasets, and object-based masks for the AVSS dataset. To address these distinct outputs, we design two specific segmentation loss functions. Our ablation studies show that removing either loss function leads to performance degradation across all settings. Specifically, $\mathcal{L}_{binary}$ ensures the overall accuracy of target segmentation, while $\mathcal{L}_{sep}$ facilitates precise object-wise segmentation. We believe this is because the self-attention mechanism in mask decoder enables effective information exchange between objects, further enhancing the overall performance.

Settings	S4		MS3	
	$\mathcal{M}_{\mathcal{J}}$	$\mathcal{M}_{\mathcal{F}}$	$\mathcal{M}_{\mathcal{J}}$	$\mathcal{M}_{\mathcal{F}}$
$\mathcal{L}_{binary}$	83.9	0.902	62.3	0.691
$\mathcal{L}_{sep}$	82.4	0.904	65.9	0.742
$\mathcal{L}_{sep} + \mathcal{L}_{binary}$	84.8	0.912	68.2	0.759

Table 4. **Ablation studies of our TAViS for segmentation loss design.**

3. Visual Comparison with State-of-the-art Methods

As not all previous methods provide their pre-trained weights, we compare our approach with methods that have publicly available pre-trained weights across all datasets. In this section, we present additional visual comparison results on the S4,

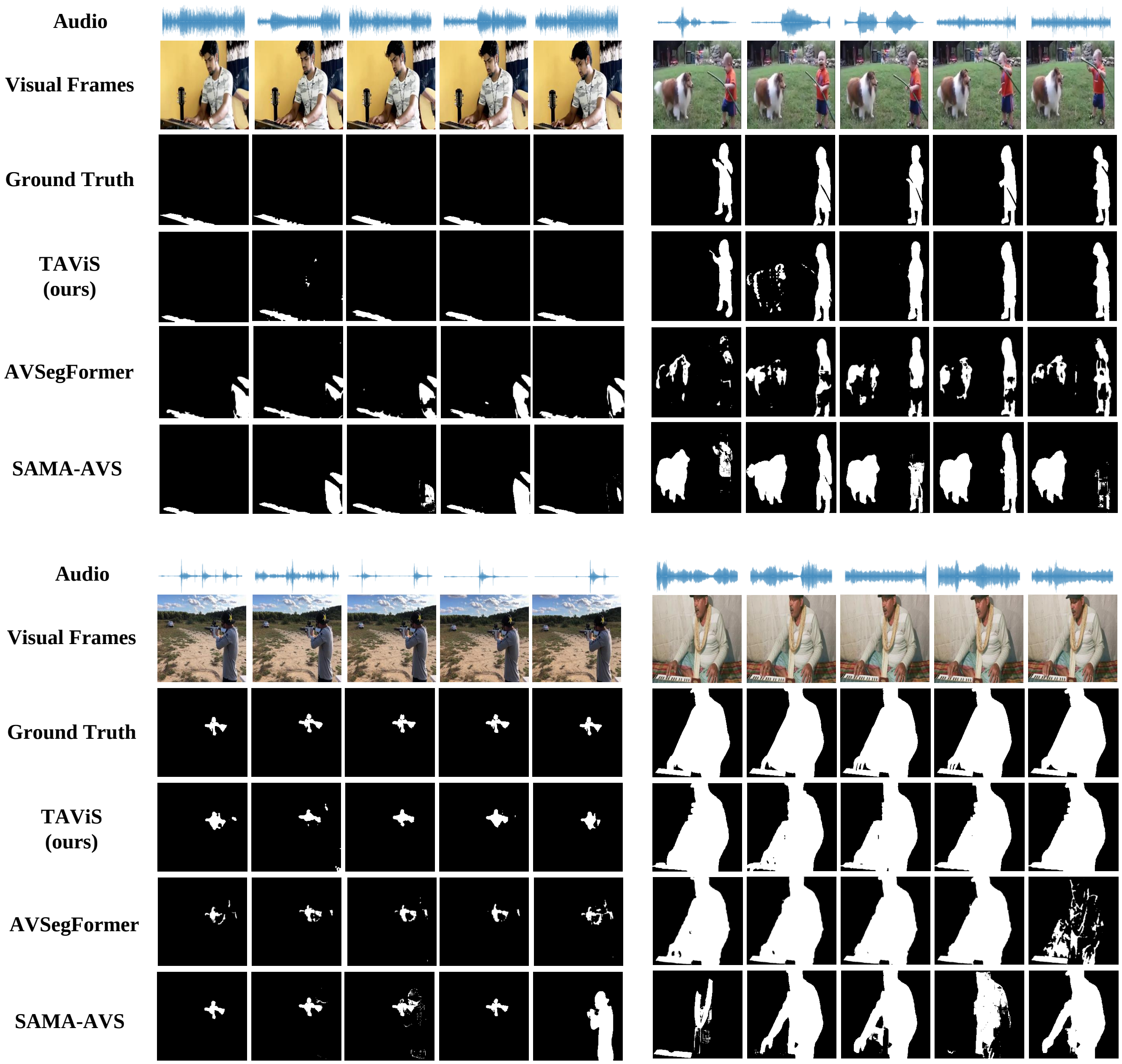


Figure 2. **Qualitative comparison of our model against state-of-the-art AVS methods on MS3 dataset.**

MS3, and AVSS datasets with these methods [1, 3], highlighting the effectiveness of our TAViS model. As shown in Figure 3, Figure 2, and Figure 4, our model demonstrates exceptional performance across various challenging scenarios, including handling significantly small or large sounding objects, multiple sounding objects, and occluded sounding objects, where existing methods often struggle.

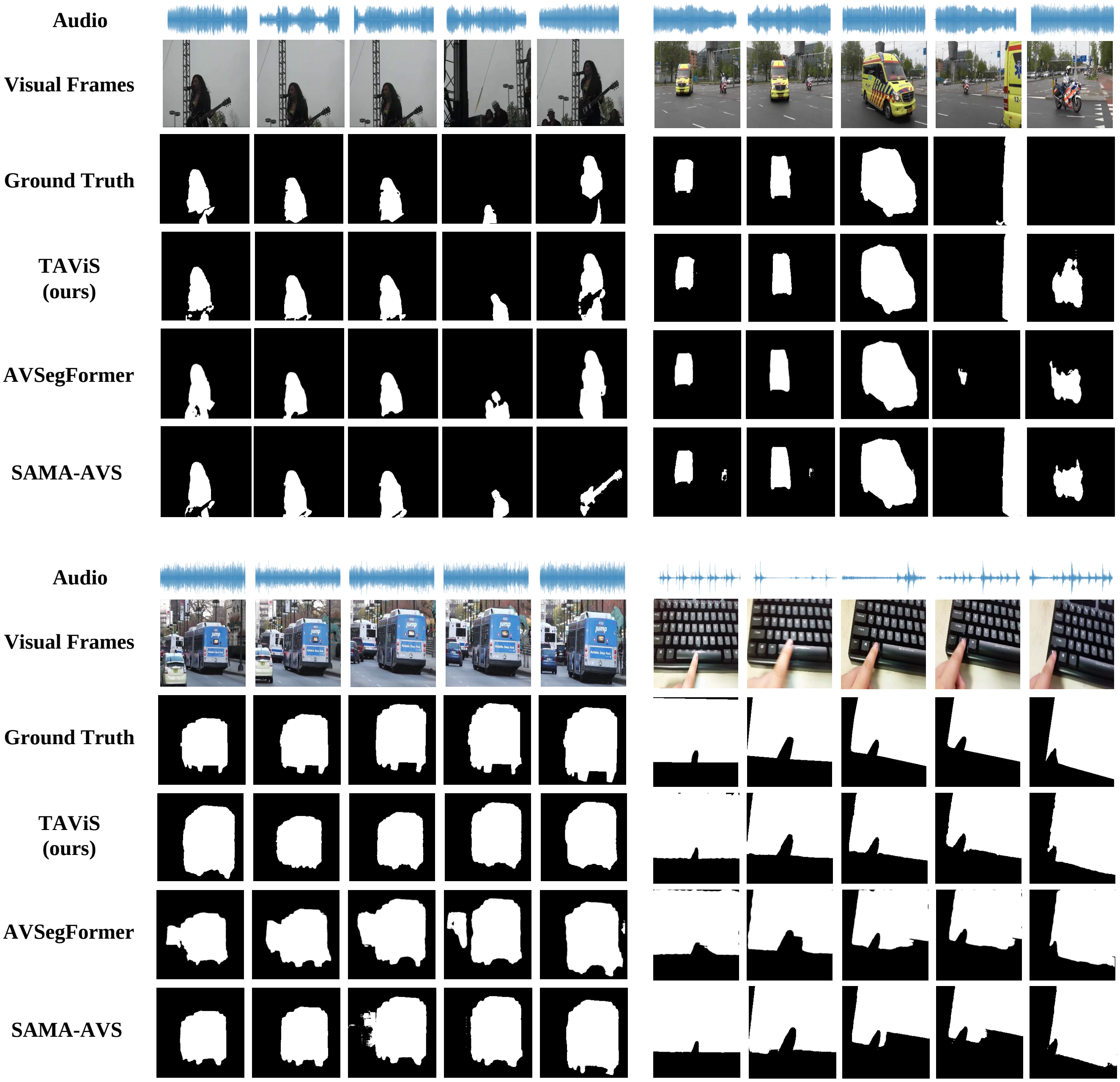


Figure 3. **Qualitative comparison of our model against state-of-the-art AVS methods on S4 dataset.**

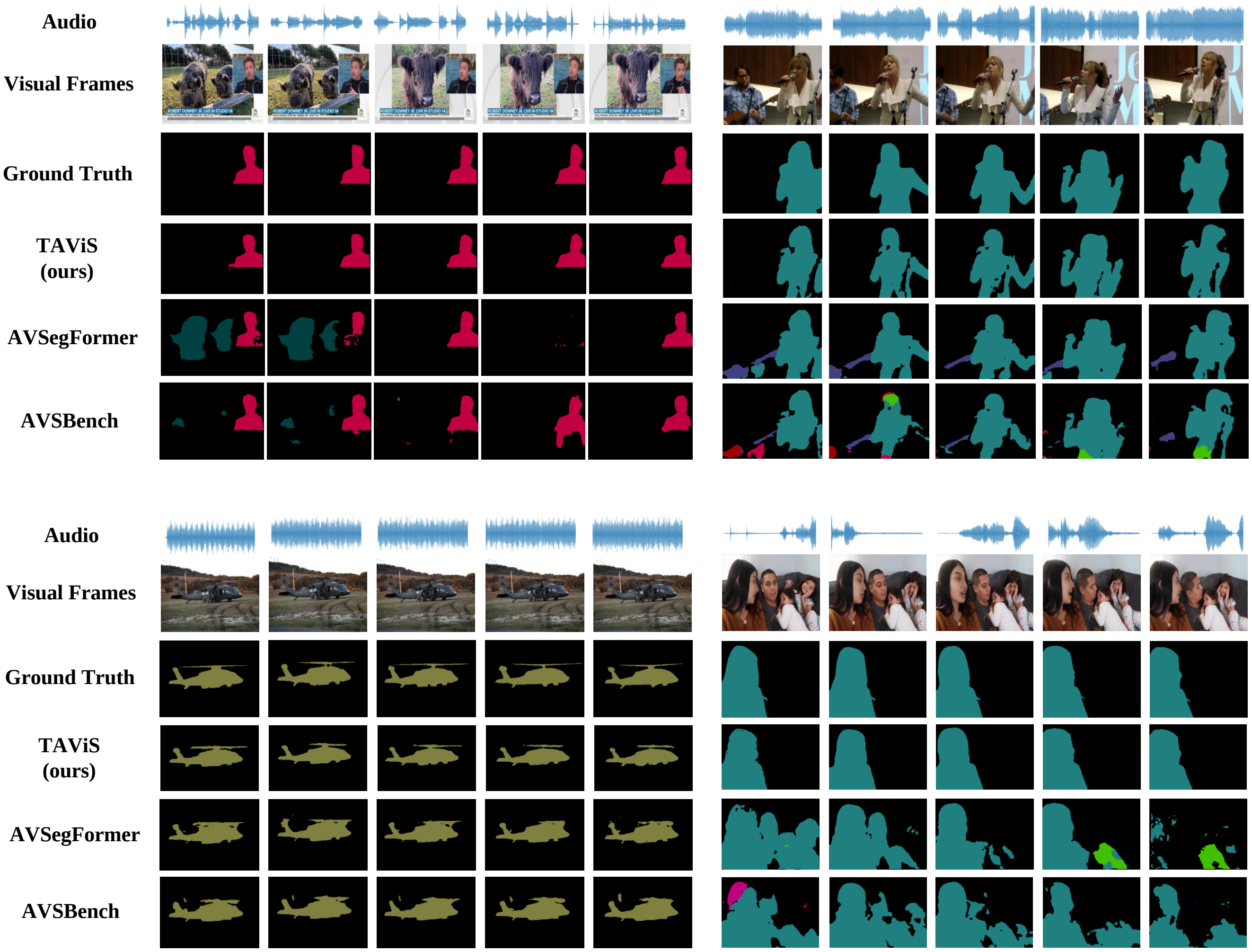


Figure 4. **Qualitative comparison of our model against state-of-the-art AVS methods on AVSS dataset.**

References

[1] Shengyi Gao, Zhe Chen, Guo Chen, Wenhai Wang, and Tong Lu. Avsegformer: Audio-visual segmentation with transformer. In *AAAI*, volume 38, pages 12155–12163, 2024. 3

[2] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *CVPR*, pages 7061–7070, 2023. 1

[3] Jinxiang Liu, Yu Wang, Chen Ju, Chaofan Ma, Ya Zhang, and Weidi Xie. Annotation-free audio-visual segmentation. In *WACV*, pages 5604–5614, 2024. 3

[4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1

[5] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *CVPR*, pages 16793–16803, 2022. 1