Can Textual Semantics Mitigate Sounding Object Segmentation Preference? (Supplementary Material)

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

[1] Gaoling School of Artificial Intelligence, Renmin University of China, China
yaoting.wang@outlook.com
dihu@ruc.edu.cn
[2] Beijing University of Posts and Telecommunications, Beijing, China
{sunpeiwen,zhhg}@bupt.edu.cn
[3] University of Edinburgh, Scotland, UK
yuanchao.li@ed.ac.uk
[4] Engineering Research Center of Next-Generation Search and Recommendation

A Detailed Prompts

Our detailed prompt template with CoT instructions for LLMs to reason the potential sounding objects is shown in Fig. A1. To enhance the comprehensiveness of instruction examples, we generate examples based on various quantitative and scenario settings. These include scenarios ranging from no possible sounding object to multiple sounding objects, as well as simple sounding scenarios with static objects to complex sounding scenarios with multiple sound sources.

B Qualitative Segmentation Results

To demonstrate the effectiveness of our model, we showcase its performance under various conditions by presenting qualitative segmentation examples. These include the results on the AVS task, depicted in Fig. B2, as well as the results on the AVSS task, shown in Fig. B3. Due to the unavailability of published work on AVS, which hinders result reproduction and comparison, we still employ AVSBench as our baseline for comparison.

Comparison on AVS: From Fig. B2, it is evident that our segmentation results show a significant improvement compared to AVSBench, demonstrating a substantial enhancement of audio-visual grounding.

Comparison on AVSS: From Fig. B3, we can observe that our model achieves

† Equal contribution.
✉ Corresponding author.

I will provide you with a scene description that includes both visual and auditory elements. You need to analyze the semantics and grammar of the given sentences step-by-step, to identify which objects may emit sounds. Note, your reasoning should follow the **sub-problems** outlined below. Which objects exist in the description? What are categories of the objects? Which categories can make sounds? Which group of sounding objects is able to make sounds simultaneously in description? Which group of objects is more possible?
Example 1: Inside a white room a boy is playing the trumpet.
Question: Which objects may make sounds?
Reasoning: The objects are boy, trumpet, and room. A boy is a human, and humans can make sounds by speaking and singing; A trumpet is an instrument, and the instrument can make sounds; A room is a static object, and the static object cannot make sounds; Boy and trumpet are unable to make sounds simultaneously; The trumpet is the most possible group to make sounds according to the description.
Answer: ['Trumpet'].
Example 2: A woman is walking with a dog in a garden.
Question: Which objects may make sounds?
Reasoning: The objects are woman, dog, and garden. A woman is a human, and humans can make sounds by speaking and singing; A dog is an animal, and animals can make sounds; A garden is an environment, and an environment cannot make sounds; Woman and dog are able to make sounds simultaneously. Woman and dog are the most possible group to make sounds according to the description.
Answer: ['woman', 'dog'].
Example 3: Several cars driving on the road, the sky is blue, and the distant mountains are blurred.
Question: Which objects may make sounds?
Reasoning: The objects are car, road, sky, and mountain. A car is a vehicle, and the vehicle can make sounds during driving; A road is an environment, and the environment cannot make sounds; The sky is an environment, and the environment cannot make sounds; The mountain is an environment, and the environment cannot make sounds. Cars are able to make sounds simultaneously. Cars are the most possible group to make sounds according to the description.
Answer: ['car'].
Example 4: A man is playing the piano and another is playing the guitar.
Question: Which objects may make sounds?
Reasoning: The Objects are man, people, piano, and guitar. A man is a human, and humans can make sounds by speaking and singing; A piano is an instrument, and the instrument can make sounds; A guitar is an instrument, and the instrument can make sounds; Piano and guitar are able to make sounds simultaneously. Two men are able to make sounds simultaneously. Guitar and man are able to make sounds simultaneously. Piano and guitar are the most possible group to make sounds according to the description.
Answer: ['piano', 'guitar'].
Example 5: A pencil box is put on the bench.
Question: Which objects may make sounds?
Reasoning: The Objects are: pencil box, and bench. A pencil box is a static object, and the static object cannot make sounds; A bench is a static object, and the static object cannot make sounds;
Answer: [].

Fig. A1: Few-shot prompt with CoT instructions. We feed these prompts to LLMs before each time we generate the reasoning results. This template generates the best result of Tab. G3.

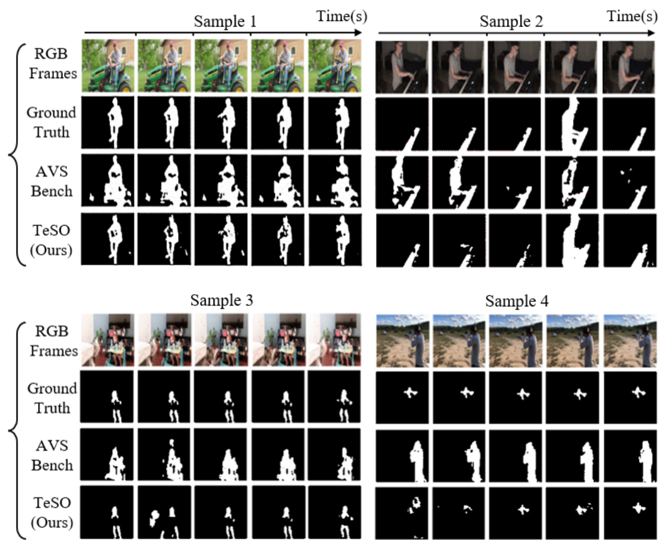

Fig. B2: Qualitative comparison on AVS task. TeSO shows brilliant audio-visual grounding compared with the baseline.

great performance gains on AVSS. Besides the noticeable improvements in semantic segmentation results, it is crucial to remind that our approach tries to mitigate the segmentation preference of audible objects.

C Effectiveness of Audio Control

More examples of muted or noise-only audio are provided here to show the effectiveness of audio control. From Fig. J4, it is evident that our segmentation results achieved by TeSO exhibit almost entirely blank masks, indicating effective "guidance" by the muted or noise-only audio. As we mentioned in Sec. 4.4, our experiments reveal that the comparable results of some methods may to a certain extent rely on the segmentation preference of audible objects built during the training because they still segment audible objects (as shown in Fig. J4) and get comparable results even with pure Gaussian white noise.

D Effectiveness of Text Guidance

As shown in Fig. J5, we present the qualitative results of introducing text cues to enhance the audio-visual correlation using a heatmap instead of the final binary mask, to show the degree of the model's attention on every pixel. By employing language as a bridge, we elevate audio-visual correlation by leveraging text cues as guidance, thereby enabling the model to better perceive the audio guidance.

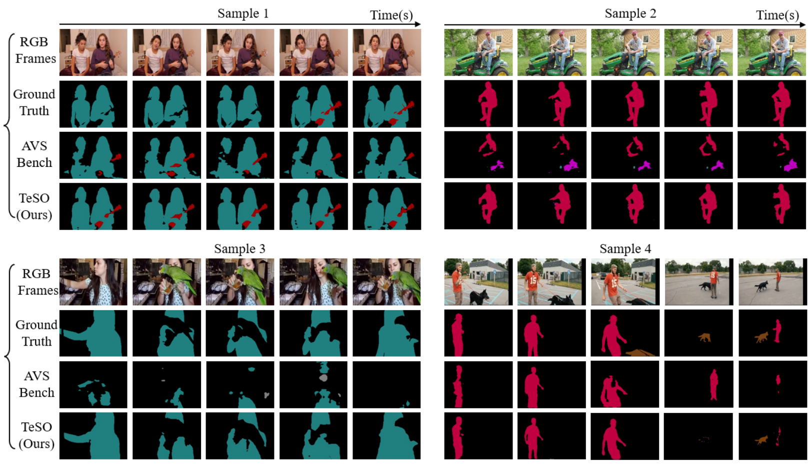

Fig. B3: Qualitative comparison on AVSS task. TeSO has better semantic perception capability compared with other method.

E Ablation of Captioner and Reasoner

As shown in Tab. D1, (a) and (b) indicate that using a text-only LLM LLaMA-2 as a reasoner outperforms a multimodal LLM LLaVA-1.5. (a) and (c) demonstrate that LLaVA-1.5 is more effective as a frame-level dense captioner. (a) and (d) show that benefits from the reasoning ability, LLM reasoner outperforms a naive NLTK noun parser. (a) and (e) suggest that our frame-level reasoning approach is more accurate than using manual video-level noisy labels.

F Ablation of Audio Backbone.

We use the BEATS$_{iter3}$ as an alternative to extract audio features. As shown in Tab. F2, we achieved a modest improvement of 0.24%. However, we use VGGish for a fair comparison with most of the other methods.

Table D1: Captioner and reasoner ablation on V1M-Test. The cooperation of LLaVA-1.5 and LLaMA-2 (the first column) works best.

	(a)	(b)	(c)	(d)	(e)
Captioner	LLaVA-1.5-7b	LLaVA-1.5-7b	VideoLLaMA-7b	LLaVA-1.5-7b	/
Reasoner	LLaMA-2-7b	LLaVA-1.5-7b	LLaMA-2-7b	NLTK	Manual video-level noisy label
mIoU	**66.02**	65.41	64.27	63.87	65.53
F-score	**0.801**	0.790	0.788	0.783	0.792

Table F2: Ablation of audio backbone on V1M-Test. Model with BEATs audio feature gets better performance.

Method	Backbone	mIoU(%)	Fscore
BAVS	BEATs	59.63	0.659
Ours	VGGish	66.02	0.801
Ours	BEATs	**66.26**	**0.803**

Table G3: Text assistance and prompt template. We implement diverse prompt templates on LLaMA-2-7b-Q4 to generate semantic text labels with varying levels of quality in V1M. Subsequently, we assess the text's capability to control the segmentation results when subjected to different quality conditions of semantic text labels.

No.	Category	CoT	Few shot	Template	mIoU(%)	F-score
1	instructive	✔	few-shot	$N\times$ templates shown in Fig. 3.	**66.02**	**0.801**
2		✔	one-shot	$1\times$ template shown in Fig. 3.	65.33	0.792
3		✔	zero-shot	"Let's think step by step to obtain sounding objects in the caption."	64.76	0.787
4		✘	few-shot	$N\times$ questions of sounding objects with direct answers.	64.80	0.785
5		✘	one-shot	$1\times$ question of sounding objects with direct answer.	64.08	0.773
6		✘	zero-shot	"Please tell me the sounding objects in the caption."	63.83	0.775
7	misleading	✘	zero-shot	"Please tell me the most possible sounding object in the caption."	62.98	0.753
8		✘	zero-shot	"Please tell me any objects in the caption."	63.39	0.765
9		✘	zero-shot	"Please tell me the background objects in the caption."	61.52	0.740
10	irrelevant	✘	zero-shot	"Tell me any random object."	60.86	0.722

G Ablation of Different Prompts

We experienced with different prompt templates to explore the best prompting method. The alternative methods are divided into three categories based on instructive (No.1-6), misleading (No.7-9), and irrelevant (No.10) effects on segmentation, as shown in Tab. G3.

The inductive approach involves employing various prompt methods to acquire informative guidance. By examining the impact of this instructive information on segmentation, we can analyze the significance of incorporating CoT within the method. Utilizing a few-shot template with CoT can lead to a performance improvement of up to 1.22% (No.1 vs. No.4). Furthermore, we devise an effective few-shot (No.1) template that yields a performance improvement of up to 1.25% compared to the zero-shot (No.3) inference process. Finally, our method observes that the combination of few-shot templates and CoT leads to optimal performance.

In a similar context, we conduct experiments to investigate the impact of misleading text labels on the model. Specifically, we explore the effects of missing (No.7), ambiguous (No.8), and erroneous (No.9) text semantic information on the model's performance. The presence of such problematic semantic information results in a significant 4.50% decrease in mIoU. Moreover, the introduction of irrelevant information (No.10) causes the largest drop in performance, amounting to 5.16% performance.

Table H4: Performance on AVS-V3 for testing the generalization on unseen objects.

	0-shot		1-shot		3-shot		5-shot	
Method	mIoU(%)	F-score	mIoU(%)	F-score	mIoU(%)	F-score	mIoU(%)	F-score
AVSBench	53.00	0.707	56.11	0.754	63.22	0.767	63.87	0.783
AVSegFormer	54.26	0.715	58.30	0.764	64.19	0.774	65.17	0.785
GAVS	54.71	0.722	62.89	0.768	66.28	0.774	67.75	0.795
TeSO (ours)	**61.06**	**0.743**	**65.41**	**0.775**	**69.29**	**0.791**	**72.15**	**0.810**

H Multi-instance Scenario

The multi-instance scenario (*e.g.*, two men in the same frame) is a challenging problem in the field of AVS, especially since the existing mono-channel audio input cannot provide sufficient audio spatial information to the model. While our focus in this work is on the segmentation preference problem caused by weak audio guidance, we believe that our method has the potential to address such issues because visual scene descriptions can provide guidance that is difficult to obtain through audio features. For example, "a man in red clothes on the right is singing a song while the man in black is smiling" can help eliminate the ambiguity in multi-instance scenes. In future work, we hope to explore the multi-instance problem in AVS using our framework.

I Generalization

Benefiting from the strong generalization of LLMs, it is possible to perform visual scene understanding on almost any scene and infer potential sounding objects as text cues. Therefore, we explore whether the model can exhibit better segmentation generalization performance for unseen object categories with the assistance of generalizable text cues from LLMs. As shown in Tab. H4, our experiments on the AVS-V3 dataset demonstrate that our approach can significantly improve the generalization performance compared to other methods. This improvement may be attributed to the generalizable text cues reasoned by LLMs from scene descriptions, establishing a better audio-visual correlation.

J Versatility

Our text-guided method is generally applicable to various visual foundation models with a transformer-based decoder. As depicted in Tab. J5, we present the outcomes obtained by employing two different popular visual foundation models, namely SAM (Segment Anything Model) and Mask2Former, in their base versions. Even when using SAM as the foundation model, our method remains effective and produces comparable results.

Table J5: Performance of TeSO on AVS-Benchmarks with different visual foundation models. TeSO-M2F uses Mask2Former as the visual foundation model while TeSO-SAM uses SAM as the visual foundation model.

Method	Audio-backbone	Visual-backbone	V1S		V1M		AVSS-binary	
			mIoU(%)	F-score	mIoU(%)	F-score	mIoU(%)	F-score
AVSBench	VGGish	PVT-v2	78.70	0.879	54.00	0.645	62.45	0.756
BAVS	Beats	Swin-Base	82.68	0.898	59.63	0.659	55.45	0.640
TeSO-SAM (ours)	VGGish	ViT-Base	82.68	0.903	64.97	0.794	68.16	0.808
TeSO-M2F (ours)	VGGish	Swin-Base	**82.84**	**0.917**	**66.02**	**0.801**	**68.53**	**0.813**

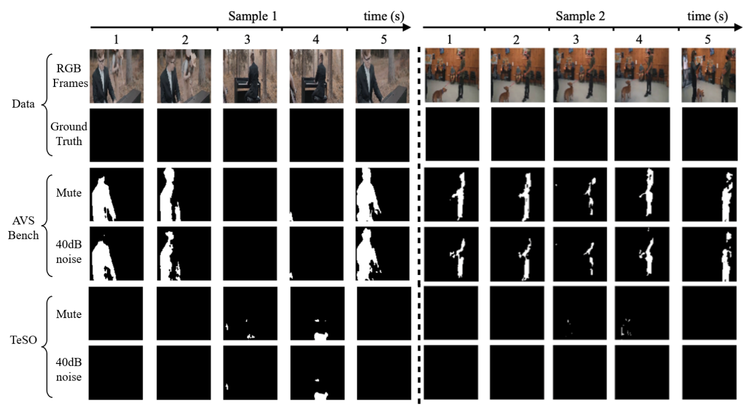

Fig. J4: Comparison of audio control on AVS results with muted and 40dB pure White Gaussian Noise (WGN) audio input. The ground truth labels are all blank because the audio input is manually set wrong with muted or noise-only audio.

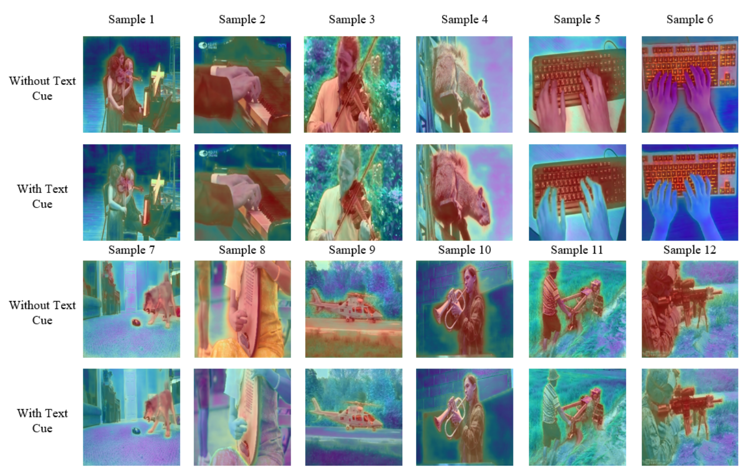

Fig. J5: Comparison of whether text cues are used in TeSO. TeSO enhances audio-visual correlation by using text cues as a bridge between audio and visual modalities. Brighter (or redder) colors indicate the model's increased attention to these pixels.