Can Textual Semantics Mitigate Sounding Object Segmentation Preference?

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

Abstract. The Audio-Visual Segmentation (AVS) task aims to segment sounding objects in the visual space using audio cues. However, in this work, it is recognized that previous AVS methods show a heavy reliance on detrimental segmentation preferences related to audible objects, rather than precise audio guidance. We argue that the primary reason is that audio lacks robust semantics compared to vision, especially in multi-source sounding scenes, resulting in weak audio guidance over the visual space. Motivated by the the fact that text modality is well explored and contains rich abstract semantics, we propose leveraging text cues from the visual scene to enhance audio guidance with the semantics inherent in text. Our approach begins by obtaining scene descriptions through an off-the-shelf image captioner and prompting a frozen large language model to deduce potential sounding objects as text cues. Subsequently, we introduce a novel semantics-driven audio modeling module with a dynamic mask to integrate audio features with text cues, leading to representative sounding object features. These features not only encompass audio cues but also possess vivid semantics, providing clearer guidance in the visual space. Experimental results on AVS benchmarks validate that our method exhibits enhanced sensitivity to audio when aided by text cues, achieving highly competitive performance on all three subsets. Project page: https://github.com/GeWu-Lab/Sounding-Object-Segmentation-Preference

Keywords: Audio-Visual Segmentation · Textual Guidance · Multimodal learning

---

† Equal contribution.
✉ Corresponding author.

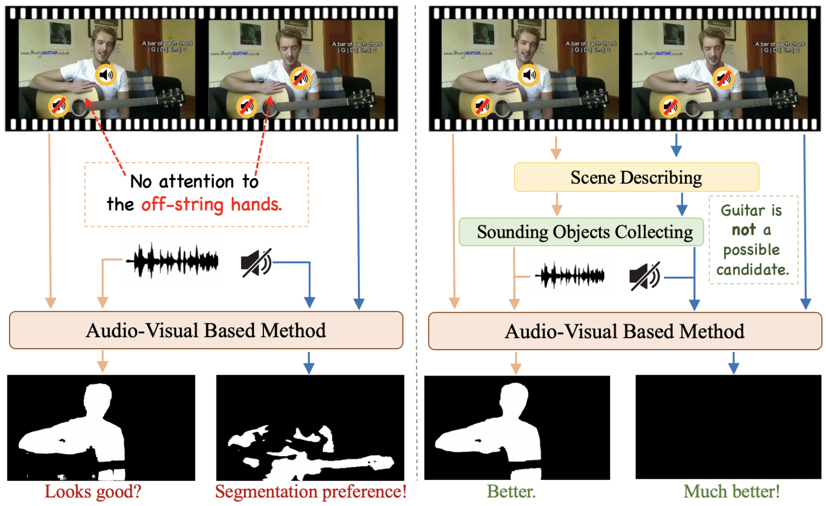

Fig. 1: The previous methods *(left)* achieve satisfactory results in receiving normal audio input. However, even when the sound is completely silent, they still segment vast pixels to represent the guitar regarding the segmentation preference of audible objects built during the training. In contrast, our approach *(right)* utilizes a frozen LLM to reason from scene descriptions that the male is not playing the guitar, as his hands are off the strings. Guided by this semantic information, our method produces more precise segmentation with finer audio guidance.

1 Introduction

In human perception, visual and auditory senses are closely related, and audio can provide vision with additional and dynamic scene information to enhance visual understanding. In the context of audio-visual understanding tasks, Audio-Visual Localization (AVL) enables the localization of sounding objects within visual scenes using audio references [15, 36]. However, the coarse localization of sounding objects is no longer sufficient to meet the practical demands of complex scene understanding, such as in autonomous driving [12, 49] and augmented reality [33]. To address this challenge, tasks are gradually shifting from bounding-box or rough heat-map localization to finer-grained pixel-level segmented localization, which is known as Audio-Visual Segmentation (AVS).

The existing works on AVS can be broadly categorized into fusion-based methods [8, 16, 20, 21, 29, 47] and prompt-based methods [30, 34, 40]. The former primarily focuses on localizing sounding objects by fusing audio and visual features, while the latter emphasizes constructing effective audio prompts for the visual foundation model. However, most of these methods fail to establish compact audio-visual correlations to achieve clear audio guidance in the visual space. As depicted in the left of Fig. 1, previous methods could achieve satisfactory results with normal audio input. Nevertheless, when the audio is silent, these methods still segment many pixels to represent potential sounding objects based on the **segmentation preferences** of audible objects[1].

Furthermore, from the experiment with manually muted audio input, we observe that most of the behavior of AVS baselines depends on the segmentation preference established during the training process rather than relying on reliable audio-visual correlation. For instance, these models tend to take a shortcut by segmenting out the muted guitars because the audio guidance is weak, and they simply learned with the easy-to-learn visual feature alone (*i.e.*, shortcut learning) that guitars are often associated with sound-emitting during the training phase. The unreliable audio-visual correlation can be attributed to two main factors. Firstly, the scarcity of training data for AVS is a significant challenge due to the demanding nature of pixel-level annotation. Secondly, the audio modality itself presents inherent complexity and ambiguity [16, 21, 29], especially in scenarios involving multiple audio sources that may be intertwined.

Given the challenges discussed above, our approach aims to build a robust audio-visual correlation by integrating the text modality, which inherently provides meaningful and robust semantic information about objects and their interactions. **Firstly**, in the Scene Describing phase, we utilize an off-the-shelf image captioner (LLaVA-1.5 [28]) to generate detailed dense captions for scene understanding. **Then**, in the Sounding Objects Reasoning phase, we instruct the frozen large language models (LLMs, *e.g.*, LLaMA2 [39]) as a text cues capturer to collect potential sounding objects from the scene description as text cues. Notably, we introduce an elaborate few-shot prompt template to guide the text cues capturer to reason with Chain-of-Thought (CoT) instructions. **After that**, in the Semantics-Driven Audio Modeling (SeDAM) module, we project the audio feature into latent audio features. However, instead of directly using these latent features to interact with visual features for mask decoding, we leverage these latent features to interact with text cues. Specifically, the latent features are attended to and collected to form sounding object features with semantics provided by the text cues through the crossmodal transformer. We further propose the Prompting Mask Queries with Semantics (PMQS) module to introduce the sounding object features into pre-trained mask queries for later segmentation. **Finally**, we use simple but effective bottleneck adapters in the Audio-Prompted Decoding phase for better segmentation quality. The experimental results demonstrate that our proposed method improves audio-visual grounding by incorporating guidance from both audio and text cues.

Experiments in Sec. 4.4 reveal the advantage of our method, as verified by its increased sensitivity to the changes in audio input compared to networks without enhancing audio guidance with text cues. In summary, our contributions are threefold:

- We recognize the behavior of AVS model is influenced by the preference segmentation phenomenon, which can not be solved easily by previous audio-

[1] To be clear, we define audible objects as objects ***capable*** of producing sound, while sounding objects are defined as objects that are ***currently*** producing sound.

visual based methods. A novel strategy introduces textual semantics as the bridge into the AVS task, establishing a better audio-visual correlation.
- We introduce the SeDAM module to form sounding object features with audio and text cues. We also suggest a PMQS module to prompt the pre-trained mask queries with these sounding object features for sounding object segmentation using a visual foundation model.
- Our experiments show that our method can be more sensitive to the audio guidance, manifested in its high sensitivity to the changes in audio input.

2 Related Works

2.1 Audio-Visual Localization and Segmentation

Traditional AVL tasks [15,35,36] predict the positions of sounding objects using bounding-box or coarse heat-maps through unsupervised learning of audio-visual correlation. In recent years, driven by the growing demand for more precise localization in industries such as autonomous driving [12,49] and augmented reality [33], the AVL task has evolved towards a finer-grained AVS task that localizes sounding objects at a pixel-level. The existing AVS works can be broadly categorized into fusion-based [8, 16, 20, 21, 29, 43, 47] and prompt-based [30, 32, 40] The pioneering AVS work [47] is fusion-based and adopts a multi-stage strategy to integrate audio with multi-scale visual features. Liu *et al*. [29] address the issues of inadequate fusion of audio-visual features with an audio-aware query-enhanced transformer (AuTR). In comparison, AVSegFormer [8] directly decodes the fused feature with the audio query. Wang *et al*. [40] innovatively prompt the visual foundation model with audio input, harnessing the abundant visual before achieving generalizable AVS in zero-shot and few-shot scenarios. However, these methods have not effectively established a reliable audio-visual correlation, preventing the models from perceiving robust audio guidance. On the contrary, our experiments in Sec. 4.4 also reveal that some AVS baselines' competitive performance may stem from segmentation preferences formed during the training.

2.2 Text Aided Scene Understanding

Before the explosion of LLM, textual semantic information had been widely used for visual understanding tasks. Zhan *et al*. [38] and Sharma *et al*. [37] improve Visual Question Answering by using the extra semantic information of image captions, while Hur and Park [17] utilize the image captioner to help zero-shot image classification. Wang *et al*. [41] incorporating text with audio and visual cues to solve reference AVS task.

In recent years, the community has witnessed an emergent interest in strong off-the-shelf capabilities in open-world visual understanding [24–26, 45]. Notable examples, such as GPT4(V) [44] and LLaVA [27] have showcased remarkable linguistic and visual capabilities in zero-shot scene understanding [7, 42] without any training requirement. Innovatively, in the field of AVS, BAVS [23] replaces

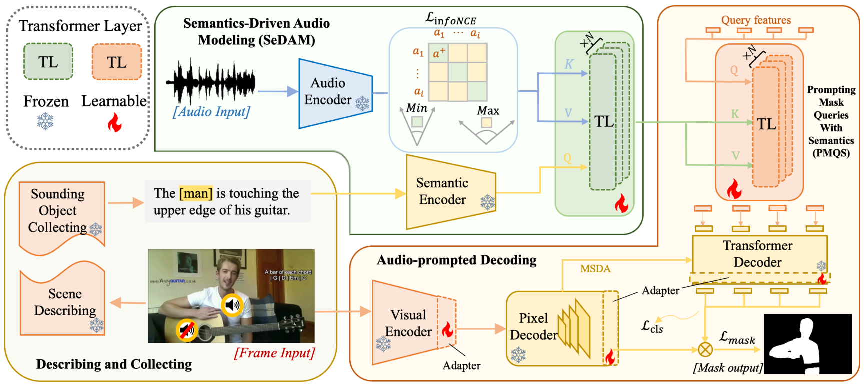

Fig. 2: Overall pipeline of the proposed TeSO method. We utilize an off-the-shelf image captioner for dense scene describing and employ a frozen LLM to reason out potential sounding objects as text cues. These semantic text cues are then aggregated with audio features in the SeDAM module to form the sounding object features. Subsequently, we introduce the sounding object features into pre-trained mask queries in the PMQS module. Finally, we use adapters to tune the visual-only mask decoder for AVS in the Audio-prompted Decoding phase. "MSDA" is the multi-scale deformable attention proposed by Zhu *et al.* [48].

audio input with textual audio tags from a large pre-trained vision (UniDiffuser [2]) and audio (BEATs [3]) foundation model, establishing a reliable AVS system with foundation knowledge. However, due to the significant gap in class distribution between the pre-training dataset and the AVS dataset, as well as the presence of multi-source scenarios in the AVS task, the extracted audio tags are not ideal to provide accurate semantics. In contrast, we utilize the text modality with robust semantic features to enhance rather than replace the audio guidance and address the segmentation preference issue.

3 Text-guided Sounding Object Segmentation

As shown in Fig. 2, we introduce our **Te**xt-guided **S**ounding **O**bject Segmentation (TeSO) method in this section. Our experiments about muted audio input imply that audio does not exert reasonable guidance over the visual space for some AVS baselines, resulting in the models relying more on the segmentation preferences established during the training process. In essence, they learn which objects are most likely audible during training and take shortcuts during inference. To address the above issue, our method aims to enhance the audio-visual correlation by leveraging the text modality, which inherently possesses robust semantic information, to obtain finer-grained audio guidance. We begin by acquiring detailed scene descriptions through an off-the-shelf image captioner. Subsequently, a frozen LLM works as the text cues capturer, collecting potential sounding objects as text cues from the scene descriptions with CoT instructions.

Finally, we introduce a novel SeDAM module with a dynamic mask to seamlessly integrate audio features and text semantics through a crossmodal transformer, providing finer audio guidance with text semantic cues.

3.1 Multimodal Representation

Visual: Following previous works [40,47], we sample frames from the video at 1-second intervals. Mask2Former [4] serves as our visual foundation model, and we extract visual features $F_V \in \mathbb{R}^{d_V \times H \times W}$ from a pre-trained Swin transformer [6]. We use simple two-layer MLP adapters to tune the visual encoder for better visual representations.
Audio: Similar to the video processing, we split the audio into clips at 1-second intervals. We extract audio features $F_{As} = \{F_{A_1}, F_{A_2}, ..., F_{A_T}\} \in \mathbb{R}^{T \times d_A}$ using VGGish [9, 13], where T represents the length of the audio in seconds, matching the number of video frames. Therefore, each video frame corresponds to $F_A = F_{As}[i]$. We freeze the pre-trained parameters for a better comparison with previous works.
Text: For each video frame, we complement the captured text cues into a sentence using a template ("This is a {_}.") to extract text features $F_T \in \mathbb{R}^{N_T \times d_T}$ with ImageBind [10], where N_T is the number of text cues.

3.2 Text Cues Generation

To obtain more precise text cues, we use a two-stage inference flow. Firstly, we employ an off-the-shelf image captioner to generate dense scene descriptions. Secondly, we ask a text cues capturer to collect potential sounding objects from the generated descriptions.

Scene Describing. As shown in the top-left of Fig. 3, in this phase, we employ the commonly used and straightforward prompt: *"Please carefully understand the image content and provide as rich a description as possible."* This instructs the image captioner to generate a detailed description of the visual scene, a task commonly referred to as dense captioning [18]. This process enables the image captioner to produce comprehensive and informative descriptions that not only describe the objects like "a human and guitar in the image" but also capture their interactions and relationships.

Sounding Object Collecting. To extract potential sounding objects from the generated dense descriptions, we design a task-specific CoT prompt template with few-shot demonstrations, as illustrated in Fig. 3. Specifically, we carefully craft examples that cover various scenarios and provide reasoning approaches for each example. The process begins by identifying the objects present in the dense scene descriptions, followed by step-by-step common-sense reasoning according to the CoT instructions, as depicted in the lower right of Fig. 3. During the CoT reasoning, the text cues capturer first categorizes the sample objects and

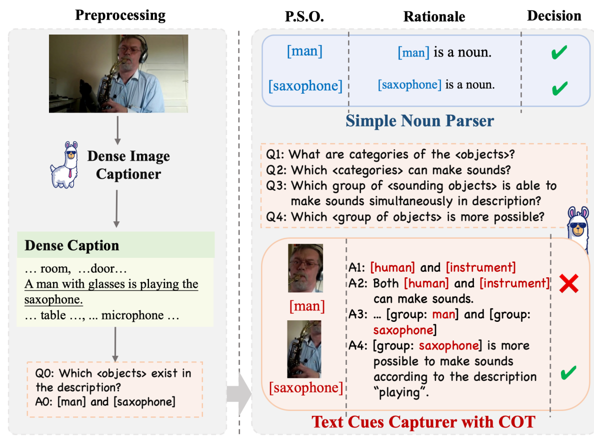

Fig. 3: "P.S.O" stands for "Potential Sounding Object". A frozen LLM reasoner works as a text cues capturer by considering the interaction of audible objects. For instance, a person may sing a song while playing the guitar, but he would not sing along with a saxophone. In contrast, a noun parser simply captures any nouns that are present.

assesses whether objects in that category are likely making sounds. Subsequently, the capturer analyzes which group of potential sounding objects can produce sounds simultaneously. Finally, the group with higher confidence is captured as the text cues of length N_T and encoded in F_T for further guidance. The right part of Fig. 3 also illustrates why we use the frozen LLM as a text cues capturer instead of using a naive noun parser. It is important to note that the use of LLM is solely for generating high-quality text cues, the LLM itself is not engaged in the training process for a fairness concern.

3.3 Semantic-Driven Audio Modeling

After obtaining potential sounding objects, we can enhance the semantics of audio features with these text cues in this SeDAM module. To allow semantic text cues to capture different aspects of audio features F_A, we project F_A into distinct latent features $F_{A_L} = \{a_i\}_{i=1}^{N_L} \in \mathbb{R}^{N_L \times d_{A_L}}$, where N_L is the number of latent features. It is worth noting that, we incorporate an auxiliary loss term $\mathcal{L}_{infoNCE}$ aims at ensuring the distinction among latent audio features. This way, we project audio features into the latent space, where each unique a_i represents a latent audio feature to be captured by semantic text cues, forming clear and loud sounding object features with robust semantics for finer guidance.

To incorporate the semantics of text cues, we employ text cues as the query to attend and combine the relevant latent features through crossmodal multi-head attention, formulated as:

$$F_{\widetilde{A_L}} = MultiHead(F_T, F_{A_L}, F_{A_L}), \tag{1}$$

where $F_{\widetilde{A}_L} \in \mathbb{R}^{N_T \times d_A}$ is the attended latent audio features and will be combined with text cues F_T as follows:

$$F_{A_T} = FC(LN(F_T + F_{\widetilde{A}_L})), \tag{2}$$

where $F_{A_T} \in \mathbb{R}^{N_T \times d_A}$ is the combined feature with both audio information and semantic text cues, and FC and LN for linear layer and layer normalization. Note that the LLM reasoner may generate a different number of text cues, and we either pad with a zero matrix or truncate to N_T. It is noteworthy that we propose task-specific CoT instructions to prompt the text cues capturer to infer more accurate potential sounding objects as the text cues. Additionally, we also introduce a dynamic mask based on attention score to mitigate the potential noisy information.

We first compute the average pooling attention scores $\zeta \in \mathbb{R}^{N_T}$ for text cues on the attention head and key dimension:

$$\xi = Pool\left(\frac{F_T F_{A_L}^{\mathrm{T}}}{\sqrt{d_{A_L}}}\right), \tag{3}$$

where $\sqrt{d_{A_L}}$ is the scaling factor for dot production. Then we form the attention mask $\mathcal{M} = \{m_1, m_2, m_3, ..., m_{N_T}\}$ as follows:

$$\mathcal{M}_i = \begin{cases} 0 & \text{if } \xi_i > 0 \\ -\infty & \text{otherwise} \end{cases}, \tag{4}$$

where $\mathcal{M} \in \mathbb{R}^{N_T}$, will be used as a key padding mask for the PMQS module in the next section to ensure that the attention weight of redundant text cues approaches zero.

3.4 Prompting Mask Queries with Semantics

Current segmentation foundation models [4, 5, 19] adopt transformer-based decoder with mask queries as the mainstream approach. To prompt such foundation models, we propose the PMQS module, aiming to introduce semantic information to mask query features $F_Q = \{q_1, q_2, ..., q_{N_Q}\}$, where $F_Q \in \mathbb{R}^{N_Q \times d_V}$, and N_Q is the number of mask queries in crossmodal attention. We formulate this updated expression as follows:

$$\phi_{ca}(F_Q, F_{A_T}, F_{A_T}) = Softmax\left(\frac{F_Q F_{A_T}^{\mathrm{T}}}{\sqrt{d_{A_T}}} + \mathcal{M}\right) F_{A_T}, \tag{5}$$

where $\mathcal{M}$ is added to the attention score to mask the redundant text cues for each query. Then we obtain the updated mask query features $\widetilde{F}_Q \in \mathbb{R}^{N_Q \times d_V}$ as:

$$\widetilde{F}_Q = F_Q + \phi_{ca}(F_Q, F_{A_T}, F_{A_T}). \tag{6}$$

3.5 Audio-Prompted Decoding

As shown on the right side of Fig. 2, the updated mask query features $\widetilde{F}_Q$ are forwarded into the transformer-based decoder, generating mask queries $M_Q \in \mathbb{R}^{N_Q \times H \times W}$ and class queries $C_Q \in \mathbb{R}^{N_Q \times N_C}$, where N_C is the number of class queries pre-defined in the segmentation foundation model. Subsequently, we can obtain the final segmentation $M_{pred} \in \mathbb{R}^{N_C \times H \times W}$:

$$M_{pred} = Einsum(C_Q, \widetilde{F}_Q), \tag{7}$$

where $Einsum(\cdot)$ is the Einstein Summation Convention.

Notably, to enhance the quality of the segmentation, we use two-layer bottleneck MLP adapters [14] to lightly fine-tune the output projection of the following attentions: a) Swin Window Attention (W-MSA) [31], b) Deformable Attention in pixel decoder [48] and c) Masked Attention in transformer decoder [4].

3.6 Learning Objectives

Segmentation Loss. During the model training process, we employ the binary cross-entropy loss and dice loss to optimize the mask quality:

$$\mathcal{L}_{mask} = \lambda_{bce} \cdot \mathcal{L}_{bce} + \lambda_{dice} \cdot \mathcal{L}_{dice}. \tag{8}$$

Further, we adopt the mask classification loss [5] to compose the final segmentation loss:

$$\mathcal{L}_{seg} = \mathcal{L}_{mask} + \lambda_{cls} \cdot \mathcal{L}_{cls}. \tag{9}$$

Latent Component Loss. This loss term aims to explicitly guide the model to learn distinct representations for each latent audio component, helping prevent redundancy and overlapping information across the latent components.

$$\mathcal{L}_{infoNCE} = -\log \left(\frac{\exp(sim(a, a^+))}{\sum_{i=1}^{N_L} \exp(sim(a, a_i))} \right), \tag{10}$$

where a^+ and $sim(\cdot)$ denote the positive example and the cosine similarity.
Total Loss The final loss function is composed of the weighted sum of the two kinds of losses above:

$$\mathcal{L} = \mathcal{L}_{seg} + \lambda_{infoNCE} \cdot \mathcal{L}_{infoNCE}. \tag{11}$$

4 Experiments

4.1 Implementation Details

Dataset. Our proposed method is evaluated on the AVS Benchmarks [47], which contains three subsets. Firstly, the single-source subset (V1S) contains 4932 videos over 23 categories, covering various sounds, such as humans, animals, vehicles, and musical instruments. For videos in the training split of this subset, only the first sampled frame is annotated. Secondly, the multi-source subset (V1M) contains 424 videos that include two or more categories from V1S, and all sounding objects are visible in the frames. Finally, the AVSBench-semantic (AVSS) subset, an extension of V1S and V1M, contains 12,356 videos. These newly collected videos are trimmed to 10 seconds, differing from the 5 seconds in V1S and V1M.

Setting. We conduct training and evaluation using the VGGish backbone pretrained on Youtube-8M [1] and Swin-base Transformer backbone pretrained on semantic-ADE20K [46] by Mask2Former [4]. The number N of the crossmodal transformer layers is set to 4, and the parameters $\lambda_{bce}, \lambda_{dice}, \lambda_{cls}, \lambda_{infoNCE}$ in the loss are set to $5, 5, 2, 1$. The AdamW optimizer is adopted with a learning rate of 1e-4 for the visual encoder adapters and 1e-3 for other learnable parameters, and the training epoch is roughly set to 60.

Metrics. To conduct a comprehensive evaluation of our model, we carry out tests using mean Intersection over Union (mIoU) and F-score as the performance metrics, following previous works [8, 47].

4.2 Comparison Results

When compared to methods that do not incorporate text as additional semantic information, our method, which leverages captioning and reasoning, demonstrates strong competitiveness. As illustrated in Sec. 4.2, our method achieves comparable results across all V1S, V1M, and AVSS-binary subsets, showcasing average performance improvements of up to 1.25% in mIoU and 2.7% in F-score, respectively. Furthermore, qualitative analysis indicates that our predicted masks exhibit superior quality, as demonstrated in sub-figure (1) of Fig. 4 (more examples in supplementary materials).

It is necessary to mention that our proposed TeSO does not show significant improvements on V1S over certain existing methods, such as BAVS [23], which also utilizes text semantic information. Apart from marginal effects, this phenomenon can also be attributed to the singular sound source and simplistic nature of the data scenarios present in V1S. On V1M, however, our performance significantly surpasses that of other methods, including BAVS.

In addition to the AVS task, we also conduct experiments on the Audio-Visual Semantic Segmentation (AVSS) task, as shown in Tab. 2. Our model exceeds the best performance in previous models by 5.37% on mIoU, which further demonstrates that our model is capable of establishing a better audio-visual correlation.

4.3 Bottlenecks and Fairness

Our method (TeSO) collects text cues from dense image descriptions at $\text{REC}_{\#1}$ (80.41%) and a filter with a dynamic mask at $\text{REC}_{\#2}$ (80.95%), indicating a robust startup for the text-guided AVS framework, in comparison with another two-stage model, BAVS.

Table 1: Performance on AVS-Benchmarks. The underscore is used to indicate suboptimal results. BAVS uses large pre-trained visual and audio foundation models to replace audio input with text tag. In comparison, TeSO (ours) enhances rather than replaces audio guidance based on the robust semantics of text. It is evident that our method has a good F-score result while ensuring mIoU.

Method	Audio-backbone	Visual-backbone	V1S		V1M		AVSS-binary	
			mIoU(%)	F-score	mIoU(%)	F-score	mIoU(%)	F-score
AVSBench [47]	VGGish	PVT-v2	78.70	0.879	54.00	0.645	62.45	0.756
AVSegFormer [8]	VGGish	PVT-v2	82.06	0.899	58.36	0.693	64.34	0.759
AVSC [22]	VGGish	PVT-v2	81.29	0.886	59.50	0.657	-	-
AVS-BG [11]	VGGish	PVT-v2	81.71	<u>0.904</u>	55.10	0.668	-	-
AQFormer [16]	VGGish	PVT-v2	81.60	0.894	61.10	0.721	-	-
†CATR [20]	VGGish	PVT-v2	81.40	0.896	59.00	0.700	-	-
AV-SAM [34]	ResNet18	ViT-Base	40.47	0.566	-	-	-	-
SAMA [30]	VGGish	ViT-Huge	81.53	0.886	63.14	0.691	-	-
GAVS [40]	VGGish	ViT-Base	80.06	0.902	63.70	<u>0.774</u>	<u>67.70</u>	<u>0.788</u>
AuTR [29]	VGGish	Swin-Base	80.40	0.891	56.20	0.672	-	-
MUTR [43]	VGGish	Video-Swin-Base	81.60	0.897	<u>64.00</u>	0.735	-	-
BAVS [23]	‡BEATs	Swin-Base	<u>82.68</u>	0.898	59.63	0.659	55.45	0.640
TeSO (ours)	VGGish	Swin-base	**83.27**	**0.933**	**66.02**	**0.801**	**68.53**	**0.813**

† : To make a fair comparison, the results of CATR here are without supplemented annotation of the training set.

‡ : BEATs is a stronger audio backbone, trained on AudioSet dataset with 90M parameters. In comparison, VGGish is trained on Youtube-8M with 70M parameters.

Table 2: Performance on AVSS dataset. Our method showcases a substantial improvement in the semantic subset, achieved by leveraging robust text semantics to enhance audio guidance.

Method	Backbone	AVSS	
		mIoU(%)	F-Score
AVSBench	PVT-v2	29.80	0.352
CATR	PVT-v2	32.80	0.385
BAVS	Swin-base	33.59	0.375
TeSO (Ours)	Swin-base	**38.96**	**0.451**

Table 3: Potential sounding objects recognition (REC) performance at different model stages and the corresponding segmentation results. Experiments are conducted on AVS-V1M test set.

Method	$\text{REC}_{(\#1)}$	$\text{REC}_{(\#2)}$	SEG
	AP(%)	AP(%)	mIoU(%)
BAVS	24.22	79.69	59.63
TeSO_p (ours)	77.96	78.33	63.87
TeSO_n (ours)	77.82	78.27	63.31
TeSO (ours)	**80.41**	**80.95**	**66.02**

In addition, we develop several variants of TeSO to investigate its performance sources and bottlenecks, as well as its fairness in comparison with other methods. TeSO_p denotes the variant using one-shot prompts, while TeSO_n denotes the variant employing a noun parser instead of a frozen LLM for text cues collection. As shown in Tab. 3, the results demonstrate that both the number (shots) of demonstrations and the method of text cues collection significantly impact the model's ability to accurately identify audio semantics and perform segmentation (SEG). Thus, utilizing few-shot prompts and off-the-shelf foundation models are both essential for ensuring the acquisition of high-quality text cues (*i.e.*, the bottleneck).

However, we would like to emphasize that, even without using LLM to capture text cues, our method variant (TeSO_n) still fairly achieves competitive performance (63.31% mIoU) compared with BAVS (59.63% mIoU) and state-of-the-art method GAVS (63.70% mIoU) in the complex multi-source scenario (AVS-V1M), and significantly outperforms other methods. This suggests that

Table 4: Impact on mIoU and F-score by muted audio or noise-only audio on AVS-V1M test set. TeSO (ours) exhibits the most intensive attenuation up to 48.8%. It is noteworthy that the other models still generate "satisfactory" results even in silent or purely noisy environments. This observation suggests that the models might rely more on the segmentation preferences formed during the training rather than truly learning efficient audio guidance.

	Mute		WGN-10dB		WGN-40dB	
Method	mIoU% (Δ_m%)	F-score (Δ_f%)	mIoU% (Δ_m%)	F-score (Δ_f%)	mIoU% (Δ_m%)	F-score (Δ_f%)
AVS-Bench	49.15 (6.6)	0.622 (4.9)	46.76 (11.1)	0.618 (5.5)	47.46 (9.8)	0.610 (6.7)
GAVS	42.53 (29.2)	0.695 (8.4)	41.51 (34.8)	0.692 (10.6)	41.87 (30.3)	0.696 (8.4)
BAVS	45.06 (26.1)	0.644 (5.8)	43.27 (27.4)	0.601 (8.8)	42.93 (31.6)	0.595 (9.7)
TeSO (ours)	37.89 (**42.6**)	0.671 (**16.2**)	34.20 (**49.6**)	0.661 (**17.5**)	34.74 (**47.4**)	0.656 (**18.1**)

leveraging text-enhanced audio for visual control with robust semantics is a viable and promising direction.

4.4 Audio Control

Based on our observations, we notice a distinct shape of the predicted mask in sub-figure (2) of Fig. 4 when some popular AVS baselines are used for inference. Interestingly, this shape remains distinguishable even when the inference is conducted with muted audio. This suggests that the network can predict rough masks without relying on audio input, indicating that the audio-visual grounding is not effectively guided by audio cues. As a result, this undesirable phenomenon can be attributed to the segmentation preference, which is a result of these models lacking proper audio control.

To evaluate the influence of audio control on segmentation, various models are subjected to inference using muted audio and noise-only audio (at 10 dB and 40 dB) instead of the corresponding audio signals. It should be noted that the visual features remained unchanged throughout the evaluation. As shown in Tab. 4, the disparity Δ_m (and Δ_f) refers to the difference in mIoU (and F-score) values before and after the audio substitution is applied as an evaluation metric for the sensitivity of audio changes. A larger drop in mIoU and F-score indicates a greater degree of audio control over the visual space. In addition to demonstrating strong segmentation capabilities in the AVS task, TeSO (ours) exhibits the highest level of attenuation, reaching up to 47.4%, when subjected to muted and noise-only audio. Additionally, qualitative analysis revealed that the predicted masks remained predominantly blank, as depicted in sub-figure (2) of Fig. 4 (more examples in supplementary materials). This observation indicates that our model has better audio guidance over the visual space, enabling it to play a crucial role in audio-visual grounding.

4.5 Ablation Study

As shown in Tab. 5, we conduct ablation experiments targeting the designed TeSO modules and the guidance from different modalities. To explore the guiding ability, we progressively remove either the audio or text guidance for TeSO.

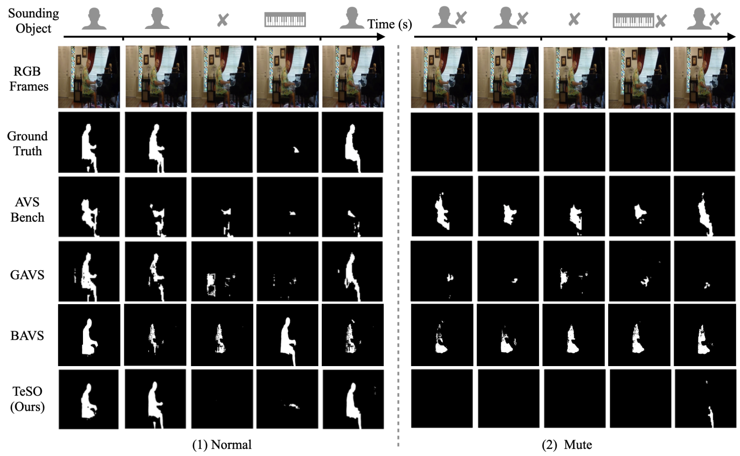

Fig. 4: Examples of the impact of normal audio input and all-mute audio on popular methods. In normal scenarios, our method shows better masks than previous methods. In all-mute scenarios, our approach exhibits strong sensitivity towards audio inputs, as it is capable of generating blank masks for silent audio clips.

The experiments indicate that control from both audio and text inputs for visual space can establish a strong audio-visual correlation. Furthermore, by utilizing our proposed TeSO, we achieve enhanced audio guidance and reach a 66.02% mIoU while incorporating precise semantic information from the text modality. Moreover, through the utilization of our proposed TeSO method that incorporates robust semantic information from the text modality, we achieved improved audio guidance and attained a mIoU of 66.02%.

As shown in Tab. 6, we also implement a baseline that directly filters the inferred semantic labels during the model evaluation process by intersecting them with the potential sounding object collected by the frozen LLM capturer. This implies that no text cues are utilized to enhance the audio guidance for better

Table 5: Ablation study on AVS-V1M. Our method benefits from using both audio and text as the guiding information for sounding objects.

Control	mIoU(%)	F-score
TeSO	**66.02**	**0.801**
(-) $\mathcal{L}_{infoNCE}$	65.34	0.793
(-) Dynamic Mask	64.78	0.781
(-) PMQS	62.90	0.774
(-) SeDAM	62.22	0.760
(-) Text	62.28	0.768
(-) Audio	42.11	0.692
(-) Text & Audio	37.35	0.656

Table 6: Comparison of our proposed baseline and TeSO. The baseline method employs LLM to filter the predicted results but does not enhance the audio guidance with text cues. Experiments are conducted on AVS-V1M test set.

	Normal	Mute	WGN-40dB
	mIoU%	mIoU% (Δ_m%)	mIoU% (Δ_m%)
Baseline	64.15	43.37 (32.4)	43.09 (32.8)
TeSO	**66.02**	37.89 (**42.6**)	34.74 (**47.4**)

control over the visual space. As expected, the sensitivity (Δ_m) of model performance to changes in audio input has decreased compared with TeSO, indicating the effectiveness of our method.

5 Discussion

Noisy text cues. Besides dense image captions for collecting semantic text cues, we also constructed a dynamic mask in Eq. (4) to further select effective semantics based on audio correspondence. We use audio recognition results to determine the accuracy of text cues. Please refer to Tab. 3 to compare the audio recognition results based on text cues collection stage ($\mathrm{REC}_{\#1}$) and dynamic mask filtering stage ($\mathrm{REC}_{\#2}$). Our approach achieves higher audio recognition performance, compared with using the audio foundation model in BAVS. This indicates the reliability and accuracy of our strategy for obtaining text cues.
Prompt with CoT. As shown in the Fig. 3, we utilize task-specific prompt templates with multiple scenario demonstrations and CoT instructions to assist the frozen LLM in collecting more accurate potential sounding objects as text cues. For the detailed prompt templates with CoT instructions and related experimental information, please refer to the Prompt Template part and Text Assistance part in the supplementary materials.
Versatility. We explore an alternative visual foundation model like the Segment Anything Model [19] for segmentation. Conducting experiments using a ViT-based backbone with minimal adjustments, our model achieves an impressive 65.36% mIoU on V1M, still getting comparable performance. This demonstrates the versatility and effectiveness of our model. Additional experiment details and analysis can be found in the supplementary materials.

6 Conclusion

This study represents a pioneering effort to incorporate text cues into the AVS task. Unlike previous approaches that solely rely on audio assistance, our TeSO method establishes a stronger audio-visual correlation by harnessing the robust semantics unique to the text modality By integrating CoT instructions and common-sense reasoning, we enhance the ability to capture potential sounding objects from detailed scene descriptions, thereby optimizing the utilization of text cues. Furthermore, we have developed a dynamic mask based on attention scores to effectively handle noisy information present in the text cues. Our comprehensive experiments highlight the effectiveness of our approach, demonstrating comparative performance across all AVS Benchmarks. In particular, more observation experiments validate that enhancing audio with text cues grants superior audio guidance. In summary, our work provides a new direction for building reliable audio-visual correlation with textual semantics, and we hope that it can help the community construct more robust audio-visual systems.

Acknowledgements

This research was supported by National Natural Science Foundation of China (NO.62106272), and Public Computing Cloud, Renmin University of China.

References

1. Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B., Vijayanarasimhan, S.: Youtube-8m: A large-scale video classification benchmark. arXiv preprint arXiv:1609.08675 (2016)
2. Bao, F., Nie, S., Xue, K., Li, C., Pu, S., Wang, Y., Yue, G., Cao, Y., Su, H., Zhu, J.: One transformer fits all distributions in multi-modal diffusion at scale. arXiv preprint arXiv:2303.06555 (2023)
3. Chen, S., Wu, Y., Wang, C., Liu, S., Tompkins, D., Chen, Z., Wei, F.: Beats: Audio pre-training with acoustic tokenizers. arXiv preprint arXiv:2212.09058 (2022)
4. Cheng, B., Choudhuri, A., Misra, I., Kirillov, A., Girdhar, R., Schwing, A.G.: Mask2former for video instance segmentation (2021)
5. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1290–1299 (2022)
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
7. Espejel, J.L., Ettifouri, E.H., Alassan, M.S.Y., Chouham, E.M., Dahhane, W.: Gpt-3.5, gpt-4, or bard? evaluating llms reasoning ability in zero-shot setting and performance boosting through prompts. Natural Language Processing Journal **5**, 100032 (2023)
8. Gao, S., Chen, Z., Chen, G., Wang, W., Lu, T.: Avsegformer: Audio-visual segmentation with transformer (2023)
9. Gemmeke, J.F., Ellis, D.P., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M.: Audio set: An ontology and human-labeled dataset for audio events. In: 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 776–780. IEEE (2017)
10. Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I.: Imagebind: One embedding space to bind them all. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15180–15190 (2023)
11. Hao, D., Mao, Y., He, B., Han, X., Dai, Y., Zhong, Y.: Improving audio-visual segmentation with bidirectional generation. arXiv preprint arXiv:2308.08288 (2023)
12. Van der Heiden, R.M., Janssen, C.P., Donker, S.F., Hardeman, L.E., Mans, K., Kenemans, J.L.: Susceptibility to audio signals during autonomous driving. PloS one **13**(8), e0201963 (2018)
13. Hershey, S., Chaudhuri, S., Ellis, D.P., Gemmeke, J.F., Jansen, A., Moore, R.C., Plakal, M., Platt, D., Saurous, R.A., Seybold, B., et al.: Cnn architectures for large-scale audio classification. In: 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 131–135. IEEE (2017)

14. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for nlp. In: International Conference on Machine Learning. pp. 2790–2799. PMLR (2019)
15. Hu, D., Wei, Y., Qian, R., Lin, W., Song, R., Wen, J.R.: Class-aware sounding objects localization via audiovisual correspondence. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(12), 9844–9859 (2021)
16. Huang, S., Li, H., Wang, Y., Zhu, H., Dai, J., Han, J., Rong, W., Liu, S.: Discovering sounding objects by audio queries for audio visual segmentation (2023)
17. Hur, C., Park, H.: Zero-shot image classification with rectified embedding vectors using a caption generator. Applied Sciences **13**(12), 7071 (2023)
18. Johnson, J., Karpathy, A., Fei-Fei, L.: Densecap: Fully convolutional localization networks for dense captioning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4565–4574 (2016)
19. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. arXiv:2304.02643 (2023)
20. Li, K., Yang, Z., Chen, L., Yang, Y., Xun, J.: Catr: Combinatorial-dependence audio-queried transformer for audio-visual video segmentation. arXiv preprint arXiv:2309.09709 (2023)
21. Ling, Y., Li, Y., Gan, Z., Zhang, J., Chi, M., Wang, Y.: Hear to segment: Unmixing the audio to guide the semantic segmentation (2023)
22. Liu, C., Li, P., Qi, X., Zhang, H., Li, L., Wang, D., Yu, X.: Audio-visual segmentation by exploring cross-modal mutual semantics (2023)
23. Liu, C., Li, P., Zhang, H., Li, L., Huang, Z., Wang, D., Yu, X.: Bavs: Bootstrapping audio-visual segmentation by integrating foundation knowledge. arXiv preprint arXiv:2308.10175 (2023)
24. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning (2023)
25. Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J.: Llava-next: Improved reasoning, ocr, and world knowledge (January 2024), `https://llava-vl.github.io/blog/2024-01-30-llava-next/`
26. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning (2023)
27. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. arXiv preprint arXiv:2304.08485 (2023)
28. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36** (2024)
29. Liu, J., Ju, C., Ma, C., Wang, Y., Wang, Y., Zhang, Y.: Audio-aware query-enhanced transformer for audio-visual segmentation (2023)
30. Liu, J., Wang, Y., Ju, C., Zhang, Y., Xie, W.: Annotation-free audio-visual segmentation. arXiv preprint arXiv:2305.11019 (2023)
31. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
32. Ma, J., Sun, P., Wang, Y., Hu, D.: Stepping stones: A progressive training strategy for audio-visual semantic segmentation. IEEE European Conference on Computer Vision (ECCV) (2024)
33. Majumder, S., Al-Halah, Z., Grauman, K.: Move2hear: Active audio-visual source separation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 275–285 (2021)

34. Mo, S., Tian, Y.: Av-sam: Segment anything model meets audio-visual localization and segmentation. arXiv preprint arXiv:2305.01836 (2023)
35. Park, S., Senocak, A., Chung, J.S.: Marginnce: Robust sound localization with a negative margin. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023)
36. Senocak, A., Oh, T.H., Kim, J., Yang, M.H., Kweon, I.S.: Learning to localize sound source in visual scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4358–4366 (2018)
37. Sharma, H., Jalal, A.S.: Image captioning improved visual question answering. Multimedia tools and applications **81**(24), 34775–34796 (2022)
38. Shi, Z., Zhou, X., Qiu, X., Zhu, X.: Improving image captioning with better use of captions. arXiv preprint arXiv:2006.11807 (2020)
39. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C.C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P.S., Lachaux, M.A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E.M., Subramanian, R., Tan, X.E., Tang, B., Taylor, R., Williams, A., Kuan, J.X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., Scialom, T.: Llama 2: Open foundation and fine-tuned chat models (2023)
40. Wang, Y., Liu, W., Li, G., Ding, J., Hu, D., Li, X.: Prompting segmentation with sound is generalizable audio-visual source localizer. arXiv preprint arXiv:2309.07929 (2023)
41. Wang, Y., Sun, P., Zhou, D., Li, G., Zhang, H., Hu, D.: Ref-avs: Refer and segment objects in audio-visual scenes. IEEE European Conference on Computer Vision (ECCV) (2024)
42. Wu, W., Yao, H., Zhang, M., Song, Y., Ouyang, W., Wang, J.: Gpt4vis: What can gpt-4 do for zero-shot visual recognition? arXiv preprint arXiv:2311.15732 (2023)
43. Yan, S., Zhang, R., Guo, Z., Chen, W., Zhang, W., Li, H., Qiao, Y., He, Z., Gao, P.: Referred by multi-modality: A unified temporal transformer for video object segmentation. arXiv preprint arXiv:2305.16318 (2023)
44. Yang, Z., Li, L., Lin, K., Wang, J., Lin, C.C., Liu, Z., Wang, L.: The dawn of lmms: Preliminary explorations with gpt-4v (ision). arXiv preprint arXiv:2309.17421 **9**(1), 1 (2023)
45. Zhang, H., Li, X., Bing, L.: Video-llama: An instruction-tuned audio-visual language model for video understanding (2023)
46. Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., Torralba, A.: Semantic understanding of scenes through the ade20k dataset. International Journal of Computer Vision **127**(3), 302–321 (2019)
47. Zhou, J., Wang, J., Zhang, J., Sun, W., Zhang, J., Birchfield, S., Guo, D., Kong, L., Wang, M., Zhong, Y.: Audio-visual segmentation. In: European Conference on Computer Vision (2022)
48. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020)
49. Zürn, J., Burgard, W.: Self-supervised moving vehicle detection from audio-visual cues (2022)