

Object-aware Sound Source Localization via Audio-Visual Scene Understanding

–Supplementary Material–

This manuscript provides additional implementation details and additional results of our proposed method. In Section 1, we elaborate on the additional implementation details of our method. Section 2 presents additional experimental results. Moreover, Section 3 shows additional visualization results. Note that [PXX] indicates the reference in the main paper.

1. Additional Implementation Details

As mentioned in the main paper, we utilize the ResNet-18 [P12] for the audio encoder. At this time, since the audio spectrogram has only one channel, we modify the first convolution layer of the audio encoder to have an input channel of 1 and an output channel of 64, utilizing a kernel size of 7, stride of 2, and padding of 3. Additionally, we employ the Adam optimizer, setting the parameters (β_1, β_2) to (0.9, 0.999), which are the standard values for Adam. Following [P3], we set the hyperparameters $\alpha_p = 0.65$, $\alpha_n = 0.4$, and $\omega = 0.03$ mentioned in Section 3.3 of main paper. For our Object Region Isolation loss \mathcal{L}_{ori} in Section 3.4 of main paper, we utilize the Sinkhorn algorithm [P6] with a maximum of 100 iterations. Since a reference associated map \mathbf{S}_r is a 2D spatial region, we incorporate both the pixel intensity differences and the Euclidean distance between the spatial coordinates of each element during the distance matrix computation.

We utilize a Multimodal Large Language Model to generate foreground captions for sound-making objects and background captions for non-sound-making objects in diverse and complex scenarios. As shown in Table 5, we provide prompts to guide audio-visual understanding across the following scenarios: (1) Scenarios with multiple objects, including a sound-making one, (2) Scenarios with visually similar objects, distinguishing sound-making ones, and (3) Scenarios with multiple sound-making elements. These generated foreground and background captions are employed to facilitate the learning of fine-grained audio-visual correspondence. Examples of the generated captions are provided in Section 3.

Algorithm 1 provides additional details on the Object Region Isolation loss \mathcal{L}_{ori} described in Section 3.4 in main paper. It clarifies how spatial distinctiveness is enforced between sound-making object regions and background regions by minimizing overlaps through Wasserstein Distance computation using the Sinkhorn algorithm.

2. Additional Experiments

Evaluating Generalization Across Different MLLMs.

We evaluate the generalization capability of our method

Algorithm 1 Object Region Isolation Loss Function

Require: $\mathbf{F}_v \in \mathbb{R}^{w \times h \times c}$, $\mathbf{F}_r \in \mathbb{R}^{(K+1) \times c}$
Ensure: $\bar{\mathbf{S}}_r = \text{Sim}(\mathbf{F}_v, \mathbf{F}_r) \in \mathbb{R}^{w \times h \times (K+1)}$, $\mathcal{L} = 0$
for $i, j \in K + 1, i \neq j$ **do**
 $\mathcal{L} = \mathcal{L} + \text{Sinkhorn}(\bar{\mathbf{S}}_r[:, i], 1 - \bar{\mathbf{S}}_r[:, j])$
end for
Results: $\mathcal{L}_{ori} = \mathcal{L}$

Method	MLLM	CAP(%)	CloU@0.3(%)	AUC(%)
NoPrior [P18]	–	32.5	46.9	29.2
Proposed Method	LLaVa-NeXT [1]	41.4	49.8	42.2
	Qwen2-VL [3]	43.1	54.4	43.9
	InternVL2 [P5]	45.9	55.2	44.8

Table 1. Experimental results on the VGGSound-Duet test set using different Multimodal Large Language Models (MLLMs).

Method	Text Encoder	CAP(%)	CloU@0.3(%)	AUC(%)
NoPrior [P18]	–	32.5	46.9	29.2
Proposed Method	CLIP [4]	43.7	54.0	42.7
	BERT [5]	45.9	55.2	44.8

Table 2. Experimental results on the VGGSound-Duet test set using different text encoders for generated caption.

by conducting experiments with various Multimodal Large Language Models (MLLMs). In addition to InternVL2.0-8B [P5], which is used in the main paper, we include two widely adopted MLLMs in recent research: LLaVA-NeXT [1] with Mistral-7B [2] and Qwen2-VL-7B [3]. These two models are also guided using the same prompt to generate foreground and background captions.

As shown in Table 1, we test the models on the VGGSound-Duet [P2] test set, where InternVL2.0-8B achieves the best performance, followed by Qwen2-VL and LLaVA-NeXT. While performance slightly varies depending on the chosen MLLM, all models consistently outperform the current state-of-the-art method, NoPrior [P18]. These results demonstrate that our approach is effective across diverse MLLMs, consistently showing superior performance in generating fine-grained audio-visual correspondences.

Performance Variation with Different Text Encoders.

We present an additional experiment to validate the robustness of our approach with various text encoders for generated captions in Section 3.2 of the main paper. While the main paper used BERT as the default text encoder, we additionally adopt CLIP [4], widely used for image-text multimodal learning, to further investigate the generalization ability of our method across different text encoders.

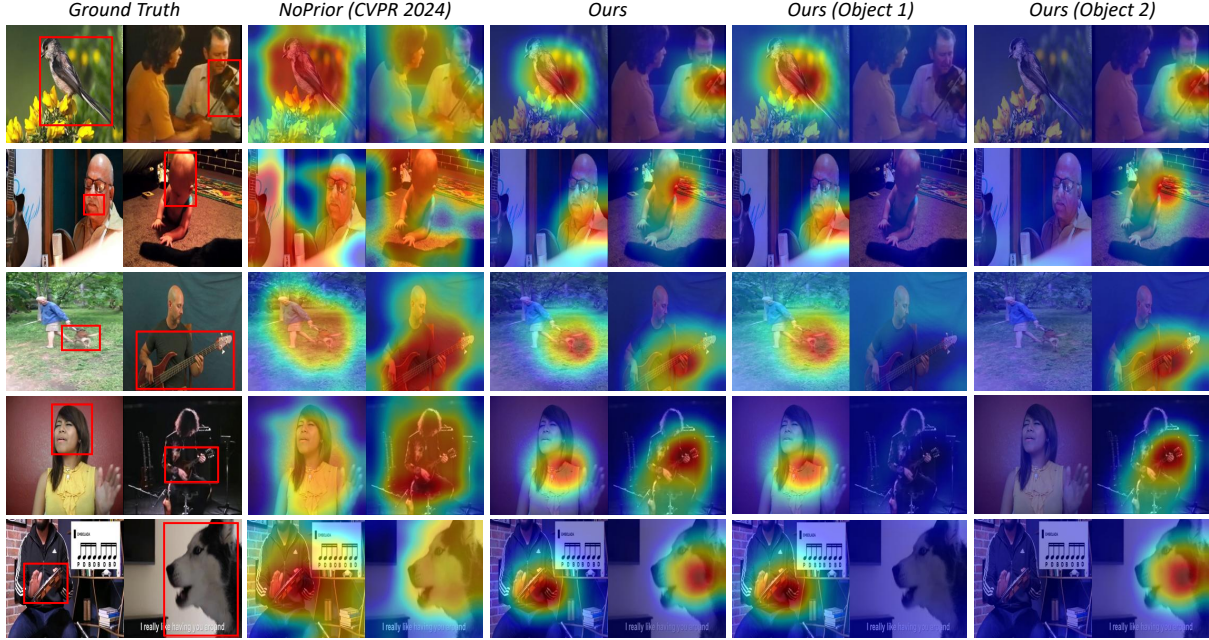


Figure 1. Additional visualization results for VGGSound-Duet test set.

Method	λ_1	λ_2	CAP(%)	CloU@0.3(%)	AUC(%)
NoPrior [P18]	—	—	32.5	46.9	29.2
		10	42.2	51.1	40.6
	1	1	45.4	53.5	44.4
		0.1	45.9	55.2	44.8
Proposed Method		0.01	45.2	51.5	42.2
	10		43.1	53.3	42.9
	1	0.1	45.9	55.2	44.8
	0.1		43.2	51.3	42.1
	0.01		42.6	50.1	41.2

Table 3. Experimental results on the VGGSound-Duet test set using different balancing parameters λ_1 and λ_2 .

We utilize the VGGSound-Duet test set to compare performance. As shown in Table 2, our method significantly outperforms the existing approaches across both tested text encoders. These results indicate the effectiveness and generalization ability of our model in incorporating different text encoder models.

Experiment on Balance Parameters λ_1 and λ_2 used in Loss Function. To evaluate the effect of the balancing parameters λ_1 and λ_2 in our total loss function in Section 3.5 of main paper, we perform an additional study. As shown in Table 3, our model achieves optimal performance when $\lambda_1 = 1$ and $\lambda_2 = 0.1$. Remarkably, our method consistently surpasses the existing approach across a range of balancing parameters. These results demonstrate that even when varying the hyperparameters, our method maintains consistently high performance with minimal

Method	MLLM	Training		Inference	
		time (s) (per image)	time (s) (per iter) memory (GB)	time (s) (per image)	memory (GB)
T-VSL [P23]	—	—	1.53 22.0	0.051	0.81
Ours		0.92	1.13 16.63	0.044	0.35

Table 4. The comparisons of training time, inference time, and the number of parameters.

variation, indicating that our approach does not heavily rely on hyperparameter tuning.

Computational Cost. We compare the computational efficiency of our method with T-VSL [P23], which uses AudioCLIP as its backbone. As shown in Table 4, despite incorporating MLLMs during training, our approach maintains comparable efficiency to state-of-the-art methods. For the MUSIC dataset (50K samples, 100 epochs), T-VSL requires 47.8h for training, while our method takes 48.1h. Importantly, MLLMs are only used at the beginning of training for caption generation and not during inference, ensuring real-world applicability. Our method demonstrates improved efficiency with 13.7% faster inference time and lower memory usage compared to T-VSL [P23], due to our simpler architecture and the absence of MLLMs at test time.

3. Additional Visualization Results

Visualization Results on VGGSound-Duet. We provide additional visualization results in Figure 1 to demonstrate the effectiveness of our method in achieving fine-grained audio-visual localization. Our approach identifies sound-

making objects, leveraging the Object-aware Contrastive Alignment loss \mathcal{L}_{oca} to focus exclusively on sound-making regions. Building on this, the Object Region Isolation loss \mathcal{L}_{ori} enhances the ability of model to separate multiple sound-making objects in multi-source scenarios, ensuring precise isolation of each source. Together, these two loss functions enable the model to handle diverse audio-visual scenes effectively. These results highlight the effectiveness of our approach in improving audio-visual localization accuracy across a wide range of challenging scenarios.

Quality of the Generated Captions. We visualize the generated foreground and background captions, along with the corresponding image and audio class information. The captions are generated using the InternVL 2.0-8B model, guided by a carefully crafted prompt (refer to Table 5). Foreground captions describe the sound-making objects corresponding to the provided audio class, while background captions capture the silent objects visible in the image. Figure 2 demonstrates the consistency and relevance of the captions in relation to the audio-visual scenes, highlighting how effectively MLLMs capture both sound-making and silent objects in diverse scenarios.

Video Demo. We provide supplementary video materials that offer a more in-depth explanation of our method for localizing sound-making objects in complex environments. These videos demonstrate the real-time applicability and robustness of our approach under various conditions. We provide the results of our method with some examples from the VGGSound-Single and VGGSound-Duet datasets, illustrating its ability to distinguish sound-making objects from silent ones effectively. Please refer to the video titled “CVPR2025_SubmissionID_698_Supp_Demo.mp4”.

References

- [1] Liu, Haotian, et al. Visual instruction tuning. In *NeurIPS*, 2024.
- [2] Jiang, Albert Q., et al. Mistral 7B. In *arXiv preprint*, 2023.
- [3] Wang, Peng, et al. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. In *arXiv preprint*, 2024.
- [4] Radford, Alec, et al. Learning transferable visual models from natural language supervision. In *ICML*. PMLR, 2021.
- [5] Devlin, Jacob, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *arXiv preprint*, 2019.
- [P2] Honglie Chen, et al. Vggsound: A large-scale audio-visual dataset. In *ICASSP*, 2020.
- [P3] Honglie Chen, et al. Localizing visual sounds the hard way. In *CVPR*, 2021.
- [P5] Zhe Chen, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, 2024.
- [P6] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *NeurIPS*, 2013.
- [P12] Kaiming He, et al. Deep residual learning for image recognition. In *CVPR*, 2016.
- [P18] Dongjin Kim, et al. Learning to visually localize sound sources from mixtures without prior source knowledge. In *CVPR*, 2024.
- [P23] Tanvir Mahmud, et al. T-vsl: Text-guided visual sound source localization in mixtures. In *CVPR*, 2024.

Analyze the provided image along with its associated class label, which identifies an object or element in the image that emits sound. The scene is complex, containing multiple objects, and requiring categorization based on the examples below.

Instructions:

1. Identify foreground (sound-related) elements: These are objects in the image emitting sounds that match the class description.
2. Identify background (sound-unrelated) elements: These are distinct objects visible in the image but unrelated to the sound described by the class.
3. Focus strictly on what is visible in the image. Do not infer or describe unseen objects.

Output Format:

The response must always be in JSON format with structured sentences that start with ‘an image of...’. If there are two or more class labels (separated by commas), the foreground must be provided as a list of sound-making elements.

Examples:

(1) Scenario with multiple objects, including a sound-making one

Input:

- image: example_image_1
- class label: man_blowing_whistle

Output:

```
{
  "foreground": "an image of a man blowing a whistle",
  "background": "an image of mountains, desert landscape, and sky"
}
```

(2) Scenario with visually similar objects, distinguishing sound-making ones

Input:

- image: example_image_2
- class label: acoustic_guitar

Output:

```
{
  "foreground": "an image of a man playing guitar",
  "background": "an image of non-playing guitars, drum-set, and amp"
}
```

(3) Scenario with multiple sound-making elements

Input:

- image: example_image_3
- class label: clarinet, violin

Output:

```
{
  "foreground": ["an image of playing clarinet", "an image of playing violin"],
  "background": "an image of the kitchen, curtains, and piano in the background"
}
```

Now, process the provided input following the same structure and RETURN ONLY the JSON FORMAT.

Table 5. Guiding prompt for foreground and background caption generation using a Multimodal Large Language Model (MLLM)




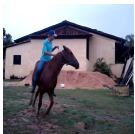
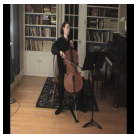

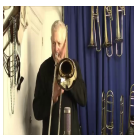
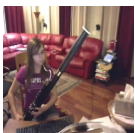
Image	Audio Class	Generated Captions
	<i>femele_singing</i>	Foreground : "an image of singing female" Background : "an image of dining table, chairs, windows, and lamp"
	<i>race_car_auto_racing</i>	Foreground : "an image of a race car racing on a track" Background : "an image of spectator, fence, tree, building, and banner"
	<i>playing_erhu</i>	Foreground : "an image of playing erhu" Background : "an image of framed pictures, lamp, Buddha statue, cupboard, decorative screens, calligraphy scroll, purple tablecloth, vase, and door"
	<i>horse_clip-clop</i>	Foreground : "an image of a horse making a clip-clop sound" Background : "an image of a man wearing blue jeans and a blue shirt, an old beige house, large puddle on the ground, a small white bucket, and a wooden wheelbarrow"
	<i>playing_cello</i>	Foreground : "an image of playing cello" Background : "an image of room, piano, books, and carpet"
	<i>motorboat_speedboat_acceleration</i>	Foreground : "an image of a motorboat" Background : "an image of trees, a fence, and a white truck"
	<i>playing_trombone</i>	Foreground : "an image of a man playing the trombone" Background : "an image of brass musical instruments, room curtain, shelf, and a microphone"
	<i>playing_basson</i>	Foreground : "an image of girl playing bassoon" Background : "an image of couch, books, computer desk, laptop, and table"

Figure 2. Visualization results of generated captions on VGGSound train set.