Generate, Analyze, and Refine: Training-Free Sound Source Localization via MLLM Meta-Reasoning
–Supplementary Material–

This supplementary material provides additional implementation details and extended experimental results for the proposed method. First, it presents experimental results analyzing the impact of threshold adjustments at each stage, with the number of iterations in the Analysis stage fixed at 5, based on the VGGSound dataset. Next, it shows comparative results between various prompt variations and the proposed method, and further explains the effectiveness of the approach through additional visualization materials. Finally, it discloses the detailed prompts used throughout the entire Generate-Analysis-Refinement process.

Table of Contents

1. Additional Experimental Results

We analyzed the impact of the Audio Confidence threshold (A_C) in Stage 1 (Generation) with the number of iterations fixed at $N = 5$. Table S.1 presents results on the VGGSound-Single [P16] dataset. Performance remains stable despite variations in the audio confidence (A_C) and Audio-Visual Consistency (AV_C) thresholds, with the best performance achieved at A_C=0.75 and AV_C=0.5. For single-sound source datasets VGGSound-Single, while the SOTA performance is AP 51.7, IoU@0.5 47.3, and AUC 44.9, our proposed method significantly surpasses these with AP 60.5, IoU@0.5 60.2, and AUC 55.2. These results demonstrate that the proposed method is robust to threshold variations while consistently maintaining strong performance on the VGGSound-Single dataset.

Table S.1. Analysis of the impact of Audio Confidence (A_C) and Audio-Visual Consistency (AV_C) thresholds on performance. The number of iterations in the Stage 2 (Analysis) is fixed at $N = 5$. Evaluated on the VGGSound-Single dataset.

		VGGSound-Single [P16]		
A_C	AV_C	**AP**	**IoU@0.5**	**AUC**
0.5	0.5	60.1	60.0	55.0
0.75	0.5	**60.5**	**60.2**	**55.2**
0.5	0.75	60.1	60.0	55.0
0.75	0.75	60.2	60.1	55.1

2. Prompt Variation Comparison

In this experiment, we design four different prompt-based methods that apply varying conditions and constraints to perform Sound Source Localization in a more fine-grained.

Method 1 (Direct Estimation): This method represents the simplest approach, directly generating multiple candidate bounding boxes from the image and audio. The generated candidates are self-examined to assess the appropriateness of bounding box sizes and identify positional errors, with suggestions for improvements. Finally, based on the inspection results, the bounding boxes are refined to select the optimal candidate. When refinement is needed, a conservative rule of adjusting by at least 1 pixel is applied, serving as a basic calibration that quickly validates the initial box.

Method 2 (Class-Conditional Refinement): This method applies stronger structural constraints than the Method 1. First, an initial bounding box is estimated from the image and audio, and separately, the audio source class (*e.g.,* "violin", "dog barking") is extracted using only the audio. Subsequently, refinement is performed by considering both the initial bounding box and the extracted audio class together, ensuring that the bounding box logically aligns with the audio source class.

Method 3 (Anchor-Guided Refinement): This method extends the Method 2 by providing more detailed analysis information. Beyond the audio class and initial bounding box, it explicitly identifies visual sub-parts (anchors) that generate the sound source. For example, in the case of a violin, anchors such as "bow-string contact point" and "violin body" are identified. The model analytically interprets the relationships among the audio class, initial bounding box, and visible anchors to perform refinement. This method focuses on identifying and utilizing fine-grained parts of the sound source.

Our Method (Generation-Analysis-Refinement): The final our method extends the Method 3 and represents the final approach proposed in this paper. In this method, all meta-analysis information including Audio-Visual Consistency (AV_C), role tags, and anchor votes is provided as input, designed to enable the model to comprehensively verify judgments from previous stages. Additionally, we

Table S.2. Performance comparison of various prompt variation methods on single-sound source datasets. Method 1 performs basic refinement with minimal adjustments. Method 2 incorporates audio class information for refinement. Method 3 leverages detailed analysis information including visual anchors. Ours represents the proposed meta-analysis-based method with varying iteration counts (N=1, 3, 5). Evaluated on VGGSound-Single and MUSIC-Solo datasets using CAP, CIoU@0.3, and AUC metrics.

Method	VGGSound-Single [P16]			MUSIC-Solo [P13]		
	AP	IoU@0.5	AUC	AP	IoU@0.5	AUC
Method 1	52.0	46.5	44.5	81.4	96.5	78.8
Method 2	59.5	59.0	54.2	82.7	98.9	80.2
Method 3	60.0	59.7	54.9	81.6	97.6	79.1
Ours ($N = 1$)	60.1	60.0	55.0	78.8	96.3	76.8
Ours ($N = 3$)	60.2	60.1	55.0	78.9	96.2	76.9
Ours ($N = 5$)	**60.5**	**60.2**	**55.2**	**80.6**	**98.5**	**78.2**

Table S.3. Performance comparison of various prompt variation methods on multi-sound source datasets. Method 1 performs basic refinement with minimal adjustments. Method 2 incorporates audio class information for refinement. Method 3 leverages detailed analysis information including visual anchors. Ours represents the proposed meta-analysis-based method with varying iteration counts (N=1, 3, 5). Evaluated on VGGSound-Duet and MUSIC-Duet datasets using CAP, CIoU@0.3, and AUC metrics.

Method	VGGSound-Duet [P16]			MUSIC-Duet [P13]		
	CAP	CIoU@0.3	AUC	CAP	CIoU@0.3	AUC
Method 1	44.7	57.0	37.7	46.5	77.9	45.1
Method 2	32.9	23.0	26.5	44.7	36.1	44.6
Method 3	45.5	60.4	39.5	53.6	76.9	49.4
Ours ($N = 1$)	43.4	58.9	38.1	54.7	80.8	51.4
Ours ($N = 3$)	43.5	59.5	38.2	54.7	80.8	51.4
Ours ($N = 5$)	**47.2**	**77.6**	**45.8**	**56.7**	**82.7**	**53.2**

analyze the progressive improvement effect by adjusting the number of iterations N in the refinement stage (N=1, 3, 5). Through this, we systematically compare the performance of each method and demonstrate the superiority of the proposed approach.

Using the above-mentioned four methods described above, we compare performance across single-sound and multi-sound source settings. Table S.2 and Table S.3 summarize the prompt variation experiment results for single-sound source and multi-sound source datasets. Ours ($N = 5$) demonstrates the best overall performance, achieving 60.5% AP on VGGSound-Single [P16], 80.6% AP on MUSIC-Solo [P13], 47.2% CAP on VGGSound-Duet [P16], and 56.7% CAP on MUSIC-Duet. Performance progressively improves from Method 1 to the proposed method, and also consistently enhances as the number of iterations *N* increases. This clearly confirms the effectiveness of integrating meta-analysis information and iterative refinement.

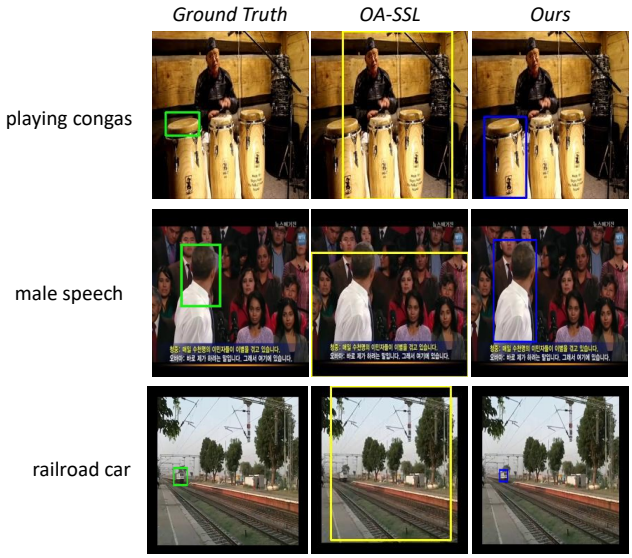

Figure S.1. Visualization of sound source localization results in VGGSound-Single [P16] dataset. Each row represents a different example, and each column shows the original image, Ground Truth (actual sound source location), prediction results of the OA-SSL [P38] method, and prediction results of the proposed method (Ours).

3. Additional Visualization Results

Figures S.1, S.2, S.3 visually compare the sound source localization results of the proposed method and the existing method (OA-SSL [P38]). In each example, Ground Truth represents the actual location of the sound source, OA-SSL shows the prediction results of the existing method, and Ours indicates the results of the proposed method.

Figure S.1 shows single-source results, where our method consistently produces tighter and more correctly positioned bounding boxes than OA-SSL [P38] across all examples. Figure S.2 further demonstrates improved precision on MUSIC-Solo [P13] with saxophone and flute cases, where our method more accurately aligns with the actual sound-producing regions. Figure S.3(a) illustrates multi-source scenarios involving two instruments. While OA-SSL struggles with scale and placement, our approach more clearly separates and localizes each source. Figure S.3(b) presents more challenging multi-source scenes with visually separated sources. Our method maintains accurate and compact localization, whereas OA-SSL often generates overly large regions. Overall, our approach yields consistently tighter and more reliable localization than OA-SSL across both single-source and multi-source settings.

4. Prompts for Proposed Method

In this study, we design a structured prompt framework to process visual and audio information in a step-by-step man-

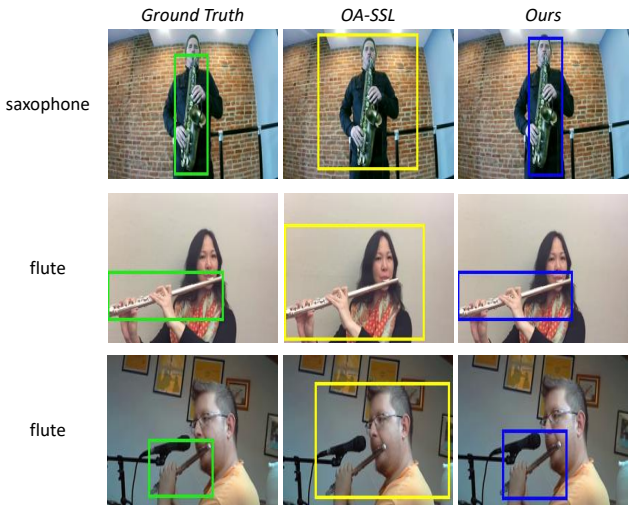

Figure S.2. Visualization of sound source localization results on the MUSIC-Solo [P13] dataset. Each row shows a different instrument example. From left to right, we present the Ground Truth bounding box, the prediction produced by OA-SSL [P38], and the prediction generated by our proposed method (Ours).

ner. Stage 1 (Generation) consists of two sub-stages: Stage 1 (Generation): Audio-Visual Localization Table S.4 estimates the location of the primary sound source object by utilizing both image and audio, while Stage 1 (Generation): Audio Classification Table S.5 classifies sound events based solely on audio. Stage 2 (Analysis) Table S.6 verifies whether the predicted sound source actually matches visually based on the information generated in Stage 1 (Generation), and quantifies this as Audio-Visual Consistency (AV_C). Finally, Stage 3 (Refinement) Table S.7 refines the bounding box based on these analysis results, achieving meaningful improvements with minimal changes. These three stages of the prompt framework are combined to enable step-by-step and interpretable single-sound source and multi-sound source localization without training.

References

[P13] Zhao Hang, Gan Chuang, Rouditchenko Andrew, Vondrick Carl, McDermott Josh, and Torralba Antonio. The sound of pixels. In *ECCV*, 2018.

[P16] Chen Honglie, Xie Weidi, Vedaldi Andrea, and Zisserman Andrew. Vggsound: A large-scale audio-visual dataset. In *ICASSP*, 2020.

[P38] Sung Jin Um, Dongjin Kim, Sangmin Lee, and Jung Uk Kim. In *CVPR*, 2025.

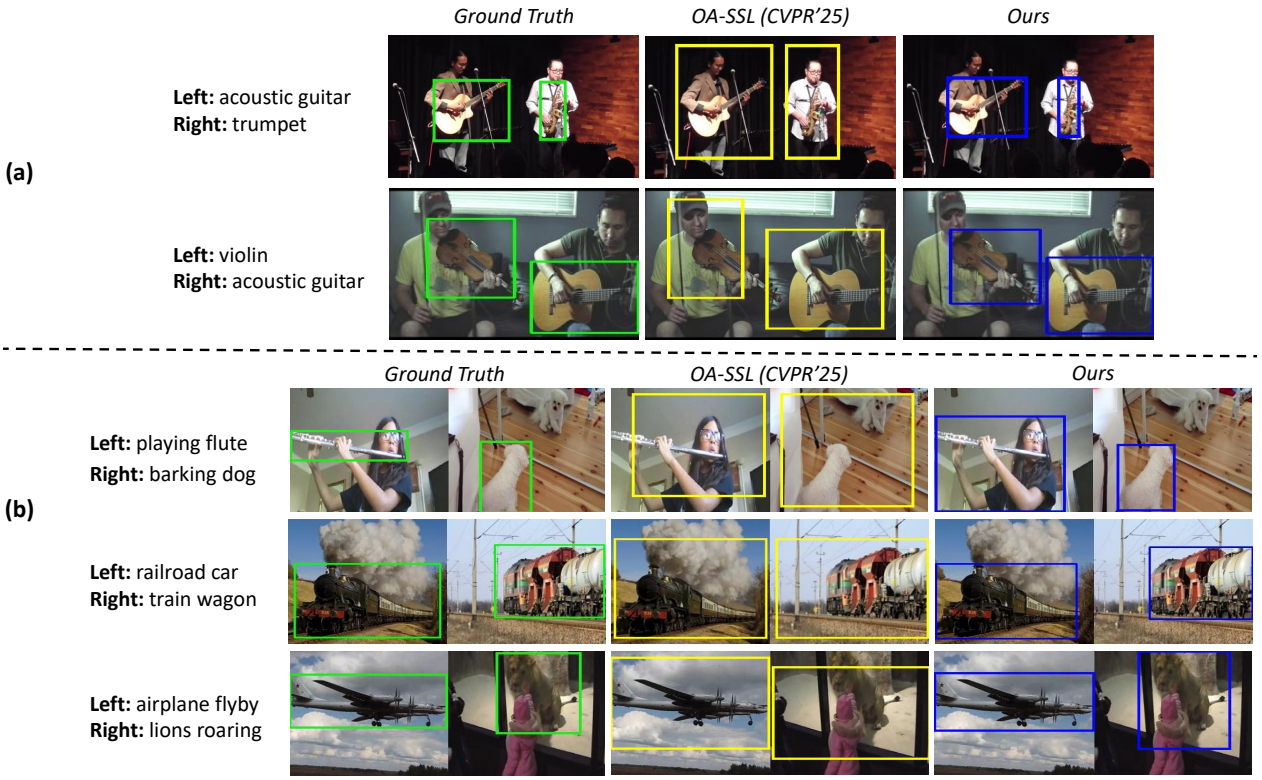

Figure S.3. Visualization of sound source localization results in VGGSound-Duet [P16] and MUSIC-Duet [P13] datasets. **(a)** Results of simultaneously localizing two instrumental sound sources in VGGSound-Duet [P16]. **(b)** Results in more complex multi-sound source environments in MUSIC-Duet [P13]. Each example includes the original image, Ground Truth, OA-SSL [P38] prediction, and the proposed method prediction (Ours).

Table S.4. Stage 1 (Generation) Prompt: Audio-Visual Localization

Prompt:

You are an assistant for audio-visual sound source localization (SSL).

TASK (Stage A):
Given an IMAGE and an AUDIO clip from the same scene:
1) Locate exactly one *main* sound-emitting object in the image and output its bounding box as $[x1, y1, x2, y2]$.
2) Provide a concise visual description of the sound-emitting object.

STRICT OUTPUT:

```
{
  "bbox": [x1, y1, x2, y2],
  "description": "visual description of the
                  sound-emitting object"
}
```

- The bbox must be four integers in the original image coordinates (x1¡x2, y1¡y2).
- Do not output any text or fields outside the JSON object.

Table S.5. Stage 1 (Generation) Prompt: Audio Classification

Prompt:

You are an audio classification expert.

TASK (Stage B):
Listen to the AUDIO and classify the dominant audio event using a short, lowercase class name (e.g., “violin”, “piano”, “dog barking”, “engine”, “drum set”).
You must also provide a confidence score in the range $[0.0, 1.0]$.

STRICT OUTPUT:

```
{
  "audio_class": "<concise class name>",
  "audio_confidence_score": <float>
}
```

- The class name must be lowercase and concise.
- The confidence must be a float between 0.0 and 1.0.
- Do not include any text outside the JSON.

Table S.6. Stage 2 (Analysis) Prompt

Prompt:

You must verify whether the sound suggested by the AUDIO is actually *visibly supported* within the IMAGE.
You must rely only on the given image–audio pair and must not hallucinate unseen content.

Context:
- previous_bbox
- audio_class
- audio_confidence_score
- image size $[W \times H]$

Definitions:

- **anchor_votes**: propose 0–5 concise, lowercase visual anchors that represent visible causes of the sound indicated by the audio class.
  Examples:
  - applause → "hands_clapping"
  - violin → "bow_on_strings", "violin_body"
  - dog barking → "dog_mouth_open"

  Format:
  `{"anchor":"<token_with_underscores>", "score": s}`

where $s \in [0, 1]$.

- **role_tags**: up to four short tokens summarizing the visual roles or cues relied upon.
- **av_consistency**: audio–visual consistency score $[0, 1]$, based on
  (i) alignment between audio class and visible evidence,
  (ii) spatial proximity to previous bbox,
  (iii) clarity of the visible cues.
- **keep**: true only when refinement can be safely skipped.

STRICT OUTPUT:

```
{
  "av_consistency": <float>,
  "role_tags": [...],
  "anchor_votes": [...],
  "keep": <true|false>
}
```

Table S.7. Stage 3 (Refinement) Prompt

Prompt:

You refine the bounding box of the main sound-emitting object by integrating IMAGE, AUDIO, and Stage 2 analysis results.

Context:
- previous_bbox
- audio_class
- image size $W \times H$
- av_consistency, role_tags, anchor_votes, keep

Refinement Rules:
1) Produce a final bbox that best matches the audio class and verified visual anchors, while minimizing unnecessary change.
2) The bbox must remain inside the image bounds $[0, W-1] \times [0, H-1]$ and satisfy $x1 < x2, y1 < y2$.
3) Unless the previous box is clearly incorrect, limit coordinate adjustments to within ±MAX_DELTA_PX per side.
4) Optionally describe the modification using an "ops" field: delta, expand, shrink, or recenter.
5) Provide a factual refined_description consisting of 2–4 sentences describing the scene and its relation to the audio class.

STRICT OUTPUT:

```
{
  "bbox": [x1, y1, x2, y2],
  "changed": true/false,
  "ops": {...} | null,
  "refined_description": "..."
}
```