STAR-AVL: STRUCTURED TEMPORAL-AWARE REASONING FOR TRAINING-FREE VIDEO-CENTRIC AUDIO-VISUAL LOCALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Video-centric Audio-Visual Localization (AVL) aims to identify the spatial location of sound-emitting objects by jointly interpreting audio and temporally changing visual information. Existing methods mainly rely on feature-level similarity between audio and visual embeddings, which limits their ability to explicitly reason about the audio-visual context and re-evaluate predictions using temporal information. To address these limitations, we propose STAR-AVL (**S**tructured **T**emporal-**A**ware **R**easoning for **A**udio-**V**isual **L**ocalization), a training-free structured reasoning framework built on Multimodal Large Language Models (MLLMs). STAR-AVL consists of two complementary components, Audio-Visual Reasoning and Temporal Context Calibration, which are together organized into four reasoning steps. Audio-Visual Reasoning identifies the actual sound-emitting objects by verifying the audio-visual evidence, while Temporal Context Calibration resolves the remaining localization ambiguity by leveraging the surrounding temporal context. On the AVATAR dataset, STAR-AVL consistently outperforms existing video-centric AVL and CoT reasoning methods. Additional experiments on VGGSound-Duet demonstrate its generalization to image-centric AVL settings. The source code will be publicly available.

1 INTRODUCTION

Audio-Visual Localization (AVL) aims to identify the spatial location of sound-emitting objects by jointly using audio and visual information. Unlike traditional image-centric AVL (Senocak et al., 2018; Hu et al., 2019; Mo & Morgado, 2022a; Hu et al., 2022; Liu et al., 2022), which mainly learns audio-visual correspondence from a single image frame and an audio signal, video-centric AVL treats the input as a sequence of consecutive frames and analyzes the relationship between sound events and visual information over time (Choi et al., 2025). This temporal perspective is important because it is often difficult to identify the actual sound source from a single frame in complex scenes. Therefore, how to exploit temporal information is a key challenge in video-centric AVL.

Prior video-centric AVL (Choi et al., 2025) models temporal and spatial information across consecutive frames for richer video representations. However, its localization relies on audio-visual feature similarity, which captures category-level correspondence rather than whether an object is producing sound. Since sound emission depends on the state or action of an object, similarity alone can lead to incorrect localization. For example, as shown in Figure 1(a), not only the guitar being played but also a silent guitar in the background can be incorrectly predicted as a sound source. Therefore, video-centric AVL needs to go beyond audio-visual feature similarity and jointly understand the audio-visual context by integrating complementary audio, visual, and temporal information.

Recent Multimodal Large Language Models (MLLMs) offer a promising way to achieve such contextual understanding by jointly processing audio and visual information and reasoning about their relationship. In particular, Chain-of-Thought (CoT) reasoning (Wei et al., 2022; Kojima et al., 2022) allows MLLMs to reach a prediction through intermediate steps, naturally integrating different information sources for sound source localization. Existing CoT methods organize reasoning by decomposing a problem into easier sub-problems (Zhou et al., 2023), abstracting underlying principles (Zheng et al., 2024), or alternating reasoning and reflection over multiple turns (Ni et al., 2025).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

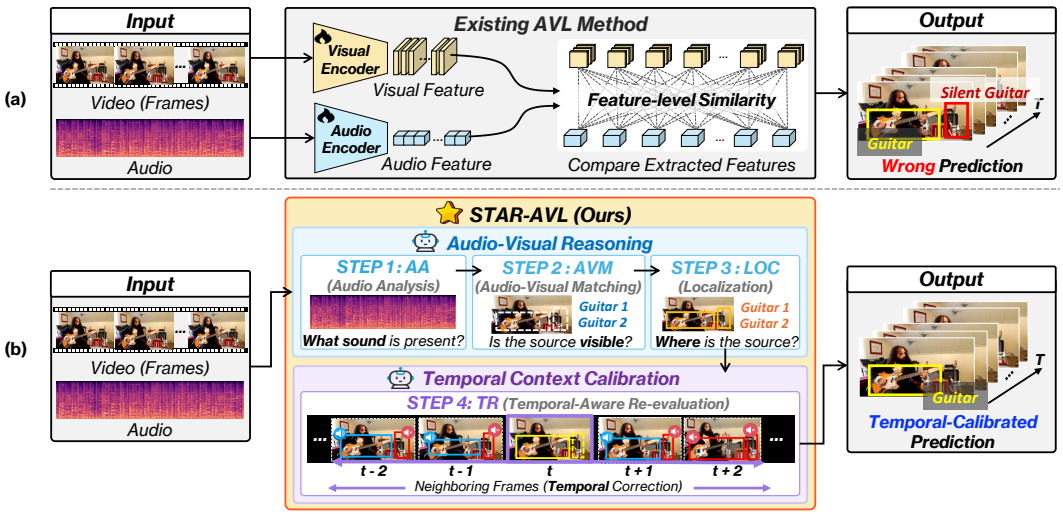


Figure 1: Overview of (a) existing feature matching-based AVL methods and (b) our STAR-AVL. Existing methods localize sound sources based on audio-visual feature similarity, which may incorrectly highlight both the sounding guitar and a silent guitar in the background. In contrast, STAR-AVL structures video-centric AVL into four steps for audio analysis, audio-visual matching, localization, and temporal-aware re-evaluation, enabling more reliable sound source localization.

However, these methods do not explicitly consider which modality is required at each reasoning step, nor do they exploit temporal information when re-evaluating their predictions.

Accordingly, applying MLLM-based reasoning to video-centric AVL raises two key challenges: (*i*) how to **define each reasoning step** according to its role and required modality, and (*ii*) how to leverage **temporal information** to re-evaluate localization in the current frame.

For challenge (*i*), different reasoning steps require *different types of information*. For example, identifying what sound is present can be performed using audio alone, whereas determining which object is actually producing the sound requires both audio and visual information. Thus, each step should be explicitly defined by its role and the modality required to perform that role. This modality-aware organization is essential for ensuring that each decision is made using the appropriate information.

For challenge (*ii*), reliable localization requires *temporal context* beyond the current frame. Since the state and position of an object can change over time, a prediction made from the current frame alone may be ambiguous or inconsistent with the surrounding video. Therefore, the prediction should be explicitly re-evaluated against neighboring frames and revised when temporal evidence indicates an inconsistency. Together, these two challenges highlight the need for a structured framework that incorporates both modality-aware reasoning and temporal-aware re-evaluation.

To address these challenges, we propose **STAR-AVL** (**S**tructured **T**emporal-**A**ware **R**easoning for **A**udio-**V**isual **L**ocalization), a training-free structured reasoning framework for video-centric AVL (Figure 1(b)). It consists of two components, **Audio-Visual Reasoning** and **Temporal Context Calibration**. For challenge (*i*), we devise Audio-Visual Reasoning to explicitly define the role and required modality of each step: **Step 1**, Audio Analysis (AA), identifies the sound from audio; **Step 2**, Audio-Visual Matching (AVM), verifies whether a visible object produces the sound using both audio and visual information; and **Step 3**, Localization (LOC), localizes the verified sound-emitting object. For challenge (*ii*), we introduce Temporal Context Calibration (**Step 4**, Temporal-aware Re-evaluation (TR)) to re-evaluate the current localization using neighboring frames and revise it when necessary. Even when errors from earlier steps propagate to the initial prediction, TR can correct them based on the surrounding temporal context. Together, these complementary components enable STAR-AVL to localize sound sources through audio-visual understanding and temporal consistency.

The main contributions of this paper are as follows:

- We formulate video-centric AVL as a structured reasoning problem, beyond both existing audio-visual feature matching and multimodal CoT reasoning, and propose STAR-AVL, a training-free framework for this task.

- We build a structured reasoning framework with two complementary goals. Audio-Visual Reasoning identifies actual sound-emitting objects, while Temporal Context Calibration addresses the remaining localization ambiguity by leveraging surrounding temporal context.
- Comprehensive experimental results show that STAR-AVL outperforms existing AVL methods, MLLM baselines, and CoT reasoning methods on AVATAR, and further generalizes well to image-centric AVL settings on VGGSound-Duet.

2 Related Work

2.1 Audio-Visual Localization

Audio-Visual Localization (AVL) has progressed mainly in single-source and multi-source settings. Single-source methods have evolved from attention-based dual-stream architectures (Senocak et al., 2018; Owens & Efros, 2018; Arandjelovic & Zisserman, 2018) to contrastive learning (Mo & Morgado, 2022a;b; Sun et al., 2023; Park et al., 2026) and other approaches using pseudo-label refinement, optical flow, transformation invariance, and semantic alignment (Song et al., 2022; Lin et al., 2023; Fedorishin et al., 2023; Um et al., 2023; Senocak et al., 2022a; 2025). Multi-source methods address multiple sounding objects through audio-visual matching, mixture-based localization, and object-level scene understanding (Hu et al., 2020; 2022; Mo & Tian, 2023; Kim et al., 2024; Um et al., 2025). Despite these advances, most methods remain image-centric, matching a single frame to audio (Senocak et al., 2022b; Park et al., 2024; Sato et al., 2025) without explicit reasoning or verification. Recently, GAR (Park & Kim, 2026) proposes training-free MLLM reasoning to AVL, but still operates on a single image-audio pair without temporal information.

Recognizing this limitation, AVATAR (Choi et al., 2025) recently introduced a video-centric benchmark with TAVLO, which models temporal and spatial information separately. However, TAVLO still relies on audio-visual feature matching and cannot verify whether a localized object actually is the sound-emitting object. This motivates reasoning beyond feature correspondence.

2.2 Reasoning with Multimodal Large Language Models

Multimodal Large Language Models (MLLMs) have demonstrated strong capabilities in cross-modal understanding (Xu et al., 2025a;b; Liu et al., 2025b) and structured reasoning (Wen et al., 2026; Hui et al., 2026; Lin et al., 2025; Li et al., 2025b). In particular, Chain-of-Thought (CoT) reasoning allows models to interpret complex information through intermediate reasoning steps rather than directly predicting the final output (Wei et al., 2022; Wang et al., 2023b). Since its introduction in text-only settings, it has been extended to vision-language settings (Zhang et al., 2024; Wei et al., 2024; Jiang et al., 2025; Zheng et al., 2025; Ni et al., 2025) and, more recently, to audio (Xie et al., 2025). CoT reasoning that jointly uses audio and visual information, however, has only begun to emerge in generation (Liu et al., 2025a), question answering (Xing et al., 2025), and modality-specific reasoning for audio-visual LLMs (Li et al., 2026), and none of these addresses the problem of identifying the spatial location of the sound-emitting object.

Several strategies have been proposed to better organize reasoning, including decomposition (Ning et al., 2024), planning (Wang et al., 2023a), abstraction (Zheng et al., 2024), refinement (Madaan et al., 2023; Wu et al., 2026), and verification (Weng et al., 2023; Dhuliawala et al., 2024). However, these strategies are largely task-independent and, having been developed mainly for text and vision-language settings (Prasad et al., 2024; Gadetsky et al., 2025), do not explicitly distinguish audio and visual evidence. More importantly, they do not verify whether the localized object is actually emitting the sound or whether the prediction is temporally consistent with neighboring frames. In contrast, we formulate video-centric AVL as a structured reasoning problem and organize audio and visual information into a training-free, modality-aware process that verifies the sound-emitting object across modalities and over time.

3 Proposed Method

3.1 Problem Formulation

Given a video sample, we denote its sampled visual frames as $\mathcal{V} = \{\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_T\}$, where T is the number of frames and $\mathcal{V}_t$ represents the visual frame at the t-th time step. For each frame $\mathcal{V}_t$, we

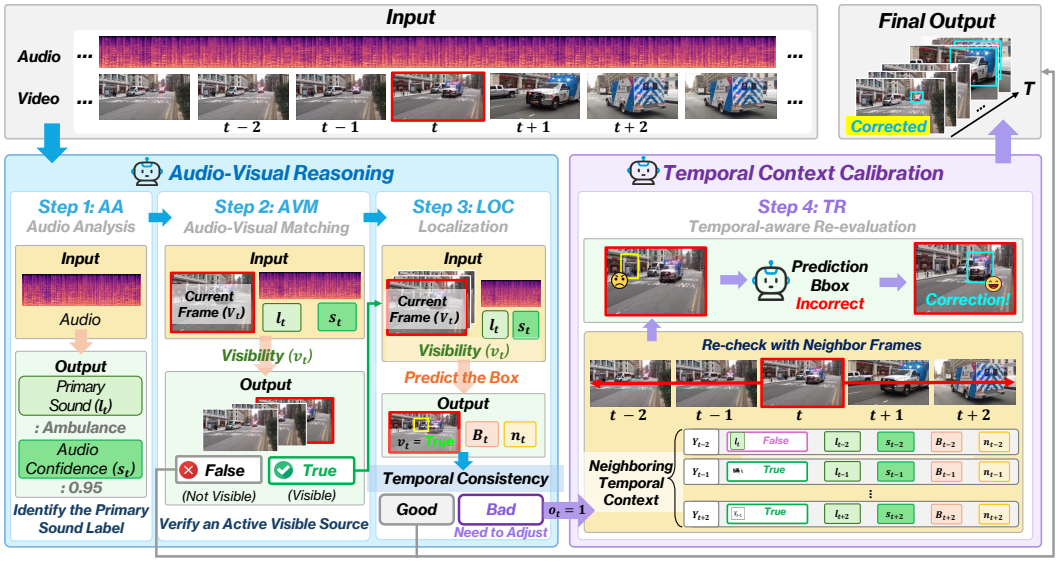


Figure 2: Overview of the proposed STAR-AVL framework. The four reasoning steps consist of two components: Audio-Visual Reasoning (AA, AVM, LOC) for audio understanding, cross-modal verification, and spatial localization. Temporal Context Calibration, which comprises TR that verifies and re-evaluates the prediction against neighboring frames when necessary.

extract a temporally aligned audio segment from the original audio track. The corresponding audio segments are denoted as $\mathcal{A} = \{\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_T\}$, where $\mathcal{A}_t$ represents the audio segment aligned with $\mathcal{V}_t$. Thus, the frame-level audio-visual input at time step t is defined as $(\mathcal{V}_t, \mathcal{A}_t)$.

To incorporate temporal context, we also use the prediction history from previous frames, denoted as $\mathcal{P}_{1:t-1} = \{Y_1^{AAL}, Y_2^{AAL}, \ldots, Y_{t-1}^{AAL}\}$, where each element is the complete frame-level output of the corresponding frame. For each frame, the model produces the following output:

$$Y_t = (l_t, s_t, v_t, \mathcal{B}_t, n_t, r_t), \tag{1}$$

where l_t denotes the primary sound label identified from $\mathcal{A}_t$, $s_t \in [0, 1]$ is the audio confidence score, and $v_t \in \{0, 1\}$ indicates whether a visible object that clearly corresponds to the primary sound exists in $\mathcal{V}_t$. $\mathcal{B}_t$ denotes the set of predicted bounding boxes for visible sound-emitting entities, n_t is the number of localized sound-emitting entities, and r_t is a brief reasoning explanation generated by the model. If $v_t = 0$, $\mathcal{B}_t = \emptyset$ and $n_t = 0$.

3.2 Overall Architecture

We present STAR-AVL, a training-free multi-step reasoning framework for video-centric AVL. It leverages the audio-visual reasoning capabilities of MLLMs to localize sounding objects in each video frame. As shown in Figure 2, the framework consists of two components, Audio-Visual Reasoning and Temporal Context Calibration, which together comprise four progressive reasoning steps. Audio-Visual Reasoning consists of ***Step 1*** (Audio Analysis (AA)), ***Step 2*** (Audio-Visual Matching (AVM)), and ***Step 3*** (Localization (LOC)), while Temporal Context Calibration consists of ***Step 4*** (Temporal-aware Re-evaluation (TR)). Given the frame-level input $(\mathcal{V}_t, \mathcal{A}_t)$, AA step first analyzes the aligned audio segment to identify the primary sound label. AVM step then determines whether a visible object corresponding to the primary sound exists in the current frame. LOC step predicts bounding boxes for visible sound-emitting entities. Finally, TR step selectively re-evaluates the prediction using temporal context. Details are in the following subsections.

3.3 Audio-Visual Reasoning

Step 1 (Audio Analysis (AA)). AA step analyzes the audio segment $\mathcal{A}_t$ temporally aligned with the current video frame $\mathcal{V}_t$ and identifies the primary sound event of each distinct sound source. This step provides the semantic cue required for matching audio with visible objects in the frame. For example, the model first determines whether the audio contains a guitar sound, speech, or other

dominant sound events. This process is formulated as:

$$f_{\mathrm{AA}}(\mathcal{A}_t) = (l_t, s_t), \tag{2}$$

where $f_{\mathrm{AA}}(\cdot)$ denotes the audio analysis function, l_t is the primary sound label inferred from $\mathcal{A}_t$, and $s_t \in [0, 1]$ is the audio analysis confidence score. The label l_t is represented as an open-vocabulary label, such as "guitar", "speech", "engine", or "dog barking". Therefore, the model is not restricted to a fixed category set, but can generate a sound label based on the audio content. We design the AA step not to directly predict bounding boxes, but to provide sound-level semantic cues for the AVM step. This follows prior audio-visual works that emphasize the importance of identifying audible events for establishing audio-visual correspondence (Gemmeke et al., 2017; Tian et al., 2018).

Step 2 (Audio-Visual Matching (AVM)). Next, AVM step determines whether the current video frame $\mathcal{V}_t$ contains a visible object that corresponds to the primary sound l_t identified in AA step. The key idea of this step is to verify whether the object is an active sound source in the current scene, rather than merely checking whether a visually matching category appears in the frame. Since sound emission depends on the state or action of an object rather than its category, category-level matching alone cannot distinguish an active source from a silent one. This process is defined as:

$$v_t = f_{\mathrm{AVM}}(\mathcal{V}_t, \mathcal{A}_t, l_t, \mathcal{P}_{1:t-1}), \tag{3}$$

where $f_{\mathrm{AVM}}(\cdot)$ denotes the audio-visual matching function for visibility estimation, and $\mathcal{P}_{1:t-1}$ denotes the prediction history accumulated from previous frames. The prediction history is used as temporal context to support the current frame decision. For example, if previous frames repeatedly indicate that no visible object matches the primary sound, the model avoids forcing an ambiguous object in the frame to be selected as the sound source. The visibility decision is summarized as:

$$v_t = \begin{cases} 1, & \textit{visible object clearly matches } l_t \textit{ and is likely active}, \\ 0, & \textit{otherwise} \end{cases}. \tag{4}$$

Thus, $v_t = 1$ only when the current frame provides clear evidence of an active visible source corresponding to l_t; otherwise, the model sets $v_t = 0$.

Step 3 (Localization (LOC)). LOC step predicts bounding boxes based on the visibility decision v_t obtained from AVM step (step 2). Given $v_t = 1$, the model is guided to estimate the number of visible entities that actively produce the primary sound in $\mathcal{V}_t$ and predict one bounding box for each entity. Given $v_t = 0$, the model is guided to return an empty bounding box set, since no active visible source has been verified. The bounding box set is defined as:

$$\mathcal{B}_t = \begin{cases} \{b_t^i\}_{i=1}^{n_t}, & v_t = 1, \\ \emptyset, & v_t = 0 \end{cases}, \tag{5}$$

where n_t is the number of visible sound-emitting entities in the current frame. Each bounding box is represented as:

$$b_t^i = [x_1^i, y_1^i, x_2^i, y_2^i], \tag{6}$$

where (x_1^i, y_1^i) and (x_2^i, y_2^i) denote the top-left and bottom-right coordinates of the bounding box, respectively. All coordinates are in pixels. When multiple sound-emitting entities are visible, the model predicts one bounding box for each entity instead of merging them into a single box, so that each entity is localized separately and silent regions between them are not included.

The predicted box should include the entire sound-emitting entity, including parts involved in sound generation, since the sound is emitted by the entity as a whole rather than by the object alone. For example, for guitar sound, the box may include the guitarist and guitar. This allows the model to localize not only isolated objects but also sound-emitting human-object interactions.

Given the audio label l_t, audio confidence score s_t, and the visibility decision v_t, the LOC step predicts the bounding box set $\mathcal{B}_t$ and the number of localized sound-emitting entities n_t:

$$(\mathcal{B}_t, n_t) = f_{\mathrm{LOC}}(\mathcal{V}_t, \mathcal{A}_t, l_t, s_t, v_t, \mathcal{P}_{1:t-1}), \quad \text{if } v_t = 1, \tag{7}$$

where $f_{\mathrm{LOC}}(\cdot)$ denotes the function that predicts bounding boxes for visible sound-emitting entities.

We denote this initial prediction as Y_t^{AAL}, where the superscript AAL is formed from the initials of the three reasoning steps that produce it (**A**A, **A**VM, and **L**OC):

$$Y_t^{\mathrm{AAL}} = (l_t, s_t, v_t, \mathcal{B}_t, n_t, r_t), \tag{8}$$

where r_t denotes a brief reasoning explanation generated by the model during the reasoning process.

After obtaining Y_t^{AAL}, LOC performs a lightweight temporal consistency check over $\mathcal{P}_{1:t-1}$ to decide whether the current frame should be re-evaluated by TR:

$$o_t = g_{\mathrm{LOC}}(Y_t^{\mathrm{AAL}}, \mathcal{P}_{1:t-1}), \tag{9}$$

where $g_{\mathrm{LOC}}(\cdot)$ denotes the temporal consistency checking function, and $o_t \in \{0, 1\}$ indicates whether the current frame is selected as a TR candidate. Overall, the three-step reasoning process assigns each step a distinct role and the information required to fulfill it, allowing an object to be localized only after it is verified as the actual sound source.

3.4 Temporal Context Calibration

Step 4 (Temporal-aware Re-evaluation (TR)). Although Audio-Visual Reasoning verifies the sound-emitting object, its prediction is made from a single frame and its past history, so it can be ambiguous. Therefore, we devise a TR step to identify predictions with low temporal consistency and re-evaluate them using preceding and subsequent frames. TR step is selectively applied to the t-th frame only when the candidate indicator o_t is set to 1 after the temporal consistency check. A frame is marked as a TR candidate when its prediction shows an abrupt change compared with the prediction history, such as a large shift of the bounding box, a change in the number of localized entities, or an inconsistent visibility decision. These candidates are identified by the MLLM based on the current audio-visual evidence and the prediction history, rather than by a fixed threshold.

Specifically, our model re-evaluates the current t-th frame $\mathcal{V}_t$ and audio $\mathcal{A}_t$ by using temporal contexts from neighboring frames. This process is defined as:

$$Y_t^{\mathrm{final}} = \begin{cases} f_{\mathrm{TR}}(\mathcal{V}_t, \mathcal{A}_t, Y_t^{\mathrm{AAL}}, \mathcal{C}_t^-, \mathcal{C}_t^+), & o_t = 1, \\ Y_t^{\mathrm{AAL}}, & o_t = 0 \end{cases}, \tag{10}$$

$$\mathcal{C}_t^- = \{Y_{t-2}^{\mathrm{AAL}}, Y_{t-1}^{\mathrm{AAL}}\}, \quad \mathcal{C}_t^+ = \{Y_{t+1}^{\mathrm{AAL}}, Y_{t+2}^{\mathrm{AAL}}\}, \tag{11}$$

where $f_{\mathrm{TR}}(\cdot)$ denotes the temporal-aware re-evaluation function, $\mathcal{C}_t^-$ denotes the preceding temporal context of the current frame t, consisting of up to two adjacent frames and their initial predictions. For boundary frames (*i.e.*, first and last frames), only available neighboring frames are used.

If $o_t = 1$, TR step refines the initial prediction; otherwise, Y_t^{AAL} is used as the final output without re-evaluation, since a temporally consistent prediction does not need correction and skipping TR avoids unnecessary computation. At this step, TR step may either suppress an unreliable localization by setting $v_t = 0$ and $\mathcal{B}_t = \emptyset$, or correct the bounding boxes and entity count when a visible sound source is present. As a result, a localization that is difficult to determine from a single frame is corrected using the temporal context of neighboring frames. This also allows TR to correct an incorrect initial prediction from Audio-Visual Reasoning, as shown in Figure 7.

4 Experiments

4.1 Datasets and Evaluation Metrics

Datasets. We evaluate STAR-AVL on two datasets. AVATAR (Choi et al., 2025) is a video-centric AVL benchmark containing 5,000 test videos, 24,266 annotated frames, and 80 categories. Each video is 10 seconds long. Following (Choi et al., 2025), we use up to five annotated frames with their audio. AVATAR includes various scenarios including single-sound, mixed-sound, and multi-entity. We additionally evaluate the cross-event subset, where the dominant sound-emitting object changes over time, to see temporal adaptability. We also adopt the VGGSound-Duet (Mo & Tian, 2023), constructed from VGGSound (Chen et al., 2020), to evaluate generalization to image-centric multi-source localization. Following (Mo & Tian, 2023; Kim et al., 2024), we concatenate frames from two different videos and mix their audio, requiring the model to localize both sound sources.

Evaluation Metrics. For the AVATAR dataset (Choi et al., 2025), we follow the video-centric AVL evaluation protocol and report Consensus Intersection over Union (CIoU), which measures the spatial overlap with the ground-truth bounding box, and Area Under Curve (AUC), which evaluates

Table 1: Comparison with existing AVL methods on the AVATAR dataset. The MLLM baseline denotes single-step MLLM baselines that directly predict sound source locations without the multi-step reasoning pipeline. **Bold**/underlined fonts denote the best/second-best performance, respectively.

Method	Single-sound		Mixed-sound		Multi-entity	
	CIoU(%)↑	AUC(%)↑	CIoU(%)↑	AUC(%)↑	CIoU(%)↑	AUC(%)↑
Audio-Visual Models						
SLAVC(144k) (NeurIPS'22) (Mo & Morgado, 2022b)	9.07	10.60	6.31	7.88	6.41	7.96
EZ-VSL(10k) (ECCV'22) (Mo & Morgado, 2022a)	9.66	11.07	8.16	9.35	6.87	8.32
EZ-VSL(144k) (ECCV'22) (Mo & Morgado, 2022a)	10.92	12.22	6.97	8.34	5.80	7.42
EZ-VSL(full) (ECCV'22) (Mo & Morgado, 2022a)	12.17	13.38	7.67	8.91	6.96	8.40
SSL-TIE(144k) (ACM MM'22) (Liu et al., 2022)	13.10	**14.23**	5.19	6.76	5.50	7.12
TAVLO (ICCV'25) (Choi et al., 2025)	**13.42**	14.08	**14.13**	**14.52**	**12.08**	**12.69**
MLLMs						
Ola-7B (Liu et al., 2025b)	13.63	14.45	16.77	17.44	12.40	13.23
Interactive-Omni-8B (Tong et al., 2025)	24.95	25.66	19.71	20.32	16.70	17.75
Baichuan-Omni-1.5 (Li et al., 2025a)	28.49	28.78	28.61	28.80	22.29	22.73
MiniCPM-o 2.6 (Yao et al., 2024)	33.11	33.58	21.97	22.64	22.02	22.72
Qwen2.5-Omni-3B (Xu et al., 2025a)	43.13	43.41	30.53	30.80	29.22	29.88
Qwen2.5-Omni-7B (Xu et al., 2025a)	58.59	58.92	44.20	44.53	37.75	38.34
Qwen3-Omni-30B-A3B-Instruct (Xu et al., 2025b)	72.83	72.87	55.06	55.10	54.90	55.12
STAR-AVL (Qwen2.5-Omni-7B)	74.05	74.08	56.20	56.24	57.65	57.84
STAR-AVL (Qwen3-Omni-30B-A3B-Instruct)	**78.71**	**78.74**	**61.15**	**61.23**	**58.10**	**58.27**

Table 2: Comparison with various CoT reasoning methods and STAR-AVL on the AVATAR dataset.

Method	Single-sound		Mixed-sound		Multi-entity		Inference Time↓
	CIoU(%)↑	AUC(%)↑	CIoU(%)↑	AUC(%)↑	CIoU(%)↑	AUC(%)↑	
Least-to-Most (Zhou et al., 2023)	70.79	70.93	49.44	49.73	40.97	41.46	3.84
Step-Back (Zheng et al., 2024)	68.38	68.41	49.01	49.10	47.23	47.58	6.33
Visual-O1 (Ni et al., 2025)	72.87	72.97	50.49	50.65	48.39	48.80	14.70
STAR-AVL (Qwen2.5-Omni-7B)	**74.05**	**74.08**	**56.20**	**56.24**	**57.65**	**57.84**	**3.24**

localization across different thresholds. For the VGGSound-Duet setting (Mo & Tian, 2023), we follow the standard multi-source AVL evaluation protocol (Mo & Tian, 2023; Kim et al., 2024) and report Class-aware Average Precision (CAP), CIoU@0.3, and AUC.

4.2 Implementation Details

The proposed STAR-AVL uses Qwen2.5-Omni-7B (Xu et al., 2025a) and Qwen3-Omni-30B-A3B-Instruct (Xu et al., 2025b) as the backbone MLLMs. We additionally evaluate STAR-AVL with Ola-7B (Liu et al., 2025b), Interactive-Omni-8B (Tong et al., 2025), Baichuan-Omni-1.5 (Li et al., 2025a), MiniCPM-o 2.6 (Yao et al., 2024), and Qwen2.5-Omni-3B (Xu et al., 2025a) to examine backbone generalization. All experiments use a single NVIDIA RTX 4090 GPU, except Qwen3-Omni-30B-A3B-Instruct (Xu et al., 2025b), which uses four RTX 4090 GPUs.

4.3 Comparison to Prior Works

Comparison with Existing AVL Methods. We compare STAR-AVL with image-centric AVL methods (Mo & Morgado, 2022b;a; Liu et al., 2022), a video-centric AVL method (Choi et al., 2025), and single-step MLLM baselines on the AVATAR dataset. As shown in Table 1, STAR-AVL achieves 74.05%, 56.20%, and 57.65% CIoU with Qwen2.5-Omni-7B and 78.71%, 61.15%, and 58.10% CIoU with Qwen3-Omni-30B-A3B-Instruct on Single-sound, Mixed-sound, and Multi-entity, respectively. It outperforms all audio-visual models in every scenario and improves over the single-step MLLM baseline, showing the effect of the multi-step reasoning pipeline itself.

Comparison with CoT Reasoning Methods. We compare STAR-AVL with representative CoT methods using the same Qwen2.5-Omni-7B backbone. As shown in Table 2, STAR-AVL outperforms them in every scenario, reaching 57.65% CIoU on Multi-entity compared with 48.39% for Visual-O1. This demonstrates the benefit of structuring the reasoning by the role and required information of each step, together with temporal re-evaluation. Note that, even without the verification process of TR (Table 3), STAR-AVL outperforms all CoT methods on Mixed-sound (53.23%) and Multi-entity (51.95%), showing that the improvement already comes from the structured multi-step reasoning itself. STAR-AVL takes 3.24 seconds per frame (1.09s for the single-step MLLM baseline), the lowest latency among multi-step reasoning methods while providing the strongest lo-

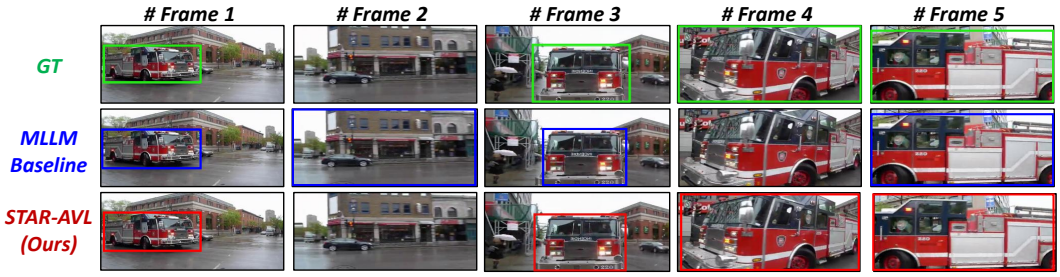


Figure 3: Visualization results of STAR-AVL and the MLLM baseline on the AVATAR. The MLLM baseline directly predicts localization results without the proposed multi-step reasoning.

Table 3: Effect of each reasoning step. The Localization (LOC) step is used in all settings to produce localization outputs, while Audio Analysis (AA) step, Audio-Visual Matching (AVM) step, and Temporal-aware Re-evaluation (TR) step are progressively added to evaluate their contributions.

Steps				Single-sound		Mixed-sound		Multi-entity	
LOC	AA	AVM	TR	CIoU(%)↑	AUC(%)↑	CIoU(%)↑	AUC(%)↑	CIoU(%)↑	AUC(%)↑
✓	–	–	–	58.59	58.92	44.20	44.53	37.75	38.34
✓	✓	–	–	64.77	64.86	50.28	50.37	51.71	51.90
✓	✓	✓	–	66.13	66.22	53.23	53.33	51.95	52.22
✓	✓	✓	✓	**74.05**	**74.08**	**56.20**	**56.24**	**57.65**	**57.84**

Table 4: Generalization of STAR-AVL across different MLLM backbones on the AVATAR dataset.

Method	Single-sound		Mixed-sound		Multi-entity	
	CIoU(%)↑	AUC(%)↑	CIoU(%)↑	AUC(%)↑	CIoU(%)↑	AUC(%)↑
Ola-7B (Liu et al., 2025b)	13.63	14.45	16.77	17.44	12.40	13.23
STAR-AVL (Ola-7B)	**26.26**	**26.76**	**22.59**	**23.07**	**21.81**	**22.33**
Interactive-Omni-8B (Tong et al., 2025)	24.95	25.66	19.71	20.32	16.70	17.75
STAR-AVL (Interactive-Omni-8B)	**34.12**	**34.60**	**24.65**	**25.24**	**19.43**	**20.23**
Baichuan-Omni-1.5 (Li et al., 2025a)	28.49	28.78	28.61	28.80	22.29	22.73
STAR-AVL (Baichuan-Omni-1.5)	**40.89**	**41.05**	**34.41**	**34.54**	**29.69**	**29.85**
MiniCPM-o 2.6 (Yao et al., 2024)	33.11	33.58	21.97	22.64	22.02	22.72
STAR-AVL (MiniCPM-o 2.6)	**40.20**	**40.31**	**34.56**	**34.65**	**29.35**	**29.55**
Qwen2.5-Omni-3B (Xu et al., 2025a)	43.13	43.41	30.53	30.80	29.22	29.88
STAR-AVL (Qwen2.5-Omni-3B)	**51.53**	**51.78**	**46.69**	**46.82**	**41.66**	**42.03**
Qwen2.5-Omni-7B (Xu et al., 2025a)	58.59	58.92	44.20	44.53	37.75	38.34
STAR-AVL (Qwen2.5-Omni-7B)	**74.05**	**74.08**	**56.20**	**56.24**	**57.65**	**57.84**
Qwen3-Omni-30B-A3B-Instruct (Xu et al., 2025b)	72.83	72.87	55.06	55.10	54.90	55.12
STAR-AVL (Qwen3-Omni-30B-A3B-Instruct)	**78.71**	**78.74**	**61.15**	**61.23**	**58.10**	**58.27**

calization performance. In contrast, CoT methods are slower because they generate intermediate reasoning text at every step and carry it over to the next, so the context grows as reasoning proceeds.

4.4 Visualization Results

Figure 3 shows qualitative visualization results using Qwen2.5-Omni-7B as the backbone MLLM. The MLLM baseline incompletely predicts the "fire truck" region or fails to predict in some frames, whereas STAR-AVL more consistently localizes the fire truck associated with the "siren" sound. Additional visualization results are in the supplementary material (see Section A.1).

4.5 Ablation Study

To see the contribution of each reasoning step, we conduct an ablation study using Qwen2.5-Omni-7B, as shown in Table 3. LOC is included in all configurations since it produces the localization output. The LOC-only setting localizes the sound-emitting object directly from the current frame and audio, and is equivalent to the single-step MLLM baseline in Table 1. Performance consistently improves as AA, AVM, and TR are added, showing that each step contributes complementarily to localization. Visualization examples of TR are in the supplementary material (see A.1.4).

4.6 Discussion

Generalization across MLLM Backbones. To evaluate the generalization of STAR-AVL across different MLLM backbones, we apply it to multiple MLLMs and compare each against its corresponding single-step baseline in Table 4. STAR-AVL consistently improves performance across

Table 5: Cross-event results on AVATAR dataset. CIoU and AUC are reported on the full test dataset and the cross-event subset to evaluate temporal adaptability.

Method	Total (CIoU (%)↑)	Cross-event (CIoU (%)↑)	Total (AUC (%)↑)	Cross-event (AUC (%)↑)
Audio-Visual Models				
SLAVC(144k) (NeurIPS'22) (Mo & Morgado, 2022b)	8.12	4.78	9.67	6.66
EZ-VSL(10k) (ECCV'22) (Mo & Morgado, 2022a)	8.95	5.08	10.33	6.73
EZ-VSL(144k) (ECCV'22) (Mo & Morgado, 2022a)	9.38	5.19	10.74	6.86
EZ-VSL(full) (ECCV'22) (Mo & Morgado, 2022a)	10.50	5.26	11.75	7.04
SSL-TIE(144k) (ACM MM'22) (Liu et al., 2022)	10.39	5.03	11.68	6.68
TAVLO(10k) (ICCV'25) (Choi et al., 2025)	**13.37**	**13.04**	**13.98**	**13.61**
MLLMs				
Ola-7B (Liu et al., 2025b)	24.17	18.72	24.73	19.56
Interactive-Omni-8B (Tong et al., 2025)	24.68	22.06	25.55	22.85
Baichuan-Omni-1.5 (Li et al., 2025a)	36.90	30.77	37.12	31.27
MiniCPM-o 2.6 (Yao et al., 2024)	30.97	25.91	31.20	24.96
Qwen2.5-Omni-3B (Xu et al., 2025a)	40.05	31.48	40.42	32.11
Qwen2.5-Omni-7B (Xu et al., 2025a)	51.25	44.40	51.68	44.92
Qwen3-Omni-30B-A3B-Instruct (Xu et al., 2025b)	64.94	53.24	65.07	53.65
STAR-AVL (Qwen2.5-Omni-7B)	66.32	50.18	66.44	50.62
STAR-AVL (Qwen3-Omni-30B-A3B-Instruct)	**70.30**	**56.86**	**70.43**	**57.29**

Table 6: Comparison of our STAR-AVL and image-centric AVL methods on the VGGSound-Duet.

Method	VGGSound-Duet		
	CAP (%) ↑	CIoU@0.3 (%) ↑	AUC (%) ↑
Attention 10k (CVPR'18) (Senocak et al., 2018)	–	11.50	15.20
OTS (ECCV'18) (Arandjelovic & Zisserman, 2018)	10.50	12.20	15.80
DMC (CVPR'19) (Hu et al., 2019)	–	13.80	17.10
CoarseToFine (ECCV'20) (Qian et al., 2020)	–	14.70	18.50
EZ-VSL (ECCV'22) (Mo & Morgado, 2022a)	–	20.50	20.20
Mix-and-Localize (CVPR'22) (Hu et al., 2022)	16.30	21.10	20.50
AVGN (CVPR'23) (Mo & Tian, 2023)	21.90	26.20	23.80
NoPrior (CVPR'24) (Kim et al., 2024)	32.50	46.90	29.20
OA-SSL (CVPR'25) (Um et al., 2025)	<u>45.90</u>	55.20	<u>44.80</u>
GAR (CVPR'26) (Park & Kim, 2026)	43.40	<u>58.90</u>	38.10
STAR-AVL (Qwen2.5-Omni-7B)	**53.08**	**70.67**	**47.44**

all backbones and scenarios. These results indicate that the gains are attributable to the proposed structured reasoning framework rather than a specific MLLM backbone.

Analysis of Cross-Event Videos. We evaluate STAR-AVL on the cross-event subset of AVATAR (Choi et al., 2025), where the active sound-emitting object changes over time within the same video. As shown in Table 5, STAR-AVL improves over the single-step MLLM baselines on this subset with both backbones, achieving 50.18% CIoU and 50.62% AUC with Qwen2.5-Omni-7B and 56.86% CIoU and 57.29% AUC with Qwen3-Omni-30B-A3B-Instruct. This shows that re-evaluating the localization with temporal context helps when the sound source changes over time.

Generalization to Image-Centric AVL Benchmarks. Although designed for video-centric AVL, STAR-AVL can be applied to image-centric AVL benchmarks as one-second short videos. As shown in Table 6, STAR-AVL achieves strong performance on VGGSound-Duet (Mo & Tian, 2023) despite the limited temporal information, showing that it generalizes beyond video-centric scenarios.

Limitations. As STAR-AVL builds on the intermediate predictions of the backbone MLLM, quality of early predictions can influence subsequent steps. The TR step mitigates this using the temporal context of neighboring frames (see Figure 7), although its effect may be limited when neighboring frames are similarly unreliable. Future work can explore strategies to reduce such dependency.

5 Conclusion

In this paper, we present STAR-AVL, a training-free structured reasoning framework for video-centric AVL. STAR-AVL consists of Audio-Visual Reasoning, which progressively analyzes audio cues, matches them with visual content, and localizes sound-emitting objects, and Temporal Context Calibration, which re-evaluates localization using neighboring frames for temporal consistency. By structuring reasoning according to the role of each step, STAR-AVL enables more reliable localization. Extensive experiments show consistent improvements over video-centric AVL and CoT methods across different MLLM backbones and challenging scenarios, including cross-event videos.

AI USE STATEMENT

In this paper, we used generative AI tools to assist with translation, as parts of the manuscript were originally written in the native language of the authors. We also used these tools to improve the readability of the manuscript, including grammar correction and sentence-level rephrasing. These tools were not used for any part of the research itself, such as developing methods, designing experiments, or interpreting results.

We reviewed all AI-assisted work, checking the resulting text against the intended meaning of the original drafts and the experimental results reported in this paper. We confirmed that no claim, numerical result, or citation was introduced or altered in this process. We take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

We have made several efforts to ensure that our results can be reproduced. The four reasoning steps of STAR-AVL are described in Section 3, and the complete prompt used at each step is provided in Appendix A.3. The datasets, evaluation metrics, backbone models, and hardware used in our experiments are described in Section 4. Our method is training-free and relies only on publicly available datasets and publicly released multimodal large language models, so no training data or model checkpoints are required to reproduce our results. The source code will be made publicly available.

REFERENCES

Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *Proceedings of the European Conference on Computer Vision*, pp. 435–451, 2018.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 721–725. IEEE, 2020.

Hahyeon Choi, Junhoo Lee, and Nojun Kwak. What's making that sound right now? video-centric audio-visual localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20095–20104. IEEE, 2025.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-verification reduces hallucination in large language models. In *Findings of the association for computational linguistics: ACL 2024*, pp. 3563–3578, 2024.

Dennis Fedorishin, Deen Dayal Mohan, Bhavin Jawade, Srirangaraj Setlur, and Venu Govindaraju. Hear the flow: Optical flow-based self-supervised visual sound source localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2277–2286. IEEE, 2023.

Artyom Gadetsky, Andrei Atanov, Yulun Jiang, Zhitong Gao, Ghazal Hosseini Mighan, Amir Zamir, and Maria Brbic. Large (vision) language models are unsupervised in-context learners. In *International Conference on Learning Representations*, volume 2025, pp. 25046–25070, 2025.

Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 776–780. IEEE, 2017.

Di Hu, Feiping Nie, and Xuelong Li. Deep multimodal clustering for unsupervised audiovisual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9240–9249. IEEE, 2019.

Di Hu, Rui Qian, Minyue Jiang, Xiao Tan, Shilei Wen, Errui Ding, Weiyao Lin, and Dejing Dou. Discriminative sounding objects localization via self-supervised audiovisual matching. *Advances in Neural Information Processing Systems*, 33:10077–10087, 2020.

Xixi Hu, Ziyang Chen, and Andrew Owens. Mix and localize: Localizing sound sources in mixtures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10473–10482. IEEE, 2022.

Wenjun Hui, Zhenfeng Zhu, Shuai Zheng, Ming-Ming Cheng, Huchuan Lu, and Yao Zhao. Mrcnet: Motion reasoning chain for cross modal video camouflaged object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Jingjing Jiang, Chao Ma, Xurui Song, Hanwang Zhang, and Jun Luo. Corvid: Improving multimodal large language models towards chain-of-thought reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3034–3046. IEEE, 2025.

Dongjin Kim, Sung Jin Um, Sangmin Lee, and Jung Uk Kim. Learning to visually localize sound sources from mixtures without prior source knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26457–26466. IEEE, 2024.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

Xuanchen Li, Yuheng Lu, Chenrui Cui, Tianrui Wang, Zikang Huang, Yu Jiang, Long Zhou, Longbiao Wang, and Jianwu Dang. Separate first, fuse later: Mitigating cross-modal interference in audio-visual llms reasoning with modality-specific chain-of-thought. *arXiv preprint arXiv:2605.09906*, 2026.

Yadong Li, Jun Liu, Tao Zhang, Song Chen, Tianpeng Li, Zehuan Li, Lijun Liu, Lingfeng Ming, Guosheng Dong, Da Pan, et al. Baichuan-omni-1.5 technical report. *arXiv preprint arXiv:2501.15368*, 2025a.

Yunxin Li, Zhenyu Liu, Zitao Li, Xuanyu Zhang, Zhenran Xu, Xinyu Chen, Haoyuan Shi, Shenyuan Jiang, Xintong Wang, Jifang Wang, et al. Perception, reason, think, and plan: A survey on large multimodal reasoning models. *arXiv preprint arXiv:2505.04921*, 2025b.

Yan-Bo Lin, Hung-Yu Tseng, Hsin-Ying Lee, Yen-Yu Lin, and Ming-Hsuan Yang. Unsupervised sound localization via iterative contrastive learning. *Computer Vision and Image Understanding*, 227:103602, 2023.

Zhiyu Lin, Yifei Gao, Xian Zhao, Yunfan Yang, and Jitao Sang. Mind with eyes: from language reasoning to multimodal reasoning. *arXiv preprint arXiv:2503.18071*, 2025.

Huadai Liu, Kaicheng Luo, Jialei Wang, Wen Wang, Qian Chen, Zhou Zhao, and Wei Xue. Thinksound: Chain-of-thought reasoning in multimodal LLMs for audio generation and editing. In *Advances in Neural Information Processing Systems*, 2025a.

Jinxiang Liu, Chen Ju, Weidi Xie, and Ya Zhang. Exploiting transformation invariance and equivariance for self-supervised sound localisation. In *Proceedings of the ACM International Conference on Multimedia*, pp. 3742–3753, 2022.

Zuyan Liu, Yuhao Dong, Jiahui Wang, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Ola: Pushing the frontiers of omni-modal language model. *arXiv preprint arXiv:2502.04328*, 2025b.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594, 2023.

Shentong Mo and Pedro Morgado. Localizing visual sounds the easy way. In *Proceedings of the European Conference on Computer Vision*, pp. 218–234. Springer, 2022a.

Shentong Mo and Pedro Morgado. A closer look at weakly-supervised audio-visual source localization. *Advances in Neural Information Processing Systems*, 35:37524–37536, 2022b.

Shentong Mo and Yapeng Tian. Audio-visual grouping network for sound localization from mixtures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10565–10574. IEEE, 2023.

Minheng Ni, Yutao Fan, Lei Zhang, and Wangmeng Zuo. Visual-o1: Understanding ambiguous instructions via multi-modal multi-turn chain-of-thoughts reasoning. In *International Conference on Learning Representations*, volume 2025, pp. 65065–65091, 2025.

Xuefei Ning, Zinan Lin, Zixuan Zhou, Zifu Wang, Huazhong Yang, and Yu Wang. Skeleton-of-thought: Prompting llms for efficient parallel generation. In *International Conference on Learning Representations*, volume 2024, pp. 917–967, 2024.

Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision*, pp. 639–658. Springer, 2018.

Sooyoung Park, Arda Senocak, and Joon Son Chung. Can clip help sound source localization? In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5699–5708. IEEE, 2024.

Sooyoung Park, Arda Senocak, and Joon Son Chung. Hearing and seeing through clip: A framework for self-supervised sound source localization. *International Journal of Computer Vision*, 134(4): 179, 2026.

Subin Park and Jung Uk Kim. Generate, analyze, and refine: Training-free sound source localization via mllm meta-reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15752–15761, June 2026.

Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. Rephrase, augment, reason: Visual grounding of questions for vision-language models. In *International Conference on Learning Representations*, 2024.

Rui Qian, Di Hu, Heinrich Dinkel, Mengyue Wu, Ning Xu, and Weiyao Lin. Multiple sound sources localization from coarse to fine. In *Proceedings of the European Conference on Computer Vision*, pp. 292–308. Springer, 2020.

Tomoya Sato, Yusuke Sugano, and Yoichi Sato. Audio-visual localization based on spatial relative sound order. *Machine Vision and Applications*, 36(4):90, 2025.

Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4358–4366. IEEE, 2018.

Arda Senocak, Hyeonggon Ryu, Junsik Kim, and In So Kweon. Learning sound localization better from semantically similar samples. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4863–4867. IEEE, 2022a.

Arda Senocak, Hyeonggon Ryu, Junsik Kim, and In So Kweon. Less can be more: Sound source localization with a classification model. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 577–586. IEEE, 2022b.

Arda Senocak, Hyeonggon Ryu, Junsik Kim, Tae-Hyun Oh, Hanspeter Pfister, and Joon Son Chung. Toward interactive sound source localization: Better align sight and sound! *IEEE transactions on pattern analysis and machine intelligence*, 47(9):7643–7659, 2025.

Zengjie Song, Yuxi Wang, Junsong Fan, Tieniu Tan, and Zhaoxiang Zhang. Self-supervised predictive learning: A negative-free method for sound source localization in visual scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3222–3231, 2022.

Weixuan Sun, Jiayi Zhang, Jianyuan Wang, Zheyuan Liu, Yiran Zhong, Tianpeng Feng, Yandong Guo, Yanhao Zhang, and Nick Barnes. Learning audio-visual source localization via false negative aware contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6420–6429. IEEE, 2023.

Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision*, pp. 252–268. Springer, 2018.

Wenwen Tong, Hewei Guo, Dongchuan Ran, Jiangnan Chen, Jiefan Lu, Kaibin Wang, Keqiang Li, Xiaoxu Zhu, Jiakui Li, Kehan Li, et al. Interactiveomni: A unified omni-modal model for audio-visual multi-turn dialogue. *arXiv preprint arXiv:2510.13747*, 2025.

Sung Jin Um, Dongjin Kim, and Jung Uk Kim. Audio-visual spatial integration and recursive attention for robust sound source localization. In *Proceedings of the ACM International Conference on Multimedia*, pp. 3507–3516, 2023.

Sung Jin Um, Dongjin Kim, Sangmin Lee, and Jung Uk Kim. Object-aware sound source localization via audio-visual scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8342–8351. IEEE, 2025.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 2609–2634, 2023a.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023b.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

Lai Wei, Wenkai Wang, Xiaoyu Shen, Yu Xie, Zhihao Fan, Xiaojin Zhang, Zhongyu Wei, and Wei Chen. Mc-cot: A modular collaborative cot framework for zero-shot medical-vqa with llm and mllm integration. *arXiv preprint arXiv:2410.04521*, 2024.

Zhihao Wen, Wenkang Wei, Yuan Fang, Xingtong Yu, Hui Zhang, Weicheng Zhu, and Xin Zhang. Star-kvqa: Structured reasoning traces for implicit-knowledge visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5114–5124, June 2026.

Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Shengping Liu, Bin Sun, Kang Liu, and Jun Zhao. Large language models are better reasoners with self-verification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 2550–2575, 2023.

Shiyu Wu, Mingzhen Sun, Weining Wang, Yequan Wang, and Jing Liu. Visualprompter: Semantic-aware prompt optimization with visual feedback for text-to-image synthesis. In *International Conference on Learning Representations*, volume 2026, pp. 30578–30600, 2026.

Zhifei Xie, Mingbao Lin, Zihang Liu, Pengcheng Wu, Shuicheng Yan, and Chunyan Miao. Audio-reasoner: Improving reasoning capability in large audio language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 23840–23862, 2025.

Zhenghao Xing, Xiaowei Hu, Chi-Wing Fu, Wenhai Wang, Jifeng Dai, and Pheng-Ann Heng. Echoink-r1: Exploring audio-visual reasoning in multimodal llms via reinforcement learning. *arXiv preprint arXiv:2505.04623*, 2025.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025a.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025b.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

Zhuosheng Zhang, Aston Zhang, Mu Li, hai zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *Transactions on Machine Learning Research*, 2024.

Huaixiu Steven Zheng, Swaroop Mishra, Xinyun Chen, Heng-Tze Cheng, Ed H Chi, Quoc V Le, and Denny Zhou. Take a step back: Evoking reasoning via abstraction in large language models. In *International Conference on Learning Representations*, volume 2024, pp. 20279–20316, 2024.

Jiyang Zheng, Jialiang Shen, Yu Yao, Min Wang, Yang Yang, Dadong Wang, and Tongliang Liu. Chain-of-focus prompting: Leveraging sequential visual cues to prompt large autoregressive vision models. In *International Conference on Learning Representations*, volume 2025, pp. 23607–23626, 2025.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H. Chi. Least-to-most prompting enables complex reasoning in large language models. In *International Conference on Learning Representations*, 2023.

A APPENDIX

Contents

A.1 Additional Visualization Results
A.1.1 STAR-AVL (Qwen2.5-Omni-7B) on AVATAR
A.1.2 STAR-AVL (Qwen3-Omni-30B-A3B-Instruct) on AVATAR
A.1.3 STAR-AVL (Qwen2.5-Omni-7B) on VGGSound-Duet
A.1.4 Effect of Temporal Context Calibration on AVATAR
A.2 Prompts for CoT Reasoning Methods
A.2.1 CoT Method 1: Least-to-Most
A.2.2 CoT Method 2: Step-Back
A.2.3 CoT Method 3: Visual-O1
A.3 Prompts for Proposed Method

A.1 ADDITIONAL VISUALIZATION RESULTS

In this section, we provide additional visualization results to further validate the generality of the proposed STAR-AVL. We first compare STAR-AVL with the single-step MLLM baseline on the AVATAR dataset (Choi et al., 2025) using two different base models, Qwen2.5-Omni-7B (Xu et al., 2025a) (Figure 4) and Qwen3-Omni-30B-A3B-Instruct (Xu et al., 2025b) (Figure 5), to examine whether the benefit of the multi-step reasoning process holds across different backbone capacities. We then compare STAR-AVL with GAR (Park & Kim, 2026) on the VGGSound-Duet dataset (Figure 6) to verify that the proposed method generalizes beyond the AVATAR dataset. In Figures 4 and 5, the single-step MLLM baseline directly predicts audio-visual localization results without the proposed multi-step reasoning process. Finally, we compare the initial and final predictions of STAR-AVL before and after the temporal-aware Re-evaluation (TR) step on the AVATAR dataset 7 to show how Temporal Context Calibration corrects single-frame predictions using the temporal context of neighboring frames.

A.1.1 STAR-AVL (QWEN2.5-OMNI-7B) ON AVATAR

Figure 4 compares the qualitative visualization results of the proposed STAR-AVL with an MLLM baseline on the AVATAR dataset (Choi et al., 2025). For a fair comparison, both STAR-AVL and the baseline adopt Qwen2.5-Omni-7B (Xu et al., 2025a) as their base model, so that the observed differences stem solely from our proposed components rather than from model capacity. Each example consists of five temporally sampled frames. Green boxes indicate ground-truth annotations (GT), blue boxes indicate predictions from the MLLM baseline, and red boxes indicate predictions from the proposed method.

Figure 4(a) shows a case where two performers appear simultaneously and play visually similar instruments. The MLLM baseline tends to include only one instrument with an overly large bounding box in some frames or fails to accurately distinguish the actual sound-emitting objects. In contrast, STAR-AVL consistently detects the two sound-emitting instruments in each frame and maintains bounding boxes close to the GT locations even as the object positions change over time. This demonstrates that the proposed step-wise reasoning process does not rely solely on single-frame visual cues, but instead leverages audio-visual correspondence and temporal context to produce more consistent localization results.

Figure 4(b) presents a multi-entity scenario in which multiple performers appear simultaneously. In this case, the MLLM baseline mainly focuses on the most salient foreground object or only a subset of performers, failing to sufficiently capture multiple sound-emitting objects. In particular, when several instruments coexist in the same scene and appear with occlusion or different scales, MLLM baseline prediction can easily miss some objects. In contrast, STAR-AVL more comprehensively localizes multiple performers and instruments, producing stable predictions for multiple sound sources across frames.

Figure 4(c) shows an outdoor scene in which a person fires a gun, where gunfire is the primary sound. In addition to the sound-emitting person and gun, the scene contains visually salient objects such



Figure 4: Visualization results of STAR-AVL (Qwen2.5-Omni-7B) and the single-step MLLM baseline (Qwen2.5-Omni-7B) on the AVATAR dataset. The single-step MLLM baseline directly predicts audio-visual localization results without using the proposed multi-step reasoning process.

as a camera and a tripod. The MLLM baseline predicts the "camera or tripod", which is unrelated to the sound, as the sound source, because it places a box on a salient object in the frame without verifying which object is actually emitting the sound. In contrast, STAR-AVL identifies gunfire as the primary sound in the AA step, verifies in the AVM step whether a visible object that corresponds to this sound and is an active sound source is present, and then localizes the gun as a single sound-emitting entity in the LOC step. As a result, visually salient but silent objects are excluded, and boxes close to the actual sound-emitting object are consistently predicted across frames.

Overall, the qualitative results show that STAR-AVL provides more accurate and temporally consistent localization performance than direct MLLM-based prediction. In particular, under complex video-centric AVL scenarios where visually similar objects exist or multiple sound-emitting entities appear simultaneously, the proposed multi-step reasoning structure enables more reliable inference of audio-visual correspondence.

A.1.2 STAR-AVL (Qwen3-Omni-30B-A3B-Instruct) on AVATAR

Figure 5 compares the qualitative visualization results of the proposed STAR-AVL with an MLLM baseline on the AVATAR dataset (Choi et al., 2025). For a fair comparison, both STAR-AVL and the baseline adopt Qwen3-Omni-30B-A3B-Instruct (Xu et al., 2025b) as their base model, so that the observed differences stem solely from our proposed components rather than from model capacity. Each example consists of five temporally sampled frames. Green boxes indicate ground-truth annotations (GT), blue boxes indicate predictions from the MLLM baseline, and red boxes indicate predictions from the proposed method.

Figure 5(a) shows a scene in which three dancers perform simultaneously. The MLLM baseline predicts a small box on an irrelevant background region or detects only one dancer in some frames, missing the remaining entities. In contrast, STAR-AVL detects all three dancers close to the GT in every frame and consistently maintains multiple sound-emitting entities.

Figure 5(b) shows a guitar performance in which the sound-emitting object occupies most of the frame. The MLLM baseline predicts a very small box on only a small part of the guitar, and then suddenly expands the box to the entire frame in Frame 4, showing unstable predictions of the object



Figure 5: Visualization results of STAR-AVL (Qwen3-Omni-30B-A3B-Instruct) and the single-step MLLM baseline (Qwen3-Omni-30B-A3B-Instruct) on the AVATAR dataset. The single-step MLLM baseline directly predicts audio-visual localization results without using the proposed multi-step reasoning process.

scale. In contrast, STAR-AVL stably predicts a box covering the entire guitar across all frames, accurately capturing the spatial extent of the sound-emitting object.

Figure 5(c) shows two dancers performing on a stage. The MLLM baseline predicts the whole frame as a single box in the early frames, and even when it separates the two dancers in later frames, the boxes remain overly loose, so the predictions vary largely across frames. In contrast, STAR-AVL detects each dancer with a tight box from the first frame and tracks them stably as they move, demonstrating both accurate entity separation and temporal consistency.

Overall, STAR-AVL consistently localizes multiple sound-emitting entities, predicts boxes with appropriate scales, and maintains temporal consistency across frames, even with a larger backbone. This demonstrates that the proposed multi-step reasoning process improves audio-visual localization regardless of the underlying base model.

A.1.3 STAR-AVL (Qwen2.5-Omni-7B) on VGGSound-Duet

Figure 6 compares the visualization results of the proposed STAR-AVL with GAR (Park & Kim, 2026) on the VGGSound-Duet dataset (Mo & Tian, 2023). Following the multi-source AVL protocol (Mo & Tian, 2023; Kim et al., 2024), each sample is constructed by concatenating frames from two different videos and mixing their audio, so that the model is required to localize the sound-emitting object in each of the two sub-frames. STAR-AVL adopts Qwen2.5-Omni-7B (Xu et al., 2025a) as its base model and treats each sample as a one-second short video. Green boxes indicate ground-truth annotations (GT), blue boxes indicate predictions from GAR, and red boxes indicate predictions from the proposed method.

Figure 6(a) shows a lawn mowing scene on the left and a giggling person on the right. GAR (Park & Kim, 2026) localizes the lawn mower but predicts the baby in the small sub-window instead of the giggling person. In contrast, STAR-AVL correctly localizes both the lawn mower and the giggling person. Figure 6(b) shows a theremin performance on the left and a rattling snake on the



Figure 6: Visualization results of STAR-AVL (Qwen2.5-Omni-7B) and GAR (Park & Kim, 2026) on the VGGSound-Duet dataset. Green, blue, and red boxes indicate ground-truth annotations, GAR predictions, and STAR-AVL predictions, respectively.

right. GAR (Park & Kim, 2026) localizes an irrelevant background region instead of the theremin, whereas STAR-AVL accurately localizes both the theremin and the snake.

Figure 6(c) shows a flute performance on the left and a barking dog on the right. GAR (Park & Kim, 2026) predicts an overly large box on the left frame and localizes the dog's reflection in the mirror instead of the actual dog. In contrast, STAR-AVL tightly localizes the flute and the actual dog. Figure 6(d) shows a speaking woman on the left and marching people on the right. GAR (Park & Kim, 2026) localizes both persons in the left frame, including the non-speaking one, and misses the marching people. In contrast, STAR-AVL correctly localizes only the speaking woman and the marching people.

Overall, although STAR-AVL is structured for video-centric AVL, it also achieves accurate localization results on the image-centric AVL benchmark, which is consistent with the quantitative results in Table 6. This demonstrates that the proposed multi-step reasoning process generalizes well to image-centric AVL as well.

A.1.4 Effect of Temporal Context Calibration on AVATAR

Figure 7 shows how the Temporal-aware Re-evaluation (TR) step improves the initial prediction. The initial prediction is Y_t^{AAL} (Eq. 8) obtained through Audio-Visual Reasoning, *i.e.*, Step 1 (AA), Step 2 (AVM), and Step 3 (LOC), and is predicted before Step 4 (TR) is applied, i.e., without the neighboring-frame context used in TR. The final prediction is Y_t^{final} (Eq. 10) obtained through Temporal Context Calibration, Step 4 (TR), which re-evaluates the initial prediction by referring to the initial predictions of neighboring frames. Both examples use Qwen2.5-Omni-7B (Xu et al., 2025a) as the backbone.

Figure 7(a) shows an ambulance passing by with its siren on. Although the ambulance occupies most of the frame over time, the initial prediction places a small box on only part of the vehicle front, covering a much smaller region than the actual sound-emitting entity. TR refers to the neighboring frames, in which the ambulance gradually becomes larger, and expands the box to cover the entire vehicle. As a result, the final prediction is close to the GT in every frame.

Figure 7(b) shows two vehicles driving simultaneously. In Frame 1 and Frame 2, the initial prediction localizes only the left vehicle and misses the right one, so the number of entities is inconsistent with the neighboring frames. Based on the two entities consistently localized from Frame 3 onward, TR re-examines the missing entity and localizes the right vehicle, which is visually confirmed in

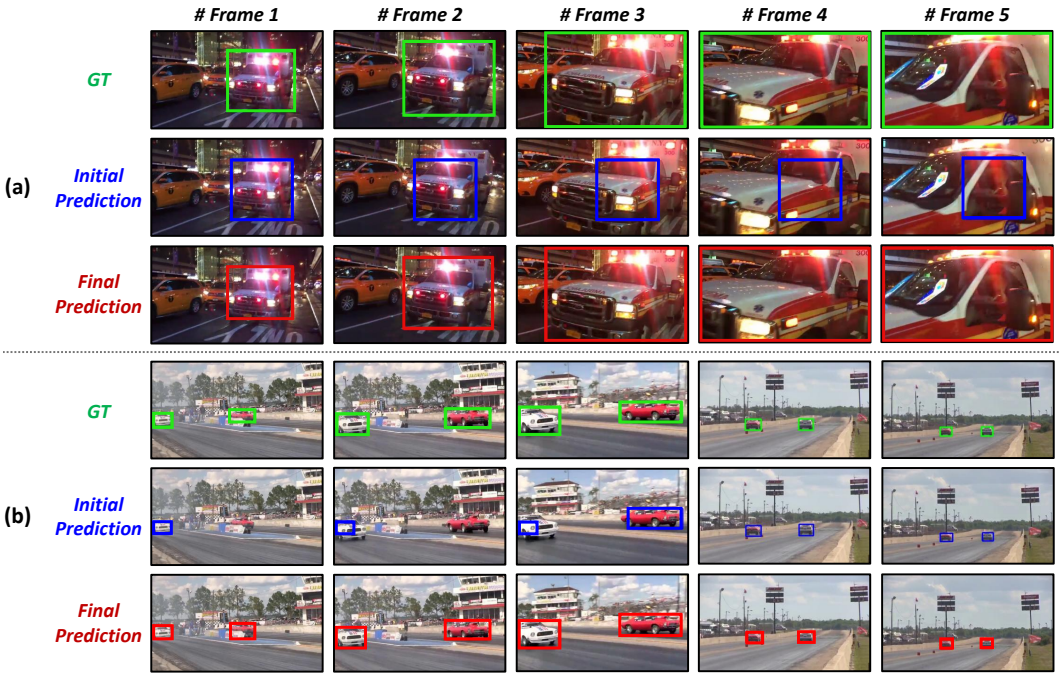


Figure 7: Visualization of initial and final predictions of STAR-AVL (Qwen2.5-Omni-7B) on the AVATAR dataset. Initial prediction denotes the output of Audio-Visual Reasoning (AA, AVM, and LOC steps) before TR, and final prediction denotes the output after Temporal Context Calibration (TR step) using neighboring frames.

the current frame. This is possible because TR does not simply copy the predictions of neighboring frames, but uses them as a reference to re-examine the audio-visual evidence in the current frame.

Overall, even when the initial prediction fails to capture the box size (Figure 7(a)) or the number of entities (Figure 7(b)) due to the ambiguity of a single frame, TR corrects it using the temporal context of neighboring frames. This is consistent with the performance improvement obtained by adding TR in Table 3.

A.2 Prompts for CoT Reasoning Methods

We describe how each CoT reasoning method in Table 2 is adapted to video-centric AVL, including the stages each method runs, what is given to the model at each stage, and where our adaptation departs from the original template. All methods use the same backbone, Qwen2.5-Omni-7B (Xu et al., 2025a), and are run without worked examples.

A.2.1 CoT Method 1: Least-to-Most

Least-to-Most (Zhou et al., 2023) does not solve a hard question in one shot but first reduces it to a sequence of easier sub-questions, then solves them in order so that answers obtained earlier serve as context for the questions that follow. The original paper proceeds in two stages, and we keep both stages and their order. The original method is text-only, so porting it to our setting requires one decision the original paper does not make: where the non-textual inputs go. We attach the frame and the corresponding audio to every call in both stages, so that each sub-question is answered with the same perceptual evidence available to the original question, and the textual scaffolding otherwise follows the original.

In the Decomposition step, the model lists the sub-questions that should be answered before the original question can be solved; no bounding box is requested. Following the convention established by the exemplars of the original paper, the last sub-question is the original question itself. In the Sequential solving step, the sub-questions are answered one at a time, each in a separate call, with

the frame, the audio, and every (question, answer) pair solved so far prepended to the prompt. Since the last sub-question is the original question, its answer is the final prediction, and no further call aggregates the intermediate answers; this too follows the original paper.

We depart from the original template in four ways. (*i*) We fill the problem position with our localization question. (*ii*) We drop the worked exemplars. In the original paper they are what conveys the decomposition pattern, implicitly and by example; keeping them would require us to hand-write decompositions for audio-visual localization, which would leak our own view of the task into the prompt. We instead state the axis of decomposition in a single sentence in the decomposition prompt: from the sound, to the visible objects that could produce it, and only then to their location. (*iii*) We fix the number of sub-questions to 3 and require one numbered sub-question per line, so that the decomposition can be parsed without a second model call. (*iv*) When the model returns fewer than three sub-questions, we fill each missing slot with a generic sub-question covering what is audible and which visible objects could produce it.

The value 3 in (*iii*) is the smallest value that leaves room for the two judgments our task should go through, what is audible and which visible objects could be producing it, once the final slot is taken by the original question. We keep it at the minimum rather than tuning it: each additional sub-question adds one call per sample to the baseline with the most calls per sample in our comparison, and with the final slot fixed to the original question, a larger budget would mostly add sub-questions along the same two axes.

A.2.2 CoT Method 2: Step-Back

Step-Back (Zheng et al., 2024) answers a concrete instance not directly but by first abstracting the concepts and principles the problem rests on, then reasoning from those principles. The original paper instantiates the abstraction step differently per domain: for knowledge-intensive QA it turns the abstraction into a step-back query used for retrieval, while for STEM reasoning it asks the model to state the underlying concepts and principles from its own knowledge. We follow the STEM instantiation, since AVL offers no corpus to retrieve the principles behind a given frame from.

In the Abstraction (AB) step, the model receives the frame and its audio and is asked only to extract the principles; no bounding box is requested. In the Reasoning (RE) step, the same frame and audio are given again together with the principles produced by AB, and the model solves the problem step by step by following them. The final bounding box is read only from the RE output.

We change three things in the original template. We fill the domain position (instantiated as physics and chemistry in the original) with audio-visual scene understanding, fill the problem position with our localization question, and supply the frame and its audio alongside the text in both steps since the principles should be grounded in the actual scene. Unlike the original, which conditions the abstraction step on few-shot demonstrations, we run both steps zero-shot so that the comparison with the other baselines stays controlled.

A.2.3 CoT Method 3: Visual-O1

Visual-O1 (Ni et al., 2025) handles instructions whose meaning is not clear from the image alone, and resolves them through multi-modal reasoning over several turns. Rather than answering in one pass, the model alternates between reasoning about the instruction and looking back on that reasoning, accumulating both into an experience, and finally derives the answer from the accumulated experience. The instruction in AVL is unclear in the same sense, since the image alone does not tell which of the visible objects is producing the sound.

In each Reasoning round, the model receives the frame, its audio, and the experience accumulated so far, and advances the reasoning by one step, with no coordinates requested; in the first round the experience is empty and the field is omitted. In each Reflection round, the reasoning just produced is examined against the image and the audio, again with no coordinates requested, and both results are appended to the experience. After three rounds, the Summarize step turns the whole process into the final prediction.

We make two changes to the original template. We fill the instruction position with our localization question, and we prepend a system line that states the audio-visual localization role together with a short per-round directive that asks each round to build on the accumulated experience rather than

repeat it. One difference remains: the original paper runs ten rounds, which we reduce to three for computational reasons. Even with three rounds, Visual-O1 requires 14.70 s/frame (Table 2), about 4.5 slower than STAR-AVL; the reduction was made for cost, not to limit its accuracy.

A.3 Prompts for Proposed Method

This section presents the prompt design used in STAR-AVL. The proposed method does not solve the video-centric audio-visual localization problem through a single direct prediction. Instead, it organizes the task into two components, **Audio-Visual Reasoning** and **Temporal Context Calibration**, which together comprise four reasoning steps. Audio-Visual Reasoning consists of ***Step 1*** (Audio Analysis (AA)), ***Step 2*** (Audio-Visual Matching (AVM)), and ***Step 3*** (Localization (LOC)), which respectively determine which sound should be localized in the current frame, whether the corresponding sound source is actually visible in the image, and where that source is located. Temporal Context Calibration consists of ***Step 4*** (Temporal-aware Re-evaluation (TR)), which verifies whether the predicted location is temporally consistent with neighboring frames and re-evaluates it when necessary. Each step uses the output of the preceding steps as an input condition, so that the prompt of every step states both the decision it is responsible for and the information available for making it.

Tables 7– 10 show the prompts for each reasoning step used in the proposed method. Specifically, Table 7 presents the AA step prompt for audio analysis, Table 8 presents the AVM step prompt for visual visibility judgment, Table 9 presents the LOC step prompt for bounding box localization, and Table 10 presents the TR step prompt for Temporal-aware Re-evaluation.

Prompt:

You are an expert in Audio-Visual Localization (AVL).

TASK: Audio Analysis (AA)
Given an AUDIO clip from the same scene as the image frame:
1) Listen carefully to the given audio clip.
2) Identify the primary sound of each distinct sound source.
3) Focus on the dominant sound that should be localized in the image.

Examples of primary sounds include guitar, speech, engine, footsteps, dog barking, singing, piano, drums, or vehicle sound.

STRICT OUTPUT:

```
{
  "primary_sound": "<short sound label>",
  "audio_confidence": <0.0 ~ 1.0>,
  "reasoning": "<brief explanation of what sound was detected>"
}
```

- The primary sound must be a short and concise label.
- The audio confidence must be a value between 0.0 and 1.0.
- Do not output any text or fields outside the JSON object.

Table 7: Audio Analysis (AA) Step

Prompt:

You are an expert in Audio-Visual Localization (AVL).

TASK: Audio-Visual Matching (AVM)
Given the current IMAGE frame, its corresponding AUDIO, the *primary_sound* identified in the AA step, and the prediction history of the previous frames:
{prediction_history} *(the sound label and visibility decision of each previous frame, inserted verbatim; omitted for the first frame)*
1) Decide whether the object producing the primary sound is clearly visible in the image.
2) Use the primary sound label as the reference for visual matching, and the audio to check whether that sound is still active in the current frame.
3) Do not assume that a matching object is visible unless there is clear visual evidence.
4) Use the prediction history only as context for the current decision, never as a replacement for it.

Set `sound_source_visible = false` when:
1. No visible object matches the primary sound category.
2. The sound source is likely outside the image.
3. The image contains only background or unrelated objects.
4. The visual evidence is uncertain or ambiguous.
5. The prediction history repeatedly reports no visible source for the same primary sound, and the current frame gives no clearer evidence than those frames.

Set `sound_source_visible = true` only when:
1. A visible object clearly matches the primary sound.
2. The object is a plausible active sound-emitting source.

STRICT OUTPUT:

```
{
  "sound_source_visible": <true | false>,
  "reasoning": "<brief explanation>"
}
```

- The value of `sound_source_visible` must be either true or false.
- Do not output any text or fields outside the JSON object.

Table 8: Audio-Visual Matching (AVM) Step

Prompt:

You are an expert in Audio-Visual Localization (AVL).

TASK: Localization (LOC)

This step is executed only when `sound_source_visible = true`.

Given the current IMAGE frame, its corresponding AUDIO, the *primary_sound* and *audio_confidence* from the AA step, the visibility decision from the AVM step, and the prediction history of the previous frames:
{prediction_history} *(the bounding boxes and entity count of each previous frame, inserted verbatim; omitted for the first frame)*
1) Predict the bounding box coordinates of all visible sound-emitting entities.
2) Use pixel coordinates in the original image.
3) Return an empty list if no sound source is visible.

Rules:
1. Predict one tight bounding box per sound-emitting entity.
2. Do not merge multiple entities into one bounding box.
3. Cover the entire sound-emitting entity, including the object used to produce the sound.
4. For example, include both the guitarist and guitar, the speaker and microphone, or the drummer and drum.
5. Do not include silent background objects.
6. If no sound source is visible, return `bboxes = []`.
7. Use the audio to judge how many entities are actually contributing to the primary sound, and be more conservative when `audio_confidence` is low.
8. Use the prediction history only as context for the number and the rough position of the entities. Do not copy a box from a previous frame without visual evidence in the current frame.
9. Compare the current prediction with the prediction history. If the bounding box shifts abruptly, the number of entities changes, or the visibility decision is inconsistent with previous frames; otherwise false.

STRICT OUTPUT:

```
{
  "bboxes": [[x1, y1, x2, y2], ...],
  "n_entities": <integer>,
  "need_temporal_re-evaluation": <true | false>,
  "reasoning": "<brief explanation>"
}
```

- Each bbox must be four integers in the original image coordinates.
- Each bbox must satisfy `x1 < x2` and `y1 < y2`.
- The value of `n_entities` must match the number of predicted boxes.
- Do not output any text or fields outside the JSON object.

Table 9: Localization (LOC) Step

Prompt:

You are an expert in Audio-Visual Localization (AVL).

TASK: Temporal-aware Re-evaluation (TR)
Given the current IMAGE frame, its corresponding AUDIO, the prediction obtained from the AA, AVM, and LOC steps for this frame, and the initial predictions of the neighboring frames:
{preceding_context} *(the initial predictions of up to two preceding frames, inserted verbatim)*
{subsequent_context} *(the initial predictions of up to two following frames, inserted verbatim)*
1) Re-evaluate whether the current prediction is temporally consistent.
2) Compare the current prediction with the sound source location, object category, and visibility decision from neighboring frames.
3) If the current bounding box is unusually small, unusually large, or spatially unstable compared with neighboring frame predictions, adjust it.
4) If the sound source is not clearly visible in the current frame, return an empty bounding box list.

Temporal-aware re-evaluation rules:
1. If the current prediction is consistent with neighboring frames, keep the original prediction.
2. If the current bbox points to the same object but its position or size is unstable, adjust it using neighboring frame information.
3. If the current bbox includes an incorrect object, correct it to the proper sound-emitting object.
4. Do not create a bbox based only on neighboring frames without clear visual evidence in the current frame.
5. If the sound source is not visible, return `bboxes = []`.
6. Use whichever of the preceding and following contexts is available. For the first and the last frames, reason from the available side only.

STRICT OUTPUT:

```
{
  "final_bboxes": [[x1, y1, x2, y2], ...],
  "reasoning": "<brief explanation>"
}
```

- Each bbox must be four integers in the original image coordinates.
- Each bbox must satisfy `x1 < x2` and `y1 < y2`.
- Do not output any text or fields outside the JSON object.

Table 10: Temporal-aware Re-evaluation (TR) Step