CONDITIONAL CAUSAL LEARNING FOR MULTIMODAL SOUND SOURCE LOCALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal sound source localization identifies image regions responsible for an input sound without spatial annotations. Most self-supervised methods treat synchronized audio-visual observations as positives and mismatched pairs as negatives. However, synchronized modalities encode both the sounding event and recurrent scene context, allowing pair discrimination to exploit contextual regularities without identifying the actual sound source. This issue is further complicated by modal asymmetry: audio captures event content without image-plane position, whereas vision preserves spatial structure together with latent appearance and acquisition biases. We propose Conditional Causal Learning (CCL), which formulates causal adjustment as a conditional design spanning representation learning and contrastive optimization. Rather than applying a uniform correction to heterogeneous biases, CCL selects conditioning information according to bias observability and its role in the learning process. Specifically, CCL suppresses context-dependent acoustic variation, preserves audio-consistent spatial information across alternative visual contexts, and reduces contextual bias in contrastive pair construction. Coordinating these adjustments limits shortcut dependence while retaining the semantic and spatial information required for localization. Extensive experiments across multiple benchmarks and perturbation settings demonstrate the effectiveness and robustness of CCL.

1 INTRODUCTION

Multimodal sound source localization (SSL) identifies the visual regions responsible for a given sound from an image or a video. Audio conveys the sounding event but lacks image-plane position, whereas vision provides spatial structure together with silent objects and background context. The high cost of spatial annotations has therefore motivated self-supervised learning from naturally synchronized audio and visual observations Senocak et al. (2018). Synchronization offers scalable pair-level supervision, yet SSL ultimately requires region-level prediction. This mismatch creates a fundamental ambiguity: audio-visual co-occurrence alone does not reveal which visual region actually produces the sound.

In Figure 1(a) and (b), correlation-driven methods Kim et al. (2025); Nguyen & Park (2026); Park et al. (2026) typically treat synchronized observations as positive pairs and mismatched observations as negatives. A synchronized pair may share both the sounding event and a recurring scene, whereas a random negative often shares neither. A concert scene, for example, may suggest the presence of an instrument without identifying which visible instrument generates the sound. The objective can therefore be satisfied through recurring contextual cues rather than precise source correspondence, allowing pair-level shortcuts to propagate into localization maps. Stronger encoders, refined negative sampling, and dense spatial features improve pair discrimination and spatial readout Chen et al. (2021); Senocak et al. (2022a); Hamilton et al. (2024), but synchronization alone still leaves the sound-producing region underdetermined. Figure 1(c) instead motivates a causal formulation that separates physical audio-visual correspondence from contextual dependence before alignment.

Causal debiasing provides a natural perspective for resolving this ambiguity. However, existing formulations for supervised visual recognition and multimodal reasoning commonly rely on labeled outcomes or task-specific counterfactuals Zhang et al. (2020); Yang et al. (2021); Niu et al. (2021). Self-supervised SSL lacks such an external anchor: the synchronized observation serves simultane-

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.

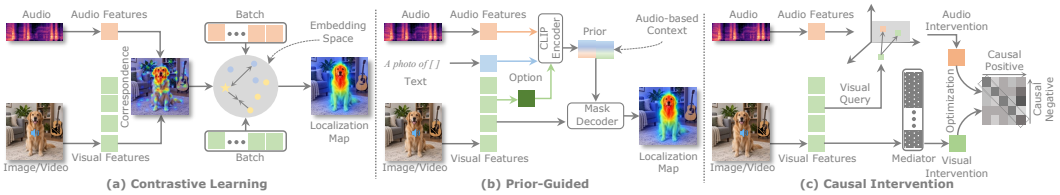

Figure 1: Comparison between correlation-driven SSL and CCL. Existing methods learn directly from observational audio-visual agreement, whereas CCL separates physical correspondence from contextual dependence before cross-modal alignment.

ously as the input and the training signal. Physical source correspondence and agreement induced by recurring context can therefore both explain the same observed pair, and improving similarity estimation alone cannot distinguish them. This issue enters SSL through two coupled stages. At the representation level, audio and vision contain confounding factors with different observability. At the optimization level, synchronized positives and random negatives make contextual similarity predictive of the pair label. Debiased contrastive learning alleviates class collision among negatives Chuang et al. (2020), but does not directly address the dependence between context and pair construction. A causal treatment therefore requires both modality-specific adjustment and context-aware optimization. The remaining challenge is how to align these corrections across modalities. Recurring cross-modal configurations provide an observable proxy for contextual factors affecting audio, enabling conditional adjustment while preserving information specific to the current sound. In contrast, visual acquisition, composition, and selection factors remain largely latent, making the same direct adjustment unsuitable for dense visual representations. We therefore adopt an ordered conditional design: audio is adjusted first using observable contextual information, and the corrected audio then guides visual mediation across alternative contexts. This ordering reduces contextual acoustic influence on visual localization while preserving source-relevant spatial structure. Context-aware pair weighting further prevents contrastive optimization from reintroducing the same dependence. In this way, representation correction and pair optimization form a coordinated causal pipeline rather than independent debiasing steps.

Technically, we introduce CCL to instantiate this formulation. Its Structural Causal Model (SCM) Pearl (2009) treats physical correspondence as the causal effect of interest while separating recurring multimodal context, latent visual confounding, and context-dependent pair selection. Guided by this structure, Confidence-Calibrated Audio Intervention (CAI) adjusts audio representations using observable context and retrieval confidence. Prediction-Level Visual Intervention (PVI) conditions visual mediation and alternative-context marginalization on the corrected audio. Causal Contrastive Optimization (CCO) then incorporates contextual pair weights into contrastive learning. These components operate on intermediate features and pairwise scores, allowing CCL to integrate naturally with existing SSL frameworks without additional localization annotations or module modifications. To the best of our knowledge, CCL is the first causal identification framework for self-supervised SSL. Our contributions are summarized as follows.

- We formulate SSL with a causal model that separates physical correspondence from co-occurrence context, latent visual confounding, and context-dependent pair selection.
- We derive observability-aware audio and visual interventions that combine direct adjustment, conditional mediation, and inverse-propensity weighting to emphasize context-atypical positives and context-matched negatives.
- We conduct extensive experiments across standard, open-set, extended, and segmentation benchmarks, and further verify the robustness under audio and visual noise perturbations.

2 Related Work

Multimodal Sound Source Localization. Early correspondence-learning methods show that annotation-free spatial grounding can emerge from audio-visual co-occurrence, while localization-specific contrastive objectives further improve dense spatial readout Senocak et al. (2018); Chen et al. (2021). Subsequent studies reduce supervision ambiguity through source decomposition,

grouping, and false-negative-aware learning Um et al. (2023); Sun et al. (2023); Mo & Tian (2023); Mahmud et al. (2024); Kim et al. (2024; 2025). CLIP-compatible representations and dense audio-visual features further enhance semantic transfer through stronger vision-language priors and finer cross-modal alignment Park et al. (2024); Hamilton et al. (2024); Um et al. (2025); Park et al. (2026); Nguyen & Park (2026). Nevertheless, plausible localization can persist under negative audio, mismatched modalities, or explicit modality conflict Juanola et al. (2025); Senocak et al. (2025); Jia et al. (2026), indicating that existing models may still exploit contextual regularities rather than source-specific correspondence.

3 METHODOLOGY

3.1 STRUCTURAL CAUSAL MODEL AND OBSERVATIONAL BIAS

Given an audio clip $\mathbf{A}$ and its corresponding visual frame $\mathbf{V}$, the goal is to predict a spatial localization map $\mathbf{Y}$ that highlights the sound-producing regions. Conventional methods directly model the observational distribution $P(\mathbf{Y} \mid \mathbf{A}, \mathbf{V})$. To describe the assumed data-generation process, we define the SCM in Figure 2. The direct paths $\mathbf{A} \rightarrow \mathbf{Y}$ and $\mathbf{V} \rightarrow \mathbf{Y}$ describe the physical correspondence intended to determine the sound source location. In contrast, $\mathbf{A} \leftarrow \mathbf{Z} \rightarrow \mathbf{V}$ and $\mathbf{A} \leftarrow \mathbf{Z} \rightarrow \mathbf{Y}$ capture recurring dataset-level configurations, such as instruments that frequently co-occur in concert halls. The path $\mathbf{V} \leftarrow \mathbf{U} \rightarrow \mathbf{Y}$ represents the effect of unobserved confounders, such as visual composition and selection factors. Marginalizing the context $\mathbf{Z}$ and the unobserved visual factors $\mathbf{U}$ gives:

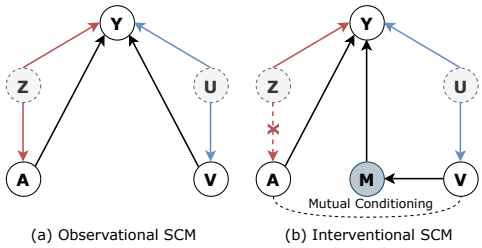

Figure 2: Observational and interventional SCMs.

$$P(\mathbf{Y} \mid \mathbf{A}, \mathbf{V}) = \mathbb{E}_{(\mathbf{Z},\mathbf{U})|\mathbf{A},\mathbf{V}}[P(\mathbf{Y} \mid \mathbf{A}, \mathbf{V}, \mathbf{Z}, \mathbf{U})]\,. \tag{1}$$

This marginalization imposes no conditional-independence assumption between $\mathbf{Z}$ and $\mathbf{U}$, and the joint posterior varies with the inputs. A correlation-driven model can consequently absorb both paths as shortcuts. The two confounders nevertheless differ in observability. Repeated cross-modal configurations provide the empirical proxy $\widehat{\mathbf{Z}} = \psi_{\mathbf{Z}}(\mathbf{A}, \mathbf{V})$, where $\psi_{\mathbf{Z}}$ denotes the learned proxy mapping. In contrast, $\mathbf{U} \notin \mathcal{O}$ for the observed set $\mathcal{O} = \{\mathbf{A}, \mathbf{V}\}$. The proxy $\widehat{\mathbf{Z}}$ supports direct adjustment, while replacing $\mathbf{U}$ with an observable surrogate does not guarantee that its back-door path is blocked. CCL therefore conditions audio adjustment on the proxied context and visual mediation on the corrected audio.

3.2 CONFIDENCE-CALIBRATED AUDIO INTERVENTION

To estimate the causal contribution of audio, we adjust for the back-door path $\mathbf{A} \leftarrow \mathbf{Z} \rightarrow \mathbf{Y}$. While $\mathbf{Z}$ is not explicitly labeled, its cross-modal context can be approximated from the training distribution. Applying $\mathtt{do}(\mathbf{A})$ cuts the incoming edge from $\mathbf{Z}$ to $\mathbf{A}$. Conditional adjustment with the observed visual context gives:

$$P(\mathbf{Y} \mid \mathtt{do}(\mathbf{A}), \mathbf{V}) = \mathbb{E}_{\mathbf{Z}|\mathbf{V}}[P(\mathbf{Y} \mid \mathbf{A}, \mathbf{V}, \mathbf{Z})]\,. \tag{2}$$

Conditioning on $\mathbf{V}$ leaves the adjustment distribution as $P(\mathbf{Z} \mid \mathbf{V})$, rather than the marginal $P(\mathbf{Z})$. Given $\mathbf{Z}$, the interventional conditional reduces to its observational counterpart. For a tractable feature operation, we approximate this conditional distribution with a co-occurrence dictionary $\mathcal{Z} = \{\mathbf{z}_k\}_{k=1}^{K_Z}$, where K_Z is the number of context prototypes. Audio-only clustering primarily organizes acoustic conditions, whereas the confounder of interest describes recurring relations between sounds and scenes. We therefore apply K-means to the concatenation of the global audio feature $\mathbf{f}_{\mathbf{A}} \in \mathbb{R}^d$ and pooled visual feature $\mathbf{f}_{\mathbf{V}}^{\mathrm{g}} \in \mathbb{R}^d$. Each centroid is partitioned as $\mathbf{z}_k = [\mathbf{z}_k^{\mathbf{A}}, \mathbf{z}_k^{\mathbf{V}}]$, pairing a visual context prototype with its associated acoustic prior. A visual query produces the posterior weight $\alpha_k(\mathbf{V}) = \mathtt{softmax}_k((\mathbf{f}_{\mathbf{V}}^{\mathrm{g}})^\top \mathbf{z}_k^{\mathbf{V}} / \sqrt{d})$. Using vision for retrieval mirrors $P(\mathbf{Z} \mid \mathbf{V})$ and prevents the audio representation under correction from choosing its own context estimate. Let $N_k^{\mathcal{Z}}$ be the number of training pairs assigned to cluster k and let $p_k = N_k^{\mathcal{Z}} / \sum_{\ell=1}^{K_Z} N_\ell^{\mathcal{Z}}$ be

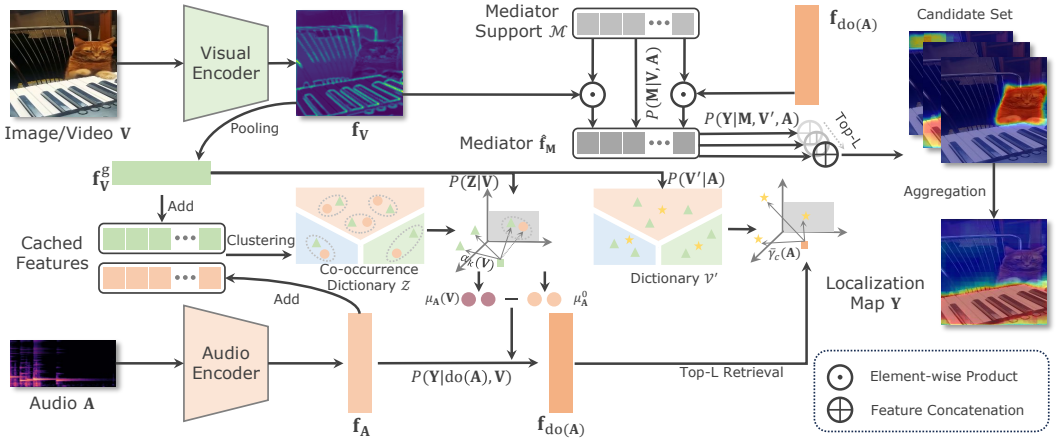

Figure 3: Pipeline of CCL. Audio and visual encoders extract a global audio feature, dense visual features, and a pooled visual context. The pooled context queries the co-occurrence dictionary to obtain corrected audio, which gates mediator prototypes and retrieves the Top-L alternative visual contexts. Fusing the location-wise mediator with each retrieved context produces candidate localization maps, whose weighted aggregation gives the final prediction and pair scores for optimization.

its empirical frequency. A fixed subtraction treats every retrieval as equally reliable and may remove source-specific information when the context posterior is diffuse. We therefore separate the conditional prior $\mu_\mathbf{A}(\mathbf{V}) = \sum_{k=1}^{K_Z} \alpha_k(\mathbf{V})\mathbf{z}^\mathbf{A}_k$ from the corpus-wide prior $\mu^0_\mathbf{A} = \sum_{k=1}^{K_Z} p_k \mathbf{z}^\mathbf{A}_k$. The difference $\Delta\mu_\mathbf{A}(\mathbf{V}) = \mu_\mathbf{A}(\mathbf{V}) - \mu^0_\mathbf{A}$ retains only the scene-specific acoustic deviation. The normalized retrieval entropy $\mathcal{H}_\mathbf{Z}(\mathbf{V}) = -\sum_{k=1}^{K_Z} \alpha_k(\mathbf{V}) \log(\alpha_k(\mathbf{V}) + \varepsilon_\mathrm{H})/\log K_Z$ determines the removal magnitude, where $\varepsilon_\mathrm{H} > 0$ provides numerical stability. The confidence is $r_\mathbf{A}(\mathbf{V}) = [\mathtt{clip}_{[0,1]}(1 - \mathcal{H}_\mathbf{Z}(\mathbf{V}))]^{\eta_\mathbf{A}}$, where $\mathtt{clip}_{[0,1]}$ restricts its input to $[0,1]$ and $\eta_\mathbf{A} > 0$ controls the calibration. A concentrated retrieval produces a large correction, whereas an uncertain retrieval preserves more of the input audio. The resulting feature-level intervention is:

$$\mathbf{f}_{\mathrm{do}(\mathbf{A})} = \mathtt{LN}\left(\mathbf{f}_\mathbf{A} - r_\mathbf{A}(\mathbf{V})\Delta\mu_\mathbf{A}(\mathbf{V})\right), \tag{3}$$

where $\mathtt{LN}$ denotes layer normalization. Centering by $\mu^0_\mathbf{A}$ prevents CAI from suppressing acoustic structure shared across the corpus, while entropy calibration reduces aggressive correction when the context dictionary is ambiguous. Both the prior and its confidence are determined by the visual query, preventing the audio representation under correction from selecting the direction or strength of its own subtraction.

3.3 Prediction-Level Visual Intervention

Unlike the audio confounder $\mathbf{Z}$, the visual confounder $\mathbf{U}$ does not admit a reliable proxy for direct adjustment. We therefore introduce a mediator $\mathbf{M}$ and use conditional front-door adjustment. The construction assumes that $\mathbf{M}$ intercepts every directed path from $\mathbf{V}$ to $\mathbf{Y}$, that $\mathbf{V} \to \mathbf{M}$ is unconfounded after conditioning on $\mathbf{A}$, and that conditioning on $(\mathbf{V}, \mathbf{A})$ blocks the back-door paths from $\mathbf{M}$ to $\mathbf{Y}$. Under these conditions,

$$P(\mathbf{Y} \mid \mathrm{do}(\mathbf{V}), \mathbf{A}) = \mathbb{E}_{\mathbf{M}|\mathbf{V},\mathbf{A}} \mathbb{E}_{\mathbf{V}'|\mathbf{A}}[P(\mathbf{Y} \mid \mathbf{M}, \mathbf{V}', \mathbf{A})]. \tag{4}$$

The outer expectation retains input-specific spatial information, whereas the inner expectation marginalizes the outcome over alternative visual contexts. Here, $\mathbf{V}'$ is an integration variable over the support of the visual data, not a second frame paired with the current input. Marginalizing context alone discards the spatial route required for localization, while mediation alone leaves the outcome tied to the current image distribution. We implement Eq. (4) through three connected steps.

Global Visual Expectation $\mathbb{E}_{\mathbf{V}'|\mathbf{A}}$. The inner expectation over alternative visual contexts uses the same sample-level feature $\mathbf{f}^{\mathrm{g}}_\mathbf{V} \in \mathbb{R}^d$ introduced in CAI, obtained by globally pooling the current visual feature map. K-means over the training set $\{\mathbf{f}^{\mathrm{g}}_{\mathbf{V},t}\}_{t=1}^{N_{\mathrm{tr}}}$ constructs the dictionary $\mathcal{V}' = \{\mathbf{v}'_c\}_{c=1}^{C}$, where N_{tr} is the number of training pairs and C is the number of visual-context prototypes. Each

$\mathbf{v}'_c \in \mathbb{R}^d$ is a dataset-level centroid in the same feature space as $\mathbf{f}^{\mathrm{g}}_{\mathbf{V}}$. The dictionary $\mathcal{V}'$ provides a discrete support approximation for the integration variable $\mathbf{V}'$. Corrected audio retrieves the alternative-context distribution through $\gamma_c(\mathbf{A}) = \mathtt{softmax}_c((\mathbf{v}'_c)^\top \mathbf{f}_{\mathtt{do}(\mathbf{A})}/(\tau_{\mathbf{V}'}\sqrt{d}))$, with temperature $\tau_{\mathbf{V}'} > 0$. Let $\mathcal{I}_L(\mathbf{A})$ contain the indices of the L largest weights, where $1 \le L \le C$, and let $\overline{\gamma}_c(\mathbf{A}) = \gamma_c(\mathbf{A})/\sum_{a\in\mathcal{I}_L(\mathbf{A})}\gamma_a(\mathbf{A})$ for $c \in \mathcal{I}_L(\mathbf{A})$. We retain these contexts and the associated probabilities so that the context expectation is evaluated after nonlinear prediction.

Audio-Gated Mediator Expectation $\mathbb{E}_{\mathbf{M}|\mathbf{V},\mathbf{A}}$. The outer mediator expectation uses the dense visual feature map $\mathbf{f}_{\mathbf{V}} \in \mathbb{R}^{N\times d}$, where N is the number of spatial locations. A discrete parameterization of the mediator support is defined once as $\mathcal{M} = [\mathbf{m}_1^\top; \ldots; \mathbf{m}_{K_M}^\top] \in \mathbb{R}^{K_M\times d}$, where K_M is the number of mediator prototypes and the s-th row contains the row-normalized prototype $\mathbf{m}_s \in \mathbb{R}^d$. At location n, the term $\mathbf{f}_{\mathbf{V},n}^\top \mathbf{m}_s/\sqrt{d}$ measures visual compatibility, while $g_s(\mathbf{A}) = \sigma(\mathbf{m}_s^\top \mathbf{f}_{\mathtt{do}(\mathbf{A})}/\sqrt{d})$ measures corrected-audio support, with σ denoting the sigmoid function. We combine both terms in log-probability space and form $q_{n,s}(\mathbf{V},\mathbf{A}) = \mathtt{softmax}_s((\mathbf{f}_{\mathbf{V},n}^\top \mathbf{m}_s/\sqrt{d} + \lambda_{\mathrm{g}}\log(g_s(\mathbf{A}) + \varepsilon_{\mathrm{g}}))/\tau_{\mathrm{M}})$, where $\lambda_{\mathrm{g}} \ge 0$ controls the audio gate, $\tau_{\mathrm{M}} > 0$ is the mediator temperature, and $\varepsilon_{\mathrm{g}} > 0$. The normalized mediator expectation is:

$$\widehat{\mathbf{f}}_{\mathbf{M},n} = \sum_{s=1}^{K_M} q_{n,s}(\mathbf{V},\mathbf{A})\mathbf{m}_s. \tag{5}$$

The posterior $q_{n,s}$ directly approximates $P(\mathbf{M} \mid \mathbf{V},\mathbf{A})$. Its normalization separates mediator selection from feature magnitude, so variations in total audio-gate activation do not introduce an uncontrolled scale into $\widehat{\mathbf{f}}_{\mathbf{M},n}$. The same mediator matrix also defines the diversity regularizer,

$$\mathcal{L}_{\mathrm{ICM}} = \left\|\mathcal{M}\mathcal{M}^\top - \mathbf{I}_{K_M}\right\|_F^2, \tag{6}$$

where $\mathbf{I}_{K_M}$ is the identity matrix and $\|\cdot\|_F$ is the Frobenius norm. This penalty acts on the prototypes used in Eq. (5), linking mediator construction to prototype diversity.

Causal Feature Recombination $P(\mathbf{Y} \mid \mathbf{M},\mathbf{V}',\mathbf{A})$. The outer mediator expectation is approximated by its posterior mean before nonlinear prediction. A multilayer perceptron (MLP) combines this mean with each retained context separately as $\mathbf{h}_n^{(c)} = \mathtt{MLP}([\widehat{\mathbf{f}}_{\mathbf{M},n}, \mathbf{v}'_c])$. Its compatibility with corrected audio is $y_n^{(c)} = (\mathbf{h}_n^{(c)})^\top \mathbf{f}_{\mathtt{do}(\mathbf{A})}/(\|\mathbf{h}_n^{(c)}\|_2\|\mathbf{f}_{\mathtt{do}(\mathbf{A})}\|_2 + \varepsilon_{\cos})$, where $\varepsilon_{\cos} > 0$ provides numerical stability. The context-specific score $y_n^{(c)}$ evaluates this approximation under alternative context $\mathbf{v}'_c$. We then average these scores to approximate the inner context expectation in Eq. (4):

$$\widehat{Y}_n = \sum_{c\in\mathcal{I}_L(\mathbf{A})} \overline{\gamma}_c(\mathbf{A}) y_n^{(c)}, \tag{7}$$

where each $y_n^{(c)}$ retains the same local mediator representation but evaluates it under a different audio-compatible context. Averaging these scores marginalizes alternative contexts after the nonlinear predictor. Contrastive training additionally requires a scalar score for an arbitrary candidate pair $(\mathbf{A}_i, \mathbf{V}_j)$. We compute $\mathbf{f}_{\mathtt{do}(\mathbf{A}),i}$ once from the native synchronized pair $(\mathbf{A}_i, \mathbf{V}_i)$ and reuse its audio gate and alternative-context probabilities when comparing it with candidate frames. For each retained context, recombining the mediator mean for $\mathbf{V}_j$ with $\mathbf{v}'_c$ and evaluating the cosine score against $\mathbf{f}_{\mathtt{do}(\mathbf{A}),i}$ gives $y_{ij,n}^{(c)}$. A smooth log-sum-exp aggregation retains strong localized responses while remaining differentiable,

$$\widehat{S}_{i,j} = \sum_{c\in\mathcal{I}_L(\mathbf{A}_i)} \overline{\gamma}_c(\mathbf{A}_i)\tau_{\mathrm{S}} \log\left[\frac{1}{N}\sum_{n=1}^{N} \exp\left(y_{ij,n}^{(c)}/\tau_{\mathrm{S}}\right)\right], \quad \tau_{\mathrm{S}} > 0. \tag{8}$$

For synchronized pairs, the unaggregated scores $\widehat{Y}_n$ form the localization map. The scalar $\widehat{S}_{i,j}$ in Eq. (8) connects prediction-level context marginalization to pair optimization.

3.4 Causal Contrastive Optimization

Standard pair construction induces a context-dependent selection process. Native positives usually share both physical correspondence and scene context, whereas random negatives usually differ in both. We first define contextual overlap at the distribution level as $\mathcal{D}(\mathbf{A}_i, \mathbf{V}_j) = \mathtt{Sim}\big(P(\mathbf{Z} \mid$

$\mathbf{A}_i), P(\mathbf{U} \mid \mathbf{V}_j)\big)$. $\texttt{Sim}$ measures the similarity between the audio-side and vision-side context posteriors. Using the audio partition of the co-occurrence dictionary, we define $\alpha_k^{\mathbf{A}}(\mathbf{A}_i) = \texttt{softmax}_k(\mathbf{f}_{\mathbf{A},i}^\top \mathbf{z}_k^{\mathbf{A}}/\sqrt{d})$. The visual counterpart is $\alpha_k^{\mathbf{V}}(\mathbf{V}_j) = \alpha_k(\mathbf{V}_j)$, using the visual-query weights defined in CAI. These scalar weights form the posterior vectors $\boldsymbol{\alpha}^{\mathbf{A}}(\mathbf{A}_i), \boldsymbol{\alpha}^{\mathbf{V}}(\mathbf{V}_j) \in \mathbb{R}^{K_Z}$. Their dot product gives:

$$\mathcal{D}_{ij} = \mathcal{D}(\mathbf{A}_i, \mathbf{V}_j) = \sum_{k=1}^{K_Z} \alpha_k^{\mathbf{A}}(\mathbf{A}_i)\alpha_k^{\mathbf{V}}(\mathbf{V}_j). \tag{9}$$

The value $\mathcal{D}_{ij} \in [0, 1]$ increases when $\mathbf{A}_i$ and $\mathbf{V}_j$ induce similar context distributions. A synchronized pair with low contextual overlap is termed a *causal hard positive*; it preserves physical correspondence under atypical context. A mismatched pair with high contextual overlap is termed a *causal hard negative*; it shares context without physical correspondence. Let J_{ij}^+ and J_{ij}^- indicate that the pair enters the objective as a positive or negative, respectively. Let $\pi^+(\omega)$ and $\pi^-(\omega)$ denote the corresponding selection propensities at overlap value ω, and let ℓ_{ij}^+ and ℓ_{ij}^- denote the pair losses. Under conditional exchangeability given $\mathcal{D}_{ij}$ and nonzero propensities, inverse probability weighting (IPW) Robins et al. (2000) gives:

$$\mathcal{R}_{\texttt{IPW}} = \mathbb{E}\left[\frac{J_{ij}^+}{\pi^+(\mathcal{D}_{ij})}\ell_{ij}^+ + \frac{J_{ij}^-}{\pi^-(\mathcal{D}_{ij})}\ell_{ij}^-\right]. \tag{10}$$

Native positives with atypical context and random negatives with shared context are rare under the original sampler. Increasing the influence of both groups limits the use of scene agreement as a surrogate pair label. Since the selection propensities are unavailable, we use monotone inverse-propensity surrogates. For a batch of size B, define the raw positive weight $u_{i,i}^+ = (\mathcal{D}_{ii} + \varepsilon_{\text{w}})^{-1}$ and raw negative weight $u_{i,j}^- = \mathcal{D}_{ij}$ for $i \neq j$, where $\varepsilon_{\text{w}} > 0$ provides numerical stability. The batch means are $\overline{u}^+ = B^{-1}\sum_{a=1}^B u_{a,a}^+$ and $\overline{u}^- = [B(B-1)]^{-1}\sum_{a=1}^B \sum_{b\neq a} u_{a,b}^-$. We use $w_{i,i}^+ = u_{i,i}^+/\overline{u}^+$ and $w_{i,j}^- = u_{i,j}^-/(\overline{u}^- + \varepsilon_{\text{w}})$ to keep the average scales stable across batches. Exact IPW recovery requires these overlap functions to be proportional to the corresponding inverse selection propensities; otherwise, the weights provide a rank-preserving correction toward rare contextual configurations. The pairwise score $\widehat{S}_{i,j}$ from Eq. (8) enters CCO through $E_{i,j} = \exp(\widehat{S}_{i,j}/\tau_{\text{CI}})$, where $\tau_{\text{CI}} > 0$ is the contrastive temperature. Substituting the normalized IPW-inspired weights and $E_{i,j}$ gives the *causal InfoNCE loss*:

$$\mathcal{L}_{\text{CI}} = -\frac{1}{B}\sum_{i=1}^B \log \frac{w_{i,i}^+ E_{i,i}}{w_{i,i}^+ E_{i,i} + \sum_{j\neq i} w_{i,j}^- E_{i,j}}. \tag{11}$$

The final objective is $\mathcal{L} = \mathcal{L}_{\text{CI}} + \lambda_{\text{ICM}}\mathcal{L}_{\text{ICM}}$, where $\lambda_{\text{ICM}} \geq 0$. CAI first suppresses a confidence-calibrated acoustic shortcut, PVI marginalizes alternative contexts at the prediction level, and CCO limits the reintroduction of contextual bias through pair construction.

4 Experiments

Datasets. Following Park et al. (2026), we use VGGSound 144K Chen et al. (2020) for training and evaluate on VGG-SS Chen et al. (2021) and SoundNet-Flickr Senocak et al. (2018), which contain 5,000 and 250 box-annotated samples, respectively. We further assess transferability and robustness on AVSBench Zhou et al. (2022) and the Extended VGG-SS/SoundNet-Flickr benchmarks Mo & Morgado (2022a). AVSBench provides pixel-level masks for the single-source S4 and multi-source MS3 subsets with 740 and 64 videos, respectively, while the extended benchmarks include non-audible frames, non-visible sources, and mismatched audio-visual pairs.

4.1 Main Results

Quantitative Comparison. Each table separates conventional encoders without CLIP from CLIP-based foundation models. CCL performs strongly in both groups, indicating consistent gains across learned and pretrained representations. Results on VGG-SS and SoundNet-Flickr further suggest reduced dependence on dataset-specific audio-visual co-occurrence. The open-set splits in Table 3

Table 1: Results on VGG-SS and SoundNet-Flickr. † denotes methods with OGL Mo & Morgado (2022b).

Method	VGG-SS		SoundNet-Flickr	
	cIoU ↑	AUC ↑	cIoU ↑	AUC ↑
LVS Chen et al. (2021)	34.40	38.20	71.90	58.20
EZ-VSL† Mo & Morgado (2022b)	38.85	39.54	83.94	63.60
SLAVC† Mo & Morgado (2022a)	39.80	-	86.00	-
MarginNCE Park et al. (2023)	38.25	39.06	83.94	63.20
FNAC Sun et al. (2023)	39.50	39.66	84.73	63.76
FNAC† Sun et al. (2023)	41.85	40.80	85.14	64.30
Alignment Senocak et al. (2025)	39.94	40.02	79.60	63.44
Alignment† Senocak et al. (2025)	42.64	41.48	82.40	64.60
JSA Kim et al. (2025)	40.62	41.66	83.60	64.68
JSA† Kim et al. (2025)	42.56	42.46	84.80	64.82
CCL (Ours)	43.72	42.15	87.28	65.48
CCL† (Ours)	**45.63**	**43.28**	**88.36**	**66.21**
WAV2CLIP Wu et al. (2022)	37.71	39.93	26.00	29.60
AudioCLIP Guzhov et al. (2022)	44.15	46.23	47.20	45.22
CLIPSeg Lüddecke & Ecker (2022)	49.09	45.75	68.00	54.96
AclCLIP Park et al. (2024)	49.46	46.32	**80.80**	64.62
AclCLIP++ Park et al. (2026)	52.11	46.91	76.00	63.20
CCL (Ours)	**53.84**	**48.12**	79.07	**65.73**

Table 2: Results on the extended VGG-SS and SoundNet-Flickr benchmarks.

Method	Extended VGG-SS			Extended Flickr		
	AP ↑	max-F1 ↑	LocAcc ↑	AP ↑	max-F1 ↑	LocAcc ↑
SLAVC Mo & Morgado (2022a)	32.95	40.00	37.79	51.63	59.10	83.60
MarginNCE Park et al. (2023)	30.58	36.80	38.25	57.99	61.80	83.94
FNAC Sun et al. (2023)	23.48	33.70	39.50	50.40	62.30	84.73
Alignment Senocak et al. (2025)	34.73	40.70	39.94	64.43	66.90	79.60
CCL (Ours)	**36.84**	**42.62**	**43.72**	**66.18**	**68.12**	**87.28**
WAV2CLIP Wu et al. (2022)	26.67	33.00	37.71	20.99	24.80	29.60
AudioCLIP Guzhov et al. (2022)	23.79	32.80	44.15	34.00	38.80	45.22
CLIPSeg Lüddecke & Ecker (2022)	34.96	41.00	49.09	55.14	57.00	68.00
AclCLIP Park et al. (2024)	40.79	49.10	49.46	76.07	73.20	80.80
AclCLIP++ Park et al. (2026)	46.72	51.90	52.11	75.21	72.40	76.00
CCL (Ours)	**48.13**	**53.20**	**55.84**	**77.34**	**74.62**	**84.67**

Table 3: Results on open-set localization. We use the training and test splits of Mo & Morgado (2022b).

Method	Heard 110		Unheard 110	
	cIoU ↑	AUC ↑	cIoU ↑	AUC ↑
LVS Chen et al. (2021)	28.90	36.20	26.30	34.70
EZ-VSL Mo & Morgado (2022b)	31.86	36.19	32.66	36.72
EZ-VSL† Mo & Morgado (2022b)	37.25	38.97	39.57	39.60
SLAVC Mo & Morgado (2022a)	35.84	-	36.50	-
SLAVC† Mo & Morgado (2022a)	38.22	-	38.87	-
FNAC† Sun et al. (2023)	39.54	39.83	42.91	41.17
Alignment Senocak et al. (2025)	38.31	39.05	39.11	39.80
Alignment† Senocak et al. (2025)	41.85	40.93	42.94	41.54
JSA Kim et al. (2025)	38.46	40.15	40.35	41.25
JSA† Kim et al. (2025)	41.50	41.60	43.11	42.04
CCL (Ours)	43.68	42.53	45.21	43.17
CCL† (Ours)	**45.74**	**43.66**	**46.83**	**44.15**
CLIPSeg Lüddecke & Ecker (2022)	49.73	45.35	43.65	41.05
AclCLIP Park et al. (2024)	48.44	45.06	41.98	41.55
AclCLIP++ Park et al. (2026)	49.15	45.33	43.06	40.24
CCL (Ours)	**52.36**	**47.24**	**47.28**	**44.31**

Table 4: Results on AVSBench.

Method	S4		MS3	
	mIoU ↑	F-score ↑	mIoU ↑	F-score ↑
SLAVC Mo & Morgado (2022a)	28.10	34.60	24.37	25.56
MarginNCE Park et al. (2023)	33.27	45.33	27.31	31.56
FNAC Sun et al. (2023)	27.15	31.40	21.98	22.50
Alignment Senocak et al. (2025)	29.60	35.90	-	-
JSA Kim et al. (2025)	31.33	44.30	-	-
CCL (Ours)	**35.62**	**46.85**	**29.18**	**32.74**
WAV2CLIP Wu et al. (2022)	28.70	35.35	25.09	23.84
AudioCLIP Guzhov et al. (2022)	36.57	42.15	27.06	26.48
CLIPSeg Lüddecke & Ecker (2022)	49.82	56.43	42.57	46.72
AclCLIP Park et al. (2024)	59.76	69.03	41.08	46.67
AclCLIP++ Park et al. (2026)	59.80	69.80	43.60	48.40
CCL (Ours)	**62.24**	**72.87**	**45.38**	**51.21**

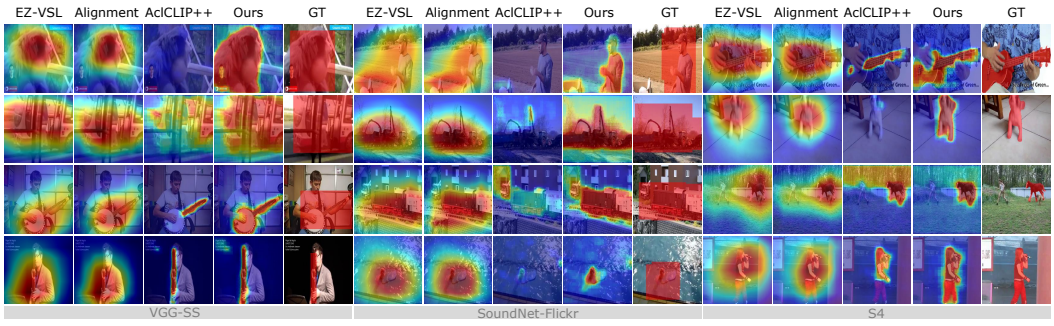

Figure 4: Qualitative comparison using samples from VGG-SS, SoundNet-Flickr, and S4.

compare categories seen and unseen during training. Improvements on unheard categories indicate that CCL better transfers the relation between sound and source region rather than relying on category-specific scene priors. Gains without OGL further distinguish this effect from object-guided refinement. The extended benchmarks in Table 2 decouple visual plausibility from acoustic support through non-audible, non-visible, and mismatched samples. CCL improves rejection on invalid pairs while retaining localization on valid ones, indicating greater reliance on source correspondence than on visually plausible sounding objects. AVSBench provides a stricter pixel-level evaluation in Table 4. S4 considers a single sound source, whereas MS3 contains multiple concurrent sources, where dominant sounds or scene priors can obscure weaker ones. CCL remains effective on MS3, with a smaller margin reflecting the increased difficulty of disentangling multiple sounding events.

Table 5: Core component ablation.

Components			VGG-SS		AVS (S4)		AVS (MS3)	
CAI	PVI	CCO	cIoU ↑	AUC ↑	mIoU ↑	F-score ↑	mIoU ↑	F-score ↑
✗	✗	✗	39.78	38.54	30.41	40.66	24.85	26.49
✓	✗	✗	41.35	39.84	31.94	42.74	26.17	28.71
✗	✓	✗	40.82	39.29	31.66	42.07	26.44	28.80
✗	✗	✓	40.36	38.90	31.15	41.47	25.52	27.53
✓	✓	✗	42.74	41.27	34.48	45.64	28.36	31.59
✓	✓	✓	**43.72**	**42.15**	**35.62**	**46.85**	**29.18**	**32.74**

Table 6: Modality intervention direction. B.D. and F.D. denote back-door and front-door.

Intervention Configuration				VGG-SS		AVS (S4)		AVS (MS3)	
Audio B.D.	Audio F.D.	Visual B.D.	Visual F.D.	cIoU ↑	AUC ↑	mIoU ↑	F-score ↑	mIoU ↑	F-score ↑
✗	✗	✗	✗	40.36	38.90	31.15	41.47	25.52	27.53
✗	✓	✓	✗	41.53	40.03	32.42	42.97	26.46	29.16
✗	✓	✗	✓	42.04	40.58	33.31	44.03	27.17	30.02
✓	✗	✓	✗	42.26	40.80	33.84	44.57	27.52	30.56
✓	✗	✗	✓	**43.72**	**42.15**	**35.62**	**46.85**	**29.18**	**32.74**

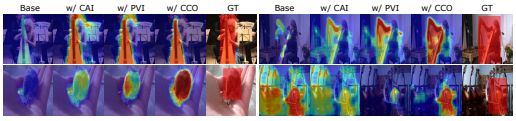

Figure 5: Qualitative changes produced by the core CCL components.

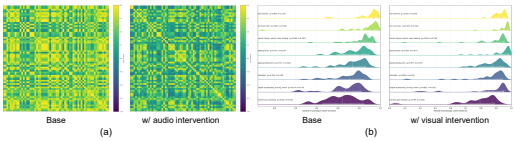

Figure 6: Prototype-similarity analysis of visual and audio interventions. (a) Inter-class prototype similarity. (b) Intra-class sample-to-prototype similarity.

Qualitative Comparison. Figure 4 further illustrates the localization behavior of CCL. First, CCL produces more fine-grained spatial responses with diverse activation shapes, rather than concentrating predictions around a single dominant region. Second, it reduces under-activation by improving sensitivity to acoustic source cues and recovering more complete sounding regions, as shown by the monkey in VGG-SS, the speaking person in Flickr, and the hand playing a guitar in S4. Third, it alleviates over-activation induced by strong contextual coupling between sound sources and surrounding objects. For example, the frequent co-occurrence of horses and trees in S4 can cause conventional methods to activate nearby trees. More generally, correlation-driven maps tend to follow semantic objectness and highlight visually plausible foreground regions that do not explain the sound. In contrast, CCL concentrates on regions whose spatial explanations remain consistent after removing the scene-conditioned acoustic prior.

4.2 Analysis and Discussion

Crucial Components. In Table 5, CAI first removes the most transferable shortcut, as global acoustic representations may encode scene context without resolving the responsible region. PVI complements this correction by preserving audio-supported spatial patterns while averaging over alternative visual contexts. CCO further reduces bias from pair construction. The consistent ordering on S4 and MS3 confirms that these corrections remain complementary from dominant to concurrent sources. Without pair-level correction, contrastive optimization may reinforce contextual regularities already suppressed in the feature space.

Qualitative and Representation Analysis. Figure 5 shows that progressively adding the CCL components shifts activation from false-positive or false-negative regions toward the actual sounding source, with the effect becoming more evident when multiple plausible sources are present, as in the guitar example. Figure 6 provides a complementary representation-level view. Lower cross-class prototype similarity indicates reduced ambiguity among audio-guided foreground features, while stronger sample-to-prototype alignment reflects more consistent intra-class structure. Together, the two analyses connect improved localization with greater inter-class separation and intra-class compactness.

Table 7: Audio dictionary construction.

Components			VGG-SS		AVS (S4)		AVS (MS3)	
Joint A-V Dictionary	Visual Query	Entropy	cIoU ↑	AUC ↑	mIoU ↑	F-score ↑	mIoU ↑	F-score ↑
✗	✗	✓	41.52	40.00	32.54	43.49	26.50	29.31
✓	✗	✓	42.56	41.11	34.34	45.41	27.91	31.13
✓	✓	✗	43.01	41.53	34.98	46.20	28.60	32.02
✓	✓	✓	**43.72**	**42.15**	**35.62**	**46.85**	**29.18**	**32.74**

Intervention Direction. In Table 6, applying front-door adjustment to both modalities preserves visual marginalization but introduces an unnecessary bottleneck for audio. Applying back-door adjustment to both modalities better controls acoustic context, yet a single visual proxy cannot account for spatially heterogeneous latent bias. The reversed assignment combines these two mismatches

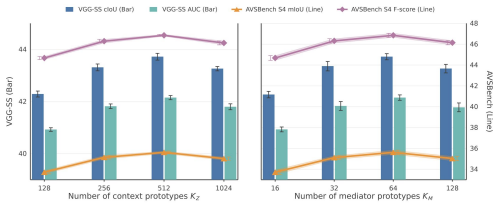

Figure 7: Capacity sensitivity of the context dictionary K_Z and the mediator bank K_M.

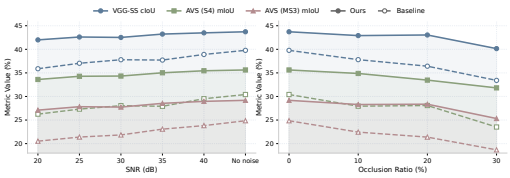

Figure 8: Modality robustness at different SNRs and occlusion ratios.

and performs worse than either partially aligned configuration. The proposed direction instead uses the empirical cross-modal proxy for audio and reserves prediction-level mediation for latent visual nuisance. This ordering follows confounder observability rather than modal symmetry.

Cross-Modal Confounder Identification. In Table 7, audio-only clustering may entangle source identity with environmental regularity, while joint clustering better captures recurring cross-modal configurations. A visual query estimates scene-expected sound without letting the corrected audio select prototypes containing target-source information. Entropy calibration further downweights unreliable corrections under diffuse retrieval, helping preserve source-related content. Together, these choices retain acoustic cues less predictable from visual context.

Context-Aware Pair Reweighting. In Table 8, native positives over-represent agreement between source and context, while random negatives rarely share context. Downweighting predictable positives prevents scene agreement from becoming part of the target. Emphasizing context-matched negatives forces source discrimination within similar environments. We analyze the two corrections as complementary controls over optimization pressure. Joint use prevents contrastive training from reconstructing a shortcut already removed from the feature space.

Table 8: Ablation of context-aware pair reweighting.

Components		VGG-SS		AVS (S4)		AVS (MS3)	
Positive IPW	Negative IPW	cIoU ↑	AUC ↑	mIoU ↑	F-score ↑	mIoU ↑	F-score ↑
✗	✗	42.74	41.27	34.48	45.64	28.36	31.59
✗	✓	43.09	41.58	34.88	46.04	28.66	32.02
✓	✗	43.31	41.80	35.13	46.35	28.80	32.23
✓	✓	**43.72**	**42.15**	**35.62**	**46.85**	**29.18**	**32.74**

Hyperparameter Sensitivity. In Figure 7, the context dictionary controls the granularity of contextual variation. Too few prototypes merge environments with different acoustic priors, whereas too many fragment recurring contexts and may capture source-specific details. The mediator bank governs the diversity of spatial source patterns. Limited capacity merges distinct patterns, while excessive capacity encourages near instance-specific assignments. Intermediate values therefore provide sufficient contextual diversity without turning either latent structure into an identity lookup.

Robustness Analysis. In Figure 8, we perturb audio and vision independently. Zero-mean Gaussian noise is scaled per waveform to the target signal-to-noise ratio (SNR), while random 16×16 visual patches are replaced by the channel-wise image mean at the specified occlusion ratio. Audio noise directly affects the corrected audio used for mediator gating and alternative-context retrieval in PVI. With vision fixed, the scene-conditioned prior and retrieval confidence in CAI remain unchanged, so CAI continues to suppress contextual bias rather than adapting the correction to acoustic corruption. Visual masking instead affects both the CAI query and PVI spatial assignments. When masking increases retrieval uncertainty, confidence calibration reduces the correction, while context averaging helps preserve predictions for source regions that remain visible. The gradual degradation across both perturbation types indicates stable robustness to modality corruption.

5 Conclusion

We propose CCL, a conditional causal framework that aligns bias correction with confounder observability. Through stage-specific causal adjustment, CCL focuses representation learning and contrastive supervision on physical audio-visual correspondence rather than recurring contextual regularities. It reframes SSL from maximizing pair agreement to locating the visual region that explains the sound, while retaining annotation-free training and compatibility with existing frameworks.

AI USE STATEMENT

Generative AI tools were used in this work for language refinement, primarily to improve the clarity, fluency, and grammatical quality of sections such as the abstract, introduction, methodology, experiments, and appendix. Their use was limited to polishing the English presentation and did not extend to activities including experimental design or scientific interpretation.

REFERENCES

Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. Self-supervised learning of audio-visual objects from video. In *European Conference on Computer Vision*, pp. 208–224. Springer, 2020.

Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 609–617. IEEE, 2017.

Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 435–451, 2018.

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

Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Localizing visual sounds the hard way. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16862–16871. IEEE, 2021.

Ziyang Chen, Shengyi Qian, and Andrew Owens. Sound localization from motion: Jointly learning sound direction and camera rotation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7863–7874. IEEE, 2023.

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

Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. *Advances in neural information processing systems*, 33:8765–8775, 2020.

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 *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2277–2286. IEEE, 2023.

Yuxin Guo, Shijie Ma, Hu Su, Zhiqing Wang, Yuhao Zhao, Wei Zou, Siyang Sun, and Yun Zheng. Dual mean-teacher: An unbiased semi-supervised framework for audio-visual source localization. *Advances in Neural Information Processing Systems*, 36:48639–48661, 2023.

Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio. In *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 976–980. IEEE, 2022.

Mark Hamilton, Andrew Zisserman, John R Hershey, and William T Freeman. Separating the "chirp" from the "chat": Self-supervised visual grounding of sound and language. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13117–13127. IEEE, 2024.

Di Hu, Feiping Nie, and Xuelong Li. Deep multimodal clustering for unsupervised audiovisual learning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10473–10482. IEEE, 2022.

Chao Huang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. Egocentric audio-visual object localization. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22910–22921. IEEE, 2023.

Yanhao Jia, Ji Xie, S Jivaganesh, Li Hao, Xu Wu, and Mengmi Zhang. Seeing sound, hearing sight: Uncovering modality bias and conflict of ai models in sound localization. *Advances in neural information processing systems*, 38:148468–148499, 2026.

Xavier Juanola, Gloria Haro, and Magdalena Fuentes. A critical assessment of visual sound source localization models including negative audio. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. 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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26457–26466. IEEE, 2024.

Inho Kim, Youngkil Song, Jicheol Park, Won Hwa Kim, and Suha Kwak. Improving sound source localization with joint slot attention on image and audio. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3121–3130. IEEE, 2025.

Sizhe Li, Yapeng Tian, and Chenliang Xu. Space-time memory network for sounding object localization in videos. *arXiv preprint arXiv:2111.05526*, 2021.

Tingle Li, Siddharth Gururani, Kevin J Shih, Gantavya Bhatt, Sang-gil Lee, Zhifeng Kong, Arushi Goel, Gopala Anumanchipalli, and Ming-Yu Liu. Benchmarking single-factor physical video-to-audio generation. *arXiv preprint arXiv:2605.30339*, 2026.

Bing Liu, Dong Wang, Xu Yang, Yong Zhou, Rui Yao, Zhiwen Shao, and Jiaqi Zhao. Show, deconfound and tell: Image captioning with causal inference. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18020–18029. IEEE, 2022a.

Jinxiang Liu, Chen Ju, Weidi Xie, and Ya Zhang. Exploiting transformation invariance and equivariance for self-supervised sound localisation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 3742–3753, 2022b.

Ruyang Liu, Hao Liu, Ge Li, Haodi Hou, TingHao Yu, and Tao Yang. Contextual debiasing for visual recognition with causal mechanisms. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12745–12755. IEEE, 2022c.

Tianyu Liu, Peng Zhang, Wei Huang, Yufei Zha, Tao You, and Yanning Zhang. Induction network: Audio-visual modality gap-bridging for self-supervised sound source localization. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 4042–4052, 2023.

Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7076–7086. IEEE, 2022.

Jie Ma, Min Hu, Pinghui Wang, Wangchun Sun, Lingyun Song, Hongbin Pei, Jun Liu, and Youtian Du. Look, listen, and answer: Overcoming biases for audio-visual question answering. *Advances in Neural Information Processing Systems*, 37:9507–9531, 2024.

Tanvir Mahmud, Yapeng Tian, and Diana Marculescu. T-vsl: Text-guided visual sound source localization in mixtures. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26732–26741. IEEE, 2024.

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

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

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

Khanh Binh Nguyen and Chae Jung Park. Souple: Enhancing audio-visual localization and segmentation with learnable prompt contexts. *arXiv preprint arXiv:2603.22732*, 2026.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual vqa: A cause-effect look at language bias. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12695–12705. IEEE, 2021.

Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H Adelson, and William T Freeman. Visually indicated sounds. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2405–2413, 2016.

Takashi Oya, Shohei Iwase, Ryota Natsume, Takahiro Itazuri, Shugo Yamaguchi, and Shigeo Morishima. Do we need sound for sound source localization? In *Asian Conference on Computer Vision*, pp. 119–136. Springer, 2020.

Sooyoung Park, Arda Senocak, and Joon Son Chung. Marginnce: Robust sound localization with a negative margin. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Sooyoung Park, Arda Senocak, and Joon Son Chung. Can clip help sound source localization? In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 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. *arXiv preprint arXiv:2604.06824*, 2026.

Judea Pearl. *Causality*. Cambridge university press, 2009.

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

James M Robins, Miguel Angel Hernan, and Babette Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560, 2000.

Hyeonggon Ryu, Seongyu Kim, Joon Son Chung, and Arda Senocak. Seeing speech and sound: Distinguishing and locating audios in visual scenes. *arXiv preprint arXiv:2503.18880*, 2025.

Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *2018 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 *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 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 *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 577–586. IEEE, 2022b.

Arda Senocak, Hyeonggon Ryu, Junsik Kim, Tae-Hyun Oh, Hanspeter Pfister, and Joon Son Chung. Sound source localization is all about cross-modal alignment. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7743–7753. IEEE, 2023.

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.

Arda Senocak, Sooyoung Park, Tae-Hyun Oh, and Joon Son Chung. How far can we go with synthetic data for audio-visual sound source localization? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22931–22940, 2026.

Nikhil Singh, Chih-Wei Wu, Iroro Orife, and Mahdi Kalayeh. Looking similar, sounding different: Leveraging counterfactual cross-modal pairs for audiovisual representation learning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26897–26908. IEEE, 2024.

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. *arXiv preprint arXiv:2203.13412*, 2022.

Zengjie Song, Jiangshe Zhang, Yuxi Wang, Junsong Fan, and Zhaoxiang Zhang. Enhancing sound source localization via false negative elimination. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10499–10514, 2024.

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 *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6420–6429. IEEE, 2023.

Yuewen Sun, Lingjing Kong, Guangyi Chen, Loka Li, Gongxu Luo, Zijian Li, Yixuan Zhang, Yujia Zheng, Mengyue Yang, Petar Stojanov, et al. Causal representation learning from multi-modal biomedical observations. *ArXiv*, pp. arXiv–2411, 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 31st 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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8342–8351. IEEE, 2025.

Jingyao Wang, Siyu Zhao, Wenwen Qiang, Jiangmeng Li, Changwen Zheng, Fuchun Sun, and Hui Xiong. Towards the causal complete cause of multi-modal representation learning. In *International Conference on Machine Learning*, pp. 65738–65775. PMLR, 2025.

Jun Wang, Fuyuan Cao, Zhixin Xue, Xingwang Zhao, and Jiye Liang. Cmob: Modality valuation via causal effect for balanced multimodal learning. *Advances in Neural Information Processing Systems*, 38:82585–82607, 2026.

Liuyi Wang, Zongtao He, Ronghao Dang, Mengjiao Shen, Chengju Liu, and Qijun Chen. Vision-and-language navigation via causal learning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13139–13150. IEEE, 2024.

Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. Wav2clip: Learning robust audio representations from clip. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4563–4567. IEEE, 2022.

Xinyi Wu, Zhenyao Wu, Lili Ju, and Song Wang. Binaural audio-visual localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 2961–2968, 2021.

Hanyu Xuan, Zhiliang Wu, Jian Yang, Yan Yan, and Xavier Alameda-Pineda. A proposal-based paradigm for self-supervised sound source localization in videos. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1019–1028. IEEE, 2022.

Hanyu Xuan, Zhiliang Wu, Jian Yang, Bo Jiang, Lei Luo, Xavier Alameda-Pineda, and Yan Yan. Robust audio-visual contrastive learning for proposal-based self-supervised sound source localization in videos. *IEEE transactions on pattern analysis and machine intelligence*, 46(7):4896–4907, 2024.

Dizhan Xue, Shengsheng Qian, and Changsheng Xu. Variational causal inference network for explanatory visual question answering. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2515–2525. IEEE, 2023.

Xu Yang, Hanwang Zhang, Guojun Qi, and Jianfei Cai. Causal attention for vision-language tasks. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9842–9852. IEEE, 2021.

Heeseung Yun, Ruohan Gao, Ishwarya Ananthabhotla, Anurag Kumar, Jacob Donley, Chao Li, Gunhee Kim, Vamsi Krishna Ithapu, and Calvin Murdock. Spherical world-locking for audio-visual localization in egocentric videos. In *European Conference on Computer Vision*, pp. 256–274. Springer, 2024.

Chuanqi Zang, Hanqing Wang, Mingtao Pei, and Wei Liang. Discovering the real association: Multimodal causal reasoning in video question answering. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19027–19036. IEEE, 2023.

Mingfeng Zha, Tianyu Li, Guoqing Wang, Peng Wang, Yangyang Wu, Yang Yang, and Heng Tao Shen. Implicit counterfactual learning for audio-visual segmentation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22349–22360. IEEE, 2025.

Dong Zhang, Hanwang Zhang, Jinhui Tang, Xian-Sheng Hua, and Qianru Sun. Causal intervention for weakly-supervised semantic segmentation. *Advances in neural information processing systems*, 33:655–666, 2020.

Jinxing Zhou, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, and Yiran Zhong. Audio–visual segmentation. In *European Conference on Computer Vision*, pp. 386–403. Springer, 2022.

Bingwen Zhu, Yuqian Fu, Qiaole Dong, Guolei Sun, Tianwen Qian, Yuzheng Wu, Danda Pani Paudel, Yanwei Fu, and Xiangyang Xue. Egosound: Benchmarking sound understanding in egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25589–25598, 2026.

A Extended Related Work **15**

B Training Procedure **16**

C More Experimental Setup **16**

C.1 Implementation Details 16

C.2 Evaluation Metrics 18

D More Comparison Results **19**

D.1 Different Training Bases 19

D.2 Visualization under Different Conditions 19

E More Ablation Studies **20**

E.1 Additional Component Visualizations 20

E.2 Alignment between Confounder Observability and Intervention Direction 21

E.3 Ablation of Visual Intervention 22

F Additional Experimental Analysis **22**

F.1 Model Efficiency 22

F.2 Sensitivity to the Retained-Context Budget 22

F.3 Failure Cases and Limitations 22

G Expanded Causal Derivations **23**

G.1 Observational Marginalization 23

G.2 Conditional Back-Door Adjustment for Audio 23

G.3 Conditional Front-Door Adjustment for Vision 23

G.4 Selection Reweighting for CCO 24

A Extended Related Work

Multimodal SSL. Foundational SSL studies show that audio-visual co-occurrence can support spatial grounding without location annotations Owens et al. (2016); Arandjelovic & Zisserman (2017; 2018); Hu et al. (2020). Subsequent methods refine this supervision through transformation consistency, object proposals, negative-free prediction, and related correspondence objectives Liu et al. (2022b); Xuan et al. (2022); Song et al. (2022); Liu et al. (2023); Xuan et al. (2024); Song et al. (2024). Although these designs improve pair discrimination and localization readout, synchronized observations still provide both the supervision signal and its surrounding context. Another line of work modifies the spatial prior or supervision regime. Class activation maps expose sounding regions from category supervision, while vision-only controls reveal how strongly appearance alone can predict localization Senocak et al. (2022b); Oya et al. (2020). Cross-modal alignment analysis further shows that stable image-level representations can rival specialized localization designs Senocak et al. (2023), while temporal memory and optical flow propagate localization support across ambiguous frames Li et al. (2021); Fedorishin et al. (2023); Choi et al. (2025). Weakly supervised objectives combine image-level alignment with object-guided refinement, whereas semi-supervised mean teachers exploit limited spatial annotations to produce sharper maps Mo & Morgado (2022b;a); Guo et al. (2023). Training composition provides a complementary direction.

Source separation and mixture-based composition expose multiple sounding objects within one observation Qian et al. (2020); Hu et al. (2022), while semantically related sampling and negative margins reduce class collision and excessive separation Senocak et al. (2022a); Park et al. (2023). Mixed-scene analysis further distinguishes speech from more general acoustic events Ryu et al. (2025). Recent studies also explore data synthesis and external foundation knowledge to improve localization Senocak et al. (2026); Park & Kim (2026), as well as localization across different camera viewpoints and sound configurations Huang et al. (2023); Chen et al. (2023); Yun et al. (2024); Wu et al. (2021). Collectively, these methods vary spatial continuity, supervision strength, pair construction, and acoustic composition, yet contextual regularities can remain predictive of correspondence. This motivates separating contextual dependence from source-specific grounding.

Causal Learning for Visual and Multimodal Understanding. Causal learning distinguishes observational association from interventional effects Pearl (2009). In visual recognition, context is commonly modeled as a confounder of observed features and predictions, with back-door adjustment and interventional attention used to suppress contextual shortcuts while preserving object information Liu et al. (2022c); Zhang et al. (2020). Vision-language methods extend this principle to token and region co-occurrence through deconfounded pretraining, causal attention, counterfactual question answering, and causal captioning Yang et al. (2021); Niu et al. (2021); Liu et al. (2022a). Video question answering and navigation further model confounding across visual content, language, and temporal history Zang et al. (2023); Xue et al. (2023); Wang et al. (2024). Audio-related settings introduce stronger cross-modal asymmetry. Audio-visual question answering isolates answer priors that persist after multimodal fusion Ma et al. (2024), while counterfactual audiovisual pairs vary sound under similar visual content to reduce appearance-driven correspondence Singh et al. (2024). Implicit counterfactual learning for AVS further suppresses category-level visual priors that remain predictive after audio replacement Zha et al. (2025). At a broader multimodal level, causal-effect valuation measures sample-specific modality contributions, while causal representation analysis studies identifiability, shared structure, and complete causes across coupled observations Wang et al. (2026); Sun et al. (2025); Wang et al. (2025). These directions complement sound-localization studies of modality conflict Jia et al. (2026). Recent benchmarks also introduce controlled interventions for audio-related causal analysis. FlatSounds varies individual physical factors to test whether video-to-audio models follow the generating process rather than textual or semantic shortcuts Li et al. (2026), while EgoSound evaluates spatial, cross-modal, and causal reasoning in egocentric sound understanding Zhu et al. (2026). These studies characterize causal dependence in multimodal decisions, but do not identify spatial interventions for self-supervised SSL. CCL addresses this gap through modality-specific causal adjustment and context-aware pair reweighting conditioned on cross-modal context.

B Training Procedure

We summarize the dependency among CAI, PVI, and CCO in Algorithm 1. CAI first uses retrieval entropy to calibrate removal of the context-specific acoustic prior. PVI then forms a normalized audio-visual mediator posterior and marginalizes predictions over the top-L alternative visual contexts. CCO finally reweights contrastive pairs according to contextual overlap. Here, $\mathcal{L}_{\mathrm{CI}}$ is the causal InfoNCE loss and $\mathcal{L}_{\mathrm{ICM}}$ is the mediator orthogonality loss; no additional CAI loss is used.

C More Experimental Setup

C.1 Implementation Details

Conventional Setting. We evaluate CCL with the conventional two-stream frameworks EZ-VSL, FNAC, and JSA Mo & Morgado (2022b); Sun et al. (2023); Kim et al. (2025). Each framework employs ResNet-18 encoders for vision and audio, with ImageNet-1K initialization for the visual branch and independent initialization for the audio branch. The resulting features are mapped into a 512-dimensional shared space. Training frames are spatially augmented, while evaluation frames are resized to 224×224. Audio segments are converted into log-spectrograms using the preprocessing protocol of the corresponding framework.

Algorithm 1: Training Procedure of CCL

Input: Training set $\mathcal{D}_{\mathrm{tr}}$; capacities K_Z, K_M, C; retained contexts L; total training iterations T
Output: Trainable projection layers, MLP, and mediator bank $\mathcal{M}$

// Construct global dictionaries
foreach $(\mathbf{A}_i, \mathbf{V}_i) \in \mathcal{D}_{\mathrm{tr}}$ **do**
  extract and cache $\mathbf{f}_{\mathbf{A},i}$ and $\mathbf{f}^{\mathrm{g}}_{\mathbf{V},i}$
$\mathcal{Z} \leftarrow \mathtt{KMeans}(\{[\mathbf{f}_{\mathbf{A},i}, \mathbf{f}^{\mathrm{g}}_{\mathbf{V},i}]\}_i, K_Z)$
split $\mathbf{z}_k \in \mathcal{Z}$ into $[\mathbf{z}^{\mathbf{A}}_k, \mathbf{z}^{\mathbf{V}}_k]$ and record cluster frequencies p_k
$\mu^0_{\mathbf{A}} \leftarrow \sum_k p_k \mathbf{z}^{\mathbf{A}}_k$
$\mathcal{V}' \leftarrow \mathtt{KMeans}(\{\mathbf{f}^{\mathrm{g}}_{\mathbf{V},i}\}_i, C)$; initialize row-normalized $\mathcal{M} \in \mathbb{R}^{K_M \times d}$
// Optimize CCL
for $t \leftarrow 1$ **to** T **do**
  sample a mini-batch of size B and reindex it by $\mathcal{B} = \{1, \ldots, B\}$
  row-normalize the mediator prototypes used in the current forward pass
  foreach $i \in B$ **do**
    extract current audio and visual features
    compute $\alpha_k(\mathbf{V}_i)$ and $r_{\mathbf{A}}(\mathbf{V}_i)$ as defined in CAI
    $\mu_{\mathbf{A}}(\mathbf{V}_i) \leftarrow \sum_k \alpha_k(\mathbf{V}_i)\mathbf{z}^{\mathbf{A}}_k$
    $\mathbf{f}_{\mathrm{do}(\mathbf{A}),i} \leftarrow \mathrm{LN}(\mathbf{f}_{\mathbf{A},i} - r_{\mathbf{A}}(\mathbf{V}_i)[\mu_{\mathbf{A}}(\mathbf{V}_i) - \mu^0_{\mathbf{A}}])$
    compute $q_{i,n,s} \equiv q_{n,s}(\mathbf{V}_i, \mathbf{A}_i)$ as defined before Eq. (5)
    $\widehat{\mathbf{f}}_{\mathbf{M},i,n} \leftarrow \sum_s q_{i,n,s}\mathbf{m}_s$
    compute $\gamma_c(\mathbf{A}_i)$, retain $\mathcal{I}_L(\mathbf{A}_i)$, and renormalize to $\overline{\gamma}_c(\mathbf{A}_i)$ as defined before Eq. (7)
    compute context-specific maps $\{y^{(c)}_{i,i,n} \mid c \in \mathcal{I}_L(\mathbf{A}_i)\}$
    $\widehat{Y}_{i,n} \leftarrow \sum_{c \in \mathcal{I}_L(\mathbf{A}_i)} \overline{\gamma}_c(\mathbf{A}_i) y^{(c)}_{i,i,n}$
  foreach $(i, j) \in B \times B$ **do**
    reuse $\mathbf{f}_{\mathrm{do}(\mathbf{A}),i}$ and its retained context weights, recompute the mediator posterior for $\mathbf{V}_j$, and obtain $\widehat{S}_{i,j}$ using Eq. (8)
    compute contextual overlap $\mathcal{D}_{ij}$ using Eq. (9)
  compute all mean-normalized $\{w^+_{i,i}\}_{i=1}^B$ and $\{w^-_{i,j}\}_{i \neq j}$ from the complete overlap matrix as defined before Eq. (11)
  $\mathcal{L} \leftarrow \mathcal{L}_{\mathrm{CI}} + \lambda_{\mathrm{ICM}}\mathcal{L}_{\mathrm{ICM}}$; update all trainable parameters

Table 9: Quantitative results of models trained on the VGGSound 10K dataset.

Method	VGG-SS		SoundNet-Flickr	
	cIoU ↑	AUC ↑	cIoU ↑	AUC ↑
LVS Chen et al. (2021)	-	-	61.80	53.60
EZ-VSL Mo & Morgado (2022b)	25.84	33.68	63.85	54.44
FNAC Sun et al. (2023)	37.29	38.99	85.74	63.66
JSA Kim et al. (2025)	37.84	39.82	86.00	64.90
CCL (Ours)	42.16	42.75	88.84	67.60
EZ-VSL† Mo & Morgado (2022b)	38.71	39.80	78.71	61.53
FNAC† Sun et al. (2023)	40.69	40.42	82.13	63.64
JSA† Kim et al. (2025)	41.02	41.29	87.60	65.42
CCL† (Ours)	44.11	43.85	89.82	68.35

CLIP-Based Setting. We further evaluate CCL with the CLIP-based pipelines adopted by AclCLIP and SOUPLE Park et al. (2024); Nguyen & Park (2026). A frozen ViT-B/16 CLIP encoder processes the center video frame, while BEATs encodes a ten-second audio segment sampled at 16 kHz. CLIPSeg provides spatial grounding from the resulting audio-conditioned representation. Frames are resized to 352×352. The pretrained encoders remain fixed, and optimization is restricted to lightweight task-specific layers and the parameters introduced by CCL.

Table 10: Quantitative results of models trained on the SoundNet-Flickr 10K dataset.

Method	VGG-SS		SoundNet-Flickr	
	cIoU ↑	AUC ↑	cIoU ↑	AUC ↑
Attention Senocak et al. (2018)	-	-	43.60	44.90
CoarseToFine Qian et al. (2020)	-	-	52.20	49.60
AVObject Afouras et al. (2020)	-	-	54.60	50.40
LVS Chen et al. (2021)	-	-	58.20	52.50
EZ-VSL Mo & Morgado (2022b)	19.86	30.96	62.24	54.74
FNAC Sun et al. (2023)	35.27	38.00	84.33	63.26
JSA Kim et al. (2025)	35.11	38.66	85.20	65.26
CCL (Ours)	39.67	39.74	87.35	66.93
EZ-VSL† Mo & Morgado (2022b)	37.61	39.21	81.93	62.58
FNAC† Sun et al. (2023)	40.97	40.38	84.73	64.34
JSA† Kim et al. (2025)	39.16	40.51	87.60	66.50
CCL† (Ours)	42.00	41.15	89.19	67.43

Table 11: Quantitative results of models trained on the SoundNet-Flickr 144K dataset.

Method	VGG-SS		SoundNet-Flickr	
	cIoU ↑	AUC ↑	cIoU ↑	AUC ↑
Attention Senocak et al. (2018)	-	-	66.00	55.80
DMC Hu et al. (2019)	-	-	67.10	56.80
LVS Chen et al. (2021)	-	-	69.90	57.30
HardPos Senocak et al. (2022a)	-	-	75.20	59.70
EZ-VSL Mo & Morgado (2022b)	30.27	35.92	72.69	58.70
FNAC Sun et al. (2023)	33.93	37.29	78.71	59.33
Alignment Senocak et al. (2025)	-	-	85.20	62.20
JSA Kim et al. (2025)	37.01	39.42	86.00	65.16
CCL (Ours)	41.83	40.67	88.50	65.95
EZ-VSL† Mo & Morgado (2022b)	41.01	40.23	83.13	63.06
FNAC† Sun et al. (2023)	41.10	40.44	83.93	63.06
Alignment† Senocak et al. (2025)	-	-	84.00	64.16
JSA† Kim et al. (2025)	40.15	40.75	89.20	64.50
CCL† (Ours)	43.70	41.98	90.40	66.73

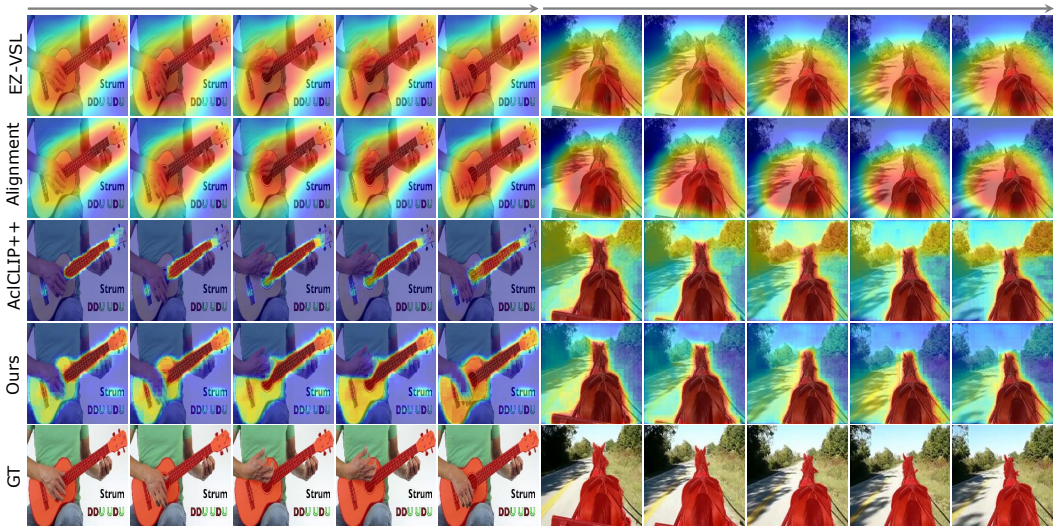

Figure 9: Qualitative comparison on AVSBench S4. CCL covers the sounding guitar and reduces background activation in the horse scene.

In both settings, we construct the joint dictionary from concatenated audio and pooled visual features and the alternative-context dictionary from pooled visual features. We use $K_Z = 512$, $K_M = 64$, and $L = 8$ unless stated otherwise. The dictionaries remain fixed after construction, while mediator prototypes are normalized during each forward pass.

C.2 Evaluation Metrics

Standard and Open-Set Localization. For VGG-SS and SoundNet-Flickr Senocak et al. (2018); Chen et al. (2021), IoU measures overlap with the consensus map. cIoU is the percentage of samples with IoU at least 0.5. AUC summarizes success rates over thresholds from 0 to 1 at intervals of 0.05. Heard and unheard splits use the same parameters.

Extended Localization. Extended VGG-SS and Extended SoundNet-Flickr Mo & Morgado (2022a) jointly assess source presence, correspondence confidence, and localization. A true positive requires a visible source, confidence above the current threshold, and IoU at least 0.5. Sweeping the confidence threshold yields AP and max-F1. LocAcc reports localization success at IoU 0.5 on visible-source samples without confidence filtering.

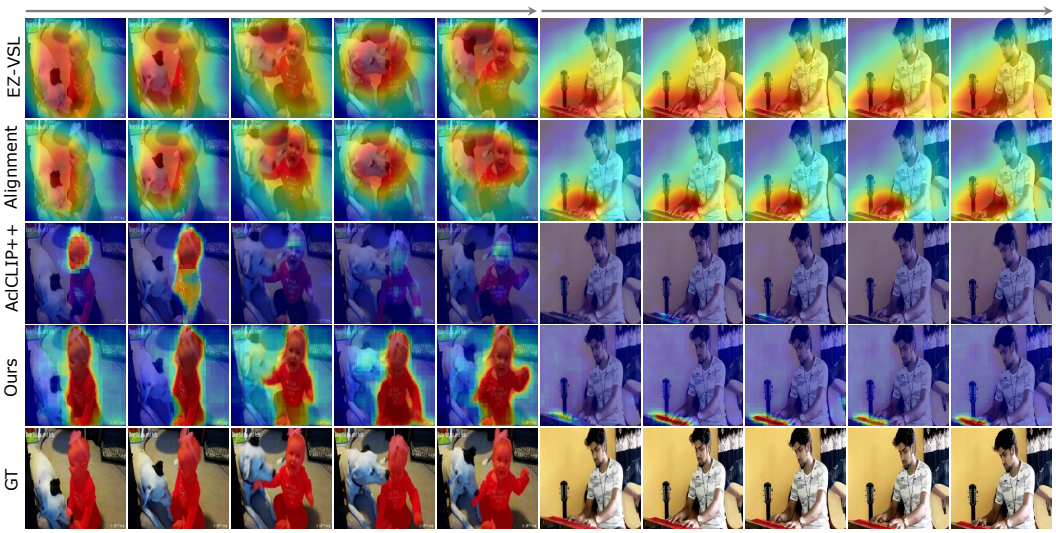

Figure 10: Qualitative comparison on AVSBench MS3. CCL separates the sounding dog and person from nearby visual candidates in multi-source scenes.

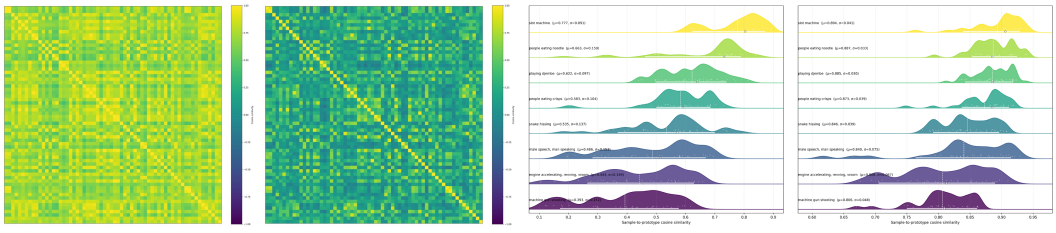

Figure 11: Extended prototype analysis using EZ-VSL as the reference. (a) Similarity across class prototypes. (b) Similarity between samples and class prototypes.

Audio-Visual Segmentation. On AVSBench S4 and MS3 Zhou et al. (2022), mIoU averages pixel overlap over all frames. F-score uses F_β with $\beta^2 = 0.3$ and reports the maximum mean value over 255 uniformly spaced thresholds in $[0, 1)$.

D More Comparison Results

D.1 Different Training Bases

We further report results using VGGSound 10K, SoundNet-Flickr 10K, and SoundNet-Flickr 144K in Tables 9, 10, and 11, following prior protocols Mo & Morgado (2022b); Sun et al. (2023); Kim et al. (2025). With only 10K training pairs, recurring audio-scene relations are estimated from a narrower sample and may exert stronger influence on cross-modal matching. CCL remains effective when trained on either VGGSound or SoundNet-Flickr, improving both in-domain evaluation and transfer to the other benchmark. This consistency across training sources suggests reduced reliance on dataset-specific correlations. The gains also persist with SoundNet-Flickr 144K, where broader sampling weakens incidental correlations but does not eliminate systematic contextual bias.

D.2 Visualization under Different Conditions

In Figures 9 and 10, we compare localization under single-source and multi-source settings. On S4, CCL recovers more complete guitar regions while suppressing background activation in the horse scene. On MS3, localization remains focused on the sounding dog and person despite nearby visual

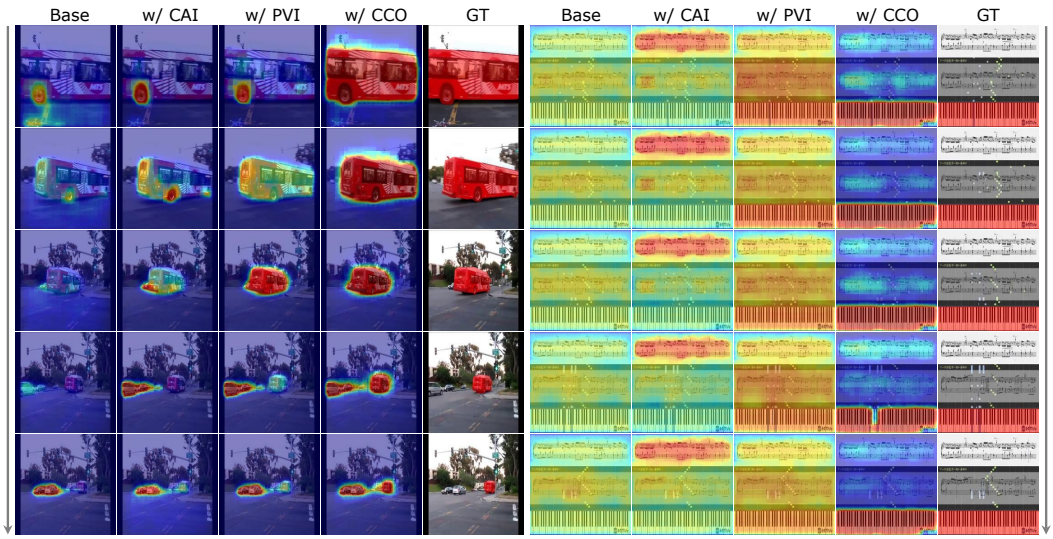

Figure 12: Progressive localization responses after adding the CCL components on AVSBench S4.

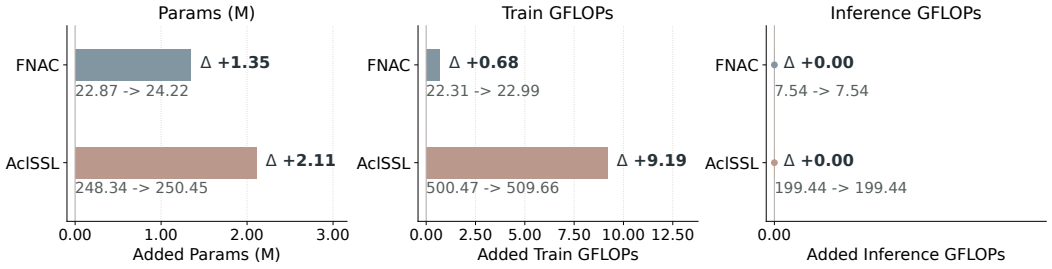

Figure 13: Efficiency across two model settings.

distractors. These results show that conditional adjustment improves both source completeness and separation under concurrent sounds.

E More Ablation Studies

E.1 Additional Component Visualizations

In Figure 11, we extend the representation analysis using EZ-VSL as the reference. Prototype matrices characterize inter-class overlap, while sample-to-prototype distributions measure within-class consistency. Clearer prototype separation and tighter class alignment are consistent with the improved spatial localization patterns. In Figure 12, we further present a video-level progressive ablation by incrementally introducing each component. For the rapidly moving bus, the model better captures temporal correspondence across frames. In the piano scene, where the keyboard frequently co-occurs with the musical staff, the activation progressively shifts from the correlated staff region toward the actual sounding keyboard.

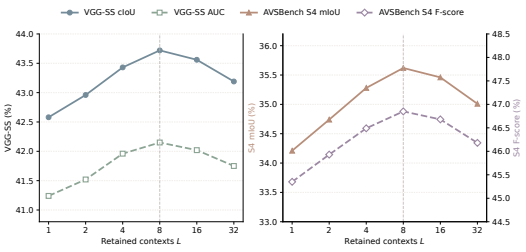

Figure 14: Sensitivity to the number of retained alternative contexts L.

Table 12: Component analysis of visual intervention.

Components				VGG-SS		AVS (S4)		AVS (MS3)	
Global Expectation	Audio Gate	Debiased Conditioning	Audio-Queried Context	cIoU ↑	AUC ↑	mIoU ↑	F-score ↑	mIoU ↑	F-score ↑
✗	✓	✓	✗	42.05	40.45	33.29	44.22	27.03	30.11
✓	✗	✓	✓	42.68	41.16	34.45	45.56	28.07	31.24
✓	✓	✗	✓	42.38	40.83	34.06	45.19	27.72	30.84
✓	✓	✓	✗	41.85	40.28	32.99	44.01	26.80	29.85
✓	✓	✓	✓	**43.72**	**42.15**	**35.62**	**46.85**	**29.18**	**32.74**

E.2 Alignment between Confounder Observability and Intervention Direction

In Table 6, pair reweighting and mediator regularization remain fixed while the modal intervention assignment changes.

Reversed intervention configuration. The reversed variant first applies back-door residualization to vision and then uses the corrected visual representation to condition audio front-door mediation. This order mirrors the dependency from CAI to PVI in CCL. For visual back-door adjustment, the audio query defines $\rho_k(\mathbf{A}) = \texttt{softmax}_k(\mathbf{f}_{\mathbf{A}}^\top \mathbf{z}_k^{\mathbf{A}}/\sqrt{d})$. These weights form $\boldsymbol{\rho}(\mathbf{A}) = [\rho_1(\mathbf{A}), \ldots, \rho_{K_Z}(\mathbf{A})]^\top \in \mathbb{R}^{K_Z}$. The conditional visual prior and corpus-wide visual prior are

$$\mu_{\mathbf{V}}(\mathbf{A}) = \sum_{k=1}^{K_Z} \rho_k(\mathbf{A}) \mathbf{z}_k^{\mathbf{V}}, \qquad \mu_{\mathbf{V}}^0 = \sum_{k=1}^{K_Z} p_k \mathbf{z}_k^{\mathbf{V}}. \tag{12}$$

The normalized entropy is $\mathcal{H}_{\mathbf{Z}}^{\mathbf{A}}(\mathbf{A}) = -\sum_{k=1}^{K_Z} \rho_k(\mathbf{A}) \log(\rho_k(\mathbf{A}) + \varepsilon_{\mathrm{H}})/\log K_Z$. The confidence $r_{\mathbf{V}}(\mathbf{A}) = [\texttt{clip}_{[0,1]}(1 - \mathcal{H}_{\mathbf{Z}}^{\mathbf{A}}(\mathbf{A}))]^{\kappa_{\mathbf{V}}}$ controls the removal magnitude, where $\kappa_{\mathbf{V}} > 0$. Let $\mathbf{1}_N \in \mathbb{R}^{N\times 1}$ be the all-ones column vector. Broadcasting the centered prior deviation over the spatial map yields

$$\mathbf{f}_{\mathbf{V}}^{\mathrm{BD}} = \texttt{LN}\big(\mathbf{f}_{\mathbf{V}} - \mathbf{1}_N r_{\mathbf{V}}(\mathbf{A})[\mu_{\mathbf{V}}(\mathbf{A}) - \mu_{\mathbf{V}}^0]^\top\big). \tag{13}$$

This operation matches the confidence calibration and prior centering used by CAI. The joint dictionary remains only a proxy for visual nuisance, so the correction does not make the original confounder $\mathbf{U}$ observable. For audio front-door adjustment, a learnable mediator bank $\mathcal{M}_{\mathbf{A}} = \{\mathbf{m}_l^{\mathbf{A}}\}_{l=1}^{K_M}$ discretizes acoustic patterns. After projection into the shared embedding space, global average pooling gives the corrected visual summary $\mathbf{f}_{\mathbf{V}}^{\mathrm{BD,g}} = N^{-1}(\mathbf{f}_{\mathbf{V}}^{\mathrm{BD}})^\top \mathbf{1}_N \in \mathbb{R}^d$. It supplies the gate $g_l^{\mathbf{V}}(\mathbf{A}, \mathbf{V}) = \sigma((\mathbf{m}_l^{\mathbf{A}})^\top \mathbf{f}_{\mathbf{V}}^{\mathrm{BD,g}}/\sqrt{d})$. Combining this gate with audio compatibility gives the mediator posterior $q_l^{\mathbf{A}}(\mathbf{A}, \mathbf{V}) = \texttt{softmax}_l((\mathbf{f}_{\mathbf{A}}^\top \mathbf{m}_l^{\mathbf{A}}/\sqrt{d} + \lambda_{\mathrm{g}}^{\mathbf{A}} \log(g_l^{\mathbf{V}}(\mathbf{A}, \mathbf{V}) + \varepsilon_{\mathrm{g}}))/\tau_{\mathrm{M}}^{\mathbf{A}})$, where $\lambda_{\mathrm{g}}^{\mathbf{A}} \geq 0$ controls the visual gate and $\tau_{\mathrm{M}}^{\mathbf{A}} > 0$ is the audio-mediator temperature. The posterior mean is

$$\widehat{\mathbf{f}}_{\mathbf{M_A}} = \sum_{l=1}^{K_M} q_l^{\mathbf{A}}(\mathbf{A}, \mathbf{V}) \mathbf{m}_l^{\mathbf{A}}. \tag{14}$$

K-means over global audio features constructs the alternative-audio dictionary $\mathcal{A}' = \{\mathbf{a}'_c\}_{c=1}^{C_{\mathbf{A}}}$, where $C_{\mathbf{A}}$ is the number of alternative-audio prototypes. The corrected visual summary retrieves $\delta_c(\mathbf{A}, \mathbf{V}) = \texttt{softmax}_c((\mathbf{a}'_c)^\top \mathbf{f}_{\mathbf{V}}^{\mathrm{BD,g}}/(\tau_{\mathbf{A}'}\sqrt{d}))$, where $\tau_{\mathbf{A}'} > 0$ is the retrieval temperature. Let $\mathcal{I}_L^{\mathbf{A}}(\mathbf{A}, \mathbf{V})$ contain the indices of the L largest weights, where $1 \leq L \leq C_{\mathbf{A}}$. For $c \in \mathcal{I}_L^{\mathbf{A}}(\mathbf{A}, \mathbf{V})$, the retained weight is $\overline{\delta}_c(\mathbf{A}, \mathbf{V}) = \delta_c(\mathbf{A}, \mathbf{V})/\sum_{a\in\mathcal{I}_L^{\mathbf{A}}(\mathbf{A},\mathbf{V})} \delta_a(\mathbf{A}, \mathbf{V})$. Each retained prototype is recombined as $\mathbf{h}_{\mathbf{A}}^{(c)} = \texttt{MLP}_{\mathbf{A}}([\widehat{\mathbf{f}}_{\mathbf{M_A}}, \mathbf{a}'_c])$. Let $\mathbf{f}_{\mathbf{V},n}^{\mathrm{BD}}$ denote the corrected visual feature at location n. Its context-specific score is $y_{\mathbf{A},n}^{(c)} = (\mathbf{h}_{\mathbf{A}}^{(c)})^\top \mathbf{f}_{\mathbf{V},n}^{\mathrm{BD}}/(\|\mathbf{h}_{\mathbf{A}}^{(c)}\|_2 \|\mathbf{f}_{\mathbf{V},n}^{\mathrm{BD}}\|_2 + \varepsilon_{\cos})$. Prediction-level marginalization then gives

$$\widehat{Y}_n^{\mathrm{AFD}} = \sum_{c\in\mathcal{I}_L^{\mathbf{A}}(\mathbf{A},\mathbf{V})} \overline{\delta}_c(\mathbf{A}, \mathbf{V}) y_{\mathbf{A},n}^{(c)}. \tag{15}$$

The reversed branch now matches PVI in posterior normalization, top-L context retrieval, and prediction-level averaging. The construction remains an architectural analogue rather than an identified audio front-door effect, since a learned acoustic mediator does not establish interception of every directed audio effect. The dual front-door configuration retains appropriate visual marginalization but compresses audio through an unnecessary mediator. The dual back-door configuration corrects the dominant acoustic context but leaves the latent visual path only approximately controlled. The reversed configuration contains both mismatches and performs below these partially aligned variants. We attribute the advantage of the proposed direction to using direct adjustment where a cross-modal proxy is available and mediation where visual nuisance remains latent.

E.3 Ablation of Visual Intervention

In Table 12, removing alternative-context marginalization leaves local mediator assignments tied to the current scene. Without audio gating, visually compatible but silent patterns can enter the mediator, whereas raw-audio conditioning may propagate the original co-occurrence prior downstream. Using a visual query similarly favors contexts close to the current image and can reintroduce nuisance variation. Corrected-audio conditioning avoids these effects by guiding both mediator selection and context retrieval before averaging separately evaluated predictions.

F Additional Experimental Analysis

F.1 Model Efficiency

In Figure 13, we compare the parameter count and per-pair computation before and after incorporating CCL. For FNAC, CCL introduces 1.35M additional parameters (22.87M to 24.22M) and 0.68 training GFLOPs (22.31 to 22.99), corresponding to increases of approximately 5.9% and 3.0%, respectively. For AclSSL, it adds 2.11M parameters (248.34M to 250.45M) and 9.19 training GFLOPs (500.47 to 509.66), corresponding to increases of approximately 0.85% and 1.8%, respectively. The larger absolute overhead in AclSSL is expected from its higher-resolution visual features, while the relative training cost remains limited in both settings. Since CCL is applied only during training, the inference pathways remain unchanged at 7.54 GFLOPs for FNAC and 199.44 GFLOPs for AclSSL. Offline dictionary construction is excluded from the reported per-pair computation.

F.2 Sensitivity to the Retained-Context Budget

In Figure 14, we examine the impact of retaining different numbers of alternative visual contexts for marginalization. A small L provides insufficient coverage of plausible scene variations, whereas an excessively large L introduces weakly related contexts and dilutes audio-supported predictions. Both localization and segmentation performance improve as L increases toward 8 and decline beyond this point. This consistent trend indicates that selective context coverage is preferable to either a narrow context subset or exhaustive averaging over alternatives.

F.3 Failure Cases and Limitations

Figure 15 illustrates two representative failure settings. In scenes with multiple people, visually similar human regions become competing candidates, leading to responses on non-sounding individuals. In soccer and basketball scenes, the soundtrack is strongly correlated with venue layout, player activity, and crowd presence. As a result, the response may spread across the field or court, including spectator regions and central playing areas, rather than isolate the actual source. These cases reveal remaining limitations in fine-grained instance selection and in disentangling localized sources from broad scene-level correlations.

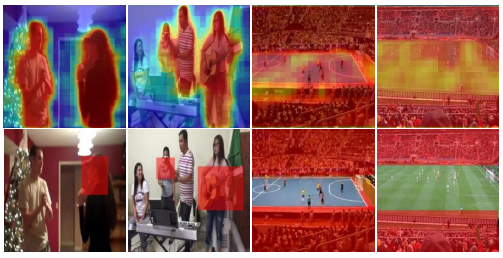

Figure 15: Failure cases in multi-person scenes and sports venues. Predictions may activate non-sounding people or spread across the playing field and spectator regions.

Fixed dictionaries may also miss fine-grained event variations within the same sports context, while a single normalized mediator mixture can conflate concurrent candidate sources. Future work may explore adaptive context updates, source-specific mediator mixtures, and temporal motion cues to better address these cases.

G Expanded Causal Derivations

This section provides self-contained probability derivations for the causal adjustments in CCL. We use discrete variables for clarity. For a continuous variable, the corresponding summation is replaced by an integral over its support.

G.1 Observational Marginalization

The observational predictor averages over the joint posterior of the proxied context $\mathbf{Z}$ and the unobserved visual factor $\mathbf{U}$ as follows:

$$P(\mathbf{Y} \mid \mathbf{A}, \mathbf{V}) = \sum_{\mathbf{Z}} \sum_{\mathbf{U}} P(\mathbf{Y} \mid \mathbf{A}, \mathbf{V}, \mathbf{Z}, \mathbf{U}) P(\mathbf{Z}, \mathbf{U} \mid \mathbf{A}, \mathbf{V}). \tag{16}$$

This expression retains the joint posterior $P(\mathbf{Z}, \mathbf{U} \mid \mathbf{A}, \mathbf{V})$ and introduces no independence assumption between the two confounders.

G.2 Conditional Back-Door Adjustment for Audio

Intervening on the audio variable removes incoming edges to $\mathbf{A}$. Applying the law of total probability over $\mathbf{Z}$ gives:

$$P(\mathbf{Y} \mid \mathtt{do}(\mathbf{A}), \mathbf{V}) = \sum_{\mathbf{Z}} P(\mathbf{Y} \mid \mathtt{do}(\mathbf{A}), \mathbf{V}, \mathbf{Z}) P(\mathbf{Z} \mid \mathtt{do}(\mathbf{A}), \mathbf{V}). \tag{17}$$

In the intervened graph, changing $\mathbf{A}$ does not change the conditional distribution of its former parent $\mathbf{Z}$. Hence,

$$P(\mathbf{Z} \mid \mathtt{do}(\mathbf{A}), \mathbf{V}) = P(\mathbf{Z} \mid \mathbf{V}). \tag{18}$$

Conditioning on $\mathbf{Z}$ and $\mathbf{V}$ blocks the back-door paths from $\mathbf{A}$ to $\mathbf{Y}$. Action and observation are consequently exchangeable in the first factor:

$$P(\mathbf{Y} \mid \mathtt{do}(\mathbf{A}), \mathbf{V}, \mathbf{Z}) = P(\mathbf{Y} \mid \mathbf{A}, \mathbf{V}, \mathbf{Z}). \tag{19}$$

Substitution into (17) yields:

$$P(\mathbf{Y} \mid \mathtt{do}(\mathbf{A}), \mathbf{V}) = \sum_{\mathbf{Z}} P(\mathbf{Y} \mid \mathbf{A}, \mathbf{V}, \mathbf{Z}) P(\mathbf{Z} \mid \mathbf{V}). \tag{20}$$

The dictionary weight $\alpha_k(\mathbf{V})$ approximates the discrete mass $P(\mathbf{Z} = \mathbf{z}_k \mid \mathbf{V})$ in (20). Exact adjustment with a learned proxy additionally requires the dictionary index to retain the variation in $\mathbf{Z}$ needed to block the audio back-door paths. An insufficient proxy leaves residual confounding even when retrieval is confident. CAI implements a conservative feature-level surrogate rather than the probability-level adjustment itself. The centered deviation $\mu_{\mathbf{A}}(\mathbf{V}) - \mu_{\mathbf{A}}^0$ removes only context-specific variation, while $r_{\mathbf{A}}(\mathbf{V})$ shrinks the correction when the proxy posterior is diffuse. Prior centering and confidence shrinkage improve robustness but do not constitute additional identification steps.

G.3 Conditional Front-Door Adjustment for Vision

Let $\mathbf{M}$ intercept the directed visual effect. We additionally assume that the path from $\mathbf{V}$ to $\mathbf{M}$ is unconfounded conditional on $\mathbf{A}$ and that $(\mathbf{V}, \mathbf{A})$ blocks the back-door paths from $\mathbf{M}$ to $\mathbf{Y}$. Expanding over the observed mediator under intervention gives:

$$P(\mathbf{Y} \mid \mathtt{do}(\mathbf{V}), \mathbf{A}) = \sum_{\mathbf{M}} P(\mathbf{M} \mid \mathtt{do}(\mathbf{V}), \mathbf{A}) P(\mathbf{Y} \mid \mathtt{do}(\mathbf{V}), \mathbf{M}, \mathbf{A}). \tag{21}$$

The conditional absence of confounding on $\mathbf{V} \to \mathbf{M}$ gives:

$$P(\mathbf{M} \mid \mathtt{do}(\mathbf{V}), \mathbf{A}) = P(\mathbf{M} \mid \mathbf{V}, \mathbf{A}). \tag{22}$$

Blocking the mediator back-door paths permits replacing the observation of $\mathbf{M}$ with $\mathtt{do}(\mathbf{M})$ while retaining $\mathtt{do}(\mathbf{V})$. Full mediation then removes the intervention on $\mathbf{V}$. These two steps yield:

$$P(\mathbf{Y} \mid \mathtt{do}(\mathbf{V}), \mathbf{M}, \mathbf{A}) = P(\mathbf{Y} \mid \mathtt{do}(\mathbf{M}), \mathbf{A}). \tag{23}$$

The remaining mediator effect is expanded over an alternative visual variable $\mathbf{V}'$:

$$P(\mathbf{Y} \mid \mathtt{do}(\mathbf{M}), \mathbf{A}) = \sum_{\mathbf{V}'} P(\mathbf{Y} \mid \mathtt{do}(\mathbf{M}), \mathbf{V}', \mathbf{A}) P(\mathbf{V}' \mid \mathtt{do}(\mathbf{M}), \mathbf{A}). \tag{24}$$

There is no directed edge from $\mathbf{M}$ to $\mathbf{V}$. Conditioning on $\mathbf{V}'$ and $\mathbf{A}$ also blocks the back-door paths from $\mathbf{M}$ to $\mathbf{Y}$. These relations give:

$$P(\mathbf{V}' \mid \mathtt{do}(\mathbf{M}), \mathbf{A}) = P(\mathbf{V}' \mid \mathbf{A}). \tag{25}$$

The outcome term admits the corresponding action and observation exchange once $\mathbf{V}'$ and $\mathbf{A}$ block the mediator back-door paths:

$$P(\mathbf{Y} \mid \mathtt{do}(\mathbf{M}), \mathbf{V}', \mathbf{A}) = P(\mathbf{Y} \mid \mathbf{M}, \mathbf{V}', \mathbf{A}). \tag{26}$$

Combining the preceding relations produces

$$P(\mathbf{Y} \mid \mathtt{do}(\mathbf{V}), \mathbf{A}) = \sum_{\mathbf{M}} P(\mathbf{M} \mid \mathbf{V}, \mathbf{A}) \sum_{\mathbf{V}'} P(\mathbf{Y} \mid \mathbf{M}, \mathbf{V}', \mathbf{A}) P(\mathbf{V}' \mid \mathbf{A}). \tag{27}$$

The normalized mediator posterior $q_{n,s}$ approximates $P(\mathbf{M} \mid \mathbf{V}, \mathbf{A})$ at each location. PVI uses the posterior mean $\widehat{\mathbf{f}}_{\mathbf{M},n}$ before nonlinear prediction, so the outer mediator expectation is a feature-level approximation. Moving this expectation before prediction would be exact for an outcome mapping affine in the mediator state. Under the current nonlinear predictor, the approximation is most accurate when the mediator posterior is concentrated. The visual dictionary and corrected-audio retrieval approximate $P(\mathbf{V}' \mid \mathbf{A})$. Top-$L$ truncation replaces this distribution with a renormalized finite-support approximation. PVI evaluates every retained context after the nonlinear predictor and averages the resulting scores, preserving prediction-level evaluation of the inner context expectation. The derivation identifies the visual effect only under the stated front-door assumptions; learning a prototype mediator encourages the required bottleneck but does not establish the graphical conditions from data alone.

G.4 Selection Reweighting for CCO

For contextual overlap $\omega \in [0, 1]$, let $\varphi_{\mathrm{s}}^{+}(\omega)$ and $\varphi_{\mathrm{s}}^{-}(\omega)$ denote the overlap densities induced by the positive and negative samplers, and let $\varphi_{\mathrm{t}}^{+}(\omega)$ and $\varphi_{\mathrm{t}}^{-}(\omega)$ denote target densities that assign greater mass to context-atypical positives and context-matched negatives. Under conditional exchangeability given the overlap and support inclusion, the corresponding importance ratios are:

$$w^{+}(\omega) = \frac{\varphi_{\mathrm{t}}^{+}(\omega)}{\varphi_{\mathrm{s}}^{+}(\omega)}, \qquad w^{-}(\omega) = \frac{\varphi_{\mathrm{t}}^{-}(\omega)}{\varphi_{\mathrm{s}}^{-}(\omega)}. \tag{28}$$

Weighting each conditional pair loss by the appropriate ratio converts the sampler-induced risk into the target risk. CCO does not estimate the four densities explicitly. The implemented functions $(\omega + \varepsilon_{\mathrm{w}})^{-1}$ for positives and ω for negatives preserve the desired ordering of the importance ratios, while batch-mean normalization controls scale. The resulting objective is an exact IPW estimator only when these functions are proportional to the true ratios. In the general case, CCO is an IPW-motivated contextual reweighting rule rather than an unbiased estimator of a fully specified target population.