UNLOCKING SPATIAL GROUNDING IN LARGE AUDIO-VISUAL RETRIEVAL MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Weak supervision sets a practical regime for audio-visual sound source localization as dense spatial annotations are costly to obtain at scale. The task, however, remains challenging, as models must locate sound sources from temporally aligned audio-visual data without pixel-level supervision. Recent large-scale audio-visual retrieval models, trained at unprecedented scale, encode rich multimodal structure. We show their latent representations, though optimized for global alignment, can nonetheless enable fine-grained spatial grounding. While spatial detail is progressively lost in the upper layers of retrieval backbones due to global pooling, intermediate visual tokens retain highly structured spatial information. To exploit this, we introduce LAIP (*Localization via Audio-Informed Pooling*), a framework that employs a lightweight *Audio-informed Spatial Pooling* (AiSP) to replace the standard global aggregation module. By querying intermediate visual tokens with audio aligned at the frame level, LAIP recovers localized spatial information that is otherwise discarded by the retrieval pipeline, with the largest gains observed for PE-AV, a stack with underlying temporal aggregation. Our approach achieves state-of-the-art performance on AVSBench and AVATAR, nearly doubling previous results on the latter, improving average CIoU from 13.21 to 26.22.

1 INTRODUCTION

Audio-visual sound source localization aims to identify, in each video frame, the spatial regions of active sound Mo & Morgado (2022b;a). As a fundamental problem in audio-visual understanding, it underpins applications in video editing, robotics, and multimodal retrieval. Yet, it is inherently challenging: in real-world scenes, visual content is often dense and cluttered, while sound sources are typically sparse. Consequently, models must pinpoint the regions responsible for the audio, such as *a ringing phone on a cluttered desk filled with silent objects*, or *a moving car honking in dense traffic*. This sparsity makes the task technically difficult, as it requires precise spatial disambiguation under weak supervision (only having access to the audio associated with the video), and dense spatial annotations are rarely available. Obtaining such annotations for videos is burdensome: annotators must delineate sounding regions frame by frame across long clips.

In parallel, large-scale contrastive pretraining has produced strong multimodal retrieval models across image–text (Radford et al., 2021; Zhai et al., 2023), image–any-modality (Girdhar et al., 2023), and audio–visual settings (Vyas et al., 2025) by aligning global embeddings of those modalities over large dataset. These models are appealing for weakly supervised localization because they learn rich semantic structure from large collections of paired data, potentially avoiding the need to train a localizer from scratch on small, curated datasets.

However, this promise comes with challenges illustrated in Figure 1. Retrieval models are optimized for global alignment—matching audio to an entire frame or video—rather than for local matching of the specific region that emits the sound. It results in frame ( ) and video features ( ) with poor spatiality. In principle, dense spatial annotations could recover locality through fine-tuning, but collecting them is costly at scale. As a result, much of the prior work in the weakly supervised regime relies on localization-specific models trained directly for the task (Sun et al., 2023; Mo & Morgado, 2022b; Choi et al., 2025). This leaves open a central question: *can large pretrained retrieval models be adapted to recover fine-grained spatial grounding from such weak supervision alone?*

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.

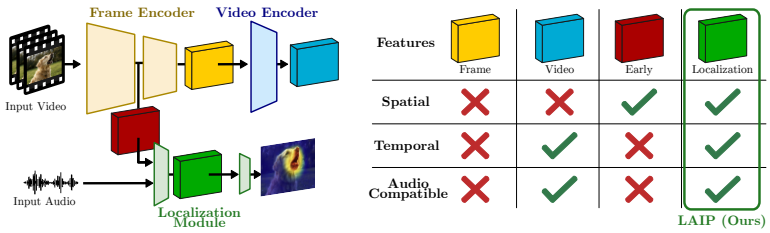


Figure 1: Global alignment suppresses spatial information in frame and video-level features. Earlier frame-encoder layers retain locality but are incompatible with the global multimodal space. LAIP learns locally grounded representations compatible with the audio space.

Our hypothesis is that they can. Although global pooling progressively suppresses spatial detail in the upper layers of retrieval backbones, intermediate visual tokens still retain structured local information. This suggests intervening before the final aggregation stage ( ) to recover audio–visual correspondence from within a pretrained retrieval model, rather than learning localization entirely from scratch.

To test this hypothesis, we instantiate it on PE-AV Vyas et al. (2025), a retrieval backbone composed of pretrained audio, frame, and video encoders. We propose Localization via Audio-Informed Pooling (LAIP), which inserts an *Audio-informed Spatial Pooling (AiSP)* mechanism between the frame and video encoders: audio embeddings query dense visual tokens before temporal aggregation, yielding a sound-conditioned visual representation while remaining compatible with the original retrieval pipeline.

To train under this constraint, we adopt a SigLIP-style contrastive objective inspired by FLAIR Xiao et al. (2025), which aligns the pooled video representation with its corresponding audio embedding while treating mismatched audio-conditioned views as negatives. This preserves cross-modal discrimination while remaining compatible with our pooled-token formulation. Although localization is not directly supervised, the attention maps induced by audio-informed spatial pooling provide strong spatial cues and lead to state-of-the-art sound source segmentation performance.

At inference time, retrieval can still be performed using the standard PE-AV forward pass, as our approach reuses the same audio and visual representations for localization. However, this design changes the supervision interface: after AiSP, each frame is represented by a single pooled token passed to the video encoder, whereas standard localization methods such as EZ-VSL and its extensions require dense spatial tokens for region-level matching Mo & Morgado (2022b); Choi et al. (2025).

Overall, our contributions are as follows:

- We show that large-scale audio–visual models pretrained for retrieval contain exploitable spatial information in intermediate visual tokens, making it possible to extend their capabilities to localization without dense spatial annotations.
- We introduce LAIP, a framework that employs a lightweight audio-informed architecture to inject temporally aligned audio into the visual backbone prior to video pooling, enabling stronger cross-modal synergy while preserving compatibility with pretrained retrieval architectures.
- We show that, with the right architecture and regularization, global audio–video supervision is sufficient to learn rich spatial audio–visual representations without spatial annotations. Our audio-conditioned training strategy yields a new state of the art on the reference AVSBench and AVATAR benchmarks, almost doubling the performance of previous models on the latter.

2 Related Work

Sound source localization Audio-visual sound source localization aims to identify the visual region responsible for a sound. Weakly supervised methods show that synchronized video–audio pairs already provide a useful grounding signal, while also exposing recurring failures such as attending to correlated but silent objects or mishandling off-screen sources Senocak et al. (2018); Mo & Morgado (2022b;a); Liu et al. (2022); Park et al. (2023); Sun et al. (2023); Senocak et al. (2023). The task has since expanded to segmentation-oriented and open-vocabulary settings Zhou et al. (2022); Mo & Tian (2023); Lee et al. (2025); Guo et al. (2024); Malard et al. (2026).

EZ-VSL Mo & Morgado (2022b) remains a key reference point, formulating localization as multiple-instance learning. Later works revisited protocols Mo & Morgado (2022a), reduced false negatives Sun et al. (2023), or strengthened semantic alignment Senocak et al. (2023). However, most methods remain frame-centric, aligning one audio representation with one image. This is restrictive for realistic videos, where sources move, appear off-screen, or switch across objects. AVATAR Choi et al. (2025) addresses this gap with a video-centric benchmark and TAVLO, which combines spatial and temporal attention under a temporal extension of the EZ-VSL objective.

Our work is complementary to these localization-specific architectures. Instead of preserving spatial outputs throughout the model, we start from a retrieval encoder that compresses each frame or clip into pooled tokens before alignment, and ask whether localization can be recovered from this interface. PE-AV Vyas et al. (2025) provides a concrete instance of this setting.

Large-scale learning of multimodal representations Recent multimodal representation learning follows a scaling paradigm in which large encoders are trained on massive paired datasets to align modalities in a shared embedding space. CLIP Radford et al. (2021) established this recipe for image–text learning, SigLIP Zhai et al. (2023) replaced the softmax contrastive loss with a pairwise sigmoid loss, ImageBind Girdhar et al. (2023) extended the shared-embedding perspective to additional modalities, and PE-AV Vyas et al. (2025) brought the same philosophy to joint audio, video, and text learning using roughly 100 million curated audio–video pairs. However, these models are optimized for global retrieval: they align modalities through pooled representations rather than dense spatial correspondences, and therefore do not directly expose the spatial outputs required for sound source localization.

FLAIR Xiao et al. (2025) addresses an analogous problem in image–text learning by using the text embedding to pool local visual tokens and recover text-conditioned image features from a globally trained model. Our setting is the audio-visual analogue: can audio pool local visual tokens so that a retrieval model becomes spatially informative?

3 METHOD

3.1 PROBLEM FORMULATION AND OVERVIEW

Audio-visual sound source localization (AV-SSL) involves identifying the spatial regions within a video frame x_t that correspond to the synchronized audio signal at time t. We formulate this task as the extraction of a per-frame spatial probability map over the visual tokens of x_t to isolate the sounding sources. We build on top of a large-scale audio-visual retrieval model composed of three transformer-based encoders: an audio encoder, an image frame encoder, and a video encoder. PE-AV is particularly relevant for our setup as it contains very rich pretrained audio-visual semantics, but it exhibits a clear architectural bottleneck: dense frame-level visual tokens are ultimately compressed into a single token before temporal modeling. Our goal is to show that this retrieval architecture can be adapted into an audio source localizer through a principled modification of that interface. Let $c_t^L \in \mathbb{R}^F$ denote the classification token [CLS] from the final layer L of the frame encoder, F being its dimension. The PE-AV architecture constructs a video-level representation by processing the sequence of frame-level [CLS] tokens $\{c_t^L\}_{t=1}^T$ through the video encoder. In parallel, the audio encoder produces a sequence of audio tokens $\{\tilde{a}_r\}_{r=1}^R$ at its native temporal resolution. Consequently, downstream temporal modeling operates exclusively on a single aggregated token per frame, rather than the underlying spatial grid. While this *[CLS]-based* design is effective for global retrieval tasks, it introduces an inherent information bottleneck for AV-SSL, as the video encoder is disconnected from the localized visual features necessary for precise spatial grounding.

Consistent with the observations in Bolya et al. (2025), we posit that spatial detail is progressively lost across the upper layers of the frame encoder, so that the final representation no longer retains the local information required for precise grounding. We therefore perform sound source localization from an intermediate layer ℓ of the frame encoder. Let $V_t^\ell \in \mathbb{R}^{HW \times F}$ denote the corresponding spatial tokens for frame t; in practice, we use the 16th layer, i.e., eight layers before the final 24th layer. Crucially, the audio and image-frame encoders operate at different temporal resolutions. We therefore transform the native audio tokens $\{\tilde{a}_r\}_{r=1}^R$ into frame-aligned representations using a lightweight local aggregator (a two-layer 1D convolutional network followed by temporal average pooling), which yields a synchronized audio embedding $a_t \in \mathbb{R}^F$ for each visual frame x_t. The proposed method, is illustrated in Figure 2.

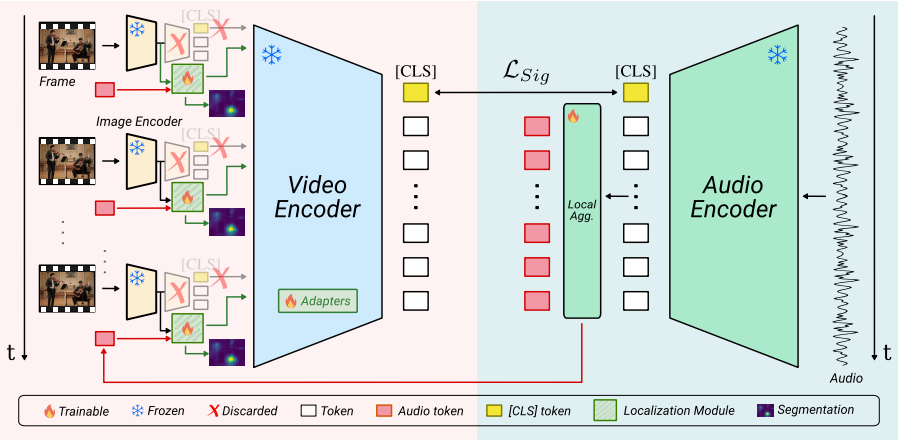


Figure 2: **LAIP: Localization via Audio-Informed Pooling**. The **audio tokens** are first forwarded to a **local aggregator** to match visual frame rate, and then used to pool the frame representation through our proposed **localization module**. The pooled representations are used as input to the **video encoder**, which outputs the **final representation** used to compute the sigmoid loss.

3.2 Localization Module

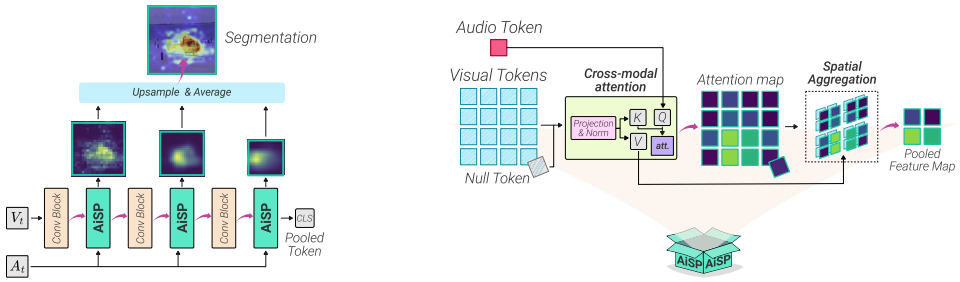
Rather than attaching a separate localization head on top of frozen PE-AV features, LAIP modifies the interface between the frame encoder and the video encoder, by adding a localization module (illustrated in Figure 3a). Its core building block is the Audio-informed Spatial Pooling (AiSP) modules, which progressively pool the intermediate visual tokens into an audio-informed [CLS] token $\tilde{c}_t^L$ while exposing attention maps for localization. This compatibility with the original PE-AV video encoder acts as a form of *implicit distillation*: the model is encouraged to extract localized spatial information while still producing per-frame embeddings that the pretrained temporal stack can consume.

A straightforward approach would be to apply a single cross-attention layer between the audio representation and the full set of spatial visual tokens. However, large-scale vision transformers are known to produce *register tokens*, i.e., outlier tokens with disproportionately high norms and overly globalized semantics Darcet et al. (2024), which can dominate cross-attention weights and degrade spatial precision. To circumvent this, LAIP adopts a hierarchical design inspired by the classic convolution-pooling paradigm LeCun et al. (2002): we stack three AiSP modules, each followed by a lightweight convolutional adaptation block, and progressively reduce the spatial resolution of the visual tokens until they collapse into the replacement token $\tilde{c}_t^L$.

This design keeps the pooling process spatially grounded, as the audio signal selectively attends to tokens corresponding to active sounding regions rather than to registers or tokens that have already aggregated global context. It also yields multi-scale attention maps across the successive pooling stages. At inference time, we combine these maps to obtain the final localization.

Audio-informed Spatial Pooling (AiSP) This pooling layer (depicted in Figure 3b) is designed with two objectives: (i) reducing the spatial resolution of the visual feature maps, and (ii) suppressing regions corresponding to non-sounding objects. To this end, we first estimate an attention map that localizes the sounding object at the current resolution, and then perform pooling conditioned on this localization to refine the features passed to subsequent layers. Since the model is trained through a cross-modal global objective, the underlying intuition is that improved localization of sounding objects yields more informative visual features for the temporal transformer. Consequently, the architecture is implicitly encouraged to produce accurate localization, leading to better alignment between visual and audio representations.

Concretely, the AiSP operator down-samples the spatial token map by a factor of K along each spatial dimension. Using the notation from Section 3.1, we consider the frame-level visual tokens as an $H \times W$ grid, denoted by $V_t \in \mathbb{R}^{H \times W \times F}$ for readability, and query them with the temporally aligned audio token a_t.



(a) Localization module.

(b) AiSP module.

Figure 3: Overview of our localization and AiSP modules, with $K = 2$. Starting from PE visual tokens extracted from the image branch at the 16-th frame-encoder layer, a frame-aligned audio token progressively pools the spatial grid into one sound-conditioned visual token per frame.

We then flatten V_t into a sequence of HW spatial tokens. Ignoring for now the optional null token introduced below, keys and values are obtained by linear projection and normalization, yielding $\{k_i, v_i\}_{i=1}^{HW}$, while the query is obtained from the linearly projected and normalized audio token a_t as $q_a \in \mathbb{R}^F$. We then compute an attention matrix as

$$\alpha_i = \frac{\exp\left(\frac{q_a^\top k_i}{\sqrt{F}}\right)}{\sum_{j=1}^{HW} \exp\left(\frac{q_a^\top k_j}{\sqrt{F}}\right)}, \quad i \in \{1, \ldots, HW\},$$

such that $\sum_{i=1}^{HW} \alpha_i = 1$. Furthermore, as α reflects the image regions attended to by the audio modality, we employ it as a segmentation map.

The spatial attention weights $\{\alpha_{h,w}\}$ are then reshaped into an attention map $M \in \mathbb{R}^{H \times W}$, preserving the two-dimensional structure of the frame. For each non-overlapping $K \times K$ spatial block $\mathcal{B}_{u,v}$, we define the pooled feature as:

$$Y_{u,v} = \sum_{(h,w) \in \mathcal{B}_{u,v}} \alpha_{h,w}\, v_{h,w},$$

where $v_{h,w}$ denotes the projected and normalized visual token at location (h, w). This yields a pooled feature map $Y \in \mathbb{R}^{\frac{H}{K} \times \frac{W}{K} \times F}$. Unlike locally normalized attention pooling, this formulation preserves globally normalized cross-attention scores while restricting the aggregation to local spatial neighborhoods (but does not preserve the mass in each spatial block, as some of them can contain no audio-related information at all). Overall, this defines $P_K(V_t, a_t) : \mathbb{R}^{H \times W \times F} \to \mathbb{R}^{\frac{H}{K} \times \frac{W}{K} \times F}$.

Inspired by FLAIR, we propose appending a learnable token, $x_\emptyset$, to the key tokens, before the pooling. It gives the audio query an explicit fallback option, allowing it to attend away from the image tokens when none of them is relevant. This is particularly useful in the audio-visual setting, where the audio content may differ substantially from the visible scene.

3.3 Training and Inference with Global Supervision

Let $O^{(i,j)} \in \mathbb{R}^F$ denote the output of the video encoder when the visual tokens of video i are aggregated by LAIP using the temporal tokens of audio j as queries. Following Xiao et al. (2025), we train all B^2 audio–video combinations with a sigmoid-based contrastive objective:

$$\mathcal{L}_{\text{SIG}} = -\frac{1}{B^2} \sum_{i=1}^{B} \sum_{j=1}^{B} \log \sigma\Big(y_{ij} \Big[\tau \left\langle O^{(i,j)}, A_g^{(j)} \right\rangle + b\Big]\Big). \tag{1}$$

Here, $y_{ij} = +1$ when $i = j$ and $y_{ij} = -1$ otherwise. Moreover, σ denotes the sigmoid function, τ is a learnable logit scale, b is a learnable bias, and $\langle \cdot, \cdot \rangle$ is the inner product. The visual pooling that produces $O^{(i,j)}$ is conditioned by the frame-aligned tokens $\{a_t^{(j)}\}_{t=1}^{T}$, while $O^{(i,j)}$ is compared

with the global representation $A_g^{(j)}$ produced by the same audio encoder. The pair is positive only when audio j is associated with video i.

This objective preserves global discriminative power while specifically regularizing the sound-conditioned pooling mechanism that generates $O^{(i,j)}$. As in FLAIR, a query from one modality first conditions the visual stream before global contrastive matching; here the conditioning signal is audio rather than text, which distinguishes our loss from the standard SigLIP objective Zhai et al. (2023) used in PE-AV.

While $\mathcal{L}_{\text{SIG}}$ supervises the final sound-conditioned video representation, it does not directly constrain the intermediate attention maps produced by the three-stage AiSP hierarchy. We therefore enforce consistency across resolutions. For frame t, the first two single-head stages produce maps $M_{1,t}$ and $M_{2,t}$, while the final, strongest-reduction stage produces 8 maps $M_{3,t}^h$. The per-frame loss is:

$$\mathcal{L}_{\text{MRes}}^t = \|M_{1,t} - \text{Up}(M_{2,t})\|_F + \min_h \left\|M_{2,t} - \text{Up}(M_{3,t}^h)\right\|_F, \tag{2}$$

where Up is up-sampling and $\|\cdot\|_F$ is the Frobenius norm. The training loss $\mathcal{L}_{\text{MRes}}$ averages $\mathcal{L}_{\text{MRes}}^t$ over the frames and batch. At inference, the same multi-resolution criterion selects the final-stage head for each frame:

$$h_t^\star = \arg\min_h \left\|M_{2,t} - \text{Up}(M_{3,t}^h)\right\|_F, \tag{3}$$

then produce the final heatmap for frame t by averaging $M_{1,t}$, $\text{Up}(M_{2,t})$, and $\text{Up}(M_{3,t}^{h_t^\star})$ at a common resolution.

During preliminary experiments, we observed that the model tended to allocate excessive attention to the learnable null token $x_\emptyset$, particularly for negative audio-video pairs (more details in Appendix C). Because the null token participates in the attention normalization but is excluded from the pooled output, high attention to the null token reduces the total weight assigned to visual tokens. Accumulated across successive AiSP stages, this effect can drive the conditioned representation $O^{(i,j)}$ toward zero and degrade its quality. We therefore directly penalize null attention. Let $\alpha_{\emptyset,s}^{i,j,t}$ denote the attention assigned to $x_\emptyset$ at stage s and frame t for video i queried with audio j. We penalize its accumulation across all negative pairs, stages, and frames as follows. $\mathcal{L}_\emptyset = \frac{1}{ST}\sum_{\substack{i\neq j\\ s,t}} \alpha_{\emptyset,s}^{i,j,t}$, and incorporate it in the final objective:

$$\mathcal{L} = \mathcal{L}_{\text{SIG}} + \lambda\mathcal{L}_{\text{MRes}} + \mu\mathcal{L}_\emptyset. \tag{4}$$

This anti-collapse regularizer forces the negative branch to retain visual evidence, with μ and λ controlling the strength of the null-attention and multi-resolution regularizers, respectively. Finally, small adapters (Houlsby et al., 2019) bridge the difference between the pretrained frame token c_t^L and our audio-informed $\tilde{c}_t^L$; all other pretrained weights remain frozen. LAIP preserves PE-AV's original retrieval forward pass: for retrieval, it bypasses the local audio aggregator, AiSP modules, and adapters, and feeds the original final-layer tokens $\{c_t^L\}_{t=1}^T$ to the unchanged frozen video encoder. Because this path has exactly the same operations and weights as PE-AV, its retrieval embeddings and scores are identical. The audio-conditioned path is used only for localization.

4 Experiments

4.1 Datasets and implementation details

We evaluate LAIP on three benchmarks. Our primary benchmark is AVATAR (Choi et al., 2025), a video-centric audio-visual localization dataset designed to capture temporal dynamics beyond frame-centric AVL settings. AVATAR contains 5,000 videos with an average duration of 10 seconds and 24,266 annotated frames across 80 target categories, with instance-segmentation annotations covering four scenarios: single-sound, mixed-sound, multi-entity, and off-screen.

We also evaluate on AVSBench (Zhou et al., 2022), a benchmark for sound-prompted audio-visual segmentation and localization that provides pixel-level masks for sounding objects. We report results on its single-source (S4) and multi-source (MS3) test sets. Finally, we also evaluate our method on ADE-SP, the sound-prompted ADE20k benchmark introduced in (Hamilton et al., 2024), which pairs ADE20k images and segmentation masks with audio from VGGSound. Evaluation protocol

and metrics are further discussed in Appendix B. Following the experimental protocol of (Choi et al., 2025), we train on the same 10,000 high-frame-rate videos selected from VGGSound (Chen et al., 2020); using this subset keeps our setup directly comparable to prior results on AVATAR. We instantiate the multi-resolution AiSP module with pooling factors $K = \{2, 2, 6\}$. Each adaptation block uses convolutions with kernel size 3, for a total of $\sim 35M$ added parameters (out of $1.7B$ in PE-AV). We train for 10 epochs with a learning rate of $1e{-}4$ and a batch size of 10. For the main experiments, we set $\mu = 0.01$ and $\lambda = 100$ to obtain losses of a similar order of magnitude. We use adapters of size 512, applied only in the MLP blocks, for a total of $7M$ parameters. For computational efficiency, we use the PE-AV-large variant that processes 16 frames per video instead of the version that uses all frames. We report incremental compute and memory costs of LAIP in Appendix F.

4.2 Quantitative results

AVATAR Benchmark Table 1 reports results on all four AVATAR scenarios. LAIP substantially outperforms static sound-source localization baselines that rely on a single frame, as well as TAVLO, the strongest temporal baseline in Table 1. Notably, despite TAVLO operating at a substantially higher frame rate, LAIP remains clearly superior on the *multi-entity* scenario, which demands the most temporal reasoning. To isolate LAIP from the strength of PE-AV pretraining, we reimplement both EZ-VSL, FNAC and TAVLO on the same frozen PE-AV intermediate features, with comparable trainable capacity and optimization budget. Both remain far below LAIP, indicating that access to the same strong pretrained representation is insufficient to explain our gains. Further details on these experiments are provided in Appendix D. LAIP obtains a lower TN than the best baselines in the *off-screen* scenario. As noted in Choi et al. (2025), however, the TN metric used for this setting suffers from a known bias. Specifically, the threshold underlying the TN metric is computed over predictions from the entire test set across all evaluation settings, by selecting the top 10% of heatmap activations. Consequently, methods with high activations in the other three settings may result in a threshold that artificially reduces the measured false-positive rate on off-screen examples. Overall, these results establish a new state-of-the-art range on AVATAR. We attribute this improvement to LAIP's ability to leverage PE-AV pretraining, which equips the model with strong audio-visual semantic representations prior to localization-specific pattern extraction. Applying LAIP to intermediate ImageBind or LanguageBind features also consistently outperforms TAVLO across the three localization scenarios, demonstrating that the proposed pooling mechanism is not specific to PE-AV. PE-AV outperforms all other retrieval models largely because it includes a pretrained temporal encoder, whereas ImageBind and LanguageBind require an additional video encoder trained from scratch, as they lack temporal modeling. However, when the PE-AV temporal encoder is likewise trained from scratch, its performance drops to a similar level as ImageBind and LanguageBind (see Appendix B). The advantage of intermediate representations further holds for ImageBind, where LAIP substantially outperforms the same model applied to final-layer features. Together with the PE-AV layer-wise analysis (see Appendix D), this suggests that the spatial–semantic trade-off exploited by LAIP generalizes across retrieval backbones. Complementary results and additional implementation details are provided in Appendix D. These findings highlight the utility of LAIP to unlock fine-grained spatial grounding from pretrained retrieval representations. Notably, although PE-AV was not designed for spatial localization, LAIP transforms its learned audio-visual semantics into spatially grounded predictions, substantially extending the capabilities of the original representation.

AVSBench and ADE20k Tables 2 and 3 show that the same model also transfers well to other sound source localization benchmarks, even though LAIP is only trained on a very small subset of VGGSound ($\sim$10k videos). On AVSBench, LAIP achieves the best F-score on both S4 and MS3 by a large margin, and also obtains the best mask-IoU on the more challenging MS3 split. The gap between F-score and mask-IoU is informative. In AVSBench, the F-score is computed with an adaptive threshold, whereas mask-IoU uses a fixed threshold of 0.5. Because our method produces attention maps rather than dense segmentation logits, its predictions are naturally sharp and peaky: this structure is then penalized by having an arbitrarly fixed threshold at 0.5. This is also why LAIP is especially strong in F-score while remaining only competitive on S4 mask-IoU. For all AVSBench results, we apply min–max normalization to rescale each predicted map to $[0, 1]$ before evaluation.

On ADE-SP, LAIP outperforms all baselines in both m-IoU and mAP, reaching 33.35 m-IoU and 53.57 mAP. These gains are particularly notable because several competing methods are explicitly designed for sound-prompted segmentation, whereas our model is primarily built to recover localization from

Table 1: Segmentation performance comparison across AVATAR scenarios.

Method	(1) Single-sound		(2) Mixed-sound		(3) Multi-entity		(4) Off-screen
	CIoU(%) ↑	AUC(%) ↑	CIoU(%) ↑	AUC(%) ↑	CIoU(%) ↑	AUC(%) ↑	TN(%) ↑
EZ-VSL(10k) (Mo & Morgado, 2022b)	9.66	11.07	8.16	9.35	6.87	8.32	**96.91**
EZ-VSL(144k) (Mo & Morgado, 2022b)	10.92	12.22	6.97	8.34	5.80	7.42	96.47
FNAC (144k) (Sun et al., 2023)	11.17	13.21	5.87	7.73	3.77	5.79	92.84
FNAC (10k) (Sun et al., 2023)	13.00	14.19	10.99	12.08	10.31	11.45	87.52
TAVLO(10k) (Choi et al., 2025)	13.42	14.08	14.13	14.52	12.08	12.69	91.18
EZ-VSL on PE features (10k) (Mo & Morgado, 2022b)	13.08	13.97	12.12	12.78	11.10	11.91	92.62
FNAC on PE features (10k) (Sun et al., 2023)	13.55	13.97	10.98	11.25	8.89	9.42	93.47
TAVLO on PE features (10k) (Choi et al., 2025)	7.51	8.01	7.50	7.75	5.78	6.18	91.33
LAIP on ImageBind (10k) (Girdhar et al., 2023)	18.29	18.56	16.14	16.30	15.60	16.02	90.05
LAIP on LanguageBind (10k) (Zhu et al., 2024)	17.64	18.08	15.73	15.91	14.22	14.74	91.00
LAIP (Ours, 10k)	**27.63**	**27.77**	**27.35**	**27.40**	**23.69**	**23.85**	90.94

Table 2: Quantitative results on the AVSBench test sets.

Method	S4		MS3	
	mask-IoU ↑	F-score ↑	mask-IoU ↑	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., 2023)	29.60	35.90	-	-
TACO (Malard et al., 2026)	29.68	41.91	25.88	30.72
LAIP (Ours)	**39.31**	**65.18**	**30.77**	**49.00**

a retrieval-trained backbone. We believe this transfer performance stems from the same property observed on AVATAR: the pretrained PE-AV features provide strong semantic alignment, while AiSP extracts spatial evidence without destroying the retrieval representation.

Interestingly, our method also outperforms methods like TACO (reported without the segmenter for fair comparison), and DenseAV, which also rely on pretrained models and explicitly target sound source localization, while leaving retrieval aside. This highlights the effectiveness of our method and suggests that using an aligned audio-visual backbone is better than using separate audio and image backbones.

Table 3: Quantitative results on the ADE-SP dataset.

Method	Retrieval model	m-IoU↑	mAP↑
DAVENet (Hsu et al., 2019)	✗	17.0	16.8
DenseAV (Hamilton et al., 2024)	✗	25.5	32.4
TACO (Malard et al., 2026)	✗	27.74	35.75
ImageBind (Girdhar et al., 2023)	✓	18.3	18.1
CAVMAE (Gong et al., 2023)	✓	20.6	21.2
CAV-MAE Sync (Araujo et al., 2025)	✓	22.7	22.6
LAIP (Ours)	✓	**33.35**	**53.57**

Finally, CAV-MAE Sync (Araujo et al., 2025) is the closest work to ours in terms of applications, since it also aims to support both sound source localization and retrieval. However, it was trained from scratch on a limited amount of data compared to PE-AV, while our method preserves compatibility with the pretrained stack and therefore benefits from higher-quality features. This difference is reflected in the results: LAIP substantially outperforms CAV-MAE Sync on ADE-SP.

Ablations Table 4 shows the performance of different ablation on the AVATAR benchmark (the average scores over the scenarios are reported, but the full Tables are available in Appendix). Removing the null-token regularization ($\mu = 0$, while retaining the null token) degrades performance from $26.22/26.34$ to $23.01/23.27$ CIoU/AUC. Removing the null token entirely gives a similar $23.11/23.25$, showing that the token alone provides no benefit without the anti-collapse regularizer. Removing both μ and the multi-resolution consistency term λ leads to a larger drop to $19.45/19.53$. These results show that preventing the attention to null token to collapse and enforcing cross-scale consistency are both important. Additional sensitivity experiments on the loss weights, number of heads, adapter size, and pooling schedule are reported in Appendix E. Additionally, replacing the CNN layers with transformer encoder layers in the AiSP modules also hurts performance ($21.61/21.75$), suggesting that local convolutions are better suited to progressive spatial refinement and alignment (as observed in Yu et al. (2023)). Table 4 (right) further shows that the gains cannot

Table 4: AVATAR ablations (averaged over scenarios): regularization (left) and baselines (right).

Method	CIoU(%) ↑	AUC(%) ↑
LAIP (Ours)	**26.22**	**26.34**
$\mu = 0$	23.01	23.27
No null token	23.11	23.25
$\mu = \lambda = 0$	19.45	19.53
$\lambda = 0$	17.52	17.57

Method	CIoU(%) ↑	AUC(%) ↑
LAIP (Ours)	**26.22**	**26.34**
Gradient-based	4.14	4.89
Attention pooling	14.34	14.74
LAIP on last layer	15.92	16.38
Transformer LAIP	21.61	21.75

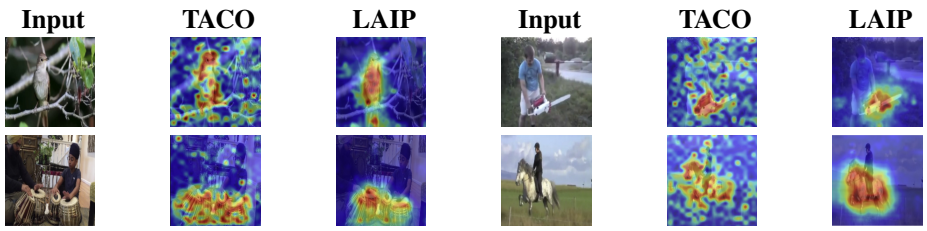


Figure 4: TACO–LAIP comparison on S4. LAIP produces smoother maps with fewer outliers.

be recovered through simpler ways of extracting localization from PE-AV. Gradient attribution based on visual-token norms performs poorly (4.14/4.89). Using a single attention-pooling layer after the same CNN blocks is also substantially weaker (14.34/14.74 versus 26.22/26.34), showing that cross-attention alone is insufficient. Finally, applying LAIP to the final PE-AV layer degrades performance to 15.92/16.38; the complete layer-wise analysis in Appendix C confirms a clear intermediate-layer optimum. These experiments validate empirically that LAIP's performance stems from progressively extracting audio-conditioned spatial information from the appropriate intermediate representation, not merely from layer selection, cross-attention, or strong PE-AV features.

4.3 Qualitative analysis

Figure 4 compares the non-thresholded localization maps produced by TACOMalard et al. (2026) (the best competitor both in Table 2 and Table 3) and LAIP. This comparison is particularly informative because both methods build on large-scale pretrained models—OpenCLIPCherti et al. (2023) and CLAPElizalde et al. (2024) for TACO, and PE-AVVyas et al. (2025) for LAIP—but they exploit them in different ways. TACO discovers correspondences between independently pretrained backbones through inference-time optimization, whereas LAIP learns an explicit localizer module during training. Although both methods usually identify the correct sound source, the maps produced by TACO contain many outliers, likely because of irregularities in the image-token representation. In contrast, LAIP produces smoother and more spatially coherent maps thanks to its multi-resolution pooling mechanism.

5 Conclusion and limitations

We adapt a large-scale audio–visual retrieval model for sound source localization without dense spatial supervision. LAIP uses lightweight Audio-informed Spatial Pooling (AiSP) modules to recover spatial information from intermediate visual representations while preserving compatibility with the pretrained retrieval pipeline. Applied to PE-AV, this yields strong results across AVATAR, AVSBench, and ADE-SP, while experiments with ImageBind and LanguageBind further show that the proposed pooling mechanism is not specific to a single backbone. Overall, our results suggest that global retrieval representations need not be discarded for localization: spatial information can be recovered before global aggregation and exploited within the pretrained architecture.

LAIP's main limitation is retrieval scalability: audio-conditioned pooling prevents precomputed video embeddings. Standard retrieval remains available through the unchanged PE-AV backbone; future work should enable query-independent LAIP representations.

AI USE STATEMENT

Generative AI tools were used to assist with adapting the manuscript to the ICLR LaTeX format, help refining the written text, as well as implementing or adapting some baselines. The authors reviewed the AI-assisted changes and take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

The method architecture and objectives are specified in Sections 3.1 and 3.3. Dataset protocols, implementation details, evaluation metrics, and expanded ablations are provided in the experiments and appendix. Code will also be release upon acceptance.

REFERENCES

Edson Araujo, Andrew Rouditchenko, Yuan Gong, Saurabhchand Bhati, Samuel Thomas, Brian Kingsbury, Leonid Karlinsky, Rogerio Feris, James R Glass, and Hilde Kuehne. Cav-mae sync: Improving contrastive audio-visual mask autoencoders via fine-grained alignment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025.

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Abdul Rasheed, et al. Perception encoder: The best visual embeddings are not at the output of the network. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

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.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023.

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

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International Conference on Learning Representations (ICLR)*, 2024.

Benjamin Elizalde, Soham Deshmukh, and Huaming Wang. Natural language supervision for general-purpose audio representations. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023.

Yuan Gong, Andrew Rouditchenko, Alexander H Liu, David Harwath, Leonid Karlinsky, Hilde Kuehne, and James R Glass. Contrastive audio-visual masked autoencoder. In *The Eleventh International Conference on Learning Representations*, 2023.

Ruohao Guo, Liao Qu, Dantong Niu, Yanyu Qi, Wenzhen Yue, Ji Shi, Bowei Xing, and Xianghua Ying. Open-vocabulary audio-visual semantic segmentation. In *Proceedings of the 32nd ACM International conference on multimedia*, 2024.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pp. 2790–2799. PMLR, 2019.

Wei-Ning Hsu, David Harwath, and James Glass. Transfer learning from audio-visual grounding to speech recognition. In *Proc. Interspeech 2019*, 2019.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 2002.

Kyungbok Lee, You Zhang, and Zhiyao Duan. Audio visual segmentation through text embeddings. In *2025 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2025.

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, 2022.

Hugo Malard, Michel Olvera, Stéphane Lathuilière, and Slim Essid. Taco: Training-free sound prompted segmentation via semantically constrained audio-visual co-factorization. *Transactions on Machine Learning Research*, 2026.

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

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

Shentong Mo and Yapeng Tian. Av-sam: Segment anything model meets audio-visual localization and segmentation. *arXiv preprint arXiv:2305.01836*, 2023.

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)*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PmLR, 2021.

Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018.

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 *Proceedings of the IEEE/CVF international conference on computer vision*, 2023.

Weixuan Sun, Jiayi Zhang, Jianyuan Wang, Zheyuan Liu, Yiran Zhong, Tianpeng Feng, Yandong Guo, Yanhao Zhang, and Nick Barnes. Learning audio-visual source localization via false negative aware contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023.

Apoorv Vyas, Heng-Jui Chang, Cheng-Fu Yang, Po-Yao Huang, Luya Gao, Julius Richter, Sanyuan Chen, Matt Le, Piotr Dollár, Christoph Feichtenhofer, et al. Pushing the frontier of audiovisual perception with large-scale multimodal correspondence learning. *arXiv preprint arXiv:2512.19687*, 2025.

Rui Xiao, Sanghwan Kim, Mariana-Iuliana Georgescu, Zeynep Akata, and Stephan Alaniz. Flair: Vlm with fine-grained language-informed image representations. In *Proceedings of the computer vision and pattern recognition conference*, 2025.

Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. *Advances in Neural Information Processing Systems*, 36, 2023.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2023.

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*, 2022.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, WANG HongFa, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, et al. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. In *International Conference on Learning Representations*, volume 2024, pp. 9588–9608, 2024.