AV-SSAN: Audio-Visual Selective DOA Estimation Through Explicit Multi-Band Semantic-Spatial Alignment

Yu Chen[1,5], Hongxu Zhu[2], Jiadong Wang[3], Kainan Chen[4], Xinyuan Qian*[1]
[1]University of Science and Technology Beijing (USTB), China
[2]Fano, Hong Kong
[3]Technical University of Munich, Germany
[4]Eigenspace GmbH, Germany
[5]School of Data Science, The Chinese University of Hong Kong (Shenzhen), China

Abstract

Audio-visual sound source localization (AV-SSL) estimates the position of sound sources by fusing auditory and visual cues. Current AV-SSL methodologies typically require spatially-paired audio-visual data and cannot selectively localize specific target sources. To address these limitations, we introduce Cross-Instance Audio-Visual Localization (CI-AVL), a novel task that localizes target sound sources using visual prompts from different instances of the same semantic class. CI-AVL enables selective localization without spatially paired data. To solve this task, we propose AV-SSAN, a semantic-spatial alignment framework centered on a Multi-Band Semantic-Spatial Alignment Network (MB-SSA Net). MB-SSA Net decomposes the audio spectrogram into multiple frequency bands, aligns each band with semantic visual prompts, and refines spatial cues to estimate the direction-of-arrival (DoA). To facilitate this research, we construct VGGSound-SSL, a large-scale dataset comprising 13,981 spatial audio clips across 296 categories, each paired with visual prompts. AV-SSAN achieves a mean absolute error of 16.59° and an accuracy of 71.29%, significantly outperforming existing AV-SSL methods.

Introduction

Sound Source Localization (SSL) estimates the DoA of sound sources from multichannel audio. Classical methods, such as GCC-PHAT (Knapp and Carter 1976), MUSIC (Schmidt 1986), and SRP-PHAT (DiBiase, Silverman, and Brandstein 2001), rely on spatial spectrum estimation or time-delay analysis. While effective in controlled settings, they degrade severely in the presence of multiple sources, high background noise, or strong reverberation.

To overcome these limitations, recent works leverage deep neural networks (DNNs) to learn spatial representations either to enhance classical methods (He, Motlicek, and Odobez 2018; Shmuel et al. 2023; Pertilä and Cakir 2017) or in an end-to-end form (Chakrabarty and Habets 2017; Li, Zhang, and Li 2018; Xiao and Das 2024). However, these audio-only models are fragile in acoustically challenging scenes or when the target sound source is inactive.

Visual modalities offer a complementary cue to SSL. Recent AV-SSL frameworks (Qian et al. 2021b,a, 2022b,a; Jiang et al. 2023; Berghi et al. 2024; Zhao et al. 2023; Wu et al. 2023) integrate spatial audio and visual context to improve localization performance. However, these methods have two major limitations: (1) They require tightly spatially-aligned audio-visual inputs where the visible object directly corresponds to the sounding source, a condition rarely met in real-world data. (2) They localize all active sources but lack the ability to selectively localize a specific target source of interest.

To address these limitations, we propose a new task: **Cross-Instance Audio-Visual Localization (CI-AVL)**. CI-AVL aims to localize a target source using a visual prompt derived from a different instance of the same semantic class (e.g., localizing a barking dog with an image of another dog). This setting enables selective localization without requiring explicitly paired audio-visual data. However, CI-AVL presents a unique challenge: the visual prompt is only semantically associated with the sound source but spatially unrelated, making direct fusion strategies employed by prior AV-SSL methods ineffective.

This challenge stems from a fundamental semantic-spatial misalignment: existing methods primarily focus on spatial alignment ("where"), while overlooking semantic alignment ("what") between audio and visual modalities. In contrast, human perception follows a hierarchical process: we first semantically recognize the sound source ("what"), then localize its position ("where") (Taevs et al. 2010; van der Heijden et al. 2019). Inspired by this, we hypothesize that effective target-aware localization requires a two-stage alignment mechanism: (1) aligning the visual prompt and audio semantically to isolate the target's identity, and (2) localizing the target source conditioned on the identity.

To this end, we introduce AV-SSAN, which first semantically aligns a cross-instance visual prompt with mixed audio and then spatially localizes the target source. Inspired by the frequency-dependent characteristics of spatial hearing, AV-SSAN incorporates an MB-SSA Net. It decomposes the spectrogram into different frequency resolutions, semantically aligns each band with the visual prompt, and then fuses them via an attention-guided refiner to yield DoA estimates.

In addition, we introduce VGGSound-SSL, a large-scale dataset constructed from VGGSound (Chen et al. 2020). It contains 13,981 spatial audio clips across 296 categories, each paired with semantically matched visual prompts. Ex-

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.
* Corresponding Author: *Xinyuan Qian*

tensive experiments show that AV-SSAN outperforms other AV-SSL baselines. Our contributions are listed as follows:

- We formulate CI-AVL, a novel task designed to enable selective localization of a target source. It utilizes semantic visual prompts derived from a different instance of the same class, thereby relaxing the requirement of explicitly paired audio-visual data.
- We propose AV-SSAN, an innovative framework that performs explicit Semantic-Spatial Alignment, enabling identity-aware localization by bridging visual semantics with spatial audio features.
- We propose an MB-SSA Net module, which introduces frequency-aware modeling into the alignment process. It mimics the frequency-dependent nature of spatial hearing, using a tri-band decomposition design with semantic-guided band fusion and spatial refinement.
- We construct VGGSound-SSL, a large-scale AV-SSL dataset comprising 13,981 spatial audio clips across 296 sound event categories paired with semantic visual prompts. This dataset offers a valuable benchmark for future research in identity-aware localization.

Related Work

Audio-only SSL

Traditional SSL methods combine handcrafted spatial features with deep learning models. GCC-MLP (He, Motlicek, and Odobez 2018) feeds GCC-PHAT features into multilayer perceptrons, while DR-MUSIC (Shmuel et al. 2023) enhances covariance matrix estimation for the MUSIC algorithm. Cross3D (Pertilä and Cakir 2017) processes SRP-PHAT feature using 3D CNNs for localization in reverberant environments.

Fully end-to-end models directly operate on spectral inputs. CNN-based methods (Chakrabarty and Habets 2017) exploit Short-Time Fourier Transform (STFT) phase cues, while SELDNet (Adavanne et al. 2018) integrates magnitude and phase using CRNNs. Later extensions incorporate temporal modeling via LSTMs (Li, Zhang, and Li 2018), and more recently, TF-Mamba (Xiao and Das 2024) replaces RNNs with Mamba (Jiang, Han, and Mesgarani 2025), improving performance in complex acoustic scenes.

Despite these advances, audio-only methods lack high-level semantic understanding and fail to disambiguate overlapping sources or occluded sources.

Audio-Visual SSL

Recent audio-visual methods integrate visual context to enhance robustness and resolve spatial ambiguities. A prevalent strategy is fusing visual embeddings with GCC-PHAT features. Specific approaches include: MLP-AVC (Qian et al. 2021b), which models visual priors as multivariate Gaussians; AVMLP (Qian et al. 2021a), which utilizes de-emphasis maps to suppress distractors; DGB (Qian et al. 2022b), which explores cross-modal latent spaces via generative models; CMAF (Qian et al. 2022a), which employs dynamic attention for audio-visual alignment; and AVST (Zhao et al. 2023), which disentangles modality encoding using Vision Transformers (Dosovitskiy et al. 2020) prior to multimodal fusion.

Beyond spatial features, richer spectral representations are also employed. Some works (Wu, Hu, and Wang 2023; Wu et al. 2023) integrate STFTs and Gaussian-encoded visual cues through CNN or Transformers (Vaswani et al. 2017). AV-SELD (Berghi et al. 2024) combines Log-mel spectrograms and Intensity Vectors with visual features, fusing them via Conformer (Gulati et al. 2020) blocks.

Although these methods achieve promising results, their reliance on tightly synchronized audio-visual pairs limits their applicability. Moreover, they estimate all active sources indiscriminately and lack the ability to isolate a target source. These limitations stem from their emphasis on spatial correlation, while ignoring semantic alignment, which is crucial for selective localization.

Selective and Prompt-based SSL

Recent works explored selective localization using external semantic cues. Class-conditioned SELD models (Slizovskaia et al. 2022; Shimada et al. 2024) incorporate class labels to guide attention toward target categories. Text-queried SSL (Zhao et al. 2024) fuses textual prompts with spatial audio for selective reasoning. LocSelect (Chen et al. 2024) uses a reference audio to localize the corresponding speaker in mixtures. However, these approaches depend on either class labels or same-instance queries, limiting generalization to new categories or modalities.

In summary, Audio-only methods leverage handcrafted spatial features but degrade under noisy and reverberant conditions. AV-SSL methods incorporate visual signals to enhance robustness, but they are constrained by paired data and the absence of semantic-level alignment mechanisms. Recent advances explore selective localization using text or reference audio, but are constrained to predefined sound categories and overlook visual prompt-based conditioning. These challenges highlight the need for a semantically aware SSL framework capable of selectively extracting the target sound source, which leads to the introduction of CI-AVL.

Methodology

We introduce the AV-SSAN to address the core challenge of cross-instance semantic-spatial alignment for CI-AVL. As shown in Figure 1, AV-SSAN comprises three modules: 1) a SBF Net that fuses visual prompts and semantic audio embeddings, 2) an MB-SSA Net that achieves multi-band semantic-spatial alignment, and 3) a DoA Prediction Module that estimates target DoA.

Problem Formulation

Given a two-channel audio signal captured by a microphone pair $\mathbf{x}$ and a user-specified prompt image $\mathbf{I}$, CI-AVL aims to predict the DoA of the sound source semantically matching the prompt. We formulate this as a regression task, discretizing the DoA into 180 classes: $\hat{\theta} = \{j \mid 1 \leq j \leq 180, j \in \mathbb{Z}\}$. To capture spatial continuity, we model the posterior probability distribution of $\hat{\theta}$ using a Gaussian-like

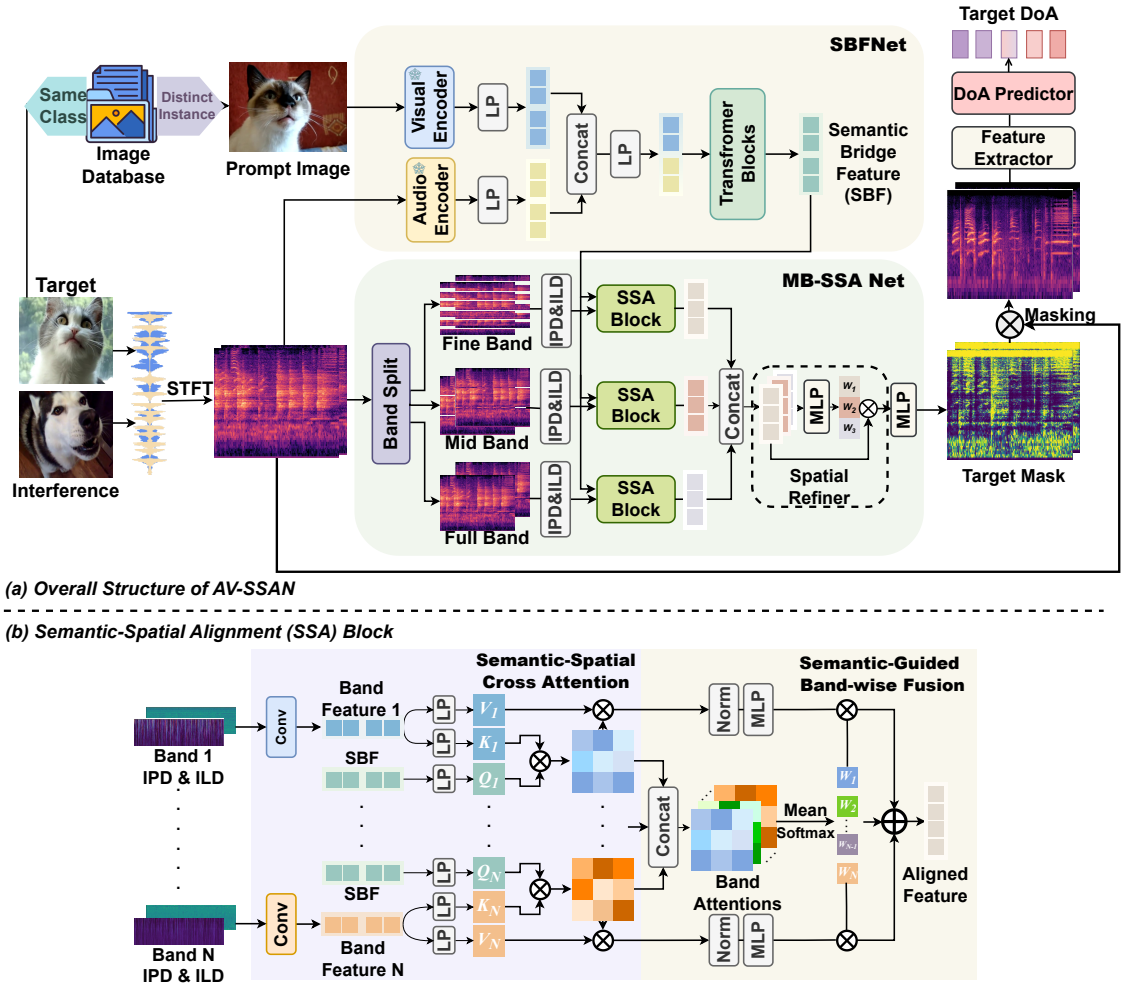

Figure 1: The architecture of (a) our proposed AV-SSAN and (b) Semantic-Spatial Alignment (SSA) block.

vector. $p(\theta) = \exp\left(-\frac{|\hat{\theta}-\theta|}{\sigma_\theta}\right)$ instead of one-hot encoding. $p(\theta)$ is centered on the ground truth θ with a standard deviation σ_θ. The normalization factor $\frac{1}{\sqrt{2\pi}\sigma_\theta}$ is omitted as it does not influence the model's predictions.

We use a deep neural network $\mathcal{F}$ to map the multimodal inputs to the DoA distribution $\hat{p}(\theta) = \mathcal{F}(\mathbf{I}, \mathbf{x}|\mathbf{\Omega})$ where $\mathbf{\Omega}$ are learnable parameters and the DoA is determined as the direction with the highest probability $\hat{\theta} = \arg\max_{\forall\theta} \hat{p}(\theta)$.

SBF Net

We first establish semantic correspondence between visual and audio modalities. We leverage a pretrained CLIP encoder (Radford et al. 2021) to extract the visual prompt representation $\mathbf{F}_V \in \mathbb{R}^{d_V}$ and a VGGish network (Hershey et al. 2017) to extract the semantic audio embedding $\mathbf{F}_A \in \mathbb{R}^{t \times d_A}$.

After temporal broadcasting and linear projection to a shared space, $\mathbf{F}_V^{'}$ and $\mathbf{F}_A^{'}$ are concatenated and processed by a Transformer encoder:

$$\mathbf{SBF} = \text{Transformer}(\text{LP}(\mathbf{F}_V^{'} \copyright \mathbf{F}_A^{'})) \quad (1)$$

where © denotes concatenation along the feature dimension and LP is a linear projection layer.

This SBF captures the semantic information of the target sound source, serving as a semantic cue for the subsequent target source's spatial characteristics disentanglement.

MB-SSA Net

Auditory spatial perception is inherently frequency-dependent (Brughera, Dunai, and Hartmann 2013). Motivated by psychoacoustic principles and prior empirical findings (Wang, Yang, and Li 2024), we design MB-SSA Net to model semantic-spatial alignment across different frequency bands. The architecture comprises three parts: 1) Tri-band Spatial Feature Extraction Module, 2) Semantic-Spatial Alignment (SSA) block, and 3) Spatial Refiner.

Tri-band Spatial Feature Extraction. Human auditory perception distinguishes spatial cues across frequency

Dataset	Samples	Duration	Sound Event	Annotations
AV16.3	43	∼2hrs	$2^{\diamondsuit}$	DoA
SSLR	6,622	∼25hrs	$2^{\diamondsuit}$	DoA, VAD
CAV3D	20	∼25hrs	$2^{\diamondsuit}$	DoA, VAD
AVRI	43	∼8hrs	$2^{\diamondsuit}$	DoA, VAD
STARSS23	168	∼7hrs	13	DoA, VAD, Object Category
VGGSound-SSL (ours)	13,981	∼39hrs	296	DoA, Object Category

Table 1: Comparison of VGGSound-SSL with existing datasets for AV-SSL. We use $\diamondsuit$ for datasets limited to male and female speech (counted as two distinct sound events). Annotations cover Object Category (sound event label), VAD (voice activity detection label), and DoA

bands (Strutt 1907). Motivated by this, we decompose spectrograms into three frequency resolutions: fine bands (32-bin width), mid bands (128-bin width), and the full spectrum. For each sub-band, we compute Interaural Phase Difference (IPD) and Interaural Level Difference (ILD):

$$\mathrm{IPD}^b(t,f) = \angle X_1^b(t,f) \; - \; \angle X_2^b(t,f) \tag{2}$$

$$\mathrm{ILD}^b(t,f) = 20\,\log_{10}\!\left(\frac{\left|X_1^b(t,f)\right|}{\left|X_2^b(t,f)\right| + \epsilon}\right) \tag{3}$$

where b denotes the band type (fine, mid, or full), ϵ is a small constant for stability, $\angle X$ represents the phase, and $\left|X\right|$ the magnitude of the spectrogram. This process yields the tri-band spatial features $\mathbf{X}^{\text{fine}} \in \mathbb{R}^{2\times T\times \lfloor \frac{F}{32} \rfloor \times 32}$, $\mathbf{X}^{\text{mid}} \in \mathbb{R}^{2\times T\times \lfloor \frac{F}{128} \rfloor \times 128}$, and $\mathbf{X}^{\text{full}} \in \mathbb{R}^{2\times T\times F\times 1}$.

SSA block. Leveraging the SBF rich in target semantics, we design the SSA block to selectively isolate the spatial characteristics of the target sound source. As depicted in Figure 2(b), the SSA block comprises two key components: a Semantic-Spatial Cross-Attention module and a Semantic-Guided Band-wise Fusion module.

For each sub-band spatial feature $\mathbf{X}^b_{:,i} \in \mathbb{R}^{2\times T\times d_b}$, where b denotes the band type (fine, mid, or full), $i = 1,\ldots,n$ represents the patch index, and d_b is the band width, we apply shared convolutional layers to encode it into $\mathbf{X}'^b_{:,i} \in \mathbb{R}^{2\times T\times 64}$. The encoded feature $\mathbf{X}'^b_{:,i}$ is then processed through a cross-attention mechanism, with the SBF serving as the query to extract target-related spatial information:

$$\mathbf{q}_i^b = \mathbf{W}^q_{b,i}\mathrm{SBF}, \quad \mathbf{k}_i^b = \mathbf{W}^k_{b,i}\mathbf{X}'^b_{:,i}, \quad \mathbf{v}_i^b = \mathbf{W}^v_{b,i}\mathbf{X}'^b_{:,i}, \tag{4}$$

where $\mathbf{W}^q_{b,i}$, $\mathbf{W}^k_{b,i}$, and $\mathbf{W}^v_{b,i}$ are learnable projection matrices. The attention output is computed as:

$$\mathbf{Z}_i^b = \mathrm{softmax}\left(\frac{\mathbf{q}_i^b(\mathbf{k}_i^b)^\top}{\sqrt{D}}\right)\mathbf{v}_i^b, \tag{5}$$

where D is the feature dimension.

Different sound events often exhibit energy concentration in distinct frequency ranges (e.g., whispers in high frequencies, bass in low). Motivated by this observation, we hypothesize that target-specific spatial characteristics may similarly vary across frequency bands, and that their semantic cross-attention weights should reflect this trend. To exploit this frequency-dependent behavior, we introduce a Semantic-Guided Band-wise Fusion module, which adaptively aggregates sub-band features based on their semantic relevance. This allows the model to assign greater weights to the sub-band features that are most informative for the given target. Specifically, we compute a semantic attention map for each patch as:

$$\mathbf{A}_i^b = \mathbf{q}_i^b(\mathbf{k}_i^b)^\top. \tag{6}$$

These attention maps are concatenated to form $\mathbf{A}^b = [\mathbf{A}_1^b,\ldots,\mathbf{A}_n^b]$. A scalar band importance vector $\boldsymbol{\beta}^b = [\beta_1^b,\ldots,\beta_n^b]$ is derived by applying mean pooling and softmax over $\mathbf{A}^b$. The final aggregated feature is computed as:

$$\mathbf{Z}^b = \sum_{i=1}^{n} \beta_i^b \mathbf{Z}_i^b. \tag{7}$$

Spatial Refiner. To unify the tri-band features $[\mathbf{Z}^{\text{fine}}, \mathbf{Z}^{\text{mid}}, \mathbf{Z}^{\text{full}}] \in \mathbb{R}^{3\times T\times 64}$, we employ an MLP that predicts temporal importance scores across bands. After softmax normalization, we compute a weighted sum:

$$\mathbf{Z}' = \sum_{b\in\{\text{fine,mid,full}\}} \alpha^b(t)\cdot \mathbf{Z}^b(t) \tag{8}$$

The fused feature $\mathbf{Z}' \in \mathbb{R}^{T\times 64}$ is passed through an MLP to predict a TF mask:

$$\mathbf{F}_{\text{Mask}} = \mathrm{MLP}(\mathbf{Z}') \tag{9}$$

This mask is applied to the original spectrogram:

$$\mathbf{X}m = \mathbf{X} \otimes \mathbf{F}_{\text{Mask}} \tag{10}$$

where $\otimes$ denotes element-wise multiplication.

DoA Prediction

Using the masked spectrogram $\mathbf{X}_m$ as input, an MLP followed by a softmax layer predicts the DoA posterior:

$$\hat{p}(\theta) = \mathrm{Softmax}(\mathrm{MLP}(\mathbf{X}_m)) \tag{11}$$

Training Loss

To optimize reconstruction loss between the masked spectrogram $\mathbf{X}_m$ and the ground truth $\mathbf{X}_{gt}$, we adopt a Mean Squared Error (MSE) loss:

$$\mathcal{L}_{recon} = \|\mathbf{X}_m - \mathbf{X}_{gt}\|_2^2 \tag{12}$$

Model	Modality	0 dB		-5 dB		-10 dB	
		MAE(°) ↓	ACC(%) ↑	MAE(°) ↓	ACC(%) ↑	MAE(°) ↓	ACC(%) ↑
SELDNet	Audio Only	32.19	51.27	32.82	49.91	33.49	44.67
GCC-MLP	Audio Only	30.08	54.89	30.94	52.98	31.22	51.69
MLP-AVC	Audio-Visual	28.92	55.28	29.13	51.63	29.97	50.74
AVMLP	Audio-Visual	25.42	59.34	30.07	55.24	31.72	51.62
DGB	Audio-Visual	19.09	63.87	30.38	56.63	30.98	52.87
AVST	Audio-Visual	21.53	64.14	26.94	58.71	27.21	55.51
AVSELD	Audio-Visual	18.95	67.56	22.81	60.86	24.69	57.21
CMAF	Audio-Visual	18.65	67.71	22.17	61.90	24.52	57.51
AV-SSAN (Ours)	Audio-Visual	**16.59**	**71.29**	**19.77**	**65.28**	**23.08**	**60.19**

Table 2: Experimental results under different SNRs on VGGSound-SSL.

An MSE loss is used for posterior probability-based DoA estimation, optimizing the predicted DoA distribution:

$$\mathcal{L}_{DoA} = \sum_{\theta=1}^{180} \|\hat{p}(\theta) - p(\theta)\|_2^2 \tag{13}$$

The final joint loss function is formulated as:

$$\arg\min \mathcal{L} = \mathcal{L}_{recon} + \mathcal{L}_{DoA} \tag{14}$$

Model	MAE(°)	ACC(%)
MLP-AVC	36.85	41.23
AVMLP	35.19	40.12
DGB	36.27	42.76
AVST	33.28	45.34
AVSELD	28.54	51.64
CMAF	30.03	50.87
AV-SSAN (Ours)	**27.46**	**52.31**

Table 3: Experimental results on STARSS23.

Experiments & Discussions

Dataset Construction

Table 1 provides a review of existing AV-SSL datasets. AV16.3 (Lathoud, Odobez, and Gatica-Perez 2004), CAV3D (Qian et al. 2019), SSLR (He, Motlicek, and Odobez 2018), and AVRI (Qian et al. 2022a) offer 3D location annotations with real-recorded spatial audio. However, they primarily contain human speech, restricting their utility in more general localization tasks involving diverse sound events. While STARSS23 (Shimada et al. 2023) includes 13 sound events with spatial annotations, its prevalence of overlapping, speech-dominated sources renders it less suitable for evaluating disentangled audio-visual spatial localization.

Our proposed VGGSound-SSL dataset, derived from VGGSound (Chen et al. 2020), comprises two-channel spatial audio and semantically aligned visual prompts for 296 distinct sound events. Its construction pipeline involves two primary stages—spatial audio synthesis and prompt image generation.

Spatial Audio Synthesis. We extract 10-second single-channel audio clips from VGGSound videos and resample them to 16 kHz. To simulate spatial audio, each audio segment is convolved with a randomly selected room impulse response (RIR) using GPU-RIR (Diaz-Guerra, Miguel, and Beltran 2021). A total of 10,000 RIRs are synthesized by varying critical parameters, including room dimensions, sound source positions, and reverberation times (T60).

Prompt Image Generation. To provide semantically consistent visual cues, we extract frames from each video and compute their CLIP embeddings. Concurrently, text embeddings for each sound class are obtained using CLIP. The frame exhibiting the highest image-text similarity is then selected as the prompt image, thus ensuring strong semantic alignment between the sound source and its visual reference. All selected prompt images were manually verified to ensure semantic consistency and reduce selection bias.

All selected prompt images are organized into class-specific pools. During training and inference, a prompt image of the target sound is randomly sampled from its corresponding pool, excluding the image from the current instance.

Baselines

We compare our proposed method with audio-only baselines, including 1) SELDNet (Adavanne et al. 2018) and 2) GCC-MLP (He, Motlicek, and Odobez 2018), which estimate DoA using only the mixed spatial audio signals. For audio-visual methods, we evaluate 3) MLP-AVC (Qian et al. 2021b), 4) AVMLP (Qian et al. 2021a), 5) DGB (Qian et al. 2022b), 6) AVST (Zhao et al. 2023), 7) AVSELD (Berghi et al. 2024), and 8) CMAF (Qian et al. 2022a).

Training and Evaluation

We train all models on the proposed VGGSound-SSL dataset. Each training sample is formed by randomly selecting two audio segments from distinct sound classes. One is selected as the target source, and the other as the interfering source. The two segments are mixed at an SNR of 0 dB. For visual prompting, we randomly selected an image corresponding to a different instance within the same category.

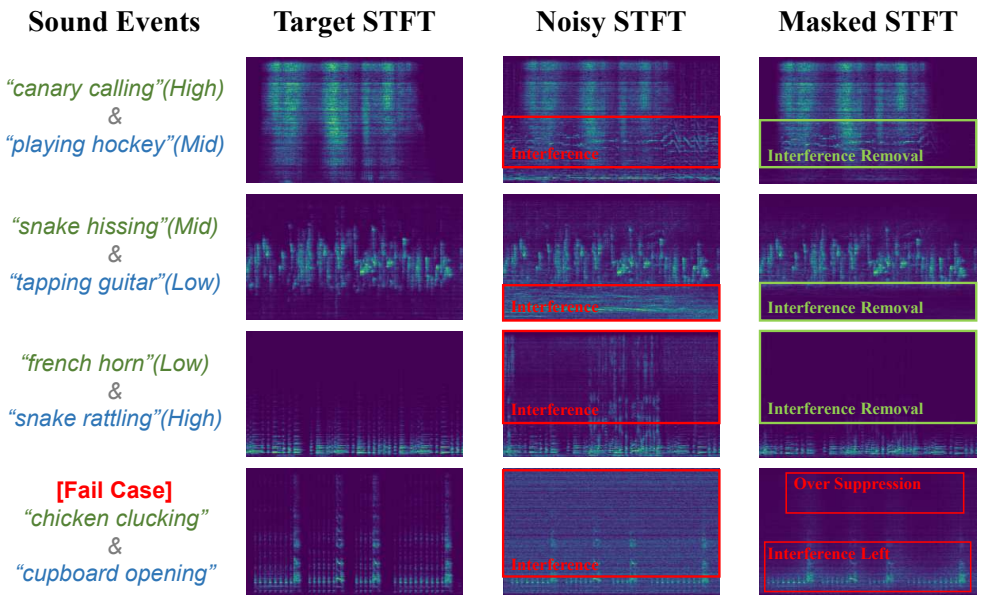

Figure 2: Visualizations of AV-SSAN's outputs on VGG-SSL. Each row shows a pair of overlapping sound events categorized by their dominant frequency regions: high (H), mid (M), or low (L).

To assess generalization, we further train models on the STARSS23 dataset, which contains real-world spatial audio recordings. For each 10-second segment, the most frequent sound event is selected as the target, with remaining events treated as interference. Since STARSS23 lacks visual counterparts, we use category-matched images from VGGSound-SSL as semantic prompts.

We evaluate performance using Mean Absolute Error (MAE) and Accuracy (ACC). MAE quantifies the average angular error, while ACC reports the percentage of predictions within a 5° tolerance. Lower MAE and higher ACC indicate stronger localization capability.

All the experiments are conducted on two RTX-4090 GPUs. Our model was trained for 140k steps with a batch size of 32, using the AdamW optimizer with a learning rate of 5e-3. For STFT computations, a frame size of 64 ms and a hop size of 32 ms were employed.

Results and Analysis

Table 2 presents performance comparisons on VGG-SSL across varying SNR levels. At SNR = 0 dB, our proposed AV-SSAN achieves the lowest MAE of 16.59° and the highest ACC of 71.29%, outperforming all competing methods.

As expected, audio-only models such as SELDNet and GCC-MLP yield ACC near 50%. This aligns with the binary choice setting of the task: without visual guidance, the model must choose between two plausible sources. The result confirms the necessity of visual guidance to resolve spatial ambiguity.

Compared to prior audio-visual methods, AV-SSAN shows superior accuracy. This validates the effectiveness of our Multi-band Semantic-Spatial Alignment framework, which explicitly aligns visual semantics with spatial auditory patterns across frequency bands.

To evaluate robustness in noisy environments, we conduct a generalization experiment. All models are trained at 0 dB and tested under SNRs ranging from -5 dB to –10 dB. As shown in Table 2, AV-SSAN retains strong performance and outperforms all baselines. This indicates that our method captures discriminative spatial cues resilient to interference, demonstrating strong generalization across noise conditions.

We further assess transferability on the STARSS23 dataset, which features real-world acoustic environments. As shown in Table 3, all models experience performance drops due to the increased acoustic complexity of real-world scenes, where multiple overlapping sound events create more challenging interference patterns. AV-SSAN remains the best performer, demonstrating stronger transferability to in-the-wild conditions.

Figure 2 provides qualitative examples of VGG-SSL. Our method effectively suppresses interference while preserving the target, showing the benefits of multi-band alignment. We also include a failure case where the target energy is over-suppressed and interference remains. We attribute this to an energy imbalance between the target and interference, which hampers the selective localization.

Ablation Study

We conduct ablation studies on the VGG-SSL dataset to isolate the contribution of each component in our proposed AV-SSAN framework. Results are summarized in Table 4.

Explicit Semantic-Spatial Alignment. We first evaluate the importance of semantic and spatial alignment. Simply applying cross-attention between cross-instance visual prompts and mixed spatial audio yields near-random performance of 51.38%. These results suggest a strong modality mismatch between the two modalities. Incorporating Semantic Audio representations reduces ambiguity, improving

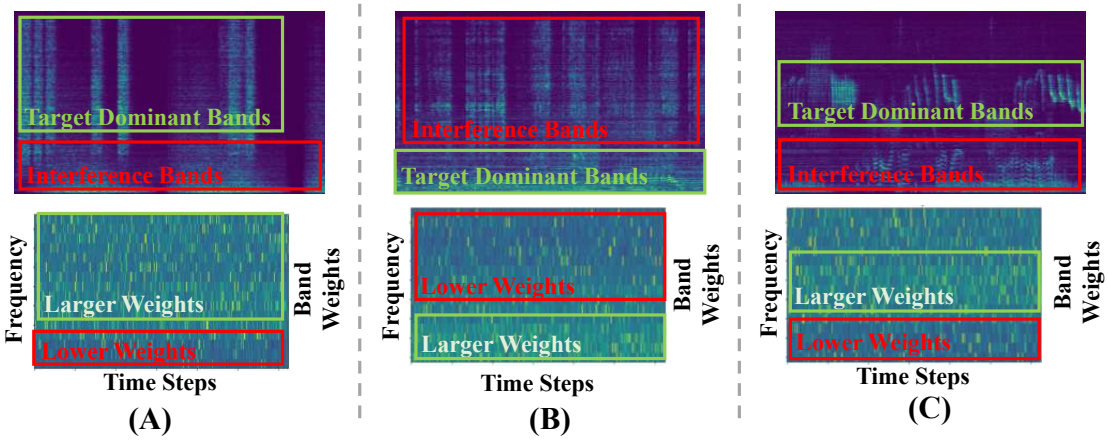

Figure 3: Visualizations of the ablation study on band attention. Our model assigns higher weights to frequency bands dominated by the target and suppresses interference-dominated bands.

V	A	MB	BA	Refiner	MAE (°)	ACC (%)
√	×	×	×	×	28.54	51.38
√	√	×	×	×	22.21	61.99
√	√	√	×	×	19.19	68.04
√	√	√	√	×	17.98	69.48
√	√	√	√	√	16.59	71.29

Table 4: Ablation study on VGG-SSL. V: visual prompt; A: semantic audio; MB: multi-band modeling; BA: band attention; Refiner: spatial refiner.

ACC by 10.61% and lowering MAE by 6.33°. It confirms that bridging visual semantics and spatial acoustics at a semantic level is critical for disambiguating the target.

Multi-band Semantic-Spatial Alignment. Adding multi-band modeling further improves performance, increasing ACC by 6.05%. This validates our design to align spatial features across multiple frequency bands. By decomposing the audio into coarse-to-fine frequency resolutions, the model better captures target-specific spatial patterns that may be frequency-dependent.

Band Attention. Introducing Band Attention yields additional gains, with ACC reaching 69.48% and the MAE dropping to 17.98°. This confirms that not all frequency bands contribute equally. Our attention mechanism allows the model to prioritize sub-bands that are more discriminative for a given target class, thus enhancing localization performance.

Figure 4 provides a visualization of the learned band attentions. Patches corresponding to the target exhibit stronger responses, while interference regions are suppressed. This validates our hypothesis: semantic-spatial cross attention should reflect frequency-specific discriminability, and band attention enables this behavior to emerge dynamically.

Spatial Refiner. Finally, adding the Spatial Refiner leads to further improvements of 1.8% in ACC and 1.39° in MAE. It integrates refined spatial features from attended sub-bands, enabling the model to make more globally consistent predictions based on the band-wise aligned information.

Conclusion

We introduce CI-AVL, a novel and challenging task that localizes a target sound source using semantically related but spatially unpaired visual prompts. To solve this task, we propose AV-SSAN, a selective localization framework that performs multi-band semantic-spatial alignment between visual semantics and spatial audio cues. To support research in this direction, we construct VGG-SSL, a large-scale dataset with 13,981 spatial audio clips and class-consistent prompt images. Extensive experiments show that our method outperforms SOTA approaches, achieving a MAE of 16.59° and an ACC of 71.29%.

Acknowledgements

This work is supported by National Natural Science Foundation of China (No. 62306029) and Shenzhen Research Institute of Big Data (K00120240007).

References

Adavanne, S.; Politis, A.; Nikunen, J.; and Virtanen, T. 2018. Sound event localization and detection of overlapping sources using convolutional recurrent neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 13(1): 34–48.

Berghi, D.; Wu, P.; Zhao, J.; Wang, W.; and Jackson, P. J. 2024. Fusion of audio and visual embeddings for sound event localization and detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 8816–8820. IEEE.

Brughera, A.; Dunai, L.; and Hartmann, W. M. 2013. Human interaural time difference thresholds for sine tones: The high-frequency limit. *The Journal of the Acoustical Society of America*, 133(5): 2839–2855.

Chakrabarty, S.; and Habets, E. A. 2017. Broadband DOA estimation using convolutional neural networks trained with noise signals. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 136–140. IEEE.

Chen, H.; Xie, W.; Vedaldi, A.; and Zisserman, A. 2020. Vggsound: A large-scale audio-visual dataset. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 721–725. IEEE.

Chen, Y.; Qian, X.; Pan, Z.; Chen, K.; and Li, H. 2024. LocSelect: Target Speaker Localization with an Auditory Selective Hearing Mechanism. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 8696–8700. IEEE.

Diaz-Guerra, D.; Miguel, A.; and Beltran, J. R. 2021. gpuRIR: A python library for room impulse response simulation with GPU acceleration. *Multimedia Tools and Applications*, 80(4): 5653–5671.

DiBiase, J. H.; Silverman, H. F.; and Brandstein, M. S. 2001. Robust localization in reverberant rooms. In *Microphone arrays: signal processing techniques and applications*, 157–180. Springer.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Gulati, A.; Qin, J.; Chiu, C.-C.; Parmar, N.; Zhang, Y.; Yu, J.; Han, W.; Wang, S.; Zhang, Z.; Wu, Y.; et al. 2020. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*.

He, W.; Motlicek, P.; and Odobez, J.-M. 2018. Deep neural networks for multiple speaker detection and localization. In *IEEE International Conference on Robotics and Automation*, 74–79. IEEE.

Hershey, S.; Chaudhuri, S.; Ellis, D. P.; Gemmeke, J. F.; Jansen, A.; Moore, R. C.; Plakal, M.; Platt, D.; Saurous, R. A.; Seybold, B.; et al. 2017. CNN architectures for large-scale audio classification. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 131–135. IEEE.

Jiang, X.; Han, C.; and Mesgarani, N. 2025. Dual-path mamba: Short and long-term bidirectional selective structured state space models for speech separation. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1–5. IEEE.

Jiang, Y.; Chen, H.; Du, J.; Wang, Q.; and Lee, C.-H. 2023. Incorporating lip features into audio-visual multi-speaker doa estimation by gated fusion. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1–5. IEEE.

Knapp, C.; and Carter, G. 1976. The generalized correlation method for estimation of time delay. *IEEE Transactions on acoustics, speech, and signal processing*, 24(4): 320–327.

Lathoud, G.; Odobez, J.-M.; and Gatica-Perez, D. 2004. AV16. 3: An audio-visual corpus for speaker localization and tracking. In *International Workshop on Machine Learning for Multimodal Interaction*, 182–195. Springer.

Li, Q.; Zhang, X.; and Li, H. 2018. Online direction of arrival estimation based on deep learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2616–2620. IEEE.

Pertilä, P.; and Cakir, E. 2017. Robust direction estimation with convolutional neural networks based steered response power. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 6125–6129. IEEE.

Qian, X.; Brutti, A.; Lanz, O.; Omologo, M.; and Cavallaro, A. 2019. Multi-speaker tracking from an audio–visual sensing device. *IEEE Transactions on Multimedia*, 21(10): 2576–2588.

Qian, X.; Brutti, A.; Lanz, O.; Omologo, M.; and Cavallaro, A. 2021a. Audio-visual tracking of concurrent speakers. *IEEE Transactions on Multimedia*, 24: 942–954.

Qian, X.; Madhavi, M.; Pan, Z.; Wang, J.; and Li, H. 2021b. Multi-target DoA estimation with an audio-visual fusion mechanism. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 4280–4284. IEEE.

Qian, X.; Wang, Z.; Wang, J.; Guan, G.; and Li, H. 2022a. Audio-visual cross-attention network for robotic speaker tracking. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31: 550–562.

Qian, X.; Zhang, Q.; Guan, G.; and Xue, W. 2022b. Deep audio-visual beamforming for speaker localization. *IEEE Signal Processing Letters*, 29: 1132–1136.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PmLR.

Schmidt, R. 1986. Multiple emitter location and signal parameter estimation. *IEEE Transactions on Antennas and Propagation*, 34(3): 276–280.

Shimada, K.; Politis, A.; Sudarsanam, P.; Krause, D. A.; Uchida, K.; Adavanne, S.; Hakala, A.; Koyama, Y.; Takahashi, N.; Takahashi, S.; et al. 2023. STARSS23: An audio-visual dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events. *Advances in Neural Information Processing Systems*, 36: 72931–72957.

Shimada, K.; Uchida, K.; Koyama, Y.; Shibuya, T.; Takahashi, S.; Mitsufuji, Y.; and Kawahara, T. 2024. Zero-and few-shot sound event localization and detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 636–640. IEEE.

Shmuel, D. H.; Merkofer, J. P.; Revach, G.; Van Sloun, R. J.; and Shlezinger, N. 2023. Deep root MUSIC algorithm for data-driven DoA estimation. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1–5. IEEE.

Slizovskaia, O.; Wichern, G.; Wang, Z.-Q.; and Le Roux, J. 2022. Locate this, not that: Class-conditioned sound event doa estimation. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 711–715. IEEE.

Strutt, J. W. 1907. On our perception of sound direction. *Philosophical Magazine*, 13(74): 214–32.

Taevs, M.; Dahmani, L.; Zatorre, R. J.; and Bohbot, V. D. 2010. Semantic elaboration in auditory and visual spatial memory. *Frontiers in Psychology*, 1: 228.

van der Heijden, K.; Rauschecker, J. P.; de Gelder, B.; and Formisano, E. 2019. Cortical mechanisms of spatial hearing. *Nature Reviews Neuroscience*, 20(10): 609–623.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Wang, Y.; Yang, B.; and Li, X. 2024. IPDnet: A universal direct-path IPD estimation network for sound source localization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Wu, Y.; Hu, R.; and Wang, X. 2023. Multi-speaker Direction of Arrival Estimation Using Audio and Visual Modalities with Convolutional Neural Network. In *IEEE International Conference on Multimedia and Expo*, 636–641. IEEE.

Wu, Y.; Hu, R.; Wang, X.; and Ke, S. 2023. Multi-speaker DoA estimation using audio and visual modality. *Neural Processing Letters*, 55(7): 8887–8901.

Xiao, Y.; and Das, R. K. 2024. Tf-mamba: A time-frequency network for sound source localization. *arXiv preprint arXiv:2409.05034*.

Zhao, J.; Qian, X.; Xu, Y.; Liu, H.; Cao, Y.; Berghi, D.; and Wang, W. 2024. Text-queried target sound event localization. In *European Signal Processing Conference*, 261–265. IEEE.

Zhao, J.; Xu, Y.; Qian, X.; and Wang, W. 2023. Audio visual speaker localization from egocentric views. *arXiv preprint arXiv:2309.16308*.