Seeing Sound, Hearing Sight: Uncovering Modality Bias and Conflict of AI models in Sound Localization

Yanhao Jia[1], Ji Xie[1], S Jivaganesh[1], Hao Li[2], Xu Wu[1,3], Mengmi Zhang[1]†

[1] Nanyang Technological University, Singapore
[2] Peking University, China [3] Shenzhen University, China
† Corresponding author; address correspondence to mengmi.zhang@ntu.edu.sg

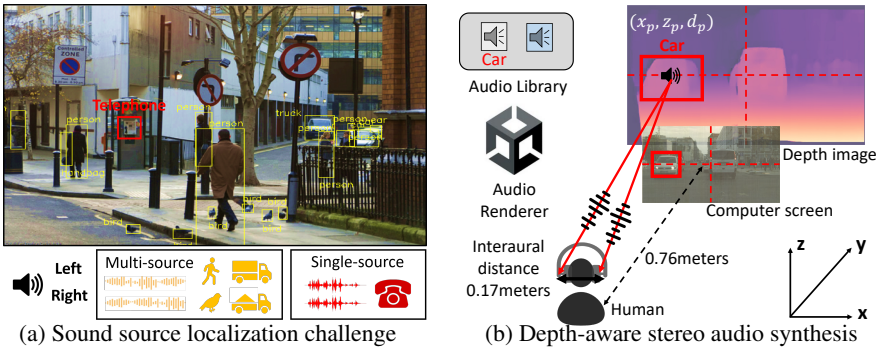
Abstract

Imagine hearing a dog bark and instinctively turning toward the sound—only to find a parked car, while a silent dog sits nearby. Such moments of sensory conflict challenge perception, yet humans flexibly resolve these discrepancies, prioritizing auditory cues over misleading visuals to accurately localize sounds. Despite the rapid advancement of multimodal AI models that integrate vision and sound, little is known about how these systems handle cross-modal conflicts or whether they favor one modality over another. Here, we systematically and quantitatively examine modality bias and conflict resolution in AI models for Sound Source Localization (SSL). We evaluate a wide range of state-of-the-art multimodal models and compare them against human performance in psychophysics experiments spanning six audiovisual conditions, including congruent, conflicting, and absent visual and audio cues. Our results reveal that humans consistently outperform AI in SSL and exhibit greater robustness to conflicting or absent visual information by effectively prioritizing auditory signals. In contrast, AI shows a pronounced bias toward vision, often failing to suppress irrelevant or conflicting visual input, leading to chance-level performance. To bridge this gap, we present EchoPin, a neuroscience-inspired multimodal model for SSL that emulates human auditory perception. The model is trained on our carefully curated AudioCOCO dataset, in which stereo audio signals are first rendered using a physics-based 3D simulator, then filtered with Head-Related Transfer Functions (HRTFs) to capture pinnae, head, and torso effects, and finally transformed into cochleagram representations that mimic cochlear processing. To eliminate existing biases in standard benchmark datasets, we carefully controlled the vocal object sizes, semantics, and spatial locations in the corresponding images of AudioCOCO. EchoPin outperforms existing models trained on standard audio-visual datasets. Remarkably, consistent with neuroscience findings, it exhibits a human-like localization bias, favoring horizontal (left–right) precision over vertical (up–down) precision. This asymmetry likely arises from HRTF-shaped and cochlear-modulated stereo audio and the lateral placement of human ears, highlighting how sensory input quality and physical structure jointly shape precision of multimodal representations. All code, data, and models are available here.

1 Introduction

Sound source localization (SSL) is the task of identifying the spatial origin of a sound within a visual scene. It plays a fundamental role in both biological perception [1] and artificial intelligence (AI) [2], enabling systems to connect what they hear with what they see. In natural environments, visual inputs interact with auditory cues, often dominating or recalibrating sound perception—a phenomenon exemplified by classic cross-modal illusions [3]. In practice, accurate SSL is critical for

39th Conference on Neural Information Processing Systems (NeurIPS 2025).



(a) Sound source localization challenge (b) Depth-aware stereo audio synthesis

Figure 1: **(a) Sound source localization challenge in naturalistic images.** In the multi-source scenario (orange panel), multiple objects in the scene—such as pedestrians, birds, cars, and trucks (highlighted by yellow boxes)—emit sounds, whereas in the single-source scenario (red panel), only one object, a telephone (red box), produces sound. In both settings, the task is to localize all sounding objects based on two-channel stereo audio. To allow systematic and controllable benchmarking of human and AI performance, we focus on the single-source localization task (red panel), which remains challenging due to scene clutter, occlusions, and ambiguous visual cues. **(b) Depth-aware stereo audio synthesis.** In the 3D simulator, a human listener (interaural distance: 0.17m) is placed at the origin, facing the RGB image on the screen. This image and its depth image are aligned in the same direction. Using a spatial audio renderer and audio from our library, stereo audio of the target car (red box) in the RGB image can be synthesized (see **Sec 2.1**).

real-world systems such as autonomous vehicles anticipating hazards [4], assistive technologies for visually impaired users [5], and robots operating in human-centered environments [6, 7, 8, 9, 10]. These applications demand robust auditory inference under noisy, cluttered, and ambiguous sensory conditions. Consider a real-life example in **Fig. 1a** depicting a busy traffic intersection: cars honk, people converse, birds chirp, and an ambulance siren approaches from behind a building. Humans must rapidly detect and localize the siren despite competing sounds, partial visual occlusion, and environmental noise. Such everyday scenes highlight the challenges of SSL: resolving ambiguous visual or auditory cues, handling conflicting signals across modalities, and focusing attention on the most relevant source amidst distractions. Building AI systems capable of performing SSL tasks in these conditions remains an open problem.

Recent AI research has developed multimodal models for SSL [11, 12, 13, 14, 15, 2, 16]. Methods such as contrastive learning [17, 18, 19] aim to align audio-visual embeddings by maximizing the similarity between matching pairs and minimizing it between mismatched ones. Other approaches leverage cross-modal attention [20, 21, 22] in transformer architectures [23, 24] to allow sound features to dynamically query relevant visual regions. Some works [25, 26, 27] also incorporate object priors, leveraging knowledge of what typically makes sound to guide localization. However, despite these advances [28], little is known about how such models behave when the modalities conflict or when biases emerge—such as favoring visual cues over auditory ones.

To address this gap, we systematically evaluate model behavior under six controlled audio-visual conditions: (1) Congruent — audio and visual cues align in both semantics and location; (2) conflicting visual cues, where the visual scene misleads localization; (3) absent visual cues, where the sounding object is completely occluded in the visual scene; and (4) vision-only and (5) audio-only conditions, where either vision or audio is entirely omitted. Moreover, similar to the cocktail party problem [29, 30], we further extend the single-source stereo audio to (6) the multi-instance SSL [31], where multiple instances of the same semantic category are present, potentially distracting the model from correctly localizing the target instance. These manipulations allow us to probe how models resolve cross-modal ambiguity and characterize their reliance on each modality. To benchmark these model behaviors, we additionally conduct human behavioral studies under the same experimental conditions. Results show a clear performance gap: humans consistently outperform AI models in handling both congruent and incongruent conditions.

While numerous SSL datasets [32, 33, 34] exist, they often suffer from limitations that hinder robust multimodal alignment in AI models. Typically, these datasets [35, 36, 37, 38, 39, 40, 41] consist of scenes with a single, large, centrally placed sounding object, making them vulnerable to visual shortcut learning, where models perform well without truly integrating audio information [42, 43]. Others [44, 45, 46] are constructed by pairing unrelated images and sounds from independent

datasets [47, 48, 49, 50, 51, 52, 53, 54, 55], leading to weak cross-modal entanglement. More recent efforts [56, 57, 58] synthesize audio based on physics-informed rules [59, 60] or generative AI models [61, 62], but they still frequently rely on mono audio, neglecting the richer spatial cues provided by stereo audio signals.

Inspired by neuroscience findings highlighting the role of inter-channel differences in spatial hearing [63], we propose a method that leverages 3D simulation engines to generate stereo audio from images by integrating separate image and sound datasets. Unlike neuroscience approaches that require physically setting up microphones in real-world spaces, which is both time-consuming and costly, our simulation-based method offers a scalable and efficient alternative for producing spatialized audio paired with complex visual scenes. With this method, we contribute a large-scale AudioCOCO dataset, comprising 28,224 image-audio pairs with ground truth annotations. By simulating stereo audio that adheres to physical principles of sound propagation and spatial cues, AudioCOCO provides realistic and diverse audio-visual scenes.

Human sound localization relies on a cascade of acoustic transformations shaped by the ear, head, and torso. Direction-dependent spectral notches introduced by the pinna help resolve elevation and front–back ambiguities [1, 64, 65]. These cues are formalized in the Head-Related Transfer Function (HRTF), which encodes interaural time (ITD) and level differences (ILD) alongside fine-grained spectral features. At the cochlea, sounds are further decomposed into frequency-selective channels that preserve ITD/ILD while transforming HRTF-induced modulations into neural representations.

To model these biological mechanisms, we introduce EchoPin, a neuroscience-inspired model for SSL. EchoPin pre-processes stereo audio using Head-Related Transfer Function (HRTF)–based filtering and cochleagram representations derived from Equivalent Rectangular Bandwidth (ERB) filters. These designs capture the tonotopic organization and temporal dynamics of the auditory periphery [66, 67] more faithfully than conventional mel-spectrograms [68, 69, 70]. EchoPin then employs dual encoders to jointly process audio and visual inputs, trained with contrastive learning on our AudioCOCO dataset. Experimental results suggest that EchoPin outperforms existing models. Notably, without any human behavioral supervision, EchoPin reproduces a human-like horizontal localization bias [71], an emergent property that was not prominent in previous AI systems. We attribute this to EchoPin's 3D stereo audio pipeline, which integrates interaural spacing, HRTF filtering, and ERB-based cochlear processing to jointly enhance sensory fidelity and multimodal alignment. We highlight our key contributions below:

1. We introduce a unified framework to systematically benchmark audio-visual localization models under modality conflicts, absence, and misalignment. We propose a scalable pipeline that synthesizes two-channel stereo audio for static images via 3D simulation, and construct a large-scale, naturalistic, and spatially grounded audio-visual dataset, named as AudioCOCO.

2. We design and conduct psychophysics experiments to assess human strategies in resolving audio-visual conflicts, providing a strong baseline for AI-human comparison. We provide a detailed analysis of modality conflicts and biases in existing SSL models and humans, highlighting key performance and behavioral differences under challenging multimodal conditions.

3. We introduce EchoPin, a neuroscience-inspired model trained on our curated AudioCOCO dataset. It features a dual-encoder architecture that processes stereo audio–visual pairs. The stereo audio is pre-processed using HRTF-based spatial filtering and cochlear-inspired frequency decomposition. Experimental results show that precise audio–visual alignment emerges from high-fidelity sensory inputs and biologically grounded ear-structure priors. EchoPin not only achieves superior localization accuracy but also exhibits human-like localization biases, favoring horizontal over vertical precision.

2 Experiments

2.1 AudioCOCO dataset

Image selection. We use the MSCOCO [47] dataset for its broad coverage of everyday objects and apply our selection criteria below to all images from the standard training and test sets. From its dataset annotations, we manually select 12 audible object categories encompassing humans, animals, vehicles, and electronic devices. Recognizing that larger objects may be easier to detect and localize, we further categorize sounding objects by their relative size in the image. Object size is defined as the ratio of the object's segmentation mask area to the total image area, independent of their

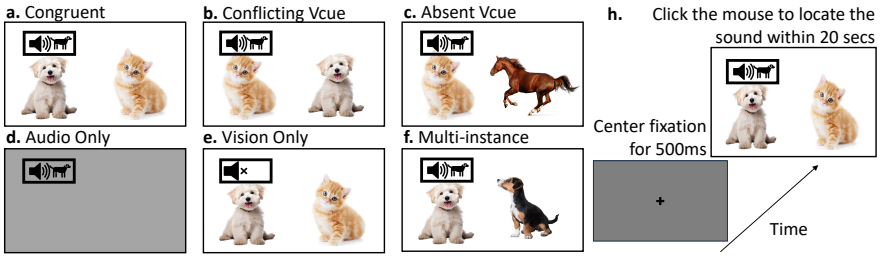


Figure 2: **Overview of congruent and manipulated vision-audio conditions in our UniAV framework and task schematic**. An example of the vision-audio congruent condition (a) is shown, where a dog sound is played (black box) with a matching visual source. Visual and auditory modifications for the other five experimental conditions (b-f) are also displayed. See **Sec 2.2** for more details. (h) Each trial began with a fixation cross (500 ms), followed by the presentation of an image-audio pair from either of the six conditions (a-f). Participants were instructed to use the computer mouse to click on the perceived location of the sound source within 20 seconds.

real-world scale. We define three size bins: Size1 (0–5%), Size2 (5–15%), and Size3 (15–30%). Objects occupying more than 30% of the image are excluded, as they make the localization task trivially easy. To maintain class balance, we limit the number of images per object size per category to 150 within each training or test set. For experimental control, each image contains only one target sounding object, without any other instances of the same category. For the multi-instance SSL (**Sec 2.2**), we select 150 images per object size per category, each containing 2–5 instances of the same semantic category, with exactly one designated as the sounding target. See **Supp Fig. S1**, **Supp Fig. S2**, and **Supp Fig. S3** for the distribution of image counts by category and their target spatial locations in both the training and test sets. After applying the selection criteria, we randomly sample 4,953 qualified images from the MSCOCO training set for training, and 5,500 images from the official test set—comprising 2,840 single-object and 2,660 multi-object scenes—for testing.

Audio selection. Prior work often pairs MSCOCO images with external audio datasets like VGGSound [49] or FSDnoisy18k [50], but these audio sources often lack spatial and temporal consistency. For example, moving sound sources (e.g., a dog running across the video frames) cause spatial shifts that make them unsuitable for generating realistic stereo audio on static images. Additionally, recordings may include background noise or mixed sounds from multiple objects, reducing semantic clarity and overall quality. To address these issues, we apply three filtering steps to the VGGSound videos, retaining 3,533 clips from the standard training set and 727 clips from the standard test set that contain high-quality audio, spatially and semantically aligned with the visual content. See **Supp Sec. S1**, **Supp Fig. S4**, and **Supp Fig. S5** for detailed filtering steps and results.

Next, we randomly pair the selected MSCOCO images with high-quality audios from VGGSound to construct the AudioCOCO dataset. The dataset comprises 9,360 audio–image pairs in the training set and 18,864 pairs in the test set, which are further divided into six experimental conditions (**Sec. 2.2**).

Depth-aware stereo audio synthesis. To generate spatialized stereo audio for their corresponding sounding objects on static images, we employ Unity [72] as our 3D simulation engine, which allows us to control precise object locations and simulate realistic stereo sound based on the spatial layout of visual scenes. As illustrated in **Fig. 1b**, we define a Cartesian coordinate system within Unity. The listener is positioned at the origin (0, 0, 0). The computer screen displays the image to listener, is placed parallel to the x–z plane and aligned along the positive y-axis, with a fixed physical distance of 0.76 meters from the listener in the human psychophysics experiment (**Sec 2.3**). To estimate the depth of objects within the 2D image, we utilize the DepthAnything model [73], which outputs a relative depth map with values ranging from 0 to 10. However, without access to the original camera intrinsics of MSCOCO images, determining absolute scene scale in Unity is nontrivial. To address this, we normalize the depth values d_p of the image as $d_p^{\text{norm}} = d_{p,\max} - d_p$ where $d_{p,\max}$ are the maximum depth values in the image. Sound loudness decreases logarithmically with distance from the listener. To maintain sound amplitudes within a comfortable and perceptible range, we rescale the normalized depth by 0.5. The final y-coordinate in Unity y_u for the sounding object is computed as $y_u = d_p^{\text{norm}}/2 + 0.76$. For the x and z coordinates of the sounding object in Unity, we map the object's pixel location from the 2D image to the physical dimensions of the monitor. Given that the display has a resolution of 90 pixels per inch, we compute $x_u = x_p/90$ and $z_u = z_p/90$, where x_p and z_p are the pixel coordinates of the target object's center in the image.

Once the sounding object's 3D position (x_u, y_u, z_u) is established in Unity, we place a static audio source at this location. Using an interaural distance of 0.17 meters, reflecting the typical distance between human ears, Unity simulates realistic two-channel stereo sound based on the spatial relationship between the listener and the sound source. This procedure allows us to synthesize spatially grounded, depth-aware stereo audio for each image-sound pair.

2.2 Experimental conditions in the AudioCOCO test set

As shown in **Fig. 2**, the AudioCOCO test set includes six experimental conditions, totaling 18,864 image–audio pairs to systematically probe modality biases and conflicts in humans and AI models. Each condition contains 2,900 pairs, except MultiInstLoc, which includes 4,364. All models are trained only on congruent conditions to evaluate generalization, while these six conditions are used exclusively for testing.

Audio-visual Congruent (Congruent) represents the ideal scenario where both the audio semantics and localization align perfectly with the corresponding visual target's semantics and location. This should serve as the upper bound for performance in both humans and AI models. For instance, as shown in **Fig. 2(a)**, a dog sound is played at the same location as the dog in the image.

Conflicting Visual Cue (ConflictVcue) examines the scenario where the semantics of both the visual and audio cues belong to the correct category but are spatially misaligned. In **Fig. 2(b)**, a dog sound is played at the location of a cat, while the silent dog is visually present at a different location. Among all the image-audio pairs in Congruent condition, we randomly choose an object from a non-target category as the sound source. We do not limit the distance between the distractor and the target.

Absent Visual Cue (AbsVcue) explores the case when a target sound is present but the visual scene contains non-relevant objects, and no visual cue matches the sound. For instance, in **Fig. 2(c)**, a dog sound might be played on the cat, but no dog is visually present. From the image-audio pairs in the congruent condition, we randomly select a target sound to play at a randomly selected sounding object in the scene where no relevant objects aligning with the semantics of the sound source exists. This condition is more stringent than Conflicting Visual Cue, as it lacks any visual cues altogether.

Audio Only (AOnly) represents the extreme case where no meaningful visual information is provided, and the sound source is randomly placed anywhere within the image. The image could be a blank gray image with pixel values set to 128 (**Fig. 2(d)**) or a Gaussian noise image with a mean of 0 and a standard deviation of 1. For example, a dog sound could be randomly played on the left side of a pure gray background.

Vision Only (VOnly) exploits multi-modal biases or priors. In this condition, only visual scenes are provided to the AI models, and they must localize the sounding object despite the absence of any meaningful sound. The audio could either be completely silent or filled with random Gaussian noise (mean 0, standard deviation 1). For example, the same visual stimulus as in the congruent condition is presented, but the dog sound is replaced with silence or noise (**Fig. 2(e)**).

Multi-Instance Localization (MultiInstLoc) follows the same motivation as the cocktail party problem [29] and features several objects of the same category, but the audio corresponds to only one specific instance, testing localization accuracy in a multi-instance scenario. For example, as illustrated in **Fig. 2(f)**, a dog sound is played at the location of the left dog, while both dogs are visually present in the scene. This condition follows the same setup as the Congruent condition, but is more challenging due to the presence of multiple visually relevant objects. We selected images from the test set of the MSCOCO dataset, where the number of object instances within a given image is restricted to between 2 and 5 for the target category.

2.3 Human psychophysics experiment

We conducted in-lab psychophysics experiments on the AudioCOCO test set with 14 participants, collecting a total of 2,100 trials. All the experiments are conducted with the subjects' informed consent and according to protocols approved by the Institutional Review Board of our institution. Every experiment lasted approximately 40 minutes. The experimental setup is schematically illustrated in **Fig. 2(h)**. Each trial began with a fixation cross displayed for 500 milliseconds, followed by the presentation of an image and audio pair drawn from one of the six experimental conditions (see **Sec 2.2**). The 6-second audio clip paired with the image stimulus continuously loops until

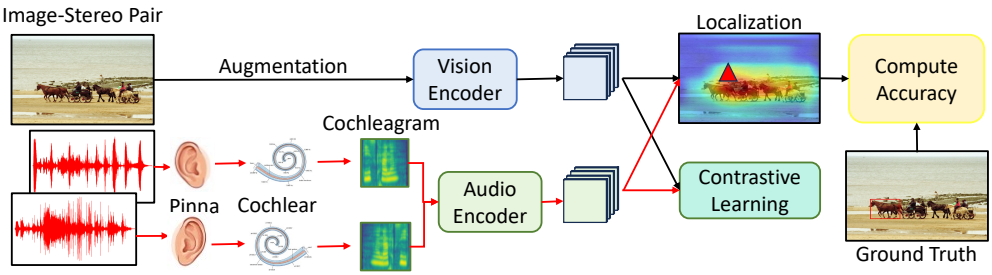


Figure 3: **Overview of our neuroscience-inspired EchoPin model.** EchoPin takes as input a static image paired with a two-channel stereo audio signal. The stereo waveforms are first filtered using the Head-Related Transfer Function (HRTF) to simulate sound filtering by the pinnae, and then converted into cochleagrams to mimic auditory processing in the human cochlea. The audio and visual streams are independently processed through dedicated encoders. During training, semantic alignment between the two modalities is enforced using a contrastive loss applied to paired audio and visual embeddings, while localization alignment is achieved by regressing the predicted sound source location (indicated by a red triangle) from the multimodal feature similarity map to the ground-truth location (red bounding box).

the trial concludes. Participants were instructed to use a computer mouse to click on the perceived location of the sound source within a time limit of 20 seconds, while wearing stereo headphones throughout the experiment. All trials were randomly sampled, and their presentation sequences were shuffled to minimize order effects. If participants failed to respond within the allotted 20 seconds, the trial automatically ended, and the next trial commenced. Instead of relying on the pre-rendered image-audio pairs, we conducted an audio calibration procedure at the start of the experiment to account for individual differences in auditory perception. To achieve this, we implemented real-time stereo audio synthesis in Unity. Following calibration, an audio validation task was conducted to ensure successful calibration and adequate spatial hearing accuracy from the participants. See **Supp Sec. S2 and Supp Fig.S6** for additional details on the human psychophysics experiments.

2.4 Our neuroscience-inspired EchoPin model

We introduce EchoPin, a neuroscience-inspired model for SSL (**Fig. 3**). The model takes image–audio pairs as input and emulates the human auditory periphery to decompose sounds in raw waveforms into frequency components. These representations are then aligned with visual features through a dual-encoder architecture for auditory and visual processing.

In human auditory neuroscience, incoming sound waveforms are directionally shaped by the pinna, head, and torso. These spectral transformations encode elevation-specific notches and interaural differences, which are essential for resolving front–back and vertical ambiguities [1]. For implementation, we use pre-measured Head-Related Transfer Functions (HRTFs) from human ears, developed in Unity simulations and based on the extensive KEMAR dummy head dataset [74]. KEMAR is equipped with microphones in the ear canals to capture how sounds from different directions are filtered by the head and pinnae. The dataset includes left and right ear impulse responses recorded from a Realistic Optimus Pro 7 loudspeaker positioned 1.4 meters from KEMAR, covering 710 spatial positions with elevations from $-40°$ to $+90°$.

Next, the HRTF-filtered time-domain sound waveform is passed through a cochlear-inspired frequency decomposition, converting it into a cochleagram using the PyCochleagram library [74]. The cochleagrams are constructed via Equivalent Rectangular Bandwidth (ERB) filterbanks, capturing the tonotopic and temporal resolution of sound across frequency channels. This representation retains key auditory features, including pitch, timbre, and spatial cues. The resulting 10-second stereo waveform, sampled at 16 kHz, is transformed into cochleagrams, yielding a tensor of size $66 \times 160{,}000 \times 2$, where the dimensions correspond to 66 ERB filters (after truncating 10 high-frequency channels for efficiency), 160k temporal samples, and two binaural channels. The binaural channels are first integrated using 1D convolution kernels to merge information across ears and then fed into the dual-encoder architecture of IS3 [45], allowing separate audio and visual processing streams.

During training, we initialize all weights from the pre-trained IS3 model, except for the first 1D-convolution layer in the audio encoder described above, and then optimize all parameters end-to-end using supervised learning. Two losses in the IS3 [45] model are employed: (i) a

Triplet loss to enforce semantic alignment by pulling matched audio–visual embeddings closer than mismatched ones, and (ii) a CIoU loss to penalize spatial deviation between predicted and ground-truth sounding-object bounding boxes. This combination enables EchoPin to jointly capture what is sounding and where it is located. See **Supp Sec. S3** for extra implementation details.

Model variants of EchoPin. To study the effects of design components in EchoPin, we introduce two model variants: EchoPin-M (Mono) averages the two HRTF-filtered stereo channels into a single time-domain waveform, allowing us to examine the impact of mono versus stereo audio on SSL tasks. EchoPin-S (Stereo) uses the HRTF-filtered stereo waveforms as input, but processes them with standard mel-spectrograms instead of cochleagrams. Comparing these variants with the full EchoPin model allows us to examine how stereo structure captured by the pinnae and frequency-specific cochlear modulation affect SSL performance. See **Tab. 1(b)** and **Sec. 3** for results and discussion.

2.5 Baseline methods and evaluation metrics

We benchmark EchoPin and the state-of-the-art multimodal models, including SSLTI [75], LVS [44], FNAC [43], CAVP [42], AVSegformer [23], IS3 [45], ImageBind [76], and LanguageBind [33], using the same stimuli as in our human psychophysics experiments. While humans can leverage the pinna, head, and torso to encode elevation-specific auditory cues, models lack these physical structures. To ensure fair comparisons across models and with human participants, all AudioCOCO audios are HRTF-filtered to approximate human auditory processing, and these filtered sounds are used for all model evaluations. In the main text, we provide brief overviews of IS3 [45] and a random baseline, and report their performance alongside our proposed EchoPin model. Detailed descriptions of the other models and their extended experimental results are provided in **Supp. Sec. S3**.

IS3 [45] is a dual-stream architecture with 2D CNNs, which processes visual and monaural auditory inputs separately using dedicated encoders before fusing the features for contrastive learning during training. IS3 also includes an Intersection-over-Union (IoU) loss and a semantic alignment loss to improve localization accuracy during supervised training. The model is trained on the FlickrSoundNet [58] and VGG-Sound [49] datasets. **Random** is a chance model that randomly selects a location on the image as the predicted sound source location. It serves as a lower bound for SSL without using any audio-visual information.

Predicting target sound locations. For IS3, EchoPin, and other 2D CNN–based models, feature maps from the final layers of the visual and audio encoders are extracted, and cosine similarity is computed between them to generate a similarity heatmap. The predicted sound location is taken as the point with the highest activation on this heatmap. See **Supp. Fig. S7** for an illustration of how transformer-based baselines predict target sound source locations.

Evaluation metrics. To disentangle spatial localization from semantic alignment between visual and audio modalities, we define two metrics: **Audio Accuracy (A-Acc)** measures whether the model or human localizes the true sound source regardless of matching semantics. A-Acc = 1 if the peak activation falls within the bounding box of the sounding object; 0 otherwise. **Vision Accuracy (V-Acc)** measures alignment with visual semantics. V-Acc = 1 if the peak activation falls within any object of the correct category, even if it might not be the actual sound source, such as in MultiInstLoc conditions. In VOnly condition, V-Acc = 1 if the activation overlaps with any object from the 12 sound-emitting categories in AudioCOCO.

To robustly evaluate model performance, we consider three complementary factors that could influence results in **Supp Sec. S5**. First, A-Acc can be biased by object size, so we introduce a chance-corrected A-Acc (**Supp Tab. S5**; **Supp Tab. S1**). Second, human clicks may fall near but not inside the target, so we treat clicks within a thresholded radius as correct in **Supp. Tab. S2**. None of these metric variants alter the conclusions. Finally, we evaluate the predicted target object bounding boxes by all the models using corrected Intersection over Union (cIoU, [77]), where EchoPin continues to outperform other baselines (**Supp Sec. S5** and **Supp Tab. S3**).

3 Results

We report results from both human participants and AI models across all experimental conditions and object sizes. For brevity, the main text focuses on comparisons between humans and the two

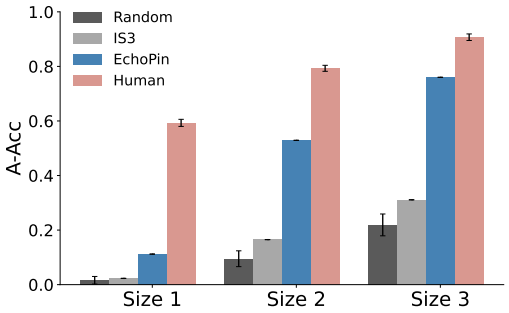


Figure 4: **Object size matters for humans and AI models in the congruent condition**. Accuracy increases with object sizes for both humans and AI models, with humans outperforming AI models, especially for small targets. Here and in subsequent figures, error bars represent Standard Error Mean (SEM).

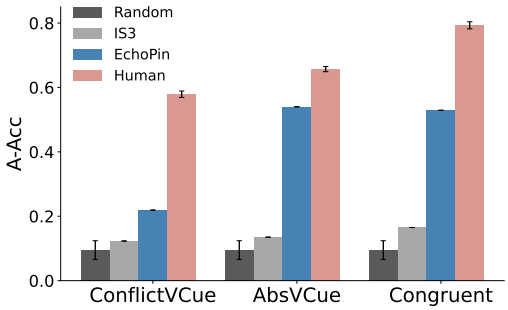


Figure 5: **ConflictVCue and AbsVCue hurt SSL performance for both humans and AI models**. The accuracy of humans and AI models under ConflicVcue and AbsVcue conditions is shown for object size 2, with results from the congruent condition included for comparison. See **Supp Tab.** S4 for the results on object sizes 1 and 3.

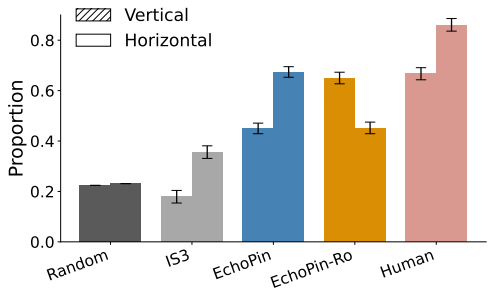


Figure 6: **EchoPin shows human-like horizontal-vertical asymmetry in SSL accuracy, while other models do not.** We report the proportion of trials (object size 2) where the predicted sound location falls within 6 degrees of visual angle from the ground truth, separately along the horizontal (textured bars) and vertical (plain bars) directions.

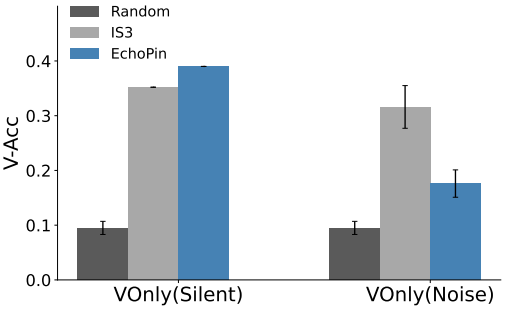


Figure 7: **AI models show a bias towards objects that emit sounds, even in the absence of sound or in the presence of noisy sound.** The V-Acc of AI models is presented in the VOnly condition, where either no sound or only noisy sound is present. Despite the lack of auditory cues, V-Acc of AI models on object size 2 remains higher than random (dark gray).

best-performing models, IS3 and EchoPin. Additional qualitative analyses for all other models are presented in **Supp. Tab.** S4.

Object size matters for humans and AI. As illustrated in **Fig.** 4, human A-Acc steadily increases with object size, suggesting that larger sounding objects are easier for humans to localize. EchoPin exhibits a similar trend, with A-Acc rising substantially from 11.2% to 76.0% as object size grows. Across all object sizes, EchoPin consistently outperforms IS3 and the other models. This difference likely arises because IS3 is trained on datasets biased toward large, centered objects, whereas AudioCOCO provides a wider range of object sizes and spatial locations. While both humans and AI models perform above chance for large objects, only humans and EchoPin maintain strong, consistent performance across all object sizes, including smaller ones—performance that IS3 fails to achieve.

Conflict cues harm more than the lack of cues for AI. As shown in Fig. 5, both the ConflictVCue and AbsVCue conditions impair SSL performance in humans compared to the congruent conditions. While humans and EchoPin still perform significantly above chance, IS3 drops to near-chance levels under the ConflictVCue condition. This suggests that humans and EchoPin are more robust in SSL tasks involving conflicting or missing visual cues, whereas IS3 relies heavily on visual information. Notably, unlike humans, EchoPin shows a greater performance decline when visual cues are conflicting but not when they are absent, indicating that it is more easily misled by incongruent visual information yet remains stable in the absence of such cues.

AI fails when there are only auditory cues but humans can. As shown in **Supp. Fig.** S8, humans achieve above-chance A-Acc even without visual information (e.g., gray or Gaussian noise

(a) Multi-Instance Localization

Acc(%)		Rand	IS3	CAVP	AVSeg	EchoPin	Human
A-Acc	**Size1**	1.6	<u>4.8</u>	2.9	2.5	4.5	**25.7**
	Size2	9.1	7.9	7.5	7.3	<u>24.1</u>	**36.4**
	Size3	21.3	22.4	20.4	20.2	<u>47.1</u>	**38.6**
V-Acc	**Size1**	8.4	11.9	10.5	11.2	<u>37.5</u>	**60.9**
	Size2	17.8	24.1	23.0	23.7	<u>53.8</u>	**82.8**
	Size3	26.2	40.9	39.5	40.9	<u>64.2</u>	**89.1**

(b) Overall Performance

Acc(%)		Rand	IS3	EchoPin-M/S	EchoPin
Mono	**Size1**	1.6	3.0	3.6	-
	Size2	9.4	13.9	15.8	-
	Size3	19.8	28.7	31.4	-
Stereo	**Size1**	1.6	-	<u>5.3</u>	**9.7**
	Size2	9.4	-	<u>17.0</u>	**31.3**
	Size3	19.8	-	<u>35.2</u>	**47.6**

Table 1: **Multi-instance SSL remains a challenging task for both humans and AI models.** The table (a) on the left summarizes the audio and visual localization accuracy of humans and AI models under the multi-instance condition across all object sizes. **For AI models, both input data quality and the use of stereo audio substantially impact SSL performance.** Table (b) on the right summarizes the average A-Acc across the Congruent, ConflictVcue, AbsVcue, and AOnly conditions for models trained with either mono or stereo audio. The second-to-last column shows the results of EchoPin-M (Rows 1–3) and EchoPin-S (Rows 4–6). See **Sec. 2.4** for details on these variants. (–) indicates that the results are not applicable due to the model configurations. In both tables, best is in bold and the second best is underlined.

backgrounds), confirming that they can perform SSL based solely on auditory cues. However, their performance remains lower than in the congruent condition, indicating that congruent visual information facilitates SSL. Similar trends are observed in IS3 and EchoPin, with EchoPin consistently outperforming IS3 under both AOnly conditions (gray or Gaussian background). Interestingly, both humans and EchoPin perform better with gray than Gaussian backgrounds, suggesting that incongruent visual noise can distract attention and interfere with SSL, whereas a neutral (gray) background minimizes interference.

EchoPin shows human-like asymmetry in auditory spatial precision. In neuroscience, auditory spatial precision is known to exhibit a horizontal–vertical asymmetry: humans localize sounds more accurately along the azimuth (horizontal) than the elevation (vertical) axis [71]. To quantify this effect, we measured the proportion of trials where predictions fell within six degrees of visual angle from ground-truth locations, separately for horizontal and vertical dimensions (**Fig. 6**). As expected, humans showed a strong horizontal advantage, localizing targets in 86.1% of trials horizontally but only 66.7% vertically. Remarkably, EchoPin exhibited a similar asymmetry pattern despite being trained without any human behavioral data. In contrast, IS3, which relies solely on monaural Mel-spectrograms, also showed asymmetry but to a much lesser degree. Although both models have benefited from spatial filtering by the pinnae, the observed asymmetry in EchoPin arises from its biologically grounded auditory frequency decomposition in the cochlea and its stereo audio perception. To further validate this, we introduced an EchoPin variant, EchoPin-Ro, in which the interaural axis was rotated by 90 degrees in Unity, effectively simulating vertically aligned ears. When audio was re-rendered under this configuration, the model's asymmetry was reversed, confirming the structural origin of this effect.

AI biases toward sound-emitting objects. Previous studies [42, 45] show that AI models often exploit visual shortcuts by favoring large or centered objects. To examine additional behavioral biases, we evaluate models under the VOnly condition. Without meaningful sound, any object could plausibly be the sound source. As shown in **Fig. 7**, models tend to localize sounds to vocal objects (e.g., people, animals) rather than irrelevant ones (e.g., sky, trees), resulting in above-chance V-Acc. This indicates that models encode prior knowledge of which objects typically produce sound. Furthermore, for all the models, Gaussian noise audio input leads to lower V-Acc than the absence of audio, indicating that even noisy audio can reduce the models' over-reliance on visual signals.

Multi-instance SSL remains challenging for humans and AI. As shown in **Tab. 1a**, V-Acc is high for both EchoPin and humans, indicating strong alignment between audio and visual semantics. This allows them to use audio cues to identify all visual objects with matching semantics. However, in scenes with multiple object instances, successful localization requires fine-grained SSL, beyond semantic matching. In these cases, A-Acc drops for both humans and EchoPin. Compared to the Congruent condition (with a single sounding object), multi-instance SSL is notably harder, as it demands precise spatial disambiguation. Despite this, humans still perform above chance, especially for small objects. Similarly, EchoPin achieves high A-Acc on object sizes 2 and 3. However, it still

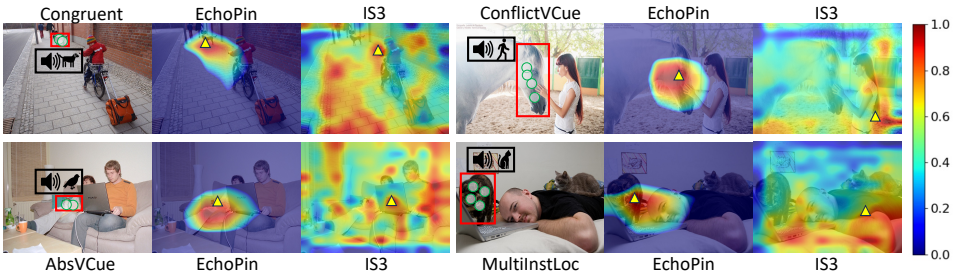


Figure 8: **IS3 struggles to localize sound sources, whereas humans and EchoPin perform well across all four experimental conditions.** The leftmost images (Columns 1 and 4) show the correct localization results made by human participants. Red boxes mark the ground truth sound source locations, while green circles indicate mouse click responses from multiple participants. The middle columns (2 and 5) and rightmost columns (3 and 6) display heatmaps predicted by EchoPin and IS3, respectively. Yellow triangles on the heatmaps denote the predicted sound source locations from each model. See the colorbar for the activation values of the heatmaps.

lags behind human performance, with the gap more pronounced in small object size. Despite this, both EchoPin and humans still perform above chance, especially for small objects. This demonstrates an ability to localize sounds at fine spatial resolution for both humans and EchoPin. Among all models, EchoPin performs best. It even outperforms strong baselines like CAVP and AVSeg, which are trained on large-scale, standard audio-visual datasets.

Training data quality and stereo input are important for SSL. We report the average A-Acc across all object sizes for Random, IS3, and the EchoPin variants under four experimental conditions in **Tab. 1b**. EchoPin consistently achieves the highest performance across all conditions. Notably, **EchoPin-M** surpasses IS3 despite using fewer fine-tuning examples, underscoring the importance of high-quality training data. Moreover, **EchoPin-S** further improves over **EchoPin-M**, highlighting the advantage of incorporating human-like stereo configurations for spatial localization.

We further visualize the predicted sound source locations from human participants, IS3, and EchoPin in **Fig. 8**. EchoPin localizes sound sources more accurately and often aligns closely with human judgments. In contrast, IS3 struggles with small or peripheral targets—for instance, it fails to localize a dog in the top-left corner under the Congruent condition and frequently misattributes sounds to other vocal objects (e.g., person, motorbike). Although EchoPin markedly improves SSL robustness, it is still inferior to human performance in complex scenes. Failure cases and additional comparisons with other baselines, such as CAVP, are provided in **Supp. Sec. S4**, **Supp. Fig. S9**, and **Supp. Fig. S10**.

4 Discussion

We systematically and quantitatively examine modality biases and conflicts in SSL across humans and AI models. Our study covers six audiovisual conditions, including congruent cues, conflicting signals, and cases with missing audio or visual input in natural scenes. Human listeners show strong robustness: although conflicting or absent visual cues reduce performance, they can still accurately localize even small sound sources under challenging or multi-instance conditions, and even in the absence of visual input. In contrast, current multimodal AI models rely heavily on vision—misattributing sounds to large, centered, and salient objects, and suffering steep performance drops when visual cues are removed. Conflicting visuals further degrade their accuracy, with most models performing near chance for small or visually absent sound sources.

We identify two primary causes of these limitations: (1) low-quality, visually biased audiovisual datasets, and (2) monaural audio inputs lacking spatial fidelity. To overcome these issues, we curate AudioCOCO, a high-quality dataset built through rigorous filtering and 3D physical simulation. By integrating depth maps, HRTF-based filtering that mimics pinna effects, cochlear-inspired frequency decomposition and modulation, and physically grounded 3D sound propagation, AudioCOCO produces realistic, spatialized stereo audio aligned with human auditory processing. Building on this, we introduce EchoPin, a neuroscience-inspired model trained on AudioCOCO. Despite fewer training examples, EchoPin surpasses state-of-the-art models across all conditions and exhibits human-like localization biases, such as stronger precision along the horizontal plane. Ablation studies confirm the importance of both high-quality datasets and stereo auditory input for capturing spatial cues. This work underscores the value of designing models and datasets that respect the physical constraints of

sensory systems. Future directions include scaling AudioCOCO to incorporate temporal dynamics from videos and improving realism in simulated sound rendering, such as the effect of refraction.

Acknowledgements

This research is supported by the National Research Foundation, Singapore under its NRFF award NRF-NRFF15-2023-0001 and Mengmi Zhang's Startup Grant from Nanyang Technological University, Singapore. We would also like to thank Qing Lin and Shuangpeng Han for their valuable advice and feedback on the project.

References

[1] Kiki van der Heijden, Josef P Rauschecker, Beatrice de Gelder, and Elia Formisano. Cortical mechanisms of spatial hearing. *Nature Reviews Neuroscience*, 20(10):609–623, 2019. 1, 3, 6

[2] Simon Jenni, Alexander Black, and John Collomosse. Audio-visual contrastive learning with temporal self-supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 7996–8004, 2023. 1, 2

[3] Uta Noppeney. Perceptual inference, learning, and attention in a multisensory world. *Annual review of neuroscience*, 44(1):449–473, 2021. 1

[4] Jianwu Fang, Jiahuan Qiao, Jie Bai, Hongkai Yu, and Jianru Xue. Traffic accident detection via self-supervised consistency learning in driving scenarios. *IEEE Transactions on Intelligent Transportation Systems*, 23(7):9601–9614, 2022. 2

[5] George Dimas, Eirini Cholopoulou, and Dimitris K Iakovidis. Self-supervised soft obstacle detection for safe navigation of visually impaired people. In *2021 IEEE International Conference on Imaging Systems and Techniques (IST)*, pages 1–6. IEEE, 2021. 2

[6] Hongpeng Chen, Shufei Li, Junming Fan, Anqing Duan, Chenguang Yang, David Navarro-Alarcon, and Pai Zheng. Human-in-the-loop robot learning for smart manufacturing: A human-centric perspective. *IEEE Transactions on Automation Science and Engineering*, 2025. 2

[7] Aaqib Saeed, Tanir Ozcelebi, and Johan Lukkien. Multi-task self-supervised learning for human activity detection. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(2):1–30, 2019. 2

[8] Daniel Dworakowski, Angus Fung, and Goldie Nejat. Robots understanding contextual information in human-centered environments using weakly supervised mask data distillation. *International Journal of Computer Vision*, 131(2):407–430, 2023. 2

[9] Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19108–19118, 2022. 2

[10] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9308–9316, 2019. 2

[11] Mariana-Iuliana Georgescu, Eduardo Fonseca, Radu Tudor Ionescu, Mario Lucic, Cordelia Schmid, and Anurag Arnab. Audiovisual masked autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16144–16154, 2023. 2

[12] Po-Yao Huang, Vasu Sharma, Hu Xu, Chaitanya Ryali, Yanghao Li, Shang-Wen Li, Gargi Ghosh, Jitendra Malik, Christoph Feichtenhofer, et al. Mavil: Masked audio-video learners. *Advances in Neural Information Processing Systems*, 36:20371–20393, 2023. 2

[13] Yanhao Jia, Xinyi Wu, Hao Li, Qinglin Zhang, Yuxiao Hu, Shuai Zhao, and Wenqi Fan. Uni-retrieval: A multi-style retrieval framework for stem's education, 2025. 2

[14] Sizhe Li, Yapeng Tian, and Chenliang Xu. Space-time memory network for sounding object localization in videos. *arXiv preprint arXiv:2111.05526*, 2021. 2

[15] Hao Li, Yanhao Jia, Jin Peng, Zesen Cheng, Kehan Li, Jialu Sui, Chang Liu, and Li Yuan. Freestyleret: Retrieving images from style-diversified queries. In *Computer Vision – ECCV 2024*, pages 258–274, Cham, 2025. Springer Nature Switzerland. 2

[16] 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*, 2025. 2

[17] 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*, pages 13117–13127, 2024. 2

[18] Shentong Mo and Pedro Morgado. A closer look at weakly-supervised audio-visual source localization. *Advances in Neural Information Processing Systems*, 35:37524–37536, 2022. 2

[19] Rui Qian, Di Hu, Heinrich Dinkel, Mengyue Wu, Ning Xu, and Weiyao Lin. Multiple sound sources localization from coarse to fine. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 292–308. Springer, 2020. 2

[20] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 2

[21] Tianhong Li, Peng Cao, Yuan Yuan, Lijie Fan, Yuzhe Yang, Rogerio S Feris, Piotr Indyk, and Dina Katabi. Targeted supervised contrastive learning for long-tailed recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6918–6928, 2022. 2

[22] Yan-Bo Lin, Hung-Yu Tseng, Hsin-Ying Lee, Yen-Yu Lin, and Ming-Hsuan Yang. Unsupervised sound localization via iterative contrastive learning. *Computer Vision and Image Understanding*, 227:103602, 2023. 2

[23] Shengyi Gao, Zhe Chen, Guo Chen, Wenhai Wang, and Tong Lu. Avsegformer: Audio-visual segmentation with transformer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 12155–12163, 2024. 2, 7, 5, 6

[24] Kexin Li, Zongxin Yang, Lei Chen, Yi Yang, and Jun Xiao. Catr: Combinatorial-dependence audio-queried transformer for audio-visual video segmentation. In *Proceedings of the 31st ACM international conference on multimedia*, pages 1485–1494, 2023. 2

[25] Shentong Mo and Pedro Morgado. Localizing visual sounds the easy way. In *European Conference on Computer Vision*, pages 218–234. Springer, 2022. 2

[26] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1029–1038, 2022. 2

[27] Arda Senocak, Hyeonggon Ryu, Junsik Kim, and In So Kweon. Less can be more: Sound source localization with a classification model. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3308–3317, 2022. 2

[28] Yuan Gong, Andrew Rouditchenko, Alexander H Liu, David Harwath, Leonid Karlinsky, Hilde Kuehne, and James Glass. Contrastive audio-visual masked autoencoder. *arXiv preprint arXiv:2210.07839*, 2022. 2

[29] Simon Haykin and Zhe Chen. The cocktail party problem. *Neural computation*, 17(9):1875–1902, 2005. 2, 5

[30] Peiwen Sun, Honggang Zhang, and Di Hu. Unveiling and mitigating bias in audio visual segmentation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7259–7268, 2024. 2

[31] Yejiang Wang, Yuhai Zhao, Zhengkui Wang, and Meixia Wang. Robust self-supervised multi-instance learning with structure awareness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 10218–10225, 2023. 2

[32] Sangho Lee, Jiwan Chung, Youngjae Yu, Gunhee Kim, Thomas Breuel, Gal Chechik, and Yale Song. Acav100m: Automatic curation of large-scale datasets for audio-visual video representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10274–10284, 2021. 2

[33] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, Wang HongFa, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Cai Wan Zhang, Zhifeng Li, Wei Liu, and Li Yuan. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment, 2023. 2, 7, 5

[34] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127:302–321, 2019. 2

[35] Jinxing Zhou, Xuyang Shen, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, et al. Audio-visual segmentation with semantics. *International Journal of Computer Vision*, pages 1–21, 2024. 2, 5, 6

[36] Yaoting Wang, Peiwen Sun, Dongzhan Zhou, Guangyao Li, Honggang Zhang, and Di Hu. Ref-avs: Refer and segment objects in audio-visual scenes. *IEEE European Conference on Computer Vision (ECCV)*, 2024. 2

[37] 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. 2

[38] Lucas Goncalves, Prashant Mathur, Chandrashekhar Lavania, Metehan Cekic, Marcello Federico, and Kyu J. Han. Peavs: Perceptual evaluation of audio-visual synchrony grounded in viewers' opinion scores, 2024. 2

[39] Juan F Montesinos, Venkatesh S Kadandale, and Gloria Haro. A cappella: Audio-visual singing voice separation. In *32nd British Machine Vision Conference, BMVC 2021*, 2021. 2

[40] Ethan Manilow, Gordon Wichern, Prem Seetharaman, and Jonathan Le Roux. Cutting music source separation some Slakh: A dataset to study the impact of training data quality and quantity. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2019. 2

[41] Justin Salamon, Duncan MacConnell, Mark Cartwright, Peter Li, and Juan Pablo Bello. Scaper: A library for soundscape synthesis and augmentation. In *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 344–348. IEEE, 2017. 2

[42] Yuanhong Chen, Yuyuan Liu, Hu Wang, Fengbei Liu, Chong Wang, Helen Frazer, and Gustavo Carneiro. Unraveling instance associations: A closer look for audio-visual segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26497–26507, June 2024. 2, 7, 9, 5, 6

[43] 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*, pages 6420–6429, 2023. 2, 7, 4

[44] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Localizing visual sounds the hard way. In *CVPR*, 2021. 2, 7, 4

[45] 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*, pages 7777–7787, 2023. 2, 6, 7, 9, 4

[46] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011. 2

[47] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 3, 1

[48] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3

[49] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020. 3, 4, 7, 2

[50] Eduardo Fonseca, Manoj Plakal, Daniel PW Ellis, Frederic Font, Xavier Favory, and Xavier Serra. Learning sound event classifiers from web audio with noisy labels. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 21–25. IEEE, 2019. 3, 4

[51] Nicolas Turpault, Romain Serizel, Ankit Parag Shah, and Justin Salamon. Sound event detection in domestic environments with weakly labeled data and soundscape synthesis. In *Workshop on Detection and Classification of Acoustic Scenes and Events*, 2019. 3

[52] Eric Guizzo, Christian Marinoni, Marco Pennese, Xinlei Ren, Xiguang Zheng, Chen Zhang, Bruno Masiero, Aurelio Uncini, and Danilo Comminiello. L3das22 challenge: Learning 3d audio sources in a real office environment. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 9186–9190. IEEE, 2022. 3

[53] Yuan Gong, Jin Yu, and James Glass. Vocalsound: A dataset for improving human vocal sounds recognition. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 151–155, 2022. 3

[54] Archontis Politis, Kazuki Shimada, Parthasaarathy Sudarsanam, Sharath Adavanne, Daniel Krause, Yuichiro Koyama, Naoya Takahashi, Shusuke Takahashi, Yuki Mitsufuji, and Tuomas Virtanen. Starss22: A dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events. *arXiv preprint arXiv:2206.01948*, 2022. 3

[55] Heinrich W Löllmann, Christine Evers, Alexander Schmidt, Heinrich Mellmann, Hendrik Barfuss, Patrick A Naylor, and Walter Kellermann. The locata challenge data corpus for acoustic source localization and tracking. In *2018 IEEE 10th sensor array and multichannel signal processing workshop (SAM)*, pages 410–414. IEEE, 2018. 3

[56] Valentina Sanguineti, Pietro Morerio, Alessio Del Bue, and Vittorio Murino. Audio-visual localization by synthetic acoustic image generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2523–2531, 2021. 3

[57] Eduardo Fonseca, Manoj Plakal, Daniel PW Ellis, Frederic Font, Xavier Favory, and Xavier Serra. Learning sound event classifiers from web audio with noisy labels. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 21–25. IEEE, 2019. 3

[58] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. *Advances in neural information processing systems*, 29, 2016. 3, 7, 4

[59] Josh H McDermott and Eero P Simoncelli. Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis. *Neuron*, 71(5):926–940, 2011. 3

[60] Junwoo Park, Youngwoo Cho, Gyuhyeon Sim, Hojoon Lee, and Jaegul Choo. Enemy spotted: In-game gun sound dataset for gunshot classification and localization. In *2022 IEEE Conference on Games (CoG)*, pages 56–63. IEEE, 2022. 3

[61] Hang Zhou, Xudong Xu, Dahua Lin, Xiaogang Wang, and Ziwei Liu. Sep-stereo: Visually guided stereophonic audio generation by associating source separation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 52–69. Springer, 2020. 3

[62] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE international conference on computer vision*, pages 609–617, 2017. 3

[63] Kevin JP Woods, Max H Siegel, James Traer, and Josh H McDermott. Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, & Psychophysics*, 79:2064–2072, 2017. 3

[64] Muhammad SA Zilany, Ian C Bruce, Paul C Nelson, and Laurel H Carney. A phenomenological model of the synapse between the inner hair cell and auditory nerve: long-term adaptation with power-law dynamics. *The Journal of the Acoustical Society of America*, 126(5):2390–2412, 2009. 3

[65] Richard McWalter and Josh H McDermott. Illusory sound texture reveals multi-second statistical completion in auditory scene analysis. *Nature communications*, 10(1):5096, 2019. 3

[66] Mark R Saddler, Ray Gonzalez, and Josh H McDermott. Deep neural network models reveal interplay of peripheral coding and stimulus statistics in pitch perception. *Nature communications*, 12(1):7278, 2021. 3

[67] Mark R Saddler and Josh H McDermott. Models optimized for real-world tasks reveal the task-dependent necessity of precise temporal coding in hearing. *Nature Communications*, 15(1):10590, 2024. 3

[68] Kazuki Shimada, Archontis Politis, Iran R Roman, Parthasaarathy Sudarsanam, David Diaz-Guerra, Ruchi Pandey, Kengo Uchida, Yuichiro Koyama, Naoya Takahashi, Takashi Shibuya, et al. Stereo sound event localization and detection with onscreen/offscreen classification. *arXiv preprint arXiv:2507.12042*, 2025. 3

[69] Hogeon Yu. A two-step learning framework for enhancing sound event localization and detection. *arXiv preprint arXiv:2507.22322*, 2025. 3

[70] Da Mu, Zhicheng Zhang, Haobo Yue, Zehao Wang, Jin Tang, and Jianqin Yin. Seld-mamba: Selective state-space model for sound event localization and detection with source distance estimation. *arXiv preprint arXiv:2408.05057*, 2024. 3

[71] Andrew Francl and Josh H McDermott. Deep neural network models of sound localization reveal how perception is adapted to real-world environments. *Nature human behaviour*, 6(1):111–133, 2022. 3, 9

[72] Philipp Bomatter, Mengmi Zhang, Dimitar Karev, Spandan Madan, Claire Tseng, and Gabriel Kreiman. When pigs fly: Contextual reasoning in synthetic and natural scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 255–264, 2021. 4

[73] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024. 4

[74] Bill Gardner, Keith Martin, et al. Hrft measurements of a kemar dummy-head microphone, 1994. 6

[75] 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*, pages 3742–3753, 2022. 7, 4

[76] 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 *CVPR*, 2023. 7, 5

[77] Zhaohui Zheng, Ping Wang, Dongwei Ren, Wei Liu, Rongguang Ye, Qinghua Hu, and Wangmeng Zuo. Enhancing geometric factors in model learning and inference for object detection and instance segmentation. *IEEE transactions on cybernetics*, 52(8):8574–8586, 2021. 7

[78] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020. 2

[79] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 4

[80] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational visual media*, 8(3):415–424, 2022. 5

[81] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *ICASSP 2017 - 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 131–135. IEEE, 2017. 5

[82] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021. 5

[83] Yuan Gong, Yu-An Chung, and James Glass. AST: Audio Spectrogram Transformer. In *Proc. Interspeech 2021*, pages 571–575, 2021. 5

[84] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 5

[85] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 5

[86] 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*, pages 2818–2829, 2023. 5

[87] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017. 5

[88] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 5

[89] Diederik Kinga, Jimmy Ba Adam, et al. A method for stochastic optimization. In *International conference on learning representations (ICLR)*, volume 5. California;, 2015. 6

[90] 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*, pages 4358–4366, 2018. 7

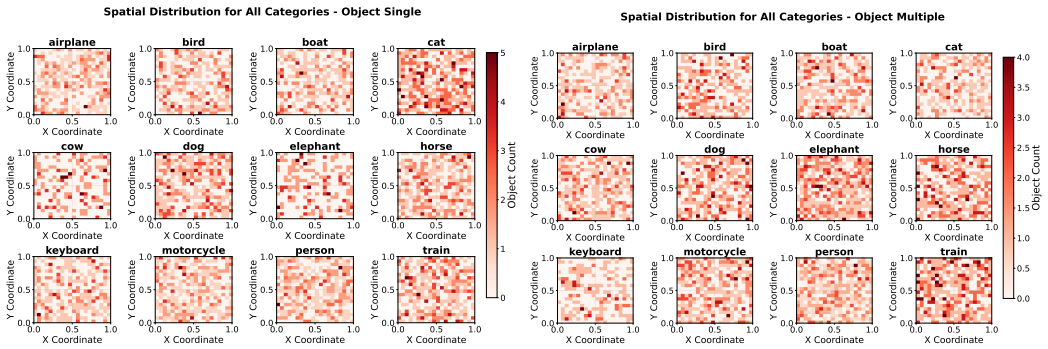
Supplementary Material

In the supplementary material, we provide additional information of our AudioCOCO dataset in **Sec.**S1 and human experiment setup in **Sec.**S2. Moreover, we include extra experimental results, an ablation study, and visualization results for predicted position bias, which are discussed in **Sec.**S4 and **Sec.**S5.

S1 Details about the AudioCOCO dataset

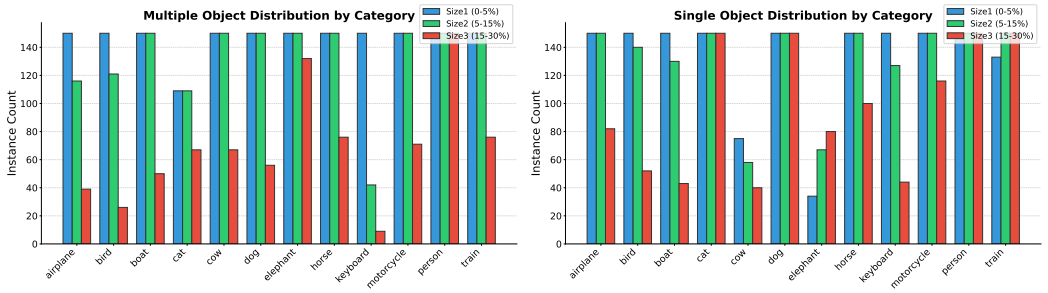
More details on image selection. The MSCOCO 2014 dataset [47] offers annotations for 80 object categories, including both bounding boxes and segmentation masks. To ensure a fair comparison—especially given that some model backbones are pre-trained on MSCOCO—we utilized the MSCOCO 2014 validation set rather than the training set. From this, we selected 12 common vocal categories frequently encountered in daily life: person, motorbike, train, boat, elephant, bird, cat, dog, horse, sheep, cow, and keyboard. A total of 29,737 images containing at least one vocal object were extracted to form the pool of image candidates for our dataset.

To investigate spatial distribution, we visualized the occurrence frequency of these categories across the 29,737 images in **Fig.** S1b and **Fig.** S1a, grouping instances by three object area size bins defined in the main draft: object size1 (0-5%), size2 (5-15%), and size3 (15-30%). For smaller object sizes, category locations were relatively uniformly distributed across the image space. However, for larger objects, some categories—such as bus, train, and truck—exhibited a strong center bias. To correct for this, we filtered the dataset to ensure that no position in the final heatmap exceeded 50% in frequency.



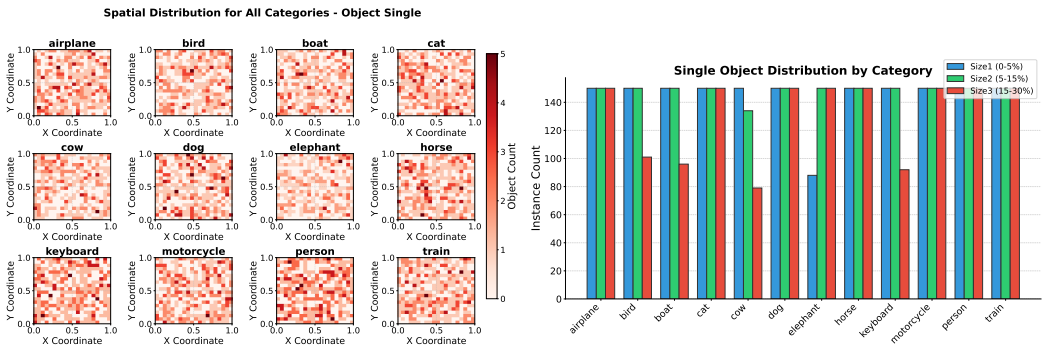
(a) The spatial distribution of single objects by category. (b) The spatial distribution of multi-objects by category.

Figure S1: The visualization results of images' spatial distributions in the AudioCOCO test set are presented for each object category, where the color bar indicates the frequency of object occurrences at each spatial location within the images. Darker regions correspond to higher frequencies.



(a) The count distribution of multi-objects by category. (b) The count distribution of single objects by category.

Figure S2: The visualization results of the image count distributions in the AudioCOCO test set are presented, where blue, green, and red represent object size1, size2, and size3, respectively. This visualization illustrates how instances are distributed across different object size categories, highlighting the dataset's balanced distribution in terms of object sizes.



(a) The spatial distribution of single objects by category for AudioCOCO's training set.

(b) The count distribution of single objects by category for AudioCOCO's training set.

Figure S3: The visualization results of the image count distribution and spatial distribution for the AudioCOCO training set. For the count distribution, object sizes are color-coded: blue for Size1, green for Size2, and red for Size3. This plot illustrates the number of object instances across different size categories, confirming that the dataset maintains a balanced distribution in terms of object size. For the spatial distribution, the heatmap reflects the frequency of object occurrences at each spatial location in the images, with the color bar indicating frequency. Darker regions correspond to areas of higher object density, highlighting positional biases or spread across the dataset.

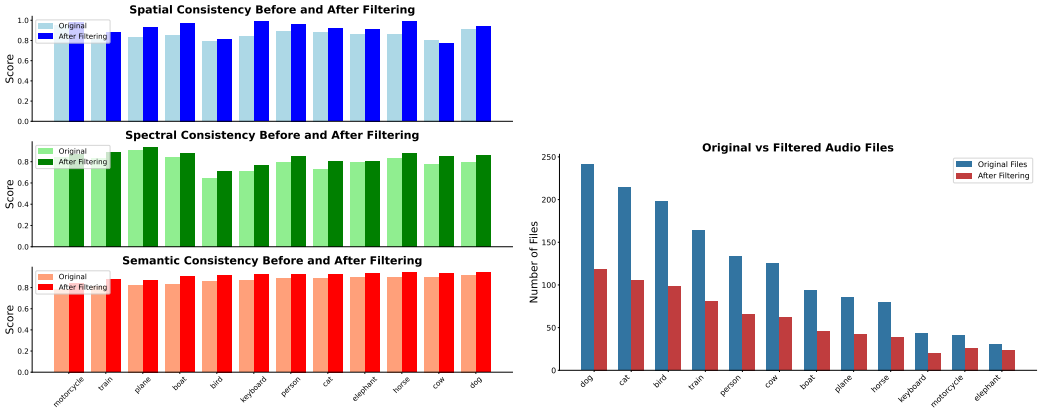
Additionally, given that the image distribution over all object categories in COCO is not uniform, we performed random sampling to cap the number of instances to 150 per category per object size bin, promoting balanced representation within the AudioCOCO dataset. The final distributions after filtering are shown in **Fig.** S2a, **Fig.** S2b, and **Fig.** S3, demonstrating that AudioCOCO achieves a balanced dataset across object area sizes, object center positions, and categories.

More details on audio selection. To ensure that only high-quality, semantically aligned clips are retained, we apply the following three filtering steps to the videos from VGGSound [49]. First, we introduce Semantic Consistency (SeC) and select audios that are representative of the semantic object categories. Specifically, we take all audio files belonging to the same semantic category, extract audio features from the last layer of the Wav2vec [78] model, compute their pari-wise cosine similarities based on these features, and retain the top 80% audios with the highest cosine similarity scores. Second, to ensure the audio is free from noise or interference from unrelated sources, we introduce Mel-Spectrogram Similarity (MSS) as a filtering criterion within each category. For each audio clip, we compute its Mel spectrogram, average the frequency magnitudes over time, and apply a logarithmic transformation to compress high-frequency components—yielding a compact representation of the audio's overall spectral structure. We then calculate the cosine similarity between these representations and retain the top 65% of audio clips based on this MSS metric, following the SeC criteria. Third, to ensure the stereo audio corresponds to the sounding object being centered in the video frame, we introduce Spatial Consistency (SpC) for each audio-image pair. We compute the Spearman correlation between the left and right audio channels; a high correlation suggests the sound source is centrally located, producing similar waveforms in both channels. We retain the top 50% of audio clips based on this metric from the MSS-filtered set.

For audio selection, we also enforce a minimum threshold of 20 audio clips per category. If any filtering step results in a category falling below this threshold, that specific step is discarded, and the previous filtering output is retained as the final result to ensure sufficient data coverage across all categories. The final distributions after filtering are shown in **Fig.** S4.

S2 Details of human psychophysics experiments

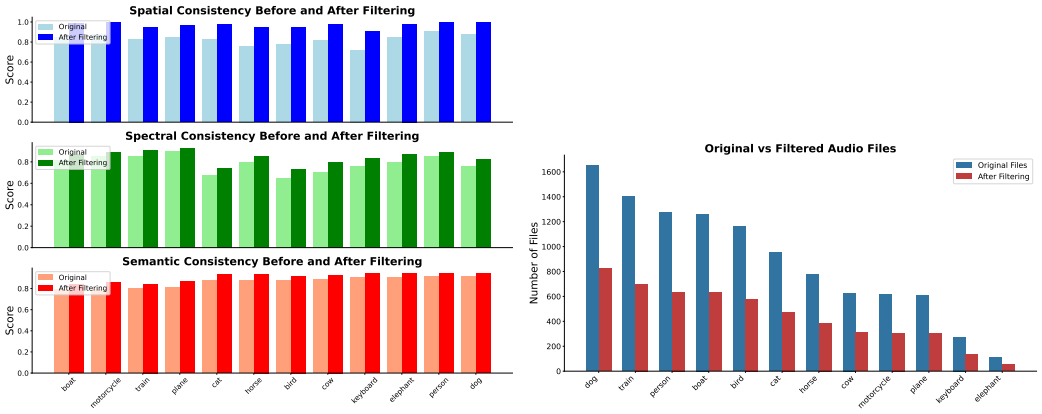
All the experiments are conducted with the subjects' informed consent and according to protocols approved by the Institutional Review Board of our institution. Participants were instructed to use a computer mouse to click on the perceived location of the sound source within a time limit of 20 seconds, while wearing stereo headphones (SENNHEISER MOMENTUM 4 with active noise cancellation) throughout the experiment.



(a) The three stages of audio filtering result for AudioCOCO test set.

(b) The comparison between original audio counts and filtering audio counts for AudioCOCO test set.

Figure S4: The visualization results of audio statistics in the AudioCOCO test set demonstrate that the audio quality across most categories has significantly improved following the filtering process. Additionally, we ensured a balanced count distribution among all categories. These outcomes highlight the effectiveness of our selection and refinement strategy in enhancing both the semantic consistency and acoustic quality of the dataset.



(a) The three stages of audio filtering result for AudioCOCO training set.

(b) The comparison between original audio counts and filtering audio counts for AudioCOCO training set.

Figure S5: The visualization results of audio statistics in the AudioCOCO training set demonstrate that the audio quality across most categories has significantly improved following the filtering process. Additionally, we ensured a balanced count distribution among all categories. These outcomes highlight the effectiveness of our selection and refinement strategy in enhancing both the semantic consistency and acoustic quality of the dataset.

Audio calibration. Instead of relying on the pre-rendered image-audio pairs, we conducted an audio calibration procedure at the start of the experiment to account for individual differences in height and auditory perception. To achieve this, we implemented real-time stereo audio synthesis in Unity, which communicates with MATLAB (hosting the human behavioral experiment) via TCP connections. The measured delay for real-time stereo sound synthesis and presentation is within 500 milliseconds, ensuring seamless interaction between the two systems.

During calibration, participants were presented with a white dot at a random location on the screen, accompanied by an audio clip spatially rendered at the dot's position. A white cross at the center of the screen served as a spatial reference (see **Supp Fig.**S6). Using the keyboard's arrow keys, participants adjusted the perceived audio source position until the sound aligned with the white dot.

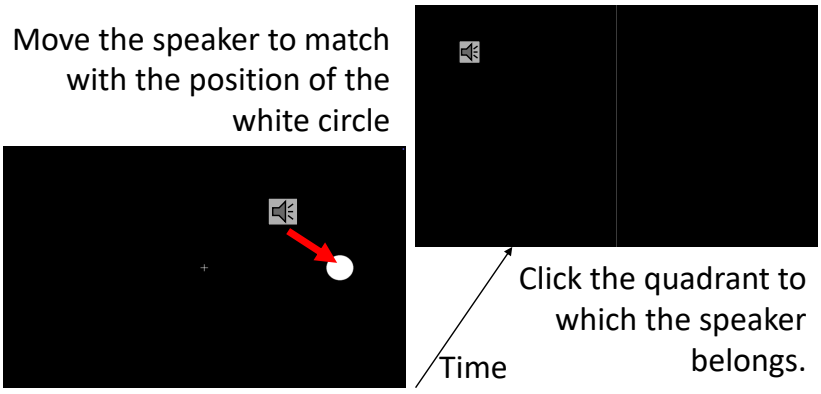


Figure S6: **The audio calibration and validation process**. During calibration, the player is at a random position close to the white dot. The red arrow represents a potential movement trajectory. During validation, the display is divided into two equal quadrants, with the audio player appearing at a random position three times within each quadrant.

Once satisfied, participants pressed the ESC key to confirm the alignment. This process was repeated six times with different white dot positions.

For each calibration step $t \in 1, 2, \ldots, 6$, we recorded the participants' adjustments of the sound source location in pixels $(\Delta x_t, \Delta z_t)$. The calibration hyperparameters α and β were computed as the mean of Δx_t and the sum of Δz_t, respectively. These hyperparameters were then applied to scale the Unity coordinates x_u and z_u, correcting for individual perceptual biases in auditory localization during the main experiment. The rationale for using the mean adjustment for Δx_t (horizontal direction) and the sum for Δz_t (vertical direction) is based on human spatial hearing characteristics—listeners typically localize horizontal (azimuth) sound sources with higher precision than vertical ones. This design allows greater tolerance for variability in the vertical dimension (altitude) during calibration, ensuring more robust alignment with participants' perceptual expectations.

Audio validation. Following calibration, an audio validation task was conducted to ensure successful calibration and adequate spatial hearing accuracy. Participants heard an audio clip played from one of two possible locations on a horizontally arranged 1x2 grid displayed on the screen (see **Supp Fig. S6**). They were instructed to click on the half of the screen from which they perceived the sound. This validation procedure was repeated six times. If a participant's spatial sound localization accuracy fell below 83%, the calibration and validation procedures were repeated to guarantee reliable data quality during the actual experiment.

Center fixation presentation before the visual stimulus onset. During the experiment, participants were instructed to fixate on a central dot before each trial began. This pre-trial fixation is a standard element in human psychophysics and cognitive neuroscience, designed to recenter attention and minimize carry-over effects across trials. By requiring participants to begin each trial from a common spatial and attentional baseline, this design ensures that any differences in response latency or eye movement patterns can be attributed to the experimental manipulation, rather than lingering attentional bias from the previous trial.

S3 More implementation details of AI models

SSLTI [75], LVS [44], FNAC [43], and IS3 [45] are SSL models based on dual-stream architectures with 2D Convolutional Neural Networks (CNNs). Each model processes visual and auditory inputs separately using dedicated encoders before fusing the features for contrastive learning during training. Both encoders are based on ResNet18 [79]. Beyond standard contrastive learning, IS3 introduces an Intersection-over-Union (IoU) loss and a semantic alignment loss to improve localization accuracy and better alignment between audio and visual modalities during supervised training. All the models are trained on the FlickrSoundNet [58] and VGG-Sound [49] datasets.

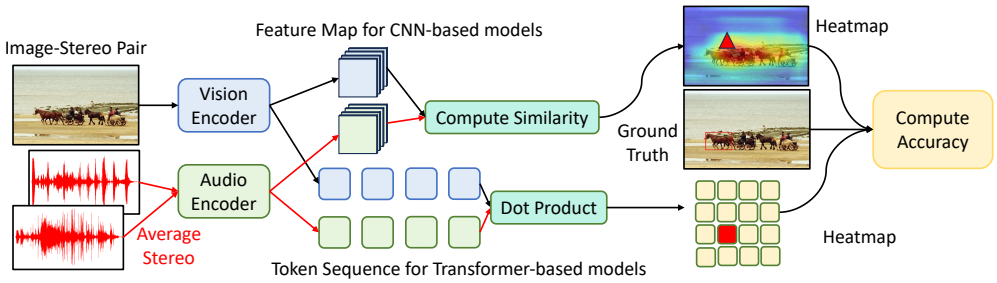


Figure S7: **Overview of SOTA multi-modal models for sound source localization (SSL).** All models receive paired images and mono audio inputs, where stereo signals are averaged into a single channel. Visual and auditory inputs are processed independently through separate encoders. For CNN-based models (indicated by blue arrows), feature maps from the final layers of each encoder are extracted, compared via cosine similarity, and used to generate a similarity heatmap. For transformer-based models, output token sequences are obtained from both encoders. Dot products on their token embeddings are calculated, and a heatmap is produced accordingly. During evaluation, SSL accuracy is determined by verifying whether the location of the maximum activation on the heatmaps (red triangles or red tokens) lies within the segmentation mask of the ground truth sounding object (red bounding box).

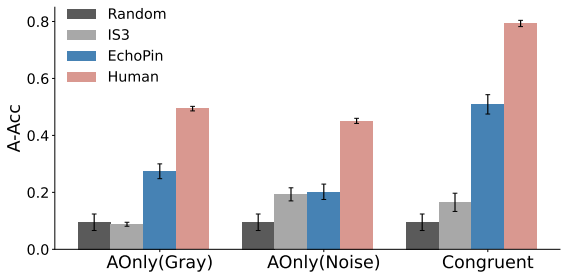


Figure S8: **While humans benefit from visual cues, they remain capable without them in the AOnly condition, unlike AI models, which fail to localize sound sources without congruent visual information.** The accuracy under AOnly conditions with both grayscale and Gaussian noise backgrounds drops for both humans and AI models compared to the congruent condition on object size 2.

CAVP [42] is a sound source segmentation model that follows a dual-stream CNN architecture. It uses PVTV2-B5 [80] as the visual encoder and either VGGish [81] or ResNet18 as the audio encoder. CAVP is trained in a fully supervised manner using cross-entropy loss for segmentation and contrastive learning to align audio-visual features. Training data includes AVSBench [35] and VPO [42] datasets.

AVSegformer [23] is an audio-visual semantic segmentation model built on a dual-stream transformer-based architecture. It employs SegFormer [82] as the visual backbone and Audio Spectrogram Transformer (AST) [83] as the audio encoder to capture fine-grained cross-modal representations. AVSegformer integrates both modality-specific and fused token embeddings through a lightweight fusion decoder for pixel-level prediction. It is trained in a fully supervised setting using a combination of cross-entropy loss for segmentation and audio-visual consistency loss. AVSegformer's training data includes three subsets of AVSBench.

ImageBind [76] and LanguageBind [33] are large-scale, multi-modal transformer models with billions of parameters. These models are trained with contrastive learning objectives across six modalities: image, audio, video, text, depth, thermal, and IMU, and embed these into a shared representation space. Both models use a Vision Transformer(ViT) [84] for visual encoding and AST [83] for audio encoding. ImageBind is trained on large-scale datasets including LAION [85, 86], SSv2 [87], and K400 [88]. LanguageBind builds on the pre-trained encoders from ImageBind and further fine-tunes them on the VIDAL-10M [33] dataset.

Implementation details of AI models. All eleven models—except EchoPin-S and EchoPin—take paired image and mono audio inputs, processing each modality independently through dedicated visual and audio encoders. For the mono audio setup, we follow the EchoPin-M design by averaging the stereo channels into a single-channel input. As shown in Fig. S7, for CNN-based models, feature

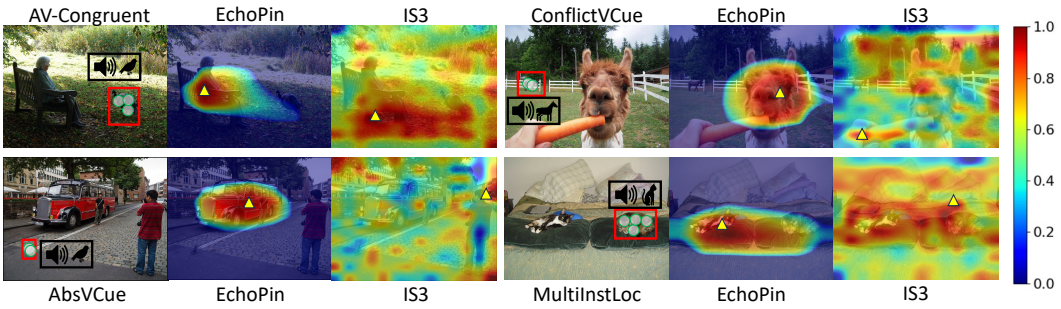


Figure S9: The figure shows the failure examples of both IS3 and EchoPin compared to humans. The leftmost images (Columns 1 and 4) show the correct localization results made by humans. Red boxes mark the ground truth sound source locations, while green circles indicate mouse click responses from multiple participants. The middle columns (2 and 5) and rightmost columns (3 and 6) display heatmaps predicted by EchoPin and IS3, respectively. Yellow triangles on the heatmaps denote the predicted sound source locations from each model. See the colorbar for the activation values of the heatmaps.

maps from the final layers of the encoders are extracted, and cosine similarity is computed between them to produce a similarity heatmap. For transformer-based models, token sequences from both encoders are retrieved, and pairwise dot products of their token embeddings are calculated to generate the heatmap. The predicted sound source location is identified as the point with the highest activation on the heatmap.

We evaluated all the current models using their publicly available, pre-trained weights and adhered to their original implementation details. Since AVSBench [35] shares some images with our AudioCOCO test set, we exclude those overlapping image-audio pairs when evaluating CAVP [42] and AVSegformer [23]. Experiments for ImageBind and LanguageBind were conducted on 8 NVIDIA A100 GPUs, whereas all other models, including EchoPin and its variants, were trained and evaluated on 4 NVIDIA A6000 GPUs. We fine-tune the EchoPin models using the Adaptive Moment Estimation (Adam) optimizer [89] with a weight decay of 1×10^{-4} for 10 epochs. The initial learning rate is set to 1×10^{-5}, and the batch size is 16. Each fine-tuning session takes approximately 16 hours to complete. To accelerate data loading, all audio waveforms are preprocessed and stored as cochleagram tensors in advance, with each .npy file occupying roughly 160 MB. All model evaluations are repeated three times with different random seeds to ensure statistical reliability.

S4 More qualitative results of humans and AI in SSL

As shown in **Fig.S9** and **Fig.S10**, we further illustrate the limitations of EchoPin relative to human performance and provide additional qualitative results for other baselines. Notably, EchoPin often fails to localize the correct sounding object when a visually salient distractor occupies a large portion of the scene. This suggests that the model remains vulnerable to visual saliency bias, tending to prioritize large or central objects even when they are not the true auditory source—a tendency that human listeners are better equipped to suppress.

Similar patterns are shown in CAVP, which exhibits modality conflict sensitivity akin to IS3. Both models are frequently misled by conflicting audio-visual cues and demonstrate a systematic bias toward the visual modality, highlighting a lack of robust auditory grounding under incongruent conditions.

S5 More quantitative results of AI models in SSL

The raw A-Acc may be confounded by the area of the ground-truth mask—since, intuitively, smaller objects are more difficult to localize while larger ones are easier. To address this potential bias, we introduce a chance-corrected gain metric, which quantifies the improvement of a human or model over a random guess, normalized by the baseline accuracy of the random guess:

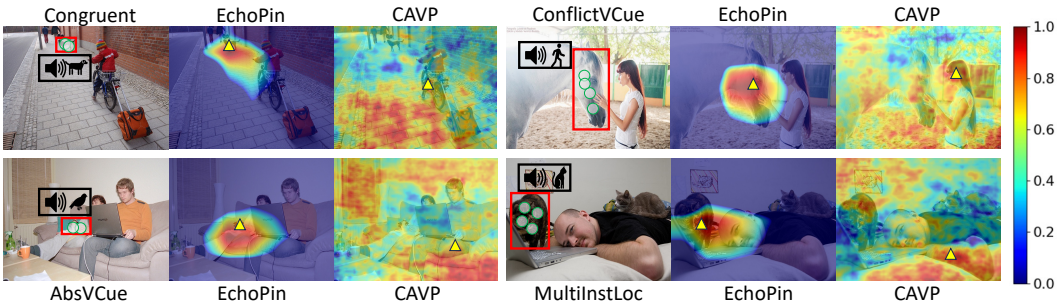


Figure S10: The figure shows the failure examples of CAVP and EchoPin compared to humans. The rightmost columns (3 and 6) display heatmaps predicted by CAVP.

$$\text{Gain}(X) = \frac{\text{Acc}_X - \text{Acc}_{\text{rand}}}{\text{Acc}_{\text{rand}}} \times 100 \tag{1}$$

The Acc_{rand} represents the accuracy achieved when predictions are sampled uniformly at random across the scene.

Normalized Percentage	**Size 1**	**Size 2**	**Size 3**
(EchoPin - Random) / Random	5.2	4.6	2.9
(Human - Random) / Random	31.9	7.3	3.6
(Human - EchoPin) / Random	26.7	3.0	0.8

Table S1: We report the chance-corrected gain for humans and models across three object sizes under the congruent condition. While human performance increases significantly with larger object sizes, models show limited improvement. This discrepancy may stem from AudioCOCO's uniform object size distribution, which limits the models' ability to exploit size-dependent cues.

This normalization removes the influence of mask size and isolates the true localization capability of humans and models. As shown in Tab.S1, even after correcting for chance, humans consistently outperform models—especially for small object sizes. For example, in the smallest size category, humans achieve a gain of 31.9%, compared to only 5.2% for models. These results suggest that object size modulates sound source localization performance in a way that cannot be fully explained by ground-truth mask area alone, highlighting deeper perceptual and representational differences between human and model behavior.

We apply the same evaluation criterion to both human participants and AI models to ensure fair comparisons: a prediction is considered correct if the peak activation (for models) or the human click falls within the bounding box of the sounding object. We also conducted an additional analysis by varying the pixel distance thresholds used to determine correctness. Specifically, a prediction is considered correct if it falls within x pixels of the ground-truth bounding box. We report the resulting A-Acc (accuracy with spatial tolerance) as a function of pixel thresholds in Tab.S2 based on congruent conditions for object size2. From these results, we observe that while larger thresholds naturally lead to higher A-Acc values, the relative performance trend between humans and models remains consistent. This further supports the validity of our evaluation methodology.

Moreover, we conducted an additional experiment by fine-tuning our EchoPin-S model using the VGG-SS and Flickr-SoundNet datasets, and report comparative results against IS3 and CAVP using the CIoU metric under the default evaluation protocol provided by [90]. As a localization-aware metric, CIoU accounts for overlap area, center distance, and aspect ratio alignment, offering a more comprehensive assessment of spatial prediction quality. From Tab. S3, we observe that benefiting from the fine-grained spatial features provided by AudioCOCO, our EchoPin-S model achieves superior performance on these standard SSL benchmarks, outperforming the baselines.

Threshold	Random	Human	EchoPin-S	EchoPin
0 (default)	9.5	79.3	17.3	50.9
10	9.6	79.9	18.5	52.0
25	9.9	80.4	19.8	52.4

Table S2: We compare performance across varying localization thresholds (0–25 pixels) under the congruent condition for object size 2. As shown, both human and model accuracy exhibit only marginal improvement with increasing thresholds, indicating that our evaluation metric is stable and not overly sensitive to small shifts in the decision boundary—thereby validating its robustness.

Method	VGG-SS	Flickr-SoundNet
IS3	42.96	84.40
CAVP	43.58	85.03
EchoPin-S	<u>43.61</u>	<u>85.25</u>
EchoPin	**45.02**	**85.87**

Table S3: We conduct comparison experiments on the VGG-SS and Flickr-SoundNet datasets. Despite the challenges posed by these large-scale, imbalanced public benchmarks, EchoPin consistently outperforms prior state-of-the-art (SOTA) methods, highlighting the robustness of spatial cues and the effectiveness of our neuroscience-inspired design.

Model Acc(%)	Congruent			Conflicting VCue			Absent VCue			Audio Only						Vision Only						Multi-Instance					
	A-Acc									Vision Gray			Vision Noise			Audio Silent			Audio Noise			A-Acc			V-Acc		
	Size1	Size2	Size3	Size1	Size2	Size3	Size1	Size2	Size3	Size1	Size2	Size3	Size1	Size2	Size3	Size1	Size2	Size3	Size1	Size2	Size3	Size1	Size2	Size3	Size1	Size2	Size3
Random	1.8	9.5	19.6	1.8	9.5	19.6	1.8	9.5	19.6	1.8	9.5	19.6	1.8	9.5	19.6	1.8	9.5	19.6	1.8	9.5	19.6	1.6	9.1	21.3	8.4	17.8	26.2
IS3	2.3	16.5	31.1	1.7	12.3	24.8	3.3	13.5	29.8	2.3	8.8	11.9	3.3	17.3	39.7	21.5	35.2	42.7	**20.5**	**31.6**	**38.9**	4.8	7.9	22.4	11.9	24.1	40.9
FNAC	1.8	12.5	26.9	1.9	7.2	13.3	5.8	10.5	18.6	1.1	4.8	8.7	1.3	5.0	9.6	6.2	16.5	43.8	5.4	12.6	11.9	0.7	3.9	26.3	13.4	25.5	52.8
SSLTIE	1.4	9.9	23.7	2.1	6.8	12.4	5.2	9.9	16.5	0.7	3.9	8.2	1.1	4.3	8.9	6.1	16.5	40.7	5.5	12.7	12.8	0.6	2.5	24.0	13.2	20.7	49.4
LVS	1.3	9.7	23.6	2.0	6.5	12.1	5.0	9.4	16.0	0.7	3.8	8.3	1.0	4.2	8.8	6.2	16.4	40.5	5.4	12.7	12.2	0.5	3.9	25.3	12.5	22.4	50.6
CAVP†	3.1	15.2	38.7	3.8	7.0	13.2	5.5	10.2	18.4	0.9	4.7	8.4	1.2	4.9	9.8	8.4	21.2	44.5	6.6	16.5	36.3	2.9	7.5	20.4	10.5	23.0	39.5
AVSegformer†	2.7	15.5	39.5	3.6	7.1	12.8	5.3	9.8	17.2	0.9	4.5	7.9	1.1	4.6	9.4	8.0	20.3	29.9	8.7	17.8	32.6	2.5	7.3	20.2	11.2	23.7	40.9
ImageBind	1.4	5.8	14.5	1.2	5.7	11.3	3.4	7.2	12.8	0.3	3.1	6.6	0.4	3.8	7.5	5.8	14.5	27.2	4.2	16.3	24.0	0.5	2.7	15.3	6.5	12.4	25.1
LanguageBind	1.1	4.6	12.7	1.1	5.4	10.7	3.2	6.9	10.6	0.2	3.0	6.5	0.4	3.7	7.6	5.6	13.2	24.9	4.1	13.8	24.8	0.4	2.1	13.0	6.2	10.8	22.5
EchoPin	11.2	50.9	76.0	16.6	21.9	26.4	11.8	54.0	77.2	10.0	16.9	17.9	4.0	20.2	40.8	**26.3**	**39.0**	**53.7**	13.5	17.6	21.9	4.5	24.1	47.1	37.5	53.8	64.2
Human	**59.3**	**79.3**	**90.7**	**40.0**	**57.9**	**67.9**	**54.3**	**65.7**	**80.0**	**9.3**	**49.4**	**76.8**	**12.3**	**45.1**	**80.3**	-			-			**25.7**	**36.4**	**38.6**	**60.9**	**82.8**	**89.1**

Table S4: **Comparison of AI models in our benchmark across six conditions and three object sizes.** Bold and underlined values indicate the best and second-best performances, respectively. Since CAVP and AVSegformer are trained on MS-COCO2017, we evaluate them on a subset of AudioCOCO and denote their results with the symbol †.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims in the abstract and introduction accurately summarize the paper's key contributions and align well with the scope and results presented throughout the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper explicitly discusses the limitations of the proposed method, acknowledging its constraints and outlining areas for future improvement in section of Discussion.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper provides sufficient details on the experimental setup, model architecture, and evaluation protocol to support reproducibility of the main results and states that the code, model, and data are publicly available.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The paper indicates that the code and data are made publicly available.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies key training and testing details, including data selection pipeline, hyperparameters and optimizer settings, allowing readers to understand how the results were obtained both in the main body and the appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper contains error bars and we test all models with 3 runs and report its mean. For human experiments, we summarize all results, randomly select 60% data, calculate the accuracy with 5 runs and report its mean.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper provides sufficient information on the computational resources used for model experiments and human experiments in the section of Experiment Details in the appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The research presented in the paper adheres to the NeurIPS Code of Ethics, with no identified ethical concerns regarding the methods, data usage, or potential societal impact.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: Our work shares the general ethical considerations common to AI research, and does not present any unique or specific societal impact that warrants separate discussion.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: For the human experiment data, all participants have approved to use their anonymous data for research activity and signed the consent form under the supervision of IRB.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper uses existing assets, which the authors have properly acknowledged. It also contributes new assets, including novel models, datasets, and benchmarks. Details of these contributions are clearly described in the paper, and all associated code, data, and models are made publicly available.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: Details of the human experiments are provided in both the main body and the Experiment Details section of the appendix.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

Justification: IRB has approved our human behavioral experiments. The human experiments pose no risks to participants.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.