Unsupervised Audio-Visual Segmentation with Modality Alignment

Swapnil Bhosale[1], **Haosen Yang**[1], **Diptesh Kanojia**[1], **Jiankang Deng**[2], **Xiatian Zhu**[1]

[1]University of Surrey, UK
[2]Imperial College London, UK

Abstract

Audio-Visual Segmentation (AVS) aims to identify, at the pixel level, the object in a visual scene that produces a given sound. Current AVS methods rely on costly fine-grained annotations of mask-audio pairs, making them impractical for scalability. To address this, we propose the **M**odality **C**orrespondence **A**lignment (MoCA) framework, which seamlessly integrates off-the-shelf foundation models like DINO, SAM, and ImageBind. Our approach leverages existing knowledge within these models and optimizes their joint usage for multimodal associations. Our approach relies on estimating positive and negative image pairs in the feature space. For pixel-level association, we introduce an audio-visual adapter and a novel pixel matching aggregation strategy within the image-level contrastive learning framework. This allows for a flexible connection between object appearance and audio signal at the pixel level, with tolerance to imaging variations such as translation and rotation. Extensive experiments on the AVSBench (single and multi-object splits) and AVSS datasets demonstrate that MoCA outperforms unsupervised baseline approaches and some supervised counterparts, particularly in complex scenarios with multiple auditory objects. In terms of mIoU, MoCA achieves a substantial improvement over baselines in both the AVSBench (S4: **+17.24%**; MS3: **+67.64%**) and AVSS (**+19.23%**) audio-visual segmentation challenges.

Introduction

Audio-visual segmentation (AVS) accurately identifies pixel-level objects producing specific sounds in videos. Previous studies focused on audio-visual signal intersection using self-supervised learning methods (Afouras et al. 2020; Rouditchenko et al. 2019), but face limitations in real-world applications like video editing and robotics. Supervised learning with pixel-level annotated video-audio pairs is intuitive (Zhou et al. 2022), but scaling annotation is challenging, as object segmentation levies significantly higher cognitive load, and effort, than classifying or bounding boxes (Zlateski et al. 2018). In complex backgrounds, ground-truth segmentation labels may overlap or be ill-defined, requiring fine-grained boundaries for differentiation among surrounding objects. In this work, we address this challenge with a more pragmatic and scalable

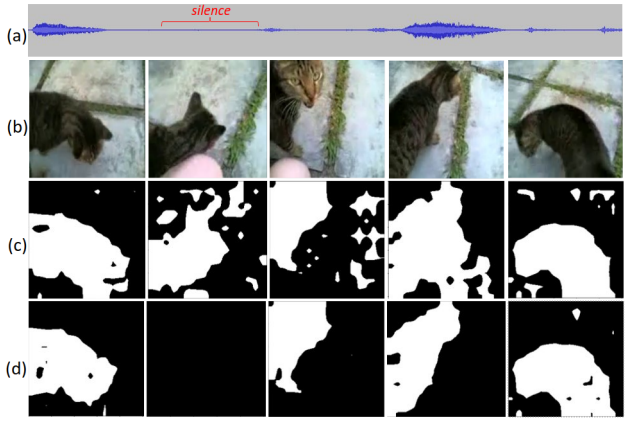

Figure 1: Emergence of coarse and noisy *audio-pixel association* in ImageBind's (Girdhar et al. 2023) multimodal feature space. (a) Raw audio waveform, (b) input RGB, (c) using frozen ImageBind features, (d) MoCA (ours) – generating finer masks and particularly only in the presence of sounding objects (see the 2^{nd} frame, MoCA generates no mask due to silent audio).

unsupervised AVS approach, eliminating the need for exhaustive pixel annotation. This choice is influenced by notable advancements in purpose-generic image-level audio-visual representation learning, as demonstrated by ImageBind (Girdhar et al. 2023). Empirically, we observe the emergence of coarse cross-modal associations between audio and image pixels, as depicted in Figure 1. However, despite this progress, the acquired knowledge remains insufficient for achieving an extremely precise unsupervised solution.

To address the mentioned limitation, we propose **M**odality **C**orrespondence **A**lignment (MoCA), an unsupervised audio-visual segmentation framework that efficiently integrates off-the-shelf foundation models using a minimal number of learnable parameters. Our approach initiates with unsupervised pairing of training images, leveraging image-level visual features such as those provided by DINO (Caron et al. 2021). This process augments the original audio-image pairing relationships by introducing additional positive and negative pairs at a coarse image level, thereby estab-

lishing the underlying *one-audio-to-multi-object-instances* mapping information. To establish precise audio-pixel associations, we introduce a novel audio-visual adapter and pixel matching aggregation within an audio-enhanced visual feature space. This strategy facilitates the correlation of audio signals and image pixels within a conventional image-level contrastive learning framework. Specifically, we aggregate the similarity of pixel pairs across both the original and additionally paired images, effectively addressing imaging variations like translation and rotation, while allowing for the emergence of desired audio-pixel associations. To further refine boundaries, we incorporate open-world object detection and segmentation models (*e.g.,* SAM (Kirillov et al. 2023)), aligning them with the derived audio-pixel associations. This integration ensures accurate delineation of object boundaries in synchronized audio-visual context. Our contributions are summarized as follows:

- We propose a novel and efficient unsupervised framework, named MoCA, which systematically incorporates the off-the-shelf capabilities of pre-trained foundation models with a minimal number of additional parameters. Pixel matching aggregation with our approach reveals audio-pixel associations based on coarse image-level pairwise information.
- We perform a comprehensive evaluation and benchmark the performance of our framework against with existing approaches, demonstrating that MoCA significantly outperforms the baselines on both the AVSBench (S4: **+17.24%**; MS3: **+67.64%**) and AVSS (**+19.23%**) datasets in terms of mIoU, narrowing the performance gap with supervised AVS alternatives.

Related Work

Audio-Visual Semantic Segmentation Existing AVS methods rely on fully supervised models to identify audible visual pixels, trained on numerous manually annotated segmentation masks (Zhou, Guo, and Wang 2022; Mao et al. 2023; Liu et al. 2023a; Hao et al. 2023; Shi et al. 2023; Mo and Tian 2023; Zhou et al. 2022). This resource-intensive process poses challenges for large-scale applications with diverse scenes and multiple audible objects. To address these challenges, we propose an unsupervised AVS approach that eliminates the need for exhaustive audio-mask pairs.

Self-Supervised Audio-Visual Feature Learning Various self-supervised approaches have been developed for learning audio-visual correspondence, primarily focusing on reconstruction by masked autoencoders (MAEs) (Georgescu et al. 2023; Nunez et al. 2023) and contrastive learning (Chen et al. 2021b; Ma et al. 2020; Guzhov et al. 2022; Wu et al. 2022). Recent methods like MAVIL combine MAE and contrastive learning to generate efficient audio-visual representations for tasks such as audio-video retrieval (Huang et al. 2022). Additionally, ImageBind (Girdhar et al. 2023) learns joint embeddings across six modalities, enhancing zero-shot capabilities for large-scale vision models. While these methods address coarse cross-modal relationships (Yang et al. 2024a), our work focuses on AVS to accurately pinpoint object boundaries at the pixel level, tackling this challenge for the first time using an unsupervised approach.

Unsupervised Image Semantic Segmentation In unsupervised image semantic segmentation, the goal is to accurately isolate objects of interest without using annotations. Existing methods focus on enhancing pre-trained self-supervised visual features. SegSort (Hwang et al. 2019) refines visual features in a spherical embedding space. MaskContrast (Van Gansbeke et al. 2021) uses saliency models to generate binary masks, contrasting learned features within and across these maps. STEGO (Hamilton et al. 2022) distills visual feature correspondence through self-correlation within the same image and cross-correlation with a similar image, along with complex post-processing. These methods address mono-modality challenges, lacking cross-modal associations between audio and pixel data, which our work aims to address.

Method

Given a video sequence $\{I_t\}_{t=1}^{T} \in \mathcal{R}^{H\times W}$ comprising T non-overlapping continuous image frames and an accompanying audio sequence $\{A_t\}_{t=1}^{T}$, our objective is to generate object segmentation masks $\{G_t\} \in \mathcal{R}^{H\times W}$ that correspond to the audio content. Here, H and W represent the height and width of each frame. These masks label individual pixels, highlighting the sound-producing object in A_t within the frame I_t. In unsupervised AVS, detailed annotations for pixel-level masks are not available. To address this challenge, we propose an unsupervised approach named MoCA.

MoCA Framework

As depicted in Figure 2, MoCA leverages the existing capabilities of foundational models, including DINO (Caron et al. 2021), ImageBind (Girdhar et al. 2023), and SAM (Kirillov et al. 2023), within a contrastive learning framework. Initially, we generate positive and negative image pairs using DINO embeddings. To seamlessly integrate these embeddings for unsupervised AVS, we propose an audio-visual adapter design with a minimal number of extra parameters. This design is integrated into a frozen multimodal foundation model, *e.g.,* ImageBind, to produce audio-enhanced image features. For accurate pixel segmentation association with the audio signal, we introduce a novel region matching aggregation strategy, facilitating the identification of object-level correspondences grounded on coarse image-level pairing supervision.

Audio-Visual Adapter (AdaAV) The objective of our audio-visual adapter is to integrate audio event-specific knowledge into frozen visual feature representations. This integration results in audio-enhanced image features that capture pixel-level correlations with audio signal. This forms the basis for exploring the inherent association between audio and visual objects.

The architecture for this design is illustrated in our supplementary. Although our subsequent discussion is grounded in the context of ImageBind (Girdhar et al. 2023), it's important to note that our design is versatile, and integration with

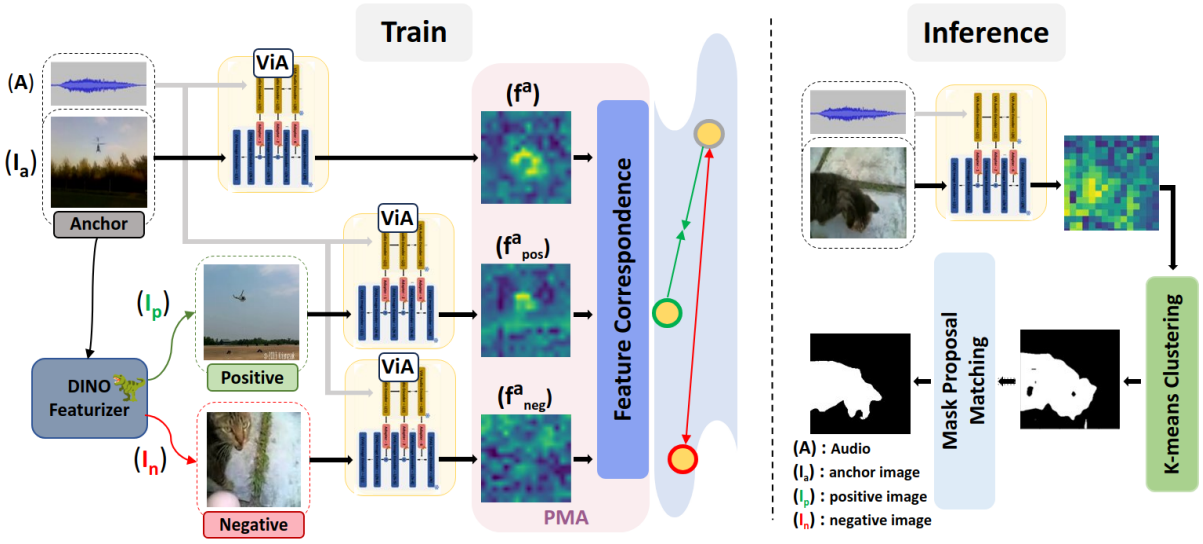

Figure 2: **Overview of our proposed MoCA**: (**Left (Train)**) We generate positive and negative images by utilizing DINO embeddings. The fusion of these images with the corresponding audio from the anchor image yields audio-enhanced image features. With our pixel matching aggregation, and a contrastive training objective, this efficiently learns proposed audio-visual adapter weights. (**Right (Inference)**) Extract audio-enhanced image features and employ k-means clustering. Optionally, we enhance object boundaries by matching the clustered feature map with mask proposals from a pre-trained SAM model. All vision-audio (ViA) model weights are shared and frozen.

other multimodal Vision-Audio (ViA) models can be similarly achieved.

Each image frame, I_t, is individually input to ImageBind's encoder. Within the encoder, I_t undergoes decomposition into n non-overlapping patches, which are then flattened into visual embeddings $\mathbf{X}_v^{(0)} \in \mathbb{R}^{n\times d}$. Simultaneously, A_t is projected into Mel-spectrograms, patched, and fed into the ImageBind audio encoder, revealing audio embeddings denoted as $\mathbf{X}_a^{(0)} \in \mathbb{R}^{k\times d}$ (where 0 signifies layer-0, *i.e.,* the input layer). With $\mathbf{X}_a^{(\ell)}$ and $\mathbf{X}_v^{(\ell)}$ representing audio and visual inputs, respectively, at layer l, both audio and image trunks employ a multi-headed attention (MHA) layer. The output from the MHA layer is then passed through a multi-layer perceptron (MLP) layer.

Our audio-visual adapter module, AdaAV, is introduced between the intermediate layers of the image and audio trunks. We describe AdaAV in detail in Supplementary. For a given intermediate layer-l, AdaAV generates an audio-enhanced image feature representation,

$$\mathbf{F}_v^{(\ell)} = \text{AdaAV}(\mathbf{X}_a^{(\ell)}, \mathbf{X}_v^{(\ell)})$$

The updated MHA and MLP operations in each layer of the image trunk are as follows:

$$\mathbf{Y}_v^{(\ell)} = \mathbf{X}_v^{(\ell)} + \text{MHA}(\mathbf{X}_v^{(\ell)}) + \text{AdaAV}(\mathbf{X}_a^{(\ell)}, \mathbf{X}_v^{(\ell)})$$
$$\mathbf{X}_v^{(\ell+1)} = \mathbf{Y}_v^{(\ell)} + \text{MLP}(\mathbf{Y}_v^{(\ell)}) + \text{AdaAV}(\mathbf{Y}_a^{(\ell)}, \mathbf{Y}_v^{(\ell)})$$

AdaAV employs a limited number (m) of latent audio tokens, denoted as $\mathbf{L}_a^{(l)} \in \mathbb{R}^{m\times d}$, to effectively integrate audio-specific knowledge into the visual representation. Here, the value of m is significantly smaller than the overall number of audio tokens. The primary purpose of these latent tokens is to succinctly encapsulate information from the audio tokens, facilitating efficient information transfer into the intermediate layer visual representations.

Distinct sets of latent tokens are utilized at each layer. A cross-modal attention block is employed to compress all tokens from the audio modality into these latent tokens. Subsequently, another cross-modal attention block is utilized to fuse information between the compressed latent tokens of the audio modality and all tokens of the image modality.

In alignment with previous research on adapters (Houlsby et al. 2019; Jaegle et al. 2021; Nagrani et al. 2021), a bottleneck module is incorporated. This module comprises a learnable down-projection layer, a non-linear activation function, and a learnable up-projection layer.

Pixel matching aggregation (PMA) To achieve AVS, it is necessary to establish cross-modal associations at the pixel level. In pursuit of this goal, we delve into pixel-level correlations involved in audio-enhanced image features. Drawing inspiration from the stereo matching literature (Prince 2012; Hartley and Zisserman 2003), we propose a *pixel matching strategy* that gauges the similarity between audio-enhanced image features generated at the pixel level. Formally, given two audio-enhanced feature tensors f^a and g^a, a matching cost is computed for each spatial location in f^a, considering a set of pixels along the corresponding scanline of g^a. This cost signifies the matching likelihood of a pixel at a specific disparity in f^a with a pixel in g^a. We further refine this process to identify optimal correspondences by pinpointing disparities with the lowest costs, signifying regions containing

potential sounding source objects.

We compute a combined matching cost as:

$$C = \lambda_{ssd} \cdot C_{ssd} + \lambda_{ncc} \cdot C_{ncc} \quad (1)$$

where $\lambda_{ssd/ncc}$ is the weight hyper-parameter, and the first term is the sum of squared differences (SSD) (Szeliski 2022; Trucco and Verri 1998):

$$C_{ssd}(f^a, g^a) = \sum_{i,j} \left(f^a(i,j) - g^a(i,j)\right)^2 \quad (2)$$

and the second term is the normalized cross-correlation (NCC) (Pratt 2007):

$$C_{ncc}(f^a, g^a) = \frac{\sum_{i,j} \left(f^a(i,j) \cdot g^a(i,j)\right)}{\sqrt{\sum_{i,j} \left(f^a(i,j)^2\right) \cdot \sum_{i,j} \left(g^a(i,j)^2\right)}} \quad (3)$$

Associating pixels between two images allows for versatile object matching, accommodating variations like translation and rotation. We propose that the dual-metrics cost offers complementary advantages: SSD captures absolute intensity differences, while NCC considers shape and scale similarity. This fusion's positive impact is shown in Table 4. Given f^a and g^a as the two audio-enhanced features, SSD captures intensity by summing squared pixel value discrepancies, while NCC computes pattern similarity by normalizing the data, making it less sensitive to amplitude variations.

AdaAV Adapter Training In our design, the AdaAV is the sole learnable module, while the audio and image encoder weights remain fixed. The learning of AdaAV weights is achieved through a contrastive training objective, which is outlined as follows:

$$L_c = \max\left(0, C(f^a, f^a_{pos}) - C(f^a, f^a_{neg}) + \alpha, 0\right) \quad (4)$$

where α is the margin parameter and C is defined in Eq. (1). Positive (f^a_{pos}) and negative (f^a_{neg}) sets are formed by leveraging the pre-trained DINO (Caron et al. 2021) as an off-the-shelf image featurizer. Specifically, global image features are extracted using DINO embeddings through global average pooling. Subsequently, we create a lookup table containing each image's K-Nearest Neighbors (KNNs) based on cosine similarity in the DINO's feature space. This augments the original audio-image pairs by introducing additional pairing information at the image level, providing the coarse *one-audio-to-multi-object-instances* association information. For any images x and their corresponding random nearest neighbors, we form the set of positive images, denoted as x_{pos}. To constitute the set of negative images x_{neg} relative to x, we randomly sample images by shuffling x while ensuring that no image matches with itself or its top KNNs. The audio embedding A corresponding to x augments both the positive set (f^a_{pos}) and negative set (f^a_{neg}), as illustrated in Figure 2.

Inference During inference, we use the AdaAV module to extract audio-enhanced features from an image with accompanying audio. We refine pixel-level associations using a cosine distance-based k-means algorithm (Hartigan and Wong 1979). Overlaying clusters with mask proposals from a pre-trained SAM model helps achieve precise object boundaries. Mask proposal generation (in supplementary), and an ablation study in Section highlights the importance of our pixel matching strategy.

Experiments

Training Data For training our AdaAV weights, we employ VGGSound dataset (Chen et al. 2020), a large scale audio-visual dataset collected from YouTube. The dataset consists of over 200k clips for 300 different sound classes, spanning a large number of challenging acoustic environments and noise characteristics of real applications.

Benchmark settings To assess our proposed method, we utilize the AVSBench dataset (Zhou et al. 2022), which includes YouTube videos split into five clips, each with a spatial segmentation mask for audible objects. AVSBench consists of two subsets: `Semi-supervised Single Source Segmentation` (S4) and fully supervised `Multiple Sound Source Segmentation` (MS3), differing by the number of audible objects. We also report scores on the `AVS-Semantic` (AVSS) dataset (Zhou et al. 2023). Our evaluation on the S4, MS3, and AVSS test splits is conducted without using audio-mask pairs from the training and validation splits, emphasizing unsupervised AVS.

Metrics We use average Intersection Over Union (M_{IoU}) and F_{score} as metrics. A higher M_{IOU} implies better region similarity, and an elevated F_{score} indicates improved contour accuracy.

Baselines The current supervised AVS training framework requires resource-intensive audio-mask pair annotations, making it unsuitable for unsupervised benchmarking. We propose an alternative approach that eliminates this need by leveraging insights from foundational models. Our method includes three integration mechanisms utilizing these models for segmentation, audio tagging, and visual object grounding.

(I) AT-GDINO-SAM In the first baseline, we adopt the Audio Spectrogram Transformer (AST) (Gong, Chung, and Glass 2021) to generate audio tags for the AVS task, utilizing its capacity to capture global context in audio sequences with multiple events. Trained on the large Audioset dataset (Gemmeke et al. 2017), AST detects diverse polyphonic audio events across 521 categories like animal sounds, instruments, and human activity. Given an audio sequence, A_t, we pad it to a maximum of 960 msec and obtain its corresponding audio tags $\{AT_i\}_{i=1}^{C_a}$ using the pre-trained AST model. The tags are ranked based on their probability scores across 521 generic classes from the Audioset ontology. Relevant audio tags are filtered using an empirically determined threshold, τ_{AT} and forwarded to a pre-trained GroundingDINO model (Liu et al. 2023b), generating bounding boxes in image frame I_t. These boxes serve as visual prompts for Segment Anything model (SAM) (Kirillov et al. 2023), to produce producing binary masks.

(II) OWOD-BIND We use a pre-trained Open World Object Detector (OWOD) (Maaz et al. 2022) model to generate class-agnostic object proposals that are further processed by a segmentation pipeline. Unlike traditional object detectors which consider unfamiliar objects as background, OWOD

models have been designed to handle unknown objects during training and inference. Given an image frame, I_t, we use OWOD to generate C_v proposal bounding box proposals, $\{BB_i\}_{i=1}^{C_v}$. These are filtered by objectness score (Maaz et al. 2022) using a threshold τ_{BB}. To link boxes and acoustic cues, both modalities need a shared latent space embedding semantics. To this end, we utilize ImageBind's (Girdhar et al. 2023) latents, and extract image and audio embeddings (post global average pooling), to rank the proposals via cosine similarity with the audio embedding. Bounding boxes above τ_{BIND} form the final mask.

(III) SAM-BIND Alternatively, instead of relying on the OWOD model to generate bounding box proposals, we can randomly position single-point input prompts in a grid across the image. From each point, SAM can predict multiple masks. These masks are refined and filtered for quality, employing non-maximal suppression (NMS) (Girshick et al. 2014) to remove duplicates.

Implementation Details

Baselines We resize all image frames to 224×224. For all our evaluation results we use τ_{AT} as 0.5 for AT-GDINO-SAM and τ_{BB} and τ_{BIND} as 0.5 and 0.7 respectively, for OWOD-BIND. For refining the mask outputs from SAM-BIND, we use IoU threshold of 0.5 for NMS.

MoCA For the frozen ViA we choose the ImageBind-Huge model with 6 and 12 transformer encoders in the audio and image trunks respectively. The resolution of the input images are 224×224. To optimize the model parameters, we employ the Adam (Loshchilov and Hutter 2017) optimizer with an initial learning rate of $1e-4$ with cosine decay. We use $\lambda_{SSD} = \lambda_{NCC} = 1$ and $\alpha = 0.3$. We train models for a maximum of 10000 iterations on a single NVIDIA RTX A5500 GPU with batch size of 8. For the mask proposal matching, we consider only the proposal masks with IoU$>$0.5, when compared with the mask generated from the fused ViA encoder (post k-means).

Benchmark Results

We present our primary AVSBench results on the test set for both S4 and MS3, and the AVSS dataset in Table 1. Below, we first provide a quantitative comparison in terms of M_{IoU} and F_{score}, followed by a more qualitative comparison in terms of the quality of the generated masks.

Comparing with unsupervised baseline approaches, we can observe that both, OWOD-BIND and SAM-BIND, outperform AT-GDINO-SAM in terms of both M_{IoU} and F_{score} by a significant margin. This is because AST itself achieves a mean average precision (mAP) of 0.485 when evaluated on the audio tagging task of Audioset, and hence the generated audio tags are prone to errors. Additionally, we believe despite AST's training data *i.e.*, Audioset follows a generic ontology, many rare events (e.g., "lawn mover", "tabla", *etc.*) are under-represented and hence are unable to cope with an open-set inference. We also observe that for MS3, OWOD-BIND achieves an absolute improvement of 0.06 and 0.08 over SAM-BIND in terms of M_{IoU} and F_{score}, respectively with a similar trend in S4 setting, with an improvement of 0.16 over both M_{IoU} and F_{score} respectively. **MoCA surpasses all the baseline approaches** by more than 10% in terms of M_{IoU} and more than 13% in terms of the F_{score} on both the S4 and MS3 settings. In contrast, to our best performing baseline OWOD-BIND, MoCA differs primarily in the latent space matching generated by the ImageBind encoders.

Comparing with supervised AVS approaches, it is crucial to understand that "supervised" in this context refers to access to datasets with audio-visual masks (AVSBench (Zhou et al. 2022)). Provisioning visual masks is generally easier than audio-visual masks. While our framework employs pre-trained supervised models, "supervised" here specifically means pre-trained using only visual masks (as in the case of SAM). By leveraging audio-visual masks, along with a consistent training distribution (train and test splits from AVSBench) and a significantly larger number of trainable parameters, existing supervised approaches achieve higher segmentation performance. Nonetheless, we see promising potential in unsupervised AVS and the robust performance offered by the MoCA framework.

Comparing with salient object detection (SOD) and sound source localization (SSL) based approaches, we observe that **MoCA performs significantly better** on both S4 and MS3 subset. It is important to acknowledge that SOD operates solely on visual data without accounting for sound, prioritizing the dataset it was trained on. Consequently, it struggles with scenarios like MS3, where the auditory elements change while the visual components remain constant. On the other hand, SSL provides a more fair comparison to our MoCA, as it incorporates audio signals to guide object segmentation. However, a significant disparity is apparent between SSL and both our MoCA and the proposed training-free baselines (AT-GDINO-SAM, OWOD-BIND, SAM-BIND).

Qualitative Comparison: We present example segmentations from MoCA and the AVS supervised benchmark approach in Figure 3. MoCA excels at resolving fine-grained details such as object boundaries (e.g., the guitar and guitarist), unlike AVSBench, which, despite using a PVTv2 architecture for high-resolution segmentation, misses these details and generates discontinuous segments. MoCA, without using audio-mask pairs during training, effectively captures overlapping objects and fine details. The baseline method struggles to filter out sounding objects, particularly when multiple objects of the same category are present. MoCA successfully highlights the "`sounding human`" in such cases. A qualitative comparison of MoCA and AVSBench on the AVSS split is provided in the Appendix.

Ablation Study

Effect of pixel matching aggregation: As discussed in Section , we align the clustered predictions from MoCA with proposal masks, to obtain finer object segmentation boundaries. From Table 2 it is evident that matching the proposals yields an overall improved segmentation performance, with maximum contribution of this performance owed to

[1] Our unsupervised setup does not generate softmax prob. across the ground-truth event labels, details and qualitative samples in supplementary.

	Approach	Mask-free training	# learnable params ↓	S4		MS3		AVSS	
				M_{IoU} ↑	F_{score} ↑	M_{IoU} ↑	F_{score} ↑	M_{IoU} ↑	F_{score} ↑
Supervised	AVSBench (Zhou et al. 2022)	✗	101M	0.78	0.87	0.54	0.64	0.29	0.35
	AVSegFormer (Gao et al. 2024)	✗	344M	0.82	0.89	0.58	0.69	0.36	0.42
	COMBO (Yang et al. 2024b)	✗	396M	0.84	0.91	0.59	0.71	0.42	0.46
SOD	iGAN (Mao et al. 2021)	✓	86M	0.61	0.77	0.42	0.54	-	-
	LGVT (Zhang et al. 2021)	✓	90M	**0.74**	**0.87**	0.40	0.59	-	-
SSL	LVS (Chen et al. 2021a)	✓	20M	0.37	0.51	0.29	0.33	-	-
	MSSL (Qian et al. 2020)	✓	21M	0.44	0.66	0.26	0.36	-	-
	EZ-VSL (Mo and Morgado 2022)	✓	18M	0.45	0.68	0.28	0.34	-	-
	Mix-Loc (Hu, Chen, and Owens 2022)	✓	19M	0.44	0.69	0.32	0.36	-	-
Baseline	AT-GDINO-SAM	✓	-	0.38	0.46	0.25	0.29	0.24	0.25
	SAM-BIND	✓	-	0.42	0.51	0.28	0.36	0.24	0.26
	OWOD-BIND	✓	-	0.58	0.67	0.34	0.44	0.26	0.29
Our	**MoCA**	✓	**1.3M**	**0.68**	0.79	**0.57**	**0.62**	**0.31**	**0.33**

Table 1: Performance comparison on the AVSBench `test` split under the S4, MS3 and the AVSS[1] setting. Grayed: *fully supervised learning based methods*.

PMA	MPM	MS3	
		M_{IoU} ↑	F_{score} ↑
✓	✗	0.52	0.59
✗	✓	0.34	0.44
✓	✓	**0.57**	**0.62**

Table 2: Effect of pixel matching aggregation (PMA) and mask proposal matching (MPM) (Sec), on AVSBench MS3 `test` split.

Method	M_{IoU}
EZ-VSL (Mo and Morgado 2022)+MPM	0.43
Mix-Loc (Hu, Chen, and Owens 2022)+MPM	0.45
MoCA W/o MPM	**0.52**
MoCA	**0.57**

Table 3: Comparing MoCA and SSL methods equipped with mask proposal matching (MPM), on the AVSBench MS3 `test` split.

	Parameter	S4		MS3	
		M_{IoU} ↑	F_{score} ↑	M_{IoU} ↑	F_{score} ↑
# (AdaAV)	2	0.60	0.69	0.39	0.47
	4	0.65	0.73	0.43	0.51
	6	**0.68**	**0.79**	**0.57**	**0.62**
Loss	SSD	0.61	0.72	0.40	0.57
	NCC	0.65	0.76	0.44	0.59
	SSD+NCC	**0.68**	**0.79**	**0.57**	**0.62**
AdaAV Pos	Interleaved	**0.68**	0.77	0.56	0.60
	first-k	0.67	0.78	0.54	0.60
	last-k	**0.68**	**0.79**	**0.57**	**0.62**

Table 4: Performance obtained on S4 and MS3 `test` splits varying: a) Number of AdaAV blocks, b) Loss functions SSD and NCC, and c) positioning the AdaAV blocks w.r.t to the ImageBind image and audio trunk.

the audio-enhanced feature map learned by MoCA. For the scores in first row, we select random points as input to the SAM model[2] and match all the segmentation maps with the closest ground truth in the MS3 test set. It can be noted that, this deprived of the mapping with MoCA generated enhanced feature, drops the performance significantly, with more than 0.2 drop compared to MoCA (row 3). Additionally, we make it more concrete to particularly establish the higher contribution of PMA over MPM by combining existing SSL approaches with MPM. As can be seen in Table 3, combining MPM with Mix-Loc (Hu, Chen, and Owens 2022) and EZ-VSL (Mo and Morgado 2022) fails to surpass MoCA and even MoCA without MPM, due to the sparse nature of SSL outputs. To further understand this, we pictorially highlight the under-segmentation (Figure 4(d)) and over-segmentation (Figure 4(f)) problems that occur when converting SSL output (heatmaps) to binary masks by employing MPM. Using this we reiterate the efficacy of pixel-level association learned by our proposed PMA strategy leading to fine-grained localization, while the MPM gives further (less significant than PMA) advantage by refining the object edges.

Loss function: We observe impact of SSD and NCC for loss computation, in Table 4. The results indicate that: (1) Using NCC, compared to only SSD, results in a higher M_{IoU} and F_{score} (+6.55%); primarily because SSD is sensitive to the intensity of the pixels when computing feature correspondences. (2) Combining SSD and NCC for computing the correlation, leads to the best segmentation performance. NCC is invariant to the shape of the potential sounding region, hence, agnostic to the scale of the sounding object among

[2] github.com/facebookresearch/segment-anything/blob/main/notebooks/automatic_mask_generator_example

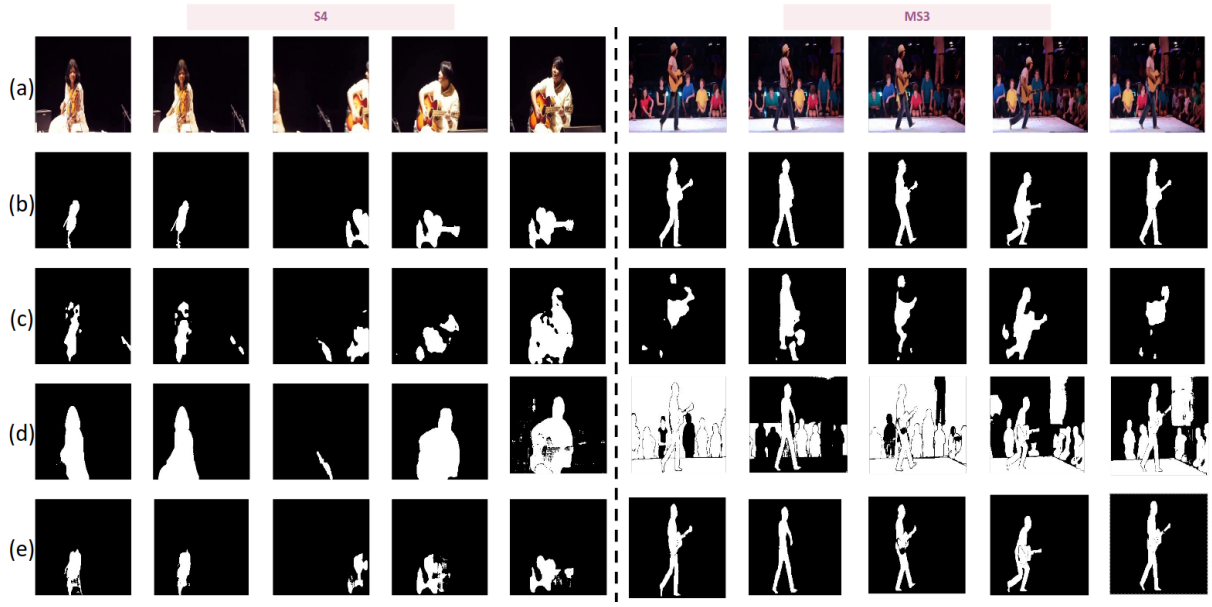

Figure 3: **Qualitative comparisons** (**left**: S4, **right**: MS3): (a) RGB frame, (b) ground truth mask, (c) AVSBench (supervised), (d) OWOD-BIND (baseline), (e) MoCA (ours) produces precise segmentation of overlapping objects **without utilizing any audio-visual masks during training**.

Figure 4: **Comparing SSL methods equipped with MPM**: (a) input frame, (b) ground truth mask, (c) Mix-Localize (Hu, Chen, and Owens 2022), (d) Mix-Localize + MPM, (e) EZ-VSL (Mo and Morgado 2022), (f) EZ-VSL + MPM, (g) MoCA (ours) – producing more precise segmentation of all overlapping sounding objects (Guitarist and guitar in this example).

the anchor, and the positive images.

Number of adapter blocks: Our proposed audio-visual adapters decide the amount of audio specific information that is added into the frozen image encoders. To examine its effect, we vary the number of AdaAV blocks from 2 to 6 (the maximum number of adapters we can add since the ImageBind audio trunk consists of 6 transformer encoder blocks). From Table 4, increasing the number of adapter blocks significantly improves both M_{IoU} and F_{score}, though the gains diminish from 4 to 6 blocks. Two AdaAV blocks result in the worst performance, with minimal audio enhancement and poor audio-guided segmentation.

Placement of AdaAV blocks: Utilizing 6 AdaAV blocks, we explore their optimal placement relative to the ImageBind audio and image encoder blocks, (Table 4). ImageBind image encoder consists of 12 transformer encoder blocks. The AdaAV blocks can be positioned along the first 6 image encoder blocks, the last 6, or interleaved, where every alternate encoder block is fused with an AdaAV block. Our observations are: (1)Interleaving and placing blocks closer to the output (last 6) yield similar performance, with a maximum deviation of 0.01 in M_{IoU} on MS3. (2) Positioning blocks closer to the output results in up to a 3% improvement in M_{IoU} on MS3, indicating better adaptability in the latter layers.

Conclusion

We introduce a more challenging unsupervised Audio-Visual Segmentation (AVS) problem, aiming to scaling its applicability. For performance benchmarks, we show that harnessing self-supervised audio-visual models leads is effective. Our proposed method, **M**odality **C**orrespondence **A**lignment (MoCA), features a novel pixel matching aggregation strategy for accurate pixel-audio associations. Extensive experiments demonstrate that MoCA outperforms well-designed methods and closely approaches fully supervised counterparts.

References

Afouras, T.; Owens, A.; Chung, J. S.; and Zisserman, A. 2020. Self-supervised learning of audio-visual objects from video. In *ECCV*.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. *arXiv preprint arXiv:2104.14294*.

Chen, H.; Xie, W.; Afouras, T.; Nagrani, A.; Vedaldi, A.; and Zisserman, A. 2021a. Localizing visual sounds the hard way. In *CVPR*.

Chen, H.; Xie, W.; Vedaldi, A.; and Zisserman, A. 2020. Vggsound: A large-scale audio-visual dataset. In *IEEE ICASSP*.

Chen, Y.; Xian, Y.; Koepke, A.; Shan, Y.; and Akata, Z. 2021b. Distilling audio-visual knowledge by compositional contrastive learning. In *CVPR*.

Gao, S.; Chen, Z.; Chen, G.; Wang, W.; and Lu, T. 2024. Avsegformer: Audio-visual segmentation with transformer. In *AAAI*.

Gemmeke, J. F.; Ellis, D. P.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R. C.; Plakal, M.; and Ritter, M. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *IEEE ICASSP*.

Georgescu, M.-I.; Fonseca, E.; Ionescu, R. T.; Lucic, M.; Schmid, C.; and Arnab, A. 2023. Audiovisual masked autoencoders. In *ICCV*.

Girdhar, R.; El-Nouby, A.; Liu, Z.; Singh, M.; Alwala, K. V.; Joulin, A.; and Misra, I. 2023. Imagebind: One embedding space to bind them all. In *CVPR*.

Girshick, R.; Donahue, J.; Darrell, T.; and Malik, J. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*.

Gong, Y.; Chung, Y.-A.; and Glass, J. 2021. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*.

Guzhov, A.; Raue, F.; Hees, J.; and Dengel, A. 2022. Audioclip: Extending clip to image, text and audio. In *IEEE ICASSP*.

Hamilton, M.; Zhang, Z.; Hariharan, B.; Snavely, N.; and Freeman, W. T. 2022. Unsupervised semantic segmentation by distilling feature correspondences. *ICLR*.

Hao, D.; Mao, Y.; He, B.; Han, X.; Dai, Y.; and Zhong, Y. 2023. Improving audio-visual segmentation with bidirectional generation. *arXiv preprint arXiv:2308.08288*.

Hartigan, J. A.; and Wong, M. A. 1979. Algorithm AS 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, 28(1): 100–108.

Hartley, R.; and Zisserman, A. 2003. *Multiple view geometry in computer vision*. Cambridge university press.

Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; De Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; and Gelly, S. 2019. Parameter-efficient transfer learning for NLP. In *ICML*.

Hu, X.; Chen, Z.; and Owens, A. 2022. Mix and localize: Localizing sound sources in mixtures. In *CVPR*.

Huang, P.-Y.; Sharma, V.; Xu, H.; Ryali, C.; Fan, H.; Li, Y.; Li, S.-W.; Ghosh, G.; Malik, J.; and Feichtenhofer, C. 2022. MAViL: Masked Audio-Video Learners. *arXiv preprint arXiv:2212.08071*.

Hwang, J.; Yu, S. X.; Shi, J.; Collins, M. D.; Yang, T.; Zhang, X.; and Chen, L. 2019. SegSort: Segmentation by Discriminative Sorting of Segments. *CoRR*, abs/1910.06962.

Jaegle, A.; Gimeno, F.; Brock, A.; Vinyals, O.; Zisserman, A.; and Carreira, J. 2021. Perceiver: General perception with iterative attention. In *ICML*.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. *arXiv preprint arXiv:2304.02643*.

Liu, C.; Li, P.; Zhang, H.; Li, L.; Huang, Z.; Wang, D.; and Yu, X. 2023a. BAVS: Bootstrapping Audio-Visual Segmentation by Integrating Foundation Knowledge. *arXiv preprint arXiv:2308.10175*.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Li, C.; Yang, J.; Su, H.; Zhu, J.; et al. 2023b. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Ma, S.; Zeng, Z.; McDuff, D.; and Song, Y. 2020. Active contrastive learning of audio-visual video representations. *arXiv preprint arXiv:2009.09805*.

Maaz, M.; Rasheed, H.; Khan, S.; Khan, F. S.; Anwer, R. M.; and Yang, M.-H. 2022. Class-agnostic object detection with multi-modal transformer. In *ECCV*.

Mao, Y.; Zhang, J.; Wan, Z.; Dai, Y.; Li, A.; Lv, Y.; Tian, X.; Fan, D.-P.; and Barnes, N. 2021. Transformer transforms salient object detection and camouflaged object detection. *arXiv preprint arXiv:2104.10127*, 1(2): 5.

Mao, Y.; Zhang, J.; Xiang, M.; Lv, Y.; Zhong, Y.; and Dai, Y. 2023. Contrastive conditional latent diffusion for audio-visual segmentation. *arXiv preprint arXiv:2307.16579*.

Mo, S.; and Morgado, P. 2022. Localizing visual sounds the easy way. In *EECCV*. Springer.

Mo, S.; and Tian, Y. 2023. AV-SAM: Segment anything model meets audio-visual localization and segmentation. *arXiv preprint arXiv:2305.01836*.

Nagrani, A.; Yang, S.; Arnab, A.; Jansen, A.; Schmid, C.; and Sun, C. 2021. Attention bottlenecks for multimodal fusion. In *NeurIPS*.

Nunez, E.; Jin, Y.; Rastegari, M.; Mehta, S.; and Horton, M. 2023. Diffusion Models as Masked Audio-Video Learners. *arXiv preprint arXiv:2310.03937*.

Pratt, W. K. 2007. *Digital image processing: PIKS Scientific inside*, volume 4. Wiley Online Library.

Prince, S. J. 2012. *Computer vision: models, learning, and inference*. Cambridge University Press.

Qian, R.; Hu, D.; Dinkel, H.; Wu, M.; Xu, N.; and Lin, W. 2020. Multiple sound sources localization from coarse to fine. In *ECCV*. Springer.

Rouditchenko, A.; Zhao, H.; Gan, C.; McDermott, J.; and Torralba, A. 2019. Self-supervised audio-visual co-segmentation. In *IEEE ICASSP*.

Shi, Z.; Wu, Q.; Li, H.; Meng, F.; and Xu, L. 2023. Cross-modal Cognitive Consensus guided Audio-Visual Segmentation. *arXiv preprint arXiv:2310.06259*.

Szeliski, R. 2022. *Computer vision: algorithms and applications*. Springer Nature.

Trucco, E.; and Verri, A. 1998. *Introductory techniques for 3-D computer vision*, volume 201. Prentice Hall Englewood Cliffs.

Van Gansbeke, W.; Vandenhende, S.; Georgoulis, S.; and Van Gool, L. 2021. Unsupervised Semantic Segmentation by Contrasting Object Mask Proposals. *arxiv preprint arxiv:2102.06191*.

Wu, H.-H.; Seetharaman, P.; Kumar, K.; and Bello, J. P. 2022. Wav2clip: Learning robust audio representations from clip. In *IEEE ICASSP*.

Yang, H.; Ma, C.; Wen, B.; Jiang, Y.; Yuan, Z.; and Zhu, X. 2024a. Recognize any regions. In *NeurIPS*.

Yang, Q.; Nie, X.; Li, T.; Gao, P.; Guo, Y.; Zhen, C.; Yan, P.; and Xiang, S. 2024b. Cooperation Does Matter: Exploring Multi-Order Bilateral Relations for Audio-Visual Segmentation. In *CVPR*.

Zhang, J.; Xie, J.; Barnes, N.; and Li, P. 2021. Learning generative vision transformer with energy-based latent space for saliency prediction. *NeurIPS*.

Zhou, J.; Guo, D.; and Wang, M. 2022. Contrastive positive sample propagation along the audio-visual event line. *IEEE TPAMI*.

Zhou, J.; Shen, X.; Wang, J.; Zhang, J.; Sun, W.; Zhang, J.; Birchfield, S.; Guo, D.; Kong, L.; Wang, M.; et al. 2023. Audio-Visual Segmentation with Semantics. *arXiv preprint arXiv:2301.13190*.

Zhou, J.; Wang, J.; Zhang, J.; Sun, W.; Zhang, J.; Birchfield, S.; Guo, D.; Kong, L.; Wang, M.; and Zhong, Y. 2022. Audio–visual segmentation. In *ECCV*.

Zlateski, A.; Jaroensri, R.; Sharma, P.; and Durand, F. 2018. On the importance of label quality for semantic segmentation. In *CVPR*.