# Revisiting Audio-Visual Segmentation with Vision-Centric Transformer

Shaofei Huang[1,2*]    Rui Ling[3]    Tianrui Hui[1†]    Hongyu Li[4]

Xu Zhou[5]    Shifeng Zhang[5]    Si Liu[4†]    Richang Hong[1]    Meng Wang[1]

[1]School of Computer Science and Information Engineering, Hefei University of Technology

[2]Institute of Information Engineering, Chinese Academy of Sciences

[3]School of Computer Science and Engineering, Beihang University

[4]School of Artificial Intelligence, Beihang University    [5]Sangfor Technologies

## Abstract

*Audio-Visual Segmentation (AVS) aims to segment sound-producing objects in video frames based on the associated audio signal. Prevailing AVS methods typically adopt an audio-centric Transformer architecture, where object queries are derived from audio features. However, audio-centric Transformers suffer from two limitations: perception ambiguity caused by the mixed nature of audio, and weakened dense prediction ability due to visual detail loss. To address these limitations, we propose a new Vision-Centric Transformer (VCT) framework that leverages vision-derived queries to iteratively fetch corresponding audio and visual information, enabling queries to better distinguish between different sounding objects from mixed audio and accurately delineate their contours. Additionally, we also introduce a Prototype Prompted Query Generation (PPQG) module within our VCT framework to generate vision-derived queries that are both semantically aware and visually rich through audio prototype prompting and pixel context grouping, facilitating audio-visual information aggregation. Extensive experiments demonstrate that our VCT framework achieves new state-of-the-art performances on three subsets of the AVSBench dataset.*

## 1. Introduction

Audio-Visual Segmentation (AVS) [43, 44] is designed to identify and generate pixel-level segmentation masks of sound-emitting objects within visual frames by utilizing audio signals from the video. This task requires models to determine regions associated with the sound source and to delineate the shape and boundaries of the sound-producing objects precisely. Compared to Sound Source Localization
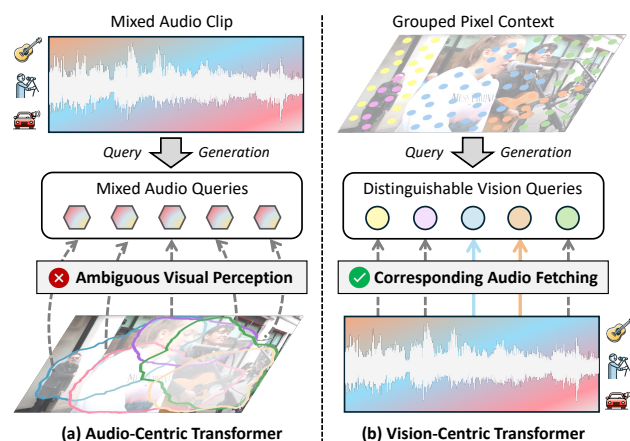


Figure 1. Motivation of our method. (a) Previous methods typically rely on audio-centric Transformers, where queries are derived from mixed sources of audio features, causing ambiguity in the sounding object perception process. (b) We propose the Vision-Centric Transformer, where queries are derived from visual features of different regions and interact with audio features to fetch corresponding audio information, achieving more accurate sound source distinguishment and object segmentation.

(SSL) [13, 14, 33], AVS places greater emphasis on fine-grained audio-visual understanding and pixel-level perception accuracy. Based on these properties, the AVS task holds potential value in various application scenarios, including video understanding and editing, augmented and virtual reality, intelligent surveillance systems, *etc*.

Prevailing AVS methods [9, 21, 37, 41] typically employ an **audio-centric** Transformer architecture [4, 35], where audio features are used as or integrated into object queries to identify the sounding objects in the video through layer-by-layer audio-visual interactions. *However, audio-centric Transformers present two main limitations.* First, audio information is typically a mixture of various sound sources in real-world scenarios, including both semantically distinct

sounds from in-screen sources and background noises from off-screen sources. For example, the sample audio shown in Figure 1 contains mixed sounds from voice, guitar, and cars, with the last one not visible in the video frames. This inherent mixture in audio causes potential confusion in perception for audio-derived queries, making them difficult to distinguish due to mutual interference. Furthermore, noises from off-screen sources may introduce additional disturbance, causing audio-derived queries to yield false-positive predictions. Second, AVS is essentially a vision-centric dense prediction task, where an object query must encompass both abstract audio semantics to identify the existence of the target sounding objects and concrete visual details to delineate their contours accurately. However, audio-derived queries initially contain abstract audio semantics only. The delayed integration of visual information may lead to the loss of concrete visual details, weakening the ability to make dense predictions.

To mitigate the aforementioned limitations of audio-centric Transformers, we centralize visual information in the query design and propose a new **Vision-Centric** Transformer (VCT) framework. Within the VCT framework, object queries are initially derived from different image regions with abundant visual details, each serving as a candidate for a sounding object. Through interactions with both audio and visual features across multiple Transformer decoder layers, these queries gradually become aware of their *corresponding sound information* and *fine-grained visual features*. The former aids in determining whether an object is sound-emitting, while the latter helps in capturing the object's contours, thereby enabling comprehensive perception. As shown in Figure 1(b), these vision-derived queries, which focus on semantically distinct visual regions, can independently integrate their corresponding sound information from mixed audio, effectively excluding unrelated or noisy sound sources during interaction with audio features in the Transformer decoder layers. By focusing on semantically relevant visual regions and sound information, vision-derived queries maintain superior discriminative ability, enabling them to better distinguish between different sounding objects in the scene and thereby alleviating the perceptual ambiguity caused by mixed audio.

Furthermore, to generate vision-derived queries that are both *semantically aware* and *visually rich*, we also introduce a novel Prototype Prompted Query Generation (PPQG) module within our VCT framework, thus laying a solid foundation for effective audio-visual information aggregation. For the semantical aspect, we first define a set of prototypes for different categories of audio events. Through contrastive learning with audio features, each prototype learns the corresponding acoustic semantic information from the audio signals. The vision-derived queries are then prompted by these prototypes, making them aware of

the potential audio events before interacting with the associated audio information. This enables more targeted fetching of relevant audio information during subsequent interactions with the audio features. For the visual aspect, since the queries are generated from visual features, they already contain a considerable amount of visual information. What we need to do is enable vision-derived queries to initially focus on different regions, thus further enhancing their discriminative ability in the subsequent Transformer decoder. To this end, we group pixel contexts into queries using an attention mechanism with a hard assignment, which encourages the perceptual focus of the queries to become more diverse. With the PPQG module, our VCT framework can better harness the advantages of the vision-centric paradigm.

Contributions of our paper are summarized as follows: (1) We propose a new Vision-Centric Transformer (VCT) framework that leverages vision-derived queries to fetch corresponding audio information and fine-grained visual features, thus realizing a comprehensive perception of sounding objects. (2) We also introduce a novel Prototype Prompted Query Generation (PPQG) module to generate vision-derived queries that focus on different image regions with the awareness of occurring audio event categories, facilitating later information aggregation. (3) Extensive experiments on three subsets of the AVSBench dataset demonstrate that our method outperforms previous state-of-the-art methods.

## 2. Related Works

### 2.1. Sound Source Localization

Sound Source Localization (SSL) [2, 3, 13, 14, 18, 31, 33] focuses on determining the location of sound-producing objects within a video by identifying visual areas that correspond to audio cues. This task aims to achieve a localized understanding of the scene at a patch level without emphasizing the exact shape of the sounding objects. Hu *et al.* [13] leverage audiovisual consistency as a self-supervised signal to match the category distribution differences between audio and visual inputs, effectively isolating sounding objects and filtering out silent ones. Unsupervised SSL is explored in EZ-VSL [31] where a multiple-instance contrastive learning framework and an object-guided localization scheme are proposed to improve localization accuracy. Kim *et al.* [18] achieve the localization of multiple sound sources without prior knowledge by iteratively identifying objects and applying object similarity-aware clustering loss. Compared to SSL, AVS methods usually require more accurate visual representations to segment the sounding objects.

### 2.2. Audio-Visual Segmentation

Different from sound source localization, Audio-Visual Segmentation (AVS) [43, 44] aims to localize the sound-
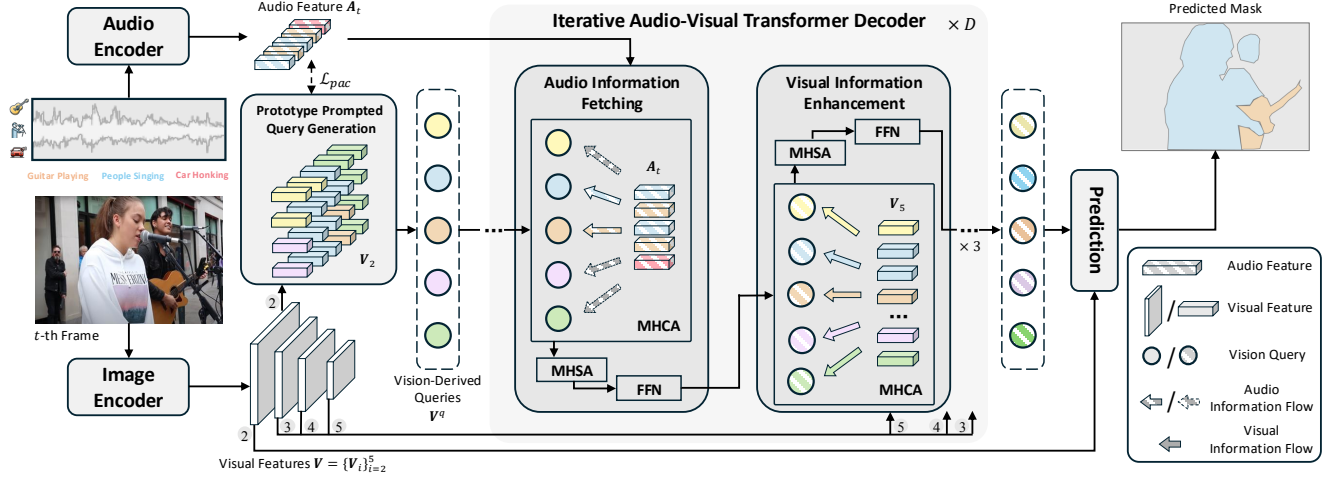
Figure 2. Overall architecture of our proposed VCT. The audio feature and multi-scale visual features are first extracted by the encoders. Taking the largest visual feature as input, our proposed Prototype Prompted Query Generation (PPQG) module generates vision-derived queries, which are then fed into the iterative audio-visual Transformer decoder. Through consecutive audio information fetching and visual information enhancement blocks, vision-derived queries are further refined to yield the final segmentation masks and categories.

emitting objects with a more precise segmentation mask rather than a coarse heat map. Some methods adopt the fusion-based architecture [10, 29, 43, 44] which first performs a multimodal fusion of features from the audio and visual encoders, then input the fused feature into a decoder that conducts per-pixel classification to generate the predicted mask. For example, TPAVI [43, 44] attentively aggregates the audio features with visual features from different stages of the encoder and performs an FPN-style fusion through the decoder. Other Transformer-based methods [6, 9, 15, 16, 20–24, 37, 41] follow the mask classification paradigm of Mask2Former [7] where queries are derived largely from audio information to discover sounding objects through Transformer decoder layers. For example, COMBO [41] enhances learnable queries by summing with audio features and performs bilateral fusion with visual features to construct more comprehensive queries. We refer to the framework of these methods as audio-centric Transformers, which suffer from ambiguous perception and visual detail loss. Apart from improvements in network architecture, recent works [25, 34, 38] also focus on the intrinsic analysis of the AVS task and data. In this paper, we propose a vision-centric Transformer framework to address the limitations of ambiguous perception and visual detail loss introduced by audio-centric Transformers.

### 2.3. Query Designs for Transformers

The three key components of the attention mechanism in Transformer [35] models are query, key, and value, where the role of the query is to search for the most important information in the input sequence for the current position. DETR [4] designs learnable object queries with im-

age features as keys and values to detect object instances. DINO [42] further introduces positional and content queries to more effectively capture both positional and categorical information for better identifying targets in complex scenes. For multimodal tasks, ReferFormer [39] and AQ-Former [15] exploit language and audio as queries respectively to distinguish the referred objects by directly aggregating visual features relevant to the multimodal cues. In this paper, we centralize visual information in the query design with a PPQG module to generate vision-derived queries that are both semantically aware and visually rich.

## 3. Method

The overall framework of our proposed VCT is illustrated in Figure 2. Given an input video $\mathcal{V} \in \mathbb{R}^{T \times H \times W \times 3}$ with $T$ frames and an audio clip $\mathcal{A}$, AVS aims to produce the segmentation masks $\mathcal{Y} \in \mathbb{R}^{T \times H \times W \times K}$ of sounding objects, where $K$ is the number of audio event categories ($K = 1$ for category-agnostic settings, $e.g.$, S4 and MS3). We first feed the audio clip and multiple video frames into an audio encoder and an image encoder respectively to obtain the corresponding audio feature and multi-scale visual features. The largest visual feature is then fed into our PPQG module to generate vision-derived queries, which interact with audio features and visual features iteratively in the Transformer decoder for further refinement. Mask and category predictions are obtained based on refined vision-derived queries.

### 3.1. Visual and Audio Feature Extraction

For each input frame, we adopt a visual encoder ($e.g.$, Swin Transformer [26]) to extract multi-scale image features $\boldsymbol{V} = \{\boldsymbol{V}_i\}_{i=2}^5$, where $\boldsymbol{V}_i \in \mathbb{R}^{H_i \times W_i \times C_i^v}$, $H_i = \frac{H}{2^i}$,

$W_i = \frac{W}{2^i}$, and $C_i^v$ are the height, width, and channel numbers of visual feature from the $i$-th stage of the visual encoder. For the audio clip associated with the input video, we follow TPAVI [43] to utilize VGGish [12] for audio feature extraction. The original audio clip $\mathcal{A}$ is first converted to a mono output $\mathcal{A}_m \in \mathbb{R}^{L \times 96 \times 64}$ with a resampling rate of 16 kHz to standardize the audio data, where $L$ varies according to the audio duration. Then, we compute the time-frequency representation of $\mathcal{A}_m$ by applying Short-Time Fourier Transform to obtain its mel-spectrogram $\mathcal{A}_s \in \mathbb{R}^{T \times 96 \times 64}$, where $T$ is the number of frames. $\mathcal{A}_s$ is further fed into the VGGish model with the final compression operation removed to extract the audio feature $\boldsymbol{A} \in \mathbb{R}^{T \times S \times C^a}$ where $S$ denotes the size of the mel-spectrogram feature, $C^a$ denotes its channel number.
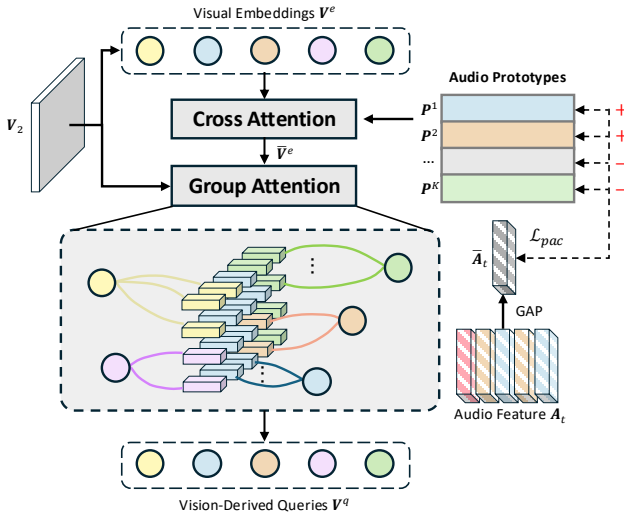


Figure 3. Details of Prototype Prompted Query Generation.

## 3.2. Prototype Prompted Query Generation

The goal of the Prototype Prompted Query Generation (PPQG) module is to generate vision-derived queries that contain both rich visual details and audio semantics. As shown in Figure 3, the computation process of our PPQG module mainly consists of three steps.

**Visual embedding aggregation.** We first perform spatial information aggregation on the high-resolution image feature $\boldsymbol{V}_2$ from the 2-nd stage of the encoder to obtain a set of visual embeddings. In detail, $\boldsymbol{V}_2$ is first projected into a hidden space by several convolution layers which convert its channel number from $C_2^v$ to $C^h$:

$$\boldsymbol{V}^h = \text{Conv}_{1 \times 1}(\delta(\text{Conv}_{3 \times 3}(\delta(\text{Conv}_{1 \times 1}(\boldsymbol{V}_2))))), \quad (1)$$

where $\delta$ denotes the ReLU activation. Then, we adjust the shape of $\boldsymbol{V}^h$ from $\mathbb{R}^{H_2 \times W_2 \times C^h}$ to $\mathbb{R}^{C^h \times H_2 W_2}$ and apply a three-layer MLP on it to aggregate spatial information by

regarding $H_2 W_2$ as the input channel dimension for MLP:

$$\boldsymbol{V}^e = \text{Reshape}(\text{MLP}(\text{Reshape}(\boldsymbol{V}^h))), \quad (2)$$

where the output of MLP has the shape of $\mathbb{R}^{C^h \times N}$, $\boldsymbol{V}^e \in \mathbb{R}^{N \times C^h}$ is a set of visual embeddings after another shape adjustment and $N$ is the number of visual embeddings.

**Audio prototype prompting.** To enable more targeted audio information fetching of vision-derived queries in the subsequent Transformer decoder, we define a group of audio prototypes to prompt them with audio event categories present in the scene during the query generation. In detail, we randomly initialize a set of learnable parameters as the audio prototypes $\boldsymbol{P} \in \mathbb{R}^{K \times C^h}$, where $K$ is the number of audio event categories, to prompt visual embeddings $\boldsymbol{V}^e$ with audio category priors by cross-attention:

$$\bar{\boldsymbol{V}}^e = \boldsymbol{V}^e + \text{Softmax}\left(\frac{(\boldsymbol{V}^e \boldsymbol{W}_1^q)(\boldsymbol{P} \boldsymbol{W}_1^k)^{\text{T}}}{\sqrt{C^h}}\right)(\boldsymbol{P} \boldsymbol{W}_1^v). \quad (3)$$

$\boldsymbol{W}_1^q$, $\boldsymbol{W}_1^k$ and $\boldsymbol{W}_1^v$ are projection weights. For clarity, we omit other operations like LN and FFN and use $\bar{\boldsymbol{V}}^e$ to denote the output of cross-attention.

Since the audio prototypes are randomly initialized, we also design a prototype-audio contrastive loss $\mathcal{L}_{\text{pac}}$ to ensure that the prototypes can learn semantic information of different audio event categories from the audio features. Concretely, let $\boldsymbol{A}_t \in \mathbb{R}^{S \times C^a}$ denotes the audio feature associated with the current frame and we perform linear projection followed by global average pooling on $\boldsymbol{A}_t$ to obtain $\bar{\boldsymbol{A}}_t \in \mathbb{R}^{C^h}$. An inner product between $\bar{\boldsymbol{A}}_t$ and $\boldsymbol{P}$ yields a matching prediction $\boldsymbol{M} \in \mathbb{R}^K$ representing the likelihood of each audio category prototype present in the current audio. The matching ground truth $\boldsymbol{M}^*$ indicating which audio event categories are actually present in the current audio can be obtained from the dataset annotation. We implement $\mathcal{L}_{\text{pac}}$ as a BCE loss between $\boldsymbol{M}$ and $\boldsymbol{M}^*$:

$$
\begin{aligned}
\mathcal{L}_{\text{bce}}(\boldsymbol{M}_k, \boldsymbol{M}_k^*) = & -\boldsymbol{M}_k^* \log(\boldsymbol{M}_k) \\
& - (1 - \boldsymbol{M}_k^*) \log(1 - \boldsymbol{M}_k),
\end{aligned} \quad (4)
$$

$$\mathcal{L}_{\text{pac}}(\boldsymbol{M}, \boldsymbol{M}^*) = \frac{1}{K} \sum_{k=1}^{K} \mathcal{L}_{\text{bce}}(\boldsymbol{M}_k, \boldsymbol{M}_k^*). \quad (5)$$

The $\mathcal{L}_{\text{pac}}$ pulls the audio features closer to the present audio prototypes, while pushing the audio features further away from non-existent audio prototypes.

**Pixel context grouping.** Inspired by GroupViT [40], to obtain more distinguishable vision-derived queries that can focus on different image regions, we adopt Gumbel-Softmax [17, 28] to group pixel context for each visual embedding with a hard and differentiable assignment:

$$\boldsymbol{R} = \text{Softmax}((\bar{\boldsymbol{V}}^e \boldsymbol{W}_2^q)(\boldsymbol{V}^h \boldsymbol{W}_2^k)^{\text{T}} + \boldsymbol{G}), \quad (6)$$

$$\hat{R} = \text{One-hot}(\arg\max_{N}(R)) + R - \text{sg}(R), \quad (7)$$

where $W_2^q$ and $W_2^k$ are linear projection weights, $G \in \mathbb{R}^{N \times H_2 W_2}$ are i.i.d random samples drawn from the Gumbel$(0, 1)$ distribution. We apply $\arg\max$ on the $N$ dimension of $R \in \mathbb{R}^{N \times H_2 W_2}$ to find the index of the visual embedding with the highest similarity to each pixel feature as the assignment target. One-hot converts the index for each pixel feature into a one-hot vector. sg is the stop gradient operator. The resulting $\hat{R} \in \mathbb{R}^{N \times H_2 W_2}$ contains the hard assignments between visual embeddings and pixel features. We use $\hat{R}$ to group the pixel context in image feature $V^h$ into prototype-prompted visual embeddings $\bar{V}^e$:

$$V^q = \bar{V}^e + (\text{Norm}(\hat{R})(V^h W_2^v))W^o, \quad (8)$$

where $W_2^v$ and $W^o$ are projection weights, Norm is normalization on the $H_2 W_2$ dimension of $\hat{R}$ by dividing the sum of values, $V^q \in \mathbb{R}^{N \times C^h}$ denotes the vision-derived queries generated by our PPQG module.

### 3.3. Iterative Audio-Visual Transformer Decoder

In our designed iterative audio-visual Transformer decoder, vision-derived queries interact with the audio feature and multi-scale visual features iteratively through two kinds of Transformer blocks, each composed of a multi-head cross-attention (MHCA) layer, a multi-head self-attention (MHSA) layer, and an FFN layer. In the audio information fetching block, each vision-derived query interacts with the audio feature of the current frame through MHCA, with $A_t$ serving as the key and value. MHCA allows each vision-derived query to obtain the corresponding sound information for the region it represents, enabling the determination of whether an object in that region is emitting sound and its audio category. In the visual information enhancement blocks, each vision-derived query sequentially interacts with three different scales of visual features from the current frame, with $V_5$, $V_4$ and $V_3$ serving as the key and value for MHCA respectively. Here, we follow Mask2Former [7] to use the segmentation mask predicted in the previous layer as the attention mask for the current layer, restricting the MHCA's attention to potential sounding image regions. Through visual information enhancement, each vision-derived query captures finer-grained visual features, allowing for more accurate mask predictions.

We use $\mathcal{U} = \{A_t, V_5, V_4, V_3\}$ as an interaction unit where one audio information fetching block and three visual information enhancement blocks are stacked in order. We repeat $\mathcal{U}$ for $D$ times and append one audio block at the end to construct our iterative audio-visual Transformer decoder. Through the above interaction, vision-derived queries gradually become aware of their corresponding sound information and fine-grained visual features, thus better distinguish-

ing between different sounding objects and mitigating the perception ambiguity caused by mixed audio.

### 3.4. Loss Functions

Following Mask2Former [7], we use a classification head and a mask head to obtain category and segmentation predictions. Our loss functions mainly consist of three parts, *i.e.*, classification loss $\mathcal{L}_{\text{cls}}$, mask loss $\mathcal{L}_{\text{mask}}$, and our proposed prototype-audio contrastive loss $\mathcal{L}_{\text{pac}}$. The classification loss is namely CE loss. Apart from BCE loss, the mask loss also contains Dice loss [30] to handle relatively small foreground areas in the images. Therefore, the overall loss function of our method is formulated as:

$$\mathcal{L} = \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{mask}}\mathcal{L}_{\text{mask}} + \lambda_{\text{pac}}\mathcal{L}_{\text{pac}}, \quad (9)$$

where $\lambda_{\text{cls}}$, $\lambda_{\text{mask}}$ and $\lambda_{\text{pac}}$ are loss coefficients.

## 4. Experiments

### 4.1. Dataset and Evaluation Metrics

We conduct experiments on the AVSBench dataset [43, 44] to train and evaluate our models. The AVSBench dataset is organized into three subsets: the Single-source subset (S4), the Multi-sources subset (MS3), and the Semantic-labels subset (AVSS). Built on the previous two subsets, the AVSS subset substantially extends the dataset in both quantity and semantic labels, which brings more challenges by requiring models not only to produce segmentation masks but also to categorize each sounding object.

Regarding evaluation metrics, we follow prior works [41, 43, 44] to adopt $\mathcal{M}_{\mathcal{J}}$ and $\mathcal{M}_{\mathcal{F}}$ for performance measurement. $\mathcal{M}_{\mathcal{J}}$ is the Jaccard index which computes the Intersection over Union (IoU) between predicted segmentation and ground truth. On the AVSS subset, $\mathcal{M}_{\mathcal{J}}$ is the mean IoU averaged over all categories. $\mathcal{M}_{\mathcal{F}}$ denotes the F-score which combines the precision and recall of the segmentation results. Its formulation is $\mathcal{M}_{\mathcal{F}} = \frac{(1+\beta^2) \cdot \text{Precision} \cdot \text{Recall}}{\beta^2 \cdot \text{Precision} + \text{Recall}}$ where $\beta^2 = 0.3$ indicating more emphasis on the recall.

### 4.2. Implementation Details

Our model is implemented with PyTorch. For a fair comparison with previous methods [27, 38, 41], we adopt ResNet-50 [11] and Pyramid Vision Transformer (PVT-v2) [36] pretrained on ImageNet [8] and Swin Transformer (Swin-Base) [26] pretrained on ImageNet-21K [32] as visual encoders. The audio encoder is VGGish [12] pretrained on YouTube-8M [1]. Following COMBO [41], we also adopt Semantic-SAM [19] to generate colored class-agnostic Maskige as segmentation prior knowledge and fuse its feature with the original visual feature in the same way as COMBO. The values of $C^h$ and $S$ are 256 and 24. The number of vision-derived queries $N = 100$. $D = 2$ in the

| Method | Reference | Backbone | Image Size | AVS-Semantic (AVSS) | | Single-Source (S4) | | Multi-Source (MS3) | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | $\mathcal{M}_{\mathcal{J}}$ | $\mathcal{M}_{\mathcal{F}}$ | $\mathcal{M}_{\mathcal{J}}$ | $\mathcal{M}_{\mathcal{F}}$ | $\mathcal{M}_{\mathcal{J}}$ | $\mathcal{M}_{\mathcal{F}}$ |
| TPAVI [43, 44] | [ECCV'22] | ResNet-50 | 224×224 | - | - | 72.8 | 84.8 | 46.9 | 57.8 |
| | | PVT-v2 | | 29.8 | 35.2 | 78.7 | 87.9 | 54.0 | 64.5 |
| AQFormer [15] | [IJCAI'23] | ResNet-50 | 224×224 | - | - | 77.0 | 86.4 | 55.7 | 66.9 |
| | | PVT-v2 | | - | - | 81.6 | 89.4 | 61.1 | <u>72.1</u> |
| ECMVAE [29] | [ICCV'23] | ResNet-50 | 224×224 | - | - | 76.3 | 86.5 | 48.7 | 60.7 |
| | | PVT-v2 | | - | - | 81.7 | 90.1 | 57.8 | 70.8 |
| CATR [21] | [ACMMM'23] | ResNet-50 | 224×224 | - | - | 74.8 | 86.6 | 52.8 | 65.3 |
| | | PVT-v2 | | 32.8 | 38.5 | 81.4 | 89.6 | 59.0 | 70.0 |
| AVSC [23] | [ACMMM'23] | ResNet-50 | 224×224 | - | - | 77.0 | 85.2 | 49.6 | 61.5 |
| | | PVT-v2 | | - | - | 80.6 | 88.2 | 58.2 | 65.1 |
| BAVS [24] | [TMM'24] | ResNet-50 | 224×224 | 24.7 | 29.6 | 78.0 | 85.3 | 50.2 | 62.4 |
| | | PVT-v2 | | 32.6 | 36.4 | 82.0 | 88.6 | 58.6 | 65.5 |
| | | Swin-B | | <u>33.6</u> | <u>37.5</u> | <u>82.7</u> | <u>89.8</u> | <u>59.6</u> | <u>65.9</u> |
| AVSBG [10] | [AAAI'24] | ResNet-50 | 224×224 | - | - | 74.1 | 85.4 | 45.0 | 56.8 |
| | | PVT-v2 | | - | - | 81.7 | 90.4 | 55.1 | 66.8 |
| GAVS [37] | [AAAI'24] | ViT-B | 224×224 | - | - | 80.1 | 90.2 | 63.7 | 77.4 |
| AVSF [9] | [AAAI'24] | ResNet-50 | 224×224 | 24.9 | 29.3 | 76.5 | 85.9 | 49.5 | 62.8 |
| | | PVT-v2 | | 36.7 | 42.0 | 82.1 | 89.9 | 58.4 | 69.3 |
| UFE [25] | [CVPR'24] | ResNet-50 | 224×224 | - | - | 79.0 | 87.5 | 55.9 | 64.5 |
| | | PVT-v2 | | - | - | 83.2 | 90.4 | **62.0** | 70.9 |
| CAVP [5] | [CVPR'24] | ResNet-50 | 224×224 | 30.4 | 35.3 | 78.8 | 88.9 | 55.8 | 67.1 |
| COMBO [41] | [CVPR'24] | ResNet-50 | 224×224 | 33.3 | 37.3 | <u>81.7</u> | 90.1 | 54.5 | 66.6 |
| | | PVT-v2 | | <u>42.1</u> | 46.1 | <u>84.7</u> | <u>91.9</u> | 59.2 | 71.2 |
| CPM [6] | [ECCV'24] | ResNet-50 | 224×224 | <u>34.5</u> | <u>39.6</u> | 81.4 | <u>90.5</u> | <u>59.8</u> | <u>71.0</u> |
| TeSO [38] | [ECCV'24] | Swin-B | 384×384 | 39.0 | 45.1 | <u>83.3</u> | <u>93.3</u> | 66.0 | 80.1 |
| SelM [20] | [ACMMM'24] | ResNet-50 | 224×224 | 31.9 | 37.2 | 76.6 | 86.2 | 54.5 | 65.6 |
| | | PVT-v2 | | 41.3 | <u>46.9</u> | 83.5 | 91.2 | 60.3 | 71.3 |
| AVSBias [34] | [ACMMM'24] | Swin-B | 384×384 | <u>44.4</u> | <u>49.9</u> | <u>83.3</u> | 93.0 | <u>67.2</u> | <u>80.8</u> |
| **VCT (Ours)** | - | ResNet-50 | 224×224 | **37.5** | **42.2** | **81.8** | **90.6** | **61.9** | **74.7** |
| | | PVT-v2 | 224×224 | **44.7** | **49.5** | **84.8** | **92.1** | **62.0** | **75.0** |
| | | Swin-B | 224×224 | **47.9** | **52.9** | **84.7** | **92.3** | **67.5** | **79.3** |
| | | Swin-B | 384×384 | **51.2** | **55.5** | **86.2** | **93.4** | **67.6** | **81.4** |

Table 1. Comparison with state-of-the-art methods on three subsets of the AVSBench dataset. Our method obtains significant performance gains on the most challenging AVSS subset. Best results in **bold** and 2nd best <u>underlined</u> for different backbones and image sizes.

iterative audio-visual Transformer decoder. $\lambda_{cls}$, $\lambda_{mask}$, and $\lambda_{pac}$ are set to 2, 5, and 1. The AdamW optimizer is used with the learning rate of $1e^{-4}$. We train our model with a batch size of 16 for ResNet-50 and 8 for PVT-v2 and Swin-B. The training iterations are 45K, 40K, and 45K for S4, MS3, and AVSS subsets respectively.

### 4.3. Comparison with State-of-the-art Methods

In Table 1, we show the performance comparison between our method and previous state-of-the-art methods on the three subsets of the AVSBench dataset. Our method outperforms the previous best-performing approaches across three subsets with different visual backbones and image sizes, demonstrating its ability to effectively differentiate multiple sound sources, accurately delineate the corresponding object shapes, and recognize their categories. In particular, on the most challenging AVSS subset, our method achieves even better $\mathcal{M}_{\mathcal{J}}$ performance with a PVT-v2 backbone and a $224 \times 224$ image size compared to AVSBias [34] using a Swin-B backbone and a $384 \times 384$ image size. This also indicates that our new architecture, which integrates vision-derived queries and the iterative audio-visual Transformer, demonstrates a clear advantage over the previous architectures that rely on audio-derived queries to interact with vi-

sual information. Moreover, on the S4 subset where performance is relatively plateaued, our method achieves 86.2 $\mathcal{M}_{\mathcal{J}}$, indicating that it is better at handling some of the hard cases compared to previous methods.

### 4.4. Ablation Studies

We conduct ablation studies on the AVSS subset with ResNet-50 as the visual backbone to verify the effectiveness of different designs in our method.

| Framework | Queries | $\mathcal{M}_{\mathcal{J}}$ | $\mathcal{M}_{\mathcal{F}}$ |
|---|---|---|---|
| ACT | Audio-Derived Queries | 33.2 | 37.0 |
| VCT | Naive Vision Queries | 35.2 | 39.3 |
| | *w/* Cross-Attention | 35.8 | 39.8 |
| | *w/* Group-Attention | 36.3 | 40.5 |
| | ***w/* Audio Prototypes** | **37.5** | **42.2** |

Table 2. Ablation study of generating vision-derived queries in our PPQG module on the AVSS subset.

**Component analysis.** We include the ablation study results of generating vision-derived queries in Table 2 to show the influence of different operations in our proposed PPQG module. "ACT" denotes the baseline model with audio-derived queries where audio features largely enhance the initial learnable queries for later audio-visual interaction. In our VCT framework, "Naive Vision Queries" represents that the vision-derived queries are generated only by the first step of visual embedding aggregation in the PPQG module. It yields better performance than audio-derived queries, which indicates that using visual features as the basis for locating sound-emitting objects, compared to introducing entangled audio information, enables better visual-audio correspondence during subsequent decoder interactions. "*w/* Cross-Attention" denotes that we use cross-attention with normal softmax to incorporate image features into naive vision-derived queries, while "*w/* Group-Attention" means replacing cross-attention with group-attention equipped by Gumbel-softmax to group pixel contexts for vision-derived queries using hard assignment. The performance improvements brought by both attention mechanisms show that incorporating additional visual region information helps to obtain more distinguishable vision-derived queries. The last row demonstrates that providing audio category priors to vision-derived queries enables them to more effectively fetch the corresponding audio information in the subsequent iterative audio-visual Transformer decoder.

**Audio prototype prompting.** Table 3 summarizes the ablation results of audio prototype prompting in our PPQG module. By comparing the first and second rows, it can be observed that without loss constraints, the randomly initialized audio prototypes fail to learn meaningful audio event category priors, resulting in no positive impact on perfor-

| Method | $\mathcal{M}_{\mathcal{J}}$ | $\mathcal{M}_{\mathcal{F}}$ |
|---|---|---|
| *w/o* Audio Prototypes | 36.3 | 40.5 |
| *w/* Audio Prototypes, *w/o* loss | 36.3 | 40.4 |
| *w/* Audio Prototypes, *w/* visual loss | 36.5 | 40.8 |
| ***w/* Audio Prototypes, *w/* PAC loss** | **37.5** | **42.2** |

Table 3. Ablation study of audio prototype prompting in PPQG.

mance. In the third row, "*w/* visual loss" denotes that we compute contrastive loss between audio prototypes and positive vision-derived queries that matched with the ground truth, bringing a small performance gain. In contrast, introducing our designed PAC loss in the last row results in an obvious performance improvement, suggesting that audio prototypes, through contrastive learning with audio features, can learn more explicit and clear audio event category priors than those learned from positive vision queries.

| | Audio Feature | $\mathcal{M}_{\mathcal{J}}$ | $\mathcal{M}_{\mathcal{F}}$ |
|---|---|---|---|
| **Visual Feature** | Multiply | 33.9 | 37.9 |
| | Concatenation | 35.3 | 39.3 |
| | Addition | 36.3 | 40.5 |
| **Audio Proto** | Replace | 36.6 | 40.8 |
| **VCT** | **Full Model** | **37.5** | **42.2** |

Table 4. Ablation study of audio feature incorporation.

**Audio feature incorporation.** We also explore different approaches to incorporate audio features into our VCT framework and present the results in Table 4. The first three rows show how we fuse audio features with visual features using multiplication, concatenation, and addition, which are then used as the key and value in the cross-attention of the Transformer decoder. The fourth row denotes that we replace the audio prototypes in PPQG with the audio feature and the PAC loss is hence removed. Our full VCT obtains the best performance compared to these approaches, indicating that pre-fetching audio features during the generation of vision-derived queries or fetching fused multimodal features in the Transformer decoder can lead to queries learning inaccurate audio-visual correspondence. Additional ablation results are included in the supplementary materials.

### 4.5. Qualitative Analysis

Figure 4 presents the qualitative comparison between our full model and the ACT baseline. In the left example, the input audio is the voiceover narration. Compared to the baseline, our full model is able to recognize the off-screen background audio rather than biasing its focus on the main objects in the video, demonstrating stronger robustness. In the right example, the baseline confuses between the girl and woman, while our full model correctly distinguishes between them and identifies the actual sound-
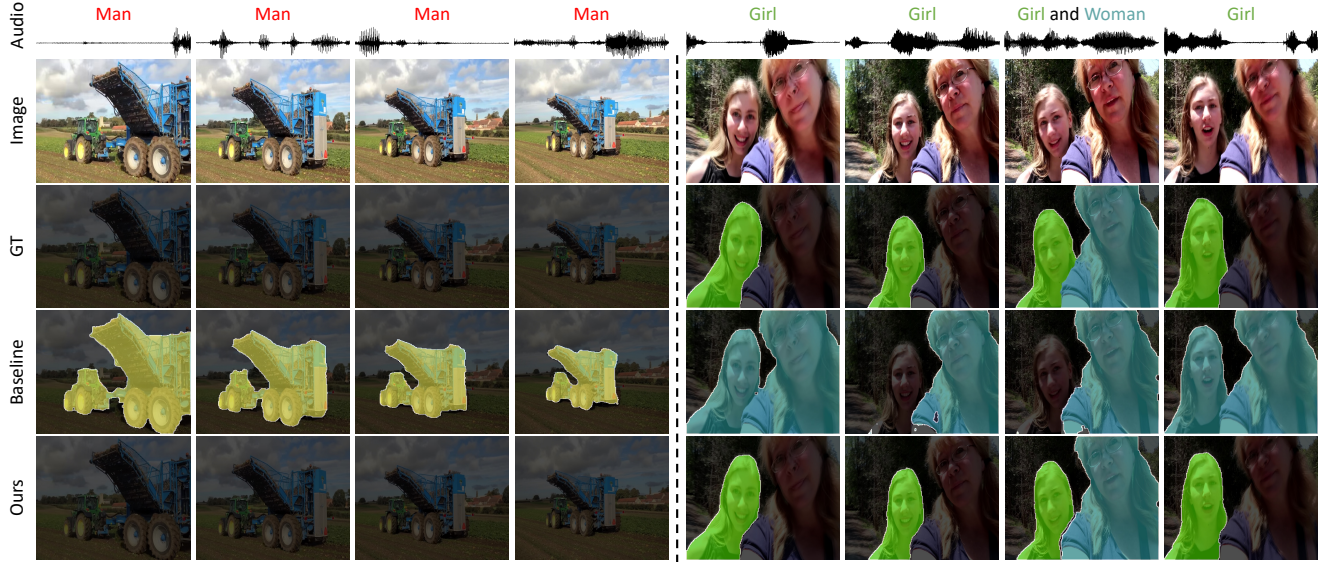
Figure 4. Qualitative comparison between our full model and the ACT baseline. Existing sounds correspond to masks of the same colors.
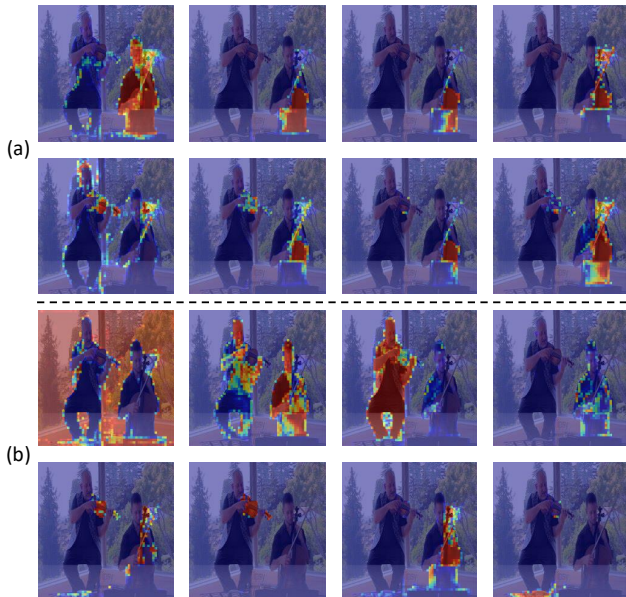


Figure 5. Visualization of logit maps from different types of queries. (a) Audio-derived queries. (b) Vision-derived queries.

emitting source in the video. This demonstrates that the vision-derived queries have a stronger ability to differentiate between different regions in the image and can fetch the corresponding information from the audio.

In addition, we also visualize the logit maps predicted by audio-derived and vision-derived queries within the Transformer decoder. For each type of 100 queries, we uniformly select one every five queries, remove those with blank logit maps, and then randomly retain 8 queries for display. As

shown in Figure 5(a), both instruments are producing sound in this example, but most of the audio-derived queries focus on the person on the right and the cello, indicating that these queries are biased towards certain sound-emitting objects due to the influence of entangled audio. In contrast, our vision-derived queries Figure 5(b) are able to focus on diverse visual regions in the image, demonstrating that vision-derived queries are more distinguishable, which forms the foundation for superior performance.

## 5. Conclusion

Given an audio clip, the goal of audio-visual segmentation (AVS) is to segment the sounding objects in the video. In this work, we proposed a new Vision-Centric Transformer (VCT) framework to overcome the limitations of audio-centric Transformers adopted by prevailing AVS methods. Concretely, our VCT framework utilizes vision-derived queries to iteratively extract both corresponding audio and fine-grained visual information, which aids in better distinguishing between different sound-emitting objects, eliminating the perception ambiguity caused by mixed audio and yielding accurate object segmentation. Additionally, we introduce a Prototype Prompted Query Generation (PPQG) module within the VCT framework, which generates vision-derived queries that contain both rich visual details and audio semantics through audio prototype prompting and pixel context grouping, facilitating later information aggregation. Extensive experiments on three subsets of the AVSBench dataset demonstrate that our method achieves new state-of-the-art performances.

## 6. Acknowledgments

## References

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016. 5

[2] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *ICCV*, 2017. 2

[3] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *ECCV*, 2018. 2

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 1, 3

[5] 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 *CVPR*, 2024. 6

[6] Yuanhong Chen, Chong Wang, Yuyuan Liu, Hu Wang, and Gustavo Carneiro. Cpm: Class-conditional prompting machine for audio-visual segmentation. In *ECCV*, 2024. 3, 6

[7] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022. 3, 5

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 5

[9] Shengyi Gao, Zhe Chen, Guo Chen, Wenhai Wang, and Tong Lu. Avsegformer: Audio-visual segmentation with transformer. In *AAAI*, 2024. 1, 3, 6

[10] Dawei Hao, Yuxin Mao, Bowen He, Xiaodong Han, Yuchao Dai, and Yiran Zhong. Improving audio-visual segmentation with bidirectional generation. In *AAAI*, 2024. 3, 6

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5

[12] 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. 4, 5

[13] Di Hu, Rui Qian, Minyue Jiang, Xiao Tan, Shilei Wen, Errui Ding, Weiyao Lin, and Dejing Dou. Discriminative sounding objects localization via self-supervised audiovisual matching. *NeurIPS*, 2020. 1, 2

[14] Xixi Hu, Ziyang Chen, and Andrew Owens. Mix and localize: Localizing sound sources in mixtures. In *CVPR*, 2022. 1, 2

[15] Shaofei Huang, Han Li, Yuqing Wang, Hongji Zhu, Jiao Dai, Jizhong Han, Wenge Rong, and Si Liu. Discovering sounding objects by audio queries for audio visual segmentation. In *IJCAI*, 2023. 3, 6

[16] Shaofei Huang, Rui Ling, Hongyu Li, Tianrui Hui, Zongheng Tang, Xiaoming Wei, Jizhong Han, and Si Liu. Unleashing the temporal-spatial reasoning capacity of gpt for training-free audio and language referenced video object segmentation. In *AAAI*, 2025. 3

[17] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016. 4

[18] Dongjin Kim, Sung Jin Um, Sangmin Lee, and Jung Uk Kim. Learning to visually localize sound sources from mixtures without prior source knowledge. In *CVPR*, 2024. 2

[19] Feng Li, Hao Zhang, Peize Sun, Xueyan Zou, Shilong Liu, Chunyuan Li, Jianwei Yang, Lei Zhang, and Jianfeng Gao. Segment and recognize anything at any granularity. In *ECCV*, 2024. 5

[20] Jiaxu Li, Songsong Yu, Yifan Wang, Lijun Wang, and Huchuan Lu. Selm: Selective mechanism based audio-visual segmentation. In *ACM MM*, 2024. 3, 6

[21] Kexin Li, Zongxin Yang, Lei Chen, Yi Yang, and Jun Xiao. Catr: Combinatorial-dependence audio-queried transformer for audio-visual video segmentation. In *ACM MM*, 2023. 1, 6

[22] Xiang Li, Jinglu Wang, Xiaohao Xu, Xiulian Peng, Rita Singh, Yan Lu, and Bhiksha Raj. Qdformer: Towards robust audiovisual segmentation in complex environments with quantization-based semantic decomposition. In *CVPR*, 2024.

[23] Chen Liu, Peike Patrick Li, Xingqun Qi, Hu Zhang, Lincheng Li, Dadong Wang, and Xin Yu. Audio-visual segmentation by exploring cross-modal mutual semantics. In *ACM MM*, 2023. 6

[24] Chen Liu, Peike Li, Hu Zhang, Lincheng Li, Zi Huang, Dadong Wang, and Xin Yu. Bavs: bootstrapping audio-visual segmentation by integrating foundation knowledge. *TMM*, 2024. 3, 6

[25] Jinxiang Liu, Yikun Liu, Fei Zhang, Chen Ju, Ya Zhang, and Yanfeng Wang. Audio-visual segmentation via unlabeled frame exploitation. In *CVPR*, 2024. 3, 6

[26] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 3, 5

[27] Juncheng Ma, Peiwen Sun, Yaoting Wang, and Di Hu. Stepping stones: A progressive training strategy for audio-visual semantic segmentation. In *ECCV*, 2024. 5

[28] Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016. 4

[29] Yuxin Mao, Jing Zhang, Mochu Xiang, Yiran Zhong, and Yuchao Dai. Multimodal variational auto-encoder based audio-visual segmentation. In *ICCV*, 2023. 3, 6

[30] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *3DV*, 2016. 5

[31] Shentong Mo and Pedro Morgado. Localizing visual sounds the easy way. In *ECCV*, 2022. 2

[32] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021. 5

[33] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *CVPR*, 2018. 1, 2

[34] Peiwen Sun, Honggang Zhang, and Di Hu. Unveiling and mitigating bias in audio visual segmentation. In *ACM MM*, 2024. 3, 6

[35] A Vaswani. Attention is all you need. *NeurIPS*, 2017. 1, 3

[36] 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. *CVM*, 2022. 5

[37] Yaoting Wang, Weisong Liu, Guangyao Li, Jian Ding, Di Hu, and Xi Li. Prompting segmentation with sound is generalizable audio-visual source localizer. In *AAAI*, 2024. 1, 3, 6

[38] Yaoting Wang, Peiwen Sun, Yuanchao Li, Honggang Zhang, and Di Hu. Can textual semantics mitigate sounding object segmentation preference? In *ECCV*, 2024. 3, 5, 6

[39] Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. In *CVPR*, 2022. 3

[40] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *CVPR*, 2022. 4

[41] Qi Yang, Xing Nie, Tong Li, Pengfei Gao, Ying Guo, Cheng Zhen, Pengfei Yan, and Shiming Xiang. Cooperation does matter: Exploring multi-order bilateral relations for audio-visual segmentation. In *CVPR*, 2024. 1, 3, 5, 6

[42] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022. 3

[43] 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 *ECCV*, 2022. 1, 2, 3, 4, 5, 6

[44] 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. *IJCV*, 2024. 1, 2, 3, 5, 6