CPM: Class-conditional Prompting Machine for Audio-visual Segmentation

Yuanhong Chen[1], Chong Wang[1], Yuyuan Liu[1], Hu Wang[2], and Gustavo Carneiro[3]

[1] Australian Institute for Machine Learning, University of Adelaide, Australia
[2] Mohamed bin Zayed University of Artificial Intelligence, United Arab Emirates
[3] Centre for Vision, Speech and Signal Processing, University of Surrey
yuanhong.chen@adelaide.edu.au

Abstract. Audio-visual segmentation (AVS) is an emerging task that aims to accurately segment sounding objects based on audio-visual cues. The success of AVS learning systems depends on the effectiveness of cross-modal interaction. Such a requirement can be naturally fulfilled by leveraging transformer-based segmentation architecture due to its inherent ability to capture long-range dependencies and flexibility in handling different modalities. However, the inherent training issues of transformer-based methods, such as the low efficacy of cross-attention and unstable bipartite matching, can be amplified in AVS, particularly when the learned audio query does not provide a clear semantic clue. In this paper, we address these two issues with the new Class-conditional Prompting Machine (CPM). CPM improves the bipartite matching with a learning strategy combining class-agnostic queries with class-conditional queries. The efficacy of cross-modal attention is upgraded with new learning objectives for the audio, visual and joint modalities. We conduct experiments on AVS benchmarks, demonstrating that our method achieves state-of-the-art (SOTA) segmentation accuracy[4].

Keywords: Audio-visual Learning · Segmentation · Multi-modal Learning

1 Introduction

The recognition and integration of auditory and visual data are fundamental to human cognitive processes, playing a critical role in facilitating meaningful communication [44]. Audio-visual segmentation (AVS) is an emerging cross-modal reasoning task that mimics such cognitive processes, aiming to localize visual objects based on audio-visual cues. AVS has many applications, such as the automatic localisation and identification of sounding objects to improve the accessibility of videos for the blind and visually impaired [34]. A major challenge in AVS is achieving effective cross-modal interaction between sound and visual

[4] This project is supported by the Australian Research Council (ARC) through grant FT190100525.

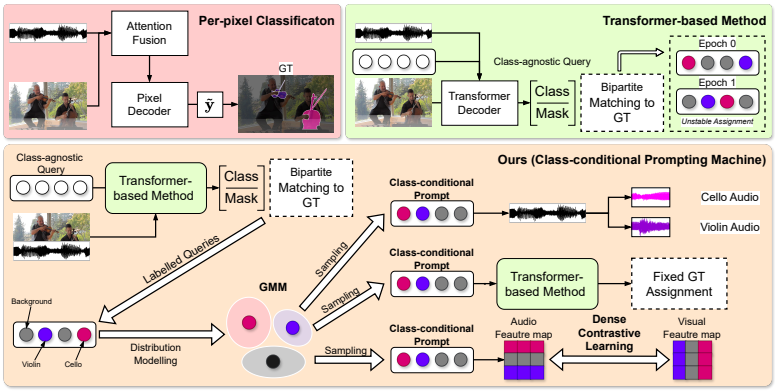


Fig. 1: Comparing conventional AVS methods [7, 14] with our CPM approach, CPM inherits the class-agnostic query from transformer-based methods and integrates class-conditional prompts sampled from the learned joint-modal data distribution to achieve three objectives: 1) learn disentangled audio partitioning, 2) facilitate semantic-guided object identification, and 3) promote more explicit audio-visual contrastive learning.

objects [46]. Many AVS methods [7, 37, 57] generally adopt a traditional per-pixel classification framework [8, 52], utilising early fusion strategies (e.g., cross-attention fusion [49]) together with an FCN decoder [4, 35] to make predictions. Such per-pixel design has achieved good performance, but it tends to under-utilise the audio data due to its lower informativeness compared with the visual data [7, 46]. Contrastive learning can mitigate this issue by matching informative audio-visual pairs [7]. However, another critical limitation of the per-pixel design is its failure to capture instance-level visual information, resulting in inconsistent segmentation predictions within or between frames in a video sequence [52].

These two issues have been addressed by transformer-based AVS methods designed to capture instance-level information and to rely on more effective contrastive learning [14, 24, 28, 30, 31, 52, 54]. Nevertheless, these AVS approaches still show slow convergence and relatively poor accuracy [23, 33, 55], primarily attributed to the *low efficacy of the cross-attention* [48] and the *unstable bipartite matching* [23]. Solutions for these problems are based on integrated masked-attention [8], which focuses on features around predicted segments, and anchor de-noising [23] to reconstruct the noisy ground-truth bounding box detection. However, these solutions may not work well in AVS because of the weak constraint provided by global audio features that contain a mixture of sound sources [14, 26, 30] resulting in more instability during training. It is worth noting that all methods above have the common issue of relying on class-agnostic prompts that provide little guidance to the bipartite matching process, thereby reducing training efficacy.

In this paper, we introduce the Class-conditional Prompting Machine (CPM), an audio-visual segmentation training approach that leverages class-conditional prompts to enhance bipartite matching stability and improve cross-modal attention efficacy. To enhance bipartite matching, we introduce a novel learning method, combining class-agnostic queries [8, 14, 26] with class-conditional queries,

sampled from our iteratively updated generative model of class-specific embeddings, with the former queries matched to ground-truth labels using the Hungarian Algorithm [1], followed by individual processing of class-conditional queries in two modalities. To improve cross-modal attention efficacy, new learning objectives are proposed for both audio and visual modalities. In the audio modality (audio conditional prompting, denoted as ACP), the original spectrogram is corrupted with auditory off-the-screen noise and class-conditional queries are employed to reconstruct the original spectrogram, while in the visual modality (visual conditional prompting, denoted as VCP), noisy class-conditional queries sampled from our generative model are used to probe semantically similar content in the image space. To further upgrade the performance of cross-modal attention, we introduce a new prompting-based audio-visual contrastive learning (PCL) task, guided by class-specific queries to densely constrain the cross-modal representations. To summarise, our main contributions are:

- A new AVS training approach to enhance bipartite matching stability and enhance the efficacy of cross-modal attention. Our core innovation lies in the development of a **C**lass-conditional **P**rompting **M**achine (CPM).
- To improve the bipartite matching, we propose a new AVS learning strategy that combines class-agnostic queries with class-conditional queries, sampled from our iteratively updated generative model of class-specific embeddings.
- The efficacy of cross-modal attention is upgraded with new learning objectives for the audio, visual and joint modalities. For **audio**, we present ACP that perturbs the original spectrogram and uses class-conditional queries to reconstruct the spectrogram; for **visual**, we introduce VCP that explores noisy class-conditional queries sampled from our generative model to probe the corresponding semantic in visual feature map. To enhance the cross-modal attention efficacy further, we propose PCL, consisting of a new prompting-based **joint** (audio-visual) contrastive learning task.

We firstly show the effectiveness of our CPM model through rigorous evaluation on established benchmarks such as AVSBench-Objects [57] and AVSBench-Semantics [56]. Furthermore, we extend our evaluation by including VPO synthetic benchmarks [7], aiming to enhance our comprehension of AVS methods' capacity to capture audio-visual correlations. Our findings across these benchmarks consistently demonstrate that our approach yields better classification accuracy compared to existing methods.

2 Related Works

2.1 Transformer-based architecture

Transformer-based methods have shown promising performance in detection [1] and segmentation [8,9] benchmarks. The fundamental concept is to leverage the object query to probe image features from the output of transformer encoders and bipartite graph matching to perform set-based box/mask prediction. It is

also evidenced that such a framework can benefit multi-modal learning due to its attention mechanism and flexibility in data modelling [53]. Despite its successful application in various domains [1, 8], such frameworks have also been reported to exhibit a poorer convergence rate compared to traditional CNN-based methods [8, 23, 33]. In exploring the reasons behind the poor convergence, previous studies have emphasized enhancing the interpretability of the learnable query [33, 39] (often referring to them as anchor points/boxes). Alternatively, the utilisation of denoising methods [23, 55] has also been adopted to facilitate the learning of bounding box offset [23] and improve the utilisation of the transformer decoder layers [55]. Furthermore, masked attention has proven effective in enhancing both convergence rate and model performance [8]. The successful implementation of denoising methods and masked attention mechanisms has inspired us to devise a mitigation strategy for bipartite matching that promotes improved cross-modal understanding within the cross-attention process.

2.2 Audio-visual Segmentation (AVS)

Audio-visual Segmentation is a dense classification task or detection of sounding visual objects in videos, using image sequences and audio cues. Zhou et al. [57] introduced the AVSBench-Object and AVSBench-Semantics benchmarks [56], which enable the evaluations of single-source and multi-source AVS tasks for salient and multi-class object segmentation. To address concerns related to the high annotation cost and dataset diversity, Chen et al. [7] introduced a cost-effective strategy to build a relatively unbiased AVS dataset, named Visual Post-production (VPO). Mainstream AVS methods use audio as the reference query [7, 14, 19, 24, 29–32] or prompt [43, 51]. For example, some methods adopt MaskFormer [9] or segment anything model (SAM) [22] to perform image segmentation using audio queries or encoded audio prompts and cross-attention layers. These methods benefit from the attention mechanism's ability and mask-classification's features to capture long-range dependencies and enhance image segmentation ability [9], spatial-temporal reasoning [24] and task-related features [14, 24, 29, 30, 37, 57]. The adaptation of the Maskformer-based framework for AVS relies on methods to encourage the audio-visual semantic alignment [24, 30, 31, 54]. Such strategy mitigates the poor audio semantic information caused by modality imbalance [5] and learns disentangled multi-modal representations [26, 54] via vector quantization [15] and bidirectional attention mechanism. Even though the query-based transformer architecture [1] has shown great success in semantic segmentation tasks, certain weaknesses such as low efficacy of cross-attention and unstable bipartite matching process [23] are still only partially addressed. These two issues may be exacerbated in AVS due to the poor audio semantic information.

2.3 Audio-visual Contrastive Learning

Contrastive Learning has shown promising results in audio-visual learning (AVL) methods [2, 7, 18, 40, 41]. These methods bring together augmented representations from the same instance as positives while separating representation pairs

from different instances as negatives within a batch. A reported issue with current AVL contrastive learning is its reliance on self-supervision [6] to connect audio and visual representations of the same class. To overcome this issue, CAVP [7] propose a supervised contrastive learning [21, 25, 50] method that mines informative contrastive pairs from arbitrary audio-visual pairings to constrain the learning of audio-visual embeddings. Nevertheless, these previous approaches predominantly leverage global audio representations, thereby limiting the model's capacity to discern individual audio sources. Consequently, this constraint reduces the model's effectiveness in scenarios involving multiple sound sources. In our work, we employ class-specific queries to retrieve the corresponding audio representations, which facilitates a more explicit form of contrastive learning between audio and visual modalities.

3 Method

We denote a multi-class audio-visual dataset as $\mathcal{D} = \{(\mathbf{a}_i, \mathbf{x}_i, \mathbf{y}_i, \mathbf{t}_i))\}_{i=1}^{|\mathcal{D}|}$, where $\mathbf{x}_i \in \mathcal{X} \subset \mathbb{R}^{H\times W\times 3}$ is an RGB image with resolution $H \times W$, $\mathbf{a} \in \mathcal{A} \subset \mathbb{R}^{T\times F}$ denotes the magnitude spectrogram of the audio data with T time and F frequency bins, $\mathbf{y}_i \in \mathcal{Y} \subset \{0,1\}^{H\times W\times C}$ denotes the pixel-level ground truth for the C classes (the background class is included in these C classes), and $\mathbf{t}_i \in \mathcal{Y} \subset \{0,1\}^{C}$ is a multi-label ground truth audio annotation. We use Mask2former [8] as the segmentation framework.

3.1 Preliminaries using Class-agnostic Queries

Like other Maskformer-based methods [14,30,31,52], we aim to learn the parameters $\theta \in \Theta$ for the model $f_\theta : \mathcal{X} \times \mathcal{A} \rightarrow [0,1]^{H\times W\times C}$, which comprises the image and audio backbones that extract features with $\mathbf{u}_a = f_\gamma(\mathbf{a})$ and $\mathbf{u}_v = f_\phi(\mathbf{x})$, respectively, where $\gamma, \phi \in \theta$, and $\mathbf{u}_a, \mathbf{u}_v \in \mathcal{U}$, with $\mathcal{U}$ denoting a unified feature space. A set of learnable query features (comprising the object query feature and positional query embeddings) are defined as the joint-modal output embeddings similar to the Perceiver model [20]. We define the class-agnostic query feature as $\mathbf{q} \in \mathcal{Q} \subset \mathbb{R}^{N\times D_q}$, where N denotes the number of class-agnostic queries, and D_q represents the dimensionality of the feature space. As depicted in Fig. 2, given $\mathbf{q}$, we aim to group pixels with matched semantic information from $\mathbf{u}_a$ and $\mathbf{u}_v$ through consecutive transformer decoder layers [8, 49] and generate N mask embeddings $\tilde{\mathbf{q}} \in \mathcal{Q}$ and pixel embeddings $\tilde{\mathbf{u}}_v \in \mathbb{R}^{H\times W\times D_q}$. Then, the model independently predicts the embeddings into N set of class predictions (via $\mathsf{Softmax}(\cdot)$) and mask predictions (via $\mathsf{Sigmoid}(\cdot)$) denoted as $\mathcal{E}_{\mathsf{pred}} = \{(\mathbf{m}_i, \mathbf{p}_i)\}_{i=1}^{N}$, where $\mathbf{m} \in \mathcal{M} \subset \{0,1\}^{H\times W}$ and $\mathbf{p} \in \mathcal{P} \subset \{0,1\}^{C}$. We denote the ground-truth set derived from the training set $\mathcal{D}$ with $\mathcal{E}_{\mathsf{gt}} = \{(\mathbf{m}_i^{\mathsf{gt}}, \mathbf{p}_i^{\mathsf{gt}})\}_{i=1}^{N^{\mathsf{gt}}}$. During training, we use the Hungarian algorithm to perform optimal matching between $\mathcal{E}_{\mathsf{pred}}$ and $\mathcal{E}_{\mathsf{gt}}$ [1,8] to facilitate the label assignment. Since $|\mathcal{E}_{\mathsf{pred}}| > |\mathcal{E}_{\mathsf{gt}}|$, we pad $\mathcal{E}_{\mathsf{gt}}$ with no-object class $\varnothing$ to achieve one-to-one matching [8].

The loss ℓ to train the model $f_\theta(.)$ includes a cross-entropy query classification loss ℓ_{ce}, and a binary mask loss $\ell_{mask} = \ell_{focal} + \ell_{dice}$, which combines a focal loss

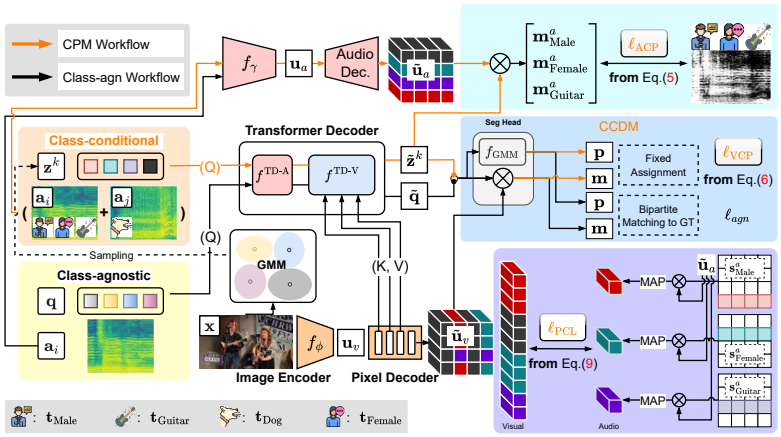


Fig. 2: Illustration of our CPM method. Starting with a scene with a mixture of sound sources including Male, Female and Guitar, the training alternates between the use of learnable class-agnostic queries, and queries sampled from a list of class-specific query features from the GMM, denoted as $\mathbf{z}^{\text{male}}$, $\mathbf{z}^{\text{female}}$ and $\mathbf{z}^{\text{guitar}}$. The overall training objective is composed of three learning tasks: 1) in audio conditional prompting (ACP), we aim to use $\mathbf{z}^{\text{male}}$, $\mathbf{z}^{\text{female}}$ and $\mathbf{z}^{\text{guitar}}$ to recover the original magnitude spectrogram $\mathbf{a}_i$ from the noise spectrogram (i.e., $\mathbf{a}_i + \mathbf{a}_j$) that is corrupted by another Dog audio signal $\mathbf{a}_j$; 2) a visual conditional prompting (VCP) that aim to probe the corresponding pixels w.r.t to the class-specific query features; and 3) a contrastive learning task that target to densely constrain the audio and visual representations. For training, both the CPM Workflow, indicated by the **orange arrow**, and the Class-agnostic Workflow, marked by the **black arrow**, are utilized. However, only the Class-agnostic Workflow is used for inference.

ℓ_{focal} and a dice loss ℓ_{dice} [8,14,30]. The overall training loss for the class-agnostic query feature is defined as $\ell_{agn} = \ell_{ce} + \ell_{mask}$. During testing, $\mathbf{p}$ and $\mathbf{m}$ are merged via the multiplication $\arg\max_{i:c_i \neq \varnothing} \mathbf{p}(c_i) \cdot \mathbf{m}_i$, where c_i is the class label with maximum likelihood $c_i = \arg\max_{c \in \{1,\dots,C,\varnothing\}} \mathbf{p}_i(c)$ for each probability-mask pair indexed by i.

3.2 Class-conditional Prompting Machine (CPM)

The proposed CPM training builds upon the class-agnostic training method from Sec. 3.1 by introducing a new training stage based on class-conditional prompting. This allows the probing of both the magnitude spectrogram and the image feature map, aiming to mitigate unstable training issues and improve cross-attention efficacy. In some cases, the prompts can be manually crafted (i.e., text label, bounding box) based on domain knowledge or specific task requirements [22]. Prompts can also be automatically learned to form a fixed set of prototypical embeddings for each class [42]. However, it is impossible to manually define class conditional prompts in high-dimensional latent spaces, and using a limited-size set of learned prompts may fail to capture the comprehensive distribution of class-specific prompts. Hence, it is desirable that these

class-specific prompts can be sampled from a generative model [3,27]. that comprehensively represents the respective class. We adopt the Gaussian Mixture Models (GMMs) as such generative model [45], which improves the intra-class variability and increases robustness to class imbalances when compared to the alternative approaches mentioned before. Before delving into the methodology of the CPM, we first introduce the generation process of the class-conditional query features.

Class Conditional Distribution Modelling (CCDM) Instead of constructing a posterior $p(c|\tilde{\mathbf{q}})$ with $\mathsf{Softmax}$ classifier for the mask embeddings $\tilde{\mathbf{q}}$, the generative classifier employs the Bayes rule for label prediction, estimating the class-conditional distribution $p(\tilde{\mathbf{q}}|c)$ alongside the class prior $p(c)$ [27], with:

$$p(c|\tilde{\mathbf{q}}) = \frac{p(\tilde{\mathbf{q}}|c)p(c)}{\sum_{c'} p(\tilde{\mathbf{q}}|c')p(c')}, \tag{1}$$

where the class probability prior $p(c)$ is uniform. To enable GMM modelling in Maskformer architecture [8, 9], we replace the $\mathbf{p} = \mathsf{Softmax}(\tilde{\mathbf{q}})$ activation function with $f_{\mathsf{GMM}}(\tilde{\mathbf{q}}) = p(c|\tilde{\mathbf{q}})$ defined in (1) which maps the input mask embeddings $\tilde{\mathbf{q}}$ to probability data density over C classes. Assuming that after the Hungarian matching between the predictions set $\{(\mathbf{m}_i, \mathbf{p}_i)\}_{i=1}^{N}$ and $\mathbf{y}$, we get a new one-to-one matched ground-truth set $\tilde{\mathbf{y}}$ w.r.t each prediction by padding the $\mathbf{y}$ with no-object class $\varnothing$. Subsequently, we can form the GMM training dataset $\mathcal{F} = \{(\tilde{\mathbf{q}}_n, \tilde{\mathbf{y}}_n)\}_{n=1}^{F}$ by pairing the mask embeddings with the assigned label. In our method, the goal of GMM is to model the data distribution of the joint-modal mask embedding $\tilde{\mathbf{q}}$ for each class C in the D_q-dimensional space by employing a weighted mixture of M multivariate Gaussians, defined as follows:

$$p(\tilde{\mathbf{q}}|c) = \sum_{m=1}^{M} p(m|c;\boldsymbol{\pi}_c)p(\tilde{\mathbf{q}}|c,m;\boldsymbol{\mu}_c,\boldsymbol{\Sigma}_c) = \sum_{m=1}^{M} \boldsymbol{\pi}_{cm}\mathcal{N}(\tilde{\mathbf{q}};\boldsymbol{\mu}_{cm},\boldsymbol{\Sigma}_{cm}), \tag{2}$$

where $\sum_m \pi_{cm} = 1$ represents the mixing coefficients, $\boldsymbol{\mu}_{cm} \in \mathbb{R}^{D_q}$ denotes the mean vector, and $\boldsymbol{\Sigma}_{cm} \in \mathbb{R}^{D_q \times D_q}$ is the covariance matrix. The optimisation of the GMM parameters is performed by the Expectation Maximisation (EM) algorithm [10]. In the **E-step**, we iterate through $\mathcal{F}$, computing the responsibilities using the current parameters of the GMM for each class c:

$$\gamma_{c,n}^{(t)}(m) = \frac{\boldsymbol{\pi}_{cm}^{(t-1)}\mathcal{N}(\tilde{\mathbf{q}}_n;\boldsymbol{\mu}_{cm}^{(t-1)},\boldsymbol{\Sigma}_{cm}^{(t-1)})}{\sum_{m'=1}^{M}\boldsymbol{\pi}_{cm'}^{(t-1)}\mathcal{N}(\tilde{\mathbf{q}}_n;\boldsymbol{\mu}_{cm'}^{(t-1)},\boldsymbol{\Sigma}_{cm'}^{(t-1)})}. \tag{3}$$

We re-estimate the parameters using the calculated responsibility in the **M-step**.

$$\begin{aligned}
\boldsymbol{\mu}_{cm}^{(t)} &= \frac{1}{N_{cm}^{(t)}} \sum_{(\tilde{\mathbf{q}}_n,\tilde{\mathbf{y}}_n)\in\mathcal{F}_c} \gamma_{c,n}^{(t)}(m)\ \tilde{\mathbf{q}}, \quad \boldsymbol{\pi}_{cm}^{(t)} = \frac{N_{cm}^{(t)}}{|\mathcal{F}|}, \\
\boldsymbol{\Sigma}_{cm}^{(t)} &= \frac{1}{N_{cm}^{(t)}} \sum_{(\tilde{\mathbf{q}}_n,\tilde{\mathbf{y}}_n)\in\mathcal{F}_c} \gamma_{c,n}^{(t)}(m)\ (\tilde{\mathbf{q}}_n - \boldsymbol{\mu}_{cm}^{(t)})(\tilde{\mathbf{q}}_n - \boldsymbol{\mu}_{cm}^{(t)})^{\mathsf{T}},
\end{aligned} \tag{4}$$

where $N_{cm}^{(t)} = \sum_{(\tilde{\mathbf{q}}_n, \tilde{\mathbf{y}}_n) \in \mathcal{F}_c} \gamma_{c,n}^{(t)}(m)$. We maintain an external memory bank containing GMM training samples, defined above with the set $\mathcal{F}$, and the GMM parameters are robustly estimated with momentum [27].

Audio Conditional Prompting (ACP) Motivated by the concept of "mix-and-separate" [13, 18], we design the ACP process to improve cross-attention interaction for the dense audio feature representations. We consider an off-the-screen noise dataset $\mathcal{D}_r = \{(\mathbf{a}_j)\}_{j=1}^{|D_r|}$. For each training iteration, we sample a clean audio sample $\mathbf{a}_i$ from the $\mathcal{D}$ and an off-the-screen noisy audio sample $\mathbf{a}_j$ from $\mathcal{D}_r$ to build the mixture of magnitude spectrogram $\mathbf{a}_p = \mathbf{a}_i + \mathbf{a}_j$. The whole ACP process is illustrated in the first section of Fig. 2. Before delving into the audio recovery process, we first sample a set of class-conditional prompts $\mathbf{z}^k \sim \{\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}_{k \in \mathcal{K}}$ (where $\mathcal{K} = \{k | \mathbf{t}_i(k) = 1\}$ represent the indices for ground truth labels) via the GMM model from the last iteration according to the target semantic classes $\mathbf{t}_i$ that we want to recover. The major process of the audio recovery process shares a similar concept with the visual part described in Sec. 3.1; the difference is that we use the joint-modal mask embeddings $\tilde{\mathbf{z}}^k$ to search the semantically similar sound source on the decoded dense audio feature map $\tilde{\mathbf{u}}_a$ and generate their corresponding mask predictions $\{\mathbf{m}_k^a | k \in \mathcal{K}\}$.

The learning objective of the ACP process is to encourage the consistency between the predicted masks $\mathbf{m}_k^a$ with the derived ground-truth spectrogram ratio mask $\frac{\mathbf{a}_i}{\mathbf{a}_p}$. The AVS datasets generally do not provide the separated sound source data for each semantic class (i.e., in Fig. 2, we do not have the ground-truth audio data of Male, Female and Guitar), hence we adopted a summation operation over the predicted masks to integrate all mixed sound sources. Finally, a Sigmoid activation function $\sigma(\cdot)$ is applied to constrain the prediction into a binary mask. We denoted the mean squared error (MSE) loss of the ACP process as follows:

$$\ell_{\text{ACP}} = \left\| \sigma \left(\sum_{k \in \mathcal{K}} \mathbf{m}_k^a \right) - \frac{\mathbf{a}_i}{\mathbf{a}_p} \right\|_2 . \tag{5}$$

Visual Conditional Prompting (VCP) The VCP module is processed simultaneously with ACP. The design of VCP aims to bypass bipartite matching via the generated class-conditional prompts $\mathbf{z}^k$ to ease the model training as it can provide a more stable learning target for each query feature to mitigate the instability brought by the bipartite matching with the class-agnostic queries. Given a set of class-conditional prompts derived from the ground-truth image labels, our training objective of VCP is to correctly segment the corresponding image regions as well as successfully classify these prompts after the consecutive transformer decoder layers. The loss function is similar to ℓ_{agn} in Sec. 3.1, where we replace $\mathbf{q}$ with $\mathbf{z}^k$, denoted as $\ell_{\text{VCP}} = \ell_{ce} + \ell_{mask}$ (6).

Prompting-based Contrastive Learning (PCL) The ability to learn discriminative feature representation is critical for the audio-visual system. One limitation of the previous audio-visual contrastive learning methods [7,36] is that

they can only leverage the global audio representation due to the lack of class conditional prompts for class-specific feature disentanglement. By taking advantage of class-conditional distribution modelling, we can overcome this limitation by utilising the predicted spectrogram saliency mask $\mathbf{s}_k^a = \sigma(\mathbf{m}_k^a)$ and its associated class label k of each sound source, denoted as $\{\mathbf{s}_k^a | \mathbf{s}_k^a \in \mathcal{M} \subseteq [0,1]^{T\times F}, k \in \mathcal{K}\}$. To disentangle the class-specific representations for the audio feature map, we iteratively apply the masked average pooling (MAP) [47,58] to $\mathbf{u}_a$ based on $|\mathcal{K}|$ saliency masks $\mathbf{s}_k^a$ for the feature extraction process. We first threshold $\mathbf{s}_k^a$ and convert it to a binary mask via $\tilde{\mathbf{s}}_k^a = \mathbb{1}(\mathbf{s}_k^a > \bar{\mathbf{s}}_k^a)$, where $\bar{\mathbf{s}}_k^a$ represent the mean value of the saliency mask. Then, for the k-th category, we extract its region-level mean feature $\mathbf{f}(k) \in \mathbb{R}^{D_q}$ by MAP defined as follows:

$$\mathbf{f}(k) = \frac{\sum_{\psi\in\Psi} \tilde{\mathbf{s}}_k^a \mathbf{u}_a(\psi)}{\sum_{\psi} \tilde{\mathbf{s}}_k^a}, \tag{7}$$

where Ψ is the lattice of size $T \times F$. We combine the $\mathbf{f}(k)$ with the ground-truth class label k to form the anchor set $\mathcal{E}_{\text{anch}} = \{(\mathbf{f}(k), k) | \mathbf{f}(k) \in \mathcal{U}, k \in \mathcal{K}\}$, with $\mathcal{U}$ denoting the audio and visual feature spaces defined in Sec. 3.1. Since the pixel-level label $\mathbf{y}$ is available, we can directly form the visual contrastive set $\mathcal{E}_{\text{cont}} = \{(\mathbf{u}_v(\omega), \mathbf{y}_v(\omega)) | \mathbf{u}_v(\omega) \in \mathcal{U}, \mathbf{y}(\omega) \in \mathcal{Y}\}$. Hence, we can derive the positive and negative sets as:

$$\begin{aligned} \mathcal{P}(\mathbf{u}_v(\omega)) &= \{\mathbf{u}_v(\omega) | \mathbf{u}_v(\omega) \in \mathcal{U}, \mathbf{y}(\omega, k) = 1\}, \\ \mathcal{N}(\mathbf{u}_v(\omega)) &= \{\mathbf{u}_v(\omega) | \mathbf{u}_v(\omega) \in \mathcal{U}, \mathbf{y}(\omega, k) = 0\}. \end{aligned} \tag{8}$$

Adopting the supervised InfoNCE [21] as the objective function to pull the anchor $\mathbf{f} \in \mathcal{E}_{\text{anch}}$ and respective positive visual features closer while repelling anchors and their negative visual features, we define the following loss:

$$\ell_{\text{PCL}}(\mathbf{f}) = \frac{1}{|\mathcal{P}(\mathbf{u}_v(\omega))|} \sum_{\mathbf{f}_p \in \mathcal{P}(\mathbf{u}_v(\omega))} -\log \frac{\exp(\mathbf{f}\cdot\mathbf{f}_p/\tau)}{\exp(\mathbf{f}\cdot\mathbf{f}_p/\tau) + \sum_{\mathbf{f}_n \in \mathcal{N}(\mathbf{u}_v(\omega))} \exp(\mathbf{f}\cdot\mathbf{f}_n/\tau)}, \tag{9}$$

where $\mathbf{f}$ is an anchor feature, and τ is the temperature hyperparameter The combination of all sub-objective ℓ_{ACP}, ℓ_{VCP} and ℓ_{PCL} form the CPM loss $\ell_{\text{CPM}} = \ell_{\text{ACP}} + \ell_{\text{VCP}} + \ell_{\text{PCL}}$.

Overall training The **overall training objective** is $\ell = \ell_{agn} + \lambda \ell_{\text{CPM}}$, where λ is the weight coefficient.

4 Experiments

4.1 Implementation Details

Evaluation Protocols We utilize standard evaluation protocols from AVS-Bench datasets [56,57] for single-source (SS) and multi-source (MS) scenarios with binary labels, as well as for AVSBench-Semantics with multi-class labels. Image sizes are standardized to 224 × 224 for fair comparison. Additionally, we

Table 1: Quantitative (mIoU, F_β) audio-visual segmentation results (%) for the AVS-Bench test sets [56,57] (resized to 224×224) with ResNet50 [17] backbone. Best results in **bold**, 2^{nd} best underlined. Improvements against the 2^{nd} best are in the last row.

Eval. Methods	Method	AVSBench-Obj.(SS)		AVSBench-Obj.(MS)		AVSBench-S	
		mIoU ↑	F_β ↑	mIoU ↑	F_β ↑	mIoU ↑	F_β ↑
Per-image [56,57]	TPAVI [57]	72.79	84.80	47.88	57.80	20.18	25.20
	AVSBG [16]	74.13	85.40	44.95	56.80	-	-
	ECMVAE [37]	76.33	86.50	48.69	60.70	-	-
	DiffusionAVS [36]	75.80	86.90	49.77	62.10	-	-
	CATR [24]	74.80	86.60	52.80	65.30	-	-
	AuTR [31]	75.00	85.20	49.40	61.20	-	-
	AQFormer [19]	77.00	86.40	55.70	66.90	-	-
	AVSegFormer [14]	76.54	84.80	49.53	62.80	24.93	29.30
	AVSC [29]	77.02	85.24	49.58	61.51	-	-
	BAVS [30]	77.96	85.29	50.23	62.37	24.68	29.63
	AVSAC [5]	76.90	86.95	53.95	65.81	25.43	29.71
	QSD [26]	77.60	86.00	59.60	63.50	<u>46.60</u>	-
	COMBO [54]	81.70	90.10	54.50	66.60	33.30	37.30
Per-dataset [12]	AVSegFormer* [14]	79.80	91.41	53.28	64.76	41.64	55.05
	CAVP [7]	<u>85.77</u>	<u>92.86</u>	<u>62.39</u>	<u>73.62</u>	44.70	<u>57.76</u>
	CPM	**87.64**	**93.53**	**65.22**	**75.22**	**48.39**	**64.01**

adopt the original image resolution for testing, following [7], to demonstrate optimal model performance. Evaluation is extended to the Visual Post-production (VPO) benchmark [7] for challenging cases. We employ mean Intersection over Union (mIoU) [11] and F_β score with $\beta^2 = 0.3$ [38, 57] to assess segmentation quality, precision, and recall performance at pixel level. Models are stratified into CNN-based pre-pixel classification and transformer-based mask classification to demonstrate architectural capabilities. Official training splits are used for AVSBench datasets [56, 57] and VPO [7], with results reported on the respective testing sets. Training is performed on the entire AVSBench-Semantics dataset, and testing is conducted on subsets as well as the entire testing set to demonstrate partitioned model performance. For further details on training and inference, please refer to the *Supplementary Material*.

Results We collected the experimental results from the existing benchmark [7] and updated it with recent works [5, 26, 54]. We modified the AVSegFormer [14] with Mask2former [8] and denoted it as *AVSegFormer** in the tables to encourage a fair comparison with our method. Please note that Tab. 1 includes two evaluation protocols. We employ standard semantic segmentation protocols to compute both mIoU and F_β same as PascalVOC, initially mentioned in [57].

4.2 Performance on Low-resolution AVSBench Videos

We adopt established methodologies [7, 30, 31, 37] for conducting performance evaluations on down-sampled image benchmarks, such as AVSBench-Objects [57], which includes single-source (SS) and multi-source (MS) splits with binary annotations, as well as the AVSBench-Semantics dataset [56], which contains multi-class annotations. We compare the performance of state-of-the-art (SOTA) methods with our CPM in Tab. 1 using mIoU and F_β. The results demonstrate that our model surpasses the second-best CAVP [7] in terms of mIoU by 1.87% on AVSBench-Object (SS) [57], 2.83% on AVSBench-Object (MS) [57] and 1.79% on AVSBench-Semantics [57] using the ResNet-50 [17] backbone.

Table 2: Quantitative (mIoU, F_β) audio-visual segmentation results (in %) for the AVSBench-Semantic (AVSS) test sets [56] (original resolution) with ResNet50 [17] backbone. Best results in **bold**, 2^{nd} best underlined. Improvements against the 2^{nd} best are in the last row.

D-ResNet50 [17]	Method	AVSS (SS)		AVSS (MS)		AVSS	
		mIoU ↑	F_β ↑	mIoU ↑	F_β ↑	mIoU ↑	F_β ↑
Per-pixel Classification	TPAVI [57]	42.10	61.46	26.33	40.99	43.39	59.24
	CAVP [7]	56.91	69.15	38.61	52.92	50.75	64.57
Transformer	AVSegFormer [14]	46.25	59.76	27.21	41.38	41.48	56.21
	AVSegFormer* [14]	50.52	63.75	31.40	42.81	45.80	59.16
	CPM	**61.71**	**72.94**	**43.11**	**56.28**	**57.25**	**70.54**
Improvement	**CPM**	**+4.80**	**+3.79**	**+4.50**	**+3.36**	**+6.50**	**+5.97**

Table 3: Quantitative (mIoU, F_β) audio-visual segmentation results (in %) for the VPO test sets (original resolution) with ResNet50 [17] backbone. Best results in **bold**, 2^{nd} best underlined. Improvements against the 2^{nd} best are in the last row.

D-ResNet50 [17]	Method	VPO (SS)		VPO (MS)		VPO (MSMI)	
		mIoU ↑	F_β ↑	mIoU ↑	F_β ↑	mIoU ↑	F_β ↑
Per-pixel Classification	TPAVI [57]	52.75	69.54	54.30	71.95	51.73	68.85
	CAVP	62.31	78.46	64.31	78.92	60.36	**75.60**
Transformer	AVSegFormer [14]	57.55	73.03	58.33	74.28	54.22	70.39
	AVSegFormer* [14]	60.51	74.81	62.91	77.33	56.24	72.67
	CPM	**67.09**	**79.88**	**65.91**	**79.90**	**60.55**	75.58
Improvements	**CPM**	**+4.78**	**+1.42**	**+1.60**	**+0.98**	**+0.19**	**−0.02**

4.3 Performance on Original AVSBench Videos

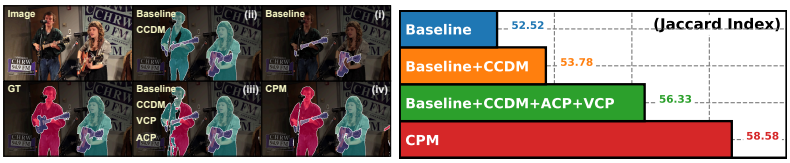
The benchmark mentioned above adopted low-resolution benchmarks using significantly resized input images (from 720p to 224 × 224). While this simplifies model training, disregarding the original image aspect ratio can lead to a degradation in model performance, which is not advisable for segmentation tasks. Therefore, we follow the evaluation method outlined in [7] for training and testing on the raw AVSBench videos. This includes using random resized crops for training and performing frame-by-frame evaluation during testing. The initial published version of AVSBench [57] only offers single-source and multi-source partitioning. However, this partitioning was not conducted in the later version [56], resulting in some cases being overlooked. We re-organised the previous SS and MS with the newly added video data, resulting in 1278 single-source videos (AVSS-SS) and 276 multi-source videos (AVSS-MS). We compare the results with mIoU and F_β to demonstrate the comprehensive model performance in Tab. 2. Our method shows a significant mIoU improvement of 4.80% on AVSS-SS, 4.50% on AVSS-MS and +6.50% on the entire AVSS dataset. To further demonstrate the effectiveness of our method, we show a visualisation of 6-second video clip in Fig. 5 that displays a qualitative comparison between TPAVI, AVSegFormer, CAVP and our CPM. Our method can successfully approximate the ground truth segmentation of the target sound source within a group of other semantic objects.

4.4 Performance on VPO Dataset

We also compare the model performance on the VPO benchmark [7], equipped with synthesized stereo audios. We adopted ResNet-50 [17] backbone for all

Table 4: Ablation study of the model components on AVSBench-Semantics [31].

Method					Metrics	
Baseline	CCDM	VCP	ACP	PCL	mIoU ↑	F_β ↑
✔					53.04	65.29
✔	✔				54.12	66.65
✔	✔	✔			55.79	68.17
✔	✔		✔		55.07	68.03
✔	✔	✔	✔		56.41	69.07
✔	✔	✔	✔	✔	**57.25**	**70.54**



(a) Visual comparison. **(b)** Coverage score comparison.

Fig. 3: Qualitative (3a) and quantitative (3b) comparisons between model components in a multi-source scenario (i.e., male singing, female singing and guitar)

three subsets, and the results are shown in Tab. 3. To facilitate stereo audio encoding, we adopt the approach outlined in [7] by adjusting the number of input channels in the first layer to 2. Our approach surpasses the SOTA method CAVP [7] by 4.78% and 1.60% in terms of mIoU on VPO (SS) and VPO (MS), respectively. However, we observe a marginal improvement of 0.19% on the VPO (MSMI) setup compared with the SS and MS subsets. The possible explanation for this phenomenon may stem from the overly simplistic integration of stereo audio within the transformer architecture, as the foundational AVS transformer architectures [14, 24, 26, 30] did not account for a distinct positional encoding scheme tailored for stereo data. We will explore the design of a dedicated stereo AVS transformer architecture as part of our future work.

4.5 Ablation Study

Ablation of Key Components We first perform the key components analysis of CPM on AVSBench-Semantics [56] in Tab. 4. The baseline is AVSegformer* [14] in the 1st row. We replace the Softmax classifier in baseline with CCDM defined in (1), and observe an mIoU improvement of +1.08%. By integrating the model with VCP (3rd row), using training loss in (6), and ACP (4th row), using loss in (5), separately, we achieve an improvement of +1.26% and +1.39%, respectively. Subsequently, when we apply both ACP and VCP methods (5th row) the mIoU performance improves +2.04% compared to the 2nd row. The final row displays the complete CPM method, incorporating the dense contrastive learning approach we introduced, which leads to an additional mIoU improvement of +1.09%.

Ablation of Bipartite Matching Stability To study the stability of the bipartite matching process, we design an entropy-based stability score STS to quantify such performance on AVSBench-Semantics [56]. Intuitively, the STS measures the average assignment consistency after bipartite matching across all the classes. During the training process, we collect the assigned label $\tilde{\mathbf{y}}_i \in$

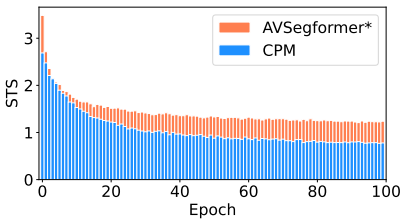


Fig. 4: Matching stability (STS ↓) comparison on AVSBench-Semantics [31]

Table 5: Ablation of GMM modelling on AVSBench-Semantics [31].

Method	mIoU ↑	F_β ↓
Point Rep.	55.96	68.58
Dist Rep.	57.25	70.54
M	mIoU	F_β
1	55.83	68.68
3	57.25	70.54
5	56.31	69.04
7	56.05	68.59

$\{0, ..., C\}^N$ to the class-agnostic query $\tilde{\mathbf{q}}_n$ produced by the Hungarian algorithm for each training image, resulting in $\mathbf{R} = [\tilde{\mathbf{y}}_0 \oplus \tilde{\mathbf{y}}_1 \oplus, ..., \oplus \tilde{\mathbf{y}}_D]$, where $\oplus$ represents the concatenation operator and $\mathbf{R}$ has size $D \times N$. We then use the indicator function to extract the assignment information over all semantic classes. We denote the resulting set as $\{\mathbf{s}_0, \mathbf{s}_1, ..., \mathbf{s}_c\}$, where $\mathbf{s}_c(n) = \sum_d^D \mathbb{1}(\mathbf{R}(d, n) = c)$ for $n \in \{1, ..., N\}$, which forms $\mathbf{s}_c \in [0, 1]^N$ after normalising it to be a probability distribution. Finally, we calculate the stability score $\mathsf{STS} = \frac{1}{C} \sum_{c=1}^{C} H(\mathbf{s}_c)$, where $H(\cdot)$ computes the entropy. We compare our CPM with AVSegformer* over 100 training epochs as illustrated in Fig. 4. Our findings reveal that our CPM consistently improves the stability score (showing lower average entropy) throughout the training process in comparison to AVSegformer*, thereby validating the efficacy of our approach in enhancing bipartite matching stability. Please refer to the *Supplementary Material* for the Pseudo-code of STS.

Ablation of Audio Contribution While experimenting with the baseline method mentioned above, we empirically observed that the classification after the audio transformer decoder (i.e., $f^{\text{TD-A}}$ in Fig. 2) has low accuracy, with noisy predictions. We also found that these noisy predictions are progressively refined by the vision transformer decoder layer based solely on visual clues. Such an observation illustrates that in some hard cases, we may fail to extract useful semantic information from audio, leading to the model being overly reliant on the visual content, which may cause erroneous testing predictions, as shown in Fig. 3a-(i), (ii), (iii). To monitor the amount of valid information retained by the query following interaction with the audio branch, we introduce a classification coverage score. This score is computed by comparing the predictions generated after passing through the audio transformer decoder with the ground-truth class label, using the Jaccard index. We show qualitative results in Fig. 3a for the designated scenario and quantitative results in Fig. 3b for the AVSBench-Semantics testing set. The comparison starts with AVSegFormer as the baseline method with progressive addition of our proposed modules, similar to Tab. 4. Notice that each CPM sub-module contributes to refining key semantic information from the audio modality, as evidenced by the segmentation mask predictions and the improvement in Jaccard indexes. This shows the effectiveness of CPM.

Fig. 5: Qualitative audio-visual segmentation results on AVSBench-Semantics [56] by TPAVI [57], AVSegFormer [14], CAVP [7] and our CPM, which can be compared with the ground truth (GT) Ambulance of the first row.

Ablation of Distribution Representation Table 5 provides a study of the GMM data distribution representation on AVSBench-Semantics [56]. The first row (Point Rep.) replaces the GMM module with a linear SoftMax for the classification output, introducing a learnable $C \times D_q$ feature map (i.e., D_q-dimensional feature embeddings for C semantic classes for the audio and visual prompting task). The second row (Dist Rep.) shows our approach with the CCDM module, showing that the utilisation of the distribution modelling leads to considerable improvements of 1.29% and 1.96% on mIoU and F_β respectively. The next rows of Tab. 5 show a hyper-parameter analysis on the number of GMM components (from $M = 1$ to $M = 7$ components). Results reveal that 3 GMM components enable the best fitting for the data distribution, leading to the top performance with a maximum improvement of 1.42% and 1.95% on mIoU and F_β.

5 Discussion and Conclusion

We introduced CPM, a new audio-visual training method designed for the Maskformer based framework to enhance bipartite matching stability and improve the efficacy of cross-modal attention for audio-visual segmentation. We proposed a class-conditional prompting learning strategy that combines class-agnostic queries with class-conditional queries, sampled from our iteratively updated generative model. The generated class-conditional queries are utilised to probe both the magnitude spectrogram and the image feature map, aiming to remove off-the-screen noise and bypass bipartite matching to produce a more stable learning process. Lastly, we extend the class-conditional queries to a new prompting-based audio-visual contrastive learning to explicitly constrain the cross-modal representations. SOTA results on AVS benchmarks suggest that CPM can be a valuable resource for future AVS research.

Limitations and future work. We acknowledge that the current adaptation of stereo audio into the transformer-based method has limitations, as it encodes positional and semantic information jointly within the transformer block. This contrasts with the typical approach of separately encoding these two types of information in the transformer-based framework. Our future work will concentrate on optimizing this framework through the integration of spatial reasoning.

References

1. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020)
2. Chen, H., Xie, W., Afouras, T., Nagrani, A., Vedaldi, A., Zisserman, A.: Localizing visual sounds the hard way. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16867–16876 (2021)
3. Chen, J., Lu, J., Zhu, X., Zhang, L.: Generative semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7111–7120 (2023)
4. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV). pp. 801–818 (2018)
5. Chen, T., Tan, Z., Gong, T., Chu, Q., Wu, Y., Liu, B., Lu, L., Ye, J., Yu, N.: Bootstrapping audio-visual segmentation by strengthening audio cues. arXiv preprint arXiv:2402.02327 (2024)
6. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
7. Chen, Y., Liu, Y., Wang, H., Liu, F., Wang, C., Carneiro, G.: A closer look at audio-visual semantic segmentation. arXiv e-prints pp. arXiv–2304 (2023)
8. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1290–1299 (2022)
9. Cheng, B., Schwing, A., Kirillov, A.: Per-pixel classification is not all you need for semantic segmentation. Advances in Neural Information Processing Systems **34**, 17864–17875 (2021)
10. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the em algorithm. Journal of the royal statistical society: series B (methodological) **39**(1), 1–22 (1977)
11. Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. International journal of computer vision **111**, 98–136 (2015)
12. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision **88**, 303–338 (2010)
13. Gao, R., Grauman, K.: Co-separating sounds of visual objects. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3879–3888 (2019)
14. Gao, S., Chen, Z., Chen, G., Wang, W., Lu, T.: Avsegformer: Audio-visual segmentation with transformer. arXiv preprint arXiv:2307.01146 (2023)
15. Gray, R.: Vector quantization. IEEE Assp Magazine **1**(2), 4–29 (1984)
16. Hao, D., Mao, Y., He, B., Han, X., Dai, Y., Zhong, Y.: Improving audio-visual segmentation with bidirectional generation. arXiv preprint arXiv:2308.08288 (2023)
17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)

18. Hu, X., Chen, Z., Owens, A.: Mix and localize: Localizing sound sources in mixtures. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10483–10492 (2022)
19. Huang, S., Li, H., Wang, Y., Zhu, H., Dai, J., Han, J., Rong, W., Liu, S.: Discovering sounding objects by audio queries for audio visual segmentation. arXiv preprint arXiv:2309.09501 (2023)
20. Jaegle, A., Borgeaud, S., Alayrac, J.B., Doersch, C., Ionescu, C., Ding, D., Koppula, S., Zoran, D., Brock, A., Shelhamer, E., et al.: Perceiver io: A general architecture for structured inputs & outputs. arXiv preprint arXiv:2107.14795 (2021)
21. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. Advances in neural information processing systems **33**, 18661–18673 (2020)
22. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)
23. Li, F., Zhang, H., Liu, S., Guo, J., Ni, L.M., Zhang, L.: Dn-detr: Accelerate detr training by introducing query denoising. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13619–13627 (2022)
24. Li, K., Yang, Z., Chen, L., Yang, Y., Xun, J.: Catr: Combinatorial-dependence audio-queried transformer for audio-visual video segmentation. arXiv preprint arXiv:2309.09709 (2023)
25. Li, T., Cao, P., Yuan, Y., Fan, L., Yang, Y., Feris, R.S., Indyk, P., Katabi, D.: Targeted supervised contrastive learning for long-tailed recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6918–6928 (2022)
26. Li, X., Wang, J., Xu, X., Peng, X., Singh, R., Lu, Y., Raj, B.: Towards robust audiovisual segmentation in complex environments with quantization-based semantic decomposition. arXiv preprint arXiv:2310.00132 (2023)
27. Liang, C., Wang, W., Miao, J., Yang, Y.: Gmmseg: Gaussian mixture based generative semantic segmentation models. Advances in Neural Information Processing Systems **35**, 31360–31375 (2022)
28. Liu, C., Ding, H., Jiang, X.: Gres: Generalized referring expression segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23592–23601 (2023)
29. Liu, C., Li, P., Qi, X., Zhang, H., Li, L., Wang, D., Yu, X.: Audio-visual segmentation by exploring cross-modal mutual semantics (2023)
30. Liu, C., Li, P., Zhang, H., Li, L., Huang, Z., Wang, D., Yu, X.: Bavs: Bootstrapping audio-visual segmentation by integrating foundation knowledge. arXiv preprint arXiv:2308.10175 (2023)
31. Liu, J., Ju, C., Ma, C., Wang, Y., Wang, Y., Zhang, Y.: Audio-aware query-enhanced transformer for audio-visual segmentation. arXiv preprint arXiv:2307.13236 (2023)
32. Liu, J., Wang, Y., Ju, C., Zhang, Y., Xie, W.: Annotation-free audio-visual segmentation. arXiv preprint arXiv:2305.11019 (2023)
33. Liu, S., Li, F., Zhang, H., Yang, X., Qi, X., Su, H., Zhu, J., Zhang, L.: Dab-detr: Dynamic anchor boxes are better queries for detr. arXiv preprint arXiv:2201.12329 (2022)
34. Liu, X., Carrington, P., Chen, X., Pavel, A.: What makes videos accessible to blind and visually impaired people? In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. pp. 1–14 (2021)

35. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015)
36. Mao, Y., Zhang, J., Xiang, M., Lv, Y., Zhong, Y., Dai, Y.: Contrastive conditional latent diffusion for audio-visual segmentation. arXiv preprint arXiv:2307.16579 (2023)
37. Mao, Y., Zhang, J., Xiang, M., Zhong, Y., Dai, Y.: Multimodal variational auto-encoder based audio-visual segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 954–965 (2023)
38. Martin, D.R., Fowlkes, C.C., Malik, J.: Learning to detect natural image boundaries using local brightness, color, and texture cues. IEEE transactions on pattern analysis and machine intelligence **26**(5), 530–549 (2004)
39. Meng, D., Chen, X., Fan, Z., Zeng, G., Li, H., Yuan, Y., Sun, L., Wang, J.: Conditional detr for fast training convergence. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3651–3660 (2021)
40. Mo, S., Morgado, P.: A closer look at weakly-supervised audio-visual source localization. arXiv preprint arXiv:2209.09634 (2022)
41. Mo, S., Morgado, P.: Localizing visual sounds the easy way. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII. pp. 218–234. Springer (2022)
42. Mo, S., Tian, Y.: Audio-visual grouping network for sound localization from mixtures. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10565–10574 (2023)
43. Mo, S., Tian, Y.: Av-sam: Segment anything model meets audio-visual localization and segmentation. arXiv preprint arXiv:2305.01836 (2023)
44. Murray, M.M., Wallace, M.T.: The neural bases of multisensory processes. Frontiers in Neuroscience (2011)
45. Reynolds, D.A., et al.: Gaussian mixture models. Encyclopedia of biometrics **741**(659-663) (2009)
46. Senocak, A., Ryu, H., Kim, J., Oh, T.H., Pfister, H., Chung, J.S.: Sound source localization is all about cross-modal alignment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7777–7787 (2023)
47. Siam, M., Oreshkin, B.N., Jagersand, M.: Amp: Adaptive masked proxies for few-shot segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5249–5258 (2019)
48. Sun, Z., Cao, S., Yang, Y., Kitani, K.M.: Rethinking transformer-based set prediction for object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3611–3620 (2021)
49. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
50. Wang, W., Zhou, T., Yu, F., Dai, J., Konukoglu, E., Van Gool, L.: Exploring cross-image pixel contrast for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7303–7313 (2021)
51. Wang, Y., Liu, W., Li, G., Ding, J., Hu, D., Li, X.: Prompting segmentation with sound is generalizable audio-visual source localizer. arXiv preprint arXiv:2309.07929 (2023)
52. Wu, J., Jiang, Y., Sun, P., Yuan, Z., Luo, P.: Language as queries for referring video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4974–4984 (2022)

53. Xu, P., Zhu, X., Clifton, D.A.: Multimodal learning with transformers: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
54. Yang, Q., Nie, X., Li, T., Gao, P., Guo, Y., Zhen, C., Yan, P., Xiang, S.: Cooperation does matter: Exploring multi-order bilateral relations for audio-visual segmentation. arXiv preprint arXiv:2312.06462 (2023)
55. Zhang, H., Li, F., Xu, H., Huang, S., Liu, S., Ni, L.M., Zhang, L.: Mp-former: Mask-piloted transformer for image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18074–18083 (2023)
56. Zhou, J., Shen, X., Wang, J., Zhang, J., Sun, W., Zhang, J., Birchfield, S., Guo, D., Kong, L., Wang, M., et al.: Audio-visual segmentation with semantics. arXiv preprint arXiv:2301.13190 (2023)
57. Zhou, J., Wang, J., Zhang, J., Sun, W., Zhang, J., Birchfield, S., Guo, D., Kong, L., Wang, M., Zhong, Y.: Audio–visual segmentation. In: European Conference on Computer Vision. pp. 386–403. Springer (2022)
58. Zhou, T., Zhang, M., Zhao, F., Li, J.: Regional semantic contrast and aggregation for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4299–4309 (2022)