AVSegFormer: Audio-Visual Segmentation with Transformer

Shengyi Gao[1], Zhe Chen[1], Guo Chen[1], Wenhai Wang[2], Tong Lu[1†]
[1]Nanjing University [2]The Chinese University of Hong Kong

ABSTRACT

The combination of audio and vision has long been a topic of interest in the multi-modal community. Recently, a new audio-visual segmentation (AVS) task has been introduced, aiming to locate and segment the sounding objects in a given video. This task demands audio-driven pixel-level scene understanding for the first time, posing significant challenges. In this paper, we propose AVSegFormer, a novel framework for AVS tasks that leverages the transformer architecture. Specifically, we introduce audio queries and learnable queries into the transformer decoder, enabling the network to selectively attend to interested visual features. Besides, we present an audio-visual mixer, which can dynamically adjust visual features by amplifying relevant and suppressing irrelevant spatial channels. Additionally, we devise an intermediate mask loss to enhance the supervision of the decoder, encouraging the network to produce more accurate intermediate predictions. Extensive experiments demonstrate that AVSegFormer achieves state-of-the-art results on the AVS benchmark. The code is available at https://github.com/vvvb-github/AVSegFormer.

CCS CONCEPTS

• **Computing methodologies** → **Video segmentation**; Scene understanding; • **Information systems** → *Video search*.

KEYWORDS

audio-visual, multi-modal segmentation, transformer

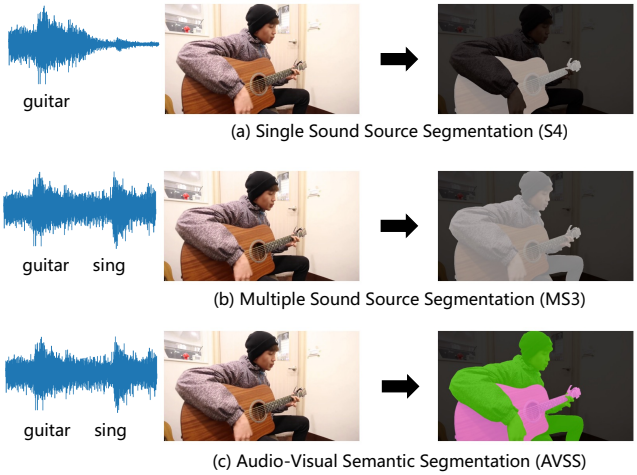

Figure 1: Illustration of audio-visual segmentation (AVS). AVS aims to segment sounding objects from video frames according to a given audio. In the S4 sub-task, the input audio only contains one sound source, while in MS3 the input audio has multiple sound sources. Besides, S4 and MS3 only require binary segmentation, whereas AVSS requires more difficult multiple-category semantic segmentation.

1 INTRODUCTION

Audio and vision are closely intertwined modalities that play crucial roles in the perception of the world. For example, we rely heavily on auditory and visual cues to comprehend and navigate our surroundings. These underlying connections motivated many audio-visual tasks, such as audio-visual correspondence [3, 4], audio-visual event localization [28, 30, 45], audio-visual video parsing [29, 44, 50], and sound source localization [3, 4, 42]. However, due to the lack of pixel-level annotations for these tasks, they are often limited to the frame/temporal level, which makes them become audio-informed image classification problems eventually.

Recently, a novel audio-visual segmentation (AVS) [56] task has been proposed, which aims to segment sounding objects from video frames corresponding to a given audio. It includes three sub-tasks: single sound source segmentation (S4), multiple sound source segmentation (MS3), and audio-visual semantic segmentation (AVSS). Figure 1 illustrates the objectives of the three sub-tasks. Compared to previous audio-visual tasks, the AVS task presents a greater challenge as it requires the network to not only locate the audible frames but also delineate the shape of the sounding objects [56, 57]. This demands the alignment of multiple modalities and necessitates a detailed understanding of the scenarios. These unique characteristics of the AVS task make many existing methods [6, 35, 38] designed for other audio-visual tasks may be sub-optimal on AVS. Therefore, designing new methods tailored for AVS becomes necessary.

For the brand-new AVS task, AVSBench [56] designed a strong baseline method that achieves state-of-the-art audio-visual segmentation performance. Figure 2(a) illustrates its network architecture, which incorporates a modality fusion module before the convolution-based decoder (*i.e.*, Semantic FPN [25]) to enable audio-visual segmentation. This design is simple yet effective, but some inherent flaws of convolution still limit it. (1) The effective receptive field of convolutions is relatively small. Even with a deep decoder, the audio feature still cannot capture long-range visual dependencies, which restricts its performance. (2) Convolution is an operator with static weights, which is difficult to provide different visual features conditioned by the input audio.

To remedy these issues, we propose **AVSegFormer**, a novel framework for audio-visual segmentation with transformers. The brief architecture is shown in Figure 2(b). First, we introduce audio queries along with learnable queries into the segmentation decoder, enabling the network to selectively attend to interested visual features. Second, we present an audio-visual mixer, which can dynamically adjust visual features by amplifying relevant and suppressing irrelevant spatial channels, allowing visual features to adapt to diverse audio features. Third, an intermediate mask loss is designed to enhance the supervision of the decoder, which

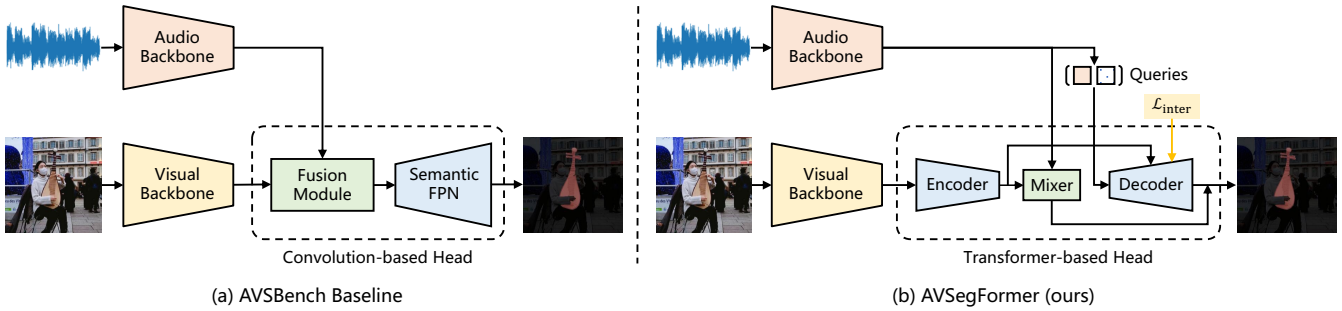

Figure 2: Overview of the AVSBench baseline [57] and our AVSegFormer. (a) The baseline method combines a modality fusion module with the convolution-based decoder (*i.e.*, Semantic FPN [25]) for audio-visual segmentation. (b) The proposed AVSegFormer performs audio-visual segmentation with transformer-based encoder-decoder architecture. It has three key designs: (1) audio and learnable queries, (2) the audio-visual mixer, and (3) the intermediate mask loss $\mathcal{L}_{\text{inter}}$.

encourages the network to produce more accurate intermediate predictions and helps refine the final segmentation outputs.

We evaluate AVSegFormer on the three sub-tasks of AVS, including S4, MS3, and AVSS, with widely-used backbones ResNet-50 [19] and PVTv2 [48]. Our experimental results demonstrate that AVSegFormer significantly outperforms existing state-of-the-art methods, such as LGVT [55], SST [16], iGAN [36], and AVSBench baseline [57]. Specifically, AVSegFormer-R50 achieves 76.45, 49.53, and 24.93 mIoU on S4, MS3, and AVSS, surpassing the AVSBench-R50 by 3.66, 1.65, and 4.75 mIoU, respectively. Furthermore, using PVTv2 as the backbone, AVSegFormer yields consistently higher segmentation performance on all three sub-tasks, setting new state-of-the-art records of 82.06, 58.36, and 36.66 mIoU.

Overall, our contributions to this work are four-fold.

- We employ audio queries and learnable queries to selectively attend to relevant visual features, overcoming the limitations of previous convolutional-based approaches.
- We design an audio-visual mixer, which can amplify relevant visual features and suppress irrelevant visual features in response to audio cues.
- We present an intermediate mask loss that provides additional supervision, encouraging more accurate intermediate results and improving final prediction.
- We conduct extensive experiments on three sub-tasks of AVS. These results demonstrate that our method achieves state-of-the-art performance on the AVS benchmark.

2 RELATED WORKS

2.1 Multi-Modal Tasks

In recent years, multi-modal tasks have gained significant attention in the research community due to their potential for advancing the understanding of complex real-world scenarios. These tasks aim to exploit the complementary information from different modalities, such as vision, audio, and text, to improve the overall performance of various applications. Among these, text-visual tasks have attracted considerable interest from researchers. Numerous works focus on related tasks, such as visual question answering [1, 2, 8, 49] and visual grounding [12, 14, 24, 31]. In addition to text-visual tasks, audio-visual tasks are emerging as hot spots in the multi-modal field. Several related tasks include audio-visual correspondence [3, 4], audio-visual event localization [28, 30, 45], and sound source localization [3, 4, 42]. Concurrently, some works [47, 59] have proposed unified architectures capable of handling nearly all types of modalities using a consistent format for embeddings.

Most of these works are based on transformer architecture [46], which demonstrates a strong cross-modal capability. Their achievements highlight the reliability of transformers in the multi-modal field. As a recently proposed multi-modal task, audio-visual segmentation (AVS) [56, 57] shares many commonalities with the aforementioned tasks. The pioneering works in these areas have significantly inspired our research of AVSegFormer.

2.2 Sound Source Localization

Sound source localization (SSL) is an important problem in the audio-visual multi-modal community and is also the most related task to audio-visual segmentation. It aims to locate the regions for sound sources in a video sequence, but the results are usually represented as heat maps. The major challenge in the SSL problem is dealing with multiple sound sources. In prior arts, Hu et al. [21] divided audio-visual features into multiple clustering centers and used center distance as a supervised signal to locate paired audio-visual information. Qian et al. [38] trained an audio-visual correspondence model to extract coarse feature representations of audio and visual signals and used Grad-CAM [41] to locate specific categories of features. Hu et al. [22] adopted a two-stage approach by first learning audio-visual semantics under a single sound source and then using this knowledge to help locate multiple sound sources. Besides, Rouditchenko et al. [40] tackled the problem by unraveling the concept of categories in neural networks.

Although these SSL methods indicate which areas in the image emit sound, they do not clearly depict the shape of the object, which is another challenge in the AVS task. Furthermore, the above methods all rely on unsupervised learning when capturing the shape of the detected object, which may result in inaccurate localization. Nevertheless, handling multiple sound sources in SSL methods offers valuable insights and references for our work.

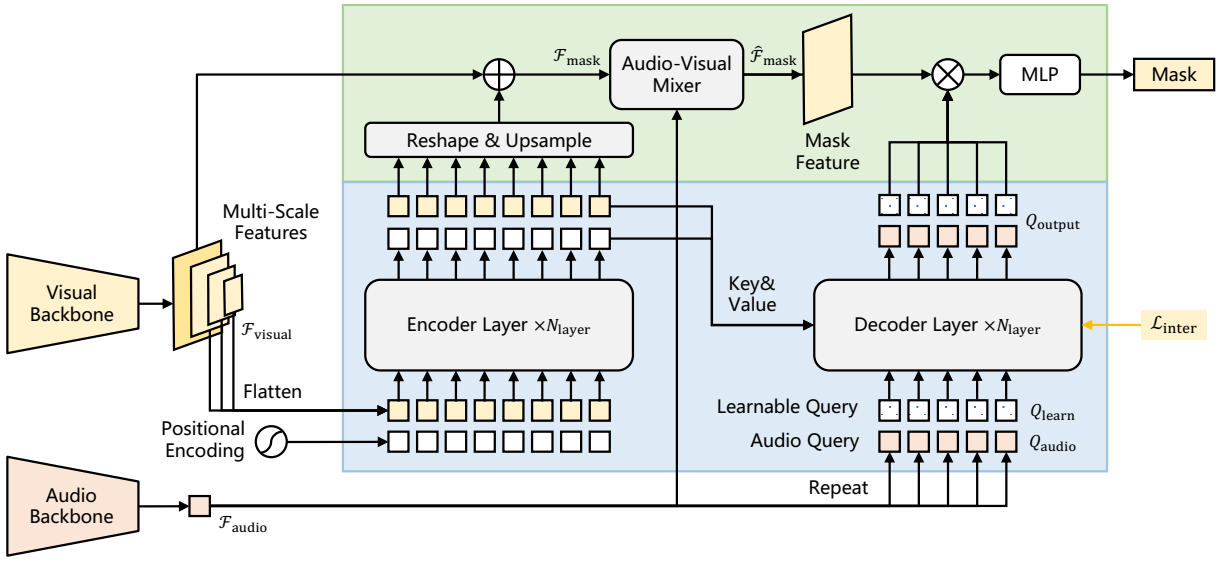

Figure 3: Overall architecture of AVSegFormer. We propose three key components in this framework: (1) Audio features and learnable queries are utilized as decoder inputs, enabling the model to focus on relevant features of sounding objects. (2) An audio-visual mixer is applied to dynamically adjust visual features based on auditory information, improving the model's adaptability. (3) An intermediate mask loss $\mathcal{L}_{\text{inter}}$ is designed to enhance the effectiveness of the training process. The entire transformer structure is depicted in blue, while the mask generation process is shown in green.

2.3 Vision Transformer

During the past few years, Transformer [46] has experienced rapid development in natural language processing. Following this success, the Vision Transformer (ViT) [15] emerged, bringing the transformer into the realm of computer vision and yielding impressive results. Numerous works [9, 32, 48] have built upon ViT, leading to the maturation of vision transformers. As the performance of vision transformers continues to advance, they are increasingly replacing CNNs as the mainstream paradigm in the field of computer vision, especially in object detection and image segmentation tasks.

Carion et al. [5] proposed the DETR model and designed a novel bipartite matching loss based on the transformer architecture, paving the way for new research directions in vision transformers. Subsequently, improved frameworks such as Deformable DETR [58] and DINO [54] are proposed, introducing mechanisms like deformable attention and denoise training. These arts take vision transformers to new heights. The remarkable performance of vision transformers has also inspired us to apply this paradigm to AVS tasks, anticipating further advancements in the field.

2.4 Image Segmentation

Image segmentation is a critical visual task that involves partitioning an image into distinct segments or regions. It includes three different tasks: instance segmentation [7, 18], semantic segmentation [33, 39, 51], and panoptic segmentation [25, 27, 52]. Instance segmentation predicts the mask of each object instance and its corresponding category, while semantic segmentation needs to classify each pixel in the image into different semantic categories. Panoptic segmentation unifies instance and semantic segmentation tasks and predicts the mask of each object instance or background segment. Early research proposed specialized models for these tasks, such as Mask R-CNN [18] and HTC [7] for instance segmentation, or FCN [33] and U-Net [39] for semantic segmentation. After panoptic segmentation was proposed, some related research [25, 27, 52] were conducted and designed universal models for both tasks.

The recent introduction of the transformer has led to the development of new models that can unify all the segmentation tasks. Mask2Former [10] is one such model that introduces mask attention into the transformer and improves MaskFormer [11]. Mask DINO [26] is a unified transformer-based framework for both detection and segmentation. Recently, OneFormer [23] presented a new multi-task universal image segmentation framework with transformers. These models have brought image segmentation to a new level, demonstrating the potential of transformer architecture in vision tasks. Considering that the AVS task involves segmentation, these methods have significantly contributed to our work.

3 METHOD

3.1 Overall Architecture

Figure 3 illustrates the overall architecture of our method. In contrast to previous CNN-based segmentation methods [56, 57], we design a query-based framework to leverage the transformer architecture. Specifically, our model combines audio queries with learnable queries, allowing it to adjust its focus on visual features dynamically. Additionally, we design an audio-visual mixer and an intermediate mask loss $\mathcal{L}_{\text{inter}}$ as auxiliary components. The audio-visual mixer aids in amplifying relevant features and suppressing irrelevant ones, while the intermediate mask loss helps supervise intermediate predictions for enhancing performance.

The pipeline of AVSegFormer consists of four stages. First, a visual backbone and an audio backbone are employed to extract features from video and audio frames, respectively. Second, the transformer encoder refines the visual features and generates an initial mask feature, which serves as the basis for predicting the final mask. Third, an audio-visual mixer is utilized to amplify feature channels relevant to sounding objects while suppressing those that are irrelevant. Lastly, the transformer decoder incorporates audio queries and learnable queries, capturing richer features about sounding objects and predicting the final mask.

3.2 Multi-Modal Representation

Visual encoder. We follow the feature extraction process adopted in previous methods [56, 57], which involves using a visual backbone and an audio backbone to extract video and audio features, respectively. For videos, the dataset provides pre-extracted frame images, making the process similar to image feature extraction. Specifically, the input video frames are denoted as $x_{\text{visual}} \in \mathbb{R}^{T\times 3\times H\times W}$, in which T denotes the number of frames. Then, we use a visual backbone (*e.g.*, ResNet-50 [19]) to extract hierarchical visual features $\mathcal{F}_{\text{visual}}$, which can be written as:

$$\mathcal{F}_{\text{visual}} = \{\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3, \mathcal{F}_4\}, \tag{1}$$

in which $\mathcal{F}_i \in \mathbb{R}^{T\times C_i\times \frac{H}{2^{i+1}}\times \frac{W}{2^{i+1}}}$ and $i \in [1, 2, 3, 4]$. C_i represents the number of output channels of the $i-$stage of the visual backbone. In other words, $\mathcal{F}_{\text{visual}}$ is a list of multi-scale features, where each feature map is half resolution of the previous one.

Audio encoder. The process of audio feature extraction follows the VGGish [20] method. Initially, the audio is resampled to 16kHz mono audio $x_{\text{audio}} \in \mathbb{R}^{N_{\text{samples}}\times 96\times 64}$, where N_{samples} is related to the audio duration. Then, we perform a short-time Fourier transform to obtain a mel spectrum. The mel spectrum is calculated by mapping the spectrum to a 64th-order mel filter bank and subsequently fed into the VGGish model to obtain the audio features $\mathcal{F}_{\text{audio}} \in \mathbb{R}^{T\times d_{\text{model}}}$, where T represents the number of frames and d_{model} defaults to 128 in the VGGish.

Feature transformation. For the convenience of subsequent usage, we use multiple linear layers to unify the number of channels for all features. Specifically, all features extracted by the visual backbone and audio backbone are transformed into D dimensions, which also equals the embedding dimension of the transformer encoder and decoder. Typically, D is set to 256 by default.

3.3 Transformer Encoder

The transformer encoder is responsible for multi-scale feature fusion and mask feature generation. Specifically, we collect the visual features of three resolutions (*i.e.*, 1/8, 1/16, and 1/32), and then flatten and concatenate them as the input for the transformer encoder. After that, the output features of the transformer encoder are reshaped to their original shapes, and the 1/8-scale features are taken out separately and 2× upsampled. Finally, we add the upsampled features to the 1/4-scale features from the visual backbone and obtain the initial mask features $\mathcal{F}_{\text{mask}} \in \mathbb{R}^{T\times D\times h\times w}$, where $h = \frac{H}{4}, w = \frac{W}{4}$, and D is the embedding dimension of the transformer encoder and decoder.

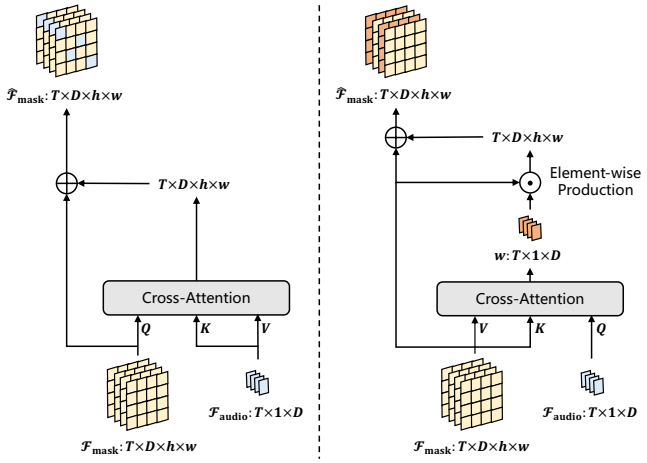

Figure 4: Architecture of the audio-visual mixer. (a) Our initial design incorporates a cross-attention mixer, which fails to deliver satisfactory results. (b) We ultimately adopt the design of channel-attention mixer, which demonstrates significantly improved performance.

3.4 Audio-Visual Mixer

As illustrated in Figure 3, the segmentation mask is generated based on the mask feature, which plays a crucial role in the final prediction results. However, since the audio features can vary widely, a static network may not be able to capture all of the relevant information. This limitation may hinder the model's ability to identify potential sounding objects accurately.

To address this issue, we propose an audio-visual mixer as shown in Figure 4(b). The design of this module is based on channel attention, which allows the model to selectively amplify or suppress different visual channels depending on the audio input, improving its ability to capture complex audio-visual relationships. Specifically, the mixer learns a set of weights ω through audio-visual cross-attention, and applies them to highlight the relevant channels. The whole process can be represented as follows:

$$\begin{aligned} \omega &= \frac{\mathcal{F}_{\text{audio}}\mathcal{F}_{\text{mask}}{}^{T}}{\sqrt{D/n_{\text{head}}}}\mathcal{F}_{\text{mask}}, \\ \hat{\mathcal{F}}_{\text{mask}} &= \mathcal{F}_{\text{mask}} + \mathcal{F}_{\text{mask}} \odot \omega. \end{aligned} \tag{2}$$

Here, $\mathcal{F}_{\text{audio}}$ and $\mathcal{F}_{\text{mask}}$ represent the input audio features and the initial mask features, and $\hat{\mathcal{F}}_{\text{mask}}$ denotes the mixed mask features. Besides, n_{head} means the number of attention heads, which is set to 8 by default following common practice.

3.5 Transformer Decoder

Audio query. The transformer decoder is designed to learn semantic-rich features of the sounding objects. We repeat the audio feature $\mathcal{F}_{\text{audio}} \in \mathbb{R}^{T\times 1\times D}$ to the number of queries N_{query}, and employ it as the audio queries $Q_{\text{audio}} \in \mathbb{R}^{T\times N_{\text{query}}\times D}$. As the decoding process continues, the queries continuously aggregate visual information from the encoder's outputs. The output queries Q_{output} ultimately

combine the auditory and visual modalities and contain richer features of the sounding objects.

Learnable query. To enhance the model's ability to capture in-depth information about the sounding object, we add learnable queries $Q_{\text{learn}} \in \mathbb{R}^{T\times N_{\text{query}}\times D}$ to the audio queries Q_{audio}. The learnable queries enhance our model's adaptability for various AVS tasks and datasets. Specifically, it enables the model to learn dataset-level contextual information, and adjust the attention allocated to different target categories. Furthermore, learnable queries empower the model with a more robust capability to process target semantics, ultimately improving segmentation accuracy.

Mask generation. To generate the segmentation masks, we multiply the mask feature $\hat{\mathcal{F}}_{\text{mask}} \in \mathbb{R}^{T\times D\times h\times w}$ obtained from the audio-visual mixer with the output queries Q_{output} from the decoder. Then, an MLP is used to integrate different channels. Additionally, we introduce a residual connection to ensure that the fusion of auditory information does not result in excessive loss of visual information. Finally, the model predicts the segmentation mask $\mathcal{M}$ through a simple linear layer:

$$\mathcal{M} = \text{Linear}(\hat{\mathcal{F}}_{\text{mask}} + \text{MLP}(\hat{\mathcal{F}}_{\text{mask}} \cdot Q_{\text{output}})). \tag{3}$$

Here, MLP$(\cdot)$ represents the MLP process, and Linear$(\cdot)$ means the linear layer. The output $\mathcal{M} \in \mathbb{R}^{T\times N_{\text{class}}\times h\times w}$ is the predicted segmentation mask, with the dimension N_{class} denotes the number of semantic classes.

3.6 Loss Function

Intermediate mask loss. With the introduction of the deep transformer decoder, achieving satisfying performance only by supervising the final prediction becomes more challenging. To address this issue, we design an intermediate mask loss $\mathcal{L}_{\text{inter}}$ as the auxiliary loss, which supervises each layer of the transformer decoder.

Specifically, the intermediate mask loss is based on the cross-attention operation of each decoder layer. First, we utilize the input queries $Q \in \mathbb{R}^{T\times N_{\text{query}}\times D}$ and keys $K \in \mathbb{R}^{T\times \Sigma h_i w_i\times D}$ to compute the attention map $\mathcal{A} = \frac{QK^T}{\sqrt{D/n_{\text{head}}}} \in \mathbb{R}^{T\times N_{\text{query}}\times \Sigma h_i w_i}$. Then, we multiply the attention map $\mathcal{A}$ with the output queries of the current decoder layer and feed it through a simple FPN [43] to predict a mask $\hat{\mathcal{M}} \in \mathbb{R}^{T\times N_{\text{class}}\times h\times w}$, as shown in Figure 5(c). Finally, we calculate the Dice loss [37] between this predicted mask and the ground truth to obtain the intermediate mask loss $\mathcal{L}_{\text{inter}}$.

Total loss. The total loss function comprises two parts: IoU loss and mask loss. As shown in Figure 5(a), the IoU loss $\mathcal{L}_{\text{IoU}}$ is calculated by comparing the final segmentation mask with the ground truth. Here, we also use Dice loss [37] for supervision. Considering that in AVS tasks, the proportion of segmented objects occupying the entire image is relatively small, the model can better focus on the foreground and reduce interference from the background by using Dice loss. Thus, the total loss of our method is as follows:

$$\mathcal{L} = \mathcal{L}_{\text{IoU}} + \lambda \sum_{i=1}^{N_{\text{layer}}} \mathcal{L}^i_{\text{inter}}. \tag{4}$$

Here, N_{layer} represents the number of decoder layers, and λ is a coefficient that controls the effect of the auxiliary loss. We choose $\lambda = 0.1$ in our experiments as it performs best.

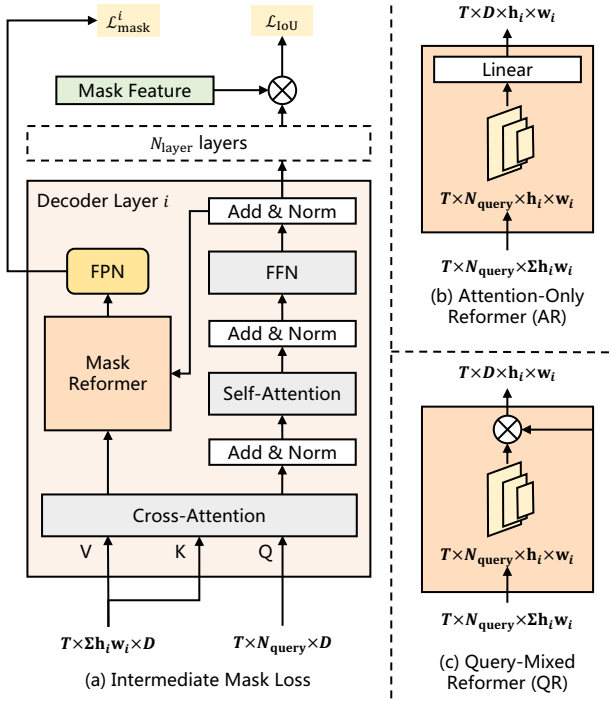

Figure 5: The design of intermediate mask loss. (a) We generate features for predicting auxiliary masks using a mask reformer module. (b) Initially, we designed an attention-only reformer that solely relies on attention output for mask reforming, resulting in minimal improvements. (c) As an enhanced design, we propose a query-mixed reformer that combines attention masks with queries, which ultimately proves more effective and is finally adopted in our approach.

4 EXPERIMENTS

4.1 Dataset

AVSBench-Object [57] is an audio-visual dataset specifically designed for the audio-visual segmentation task, containing pixel-level annotations. The videos are downloaded from YouTube and cropped to 5 seconds, with one frame per second extracted for segmentation. The dataset includes two subsets: a semi-supervised single sound source subset for single sound source segmentation (S4), and a fully supervised multi-source subset for multiple sound source segmentation (MS3). **S4 subset:** The S4 subset contains a total of 4, 932 videos, with 3, 452 videos for training, 740 for validation, and 740 for testing. The target objects cover 23 different categories, including humans, animals, vehicles, and musical instruments. Besides, this subset is trained in a semi-supervised manner, where each video contains five frames, but only the first frame is annotated. **MS3 subset:** The MS3 subset includes 424 videos, with 286 training, 64 validation, and 64 testing videos, covering the same categories as the S4 subset. Additionally, unlike S4, the MS3 subset is full-supervised with all five frames annotated in training.

AVSBench-Semantic [56] is an extension of the AVSBench-Object dataset, which offers additional semantic labels that are not

Table 1: Comparison with state-of-the-art methods on the AVS benchmark. All methods are evaluated on three AVS sub-tasks, including single sound source segmentation (S4), multiple sound source segmentation (MS3), and audio-visual semantic segmentation (AVSS). The evaluation metrics are F-score and mIoU. The higher the better.

Method	Backbone	S4		MS3		AVSS		Reference
		F-score	mIoU	F-score	mIoU	F-score	mIoU	
LVS [6]	ResNet-50 [19]	51.0	37.94	33.0	29.45	–	–	CVPR'2021
MSSL [38]	ResNet-18 [19]	66.3	44.89	36.3	26.13	–	–	ECCV'2020
3DC [35]	ResNet-34 [19]	75.9	57.10	50.3	36.92	21.6	17.27	BMVC'2020
SST [16]	ResNet-101 [19]	80.1	66.29	57.2	42.57	–	–	CVPR'2021
AOT [53]	Swin-B [32]	–	–	–	–	31.0	25.40	NeurIPS'2021
iGAN [36]	Swin-T [32]	77.8	61.59	54.4	42.89	–	–	ArXiv'2022
LGVT [55]	Swin-T [32]	87.3	74.94	59.3	40.71	–	–	NeurIPS'2021
AVSBench-R50 [56]	ResNet-50 [19]	84.8	72.79	57.8	47.88	25.2	20.18	ECCV'2022
AVSegFormer-R50 (ours)	ResNet-50 [19]	**85.9**	**76.45**	**62.8**	**49.53**	**29.3**	**24.93**	–
AVSBench-PVTv2 [56]	PVTv2 [48]	87.9	78.74	64.5	54.00	35.2	29.77	ECCV'2022
AVSegFormer-PVTv2 (ours)	PVTv2 [48]	**89.9**	**82.06**	**69.3**	**58.36**	**42.0**	**36.66**	–

Table 2: AVSegFormer ablation experiments on the S4 and MS3 subsets. We report the performance of F-score (denoted as F) and mIoU. If not specified, the default settings are: the number of queries N_{query} is 300, the queries in the decoder are learnable, the audio-visual mixer is used, and the intermediate mask loss is applied. Default settings are marked in gray.

(a) Effect of the number of queries. We find that 300 queries work best than other configurations.

N_{query}	S4		MS3	
	mIoU	F	mIoU	F
1	81.1	88.6	57.1	67.7
100	81.4	88.9	57.5	68.2
200	81.9	89.6	58.0	68.4
300	82.0	89.9	58.3	69.3

(b) Effect of learnable query. Using learnable queries along with audio queries improves the performance.

learnable queries	S4		MS3	
	mIoU	F	mIoU	F
✓	82.0	89.9	58.3	69.3
×	81.8	89.6	57.2	67.9

(c) Effect of audio-visual mixer. It is shown that the channel-attention (CHA) mixer works better.

mixer	S4		MS3	
	mIoU	F	mIoU	F
–	81.4	89.0	57.4	67.9
CRA	81.7	89.7	57.8	68.2
CHA	82.0	89.9	58.3	69.3

(d) Effect of intermediate mask loss. It shows that query-mixed reformer (QR) improves more.

mask reformer	S4		MS3	
	mIoU	F	mIoU	F
–	81.1	88.4	57.3	68.3
AR	81.7	89.6	57.9	68.8
QR	82.0	89.9	58.3	69.3

available in the original AVSBench-Object dataset. It is designed for audio-visual semantic segmentation (AVSS). In addition, the videos in AVSBench-Semantic are longer, with a duration of 10 seconds, and 10 frames are extracted from each video for prediction. It combines semi-supervised and fully-supervised manners, with labels provided for the first and first five frames of videos inherited from the S4 and MS3 subsets, respectively. For the newly collected 7,000 videos, a complete 10-frame label is provided. Overall, the AVSBench-Semantic dataset has increased in size by approximately three times compared to the original AVSBench-Object dataset, with 8,498 training, 1,304 validation, and 1,554 test videos.

Metric. These benchmark datasets employ mean intersection over union (mIoU) and F-score as the evaluation metrics.

4.2 Implementation Details

Our method is evaluated on three AVS sub-tasks, including single sound source segmentation (S4), multiple sound source segmentation (MS3), and audio-visual semantic segmentation (AVSS).

Details. We train our AVSegFormer models for the three sub-tasks using an NVIDIA V100 GPU. Consistent with previous works [56, 57], we employ AdamW [34] as the optimizer, with a batch size of 4 and an initial learning rate of 2×10^{-5}. All video frames are resized to 224×224 resolution. For the S4 and MS3 subsets, each video contains 5 frames, while each video in AVSS contains 10 frames. Since the MS3 subset is quite small, we train it for 60 epochs, while the S4 and AVSS subsets are trained for 30 epochs. The encoder and decoder in our AVSegFormer both are comprised of 6 layers with an embedding dimension of 256. We set the coefficient of the proposed intermediate mask loss to 0.1 for the best performance. More detailed training settings can be found in Table 3.

4.3 Comparison with Prior Arts

To verify the effectiveness of our method, we conducted a comprehensive comparison between our AVSegFormer and existing methods on the AVS benchmark [56, 57]. For fairness, we employ the ImageNet-1K [13] pre-trained ResNet-50 [19] or PVTv2 [48] as the backbone to extract visual features, and the AudioSet [17] pre-trained VGGish [20] to extract audio features.

Comparison with methods from related tasks. Firstly, we compare our AVSegFormer with state-of-the-art methods from

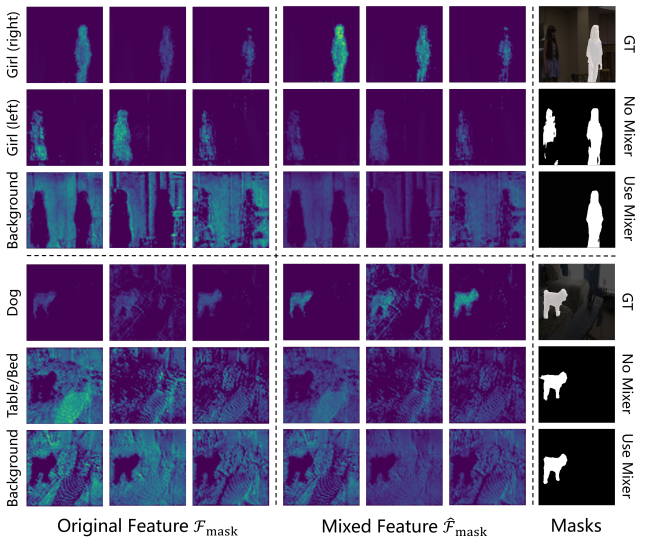

Figure 6: Comparison between the original features $\mathcal{F}_{\text{mask}}$ and the mixed features $\hat{\mathcal{F}}_{\text{mask}}$. We show two examples with 9 channels of both two features and the ground truth. As can be seen, the features of the ground truth (right girl and dog) are amplified while those of the non-sounding objects (left girl, table/bed, or background) are suppressed.

three AVS-related tasks, including sound source localization (LVS [6] and MSSL [38]), video object segmentation (3DC [35], SST [16] and AOT [53]), and salient object detection (iGAN [36] and LGVT [55]). These results are collected from the AVS benchmark [56, 57], which are transferred from the original tasks to the AVS tasks.

As shown in Table 1, our AVSegFormer exceeds these methods by large margins. For instance, on the S4 subset, AVSegFormer-R50 achieves an impressive mIoU of 76.45, which is 1.51 points higher than the best LGVT. Although LGVT has a better Swin-T [32] backbone, our AVSegFormer with ResNet-50 backbone still performs better regarding mIoU. In addition, AVSegFormer-PVTv2 produces an outstanding mIoU of 82.06 and an F-score of 89.9 on this subset, which is 7.12 mIoU and 2.6 F-score higher than LGVT, respectively. On the MS3 subset, AVSegFormer-R50 outperforms the best iGAN with 6.64 mIoU and 8.4 F-score, while AVSegFormer-PVTv2 further raised the bar with an exceptional improvement of 15.47 mIoU and 14.9 F-score. On the AVSS subset, our AVSegFormer-R50 yields 24.93 mIoU and 29.3 F-score, which are slightly lower than the best method AOT. Nevertheless, AVSegFormer-PVTv2 obtains an impressive performance of 36.66 mIoU and 42.0 F-score, surpassing AOT by 11.26 mIoU and 11.0 F-score, respectively.

Comparison with AVSBench baseline [56, 57]. Then, we compare our AVSegFormer with the AVSBench baseline, which is the current state-of-the-art method for audio-visual segmentation. As reported in Table 1, on the S4 subset, AVSegFormer-R50 achieves 3.66 mIoU and 1.1 F-score improvements over AVSBench-R50, while AVSegFormer-PVTv2 surpasses AVSBench-PVTv2 by 3.32 mIoU and 2.0 F-score. On the MS3 subset, AVSegFormer-PVTv2 surpasses AVSBench-PVTv2 with a margin of 4.36 mIoU and 4.8 F-score. On the AVSS subset, AVSegFormer-R50 and AVSegFormer-PVTv2 achieve significant results with an mIoU improvement of 4.75 and 6.89, and a substantial F-score improvement of 4.1 and 6.8. These results demonstrate that AVSegFormer outperforms the AVS-Bench baseline on all sub-tasks, becoming a new state-of-the-art method for audio-visual segmentation.

4.4 Ablation Study

In this section, we conduct ablation experiments to verify the effectiveness of each key design in the proposed AVSegFormer. Specifically, we adopt PVTv2 [48] as the backbone and conduct extensive experiments on the S4 and MS3 sub-tasks. Other training settings are the same as in Section 4.2.

Number of queries. To analyze the impact of the number of queries on the model's performance, we conducted experiments with varying numbers of queries for the decoder input, specifically 1, 100, 200, and 300. Our results reveal a positive correlation between the number of queries and the model performance, with the optimal performance obtained when the number of queries was set to 300. Table 2a presents these findings.

Effect of learnable queries. We further investigated the impact of learnable queries in the decoder inputs. As shown in Table 2b, the improvement due to the learnable queries is relatively small in the single sound source task (S4), while it brings significant improvement in the multiple sound source task (MS3). This can be attributed to the complexity of sounding objects. In S4, since there is only one sounding object that remains unchanged, its auditory features are distinct enough. In contrast, in MS3, it becomes relatively difficult for the model to learn target features solely based on audio features, due to the mixing and change of sound sources. The learnable queries partially compensate for this limitation. The results indicate that although the concept of learnable queries is simple, it can produce excellent results in complex scenarios.

Effect of audio-visual mixer. We then studied the impact of the audio-visual mixer in our model. There are two versions designed for this module, as illustrated in Figure 4. The cross-attention mixer (CRA) utilizes visual features as queries and audio features as keys/values for cross-attention, aiming to bring audio information into visual features in the early stages. The channel-attention mixer (CHA) introduced the mechanism of channel attention with audio features as queries and visual features as keys/values. As presented in Table 2c, the design of CHA brought greater performance improvement compared to CRA.

In addition, to validate our suppose that the audio-visual mixer can amplify relevant features while suppressing irrelevant ones, we also visualize features before and after the audio-visual mixer along with their predicted masks. Figure 6 presents the visualization results. We compare the original features $\mathcal{F}_{\text{mask}}$ with the mixed features $\hat{\mathcal{F}}_{\text{mask}}$ generated by the audio-visual mixer and selected 9 channels for rendering. It is evident that for the sounding object (right girl and dog), the mixer effectively enhanced its features. Meanwhile, the non-sounding objects (left girl, table/bed, or background) experienced some degree of suppression. As a result, the predicted mask without mixer may segment out the wrong target (left girl), and the mixer can lead to a more accurate prediction of

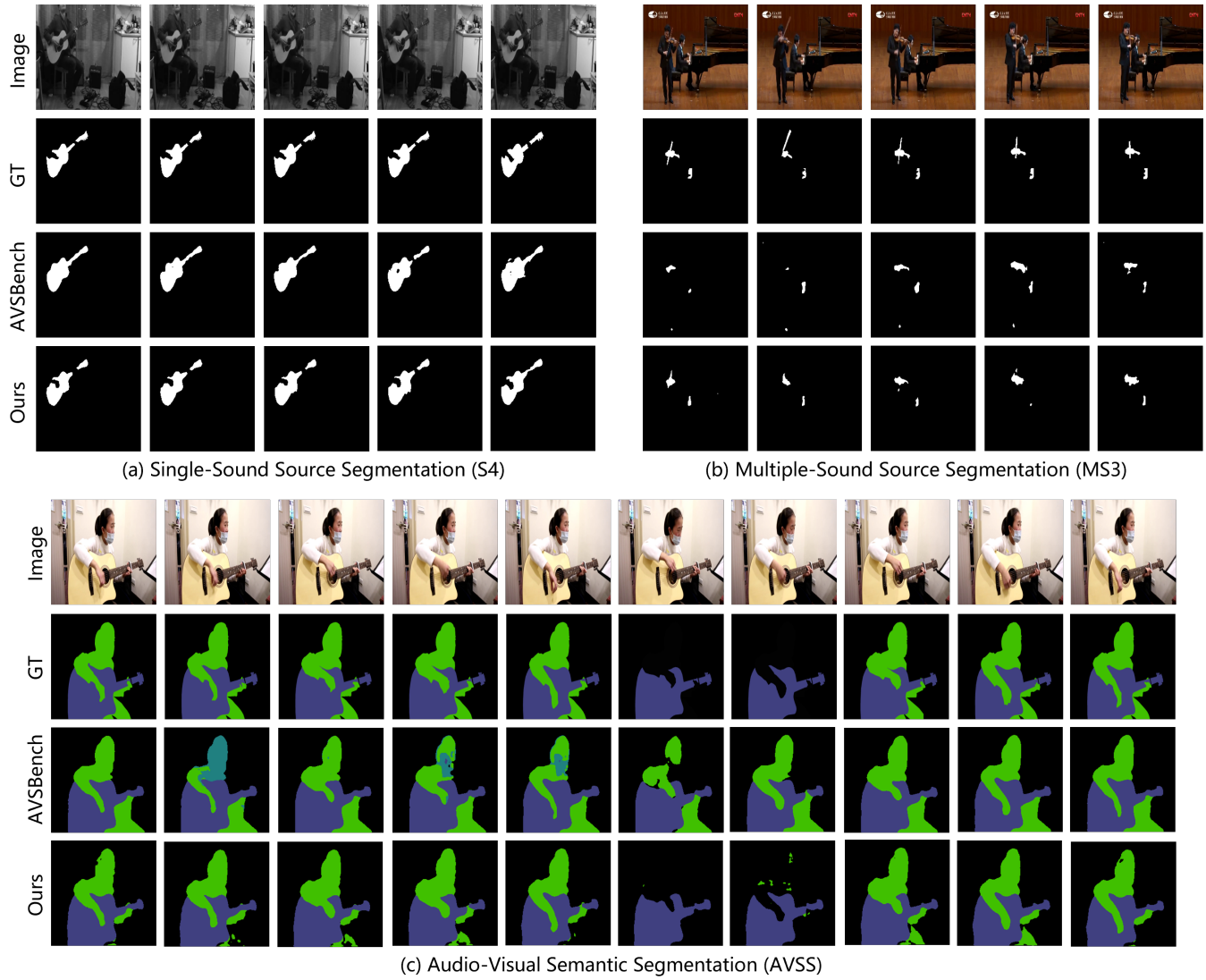

Figure 7: Qualitative results of AVSegFormer on three AVS sub-tasks. These results show that the proposed method can accurately segment the pixels of sounding objects and outline their shapes well.

the correct sounding object. These findings align with our hypothesis and further substantiate the effectiveness of the audio-visual mixer.

Effect of intermediate mask loss. We finally conducted experiments to learn the impact of the intermediate mask loss. Similarly, we designed two versions for the mask reformer module as shown in Figure 5. In the attention-only reformer (AR), we directly feed the attention map of each cross-attention layer into an FPN network to generate intermediate masks. In the query-mixed reformer (QR), we multiply the attention map with the output queries of the corresponding decoder layer, and then feed the generated features into the FPN. The experimental results presented in Table 2d demonstrate that both versions have brought some performance improvement, with the adopted QR outperforming AR.

Qualitative analysis. We also present the visualization results of AVSegFormer compared with those of AVSBench on three audio-visual segmentation tasks in Figure 7. The top line of each task displays the raw images, and the second line displays the ground truth. The last two lines display the predicted masks by AVSBench and AVSegFormer, respectively. The visualization results clearly demonstrate that our method performs better than the previous method. It has a strong ability in target localization and semantic understanding, which can accurately segment and classify the sounding objects in each task. Furthermore, in multiple sound sources scenes, the model can effectively identify the correct sound source and accurately segment the target object. These results highlight the effectiveness and robustness of our method.

5 CONCLUSION

In this paper, we propose AVSegFormer, a novel framework that leverages the power of transformers to achieve leading performance in audio-visual segmentation tasks. First, our method introduces learnable and audio queries, enabling our network to dynamically attend to relevant visual features and significantly enhance segmentation performance. Second, we design an audio-visual mixer that selectively amplifies or suppresses different visual channels, making the visual features better adapt to diverse audio inputs. Additionally, we propose an intermediate mask loss to improve the effectiveness of the training process. Our experimental results demonstrate the superior performance of AVSegFormer compared to existing state-of-the-art methods, and a series of ablation studies validate the effectiveness of our proposed components.

REFERENCES

[1] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. Neural module networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 39–48.

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*. 2425–2433.

[3] Relja Arandjelovic and Andrew Zisserman. 2017. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision*. 609–617.

[4] Relja Arandjelovic and Andrew Zisserman. 2018. Objects that sound. In *Proceedings of the European Conference on Computer Vision*. 435–451.

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *Proceedings of the 16th European Conference of Computer Vision*. 213–229.

[6] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. 2021. Localizing visual sounds the hard way. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16867–16876.

[7] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, et al. 2019. Hybrid task cascade for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4974–4983.

[8] Xianyu Chen, Ming Jiang, and Qi Zhao. 2021. Predicting human scanpaths in visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10876–10885.

[9] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. 2022. Vision transformer adapter for dense predictions. In *International Conference on Learning Representations*.

[10] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. 2022. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1290–1299.

[11] Bowen Cheng, Alex Schwing, and Alexander Kirillov. 2021. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems* 34 (2021), 17864–17875.

[12] Chaorui Deng, Qi Wu, Qingyao Wu, Fuyuan Hu, Fan Lyu, and Mingkui Tan. 2018. Visual grounding via accumulated attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7746–7755.

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 248–255.

[14] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. 2021. Transvg: End-to-end visual grounding with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1769–1779.

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

[16] Brendan Duke, Abdalla Ahmed, Christian Wolf, Parham Aarabi, and Graham W Taylor. 2021. Sstvos: Sparse spatiotemporal transformers for video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5912–5921.

[17] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing*. 776–780.

[18] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*. 2961–2969.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.

[20] 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. 2017. CNN architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing*. 131–135.

[21] Di Hu, Feiping Nie, and Xuelong Li. 2019. Deep multimodal clustering for unsupervised audiovisual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9248–9257.

[22] Di Hu, Rui Qian, Minyue Jiang, Xiao Tan, Shilei Wen, Errui Ding, Weiyao Lin, and Dejing Dou. 2020. Discriminative sounding objects localization via self-supervised audiovisual matching. *Advances in Neural Information Processing Systems* 33 (2020), 10077–10087.

[23] Jitesh Jain, Jiachen Li, MangTik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. 2022. OneFormer: One Transformer to Rule Universal Image Segmentation. *arXiv preprint arXiv:2211.06220* (2022).

[24] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. 2021. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1780–1790.

[25] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. 2019. Panoptic feature pyramid networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6399–6408.

[26] Feng Li, Hao Zhang, Shilong Liu, Lei Zhang, Lionel M Ni, Heung-Yeung Shum, et al. 2022. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. *arXiv preprint arXiv:2206.02777* (2022).

[27] Zhiqi Li, Wenhai Wang, Enze Xie, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, Ping Luo, and Tong Lu. 2022. Panoptic segformer: Delving deeper into panoptic segmentation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1280–1289.

[28] Yan-Bo Lin, Yu-Jhe Li, and Yu-Chiang Frank Wang. 2019. Dual-modality seq2seq network for audio-visual event localization. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing*. 2002–2006.

[29] Yan-Bo Lin, Hung-Yu Tseng, Hsin-Ying Lee, Yen-Yu Lin, and Ming-Hsuan Yang. 2021. Exploring cross-video and cross-modality signals for weakly-supervised audio-visual video parsing. *Advances in Neural Information Processing Systems* 34 (2021), 11449–11461.

[30] Yan-Bo Lin and Yu-Chiang Frank Wang. 2020. Audiovisual transformer with instance attention for audio-visual event localization. In *Proceedings of the Asian Conference on Computer Vision*.

[31] Shilong Liu, Yaoyuan Liang, Feng Li, Shijia Huang, Hao Zhang, Hang Su, Jun Zhu, and Lei Zhang. 2022. DQ-DETR: Dual Query Detection Transformer for Phrase Extraction and Grounding. *arXiv preprint arXiv:2211.15516* (2022).

[32] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10012–10022.

[33] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3431–3440.

[34] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

[35] Sabarinath Mahadevan, Ali Athar, Aljoša Ošep, Sebastian Hennen, Laura Leal-Taixé, and Bastian Leibe. 2020. Making a case for 3d convolutions for object segmentation in videos. *arXiv preprint arXiv:2008.11516* (2020).

[36] Yuxin Mao, Jing Zhang, Zhexiong Wan, Yuchao Dai, Aixuan Li, Yunqiu Lv, Xinyu Tian, Deng-Ping Fan, and Nick Barnes. 2021. Transformer transforms salient object detection and camouflaged object detection. *arXiv preprint arXiv:2104.10127* (2021).

[37] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 40th International Conference on 3D Vision (3DV)*. 565–571.

[38] Rui Qian, Di Hu, Heinrich Dinkel, Mengyue Wu, Ning Xu, and Weiyao Lin. 2020. Multiple sound sources localization from coarse to fine. In *Proceedings of the 16th European Conference on Computer Vision*. 292–308.

[39] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the 18th International Conference of Medical Image Computing and Computer-Assisted Intervention*. 234–241.

[40] Andrew Rouditchenko, Hang Zhao, Chuang Gan, Josh McDermott, and Antonio Torralba. 2019. Self-supervised audio-visual co-segmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing*. 2357–2361.
[41] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*. 618–626.
[42] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. 2018. Learning to localize sound source in visual scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4358–4366.
[43] Hongmei Song, Wenguan Wang, Sanyuan Zhao, Jianbing Shen, and Kin-Man Lam. 2018. Pyramid dilated deeper convlstm for video salient object detection. In *Proceedings of the European Conference on Computer Vision*. 715–731.
[44] Yapeng Tian, Dingzeyu Li, and Chenliang Xu. 2020. Unified multisensory perception: Weakly-supervised audio-visual video parsing. In *Proceedings of the European Conference on Computer Vision*. 436–454.
[45] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. 2018. Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision*. 247–263.
[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* 30 (2017).
[47] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*. 23318–23340.
[48] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. 2022. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media Journal* 8, 3 (2022), 415–424.
[49] Qi Wu, Peng Wang, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. 2016. Ask me anything: Free-form visual question answering based on knowledge from external sources. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4622–4630.
[50] Yu Wu and Yi Yang. 2021. Exploring heterogeneous clues for weakly-supervised audio-visual video parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1326–1335.
[51] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems* 34 (2021), 12077–12090.
[52] Yuwen Xiong, Renjie Liao, Hengshuang Zhao, Rui Hu, Min Bai, Ersin Yumer, and Raquel Urtasun. 2019. Upsnet: A unified panoptic segmentation network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8818–8826.
[53] Zongxin Yang, Yunchao Wei, and Yi Yang. 2021. Associating objects with transformers for video object segmentation. *Advances in Neural Information Processing Systems* 34 (2021), 2491–2502.
[54] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. 2022. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605* (2022).
[55] Jing Zhang, Jianwen Xie, Nick Barnes, and Ping Li. 2021. Learning Generative Vision Transformer with Energy-Based Latent Space for Saliency Prediction. In *2021 Conference on Neural Information Processing Systems*.
[56] Jinxing Zhou, Xuyang Shen, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, et al. 2023. Audio-Visual Segmentation with Semantics. *arXiv preprint arXiv:2301.13190* (2023).
[57] Jinxing Zhou, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, and Yiran Zhong. 2022. Audio–Visual Segmentation. In *Proceedings of the European Conference on Computer Vision*. 386–403.
[58] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. 2020. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In *International Conference on Learning Representations*.
[59] Xizhou Zhu, Jinguo Zhu, Hao Li, Xiaoshi Wu, Hongsheng Li, Xiaohua Wang, and Jifeng Dai. 2022. Uni-perceiver: Pre-training unified architecture for generic perception for zero-shot and few-shot tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16804–16815.

A APPENDIX

We listed the detailed settings of our models in Table 3.

Table 3: Detailed settings. This table provides a detailed overview of the specific settings used for each sub-task.

Settings	S4	MS3	AVSS
input resolution $H \times W$	224×224	224×224	224×224
frames T	5	5	10
embedding dimension D	256	256	256
transformer layers N_{layer}	6	6	6
number of queries N_{query}	300	300	300
mask loss coefficient λ	0.1	0.1	0.1
learnable queries	✓	✓	✓
audio-visual mixer	✓	✓	✓
batch size	4	4	4
optimizer	AdamW	AdamW	AdamW
learning rate	2×10^{-5}	2×10^{-5}	2×10^{-5}
drop path rate	0.1	0.1	0.1
epoch	30	60	30