Audio-Visual Instance Segmentation

Ruohao Guo[1], Xianghua Ying[1]*, Yaru Chen[2], Dantong Niu[3], Guangyao Li[4], Liao Qu[5], Yanyu Qi[6], Jinxing Zhou[7] Bowei Xing[1], Wenzhen Yue[1], Ji Shi[1], Qixun Wang[1], Peiliang Zhang[8], Buwen Liang[6]
[1]State Key Laboratory of General Artificial Intelligence, School of Intelligence Science and Technology, Peking University, [2]University of Surrey, [3]UC Berkeley, [4]Tsinghua University, [5]CMU, [6]China Agricultural University, [7]MBZUAI, [8]Wuhan University of Technology

Abstract

In this paper, we propose a new multi-modal task, termed audio-visual instance segmentation (AVIS), which aims to simultaneously identify, segment and track individual sounding object instances in audible videos. To facilitate this research, we introduce a high-quality benchmark named AVISeg, containing over 90K instance masks from 26 semantic categories in 926 long videos. Additionally, we propose a strong baseline model for this task. Our model first localizes sound source within each frame, and condenses object-specific contexts into concise tokens. Then it builds long-range audio-visual dependencies between these tokens using window-based attention, and tracks sounding objects among the entire video sequences. Extensive experiments reveal that our method performs best on AVISeg, surpassing the existing methods from related tasks. We further conduct the evaluation on several multi-modal large models. Unfortunately, they exhibits subpar performance on instance-level sound source localization and temporal perception. We expect that AVIS will inspire the community towards a more comprehensive multi-modal understanding. Dataset and code is available at https://github.com/ruohaoguo/avis.

1. Introduction

Vision and hearing are our primary channels of communication and sensation [9, 23, 63, 64, 66–68]. Audio-visual collaboration is beneficial for humans to better perceive and interpret the world. Humans have the ability to associate mixed sounds with object instances in complicated realistic scenarios. Imagine a cocktail-party scenario: when a group of people is speaking, we can not only locate the sound sources but also determine how many people are talking.

Inspired by this human perception, we explore instance-level sound source localization in long videos and propose a new task, namely audio-visual instance segmentation (AVIS). As can be seen in Figure 1 (c), it requests a model to simultaneously classify, segment and track sounding object instances—identify *which object categories are making sounds*, infer *where the sounding objects are*, and monitor *when they are making sounds*. This new task facilitates a wide range of practical applications, including embodied robotics, video surveillance, video editing, etc. Moreover, it can serve as a fundamental task for evaluating the comprehension capabilities of multi-modal large models.

Audio-visual instance segmentation is related to several existing tasks. For example, audio-visual object segmentation (AVOS) [65] is to separate sounding objects from the background region of a given audible video, as shown in Figure 1 (a). Unlike AVOS being tasked with binary foreground segmentation, audio-visual semantic segmentation (AVSS) [69] aims at predicting semantic maps that assign each pixel with a specific category, as shown in Figure 1 (b). To accomplish the above tasks, many works [17, 22, 35, 56] extend the image segmentation frameworks [5, 12] to the video domain, and design various audio-visual fusion modules for sound source localization. Despite promising performance in the AVSBench dataset [69], current methods still suffer from two limitations in real-world scenarios. First, these methods fail to differentiate two sounding objects with the same category, such as the woman, man, left ukulele and right ukulele depicted in Figure 1. Second, these methods focus on 5- or 10-second trimmed short videos and ignore long-range modeling abilities, which may lead to weak performance in real world.

One potential reason that the AVIS task is rarely studied is the absence of a high-quality dataset. Despite the existence of audio-visual segmentation datasets [65, 69], none are directly applicable to our proposed task, due to lacking instance-level annotations and long-form videos. Therefore, in this work, we built the first audio-visual instance segmentation dataset, namely AVISeg. The new dataset consists of 926 videos with an average duration of 61.4 seconds and 94,074 high-quality masks, covering 26 common

*Corresponding Author

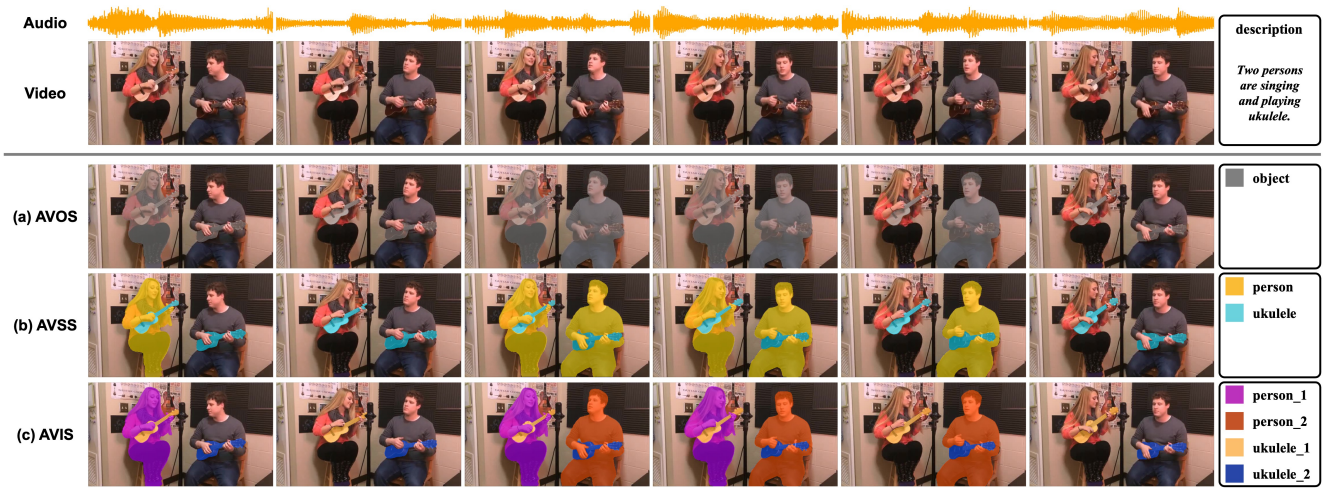

Figure 1. Comparison of different audio-visual segmentation tasks. (a) Audio-Visual Object Segmentation (AVOS) only requires binary segmentation. (b) Audio-Visual Semantic Segmentation (AVSS) associates one category with every pixel. (c) Audio-Visual Instance Segmentation (AVIS) treats each sounding object of the same class as an individual instance.

categories from 4 real-world scenarios (Music, Speaking, Machine and Animal). Our dataset can also be served as a benchmark for AVOS and AVSS tasks. Additionally, we present a novel evaluation metric, termed frame-level sound localization accuracy (FSLA), which measures the proportion of frames that are correctly predicted by the model out of the total number of frames.

In order to deal with the above AVIS task, we follow the query-based segmentation paradigm [12, 26] and propose a baseline model called AVISM. To be specific, a frame-level sound source localizer segments sounding objects within each frame independently and summarizes per-frame scenes into a small amount of object tokens. Then, a video-level sounding object tracker is designed to build frame-to-frame communications and track sounding objects throughout the entire video. To lessen computational overheads in processing long and high-resolution videos, the tracker uses the concise object tokens as a mean of conveying information rather than dense image features, and adopts window-based self-attention mechanisms to efficiently capture long-range dependencies in consecutive frames. Experimental results demonstrate the superiority of our baseline. Additionally, we make a thorough evaluation of several prominent multi-modal large models on our AVISeg dataset. Surprisingly, these self-proclaimed large models are far from satisfactory in instance-aware sound source localization and temporal perception. Our dataset emphasizes the necessity for further improvements in handling audio-visual data and long videos, providing insights for future development of multi-modal large models. Our contributions are as follows:

(1) To our best knowledge, this is the first work exploring audio-visual instance segmentation, which aims to classify, segment and track sounding objects in given audible videos.

(2) We create a high-quality video dataset to support the above task, containing 926 videos with an average length of 61.4s. Besides, we propose a novel frame-level metric for evaluating audio-visual instance segmentation.

(3) A strong baseline model is developed to localize sound source in each frame and track sounding objects in the entire video. To handle long videos, it distils image features into a small number of tokens and uses window-based attention to convey audio-visual temporal information.

(4) Extensive experiments indicate that our framework achieves state-of-the-art results under all evaluation metrics. Moreover, our dataset can also serve as a potential benchmark for evaluating various multi-modal large models.

2. Related Work

2.1. Video Instance Segmentation

Video instance segmentation (VIS) aims at simultaneous segmentation and tracking of all object instances in videos. Early methods [2, 4, 39, 55, 57] often extend CNN-based image segmentation methods [20, 25] to establish temporal consistency. For example, MaskTrack R-CNN [55] introduces an additional tracking head to Mask R-CNN [25] for object matching and association between frames. SG-Net [39] follows the anchor-free FCOS detector [49] and directly leverages the object centerness from detection to delineate the temporal coherence in video sequences. The above approaches require extra post-processing steps, such as non-maximal suppression (NMS), leading to higher computational costs and potential misdetections. Recent methods [11, 26, 28, 31, 51, 52, 58, 60] adapt Transformer-based image segmentation methods [5, 12] to the VIS task. For example, VisTR [51] builds on the query-based DETR [5]

and naturally outputs the sequence of masks for each instance without heuristic matching or hand-designed post-processing. Follow-up works, such as Mask2Former-VIS [11] and SeqFormer [52], design more querying strategies to improve the performance of segmentation and tracking. To avoid heavy computation and memory usage, IFC [28] and VITA [26] first distill dense spatio-temporal features into a small amount of tokens, and then perform inter-frame communication between tokens. This information-passing paradigm allows models for efficiently handling long and high-resolution videos with a common GPU.

2.2. Audio-Visual Segmentation

Audio-visual segmentation (AVS) focuses on localizing and segmenting sounding objects within each video frame. Zhou et al. [65, 69] introduce the first AVS dataset, namely AVSBench, which serves two different sub-tasks including audio-visual object segmentation (AVOS) and audio-visual semantic segmentation (AVSS). The former [65] requires producing binary masks of sounding objects, while the latter [69] further needs to generate semantic maps representing the object category. To address these problems, they employ a standard encoder-decoder architecture with a modified non-local block to encode space-time relation and segment sounding objects. CAVP [10] builds an AVS dataset by randomly matching the images from COCO [38] and audio files from VGGSound [6] based on the semantic classes of the objects. Inspired by DETR [5] and Mask2Former [12], recent works [17, 21, 22, 35, 56] adopt the query-based architecture decode masks for sounding objects. For example, AVSegFormer [17] trivially incorporates audio features and learnable queries, enabling the decoder to capture relevant visual semantics and predict the audio-constrained masks. COMBO [56] explores multi-order bilateral relations in modality, temporal and pixel levels for the AVSS task. Notably, a bilateral-fusion module is designed to align audio and visual modalities bi-directionally and assist the model in segmenting the sounding objects.

3. New Task

3.1. Problem Definition

Audio-visual instance segmentation (AVIS) is a challenging multi-modal task that involves localizing and segmenting sounding objects in a video, while assigning each a unique identity label to ensure consistent tracking throughout the video. In this task, we predefine a category label set as $\mathcal{C} = \{1, ..., K\}$, where K is the number of categories. Given a video sequence with T frames and its corresponding audio, suppose there are N sounding objects belonging to the category label set $\mathcal{C}$ in the video. For each sounding object o^i, let $c^i \in \mathcal{C}$ denote its category label, and let m_t^i denote its binary segmentation mask in t^{th} frame where $t \in T'$ and T' denotes the sounding time set, i.e., $T' \subseteq T$. We assume that an AVIS model outputs H instance hypotheses. For each hypothesis o^j, it needs to contain a predicted category label $\tilde{c}^j \in \mathcal{C}$, a confidence score $\tilde{s}^j \in [0, 1]$, and a sequence of predicted binary masks $\tilde{m}_t^j$. The goal of AVIS task is to minimize the difference between the ground truth and the hypotheses. This requires the AVIS model to correctly determine which instances are making sounds, accurately identify and segment these sounding instances, and reliably track them in the entire video.

3.2. Evaluation Metrics

To evaluate how well an AVIS model performs, we need to choose appropriate metrics to compare its outputs with the ground truth. In our task, we adopt two evaluation protocols including the mean Average Precision (mAP) [55] and the Higher Order Tracking Accuracy (HOTA) [43]. mAP follows the computation of the average precision-recall metric over trajectories, which is commonly used in video instance segmentation. However, mAP is not perfectly suited to our task, because it can be increased by producing many different predictions with low confidence scores and does not decrease even if non-sounding objects are predicted. HOTA performs a bijective matching at the detection level while scoring association over trajectories, which is designed for multi-object tracking task. This makes HOTA a balanced metric for measuring both detection and association. When applied to the AVIS task, it can penalize those models that predict non-sounding objects.

Besides considering the above object-based metrics, we propose a novel measure, namely frame-level sound localization accuracy (FSLA), tailored to measure the proportion of frames that are correctly predicted by the model out of the total number of frames. Specifically, we first use the Hungarian algorithm [32] to determine a one-to-one matching between ground-truth and predicted detections. For each frame, it can be treated as correct frame if it satisfies the following conditions: 1) The number of sounding objects is correct; 2) The category of the sounding objects is correct; 3) The IoU (Intersection over Union) between the ground truth and the predicted sounding objects is greater than threshold α. The final score is computed by averaging over all classes before averaging different α thresholds (0.05 to 0.95 in 0.05 intervals). The pseudo code of the FSLA metric is in the *Supp. Materials*. Compared to other metrics, our FSLA allows for easier localization of incorrect frames and offers a more intuitive explanation of the model's performance across different time periods. Additionally, it can be decomposed into a set of sub-metrics (FSLAn, FSLAs and FSLAm) which can be used for model evaluation in scenarios with no sound source, a single sound source, and multiple sound sources. This results in FSLA being able to guide how models can be improved, or understand where

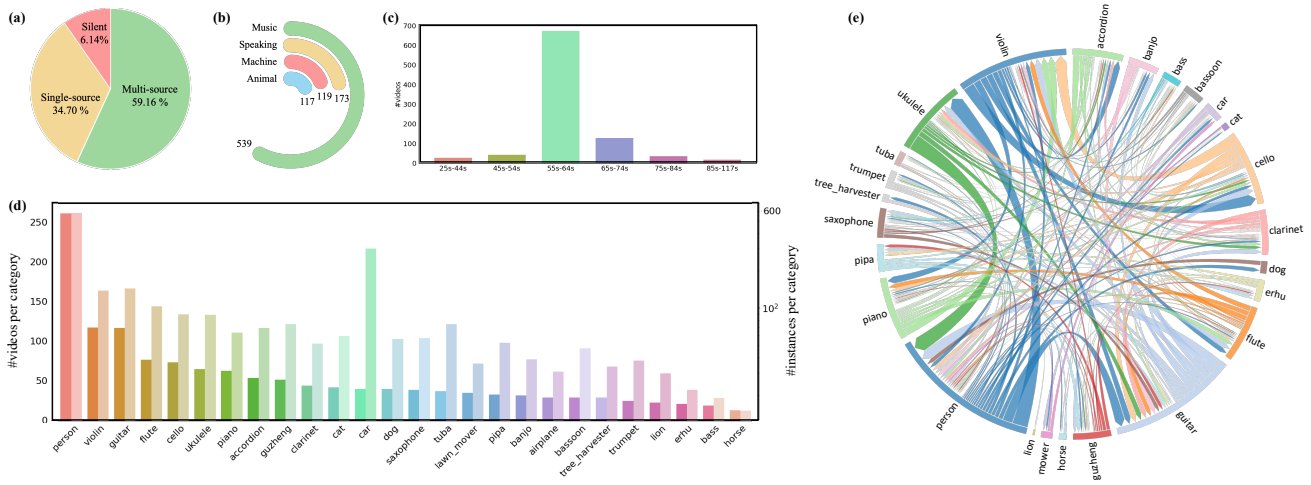

Figure 2. Illustrations of our AVISeg dataset statistics. (a) Ratio of different sound sources. (b) Number of video in 4 real-world scenarios. (c) Distribution of video lengths. (d) Number of video and objects for the 26 categories. (e) Relations between different categories.

they are likely to fail when used.

4. Dataset

To explore audio-visual instance segmentation and evaluate the proposed methods, we create a new large-scale benchmark called AVISeg. Considering that this task involves complex audio-visual interactions and requires high-quality data, we manually collect and choose 926 videos from YouTube and the publicly available datasets [33, 36, 37], e.g., MUSIC-AVQA. Our released AVISeg dataset satisfies the following criteria: 1) It focuses on long-term videos (61.4s), bringing them much closer to real applications. 2) It contains 26 common sound categories, spanning 4 dynamic scenarios: "Music", "Speaking", "Machine", and "Animal". 3) It involves some challenging cases, such as videos with silent sound sources, single sound source, and multiple sources simultaneously. These attributes impose higher demands on the model for accurate recognition, segmentation, and tracking of sounding objects.

Similar to AVSBench [65], each video is divided into 1-second clips. We then adopt an interactive semi-automatic annotation tool [1] based on ViT-H SAM model [30] to label sounding object instances belonging to the defined category set exhaustively in these videos. For example, in the first column of Figure 1, the woman is labeled as "person_1" because she is singing, while the man is not labeled since no sound is made. That is, an object will only be masked and assigned a unique identifier when it emits sound. Note that each labeled frame undergoes multiple rounds of manual review and refinement to ensure high-quality annotations.

In terms of high-level statistics, our AVISeg dataset consists of 94,074 masks on 56,871 frames, distributed in 926 videos for about 16 hours. Figure 2 (a-e) provides the statistical analysis of our dataset. In this dataset, silent frames, single-source frames and multi-source frames account for 6.14%, 34.70% and 59.16%, respectively. AVISeg covers 4 real-world scenarios, with the "Music" scenario having the largest number of videos, totaling 539. Note that a video may belong to multiple scenarios, such as the simultaneous appearance of animals and musical instruments. A comparison of the proposed AVISeg and related datasets is shown in Table 1. For training and evaluation, we randomly split the dataset into training, validation, and testing sets with 616, 105, and 205 videos, respectively.

Table 1. Comparison with other datasets from related tasks. SSL represents audio-visual event localization.

Task	Dataset	Videos	Length	Classes	Anno.
SSL	Flickr-S [48]	5,000	20.0s	50	bbox
	VGG-SS [7]	5,158	10.0s	220	bbox
AVOS	AVSBench-O [65]	5,356	5.0s	23	pixel
AVSS	AVSBench-S [69]	12,356	7.8s	70	pixel
VIS	YTVIS [55]	2,883	4.6s	40	pixel
	OVIS [45]	901	12.8s	25	pixel
AVIS	AVISeg	926	61.4s	26	pixel

5. Baseline Model

We introduce a new baseline model, termed AVISM, for the audio-visual instance segmentation task. The proposed AVISM model, built upon Mask2Former [11, 12] and VITA [26], adopts a query-based Transformer architecture to learn a set of query vectors representing sounding objects for the instance segmentation and tracking. To better model audio-visual semantic correlations in long and complicated videos, we present the frame-level audio-visual fusion mod-

[1] https://www.yatenglg.cn/isat/

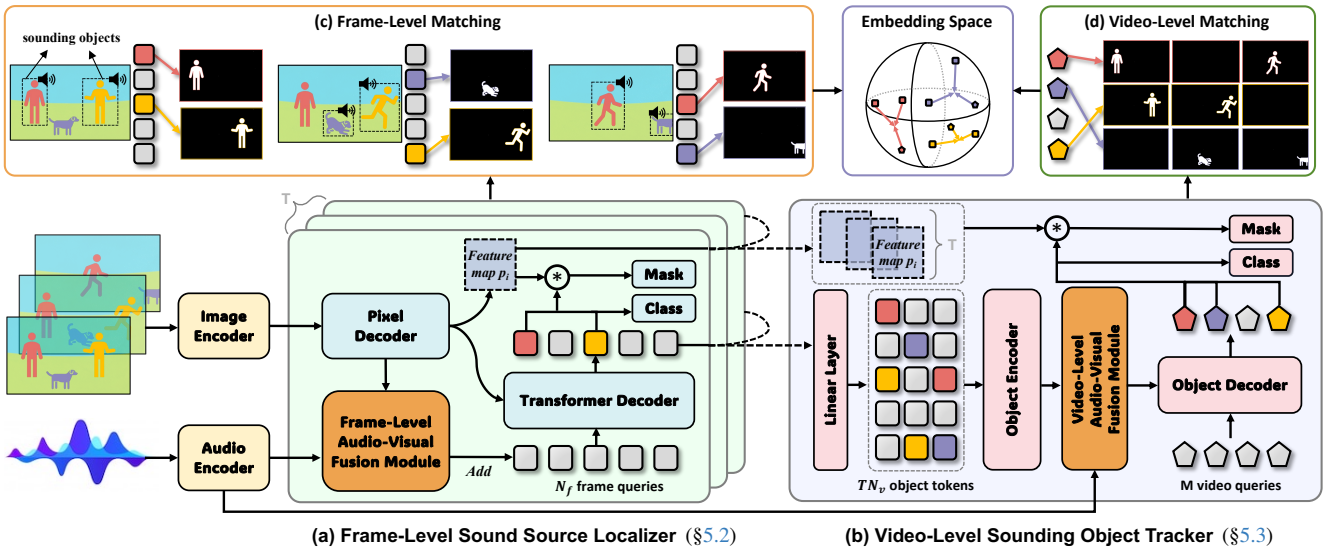

Figure 3. Overview of the proposed AVISM for audio-visual instance segmentation. (a) The frame-level sound source localizer segments sounding objects within each frame independently and condenses dense image features into frame queries. (b) The video-level sounding object tracker takes frame queries and audio features as input, and then performs temporal audio-visual communications between frames.

ule and video-level audio-visual fusion module to integrate audio and visual features. The overall framework of our baseline model is illustrated in Figure 3.

5.1. Audio-Visual Representation

Given an input video sequence that contains both visual and audio tracks, we split it into T non-overlapping visual and audio snippet pairs $\{V, A\} = \{v_i, a_i\}_{i=1}^{T}$, where each snippet spans 1 second and T represents the number of snippets as well as the video length. For each visual snippet v_i, we apply ResNet [24] or Swin Transformer [42] as the backbone to extract hierarchical features $f_{i,k}^{V} \in \mathbb{R}^{H_k \times W_k \times D_k}$. $H_k \times W_k$ denotes the output resolution of each v_i at the k$^{\text{th}}$ backbone level. The final visual representation can be formulated as $F^V = \{f_i^V\}_{i=1}^{T}$. For each audio snippet a_i, we first convert it to a mel spectrogram via the short-time Fourier transform and then encode it into an audio feature vector $f_i^A \in \mathbb{R}^D$ using a pre-trained VGGish model [18], where D is the feature dimension. The final audio representation $F^A = \{f_i^A\}_{i=1}^{T}$ is extracted offline and the VGGish model is not fine-tuned during the training process.

5.2. Frame-Level Sound Source Localizer

To accurately localize the sounding objects within each video frame, we propose the frame-level sound source localizer that establishes the spatial association between audio and visual modalities. As depicted in Figure 3 (a), we employ a multi-scale deformable attention Transformer [70], namely pixel decoder, to produce enhanced visual features $\hat{f}_i^V$ and high-resolution per-pixel embeddings p_i. Then, the frame-level audio-visual fusion module performs cross-attention computation between $\hat{f}_i^V$ and the corresponding audio feature f_i^A at multiple scales, yielding audio-to-image features $f_i^{AV} \in \mathbb{R}^C$. Inspired by the set prediction paradigm [5], we introduce N_f audio-conditioned learnable queries, which are added with f_i^{AV} to form *frame queries* $Q_f \in \mathbb{R}^{N_f \times C}$. After a Transformer decoder distills and embeds visual semantics of all frames into the frame queries, each frame query is dot-multiplied with p_i, and used for classifying and segmenting its matched sounding object.

5.3. Video-Level Sounding Object Tracker

One limitation of the above localizer is that it operates independently on each frame, with no inter-computation shared across frames. For the solution to this problem, we present the video-level sounding object tracker that builds temporal communications throughout the entire video sequence. Considering the heavy computation demands posed by processing long and high-resolution videos, our tracker takes the frame queries as inputs rather than image features, and leverages the window-based self-attention mechanisms [42] to capture long-range dependencies among frames.

As shown in Figure 3 (b), a linear layer converts $T \times N_f$ frame queries gathered from all frames into object tokens Q_o. The object encoder, similar to [26], partitions these object tokens along the temporal axis into non-overlapping local windows of size W, within which self-attention is performed. After alternately shifting the windows, object tokens $\hat{Q}_o$ from different windows can exchange object-wise information. We extend this capability of processing long videos to multi-model temporal learning, and design a video-level audio-visual fusion module (Figure 4) incor-

porating N attention layers. In each local window, it calculates cross-attention between object tokens $\hat{Q}_o$ and audio features f_i^A. As the local window shifts and the attention layer goes deeper, our model can efficiently achieve frame-to-frame audio-visual communications in long videos. Its outputs are added with $\hat{Q}_o$ and their results are referred as Q_o^{AV}. This temporal fusion benefits the global alignment of audio and object instances, while also enhancing object tracking and identity association across different frames.

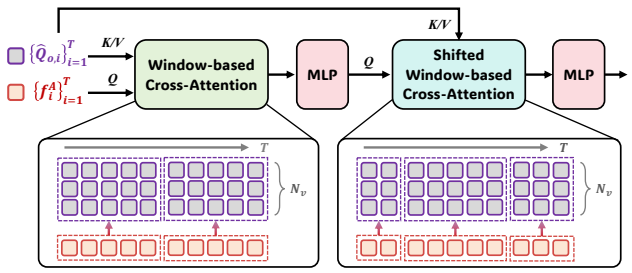

Figure 4. The architecture of our proposed video-level audio-visual fusion module. For the entire video sequence, it computes cross-attention between object tokens $\{\hat{Q}_{o,i}\}_{i=1}^T$ and audio features $\{f_i^A\}_{i=1}^T$ within local windows, and introduces cross-window connections by shifting windows.

To decode object-centric information from all object tokens, we initialize a fixed set of learnable *video queries* $Q_v \in \mathbb{R}^{N_v \times C}$, where N_v is the number of video queries. The object decoder, implemented as a standard Transformer decoder [5, 26], receives Q_o^{AV} and aggregates their semantics into video queries. At the end of the decoder, two output heads are exploited to obtain the final predictions, with each head comprising two fully-connected layers. Specifically, a class head predicts class probabilities $p \in \mathbb{R}^{K+1}$ for each video query, including a *no sounding object* $\varnothing$ class in addition to the K given classes of a dataset. Besides, object queries are input into a mask head and then dot-multiplied with p_i, resulting in the final mask logits.

5.4. Training Loss

There are three terms in the training loss as follows:

$$\mathcal{L} = \lambda_{\text{frame}}\mathcal{L}_{\text{frame}} + \lambda_{\text{video}}\mathcal{L}_{\text{video}} + \lambda_{\text{sim}}\mathcal{L}_{\text{sim}} \quad (1)$$

where λ_{frame}, λ_{video} and λ_{sim} are hyper-parameters to balance the loss terms. Their default values are set to 1, 1, 0.5, respectively. For frame-wise supervision, we first compute costs between frame queries and ground truth at each t^{th} frame using the cost function of Mask2Former [12]. Following DETR [5], the Hungarian algorithm [32] is then employed for optimal matching, as shown in Figure 3 (c). Finally, we utilize $\mathcal{L}_{\text{frame}}$ from [12] to calculate loss between the matched pairs. For video-wise supervision, we also search for optimal assignment between video queries and ground-truth sequences using the cost function of IFC [28], as shown in Figure 3 (d). These bipartitely matched pairs are used to compute the loss function $\mathcal{L}_{\text{video}}$ from [28], a simple extension of [12]. Additionally, as depicted in Figure 3 (e), we introduce the similarity loss [26, 55] to align frame queries with video queries in the embedding space, annotating pairs of equal identities as 1 and others as 0.

6. Experiment

6.1. Main Results

We compare AVISM with the state-of-the-art methods from two related tasks, including video instance segmentation (VIS) and audio-visual semantic segmentation (AVSS). For the VIS methods [11, 16, 26, 31, 52, 58, 60], only video frames are used for training, while the audio is disregarded. For the AVSS methods [17, 56], they follow the query-based detection paradigm [5] and achieve instance-level segmentation without altering the model, losses and training procedure. To make the evaluation fair, all methods utilize ResNet-50 pre-trained on ImageNet [15] as the backbone and are trained on the AVISeg dataset for 48,000 iterations.

Table 2 presents the comparison results, including three main metrics (FSLA, HOTA, mAP) and five sub-metrics (FSLAn, FSLAs, FSLAm from FSLA; AssA, DetA from HOTA). It is worth noting that our AVISM achieves the best results under all evaluation metrics. Compared to the VIS methods, AVISM incorporates audio information and leverages multi-modal contexts to localize sounding objects within video frames, which outperforms the best VITA [26]. This multi-sensory perception helps to guide our model to determine whether or which objects are making sounds. Compared to the AVSS methods, AVISM condenses per-frame scenes into a small number of frame queries and then establishes inter-frame audio-visual communication between them. Our experimental results demonstrate that using the concise frame queries, instead of dense spatio-temporal features, not only improves AVIS performance but also provides robust practicality for processing long and high-resolution videos. Furthermore, the results confirm the viability of AVISeg as a benchmark for AVIS task.

Figure 5 visualizes some sample videos with our predictions. Our AVISM model accurately localize the sounding object across both spatial and temporal dimensions, e.g., "lion" in video (d). In complex scenes with multiple sound sources, our model enables to handle the numerous mixed semantics, e.g., "person" and "ukulele" in video (a). When an object begins producing sound in the intermediate frames, AVISM is able to segment it and assign a new identity, as evidenced in video (b). This case also shows the effectiveness of our model in identifying and distinguishing objects with similar appearances or sounds. Moreover, if a sounding object disappears and reoccurs, the AVISM still

Table 2. Quantitative evaluation of different models from related tasks on the AVISeg test set. The best results are highlighted in bold.

Task	Model	Venue	Audio	FSLA	HOTA	mAP	FSLAn	FSLAs	FSLAm	AssA	DetA
VIS	Mask2Former-VIS [11]	CVPR' 22	✘	29.75	52.03	28.66	0.00	25.47	36.37	64.49	43.33
	TeViT [58]	CVPR' 22	✘	32.28	53.67	31.52	0.00	28.07	39.18	65.27	45.10
	SeqFormer [52]	ECCV' 22	✘	30.32	54.32	32.79	25.03	21.76	36.46	67.25	45.23
	VITA [26]	NeurIPS' 22	✘	38.04	57.48	36.25	15.04	27.98	47.45	69.86	48.96
	DAVIS [60]	ICCV' 23	✘	23.99	49.12	19.83	14.61	24.83	24.69	63.51	40.11
	LBVQ [16]	TCSVT' 24	✘	34.73	56.97	36.58	27.71	29.52	38.96	68.34	48.83
AVSS	AVSegFormer [17]	AAAI' 24	✔	35.66	55.74	35.72	18.58	27.51	43.08	67.13	48.51
	COMBO [56]	CVPR' 24	✔	39.49	57.39	37.84	21.91	27.18	49.63	68.87	50.12
AVIS	AVISM	CVPR' 25	✔	**42.78**	**61.73**	**40.57**	**32.22**	**29.83**	**52.40**	**71.15**	**54.97**

correctly tracks it, e.g., "tree harvester" in video (c).

Table 3. Zero-shot results of different multi-modal large models for audio-referred visual grounding on the AVISeg test set.

Model	Assistant	FSLA	HOTA	mAP
Sam4AVS [59]	-	0.00	8.18	3.93
BuboGPT-7B [62]	GPT-4	7.75	20.16	5.76
PG-Video-LLaVA-7B [44]	GPT-3.5	9.15	22.86	5.94
AL-Ref-SAM 2 [27]	GPT-4	**18.55**	**38.02**	**15.84**

6.2. Evaluations on Multi-modal Large Models

Table 3 presents the zero-shot results between different multi-modal large models (MMLMs) on AVIS task, revealing that these methods are underperforming. For instance, BuboGPT [62] and PG-Video-LLaVA [44] localize sound sources with audio-image-text aligned large language models (Vicuna [14] and LLaVA [40]), and then classifies and segments sounding objects using an off-the-shelf grounding pipeline based on GPT [1] and SAM [30]. However, BuboGPT is limited to processing a single image and one-second audio, and PG-Video-LLaVA cannot determine the exact time intervals for each sounding object. AL-Ref-SAM 2 [27] adopts Chain-of-Thought prompts to unleash GPT's temporal-spatial perception and reasoning capabilities. Although pre-trained on large-scale datasets and yielding promising results on audio-visual understanding task, these MMLMs fall short in instance segmentation and long-range modeling, resulting in poor performance on AVISeg. Our new task can provide deeper insights for multi-modal instruct tuning of MMLMs, has the potential to serve as a benchmark for evaluating their performance. More analysis can be found in *Supp. Materials*.

6.3. Ablation Studies

Impact of audio-visual fusion modules. To evaluate our proposed frame-level audio-visual fusion module (FL-AVFM) and video-level audio-visual fusion module (VL-AVFM), we first establish a baseline by disabling both modules. As evidenced in Table 4, the introduction of FL-AVFM yields substantial improvements across all metrics. These gains underscore the importance of effective audio-visual information aggregation at the frame level for enhancing per-frame object localization accuracy. Further incorporation of the VL-AVFM leads to more pronounced enhancements across all metrics, with the full configuration achieving optimal results. This observation suggests that the VL-AVFM plays a crucial role in leveraging temporal information across frames, thereby facilitating improved tracking consistency and accuracy. Our findings support the hypothesis that temporal audio-visual fusion is instrumental in resolving ambiguities during object tracking, particularly in challenging scenarios where motion cues may be insufficient for determining whether an object is producing sound. This demonstrates the potential of audio as auxiliary information to guide audio-visual instance segmentation.

Table 4. Impact of frame-level audio-visual fusion module (FL-AVFM) and video-level audio-visual fusion module (VL-AVFM).

FL-AVFM	VL-AVFM	FSLA	HOTA	mAP
		38.04	57.48	36.25
✔		39.68	59.59	39.06
✔	✔	**42.78**	**61.73**	**40.57**

Impact of local window size within video-level sounding object tracker. Table 5 presents an ablation study on local window sizes in our video-level sounding object tracker. We observe a clear trade-off between the maximum number of processable frames and tracking performance. A window size of 3 allows processing of the longest sequences (5304 frames) but yields the lowest performance across all metrics. Conversely, a window size of 12 significantly improves tracking accuracy at the cost of reduced frame capacity (1416 frames). The performance gain can be attributed to the expanded temporal receptive field, which allows the model to capture more complex inter-frame dependencies. This enhanced temporal context enables the tracker to better understand the long-term dynamics of sounding objects, leading to more accurate localization and tracking. Considering the trade-off between segmentation performance and

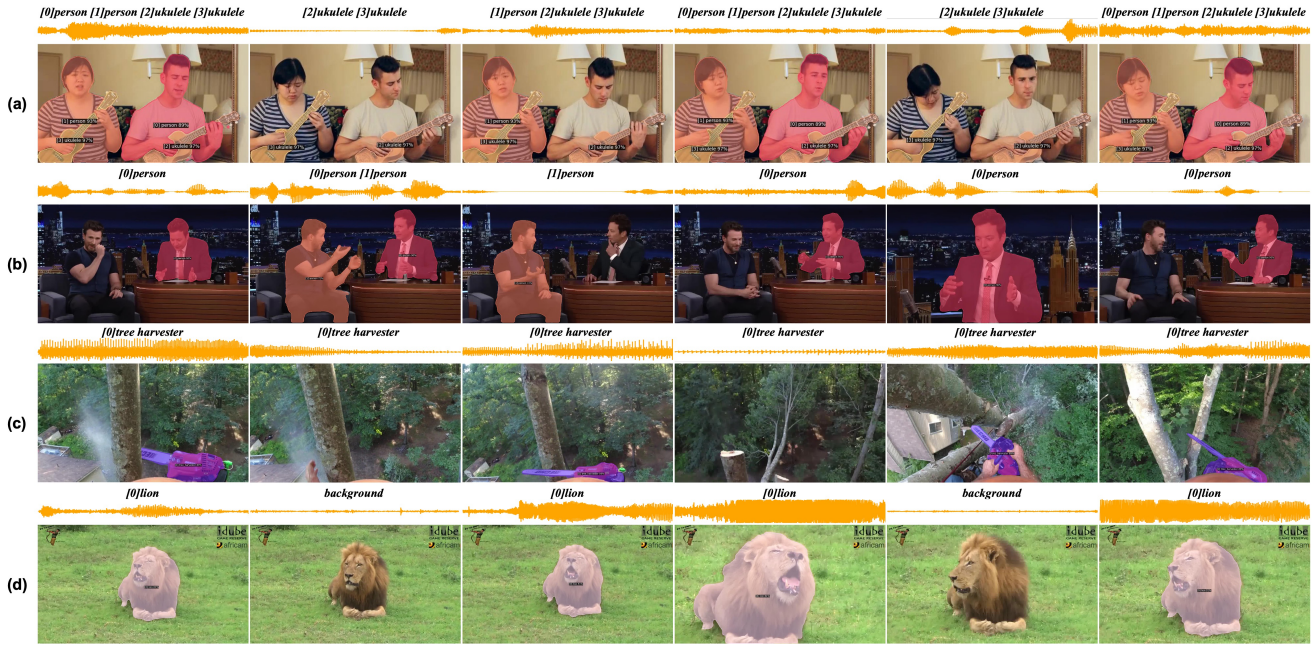

Figure 5. Sample results of our baseline model on AVISeg dataset from four scenarios: (a) Music; (b) Speaking; (c) Machine; (d) Animal. Each row have six sampled frames from a video sequence. Zoom in to see details.

the ability to process longer sequences, we chose a window size of 6 as the default, which provides a balanced compromise between accuracy and frame capacity.

Table 5. Impact of local windows size within video-level sounding object tracker. The maximum number of frames is reported on a single NVIDIA Quadro 6000 GPU.

Window Size	Max Frames	FSLA	HOTA	mAP
3	**5304**	40.83	61.13	40.14
6	2778	42.78	61.73	40.57
12	1416	**42.96**	**62.82**	**41.31**

Impact of visual backbone and pre-training dataset. We further investigate whether providing a stronger backbone and more pre-training data can further improve the model's AVIS performance. As shown in Table 6, adopting the strategy from Mask2Former [12] that using COCO for additional pre-training of our visual backbone resulted in improvements across all metrics. However, when further fine-tuned on the video instance segmentation dataset OVIS [45], despite an increase in mAP, we observe a slight decrease in FSLA. This is likely because OVIS primarily targets improving the model's video segmentation capabilities, leading to the segmentation of many non-sounding objects, thus not achieving better FSLA scores. Consequently, we opt for the IN+COCO pre-trained visual backbone for subsequent experiments. Replacing the backbone with Swin-L achieves the highest scores across all metrics.

Table 6. Impact of visual backbone and pre-training dataset.

Backbone	Pre-trained Datasets	Param.	FSLA	HOTA	mAP
R-50	IN	527.3	42.78	61.73	40.57
	IN+COCO		44.42	64.52	45.04
	IN+COCO+OVIS		43.68	64.64	45.76
R-101	IN+COCO	599.5	45.06	64.80	46.61
Swin-L	IN+COCO	1181.8	**52.49**	**71.13**	**53.46**

7. Conclusion

This paper introduces a new task of audio-visual instance segmentation with the goal of identifying, segmenting and tracking individual sounding object instances in videos. We present a high-quality dataset and a strong baseline model, providing some early explorations towards this task. In addition, we evaluate the zero-shot performance of several multi-modal large models, but they are far from satisfactory in instance-level sound source localization and long-range temporal perception. These findings underscore the need for further advancements in fine-grained and time-sensitive instruction tuning. We believe our task will innovate the community on new research ideas and directions for multi-modal understanding, and our dataset has the potential to serve as a platform for testing large models.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (NSFC) under Grant No. 62371009, and Beijing Natural Science Foundation under Grant No. L247029.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 7, 3, 5

[2] Ali Athar, Sabarinath Mahadevan, Aljosa Osep, Laura Leal-Taixé, and Bastian Leibe. Stem-seg: Spatio-temporal embeddings for instance segmentation in videos. In *European Conference on Computer Vision*, pages 158–177. Springer, 2020. 2

[3] Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. Whisperx: Time-accurate speech transcription of long-form audio. In *Interspeech*, pages 4489–4493, 2023. 4

[4] Jiale Cao, Rao Muhammad Anwer, Hisham Cholakkal, Fahad Shahbaz Khan, Yanwei Pang, and Ling Shao. Sipmask: Spatial information preservation for fast image and video instance segmentation. In *European Conference on Computer Vision*, pages 1–18. Springer, 2020. 2

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 1, 2, 3, 5, 6

[6] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 721–725. IEEE, 2020. 3

[7] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Localizing visual sounds the hard way. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16867–16876, 2021. 4

[8] Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei. Beats: Audio pre-training with acoustic tokenizers. In *International Conference on Machine Learning*, pages 5178–5193. PMLR, 2023. 5

[9] Yaru Chen, Ruohao Guo, Xubo Liu, Peipei Wu, Guangyao Li, Zhenbo Li, and Wenwu Wang. Cm-pie: Cross-modal perception for interactive-enhanced audio-visual video parsing. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8421–8425. IEEE, 2024. 1

[10] Yuanhong Chen, Yuyuan Liu, Hu Wang, Fengbei Liu, Chong Wang, Helen Frazer, and Gustavo Carneiro. Unraveling instance associations: A closer look for audio-visual segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26497–26507, 2024. 3

[11] Bowen Cheng, Anwesa Choudhuri, Ishan Misra, Alexander Kirillov, Rohit Girdhar, and Alexander G Schwing. Mask2former for video instance segmentation. *arXiv preprint arXiv:2112.10764*, 2021. 2, 3, 4, 6, 7

[12] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022. 1, 2, 3, 4, 6, 8

[13] Ho Kei Cheng, Seoung Wug Oh, Brian Price, Alexander Schwing, and Joon-Young Lee. Tracking anything with decoupled video segmentation. In *IEEE/CVF International Conference on Computer Vision*, pages 1316–1326, 2023. 4

[14] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org*, 2(3):6, 2023. 7, 3

[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009. 6

[16] Hao Fang, Tong Zhang, Xiaofei Zhou, and Xinxin Zhang. Learning better video query with sam for video instance segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 6, 7

[17] Shengyi Gao, Zhe Chen, Guo Chen, Wenhai Wang, and Tong Lu. Avsegformer: Audio-visual segmentation with transformer. In *AAAI Conference on Artificial Intelligence*, pages 12155–12163, 2024. 1, 3, 6, 7

[18] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 776–780, 2017. 5, 2

[19] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023. 3

[20] Ruohao Guo, Dantong Niu, Liao Qu, and Zhenbo Li. Sotr: Segmenting objects with transformers. In *IEEE/CVF International Conference on Computer Vision*, pages 7157–7166, 2021. 2

[21] Ruohao Guo, Dantong Niu, Liao Qu, Yanyu Qi, Ji Shi, Wenzhen Yue, Bowei Xing, Taiyan Chen, and Xianghua Ying. Instance-level panoramic audio-visual saliency detection and ranking. In *ACM International Conference on Multimedia*, pages 9426–9434, 2024. 3

[22] Ruohao Guo, Liao Qu, Dantong Niu, Yanyu Qi, Wenzhen Yue, Ji Shi, Bowei Xing, and Xianghua Ying. Open-vocabulary audio-visual semantic segmentation. In *ACM International Conference on Multimedia*, pages 7533–7541, 2024. 1, 3

[23] Ruohao Guo, Xianghua Ying, Yanyu Qi, and Liao Qu. Unitr: A unified transformer-based framework for co-object and multi-modal saliency detection. *IEEE Transactions on Multimedia*, 2024. 1

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5, 2

[25] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *IEEE/CVF International Conference on Computer Vision*, pages 2961–2969, 2017. 2

[26] Miran Heo, Sukjun Hwang, Seoung Wug Oh, Joon-Young Lee, and Seon Joo Kim. Vita: Video instance segmentation via object token association. *Advances in Neural Information Processing Systems*, 35:23109–23120, 2022. 2, 3, 4, 5, 6, 7

[27] Shaofei Huang, Rui Ling, Hongyu Li, Tianrui Hui, Zongheng Tang, Xiaoming Wei, Jizhong Han, and Si Liu. Unleashing the temporal-spatial reasoning capacity of gpt for training-free audio and language referenced video object segmentation. *arXiv preprint arXiv:2408.15876*, 2024. 7, 4, 5

[28] Sukjun Hwang, Miran Heo, Seoung Wug Oh, and Seon Joo Kim. Video instance segmentation using inter-frame communication transformers. *Advances in Neural Information Processing Systems*, 34:13352–13363, 2021. 2, 3, 6

[29] Herve Jegou, Matthijs Douze, and Cordelia Schmid. Product quantization for nearest neighbor search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(1):117–128, 2010. 6

[30] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 4, 7, 3

[31] Rajat Koner, Tanveer Hannan, Suprosanna Shit, Sahand Sharifzadeh, Matthias Schubert, Thomas Seidl, and Volker Tresp. Instanceformer: An online video instance segmentation framework. In *AAAI Conference on Artificial Intelligence*, pages 1188–1195, 2023. 2, 6

[32] Harold W Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955. 3, 6, 1

[33] Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19108–19118, 2022. 4

[34] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, pages 19730–19742. PMLR, 2023. 3

[35] Kexin Li, Zongxin Yang, Lei Chen, Yi Yang, and Jun Xiao. Catr: Combinatorial-dependence audio-queried transformer for audio-visual video segmentation. In *ACM International Conference on Multimedia*, pages 1485–1494, 2023. 1, 3

[36] Zhangbin Li, Dan Guo, Jinxing Zhou, Jing Zhang, and Meng Wang. Object-aware adaptive-positivity learning for audio-visual question answering. In *AAAI Conference on Artificial Intelligence*, pages 3306–3314, 2024. 4

[37] Zhangbin Li, Jinxing Zhou, Jing Zhang, Shengeng Tang, Kun Li, and Dan Guo. Patch-level sounding object tracking for audio-visual question answering. *arXiv preprint arXiv:2412.10749*, 2024. 4

[38] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014. 3

[39] Dongfang Liu, Yiming Cui, Wenbo Tan, and Yingjie Chen. Sg-net: Spatial granularity network for one-stage video instance segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9816–9825, 2021. 2

[40] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2024. 7

[41] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer, 2024. 3, 4

[42] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 5, 2

[43] Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. Hota: A higher order metric for evaluating multi-object tracking. *International Journal of Computer Vision*, 129: 548–578, 2021. 3, 1

[44] Shehan Munasinghe, Rusiru Thushara, Muhammad Maaz, Hanoona Abdul Rasheed, Salman Khan, Mubarak Shah, and Fahad Khan. Pg-video-llava: Pixel grounding large video-language models. *arXiv preprint arXiv:2311.13435*, 2023. 7, 3, 4, 5

[45] Jiyang Qi, Yan Gao, Yao Hu, Xinggang Wang, Xiaoyu Liu, Xiang Bai, Serge Belongie, Alan Yuille, Philip HS Torr, and Song Bai. Occluded video instance segmentation: A benchmark. *International Journal of Computer Vision*, 130(8): 2022–2039, 2022. 4, 8

[46] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 3, 4

[47] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 5

[48] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4358–4366, 2018. 4

[49] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *IEEE/CVF International Conference on Computer Vision*, pages 9627–9636, 2019. 2

[50] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 4

[51] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8741–8750, 2021. 2

[52] Junfeng Wu, Yi Jiang, Song Bai, Wenqing Zhang, and Xiang Bai. Seqformer: Sequential transformer for video instance segmentation. In *European Conference on Computer Vision*, pages 553–569. Springer, 2022. 2, 3, 6, 7

[53] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5. IEEE, 2023. 2

[54] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, 2023. 6

[55] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *IEEE/CVF International Conference on Computer Vision*, pages 5188–5197, 2019. 2, 3, 4, 6, 1

[56] Qi Yang, Xing Nie, Tong Li, Pengfei Gao, Ying Guo, Cheng Zhen, Pengfei Yan, and Shiming Xiang. Cooperation does matter: Exploring multi-order bilateral relations for audio-visual segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27134–27143, 2024. 1, 3, 6, 7, 5

[57] Shusheng Yang, Yuxin Fang, Xinggang Wang, Yu Li, Chen Fang, Ying Shan, Bin Feng, and Wenyu Liu. Crossover learning for fast online video instance segmentation. In *IEEE/CVF International Conference on Computer Vision*, pages 8043–8052, 2021. 2

[58] Shusheng Yang, Xinggang Wang, Yu Li, Yuxin Fang, Jiemin Fang, Wenyu Liu, Xun Zhao, and Ying Shan. Temporally efficient vision transformer for video instance segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2885–2895, 2022. 2, 6, 7

[59] Jiarui Yu, Haoran Li, Yanbin Hao, Jinmeng Wu, Tong Xu, Shuo Wang, and Xiangnan He. How can contrastive pre-training benefit audio-visual segmentation? a study from supervised and zero-shot perspectives. In *British Machine Vision Association*, pages 367–374, 2023. 7, 2, 4

[60] Tao Zhang, Xingye Tian, Yu Wu, Shunping Ji, Xuebo Wang, Yuan Zhang, and Pengfei Wan. Dvis: Decoupled video instance segmentation framework. In *IEEE/CVF International Conference on Computer Vision*, pages 1282–1291, 2023. 2, 6, 7

[61] Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al. Recognize anything: A strong image tagging model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1724–1732, 2024. 3, 4

[62] Yang Zhao, Zhijie Lin, Daquan Zhou, Zilong Huang, Jiashi Feng, and Bingyi Kang. Bubogpt: Enabling visual grounding in multi-modal llms. *arXiv preprint arXiv:2307.08581*, 2023. 7, 3, 4

[63] Jinxing Zhou, Liang Zheng, Yiran Zhong, Shijie Hao, and Meng Wang. Positive sample propagation along the audio-visual event line. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8436–8444, 2021. 1

[64] Jinxing Zhou, Dan Guo, and Meng Wang. Contrastive positive sample propagation along the audio-visual event line. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7239–7257, 2022. 1

[65] Jinxing Zhou, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, and Yiran Zhong. Audio–visual segmentation. In *European Conference on Computer Vision*, pages 386–403. Springer, 2022. 1, 3, 4

[66] Jinxing Zhou, Dan Guo, Ruohao Guo, Yuxin Mao, Jingjing Hu, Yiran Zhong, Xiaojun Chang, and Meng Wang. Towards open-vocabulary audio-visual event localization. *arXiv preprint arXiv:2411.11278*, 2024. 1

[67] Jinxing Zhou, Dan Guo, Yuxin Mao, Yiran Zhong, Xiaojun Chang, and Meng Wang. Label-anticipated event disentanglement for audio-visual video parsing. In *European Conference on Computer Vision*, pages 35–51. Springer, 2024.

[68] Jinxing Zhou, Dan Guo, Yiran Zhong, and Meng Wang. Advancing weakly-supervised audio-visual video parsing via segment-wise pseudo labeling. *International Journal of Computer Vision*, 132(11):5308–5329, 2024. 1

[69] Jinxing Zhou, Xuyang Shen, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, et al. Audio-visual segmentation with semantics. *International Journal of Computer Vision*, pages 1–21, 2024. 1, 3, 4

[70] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2021. 5