Benchmarking Audio Visual Segmentation for Long-Untrimmed Videos

Chen Liu[1,2,3], Peike Patrick Li[4], Qingtao Yu[5], Hongwei Sheng[1],
Dadong Wang[3], Lincheng Li[2*], Xin Yu[1*]
[1] The University of Queensland [2] NetEase Fuxi AI Lab, [3] CSIRO DATA61,
[4] Matrix Verse, [5] Australian National University
chenliu7@uqconnect.edu.au, peikeli912@gmail.com, Terry.Yu@anu.edu.au,
hongwei.sheng-1@student.uts.edu.au, Dadong.Wang@data61.csiro.au,
lilincheng@corp.netease.com, xin.yu@uq.edu.au

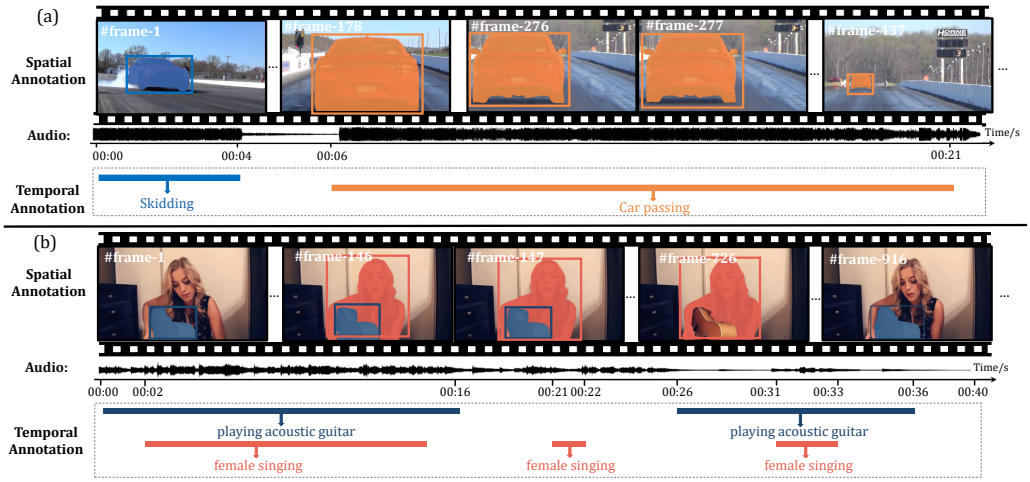


Figure 1. Samples of our **L**ong-**U**ntrimmed **A**udio-**V**isual **S**egmentation dataset. Different from previous AVS datasets, LU-AVS is crafted to explore the challenges inherent in the AVS task for long-untrimmed videos. It features detailed start- and end-sounding positions in the temporal dimension, along with comprehensive mask and bounding box annotations in the spatial dimension. The examples show that our dataset contains numerous audible segments in each video, characterized by diverse durations and varying start and end-sounding positions. Additionally, within a single video, the same objects may have notable shifts spatially and undergo deformation.

Abstract

*Existing audio-visual segmentation datasets typically focus on short-trimmed videos with only one pixel-map annotation for a per-second video clip. In contrast, for untrimmed videos, the sound duration, start- and end-sounding time positions, and visual deformation of audible objects vary significantly. Therefore, we observed that current AVS models trained on trimmed videos might struggle to segment sounding objects in long videos. To investigate the feasibility of grounding audible objects in videos along both temporal and spatial dimensions, we introduce the **L**ong-**U**ntrimmed **A**udio-**V**isual **S**egmentation dataset (LU-AVS), which includes precise frame-level annotations of sounding emission times and provides exhaustive mask annotations for all frames. Considering that pixel-level annotations are difficult to achieve in some complex scenes, we also provide the bounding boxes to indicate the sounding regions. Specifically, LU-AVS contains 10M mask annotations across 6.6K videos, and 11M bounding box annotations across 7K videos. Compared with the existing datasets, LU-AVS videos are on average 4∼8 times longer, with the silent duration being 3∼15 times greater. Furthermore, we try our best to adapt some baseline models that were originally designed for audio-visual-relevant tasks to examine the challenges of our newly curated LU-AVS. Through comprehensive evaluation, we demonstrate the challenges of LU-AVS compared to the ones containing trimmed videos. Therefore, LU-AVS provides an ideal yet challenging platform for evaluating audio-visual segmentation and localization on untrimmed long videos. The dataset is publicly available at: https://yenanliu.github.io/LU-AVS/.*

*Corresponding authors.

1. Introduction

Audio-visual segmentation (AVS) is an emerging field that segments objects in images according to the sounds from the given audio. Existing AVS datasets, such as [60, 61], commonly slice raw videos into short clips with 5 or 10-second duration. Moreover, in these short-trimmed videos, the target-sounding object is often the isolated or salient one that occurs at the start of the video clip and continues sounding throughout the whole video. As a result, the sound properties of audio are often given less emphasis, and methods developed based on these trimmed videos can be degraded into salient object segmentation, still achieving reasonable performance. However, in untrimmed videos, the start and end sounding frames of an audible object are uncertain, and the sounding duration of the target objects varies significantly among different videos. Therefore, it is necessary to investigate the AVS task on long-untrimmed videos.

To this end, we propose a large-scale Long-Untrimmed Audio-Visual Segmentation (LU-AVS) dataset. Our LU-AVS dataset comprises 6.6K untrimmed videos covering 78 categories, and 10M pixel-level annotation masks are provided to indicate the audible objects. Noticeably, in the annotation process, we found not all sounding regions can be clearly identified with masks. For example, videos in the *sailing* category are often shot from a first-person perspective, posing a challenge for mask-based sound region labeling. Consequently, we extend our dataset with bounding box annotations, extending them to both previously mask-marked categories and those challenging to annotate with masks. The extended LU-AVS dataset contains about 7.2K untrimmed videos and more than 88 categories.

To construct a high-quality audio-visual untrimmed dataset, we first select long untrimmed videos that contain objects that emit clear sound. Thus, our newly collected videos are much longer compared to the video clips in the existing AVS datasets [60, 61]. After collecting enough videos, it is observed that the audible objects have various starting and ending timestamps, possess diverse sounding durations, and appear in different spatial positions. Furthermore, a single video may feature several objects producing sounds concurrently or in an asynchronous manner. Next, we annotate the collected videos in a semi-automatic fashion with the help of a vision foundation model [25] and a tracking method [9]. Unlike existing datasets that label the first frame per second, the LU-AVS dataset provides pixel-level annotations for each audible object across all frames. In this way, sounding objects with large visual deformation will have more supervision within one second. Moreover, we ask human annotators to label sounding objects that are hard to mask with bounding boxes.

Considering objects may produce sounds at any time in videos and span an indeterminate number of frames, this imposes significant challenges to localizing objects at both spatial and temporal dimensions. With the proposed dataset, we benchmark the existing audio-visual segmentation (AVS) method and conduct a comprehensive comparison by adapting several state-of-the-art methods that were proposed for relevant tasks to the trimmed audio-visual segmentation task. We notice that segmenting the audible object requires audio-visual correlation on each individual frame while identifying the start and the end sounding frame demands the temporal context across the whole video. Moreover, the experimental results demonstrate that LU-AVS presents more challenges than existing datasets, resulting in the current AVS methods being less effective in handling untrimmed videos.

Overall, our work provides a foundation for developing more advanced audio-visual segmentation methods that require full exploration of the audio-visual correlation for audible object segmentation at a spatial and temporal level in untrimmed videos. In particular, we propose a large-scale untrimmed audio-visual segmentation dataset, **LU-AVS**. Based on LU-AVS, we develop and conduct comprehensive evaluations of the state-of-the-art audio-visual segmentation methods and the temporal sentence grounding methods, providing a reference for future works. We also develop a simple baseline approach, which points to potential solutions to some of the challenges and future research directions.

2. Related Work

Audio-Visual Segmentation. Audio-visual segmentation aims to localize audible objects by a pixel-level map for a given audio-visual pair. This task requires both audio and visual understanding and is one of the most fundamental yet challenging tasks in computer vision. Audio-visual segmentation was first introduced by Zhou *et al*. [60], and they also released the first AVS dataset AVSBench-Object. In the AVSBench-Object dataset, videos are trimmed to be five seconds in length, and each second contains only one frame of binary mask annotation. It should be noted that the masks provided by AVSBench-Object do not distinguish the categories of sounding objects. Later on, an extended semantic dataset AVSBench-Semantic dataset has been proposed in [61]. The labeled semantic maps indicate the audible object categories and each video is trimmed to a longer duration of 10 seconds.

Built upon these datasets, many works aim to improve the segmentation quality by enhancing the audio-visual interactions [17, 22, 31, 35, 36]. For instance, Huang *et al*. [22] devise a set of object queries that are conditioned on audio information and then associate each query with sounding objects. Li *et al*. [27] propose a decoupled audio-visual transformer that combines audio and visual features from the temporal and spatial dimensions. In addition to improving the segmentation mask quality, several works em-

Table 1. Statistics of publicly-available AVS datasets. Compared to the existing AVS datasets, the newly curated LU-AVS possesses a greater number of mask annotations and a broader range of categories. Additionally, it incorporates two forms of spatial annotations. For video samples where the pixel-level annotations are hard to obtain (denoted as 'Hard'), we provide the bounding box annotations. In contrast, 'Normal' samples are labeled with bounding boxes and mask annotations. More importantly, the average durations of videos and their audible segments in LU-AVS are both higher than those in other AVS datasets. This demonstrates that LU-AVS facilitates the exploration of challenges posed by untrimmed videos in the AVS task.

Dataset	Real-Data	Type	Data Amount		Category	Temporal Anntation	Annotation Amount		Avg. Video Duration (s)	Silent Proportion (%)	Avg. Segment Duration (s)
			Normal	Hard			Mask	BBox			
AVSBench-Object [60]	✓	Video	5,356	0	23	✗	26,458	✗	5	1.2	4.94
AVSBench-Semantic [61]	✓	Video	11,356	0	70	✗	82,335	✗	7.64	5.1	7.25
VPO-SS [52]	✗	Image+Audio	**12,202**	0	21	✗	12,202	✗	10	0	10
VPO-MS [52]	✗	Image+Audio	9,817	0	21	✗	13,496	✗	10	0	10
LU-AVS (Ours)	✓	Video	6,627	630	**88**	✓	**10,350,009**	**10,981,586**	**41.97**	**15.45**	**16.03**

phasize the importance of audio in this task. As suggested by Yuan *et al*. [52], Mao *et al*. [36] and Chen *et al*. [31], existing AVS datasets have a serious bias caused by the limited and less diverse data, rendering an audio-visual segmentation model to be a saliency segmentation model. To mitigate this problem, Chen *et al*. [31] postposition the audio-visual interaction process and leverage potential sounding objects to guide the audio classification. Mao *et al*. [36] emphasize the modality-specific representation by using latent space factorization to find the decouple space and the shared space of each modality. Yuan *et al*. [52] create a synthetic dataset by collecting images from the COCO dataset and audio files from the VGGSound dataset.

To fundamentally solve the data bias problem introduced by the existing AVS datasets, our work curates a large untrimmed audio-visual segmentation dataset. The diverse start and end-sounding timestamps of target objects, along with various audio durations, make our data more challenging and more closely resemble real-life scenarios.

Spatio-Temporal Video Grounding. Spatio-temporal video grounding (STVG) aims to detect the temporal boundaries and the spatial object tube at the same time, according to the given sentence [12, 26, 29, 41, 42, 45, 47, 53–58]. It was first introduced by Gao *et al*. [13] and Anne *et al*. [1], and has drawn significant attention recently.

A standard paradigm requires a pre-defined object proposal generator [32, 33, 44, 46, 48, 51, 58]. Some works design one-stage approaches but take more computing resources to process long videos [11, 21, 24, 30, 34, 47, 49, 50]. Zhang *et al*. [58] first generate object tubes based on a spatial-temporal region graph module, and then incorporate the textual clues into the graph for reasoning. Su *et al*. [41] devise a cross-modal feature learning module and leverage the cross-modal feature to generate bounding boxes for a target object and predict its starting and ending frames, thus producing a target object tube. Yang *et al*. [47] propose a transformer-based framework that models spatial-temporal interactions in its encoder and jointly performs spatial-temporal localization in its decoder. Jin *et al*. [23] introduce a spatio-temporal consistency-aware framework, explicitly constricting the grounding regions and associating them across the whole frames.

Although AVS and STVG tasks both involve determining the spatial and temporal positioning of a target object in a video based on a given prompt, such as text and audio [18, 28, 60, 61], there are still significant differences between the two tasks. For the STVG task, the sentence prompt provides clarity and precision, allowing for specific instructions, while audio may be more ambiguous and subject to interpretation. This requires the segmentation model to possess a sophisticated understanding of audio-visual associations [39, 43]. Moreover, audio, as part of the video's soundtrack, can offer real-time cues for segmentation, aligning closely with the video's content as it progresses [2], while a sentence is static and does not change over time [16].

3. LU-AVS Dataset

In this section, we introduce the newly curated long-untrimmed dataset **LU-AVS** by first presenting the video collection and annotation process in Section 3.1 and Section 3.2, respectively. Then we provide the dataset statistics and analysis in Section 3.3.

3.1. Data Source

We collect videos from the VGGSound dataset [4], a comprehensive dataset where the sounds and the visual contexts are well aligned. Initially, we select a subset from its 300 categories, prioritizing those frequently encountered in daily life and spanning various domains. Unlike the original dataset, which slices videos into 10-second clips, we adopt the original untrimmed videos using the provided YouTube URLs. It should be noted that we randomly cut the raw videos to keep them around one minute since many of them are over several hours. Afterward, we carry out a more elaborate selection process, in which the following criteria need

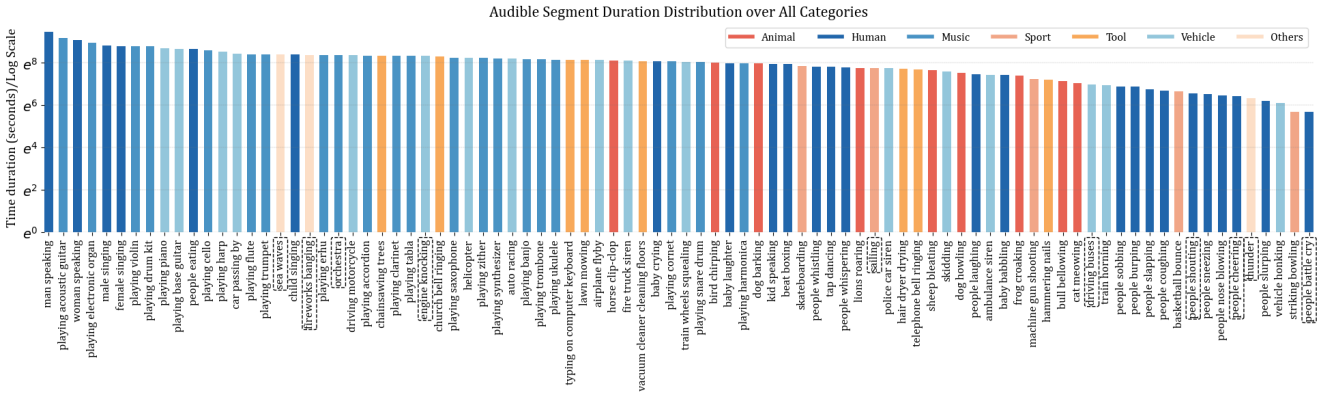


Figure 2. Sounding durations of each audible segment class in the LU-AVS dataset sorted by descending order, with colors indicating audible segment types. The category labels framed with dashed lines only include bounding box annotations, while the other categories contain mask and bounding box annotations.

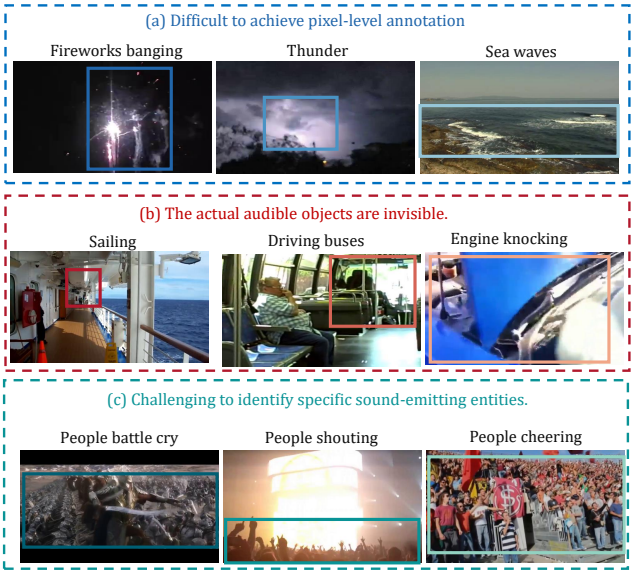


Figure 3. Examples that are hard to annotate with masks. Thus, we only provide bounding-box annotations to those categories.

to be met:

- Exclude low-quality videos in which it is hard to distinguish the sounding sources. One can clearly identify and track audible objects in the videos.
- To provide mask annotations, we also need to distinguish objects that can be explicitly masked. Examples of hard examples that cannot be masked are shown in Figure 3.

These criteria enable high-quality annotations for sounding regions and ensure the annotated objects can be tracked in videos. Finally, we select 7,257 videos of high quality to create a benchmark that is diverse and representative of a wide range of real-world scenarios.

3.2. Data Annotation

Considering not all sounding regions can be identified with masks, such as the *'fireworks banging'* in Figure 3, we use two annotation forms to describe the audible objects, *i.e.* the mask and the bounding box. Specifically, there are 6,627 videos annotated with masks, spanning 78 categories, and 7,257 videos annotated with bounding boxes, distributed across 88 categories [*]. We annotate the selected videos in two distinct steps. The first step involves identifying the category of each sounding object in a video, as well as determining the start and end frames of each audible segment. The second step entails annotating the audible object by bounding boxes and masks within each audio-visual tube. In the following section, we will introduce the process of spatial and temporal annotations.

• **Audible Tube Annotation.** We develop an annotation tool and invite five annotators to specify the start/emergence and end/vanishing frames of the audible objects in videos. In this step, the start and end frames are determined based on whether the visual and sound appear simultaneously and whether one factor disappears respectively. After that, we will filter out invalid annotations and integrate valid ones. Specifically, if one audible tube of one object has overlappings with other distinct objects more than three times and the temporal Intersection over Union (tIoU) value[†] between each other exceeds 0.8, we consider they are the valid audible tube annotations and correspond to the same sounding object. We take the mean of their start and end frames and employ the mode of the category annotations within a tube to determine the category labels.

• **Mask Annotation.** Based on the temporal annotations, we trim the long videos into audible tubes to facilitate the mask annotations. Manually annotating masks in videos is extremely expensive. Therefore, based on the large off-the-shelf visual foundation model SAM [25] and the object tracking method XMem [9], we develop a semi-automatic annotation tool to simplify the labeling procedure. Specifically, with SAM, our annotators only need to click positive and negative points within the sounding regions to generate

[*]The number of categories annotated by bounding boxes is more than the number of categories annotated by masks in Figure 2.

[†]The calculation formula is as follows: $tIoU = \frac{Intersection\ of\ the\ two\ audible\ tube\ frame\ spans}{Union\ of\ the\ two\ audible\ tube\ frame\ spans}$

the mask for the target object. To ensure mask quality, we leverage the median filter to smooth the generated masks. Additionally, annotators can also use polygons to annotate some complicated samples.

After obtaining the first labeled frame, the tracking method can automatically generate the masks of the audible object in the remaining frames. However, in many cases, the target objects are obscured or deformed, and the tracking method may fail to track and segment the object. To guarantee the annotation quality of all frames in an audible tube, annotators must re-label frames with low-quality masks until the audible object is precisely delineated throughout the entire audible tube. *More details about our developed annotator are in supplementary materials.*

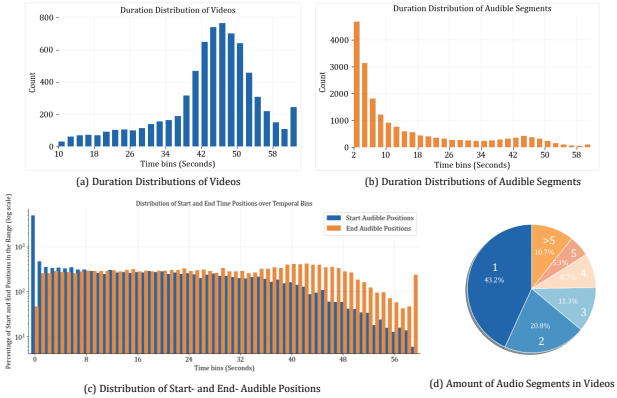
• **Bounding Box Annotation.** In addition to the mask annotations, we also provide bounding boxes with Fuxi Youling Crowdsourcing[‡] to identify the sounding regions. Commencing from the initial frame, we extract every subsequent frame at a one-second interval within a video. Each frame is annotated by three annotators, and the final annotation is determined by averaging the values of these bounding boxes. Additionally, we employ the mean shift tracking [10] and kernelized correlation filter [19] tracking methods for both forward and backward tracking of the target object within the 1-second frame intervals, thereby obtaining bounding boxes for the target object in the intermediate frames. After that, we manually check the annotation quality, and annotators re-label the frames that are difficult to obtain high-quality bounding boxes by the tracking methods.

3.3. Dataset Analysis and Statistics

• **Overview of LU-AVS Dataset.** In Table 1, we present a statistical analysis of the newly proposed **LU-AVS** dataset, using four previous audio-visual segmentation datasets as reference, including AVSBench-Object [60], AVSBench-Semantic [61], VPO-SS [52], and VPO-MS [52]. As shown in Table 1, LU-AVS contains 7.2K videos spanning 88 categories with 10M annotated masks and 11M bounding boxes. Compared to AVSBench-Semantic [61], LU-AVS has more annotation masks (10,350,009 *vs* 82,335), longer average video durations (41.97 *vs* 7.64), and longer average audible tube durations (16.03 *vs* 7.25). In addition, the proportion of silent-fragment duration to the total video duration is much higher than that of the other AVS datasets, thus imposing more challenges to AVS methods.

Figure 2 provides the statistical information on the sounding duration in each category in the LU-AVS dataset. Overall, the category distribution of the dataset spans a wide range of domains, including Animal, Human, Music, Sport, Tool, Vehicle, and others. In this dataset, the total duration of videos is 84.6 hours, of which the sounding duration is 73.3 hours. We divide the training, validation, and test sets

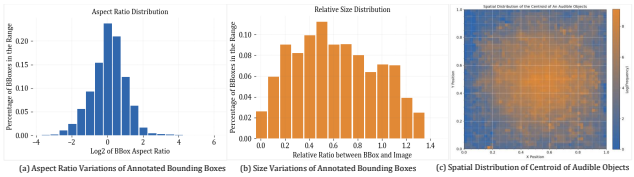
[‡]https://fuxi.163.com/solutions/data



(a) Duration Distributions of Videos (b) Duration Distributions of Audible Segments

(c) Distribution of Start- and End- Audible Positions (d) Amount of Audio Segments in Videos

Figure 4. Statistics on the temporal structure of LU-AVS. (a) and (b) indicate the duration distribution of videos and audible segments, respectively. (c) presents the distribution of the start- and end-sounding positions of audible segments in the temporal dimension. (d) shows the statistics of the number of audible segments/tubes in videos.



(a) Aspect Ratio Variations of Annotated Bounding Boxes (b) Size Variations of Annotated Bounding Boxes (c) Spatial Distribution of Centroid of Audible Objects

Figure 5. Statistics on the spatial distribution of sounding objects in our LU-AVS dataset. (a) and (b) illustrate the size and aspect ratio distributions of annotated bounding boxes, respectively. (c) represents the spatial distribution of the centroids of object masks.

in a 3: 1: 1 ratio. Moreover, to mitigate the impact of long tails, we ensure that each category has at least 50 audible segments. In the following, we provide a detailed statistical analysis of LU-AVS to further demonstrate the complexity and difficulty of the dataset.

• **Temporal Characteristics.** As suggested in Figure 4 (a) and (b), the length of video durations and audible segment durations are various. Shorter segments pose greater challenges in temporal localization, such as locating a 1.4s *'thunder'* segment in a 37s video. In contrast, longer segments present more difficulty in spatially segmenting target audible objects, such as identifying and tracking a rapidly deforming *'race car'* in a 45s audible tube. Figure 4 (c) reveals the start- and end-time distribution of audible segments is broad. Furthermore, Figure 4 (d) indicates that over half of the videos have multiple audible segments. This variety of object emergence and vanishing and the presence of several audible segments per video highlight the importance of sound recognition and increase the complexity of identifying audible objects over time.

• **Spatial Characteristics.** Figure 5 shows the distribution of spatial annotations in LU-AVS. As suggested in Figure 5 (a), bounding boxes with aspect ratios of 1:1, 1:2, and 2:1 are the most common, due to the high proportion of peo-

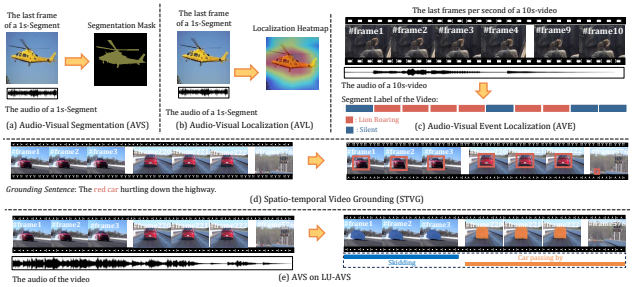


Figure 6. The task diagram of the audio-visual segmentation (AVS), audio-visual localization (AVL), audio-visual event localization (AVE), spatio-temporal video grounding (STVG), and the AVS task on LU-AVS.

ple, music instruments, and tools in the dataset. However, there are nearly 40% of the bounding boxes have unusual aspect ratios. This variability can be explained by the deformation of objects during motion, such as the significant skateboard changes during aerial spins. As illustrated in Figure 5 (b), the size of target objects is widely distributed, which increases the difficulty of audible object detection. Additionally, Figure 5 (c) indicates diverse spatial positions of audible objects across images. This diverse object spatial distribution mitigates the data bias problem and also increases the challenges of tracking objects in videos.

4. Strong Baselines for Benchmarking

To show the necessity and fully explore the challenges of LU-AVS, we investigate the performance of existing audio-visual segmentation (AVS) [36, 61], audio-visual localization (AVL) [5, 37, 38, 40], audio-visual event localization (AVE) [4, 8, 15, 43, 59], and spatio-temporal video grounding (STVG) [23, 47] methods on our LU-AVS dataset. In Figure 6, we illustrate the difference among these tasks.

• **Audio-Visual Segmentation Methods.** The goal of the AVS task is to segment audible objects in an image based on a given audio-visual pair. TPAVI [61] and ECMVAE [36] are designed based on the existing datasets with the fixed input format of 10 frames corresponding to 10s audio. To adapt these methods for untrimmed videos, we modify the input to one second of audio and five uniformly sampled frames from the segment, allowing audio-visual interactions within the per-second segment. Additionally, to obtain the temporal predictions, we consider the segment between the first and the last frame where the target object continuously appears as a predicted segment.

• **Audio-Visual Localization Methods.** The AVL task also focuses on locating audible objects in the spatial dimension. However, AVL presents sounding regions by heatmaps. For comparison, in the test stage, we convert the heatmaps into bounding boxes as [5]. Thanks to our bounding-box annotations, we can evaluate AVL on our LU-AVS dataset in a unified manner. Similar to the modification for AVS methods, we slice the videos into 1-second segments to fit the

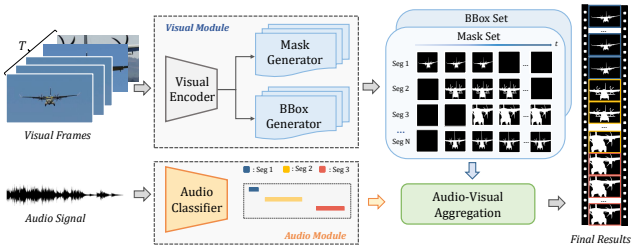


Figure 7. The overview architecture of a strong baseline. It first learns visual and audio features separately and then establishes visual and audio associations. It enables us to dissect the impacts of visual and audio branches explicitly.

AVL methods.

• **Audio-Visual Event Localization Methods.** AVE task aims at determining the audio-visual temporal segments when the target object is both audible and visible at the same time. This task does not focus on segmenting audible objects on the spatial dimension, they cannot be applied to segmenting objects in images. These methods are also developed based on the videos with fixed durations. In this setting, they usually slice the trimmed video into clips with a 1s duration and further predict the category label of these clips. To adopt them into our methods, we slice our videos into multiple one-second clips. If the sounding duration of an audible object does not exceed 0.5s, the segment is masked as silent.

• **Spatio-temporal Video Grounding Methods.** Given a query sentence, STVG methods are required to track the target object in the video both at the spatial (bounding boxes) and temporal dimensions (start and end time positions). Unlike the AVS task that may require segmenting multi-sounding objects in a video, the STVG methods only need to find and track one target object for each text-video pair. To explore the performance of STVG methods on LU-AVS, we modify these text-guided methods to adapt to our task. To be specific, we replace the text branch of these methods with the audio encoder *VGGish* [20]. Different from extracting the text feature from the whole sentence, we extract audio features for each 0.96s segment with a sliding window (stride=0.32s) to temporally interact with the corresponding visual information.

• **A Strong LU-AVS Baseline.** The LU-AVS dataset introduces unique challenges in grounding audible objects in untrimmed videos along the spatial and temporal dimensions. Different from the existing AVS datasets having videos with fixed durations, the target objects may emit or stop sounding at random positions and the sounding durations vary in LU-AVS, making it necessary to capture the audio-visual interactions throughout the entire video. Moreover, different from the typical spatio-temporal video grounding task just queries one target object according to the given sentence, the LU-AVS dataset includes videos with multiple sounding sources, increasing the difficulty of

grounding audible objects in videos.

Based on the above observation, we introduce a simple framework to provide a base reference for long video audible object grounding in the future. As depicted in Figure 7, we first employ the visual module to identify the potential-sounding objects in videos, across T frames. In detail, we utilize MaskFormer [7] and DETR [3] to generate the masks and bounding boxes, respectively. Then we integrate the potential-sounding object sequence along the time dimension according to the semantic labels of objects. Note that, in the visual branch training process, we reclassify object labels according to visual information. For instance, *'baby crying'* and *'baby laughing'* are categorized as *'baby'*. For the audio branch, we employ the VGGish as the audio feature extractor and perform a sound segment classification task. Specifically, we split the whole audio into multiple one-second duration segments and further obtain the audio label of each segment. Based on the visual and audio results obtained by the two branches, we finally aggregate them according to their semantic labels. More implementation details, experimental settings, and results are provided in the supplementary materials.

5. Experiments

5.1. Evaluation Metrics

In contrast to earlier research [52, 60, 61] that concentrated solely on the spatial segmentation of audible objects, the LU-AVS dataset extends this focus by incorporating detailed temporal annotations, specifically marking the starting and ending positions of sounds in videos. Furthermore, the dataset's comprehensive spatial annotations demand a more rigorous approach to maintain segmentation consistency for audible objects. Consequently, these enhanced complexities underline the need for developing innovative evaluation metrics that encompass both spatial and temporal dimensions. Inspired by the assessment criteria in AVS [60] and STVG tasks [6, 14], we develop new evaluation metrics for the LU-AVS task.

Specifically, for the dataset with mask annotations, we employ m_tIoU, m_tF, m_vIoU, and m_vF as evaluation metrics. m_tIoU and m_tF are the average temporal IoU and F-score between the ground-truth audible segments and the predicted audible segments, respectively. For m_vIoU and m_vF, we first define $\mathcal{S}_\mathcal{U}$ as the set of frames contained in either the predicted or ground-truth segments and $\mathcal{S}_\mathcal{I}$ as the set of frames in both predicted and ground-truth segments. We then calculate vIoU by $vIoU = \frac{1}{|\mathcal{S}_\mathcal{U}|}\sum_{t\in\mathcal{S}_\mathcal{I}} IoU(r^t, \hat{r}^t)$, where r^t and $\hat{r}^t$ represent the predicted and ground-truth regions of frame t, respectively. Similarly, vF is measured by $vF = \frac{1}{|\mathcal{S}_\mathcal{U}|}\sum_{t\in\mathcal{S}_\mathcal{I}} \mathcal{F}(r^t, \hat{r}^t)$. m_vIoU and m_vF are the average vIoU and vF of samples, respectively. For the dataset annotated by bounding boxes, we utilize m_tIoU and m_vIoU to measure the model performance.

5.2. Benchmarking Results

• **Results Analysis of AVS Methods.** We conducted a comparative analysis of the efficacy of TPAVI [61] and ECMVAE [36] on LU-AVS. As delineated in Table 2 (a), it is evident that both TPAVI and ECMVAE exhibit suboptimal performance across all evaluated metrics, with none surpassing the 15% threshold. This suggests that current AVS methods have significant limitations in processing long-untrimmed videos. Notably, the disparity between m_tIoU and m_vIoU scores highlights the inherent challenges these methods face in maintaining consistent audio-visual congruence throughout the entirety of the video.

• **Results Analysis of AVL Methods.** As indicated in Table 2 (b), there is a notable underperformance of the adapted AVL methods in terms of both m_tIoU and m_vIoU, with all recorded values falling below 14%. Predominantly, methods such as those presented in [5, 37], which are tailored for datasets like [4] featuring trimmed and consistently audible content, fail to account for silent segments. However, in the context of our dataset, silent intervals constitute a significant portion, approximately 12.18% as reported in Table 1. Consequently, models trained on these segments tend to neglect the accurate localization of audible objects.

• **Results Analysis of AVE Methods.** As original AVE methods do not emphasize spatial segmentation of audible objects, we adapt the evaluation metrics to assess how well models predict segment labels over time. As illustrated in Table 2 (c), AVE methods excel in both m_tIoU and m_vIoU metrics, surpassing 33% across all metrics. By concentrating on temporal aspects, these methods avoid the complexities of spatial video data processing, such as object movement, deformation, and occlusion. Therefore, compared to methods in other tasks, they achieve better performance.

• **Results Analysis of STVG Methods.** For STVG audio-based methods, all metrics are calculated based on bounding box annotations. As suggested by Table 2 (d), STVG audio-based methods excel in the m_tIoU metric, with all methods surpassing 15%. However, their performance drops below 8% in m_vIoU, indicating poor temporal localization. We attribute this to the weak audio-visual interactions. Specifically, unlike the query sentence interacting with all frames in a video, the audio signal interfaces with its corresponding frame in time. Furthermore, while the STVG task focuses on tracking just one target object in a video, methods developed for LU-AVS require segmenting and tracking multiple sounding objects simultaneously within videos.

• **Results Analysis of a Strong Baseline.** As shown in Table 2 (e), our method (mask-based) achieves a significant improvement (from 6.03% to 17.32%) over TPAVI on

Table 2. Benchmarking results on the LU-AVS dataset. For all the evaluation metrics, higher values indicate better performance. Notably, in AVL methods, the spatial localization results are presented by heatmaps. For comparison, we convert heatmaps to bounding boxes as [5]. Additionally, AVE methods focus on temporal localization. Here, m_tIoU represents the segmentation accuracy within the ground-truth temporal range, and m_vIoU indicates the segmentation accuracy over the temporal union between the predicted and ground-truth durations. Besides, we replace the text branch in the STVG methods with an audio branch for spatial-temporal audible object grounding.

TaskType	Method	Spatial		Temporal	m_tIoU	m_vIoU	m_tF	m_vF
		Mask	BBox					
(a) AVS	TPAVI [61]	✓		✓	14.12	6.03	14.35	6.73
	ECMVAE [36]	✓		✓	13.01	5.24	13.86	5.37
(b) AVL	EZLSL [38]		✓	✓	13.01	6.93	-	-
	LVS [5]		✓	✓	11.65	6.03	-	-
	SLAVC [37]		✓	✓	11.85	6.20	-	-
	SSPL [40]		✓	✓	10.57	5.43	-	-
(c) AVE	AVEL [43]			✓	34.28	33.56	36.21	36.07
	CPSP [4]			✓	36.72	35.36	37.28	38.14
	JoMoLD [8]			✓	37.86	36.06	38.95	38.76
	CMPAE [15]			✓	40.67	40.10	42.86	41.87
(d) STVG	TubeDETR [47]		✓	✓	15.46	6.64	-	-
	STCAT [23]		✓	✓	15.76	7.01	-	-
(e) STAG	Ours (Mask-based)	✓		✓	18.76	17.32	17.33	16.25
	Ours (BBox-based)		✓	✓	16.53	15.89	-	-

m_vIoU. This implies that the audio signal obtains more emphasis in our framework. Moreover, our method also consistently outperforms other AVS methods across all metrics. Relative to AVL methods, our method (bounding box-based) demonstrates superior performance in both the m_vIoU and the m_tIoU metrics, surpassing the EZLSL by 8.96% and 3.52% respectively. This demonstrates our method is resilient to segments devoid of sound, while AVL methods exhibit diminished efficacy in datasets characterized by a substantial presence of such silent segments.

Compared with STVG methods, our method (bounding box-based) attains 15.89% on the metric m_vIoU, which is 8.88% higher than STCAT. This further illustrates that, compared to the transformer-based multi-modal interaction way in STVG methods, our explicit audio-visual correlation method enables sound to play a more significant guiding role. Since the AVE task exclusively focuses on temporal localization while omitting spatial localization, our method, along with others, scores lower in all metrics compared to AVE methods. This highlights the importance of achieving the balance between spatial and temporal localization in untrimmed videos.

5.3. Challenges Imposed by LU-AVS Dataset

Based on the above experimental results, we summarize the dataset challenges and adaptability of existing methods as follows: (1) For long videos in LU-AVS, the sounding duration and the start- and end-sounding time positions are uncertain. Therefore, both the spatial and temporal localization of audible objects are necessary for LU-AVS. Existing AVS methods developed based on the trimmed videos struggle to achieve temporal localization, showing limited adaptability in long videos. (2) Unlike trimmed videos that feature audible objects, untrimmed videos contain a high proportion of silent segments. Hence, the existing AVL methods trained on LU-AVS tend to overlook the audible objects. This suggests the requirement for greater emphasis on audio in model development on LU-AVS. (3) Similar to STVG, the exhaustive annotations in LU-AVS pose a high demand for achieving consistent spatial and temporal localization of audible objects, requiring methods to effectively joint model spatial, temporal, and audio-visual interactions.

6. Conclusion

In this work, we propose the first large-scale long-untrimmed AVS dataset. Our LU-AVS poses significant challenges in localizing audible objects at both the spatial and temporal dimensions in untrimmed videos, thus providing an ideal benchmark for developing practical AVS methods. Moreover, our LU-AVS dataset supports various tasks, like visual-audio localization, visual-audio grounding, and event localization since it provides diverse and comprehensive annotations. Extensive experiments demonstrate that there is significant space to improve AVS performance on long untrimmed videos.

Acknowledgements. This research is supported in part by the ARC-Discovery grant (DP220100800 to XY), and the ARC-DECRA grant (DE230100477 to XY). The first author is funded by the China Scholarship Council (CSC) and the CSIRO top-up (50092128). We thank all anonymous reviewers and ACs for their constructive suggestions.

References

[1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812, 2017. 3

[2] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE international conference on computer vision*, pages 609–617, 2017. 3

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers, 2020. 7

[4] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE, 2020. 3, 6, 7, 8

[5] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Localizing visual sounds the hard way. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16867–16876, 2021. 6, 7, 8

[6] Zhenfang Chen, Lin Ma, Wenhan Luo, and Kwan-Yee K Wong. Weakly-supervised spatio-temporally grounding natural sentence in video. *arXiv preprint arXiv:1906.02549*, 2019. 7

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

[8] Haoyue Cheng, Zhaoyang Liu, Hang Zhou, Chen Qian, Wayne Wu, and Limin Wang. Joint-modal label denoising for weakly-supervised audio-visual video parsing. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 6, 8

[9] Ho Kei Cheng and Alexander G. Schwing. XMem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *ECCV*, 2022. 2, 4

[10] Dorin Comaniciu and Peter Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 24(5):603–619, 2002. 5

[11] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. Transvg: End-to-end visual grounding with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1769–1779, 2021. 3

[12] Xiang Fang, Daizong Liu, Pan Zhou, and Guoshun Nan. You can ground earlier than see: An effective and efficient pipeline for temporal sentence grounding in compressed videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2448–2460, 2023. 3

[13] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275, 2017. 3

[14] Jiyang Gao, Zhenheng Yang, Kan Chen, Chen Sun, and Ram Nevatia. Turn tap: Temporal unit regression network for temporal action proposals. In *Proceedings of the IEEE international conference on computer vision*, pages 3628–3636, 2017. 7

[15] Junyu Gao, Mengyuan Chen, and Changsheng Xu. Collecting cross-modal presence-absence evidence for weakly-supervised audio-visual event perception. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 6, 8

[16] Lianli Gao, Zhao Guo, Hanwang Zhang, Xing Xu, and Heng Tao Shen. Video captioning with attention-based lstm and semantic consistency. *IEEE Transactions on Multimedia*, 19(9):2045–2055, 2017. 3

[17] Shengyi Gao, Zhe Chen, Guo Chen, Wenhai Wang, and Tong Lu. Avsegformer: Audio-visual segmentation with transformer. *arXiv preprint arXiv:2307.01146*, 2023. 2

[18] Dongliang He, Xiang Zhao, Jizhou Huang, Fu Li, Xiao Liu, and Shilei Wen. Read, watch, and move: Reinforcement learning for temporally grounding natural language descriptions in videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8393–8400, 2019. 3

[19] João F Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista. High-speed tracking with kernelized correlation filters. *IEEE transactions on pattern analysis and machine intelligence*, 37(3):583–596, 2014. 5

[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. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017. 6

[21] Binbin Huang, Dongze Lian, Weixin Luo, and Shenghua Gao. Look before you leap: Learning landmark features for one-stage visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16888–16897, 2021. 3

[22] Shaofei Huang, Han Li, Yuqing Wang, Hongji Zhu, Jiao Dai, Jizhong Han, Wenge Rong, and Si Liu. Discovering sounding objects by audio queries for audio visual segmentation. *arXiv preprint arXiv:2309.09501*, 2023. 2

[23] Yang Jin, Zehuan Yuan, Yadong Mu, et al. Embracing consistency: A one-stage approach for spatio-temporal video grounding. *Advances in Neural Information Processing Systems*, 35:29192–29204, 2022. 3, 6, 8

[24] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021. 3

[25] 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. *arXiv preprint arXiv:2304.02643*, 2023. 2, 4

[26] Juncheng Li, Junlin Xie, Long Qian, Linchao Zhu, Siliang Tang, Fei Wu, Yi Yang, Yueting Zhuang, and Xin Eric Wang. Compositional temporal grounding with structured variational cross-graph correspondence learning. In *Proceedings*

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3032–3041, 2022. 3

[27] Kexin Li, Zongxin Yang, Lei Chen, Yi Yang, and Jun Xiao. Catr: Combinatorial-dependence audio-queried transformer for audio-visual video segmentation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1485–1494, 2023. 2

[28] Mengze Li, Han Wang, Wenqiao Zhang, Jiaxu Miao, Zhou Zhao, Shengyu Zhang, Wei Ji, and Fei Wu. Winner: Weakly-supervised hierarchical decomposition and alignment for spatio-temporal video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23090–23099, 2023. 3

[29] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Flow-grounded spatial-temporal video prediction from still images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 600–615, 2018. 3

[30] Yue Liao, Si Liu, Guanbin Li, Fei Wang, Yanjie Chen, Chen Qian, and Bo Li. A real-time cross-modality correlation filtering method for referring expression comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10880–10889, 2020. 3

[31] Chen Liu, Peike Patrick Li, Xingqun Qi, Hu Zhang, Lincheng Li, Dadong Wang, and Xin Yu. Audio-visual segmentation by exploring cross-modal mutual semantics. In *Proceedings of the 31st ACM International Conference on Multimedia*, page 7590–7598, New York, NY, USA, 2023. Association for Computing Machinery. 2, 3

[32] Jingyu Liu, Liang Wang, and Ming-Hsuan Yang. Referring expression generation and comprehension via attributes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4856–4864, 2017. 3

[33] Xihui Liu, Zihao Wang, Jing Shao, Xiaogang Wang, and Hongsheng Li. Improving referring expression grounding with cross-modal attention-guided erasing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1950–1959, 2019. 3

[34] Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. Multi-task collaborative network for joint referring expression comprehension and segmentation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 10034–10043, 2020. 3

[35] Yuxin Mao, Jing Zhang, Mochu Xiang, Yunqiu Lv, Yiran Zhong, and Yuchao Dai. Contrastive conditional latent diffusion for audio-visual segmentation. *arXiv preprint arXiv:2307.16579*, 2023. 2

[36] Yuxin Mao, Jing Zhang, Mochu Xiang, Yiran Zhong, and Yuchao Dai. Multimodal variational auto-encoder based audio-visual segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 954–965, 2023. 2, 3, 6, 7, 8

[37] Shentong Mo and Pedro Morgado. A closer look at weakly-supervised audio-visual source localization. *Advances in Neural Information Processing Systems*, 35:37524–37536, 2022. 6, 7, 8

[38] Shentong Mo and Pedro Morgado. Localizing visual sounds the easy way. In *European Conference on Computer Vision*, pages 218–234. Springer, 2022. 6, 8

[39] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. A dataset for movie description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3202–3212, 2015. 3

[40] Zengjie Song, Yuxi Wang, Junsong Fan, Tieniu Tan, and Zhaoxiang Zhang. Self-supervised predictive learning: A negative-free method for sound source localization in visual scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3222–3231, 2022. 6, 8

[41] Rui Su, Qian Yu, and Dong Xu. Stvgbert: A visual-linguistic transformer based framework for spatio-temporal video grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1533–1542, 2021. 3

[42] Zongheng Tang, Yue Liao, Si Liu, Guanbin Li, Xiaojie Jin, Hongxu Jiang, Qian Yu, and Dong Xu. Human-centric spatio-temporal video grounding with visual transformers. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(12):8238–8249, 2021. 3

[43] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European conference on computer vision (ECCV)*, pages 247–263, 2018. 3, 6, 8

[44] Peng Wang, Qi Wu, Jiewei Cao, Chunhua Shen, Lianli Gao, and Anton van den Hengel. Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1960–1968, 2019. 3

[45] Zeyu Xiong, Daizong Liu, and Pan Zhou. Gaussian kernel-based cross modal network for spatio-temporal video grounding. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 2481–2485. IEEE, 2022. 3

[46] Masataka Yamaguchi, Kuniaki Saito, Yoshitaka Ushiku, and Tatsuya Harada. Spatio-temporal person retrieval via natural language queries. In *Proceedings of the IEEE international conference on computer vision*, pages 1453–1462, 2017. 3

[47] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Tubedetr: Spatio-temporal video grounding with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16442–16453, 2022. 3, 6, 8

[48] Sibei Yang, Guanbin Li, and Yizhou Yu. Dynamic graph attention for referring expression comprehension. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4644–4653, 2019. 3

[49] Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. A fast and accurate one-stage approach to visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4683–4693, 2019. 3

[50] Zhengyuan Yang, Tianlang Chen, Liwei Wang, and Jiebo Luo. Improving one-stage visual grounding by recursive sub-query construction. In *Computer Vision–ECCV 2020: 16th*

European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16, pages 387–404. Springer, 2020. 3

[51] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1307–1315, 2018. 3

[52] Chen Yuanhong, Liu Yuyuan, Wang Hu, Liu Fengbei, Wang Chong, and Carneiro Gustavo. A closer look at audio-visual semantic segmentation. *arXiv preprint arXiv:2304.02970*, 2023. 3, 5, 7

[53] Runhao Zeng, Haoming Xu, Wenbing Huang, Peihao Chen, Mingkui Tan, and Chuang Gan. Dense regression network for video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10287–10296, 2020. 3

[54] Hanwang Zhang, Yulei Niu, and Shih-Fu Chang. Grounding referring expressions in images by variational context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4158–4166, 2018.

[55] Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Span-based localizing network for natural language video localization. *arXiv preprint arXiv:2004.13931*, 2020.

[56] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12870–12877, 2020.

[57] Yimeng Zhang, Xin Chen, Jinghan Jia, Sijia Liu, and Ke Ding. Text-visual prompting for efficient 2d temporal video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14794–14804, 2023.

[58] Zhu Zhang, Zhou Zhao, Yang Zhao, Qi Wang, Huasheng Liu, and Lianli Gao. Where does it exist: Spatio-temporal video grounding for multi-form sentences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10668–10677, 2020. 3

[59] Jinxing Zhou, Liang Zheng, Yiran Zhong, Shijie Hao, and Meng Wang. Positive sample propagation along the audio-visual event line. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8436–8444, 2021. 6

[60] 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. 2, 3, 5, 7

[61] 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. *arXiv preprint arXiv:2301.13190*, 2023. 2, 3, 5, 6, 7, 8