Towards Long Form Audio-visual Video Understanding

Wenxuan Hou 1,† , Guangyao Li 1,† , Yapeng Tian 2 , Di Hu 1,*

¹Gaoling School of Artificial Intelligence, Renmin University of China ² Department of Computer Science, The University of Texas at Dallas

Abstract

We live in a world filled with never-ending streams of multimodal information. As a more natural recording of the real scenario, long form audio-visual videos are expected as an important bridge for better exploring and understanding the world. In this paper, we propose the multisensory temporal event localization task in long form videos and strive to tackle the associated challenges. To facilitate this study, we first collect a large-scale Long Form Audio-visual Video (LFAV) dataset with 5,175 videos and an average video length of 210 seconds. Each of the collected videos is elaborately annotated with diversified modality-aware events, in a long-range temporal sequence. We then propose an event-centric framework for localizing multisensory events as well as understanding their relations in long form videos. It includes three phases in different levels: snippet prediction phase to learn snippet features, event extraction phase to extract event-level features, and event interaction phase to study event relations. Experiments demonstrate that the proposed method, utilizing the new LFAV dataset, exhibits considerable effectiveness in localizing multiple modality-aware events within long form videos. Project website: http://gewu-1ab.github.io/LFAV/

1 Introduction

Guiding the machine to perceive and understand natural scenes like human beings is a long term vision of the AI community. As a primary means for recording natural scenes, video plays an important role in approaching the above prospect. Previous works on video understanding have achieved considerable performance in many tasks, including action recognition [43, 53, 15, 32], temporal action localization (TAL) [41, 63, 5, 65], weakly supervised temporal action localization (WTAL) [35, 40, 62, 23], audio-visual event localization (AVE) [49, 59, 70], audio-visual video parsing (AVVP) [48, 58, 27], etc. However, videos they studied are usually pre-processed (e.g. trimming long videos into short meaningful clips or ignoring other modalities such as audio), which could result in the potential deviation when depicting the real scenes.

Videos captured from natural scenes have two typical characteristics: 1) Long form. They usually span several minutes, covering multiple related events in different categories. These events usually jointly contribute to depicting the main content of the video. 2) Audio-visual. Videos recorded in real-world scenarios usually comprise both audio and visual modalities. These two aspects often exhibit asynchrony, providing unique perspectives in delineating the video content, yet collaboratively facilitating video understanding. Figure 1 illustrates a video example of *a badminton game*. This video is 121-second long, consisting of various audio events and visual events, they either only occur in one modality or occur in both modalities but have different temporal boundaries. These modality-aware events, as well as their inherent relations, help to effectively infer what happens in

¹{wxhou, guangyaoli, dihu}@ruc.edu.cn, ²{yapeng.tian}@utdallas.edu

^{*}Corresponding author. †Equal contribution.

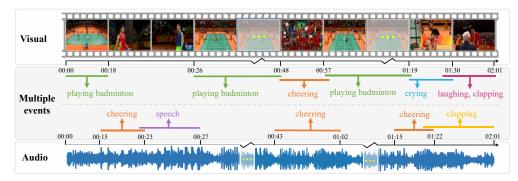


Figure 1: A long form audio-visual video example, with a length of 121 seconds. This video shows a badminton game. The audio modality contains three events: *cheering*, *clapping* and *speech*, the visual modality contains five events: *playing badminton*, *cheering*, *crying*, *laughing*, and *clapping*. The event *cheering* and *clapping* appears both in audio modality and visual modality. These multisensory events jointly facilitate the understanding of this video.

the video and then achieve a better understanding of the video content. Considering the merits of the above two characteristics, we propose to study video understanding in terms of long form and audio-visual aspect, name as *long form audio-visual video understanding*.

To achieve a better understanding of long form audio-visual videos, we propose to focus on the *multisensory temporal event localization* task, which essentially requires the model to predict the start and end time of each audio and visual event in the video. However, there remain several challenges when addressing this task. Firstly, the video contains multiple events with diverse categories, modalities, and varying lengths. Secondly, understanding the video content requires effectively modeling long-range dependencies and relations across different clips and modalities. To study the above new task, we elaborately build a large-scale Long Form Audio-visual Video (LFAV) dataset with 5,175 videos, as existing datasets are not appropriate for our proposed task. Specifically, existing datasets either just localize audio-visual events (*i.e.*, events that are both audible and visiable [49]) or only localize events in short trimmed videos. We annotate 24,875 modality-aware event labels on video-level in total. For validation and testing sets, we annotate the category and temporal boundaries of 23,666 audio and visual events. The total length, average length, and average event categories of the videos in LFAV are 302-hour, 210-second, and 3.15, respectively. We expect the LFAV dataset could facilitate the study of our proposed task in the long form audio-visual context.

To address the above challenges, we propose an event-centric framework containing three phases from snippet prediction², event extraction to event interaction. Firstly, we propose a pyramid multimodal transformer model to learn snippet-level features by executing intra-modal and cross-modal interaction within multiscale temporal windows. Secondly, we extract event-level features by refining and aggregating event-aware snippet features in structured graphs. At last, we study event relations by modeling the influence among multiple audio and visual events, then refining the event features. The three phases are jointly optimized with video-level event labels in an end-to-end fashion. Extensive experimental results show that our event-centric framework can achieve effective multisensory temporal event localization, enhancing understanding of long form audio-visual videos and advancing realistic scene perception. To summarize, our contributions are threefold:

- We direct that long form and audio-visual are two key characteristics of videos from natural scenes, then propose to explore long form audio-visual video understanding by concentrating on the proposed multisensory temporal event localization task.
- We collect a large-scale long form audio-visual video dataset, named LFAV to facilitate our study, which contains 5,175 videos with an average length of above 210 seconds, average event categories of 3.15 per video, and modality-aware annotations.
- We propose an event-centric framework to tackle multisensory temporal event localization. Experimental results show that our framework obviously surpasses comparison methods, which indicates the effectiveness of the proposed event-centric framework.

²As usual, a snippet represents a 1-second long video segment[48].

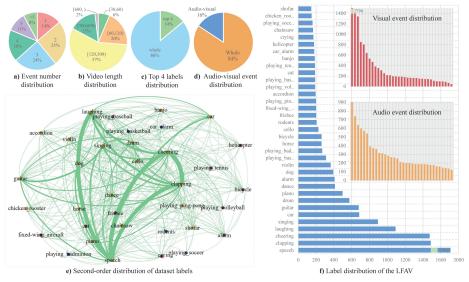


Figure 2: **Illustrations of our LFAV dataset statistics**. (a-d) Statistical analysis of label categories, including the distribution of event numbers in each video; the distribution of video length; the proportion of the top 4 event categories, the *top 4 labels* represent *speech*, *clapping*, *cheering*, and *laughing*, which are the most common human actions; the temporal proportion of events that occur on two modalities at the same time. (e) Second-order interactions between all labels, the thicker the line, the closer the association. (f) Distribution of dataset labels of each category.

2 Related Work

Long Form Video Understanding. The human vision system is one of the most important bridge for us to perceiving the world [34]. Efforts have been made to understand visual content effectively [11, 21, 4, 33] and efficiently [38, 32]. However, they mainly concentrate on images and short video clips, which struggle to accurately reflect the full picture of the ever-changing real world. In contrast, long form videos provide richer information, enabling the learning of event relations and long-range temporal dependencies. In recent years, several benchmarks of long form videos have been proposed [56, 12, 47, 68, 44, 67]. Their task definitions vary, but all require a comprehensive understanding of the entire long-form video. Many existing works that also have explored understanding long form videos in multiple views, including improving model architectures [41, 52, 39, 61, 57, 66], learning key clues or regions in the video [16, 56], aligning cross-modal information [46], etc. However, previous long-form video benchmarks and learning methods have largely ignored the audio signal, which can provide valuable or visually invisible information. Ignoring the audio channel can lead to an incomplete or biased understanding of long-form videos.

Audio-Visual Video Understanding. Inspired by the multisensory perception of humans, the community has paid more and more attention to audio-visual scene understanding in recent years. Existing works mainly contain audio-visual action recognition [28, 60, 17, 7], audio-visual question answering[1, 64, 31], egocentric audio-visual object localization [22], audio-visual segmentation [69], etc. Recently, the fine-grained temporal event localization task was introduced in the audio-visual learning community, including AVE [49, 59, 70, 2], AVVP [48, 58, 27], and denselocalizing audio-visual events [19]. These works aim to temporally recognize events in audio-visual videos. However, the AVE task and the dense-localizing audio-visual events task focus exclusively on events that are both audible and visible, overlook the prevalence of modality-aware events in real-world scenes. The AVVP task takes modality-aware events into account, but just localizes events in 10-second short trimmed clips. Besides these audio-visual event localization works, some other works also both take audio and visual modalities into account [2, 44, 24], but they mainly aim to boost the uni-modal learning under the assistance of another modality. These works are somewhat lack of the exploration of complex relations among multiple modality-aware events as well as long-range dependencies, hindering a complete and realistic understanding of natural scenarios. Compared with these previous works, our multisensory temporal event localization task is more challenging and closer to the real scene, in terms of the properties of long-range and modality-aware events.

Table 1: **Comparison with other datasets.** Our LFAV dataset is collected for the proposed multisensory temporal event localization task, where diversified domains are covered. Specifically, the LFAV dataset offers modality-aware annotations for each video, that is it points out the events are from audio, visual, or both modalities. Meanwhile, multiple events with different semantic categories per video are also annotated for better exploring the relation among events. Videos in the dataset have an average length of 210 seconds and a total length of 302 hours. *: LLP only provides modality-aware annotations in validation and testing sets.

Dataset	Domain	Year	Modality aware annotations	Multiple event annotations	Avg Length (sec.)	Total hours
GTEA [14]	Human	CVPR'11	Х	✓	130	0.58
50 Salads [45]	Human	UbiComp'13	X	✓	324	4.5
Breakfast [30]	Kitchens	CVPR'14	X	✓	162	77
ActivityNet [12]	Sports	CVPR'15	X	Х	113	648
Charades [42]	Indoor activities	ECCV'16	X	✓	30	82
EPIC-KITCHENS-100 [10]	Kitchens	IJCV'22	×	\checkmark	514	100
Long-Form VQA [68]	Sports	TIP'19	Х	Х	128	849
LVU [56]	Movies	CVPR'21	X	Х	< 180	N/A
LOGO [67]	Sports	CVPR'23	×	\checkmark	204.2	11.3
AVE [49]	Sound event	ECCV'18	Х	Х	10	11
VGGSound [6]	In the wild	ICASSP'20	X	Х	10	560
LLP [48]	Sound event	ECCV'20	X *	✓	10	32.9
UnAV-100 [19]	In the wild	CVPR'23	X	✓	42.1	126.2
LFAV (Ours)	Sports,human, instruments, etc.	-	✓	✓	210	302

3 The LFAV Dataset

3.1 Overview

Towards long form audio-visual video understanding, we build a large-scale audio-visual video dataset, named LFAV. As noted above, high-quality datasets are of considerable value for audio-visual video understanding research. The built LFAV contains 5,175 untrimmed YouTube videos spanning over 35 categories. A wide range of events (e.g., sing, crying, playing badminton and chainsaw etc.) from diverse domains (e.g., human activities, tools, instrument, sports, and traffic etc.) are included.

As shown in Tab. 1, compared to existing related datasets, our built LFAV dataset has the following advantages: 1) Videos in the LFAV dataset are usually several minutes long, the value of long-range dependencies and relations among multiple events with different lengths can be explored adequately. In contrast, most previous audio-visual datasets [49, 48, 6] are built with trimmed videos with only 10-second long, which are limited in exploring and utilizing the above properties of long form videos. 2) The LFAV dataset contains videos with multisensory events and modality-aware annotations, which provides the possibility to explore the influence and associations between audio and visual events. Previous video understanding datasets [14, 10, 12, 56] mainly focus on visual content perception. Although some of them, such as ActivityNet [12], EPIC-KITCHENS-100 [10], as well as some audio-visual datasets [49, 19], also offer audio recordings, they do not provide audio-level annotations. More seriously, the audio information is sometimes accompanied by severe noise (*e.g.* background music). In contrast, our LFAV dataset avoids the above issues and thus better supports the study on audio-visual video understanding.

3.2 Video Collection

We collect 5,175 videos from YouTube, covering five kinds of daily life to ensure the diversity, complexity, and dynamic of the real world: human-related, sports, musical instruments, tools, and animals. We also construct a label set of 35 kinds of events covering the above scenes, as shown in Tab. 2. To keep the balance of different labels, we design the following three collecting steps:

- 1) Video retrieval. The rule of this step is that each label is required to combine with one or more labels in different categories. Specifically, we use permutation and combination methods for 35 labels to ensure that all label combinations can be covered in the video as much as possible.
- 2) Video filtering. Video collectors watch the entire video under the condition of sound playback to select high-quality videos. Low-quality videos, such as videos with audio noise, or excessively

Table 2: List of 35 label categories, which belong to 5 different kinds of daily life scenes.

Daily life scene	Label categories
Musical instruments	guitar, drum, violin, piano, accordion, banjo, cello, shofar
Human-related	speech, singing, crying, laughing, clapping, cheering, dance
Animals	dog, cat, chicken rooster, horse, rodents
Traffic and tools	car, helicopter, fixed-wing aircraft, bicycle, alarm, chainsaw, car alarm
Cnarta	playing basketball, playing badminton, playing volleyball,
Sports	playing tennis, playing ping-pong, frisbee, playing soccer

trimmed videos, will not be selected to construct the dataset. In addition, collectors also need to record the categories appearing in both audio and visual modalities.

3) Label distribution regulation. To avoid seriously long-tailed distributions, we need to control the number of each event category. We count the interaction between different event categories when we have collected a certain amount of videos. Then the number of each event category and the confusion matrix among different categories will be used to guide further data collection.

3.3 Annotations

For the collected daily life videos, the event annotations contain two parts: **video-level** annotations (annotate audio event categories and visual event categories that exist in the video) and **event-level** annotations (annotate temporal boundaries of events in the video.). For most videos in the LFAV dataset (3,721 videos), we just annotate them at video-level, these videos are used as the training set. For a small part of the videos (1,454 videos), we both annotate them at video-level and event-level, these videos are used as the validation set and testing set.

To ensure the quality of video annotations and improve the reliability of evaluation results, we verify all annotations manually from two aspects: 1) All annotations are randomly assigned to annotators for one-to-one proofreading to check whether existing misplaced annotations, missing annotations, and videos with excessive noise, *etc.* 2) Spot-check the verified annotation information via sampling inspection. After verification, videos with wrong annotations will be re-annotated by other annotators, and videos with excessive noise will be directly deleted.

3.4 Statistical Analysis

Our LFAV dataset contains 5,175 videos spanning over 35 categories for over 302 hours. Fig. 2(a-d) provides the statistical analysis of our dataset. In this dataset, more than half of the videos contain at least three event categories, indicating that there are widely diverse categories of events present in long form audio-visual videos. All the videos are longer than 30-second and each of them has audio or visual events of at least 1 second. 74% of the videos are longer than two minutes, and 2% of the videos are even longer than ten minutes. Fig. 2(e) shows the interaction situation among all the labels, where the thicker the line, the more the interaction. Fig. 2(f) shows the number of label categories, and the occurrences of each category are no less than 146. Before annotations, we randomly split the dataset into training, validation, and testing sets with 3,721, 486, and 968 videos, respectively. Finally, we have 24,875 video-level event annotations on the whole dataset, and 23,666 second-level event annotations on validation and testing sets in total. More details about the statistical analysis of the LFAV dataset are in Sec. B of the Supp. Materials.

4 The Multisensory Temporal Event Localization Task

Task Definition. The multisensory temporal event localization task in long form audio-visual videos aims at precisely localizing modality-aware events in videos with several minutes long. Some previous works on weakly supervised temporal action localization divide the video into several non-overlapping snippets, then snippets are classified by multiple instance learning and aggregated to events [52, 62]. We follow this paradigm to define the task, the output of the task is event categories of all snippets. Then event-level predictions can be directly generated by concatenating consecutive snippets with the same class of events. Concretely, for a long form video with a length of T-second, we first divide it into T non-overlapping audio snippets $\{A_t\}_{t=1}^T$ and visual snippets $\{V_t\}_{t=1}^T$, each audio and visual snippet is 1-second long and T is usually up to several hundred. The goal of the task

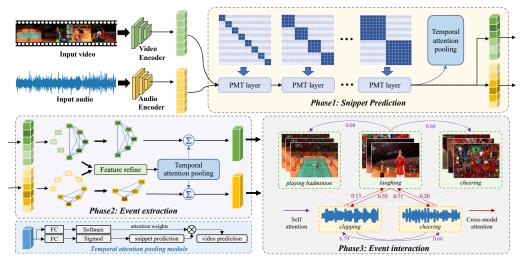


Figure 3: **Our event-centric framework. Top:** In the first phase of snippet prediction, we propose a pyramid multimodal transformer to generate the snippet features as well as their category prediction. **Middle left:** In the second phase of event extraction, we build an event-aware graph to refine the snippet features and then aggregate the event-aware snippet features into event features. **Bottom right:** In the third phase of event interaction, we model the event relations in both intra-modal and cross-modal scenarios and then refine the event feature by referring to its relation to other events. **Bottom left:** The architecture of temporal attention pooling.

is to generate snippet-level event predictions $\{p_t^a\}_{t=1}^T$ and $\{p_t^v\}_{t=1}^T$, where $p_t^a \in R^C$ and $p_t^v \in R^C$, they are audio and visual predictions of the t-th snippets, respectively. C is the number of categories. The audio event labels $y^a \in \{0,1\}^C$ and visual event labels $y^v \in \{0,1\}^C$ are both available during training, they are multi-label binary vectors that indicate the categories of contained modality-aware events in video-level but without snippet-level timestamps.

Evaluation Metrics. To achieve a more comprehensive evaluation, we use two metrics to evaluate snippet-level and event-level performance, respectively. For snippet-level, we use mAP as the evaluation metric because snippet-level prediction can be regarded as a multi-label classification task [55, 9, 8] for each snippet. For event-level, we follow [48] and use F1-score as the evaluation metric. We do not use mAP as the event-level evaluation metric, as our framework is not an event proposal-based method. We set 4 sub-metric for each evaluation metric to evaluate the performance of different types of events and their average value, as shown in Tab. 3. For example, the audio sub-metric evaluates the localization performance of all events in the audio modality.

Task Challenges. Our proposed task takes several distinct challenges that should be addressed. First, a long form audio-visual video usually contains multiple events with various semantic categories, varies wildly in temporal length, and is depicted in huge different modalities. These issues practically play as barriers to achieving effective modeling of video content. Second, modeling the inherent long-range dependencies is a key to understanding the full picture of a long video, but becoming more difficult when the scene in the long video changes dynamically or the audio modality and visual modality influence each other. Hence, a hit-to-the-point video understanding method needs to be proposed to step across these challenges.

5 Method

In this section, we propose an event-centric framework to solve the above challenges and aim to achieve a better understanding of long form audio-visual videos. Our framework contains three phases, *i.e.*, snippet prediction in Sec. 5.1, event extraction in Sec. 5.2, and event interaction in Sec. 5.3, as shown in Fig. 3. We will give a concrete introduction in this section. Due to the limitation of space, more details of the proposed framework are in the Sec. C of the *Supp. Materials*.

We use the pre-trained VGGish [18] model to extract audio features for each snippet, and use the pre-trained ResNet18 [21] and R(2+1)D-18 [50] models to extract visual features for each snippet. The audio and visual feature are represented as $\{a_t\}_{t=1}^T$ and $\{v_t\}_{t=1}^T$, respectively.

5.1 Snippet Prediction

In the proposed event-centric framework, we first focus on learning effective snippet features as the precondition of event localization. Concretely, we propose to capture the contained audio and visual events with different temporal lengths via pyramid window modules, and learn the interaction among the snippets of both modalities by multimodal attention module. These two modules constitute the *Pyramid Multimodal Transformer* (PMT), as shown in Fig. 4.

Pyramid window module. Inspired by the multiscale technique [20, 54], we use pyramid windows to capture events with different temporal scales then learn snippet features only within specific windows. Concretely, for the l-th PMT layer with a window size of 2^l , the snippet features within the window are interacted with multimodal attention, then the outputs are used as the inputs of snippet interaction in the (l+1)-th layer, but with a window size of 2^{l+1} , as the shown in Phase 1 of Fig. 3. In practice, we set six layers in the PMT, thus the window size ranges from 2 to 64.

Multimodal attention. To learn the interaction among snippets from the same or different modalities, we perform multimodal attention in each pyramid window. As shown in Fig. 4, each multimodal attention module contains four attention units, two of them are for audio and visual modeling, while the other two are for crossmodal interaction at the snippet level. Then, the updated snippet features are used as the inputs for the next layer.

To avoid the potential partition of an event caused by the pyramid window module³, we also propose a kind of snippet shift strategy to capture the events across different pyramid windows. As shown in Fig. 4, we shift the top $2^l/2$ snippet features in the first window to the end of the video after the first multimodal attention module, then feed the updated snippet sequence to the second attention module. At the end of

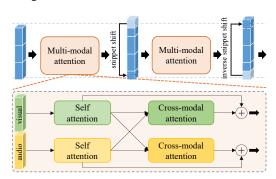


Figure 4: Architecture of the PMT layer. Top: The architecture of one PMT layer, including two multimodal attention modules, a snippet shift operation, and an inverse snippet shift operation; Bottom: The detailed architecture of an attention module, including two self-attention units and two cross-modal attention units.

the PMT layer, we restore the temporal order of snippets and obtain the updated snippet features. With the stacked PMT layers, we could learn effective representation of audio and visual snippets, by considering the event with different temporal lengths and multimodal information. Then, we use a *Temporal Attention Pooling* (TAP⁴) module to obtain the video-level and snippet-level prediction, its architecture is shown in the bottom left of Fig. 3. The video-level prediction with binary cross-entropy loss is used for training this phase.

5.2 Event Extraction

For a long form audio-visual video, modeling its long-range dependency only at the snippet-level could not be effective because of its complex dynamic scenes and modality interaction. Hence, we propose to model the long-range dependency by excavating the unimodal events then exploring their inherent relations in long videos. In this section, we extract event features based on the preliminary snippet features and their category predictions from the snippet prediction phase, as Fig. 3 (*Phase 2*).

An event usually consists of several snippets which are semantically related, but the remaining amounts of snippets could be in different semantics or backgrounds, especially for an extremely long snippet sequence. Hence, to obtain high-quality event representation as well as address the challenges in modeling the long-range dependency, we use an event-aware graph structure to model the long form video then refine event-aware snippet features. Concretely, we construct a graph for each event category of each modality, where each snippet is a node of the graph and the nodes are connected according to two kinds of edges: temporal edges and semantic edges. For temporal edges, we connect every two adjacent nodes, because successive snippets are usually considered to be similar

³Same-event snippets in different pyramid windows cannot fully interact with each other.

⁴More details about TAP are in Sec. C.1 of the Supp. Materials.

Table 3: Experiment results of our framework and comparison methods. Ours (S) means our framework with only the snippet prediction phase. Ours (S+E) means our framework with the snippet prediction and event extraction phase. Ours (All) means our whole framework with all three phases. Avg. indicates the average value of the results in audio, visual, and audio-visual.

Method	F	F1-score (%) (event-level)			mAP (%) (snippet-level)			
Method	Audio	Visual	Audio-Visual	Avg.	Audio	Visual	Audio-Visual	Avg.
STPN [35]	7.92	11.02	10.47	9.80	28.30	48.87	38.66	43.02
RSKP [23]	7.64	8.39	11.79	9.27	29.42	47.43	22.25	33.03
LongFormer [3]	14.16	15.92	13.40	14.49	37.11	46.03	35.12	39.42
Transformer-LS [71]	16.55	22.41	17.87	18.94	37.91	50.10	37.77	41.92
ActionFormer [66]	13.14	18.33	13.46	14.97	28.40	43.09	33.65	35.05
AVE [49]	16.42	11.97	13.71	13.71	42.41	47.44	39.84	43.23
AVSlowFast [60]	19.68	20.16	15.67	18.50	50.03	62.50	48.41	53.65
HAN [48]	20.96	24.13	18.07	20.87	48.27	63.26	47.42	52.98
PSP [70]	17.40	26.32	15.60	19.78	37.27	60.95	45.84	48.01
DHHN [27]	21.44	18.82	14.23	18.16	49.95	59.59	47.74	52.42
SlowFast [15]	19.42	19.72	14.94	18.03	47.46	62.18	46.75	52.13
MViT [13]	18.60	27.68	17.67	21.31	47.48	64.55	48.37	53.47
MeMViT [57]	21.62	29.21	21.81	24.22	46.37	64.36	47.67	52.80
Ours (S)	24.81	30.26	23.06	26.04	48.76	62.26	47.00	52.67
Ours (S+E)	25.76	31.46	25.36	27.53	49.31	63.38	47.97	53.55
Ours (All)	25.36	32.44	26.61	28.14	51.66	64.10	50.11	55.29

in semantics. For semantic edges, the nodes are fully connected when they all have high confidence in the same event category. The confidence comes from the snippet-level predictions in the first phase.

Based on the built event-aware graph, we propose to refine the event-aware snippet features via a graph attention based model. Note that all the graphs share the same weights, either the kinds of modalities or the event categories they belong to. Then, based on the refined event-aware snippet features, we employ the TAP module to obtain their importance (*i.e.*, attention weight) of the specific event, then perform weighted aggregation over the snippet features belonging to the same event to generate the final event feature. Like the snippet prediction phase, we also use the binary cross-entropy loss to train this phase based on the prediction of the TAP module.

5.3 Event Interaction

As mentioned above, learning the relations among events is crucial in understanding long form videos, especially when faced with the requirements of modeling long-range dependency and different modalities. Considering these events are not independent with each other, we propose to learn the event relations by interacting events from two aspects, including intra-modal event interaction and cross-modal event interaction. Based on the extracted event features in the second phase, event-aware self-attention and cross-modal attention at video-level are performed to explore the potential event relations, which are accordingly used to refine the event features. We then perform an event-level binary cross-entropy loss over the refined event features, under the video-level category label.

By cascading the introduced three phases, we could achieve an event-centric video understanding framework. These three phases are jointly trained end-to-end, and the third phase of event interaction is not considered during inference. More details about the method are in Sec.C of the *Supp. Materials*.

6 Experiments

6.1 Experimental Settings

Implementation details. We set batch size to 16 and use Adam to optimize the network, the initial learning rate of the three phases are 1e-4, 1e-4, and 2e-4, respectively. We train the model for 30 epochs, and the learning rate is reduced by a factor of 0.1 for every 10 epochs. The number of snippets of all videos is adjusted to 200 for effective training.

Comparison methods. To validate the superiority of our proposed framework, we choose 13 related methods for comparison, including weakly supervised temporal action localization methods: STPN [35], RSKP [23]; long sequence modeling methods: Longformer [3], Transformer-LS [71],

Table 4: Ablation study on feature interaction, where S-att represents self attention and C-att denotes cross-modal attention.

Snippe	t interaction	F1-score (%)	mAP (%)		
S-att	C-att	Avg.	Avg.		
X	Х	9.66	38.71		
\checkmark	X	21.45	48.63		
X	\checkmark	17.45	51.85		
\checkmark	\checkmark	26.04	52.67		

Table 5: Ablation study on event interaction, where S-att and C-att denote self and cross-modal attention, respectively.

Event interaction		F1-score (%)	mAP (%)
S-att	C-att	Avg.	Avg.
×	Х	27.52	54.19
\checkmark	×	27.65	54.79
X	\checkmark	28.07	54.27
\checkmark	\checkmark	28.14	55.29

ActionFormer [66]; audio-visual learning methods: AVE [49], AVSlowFast [60], HAN [48], PSP [70], DHHN [27]; video classification methods: SlowFast [15], MViT [13], and MeMViT [57].

6.2 Results and Analysis

Comparison to other methods. Experimental results of comparison methods and our methods are shown in Tab. 3. There are three points we could pay attention to. Firstly, temporal action localization [35, 23] and long sequence modeling methods [3, 71, 66] aim to effectively localize action events in untrimmed videos or model long sequences. But they ignore the valuable cooperation among audio and video modality, which is important in achieving more comprehensive video event understanding. Secondly, although some methods [49, 70, 48, 27] take the audio signal into account, they are consistently worse than our method. This could be because they mainly aim at understanding trimmed short videos, resulting in limited modeling of long-range dependencies and event interactions. Thirdly, our proposed method outperforms all the comparison ones obviously, although some recent video classification methods [13, 57] achieve slightly better results on visual mAP, their overall performance still lags obviously behind our proposed method, showing that our proposed event-centric framework can localize both audio and visual events in long form audio-visual videos better. Additionally, we notice that localizing audio-visual events is more challenging because it requires precise localization in both modalities.

Effectiveness of three phases. As mentioned above, our full method consists of three progressive phases. The performance of the snippet prediction phase has already surpassed most comparison methods, then the subsequent phases can further improve localization performance. Results are shown in the last three rows of Tab. 3, which indicate the potential importance of decoupling a long form audio-visual video into multiple uni-modal events with different lengths and modeling their inherent relations in both uni-modal and cross-modal scenarios.

Effectiveness of feature interaction. We further explore the effectiveness of snippet-level and event-level feature interaction. Results are shown in Tab. 4 and Tab. 5, respectively. We can get two conclusions from these results: 1) Both kinds of interactions benefit from intra-modal and cross-modal attention, which not only indicates the importance of the uni-modal sequence but also shows the meaning of cross-modal relations in achieving effective long form audio-visual video understanding. 2) The single snippet prediction phase without any feature interaction can be viewed as a normal classification method, whose results are in the third row of Tab. 4. Compared with our full model, its F1-score drop and mAP drop are 18.48% and 16.58%, respectively. This huge performance drop indicates that the inherent long-range dependency and cross-modal relations are crucial to solving our task. More experimental results and analysis are in Sec. D.3 and D.4 of the *Supp. Materials*.

7 Conclusion

In this paper, we pose and tackle a challenging multisensory temporal event localization task in long form videos. We collect the LFAV dataset to facilitate our research and propose an event-centric framework to solve the task. Experimental results demonstrate that capturing multiple audio and visual events with different temporal lengths and modeling long-range dependencies by event interaction is crucial for the problem. We hope our work could be a meaningful exploration towards more natural machine perception and bring some inspiration to audio-visual video understanding, *e.g.*, audio-visual video dense captioning, reasoning over scene dynamics, *etc*.

References

- [1] Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K Marks, Chiori Hori, Peter Anderson, et al. Audio visual scene-aware dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7558–7567, 2019.
- [2] Anurag Bagchi, Jazib Mahmood, Dolton Fernandes, and Ravi Kiran Sarvadevabhatla. Hear me out: Fusional approaches for audio augmented temporal action localization. *arXiv preprint arXiv:2106.14118*, 2021.
- [3] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv* preprint arXiv:2004.05150, 2020.
- [4] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 6299–6308, 2017
- [5] Yu-Wei Chao, Sudheendra Vijayanarasimhan, Bryan Seybold, David A Ross, Jia Deng, and Rahul Sukthankar. Rethinking the faster r-cnn architecture for temporal action localization. In *proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1130–1139, 2018.
- [6] 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.
- [7] Jiawei Chen and Chiu Man Ho. Mm-vit: Multi-modal video transformer for compressed video action recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1910–1921, 2022.
- [8] Tianshui Chen, Muxin Xu, Xiaolu Hui, Hefeng Wu, and Liang Lin. Learning semantic-specific graph representation for multi-label image recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 522–531, 2019.
- [9] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5177–5186, 2019.
- [10] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, 130(1):33–55, 2022.
- [11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pages 248–255. Ieee. 2009.
- [12] Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970, 2015.
- [13] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6824–6835, 2021.
- [14] Alireza Fathi, Xiaofeng Ren, and James M Rehg. Learning to recognize objects in egocentric activities. In *CVPR 2011*, pages 3281–3288. IEEE, 2011.
- [15] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In Proceedings of the IEEE/CVF international conference on computer vision, pages 6202–6211, 2019.
- [16] Edward Fish, Jon Weinbren, and Andrew Gilbert. Two-stream transformer architecture for long video understanding. *arXiv* preprint arXiv:2208.01753, 2022.
- [17] Ruohan Gao, Tae-Hyun Oh, Kristen Grauman, and Lorenzo Torresani. Listen to look: Action recognition by previewing audio. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10457–10467, 2020.
- [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 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP), pages 776–780. IEEE, 2017.
- [19] Tiantian Geng, Teng Wang, Jinming Duan, Runmin Cong, and Feng Zheng. Dense-localizing audio-visual events in untrimmed videos: A large-scale benchmark and baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22942–22951, 2023.
- [20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9):1904–1916, 2015.
- [21] 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.
- [22] Chao Huang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. Egocentric audio-visual object localization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 22910–22921, 2023.

- [23] Linjiang Huang, Liang Wang, and Hongsheng Li. Weakly supervised temporal action localization via representative snippet knowledge propagation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3272–3281, 2022.
- [24] Jaesung Huh, Jacob Chalk, Evangelos Kazakos, Dima Damen, and Andrew Zisserman. EPIC-SOUNDS: A Large-Scale Dataset of Actions that Sound. In *IEEE International Conference on Acoustics, Speech, & Signal Processing (ICASSP)*, 2023.
- [25] Vladimir Iashin and Esa Rahtu. A better use of audio-visual cues: Dense video captioning with bi-modal transformer. *arXiv preprint arXiv:2005.08271*, 2020.
- [26] Vladimir Iashin and Esa Rahtu. Multi-modal dense video captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pages 958–959, 2020.
- [27] Xun Jiang, Xing Xu, Zhiguo Chen, Jingran Zhang, Jingkuan Song, Fumin Shen, Huimin Lu, and Heng Tao Shen. Dhhn: Dual hierarchical hybrid network for weakly-supervised audio-visual video parsing. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 719–727, 2022.
- [28] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5492–5501, 2019.
- [29] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In Proceedings of the IEEE international conference on computer vision, pages 706–715, 2017.
- [30] Hilde Kuehne, Ali Arslan, and Thomas Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 780–787, 2014.
- [31] Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19108–19118, 2022.
- [32] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7083–7093, 2019.
- [33] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022.
- [34] David Marr. Vision: A computational investigation into the human representation and processing of visual information. MIT press, 2010.
- [35] Phuc Nguyen, Ting Liu, Gautam Prasad, and Bohyung Han. Weakly supervised action localization by sparse temporal pooling network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6752–6761, 2018.
- [36] Tanzila Rahman, Bicheng Xu, and Leonid Sigal. Watch, listen and tell: Multi-modal weakly supervised dense event captioning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8908–8917, 2019.
- [37] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
- [38] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.
 [39] Zheng Shou, Jonathan Chan, Alireza Zareian, Kazuyuki Miyazawa, and Shih-Fu Chang. Cdc:
- [39] Zheng Shou, Jonathan Chan, Alireza Zareian, Kazuyuki Miyazawa, and Shih-Fu Chang. Cdc: Convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 5734–5743, 2017.
- [40] Zheng Shou, Hang Gao, Lei Zhang, Kazuyuki Miyazawa, and Shih-Fu Chang. Autoloc: Weakly-supervised temporal action localization in untrimmed videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 154–171, 2018.
- [41] Zheng Shou, Dongang Wang, and Shih-Fu Chang. Temporal action localization in untrimmed videos via multi-stage cnns. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1049–1058, 2016.
- [42] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*, pages 510–526. Springer, 2016.
- [43] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27, 2014.
- [44] Mattia Soldan, Alejandro Pardo, Juan León Alcázar, Fabian Caba, Chen Zhao, Silvio Giancola, and Bernard Ghanem. Mad: A scalable dataset for language grounding in videos from movie audio descriptions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5026–5035, 2022.
- [45] Sebastian Stein and Stephen J McKenna. Combining embedded accelerometers with computer vision for recognizing food preparation activities. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pages 729–738, 2013.
- [46] Yuchong Sun, Hongwei Xue, Ruihua Song, Bei Liu, Huan Yang, and Jianlong Fu. Long-form videolanguage pre-training with multimodal temporal contrastive learning. *arXiv preprint arXiv:2210.06031*, 2022.

- [47] Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. Coin: A large-scale dataset for comprehensive instructional video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1207–1216, 2019.
- [48] Yapeng Tian, Dingzeyu Li, and Chenliang Xu. Unified multisensory perception: Weakly-supervised audio-visual video parsing. In *European Conference on Computer Vision*, pages 436–454. Springer, 2020.
- [49] 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.
- [50] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.
- [51] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.
- [52] Limin Wang, Yuanjun Xiong, Dahua Lin, and Luc Van Gool. Untrimmednets for weakly supervised action recognition and detection. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 4325–4334, 2017.
- [53] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016.
- [54] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 568–578, 2021.
- [55] Yunchao Wei, Wei Xia, Min Lin, Junshi Huang, Bingbing Ni, Jian Dong, Yao Zhao, and Shuicheng Yan. Hcp: A flexible cnn framework for multi-label image classification. *IEEE transactions on pattern analysis and machine intelligence*, 38(9):1901–1907, 2015.
- [56] Chao-Yuan Wu and Philipp Krahenbuhl. Towards long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1884–1894, 2021.
- [57] Chao-Yuan Wu, Yanghao Li, Karttikeya Mangalam, Haoqi Fan, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13587–13597, 2022.
- [58] Yu Wu and Yi Yang. Exploring heterogeneous clues for weakly-supervised audio-visual video parsing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1326–1335, 2021
- [59] Yu Wu, Linchao Zhu, Yan Yan, and Yi Yang. Dual attention matching for audio-visual event localization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6292–6300, 2019.
- [60] Fanyi Xiao, Yong Jae Lee, Kristen Grauman, Jitendra Malik, and Christoph Feichtenhofer. Audiovisual slowfast networks for video recognition. *arXiv preprint arXiv:2001.08740*, 2020.
- [61] Mingze Xu, Yuanjun Xiong, Hao Chen, Xinyu Li, Wei Xia, Zhuowen Tu, and Stefano Soatto. Long short-term transformer for online action detection. Advances in Neural Information Processing Systems, 34:1086–1099, 2021.
- [62] Zichen Yang, Jie Qin, and Di Huang. Acgnet: Action complement graph network for weakly-supervised temporal action localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3090–3098, 2022.
- [63] Jun Yuan, Bingbing Ni, Xiaokang Yang, and Ashraf A Kassim. Temporal action localization with pyramid of score distribution features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3093–3102, 2016.
- [64] Heeseung Yun, Youngjae Yu, Wonsuk Yang, Kangil Lee, and Gunhee Kim. Pano-avqa: Grounded audiovisual question answering on 360deg videos. In *Proceedings of the IEEE/CVF International Conference* on Computer Vision, pages 2031–2041, 2021.
- [65] Runhao Zeng, Wenbing Huang, Mingkui Tan, Yu Rong, Peilin Zhao, Junzhou Huang, and Chuang Gan. Graph convolutional networks for temporal action localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7094–7103, 2019.
- [66] Chen-Lin Zhang, Jianxin Wu, and Yin Li. Actionformer: Localizing moments of actions with transformers. In *European Conference on Computer Vision*, pages 492–510. Springer, 2022.
- [67] Shiyi Zhang, Wenxun Dai, Sujia Wang, Xiangwei Shen, Jiwen Lu, Jie Zhou, and Yansong Tang. Logo: A long-form video dataset for group action quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2405–2414, 2023.
- [68] Zhou Zhao, Zhu Zhang, Shuwen Xiao, Zhenxin Xiao, Xiaohui Yan, Jun Yu, Deng Cai, and Fei Wu. Long-form video question answering via dynamic hierarchical reinforced networks. *IEEE Transactions on Image Processing*, 28(12):5939–5952, 2019.
- [69] Jinxing Zhou, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, and Yiran Zhong. Audio-visual segmentation. arXiv preprint arXiv:2207.05042, 2022.
- [70] 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.

[71] Chen Zhu, Wei Ping, Chaowei Xiao, Mohammad Shoeybi, Tom Goldstein, Anima Anandkumar, and Bryan Catanzaro. Long-short transformer: Efficient transformers for language and vision. *Advances in Neural Information Processing Systems*, 34:17723–17736, 2021.

A Dataset Examples

Dataset examples can be found on the website of our project: http://gewu-lab.github.io/LFAV/

B Auxiliary Statistical Analysis

We show the number of video-level and event-level labels of each category in Tab. 6 and Tab. 7, respectively. For video-level, the LFAV dataset contains 24,875 video-level annotations, including 11,404 visual event annotations and 13,471 audio event annotations, each category occuring in at least 146 videos. For event-level, the label distribution is similar to video-level (see Fig. 2(f) in the *main paper*), but event-level labels just exist in the validation set and testing set. The LFAV dataset contains 23,666 event-level event annotations, including 11,331 visual event annotations and 12,335 audio event annotations. We also show the third-order interactions among different video-level label categories in Fig. 5. Almost all categories have dense interactions with other categories, and some of them have closer relations (*e.g.*, *clapping*, *laughing*, and *speech*). These statistical results illustrate the diversity of the collected videos.

Table 6: Number of video-level labels of each category. For each category, we stat number of videos that it occurs (*i.e.*, occurs in at least one modality of the video), number of video-level visual labels, and number of video-level audio labels. For example, for category *clapping*, it occurs in 1,486 videos in total, number of video-level visual labels and audio labels are 899 and 1,341, respectively.

Label			Training		,	Validatio	n		Testing			Total	
id	Categories	video	visual	audio	video	visual	audio	video	visual	audio	video	visual	audio
01	speech	1972	626	1938	322	181	296	614	342	562	2908	1149	2796
02	clapping	951	564	851	158	95	148	377	240	342	1486	899	1341
03	cheering	1048	397	985	130	44	126	296	99	286	1474	540	1397
04	laughing	629	430	457	158	100	115	305	208	233	1092	738	805
05	singing	545	356	531	109	74	102	240	166	219	894	596	852
06	car	459	440	207	88	78	36	137	119	65	684	637	308
07	guitar	401	326	382	96	82	89	181	161	171	678	569	642
08	drum	377	181	349	65	40	61	132	80	124	574	301	534
09	piano	294	211	279	66	51	61	137	107	130	497	369	470
10	dance	255	240	81	44	43	3	113	108	24	412	391	108
11	alarm	347	255	331	28	15	24	33	14	28	408	284	383
12	dog	266	249	188	47	39	28	80	73	36	393	361	252
13	violin	234	192	225	48	46	47	82	78	80	364	316	352
14	playing basketball	267	267	179	13	14	11	32	32	26	312	313	216
15	playing badminton	193	192	124	26	26	25	50	49	45	269	267	194
16	horse	220	220	149	19	19	8	23	23	12	262	262	169
17	bicycle	194	191	24	21	19	3	47	41	17	262	251	44
18	cello	137	127	129	28	27	27	62	57	59	227	211	215
19	rodents	166	166	161	15	15	12	25	23	19	206	204	192
20	frisbee	184	181	90	6	5	2	12	11	3	202	197	95
21	fixed-wing aircraft	170	164	157	11	10	11	20	20	18	201	194	186
22	playing ping-pong	181	181	147	1	1	1	18	17	17	200	199	165
23	accordion	111	108	108	24	24	21	59	56	59	194	188	188
24	playing volleyball	147	147	71	14	14	9	28	27	17	189	188	97
25	playing baseball	151	151	50	7	7	6	31	30	12	189	188	68
26	cat	130	129	78	18	15	4	40	39	15	188	183	97
27	playing tennis	161	161	126	6	6	6	21	21	18	188	188	150
28	banjo	132	125	127	10	10	10	44	42	42	186	177	179
29	car_alarm	149	107	140	12	8	8	18	6	14	179	121	162
30	helicopter	100	99	74	28	28	27	50	49	47	178	176	148
31	crying	136	121	117	13	8	10	21	18	16	170	147	143
32	chainsaw	105	104	100	26	24	26	35	34	35	166	162	161
33	playing soccer	118	117	61	17	17	13	26	26	13	161	160	87
34	chicken_rooster	135	122	131	4	4	3	11	10	8	150	136	142
35	shofar	118	116	107	15	14	14	13	12	12	146	142	133
	Total	11183	7763	9254	1693	1203	1393	3413	2438	2824	16289	11404	13471

Table 7: Number of event-level labels of each category. And the label id is the same as in Tab. 6.

Catananian	Valid	ation	Tes	ting	Total		
Categories	visual	audio	visual	audio	visual	audio	
speech	629	1151	1255	2217	1884	3368	
clapping	275	437	662	1116	937	1553	
cheering	132	469	280	1024	412	1493	
laughing	338	394	597	831	935	1225	
singing	237	259	655	556	892	815	
car	269	55	373	163	642	218	
guitar	274	137	670	356	944	493	
drum	156	92	372	234	528	326	
piano	131	123	339	233	470	356	
dance	105	4	225	32	330	36	
alarm	28	66	41	63	69	129	
dog	100	64	207	137	307	201	
violin	232	95	239	178	471	273	
playing basketball	25	13	101	68	126	81	
playing badminton	52	68	109	168	161	236	
horse	44	12	50	17	94	29	
bicycle	70	3	110	43	180	46	
cello	145	46	188	95	333	141	
rodents	27	27	25	26	52	53	
frisbee	9	2	38	7	47	9	
fixed-wing aircraft	32	27	36	23	68	50	
playing ping-pong	11	11	34	91	45	102	
accordion	67	33	147	93	214	126	
playing volleyball	72	70	154	118	226	188	
playing baseball	8	7	77	23	85	30	
cat	50	16	75	32	125	48	
tennis	9	6	53	38	62	44	
banjo	38	14	117	135	155	149	
car alarm	16	17	23	36	39	53	
helicopter	46	46	85	66	131	112	
crying	15	14	28	29	43	43	
chainsaw	64	48	62	69	126	117	
playing soccer	38	20	88	66	126	86	
chicken_rooster	4	14	21	31	25	45	
shofar	18	19	29	42	47	61	
Total	3766	3879	7565	8456	11331	12335	

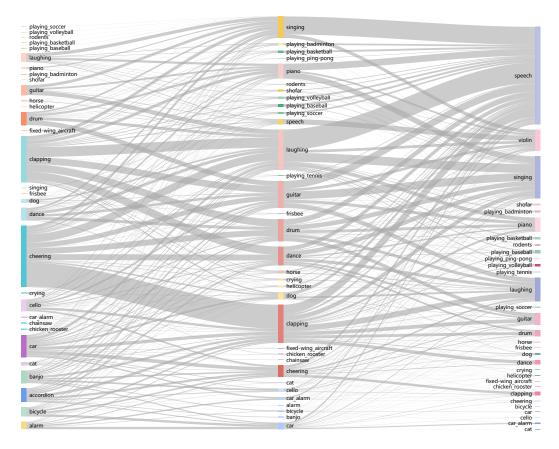


Figure 5: Sankey diagram of video-level label in LFAV dataset, which shows the third-order interactions of video-level labels.

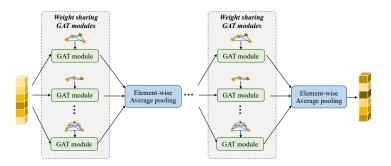


Figure 6: The architecture of the GAT based model. Model architecture of audio and visual modalities are the same, this figure shows the model architecture of one modality.

Auxiliary Explanation of the Method C

Temporal Attention Pooling

We can obtain the video-level and snippet-level prediction by the Temporal Attention Pooling (TAP) module. Inputs of the TAP module are audio features $\{a_t\}_{t=1}^T$ and visual features $\{v_t\}_{t=1}^T$, then outputs of the TAP module are video-level audio prediction p^a , video-level visual prediction p^v , snippet-level audio prediction $\{p_t^a\}_{t=1}^T$, and snippet-level visual prediction $\{p_t^v\}_{t=1}^T$. Video-level predictions are used to train the model, snippet-level predictions are used to construct event graphs during training and evaluate model performance during validation and testing. Snippet-level predictions are calculated as:

$$p_t^a = Sigmoid(FC(a_t)),$$

$$p_t^v = Sigmoid(FC(v_t)),$$
(1)

$$p_t^v = Sigmoid(FC(v_t)), \tag{2}$$

where FC represents a fully connected layer, the audio modality and visual modality share the same fully connected layer. For video-level predictions, another fully connected layer FC' is used to obtain normalized attention weights of each audio and visual snippet at first:

$$w_t^a = \frac{exp(FC'(a_t))}{\sum_{j=1}^T exp(FC'(a_j))},$$
(3)

$$w_t^v = \frac{exp(FC'(v_t))}{\sum_{j=1}^T exp(FC'(v_j))},$$
(4)

where w_t^a represents the weight of t-th audio snippet and w_t^v represents the weight of t-th visual snippet. Two modalities also share the same fully connected layer. Then video-level predictions are calculated as:

$$p^a = \sum_{t=1}^T w_t^a \odot p_t^a, \tag{5}$$

$$p^v = \sum_{t=1}^T w_t^v \odot p_t^v, \tag{6}$$

where \odot represents the element-wise multiplication.

Compared with the MMIL Pooling in HAN [48], we do not perform modality-wise attention because we have independent audio and visual labels during training.

C.2 Graph Attention Network Based Model

We propose a graph attention network (GAT) [51] based model to refine event-aware snippet features, Fig. 6 shows its detailed architecture. For a layer in the model, all GAT modules in the same layer share the same weights but use category-aware graph structures to aggregate snippet features in different events. For each GAT module, we obtain refined event-aware snippet features. Then the output of the layer is the average output of all GAT modules.

C.3 Snippet Reweighting

In the event interaction phase, we use refined event features to reweight each snippet, then obtain updated video-level predictions for training. Concretely, the attention weight of each snippet is reset according to the cosine similarity between the event feature and the snippet feature. Suppose \hat{a}^i and \hat{v}^i are event features of the *i*-th category of the audio modality and the visual modality, respectively. For the *i*-th category of the audio modality, weight of the *t*-th snippet is the *i*-th element of w_t^a , it is represented as $w_t^a[i]$ and recalculated as:

$$w_t^a[i] = \frac{exp(\cos(\hat{a}^i, a_t))}{\sum_{j=1}^T exp(\cos(\hat{a}^i, a_j))},$$
(7)

where cos represents the cosine similarity function, then the updated video-level prediction of i-th category of the audio modality is calculated as:

$$p^{a}[i] = \sum_{t=1}^{T} p_{t}^{a}[i] \odot w_{t}^{a}[i].$$
 (8)

For the visual modality, the same calculation will be also performed to update video-level prediction. For categories that have not extracted events, we do not reweight snippets and keep previous video-level predictions.

C.4 Loss Function

Suppose $p_{(j)}^a$ and $p_{(j)}^v$ represent video-level audio and video predictions of the j-th phase, respectively. For each phase, a binary cross-entropy loss is performed to optimize the model:

$$\mathcal{L}_{j} = BCE(p_{(j)}^{a}, y^{a}) + BCE(p_{(j)}^{v}, y^{v}), \ j = 1, 2, 3,$$
(9)

where y^a and y^v are audio and visual event labels, respectively. \mathcal{L}_j is the video-level loss of the j-th phase. An event-level binary cross-entropy loss (event loss) is also performed in the event interaction phase. Event predictions are first calculated as:

$$\hat{p}^{a,i} = Sigmoid(FC(\hat{a}^i))[i], \tag{10}$$

$$\hat{p}^{v,i} = Sigmoid(FC(\hat{v}^i))[i], \tag{11}$$

where $\hat{p}^{a,i}$ and $\hat{p}^{v,i}$ are audio event prediction and visual event prediction of the *i*-th category, FC is the fully connected layer which used to obtain snippet-level predictions in the TAP module of event extraction phase. Suppose n_a is the number of extracted audio events, n_v is the number of extracted visual events, S_a and S_v are sets of category indexes of extracted events in two modalities, respectively. Then the event loss \mathcal{L}_e is calculated as:

$$\mathcal{L}_{ea} = \frac{1}{n_a} \sum_{i \in S_a} BCE(\hat{p}^{a,i}, y^a[i]), \tag{12}$$

$$\mathcal{L}_{ev} = \frac{1}{n_v} \sum_{i \in S_v} BCE(\hat{p}^{v,i}, y^v[i]), \tag{13}$$

$$\mathcal{L}_e = \mathcal{L}_{ea} + \mathcal{L}_{ev},\tag{14}$$

the total training loss is:

$$\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3 + \alpha \mathcal{L}_e, \tag{15}$$

where α is the weight of event loss.

D Auxiliary Experimental Results

D.1 Detailed Training Settings

The detailed training settings are shown in Tab. 8.

D.2 Settings of Comparison Methods

We modify some of the comparison methods to make them possible to solve the multisensory temporal event localization task. For STPN [35], the weight of sparse loss is set to 0 to achieve better performance. For Longformer [3] and Transformer-LS [71], we use HAN as the backbone and replace attention layers in the HAN model with attention layers in Longformer or Transformer-LS, their window size is set to 50. For ActionFormer [66], we use its backbone and FPN decoder to extract multi-scale features and use the upsampling strategy proposed in U-Net [37] to fuse multi-scale features. The number of transformer blocks is set to 5 and the last 3 transformer layers are downsampling layers. For Longformer, Transformer-LS, and ActionFormer, the number of snippets of all videos is adjusted to 200 while testing.

Table 8: Detailed training settings of the framework, the learning rate step is set to (10; 0.1), means the learning rate is reduced by a factor of 0.1 for every 10 epochs.

Hyperparameter	Value
batch size	16
train epochs	30
initial learning rate (snippet prediction phase)	1e-4
initial learning rate (event extraction phase)	1e-4
initial learning rate (event interaction phase)	2e-4
learning rate step	10; 0.1
event loss weight	0.3
feature dim	512
depth of PMT	6
number of heads of PMT	4
dropout ratio of PMT	0.2
confidence threshold of semantic edges	0.5
depth of graph network	2
number of heads of graph network	1
dropout ratio of graph network	0
dropout ratio of event interaction layers	0

For methods proposed for audio-visual event localization [49] or audio-visual video parsing [48] tasks AVE [49], PSP [70], HAN [48], and DHHN [27], we use both audio and visual labels for training and use the same loss to HAN. For all comparison methods, predictions of audio-visual events are the multiplication of audio events predictions and visual events predictions [48].

D.3 More Ablation Studies

Effectiveness of Audio-visual Joint Training. As mentioned in the *main* paper, ablation studies about feature interaction have shown that crossmodal attention improves video understanding through audio and visual cooperation. To further illustrate it, we show the AP boosting of typical categories when considering snippetlevel cross-modal attention in Fig. 7. Audio and visual take distinct viewpoints in describing the video content but facilitate the understanding of the video cooperatively. For example, for category singing, audio provides the rhythm of the song, and visual provides the pose of the singer; for category *playing_tennis*, audio provides the impact sound of tennis, and visual provides the motion of players.

Comparison of Audio Supervision.

We use visual annotations as audio supervision to train our model (*i.e.*, regard visual labels as both audio and visual supervision during training). The performance of localizing audio-specific events (*i.e.*, events only occur in the audio modality and do not occur in the visual modality, as shown in Fig. 8) with different audio super-

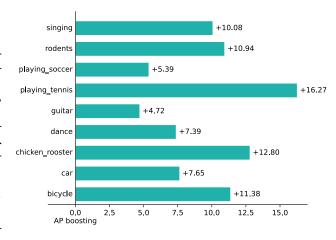


Figure 7: AP boosting of some typical categories when considering the joint learning of audio and visual.

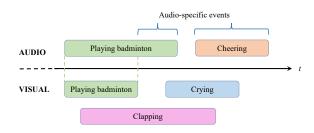


Figure 8: An example of audio-specific events.

vision are in Tab. 9. When we view visual annotations as audio supervision, the snippet-level recall of audio-specific events is even far below 50%, showing that mismatched audio supervision will result in the collapse of audio-specific events localization performance.

Effectiveness of Snippet Shift. We explore the effectiveness of the snippet shift strategy in the snippet prediction phase, results are shown in Tab. 10. The snippet shift strategy is beneficial to event-aware snippet interaction. Notice that snippet shift is a computation-efficient operation, hence, the performance improvement from snippet shift is almost computationally free.

Table 9: Ablation study on audio supervision. We choose snippet-level recall as the metric to achieve a more accurate localization performance of audiospecific events, the threshold is set to 0.5.

Audio supervision	Recall (%)
Visual annotations	30.60
Audio annotations	63.09

Effectiveness of Event Loss. We also explore

the effectiveness of event loss in the event interaction phase. The event loss is performed over snippets of extracted events. Corresponding results are shown in Tab. 11, which indicates the importance of event loss in localizing multiple events in long form videos.

D.4 Visualization Results

We visualize the event-level localization results in the videos, two examples are shown in Fig. 9. Compared with the audio-visual video parsing method HAN [48], our proposed method achieves better localization results. In some situations (e.g., event guitar in both audio and visual modality of video 01, and event speech in the audio modality of video 02), HAN tends to localize some sparse and short video clips instead of a long and complete event, which shows that HAN exists some limitations to understanding long-form videos. The possible reason is that HAN cannot learn long-range dependencies well.

We also notice that, although our proposed event-centric method has achieved the best performance among all methods, there still exist some failure cases in the shown examples. The auditory event of *drum* in *video 01* and visual event of *speech* in *video 02* are not well localized by both methods, including ours (marked with red box). There are also existing some short and sparse clips in our prediction results (marked with the black box). We can find that these multisensory events take huge different lengths and occur in a dynamic long-range scene, which makes multisensory temporal event localization become a very challenging task, especially with only video-level labels in training. Although our method has partially addressed it according to the shown results, this challenging task still needs more exploration in future work (*e.g.*, considering some post-processing methods for the prediction results to merge or correct short and sparse prediction clips).

Table 10: Ablation study on snippet shift strategy in the snippet prediction phase.

Snippet shift	F1-score Avg.	mAP Avg.
×	25.98 26.04	52.51 52.67

Table 11: Ablation study on event loss.

Event loss	F1-score Avg.	mAP Avg.
X ✓	27.64 28.14	54.70 55.29

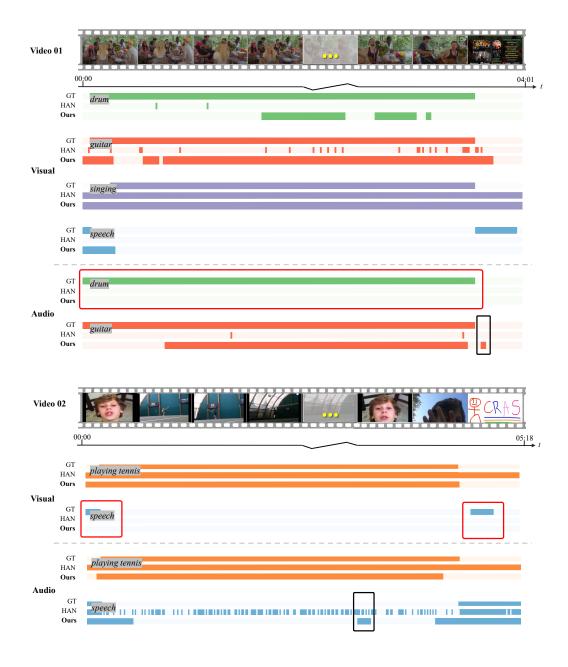


Figure 9: Two examples of visualized localization results of our method and the best comparison method HAN, red rectangles in the figure show some failure cases of our method.

E Discussion

As mentioned in the *main paper*, we hope our work could bring some inspiration to further research of audio-visual video understanding, here we give a brief discussion.

Audio-visual video dense captioning. The goal of the video dense captioning task is to detect multiple events in videos then describe them using natural language sentences [29]. The audio was exploited in previous video dense captioning approaches [25, 26, 36], while all of them did not treat audio as an independent modality. They used visual-level annotations to label the audio modality. However, we want to note that there are diverse audio and visual events and the two modalities are not always temporally correlated. Accurate multisensory temporal event localization results can help us solve a "real" audio-visual video dense captioning problem that aims to detect and describe different audio and visual events in long form videos.

Reasoning over scene dynamics. Causal relations between events are ubiquitous in the real world (e.g., In Fig. 1 in the main paper, the badminton player cries because he has won the game and he is very excited.). Due to the multisensory characteristic of events, data in different sensory channels also exist in causal relations. These casual relations can be a bridge to achieving a high-level understanding of the video. Humans usually can reason about causal relations between events easily, but for the machine, understanding and reasoning are essential difficult tasks. Our proposed event-centric framework can capture scene dynamics in videos, which serves as the cornerstone to facilitate future research in complex multisensory scene reasoning.