VidEvent: A Large Dataset for Understanding Dynamic Evolution of Events in Videos

Baoyu Liang[1, 2], **Qile Su**[1, 2], **Shoutai Zhu**[1, 2], **Yuchen Liang**[1, 2], **Chao Tong**[1, 2 *]

[1]School of Computer Science and Engineering, Beihang University, China
[2]State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, China
{liangbaoyu96, zy2423342, shoutaizhu, liangyuchen, tongchao}@buaa.edu.cn

Abstract

Despite the significant impact of visual events on human cognition, understanding events in videos remains a challenging task for AI due to their complex structures, semantic hierarchies, and dynamic evolution. To address this, we propose the task of video event understanding that extracts event scripts and makes predictions with these scripts from videos. To support this task, we introduce VidEvent, a large-scale dataset containing over 23,000 well-labeled events, featuring detailed event structures, broad hierarchies, and logical relations extracted from movie recap videos. The dataset was created through a meticulous annotation process, ensuring high-quality and reliable event data. We also provide comprehensive baseline models offering detailed descriptions of their architecture and performance metrics. These models serve as benchmarks for future research, facilitating comparisons and improvements. Our analysis of VidEvent and the baseline models highlights the dataset's potential to advance video event understanding and encourages the exploration of innovative algorithms and models.

Introduction

Events are at the center of human experience(Radvansky and Zacks 2017). We perceive events when observing, engage in events when acting, learn from events and use event knowledge to solve problems. While understanding events with our eyes is intuitive for humans, it remains a significant challenge for AI. Despite extensive research in natural language processing (NLP) on events, dealing with events in visual scenarios is still extremely difficult for AI models that are good at capturing static entities such as actions (Wang et al. 2023a; Duan et al. 2022; Shi et al. 2023) and moving objects (Wang et al. 2023b; Wei et al. 2023), but lack understanding of dynamic events.

Unlike atomic actions or objects, events typically have the following three characteristics: (1) **Complex structure**. Events are composed of different constituents including participants, tools, time, and location, thus forming complex structures, which require a comprehensive understanding of the current situation. (2) **Various semantic hierarchies**. Events usually contain different semantic levels and relations. For example, 'A singer sings on stage' can be divided into several actions such as 'person standing on stage', 'person holding a microphone' and 'person making sound with mouth', while atomic actions can hardly describe such hierarchical relation. Distinguishing these actions from high-level events is difficult and requires understanding different semantic levels. (3) **Dynamic logical evolution**. Events are dynamic and evolve over time, following logical sequences. Understanding this evolution requires sophisticated event comprehension and commonsense reasoning, which current research finds challenging.

Recent video understanding methods have made progress in tasks like predicting actions, detecting objects, and describing visual scenes (Wang et al. 2023a,b; Wei et al. 2023; Duan et al. 2022; Ko et al. 2023). However, few studies have focused on the analysis of structured and dynamic events, which are crucial to human cognition and could significantly advance computer vision (CV) research (Li et al. 2022). Situation Recognition and Grounded Situation Recognition (Sadhu et al. 2021b; Khan, Jawahar, and Tapaswi 2022) are pioneering efforts in this direction, focusing on extracting event structures. Nevertheless, these tasks mainly emphasize event structures with limited consideration of the semantic hierarchies and dynamic logical evolution of events, which are crucial for cognitive understanding. We summarize these tasks and datasets in Table 1 and further analyze them in the Supplementary Materials due to the page limits.

In this paper, we further explore the understanding of events from visual scenes and propose the task of video event understanding that extract and induct with scripts of hierarchical, and dynamically evolutionary events and as is shown in Figure 1. Video event understanding focuses on the logical evolution of events and thus encourages the extraction of highly conclusive events with high semantic levels other than the atomic actions and the prediction of the key relations that form complete logical chains.

To support the task, we publish a new large-scale video event understanding dataset called VidEvent containing over 23,000 events and more than 17,000 relations from 1,110 movie recaps videos. VidEvent is characterized with complete event structures, macro event hierarchies and sophisticated event evolutionary chains. Events are ensured to have common complete structures and broad semantic hierarchies by a meticulous annotation process under the well-established Propbank framework (Kingsbury and Palmer

*Corresponding author

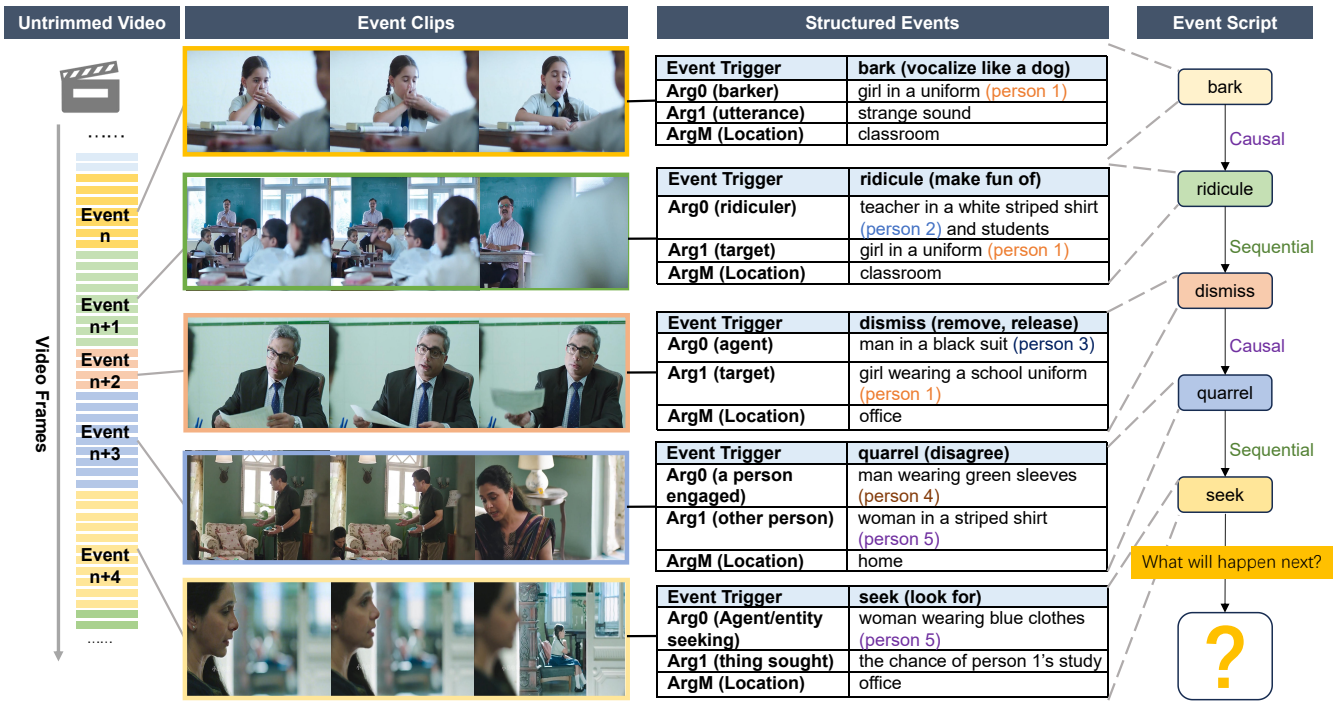

Figure 1: The task of Video Event Understanding aims to extract event scripts with complete event structures and relations, and induct with event scripts. Here is an example in our proposed dataset VidEvent supporting the task. VidEvent provides untrimmed videos segmented with different event clips that vary in length. Each event clip is carefully labeled with a structured event including the event trigger like *dismiss (remove, release)* and the event arguments like *Arg0* and *Arg1*. All the semantic roles participating in different events are co-referenced with a unique number like *person 3* across the video. The relations among these events are also provided in VidEvent to support the construction and utilization of event scripts.

2003) from highly concluded recap videos that are considered to have more complete structures, macro semantic levels, and more precise relations, compared with natural video or films. Relations among events are also carefully annotated following a strict labeling standard to form long event chains that ensures complete narrative logic.

We also provide evaluation metrics and baselines as benchmarks for future subsequent research. These benchmarks are designed with consideration of the challenges, i.e., the complex event structures, various semantic hierarchies, and dynamic logical evolution based on transformers. We also present various video understanding methods including SlowFast (Feichtenhofer et al. 2019), TimesFormer (Bertasius, Wang, and Torresani 2021), CLIP (Radford et al. 2021), ActionFormer (Zhang, Wu, and Li 2022), TriDet (Shi et al. 2023) and LLMs and provide comprehensive results for this task.

Our contributions include:

(1) We propose the task of video event understanding which includes four progressive subtasks, aiming to enhance the scene understanding capabilities in higher event level, bridge the gap of event understanding and inference between CV and NLP, and explore the leap of artificial intelligence from perception to cognition. We claim our work to be the first to support extracting highly-concluded events and analyzing long-term event evolution as far as we know.

(2) We release a large-scale dataset called VidEvent in support of the proposed task. Over 1,000 movie recaps are carefully annotated to extract over 23,000 events with higher semantic level and more than 17,000 relations with accurate evolutionary logic between events.

(3) Baseline methods and evaluation metrics are presented, along with detailed comparative results with current video understanding methods to form a comprehensive benchmark for future research.

Task: Video Event Understanding

Analyzing the dynamic complex events is still a challenging task in computer vision depict the extensive research in the area of NLP. In face of the complex structured format, various semantic hierarchy, and dynamic logical evolution of events, we propose our task of video event understanding, aiming to extract the well-structured dynamic events with different semantic levels and further analyze the relations among them to supply the logic inference over the historical development of events.

Formal Definition

We provide the following definitions to describe the task:

Definition 1 (Event Clip)**.** Given a video V with T frames, an event clip $C \subset V$ is defined as a trimmed video clip in V that describes a certain event E within the event boundaries

Task	Dataset	Event	Boundaries	Hierarchies	Evolution
AC	Kinetics (2017), ActivityNet (2015), S-S (2017), HVU (2020), Youtube8M (2016)	verb	✗	✗	✗
ACL	ActivityNet, Thumos (2017), HACS (2019), Charades (2016)	verb	✓	✗	✗
STD	AVA (2018), EPIC-Kitchens (2018), JHMDB (2013)	verb	✓	✗	✗
VD	ActivityNet, Vatex (2020), MSR-VTT (2016) , LSMDC (2016)	texts	optional	✗	✗
VQA	MSRVTT-QA (2017), VideoQA (2022), TVQA (2019)	texts	✗	✗	✗
VR	HowTo100M (2019), DiDeMo (2017), Charades-STA (2017)	texts	✗	✗	✗
VOG	ActivityNet, VidSTG (2020),VID-sentence (2019)	texts	✗	✗	✗
VSRL	VidSitu (2021a)	structures	✗	✗	pairwise
VEU	**VidEvent (Ours)**	structures	✓	✓	chain-like

Table 1: A non-exhaustive summary on the video understanding tasks and datasets.

of $\mathbf{b} = \{b, e\}$, where $b \geq 0$ and $e \leq T$ are the beginning and ending frames of E.

Definition 2 (Event Structure)**.** An event can be represented with an event structure $E = (trg, ARG)$, where an event trigger trg is a word or span expressing the event. The argument set $ARG = \{arg_0, arg_1, \ldots, arg_k\}$ contains the arguments corresponding to trg that describe the detailed elements of the event. We represent a set of E as $\mathcal{E}$.

Definition 3 (Event Relation)**.** $R = \{r_1, r_2, \ldots, r_m\}$ is defined as a set of relations, where each relation r_j is a binary relation $r_j : \mathcal{E} \times \mathcal{E} \rightarrow \{0, 1\}$. That is, $r_j(E_i, E_k) = 1$ indicates that the relation r_j holds between events E_i and E_k.

Definition 4 (Event Script with Relations)**.** An event script with relations $\mathcal{S} = (\mathcal{E}, R_{\mathcal{S}})$ can be seen as a structured representation of sequences of events that have relations in a particular context or scenario, where $\mathcal{E}$ is a set of events and $R_{\mathcal{S}} \subseteq R$ represents the set of relations that hold between the events in $\mathcal{E}$.

Definition 5 (Script Event)**.** An event E_i in one event script $\mathcal{S}$ is called a script event.

Definition 6 (Video Event Understanding)**.** The task of video event understanding is to understand the event scripts $\mathcal{S}$ from V with all the components of events $\mathcal{E}$ and relations $R_{\mathcal{S}}$, and make use of $\mathcal{S}$ to predict next unseen events $\hat{E}$. It requires a comprehensive solution to segmenting the video V into event clips C, extracting the complete event structure E from C, analyzing the relations R to form event scripts $\mathcal{S}$ and make predictions with $\mathcal{S}$.

Unlike event understanding in NLP that has clear event boundaries within sentences and explicitly presents event triggers, arguments and even relations using words, understanding events and scripts in videos poses unique challenges due to the nature of video in terms of complex structure, semantic hierarchies and dynamic evolution of events. We further discuss these challenges and our considerations in simplifying the assumptions when designing the task.

Timescale of Event Clips

Events possess rich semantic hierarchies, and the amount of information a video can convey per unit of time can vary significantly. In previous work, a fixed timescale of 1 or 2 seconds has often been used to identify atomic actions (Gu et al. 2017) or video situations (Sadhu et al. 2021b). However, a fixed timescale might inadvertently capture incidental atomic actions or fragment a cohesive event, thereby complicating event analysis and inference.Therefore, we advocate that event clips at different hierarchical levels should be represented with varying timescales. In addition, in order to better facilitate reasoning about the evolution of events, we tend to choose longer timescales for event clips that contains events of higher hierarchies, while preserving the independence of each event. This is achieved by incorporating specific criteria into the annotation guidelines and conducting rigorous annotation reviews (see Supplementary Materials for detailed annotation criteria and the pipeline). This poses challenges of segmenting the events at a holistic level rather than merely dividing event clips based on scene transitions, character changes, or atomic actions.

Event Structures and Correference of Arguments

An event is conceptualized as an event trigger paired with a set of corresponding arguments. The triggers are derived from PropBank (Kingsbury and Palmer 2003), a comprehensive semantic role labeling corpus for English verbs. PropBank offers a detailed lexicon of verbs, each associated with a range of possible argument roles. We adopt them as our event vocabulary and the corresponding roles as argument templates. In addition, co-reference of entities is critical for analyzing event evolution across an event script. Previous work (Sadhu et al. 2021a) have relied on imposing strict constraints as enforcing identical textual representations for co-referential arguments. However, this can be impractical in the context of long-duration videos or extended event scripts, where salient features of one character may vary over time. Hence we introduce a unique identifier for each person-related entity to facilitate consistent co-reference throughout the video in order to preserve the integrity of event reasoning in long event scripts.

Event Relations

Event relations are defined based on temporal, causal, conditional, coreference, and subevent relations, following the latest event relation extraction dataset MAVEN-ERE (Wang et al. 2022). We introduce an additional conditional relation to specifically capture cases where one event is a prerequisite for the occurrence of another event. If Event B appears

after Event A in a video, Event A: (1) occurs before Event B temporally in a temporal relation; (2) directly leads to the occurence of Event B in a causal relation; (3) is the essential condition, but is not directly responsible for the occurrence of Event B; (4) refers to the same event with Event B in a coreference relation; (5) includes Event B in a subevent relation. We prioritize causal and conditional relations over temporal ones because they inherently involve temporal aspects and play a more significant role in understanding event evolution. Unlike in NLP, analyzing event relation with videos lacks explicit or implicit connectives and referential cues that hint at relations between events. Therefore, a comprehensive understanding and inference of event semantics and inter-event relations are essential.

Script Event Induction in Vision

Inducting script events origins from NLP, with the aim of inferring the next possible events with current known events. Visual data can provide extensive details for understanding and predicting events in this task. However, the absence of explicit event representations and relational descriptions poses significant challenges. Given that this area remains largely unexplored, we propose a foundational taxonomy and simplify the task by introducing a single-choice setting. Formally, given the event clips $\mathcal{C} = \{C_1, C_2, \cdots, C_i\}$, a script structure $\mathcal{S} = (\mathcal{E}, R_{\mathcal{S}})$, and candidate events $\mathcal{M} = \{M_0, M_1, \cdots, M_4\}$, it predicts the next possible event

$$\hat{E} = \arg\max_{E \in \mathcal{M}} (\Pr(E|\mathcal{C}, \mathcal{S})). \quad (1)$$

We provide 5 candidate events in $\mathcal{M}$, two events are selected from a unified candidate pool, one event outside the event chain within the same video, and one event serves as a distractor by randomly replacing an event argument.

Dataset: VidEvent

To establish the database of Video Event Understanding task, we introduce VidEvent, a video event dataset composed of massive recap videos with structured annotations including diverse events, rich event arguments and precise event relations. The proposed dataset meets the demand for complex structures, various semantic hierarchy, and dynamic logical evolutionary. We further describe the detailed data and data analysis.

Data Composition

Movie Recap Videos. VidEvent boasts a collection of movie recap videos. These videos succinctly summarize and condense the plots of original movies into several minutes. In contrast to the natural movie, recap videos offer highly concise summaries, with events and scenes aligning more clearly, and the logical progression of events being more pronounced (Singh, Srivastava, and Tapaswi 2024). Typically, each sentence of the subtitles in the movie recap videos corresponds to one event, although there are cases where a single event may span several subtitles. In order to avoid the potential violation of copyrights, VidEvent only provides publicly accessible URL addresses for these videos, rather than the video files themselves.

Data Statistics	Value
Number of Videos	1,110
Average Video Length	1 min 22 s
Average Event Length	4.5 s
Annotation Coverage of Events	96%
Number of Events	23,989
Number of Arguments	80,822
Number of Relations	17,525
Average Event Chain Length	10.57
Average Coreference Chain Length	4.67

Table 2: Basic Data Statistics of VidEvent

Annotations. Each video in VidEvent is labelled with a JSON-formatted annotation. The annotations primarily consist of three types of information:

(1) Video information. Video information containing video ID and URL is provided in this field.

(2) Event clips. We include all annotated event in one video into event clips. Each event clip is annotated with one clip ID, the starting and ending time of the event, and the corresponding event structure. The entire event structure are sourced from Propbank. In addition, we add two extra arguments ArgM-LOC and ArgM-EXT that indicates the location and extent of the event's occurrence for each event to distinguish the details in visual events. All arguments are described in free texts with their salient features in the video. Each character is assigned with a unique identifier across the video, facilitating the comprehension of relations between events and characters.

(3) Relations between event clips. Each relation within a video is identified by a relation ID and includes information including the relation type, the clip ID of the head event, and the clip ID of the tail event. The head event is defined as either a triggering event or a preceding event, while the tail event is considered as an affected, included or a subsequent event.

We provide more annotation examples as well as the data collection and annotation pipeline in the Supplementary Materials.

Data Statistics and Analysis

As shown in the Table 2, VidEvent has a large number of videos and events with rich annotations and long event chains. A high annotation coverage rate in VidEvent indicates the high usability and coherence of the events. The average event chain length and coreference chain length suggest the presence of longer sequences of associations between events and complex character relations. All of these features support and challenge the long-term induction with visual events, which has not been address by current datasets.

VidEvent is a dataset that contains events of complex structures, rich hierarchy, and dynamic logical evolutionary. To validate these characteristics, we compare VidEvent with other videos datasets containing text descriptions, MSR-VTT (Xu et al. 2016), ActivityNet Captions (Krishna

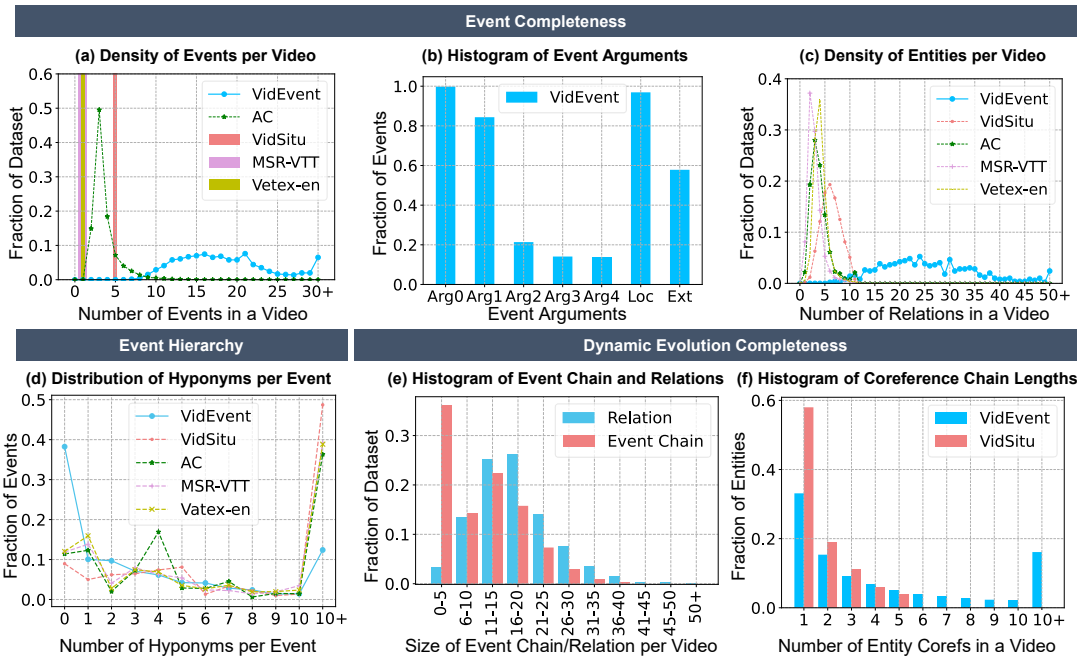

Figure 2: Data analysis. We emphasize on the **event completeness** of event structures (a-c), **event hierarchy (d)** and **relation completeness** of event relations and coreference chain lengths (e and f). AC represents ActivityNet Captions (Krishna et al. 2017).

et al. 2017), Vatex-en (Wang et al. 2019) and VidSitu (Sadhu et al. 2021b). Among them, VidSitu provides complete event structures, while the annotations in the other datasets are sentences or captions describing the video scenarios.

Event structure completeness. The factor of event structure completeness indicates whether the structures of events in the dataset are sufficiently comprehensive, containing adequate and diverse event triggers and arguments. We examine the density of events per video, as illustrated in Figure 2 (a), and the distribution of event arguments, depicted in Figure 2 (b). Compared to existing datasets, VidEvent tends to have significantly more events per video, and the distribution of its labeled event arguments closely aligns with those in Propbank, indicating the completeness of events and arguments in terms of quantity. Additionally, we investigate the density of different entities per video, as shown in Figure 2 (c). Even within a single video, VidEvent tends to feature a wider range of diverse entities, reflecting a richer and more diverse set of events and scenarios covered in the dataset. This suggests the completeness of events in terms of diversity.

Event hierarchical diversity. We introduce event hierarchical diversity to assess the breadth of semantic hierarchies encompassed by the events within the dataset as illustrated in Figure 2 (d). This is achieved by comparing the distribution of hyponyms associated with each event, as events characterized by a greater abundance of hyponyms, such as 'speak' and 'walk,' typically denote broader semantic scopes, thereby exhibiting a lower semantic hierarchy. The figure reveals that events within the VidEvent dataset demonstrate a sparse distribution of hyponyms, yet maintain a diverse range of semantic hierarchies. This observation contrasts with other datasets primarily focused on atomic actions.

Dynamic evolution completeness. VidEvent provides the annotations for the dynamic evolution of events through the logic relations among different events. To evaluate whether the logical developments of events are complete, we analyze the distribution of logical relations in VidEvent and the length of event chains as shown in Figure 2 (e). Compared with other datasets, VidEvent tends to have more extensive distribution of relations in terms of both amount and variety. We also perform an analysis on the length of coreference chain of semantic roles in VidEvent to identify whether the key characters are labeled out with the development of events. From Figure 2 (e), VidEvent has considerably longer coreference chains, indicating its completeness in event evolution.

Baselines

The task of video event understanding encompasses the recognition and induction of event scripts. We design a fundamental framework, illustrated in Figure 3, which consists of four essential steps that serve as a baseline for future comparative studies.

STEP 1 (Video event localization). Given a video V, we require the model to recognize event boundaries $B = \{(s_i, e_i)\}$, where s_i and e_i are the start and end time of one event in order to obtain the event clips C. As baselines, we provide advanced action localization models such as TriDet (Shi et al. 2023) and ActionFormer (Zhang, Wu,

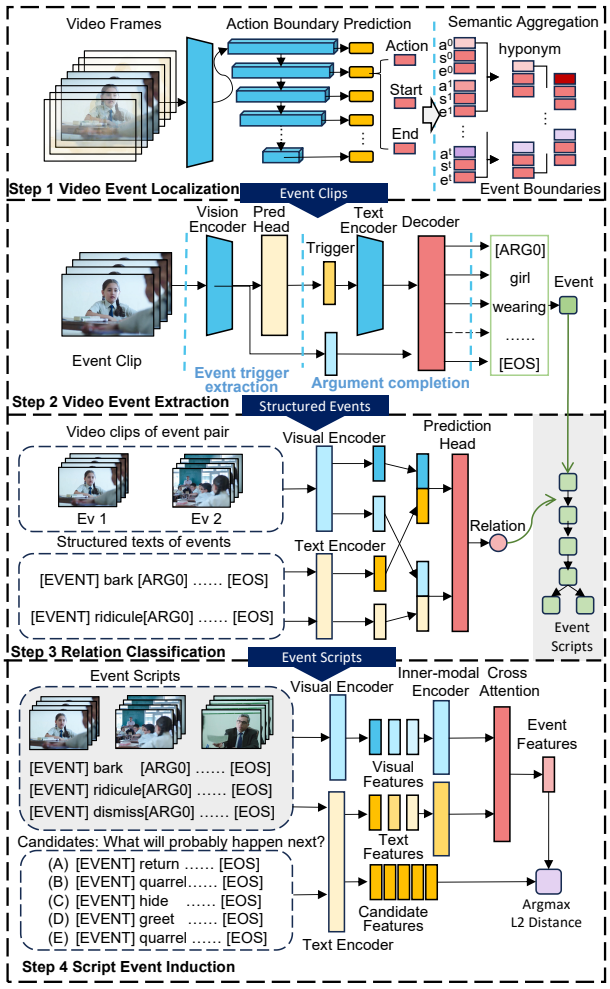

Figure 3: **The baseline models for the subtasks of video event extraction, video event localization, event ralation classification and event script induction.**

and Li 2022) to extract atomic actions and their boundaries. All these models are trained from scratch with their original loss functions. To handle the semantic hierarchies of events, we design a semantic aggregation module that iterately aggregates boundaries of atomic actions with a semantic hyponym tree (see Figure 3). We also provide an attention-based method for semantic aggregation based on TriDet. Details are provided in the Supplementary Materials.

STEP 2 (Video event extraction). For each event clip C, the model is expected to extract the complete event structure E with an event trigger trg and arguments ARG. We set trg to be a verb from the vocabulary and ARG a text sequence like

[ARG0] girl wearing a uniform [ARG1] strange sound [ARGM-LOC] classroom [EOS].

We provide a step-wise method that initially predicts event triggers from individual video clips, followed by the synthesis of event arguments using both the video clips and the event trigger as shown in Figure 3. We adapt widely-used video understanding models such as SlowFast, TimesFormer and CLIP(Feichtenhofer et al. 2019; Bertasius, Wang, and Torresani 2021; Li et al. 2022) as video encoders and transformer encoders like CLIP and RoBerta (Liu et al. 2019) as text encoders. The encoded features are combined with a projection layer, and decoded with a transformer decoder like RoBerta. We also present a similarly-designed model in a joint prediction fashion that includes event triggers into the text sequences.

STEP 3 (Event relation classification). Given two extracted events E_1 and E_2 in their text form, and their event clips C_1 and C_2, we require the model to classify the event relation r between them. To address this multimodal task, we propose a transformer-based baseline that incorporates both a visual encoder and a text encoder. The visual and text encoders can be selected from previous steps, such as SlowFast, TimesFormer, CLIP, RoBERTa, or other suitable models. Once the visual features V_1, V_2 and text features T_1, T_2 corresponding to events are extracted, these features are concatenated by event pairs. The concatenated features are then passed through a prediction head of two fully-connected layers, which is designed to predict the relation r between the two events.

STEP 4 (Script event induction). Given an event script $\mathcal{S}$ and corresponding event clips, we propose a two-branch architecture of vision and text as Figure 3. Each branch contains an encoder for video or text to extract the corresponding features of $E_i \in \hat{E}$, followed by a transformer encoder to make inner interactions among events within one modal. The features extracted by these two branches are combined with a cross-attention layer to make inter-modal interactions and to generate the representation $\hat{F}$ for the predicted event. To represent the candidate events, we use share the text encoder to extract the representations F^{M_j} of each candidate event M_j. Finally, we compute the L2 distances between $\hat{F}$ and each of the candidate event representations F^{M_j}, and choose the nearest candidate event $\hat{E}$ in distance as the answer.

To bring the predicted future event representation closer to the correct candidate event while distancing it from the incorrect candidate event, we employ a triplet contrastive loss (Schroff, Kalenichenko, and Philbin 2015) to train the model.

Experiments

Video event understanding supports evaluating in four steps: (1) video event localization, (2) video event extraction, (3) event relation classification, (4) script event induction with a given video.

Metrics

Video event localization. Video event localization is a task that combines boundary regression and type classification of events. For this task, we use mAP@IoU (Cheng and Bertasius 2022; Shi et al. 2023; Zhang, Wu, and Li 2022; Liu et al. 2022) as the evaluation metric. We report the result at IoU threshold [0.1, 0.2, 0.3, 0.4, 0.5] and the average mAP computed at [0.1:0.5:0.1].

Type	Encoder	Decoder	Event Trigger			Event Arguments		
			P@5	R@5	F1@5	METEOR	CIDEr	SPICE
Step	SF-RB	VB	0.51	0.46	0.48	0.24	0.2	0.17
Step	SF-RB	RB				0.31	0.38	0.25
Step	TF-RB	VB	0.54	0.44	0.48	0.24	0.21	0.17
Step	TF-RB	RB				0.32	0.3	0.24
Step	CLIP-CLIP	RB	0.62	0.3	0.41	0.31	0.4	0.20
Joint	SF	RB	\	\	\	0.33	0.39	0.25
Joint	TF	RB	\	\	\	0.33	0.46	0.25

Table 3: Results of video event extraction. SF, TF, RB and VB are SlowFast, TimesFormer, RoBerta and VisualBert, respectively. The encoders of SF/TF-RB use SlowFast or TimesFormer as the visual encoder and RoBerta as the text encoder, while CLIP-CLIP means that the visual and text encoders are the visual and text parts of CLIP.

Method	Backbone	mAP@					mAP
		0.1	0.2	0.3	0.4	0.5	
ActionFormer	TSP	0.83	0.81	0.75	0.59	0.38	0.67
ActionFormer	SF	0.76	0.73	0.63	0.50	0.30	0.59
TriDet	TSP	0.81	0.71	0.54	0.34	0.17	0.51
TriDet	SF	0.85	0.82	0.76	0.63	0.41	0.69
TriDet+Agg	SF	0.87	0.85	0.79	0.70	0.47	0.74

Table 4: Results of video event localization. mAP@t is the average mAP at an IoU threshold t. mAP is the average mAP at the IoU threshold of [0.1:0.5:0.1].

Video event extraction. In videos, the absence of explicit representations for event triggers and arguments in videos introduces language ambiguities, such as distinguishing between "chat" and "talk." To account for this ambiguity, we utilize Recall at 5 (R@5) and Precision at 5 (P@5) for event triggers, considering a prediction correct if any of the top 5 predictions match the ground truth. This approach helps to partially mitigate the impact of linguistic ambiguity. For evaluating the predicted arguments, we compute the average of METEOR (Banerjee and Lavie 2005), CIDEr (Vedantam, Zitnick, and Parikh 2015), and SPICE (Anderson et al. 2016) scores for each event's arguments. These metrics are designed to account for the influence of synonyms and are robust against ambiguity.

Event relation classification. Event relation classification can be considered as a 6-way classification problem. Similar to that in NLP, we report top-1 accuracy, precision, recall and F1 score to evaluate the quality of predictions.

Script event induction. Similarly to script event induction in NLP, we report the top-1 accuracy among five candidate events to evaluate the agreement between predictions and ground-truths.

Experimental Settings and Results

We introduce the experiment results regarding to the baseline framework here. Implementation details, ablation studies and comparative results with other methods including LLMs are introduced in the Supplementary Materials.

Video event localization. We report mAP@0.1 to mAP@0.5 and the averaged mAP at [0.1:0.5:0.1] on the test set as shown in Table 4. We observe that simply adding a semantic aggregation module increase the performance across all IOU thresholds. This aligns with our expectations regarding the semantic distinction between actions and events.

Method	Text	Vision	Accuracy
RoBerta	✓		0.55
TF		✓	0.23
SF		✓	0.25
TF+RoBerta	✓	✓	0.56

Table 5: Results of script event induction. SF and TF represents SlowFast and TimesFormer.

Method	Text	Vision	P	R	F1
RoBerta	✓		0.45	0.37	0.41
SF		✓	0.51	0.41	0.45
TF		✓	0.41	0.40	0.40
CLIP	✓	✓	0.46	0.38	0.42
SF+Roberta	✓	✓	0.50	0.46	0.48

Table 6: Results of event relation classification. SF and TF represents SlowFast and TimesFormer.

Video event extraction. We report P@5, R@5 and F1@5 on the test set for event trigger and METEOR, CIDEr, and SPICE for arguments. The results are shown in Table 3. The joint approaches outperform stepwise approaches sightly in all metrics, which may be attributed to the cumulative error inherent in the stepwise methodology. However, it is notable that joint approaches can hardly evaluate the performance in predicting individual event triggers, which can limit their potential applications to tasks that emphasize event triggers.

Event relation classification. We report the macro precision, recall and F1 score on test sets in Table 6. The results indicate that both visual and textual modalities demonstrate comparable performance in event relation prediction, highlighting the potential effects of the visual modality in analyzing event relations. The results of SF+Roberta corroborate this conclusion and elucidate the complementary nature of the textual and visual modalities with the highest F1 score of 0.48 among these models.

Script event induction. In our experiments, we set the length of the event chains to 3 and train with a triplet loss. Other than the baseline model using two branches, we also present the results of simplified methods on one text or vision modal by removing one branch and the cross-attention layer in the baseline architecture in Table 5. The results indicate a significant performance disparity between the textual and visual modalities in this task mainly attributed to the explicit expression of logical relations within the vast corpus used for pretraining textual models. Although multimodal learning shows only a slight improvement in performance, it also underscores the importance of improving the reasoning capabilities of the visual modality.

In summary, baselines show promise on the task of video event understanding. However, it is obvious that this task poses new challenges with a huge room for improvement.

Acknowledgments

This study is partially supported by National Natural Science Foundation of China (62176016, 72274127), Guizhou Province Science and Technology Project: Research on Q&A Interactive Virtual Digital People for Intelligent Medical Treatment in Information Innovation Environment (supported by Qiankehe[2024] General 058), Capital Health Development Research Project(2022-2-2013), Haidian innovation and translation program from Peking University Third Hospital (HDCXZHKC2023203), and Project: Research on the Decision Support System for Urban and Park Carbon Emissions Empowered by Digital Technology - A Special Study on the Monitoring and Identification of Heavy Truck Beidou Carbon Emission Reductions.

References

Abu-El-Haija, S.; Kothari, N.; Lee, J.; Natsev, A. P.; Toderici, G.; Varadarajan, B.; and Vijayanarasimhan, S. 2016. YouTube-8M: A Large-Scale Video Classification Benchmark. In *arXiv:1609.08675*.

Anderson, P.; Fernando, B.; Johnson, M.; and Gould, S. 2016. SPICE: Semantic Propositional Image Caption Evaluation. In *ECCV*.

Banerjee, S.; and Lavie, A. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 65–72. Association for Computational Linguistics.

Bertasius, G.; Wang, H.; and Torresani, L. 2021. Is space-time attention all you need for video understanding? In *ICML*, volume 2, 4.

Chen, Z.; Ma, L.; Luo, W.; and Wong, K.-Y. K. 2019. Weakly-Supervised Spatio-Temporally Grounding Natural Sentence in Video. In *ACL*.

Cheng, F.; and Bertasius, G. 2022. Tallformer: Temporal action localization with a long-memory transformer. In *ECCV*, 503–521. Springer.

Damen, D.; Doughty, H.; Farinella, G. M.; Fidler, S.; Furnari, A.; Kazakos, E.; Moltisanti, D.; Munro, J.; Perrett, T.; Price, W.; and Wray, M. 2018. Scaling Egocentric Vision: The EPIC-KITCHENS Dataset. arXiv:1804.02748.

Diba, A.; Fayyaz, M.; Sharma, V.; Paluri, M.; Gall, J.; Stiefelhagen, R.; and Van Gool, L. 2020. Large Scale Holistic Video Understanding. In *ECCV 2020*, 593–610. Springer-Verlag.

Duan, H.; Zhao, Y.; Chen, K.; Lin, D.; and Dai, B. 2022. Revisiting skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2969–2978.

Fabian Caba Heilbron, B. G., Victor Escorcia; and Niebles, J. C. 2015. ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 961–970.

Feichtenhofer, C.; Fan, H.; Malik, J.; and He, K. 2019. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6202–6211.

Gao, J.; Sun, C.; Yang, Z.; and Nevatia, R. 2017. TALL: Temporal Activity Localization via Language Query. arXiv:1705.02101.

Goyal, R.; Ebrahimi Kahou, S.; Michalski, V.; Materzynska, J.; Westphal, S.; Kim, H.; Haenel, V.; Fruend, I.; Yianilos, P.; Mueller-Freitag, M.; Hoppe, F.; Thurau, C.; Bax, I.; and Memisevic, R. 2017. The "Something Something" Video Database for Learning and Evaluating Visual Common Sense. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Gu, C.; Sun, C.; Ross, D. A.; Vondrick, C.; Pantofaru, C.; Li, Y.; Vijayanarasimhan, S.; Toderici, G.; Ricco, S.; Sukthankar, R.; Schmid, C.; and Malik, J. 2018. AVA: A Video Dataset of Spatio-temporally Localized Atomic Visual Actions. arXiv:1705.08421.

Gu, C.; Sun, C.; Vijayanarasimhan, S.; Pantofaru, C.; Ross, D. A.; Toderici, G.; Li, Y.; Ricco, S.; Sukthankar, R.; Schmid, C.; and Malik, J. 2017. AVA: A Video Dataset of Spatio-Temporally Localized Atomic Visual Actions. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6047–6056.

Hendricks, L. A.; Wang, O.; Shechtman, E.; Sivic, J.; Darrell, T.; and Russell, B. 2017. Localizing Moments in Video with Natural Language. arXiv:1708.01641.

Idrees, H.; Zamir, A. R.; Jiang, Y.-G.; Gorban, A.; Laptev, I.; Sukthankar, R.; and Shah, M. 2017. The THUMOS challenge on action recognition for videos "in the wild". *Computer Vision and Image Understanding*, 155: 1–23.

Jhuang, H.; Gall, J.; Zuffi, S.; Schmid, C.; and Black, M. J. 2013. Towards understanding action recognition. In *International Conf. on Computer Vision (ICCV)*, 3192–3199.

Khan, Z.; Jawahar, C.; and Tapaswi, M. 2022. Grounded Video Situation Recognition. *Advances in Neural Information Processing Systems*, 35: 8199–8210.

Kingsbury, P.; and Palmer, M. 2003. Propbank: the next level of treebank. In *Proceedings of Treebanks and lexical Theories*, volume 3. Citeseer.

Ko, D.; Choi, J.; Choi, H. K.; On, K.-W.; Roh, B.; and Kim, H. J. 2023. MELTR: Meta Loss Transformer for Learning to Fine-tune Video Foundation Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20105–20115.

Krishna, R.; Hata, K.; Ren, F.; Fei-Fei, L.; and Carlos Niebles, J. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, 706–715.

Lei, J.; Yu, L.; Bansal, M.; and Berg, T. L. 2019. TVQA: Localized, Compositional Video Question Answering. arXiv:1809.01696.

Li, M.; Xu, R.; Wang, S.; Zhou, L.; Lin, X.; Zhu, C.; Zeng, M.; Ji, H.; and Chang, S.-F. 2022. Clip-event: Connecting text and images with event structures. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 16420–16429.

Liu, X.; Wang, Q.; Hu, Y.; Tang, X.; Zhang, S.; Bai, S.; and Bai, X. 2022. End-to-end temporal action detection with transformer. *IEEE Transactions on Image Processing*, 31: 5427–5441.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Miech, A.; Zhukov, D.; Alayrac, J.-B.; Tapaswi, M.; Laptev, I.; and Sivic, J. 2019. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. arXiv:1906.03327.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139, 8748–8763.

Radvansky, G. A.; and Zacks, J. M. 2017. Event boundaries in memory and cognition. *Current opinion in behavioral sciences*, 17: 133–140.

Rohrbach, A.; Torabi, A.; Rohrbach, M.; Tandon, N.; Pal, C.; Larochelle, H.; Courville, A.; and Schiele, B. 2016. Movie Description. arXiv:1605.03705.

Sadhu, A.; Gupta, T.; Yatskar, M.; Nevatia, R.; and Kembhavi, A. 2021a. Visual Semantic Role Labeling for Video Understanding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sadhu, A.; Gupta, T.; Yatskar, M.; Nevatia, R.; and Kembhavi, A. 2021b. Visual semantic role labeling for video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5589–5600.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. FaceNet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 815–823.

Shi, D.; Zhong, Y.; Cao, Q.; Ma, L.; Li, J.; and Tao, D. 2023. Tridet: Temporal action detection with relative boundary modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18857–18866.

Sigurdsson, G. A.; Varol, G.; Wang, X.; Farhadi, A.; Laptev, I.; and Gupta, A. 2016. Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding. arXiv:1604.01753.

Singh, A. K.; Srivastava, D.; and Tapaswi, M. 2024. "Previously on..." from Recaps to Story Summarization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, 13635–13646. IEEE.

Vedantam, R.; Zitnick, C. L.; and Parikh, D. 2015. CIDEr: Consensus-based image description evaluation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4566–4575.

Wang, R.; Chen, D.; Wu, Z.; Chen, Y.; Dai, X.; Liu, M.; Yuan, L.; and Jiang, Y.-G. 2023a. Masked video distillation: Rethinking masked feature modeling for self-supervised video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6312–6322.

Wang, X.; Chen, Y.; Ding, N.; Peng, H.; Wang, Z.; Lin, Y.; Han, X.; Hou, L.; Li, J.; Liu, Z.; et al. 2022. MAVEN-ERE: A Unified Large-scale Dataset for Event Coreference, Temporal, Causal, and Subevent Relation Extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 926–941.

Wang, X.; Li, J.; Zhu, L.; Zhang, Z.; Chen, Z.; Li, X.; Wang, Y.; Tian, Y.; and Wu, F. 2023b. Visevent: Reliable object tracking via collaboration of frame and event flows. *IEEE Transactions on Cybernetics*.

Wang, X.; Wu, J.; Chen, J.; Li, L.; Wang, Y.-F.; and Wang, W. Y. 2019. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4581–4591.

Wang, X.; Wu, J.; Chen, J.; Li, L.; Wang, Y.-F.; and Wang, W. Y. 2020. VATEX: A Large-Scale, High-Quality Multilingual Dataset for Video-and-Language Research. arXiv:1904.03493.

Wei, X.; Bai, Y.; Zheng, Y.; Shi, D.; and Gong, Y. 2023. Autoregressive Visual Tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9697–9706.

Xu, D.; Zhao, Z.; Xiao, J.; Wu, F.; Zhang, H.; He, X.; and Zhuang, Y. 2017. Video Question Answering via Gradually Refined Attention over Appearance and Motion. In *Proceedings of the 25th ACM International Conference on Multimedia*, 1645–1653. Association for Computing Machinery.

Xu, J.; Mei, T.; Yao, T.; and Rui, Y. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5288–5296.

Zhang, C.-L.; Wu, J.; and Li, Y. 2022. Actionformer: Localizing moments of actions with transformers. In *European Conference on Computer Vision*, 492–510. Springer.

Zhang, Z.; Zhao, Z.; Zhao, Y.; Wang, Q.; Liu, H.; and Gao, L. 2020. Where Does It Exist: Spatio-Temporal Video Grounding for Multi-Form Sentences. In *CVPR*.

Zhao, H.; Torralba, A.; Torresani, L.; and Yan, Z. 2019. HACS: Human Action Clips and Segments Dataset for Recognition and Temporal Localization. arXiv:1712.09374.

Zhong, Y.; Xiao, J.; Ji, W.; Li, Y.; Deng, W.; and Chua, T.-S. 2022. Video Question Answering: Datasets, Algorithms and Challenges. arXiv:2203.01225.

Zisserman, A.; Carreira, J.; Simonyan, K.; Kay, W.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; and Suleyman, M. 2017. In *The Kinetics Human Action Video Dataset*.