A Survey of Video Datasets for Grounded Event Understanding

Kate Sanders
Johns Hopkins University
ksande25@jhu.edu

Benjamin Van Durme
Johns Hopkins University
vandurme@jhu.edu

Abstract

While existing video benchmarks largely consider specialized downstream tasks like retrieval or question-answering (QA), contemporary multimodal AI systems must be capable of well-rounded common-sense reasoning akin to human visual understanding. A critical component of human temporal-visual perception is our ability to identify and cognitively model "things happening", or events. Historically, video benchmark tasks have implicitly tested for this ability (e.g., video captioning, in which models describe visual events with natural language), but they do not consider video event understanding as a task in itself. Recent work has begun to explore video analogues to textual event extraction but consists of competing task definitions and datasets limited to highly specific event types. Therefore, while there is a rich domain of event-centric video research spanning the past 10+ years, it is unclear how video event understanding should be framed and what resources we have to study it. In this paper, we survey 105 video datasets that require event understanding capability, consider how they contribute to the study of robust event understanding in video, and assess proposed video event extraction tasks in the context of this body of research. We propose suggestions informed by this survey for dataset curation and task framing, with an emphasis on the uniquely temporal nature of video events and ambiguity in visual content.

1. Introduction

Until recently, video understanding has focused on specialized models capable of highly specific downstream tasks [4, 50, 52, 109, 115]. As general-purpose AI assistants like Gemini have begun to accept video inputs [88], it is no longer sufficient to rely on suites of video models that are tailored to perform well on a small set of video benchmarks. The next generation of video models will need to facilitate robust, generalizable, and human-like visual reasoning over unconstrained video inputs and ground them to natural language text. These systems, then, necessarily require an ability to interpret and synthesize low-level visual fea-
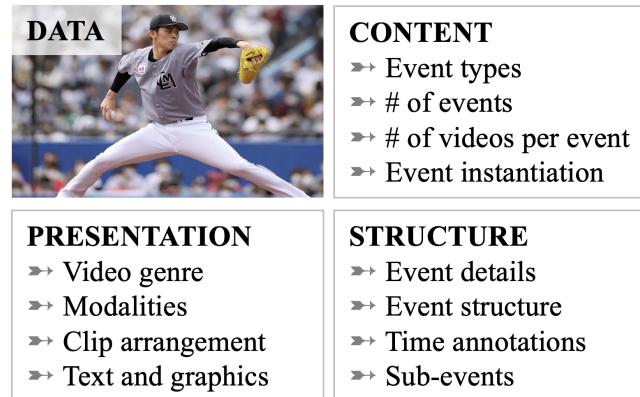

Figure 1. Illustration of an example video dataset paired with the associated topics we cover in our survey, detailed in Section 2. These topics are designed to help us answer the question of what resources we have for robust video event understanding, targeting the *content*, *presentation*, and *structure* of video events.

tures into complex semantic ideas that humans care about. A key component of this is the fundamental task of identifying and classifying "things that are happening", notions often modeled in natural language processing as "events" [72]. The purpose of this survey is to assess the breadth of video datasets and tasks introduced in the last decade to answer the question: *Given the goal of developing vision-language models that can interpret events depicted in video as humans can, what resources do we have for working towards this goal, and are they sufficient?*

There are a variety of datasets and approaches for event extraction in text documents [22, 62, 67, 124]. Some researchers have leveraged the distinct relationship between visual content and textual event semantics to explore visual event semantics for single images, ranging from k-means classification tasks to semantic role labeling and grounding [31, 85, 131]. Comparatively, the work in video event understanding is understudied and less structured, despite the richness introduced to the domain through the temporality of video. While many video benchmarks involve identifying some form of events, the domain, semantic complex-

ity, and representation of events diverge substantially across benchmarks as the primary focus is typically on their downstream task. The few datasets that do introduce event extraction as the core task propose diverging task descriptions and limited data to facilitate their study [5, 11, 95]. This makes it difficult to establish a notion of what resources the field has for the study of event understanding, and difficult to establish a notion of what an event should even be in the context of video data.

These uncertainties motivate our survey. We construct a framework to analyze video datasets with a focus on (1) what events are presented, (2) how they are presented, and (3) how they are interpreted, pictured in Figure 1. We consider a set of 105 video datasets and organize them into categories based on their intended task and construction approach. We analyze these datasets using our framework and provide a summary of our primary findings. We provide our full dataset-by-dataset analysis to the community in human- and machine-readable formats that will allow users to filter and identify datasets that meet their desired needs.[1] We then introduce a taxonomy of video event structures (Figure 2) to study the representational differences among events in different video datasets. We draw from this analysis to assess the relationships between existing video event extraction benchmarks and the surveyed datasets. Finally, we make suggestions for future dataset construction and video event understanding formulation, focusing on the temporal structure of video events, the relationship between formal event structures and natural language descriptions, and uncertainty in video understanding.

2. Survey Methodology

To guide our analysis of video understanding resources, we establish three primary dataset considerations: What events are presented (content), how they are presented (presentation), and how they are interpreted (structure).

2.1. Dataset Content

As we are concerned with studying events through video, we must assess what events are in existing datasets. We consider not only the types of events recorded, but how the recorded events occur, because naturally occurring events often have notable visual differences from staged content for TV shows, or events recreated for a benchmark. For each dataset, we ask:

- **What is the domain of events depicted?** (Sports, daily activities, social interactions, . . .)
- **How many events are depicted?** (100 activity instances, 4K videos, 70 event classes, . . .)

- **Are there multiple videos included per event?** (Yes/No)
- **How were the events instantiated?** (Natural or staged)

2.2. Dataset Presentation

Video content should not only be assessed for the semantic content it depicts but how it is presented through the audio-visual medium. This assessment largely pertains to choices surrounding the filming and editing of videos, such as filming purpose, video quality, and additional content added in post. We ask four general questions to characterize the presentation of visual events:

- **What is the purpose of the videos?** (Home videos, news broadcasts, dataset content, . . .)
- **What modalities are recorded?** (Video only, video + audio, video + depth info, . . .)
- **Are the videos compilations?** (Single video stream or concatenated clips)
- **Are text or graphics added in post?** (Yes/No)

2.3. Dataset Structure

Event-centric videos can be explained through many distinct conceptions of events. It is important to identify what semantic event structures existing benchmarks implicitly promote, as they dictate what we train vision models to care about and how to structure information. We find that existing structures can be understood through two distinct axes: how event details are selected and organized, and how the temporal nature of videos are incorporated. A sample of video event structures identified in our survey and their relationships to these axes are illustrated in Figure 2. To analyze dataset event structures, we ask the questions:

- **What event details are included?** (Actions, subjects and actions, event templates, . . .)
- **How are the details organized?** (Event structures or natural language strings)
- **How is the timing of events structured?** (No annotations, start and end points, or multiple events with temporal relationships)
- **Are hierarchical events considered?** (Yes/No)

3. Dataset Collection

Hundreds of video datasets have been released in the last decade. We limit the scope of considered datasets to those whose task description involves identifying or reasoning over at least one action, omitting datasets with applications strictly concerning "low-level" features like shapes, textures, composition, individual objects, trajectories, and attributes. This includes tasks such as object tracking, texture detection, and segmentation.

Many of the answers to questions detailed in Section 2 are highly correlated with the dataset construction method,

---

[1]Data available at `https://github.com/katesanders9/grounded-events`

AXIS 1

Axis Category	Propositional	Davidsonian	Neo-Davidsonian	Strings

| Example Dataset | Sports-1M | HOMAGE | MovieGraphs | VaTeX |

AXIS 2

Axis Category	Time-Agnostic	Temporal	Compositional	Hierarchical

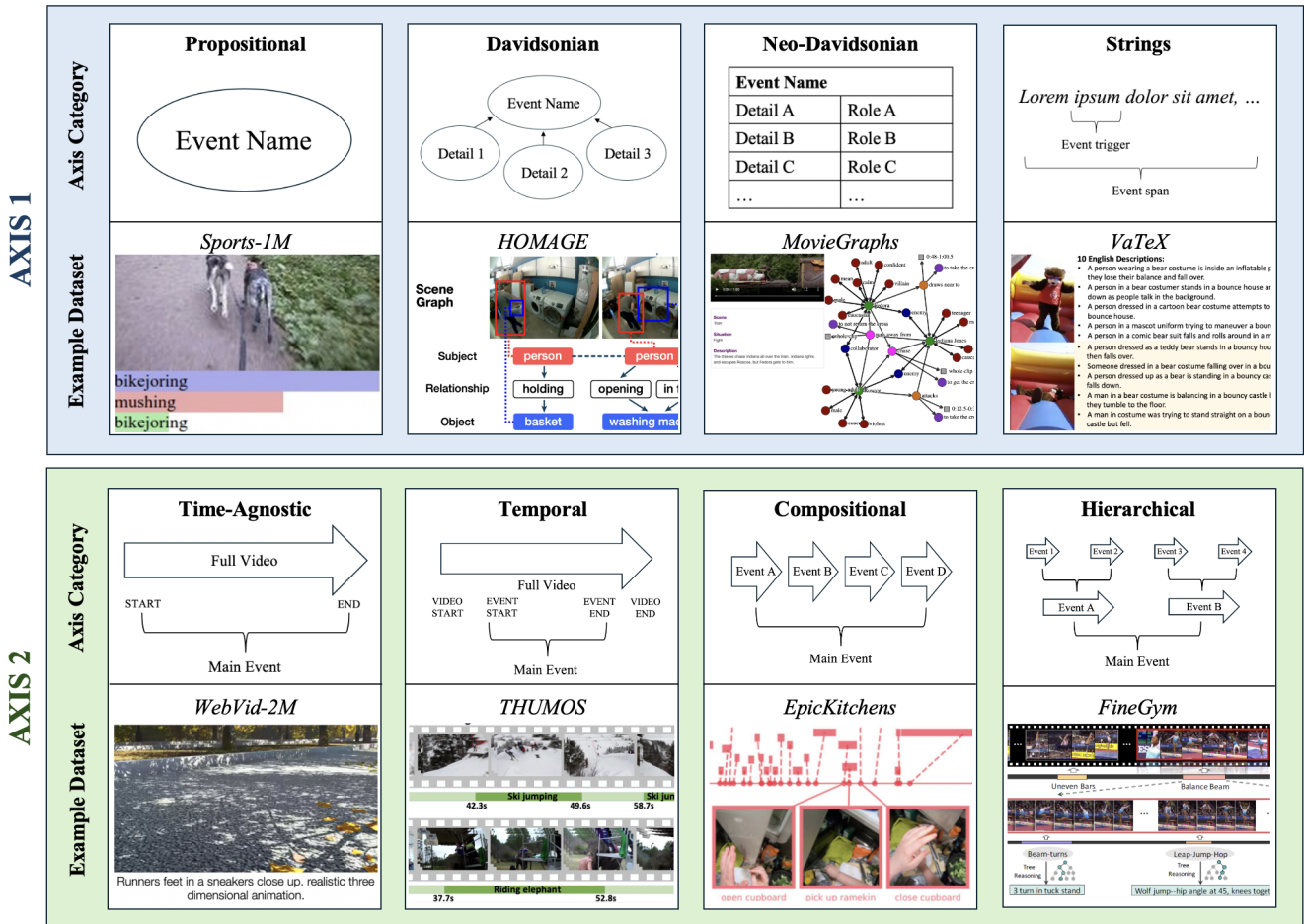| Example Dataset | WebVid-2M | THUMOS | EpicKitchens | FineGym |

Figure 2. Illustration of the two axes along which we classify event structures presented in video datasets. Alongside each category name, we include a graphic depicting the general idea of the method, a dataset that falls into this structure category, and an image of the dataset presentation to show how these methods exist in real data (dataset images are taken from the original publication). The top row shows how details are organized, and the bottom row shows how time is incorporated into the structure.

which is often dictated by the benchmark task. So, in our survey we organize datasets by general application category, and when applicable, unique construction methodology. We identify five primary application categories: Basic action recognition, classification, and localization; hierarchical action recognition, classification, and localization; scene parsing and entity relation detection; retrieval and captioning; and question-answering.

We find that videos surveyed typically belong to one of the following four production formats: Uncurated non-professional footage (YouTube uploads, vlogs, etc.); curated non-professional footage (e.g., scripted scenarios acted out by humans recruited for dataset construction); professional news footage (broadcasts covering real-world events); and professional narrative content (TV shows and movies). The latter two formats produce unique, format-specific artifacts that influence the content of the dataset

compared to unprofessional footage. In a similar vein, we find that multilingual video datasets typically follow unique and specific construction methods as well. So, we add three additional categories: movie and TV show footage; news broadcasts; and multilingual content.

Finally, a small number of datasets do not fall under any of the discussed categories, and so we classify these as miscellaneous datasets as a final ninth category.

We search through research paper and dataset literature databases to identify a set of 105 datasets that have either been released or frequently used in the last ten years. We include a list of these datasets in Table 1.

4. Video Analysis

Below are report brief summaries of our findings based on our "content" and "presentation" considerations.

Category	Description	Datasets
Action Recognition	Short videos of humans performing actions. Some datasets consider a complex definition of "actions" and include things like group activities.	UCF-101[108], Kinetics[9], MomentsInTime[75], Multi-MomentsInTime[77], Sports-1M[48], THUMOS[42], ActivityNet[8], Charades[104], Charades-Ego[103], MMAct[51], VideoLT[137], YouTube-8M[2], HVU[19], VLOG[21], DiDeMo[3], HLVU[16], Volleyball[41], NBA[129], CCV[47], FCVid[46], ShanghaiTech Campus[69], Animal Kingdom[79], Assembly101[99], Something-Something[25], HMDB51[55], AVA Actions[29], MTL-AQA[82], HowTo100M[74], Multi-HowTo100M[38]
Hierarchical Recognition	Annotation protocols structure events as "building blocks" through which complex, abstract events can be structured.	Homage [86], LEMMA [45], FineGym [100], MultiTHU-MOS [133], COIN [110], MPII Cooking2 [91], Epic Kitchens [17], YouCook2 [140], Multi-YouCook2 [94], MUSES [68], Breakfast [54], Diving48 [64], NewsNet [122]
Scene Graphs	Scene graph datasets place a strong emphasis on relationships between many entities and/or actions in videos.	MOMA [70], VidSitu [95], VM2E2 [11], TVEE [118], MovieNet [39], MovieGraphs [117], AIDA Scenario 1 [113]
Retrieval/Captioning	Videos of variable complexity, content, and genre, generally paired with unique natural language descriptions.	VaTeX [119], WebVid-2M [139], VTW [136], TGIF [65], V3C [92], MSR-VTT [127], VideoStory [33], ActivityNet-Captions [8], ActivityNet-Entities [139], TVR [59], TVC/TVR [59], TRECVID [105], EventNet [132], MultiVENT [96], Marine Video Kit [114], InternVid [120], MSVD [12], FIVR-200K [53], VITT [37], Spoken MomentsInTime [76], VideoCC3M [78], MultiVENT [96], TextVR [123], MPII [90], MAD [106], CondensedMovies [6], HowTo100M [74], Multi-HowTo100M [38], mTVR [56], VATEX [119], How2 [101], RUDDER [1], ChinaOpen [10]
Q&A	(Typically) semantically complex videos with paired natural language question-answer pairs. Sometimes the questions are multiple choice.	MSR-VTT-QA [126], HowToVQA69M [74], TVQA [57], MovieQA [111], NEXT-QA [125], NewsVideoQA [43], SocialIQ [134], EgoSchema [73], Perception Test [84], DramaQA [15], AGQA [28], STAR [121], Ego4D [26]
Movie/TV Shows	Datasets of video clips taken from content produced to convey some sort of narrative to humans. Annotations vary substantially between datasets.	HLVU [16], MPII [90], MovieNet [39], MAD [106], MovieGraphs [117], CondensedMovies [6], TVR [59], TVC [59], TVQA [57], TVQA+ [58], MovieQA [111]
News	Video clips taken from content produced specifically to convey news and current events to humans. Often OCR-heavy.	TextVR [123], NewsVideoQA [43], NewsNet [122], VM2E2 [11], TVEE [118], AcTiV [135], MultiVENT [96]
Multilingual	Datasets explicitly produced with the goal of multilinguality, either with videos produced for speakers of multiple languages or multilingual annotations.	Multi-YouCook2 [94], Multi-HowTo100M [38], mTVR [56], VATEX [119], How2 [102], 3MASSIV [32], RUDDER [1], ChinaOpen [10]
Misc	Other datasets of note that do not fall into a traditional task category. Many of these datasets facilitate multiple tasks and so their construction is unique.	Spoken Moments [76], HACS [138], TVSum [107], TEMPO [34], How-2 [102], QuerYD [81], VIOLIN [66], CrossTask [141], HD-VILA-100M [128], VALUE [61], VLEP [60], MultiHiEve [5], 3MASSIV [32], VEATIC [89], EmoMV [112], VideoInstruct [71]

Table 1. Breakdown of surveyed datasets by task type, with brief summaries describing the dataset type.

4.1. Results

Action Recognition Traditional action recognition benchmarks often largely consider simple events, building on work in pose recognition and action recognition in the image domain. However, a number of datasets have extended the concept of action recognition to more complex "actions" including longer, abstract events like "group

activity" recognition. Most of these datasets are limited in scope, considering a small domain of event types, but some include a wider domain with the aim of covering a broad range of daily human experiences.

Hierarchical Action Recognition Some action recognition datasets go beyond general event recognition by decomposing actions into hierarchies of sub-actions. Cooking video datasets in particular lend themselves well to this domain. Like the standard action recognition datasets, and especially due to the level of annotation detail necessary, most of these datasets (e.g., cooking datasets) do not consider a wide range of content. MultiTHUMOS [133] is notable in its significant range of event types.

Scene Graphs Scene graph datasets often involve visual relation extraction and include annotations for objects' relations to each other and their locations within the video - this tends to push the dataset focus toward more complex events. These datasets often include many event details and make up some of the only event-centric datasets with temporal/spatial annotations.

Retrieval and Captioning A number of video retrieval datasets focus on short, simplistic events that can be depicted in a few seconds, but others include more sophisticated content. The task limits these datasets in that the entirety of the video content must be able to be described in a single sentence, which doesn't allow for rich event semantics the way news content or movies can. However, the range of event types is often quite large.

Question-Answering Question-answering datasets are often built by re-annotating existing video datasets, often by leveraging natural language captions or narration. QA datasets lend well to narratively and socially complex video content often produced for TV, whereas other QA datasets follow more traditional formats that focus less on high-level reasoning. This latter set of QA datasets may consider a wider range of video genres than those focusing on TV content, aligning more with the range seen in captioning datasets described above.

Movie/TV Datasets Building on the overview of QA datasets, a number of other datasets depicting movies and TV shows have been introduced. Naturally, high-complexity narrative videos uniquely facilitate complex, narrative-driven reasoning work, and a high proportion of these datasets include temporal and spatial annotations.

Multilingual Content Researchers typically build multilingual datasets by either directly translating existing datasets or collecting new, organic multilingual data.

Datasets collected from organic multilingual content are often limited in genre, with a focus on studying the multilinguality of one specific domain.

News Videos It is difficult to generate text documents for video content, but news videos allow for easy "real world" text document mapping. Most of these datasets are limited in presentation format but can include rich on-screen textual data added in post.

4.2. Discussion

The vast majority of surveyed datasets consider semantically simpler events than those typically found in text-based event extraction datasets [27, 44, 116]. The datasets that do include complex event types tend to have a limited range of event genres, and the videos often come from highly edited or staged content like TV shows. While teaching models to understand low-level events is a necessary step for understanding more complex events, as vision models continue to improve it will become increasingly necessary to have datasets that present a wide range of complex event types.

On a similar note, it is difficult to collect data from a wide range of presentation formats using traditional video dataset construction approaches (i.e. not using synthetic videos generated by models), and in some tasks, it is unnecessary. This is reflected in the datasets surveyed: Very few consider multiple filming or editing methods, and many do not include multiple filming locations, setups, or participants. However, as video-language models for daily task assistance grow commonplace, it will grow increasingly important for models to be familiar with a wide range of video presentation methods to mitigate geographic, linguistic, and cultural bias. Therefore, we argue that training models on datasets constructed through multiple presentation methods is critical, which necessitates being conscientious of such considerations when constructing datasets.

Surveyed datasets with a high level of semantic complexity sometimes require event understanding beyond identifying the most likely events present in a video. Social reasoning datasets may require high-level probabilistic inference, and multiple choice QA datasets may require comparing two likely answers to identify which is *better*. This distinction between basic event understanding and high-level inference over possible events is important when considering the type of event understanding necessary for a task.

5. Structure Analysis

To ascertain how vision models are expected to understand events, we consider how events are structured in each benchmark. We find that semantic annotations make it easy to isolate the concrete event information necessary to perform a dataset task, agnostic to the annotation format

through which the information is presented. Below, we outline past work in studying event structures and connect this work to video events. The axes are illustrated in Figure 2 alongside example datasets.

5.1. Events and Language

Davidson is credited [72, 93] as an early influential proponent for considering events as spatiotemporal things, proposing a linguistic framework through which they could be modeled formally [18], alongside Reichenbach who proposed related notions involving logical forms for events [87]. Since then, linguists have built on these ideas, introducing model extensions [35, 83, 97, 98] and adapting the core ideas to tasks in natural language processing [36]. A primary application of event semantics is the extraction of events within text documents. Much of this work is based on the idea of defining events through frames (templates) [7]. A variety of datasets for this task have been introduced [14, 23, 63] and robust methods have been proposed [13, 24, 67].

Some researchers have leveraged the distinct relationship between visual content and textual event semantics to explore visual event semantics. Initially, visual event recognition was limited to k-way classification tasks [20, 30, 108], but Gupta et al. [31] expanded the definition of a visual event through the introduction of "visual semantic role labeling", or the task of mapping images to a limited taxonomy of scenarios following a subject-object-action template. A "grounded" version of this task was proposed by Yang et al. [130]. These ideas were built on by Yatskar et al. [131] and Pratt et al. [85] who presented a more complex task of mapping images to template-based scenarios, but with a much wider range of templates following the FrameNet ontology [7].

In the following sections, we outline the similarities and differences between these event structures in text and images and the event structures necessary for explaining video event benchmarks. We find that there are two primary axes along which video event interpretations diverge across datasets, which we illustrate through a set of formal semantic annotations.

5.2. Semantic Structure

The first axis considers how the details of the event are organized. We can understand this axis through a comparison to event structuring in formal semantics. Below, we characterize differences among surveyed detail organization approaches. For each category, we consider the video depicted in Figure 1 (a video of a pitcher throwing a ball at a baseball game) as a running example and outline how this video might be structured through each primary framework.

Propositional Structures The simplest semantic structures only consider the verb: Events are structured as propositions with no defined details. The video of the baseball game, therefore, would simply be structured as

$$Throwing \quad (1)$$

This modeling framework often applies to datasets depicting very simple events, such as action recognition datasets in which each video is mapped to a short label.

Davidsonian Structures More complex modeling methods include a set of details for each event in the dataset. In some benchmarks, the same set of details is considered for each event, regardless of semantic distinctions between them (generally a proto-agent, proto-patient, and predicate). The example video, then, could be structured as

$$\exists e, x, y. \; Throwing \, (pitcher \, (x) \, , baseball \, (y) \, , e) \quad (2)$$

More complex action recognition and hierarchical action recognition datasets often fall into this category, but the majority of mid-level complexity datasets require more complex structures to properly represent the events.

Neo-Davidsonian Structures A third class of structures similarly structures the details associated with events, but treats them as variable "template fields" that may differ depending on event type. Each detail associated with an event is typically labeled with a "template slot type", used to define the semantic relationship between itself and the encompassing event. Using this framework, the baseball pitch could be structured as

$$\exists e, x, y, z. \, [\, Throwing \, (e) \, \wedge \, Agent \, (pitcher \, (x) \, , e) \, \wedge$$
$$Obj \, (baseball \, (y) \, , e) \, \wedge \, At \, (third \, inning \, (z) \, , e) \, ...] \quad (3)$$

This structure covers many datasets with more complex annotations. The majority of mid-tier complexity event datasets fall into this category, as they can be structured fairly easily through this structure.

String-based Structures Finally, many datasets do not include formally organized structures of the depicted events. Instead, they are structured only using natural language descriptions, or strings. There generally is not an established structure that the strings abide by, and so there is typically no way to ensure which structuring approach is most appropriate for an entire dataset of labels. The example video could be represented as

$$\textit{The pitcher throws the ball during the third inning.} \quad (4)$$

The majority of captioning, retrieval, and QA datasets fall into this category, as videos in these datasets typically are paired with long-form natural language descriptions (or in

the case of QA, a two-part or multi-part natural language description split across questions and answers). The events in this category tend to be more complex than those covered by the other event structures.

5.3. Semantic Structure: Discussion

It should be noted that while these categories can help us categorize and understand visual event structures, they do not serve as complete explanations: Two event structures that both classify as Neo-Davidsonian structures may differ in their detail and adherence to traditional Davidsonian structuring methods, and the same applies to the temporal axis organizational framework. We argue that Davidsonian and Neo-Davidsonian event representations are superior to propositional structures, as they reduce ambiguity and underspecification. But, based on the surveyed datasets, the details of the structuring method are domain-dependent.

There is a notable divide between datasets in terms of whether they include formal event structures or natural language string descriptions. Few datasets include both. We argue that formal representations and fully string-based representations are both important for robust visual event understanding. Formal structures are useful for evaluation and organization, while natural language allows for better human alignment.

5.4. Temporal Structure

The second axis concerns how time within individual videos influences the event structure, which diverges from most work in traditional language semantics. Again, we consider four main categories into which the surveyed event structures fall, and consider how our structures can be organized using this framework.

Time-Agnostic Structures Some benchmarks do not structure the temporal position at which events exist in a given video: One video simply corresponds to one event. This can roughly be structured as

$$\exists\, e.\ Event\,(e) \tag{5}$$

Some action recognition datasets are structured this way, and many captioning and retrieval datasets involve short enough videos that time annotations are unnecessary.

Temporal Structures On the other hand, many datasets' event structures include the time with respect to the full input video at which an event takes place:

$$\exists\, e, t.\ [\,Event\,(e)\ \wedge\ At\,(e, t_1, t_2)\,] \tag{6}$$

Action localization datasets, a subset of action recognition datasets, generally use temporal event structures. Some other datasets also structure events this way, especially those using string-based methods. For these datasets,

t is sometimes loosely defined, but still defined nonetheless.

Compositional Structures Some benchmarks include multiple "sub-events" within the same video. While these sub-events are sometimes referred to as full events within the task or dataset description, we argue that through this formulation, more semantically interesting events are implicitly modeled within videos and task outputs, warranting a distinction. The structuring of these events can be illustrated as

$$\exists\, e, t.\ [\,Event\,(e_x)\ \wedge\ At\,(e_x, t_{1(x)}, t_{2(x)})\ \wedge$$
$$Event\,(e_y)\ \wedge\ At\,(e_y, t_{1(y)}, t_{2(y)})\,] \tag{7}$$

The majority of surveyed dataset structures are compositional in nature, especially those with natural language descriptions. However, string-based structures are difficult to classify into one temporal approach or the other due to the fact that there is often little structure incentivizing consistent temporal labeling across all videos in a dataset.

Hierarchical Structures Building on the ideas introduced by compositional structures, hierarchical structures also consider multiple "sub-events" within a single video [80]. However, through this structuring paradigm, the sub-events are explicitly defined as sub-events, and multiple layers can be established. The compositional complexity of these structures differs among approaches. For its formulation, we use the predicate *Comprises(x,y,z)* where sub-events *x* and *y* exist within full event *z* (this predicate can be further formalized through episodic logic [40] as $[x * z] \wedge [y * z]$):

$$\exists\, e, t.\ [\,Event\,(e_x)\ \wedge\ At\,(e_x, t_{1(x)}, t_{2(x)})\ \wedge\ Event\,(e_y)$$
$$\wedge\ At\,(e_y, t_{1(y)}, t_{2(y)})\ \wedge\ \boldsymbol{Event\,(e)}\ \wedge\ \boldsymbol{Comprises\,(e_x, e_y, e)}\,]$$
$$\tag{8}$$

Hierarchical structures are typically reserved for specialized datasets like hierarchical action recognition benchmarks. Some string-based structures also model events hierarchically, but again, in these cases, it is typically more difficult to discern and classify them as such.

5.5. Temporal Structure: Discussion

The majority of datasets emphasize the timing of events: While the timing of events and their relationships to each other are often considered an afterthought in text-centric event modeling work, it is clear that it is difficult to avoid including temporal relationships in video, and so we argue that optimal methods for structuring the temporal aspect of events and formalizing them should be explored further. Along these lines, it is sensible that many datasets are compositional or hierarchical in nature. While sub-events in

text can easily be summarized through higher-level events, no such analogue exists in the video medium: If an event is visible, a non-empty subset of the subevents constituting it must also be shown. Therefore, it makes sense for subevents to be structured as hierarchical structures.

6. Video Event Extraction Tasks

There are multiple incipient lines of work considering video event extraction as a task, each with its own distinct task definition. We outline and analyze approaches below:

6.1. Multimodal Event Extraction

Chen et al. [11] argue that video events should be considered in tandem with natural language descriptions. They define Multimodal Event Extraction (MMEE) as the task of taking in a video and text sentence and classifying the event type, event trigger text span, argument text spans, and bounding boxes for each argument.

Chen et al. provide a sound video-centric extension to the text-only event extraction task. However, it does not consider compositional or hierarchical events and requires perfect event mappings to exist between text and video. Additionally, text content is required for the task, and robust event understanding necessitates the ability to identify visual events even without ground truth text documents.

6.2. Video Semantic Role Labeling

Sadhu et al. [95] build on Visual Semantic Role Labeling [131] through Video Semantic Role Labeling (VidSRL), the task of taking in a video and predicting a set of events as defined in VSRL (an event name and a set of related roles), the set of which constitutes a "situation", and labeling the relationships between the events. Later, Khan et al. add bounding the visual roles to the task [49].

Sadhu et al.'s task of VidSRL requires no text documents as input and considers compositional events, and Khan et al.'s modifications add spatial grounding annotations. However, all events in this task are defined through 2-second intervals, which is not guaranteed to align with the temporality of the actual depicted events. Furthermore, the task only considers simple events in short clips, so VidSRL may not be optimal for long clips and higher-level events.

6.3. Multimodal Event Hierarchy Extraction

Ayyubi et al. [5] build on VidSRL to construct an event hierarchy extraction task, Multimodal Event Hierarchy Extraction (MEHE): Given a text article and video, the aim is to identify all event coreferences between the two modalities and structure their hierarchical relationships through the pairwise labels "hierarchical" and "identical".

MEHE models both hierarchical relationships between events in video and hierarchy level mismatches between events in video and text documents. The task defines events as single shots within an edited video instead of 2s clips, but this definition is not necessarily appropriate for all data. Additionally, events are only defined through single predicates, and it also requires text documents as input.

6.4. Discussion

Each task definition facilitates unique aspects of robust video event understanding, and so a combination of these definitions may serve as a more universal task, such as a combination of the low-level event modeling used in MMEE, the hierarchical event modeling explored by MEHE, the ability to map visual content to natural text of both approaches, and the text document independence of VidSRL. Such a task would theoretically enable systems to identify temporally defined atomic events, consolidate the events into high-level event hierarchies, structure events at all levels through Neo-Davidsonian semantics, and map coreferences to natural language text documents, addressing all primary event modeling points considered in our event structure analysis.

As explored in our discussion of semantically complex video datasets in Section 4.2, difficult benchmarks can require probabilistic inference to complete functions like making comparisons between multiple likely labels and achieving human-like social reasoning. So, for this genre of high-level understanding, mapping a video to a single hierarchy of event structures may be insufficient for empowering robust event understanding. We propose that in addition to the ideas introduced by these tasks, video event understanding models should be constructed to possess an understanding of event uncertainty for robust reasoning that better reflects human visual reasoning ability.

7. Conclusion

We present a comprehensive analysis of 105 event-centric video datasets, introduce formal semantic annotations to organize the implicit event structures of these datasets and assess the strengths and weaknesses of such structures. We use this survey to consider the relationship between proposed video event extraction tasks and general video event understanding research. Informed by our survey, we propose that natural language descriptions and formal event structures are both helpful for grounded event understanding, that the temporality and hierarchy of video events are critical aspects of video event modeling, and that uncertainty is a core component of modeling video events in practical and high-complexity scenarios. As video generation models improve and datasets begin to include synthetic content, we recommend that researchers producing such resources consider the visual, cultural, and presentation diversity of the generated videos.

References

[1] Jayaprakash A, Abhishek, Rishabh Dabral, Ganesh Ramakrishnan, and Preethi Jyothi. Rudder: A cross lingual video and text retrieval dataset, 2021. 4

[2] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016. 4

[3] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812, 2017. 4

[4] Evlampios Apostolidis, Eleni Adamantidou, Alexandros I. Metsai, Vasileios Mezaris, and Ioannis Patras. Video summarization using deep neural networks: A survey. *Proceedings of the IEEE*, 109(11):1838–1863, 2021. 1

[5] Hammad A Ayyubi, Christopher Thomas, Lovish Chum, Rahul Lokesh, Yulei Niu, Xudong Lin, Long Chen, Jaywon Koo, Sounak Ray, and Shih-Fu Chang. Multimodal event graphs: Towards event centric understanding of multimodal world. *arXiv preprint arXiv:2206.07207*, 2022. 2, 4, 8

[6] Max Bain, Arsha Nagrani, Andrew Brown, and Andrew Zisserman. Condensed movies: Story based retrieval with contextual embeddings, 2020. 4

[7] Collin F Baker, Charles J Fillmore, and John B Lowe. The berkeley framenet project. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*, 1998. 6

[8] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, 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. 4

[9] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019. 4

[10] Aozhu Chen, Ziyuan Wang, Chengbo Dong, Kaibin Tian, Ruixiang Zhao, Xun Liang, Zhanhui Kang, and Xirong Li. Chinaopen: A dataset for open-world multimodal learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23. ACM, Oct. 2023. 4

[11] Brian Chen, Xudong Lin, Christopher Thomas, Manling Li, Shoya Yoshida, Lovish Chum, Heng Ji, and Shih-Fu Chang. Joint multimedia event extraction from video and article. *arXiv preprint arXiv:2109.12776*, 2021. 2, 4, 8

[12] David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 190–200, 2011. 4

[13] Yunmo Chen, William Gantt, Weiwei Gu, Tongfei Chen, Aaron Steven White, and Benjamin Van Durme. Iterative document-level information extraction via imitation learning. *arXiv preprint arXiv:2210.06600*, 2022. 6

[14] Nancy Chinchor. Overview of muc-7. In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29-May 1, 1998*, 1998. 6

[15] Seongho Choi, Kyoung-Woon On, Yu-Jung Heo, Ahjeong Seo, Youwon Jang, Minsu Lee, and Byoung-Tak Zhang. Dramaqa: Character-centered video story understanding with hierarchical qa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1166–1174, 2021. 4

[16] Keith Curtis, George Awad, Shahzad Rajput, and Ian Soboroff. Hlvu: A new challenge to test deep understanding of movies the way humans do. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*, pages 355–361, 2020. 4

[17] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, pages 720–736, 2018. 4

[18] Donald Davidson. The logical form of action sentences. 1967. 6

[19] Ali Diba, Mohsen Fayyaz, Vivek Sharma, Manohar Paluri, Jürgen Gall, Rainer Stiefelhagen, and Luc Van Gool. Large scale holistic video understanding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 593–610. Springer, 2020. 4

[20] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111:98–136, 2015. 6

[21] David F Fouhey, Wei-cheng Kuo, Alexei A Efros, and Jitendra Malik. From lifestyle vlogs to everyday interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4991–5000, 2018. 4

[22] Giacomo Frisoni, Gianluca Moro, and Antonella Carbonaro. A survey on event extraction for natural language understanding: Riding the biomedical literature wave. *IEEE Access*, 9:160721–160757, 2021. 1

[23] William Gantt, Shabnam Behzad, Hannah YoungEun An, Yunmo Chen, Aaron Steven White, Benjamin Van Durme, and Mahsa Yarmohammadi. Multimuc: Multilingual template filling on muc-4. *arXiv preprint arXiv:2401.16209*, 2024. 6

[24] Jun Gao, Huan Zhao, Changlong Yu, and Ruifeng Xu. Exploring the feasibility of chatgpt for event extraction. *arXiv preprint arXiv:2303.03836*, 2023. 6

[25] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017. 4

[26] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 18995–19012, 2022. 4

[27] Ralph Grishman and Beth M Sundheim. Message understanding conference-6: A brief history. In *COLING 1996 volume 1: The 16th international conference on computational linguistics*, 1996. 5

[28] Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. Agqa: A benchmark for compositional spatio-temporal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11287–11297, 2021. 4

[29] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6047–6056, 2018. 4

[30] Abhinav Gupta, Aniruddha Kembhavi, and Larry S Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. *IEEE transactions on pattern analysis and machine intelligence*, 31(10):1775–1789, 2009. 6

[31] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015. 1, 6

[32] Vikram Gupta, Trisha Mittal, Puneet Mathur, Vaibhav Mishra, Mayank Maheshwari, Aniket Bera, Debdoot Mukherjee, and Dinesh Manocha. 3massiv: Multilingual, multimodal and multi-aspect dataset of social media short videos, 2022. 4

[33] Amirhossein Habibian, Thomas Mensink, and Cees GM Snoek. Videostory: A new multimedia embedding for few-example recognition and translation of events. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 17–26, 2014. 4

[34] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with temporal language, 2018. 4

[35] James Higginbotham. The logic of perceptual reports: An extensional alternative to situation semantics. *The journal of philosophy*, 80(2):100–127, 1983. 6

[36] Frederik Hogenboom, Flavius Frasincar, Uzay Kaymak, and Franciska De Jong. An overview of event extraction from text. *DeRiVE@ ISWC*, pages 48–57, 2011. 6

[37] Gabriel Huang, Bo Pang, Zhenhai Zhu, Clara Rivera, and Radu Soricut. Multimodal pretraining for dense video captioning. *arXiv preprint arXiv:2011.11760*, 2020. 4

[38] Po-Yao Huang, Mandela Patrick, Junjie Hu, Graham Neubig, Florian Metze, and Alexander Hauptmann. Multilingual multimodal pre-training for zero-shot cross-lingual transfer of vision-language models, 2021. 4

[39] Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. Movienet: A holistic dataset for movie understanding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 709–727. Springer, 2020. 4

[40] CHUNG HEE I-IWANG and Lenhart K Schubert. Episodic logic: A situational logic for natural language processing. 1993. 7

[41] Mostafa S Ibrahim, Srikanth Muralidharan, Zhiwei Deng, Arash Vahdat, and Greg Mori. A hierarchical deep temporal model for group activity recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1971–1980, 2016. 4

[42] Haroon Idrees, Amir R Zamir, Yu-Gang Jiang, Alex Gorban, Ivan Laptev, Rahul Sukthankar, and Mubarak Shah. The thumos challenge on action recognition for videos "in the wild". *Computer Vision and Image Understanding*, 155:1–23, 2017. 4

[43] Soumya Jahagirdar, Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. Watching the news: Towards videoqa models that can read, 2023. 4

[44] Sarthak Jain, Madeleine Van Zuylen, Hannaneh Hajishirzi, and Iz Beltagy. Scirex: A challenge dataset for document-level information extraction. *arXiv preprint arXiv:2005.00512*, 2020. 5

[45] Baoxiong Jia, Yixin Chen, Siyuan Huang, Yixin Zhu, and Song-Chun Zhu. Lemma: A multi-view dataset for le arning m ulti-agent m ulti-task a ctivities. In *European Conference on Computer Vision*, pages 767–786. Springer, 2020. 4

[46] Yu-Gang Jiang, Zuxuan Wu, Jun Wang, Xiangyang Xue, and Shih-Fu Chang. Fcvid: Fudan-columbia video dataset, 2015. 4

[47] Yu-Gang Jiang, Guangnan Ye, Shih-Fu Chang, Daniel Ellis, and Alexander C Loui. Consumer video understanding: A benchmark database and an evaluation of human and machine performance. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, pages 1–8, 2011. 4

[48] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014. 4

[49] Zeeshan Khan, CV Jawahar, and Makarand Tapaswi. Grounded video situation recognition. *Advances in Neural Information Processing Systems*, 35:8199–8210, 2022. 8

[50] Khushboo Khurana and Umesh Deshpande. Video question-answering techniques, benchmark datasets and evaluation metrics leveraging video captioning: A comprehensive survey. *IEEE Access*, 9:43799–43823, 2021. 1

[51] Quan Kong, Ziming Wu, Ziwei Deng, Martin Klinkigt, Bin Tong, and Tomokazu Murakami. Mmact: A large-scale dataset for cross modal human action understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8658–8667, 2019. 4

[52] Yu Kong and Yun Fu. Human action recognition and prediction: A survey, 2022. 1

[53] Giorgos Kordopatis-Zilos, Symeon Papadopoulos, Ioannis Patras, and Ioannis Kompatsiaris. Fivr: Fine-grained incident video retrieval. *IEEE Transactions on Multimedia*, 21(10):2638–2652, 2019. 4

[54] 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 con-*

ference on computer vision and pattern recognition, pages 780–787, 2014. 4

[55] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: A large video database for human motion recognition. In *2011 International Conference on Computer Vision*, pages 2556–2563, 2011. 4

[56] Jie Lei, Tamara L. Berg, and Mohit Bansal. Mtvr: Multilingual moment retrieval in videos, 2021. 4

[57] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L. Berg. Tvqa: Localized, compositional video question answering, 2019. 4

[58] Jie Lei, Licheng Yu, Tamara L. Berg, and Mohit Bansal. Tvqa+: Spatio-temporal grounding for video question answering, 2020. 4

[59] Jie Lei, Licheng Yu, Tamara L. Berg, and Mohit Bansal. Tvr: A large-scale dataset for video-subtitle moment retrieval, 2020. 4

[60] Jie Lei, Licheng Yu, Tamara L. Berg, and Mohit Bansal. What is more likely to happen next? video-and-language future event prediction, 2020. 4

[61] Linjie Li, Jie Lei, Zhe Gan, Licheng Yu, Yen-Chun Chen, Rohit Pillai, Yu Cheng, Luowei Zhou, Xin Eric Wang, William Yang Wang, et al. Value: A multi-task benchmark for video-and-language understanding evaluation. *arXiv preprint arXiv:2106.04632*, 2021. 4

[62] Qian Li, Jianxin Li, Jiawei Sheng, Shiyao Cui, Jia Wu, Yiming Hei, Hao Peng, Shu Guo, Lihong Wang, Amin Beheshti, and Philip S. Yu. A survey on deep learning event extraction: Approaches and applications. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–21, 2022. 1

[63] Xinyu Li, Fayuan Li, Lu Pan, Yuguang Chen, Weihua Peng, Quan Wang, Yajuan Lyu, and Yong Zhu. Duee: a large-scale dataset for chinese event extraction in real-world scenarios. In *Natural Language Processing and Chinese Computing: 9th CCF International Conference, NLPCC 2020, Zhengzhou, China, October 14–18, 2020, Proceedings, Part II 9*, pages 534–545. Springer, 2020. 6

[64] Yingwei Li, Yi Li, and Nuno Vasconcelos. Resound: Towards action recognition without representation bias. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 513–528, 2018. 4

[65] Yuncheng Li, Yale Song, Liangliang Cao, Joel Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo Luo. Tgif: A new dataset and benchmark on animated gif description, 2016. 4

[66] Jingzhou Liu, Wenhu Chen, Yu Cheng, Zhe Gan, Licheng Yu, Yiming Yang, and Jingjing Liu. Violin: A large-scale dataset for video-and-language inference, 2020. 4

[67] Jiangwei Liu, Liangyu Min, and Xiaohong Huang. An overview of event extraction and its applications. *arXiv preprint arXiv:2111.03212*, 2021. 1, 6

[68] Xiaolong Liu, Yao Hu, Song Bai, Fei Ding, Xiang Bai, and Philip HS Torr. Multi-shot temporal event localization: a benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12596–12606, 2021. 4

[69] Weixin Luo, Wen Liu, and Shenghua Gao. A revisit of sparse coding based anomaly detection in stacked rnn

[70] Zelun Luo, Wanze Xie, Siddharth Kapoor, Yiyun Liang, Michael Cooper, Juan Carlos Niebles, Ehsan Adeli, and Fei-Fei Li. Moma: Multi-object multi-actor activity parsing. *Advances in neural information processing systems*, 34:17939–17955, 2021. 4

[71] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023. 4

[72] Claudia Maienborn. Event semantics. *by C. Maienborn, K. von Heusinger, P. Portner.—Berlin*, pages 232–266, 2011. 1, 6

[73] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36, 2024. 4

[74] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips, 2019. 4

[75] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(2):502–508, 2019. 4

[76] Mathew Monfort, SouYoung Jin, Alexander Liu, David Harwath, Rogerio Feris, James Glass, and Aude Oliva. Spoken moments: Learning joint audio-visual representations from video descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14871–14881, 2021. 4

[77] Mathew Monfort, Bowen Pan, Kandan Ramakrishnan, Alex Andonian, Barry A McNamara, Alex Lascelles, Quanfu Fan, Dan Gutfreund, Rogério Schmidt Feris, and Aude Oliva. Multi-moments in time: Learning and interpreting models for multi-action video understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9434–9445, 2021. 4

[78] Arsha Nagrani, Paul Hongsuck Seo, Bryan Seybold, Anja Hauth, Santiago Manen, Chen Sun, and Cordelia Schmid. Learning audio-video modalities from image captions. In *European Conference on Computer Vision*, pages 407–426. Springer, 2022. 4

[79] Xun Long Ng, Kian Eng Ong, Qichen Zheng, Yun Ni, Si Yong Yeo, and Jun Liu. Animal kingdom: A large and diverse dataset for animal behavior understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19023–19034, 2022. 4

[80] Khanh Duy Nguyen, Zixuan Zhang, Reece Suchocki, Sha Li, Martha Palmer, Susan Windisch Brown, Jiawei Han, and Heng Ji. Resin-editor: A schema-guided hierarchical event graph visualizer and editor. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 365–372, 2023. 7

[81] Andreea-Maria Oncescu, João F. Henriques, Yang Liu, An-

drew Zisserman, and Samuel Albanie. Queryd: A video dataset with high-quality text and audio narrations, 2021. 4

[82] Paritosh Parmar and Brendan Tran Morris. What and how well you performed? a multitask learning approach to action quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 304–313, 2019. 4

[83] Terence Parsons. Events in the semantics of english: A study in subatomic semantics. 1990. 6

[84] Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Mateusz Malinowski, Yi Yang, Carl Doersch, et al. Perception test: A diagnostic benchmark for multimodal video models. *Advances in Neural Information Processing Systems*, 36, 2024. 4

[85] Sarah Pratt, Mark Yatskar, Luca Weihs, Ali Farhadi, and Aniruddha Kembhavi. Grounded situation recognition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 314–332. Springer, 2020. 1, 6

[86] Nishant Rai, Haofeng Chen, Jingwei Ji, Rishi Desai, Kazuki Kozuka, Shun Ishizaka, Ehsan Adeli, and Juan Carlos Niebles. Home action genome: Cooperative compositional action understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11184–11193, 2021. 4

[87] Hans Reichenbach. Elements of symbolic logic. 1947. 6

[88] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 1

[89] Zhihang Ren, Jefferson Ortega, Yifan Wang, Zhimin Chen, Yunhui Guo, Stella X Yu, and David Whitney. Veatic: Video-based emotion and affect tracking in context dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4467–4477, 2024. 4

[90] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. A dataset for movie description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3202–3212, 2015. 4

[91] Marcus Rohrbach, Anna Rohrbach, Michaela Regneri, Sikandar Amin, Mykhaylo Andriluka, Manfred Pinkal, and Bernt Schiele. Recognizing fine-grained and composite activities using hand-centric features and script data. *International Journal of Computer Vision*, 119:346–373, 2016. 4

[92] Luca Rossetto, Heiko Schuldt, George Awad, and Asad A. Butt. V3c - a research video collection, 2018. 4

[93] Susan Rothstein. *Events and grammar*, volume 70. Springer Science & Business Media, 2001. 6

[94] Andrew Rouditchenko, Yung-Sung Chuang, Nina Shvetsova, Samuel Thomas, Rogerio Feris, Brian Kingsbury, Leonid Karlinsky, David Harwath, Hilde Kuehne, and James Glass. C2kd: Cross-lingual cross-modal knowledge distillation for multilingual text-video retrieval, 2023. 4

[95] Arka Sadhu, Tanmay Gupta, Mark Yatskar, Ram Nevatia, and Aniruddha Kembhavi. Visual semantic role labeling

for video understanding, 2021. 2, 4, 8

[96] Kate Sanders, David Etter, Reno Kriz, and Benjamin Van Durme. Multivent: Multilingual videos of events with aligned natural text. *arXiv preprint arXiv:2307.03153*, 2023. 4

[97] Barry Schein. *Plurals and events*, volume 23. Mit Press, 1993. 6

[98] Lenhart K. Schubert. *The situations we talk about*, page 408–439. Kluwer Academic Publishers, USA, 2001. 6

[99] Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21096–21106, 2022. 4

[100] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Finegym: A hierarchical video dataset for fine-grained action understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2616–2625, 2020. 4

[101] Roshan Sharma, Shruti Palaskar, Alan W Black, and Florian Metze. End-to-end speech summarization using restricted self-attention. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8072–8076, 2022. 4

[102] Roshan Sharma, Shruti Palaskar, Alan W Black, and Florian Metze. End-to-end speech summarization using restricted self-attention. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8072–8076. IEEE, 2022. 4

[103] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Actor and observer: Joint modeling of first and third-person videos. In *proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7396–7404, 2018. 4

[104] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 510–526. Springer, 2016. 4

[105] Alan F Smeaton, Paul Over, and Wessel Kraaij. Evaluation campaigns and trecvid. In *Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, pages 321–330, 2006. 4

[106] Mattia Soldan, Alejandro Pardo, Juan León Alcázar, Fabian Caba Heilbron, Chen Zhao, Silvio Giancola, and Bernard Ghanem. MAD: A scalable dataset for language grounding in videos from movie audio descriptions. *CoRR*, abs/2112.00431, 2021. 4

[107] Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. Tvsum: Summarizing web videos using titles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5179–5187, 2015. 4

[108] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 4, 6

[109] Newton Spolaôr, Huei Diana Lee, Weber Shoity Resende Takaki, Leandro Augusto Ensina, Claudio Saddy Rodrigues Coy, and Feng Chung Wu. A systematic review on content-based video retrieval. *Eng. Appl. Artif. Intell.*, 90(C), apr 2020. 1

[110] 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. 4

[111] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering, 2016. 4

[112] Ha Thi Phuong Thao, Gemma Roig, and Dorien Herremans. Emomv: Affective music-video correspondence learning datasets for classification and retrieval. *Information Fusion*, 91:64–79, 2023. 4

[113] Jennifer Tracey, Stephanie Strassel, Jeremy Getman, Ann Bies, Kira Griffitt, David Graff, and Christopher Caruso. Aida scenario 1 practice topic source data, Oct 2023. 4

[114] Quang-Trung Truong, Tuan-Anh Vu, Tan-Sang Ha, Jakub Lokoč, Yue-Him Wong, Ajay Joneja, and Sai-Kit Yeung. Marine video kit: a new marine video dataset for content-based analysis and retrieval. In *International Conference on Multimedia Modeling*, pages 539–550. Springer, 2023. 4

[115] Elahe Vahdani and Yingli Tian. Deep learning-based action detection in untrimmed videos: A survey, 2021. 1

[116] Siddharth Vashishtha, Alexander Martin, William Gantt, Benjamin Van Durme, and Aaron Steven White. Famus: Frames across multiple sources. *arXiv preprint arXiv:2311.05601*, 2023. 5

[117] Paul Vicol, Makarand Tapaswi, Lluis Castrejon, and Sanja Fidler. Moviegraphs: Towards understanding human-centric situations from videos, 2018. 4

[118] Shuo Wang, Meizhi Ju, Yunyan Zhang, Yefeng Zheng, Meng Wang, and Guilin Qi. Cross-modal contrastive learning for event extraction. In *International Conference on Database Systems for Advanced Applications*, pages 699–715. Springer, 2023. 4

[119] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research, 2020. 4

[120] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023. 4

[121] Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. Star: A benchmark for situated reasoning in real-world videos. In *Thirty-fifth conference on neural information processing systems datasets and benchmarks track (Round 2)*, 2021. 4

[122] Haoqian Wu, Keyu Chen, Haozhe Liu, Mingchen Zhuge, Bing Li, Ruizhi Qiao, Xiujun Shu, Bei Gan, Liangsheng Xu, Bo Ren, et al. Newsnet: A novel dataset for hierarchical temporal segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10669–10680, 2023. 4

[123] Weijia Wu, Yuzhong Zhao, Zhuang Li, Jiahong Li, Hong Zhou, Mike Zheng Shou, and Xiang Bai. A large cross-modal video retrieval dataset with reading comprehension, 2023. 4

[124] Wei Xiang and Bang Wang. A survey of event extraction from text. *IEEE Access*, 7:173111–173137, 2019. 1

[125] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa:next phase of question-answering to explaining temporal actions, 2021. 4

[126] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653, 2017. 4

[127] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016. 4

[128] Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. Advancing high-resolution video-language representation with large-scale video transcriptions, 2022. 4

[129] Rui Yan, Lingxi Xie, Jinhui Tang, Xiangbo Shu, and Qi Tian. Social adaptive module for weakly-supervised group activity recognition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*, pages 208–224. Springer, 2020. 4

[130] Shaohua Yang, Qiaozi Gao, Changsong Liu, Caiming Xiong, Song-Chun Zhu, and Joyce Chai. Grounded semantic role labeling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 149–159, 2016. 6

[131] Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. Situation recognition: Visual semantic role labeling for image understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5534–5542, 2016. 1, 6, 8

[132] Guangnan Ye, Yitong Li, Hongliang Xu, Dong Liu, and Shih-Fu Chang. Eventnet: A large scale structured concept library for complex event detection in video. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 471–480, 2015. 4

[133] Serena Yeung, Olga Russakovsky, Ning Jin, Mykhaylo Andriluka, Greg Mori, and Li Fei-Fei. Every moment counts: Dense detailed labeling of actions in complex videos. *International Journal of Computer Vision*, 126:375–389, 2018. 4, 5

[134] Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. Social-iq: A question answering benchmark for artificial social intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8807–8817, 2019. 4

[135] Oussama Zayene, Jean Hennebert, Sameh Masmoudi Touj, Rolf Ingold, and Najoua Essoukri Ben Amara. A dataset

for arabic text detection, tracking and recognition in news videos- activ. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 996–1000, 2015. 4

[136] Kuo-Hao Zeng, Tseng-Hung Chen, Ching-Yao Chuang, Yuan-Hong Liao, Juan Carlos Niebles, and Min Sun. Leveraging video descriptions to learn video question answering, 2016. 4

[137] Xing Zhang, Zuxuan Wu, Zejia Weng, Huazhu Fu, Jingjing Chen, Yu-Gang Jiang, and Larry S Davis. Videolt: Large-scale long-tailed video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7960–7969, 2021. 4

[138] Hang Zhao, Antonio Torralba, Lorenzo Torresani, and Zhicheng Yan. Hacs: Human action clips and segments dataset for recognition and temporal localization, 2019. 4

[139] Luowei Zhou, Yannis Kalantidis, Xinlei Chen, Jason J Corso, and Marcus Rohrbach. Grounded video description. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6578–6587, 2019. 4

[140] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 4

[141] Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, and Josef Sivic. Cross-task weakly supervised learning from instructional videos, 2019. 4