Visual Semantic Role Labeling for Video Understanding

Arka Sadhu[1†] Tanmay Gupta[3] Mark Yatskar[2] Ram Nevatia[1] Aniruddha Kembhavi[3]
[1]University of Southern California [2]University of Pennsylvania [3]PRIOR @ Allen Institute for AI
{asadhu|nevatia}@usc.edu myatskar@seas.upenn.edu {tanmayg|anik}@allenai.org

vidsitu.org

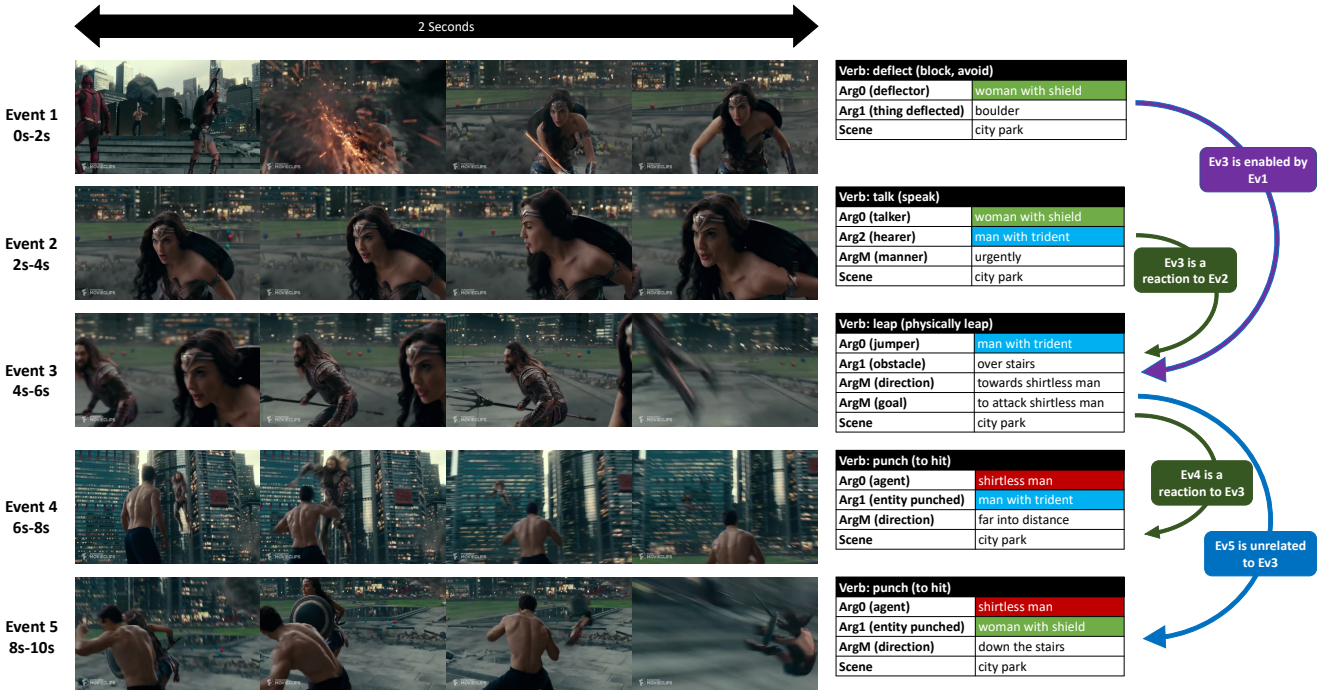

Figure 1: **A sample video and annotation from VidSitu.** The figure shows a 10-second video annotated with 5 events, one for each 2-second interval. Each event consists of a verb (like "deflect") and its arguments (like *Arg0 (deflector)* and *Arg1 (thing deflected)*). Entities that participate in multiple events within a clip are co-referenced across all such events (marked using the same color). Finally, we relate all events to the central event (Event 3). The video can be viewed at: https://youtu.be/3sP7UMxhGYw?t=20 (from 20s-30s).

Abstract

We propose a new framework for understanding and representing related salient events in a video using visual semantic role labeling. We represent videos as a set of related events, wherein each event consists of a verb and multiple entities that fulfill various roles relevant to that event. To study the challenging task of semantic role labeling in videos or VidSRL, we introduce the VidSitu benchmark, a large scale video understanding data source with 29K 10-second movie clips richly annotated with a verb and semantic-roles every 2 seconds. Entities are co-referenced across events within a movie clip and events are connected to each other via event-event relations. Clips in VidSitu are drawn from a large collection of movies (∼3K) and have been chosen to be both complex (∼4.2 unique verbs within a video) as well as diverse (∼200 verbs have more than 100 annotations each). We provide a comprehensive analysis of the dataset in comparison to other publicly available video understanding benchmarks, several illustrative baselines and evaluate a range of standard video recognition models. Our code and dataset is available at vidsitu.org.

[†]Part of the work was done during Arka's internship at PRIOR@AI2

1. Introduction

Videos record events in our lives with both short and long temporal horizons. These recordings frequently relate multiple events separated geographically and temporally and capture a wide variety of situations involving human beings interacting with other humans, objects and their environment. Extracting such rich and complex information from videos can drive numerous downstream applications such as describing videos [35, 82, 77], answering queries about them [85, 81], retrieving visual content [50], building knowledge graphs [48] and even teaching embodied agents to act and interact with the real world [84].

Parsing video content is an active area of research with much of the focus centered around tasks such as action classification [31], localization [24] and spatio-temporal detection [21]. Although parsing human actions is a critical component of understanding videos, actions by themselves paint an incomplete picture, missing critical pieces such as the agent performing the action, the object being acted upon, the tool or instrument used to perform the action, location where the action is performed and more. Expository tasks such as video captioning and story-telling provide a more holistic understanding of the visual content; but akin to their counterparts in the image domain, they lack a clear definition of the type of information being extracted making them notoriously hard to evaluate [32, 74].

Recent work in the image domain [83, 58, 22] has attempted to move beyond action classification via the task of visual semantic role labeling - producing not just the primary activity in an image or region, but also the entities participating in that activity via different roles. Building upon this line of research, we propose VidSRL – the task of recognizing spatio-temporal situations in video content. As illustrated in Figure. 1, VidSRL involves recognizing and temporally localizing salient events across the video, identifying participating actors, objects, and locations involved within these events, co-referencing these entities across events over the duration of the video, and relating how events affect each other over time. We posit that VidSRL, a considerably more detailed and involved task than action classification with more precise definitions of the extracted information than video captioning, is a step towards obtaining a holistic understanding of complex videos.

To study VidSRL, we present VidSitu, a large video understanding dataset of over 29K videos drawn from a diverse set of 3K movies. Videos in VidSitu are exactly 10 seconds long and are annotated with 5 verbs, corresponding to the most salient event taking place within the five 2 second intervals in the video. Each verb annotation is accompanied with a set of roles whose values [1] are annotated using free form text. In contrast to verb annotations which are derived from a fixed vocabulary, the free form role annotations allow the use of referring expressions (*e.g. boy wearing a blue jacket*) to disambiguate entities in the video. An entity that occurs in any of the five clips within a video is consistently referred to using the same expression, allowing us to develop and evaluate models with co-referencing capability. Finally, the dataset also contains event relation annotations capturing causation (Event Y is Caused By/Reaction To Event X) and contingency (Event X is a pre-condition for Event Y). The key highlights of VidSitu include: (i) *Diverse Situations:* VidSitu enjoys a large vocabulary of verbs (1500 unique verbs curated from PropBank [54] with 200 verbs having at least 100 event annotations) and entities (5600 unique nouns with 350 nouns occurring in at least 100 videos); (ii) *Complex Situations:* Each video is annotated with 5 inter-related events and has an average of 4.2 unique verbs, 6.5 unique entities and; (iii) *Rich Annotations:* VidSitu provides structured event representations (3.8 roles per event) with entity co-referencing and event-relation labels.

To facilitate further research on VidSRL, we provide a comprehensive benchmark that supports partwise evaluation of various capabilities required for solving VidSRL and create baselines for each capability using state-of-art architectural components to serve as a point of reference for future work. We also carefully choose metrics that provide a meaningful signal of progress towards achieving competency on each capability. Finally, we perform a human-agreement analysis that reveals a significant room for improvement on the VidSitu benchmark.

Our main contributions are: (i) the VidSRL task formalism for understanding complex situations in videos; (ii) curating the richly annotated VidSitu dataset that consists of diverse and complex situations for studying VidSRL; (iii) establishing an evaluation methodology for assessing crucial capabilities needed for VidSRL and establishing baselines for each using state-of-art components. The dataset and code are publicly available at `vidsitu.org`.

2. Related Work

Video Understanding, a fundamental goal of computer vision, is an incredibly active area of research involving a wide variety of tasks such as action classification [8, 16, 75], localization [44, 43] and spatio-temporal detection [19], video description [77, 35], question answering [85], and object grounding [61]. Tasks like detecting atomic actions at 1 second intervals [19, 79, 67] are short horizon tasks whereas ones like summarizing 180 second long videos [91] are extremely long horizon tasks. In contrast, our proposed task of VidSRL operates on 10 second video at 2 second intervals.

[1] Following nomenclature introduced in ImSitu[83], every verb (deflect) has a set of roles (Arg0 deflector, Arg1 thing deflected) which are realized by noun values. Here, we use "value" to refer to free-form text used describing the roles (woman with shield, boulder).

Task	Required Annotations	Dataset
Action Classification	Action Labels	Kinetics[31], ActivityNet [24], Moments in Time [51], Something-Something[20], HVU [14]
Action Localization	Action Labels, Temp. Segments	ActivityNet, Thumos[29], HACS [89], Tacos[59], Charades[63], COIN[69]
Spatio-Temporal Detection	Action Labels, Temp. Segments, BBoxes	AVA[21], AVA-Kinetics[39], EPIC-Kitchens [12], JHMDB[30]
Video Description	Captions, Temp. Segments	ActivityNet[24], Vatex[77], YouCook[13], MSR-VTT [82] , LSMDC [60]
Video QA	Q/A, Subtitle or Script (optional)	MSRVTT-QA[81], VideoQA[86], ActivityNetQA[85], TVQA[37], MovieQA[70]
Text to Video Retrieval	Text Query, ASR output (optional)	HowTo100M[50], TVR[38], DiDeMo[25], Charades-STA[17]
Video Object Grounding	Text Query, Temp. Segments, BBoxes	ActivityNet-SRL[61], YouCookII[90], VidSTG [88],VID-sentence[11]
VidSRL	Verbs, SRLs, Corefs, Event Relations, Temp. Segments	VidSitu

Table 1: A non-exhaustive summary of video understanding tasks, required annotations and benchmarks.

It entails producing a verb for the salient activity within each 2 second interval as well as predicting multiple entities that fulfill various roles related to that event, and finally relating these events across time.

In support of these tasks, the community has also proposed datasets [31, 24, 21], over the past few years. While early datasets were small datasets with several hundred or thousand examples[65, 36], recent datasets are massive[50] enabling researchers to train large neural models and also employ pre-training strategies[49, 92, 40]. Section 4, Table 3 and Figure 2 provide a comparison of our proposed dataset to several relevant datasets in the field. Due to space constraints, we are unable to provide a thorough description of all the relevant work. Instead we point the reader to relevant surveys on video understanding [1, 34, 87] and also present a holistic overview of tasks and datasets in Table 1.

Visual Semantic Role Labeling has been primarily explored in the image domain under situation recognition [83, 58], visual semantic role labeling [22, 41, 64] and human-object interaction [10, 9]. Compared to images, visual semantic role labeling in videos requires not just recognizing actions and arguments at a single time step but aggregating information about interacting entities across frames, co-referencing the entities participating across events.

Movies for Video Understanding: The movie domain serves as a rich data source for spatio-temporal detection [21], movie description [60], movie question answering [70] , story-based retrieval [3], generating social graphs [72] tasks, and classifying shot style [28]. In contrast to a lot of this prior work, we focus only on the visual activity of the various actors and objects in the scene, *i.e.* no additional modalities like movie-scripts, subtitles or audio are presented in our dataset.

3. VidSRL: The Task

State-of-the-art video analysis capabilities like video activity recognition and object detection yield a fairly impoverished understanding of videos by reducing complex events involving interactions of multiple actors, objects, and locations to a bag of activity and object labels. While video captioning promises rich descriptions of videos, the open-ended task definition of captioning lends itself poorly to a systematic representation of such events and evaluation thereof. The motivation behind VidSRL is to expand the video analysis toolbox with vision models that produce richer yet structured representations of complex events in videos than currently possible through video activity recognition, object detection, or captioning.

Formal task definition. Given a video V, VidSRL requires a model to predict a set of related salient events $\{E_i\}_{i=1}^{k}$ constituting a situation. Each event E_i consists of a verb v_i chosen from a set of of verbs $\mathcal{V}$ and values (entities, location, or other details pertaining to the event described in text) assigned to various roles relevant to the verb. We denote the roles or arguments of a verb v as $\{A_j^v\}_{j=1}^{m}$ and $A_j^v \leftarrow a$ implies that the j^{th} role of verb v is assigned the value a. In Fig. 1 for instance, event E_1 consists of verb v="deflect (block, avoid)" with *Arg0 (deflector)* $\leftarrow$ "woman with shield". The roles for the verbs are obtained from PropBank [54]. Finally, we denote the relationship between any two events E and E' by $l(E, E') \in \mathcal{L}$ where $\mathcal{L}$ is an event-relations label set. We now discuss simplifying assumptions and trade-offs in designing the task.

Timescale of Salient Events. What constitutes a salient event in a video is often ambiguous and subjective. For instance given the 10 sec clip in Fig. 1, one could define fine-grained events around atomic actions such as "turning" (Event 2 third frame) or take a more holistic view of the sequence as involving a "fight". This ambiguity due to lack of constraints on timescales of events makes annotation and evaluation challenging. We resolve this ambiguity by restricting the choice of salient events to one event per fixed time-interval. Previous work on recognizing atomic actions [21] relied upon 1 sec intervals. An appropriate choice of time interval for annotating events is one that enables rich descriptions of complex videos while avoiding incidental atomic actions. We observed qualitatively that a 2 sec interval strikes a good balance between obtaining descriptive events and the objectiveness needed for a systematic evaluation. Therefore, for each 10 sec clip, we annotate 5 events $\{E_i\}_{i=1}^{5}$. Appendix B.1 elaborates on this choice.

Describing an Event. We describe an event through a verb and its arguments. For verbs, we follow recent work in action recognition like ActivityNet [24] and Moments in Time [51] that choose a verb label for each video segment from a curated list of verbs. To allow for description of a wide variety of events, we select a large vocabulary of $2.2K$

visual verb from PropBank [54]. Verbs in PropBank are diverse, distinguish between homonyms using verb-senses (e.g. "strike (hit)" vs "strike (a pose)"), and provide a set of roles for each verb. We allow values of arguments for the verb to be free-form text. This allows disambiguation between different entities in the scene using referring expression such as "man with trident" or "shirtless man" (Fig. 1). Understanding of a video may require consolidating partial information across multiple views or shots. In VidSRL, while the 2 sec clip is sufficient to assign the verb, roles may require information from the whole video since some entities involved in the event may be occluded or lie outside the camera-view for those 2 secs but are visible before or after. For e.g., in Fig 1 Event 2, information about "Arg2 (hearer)" is available only in Event 3.

Co-Referencing Entities Across Events. Within a video, an entity may be involved in more than one event, for instance, "woman with shield" is involved in Events 1, 2, and 5 and "man with trident" is involved in Events 2, 3, and 4. In such cases, we expect VidSRL models to understand co-referencing *i.e.* a model must be able to recognize that the entity participating across those events is the same even though the entity may be playing different roles in those events. Ideally, evaluating coreferencing capability requires grounding entities in the video (e.g. using bounding boxes). Since grounding entities in videos is an expensive process, we currently require the phrases referring to the same entity across multiple events within each 10 sec clip to match exactly for coreference assessment. See supp. for details on how coreference is enforced in our annotation pipeline.

Event Relations. Understanding a video requires not only recognizing individual events but also how events affect one another. Since event relations in videos is not yet well explored, we propose a taxonomy of event relations as a first step – inspired by prior work on a schema for event relations in natural language [26] that includes "Causation" and "Contingency". In particular, if Event B follows (occurs after) Event A, we have the following relations: (i) *Event B is caused by Event A* (Event B is a direct result of Event B); (ii) *Event B is enabled by Event A* (Event A does not cause Event B, but Event B would not occur in the absence of Event A); (iii) *Event B is a reaction to Event A* (Event B is a response to Event A); and (iv) *Event B is unrelated to Event A* (examples are provided in supplementary).

4. VidSitu Dataset

To study VidSRL, we introduce the VidSitu dataset that offers videos with **diverse** and **complex** situations (a collection of related events) and **rich** annotations with verbs, semantic roles, entity co-references, and event relations. We describe our dataset curation decisions (Section 4.1) followed by analysis of the dataset (Section 4.2).

	Train	Valid	Test-Vb	Test-SRL	Test -ER	Total
# Movies	23626	1326	1353	1598	1317	29220
# Videos	2431	151	151	153	151	3037
# Clips	118130	6630	6765	7990	6585	146100
# Verbs Ann / Clip	1	10	10	10	10	
# Verb Ann	118130	66300	67650	79900	65850	397830
# Unique Verb Tuples	23196	1317	1341	1571	1299	28724
# Values Ann / Role	1	3	3	3	3	
# Role Ann	118130	19890	20295	23970	19755	202040

Table 2: **Statistics** on splits of VidSitu. Note that VidSitu contains multiple verb and role annotations for val and test sets for accurate evaluation.

4.1. Dataset Curation

We briefly describe the main steps in the data curation process and provide more information in Appendix B.

Video Source Selection. Videos from movies are well suited for VidSRL since they are naturally diverse (wide-range of movie genres) and often involve multiple interacting entities. Also, scenarios in movies typically play out over multiple shots which makes movies a challenging testbed for video understanding. We use videos from Condensed-Movies [3] which collates videos from MovieClips- a licensed YouTube channel containing engaging movie scenes.

Video Selection. Within the roughly 1000 hours of MovieClips videos, we select 30K diverse and interesting 10sec videos to annotate while avoiding visually uneventful segments common in movies such as actors merely engaged in dialogue. This selection is performed using a combination of human detection, object detection and atomic action prediction followed by a sampling of no more than 3 videos per movieclip after discarding inappropriate content.

Curating Verb Senses. We begin with the entire PropBank [54] vocabulary of $\sim 6k$ verb-senses. We manually remove fine-grained and non-visual verb-senses and further discard verbs that do not appear in the MPII-Movie Description (MP2D) dataset [60] (verbs extracted using a semantic-role parser [62]). This gives us a set of 2154 verb-senses.

Curating Argument Roles. We wish to establish a set of argument roles for each verb-sense. We initialize the argument list for each verb-sense using Arg0, Arg1, Arg2 arguments provided by PropBank and then expand this using frequently used (automatically extracted) arguments present in descriptions provided by the MP2D dataset.

Annotations. Annotations for the verbs, roles and relations are obtained via Amazon Mechanical Turk (AMT). The annotation interface enables efficient annotations while encouraging rich descriptions of entities and enabling a reuse of entities through the video (to preserve co-referencing). See Appendix B.2 for details.

Dataset splits. VidSitu is split into train, validation and test sets via a 80:5:15 split, ensuring that videos from the same movie end up in exactly one of those sets. Table 2 summarizes these statistics of these splits. We emphasize

Dataset	Domain	SRLs, Coref	EvRel	Videos	Clips	Descr.	Descr./Clip (Train)	Avg. Clip Len. (s)	Uniq Vbs/Vid	Uniq Ents/Vid	Avg. Roles/Event
MSR-VTT	open	Implicit	✗	7k	10k	200k	20	14.83	1.88	2.80	1.56
MPII-MD	movie	Implicit	✗	94	68k	68.3	1	3.90	1.87	2.99	2.24
ActyNet-Cap	open	Implicit	✗	20k	100k	100k	1	36.20	2.30	3.75	2.37
Vatex-en	open	Implicit	✗	41.3k	41.3k	413k	10	10.00	2.69	4.04	1.96
VidSitu	movie	**Explicit**	✓	29.2k	146k	146k	1	10.00	**4.21**	**6.58**	**3.83**

Table 3: **Dataset statistics across video description datasets.** We **highlight** key differences from previous datasets such as explicit SRL, co-reference, and event-relation annotations, and greater diversity and density of verbs, entities, and semantic roles. For a fair comparison, for all datasets we use a single description per video segment when more than one are available.

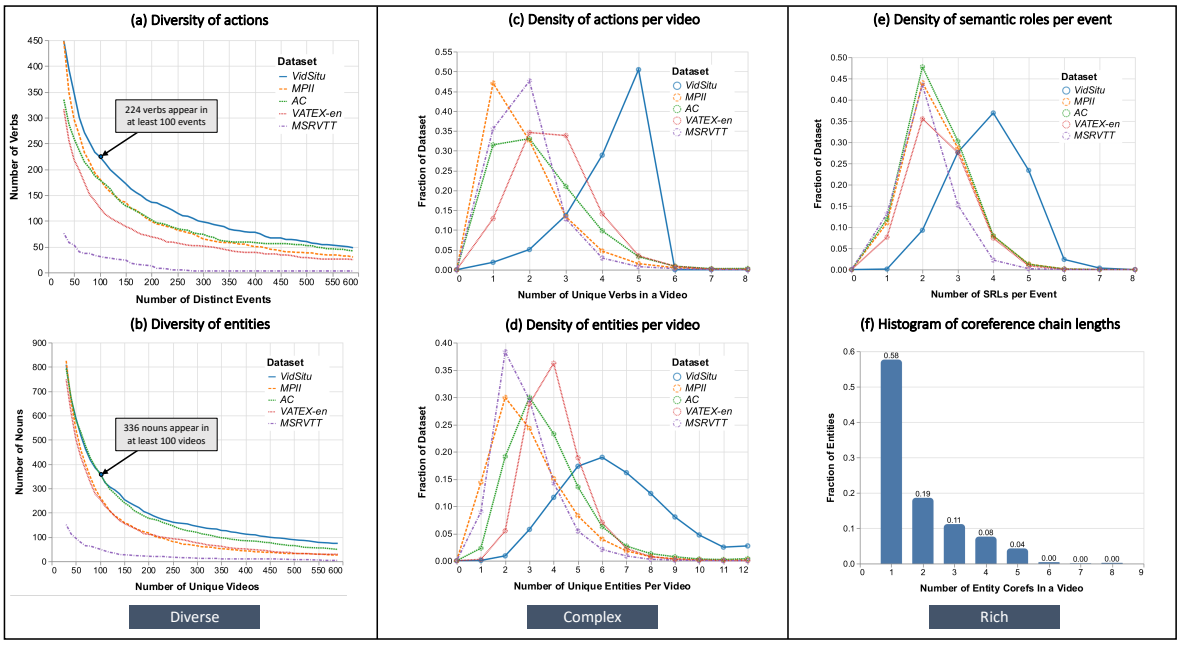

Figure 2: **Data analysis.** An analysis of VidSitu in comparison to other large scale relevant video datasets. We focus on the **diversity** of actions and entities in the dataset (a and b), the **complexity** of the situations measured in terms of the number of unique verbs and entities per video (c and d) and the **richness** of annotations (e and f).

that each of the three tasks namely **Verb Prediction**, **Semantic Role Prediction and Co-Referencing** and **Event Relation Prediction** have separate test sets.

Multiple Annotations for Evaluation Sets. Via controlled trials (see Sec 6.1) we measured the annotation disagreement rate for the train set. Based on this data, we obtain multiple annotations for validation and test sets using a 2-stage annotation process. In the first stage, we collect 10 verbs for each 2 second clip (1 verb per worker). In the second stage, we get role labels for the verb with the highest agreement from 3 different workers.

4.2. Dataset Analysis and Statistics

We present an extensive analysis of VidSitu focusing on three key elements: (i) **diversity** of events represented in the dataset; (ii) **complexity** of the situations; and (iii) **richness** of annotations. We provide comparisons to four prominent video datasets containing text descriptions – MSR-VTT [82], MPII-Movie Description [60], ActivityNet Captions [35], and Vatex-en [77] (the subset of descriptions in English). Table 3 summarizes basic statistics from all datasets. For consistency, we use one description per video segment whenever multiple annotations are available, as is the case for Vatex-en, MSR-VTT, validation set of ActivityNet-Captions and both validation and test sets of VidSitu. For datasets without explicit verb or semantic role labels, we extract these using a semantic role parser [62].

Diversity of Events. To assess the diversity of events represented in the dataset, we consider cumulative distributions of verbs[2] and nouns (see Fig. 2-a,b). For any point n on the horizontal axis, the curves show the number of verbs or nouns with at least n annotations. VidSitu not only offers greater diversity in verbs and nouns as compared to other datasets but also a large number of verbs and nouns

[2]As a fair comparison to datasets which do not have senses associated with verbs, we collapse verb senses into a single unit for this analysis.

occur sufficiently frequently to enable learning useful representations. For instance, 224 verbs and 336 nouns have at least 100 annotations. In general, since movies inherently intend to engage viewers, movie datasets such as MPII and VidSitu are more diverse than open-domain datasets like ActivityNet-Captions and VATEX-en.

Complexity of Situations. We refer to a situation as complex if it consists of inter-related events with multiple entities fulfilling different roles across those events. To evaluate complexity, Figs. 2-c,d compare the number of unique verbs and entities per video across datasets. Approximately, 80% of videos in VidSitu have at least 4 unique verbs and 70% have 6 or more unique entities, in comparison to 20% and 30% respectively for VATEX-en. Further, Fig. 2-e shows that 90% of events in VidSitu have at least 4 semantic roles in comparison to only 55% in VATEX-en. Thus, situations in VidSitu are considerably more complex that existing datasets.

Richness of Annotations. While existing video description datasets only have unstructured text descriptions, VidSitu is annotated with rich structured representations of events that includes verbs, semantic role labels, entity coreferences, and event relations. Such rich annotations not only allow for more thorough evaluation of video analysis techniques but also enable researchers to study relatively unexplored problems in video understanding such as entity coreference and relational understanding of events in videos. Fig. 2-f shows the fraction of entity coreference chains of various lengths.

5. Baselines

For a given video, VidSRL requires predicting verbs and semantic roles for each event as well as event relations. We provide powerful baselines to serve as a point of comparison these crucial capabilities. These models leverage architectures from state-of-the-art video recognition models.

Verb Prediction. Given a 2 sec clip, we require a model to predict the verb corresponding to the most salient event in the clip. As baselines, we provide state-of-art action recognition models such as I3D [8] and SlowFast [16] networks (Step 1 in Fig. 3). We consider variants of I3D both with and without Non-Local blocks [76] and for SlowFast networks, we consider variants with and without the Fast channel. For each architecture, we train a model from scratch as well as a model finetuned after pretraining on Kinetics [31]. All models are trained with a cross-entropy loss over the set of action labels. For subsequent stages, these verb classification models are frozen and used as feature extractors.

Argument Prediction Given Verbs: Given a 10 sec video and a verb for each of the 5 events, a model is required to infer entities and their roles involved in each event. To this end, we adapt seq-to-seq models [68] that consist of an encoder and a decoder (Step 2(a,b) in Fig. 3). Specif-

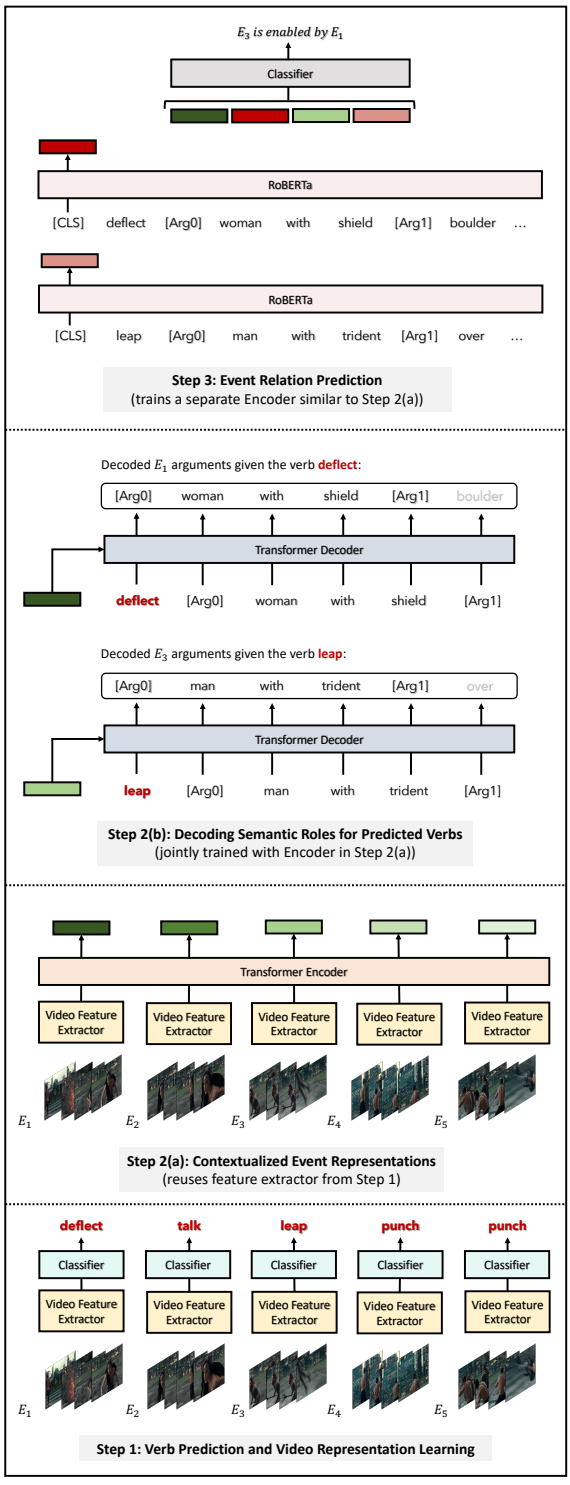

Figure 3: **Models.** The figure illustrates our baselines for verb, semantic role, and event prediction using state-of-the-art network components such as SlowFast [16] network for video feature extraction, transformers [71] for encoding events in a video and verb-conditional decoding of roles, and RoBERTa [46] language encoder for event-relation prediction.

ically, independent event features are fed through a transformer [71] encoder (TxEnc) to get contextualized event representations. Then for each event, the corresponding encoded representation and the verb are passed to a transformer decoder (TxDec) to generate the sequence of arguments and roles for that event. As an example, for Event 1 in Fig 1, we expect to generate the following sequence: [Arg0] woman with shield [Arg1] boulder [Scene] city park

The generated sequence is post-processed to obtain the argument role structure similar to those of the annotations Figure 1. We also provide language only baselines using our TxDec architecture as well as a GPT2 decoder.

Event Relation Prediction: A model must infer how the various events within a video are related given the verb and arguments. For a pair of ordered events (E_i, E_j) with $i < j$, with corresponding verbs and semantic roles, we construct a multimodal representation of each event denoted by m_i and m_j (Step 3 in Fig. 3). Each of these representations is a concatenation of visual representation from TxEnc and a language representation of the sequence of verbs, arguments, and roles obtained from a pretrained RoBERTa [46]-base language model. m_i and m_j are concatenated and fed through a classifier to predict the event relation.

6. Experiments

VidSitu allows us to evaluate performance in 3 stages: (i) verb prediction; (ii) prediction of semantic roles with coreferencing given the video and verbs for each event; and (iii) event relations prediction given the video and verbs and semantic roles for a pair of events.

6.1. Evaluation Metrics

In VidSRL, multiple outputs are plausible for the same input video. This is because of inherent ambiguity in the choice of verb used to describe the event (e.g. the same event may be described by "fight", "punch" or "hit"), and the referring expression used to refer to entities in the video (e.g. "boy with black hair" or "boy in the red shirt"). We confirm this ambiguity through a human-agreement analysis on a subset of 100 videos (500 events) with 25 verb annotations and 5 role annotations per event. Importantly, through careful manual inspection we confirm that a majority of differences in annotation for the same video across AMT workers are due to this inherent ambiguity and not due to a lack of annotation quality.

Verb Prediction. The ambiguity in verbs associated with events suggests that commonly used metrics such as Accuracy, Precision, and F1 are ill suited for the verb prediction task as they would penalize correct predictions that may not be represented in the ground truth annotations. However, recall based metrics such as Recall@k are suitable for this task. Since the large verb vocabulary in VidSitu presents a class-imbalance challenge, we use a macro-averaged Recall@k that better reflects performance across all verb-senses instead of focusing on dominant classes.

We now describe our macro-averaged Verb Recall@k metric. For any event, we only consider the set of verbs which appears at least twice within the ground-truth annotations (each event in val and test sets has 10 verb annotations). For event E_j (where j indexes events in our evaluation set), let this set of agreed-upon ground-truth be denoted by G_j. We compute recall@k for each verb-sense $v_i \in \mathcal{V}$ (where i indexes verb-senses in the vocabulary $\mathcal{V}$) as

$$R_i^k = \frac{\sum_j \mathbb{1}(v_i \in G_j) \times \mathbb{1}(v_i \in P_j^k)}{\sum_j \mathbb{1}(v_i \in G_j)} \tag{1}$$

where $\mathbb{1}$ is an indicator function and P_j^k denotes the set of top-k verb predictions for E_j. Macro-averaged verb recall@k is given by $\frac{1}{|\mathcal{V}|}\sum_i R_i^k$. We report macro-average verb recall@5 (R@5) but also report top-1 and top-5 accuracy (Acc@1/5) for completeness.

Semantic Role Prediction and Co-referencing. Given a video and verb for each event, we wish to measure the semantic role prediction performance. Through a human-agreement analysis we discard arguments such as direction (ADir) and manner (AMnr) which do not have a high inter-annotator agreement and retain Arg0, Arg1, Arg2, ALoc, and AScn for evaluation. This agreement computation is computed using the CIDEr metric by treating one of the chosen annotations as a hypothesis and remaining annotations as references for each argument. In addition to reporting a micro-averaged CIDEr score (C), we also compute macro-averaged CIDEr where the macro-averaging is performed across verb-senses (C-Vb) or argument-types (C-Arg). ROUGE-L (R-L) [42] is shown for completeness.

Since VidSitu provides entity coreference links across events and roles, we use LEA [52] a link-based co-reference metric to measure coreferencing capability. Other metrics (MUC [73], BCUBE [2], CEAFE [47]) can be found in the supp. Co-referencing in our case is done via exact string matching over the predicted set of arguments. Thus, even if the predictions are incorrect, but just the coreference is correct, LEA would give it a higher score. To address this, we propose a soft version of LEA termed LEA-soft (denoted with Lea-S) which assigns weights to cluster matches using their CIDEr score (defined in the supp.).

Event-Relation Prediction Accuracy. Event-relation prediction is a 4-way classification problem. For the subset of 100 videos, We found event relations conditioned on the verbs to have 60% agreement. For evaluation, we use the subset of event pairs for which 2 out of 3 workers agreed on the relation. We use top-1 accuracy (Acc@1) averaged across the classes as the metric for relation prediction.

Model	Vis	Enc	Val						Test					
			C	R-L	C-Vb	C-Arg	Lea	Lea-S	C	R-L	C-Vb	C-Arg	Lea	Lea-S
GPT2	✗	✗	34.67	40.08	42.97	34.45	48.08	28.1	36.48	41.33	44.27	36.51	49.38	30.24
TxDec	✗	✗	35.68	41.19	47.5	32.15	**51.76**	28.6	35.34	41.45	44.44	32.06	**52.46**	29.18
Vid TxDec	SlowFast	✗	44.78	40.61	49.97	41.24	37.88	28.69	44.95	41.12	49.46	41.98	38.91	30.21
Vid TxEncDec	SlowFast	✓	45.52	**42.66**	**55.47**	**42.82**	**50.48**	**31.99**	47.25	**43.46**	**52.92**	**45.48**	**50.88**	**33.5**
Vid TxDec	I3D	✗	**47.14**	40.67	51.61	41.29	37.89	30.38	**47.9**	41.5	51.29	43.62	38.77	31.73
Vid TxEncDec	I3D	✓	**47.06**	**42.41**	**51.67**	**42.76**	48.92	**33.58**	**48.51**	**42.96**	**53.88**	**44.53**	49.61	**35.46**
Human*			84.85	39.77	91.7	80.15	72.1	70.33	83.68	40.04	87.78	79.29	71.77	70.6

Table 4: **Semantic role prediction and co-referencing metrics.** Vis. denotes the visual features used (✗ if not used), and Enc. denotes if video features are contextualized. C: CIDEr, R-L: ROUGE-L, C-Vb: CIDEr scores averaged across verbs, C-Arg: CIDEr scores averaged over arguments. Lea-S: Lea-soft. See Section 6.1 for details.

Model	Kin.	Val			Test		
		Acc@1	Acc@5	Rec@5	Acc@1	Acc@5	Rec@5
I3D	✗	31.18	67.00	5.24	31.91	67.36	5.33
I3D+NL	✗	30.17	66.83	4.88	31.43	67.70	5.02
Slow+NL	✗	33.05	68.83	5.82	34.29	69.56	6.24
SlowFast+NL	✗	32.64	69.22	6.11	33.94	**70.54**	6.56
I3D	✓	29.65	60.77	18.21	29.87	59.10	19.54
I3D+NL	✓	**39.40**	**70.82**	17.12	**38.42**	69.27	18.46
Slow+NL	✓	29.05	58.69	**19.19**	29.03	58.77	**21.06**
SlowFast+NL	✓	**46.79**	**75.90**	**23.38**	**46.37**	**75.28**	**25.78**

Table 5: **Verb classification metrics.** Acc@K: Event Accuracy considering 10 ground-truths and K model predictions. Rec@K: Macro-Averaged Verb Recall with K predictions. Kin. denotes whether Kinetics is used.

	Verb	Args	Val Macro-Acc	Test Macro-Acc
Roberta	✓	✓	25.00	25.00
TxEnc	✓	✓	25.00	25.00
Vid TxEnc	✗	✗	31.98	31.71
Vid TxEnc	✗	✓	**32.22**	**32.03**
Vid TxEnc	✓	✓	**33.46**	**32.10**

Table 6: **Event relation classification metrics.** Macro-Averaged Accuracy on Validation and Test Sets. We evaluate only on the subset of data where two annotators agree.

6.2. Results

Verb Classification: We report macro-averaged Rec@5 (preferred metric; Sec. 6.1) and Acc@1/5 on both validation and test sets in Tab. 5. We observe verb prediction in VidSitu follows similar trends as other action recognition tasks. Specifically, SlowFast architectures outperform I3D and Kinetics pretraining significantly and consistently improves recall across all models by ≈ 10 to 16 points.

Argument Prediction: We report micro and macro-averaged version of CIDEr and ROUGE-L in Tab. 4 (see supp. for other metrics). First, video conditioned models significantly outperform video-blind baselines. Next, we observe that using an encoder to contextualize events in a video improves performance across almost all metrics. Interestingly, while SlowFast outperformed I3D in verb prediction, the reverse is true for semantic role prediction. Even so, a large gap exists between current methods and human performance.

We also evaluate coreferencing ability demonstrated by models without explicitly enforcing it during training. In Tab. 4, we report both Lea and Lea-S (preferred; Sec. 6.1) metrics and find that current techniques are unable to learn coreferencing directly from data. Among all models, only Vid TxEncDec outperformed a language only baseline (GPT2) on both val and test sets, leaving lots of room for improvement in future models.

Event Relation Prediction results are provided in Table 6. Crucially, we find video-blind baselines don't train at all and end up predicting the most frequent class "Enabled By" (hence it gets 0.25 for always predicting majority class). This suggests there exists no exploitable biases within the dataset and underscores the importance and challenge posed by event relations. In contrast, video encoder models even when given just the video without any verb description outperform video-blind baselines. Adding context in the form of verb senses and arguments yields small gains.

In summary, powerful baselines show promise on the three sub-tasks. However, it is clear that VidSitu poses significant new challenges with a huge room for improvement.

7. Conclusion

We introduce visual semantic role labeling in videos in which models are required to identify salient actions, participating entities and their roles within an event, co-reference entities across time, and recognize how actions affect each other. We also present the VidSitu dataset with diverse videos, complex situations, and rich annotations.

8. Acknowledgement

We thank the Mechanical Turk workers for doing an outstanding work in annotating the dataset - without them VidSitu and the paper would not exist. We are also grateful to the suggestions and feedback provided by the three anonymous reviewers. This research was supported, in part, by the Office of Naval Research under grant #N00014-18-1-2050.

References

[1] Nayyer Aafaq, Ajmal Mian, Wei Liu, Syed Zulqarnain Gilani, and Mubarak Shah. Video description: A survey of methods, datasets, and evaluation metrics. *ACM Computing Surveys (CSUR)*, 52(6):1–37, 2019. 3

[2] A. Bagga and B. Baldwin. Entity-based cross-document coreferencing using the vector space model. In *COLING-ACL*, 1998. 7, 16

[3] M. Bain, Arsha Nagrani, A. Brown, and Andrew Zisserman. Condensed movies: Story based retrieval with contextual embeddings. In *ACCV*, 2020. 3, 4, 9, 19

[4] C. Baker, C. Fillmore, and J. Lowe. The berkeley framenet project. In *COLING-ACL*, 1998. 9

[5] S. Banerjee and A. Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *IEEvaluation@ACL*, 2005. 16

[6] Claire Bonial, Olga Babko-Malaya, Jinho D Choi, Jena Hwang, and Martha Palmer. Propbank annotation guidelines. *Center for Computational Language and Education Research, CU-Boulder*, 2010. 9

[7] Susan Windisch Brown, Julia Bonn, James Gung, Annie Zaenen, James Pustejovsky, and Martha Palmer. VerbNet representations: Subevent semantics for transfer verbs. In *Proceedings of the First International Workshop on Designing Meaning Representations*, pages 154–163, Florence, Italy, Aug. 2019. Association for Computational Linguistics. 9

[8] J. Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733, 2017. 2, 6, 13

[9] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 381–389, 2018. 3

[10] Yu-Wei Chao, Z. Wang, Yugeng He, J. Wang, and Jia Deng. Hico: A benchmark for recognizing human-object interactions in images. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1017–1025, 2015. 3

[11] Zhenfang Chen, L. Ma, Wenhan Luo, and K. Wong. Weakly-supervised spatio-temporally grounding natural sentence in video. In *ACL*, 2019. 3

[12] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *European Conference on Computer Vision (ECCV)*, 2018. 3

[13] P. Das, C. Xu, R. F. Doell, and Corso J. J. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2013. 3

[14] Ali Diba, Mohsen Fayyaz, Vivek Sharma, Manohar Paluri, Jürgen Gall, Rainer Stiefelhagen, and Luc Van Gool. Large scale holistic video understanding. In *European Conference on Computer Vision*, pages 593–610. Springer, 2020. 3

[15] Haoqi Fan, Yanghao Li, Bo Xiong, Wan-Yen Lo, and Christoph Feichtenhofer. Pyslowfast. https://github.com/facebookresearch/slowfast, 2020. 13

[16] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6201–6210, 2019. 2, 6, 10, 13

[17] J. Gao, C. Sun, Zhenheng Yang, and R. Nevatia. Tall: Temporal activity localization via language query. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5277–5285, 2017. 3

[18] Timnit Gebru, J. Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, H. Wallach, Hal Daumé, and K. Crawford. Datasheets for datasets. *ArXiv*, abs/1803.09010, 2018. 9, 16

[19] Rohit Girdhar, J. Carreira, C. Doersch, and Andrew Zisserman. Video action transformer network. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 244–253, 2019. 2

[20] Raghav Goyal, S. Kahou, Vincent Michalski, Joanna Materzynska, S. Westphal, Heuna Kim, V. Haenel, Ingo Fründ, Peter Yianilos, Moritz Mueller-Freitag, F. Hoppe, Christian Thurau, I. Bax, and R. Memisevic. The "something something" video database for learning and evaluating visual common sense. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5843–5851, 2017. 3

[21] C. Gu, C. Sun, Sudheendra Vijayanarasimhan, C. Pantofaru, D. Ross, G. Toderici, Y. Li, S. Ricco, R. Sukthankar, C. Schmid, and J. Malik. Ava: A video dataset of spatio-temporally localized atomic visual actions. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6047–6056, 2018. 2, 3, 9, 10

[22] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *ArXiv*, abs/1505.04474, 2015. 2, 3

[23] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask r-cnn. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42:386–397, 2020. 10

[24] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–970, 2015. 2, 3, 9

[25] Lisa Anne Hendricks, O. Wang, E. Shechtman, Josef Sivic, Trevor Darrell, and Bryan C. Russell. Localizing moments in video with natural language. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5804–5813, 2017. 3

[26] Y. Hong, Tongtao Zhang, Timothy J. O'Gorman, Sharone Horowit-Hendler, Huai zhong Ji, and Martha Palmer. Building a cross-document event-event relation corpus. In *LAW@ACL*, 2016. 4, 10

[27] Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017. 20

[28] Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. Movienet: A holistic dataset for movie under-

standing. In *The European Conference on Computer Vision (ECCV)*, 2020. 3, 9

[29] 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. 3

[30] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black. Towards understanding action recognition. In *International Conf. on Computer Vision (ICCV)*, pages 3192–3199, Dec. 2013. 3

[31] W. Kay, J. Carreira, K. Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, F. Viola, T. Green, T. Back, A. Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *ArXiv*, abs/1705.06950, 2017. 2, 3, 6, 9

[32] Mert Kilickaya, Aykut Erdem, Nazli Ikizler-Cinbis, and Erkut Erdem. Re-evaluating automatic metrics for image captioning. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 199–209, Valencia, Spain, Apr. 2017. Association for Computational Linguistics. 2

[33] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015. 13

[34] Y. Kong and Yun Fu. Human action recognition and prediction: A survey. *ArXiv*, abs/1806.11230, 2018. 3

[35] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *International Conference on Computer Vision (ICCV)*, 2017. 2, 5, 12

[36] Hilde Kuehne, Hueihan Jhuang, E. Garrote, T. Poggio, and Thomas Serre. Hmdb: A large video database for human motion recognition. *2011 International Conference on Computer Vision*, pages 2556–2563, 2011. 3

[37] Jie Lei, Licheng Yu, Mohit Bansal, and T. Berg. Tvqa: Localized, compositional video question answering. In *EMNLP*, 2018. 3

[38] Jie Lei, Licheng Yu, T. Berg, and Mohit Bansal. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *ECCV*, 2020. 3

[39] Ang Li, Meghana Thotakuri, D. Ross, J. Carreira, Alexander Vostrikov, and Andrew Zisserman. The ava-kinetics localized human actions video dataset. *ArXiv*, abs/2005.00214, 2020. 3

[40] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+ language omni-representation pre-training. In *EMNLP*, 2020. 3

[41] Manling Li, Alireza Zareian, Q. Zeng, Spencer Whitehead, Di Lu, Huai zhong Ji, and Shih-Fu Chang. Cross-media structured common space for multimedia event extraction. In *ACL*, 2020. 3

[42] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. 7

[43] T. Lin, X. Liu, Xin Li, E. Ding, and Shilei Wen. Bmn: Boundary-matching network for temporal action proposal generation. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3888–3897, 2019. 2

[44] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. Bsn: Boundary sensitive network for temporal action proposal generation. In *European Conference on Computer Vision*, 2018. 2

[45] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 10, 15

[46] Y. Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692, 2019. 6, 7

[47] Xiaoqiang Luo. On coreference resolution performance metrics. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 25–32, Vancouver, British Columbia, Canada, Oct. 2005. Association for Computational Linguistics. 7, 16

[48] Louis Mahon, Eleonora Giunchiglia, B. Li, and Thomas Lukasiewicz. Knowledge graph extraction from videos. *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 25–32, 2020. 2

[49] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-End Learning of Visual Representations from Uncurated Instructional Videos. In *CVPR*, 2020. 3

[50] 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. In *ICCV*, 2019. 2, 3, 9

[51] Mathew Monfort, B. Zhou, Sarah Adel Bargal, Alex Andonian, Tom Yan, K. Ramakrishnan, L. Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, and A. Oliva. Moments in time dataset: One million videos for event understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42:502–508, 2020. 3, 9

[52] N. Moosavi and Michael Strube. Which coreference evaluation metric do you trust? a proposal for a link-based entity aware metric. In *ACL*, 2016. 7, 16

[53] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019. 13

[54] Martha Palmer, Paul Kingsbury, and Daniel Gildea. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31:71–106, 2005. 2, 3, 4, 9, 10

[55] Kishore Papineni, S. Roukos, T. Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002. 16

[56] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming

Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 13

[57] Sameer Pradhan, Xiaoqiang Luo, Marta Recasens, Eduard Hovy, Vincent Ng, and Michael Strube. Scoring coreference partitions of predicted mentions: A reference implementation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 30–35, Baltimore, Maryland, June 2014. Association for Computational Linguistics. 16

[58] Sarah Pratt, Mark Yatskar, Luca Weihs, Ali Farhadi, and Aniruddha Kembhavi. Grounded situation recognition. In *ECCV*, 2020. 2, 3

[59] Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, B. Schiele, and Manfred Pinkal. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1:25–36, 2013. 3

[60] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and B. Schiele. A dataset for movie description. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3202–3212, 2015. 3, 4, 5, 9, 10

[61] Arka Sadhu, K. Chen, and R. Nevatia. Video object grounding using semantic roles in language description. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10414–10424, 2020. 2, 3, 9

[62] Peng Shi and Jimmy Lin. Simple bert models for relation extraction and semantic role labeling. *ArXiv*, abs/1904.05255, 2019. 4, 5, 10

[63] Gunnar A. Sigurdsson, G. Varol, X. Wang, Ali Farhadi, I. Laptev, and A. Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*, 2016. 3

[64] Carina Silberer and Manfred Pinkal. Grounding semantic roles in images. In *EMNLP*, 2018. 3

[65] K. Soomro, A. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *ArXiv*, abs/1212.0402, 2012. 3

[66] Emma Strubell, Patrick Verga, Daniel Andor, David Weiss, and Andrew McCallum. Linguistically-informed self-attention for semantic role labeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5027–5038, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics. 9

[67] Chen Sun, Abhinav Shrivastava, Carl Vondrick, Kevin Murphy, Rahul Sukthankar, and Cordelia Schmid. Actor-centric relation network. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 335–351, Cham, 2018. Springer International Publishing. 2

[68] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. 6

[69] Yansong Tang, Dajun Ding, Yongming Rao, Y. Zheng, Danyang Zhang, L. Zhao, Jiwen Lu, and J. Zhou. Coin: A large-scale dataset for comprehensive instructional video analysis. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1207–1216, 2019. 3, 9

[70] Makarand Tapaswi, Y. Zhu, R. Stiefelhagen, A. Torralba, R. Urtasun, and S. Fidler. Movieqa: Understanding stories in movies through question-answering. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4631–4640, 2016. 3

[71] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 6, 7

[72] Paul Vicol, Makarand Tapaswi, Lluis Castrejon, and Sanja Fidler. Moviegraphs: Towards understanding human-centric situations from videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3, 9

[73] Marc B. Vilain, J. Burger, J. Aberdeen, D. Connolly, and L. Hirschman. A model-theoretic coreference scoring scheme. In *MUC*, 1995. 7, 16

[74] Oriol Vinyals, A. Toshev, Samy Bengio, and D. Erhan. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:652–663, 2017. 2

[75] L. Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, D. Lin, X. Tang, and L. Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, 2016. 2

[76] X. Wang, Ross B. Girshick, A. Gupta, and Kaiming He. Non-local neural networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018. 6

[77] Xin Eric Wang, Jiawei Wu, Junkun Chen, Lei Li, Y. Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4580–4590, 2019. 2, 3, 5

[78] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, Oct. 2020. Association for Computational Linguistics. 13

[79] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krähenbühl, and Ross B. Girshick. Long-

term feature banks for detailed video understanding. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 284–293, 2019. 2

[80] Yu Xiong, Qingqiu Huang, Lingfeng Guo, Hang Zhou, Bolei Zhou, and Dahua Lin. A graph-based framework to bridge movies and synopses. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 9

[81] 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 *ACM Multimedia*, 2017. 2, 3

[82] J. Xu, T. Mei, Ting Yao, and Y. Rui. Msr-vtt: A large video description dataset for bridging video and language. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5288–5296, 2016. 2, 3, 5

[83] Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. Situation recognition: Visual semantic role labeling for image understanding. In *Conference on Computer Vision and Pattern Recognition*, 2016. 2, 3, 9

[84] Sarah Young, Dhiraj Gandhi, Shubham Tulsiani, Abhinav Gupta, P. Abbeel, and Lerrel Pinto. Visual imitation made easy. *ArXiv*, abs/2008.04899, 2020. 2

[85] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *AAAI*, pages 9127–9134, 2019. 2, 3

[86] 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. In *AAAI*, 2017. 3

[87] H. Zhang, Yi-Xiang Zhang, B. Zhong, Qing Lei, Lijie Yang, Ji-Xiang Du, and Duan-Sheng Chen. A comprehensive survey of vision-based human action recognition methods. *Sensors (Basel, Switzerland)*, 19, 2019. 3, 9

[88] Zixing Zhang, Zhou Zhao, Yang Zhao, Q. Wang, H. Liu, and L. Gao. Where does it exist: Spatio-temporal video grounding for multi-form sentences. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10665–10674, 2020. 3

[89] Hang Zhao, Antonio Torralba, Lorenzo Torresani, and Zhicheng Yan. Hacs: Human action clips and segments dataset for recognition and temporal localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8668–8678, 2019. 3

[90] Luowei Zhou, Nathan Louis, and Jason J Corso. Weakly-supervised video object grounding from text by loss weighting and object interaction. In *British Machine Vision Conference*, 2018. 3, 9

[91] Luowei Zhou, Yingbo Zhou, Jason J. Corso, R. Socher, and Caiming Xiong. End-to-end dense video captioning with masked transformer. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8739–8748, 2018. 2

[92] Linchao Zhu and Y. Yang. Actbert: Learning global-local video-text representations. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8743–8752, 2020. 3