Visual Semantic Role Labeling for Video Understanding

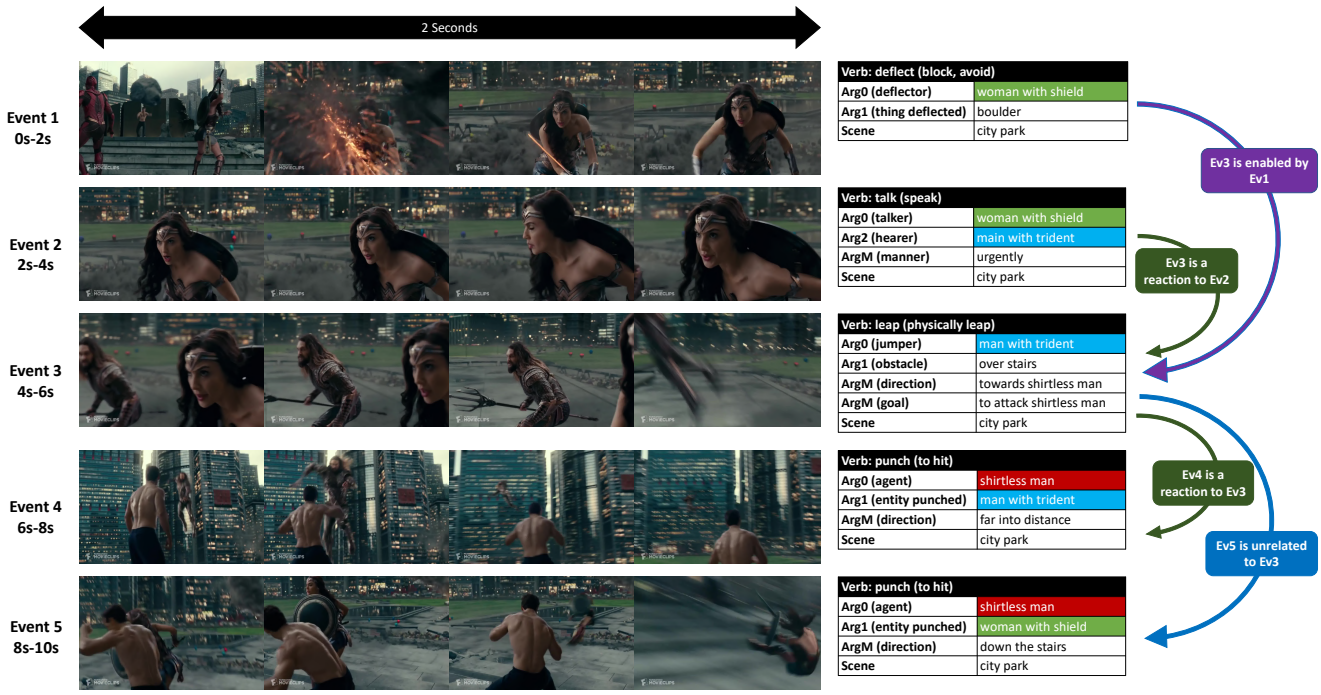


Figure 1: **A sample video and annotation from VidSitu.** The figure shows a 10-second video annotated with 5 events, one for each 2-second interval. Each event consists of a verb (like "deflect") and its arguments (like *Arg0 (deflector)* and *Arg1 (thing deflected)*). Entities that participate in multiple 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 $27K$ *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 (*$\sim 3K$*) 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 will be released publicly.*

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 [34, 81, 76], answering queries

about them [84, 80], retrieving visual content [49], building knowledge graphs [47] and even teaching embodied agents to act and interact with the real world [83].

Parsing video content is an active area of research with much of the focus centered around tasks such as action classification [30], localization [23] and spatio-temporal detection [20]. 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 [31, 73].

Recent work in the image domain [82, 57, 21] 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 27K 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 [53] 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 will be released publicly.

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, 15, 74], localization [43, 42] and spatio-temporal detection [18], video description [76, 34], question answering [84], and object grounding [60]. Tasks like detecting atomic actions at 1 second intervals [18, 78, 66] are short horizon tasks whereas ones like summarizing 180 second long videos [90] are extremely long horizon tasks. In contrast, our proposed task of VidSRL operates on 10 second video at 2 second intervals. 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 [30, 23, 20], over the past few years. While early datasets were small datasets with several hundred or thousand examples[64, 35], recent datasets are massive[49] enabling researchers to train large neural models and also employ pre-training strategies[48, 91, 39]. Section 4, Table 3 and Figure 2 provide a comparison of our proposed

[1] Following nomenclature introduced in ImSitu[82], 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[30], ActivityNet [23], Moments in Time [50], Something-Something[19]
Action Localization	Action Labels, Temp. Segments	ActivityNet, Thumos[28], HACS [88], Tacos[58], Charades[62], COIN[68]
Spatio-Temporal Detection	Action Labels, Temp. Segments, BBoxes	AVA[20], AVA-Kinetics[38], EPIC-Kitchens [12], JHMDB[29]
Video Description	Captions, Temp. Segments	ActivityNet[23], Vatex[76], YouCook[13], MSR-VTT [81] , LSMDC [59]
Video QA	Q/A, Subtitle or Script (optional)	MSRVTT-QA[80], VideoQA[85], ActivityNetQA[84], TVQA[36], MovieQA[69]
Text to Video Retrieval	Text Query, ASR output (optional)	HowTo100M[49], TVR[37], DiDeMo[24], Charades-STA[16]
Video Object Grounding	Text Query, Temp. Segments, BBoxes	ActivityNet-SRL[60], YouCookII[89], VidSTG [87],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.

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, 33, 86] 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 [82, 57], visual semantic role labeling [21, 40, 63] 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 [20], movie description [59], movie question answering [69] , story-based retrieval [3] and generating social graphs [71] tasks. 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 [53]. 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 [20] 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 $\{E_i\}_{i=1}^{5}$.

Describing an Event. We describe an event through a verb and its arguments. For verbs, we follow recent work in action recognition like ActivityNet [23] and Moments in Time [50] 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 [53]. 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 [25] 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. Since annotating videos with such rich annotations is expensive, several crucial dataset curation decisions were made to enhance the efficiency and effectiveness of the annotation process, which we describe below.

4.1. Dataset Curation

Video Source Selection. It is crucial to choose a video source that allows sampling diverse and complex situations. Instructional domain video sets [49, 68, 58, 89] typically contain a single agent in a constrained setting (*e.g*. cooking), and open-domain video sets [30, 23] while more diverse often focus on a single action (the class-label). Videos from movies are well suited for VidSRL since they are naturally diverse (wide-range of movie genres) and they often involve multiple interacting entities. Also, scenarios in movies typically play out over multiple shots which makes movies a challenging testbed for long-range 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 [53] vocabulary of $\sim 6k$ verb-senses. We keep all $3.7K$ verbs with a single sense and of the remaining verbs-senses, we discard ones that are too fine-grained (for instance the verb "go" has 23 verb senses) or non visual (e.g. "run" in the sense of running a business). To reduce this set down to ones useful for describing movies, we discard verbs that do not appear in the MPII-Movie Description (MP2D) dataset [59] (verbs extracted using a semantic-role parser [61]). This results in a final 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 coreferencing). Details on annotation interface, quality control, and reward are provided in supplementary material.

Dataset splits. VidSitu is split into train, validation and test sets via a $80:10:10$ split, ensuring that videos from the same movie end up in exactly one of those sets. Table 2 summarizes these statistics of these splits.

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 determined the number of annotations required for the validation and test sets to ensure that the metrics accurately reflected the performance of models. We obtain multiple annotations for all videos in our 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.

	Train	Val	Test	Total
# Movies	2431	294	301	3026
# Videos	23626	1804	1985	27415
# Clips	118130	9020	9925	137075
# Verbs Ann / Clip	1	10	10	
# Verb Ann	118130	90200	99250	307580
# Unique Verb Tuples	23196	1801	1929	26926
# Values Ann / Role	1	3	3	
# Role Ann	118130	27060	29775	174965

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

4.2. 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 [81], MPII-Movie Description [59], ActivityNet Captions [34], and Vatex-en [76] (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 [61].

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 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.

[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.

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 [15] networks (Step 1 in Fig. 3). We consider variants of I3D both with and without Non-Local blocks [75] 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 [30]. 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 [67] that consist of an encoder and a decoder (Step 2(a,b) in Fig. 3). Specifically, independent event features are fed through a transformer [70] 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

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**	✓	27.4k	137k	137k	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.

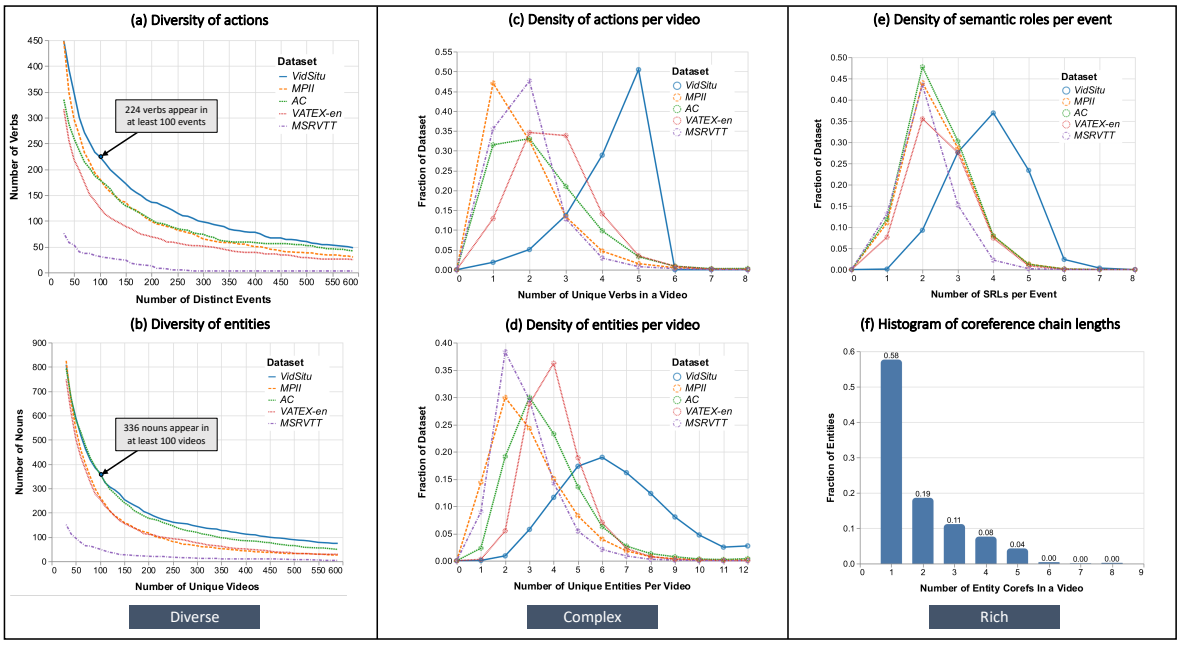


Figure 2: **Data analysis.** An analysis of VidSitu in comparison to other large scale relevant video datasets. We focus on the **diversity** of actions and entities in the dataset (a and b), the **complexity** of the situations measured in terms of the number of unique verbs and entities per video (c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

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 [45]-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 Vid-

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	✗	✗	32.76	39.59	45.10	30.30	**50.20**	26.56	38.10	41.60	43.90	37.76	**53.28**	**33.50**
TxDec	✗	✗	34.28	38.68	43.50	29.00	40.22	21.24	37.10	40.00	44.30	32.90	43.62	24.48
Vid TxDec	SlowFast	✗	42.11	38.66	**50.19**	36.83	34.81	24.79	45.08	40.21	**51.90**	40.90	36.22	27.74
Vid TxEncDec	SlowFast	✓	**43.20**	**40.49**	50.03	**38.14**	**48.99**	**29.52**	**46.34**	**41.80**	49.69	**42.22**	**51.15**	32.59
Vid TxDec	I3D	✗	40.24	39.55	46.07	36.21	36.84	25.90	44.00	41.56	50.64	41.22	38.18	29.26
Vid TxEncDec	I3D	✓	**46.89**	**42.05**	**52.89**	**42.38**	48.45	**32.97**	**48.95**	**43.30**	**52.65**	**46.16**	50.95	**35.67**
Human*			84.78	39.53	92.61	79.14	70.49	69.50	83.87	40.46	89.13	78.85	72.64	70.93

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	✗	28.36	62.01	4.65	29.34	64.59	4.61
I3D+NL	✗	28.98	65.7	4.33	29.6	67.13	4.24
Slow+NL	✗	30.49	63.99	5.27	30.77	66.53	5.08
SlowFast+NL	✗	31.3	66.83	5.83	31.48	**69.35**	5.47
I3D	✓	31.87	62.16	15.62	29.02	58.99	15.06
I3D+NL	✓	31.83	62.11	15.19	**32.99**	62.46	15.43
Slow+NL	✓	**40.07**	**71.02**	**17.06**	31.63	62.07	**17.46**
SlowFast+NL	✓	**40.3**	**70.18**	**20.55**	**42.17**	**71.04**	**21.45**

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.

Situ 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

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

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) [41] is shown for completeness.

Since VidSitu provides entity coreference links across events and roles, we use LEA [51] a link-based co-reference metric to measure coreferencing capability. Other metrics (MUC [72], BCUBE [2], CEAFE [46]) 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.

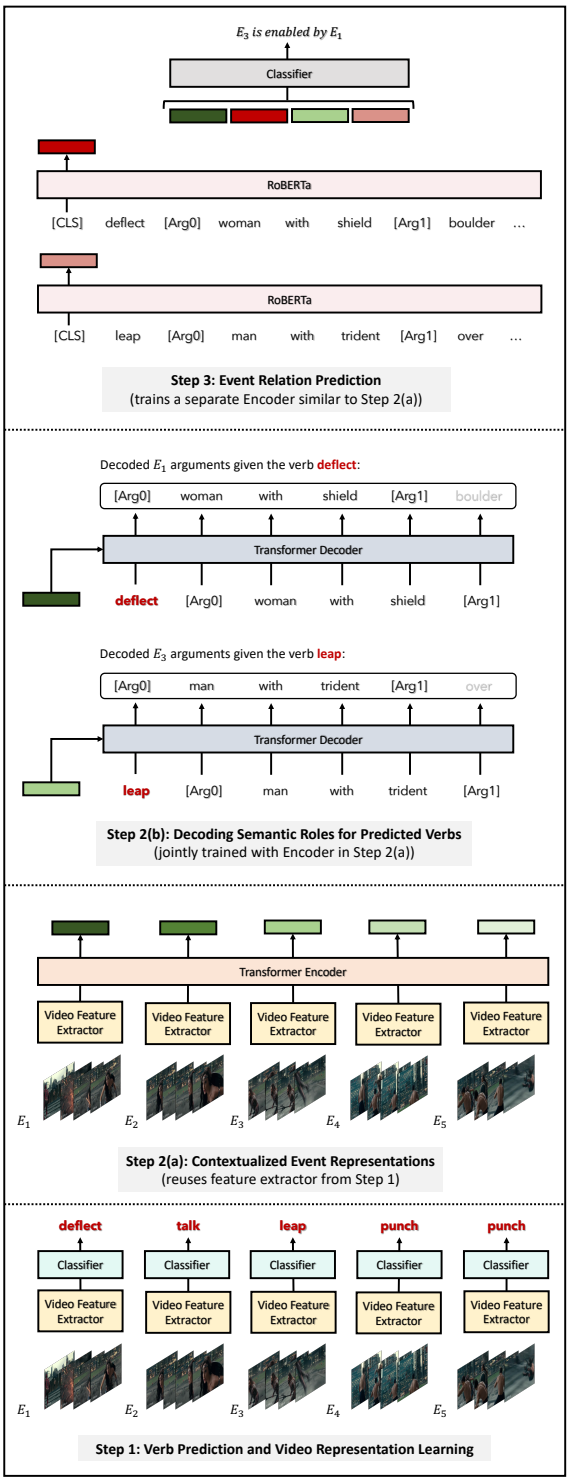


Figure 3: **Models.** The figure illustrates our baselines for verb, semantic role, and event prediction using state-of-the-art network components such as SlowFast [15] network for video feature extraction, transformers [70] for encoding events in a video and verb-conditional decoding of roles, and RoBERTa [45] language encoder for event-relation prediction.

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. Finally, we observe a large gap 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 and provides a huge room for improvement. Due to space constraints in visualizing videos, we defer **qualitative analysis** to the supp material.

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.

Appendix

Errata: In Figure 1, Event 2 Arg2 should be "man with trident" instead of "main with trident".

Appendix provides details on:

1. A Brief Summary of Semantic Roles, and their usage in our paper.
2. Details on Dataset Curation and Annotation Interface
3. Additional Dataset Statistics
4. Additional Implementation Details
5. Details on Lea-Soft along with Tables with All Metrics
6. DataSheet [17] for VidSitu
7. Qualitative Analysis of Data (this is attached as a video file in the zip folder).

A. Semantic Roles: A Brief Summary

Semantic Role Labeling attempts to abstract out at a high-level who does what to whom [65]. It is a popular natural language task which attempts at obtaining such structured outputs from natural language descriptions. As such there are multiple sources to obtain semantic roles such as FrameNet [4], PropBank [53] and VerbNet [7]. Prior work on situation recognition in images (ImSitu) [82] have curated list of verbs (situations) from FrameNet, and action recognition dataset (Moments in Time) [50] have curated action vocabulary from VerbNet. However, we qualitatively found both vocabulary to be insufficient to represent actions, and thus chose PropBank which contained action-oriented verbs. As such, PropBank has been used for video object grounding [60] but not in the context of collecting semantic roles from visual data.

PropBank contains a set of numbered semantic roles for each verb ranging from Arg0 to Arg4. Each numbered argument has a specific definition for a particular verb but some themes are similar across verbs (adapted from PropBank annotation guidelines [6][3]). For the verb "throw":

- Arg0: Agent – object performing the action. For *e.g.* "person"
- Arg1: Patient – object on which action is performed. For *e.g.* "ball"
- Arg2: Instrument, Benefactive, Attribute. For *e.g.* "towards a basket"
- Arg3: Starting Point
- Arg4: Ending Point
- ArgM: Modifier – location (LOC), manner(MNR), direction (DIR), Purpose (PRP), Goal (GOL), Temopral (TMP), Adverb (ADV)

[3] http://clear.colorado.edu/compsem/documents/propbank_guidelines.pdf

In general, we noticed that Arg3 and Arg4 were exceedingly rare for visual verbs, thus we restrict our attention to Arg0, Arg1, Arg2 for numbered arguments. For modifier arguments, we found Location (LOC) to be universally valid for all video segments. Thus, for those verbs where LOC doesn't apply usually, we additionally add a semantic role "Scene" which refers to "where" the event takes place (such as "living room", "near a lake"). Other arguments were chosen based on their appearance in MPIID dataset, and we most commonly used Manner (which suggests "how" the action takes place) and Direction (details in the Section B). For rest of the paper, we use ALoc, ADir, AMnr, and AScn to denote location, direction, manner and scene arguments respectively.

B. Dataset Collection

In this section we describe details on dataset collection including curation of verbs and arguments, followed by details on annotation interface, quality control and reward structure.

B.1. Dataset Curation

We provide more details on Dataset Curation which were omitted from Section 4.1 of the main paper.

Video Source Selection. As suggested in the Section 4.1 we aimed at a domain with two criterion: the videos should be by themselves cover diverse situations ("climb" verb should not just be associated with rocks or mountains, but also things like top of a car), and that the each video should contain complex situation (the video shouldn't depict someone doing the same task over extended period of time, which would lower chances of finding meaningful event relations and be repetitive in verbs and arguments over the entire video).

After a brief qualitative analysis, we found instruction domain videos (HowTo100M [49], YouCookII [89], COIN [68]) to have very fine-grained actions with less diversity and less complexity within small segments, open domain sources (ActivityNet [23], Moments in Time [50], Kinetics [30], HACS[86]) to be somewhat diverse but low complexity within a small segment. This led us to Movie domain which had appreciable diversity as well as complexity.

We converged on using MovieClips [3] rather than other movie sources such as MPII [59], since MovieClips already provide one-stage of filtering to provide interesting videos. While using the same movies as used in AVA[20] was an option, we found that the video retention was quite low (around 20% of the movie are removed from youtube), and the movie contained long contiguous segments with low complexity. We also note some other datasets like MovieNet [27], Movie Synopsis Dataset [79], Movie Graphs [71] do not provide movie videos and cannot be used for collecting annotations. One demerit of using movie

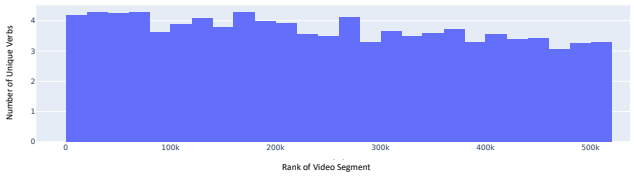


Figure 1: Bar graph showing number of unique verbs with respect to the rank of the video segment as computed via our heuristic based on predicted labels from SlowFast Network [15] trained on AVA[20].

domain is that the verb distributions are skewed towards actions like "talk", "walk", "stare". Despite this we find the videos to be reasonably complex.

Video Selection. MovieClips spans a total of $1k$ Hours which is far beyond what can be reasonably annotated. To best utilize available annotation budget, we are primarily interested in identifying video segments depicting complex situations with a high precision while avoiding visually uneventful segments common in movies such as those simply involving actors engaged in dialogue.

To avoid such segments, we use the following heuristic: a video with more atomic actions per person is likely to be more eventful. So, we divide all movieclips into 10 second videos with a stride of 5 seconds, obtain human bounding boxes from the MaskRCNN [22] object detector trained on the MSCOCO [44] dataset, predict atomic actions for each detected person using the SlowFast [15] activity recognition model trained on the AVA [20] dataset, and rank all videos by the average number of unique atomic actions per person in the video. In particular, we discard labels such as "talk", "listen", "stand" and "sit" as these atomic actions didn't correlate with complexity of situations. Since "action" sequences like "fight scenes" are favored by our ranking measure, we use simple heuristic of removing "martial arts" actions to avoid oversampling such scenes and improve diversity of situations represented in the selected videos.

To confirm the usefulness of the proposed heuristic, we conduct an experiment where we annotate $1k$ videos chosen uniformly sampled across the entire dataset (as shown in Figure 1). Reducing number of unique verbs shows the effectiveness of our heuristic and suggests at least $80K$ videos segments (which translates to $27K$ non-overlapping video segments) can be richly annotated.

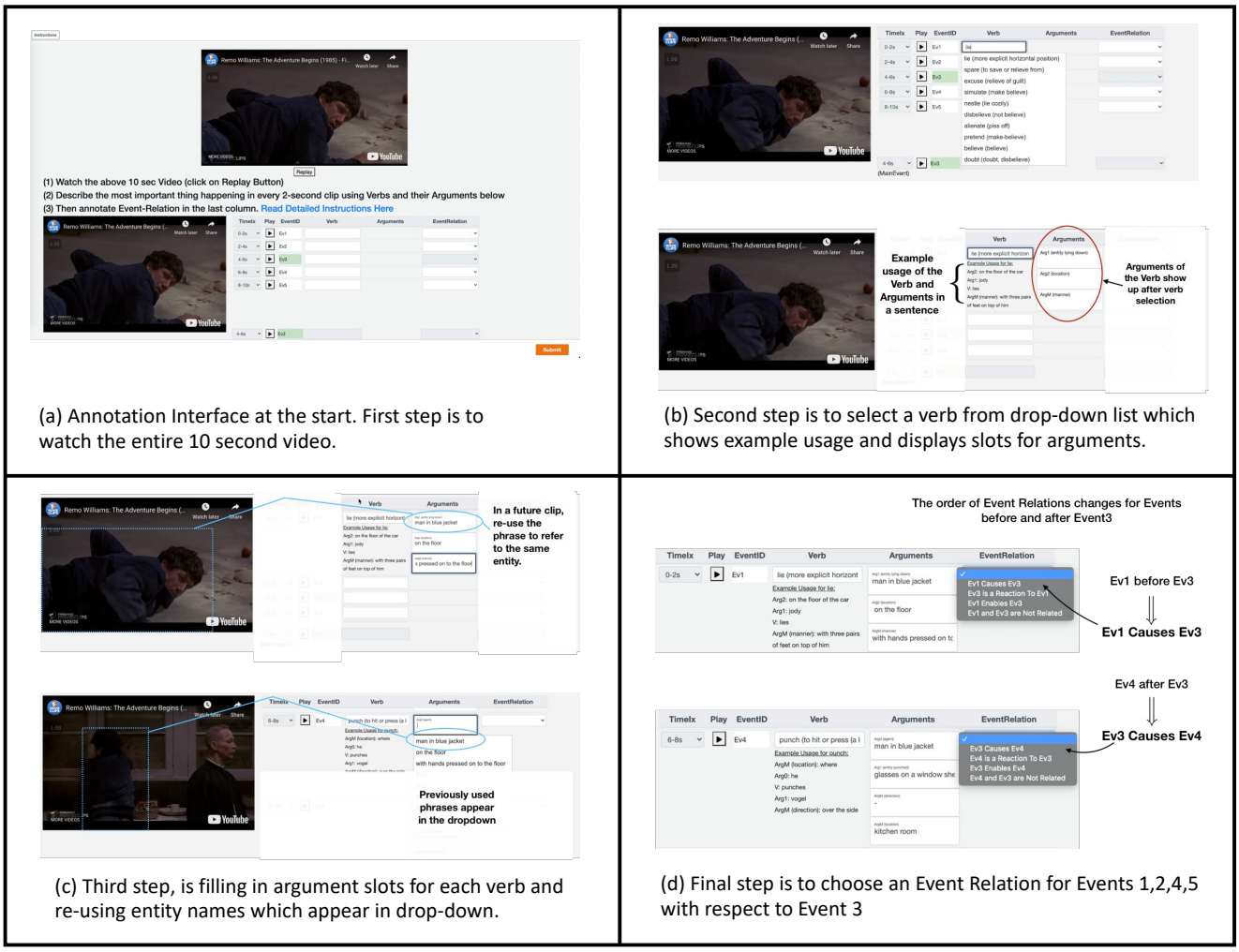
For final video selection, we randomly choose set of videos from the top-K ranks, such that the newly chosen videos don't overlap with already chosen videos, and that no more than 3 videos are uploaded from the same youtube video within a particular batch.

Curating Verb Senses. To curate verb senses, we follow a two-step process: from the initial list of $\sim 6k$ verb senses in PropBank [53], first we manually filter verb senses which share the same lemmatized verb (as previously stated "go" has 23 verb senses) to retain only "visual" verb senses (for instance we remove the verb sense of "run" which refers to running a business). We keep all $3.7K$ verbs with a single verb sense and of the remaining 2364 verbs-senses (shared across 809 verbs) we retain 629 verb senses (shared across 561 verbs). Second, to further restrict the set of verbs to those useful for describing movies, we discard verbs that do no appear at all in the MPII-Movie Description (MP2D) dataset [59]. To extract verbs from the descriptions we use a semantic-role parser [61]. This results in a final set of 2154 verb-senses.

Curating Argument Roles. Once we have curated the verb-senses from PropBank, we aim to delegate a set of argument roles for each verb-sense which would be filled based on the video. While PropBank provides numbered arguments for each verb-sense there are two issues with directly using them: first, some arguments are less relevant for visual scenes (for instance Arg1 (utterance) for "talk" is not visual), second, auxiliary arguments like direction and manner are not provided (for instance direction and manner for "look" are important to describe a scene). To address this issue, we re-use the MP2D dataset to inform us what arguments are used with the verbs. For each verb, we choose set of 5 most frequently used argument role-set and use their union. We also remove roles such as TMP (usually referring to words like "now", "then") since temporal context is implicit in our annotation structure. We also removed roles like ADV (adverb) which were too infrequent. Finally, we use the following modifier roles: "Manner", "Location", "Direction", "Purpose", "Goal", but note that "purpose" and "goal" were restricted to a small number of verbs and hence not considered for evaluation.

We further added the modifier role "Scene" which describes "where" the event takes place, and only applies to verbs which don't have "Location". For instance, "stand" has the argument role "location" which refers to "where" the person is standing and doesn't have "Scene", whereas "run" doesn't contain "location" and hence contains "Scene". In general, "Scene" refers to the "place" of the event such as "in an alleyway" or "near a beach".

Event Relations. We started with the set of three event relations namely: no relation (Events A and B are unrelated), causality (Event B is Caused By Event A *i.e.* B happens directly as a result of A) and contingency based (Event B is Enabled By Event A *i.e.* A doesn't directly cause B but B couldn't have happened without A happening first) on prior work in cross-document event relations [25]. However, we found adding an additional case of "Reaction To" for causality helpful to distinguish between event relations. For instance, in the case "X punches Y" followed by "Y falls down" would be definitely "B is Caused By A", however for the case "X punches Y" followed by "Y crouches"



(a) Annotation Interface at the start. First step is to watch the entire 10 second video.

(b) Second step is to select a verb from drop-down list which shows example usage and displays slots for arguments.

(c) Third step, is filling in argument slots for each verb and re-using entity names which appear in drop-down.

(d) Final step is to choose an Event Relation for Events 1,2,4,5 with respect to Event 3

Figure 2: Illustration of our annotation interface. (a) depicts the initial screen an annotator sees. In the first step, one needs to watch the entire 10 second video. (b) depicts the second step of choosing a verb from a drop-down which contains verb senses obtained from PropBank. After selecting a verb, an example usage is shown along with corresponding argument roles which need to be filled. (c) depicts filling the argument slots for each verb which can be phrases of arbitrary length. Each filled in phrase can be re-used in a subsequent slot, to enforce co-reference of the entities. (d) shows the final step of choosing event relations once all the arguments for all events are filled. The event relations should be classified based on causality and contingency for Events 1,2,4,5 with respect to Event 3.

it is unclear if "B is Caused By A" since Y makes a voluntary decision to crouch. As a result, we call this relation "B is a Reaction To A".

B.2. Annotation pipeline

With videos, the list of verb-sense and their roles curated, we are now ready to crowd-source annotations on Amazon Mechanical Turk (AMT).

Annotation Interface. Figure 2 shows screenshots depicting our annotation interface. For annotating a given 10 second video, the assigned worker is instructed to first watch the entire 10-second video (Figure 2 (a)). Then for every 2 second interval, the annotator selects a verb corresponding to the most salient event from our curated list of verb-senses using a search-able drop-down menu. Once the verb is chosen, slots for the corresponding roles are displayed along with an example usage (Figure 2 (b)). The worker fills in the values for each role using free-form text (typically a short phrase). When referring to an entity, we instruct the worker to use phrases that uniquely identify the entity in the full 10 second video. Furthermore, these phrases can be reused in filling semantic-roles in other events within the video, which provides the co-reference information about the entities *i.e.* co-referenced entities are

	Acc@1		Acc@5		Recall@5	
	10 A	20 A	10 A	20 A	10 A	20 A
Majority	0.20	0.21	0.66	0.75	0.03	0.02
Human	0.62	0.71	0.96	1.00	0.64	0.59

Table 1: 10A and 20A denote 10 and 20 annotations respectively. Majority denotes choosing most frequent verbs for the validation set.

	Total	Caused By	Reaction To	Enabled By	No Relation
Train Set	94016	16.94	24.05	33.76	25.25
Val Set	7216	15.06	22.8	34.67	27.47
Val Set*	5502 (76.24%)	12.4	21.25	37.15	29.21
Test Set	7940	21.73	22.95	35.05	20.28
Test Set*	6135 (77.26%)	15.14	19.43	40.83	24.6

Table 2: The distribution of Event Relations before and after filtering by taking consensus of at least two workers *i.e*. we consider only those instances where two workers agree on the event relation when given the verb.

maintained via exact-string match (Figure 2 (c)). Once all verbs and their roles are annotated, we ask the worker to label the relation of Events 1, 2, 4, and 5 with respect to Event 3 (Figure 2 (d)). Note that the order of causality and contingency is different for Events 4,5 compared to Events 1,2 respecting the temporal order.

Worker Qualification and Quality Control. To ensure that annotators have understood the task requirements, we put up a qualification task where a worker has to successfully annotate 3 videos. These annotations are manually verified by the first author who then provides feedback on their annotations. To filter potential workers, we restrict to more than 95% approval rate and having done at least 500 tasks. In total we qualified around 120 annotators, with at least 60 workers annotating more than 30 videos every batch of $2K$ videos.

In addition to manual qualification, we put automated checks one average number of unique verbs provided within a video, and average description lengths. We further manually inspect around 3 random samples from every annotator after every $3K - 5K$ videos and provide constant feedback.

Annotating Validation and Test Sets.

We ran a controlled experiments using 100 videos and annotated 25 verbs for each event. We report the human agreement in Table 1. To compute human agreement score for any event, we use one human annotation (out of 25) as a prediction and the remaining 10 or 20 annotations as ground-truths (denoted by 10A or 20A). The final score is the average over all possible prediction/ground-truth partitions. Essentially, we find that even moving from 10 to 20 annotations, the human agreement improves from 62% to 71% which suggests even at higher number of annotations, we receive verbs which are suggested by a single annotator (and hence no agreement). This rules out metrics like accuracy, precision, or F1 scores because they would penalize predictions that may be correct but are not present in a reasonably sized set of ground truth annotations. This analysis leads us to the metric Recall@5 which measures if the verbs most agreed upon by humans are indeed recalled by the model in its top-5 predictions.

Furthermore, this prompts us to collect the annotations for validation and test set in two-stages, in the first stage we collect 9 additional annotations for verb and then in the second-stage 3 annotations for argument roles and event relations given the verb (we choose the set of verbs chosen by the annotator with the highest agreement, followed by highest number of unique verbs within the video). We find this two-stage process to be of similar cost of obtaining 5 independent annotations but with the added advantage of being comparable across annotations. In total we annotation 3789 videos for validation and test sets.

Reward. We set the reward for annotating one 10-second video (for training videos) to $\$0.75$ after estimating the average time of completing an annotation to be around 5mins. This translates to around $\$9$/hour. Overall, we received generous reviews for the reward on popular turk management website. For validation and test sets, we set the reward to $\$0.2$ for the first stage (collecting only verbs from 9 annotators and $\$0.7$ for the second-stage (collecting argument and event relations from 3 annotators). As a result, the cost for annotating a single video in the validation and test set turns out to be $\$3.9$ ($0.2 \times 9 + 0.7 \times 3$) which is around $5.2\times$ the cost of annotating a single training video. Total cost for the process comes around $\$36.7K$ (note: this doesn't account for pilot experiments, qualifications, and discarded annotations due to human errors).

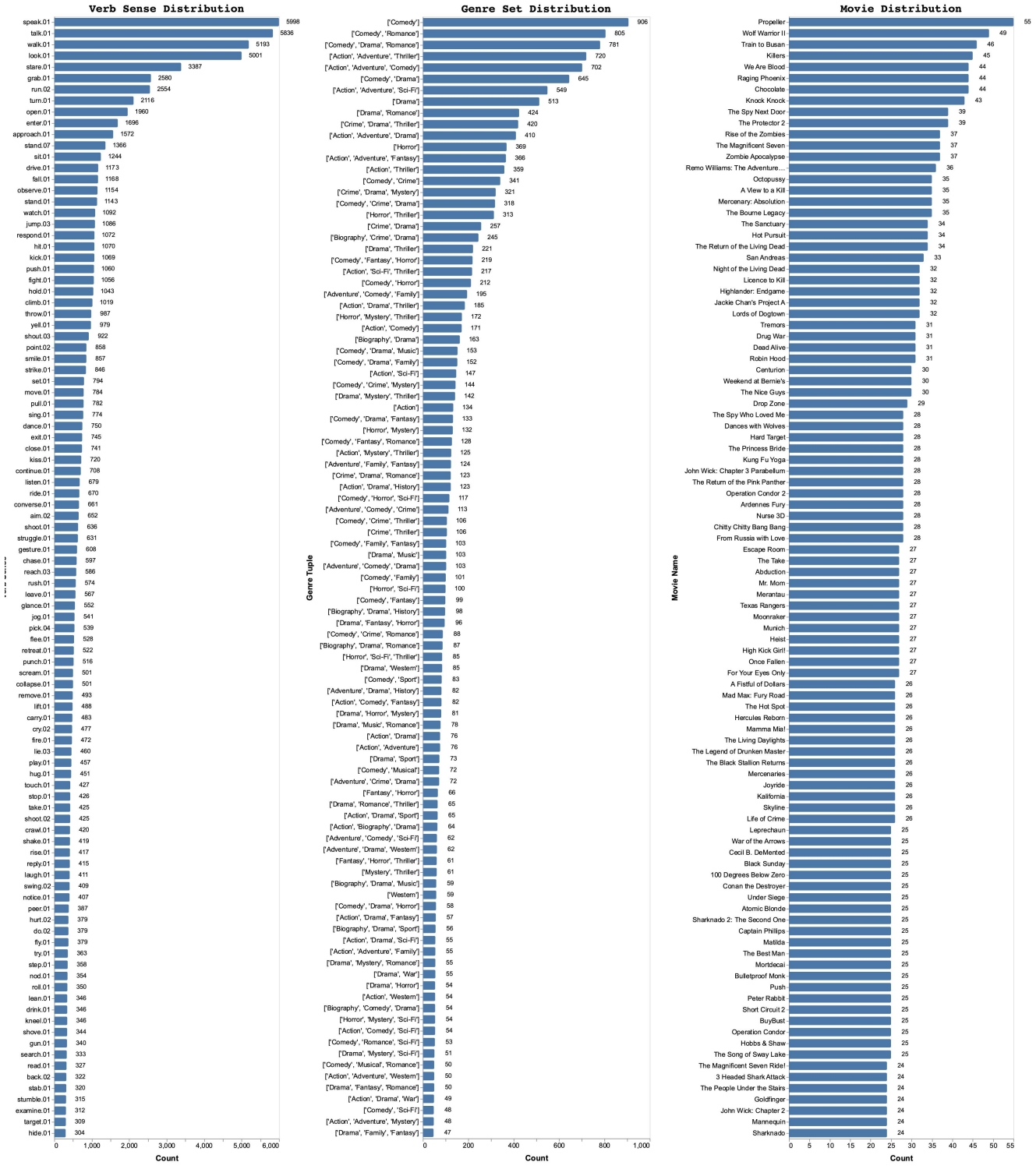
Collection Timeline. Collecting the entire training set was done over a period of about 1.2 months, and an additional 1 month for collecting the validation and test sets.

C. Additional Dataset Statistics

In this section we report additional dataset statistics not included in Section 4.2 due to space constraints.

In Table 2 we report the distributions of Event Relations before and after filtering for validation and test sets. For filtering, we use consensus of two workers *i.e*. at least two workers agree on the argument relation which we use as the ground-truth. We largely find that the consensus on Caused By and Reaction To is low, but Enabled By and No Relations are higher.

Next, we plot the distributions for the 100 most frequent verbs, genres and chosen movies in Figure 3. For verbs and genres we find Zipf's law in action. For verbs, we find most common verbs such as "talk", "speak", "walk", "look"



(a) Counts of 100 Most Frequent Verb Senses (b) Counts of 100 Most Frequent Genre Tuples (c) Counts of 100 Most Frequent Movies

Figure 3: Distribution of 100 most frequent verbs (a), genre tuples (b), and movies (c). Note that for (a), the count represents the number of events belonging to the particular verb, whereas for (b), (c) it represents the number of video segments belonging to a particular genre or movie.

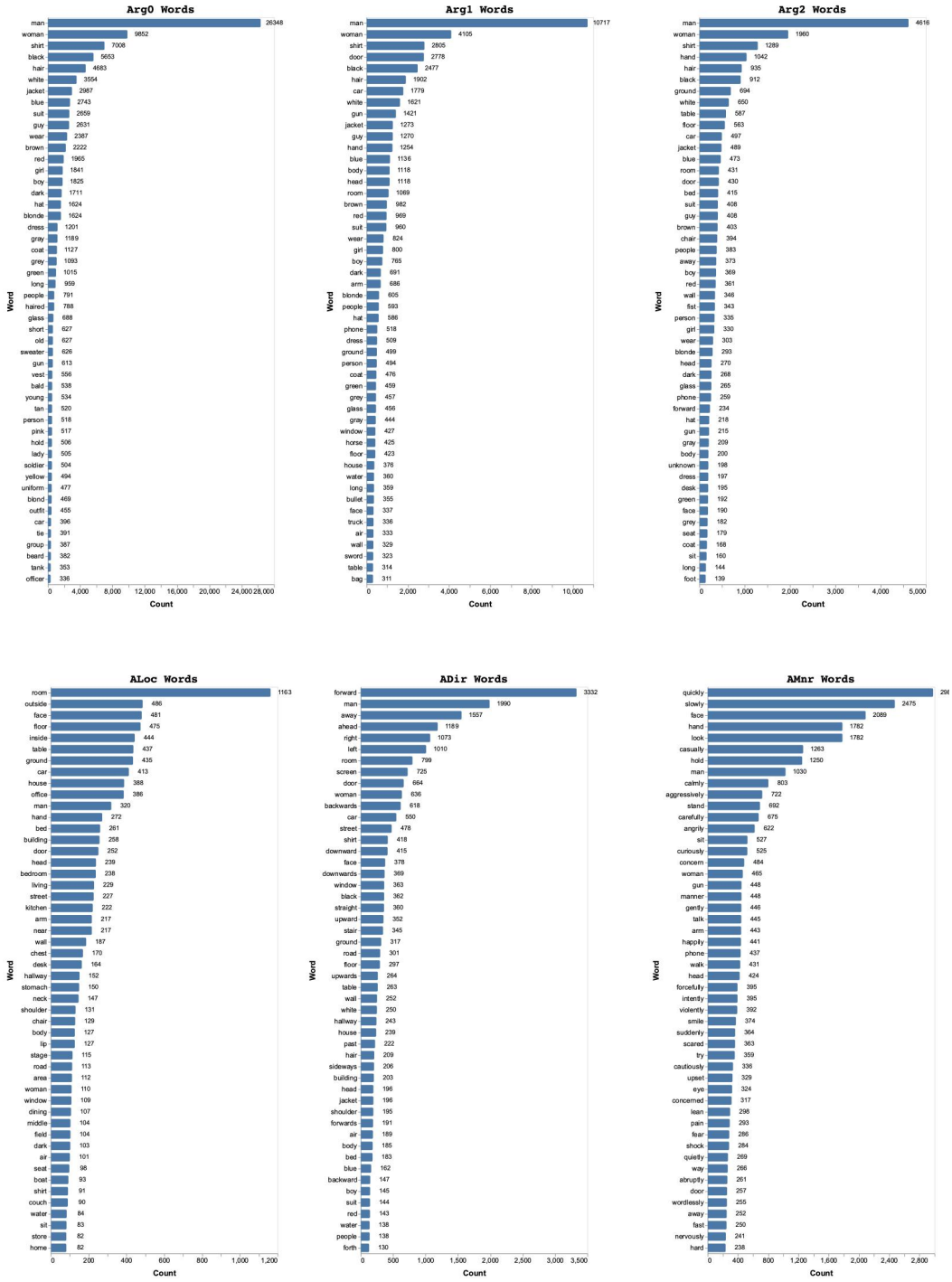


Figure 4: 50 Most frequent words (after removing stop-words) for Arg0, Arg1, Arg2, ALoc (location), ADir (direction) and AMnr(Manner).

which are also part of frequent atomic actions despite explicitly not scoring them. This is an inherent effect due to the movie domain where dialogue is a large focus. For genres we find that "Comedy", "Drama", "Action", "Romance" are the most frequent which tend to have more movements than "Mystery", "Thriller" which have less movements on actors with often extended still-frames.

In Figure 4 we plot the top 50 most frequent words

within the argument (after removing stop-words). We find "man", "woman" are the most frequent word in all of Arg0, Arg1, Arg2 which is not surprising since the movies are human-centric. We note the over-abundance of "man" compared to "woman" is an amplification of the biases present in the movie. Interestingly, the distribution is less skewed for Location, Direction, and Manner

D. Implementation Details

We detail some of the implementation details for our models. All implementations are coded in PyTorch [55]. Unless otherwise mentioned we use Adam [32] optimizer with learning rate of $1e^{-4}$.

D.1. Verb Prediction Models

All our implementations for verb prediction models such as I3D[8], Slow-only and SlowFast networks [15] is based on the excellent repository SlowFast [14]. We use the checkpoints from the repository for kinetics pre-trained models. All models are trained with a batch size of 8 for 10 epochs, and the model with best recall@5 is chosen for testing. For classification, we use a set of 1560 verbs composed two MLP projections (first projects to half the input dimension, the second to 1560 verbs) separated with a ReLU activation. For inference, we choose the top-5 scoring verbs. Training requires considerable GPU space, and on 8 TITAN GPUs, with batch size of 8 each epoch takes around 1 hour, with total being 10 hours.

D.2. Argument Prediction Models

We extract the features from underlying base networks which is 2048 and 2304 for I3D and SlowFast respectively. For transformers, we use the implementation provided in Fairseq library [52] ([4] and for GPT2 (medium) and Roberta (base) we use the implementation by HuggingFace transformer library [77] [5]. For tokenization and vocabulary, we utilize Byte-Pair Encoding and add special argument tokens such as $[Arg0]$ to encode the phrases.

For both transformer encoder and decoder we use 3 layers with 8 attention heads. The decoder uses the last encoder layer outputs as encoder attention for subsequent decoding. For training, we use cross-entropy loss over the predicted sequence. For sequence generation, we use greedy-decoding with temperature 1.0 as we didn't find improvements using beam-search or using different temperature.

For training, we used a batch size of 16 for all models other than GPT2 for which we could only use a batch size of 8 due to memory restrictions. Training time for GPT2 is around 10 hours over 8 GPUs (recall that GPT2 medium has 24 transformer layers and 16 attention heads). All other models take around 15 mins per epoch with batch size of 16 on a single TITAN GPU with total time around 3 hours for 10 epochs which we found sufficient for convergence.

For computing natural language generation metrics like ROUGE, CIDEr we use the official MSCOCO Captions implementation [44] [6]. For co-reference metrics, we use the implementation provided in coval [51] [7]

E. Evaluation Metrics

In this section, we provide details on LEA as well as our proposed LEA-soft. We further report additional metrics such BLEU [54] and METEOR [5], and coreference metrics. We also report per-argument scores for the baselines.

E.1. Co-Reference Metrics

We primarily use the metric LEA [51] which is a link-based metrics. We also note there exists other metrics such as MUC [72], BCUBE [2], CEAFE[46]. We point the reader to a seminal paper on visualizing these metrics [56] for a brief overview of MUC, BCUBE and CEAFE, and [51] for comparison of other metrics with LEA.

LEA and LEA-soft As noted in the paper [51], LEA computes an importance score and resolution score for each entity given as

$$\frac{\sum_{e_i \in E} imp(e_i) \times res(e_i)}{\sum_{e_i \in E} imp(e_i)} \tag{E.1}$$

The final score is the F1-measure computed based on recall (entities are ground-truths) and precision (entities are predictions). As noted earlier, LEA doesn't consider if the proposed entity by itself is correct and thus even incorrect entity predictions could lead high co-reference score as long as the co-referencing is correct. We address this using LEA-soft which additionally weights the importance of each entity during precision computation with the sum of cider scores in the numerator and len of cider scores in the denominator.

As a result, we have

$$Prec_{LEA} = \frac{\sum_{e_i \in E} imp(e_i) \times res(e_i)}{\sum_{e_i \in E} imp(e_i)} \tag{E.2}$$

$$Prec_{LEA-soft} = \frac{\sum_{e_i \in E} (\sum_{e_i} C(e_i)) \times imp(e_i) \times res(e_i)}{\sum_{e_i \in E} |e_i| \times imp(e_i)} \tag{E.3}$$

where $C(e_i)$ denotes the cider score for the i^{th} entity. We keep the recall computation unchanged and use the modified precision to compute the final F1-Score for LEA-soft. Since we have multiple ground-truth reference, we compute the F1-score for each ground-truth reference individually and average over the 3 ground-truths.

[4] https://github.com/pytorch/fairseq/
[5] https://github.com/huggingface/transformers
[6] https://github.com/tylin/coco-caption
[7] https://github.com/ns-moosavi/coval

	cider	Arg0	Arg1	Arg2	ALoc	AScn	ADir	AMnr
GPT2	0.39	0.40	0.39	0.45	0.43	0.22	0.37	0.15
Human	0.70	0.73	0.74	0.73	0.90	0.96	0.40	0.15

Table 3

E.2. Evaluation of Arguments

We examine the cider scores for different arguments over a set of 100 videos (same used for verb prediction results). To compare semantic role values, which are free-form text phrases, we compute CIDEr metric treating one of the chosen annotations as a hypothesis and remaining annotations as references for each argument. Table 3 compares CIDEr scores for all semantic roles and scores by argument type for a GPT2 based language only baseline that generates the sequence of roles and values given the verb for an event. We find that human-agreement is high for all arguments except direction (ADir) and manner (AMnr). For both "direction" (ADir) and "manner" (AMnr), we find that both language-only baseline and human agreements are poor. On further inspection, we find that the argument "manner" describes "how" the event took place is open to subjective interpretation, and the argument "direction" has a wide range of correct values (e.g. for "walk" directions "forward", "down the path", and "through the trees") may all be correct. For a reliable evaluation, we evaluate argument prediction performance only on arguments that achieved high human-agreement *i.e.* Arg0, Arg1, Arg2, ALoc, and AScn, and leave the evaluation of Direction and Manner for future work.

E.3. All Metrics

We report BLEU@1, BLUE@2, METEOR, ROUGE, and CIDEr for both val (Table 4) and test set (Table 5). For each metric we further report macro-averaged scores across verbs and arguments, and report per argument scores. Note that only CIDEr is able to take advantage of the macro-averaged scores due to its inverse document frequency re-weighting. Finally, we report the co-reference metrics MUC, BCUBE, CEAFE , LEA and our proposed metric LEA-Soft.

F. VidSitu DataSheet

The seminal work datasheets for datasets [17] outlines a list of questions to encourage transparency, accountability and mitigate unwanted biases. Here, we provide a datasheet for VidSitu closely following the guidelines in prior work. For simplicity and readability, we paste the questions verbatim, and omit certain questions due to double-blind anonymity.

F.1. Motivation

- **For what purpose was the dataset created?** The main motivation to create the dataset is to bridge the research gap between learning atomic actions and generating holistic captions. In particular, the dataset opens path for the task of Visual Semantic Role Labeling in Videos which in addition to action-recognition, emphasizes how various objects interact within an action, how various objects interact over time-period across multiple actions, co-referencing of these objects over time, and how various actions affect each other.
- **Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?** We omit this question due to anonymity reasons.
- **Who funded the creation of the dataset?** We omit this question due to anonymity reasons.

F.2. Composition

- **What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?** Each instance consists of a 10-second video obtained from a movie-clip available on YouTube. These are usually human-centric and hence primarily contain videos of people interacting in diverse and complex situations.
- **How many instances are there in total (of each type, if appropriate)?** In total there are $27.4K$ instances distributed across training ($23.62K$), validation ($1.80K$) and testing ($1.98K$)
- **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?** This question doesn't pertain to our dataset.
- **What data does each instance consist of?** Each instance is a 10-second video (mp4 video) available from YouTube.
- **Is there a label or target associated with each instance?** Each instance (10 second video) is annotated at 2-second intervals with a verb describing the event, corresponding argument roles for the verb co-referenced across the video, and event relations across the various verbs with respect to the middle event (Event 3 spanning from 4-6 seconds).
- **Is any information missing from individual instances?** No, every instance has the same annotations.
- **Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?** We provide information about which instances are derived from the same $2-3$ minutes

Model Vis Feats	GPT2 ✗	TxDec ✗	Vid TxDec SlowFast	Vid TxEncDec SlowFast	Vid TxDec I3D	Vid TxEncDec I3D	Human
B@1	40.09	43.51	42.13	41.54	42.57	44.84	43.38
B@1-Vb	38.68	39.02	37.39	37.7	38.18	40.51	40.15
B@1-Arg	40.06	43.1	40.83	40.97	41.71	44.1	41.32
B@1-Arg0	43.47	50.6	50.04	48.36	47.08	49.62	48.71
B@1-Arg1	31.23	30.64	32.37	34.59	31.17	34.5	41.65
B@1-Arg2	32.44	38.02	37.3	34.38	36.66	36.72	38.54
B@1-ALoc	45.78	48.36	42.04	47.75	44.64	49.48	36.67
B@1-AScn	47.37	47.85	42.41	39.75	48.98	50.19	41
B@2	26.8	30.4	29.2	27.85	29.16	30.82	29.86
B@2-Vb	24.25	25.26	23.9	23.97	24.45	26.11	25.69
B@2-Arg	27	29.8	27.81	27.41	28.63	30.18	28.21
B@2-Arg0	29.54	36.3	35.87	33.22	32.6	34.7	33.93
B@2-Arg1	18.71	19.94	21.69	21.9	19.78	21.83	27.92
B@2-Arg2	20.29	25.96	24.95	21.38	23.67	23.17	26.15
B@2-ALoc	32.93	32.8	27.86	33.88	32.16	35.74	25.06
B@2-AScn	33.56	33.99	28.66	26.67	34.94	35.45	27.99
M	16.75	15.77	16.63	17.57	17.2	17.96	21.99
M-Vb	15.85	14.81	15.28	15.86	15.29	16.49	22.66
M-Arg	15.29	14.64	15.21	16.25	15.58	16.81	20.57
M-Arg0	21.22	19.9	20.74	21.45	21.51	21.58	24.86
M-Arg1	15	14.07	15.28	16.35	14.9	15.57	22.53
M-Arg2	14.18	13.8	13.61	15.11	14.1	14.89	19.32
M-ALoc	13.2	12.98	12.68	13.94	12.6	15.6	16.6
M-AScn	12.85	12.46	13.73	14.38	14.8	16.39	19.54
R	39.59	38.68	38.66	40.49	39.55	42.05	39.53
R-Vb	37.2	35.87	35.11	35.72	35.24	37.27	38.89
R-Arg	38.96	37.54	36.7	39.27	37.94	41.29	37.81
R-Arg0	43.88	46.1	47.31	46.83	45.97	47.49	45.28
R-Arg1	33.95	31.28	33.03	35.38	32.73	35.03	40.66
R-Arg2	32.38	30.19	29.67	32.92	31.37	34.07	34.94
R-ALoc	41.76	39.06	33.77	39.27	35.99	44.23	32.09
R-AScn	42.82	41.05	39.75	41.95	43.66	45.61	36.08
C	32.76	34.28	42.11	43.2	40.24	46.89	84.78
C-Vb	45.16	43.54	50.19	50.03	46.07	52.89	92.61
C-Arg	30.34	29.04	36.83	38.14	36.21	42.38	79.14
C-Arg0	26.88	32.2	38.46	34.39	34.93	37.69	85.71
C-Arg1	36.17	37.77	41.7	45.29	39.83	45.6	85.62
C-Arg2	30.31	30.51	32.81	35.62	33.51	41.13	71.8
C-ALoc	34.85	25.41	34.75	36.51	36.37	43.15	71.06
C-AScn	23.47	19.32	36.45	38.91	36.44	44.35	81.54
MUC	60.9	54.07	42.69	63.96	45.89	61.44	79.32
BCUBE	75.05	67.72	67.31	72.33	67.83	73.26	85.8
CEAFE	62.89	55.62	54.33	57.61	56.79	60.46	77.07
Lea	50.2	40.22	36.84	34.81	48.45	48.99	70.49
Lea-Soft	26.56	21.24	25.9	24.79	32.97	29.52	69.5

Table 4: Semantic Role Prediction on Validation Set. B@1: Bleu-1, B@2: Bleu-2, M: METEOR, R: ROUGE-L, C: CIDEr, Metric-Vb: Macro Averaged over Verbs, Metric-Arg: Macro Averaged over arguments, Metric-Argi: Metric computed only for the particular argument.

Model Vis Feats	GPT2 ✗	TxDec ✗	Vid TxDec SlowFast	Vid TxEncDec SlowFast	Vid TxDec I3D	Vid TxEncDec I3D	Human
B@1	42.78	44.79	43.2	42.31	44.77	45.74	43.62
B@1-Vb	39.64	40.07	38.97	38.08	40.74	39.83	39.8
B@1-Arg	42.12	43.99	41.76	41.29	43.44	44.83	41.34
B@1-Arg0	47.5	52.3	51.38	50.81	50.63	51.3	49.79
B@1-Arg1	34.33	33.06	34.69	37.05	34.33	35.98	41.56
B@1-Arg2	35.79	40.21	39.07	36.55	39.58	37.39	40.07
B@1-ALoc	46.01	46.93	42.22	45.51	44.88	50.43	36.37
B@1-AScn	46.99	47.44	41.43	36.56	47.75	49.06	38.88
B@2	29.34	31.2	29.89	28.21	31.08	31.32	29.15
B@2-Vb	25.22	26.27	25.13	23.95	26.61	26.05	23.96
B@2-Arg	28.76	30.12	28.34	27.25	29.92	30.48	27.34
B@2-Arg0	33.51	37.71	37.02	35.09	35.94	35.9	34.03
B@2-Arg1	21.21	21.57	23.12	23.57	22.02	22.78	26.83
B@2-Arg2	23.83	27.97	26.8	23.21	26.73	24.48	26.44
B@2-ALoc	32.23	30.33	27.44	30.6	31.19	35.3	24.93
B@2-AScn	33.02	33.01	27.31	23.77	33.72	33.91	24.47
M	17.85	16.34	17.03	18.03	17.89	18.29	21.93
M-Vb	15.93	15.12	15.8	15.86	16.22	16.28	21.67
M-Arg	16.38	15.19	15.67	16.61	16.13	17.23	20.49
M-Arg0	22.2	20.26	20.88	22.03	22.3	21.75	25.26
M-Arg1	16.25	14.72	16.08	16.93	15.9	15.82	22.12
M-Arg2	15.8	14.82	14.66	16.22	15.04	15.33	20.11
M-ALoc	14.19	13.2	13.16	13.71	12.69	16.67	16.73
M-AScn	13.46	12.95	13.59	14.14	14.74	16.56	18.24
R	41.6	40.02	40.21	41.8	41.56	43.3	40.46
R-Vb	37.74	37.01	36.66	36.78	37.74	37.29	39.44
R-Arg	40.84	38.51	38.28	40.23	39.7	42.51	38.84
R-Arg0	46.22	47.3	48.33	48.67	48.32	48.65	46.82
R-Arg1	37.35	34.4	36.15	38.36	37.01	37.07	41.51
R-Arg2	35.27	32.49	32.41	35.35	33.96	34.97	37.25
R-ALoc	42.37	37.09	34.65	37.41	35.88	45.72	33.03
R-AScn	42.97	41.25	39.87	41.34	43.33	46.15	35.59
C	38.18	37.2	45.08	46.34	44.53	48.95	83.87
C-Vb	43.99	44.36	51.9	49.69	50.64	52.65	89.13
C-Arg	37.77	32.93	40.9	42.22	41.22	46.16	78.85
C-Arg0	29	32.24	40.74	36.65	39.53	37.92	86.6
C-Arg1	42.3	41.24	45.95	50.46	46.73	48.93	85.23
C-Arg2	39.06	35.01	37.76	41.23	38.83	43.78	73.26
C-ALoc	49.82	32.16	43.12	43.14	43.92	55.64	75.27
C-AScn	28.64	23.99	36.93	39.62	37.08	44.54	73.87
MUC	64.9	58.28	46.07	66.7	48.71	64.72	81.49
BCUBE	76	68.71	67.31	73.15	67.75	73.98	86.41
CEAFE	63.72	56.5	53.91	58.16	56.45	61.27	78.09
Lea	53.28	43.62	36.22	51.15	38.18	50.95	72.64
Lea-Soft	33.5	24.48	27.74	32.59	29.26	35.67	70.93

Table 5: Semantic Role Prediction on Test Set. B@1: Bleu-1, B@2: Bleu-2, M: METEOR, R: ROUGE-L, C: CIDEr, Metric-Vb: Macro Averaged over Verbs, Metric-Arg: Macro Averaged over arguments, Metric-Argi: Metric computed only for the particular argument.

YouTube video as well as the underlying movie (this information is obtained from Condensed-Movies [3] dataset). However, this information is not used for any of the task in the dataset except for splitting the videos in train, validation and test sets.

- **Are there recommended data splits (e.g., training, development/validation, testing)?** Yes, we provide training, validation and test sets by splitting the overall set in $80:10:10$ ratio randomly based on the movie names. We also ensure (qualitatively) that the normalized distributions of verbs, and genres are same across the splits.
- **Are there any errors, sources of noise, or redundancies in the dataset?** The main sources of errors would be the annotations themselves, however, we have made extended efforts from automatic to manual checks to remove such errors and provided constant feedback. Some redundancy may occur due to oversampling of dialogues in movies which are described with the verb "talk". Some redundancy may also occur due to use of closely related verbs such as "run" and "jog".
- **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?** Yes, the dataset provides links to YouTube videos. Since the videos are provided by a licensed channel, we expect the videos to have high online longevity.
- **Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor patient confidentiality, data that includes the content of individuals' non-public communications)?** No, our dataset is derived from movies publicly available on youtube.
- **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?** Some of the videos obtained from action, crime or horror movies may be sensitive to some viewers when viewed directly. Some videos may also contain violence and gore, and we suggest user discretion in viewing the videos.

F.3. Collection Process

- **How was the data associated with each instance acquired?** The data was directly observable in the form of embedded youtube videos.
- **What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?** We used Amazon Mechanical Turk to collect the data with a custom annotation interface. We validated them by small scale user study and taking feedbacks during worker qualification.
- **If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?** We sampled videos which had more verbs within their duration.
- **Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?** Crowd-Workers were involved in the process. They were paid $\$0.75$ for training videos and $\$0.2$ for verb annotation and $\$0.7$ for argument and event relation for videos in validation and test splits. On average it is around $\$9 - \$12$ per hour above the minimum wage. On popular websites, our pay was noted to be generous.
- **Over what timeframe was the data collected?** The data was collected over 2.2 months with initial 1.2 months for training set and rest for validation and testing.
- **Were any ethical review processes conducted (e.g., by an institutional review board)?** No, there was no ethical review process.

F.4. Preprocessing/cleaning/labeling

- **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?** Only, exact string match was performed to obtain co-referenced entities. We used spacy [26] to compute dataset statistics such as noun-diversity but it is not used over the collected data for down-stream tasks.
- **Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?** In our case, raw data is same as cleaned data.

F.5. Uses

- **Has the dataset been used for any tasks already?** We have used the data to show its usefulness for our proposed task Visual Semantic Role Labeling in Videos
- **Is there a repository that links to any or all papers or systems that use the dataset?** We omit this question due to anonymity reasons.
- **What (other) tasks could the dataset be used for?** We believe the dataset could be repurposed for many down-stream video understanding tasks such as video retrieval, video question answering, action forecasting, long-term reasoning.

- **Are there tasks for which the dataset should not be used?** The data is obtained from movies and exhibits certain stereotypes which donot hold true in real world. It also contains highly unlikely action sequences (such as a "man flying"), and thus it shouldn't be used for real-world cases and strictly used as a video understanding benchmark.

F.6. Distribution

- **Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?** The dataset would be made publicly available. We omit details due to anonymity.
- **How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?** It would be distributed on a website and github.

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, 15

[3] Max Bain, Arsha Nagrani, Andrew Brown, and Andrew Zisserman. Condensed movies: Story based retrieval with contextual embeddings, 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. 15

[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, 5, 15

[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] D. Damen, H. Doughty, G. Farinella, S. Fidler, Antonino Furnari, Evangelos Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. *ArXiv*, abs/1804.02748, 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] Haoqi Fan, Yanghao Li, Bo Xiong, Wan-Yen Lo, and Christoph Feichtenhofer. Pyslowfast. https://github.com/facebookresearch/slowfast, 2020. 15

[15] 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, 5, 8, 10, 15

[16] 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

[17] 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

[18] 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

[19] 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

[20] 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

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

[22] 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

[23] 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, 4, 9

[24] 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

[25] 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

[26] Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017. 19

[27] Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. Movienet: A holistic dataset for movie understanding. In *The European Conference on Computer Vision (ECCV)*, 2020. 9

[28] H. Idrees, A. Zamir, Yu-Gang Jiang, Alex Gorban, I. Laptev, R. Sukthankar, and M. Shah. The thumos challenge on action recognition for videos "in the wild". *ArXiv*, abs/1604.06182, 2017. 3

[29] 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

[30] 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, 4, 5, 9

[31] 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

[32] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015. 15

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

[34] 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. 1, 2, 5

[35] 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. 2

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

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

[38] 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

[39] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+language omni-representation pre-training. *ArXiv*, abs/2005.00200, 2020. 2

[40] 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

[41] 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

[42] 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

[43] 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

[44] Tsung-Yi Lin, M. Maire, Serge J. Belongie, James Hays, P. Perona, D. Ramanan, Piotr Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. *ArXiv*, abs/1405.0312, 2014. 10, 15

[45] 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, 8

[46] 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, 15

[47] Louis Mahon, Eleonora Giunchiglia, B. Li, and Thomas Lukasiewicz. Knowledge graph extraction from videos. *ArXiv*, abs/2007.10040, 2020. 2

[48] 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. 2

[49] 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, 4, 9

[50] 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

[51] 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, 15

[52] 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. 15

[53] 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

[54] Kishore Papineni, S. Roukos, T. Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002. 15

[55] 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. 15

[56] 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. 15

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

[58] 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, 4

[59] 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

[60] 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

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

[62] 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

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

[64] 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. 2

[65] Emma Strubell, Pat Verga, Daniel Andor, D. Weiss, and A. McCallum. Linguistically-informed self-attention for semantic role labeling. *ArXiv*, abs/1804.08199, 2018. 9

[66] C. Sun, Abhinav Shrivastava, Carl Vondrick, Kevin Murphy, R. Sukthankar, and C. Schmid. Actor-centric relation network. *ArXiv*, abs/1807.10982, 2018. 2

[67] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. *ArXiv*, abs/1409.3215, 2014. 5

[68] 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, 4, 9

[69] 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

[70] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, L. Kaiser, and Illia Polosukhin. Attention is all you need. *ArXiv*, abs/1706.03762, 2017. 5, 8

[71] 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

[72] Marc B. Vilain, J. Burger, J. Aberdeen, D. Connolly, and L. Hirschman. A model-theoretic coreference scoring scheme. In *MUC*, 1995. 7, 15

[73] 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

[74] 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

[75] 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. 5

[76] 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. 1, 2, 3, 5

[77] 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. 15

[78] 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

[79] 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

[80] 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

[81] 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. 1, 3, 5

[82] 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

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

[84] 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

[85] 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

[86] 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

[87] 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

[88] 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

[89] Luowei Zhou, Nathan Louis, and Jason J. Corso. Weakly-supervised video object grounding from text by loss weighting and object interaction. *ArXiv*, abs/1805.02834, 2018. 3, 4, 9

[90] 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

[91] 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. 2