This CVPR workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

In Defense of Structural Symbolic Representation for Video Event-Relation Prediction

Andrew Lu[*], Xudong Lin[*], Yulei Niu, Shih-Fu Chang
Columbia University
{ayl2148,xudong.lin,yn2338,sc250}@columbia.edu

Abstract

Understanding event relationships in videos requires a model to understand the underlying structures of events (i.e. the event type, the associated argument roles, and corresponding entities) and factual knowledge for reasoning. Structural symbolic representation (SSR) based methods directly take event types and associated argument roles/entities as inputs to perform reasoning. However, the state-of-the-art video event-relation prediction system shows the necessity of using continuous feature vectors from input videos; existing methods based solely on SSR inputs fail completely, even when given oracle event types and argument roles. In this paper, we conduct an extensive empirical analysis to answer the following questions: 1) why SSR-based method failed; 2) how to understand the evaluation setting of video event relation prediction properly; 3) how to uncover the potential of SSR-based methods. We first identify suboptimal training settings as causing the failure of previous SSR-based video event prediction models. Then through qualitative and quantitative analysis, we show how evaluation that takes only video as inputs is currently unfeasible, as well as the reliance on oracle event information to obtain an accurate evaluation. Based on these findings, we propose to further contextualize the SSR-based model to an Event-Sequence Model and equip it with more factual knowledge through a simple yet effective way of reformulating external visual commonsense knowledge bases into an event-relation prediction pretraining dataset. The resultant new state-of-the-art model eventually establishes a ***25%*** *Macro-accuracy performance boost.*

1. Introduction

Event understanding has been thoroughly explored in the past decade [1, 6, 23, 27, 28, 31, 34, 38, 40, 46, 48, 49]. An event is typically represented as a verb (indicating the event type) and several arguments, each of which has a role name and associated entity. Researchers have been devoted to extracting events and labeling the argument roles [6, 23, 34, 38, 41, 42, 48] in vision, language, and multimodal domains. These event extraction and argument role labeling models build the foundation for higher level understanding of event-relations. Relations between events [2, 3, 10, 32, 39, 47] have been thoroughly studied in the language domain. A specific relation and event coreference have also been explored in the multimodal setting [6, 23]. However, video event-relation prediction is still a new and challenging task [38] requiring both good representations of video events and commonsense knowledge to reason between events.

Structural symbolic representation (SSR) [14, 16, 23, 25, 26, 43, 45, 50] has been widely adopted on various downstream tasks such as visual question answering [50], image captioning [45], and action recognition [16] for its high interpretability and generalization ability [25, 26, 43, 50]. In this work, we refer to SSR as representations consisting of discrete tokens organized with certain structures. When applied to event-relation prediction, it is usually formulated as an event type (verb) with associated argument role names and corresponding entities. This event-argument SSR has been very effective in predicting event relations in the language domain [8]. However, SSR-based models for video events are easily biased by the dominant relations in the VidSitu dataset [38] and predict the dominant class for all classes, which contradicts the success of existing SSR-based methods on various tasks including event-relation prediction using text.

To answer why SSR-based models fail on video event-relation prediction, we first analyze if there are any possible patterns that the model could have leveraged to avoid being misled (Sec. 4.2). We discover that even if the model only memorizes the dominant relation for pairs of event types, the model is clearly not supposed to predict only *Enables* (the most frequent relation in the VidSitu dataset [38]). Based on this finding, we bootstrap the failed SSR-based model and the state-of-the-art model [38] with two simple processes: utilizing a balanced training data/objective and

[*] Equal contribution.

tuning hyper-parameters. Through these optimized training settings, the text-only SSR-based model outperforms the multimodal state-of-the-art model by 20%.

The different behaviors motivate us to carefully analyze the evaluation challenges of video event-relation prediction (Sec 4.3). We evaluate two state-of-the-art video-language models, HERO [22] and ClipBERT [19] on the VidSitu dataset [38]. By controlling inputs with the help of strong image-text contrastive models [35], we found that including oracle information is essential due to the presence of multiple events co-occurring simultaneously in the same video. Even with strong pretrained video-language models, the event-type and argument role descriptions are still more important than video feature vectors.

Based on these observations, we propose to contextualize the simple pairwise SSR-based models to an Event-Sequence model in order to leverage context information within sequences of events for more accurate event-relation prediction. Furthermore, we explore leveraging an external visual knowledge base, VisualCOMET [33], to teach the model commonsense knowledge about the evolution process of events. We propose an effective and straightforward strategy for reformulating the annotations of VisualCOMET into event sequences suitable for event-relation prediction. The contextualized Event-Sequence model with pretraining on VisualCOMET outperforms the state-of-the-art model [38] by a **25%** improvement in Macro-accuracy.

Our contributions are summarized as follows:

- We identify why SSR-based models fail on video event-relation prediction and improve Macro-accuracy performance by 20% through optimizing training settings.
- We identify the proper settings needed to evaluate video event-relation prediction models through extensive quantitative and qualitative analysis of different model variations on the VidSitu dataset.
- We propose a contextualized Event-Sequence model, coupled with a pretraining technique on VisualCOMET, to fully utilize the rich contextual information in event sequences and commonsense knowledge from the existing knowledge base. Our model significantly improves event-relation prediction accuracy compared to the state-of-the-art model from 34.2% → 59.2%.

2. Related Work

Visual Event Understanding aims to recognize, extract, and structure the actions or activities happening in images or videos. Previous studies simply represent visual events as verbs or subject-verb-object triplets [5, 7, 11, 17, 24, 38]. Recent research studies structural and semantic representations of visual events from image-text and video-text pairs such as M2E2 [23] and VideoM2E2 [6]. Importantly, the visual situation recognition task aims to identify not only the activity in an image [34, 48] or video [38], but also the entities including persons and objects (*i.e.*, semantic roles) associated with the activity. Another benchmark, VisualCOMET [33], proposes to depict person-centric images as a graph of commonsense descriptions, including before-event, intent of people, and next-event. Our work focuses on the video event-relation prediction task and investigates the roles of event type, argument roles, and video features in relation prediction. We also explore using VisualCOMET as an external visual knowledge base for pretraining.

Structural Symbolic Representation (SSR) denotes the representation consisting of discrete tokens with certain structures. SSR has been applied in various tasks in computer vision and natural language processing (NLP). For instance, the visual scene in visual question answering can be modeled as the structural representations of objects with their associated attributes and locations [50]. For event representation, spatio-temporal scene graphs [16, 37] decompose each event as a temporal sequence of spatial scene graphs. Recent NLP studies show the potential of SSR in event-relation prediction in text, *e.g.* part-of-speech (POS) and XML tags [8]. In this work, we follow VidSitu [38] to represent each video event as event type/verb and its associated argument roles and entities.

Event-relation Prediction in both Text and Video. Events in texts are often ordered by temporal relation [4], causal relation [30], or narrative order [15]. Hong *et al.* [12] define event-relations in text as 5 main types, along with 21 sub-types, covering inheritance, expansion, contingency, comparison, and temporality. As the only available video event-relation dataset in cross-document event-relations, VidSitu [38] defines four types of relations: no relation, causality, enabled, and reaction-to. Here, we follow the same definition for video event-relation prediction.

3. Technical Approach

3.1. Preliminaries

Structural Symbolic Representation (SSR). SSR generally refers to a representation where elements are discrete tokens and have certain structures, *e.g.* scene graphs [16]. To effectively represent an event, the event type (usually a verb), its associated argument roles, and the actual entities for each argument role are required. Therefore, we consider the sequence of text tokens with the following structure as the SSR of an event x: $x = \{v, a_1, e_1, a_2, e_2, ..., a_M, e_M\}$, where $v \in \mathbb{W}$ is the event type/verb, $a_m \in \mathbb{W}, e_m \in \mathbb{W}$ are the m^{th} argument role and associated entity, and M is the number of argument roles for this event. Such a sequence is essentially a traverse of the graph with v as the root node and a_m, e_m as the m^{th} edge and leaf node.

Event-Relation Prediction. Currently VidSitu [38] is the only dataset available for video event-relation pre-

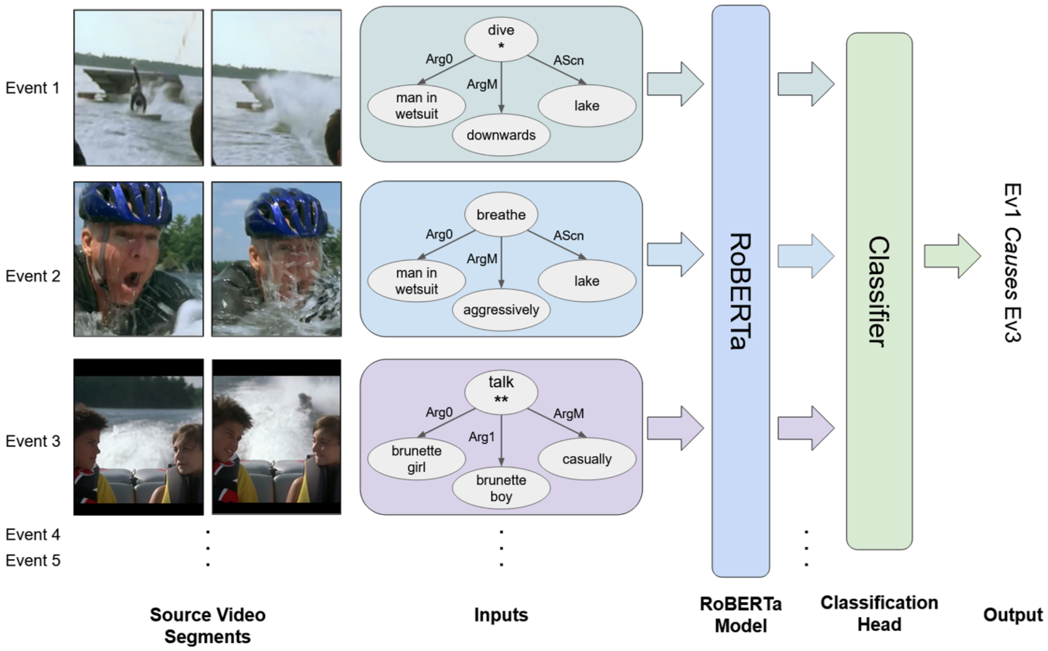


Figure 1. The pipeline of the Event-Sequence model for event-relation prediction. Special characters * and ** denote the target events the model is required to predict relation between. Detailed description is in Sec. 3.2

diction, and thus we adopt its specific settings. Each video event sequence consists of five consecutive events $\{x_1, x_2, x_3, x_4, x_5\}$, each of which is from a two-second video segment $y_i \in \mathbb{R}^{H\times W\times 3\times F}$. H, W, F are the height, width, and number of frames of the video segment. In VidSitu, only the relationship between the center event x_3 and other events $x_i, i \neq 3$ are annotated.

Baseline Model. In order to predict the relationship between two events x_i and x_j, we follow the RoBERTa variant [38] while constructing the baseline model. Based on SSR, a model $\mathcal{F} : \mathbb{W}^L \longrightarrow \mathbb{R}^C$ takes a sequence of text tokens as input and predicts a distribution over the C possible relationship classes, where L is the length of the sequence consisting of symbolic text representations of x_i and x_3. The model $\mathcal{F}$ is initialized with pretrained weights from RoBERTa [29].

Baseline + Video Features. The state-of-the-art model [38] claims that video features are more effective than directly using symbolic representations. It takes both video features and text tokens as inputs $\mathcal{G} : \mathbb{W}^L \times \mathbb{R}^{D\times F} \longrightarrow \mathbb{R}^C$ to predict the distribution over the C classes. An off-the-shelf video feature extractor H is used to extract continuous feature vectors from the video segment y_i when the event happens. The feature vector is concatenated with the output embedding from the text tokens before being fed into the final classifier $\mathcal{G}$. We denote it as *Baseline + Video Features* in the following discussion. When not specified, the video feature extractor is SlowFast [9, 38].

3.2. Contextualized Event-Sequence Model

Rich contextual information (e.g. neighboring events, location of the events, manner of the events, etc.) plays an important role in understanding the relationship between events. We extensively analyze event sequences in the VidSitu dataset and indeed find patterns, as presented in Sec. 4.2. We propose a novel contextualized Event-Sequence model for event-relation prediction and explore various ways of utilizing context.

Event-Sequence Model. Instead of only feeding the model with two events between which to predict event relations, we propose to exploit the rich contextual information in a full sequence of all the five events (shown in Figure 1),

$$p = \mathcal{F}(x_1, x_2, x_3, x_4, x_5), \tag{1}$$

where $p \in \mathbb{R}^C$ is the predicted distribution over the C classes of relations. Note that to inform the model between which two events we want to predict the relation, an extra special token "*" is added before each of them.

Event-Sequence Model + Video Features. Video features, the continuous feature vectors obtained from pretrained video feature extractors, may convey fine-grained information about the actual visual scene. We follow [38] to integrate video features as in,

$$p = \mathcal{G}(x_1, x_2, x_3, x_4, x_5, \mathcal{H}(y_i), \mathcal{H}(y_3)), \tag{2}$$

where the video features are fused with contextualized embeddings before the final classification layer.

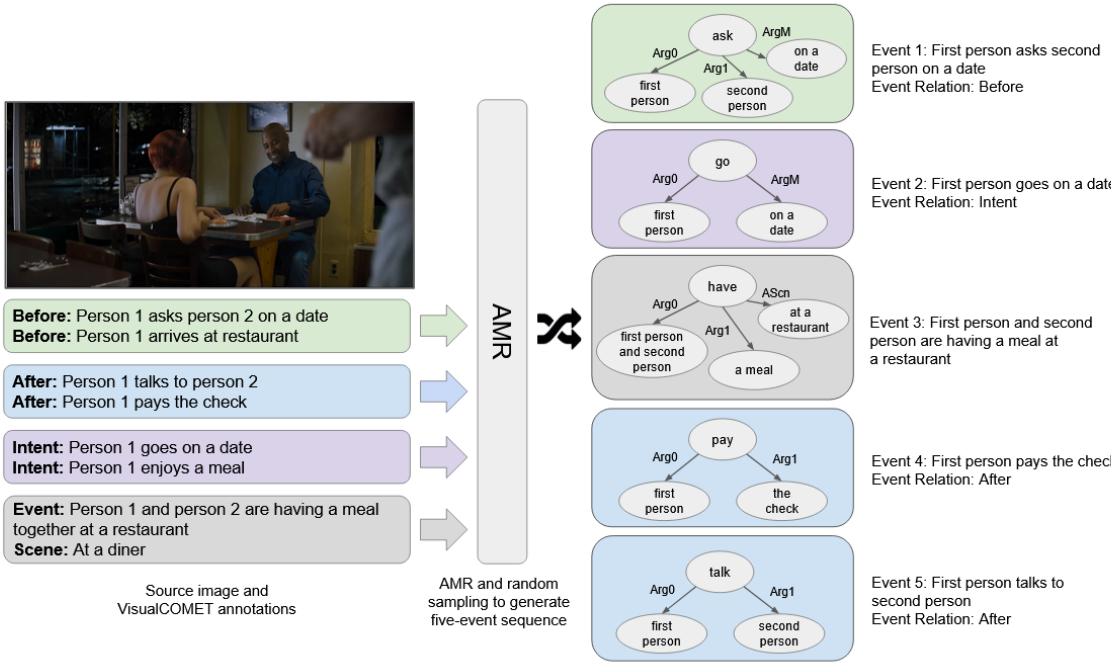

Before: Person 1 asks person 2 on a date
Before: Person 1 arrives at restaurant
After: Person 1 talks to person 2
After: Person 1 pays the check
Intent: Person 1 goes on a date
Intent: Person 1 enjoys a meal
Event: Person 1 and person 2 are having a meal together at a restaurant
Scene: At a diner
Source image and VisualCOMET annotations
AMR
AMR and random sampling to generate five-event sequence
ask
Arg0
ArgM
Arg1
first person
second person
on a date
Event 1: First person asks second person on a date
Event Relation: Before
go
Arg0
ArgM
first person
on a date
Event 2: First person goes on a date
Event Relation: Intent
have
Arg0
 AScn
Arg1
first person and second person
at a restaurant
a meal
Event 3: First person and second person are having a meal at a restaurant
pay
Arg0
Arg1
first person
the check
Event 4: First person pays the chec
Event Relation: After
talk
Arg0
Arg1
first person
second person
Event 5: First person talks to second person
Event Relation: After

Figure 2. VisualCOMET annotation conversion pipeline. We parse and restructure the full sentence annotations in VisualCOMET using AMRLib [13] into SSRs to match VidSitu annotations for pretraining. Entity co-references are converted from numeric labels to free-form text (e.g. Person 1 → First person) to more closely resemble VidSitu.

Sequence-to-Sequence Model. Inspired by recent advances of sequence-to-sequence modeling [21, 25, 36] on both language and video domains, we explore a variant of directly generating the sequence of relationships given the sequence of events as input,

$$p_{1,3}, p_{2,3}, p_{4,3}, p_{5,3} = \mathcal{S}(x_1, x_2, x_3, x_4, x_5), \quad (3)$$

where $p_{i,3}$ is the predicted distribution for the relationship between the i^{th} event and the middle event x_3. The common strategy of teacher-forcing [44] is employed to handle the conditional generation problem, which means during training

$$p_{k,3} = \mathcal{S}(x_1, x_2, x_3, x_4, x_5, l_1, ..., l_{k-1}), \quad (4)$$

the ground-truth (GT) "historical" event-relations $l_1, ..., l_{k-1}$ are used for predicting the next relation. During testing, we apply beam search to decode the actual event relation sentence. This variant does not directly use additional contextual information. Instead, it is leveraging conditional generation as a constraint to prevent the model from only predicting the dominant class as reported in [38].

Auxiliary Arguments. The model in [38] only uses base arguments (arguments tied to the verb through direct semantic relations) such the agent and the target. However, additional contextual information could be clearly provided by extra argument roles like AMnr, ADir, and AScn (manner, scene and direction). We append them after the base arguments as additional inputs to the model.

3.3. Training

The standard training objective states to use cross-entropy loss to train the model for event-relation prediction,

$$\min_{\theta} \text{-} \log p_l, \quad (5)$$

where θ is the parameter to be updated in the model, l is the GT relation, and p is the predicted relation type.

Balancing Data/Loss. In [38], the poor performance of the SSR model is attributed to the possible imbalanced relation class distribution, which leads the model to only predict the dominant relation type. To tackle this issue, we explore two solutions: re-constructing a balanced dataset or using a balanced loss.

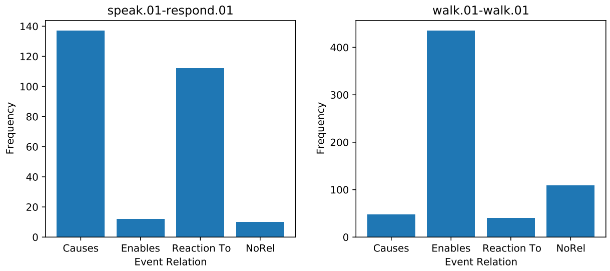
For the balanced dataset, we aim at keeping the same number of event pairs in each class by removing videos containing multiple samples of dominant relations. After this process, about 70% of the dataset is kept.

For the balanced loss, we adopt the commonly used weighted cross entropy loss for optimization,

$$\min_{\theta} -\beta_l \log p_l, \quad (6)$$

where we set β_l as the inverse of the proportion of this relation l in the training set.

VisualCOMET Pretraining. VisualCOMET [33] contains 1.4 million commonsense inferences for current visual events under three types of event relations: *Before*, *Intent* and *After*, which correspond to inferring the past, reason, or future event. We reformulate VisualCOMET for event relation prediction pretraining. The main challenge is the different annotation formats between VisualCOMET [33] and VidSitu [38]. As illustrated in Figure 2, VisualCOMET provides natural sentence annotations while VidSitu relies on SSR with verbs and argument roles. We employ AMR-Lib [13][1] to convert VisualCOMET annotations into Abstract Meaning Representation(AMR) [18] graphs where the necessary verbs and arguments are restructured to match native VidSitu annotations. Note that the AMR parsing step is crucial in improving performance on VidSitu. We take the current event as x_3 and then randomly sample from the three types of annotated events to construct an event sequence $\{x_1, x_2, x_3, x_4, x_5\}$. The relation label is then automatically generated from the type of annotations. Note that during random sampling, to simulate a real event sequence, we restrict x_1, x_2 to be either *Before* or *Intent* events and restrict x_4, x_5 to be *Intent* or *After* events.



(a) Speak-Respond

(b) Walk-Walk

Figure 3. Histograms of the event relations for two event type pairs.

4. Experiment Results and Discussions

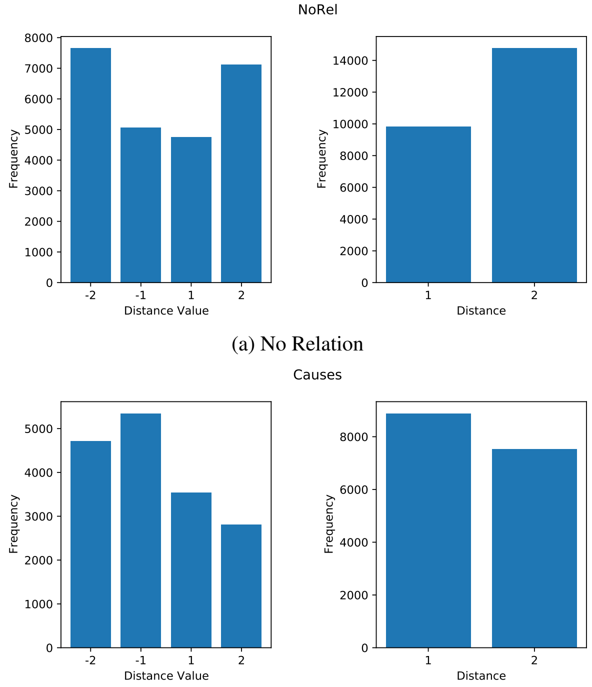
4.1. Dataset and Evaluation Metric

Our main experiments are conducted on VidSitu [38], a large-scale dataset containing 29,000 10-second video clips where each video clip is divided into five 2-second segments and the most salient verbs and arguments are annotated along with the most dominant event relation. We evaluate event-relation prediction by computing Top-1 accuracy on predicted relations as well as Top-1 accuracy macro-averaged across the four different relation classes: *Causes*, *Enables*, *Reaction To*, and *No Relation* [38].

We adopt VLEP [20] to evaluate the generalization of our model on future event prediction, which is formulated as a multiple-choice problem. We follow the official setting and report accuracy on the validation set.

4.2. Why SSR-Based Methods Fail

Preliminary Analysis. To understand why SSR baselines [38] fail on the VidSitu dataset, we check whether there are event-pairs that have a dominant relation other than *Enables*, which is the overall dominant relation in the dataset. As the two examples in Figure 3 demonstrate, we found that 66% of event type pairs satisfy this constraint, which indicates that if the model simply memorizes the dominant event relation for each event type pair, it should achieve a macro-accuracy higher than 25% [38]. This evidence motivates us to further explore different hyperparameters to investigate if the failure comes from the optimization side.



(a) No Relation

(b) Causes

Figure 4. Distribution over the relative distance for No Relation and Causes. For example, x_1 to x_3 has a distance value of -2 and a distance of 2.

To explore the benefits of using sequences of events as inputs, we study the patterns between the relation type and the relative distance from x_3 to the event. As shown in Figure 4, the distance distributions of *No Relation* and *Causes* are quite different: *No Relation* has an almost symmetric distribution w.r.t. the distance value and is significantly more frequent when the distance is 2; however, *Causes* has a clear peak at distance -1 and decays when the distance value increases, which is consistent with the temporal order. This

[1] https://github.com/bjascob/amrlib

Model	LR	Data	Val Macro Top 1 Acc(%)	Val Top 1 Acc(%)
Baseline [38]	1e-4	Original	25.00	39.43
Baseline	1e-4	Balanced	25.00	39.43
Baseline + Video Features [38]	1e-4	Original	33.73	43.71
Baseline	1e-5	Original	53.61	58.92
Baseline	1e-5	Original + balanced loss	53.98	60.47
Baseline + Video Features	1e-5	Original	44.08	49.75
Event-Sequence	1e-5	Original + balanced loss	55.38	61.59

Table 1. Comparison between baseline model trained with original unbalanced data, baseline model trained with data balanced on event relation class frequency through undersampling, and baseline model with adjusted learning rate trained on original data.

study verifies that feeding the full sequence of events as inputs is beneficial to the SSR-based models that only predict *Enables*.

Optimizing Baseline Training Configurations. This study exploits the potential training configuration issues of the baseline models that reportedly failed to predict meaningful event relations [38]. Since a moderate imbalance is observed in the distribution of event relation classes, we reevaluate the baseline model after training on the same dataset class-balanced through undersampling. This does not improve baseline model performance, which still always predicts one relation class for all events scoring 25% validation accuracy.

We then discover the reason behind the poor performance on the state-of-the-art baseline models to be a sub-optimally adjusted learning rate rather than lack of patterns in the data. As shown in Table 1, by simply adjusting the learning rate from 1e-4 to 1e-5, performance dramatically improves. Similar adjustments on state-of-the-art multi-modal models using text annotations and video features also yield a significant performance improvement. But it is noteworthy that the improvement is much less significant compared to our baseline model, which is built upon SSRs of events.

4.3. Evaluation Challenges

Simultaneous Events in Videos. Despite being rich in context, the noisy nature of video features may actually degrade performance. In complex and busy events, multiple salient verbs may emerge to dominate a scene. In such a scenario, the specific verb chosen drastically influence argument labeling of entity co-references and subsequent event-relationships. Figure 5 depicts such a scene where multiple relevant verbs and their corresponding arguments emerge. Since event relation annotations are directly tied to annotated event types and argument roles, many instances occur where predicted event relations are accurate to the scene, but evaluated as incorrect since they differ from the annotated events. We explore this further in Section 4.4.

Quantifying the Effect of Simultaneous Events. We evaluate two state-of-the-art video-language models, HERO [22] and ClipBERT [19], on video event-relation prediction. We leverage pretrained CLIP-VIT-32 [35] to perform frame selection and region selection given event and argument role as the textual input, respectively. In both models, as shown in Table 2, we observe a performance increase when frame selection and region selection are utilized, which indicates a significant presence of irrelevant frames in one video segment and also shows the negative effect cause by simultaneous events.

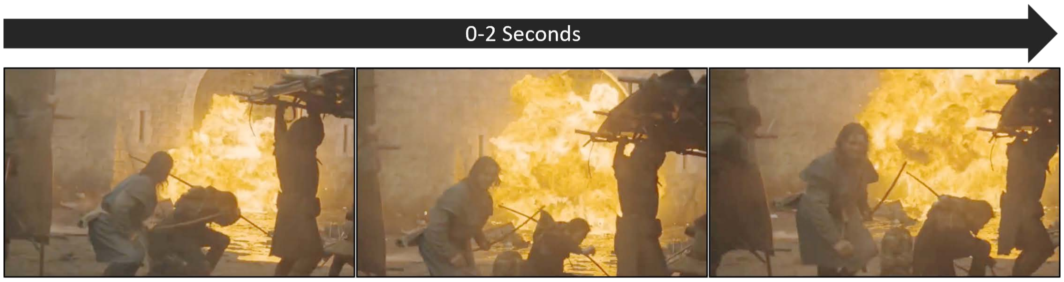
Model	Frame-level	Region-level	Val Macro Acc(%)
HERO [22]	✗	✗	42.15
HERO	✓	✗	48.03
HERO	✓	✓	48.48
ClipBERT [19]	✗	✗	47.62
ClipBERT	✓	✗	49.76
ClipBERT	✓	✓	50.52

Table 2. Performance of HERO and ClipBERT when with and without Frame Selection or Region Selection based on CLIP.

4.4. Improving SSR-Based Methods

Adding Additional Contextual Information. When predicting the relationship between two events within a five event sequence from VidSitu, baseline models only take the two target events as inputs. Richer context can be provided directly by increasing the temporal duration of the inputs. As shown in Table 3, by giving the model all five events within a sequence rather than just the two target events, we are able to leverage patterns seen in distances between events as well as events preceding and succeeding the target events, resulting in an accuracy improvement (53.85% → 55.38%).

Previous state-of-the-art baselines also receive inputs containing only verbs and base arguments (arguments tied to the verb through direct semantic relations). Providing specific context in the form of auxiliary arguments describing the scene of the event, mannerisms and goals of entities performing actions, etc., yields further performance increases (55.38% → 58.60%). When evaluating event-sequence models given only verbs and only base arguments,



Description 1	
Verb	run (walk quickly)
Arg0 (runner)	man in grey outfit
ArgM (direction)	away from explosion

Description 2	
Verb	explode (go boom)
Arg1 (thing exploding)	castle gate
ArgM (location)	castle

Figure 5. Example scene containing multiple salient events. A large explosion dominates the background while a man is running away from the explosion in the foreground. The verb and argument labels are highly dependent on which event is selected.

Model	Verbs	Base Args	Aux Args	Val Macro Acc (%)
Baseline (1e-5 lr)	✓	✓	✗	53.85
Baseline + Video Features (1e-5 lr)	✓	✓	✗	44.08
Event-Sequence Only Verb	✓	✗	✗	42.53
Event-Sequence Only Args	✗	✓	✓	34.73
Event-Sequence Baseline	✓	✓	✗	55.38
Event-Sequence All Args	✓	✓	✓	58.60
Event-Sequence + vid features (SlowFast)	✓	✓	✓	55.64
Event-Sequence + vid features (CLIP)	✓	✓	✓	45.98
HERO [22]	✓	✓	✓	42.15
ClipBERT [19]	✓	✓	✓	47.62

Table 3. Comparison of models given inputs with various forms of context.

performance is still substantially better than random guessing, showing that verbs alone without entity co-references and vice versa provide useful information for event-relation prediction.

Adding context through video features degrades performance rather than showing improvement. We observe a sizable decrease in performance in both the learning rate adjusted baseline (53.58% → 44.08%) and event-sequence input models (58.60% → 55.64%) (Table 3). We also observe a decrease in performance when switching to a CLIP-based feature extractor on video inputs instead of SlowFast. In addition, we compare with more powerful vision-language cross-modal baselines (HERO [22] and ClipBERT [19]) that have recently shown promise on other video event understanding tasks. However, neither HERO nor ClipBERT perform as well as our Event-Sequence model, which verifies its strong effectiveness.

Additional Model Constraints. We find that adding further constraints to event prediction models by employing Sequence-to-Sequence models using BART [21] does not improve performance beyond our best event-sequence RoBERTa models (Table 4). However, pretraining on VisualCOMET [33] yields further improvement, showing that commonsense knowledge of event evolution in VisualCOMET could further help our contextualized sequence model in reasoning event relations.

Model	Val Macro Top 1 Acc	Val Top 1 Acc
Event-Sequence	55.38%	61.59%
Seq-to-Seq	52.91%	57.43%
Event-Sequence + All Args	58.60%	62.01%
+ VisCom pretraining	59.21%	62.65%

Table 4. Comparison between Event-Sequence model, Sequence-to-Sequence (Seq-to-Seq) model, and Event-Sequence model pretrained on VisualCOMET data.

Performance Analyses Between Predicted and Human Annotations. To demonstrate an end-to-end approach, we train and evaluate our event-sequence model using predicted verb and argument roles generated by employing a joint feature extractor and text encoder on the video frames [38]. Results comparing the models with oracle information, predicted events, and previous state-of-the-art video only baselines [38] are summarized in Table 5.

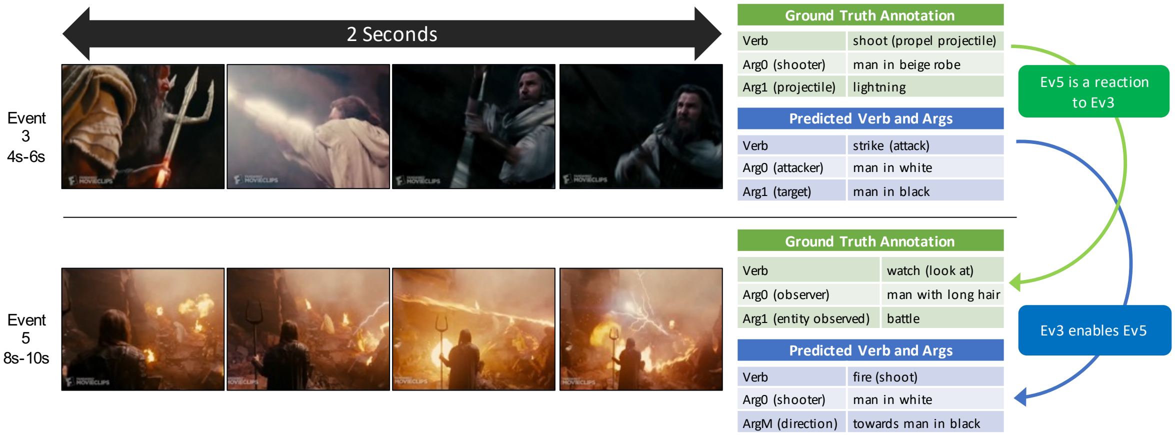


Figure 6. Example where predicted events differ from ground truth human annotations, but predicted verbs and arguments are salient since they describe different events within the same scene. The resulting event-relation prediction using predicted events is also "accurate" but different from the ground truth relation since they are relating different, equally salient event descriptions.

Input	Val Macro-Acc	Test Macro-Acc
Annotated verb + args	55.38%	54.47%
Annotated verb + Predicted args	43.30%	42.75%
Predicted verb + args	35.46%	34.94%
Vid features (SlowFast)	33.78%	30.54%

Table 5. Comparison between Event-Sequence models trained and evaluated on human annotations versus predicted events.

Notably, our event-sequence model using predicted events outperforms previous video encoder baselines showing the promise of SSRs even when they are generated. We observe a significant improvement when annotated verbs are used to generate argument roles, which verifies the effectiveness of SSR and indicates the "noisy" nature of the task and dataset.

Specifically, many scenes occur where predicted verbs and arguments describe different events than ground-truth(GT) human annotations. In the example shown in Figure 6, both GT and predicted events accurately describe the same scene in Event 3: the two men fighting. In Event 5, the predicted verb and arguments describe the background battle and subsequently classifies Event 3 as *Enabling* Event 5 since the battle is a continuation of the previous fighting. However, human annotations describe the man in the foreground watching the battle in the background, and thus the GT label is a *Reaction To* the fighting in Event 3. In this case, *Enables* is an accurate relationship descriptor but considered incorrect. Such discrepancies between annotations likely explain some of the performance degradation when using predicted events.

4.5. Applications in Downstream Tasks

We adopt the RoBERTa-based model from [20] as the baseline, which achieves 65.8% accuracy on the VLEP validation. It is a challenging task, as pretraining RoBERTa on ATOMIC [39] only improves accuracy by 0.5%. However, when using our event-sequence model as the pretrained weights for the RoBERTa model, we observe a 0.9% improvement of accuracy, which makes video event-relation prediction a better knowledge source of pretraining than the comprehensive textual knowledge base, ATOMIC. This again verifies the effectiveness and the generalization capabilities of our proposed contextualized sequence model.

5. Conclusion

In this paper, we defend the effectiveness of SSRs for video event-relation prediction and identify sub-optimal training configurations as the key reason previous models fail. We further provide in-depth analyses and find that video features are noisy because of simultaneous events and irrelevant backgrounds. Oracle event information is important to ensure proper evaluation. We propose a contextualized sequence model with a pretraining technique on VisualCOMET to demonstrate the effectiveness of SSR, which significantly outperforms the state-of-the-art models. Our model trained on video event-relation prediction can be generalized to downstream tasks such as future video event prediction.

Acknowledgement

This paper is based upon work supported by U.S. DARPA KAIROS Program No. FA8750-19-2-1004. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing official policies, either expressed or implied, of DARPA, or the U.S. Government.

References

[1] Hammad A Ayyubi, Christopher Thomas, Lovish Chum, Rahul Lokesh, Yulei Niu, Xudong Lin, Long Chen, Jaywon Koo, Sounak Ray, and Shih-Fu Chang. Multimodal event graphs: Towards event centric understanding of multimodal world. *arXiv preprint arXiv:2206.07207*, 2022. 1

[2] Allison Badgett and Ruihong Huang. Extracting subevents via an effective two-phase approach. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 906–911, 2016. 1

[3] Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. Comet: Commonsense transformers for knowledge graph construction. In *Association for Computational Linguistics (ACL)*, 2019. 1

[4] Nathanael Chambers and Dan Jurafsky. Unsupervised learning of narrative event chains. In *Proceedings of ACL-08: HLT*, pages 789–797, 2008. 2

[5] Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. Hico: A benchmark for recognizing human-object interactions in images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1017–1025, 2015. 2

[6] Brian Chen, Xudong Lin, Christopher Thomas, Manling Li, Shoya Yoshida, Lovish Chum, Heng Ji, and Shih-Fu Chang. Joint multimedia event extraction from video and article. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 74–88, 2021. 1, 2

[7] Pradipto Das, Chenliang Xu, Richard F Doell, and Jason J Corso. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2634–2641, 2013. 2

[8] Dmitriy Dligach, Timothy Miller, Chen Lin, Steven Bethard, and Guergana Savova. Neural temporal relation extraction. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 746–751, 2017. 1, 2

[9] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 3

[10] Goran Glavaš, Jan Šnajder, Parisa Kordjamshidi, and Marie-Francine Moens. Hieve: A corpus for extracting event hierarchies from news stories. In *Proceedings of 9th language resources and evaluation conference*, pages 3678–3683. ELRA, 2014. 1

[11] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015. 2

[12] Yu Hong, Tongtao Zhang, Tim O'Gorman, Sharone Horowit-Hendler, Heng Ji, and Martha Palmer. Building a cross-document event-event relation corpus. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 1–6, 2016. 2

[13] Matthew Honnibal and Ines Montani. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7(1):411–420, 2017. 4, 5

[14] Drew Hudson and Christopher D Manning. Learning by abstraction: The neural state machine. *Advances in Neural Information Processing Systems*, 32, 2019. 1

[15] Bram Jans, Steven Bethard, Ivan Vulic, and Marie-Francine Moens. Skip n-grams and ranking functions for predicting script events. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*, pages 336–344. ACL; East Stroudsburg, PA, 2012. 2

[16] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as compositions of spatio-temporal scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10236–10247, 2020. 1, 2

[17] Keizo Kato, Yin Li, and Abhinav Gupta. Compositional learning for human object interaction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 234–251, 2018. 2

[18] Shu Cai Madalina Georgescu Kira Griffitt Ulf Hermjakob Kevin Knight Philipp Koehn Martha Palmer Laura Banarescu, Claire Bonial and Nathan Schneider. Abstract meaning representation for sembankin. 2013. 5

[19] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L. Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learningvia sparse sampling. In *CVPR*, 2021. 2, 6, 7

[20] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. What is more likely to happen next? video-and-language future event prediction. *arXiv preprint arXiv:2010.07999*, 2020. 5, 8

[21] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019. 4, 7

[22] 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 preprint arXiv:2005.00200*, 2020. 2, 6, 7

[23] Manling Li, Alireza Zareian, Qi Zeng, Spencer Whitehead, Di Lu, Heng Ji, and Shih-Fu Chang. Cross-media structured common space for multimedia event extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2557–2568, 2020. 1, 2

[24] Yong-Lu Li, Siyuan Zhou, Xijie Huang, Liang Xu, Ze Ma, Hao-Shu Fang, Yanfeng Wang, and Cewu Lu. Transferable interactiveness knowledge for human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3585–3594, 2019. 2

[25] Xudong Lin, Gedas Bertasius, Jue Wang, Shih-Fu Chang, Devi Parikh, and Lorenzo Torresani. Vx2text: End-to-end learning of video-based text generation from multimodal inputs. In *Proceedings of the IEEE/CVF Conference on Com-*

puter Vision and Pattern Recognition, pages 7005–7015, 2021. 1, 4

[26] Xudong Lin, Simran Tiwari, Shiyuan Huang, Manling Li, Mike Zheng Shou, Heng Ji, and Shih-Fu Chang. Towards fast adaptation of pretrained contrastive models for multi-channel video-language retrieval. *arXiv preprint arXiv:2206.02082*, 2022. 1

[27] Jian Liu, Yubo Chen, Kang Liu, Wei Bi, and Xiaojiang Liu. Event extraction as machine reading comprehension. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1641–1651, Online, 2020. Association for Computational Linguistics. 1

[28] Jian Liu, Yubo Chen, Kang Liu, and Jun Zhao. Neural cross-lingual event detection with minimal parallel resources. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 738–748, Hong Kong, China, 2019. Association for Computational Linguistics. 1

[29] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 3

[30] Nasrin Mostafazadeh, Alyson Grealish, Nathanael Chambers, James Allen, and Lucy Vanderwende. Caters: Causal and temporal relation scheme for semantic annotation of event structures. In *Proceedings of the Fourth Workshop on Events*, pages 51–61, 2016. 2

[31] Thien Huu Nguyen, Kyunghyun Cho, and Ralph Grishman. Joint event extraction via recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 300–309, San Diego, California, 2016. Association for Computational Linguistics. 1

[32] Tim O'Gorman, Kristin Wright-Bettner, and Martha Palmer. Richer event description: Integrating event coreference with temporal, causal and bridging annotation. In *Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016)*, pages 47–56, 2016. 1

[33] Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. Visualcomet: Reasoning about the dynamic context of a still image. In *European Conference on Computer Vision*, pages 508–524. Springer, Cham, 2020. 2, 5, 7

[34] Sarah Pratt, Mark Yatskar, Luca Weihs, Ali Farhadi, and Aniruddha Kembhavi. Grounded situation recognition. In *European Conference on Computer Vision*, pages 314–332. Springer, 2020. 1, 2

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2, 6

[36] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019. 4

[37] Nishant Rai, Haofeng Chen, Jingwei Ji, Rishi Desai, Kazuki Kozuka, Shun Ishizaka, Ehsan Adeli, and Juan Carlos Niebles. Home action genome: Cooperative compositional action understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11184–11193, 2021. 2

[38] Arka Sadhu, Tanmay Gupta, Mark Yatskar, Ram Nevatia, and Aniruddha Kembhavi. Visual semantic role labeling for video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5589–5600, 2021. 1, 2, 3, 4, 5, 6, 8

[39] Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3027–3035, 2019. 1, 8

[40] Lei Sha, Feng Qian, Baobao Chang, and Zhifang Sui. Jointly extracting event triggers and arguments by dependency-bridge RNN and tensor-based argument interaction. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5916–5923. AAAI Press, 2018. 1

[41] Xiaozhi Wang, Shengyu Jia, Xu Han, Zhiyuan Liu, Juanzi Li, Peng Li, and Jie Zhou. Neural Gibbs Sampling for Joint Event Argument Extraction. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 169–180, Suzhou, China, 2020. Association for Computational Linguistics. 1

[42] Xiaozhi Wang, Ziqi Wang, Xu Han, Zhiyuan Liu, Juanzi Li, Peng Li, Maosong Sun, Jie Zhou, and Xiang Ren. HMEAE: Hierarchical modular event argument extraction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5777–5783, Hong Kong, China, 2019. Association for Computational Linguistics. 1

[43] Zhenhailong Wang, Manling Li, Ruochen Xu, Luowei Zhou, Jie Lei, Xudong Lin, Shuohang Wang, Ziyi Yang, Chenguang Zhu, Derek Hoiem, et al. Language models with image descriptors are strong few-shot video-language learners. In *Advances in Neural Information Processing Systems*. 1

[44] Ronald J Williams and David Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, 1(2):270–280, 1989. 4

[45] Ning Xu, An-An Liu, Jing Liu, Weizhi Nie, and Yuting Su. Scene graph captioner: Image captioning based on structural visual representation. *Journal of Visual Communication and Image Representation*, 58:477–485, 2019. 1

[46] Guang Yang, Manling Li, Xudong Lin, Jiajie Zhang, Shih-Fu Chang, and Heng Ji. Video event extraction via tracking visual states of arguments. *arXiv preprint arXiv:2211.01781*, 2022. 1

[47] Wenlin Yao, Zeyu Dai, Maitreyi Ramaswamy, Bonan Min, and Ruihong Huang. Weakly supervised subevent knowledge acquisition. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020. 1

[48] Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. Situation recognition: Visual semantic role labeling for image understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5534–5542, 2016. 1, 2

[49] Guangnan Ye, Yitong Li, Hongliang Xu, Dong Liu, and Shih-Fu Chang. Eventnet: A large scale structured concept library for complex event detection in video. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 471–480, 2015. 1

[50] Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. *Advances in neural information processing systems*, 31, 2018. 1, 2