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

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

A. Further Discussion

In the main paper, we discussed (1) the noisy nature of the video event-relation prediction dataset; (2) the superiority of structural symbolic representation (SSR) as a representation of event information; and (3) the performance gains from contextual information and pretraining on an external knowledge base. Due to limited space in the main paper, in this supplementary document, we include further discussions on the relation between SSR's and (1) human-annotated ground-truth labels and (2) scene graph generation.

Q1: Are SSR's always ground-truth event information?
A1: No, SSR's are **not necessarily** from oracle information. We have results obtained using event types and arguments roles detected from video features (Table 5 in the main paper). In fact, performance using detected event types and arguments outperforms models trained with only video features (35.4% vs 33.8%), showing that SSR is indeed a better representation compared to directly using visual features, even without any oracle information. However, in most of the comparisons, we use the evaluation scenario taking only oracle information as input due to the the dataset-inherent noise as discussed in Section 4.3.

Q2: What are similarities and differences between the three settings in scene graph generation?
A2: The three settings we explored (using ground-truth verb and arguments, using ground-truth verb and predicted arguments, and using predicted verb and arguments) are somewhat similar to the three common settings of scene graph generation (predicate classification, scene graph classification, and scene graph detection). We would like to share our understanding on the distinctions and shortcuts.

The ambiguity in scene graphs is spatial and object-wise; however, the ambiguity in video event relation prediction is spatial-temporal and event-wise. Zellers et al. [8] show that relationships in scene graphs can be easily predicted with only object labels as input. As shown in Table 3 in the main paper, our contextualized sequence model with SSR's significantly improves performance (42.5% $\rightarrow$ 58.6%) compared to the verb only variant. In fact, [8] is even weaker than baseline when using an unbiased metric based on Macro-Recall [7]. However, in our case, we show consistent improvement under both Macro and Micro-average seen in Table 1 and Table 4 in the main paper.

Therefore, despite sharing certain similarities with the three settings in scene graph generation, our proposed model behaves differently and also gains an advantage on video relation-prediction.

B. Full Test Set Results

We obtain test results of our best models by generating prediction files on unlabeled test sets provided by VidSitu and submitting to the VidSitu open leaderboard hosted by the Allen Institute for AI. Results are summarized in Table 1. We observe a substantial improvement in test accuracy (31.57% $\rightarrow$ 60.56%) comparing our most performant model to the previous state-of-the-art baseline. We also include state-of-the-art video-language models HERO and ClipBERT for comparison. Our proposed Event-Sequence model also achieves a quite substantial improvement over both HERO and ClipBERT when annotated verbs and arguments are provided. We note that results using Annotated verb + Predicted args are less rigorous due to the possible noise issues arising when multiple events of the same type happen in the same video segment. The results under Predicted verb and args are more unreliable since we often come across multiple events happening simultaneously within one video segment.

*Equal contribution.

Model	Val Macro Acc(%)	Test Macro Acc(%)
Annotated verb and args		
Baseline (1e-4 lr)	25.00	25.00
Baseline + Video Features (1e-4 lr, SOTA reimplimented)	33.73	31.57
Baseline + Video Features (1e-4 lr, SOTA reported by VidSitu [5])	34.15	32.98
HERO + All args	42.15	41.93
ClipBERT + All args	47.62	47.25
Event-Sequence (Ours)	55.38	54.47
Event-Sequence + All args (Ours)	58.60	58.64
Event-Sequence + All args + vid features (Ours)	55.64	56.74
Event-Sequence + All args + VisCom pretraining (Ours)	**59.30**	**60.56**
Annotated verb + Predicted args		
HERO	40.63	39.97
ClipBERT	41.20	41.17
Event-Sequence (Ours)	43.30	42.75
Predicted verb and args		
HERO	37.42	37.51
ClipBERT	37.58	37.09
Event-Sequence (Ours)	35.46	34.94

Table 1. Validation and testing accuracy of previous state-of-the-art baselines and our best performing models under three settings. For the Baseline + Video Features model, we show both results reported by VidSitu as well as our reimplementation of their best performing model. VisCom is short for VisualCOMET. All args denotes the use the additional contextual argument roles.

Representation	Pretraining Task	Val Acc(%)
Vid features	-	65.6
Vid features	Event-Relation Prediction	66.2
Verb + args	-	65.8
Verb + args	ATOMIC [6]	66.3
Verb + args	Event-Relation Prediction	**66.7**

Table 2. Comparisons on the accuracy of future video event prediction.

C. Additional Details

C.1. Frame Selection and Region Selection

We leverage a pretrained image-text contrastive model, CLIP [4] for frame selection and region selection. We first compose verb and argument role annotations into sentences using AMRLib's graph to sentence functions. As shown in Figure 1, we then use a CLIP text encoder and image encoder (ViT-Base/32) to encode the text query and associated frames in the video segment. Then by coagulating text-frame similarities, we select the most similar four frames as inputs to the video event-relation prediction model.

To obtain a region for each argument role in the video segment, as shown in Figure 2, instead of using projected embeddings of the CLS tokens in the ViT, we project the reset patch embeddings into the common embedding space between image and text. Then we can calculate the similarity map between the projected embedding map and the text embedding of the query sentence. Finally, we generate the bounding box with highest confidence based on existing tools [1]. We only use features from the generated boxes as input to the video event-relation prediction models.

C.2. VLEP

VLEP is a future event prediction dataset, where each sample is associated with two natural sentence choices. We similarly use an AMR parser to re-structure the two natural sentence choices and feed them into the RoBERTa model together with speech transcripts. To obtain the representation of the event in the current video segment, we use models pretrained on VidSitu to extract the verb and argument roles. We also adopt the feature vectors extracted by the same verb model as the representation for comparison. As show in Table 2, we observe that 1) SSR's again perform better than using continuous video features; 2) event-relation prediction is an effective pretraining task for future event prediction.

C.3. Implementation Details

We use an Nvidia V100 GPU to train and evaluate our models. We follow the experimental setting and hyper-parameters of [5] in all our experiments on the VidSitu dataset. The only hyper-parameter we tuned is the learning rate.

[1] https://github.com/shonenkov/CLIP-ODS

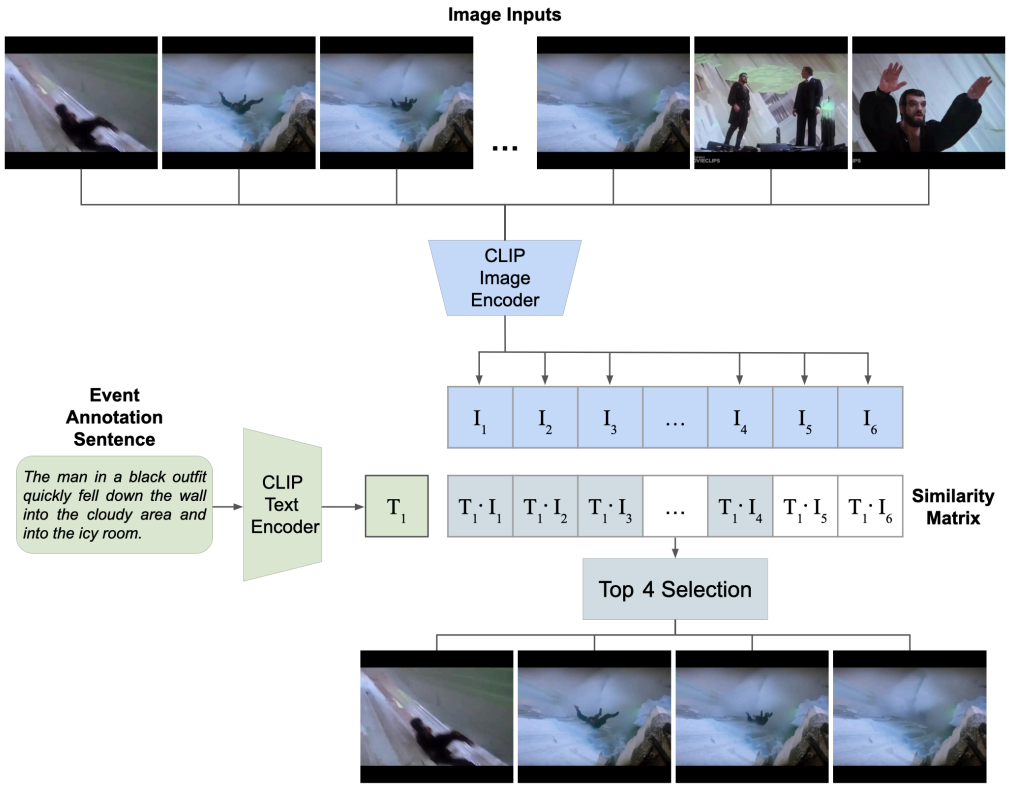


Figure 1. Frame selection pipeline with pretrained CLIP models.

Model	Val Macro Acc(%)	Val Acc(%)
Predicting over distance value + original data	25.00	39.43
Predicting over distance value + balanced data	32.16	32.97
Predicting over event type pairs + original data	29.40	34.63

Table 3. Validation accuracy of baselines using prior distributions.

We pretrain the model on VisualCOMET for 4 epochs using Adam [1] optimizer with a learning rate of $1e-5$. We use a batch size of 8. When fine-tuning from the pretrained model on VidSitu, we also use a learning rate of $1e-5$.

For experiments on VLEP, we follow [2] to train and evaluate the models when not using event-relation prediction as the pretraining task. When finetuning from our event-relation prediction model, we train the model for 6 epochs with a learning rate of $1e-5$ and a batch size of 16.

We fix the random seed in all the experiments and we do not observe significant change of accuracy ($> 0.5\%$) upon changing the random seed.

D. Baselines using Prior Distributions

We provide more details about the Preliminary Analysis that were omitted from Section 4.2 of the main paper. As shown in Figure 3, the distance distributions of each event relation type exhibit different patterns from one another. For example, the distribution of *Causes* and *Enables* w.r.t. distance value both exhibit peaks at -1 while *Reaction To* exhibits a peak at 1. On an event relation class-balanced dataset, this would suggest that within the set of related events, earlier events tend to *Cause* or *Enable* the central event, and later events tend to be a *Reaction To* the central event. Furthermore, we see that the distribution of *No Relation* w.r.t. distance shows a much higher frequency at 2 whereas the *Causes*, *Enables*, and *Reaction To* show higher frequencies at 1. This follows our intuition that events further apart temporally are more likely to be unrelated.

Since we observe numerous event relation patterns within the dataset, we present a few baselines using such prior distributions to demonstrate that our models are not just learning these dataset biases but rather utilizing contextual information contained within SSR's. Results are summarized in Table 3.

For example, by simply memorizing the dominant event relation for each distance value on the balanced training set,

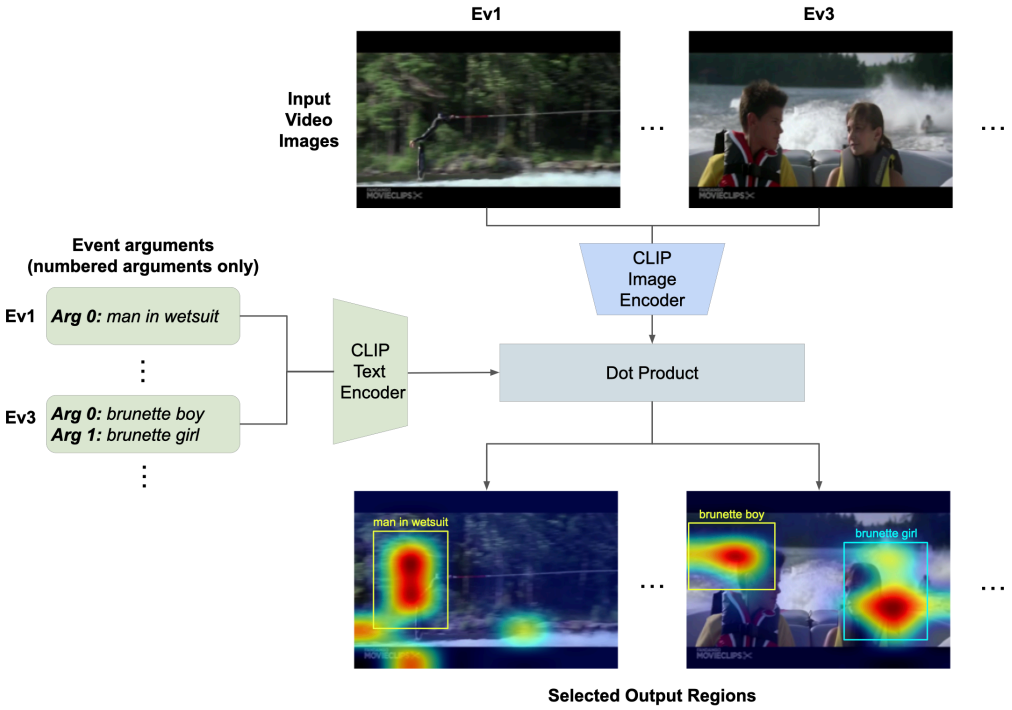


Figure 2. Region selection pipeline with pretrained CLIP models.

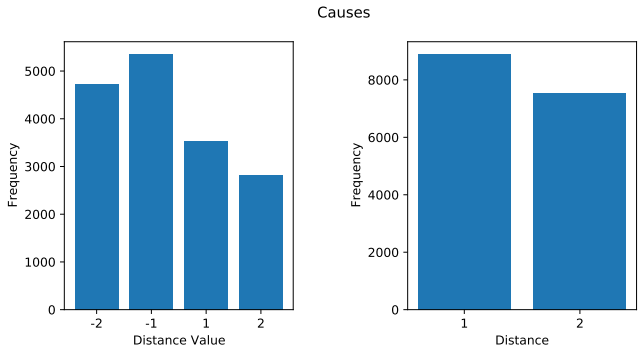
such a model scores 32.16% macro-averaged accuracy on validation. Note that on the original dataset, class imbalances cause the majority relation to dominate across distance values and the majority relation (Enables) is predicted each time (thus scoring 25%).

Similarly by memorizing the dominant event relation for each event type pair (eg. Speak-Respond $\rightarrow$ *Causes*) in the training set, such a model scores 29.40% on validation. We observed that 93% of event type pairs in the validation set were previously encountered in training and a default of random guessing is used when an unseen pair is encountered.

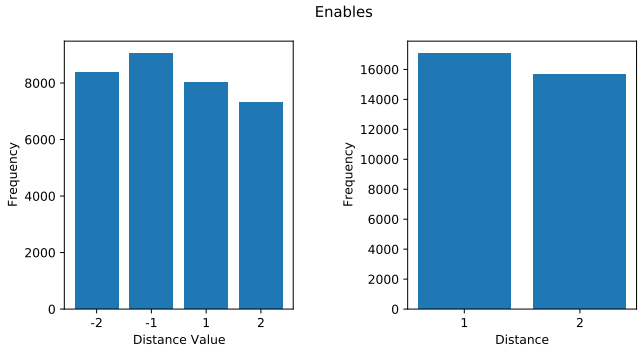
E. Broader Societal Impact and Future Work

VidSitu [5] and VisualCOMET [3] (derived from VCR) use publicly available movie clips downloaded from YouTube. VLEP [2] uses clips from TV-shows and lifestyle vlogs downloaded from YouTube. Therefore, VLEP may contain personal information but it is noteworthy that our algorithm is not designed to specifically capture personal information. We also note that certain video clips obtained from crime, action, or horror movies may contain violence and gore and viewer discretion is advised when viewing such video clips. Overall, negative societal impacts are not expected from the designed algorithms but as discussed above, the dataset used may lead to some biased or undesired results.

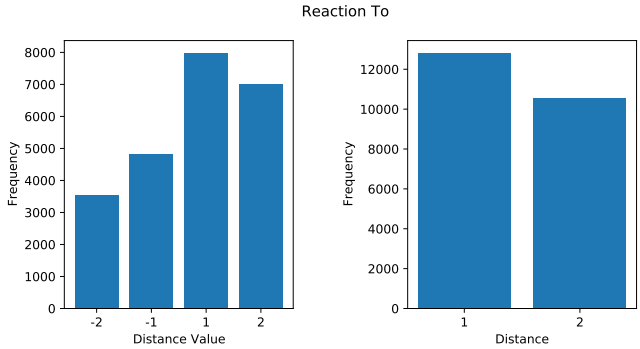
Currently, the evaluation using predicted verbs or arguments is not rigorous due to missing bounding box annotations on the event and the argument roles. In the future, we plan to add bounding box annotations to evaluate these two settings properly (using ground-truth verb and predicted arguments, and using predicted verb and arguments).



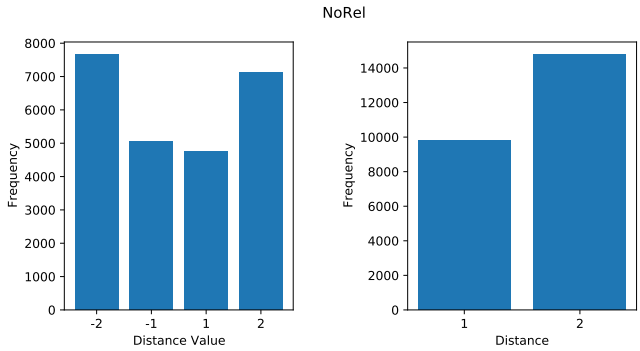
(a) Causes



(b) Enables



(c) Reaction To



(d) No Relation

Figure 3. Distribution over the relative distance for all four event relation types. For example, x_1 to x_3 has a distance value of -2 and a distance of 2. (Best viewed on a monitor when zoomed in.)

References

[1] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 3

[2] 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. 3, 4

[3] 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. 4

[4] 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

[5] 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. 2, 4

[6] 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. 2

[7] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to compose dynamic tree structures for visual contexts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6619–6628, 2019. 1

[8] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5831–5840, 2018. 1