Beyond Grounding: Extracting Fine-Grained Event Hierarchies across Modalities

Hammad Ayyubi[1], Christopher Thomas[2], Lovish Chum[1], Rahul Lokesh[3], Long Chen[4], Yulei Niu[1], Xudong Lin[1], Xuande Feng[1], Jaywon Koo[1], Sounak Ray[1], Shih-Fu Chang[1]

[1]Columbia University
[2]Virginia Tech
[3]Samsung Research America
[4]The Hong Kong University of Science and Technology
hayyubi@cs.columbia.edu

Abstract

Events describe happenings in our world that are of importance. Naturally, understanding events mentioned in multimedia content and how they are related forms an important way of comprehending our world. Existing literature can infer if events across textual and visual (video) domains are identical (via grounding) and thus, on the same semantic level. However, grounding fails to capture the intricate cross-event relations that exist due to the same events being referred to on many semantic levels. For example, the abstract event of "war" manifests at a lower semantic level through subevents "tanks firing" (in video) and airplane "shot" (in text), leading to a hierarchical, multimodal relationship between the events.

In this paper, we propose the task of extracting event hierarchies from multimodal (video and text) data to capture how the same event manifests itself in different modalities at different semantic levels. This reveals the structure of events and is critical to understanding them. To support research on this task, we introduce the ***Multi**modal **Hi**erarchical **Ev**ents* (`MultiHiEve`) dataset. Unlike prior video-language datasets, `MultiHiEve` is composed of news video-article pairs, which makes it rich in event hierarchies. We densely annotate a part of the dataset to construct the test benchmark. We show the limitations of state-of-the-art unimodal and multimodal baselines on this task. Further, we address these limitations via a new weakly supervised model, leveraging only unannotated video-article pairs from `MultiHiEve`. We perform a thorough evaluation of our proposed method which demonstrates improved performance on this task and highlight opportunities for future research. Data: https://github.com/hayyubi/multihieve

1 Introduction

Human life is eventful. We use events to describe what is happening (e.g. war, protest, etc.), to tell stories (e.g. during the war an airplane was shot down), and to depict our understanding of the world (e.g. coffin procession happens in a funeral). Thus, understanding and analyzing events is a crucial part of comprehending our world. A critical component towards this goal is to figure out the manner in which the same real-world event is manifested in multiple modalities of data.

To this end, previous studies have utilized grounding (Gao et al. 2017) to determine whether events in textual and visual domains are related identically at the same semantic level.

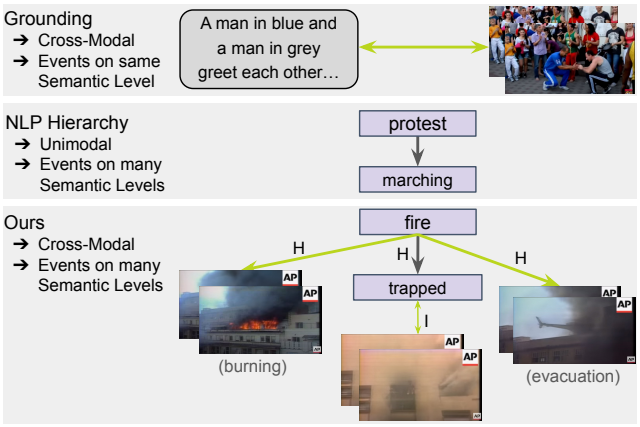

Figure 1: Illustration of our task's differences with other related tasks. Unlike prior tasks, our task is multimodal and relates events on multiple semantic scales. H: Hierarchical Relation; I: Identical Relation.

However, events in different domains can be referred to at various semantic levels, resulting in intricate hierarchical and sibling relationships. For instance, as illustrated in Figure 2, the event of "tanks firing" is a component of event "war" and denotes it at a finer semantic level. Consequently, "tanks firing" is a subevent of the parent event "war". Moreover, textual event of "airplane shot" is also a subevent of "war", and together with "tanks firing" reveal the constituents of "war" event. This creates siblings relations between "airplane shot" and "tank firing". Additionally, subevents can be further decomposed into sub-subevents, creating a hierarchy of events (see Figure 2). These event hierarchies organize events based on the semantic scale at which they occur and expose a hierarchical compositional structure, which is crucial for understanding events and their fine-grained relationships.

Much of the prior work on extracting such event hierarchies has been done in Natural Language Processing (NLP) for the text-only domain. However, as our world is multimodal, the information conveyed by a unimodal text event hierarchy is inherently limited. For example, in Figure 1, extracting "evacuation" subevent from video as a child of parent event "fire" provides us with the additional fact that relief efforts reached on time.

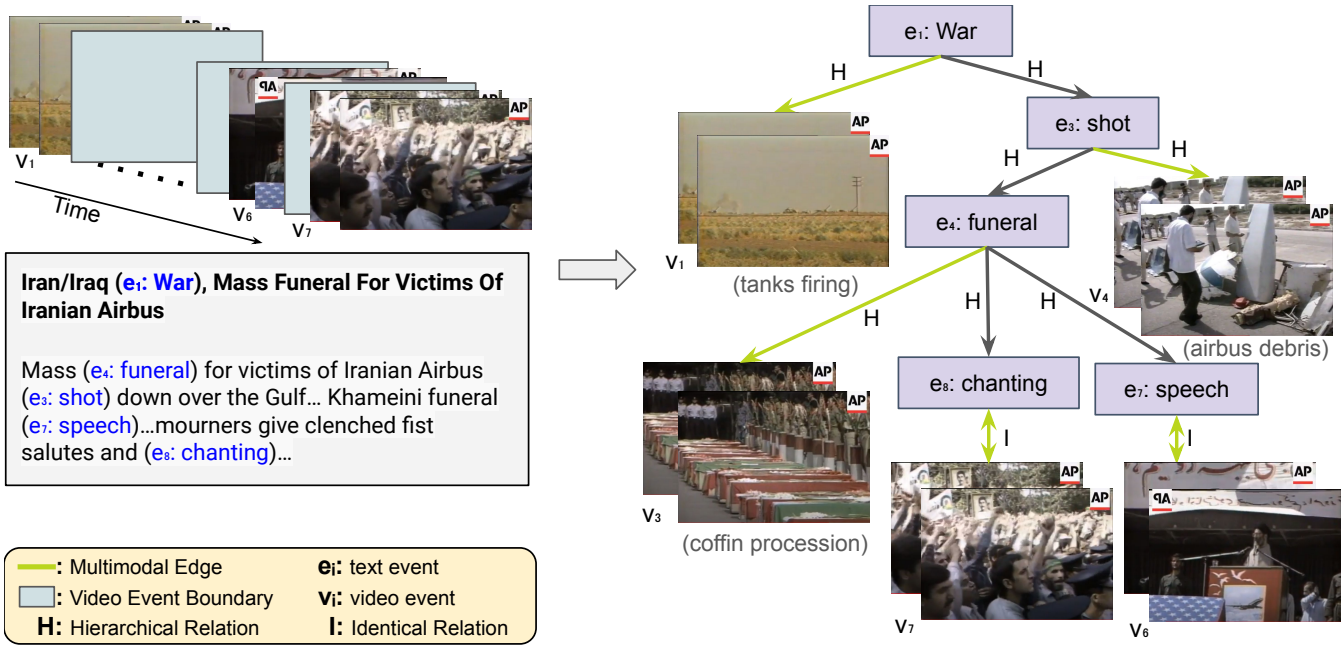

Figure 2: An example from our `MultiHiEve` dataset illustrating the proposed task of extracting event hierarchies given multimodal (text + video) content. The same legend is followed throughout the paper in all illustrations of event hierarchies.

We address these limitations through the proposed task of extracting event hierarchies from multimodal (text & video) data. Specifically, given events from paired text article and video, the task requires predicting all the multimodal hierarchical and identical event-event relationships. This output can be combined with text-only event hierarchy (from any off-the-shelf tool) to get a more holistic hierarchy.

Multimodal event hierarchies can aid many applications, such as summarization (Daumé III and Marcu 2006), story completion from multiple sources, event analysis/comparison (*e.g.* a "protest" event with "property destruction" subevent is unruly, otherwise it's peaceful), event prediction likelihood (Chaturvedi, Peng, and Roth 2017), knowledge-based information extraction (Wen et al. 2021), and multimodal knowledge graph construction (Li et al. 2020a).

To study this task, we introduce the ***Multimodal Hierarchical Events*** *(`MultiHiEve`)* dataset. `MultiHiEve` consists of approximately 100.5K pairs of news article and accompanying video. The news story in the text article mentions events on multiple semantic levels, making it ideal for the task of extracting event hierarchies. We strive to limit the socio-economic bias inherent in news media by only collecting our data from news sources rated unbiased by credible sources. We keep unannotated 100K pairs for training and densely annotate 526 pairs with multimodal hierarchical and identical relations for benchmarking and evaluation. Our annotation process is detailed and labor-intensive, requiring approximately 114 hours of expert annotator effort. Crucially, in contrast to prior text-only datasets dealing with hierarchical events, we do not limit the event types to any fixed ontology and instead consider an open world of events.

To benchmark performance on this task, we construct several baselines using state-of-the-art (SOTA) architectural components. A unimodal text-only baseline leverages ASR (automated speech recognition) and employs a SOTA NLP model (Wang et al. 2021) to find hierarchical events between a text article and its video's ASR. We also build a multimodal baseline by detecting hierarchical events in text using Wang et al. (2021) and grounding the textual subevents to video using CLIP (Radford et al. 2021). A key limitation of these baselines is that they require visual subevents to be mentioned in textual form in either the ASR or the article. To address this, we propose **M**ultimodal **A**nalysis of **S**tructured **H**ierarchical **E**vents **R**elations (MASHER), a weakly supervised model which learns to directly predict hierarchical events between text and video. By doing so, MASHER can also discover visual-only subevents (subevents not mentioned in text).

The major contributions of this work are fourfold: 1) We propose the challenging task of extracting event hierarchies from multimodal data. 2) We release `MultiHiEve` dataset to facilitate research on this task. 3) We construct several baselines and propose MASHER, a weakly supervised model, to benchmark performance. 4) We provide a detailed analysis of our dataset and methods with insights for future work.

2 Related Work

Hierarchical Event Relations in Text. Detecting hierarchical event relations (or sub-event relations) is a long-standing problem in the text domain (O'Gorman, Wright-Bettner, and Palmer 2016; Glavaš et al. 2014). Early works mainly rely on heuristic phrasal patterns. For example, Badgett and Huang (2016) found some characteristic phrases (*e.g.* "media reports" in new articles) always contain sub-events with hierarchical relations. To further enrich hierarchical event relation in-

stances, recent works (Yao et al. 2020) rely on generative language model to generate subevent knowledge among different commonsense knowledge (Bosselut et al. 2019; Sap et al. 2019), then incorporate knowledge into event ontology.

Relation Understanding in the Vision Domain. Prior work (Krishna et al. 2017b; Xu et al. 2017; Ji et al. 2020) propose scene graph methods that parse images/videos into a graph. However, the relationships studied in scene graphs are not between two events. To the best of our knowledge, the only pioneering work that has discussed the event-event relationship in video domain is VidSitu (Sadhu et al. 2021). Unfortunately, to simplify the research problem on this topic, they have made several assumptions: 1) All events are manually cutted into fixed interval (2-second). 2) All event types are "visual" only and from a fixed event ontology. On the other hand, we consider variable-length video events and focus on open-vocabulary event types (which include non-visual event types and other visual events like funeral, detain, rally etc). Besides, while they have annotated each video event with a text label, their event relations still are between events in a video. In contrast, our multimodal relations are between events in an article and a video (see Figure 1).

Multimodal Event Understanding. Since single-modality event tasks are well studied (Nguyen, Cho, and Grishman 2016; Sha et al. 2018; Liu et al. 2019, 2020; Li et al. 2017; Mallya and Lazebnik 2017; Pratt et al. 2020; Yatskar, Zettlemoyer, and Farhadi 2016; Lu et al. 2023), jointly understanding events from multiple modalities (Li et al. 2020b; Chen et al. 2021; Li et al. 2022; Zhang et al. 2017; Tong et al. 2020; Wen et al. 2021; Park et al. 2020; Reddy et al. 2022; Du et al. 2022) has attracted extensive research interests because different modalities usually provide the complementary information for comprehensively understanding the real-world complex events. Two important benchmarks (Li et al. 2020b; Chen et al. 2021) have been established for *image + text* and *video + text* settings. Li et al. (2020b) first introduced the task of jointly extracting events and labeling argument roles from both text articles and images. Chen et al. (2021) further defined the task of joint multimedia event extraction from video and text to exploit the rich dynamics from videos. However, both the works focus on event detection in comparison to the event relations task explored in this work.

3 Task

To understand (parent) events and fully comprehend what they entail, one needs to discover what (sub) events happened during the parent event. The task of extracting event hierarchy from multimodal content is aimed at revealing this compositional structure of events.

Formal Task Definition. Given a text article, T, containing events, $\{e_i\}_{i=1}^{m}$, and a video, V containing events $\{v_j\}_{j=1}^{n}$, the proposed task requires prediction of all possible hierarchical and identical event-event relations, $\{r_k\}_{k=1}^{K}$, from a text event, e_i, to a video event, v_j, among all possible $m \times n$ pairs, where $r_k \in$ {'Hierarchical', 'Identical'} and $K \leq m \times n$. We will now discuss definitions for different components of the task and justifications for task design choices.

Text Event Definition. The definition of an "event" has been defined quite thoroughly in different NLP works on information extraction (Huang et al. 2004; Reimers, Dehghani, and Gurevych 2016; GLAVAŠ and ŠNAJDER 2015). As such, we closely follow ACE Corpus's (Huang et al. 2004) definition of an event: '*a change of state or the changed state of a person, thing or entity.*' We came up with a slightly modified event definition and annotation criteria (detailed in Appendix A.1[1]) as ACE 1) addresses several linguistic nuances not required in our setting; 2) is restricted to a fixed ontology of events, whereas we annotate open-domain events.

Video Event Definition. Precisely defining what constitutes an "event" in the video domain is challenging due to the multiple granularities at which events occur in videos. For example, during a *"clash"* event, one might see a *"pulling* out baton" event and a *"throwing* a punch" event. This makes it difficult to pick salient event boundaries in video clips. Sadhu et al. (2021) circumvent this ambiguity by defining temporal event boundaries of fixed duration (two seconds). However, pre-defining the boundary duration is difficult and application specific. Additionally, a fixed duration boundary often divides salient events into multiple segments. We address these issues by defining video event boundaries to be where shot changes occur, partly following (Shou et al. 2021). From our qualitative analysis and annotator feedback, this gives us a good trade-off between ease, clarity, consistency and non-segmentation of events.

Relation Types. We define two types of event relations in this work: hierarchical and identical. These relation types are well defined in NLP (Glavaš et al. 2014) and we follow them to define the relations for our task as below:

Hierarchical: *"A parent event A is said to be hierarchically related to a subevent B, if event B is spatio-temporally contained within the event A."* For example, an "evacuation" event is a subevent of a "fire" event as it takes place during and at the same location as the fire event (see Fig. 1). Therefore, a subevent (evacuation) is a component of the parent event (fire) among multiple other subevents (burning, trapped, evacuation etc.).

Identical: *"An event A is said to be identical to another event B if both events denote exactly the same real-world events in all aspects."* For example, "trapped" event in text is identical to the video event showing people begin trapped as they both denote exactly the same event – there are no more components of trapped.

Relation direction. The multimodal relations in our event hierarchies are directed from text event to video events. The logic behind this design choice is that text events are often more abstract while video events are often atomic. For example, we are likely to observe abstract events such as war and election in text while their atomic subevents – fighting and voting – are more likely to be visible in the video.

Difference from Video Grounding. Although grounding relates similar events in text and video, it does not distinguish the type of relationship. That is, whether the video event shows all aspects of text event (*i.e.* identical) or whether it only shows "part-of" of the parent event and is thus a

[1]For all Appendix references, please see Ayyubi et al. (2023).

Dataset	Domain	W/m[2]
MSVD (Chen and Dolan 2011)	Open	54
MSR-VTT (Xu et al. 2016)	Open	38
Charades (Sigurdsson et al. 2016)	Activities	-
ActyNet Cap (Krishna et al. 2017a)	Open	27
HowTo100M (Miech et al. 2019)	Instructional	67
YT-Temporal-180M (Zellers et al. 2021)	Open	-
VidSitu (Sadhu et al. 2021)	Movie	Null
MultiHiEve	News	113

Table 1: Comparing MultiHiEve to prior video-language datasets. It is the first dataset sourced from news domain. Grayed row denotes video-only dataset. W/m: Words/minute

subevent. This has major implications. For example, in Figure 1, inferring that "burning" video subevent is identical to "fire" would imply that there was nothing else that happened during fire event and hence, relief efforts did not reach on time. On the other hand, inferring "burning" to be a subevent of "fire" indicates that there may have occured other relief/"evacuation" subevents. Further, some video subevents are visually dissimilar to its textual parent event (for example, "evacuation" is dissimilar to "fire" in Figure 1), making it difficult for grounding to relate such subevents.

4 Dataset

To support research on the proposed task, we introduce MultiHiEve – a dataset containing news articles and the associated video clips. Existing video-language datasets contain either manually annotated descriptions of video events or utterances from the video itself (see Table 1). In both cases, the text is essentially on the same semantic level as the video event. However, they lack a context or an overall story describing events on higher semantic levels that comprise the video events. In contrast, news stories provide a rich hierarchy of events, making them ideal for our task. Having $\sim$ 2x more words per minute as compared to other datasets (see Table 1), indicates this to some extent.

4.1 Data Collection and Curation

A potential drawback with news data is that they could be socio-economically biased and sensationalized. We mitigate these issues by choosing media sources rated "Center" (out of "Left", "Left Leaning", "Center", "Right Leaning" and "Right" ratings) by the media rating website allsides.com, resulting in a total of 9 news media sources (*c.f.* Appendix B.1). We scraped Youtube for news videos, associated text story and closed captions (ASR) from the official channels of these sources, collecting a total of 100.5K videos. We filtered videos whose duration was greater than 14 minutes or whose descriptions had less than 10 words. This was done to prune videos that may be too computationally expensive to process or whose descriptions may be too short to have meaningful events. We split the data two ways – 1) 100K unannotated train split for self-supervised/weakly supervised training and 2) 526 annotated test split - 249 validation set and 277 test set - for benchmarking and evaluation. We annotate a relatively small set because of the challenging and resource-consuming nature of the annotation process; two popular NLP Hierarchical event-event relations datasets (Hovy et al. 2013; Wang et al. 2021) contain 100 articles each (including train split).

[2]For datasets with duplicate descriptions per video clip (MSVD, MSR-VTT), words/min. is averaged by #descriptions .

Dataset	Modality	#Hier. Rels.	#Id. Rels.
HiEve (Glavaš et al. 2014)	Text	3648	758
IC (Hovy et al. 2013)	Text	4586	2353
MultiHiEve	Text + Vision	3077	1524

Table 2: Comparison of MultiHiEve against other hierarchical-event datasets. "Hier." denotes "Hierarchical". "Id." denotes "Identical".

4.2 Train Split

The train split contains 100K videos with a total duration of more than 4K hours. The paired text descriptions total 1.9M sentences and 28M words. The large-scale nature of the data allows for self/weakly-supervised learning on the task. We provide additional data statistics, topic distribution exploration and quantitative comparison against 12 popular video-language datasets in Appendix B.2.

4.3 Test Split

Annotation Procedure. As a first step, following the definition of a video event from Section 3, we extract video events using an off-the-shelf video segmentation model: PySceneDetect [3]. To make text event annotation easier, we provide automatically extracted text events (using (Shen et al. 2021)) to the annotators along with instructions to add or omit events according to the definition in Section 3. Next, we task the annotators to mark all possible relations, $\in$ "Hierarchical", "Identical", from the annotated text events to the provided video events in a video-article pair. We provide screenshots of our annotation tool and additional details in Appendix B.3.

We train 5 expert annotators for this task through a series of short seminars and multiple rounds of feedback and consultation with all the annotators to improve consensus. Excluding training, annotation required 114 hours in total, reflecting the labor-intensive and complex nature of the task.

Inter Annotator Agreement (IAA). We measure the quality of the annotations using IAA. Inspired by (Glavaš et al. 2014) and (GLAVAŠ and ŠNAJDER 2015), we formulate $\text{IAA}_j = \frac{\sum_r \mathbb{1}(x_{rj} \geq 2)}{|\cup_{i=1}^{5} S_j^i|}$, where $j \in$ {"Hierarchical", "Identical"}, S_j^i denotes the set of all relations annotated by annotator i as j, $r \in \cup_{i=1}^{5} S_j^i$ and x_{rj} represents # annotators who marked relation r as j. The intuition behind this formulation is to calculate the percentage of relations which have been annotated by at least 2 annotators. We obtain $\text{IAA}_{Hierarchical} = 47.5$ and $\text{IAA}_{Identical} = 48.9$. This is not far from $\text{IAA}_{Hierarchical}$ for text datasets HiEve and IC – 69 and 62 respectively. Thus, while the text-only event rela-

tion task is itself quite challenging, our new cross-modal task is even more demanding.

Following prior work for related text-only tasks (Vulić, Ponzetto, and Glavaš 2019; Glavaš et al. 2014), we consider IAA to be an upper bound on model performance because our metrics judge the model's predictions with respect to human agreement on the task. It is not clear whether a model exceeding IAA indicates a meaningful performance gain or an overfitting to annotators' subjective tendencies.

Dataset Analysis. As we are the first to propose *multimodal* hierarchical event relations analysis, we compare our dataset against two popular *text-only* hierarchical event relations datasets in Table 2. Overall, `MultiHiEve` has a comparable number of hierarchical and identical relations, but has the added novelty of being the first multimodal (text and vision) event-event relations dataset. Further, both NLP datasets limit the event types to a fixed ontology. We do not put any such constraints on either text or video event types.

5 Multimodal Hierarchical Events Detection

Acquiring a large scale labelled dataset sufficient for training a model on the proposed task is prohibitively time and resource consuming (*c.f.* Section 4.3). Thus, we instead propose a weakly supervised method which learns from pseudo labeled data. We generate pseudo labels using existing NLP and vision techniques and then use these pseudo labels for training our model. We discuss this in detail below.

5.1 Pseudo Label Generation

Event Detection in Text and Video. The first step is detect events in text and video separately. We use the same automatic methods to detect them as used on the test data: Open Domain IE (Shen et al. 2021) and open source library PySceneDetect [3] for text event and video event detection respectively.

Textual Hierarchical Relation Detection. Assume we detected m text events, $\{e_i\}_{i=1}^{m}$, in an article T and n video events, $\{v_j\}_{j=1}^{n}$, in the accompanying video V. The next step is to detect hierarchical relations among the text events, using (Wang et al. 2021), from all possible $m \times m$ pairs. Let $\{e_u e_{u_s}\}_{u=1,s=1}^{u=p,s=q}$ denote the hierarchically related event pairs, where the parent event is e_u and the subevent is e_{u_s} and $p, q \leq m$.

Video Event Retrieval The final step is to retrieve video events, $\{v_l^{u_s}\}_{l=1}^{r}$ and $r \leq n$, from video V, which depict the same real world event as the text subevent, e_{u_s}. This step essentially simplifies to a video retrieval task. As CLIP (Radford et al. 2021) model has demonstrated state-of-the-art performance in multimodal retrieval tasks (Luo et al. 2021; Fang et al. 2021), we use it for this step. We provide more details in Appendix C.1.

We use CLIP to get all possible video events which are identical to the text subevent, denoted $\{e_{u_s} v_l^{u_s}\}$. Critically, since e_u was the parent event of e_{u_s} and e_{u_s} depicts the same event as $v_l^{u_s}$, we can conclude that e_u is the parent event of $v_l^{u_s}$ by transitivity. As a result, we get a total of $\{e_u v_l^{u_s}\}$ hierarchical event pairs and $\{e_{u_s} v_l^{u_s}\}$ identical event pairs.

[3] http://scenedetect.com/en/latest/

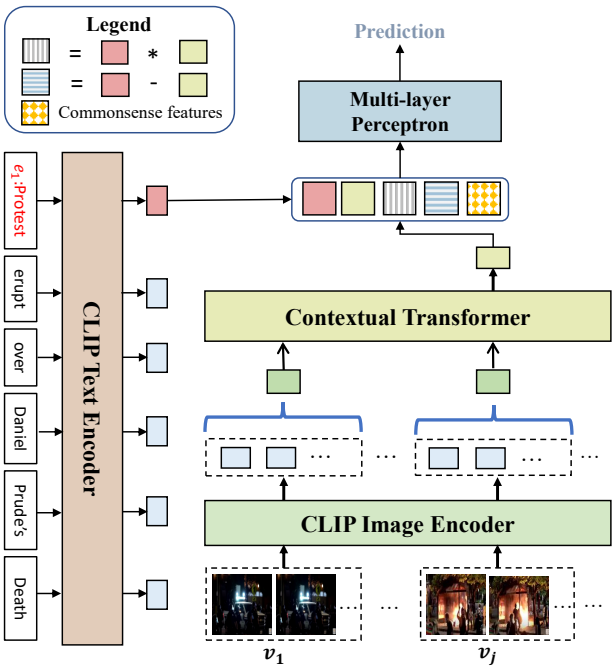

Figure 3: Overview of Proposed MASHER Model

We collect additional identical pairs by directly comparing all text events $\{e_i\}_{i=1}^{m}$ in the article T to all video events in $\{v_j\}_{j=1}^{n}$ in the paired video V using CLIP. This gives an aggregate of $\{e_{u_s} v_l^{u_s}\} \cup \{e_i v_j\}$ identical pairs.

In total, we collect 57,910 multimodal hierarchical event pairs and 39,0149 multimodal identical event pairs from the 100K video-article pairs training set. We evaluate the quality of these pseudo labels in Appendix C.2.

5.2 Training

Once we obtain the pseudo labels, we proceed to training using our model, **M**ultimodal **A**nalysis of **S**tructured **H**ierarchical **E**vents **R**elations (MASHER). The method is illustrated in Figure 3. Given a text event e_i and video event v_j having a label from the pseudo label set, r'_{ij}, we follow the procedure described below to train our model.

Input Representation and Feature Extraction. We represent text events as a word, e_i, in a sentence $se_i = [w_1, w_2, ...e_i, ..w_j, ...w_n]$. The video event, v_j, is a video clip in a video consisting of n video events, $\{v_j\}_{j=1}^{n}$. v_j is comprised of a stack of frames sampled uniformly at f_s frames per second, $v_j = \{F_y^j\}_{y=1}^{Z}$. We use CLIP to extract text event features, $ft_i = f'_t(se_i)$ as well as video event features, $fv_j = \frac{1}{Z}\sum_{y=1}^{Z} f_i(F_y^j)$, where f_i is CLIP's image encoder and f'_t is a modification of CLIP's text encoder to capture additional context, f_t (*c.f.* Appendix C.1).

Contextualizing Video Event Features So far, we have extracted video event features independent of other events in the video. This is a limitation since a video event such as building *destruction* needs to be contextualized with respect to other events in the video to ascertain whether it happened because of, say, a *"storm"* event or a *"earthquake"* event. As

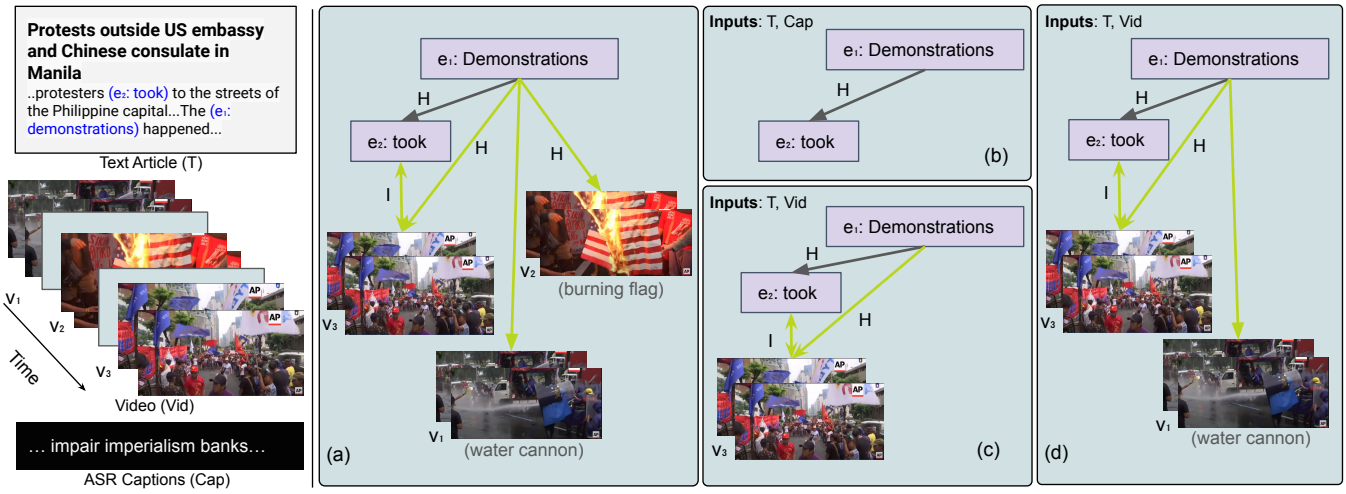

Figure 4: The left most column shows the inputs and the rest are outputs. (a): Ground Truth, (b) Text Base. (c) MM Base. (d) MASHER. The text-text event relations are derived using the method described in Section 5.1.

	Hierarchical			Identical			Avg F_1
	P	R	F_1	P	R	F_1	
Prior Base.	4.7/2.0	4.7/2.0	4.7/2.0	2.0/1.2	2.0/1.2	2.0/1.2	3.4/1.6
Text Base.	5.9/2.1	0.1/0.1	0.1/0.1	2.5/2.6	7.1/13.6	3.6/4.3	1.9/2.2
MM Base.	**35.7/28.0**	5.0/6.3	8.8/10.3	**8.8/7.6**	33.1/32.3	**13.9/12.4**	11.4/11.4
Video-LLaMA	4.82/2.21	13.15/13.28	7.06/3.79	1.76/1.03	4.08/4.25	2.46/1.65	4.76/2.72
MASHER	21.9/11.9	**22.1/18.8**	**22.0/14.6**	8.2/6.3	**44.5/39.0**	**13.9**/10.9	**18.0/12.8**

Table 3: Comparison with baseline models on the validation/test set.

such, we use Contextual Transformer (CT) to contextualize the event features with respect to other events in the video. CT is essentially a stack of multi-headed attention layers (Vaswani et al. 2017). All the video events' features from video V, $\{fv_j\}_{j=1}^{n}$, forms the input tokens to CT. The output is $cfv_j = CT(\{fv_j\}_{j=1}^{n})$.

Commonsense Features To aid learning the relationship between open domain text and video events, we incorporate commonsense knowledge from an external knowledge base, ConceptNet (Speer, Chin, and Havasi 2017). Inspired by (Wang et al. 2020), we extract events related by relations "HasSubevent", "HasFirstSubevent" and "HasLastSubevent" from ConceptNet as positive pairs and random events as negative pairs. We embed the event pairs using CLIP and then leverage the embeddings to train a feature extractor $CS(.,.)$, a MLP (Multi Layer Perceptron), using contrastive loss (*c.f.* Appendix C.3). Once trained, we freeze it and use it as a commonsense feature extractor while training MASHER, $cs_{ij} = CS(ft_i, cfv_j)$. Although while training MASHER, one of the events is from the visual modality, we are still able to use CS because CLIP's image embeddings and text embeddings lie in the same embedding space. We provide more analysis on this hypothesis in Appendix C.3.

Embeddings Interactions (EI) Following (Wang et al. 2020), we also add additional text event and video event feature interactions for a better representation. Specifically, (1) Subtraction of events' features (SF), $sf_{ij} = ft_i - cfv_j$ and (2) Hadamard product of events' features (MF), $mf_{ij} = ft_i * cfv_j$.

Multi Layer Perceptron (MLP) & Loss We concatenate the text event feature, ft_i, contextualized video event features, cfv_j, commonsense features, cs_{ij}, and embedding interactions, sf_{ij} and mf_{ij}, to form the input to a 2 layer MLP. The MLP is a 3-way classifier, outputting $p_{ij} \in \mathbb{R}^{1\times3}$: the probabilities for e_i and v_j being classified as "Hierarchical", "Identical" or "NoRel" (Not Related). We train the model using cross entropy loss between p_{ij} and the label, r'_{ij}.

5.3 Implementation Details

Notably, most text event and video event pairs are unrelated (94.52% in the train set). To mitigate label bias, we adjust the labels in the cross-entropy loss using the inverse ratio of their count in the train set, following Wang et al. (2021). Our best model uses a single layer of multi-headed attention in CT. We train our model for 15 epochs using a batch size of 1024 and a learning rate of 1e-5 on 4 NVIDIA Tesla v100 GPUs for a total training time of around 34 hours. In inference, we employ CLIP with MASHER as an ensemble to eliminate false positives for identical relations. This leverages CLIP's robust multimodal feature matching to confidently discard event pairs falsely predicted as identical. We provide ablation study of our model architecture in Section 6.2.

	Hierarchical			Identical			Avg F_1
	P	R	F_1	P	R	F_1	
MASHER Basic	12.1	**30.7**	17.4	7.2	42.8	12.4	14.9
+ CT	17.4	23.9	20.1	7.5	43.3	12.8	16.5
+ CS	13.4	29.1	18.4	7.6	44.9	13.0	15.7
+ EI	15.2	29.3	20.0	7.3	**45.1**	12.5	16.3
+ CT + CS + EI	**21.9**	22.1	**22.0**	**8.2**	44.5	**13.9**	**18.0**

Table 4: Ablation studies on components and features.

6 Experiments

Evaluation Metric We evaluate hierarchical and identical relations using Precision, Recall, and F1-score, following prior work in NLP (Hovy et al. 2013; Glavaš et al. 2014) and scene graph work in vision (Xu et al. 2017) (details in Appendix D.1). The F1-score effectively balances rewarding the model for correct relations and penalizing excessive incorrect predictions. We also report the macro average of F1-scores hierarchical and identical relations.

6.1 Baselines

Prior Baseline. (Prior Base.) We use a random weighted classifier that predicts a relation type $\in$ {"Hierarchical", "Identical", "NoRel"} for an event pair based on the prior distribution of the relation type in the annotated labels.

Text-only Baseline. (Text Base.) We construct a text-only baseline to study the limitations of text-only data in this task. To this end, we use ASR provided with video as a proxy for video events (*c.f.* Appendix D.2). Specifically, the proxy for video event v_j is the ASR found within the timestamps of v_j, denoted X_j. We extract events from X_j, $\{e'_{j_z}\}_{z=1}^{w}$ following Section 5.1. Next, we use the NLP model (Wang et al. 2021) to predict the relationship type, r_{ij_z} between a text event from the article, e_i, and proxy video events from ASR, e'_{j_z}. If any $r_{ij_z} \in$ {"Hierarchical", "Identical"}, we propagate r_{ij_z} from e_i and e'_{j_z} to e_i and v_j as e'_{j_z} is a proxy for v_j.

Multimodal Baseline. (MM Base.) We discussed a method to predict multimodal relations in Section 5.1, which used NLP and vision methods to produce pseudo labels. This is currently the best performance that NLP and vision can separately combine to give without a trained model. As such, we consider this pipeline as our multimodal baseline.

Video-LLaMA (Zhang, Li, and Bing 2023). It is a video based large language model which has demonstrated strong zero-shot results on multiple video language tasks. Consequently, we consider it as one of our baseline.

6.2 Results

Comparison against baselines The comparison between MASHER and above-mentioned baselines on the validation and test set are reported in Table 3. We also compare MASHER's and the baselines' performance on a dataset sample visually in Figure 4. From the table and the figure, we make following observations:

- For the most comprehensive metric, Avg F_1 score, MASHER outperforms all baselines with significant performance gains (*e.g.* 18.0 vs. 11.4 on the validation set).
- Text Base. performs quite poorly (Avg F_1 2.2). This is because a lot of visual events in the video are not mentioned in its ASR. This fact is also demonstrated by Figure 4.
- MM Base. performs better than Text Base. (Avg F_1 11.4 vs 2.2), stressing the importance of visual data to this task.
- MASHER achieves 4x recall over MM Base for hierarchical relations. It is because MM Base. relies on finding the subevent in text first before retrieving the matching video subevent (Section 5.1). This causes it to miss visual-only subevents while MASHER can discover those as it directly predicts multimodal relations. This fact is evident in Figure 4 – MM Base can only discover "took" (to streets) video subevent as it is mentioned in text as well. While MASHER can also detect "water cannon" visual-only subevent. We further validate this hypothesis by measuring recall on visual-only subevents. MASHER scores 15.92% while MM Base. scores 2.14%. This also explains MM Base's better precision, since it only predicts a few relations.
- Both MM Base and MASHER have low precision for identical relations due to their use of a video retrieval component that noisily predicts hierarchical relations as identical. For instance, a "meeting" text event is predicted identical to a video sub-event showing a handshake, because of the nature of the training dataset used for the video retrieval model. In contrast, our dataset annotates handshake to be a subevent of "meeting", as it's only a part-of meeting event.
- Video-LLaMA, a strong multimodal baseline, outperforms Prior and Text Base. Still, it's worse than MM Base, indicating specialized models (MM Base.) surpass generic vision-language model (Video-LLaMA) on this task.

Ablation Study and Analysis In Table 4, our ablation study examines the importance of various features in our model. Key findings include: 1) contextualizing video event features with CT enhances performance; 2) an external knowledge base (CS features) improves understanding of open-domain event-event relations; 3) employing different embedding interaction techniques (EI) enhances feature representation; and 4) the synergy of all three components (CT, CS, and EI) yields the best performance. Further ablations on inference time ensemble with CLIP and the number of layers in CT are explored in Appendix E. Additionally, Appendix F demonstrates MASHER's attention to relevant objects via Grad-CAM (Selvaraju et al. 2017) visualizations.

7 Conclusion

We proposed the novel task of extracting multimodal event hierarchies from multimedia content, a powerful way to understand, represent and reason about our world. Along with the task, we introduced `MultiHiEve` – a video-language dataset sourced from news domain and containing rich hierarchy of events. We proposed a weakly supervised model, MASHER, to predict these multimodal event relationships, achieving an improvement of around 3x on recall and 50% on F_1 score (hierarchical relations) over the strongest baseline.

We discuss the limitations and future directions of our work in Appendix H. We also discuss privacy and social bias concerns with respect to `MultiHiEve` in Appendix B.4.

Acknowledgements

This research 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 the official policies, either expressed or implied, of DARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Ayyubi, H. A.; Thomas, C.; Chum, L.; Lokesh, R.; Chen, L.; Niu, Y.; Lin, X.; Feng, X.; Koo, J.; Ray, S.; and Chang, S.-F. 2023. Beyond Grounding: Extracting Fine-Grained Event Hierarchies Across Modalities. arXiv:2206.07207.

Badgett, A.; and Huang, R. 2016. Extracting subevents via an effective two-phase approach. In *EMNLP*, 906–911.

Bosselut, A.; Rashkin, H.; Sap, M.; Malaviya, C.; Celikyilmaz, A.; and Choi, Y. 2019. Comet: Commonsense transformers for knowledge graph construction. In *ACL*.

Chaturvedi, S.; Peng, H.; and Roth, D. 2017. Story Comprehension for Predicting What Happens Next. In *EMNLP*, 1603–1614. Copenhagen, Denmark: ACL.

Chen, B.; Lin, X.; Thomas, C.; Li, M.; Yoshida, S.; Chum, L.; Ji, H.; and Chang, S.-F. 2021. Joint Multimedia Event Extraction from Video and Article. *arXiv preprint arXiv:2109.12776*.

Chen, D.; and Dolan, W. B. 2011. Collecting highly parallel data for paraphrase evaluation. In *ACL*, 190–200.

Daumé III, H.; and Marcu, D. 2006. Bayesian Query-Focused Summarization. In *ACL*, 305–312. Sydney, Australia: ACL.

Du, X.; Zhang, Z.; Li, S.; Yu, P.; Wang, H.; Lai, T.; Lin, X.; Wang, Z.; Liu, I.; Zhou, B.; et al. 2022. RESIN-11: Schema-guided event prediction for 11 newsworthy scenarios. In *NAACL: Human Language Technologies: System Demonstrations*, 54–63.

Fang, H.; Xiong, P.; Xu, L.; and Chen, Y. 2021. CLIP2Video: Mastering Video-Text Retrieval via Image CLIP. arXiv:2106.11097.

Gao, J.; Sun, C.; Yang, Z.; and Nevatia, R. 2017. Tall: Temporal activity localization via language query. In *ICCV*, 5267–5275.

Glavaš, G.; Šnajder, J.; Kordjamshidi, P.; and Moens, M.-F. 2014. HiEve: A corpus for extracting event hierarchies from news stories. In *Proceedings of 9th language resources and evaluation conference*, 3678–3683. ELRA.

GLAVAŠ, G.; and ŠNAJDER, J. 2015. Construction and evaluation of event graphs. *Natural Language Engineering*, 21(4): 607–652.

Hovy, E.; Mitamura, T.; Verdejo, F.; Araki, J.; and Philpot, A. 2013. Events are Not Simple: Identity, Non-Identity, and Quasi-Identity. In *Workshop on Events: Definition, Detection, Coreference, and Representation*, 21–28. Atlanta, Georgia: ACL.

Huang, S.; Strassel, S.; Mitchell, A.; and Song, Z. 2004. Automatic Content Extraction (ACE) Program - Task Definitions and Performance Measures. In *LREC*.

Ji, J.; Krishna, R.; Fei-Fei, L.; and Niebles, J. C. 2020. Action genome: Actions as compositions of spatio-temporal scene graphs. In *CVPR*, 10236–10247.

Krishna, R.; Hata, K.; Ren, F.; Fei-Fei, L.; and Carlos Niebles, J. 2017a. Dense-captioning events in videos. In *ICCV*, 706–715.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017b. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1): 32–73.

Li, M.; Xu, R.; Wang, S.; Zhou, L.; Lin, X.; Zhu, C.; Zeng, M.; Ji, H.; and Chang, S.-F. 2022. CLIP-Event: Connecting Text and Images with Event Structures. *arXiv preprint arXiv:2201.05078*.

Li, M.; Zareian, A.; Lin, Y.; Pan, X.; Whitehead, S.; Chen, B.; Wu, B.; Ji, H.; Chang, S.-F.; Voss, C.; Napierski, D.; and Freedman, M. 2020a. GAIA: A Fine-grained Multimedia Knowledge Extraction System. In *ACL: System Demonstrations*. ACL.

Li, M.; Zareian, A.; Zeng, Q.; Whitehead, S.; Lu, D.; Ji, H.; and Chang, S.-F. 2020b. Cross-media Structured Common Space for Multimedia Event Extraction. In *ACL*. Online: ACL.

Li, R.; Tapaswi, M.; Liao, R.; Jia, J.; Urtasun, R.; and Fidler, S. 2017. Situation Recognition with Graph Neural Networks. In *ICCV*, 4183–4192. IEEE Computer Society.

Liu, J.; Chen, Y.; Liu, K.; Bi, W.; and Liu, X. 2020. Event Extraction as Machine Reading Comprehension. In *EMNLP*, 1641–1651. Online: ACL.

Liu, J.; Chen, Y.; Liu, K.; and Zhao, J. 2019. Neural Cross-Lingual Event Detection with Minimal Parallel Resources. In *EMNLP-IJCNLP*. ACL.

Lu, A.; Lin, X.; Niu, Y.; and Chang, S.-F. 2023. In Defense of Structural Symbolic Representation for Video Event-Relation Prediction. In *CVPR-W*, 4940–4950.

Luo, H.; Ji, L.; Zhong, M.; Chen, Y.; Lei, W.; Duan, N.; and Li, T. 2021. CLIP4Clip: An Empirical Study of CLIP for End to End Video Clip Retrieval. *CoRR*, abs/2104.08860.

Mallya, A.; and Lazebnik, S. 2017. Recurrent Models for Situation Recognition. In *ICCV*, 455–463. IEEE Computer Society.

Miech, A.; Zhukov, D.; Alayrac, J.; Tapaswi, M.; Laptev, I.; and Sivic, J. 2019. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *ICCV*.

Nguyen, T. H.; Cho, K.; and Grishman, R. 2016. Joint Event Extraction via Recurrent Neural Networks. In *NAACL: Human Language Technologies*, 300–309.

O'Gorman, T.; Wright-Bettner, K.; and Palmer, M. 2016. Richer event description: Integrating event coreference with temporal, causal and bridging annotation. In *Proceedings*

of the 2nd Workshop on Computing News Storylines (CNS 2016), 47–56.

Park, J. S.; Bhagavatula, C.; Mottaghi, R.; Farhadi, A.; and Choi, Y. 2020. VisualCOMET: Reasoning about the Dynamic Context of a Still Image. In *ECCV*.

Pratt, S.; Yatskar, M.; Weihs, L.; Farhadi, A.; and Kembhavi, A. 2020. Grounded Situation Recognition. In *ECCV*, 314–332. Springer.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision.

Reddy, R. G.; Rui, X.; Li, M.; Lin, X.; Wen, H.; Cho, J.; Huang, L.; Bansal, M.; Sil, A.; Chang, S.-F.; et al. 2022. MuMuQA: Multimedia Multi-Hop News Question Answering via Cross-Media Knowledge Extraction and Grounding. In *AAAI*, volume 36.

Reimers, N.; Dehghani, N.; and Gurevych, I. 2016. Temporal Anchoring of Events for the TimeBank Corpus. 2195–2204.

Sadhu, A.; Gupta, T.; Yatskar, M.; Nevatia, R.; and Kembhavi, A. 2021. Visual Semantic Role Labeling for Video Understanding. In *CVPR*.

Sap, M.; Le Bras, R.; Allaway, E.; Bhagavatula, C.; Lourie, N.; Rashkin, H.; Roof, B.; Smith, N. A.; and Choi, Y. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *AAAI*, volume 33, 3027–3035.

Sha, L.; Qian, F.; Chang, B.; and Sui, Z. 2018. Jointly Extracting Event Triggers and Arguments by Dependency-Bridge RNN and Tensor-Based Argument Interaction. In McIlraith, S. A.; and Weinberger, K. Q., eds., *AAAI*, 5916–5923. AAAI Press.

Shen, J.; Zhang, Y.; Ji, H.; and Han, J. 2021. Corpus-based Open-Domain Event Type Induction. In *EMNLP*, 5427–5440. Online and Punta Cana, Dominican Republic: ACL.

Shou, M. Z.; Ghadiyaram, D.; Wang, W.; and Feiszli, M. 2021. Generic Event Boundary Detection: A Benchmark for Event Segmentation. *ICCV*.

Sigurdsson, G. A.; Varol, G.; Wang, X.; Farhadi, A.; Laptev, I.; and Gupta, A. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*, 510–526. Springer.

Speer, R.; Chin, J.; and Havasi, C. 2017. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. *AAAI*, 31(1).

Tong, M.; Wang, S.; Cao, Y.; Xu, B.; Li, J.; Hou, L.; and Chua, T.-S. 2020. Image enhanced event detection in news articles. In *AAAI*, volume 34, 9040–9047.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Vulić, I.; Ponzetto, S. P.; and Glavaš, G. 2019. Multilingual and cross-lingual graded lexical entailment. In *ACL*, 4963–4974.

Wang, H.; Chen, M.; Zhang, H.; and Roth, D. 2020. Joint Constrained Learning for Event-Event Relation Extraction. In *EMNLP*, 696–706. Online: ACL.

Wang, H.; Zhang, H.; Chen, M.; and Roth, D. 2021. Learning Constraints and Descriptive Segmentation for Subevent Detection. In *EMNLP*, 5216–5226. Online and Punta Cana, Dominican Republic: ACL.

Wen, H.; Lin, Y.; Lai, T.; Pan, X.; Li, S.; Lin, X.; Zhou, B.; Li, M.; Wang, H.; Zhang, H.; Yu, X.; Dong, A.; Wang, Z.; Fung, Y.; Mishra, P.; Lyu, Q.; Surís, D.; Chen, B.; Brown, S. W.; Palmer, M.; Callison-Burch, C.; Vondrick, C.; Han, J.; Roth, D.; Chang, S.-F.; and Ji, H. 2021. RESIN: A Dockerized Schema-Guided Cross-document Cross-lingual Cross-media Information Extraction and Event Tracking System. In *NAACL: Human Language Technologies: Demonstrations*, 133–143. Online: ACL.

Xu, D.; Zhu, Y.; Choy, C. B.; and Fei-Fei, L. 2017. Scene graph generation by iterative message passing. In *CVPR*, 5410–5419.

Xu, J.; Mei, T.; Yao, T.; and Rui, Y. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, 5288–5296.

Yao, W.; Dai, Z.; Ramaswamy, M.; Min, B.; and Huang, R. 2020. Weakly supervised subevent knowledge acquisition. In *EMNLP*.

Yatskar, M.; Zettlemoyer, L.; and Farhadi, A. 2016. Situation Recognition: Visual Semantic Role Labeling for Image Understanding. *CVPR*, 5534–5542.

Zellers, R.; Lu, X.; Hessel, J.; Yu, Y.; Park, J. S.; Cao, J.; Farhadi, A.; and Choi, Y. 2021. MERLOT: Multimodal Neural Script Knowledge Models. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *NeurIPS*, volume 34, 23634–23651. Curran Associates, Inc.

Zhang, H.; Li, X.; and Bing, L. 2023. Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding. arXiv:2306.02858.

Zhang, T.; Whitehead, S.; Zhang, H.; Li, H.; Ellis, J. G.; Huang, L.; Liu, W.; Ji, H.; and Chang, S. 2017. Improving Event Extraction via Multimodal Integration. In *ACM MM*.