Leveraging Text Representation and Face-head Tracking for Long-form Multimodal Semantic Relation Understanding

Raksha Ramesh*
rn2486@columbia.edu
Columbia University, Graphen, Inc.
New York, NY, USA

Vishal Anand*†
va2361@columbia.edu
Columbia University, Microsoft Corp.
Redmond, WA, USA

Zifan Chen
zc2628@columbia.edu
Columbia University, Graphen, Inc.
New York, NY, USA

Yifei Dong
yd2616@columbia.edu
Columbia University, Graphen, Inc.
New York, NY, USA

Yun Chen
yunchen@graphen.ai
Graphen, Inc.
New York, NY, USA

Ching-Yung Lin
c.lin@columbia.edu
Columbia University, Graphen, Inc.
New York, NY, USA

ABSTRACT

In the intricate problem of understanding long-form multi-modal inputs, few key-aspects in scene-understanding and dialogue-and-discourse are often overlooked. In this paper, we investigate two such key-aspects for better semantic and relational understanding - (i). head-object-tracking in addition to usual face-tracking, and (ii). fusing scene-to-text representation with external common-sense knowledge-base for effective mapping to sub-tasks of interest. The usage of head-tracking especially helps with enriching sparse entity mapping to inter-entity conversation interactions. These methods are guided by natural language supervision on visual models, and perform well for interaction and sentiment understanding tasks.

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**; **Discourse, dialogue and pragmatics**; **Information extraction**; **Tracking**; **Knowledge representation and reasoning**; **Scene understanding**.

KEYWORDS

natural language processing, object tracking, dialogue and discourse, language models, slot filling, intent detection, knowledge graphs, speaker diarization

ACM Reference Format:
Raksha Ramesh, Vishal Anand, Zifan Chen, Yifei Dong, Yun Chen, and Ching-Yung Lin. 2022. Leveraging Text Representation and Face-head Tracking for Long-form Multimodal Semantic Relation Understanding. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3503161.3551610

*Both authors contributed equally to this research.
†Corresponding author

MM '22, October 10–14, 2022, Lisboa, Portugal

ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3551610

1 INTRODUCTION

Long-form multi-modal reasoning is garnering increased attention [6] to better deduce evolving human-conversation understanding across long duration [1, 2, 14, 20]. Recent works extend BERT based architectures to process multi-modal inputs to learn joint representations[11, 16, 17]. With large-scale self-supervised pre-training, these benefit downstream vision-language tasks such as video question answering, visual common sense reasoning using few-shot learning. However, most of these focus on visual concepts and their spatial relations [3, 9]. The transferable capability of these models is seldom tested on tasks involving semantic relations understanding in long-form datasets, especially when dialogue-discourse are limited by entity-tracking - which happens since people don't often face cameras in long-form multimodal datasets.

2 APPROACH

We investigate how to leverage large-scale language-image pre-training for our objectives of long-form video understanding. We discover that our framework of zero-shot transfer on image-text models combined with multi-body tracking for entity localization can generate accurate scene representations for fine-grained interaction and sentiment prediction. We discuss the key components of our methodology in this section - first we identify and track key persons throughout the scenes, then perform zero-shot prediction with enriched prompts from scene-graphs, finally, we organize and visualize scenes as a knowledge graph for multi-hop question answering. Figure 1 summarizes the above approach.

2.1 Face Recognition and Multi-Body Tracking

Identifying scene-entities is a key aspect of long-form multimodal understanding. Face-embeddings are computed from five facial landmark points: eye-left, eye-right, nose, mouth-left, & mouth-right using additive angular margin-loss[8] to perform face-recognition.

There were significant challenges in mapping entities' names with their figures on frame - especially when they are facing away from the camera, or when less than half of their faces are visible, a very common scenario in movies. Therefore, substantial interactions between entities were not mapped and had to be discarded.

We develop a novel approach for multi-entity tracking. Movies are split into multiple shots[15] and Mask R-CNN object detector[19] extracts bounding boxes and masks for each person entities, per

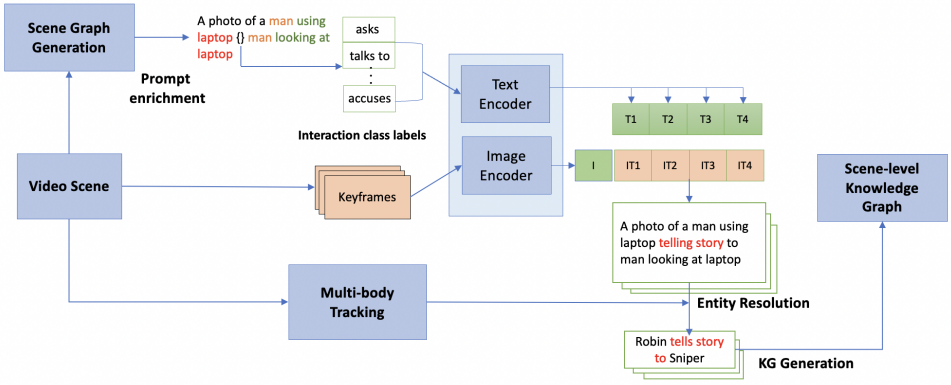

Figure 1: Schematic algorithmic flow

frame. Since a person's movement on screen is continuous within a shot, we infer each person's moving bounding boxes. For every overlap between face's bounding boxes and a person object's mask, we add the face-name to that person's voting pool and perform majority voting. (Algorithm1 performed every 6 frames per shot). If a person's face is detected even once, they are tracked throughout the shot. Figure2 illustrates a person whose back is to the camera is still identified (unachievable with conventional face detectors).

Algorithm 1: Multi-Entity Tracking Algorithm

```
Data: frame, frameID, entityList
Result: updated entityList
boundingBoxes, maskList = detectron2.predict(frame);
faceBoxes, nameList = Arcface.predict(frame);
entityCenters = getLastCentersForAllEntities(entityList);
for box,mask in boundingBoxes,maskList do
    center = getCenter(box);
    correctName = nameList[overlapFaceIndex(mask, faceBoxes)];
    correctEntity = getClosestEntity(center, entityCenters);
    if correctEntity == None then
        newEntity = createNewEntity();
        newEntity.addLastCenter(center, frameID);
        if correctName != None then
            newEntity.addName(correctName);
        end
        entityList.append(newEntity);
    else
        correctEntity.addLastCenter(center, frameID);
        if correctName != None then
            correctEntity.addName(correctName);
        end
    end
end
```

2.2 Model

We leverage CLIP[13] model pre-trained on 400k image-text pairs with a contrastive-loss and maps the image-text pairs to a common embedding space useful for downstream vision-language tasks. It

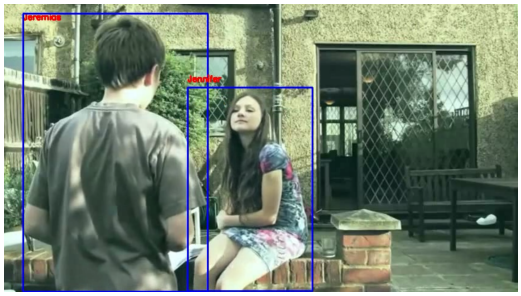

Figure 2: Body-tracking: facing away, and towards camera

comprises of vision and text transformers that serve as the video-frame and prompt encoders in our setting respectively.

(1) Generating Image-text pairs: We sample all I-frames as keyframes from every shot to depict the scene. We design a simple prompt to pair with the image: "A photo of a person {label}" where label is the class categories of 116 interactions eg: accusing, asking provided in the dataset. Similarly the prompt for emotion prediction is "A photo of a person feeling {label}". We also add object attributes and spatial relations in the prompt to help ground the entities and steer the model towards the interaction between the entities. We discuss prompt engineering [5] in Section 4.

(2) Entity mapping and localization: After obtaining the predictions for the semantic attributes, we use the multi-body tracking framework to obtain the source and target entity between whom the interaction occurs in every key-frame. Furthermore, for some shots with missing entities we use a range of shots(+/- 1) to infer the entity within the shot i.e., if an entity occurs in the previous and next shot they are most likely to occur in the current shot too.

3 EXPERIMENTS

The model is pre-trained on images paired with natural language text descriptions rather than class labels in traditional image classification tasks [7]. We attempt to enrich the prompt by adding object attributes and spatial relations to describe the images.

3.1 Scene-graph generation

The purpose of scene graphs generation is two-fold: (i). enrich the knowledge graph for holistic scene understanding, and (ii). infuse object relations to ground the scene-entities for prompt generation.

We adopt the Scene graph generation method[18] due to its performance on Visual Genome[10], which extends Mask R-CNN[12] to define relationship prediction. We sample frames from scenes where two or more entities are co-located to generate scene-graphs.

For relationship prediction, we capture triplets per frame (entity1-relation-entity2) - including entity-details (person, woman, etc.) and related attributes (hand, mouth, etc.). This step captures relationships between people-entities and surrounding objects in the frame. We then map the person entities in the triplet with the related entity name from the multi-body tracking using Cartesian distances between bounding boxes. Figure 3 shows object attributes & relations along-with entity-name where 'man' is mapped to 'Sniper'.

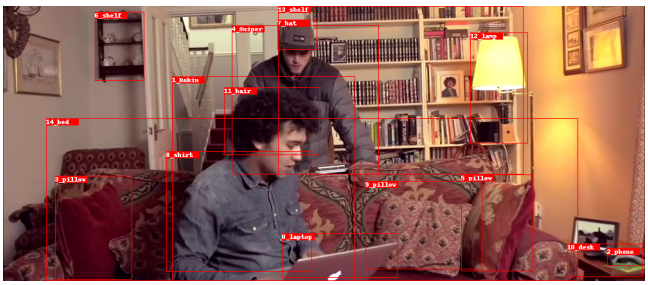

Figure 3: Scene-graph - Objects+attributes, Scene 17: Sophie

3.2 Prompt Enrichment and Entity Grounding

The traditional prompt only includes the names of two entities and the interaction between them. In order to enhance prompts in this experiment, we add a triplet generated from scene-graph to the corresponding entity as a feature, which makes our prompt template: "triplet1 - interaction - triplet2" where interaction is one of 116 types from the dataset. [1].

When choosing ideal triplets, the first step is to filter out duplicate triplets to ensure unambiguous entity references. We find that triplets from the scene graphs contain a total of 28 relations ("wearing", "behind", "holding", etc.). Among these relations, since "has" and "and" contain the most ambiguous contextual information, we delete all triplets including them. Also, the less frequently a relation occurs, the more unique the contextual meaning it expresses. Therefore, we give different priorities to the remaining 26 relations that the relation with lower frequency has higher priority. In this way, triplet with the highest priority relation will be selected for the prompt. For scene 17 in sophie in Figure 3 and Figure 4, the triplets selected for "Robin" and "Sniper" are "using laptop" and "looking at laptop", respectively. In this case, the prompt output is ["Robin using laptop {interaction_label} Sniper looking at laptop"]

[1] https://www-nlpir.nist.gov/projects/trecvid/dvu/dvu.development.dataset/

4 EVALUATION

4.1 Knowledge Graph Visualization

Each scene within a movie is represented by a knowledge graph that aggregates and organizes information from these key sources:

- Localized Entities through body-tracking
- Scene graphs for visual+spatial relations between entities
- Fine-grained inference of interactions, sentiments, location and emotions through image-language model

Figure 4 is the knowledge graph for Movie Sophie for Scene 17. The knowledge graphs is used to query and retrieve information effectively as discussed in later sections.

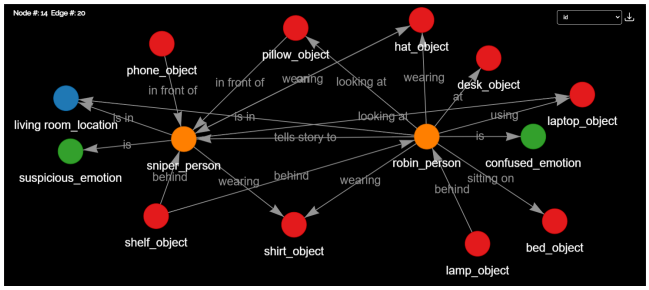

Figure 4: Scene-graph with relationship, interaction, emotion, location-prediction, and entity mapping; Scene 17: Sophie

4.2 Scene-Level Query and Answering

Six types of scene-level queries are used to evaluate our knowledge understanding. The question queries and their answers are available from NIST [2] and we follow their query guidance [3]. The S1-S6 queries are described below along with sample questions, along with our detailed approach. The performances are shared in Table1 and 2.

For question S1 and S6, semantic similarity matrix quantifies closeness between interactions or sentiments. Each word is embedded using fastText [4] and we store cosine similarity for each pair of words in a 2-dimensional matrix.

- **S1**: Given a full and inclusive set of interactions unique to a specific scene in the movie, find which scene it is.
  ***e.g. Question**: Which unique scene contains the following interactions: asks, talks to, thanks?*
  Approach: Calculate weighted similarity scores foreach scene's interactions-set and prompt's interactions-set according to interaction matrix as described in this section. Return the scene with the arg-max value.
- **S2**: Given a scene in movie and a set of interaction:other-person pair, find the correct person.
  ***e.g. Question**: Which person in scene 7 has the following interactions: SourcePerson:Kelly's husband talks to, SourcePerson:Kelly's husband Comforts?*
  Approach: (1) Generate triplets <candidate, interaction, otherEntity> according to the question's prompt for each possible answer-candidate. (2) Return arg-max candidate with

[2] https://www-nlpir.nist.gov/projects/trecvid/dvu/dvu.development.dataset
[3] https://sites.google.com/view/dvuchallenge2022/home/datasets-queries

Movie	S1-MRR	S2-Acc	S2-MRR	S3-Acc	S4-Acc	S5-Acc	S6-Acc
The_Big_Something	7.3	20.0	30.5	18.2	9.1	9.1	36.9
honey	21.4	38.3	63.7	23.3	22.5	21.7	42.1
shooters	12.1	28.0	55.8	34.2	39.5	30.0	30.0
Huckleberry_Finn	12.9	29.7	48.4	18.3	18.3	27.1	37.3
sophie	8.3	34.4	52.5	21.7	19.6	51.1	31.8
time_expired	9.5	39.1	44.5	19.6	19.6	24.0	33.8
spiritual_contact	12.3	33.3	50.6	32.1	22.6	26.3	72.7
Valkaama	14.4	17.4	36.2	15.4	15.4	22.2	50.0
Nuclear_Family	30.0	53.9	67.3	21.4	14.3	12.5	46.7
SuperHero	18.3	22.2	61.1	11.1	11.1	30.0	50.0
Average	14.7	31.6	51.1	21.5	19.2	25.4	43.1

Table 1: Training Evaluation (percentage) on Section 4.2 Tasks

Movie	S1	S2	S3	S4	S5	S6
Chained_for_life	1.7	50.0	2.5	33.3	0.0	16.7
Liberty_kid	8.3	77.8	25.0	25.0	10.0	0.0
Like_me	31.6	62.5	50.0	75.0	20.0	0.0
Little_rock	7.50	62.5	0.0	0.0	0.0	33.3
Losing_ground	25.0	38.3	0.0	0.0	30.0	16.7
Calloused_hands	3.3	100.0	50.0	50.0	30.0	16.7
Average	12.9	65.2	25.0	30.6	15.0	13.9

Table 2: Test Evaluation (percentage) - Section 4.2

maximum hits between triplets set of that scene's knowledge-graph and triplet-sets of the candidate.

- **S3**: Given a scene in a movie and an interaction between A and B, pick the correct next interaction between A and B in a specific scene from a set of possible interactions.
  e.g. Question*: In scene 3, Jeremias talks to John, what is the immediate next / following interaction between Jeremias and John in scene 5?* ***Choices****: compliments, pays, teases, leave together, demands, asks.*
  Approach: Owing to absence of temporal scale in the knowledge graph, we iterate over the constructed knowledge-graph interactions between the two given entities from the scene in prompt and return the first hit from the choice list.
- **S4**: Given a movie scene and an interaction between A and B, pick the correct prior interaction between A and B in a specific scene from a set of possible interactions.
  e.g. Question*: In scene 15, John asks Jeremias, what is the immediate prior / previous interaction between John and Jeremias in scene 14?* ***Choices****: reminds, waves at, touches, serves, asks.*
  Approach: Same method leveraged as that of S4.
- **S5**: Given a text-description for a movie scene, select the scene that best matches the description.
  e.g. Question*: Which scene best matches the following description:John calls Helena, Jeremias, Chris, and Jennifer for dinner?* ***Choices****: 1, 2, 5, 7, 13, 19, 21, 27, 29, 36*
  Approach: Generate scene descriptions by combining relationship predictions from Scene-Graph generation and inference of interactions, sentiments, location and emotions through image-language mode in a natural language-like form. Each pair in a scene is described as: [In (sentiment_tag), (emotion_tag) person_1 (relationship_tag) (interaction_tag) (emotion_tag) person_2 (relationship_tag) in (location_tag)]. Returns scene with the best description-sentence similarity.
- **S6**: Given a scene in a movie and a set of possible sentiments pick the correct sentiment.
  e.g. Question*: In scene 1, what is the correct sentiment label?* ***Choices****: escape, baby shower, frightening experience, recruiting, greeting, charity event.*
  Approach: We get top 10 mostly likely sentiment predictions for the scene asked in the prompt from the model. For each (predicted sentiment, choice sentiment) pair, we calculate the cosine similarity value stored in matrix mentioned at the beginning of this section. And we pick the choice sentiment in the pair that has the highest similarity score as the answer. If two pairs have the same similarity score, the pair has higher ranked predicted sentiment is preferred.

5 DISCUSSION

We conducted an analysis as annotators to understand which modalities are more likely to influence our understanding of interactions. Can a particular interaction be inferred through visual domain alone eg: bullies, hugs, kisses etc., or require dialog understanding as they are more nuanced like "talks", "explains", "asks" which can't be differentiated in images. By examining scenes from the training movies, we discovered about 43% of interactions can only be inferred if both dialog and visual aspects of the scene are considered together. Due to this limitation, since our model mainly relies on image frames for inference, it can benefit from dialog features. We leave the exploration of incorporating dialog encoder for further fine-tuning the image-text representation for future work.

6 CONCLUSION

The experiments confirm our hypothesis of how important novel face-head-tracking is for multi-modal semantic understanding and knowledge-graph generation. The accuracy and mean-reciprocal-ranks see a substantial increase with the new features added to our language-grounded-vision models. The added object-properties as features help steer the semantic and relational properties between entities through semantic contextual search.

REFERENCES

[1] Vishal Anand, Raksha Ramesh, Boshen Jin, Ziyin Wang, Xiaoxiao Lei, and Ching-Yung Lin. 2021. *MultiModal Language Modelling on Knowledge Graphs for Deep Video Understanding.* Association for Computing Machinery, New York, NY, USA, 4868–4872. https://doi.org/10.1145/3474085.3479220

[2] Vishal Anand, Raksha Ramesh, Ziyin Wang, Yijing Feng, Jiana Feng, Wenfeng Lyu, Tianle Zhu, Serena Yuan, and Ching-Yung Lin. 2020. *Story Semantic Relationships from Multimodal Cognitions.* Association for Computing Machinery, New York, NY, USA, 4650–4654. https://doi.org/10.1145/3394171.3416305

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV).*

[4] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics* 5 (2017), 135–146. https://doi.org/10.1162/tacl_a_00051

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901. https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[6] Keith Curtis, George Awad, Shahzad Rajput, and Ian Soboroff. 2020. HLVU: A New Challenge to Test Deep Understanding of Movies the Way Humans do. In *Proceedings of the 2020 International Conference on Multimedia Retrieval.* 355–361.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09.*

[8] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 4685–4694. https://doi.org/10.1109/CVPR.2019.00482

[9] Drew A. Hudson and Christopher D. Manning. 2019. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*

[10] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *Int. J. Comput. Vision* 123, 1 (may 2017), 32–73. https://doi.org/10.1007/s11263-016-0981-7

[11] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/file/c74d97b01eae257e44aa9d5bade97baf-Paper.pdf

[12] Francisco Massa and Ross Girshick. 2018. maskrcnn-benchmark: Fast, modular reference implementation of Instance Segmentation and Object Detection algorithms in PyTorch. https://github.com/facebookresearch/maskrcnn-benchmark. Accessed: July 25, 2022.

[13] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 8748–8763. https://proceedings.mlr.press/v139/radford21a.html

[14] Raksha Ramesh, Vishal Anand, Ziyin Wang, Tianle Zhu, Wenfeng Lyu, Serena Yuan, and Ching-Yung Lin. 2020. Kinetics and Scene Features for Intent Detection. In *Companion Publication of the 2020 International Conference on Multimodal Interaction* (Virtual Event, Netherlands) *(ICMI '20 Companion).* Association for Computing Machinery, New York, NY, USA, 135–139. https://doi.org/10.1145/3395035.3425641

[15] Robert Sklar. 2002. *Film: An international history of the medium.* Prentice Hall, 526.

[16] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy 0002, and Cordelia Schmid. 2019. VideoBERT: A Joint Model for Video and Language Representation Learning. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019.* IEEE, 7463–7472. https://doi.org/10.1109/ICCV.2019.00756

[17] Hao Tan and Mohit Bansal. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. 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).* Association for Computational Linguistics, Hong Kong, China, 5100–5111. https://doi.org/10.18653/v1/D19-1514

[18] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. 2020. Unbiased Scene Graph Generation From Biased Training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*

[19] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. 2019. Detectron2. https://github.com/facebookresearch/detectron2.

[20] Fan Yu, DanDan Wang, Beibei Zhang, and Tongwei Ren. 2020. *Deep Relationship Analysis in Video with Multimodal Feature Fusion.* Association for Computing Machinery, New York, NY, USA, 4640–4644. https://doi.org/10.1145/3394171.3416303