# SEMANTIC ASSOCIATION NETWORK FOR VIDEO CORPUS MOMENT RETRIEVAL

*Dahyun Kim** *Sunjae Yoon** *Ji Woo Hong* *Chang D. Yoo*

Korea Advanced Institute of Science and Technology (KAIST)
{dahyun.kim, dbstjswo505, jiwoohong93, cd_yoo}@kaist.ac.kr

## ABSTRACT

This paper considers Semantic Association Network (SAN) for Video Corpus Moment Retrieval (VCMR) which localizes temporal moment that best corresponds to the given text query in a corpus of videos. Collaborations among common semantics from multi-modal inputs are essential for effectively understanding video together with subtitle and text query. For this collaboration, SAN associates common semantics within the same modality (by Intra Semantic Association) and across different modalities (by Inter Semantic Association) with dedicated module referred to as Modality Semantic Association (MSA). SAN surpasses existing state-of-the-art performance on the TVR and DiDeMo benchmark datasets. Extensive ablation studies and qualitative analyses show the effectiveness of the proposed model.

***Index Terms***— Video Corpus Moment Retrieval, Video Moment Retrieval, Temporal Moment Localization, Localizing Moment, Vision Language Task

## 1. INTRODUCTION

Understanding visual semantics along with natural language is receiving increased attention. This is exemplified in the following tasks: video captioning [1], video question answering [2, 3], and video moment retrieval (VMR) [4, 5]. Here, we study Video Corpus Moment Retrieval (VCMR) [6, 7, 8] to localize temporal moment in corpus of videos that best corresponds to the given text query. To be specific, given a corpus of videos, subtitles for better understanding of videos, and text query describing unique scene, the goal is to infer video and temporal moment corresponding to the query. VCMR is an extension of VMR, which localize temporal moment corresponding to text query in a single video.

Gao *et al.* [9] firstly suggests VMR, which finds moments with a sentence describing action. For this VMR, there have been methods for improving multi-modal interaction. Xu *et al.* [10] performs query guided interaction between query and video. Yuan *et al.* [11] directly predicts the temporal moment related to the sentence in the entire video sequence by multi-modal co-attention mechanism. Recent efforts to generalize the format of VMR to perform on video corpus, VCMR [6] is considered. Linjie Li *et al.* [7] considers local to global context of multi-modal inputs using hierarchical transformer structure.

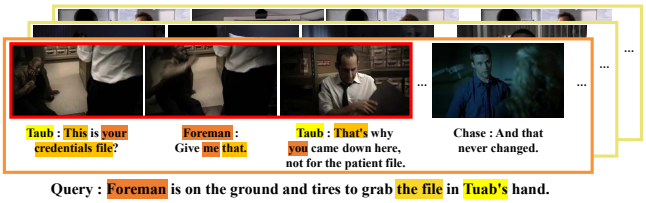


**Fig. 1**: Illustration of Video Corpus Moment Retrieval task. Ground-truth video is shown in orange box, and ground-truth temporal moment is shown in red box. Each word having same semantics is highlighted in same color.

Previous works for VCMR [6, 7, 8] have received considerable amount of attention. Collaboration among common semantics from multi-modal inputs are essential to effectively understand video based on text query. Assuming that the video includes a scene involving throwing a ball and matched subtitle of the scene is "A: Does he still practice pitching? B: Yes, he is training hard to win the game.", people can easily interact the throwing scene, "practice pitching.", and "training" with the common semantics. However, in the existing models that deal with video and natural language, there is difficulty in interacting common semantics as stated above.

For the interaction among the common semantics, this paper proposes Semantic Association Network (SAN) with Modality Semantic Association (MSA) which associates words and video frames that share common semantics within same modality and across different modalities. It consists of Intra Semantic Association (Intra-SA) and Inter Semantic Association (Inter-SA). Intra-SA associates features sharing a common semantic within same modality. It generates bipartite graphs within same modalities (between neighboring clips and between neighboring subtitles). In subtitle, "prac-

*Both authors have equally contributed.

This work was supported by Institute for Information communications Technology Planning Evaluation(IITP) grant funded by the Korea government(MSIT) (No. 2019-0-01396, Development of framework for analyzing, detecting, mitigating of bias in AI model and training data and 2021-0-01381, Development of Causal AI through Video Understanding)

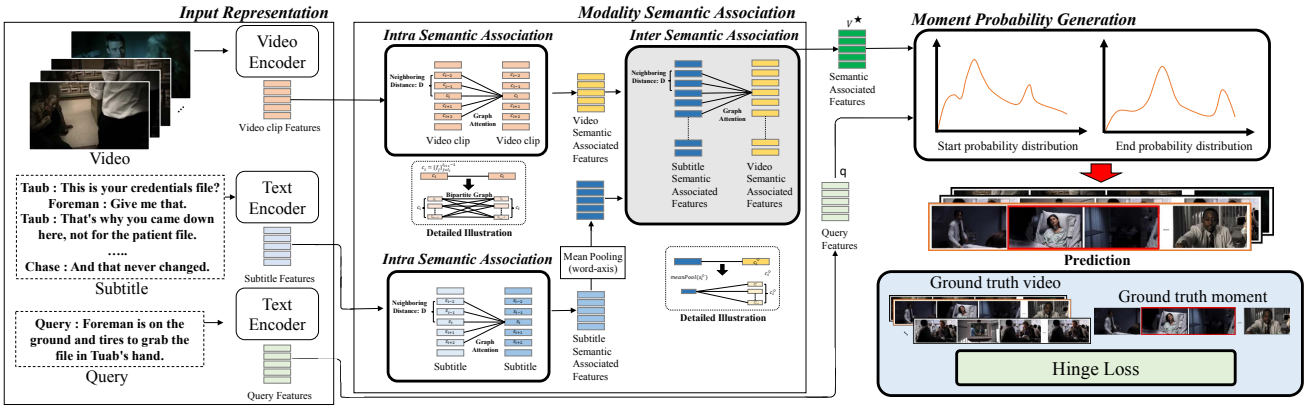


**Fig. 2**: Illustration of our proposed Semantic Association Network(SAN) for video corpus moment retrieval. SAN is composed of Modality Semantic Association (MSA) and Moment Probability Generation (MPG).

ticing pitching" and "training" are associated and in video, frames of throwing the ball are associated. Inter-SA associates features sharing common semantics across different modalities. It generates graphs across different modalities (between neighboring clip and subtitle). Unlike Intra-SA, in Inter-SA, subtitle including "practicing pitching" and frames of throwing the ball are associated across different modalities.

Our contributions can be summarized as follows. (1) We propose SAN that can associate common semantics within same modality and across different modalities. (2) We demonstrate the capability of SAN on TVR and DiDeMo benchmark datasets. (3) We show the efficiency of each module by utilizing ablation studies and qualitative analysis.

## 2. METHOD

### 2.1. Input Representation

Semantic Association Network (SAN) takes video $\mathcal{V} = \{\mathcal{F}\}_{i=1}^{L_v}$, subtitle $\mathcal{S} = \{\mathcal{S}_i\}_i^{L_s}$ ($\mathcal{S}_i = \{\mathcal{W}_{ij}^s\}_{j=1}^{L_{s_i}}$), and query $\mathcal{Q} = \{\mathcal{W}_i^q\}_{i=1}^{L_q}$, where $L_v$, $L_s$, $L_{s_i}$ and $L_q$ are respectively the number of frames in a video, subtitles in a video, words in a subtitle, and words in query. And SAN predicts temporal moment corresponding to the given query within video corpus. As video encoder, we use ResNet-152 [12] pretrained by ImageNet [13] and I3D [14] pretrained by Kinetics-600 [15] and the features are concatenated. We generate video features from the video encoder and define $V = \{f_i\}_{i=1}^{L_v} \in \mathbb{R}^{L_v \times d}$. Also, we define video clip $c_i = \{f_j\}_{j=I_i}^{I_{i+1}-1}$ which is a group of frame features matched to a single subtitle, where $I_i$ is the index of the starting frame of clip. Here, subtitle is dialogue from character in each video. Since matched frames have relations with the corresponding subtitle, multi-frames are grouped in terms of subtitle and these group is referred to as clips. As text encoder, we use RoBERTa [16] and define group of subtitles $S = \{s_i\}_{i=1}^{L_s}$ and single subtitle $s_i = \{w_{ij}\}_{j=1}^{L_{s_i}} \in \mathbb{R}^{L_{s_i} \times d}$ composed of word features. For the query, we define query $q = \{w_i^q\}_{i=1}^{L_q} \in \mathbb{R}^{L_q \times d}$ composed of word features. Each final modality is produced after positional encoding [17] and layer normalization [18].

### 2.2. Modality Semantic Association

Modality Semantic Association (MSA) associates common semantics within same modality and across different modalities. For the MSA, we first define Semantic Association (SA) using Graph Attention [19]. Graph Attention consists of nodes and edges with adjacency matrix and applies multi-head attention [17] among linked nodes. Let, the SA inputs two groups of nodes $N_x = \{n_{xi}\}_{i=1}^{L_x} \in \mathbb{R}^{L_x \times d}$ and $N_y = \{n_{yi}\}_{i=1}^{L_y} \in \mathbb{R}^{L_y \times d}$ and adjacency matrix $\mathcal{E}_{N_x,N_y} \in \mathbb{R}^{(L_x+L_y)\times(L_x+L_y)}$, which summarizes the linkages of node groups as:

$$N_x^+ = \mathrm{SA}(N_x, N_y, \mathcal{E}_{N_x,N_y}) \in \mathbb{R}^{L_x \times d}, \quad (1)$$

where subscript $+$ represents attended node. SA conducting Graph Attention with nodes of $N_x$, $N_y$ and edge $\mathcal{E}_{N_x}$ outputs attended node group $N_x^+$. For getting $N_x^+$ for Graph Attention, SA slices $N_x^+$ from all attended nodes. Founded on Semantic Association, Modality Semantic Association is composed of Intra- and Inter- Semantic Association, which is covered in following sections.

#### 2.2.1. *Intra Semantic Association*

Intra Semantic Association (Intra-SA) associates semantics within same modality. Intra-SA generates bipartite graph which is fully connected graph between two node groups. Here, two node groups are clips or subtitles within same modality. Intra-SA associates semantics of neighbor node groups (including identical node). When distance of two node groups is less than neighboring distance $D$, two node

groups are connected as neighboring node groups. We introduce neighboring distance $D$, which is how far node groups would be related as shown in Figure 2. According to D, the node group of $i$ generates bipartite graph with node groups of $i-D$ to $i+D$. Proposed Intra-SA performs semantic association on consecutive neighbors within neighboring distance D from node group $c_i$ and $s_i$. Intra Semantic Associated clips $c_i^\diamond$ and subtitle $s_i^\diamond$ are generated by averaging of all associated features as:

$$c_i^\diamond = [\sum_{j=-D}^{D} \mathrm{SA}(c_i, c_{i+j}, \mathcal{E}_{c_i,c_{i+j}})]/(2D+1), \quad (2)$$

$$s_i^\diamond = [\sum_{j=-D}^{D} \mathrm{SA}(s_i, s_{i+j}, \mathcal{E}_{s_i,s_{i+j}})]/(2D+1), \quad (3)$$

where subscript $\diamond$ represents Inter Semantic Associated node. Intra Semantic Associated clip is consist of associated frames as $c_i^\diamond = \{f_j^\diamond\}_{j=I_i}^{I_{i+1}-1}$. All of the clips $c_i^\diamond$ are grouped into $V^\diamond = \{c_j^\diamond\}_{j=1}^{L_s}$. Intra-SA features are used for associating semantics across different modalities in Inter-SA.

#### *2.2.2. Inter semantic Association*

Inter Semantic Association (Inter-SA) is for associating semantics across different modalities. Inter-SA generates graph between frame features and subtitle features. Subtitles are associated with frame features of matched clip and neighbor clips. Subtitle feature of $s_i^\diamond$ is linked with frames of clips from $c_{i-D}^\diamond$ to $c_{i+D}^\diamond$ as Inter-SA of Figure 2. We yield each subtitle feature by mean pooling on $s_i^\diamond$ over word axis and perform Inter-SA as:

$$S_{mean} = \{\mathrm{meanPool}(s_j^\diamond)\}_{j=1}^{L_s} \in \mathbb{R}^{L_s \times d}, \quad (4)$$

$$V^* = \mathrm{SA}(V^\diamond, S_{mean}, \mathcal{E}_{V^\diamond, S_{mean}}), \quad (5)$$

where $\mathrm{meanPool}()$ represents mean pooling operation and subscript $*$ represents final Intra- and Inter- Semantic Associated features.

### 2.3. Moment Probability Generation

Moment Probability Generation (MPG) generates start / end timestamps probability of target moments by calculating frame-level video-query matching score, like [6, 7]. The Intra- and Inter- Semantic Associated features $V^*$ and word-level mean pooled query feature $Q$ are introduced to produce frame-level matching score $Score_m$ and video-level matching score $Score_v$ represented by the largest score of $Score_m$ as:

$$Q = \mathrm{meanPool}(q) \in \mathbb{R}^d, \quad (6)$$

$$Score_m = V^\star \times Q \in \mathbb{R}^{L_v}, \quad (7)$$

$$Score_v = \mathrm{Max}(Score_m) \in \mathbb{R}, \quad (8)$$

where $\mathrm{Max}()$ represents max function. Each start / end probability distribution $St_p$ / $Ed_p$ is generated through two independent 1D convolution layers and Softmax, and moment-level score matrix $M_s$ in single video is obtained through matrix multiplication between the two probability distributions. Video corpus moment-level score matrix $M_v$ is obtained by multiplication between $M_s$ and $Score_v$:

$$St_p = \mathrm{Softmax}(\mathrm{Conv_s}(Score_m)) \in \mathbb{R}^{L_v}, \quad (9)$$

$$Ed_p = \mathrm{Softmax}(\mathrm{Conv_e}(Score_m)) \in \mathbb{R}^{L_v}, \quad (10)$$

$$M_s = {St_p}^T \times Ed_p \in \mathbb{R}^{L_v \times L_v}, \quad (11)$$

$$M_v = M_s \times Score_v \in \mathbb{R}^{L_v \times L_v}, \quad (12)$$

where $\mathrm{Conv_x}()$ is the independent 1D convolution layer and $\mathrm{Softmax}()$ is Softmax for probability distribution. From $M_v$ of all videos, moments of top-k score are selected as candidates to be retrieved in video corpus, where x-axis and y-axis of $M_v$ are the index of start / end frames of the moment.

### 2.4. Video Level and Moment Level Learning

SAN is trained under video-level learning for video retrieval and moment-level learning for moment retrieval. In the video-level learning, we introduce the video-level matching score $Score_v$ in the equation (10) and $Score_v$ can be obtained by positive and negative video-query pairs. We applied hinge loss $L_v$ using positive and negative pairs as :

$$\mathcal{L}_v = \max[0, \Delta_t - Score_v^p + Score_v^n], \quad (13)$$

where subscripts $p$ and $n$ means positive and negative matching scores. In moment-level learning, we utilize probabilities of start / end timestamps for moment-level loss $\mathcal{L}_m$ as:

$$\mathcal{L}_m = CE(St_{gt}, St_p) + CE(Ed_{gt}, Ed_p), \quad (14)$$

$$\mathcal{L} = \alpha\mathcal{L}_v + \beta\mathcal{L}_m, \quad (15)$$

where $gt$ means ground-truth, $CE$ is Cross entropy loss, $\mathcal{L}$ is total loss and $\alpha$ and $\beta$ are hyperparameters.

## 3. EXPERIMENTS

### 3.1. Datasets

**TV show Retrieval(TVR)** [6] includes 109K queries for 21.8K videos which is 60 to 90 seconds from 6 TV video shows of 3 genres: sitcom, medical, crime. Subtitles comprise dialogues from characters. Query is unique matched within a video and the average length of moment matched with query is 9.1 seconds. TVR is divided into 80% train, 10% val, 5% test-public, and 5% test-private.

**DiDeMo** [4] consists of 10.6K videos from Flickr and 41.2K queries matched with unique moment. The start / end timestamps of ground-truth moments are aligned at five second interval. The length of each video is 25 to 30 seconds. DiDeMo is divided 80% train, 10% val, and 10% test.

**Table 1**: Experiment results on TVR. ⋆: with pre-training.

| Method | R@1 | R@10 | R@100 |
|---|---|---|---|
| XML[6] | 3.25 | 12.49 | 29.51 |
| HERO[7] | 2.98 | 10.65 | 18.25 |
| ReLoCLNet[8] | **4.15** | 14.06 | 32.42 |
| **SAN** | 3.64 | **15.32** | **34.73** |
| HERO⋆[7] | 6.21 | 19.34 | 36.66 |
| **SAN⋆** | **7.03** | **20.24** | **38.97** |

**Table 2**: Experiment results on DiDeMo

| Method | R@1 | R@10 | R@100 |
|---|---|---|---|
| XML[6] | 1.59 | 6.71 | 25.44 |
| HERO[7] | 2.14 | 11.43 | 36.09 |
| **SAN** | **2.76** | **13.56** | **41.23** |

### 3.2. Quantitative Results

To compare the performance fairly, we train SAN with / without pretraining. With pretrainig, we pretrain SAN under HowTo100M[7] and a large-scale TV dataset [2, 20, 6, 21] as [7]. Without pretrainig, we train SAN from scratch. We use HERO[7] as baseline and results are reported on tIoU>0.7.

Table1 shows experiment results on TVR with / without pretraining. SAN gets state-of-the-art performance in R@10, R@100 except R@1 without pretraining and get state-of-the-art performance to all metrics with pretraining. The results show that Semantic Association within same modality and across different modalities helps to interpret video.

Table 2 shows experiments result on DiDeMo. SAN achieve state-of-the-art performance in all metrics. As DiDeMo does not have subtitles, we apply semantic association only with video and it shows that limited semantic association also has positive effect on video interpretation.

### 3.3. Ablation Studies

We experiment with variants of SAN to validate effectiveness of our module. The performance of Table 3 is based on SAN with pretraining. Second block of Table 3 shows ablation studies on MSA, Intra-SA, and Inter-SA. By these results, we can see that both semantic association within and across different modalities are effective for understanding semantics, and association across different modalities is more effective.

Third block of Table 3 shows ablation analysis on various neighboring distance. The result indicates that two neighboring clips and subtitle are most correlated in building effective association. We can see that when neighboring distance is too small, association between relevant semantics is not sufficient and when neighboring distance is too long, association between irrelevant semantics makes semantic confusion.

**Table 3**: Ablation study on Model Variants

| Method | Acc.(R@1) | Variance |
|---|---|---|
| **SAN** | **7.01** | **0** |
| SAN w/o MSA | 6.19 | -0.84 |
| SAN w/o Intra SA | 6.59 | -0.44 |
| SAN w/o Inter SA | 6.32 | -0.71 |
| Neighboring distance D = 1 | 6.89 | -0.14 |
| Neighboring distance **D = 2** | **7.01** | **0** |
| Neighboring distance D = 3 | 6.81 | -0.22 |
| Neighboring distance D = 4 | 6.69 | -0.34 |

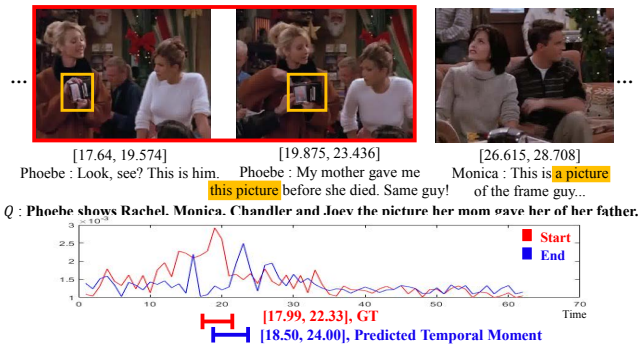


**Fig. 3**: Visualization of SAN in the validation split.

### 3.4. Qualitative Results

Figure 3 shows results of SAN in the validation split. It shows frames and subtitles closed to ground-truth moments. Frames of predicted moment are highlighted by red box. Highlighted words in subtitles have high association with frames, and boxes in frames emphasize the parts that have the meaning of the words. Red / blue graphs show start / end probability distribution predicted by given query, and red / blue bars are ground-truth / predicted temporal moment. Subtitles contain "picture" that has high association with predicted moment frames. The frames have "picture" mentioned by subtitles. These association validates that our proposed Semantic Association is implemented as we intend and has positive effect on performance gain.

## 4. CONCLUSION

In this paper, we propose Semantic Association Network (SAN) for Video Corpus Moment Retrieval (VCMR). For interaction among common semantics from multi-modal inputs, SAN adopts Modality Semantic Association (MAS) that associates common semantics within same modality and across different modality. Our experiment results show that SAN achieve state-of-the-art performance on TVR and DiDeMo benchmark datasets. Ablation studies and qualitative analysis validate efficiency of our proposed module.

## 5. REFERENCES


[1] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko, "Sequence to sequence-video to text," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4534–4542.

[2] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg, "Tvqa: Localized, compositional video question answering," *arXiv preprint arXiv:1809.01696*, 2018.

[3] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler, "Movieqa: Understanding stories in movies through question-answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4631–4640.

[4] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell, "Localizing moments in video with natural language," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5803–5812.

[5] Da Zhang, Xiyang Dai, Xin Wang, Yuan-Fang Wang, and Larry S Davis, "Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1247–1257.

[6] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal, "Tvr: A large-scale dataset for video-subtitle moment retrieval-supplementary file," .

[7] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu, "Hero: Hierarchical encoder for video+language omni-representation pre-training," in *EMNLP*, 2020.

[8] Hao Zhang, Aixin Sun, Wei Jing, Guoshun Nan, Liangli Zhen, Joey Tianyi Zhou, and Rick Siow Mong Goh, "Video corpus moment retrieval with contrastive learning," *arXiv preprint arXiv:2105.06247*, 2021.

[9] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia, "Tall: Temporal activity localization via language query," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5267–5275.

[10] Huijuan Xu, Kun He, Bryan A Plummer, Leonid Sigal, Stan Sclaroff, and Kate Saenko, "Multilevel language and vision integration for text-to-clip retrieval," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, pp. 9062–9069.

[11] Yitian Yuan, Tao Mei, and Wenwu Zhu, "To find where you talk: Temporal sentence localization in video with attention based location regression," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, pp. 9159–9166.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[14] Joao Carreira and Andrew Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[15] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al., "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[16] 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.

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[18] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[19] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio, "Graph attention networks," *arXiv preprint arXiv:1710.10903*, 2017.

[20] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal, "Tvqa+: Spatio-temporal grounding for video question answering," *arXiv preprint arXiv:1904.11574*, 2019.

[21] Jingzhou Liu, Wenhu Chen, Yu Cheng, Zhe Gan, Licheng Yu, Yiming Yang, and Jingjing Liu, "Violin: A large-scale dataset for video-and-language inference," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10900–10910.