# Cross-modal Contrastive Learning with Asymmetric Co-attention Network for Video Moment Retrieval

Love Panta[12]*, Prashant Shrestha[1]*, Brabeem Sapkota[1], Amrita Bhattarai[1],
Suresh Manandhar[2], and Anand Kumar Sah[1]
[1]IOE, Pulchowk Campus, [2]Wiseyak Solutions Pvt. Ltd.
{075bei016.love, 075bei024.prashant, 075bei011.brabeem, 075bei006.amrita,
anand.sah}@pcampus.edu.np, suresh.manandhar@wiseyak.com

## Abstract

*Video moment retrieval is a challenging task requiring fine-grained interactions between video and text modalities. Recent work in image-text pretraining has demonstrated that most existing pretrained models suffer from information asymmetry due to the difference in length between visual and textual sequences. We question whether the same problem also exists in the video-text domain with an auxiliary need to preserve both spatial and temporal information. Thus, we evaluate a recently proposed solution involving the addition of an asymmetric co-attention network for video grounding tasks. Additionally, we incorporate momentum contrastive loss for robust, discriminative representation learning in both modalities. We note that the integration of these supplementary modules yields better performance compared to state-of-the-art models on the TACoS dataset and comparable results on ActivityNet Captions, all while utilizing significantly fewer parameters with respect to baseline.*

## 1. Introduction

Recent trends in machine learning have shown a growing interest in multimodal learning, specifically in vision language tasks such as visual question answering, image-text retrieval, video grounding and so on. Video moment retrieval, also known as video grounding, aims to align a video segment semantically with a given sentence query. Numerous approaches have been proposed to address video grounding, but their results were unsatisfactory due to limitations in capturing both spatial and temporal information [8]. Transformer-based methods have dominated the vision-language landscape in recent years and have also been effectively used for video grounding [2, 25, 28, 29, 32]. One advantage of using transformers over other neural network architectures is their ability to model long sequences without losing context [22] and little need for engineering fusion approaches for effective multimodal interaction.

We opted for a single-stream transformer backbone due to their efficiency and little need for engineering on cross-modal interactions compared to dual-stream architectures. However, the effectiveness of single-stream multimodal architectures is limited by the imbalance in the length of visual and textual query in image-text pretraining which increases the learning time and reduces the performance of the model [9]. To alleviate the problem, Li *et al.* [9] proposes an asymmetric co-attention block at the beginning of the network and outputs the visual-aware text features. This asymmetry still exists when we move to the video grounding task as the video feature sequences are much longer than the accompanying textual feature sequences. Thus, we adapt this approach to a transformer-based architecture proposed by MSAT [32] for video moment retrieval.

In terms of training objectives, recent methods in image text pretraining [5, 9, 10] have extensively adapted contrastive loss together with transformers for effective multimodal interaction. However, few works have been proposed for video grounding [2, 29]. Inspired by approaches in image text pretraining, we employ additional Video Text Contrastive (VTC) loss to our architecture. Experiments on MSAT [32] have also shown the effectiveness of the loss. Additionally, we observe that VTC loss allows the decoupled attention paradigm of MSAT to be dropped without affecting performance, greatly reducing the model parameters. Moreover, MSAT architecture introduces the novel multi-stage aggregated module(MSA) on the top of their transformer module to capture the stage-specific information which is also integrated into our model.

In summary, our key contributions are three-fold:

- We evaluate the effectiveness of a recently proposed solution to the information asymmetry problem inspired by image-text pretraining on the video ground-

*Both authors contributed equally to this work

ing task.

- We employ the momentum contrastive loss for more robust feature learning across both visual and text modalities, thereby achieving better or comparable results on both datasets surpassing various state-of-arts.

- We conduct experiments to assess the effectiveness of various modules on both our architecture and the baseline, drawing conclusions about our superior performance.

## 2. Related Work

The video grounding task itself is a challenging task that requires a high-level understanding of semantic relations between video and text features [8, 27, 31]. Many approaches have been proposed so far which can be broadly categorized into proposal-based and proposal-free methods. Proposal based methods [2, 29, 32, 33] typically use a two-stage framework. This involves either utilizing a predefined set of candidate moments or generating such candidates. These candidates are then ranked by the model based on their relevance to the provided sentence. In contrast proposal-free methods [11, 14, 28, 30] aim to directly predict moment boundaries, eliminating the need for explicit proposals. These methods either directly estimate temporal boundaries or adopt a span-based approach to assign probabilities to each video index, indicating its potential as a starting or ending point for the moment.

A key engineering concern within proposal-based methods lies in the representation of proposals. Pooling-based strategies often lack the needed discrimination for precise localization. To address this, Zhang *et al.* [32] introduces a distinctive stage-specific representation method for enhancing proposal representation. From an architectural perspective, transformers have shown their effectiveness in various multimodal learning tasks, including video grounding [2, 28, 29, 31, 32]. The attention-based mechanism in transformers enables the model to efficiently capture multimodal relations, along with spatio-temporal context information, facilitating improved alignment between the text and video [20, 26].

Contrastive learning aims to learn representations that maximize the similarity scores between positive pairs thereby making the encoder more discriminative. In the following papers [1, 4], the authors introduce contrastive learning in the visual domain. Inspired by MoCo [4], ALBEF [10] introduces the multi-modal contrastive loss in image-text pretraining which shows the effectiveness of the approach on various downstream tasks. For video grounding and video corpus moment retrieval tasks, the following papers [15, 29] propose multi-modal contrastive learning which maximizes the mutual information between two modalities to learn the more robust representations in an unsupervised way. However, these approaches require large mini-batches which is computationally inefficient and results in low accuracy.

## 3. Method

Our architecture mainly consists of three main components i.e., visual language constrastive loss, transformer backbone with asymmetric co-attention and multistage-aggregated module from [32]. These components are trained end to end after freezing the feature extractors to predict the target moment given text query.

### 3.1. Feature extraction

Given a video $X$, it is divided into sequence of frames as $V = \{x_1, x_2, ..., x_f\}$. Then, pre-trained C3D [21] is employed to extract the spatio-temporal features on the bulk of frame sequences. The resulting feature vectors undergo mean pooling to standardize the segment count of features, ensuring a consistent value regardless of the video's duration. Finally, the video is represented as $V = \{v_i\}_{i=1}^{N}$ where, $N$ is the length of the video clip sequence chosen. Each video has its corresponding annotation $\{T_q, ts, te\}$, where, $T_q$ is the textual query, $t_s$ and $t_e$ are the true starting and ending index of the video moment that corresponds to the language sentence.

For the textual query, GloVe [16] embeddings are used to produce a sequence of word-level feature vectors. Thus, a sentence in a dataset is expressed as $T_q = \{w_i\}_{i=1}^{M}$, where, $w_i$ is GloVe embedding of $i^{th}$ word in a sentence $T_q$ of length $M$.

### 3.2. Visual Language Transformer Block

Following the work of mPLUG [9], we propose to use a similar transformer architecture for better vision-language understanding in the context of video moment retrieval. They utilize cross-modal skip-connections, enabling fusion at disparate abstraction levels, and creating inter-layer shortcuts to capture semantic richness in language compared to vision. In the video domain, employing the same architecture is highly likely to be beneficial, given the necessity of capturing both spatial and temporal context information from the attended language features.

The main transformer backbone consists of sequentially arranged $N$ cross-model skip-connected blocks. Each block is formed by the $S$ repeated asymmetric co-attention layers followed by a single connected self-attention block as shown in Fig. 1, where $S$ is the stride layer value. The input for the transformer backbone comprises visual and text features obtained from the feature encoder in the contrastive learning block. Subsequently, we iterate the input sequence through the asymmetric co-attention block to acquire visual aware text features. The features thus obtained

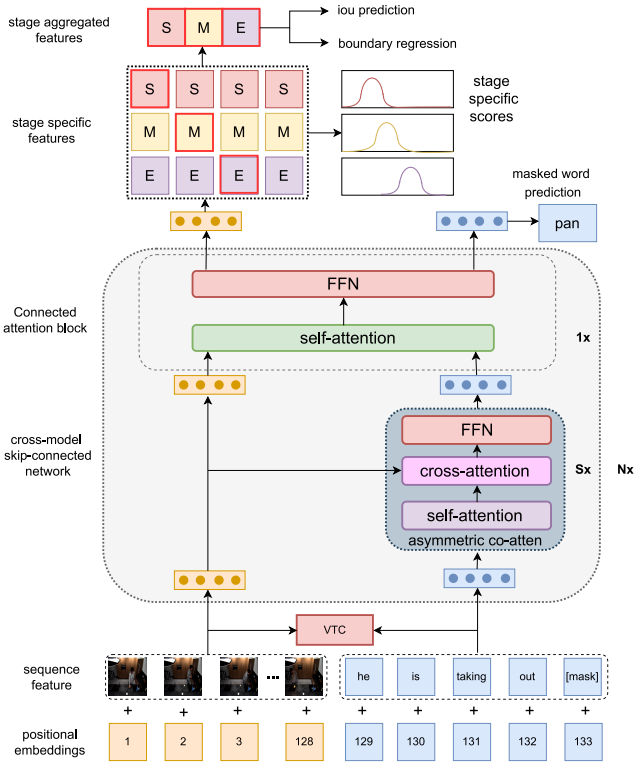


Figure 1. Cross-model Skip-connected Network

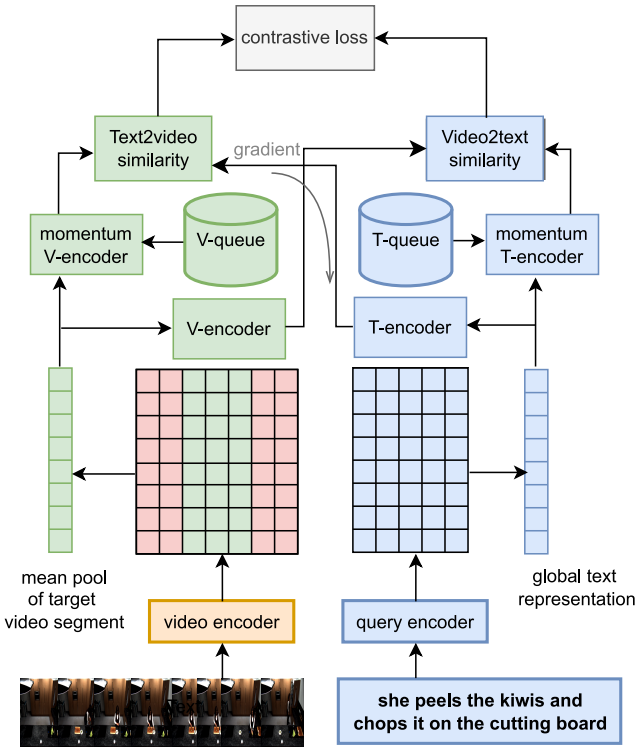


Figure 2. Video-text momentum contrastive learning

from both domains are concatenated and passed into the single self-attention block referred to as Connected Attention Block(CAB). The whole process is repeated multiple times until the semantically rich text features are given for masked language prediction task [3] whereas on the vision side, it is passed into the multi-stage aggregated module [32].

### 3.3. Proposal Generation and Ranking

We follow the approach taken by [32] and pass the encoded visual features to a multi-stage aggregated module. The candidate proposals are generated using a 2D temporal map [33]. The multi-stage aggregated module provides temporal stage-specific representations for each video clip i.e. beginning, middle and ending stages. Then, for each moment candidate, the stage-specific features are sampled and concatenated to produce proposal representations. These stage-specific representations are more discriminative than pooling-based representations. These proposals are boundary-regressed and ranked to produce the final output. More details can be found on the paper [33].

We combine the losses used by [32] together with the Contrastive loss explained in the next section to train the model.

### 3.4. Contrastive Loss

Similar to Image Text Contrastive Learning loss (ITC) adopted by [10], Video Text Contrastive Learning loss (VTC) is used before passing the features into the transformer backbone as shown in Fig. 2. The loss tries to align the features of positive pairs. The mean pooled target video segment and global text representations of the same training examples form the positive pairs while all other examples form the negative pairs.

The visual and text features obtained from the respective feature extractor are projected to the same latent dimension and then the positional embedding is added to preserve the sequence information which forms the input for the contrastive encoder block. Based on MoCo [4] and ALBEF [10], a similarity function, $s = g_v\left(\boldsymbol{v}_{\text{avg}}\right)^{\top} g_w\left(\boldsymbol{w}_{\text{avg}}\right)$ is learnt such that video segments that align with the text query have a higher similarity score. Here, $g_v$ and $g_w$ are linear transformations that map the average pooled representation to a normalized latent space. For the video modality, we first separate the features into ground truth segments and non-ground truth segments. For each batch, the average pooled latent feature of video ground truth portions together with the average pooled latent feature of the aligned textual query form the positive pairs. Two queues are maintained to store the $M$ most recent video-text representations from the momentum unimodal encoders to generate the negative

pairs. The normalized features from momentum encoders are denoted as $g'_v\left(\boldsymbol{v}'_{\text{avg}}\right)$ and $g'_w\left(\boldsymbol{w}'_{\text{avg}}\right)$. The two-way similarity is defined by $s(V,T) = g_v\left(\boldsymbol{v}_{\text{avg}}\right)^\top g'_w\left(\boldsymbol{w}'_{\text{avg}}\right)$ and $s(T,V) = g_w\left(\boldsymbol{w}_{\text{avg}}\right)^\top g'_v\left(\boldsymbol{v}'_{\text{avg}}\right)$.

For each video segment and text, the softmax-normalized video-to-text and text-to-video similarity is calculated as:

$$\begin{cases} p_m^{\text{v2t}}(V) = \dfrac{\exp\left(s\left(V,T_m\right)/\tau\right)}{\sum_{m=1}^{M}\exp\left(s\left(V,T_m\right)/\tau\right)} \\ p_m^{\text{t2v}}(T) = \dfrac{\exp\left(s\left(T,V_m\right)/\tau\right)}{\sum_{m=1}^{M}\exp\left(s\left(T,V_m\right)/\tau\right)} \end{cases} \tag{1}$$

where $\tau$ is a temperature parameter which is learnable. If $\boldsymbol{y}^{\text{v2t}}(V)$ and $\boldsymbol{y}^{\text{t2v}}(T)$ denote the ground-truth one-hot similarity, where negative pairs have a value of 0 and positive pairs have a value of 1. The video-text contrastive loss is calculated as the cross entropy between $p$ and $y$:

$$\mathcal{L}_{\text{vtc}} = \frac{1}{2}\mathbb{E}_{(V,T)\sim D}\Big[\text{H}\left(\boldsymbol{y}^{\text{v2t}}(V),\boldsymbol{p}^{\text{v2t}}(V)\right) + \text{H}\left(\boldsymbol{y}^{\text{t2v}}(T),\boldsymbol{p}^{\text{t2v}}(T)\right)\Big] \tag{2}$$

ALBEF address the weak correlation of positive pairs and the correlation of negative pairs by learning from pseudo-targets generated by the momentum model. This momentum distillation approach is also used for the VTC loss.

# 4. Experiments

## 4.1. Datasets

### 4.1.1 TACoS

The TACoS dataset was introduced in Regneri *et al.* [18] and is a popular benchmark dataset used in the literature. It contains a total of 127 videos with an average duration of about 7 minutes. The train, val and test splits use the standard split of 10,146, 4,589, and 4,083 video-segment - query pairs respectively. The videos in the TACoS dataset were built on top of the "MPII Cooking Composite Activities" video corpus (Rohrbach *et al.*, 2012, MPII Composites) [19], containing videos of various cooking activities, e.g., cutting kiwi, cleaning chopping board, etc. The actions performed are in the same kitchen but the lighting conditions do vary. TACoS is considered a challenging dataset because of the level of detail in some of the queries in the dataset.

### 4.1.2 ActivityNet Captions

The ActivityNet Captions was introduced by Krishna *et al.* [7] for dense video captioning tasks. Here, there are 20k videos and 100k annotations in total, with an average of 4.82 temporally localized sentences per video. The average duration of the videos is 2 minutes. The validation subset "val 1" is used as the validation set while the subset "val 2" is used as the test set. As in [32] and [33], for the train, val and test set of 37,417, 17,505 and 17,031 video-segment - query pairs are used respectively. The ActivityNet Captions are characteristic in that the range of domains of the videos is vast. Thus, the results on the ActivityNet Captions may be more relevant in the context of the use of the models in general settings.

One characteristic feature of the ActivityNet dataset is the decreased level of detail of the annotations. Compared to TACoS, the ground truth moments are longer while the total video duration is shorter. Thus, there is a high likelihood of a random prediction to overlap with the ground truth moment.

## 4.2. Implementation Details

AdamW optimizer [13] is used with a learning rate of $5e^{-5}$ to train the model for both datasets. We set the number of transformer blocks to be 3 and the stride length for each asymmetric co-attention block to be 2. For the VTC module, a simple linear layer is used as an encoder block for both video and query features to project it into a common latent space for calculating similarity scores. For negative sample mining, a queue size of 50,000 is used and the momentum encoders are set to have the distillation weight, $\alpha$ of 0.3, momentum parameter of 0.995 for updating the respective momentum encoders, and temperature coefficient ($\tau$) of 0.1. All hyperparameters for the proposal generation, the multi-stage aggregated module and the feature extraction are set according to [32].

## 4.3. Evaluation Metric

For the sake of comparison, R@$n$, IoU@$m$ is used for evaluation. It refers to the percentage of text queries, for which IoU of at least one of the $n$ temporal moment predictions with the ground truth exceeds $m$. For example if one of the predictions for the query, $q_i$ results in an IoU with the ground truth of over $m$, then $r\left(n,m,q_i\right) = 1$. Otherwise, $r\left(n,m,q_i\right) = 0$. Thus, R@$n$, IoU@$m$ is calculated as:

$$\text{R@}n, \text{IoU@}m = \frac{1}{N_q}\sum_{i=1}^{N_q} r\left(n,m,q_i\right) \tag{3}$$

We use $n \in \{1,5\}$ for both the datasets. For the ActivityNet Captions, $m \in \{0.5,0.7\}$ and for the TACoS dataset, $m \in \{0.3,0.5\}$.

## 4.4. Comparison with other methods

We compare our approach with previous state-of-the-art methods and the results are shown in Table 1 and Table 2.

Table 1. Comparisons with SOTA on TACoS dataset based on C3D features

| Method | R@1, IoU@0.3↑ | R@1, IoU@0.5↑ | R@5, IoU@0.3↑ | R@5, IoU@0.5↑ |
|---|---|---|---|---|
| 2D-TAN [33] | 37.29 | 25.32 | 57.81 | 45.04 |
| FIAN [17] | 33.87 | 28.58 | 47.76 | 39.16 |
| CSMGAN [12] | 33.9 | 27.09 | 53.98 | 41.22 |
| IVG [15] | 38.84 | 29.07 | - | - |
| DPIN [23] | 46.74 | 32.92 | 62.16 | 50.26 |
| SMIN [24] | 48.01 | 35.24 | 65.18 | 53.36 |
| MSAT [32] | 48.79 | 37.57 | 67.63 | 57.91 |
| STCM-Net [6] | 49.04 | 35.59 | **70.13** | 57.69 |
| OURS | **49.77** | **37.99** | 68.31 | **58.31** |

Table 2. Comparisons with SOTA on ActivityNet Captions dataset based on C3D features

| Method | R@1, IoU@0.5↑ | R@1, IoU@0.7↑ | R@5, IoU@0.5↑ | R@5, IoU@0.7↑ |
|---|---|---|---|---|
| 2D-TAN [33] | 44.51 | 26.54 | 77.13 | 61.96 |
| FIAN [17] | 47.9 | 29.81 | 77.64 | 59.66 |
| CSMGAN [12] | **49.11** | 27.09 | 77.43 | 59.63 |
| IVG [15] | 43.84 | 27.10 | - | - |
| DPIN [23] | 47.27 | 28.31 | 77.45 | 60.03 |
| SMIN [24] | 48.46 | 30.34 | **81.16** | 62.11 |
| MSAT [32] | 48.02 | **31.78** | 78.02 | 63.18 |
| STCM-Net [6] | 46.23 | 29.04 | 78.43 | 63.46 |
| OURS | 47.73 | 31.21 | 78.06 | **63.63** |

Our method performs strongly against the various baselines. In TACoS, our method outperforms the baselines in almost all the metrics whereas in ActivityNet captions, it achieves comparable performances. Compared with MSAT [32], our method achieves 0.98 point improvement for R@1,IoU@0.3 and provides small improvements over all other metrics in TACoS with considerably fewer parameters as seen in Table 5. The introduction of an asymmetric attention block reduces the overall model parameters by removing the need for decoupled attention weights. Our model requires 30 per cent fewer parameters as compared to MSAT for training the model.

2D-TAN [33] uses a 2D temporal map of features to represent the moment candidates. It then uses 2D convolution operations to consider the temporal relation between adjacent video moments for discriminative localization. DPIN [23] instead uses two interacting branches for frame level and candidate level representations to make predictions consistent with both query semantics and moment boundaries. Furthermore, FIAN [17] applies the iterative cross-modal attention network to generate visual aware sentence representations and vice versa, whereas, CSMGAN [12] utilizes the joint cross and self-modal graph attention network to capture the detailed high-level interactions.

SMIN [24] considers the boundary and content level moment representations for coarse to fine-grained cross-model interactions. To obtain additional information from textual modality, STCM-Net [6] further proposes the time concept mining network to extract time-related information from the sentence query. IVG [15] applies causal inference to remove spurious correlation between video and query features. Additionally, they apply intermodal contrastive learning to align video and text features, and intramodal video-video contrastive learning to improve visual representation. MSAT [32] uses decoupled attention weights inside multimodal transformer backbone and stage-specific representations for proposals.

In a nutshell, even though MSAT outperformed past approaches in most metrics, our modification achieved even better overall performance with significantly fewer parameters.

## 4.5. Ablation Study

Table 3. Ablation study on ActivityNet Captions

| Method | R@1, IoU@0.5↑ | R@1, IoU@0.7↑ | R@5, IoU@0.5↑ | R@5, IoU@0.7↑ |
|---|---|---|---|---|
| De-VLTrans+MSA [32] | **48.02** | **31.78** | 78.02 | 63.18 |
| ACB+De-VLTrans+MSA | 47.99 | 30.86 | 77.51 | 62.37 |
| CAB+MSA | 46.62 | 29.54 | 77.22 | 60.91 |
| ACB+CAB+MSA | 47.38 | 30.32 | 77.18 | 61.20 |
| VTC+De-VLTrans+MSA | 47.98 | 31.40 | 77.61 | 62.66 |
| **VTC+ACB+CAB+MSA** | 47.73 | 31.21 | **78.06** | **63.63** |

We perform ablation on different components of the architecture to verify the effectiveness of our modifications to the original MSAT architecture. Table 3 and 4 report the scores obtained with the different variants. While the initial design employs six distinct transformer encoder blocks, leading to higher parameter counts, our experimental setup involves the integration of only three primary transformer blocks whether in the form of De-VLTrans or CAB. This adjustment is attributed to the inclusion of supplementary components. Specifically, we study the results of the following variants.

Table 4. Ablation study on TACoS

| Method | R@1, IoU@0.3↑ | R@1, IoU@0.5↑ | R@5, IoU@0.3↑ | R@5, IoU@0.5↑ |
|---|---|---|---|---|
| De-VLTrans+MSA [32] | 48.79 | 37.57 | 67.63 | 57.91 |
| ACB+De-VLTrans+MSA | 48.76 | 36.79 | 68.58 | 57.49 |
| CAB+MSA | 46.23 | 35.30 | 65.64 | 55.74 |
| ACB+CAB+MSA | 47.44 | 35.49 | **68.71** | 56.91 |
| VTC+De-VLTrans+MSA | 47.09 | 37.78 | 67.24 | 58.08 |
| **VTC+ACB+CAB+MSA** | **49.77** | **37.99** | 68.31 | **58.31** |

- **De-VLTrans+MSA** entry reports the scores obtained by [32] without any added modifications and refers to the complete MSAT architecture.
- **ACB+De-VLTrans+MSA** refers to the use of Asymmetric Co-attention Blocks(ACB) before each **De-VLTrans** layer.
- **CAB+MSA** replaces the decoupled attention weighted self-attention layer of **De-VLTrans+MSA** with a simple connected self-attention layer.
- **ACB+CAB+MSA** refers to addition of Asymmetric Co-attention Blocks infront of **CAB+MSA** model.
- **VTC+ACB+CAB+MSA** adds an extra Video-Text Contrastive loss and required encoder and momentum encoder layers to the **ACB+CAB+MSA**.
- **VTC+De-VLTrans+MSA** only adds the Video-Text Contrastive loss to the **De-VLTrans+MSA** architecture.

Table 5 shows the number of trainable parameters of the original **De-VLTrans+MSA** and our best performing **VTC+ACB+CAB+MSA** variant. By comparing the results from above Table 3 and 4, we can clearly see that the VTC+ACB+CAB+MSA module outperforms all the other variants including the MSAT [32] architecture in TACoS and obtains comparable results in ActivityNet.

Table 5. Number of trainable parameters

| Model | TACoS | ActivityNet |
|---|---|---|
| De-VLTrans+MSA [32] | 36M | 37M |
| **OURS** | **22M** | **25M** |

Furthermore, we also evaluate the effectiveness of adding a VTC or ACB module in MSAT architecture and find that the added module doesn't improve the accuracy instead achieves comparable performances in both datasets with decreased parameter counts.

Although the addition of individual ACB or VTC components in our architecture does not improve the evaluation results, the combination of both is seen to be effective.

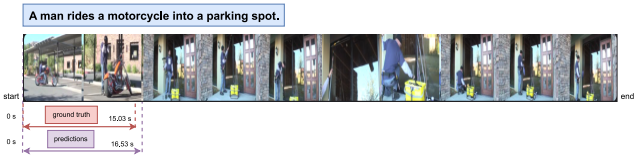


Figure 3. An example of video grounding on ActivityNet dataset

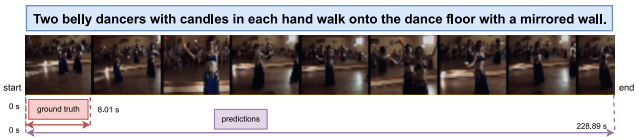


Figure 4. An example of the false predictions on ActivityNet dataset

### 4.6. Qualitative Results

We qualitatively test both our models on the corresponding dataset as shown in Fig. 3, 4 and 5. In the TACoS dataset, our model effectively captures the stage-specific information and accurately localizes the temporal moments. But, in the case of the ActivityNet dataset, even if, the prediction is quite good shown in Fig. 3, the bias in the dataset poses a challenge to generalize to the broader domain of activity recognition tasks which can be well depicted in Fig. 4. The false temporal localization is due to the reason that our model is not able to distinguish the words "walk" and "dance" and therefore, generalize the whole video as dancing instead of recognizing the walking moment as separate actions.

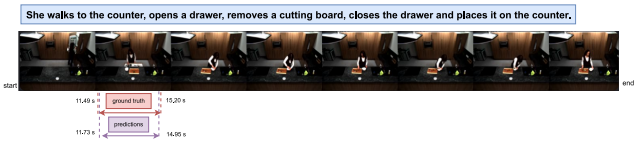


Figure 5. An example of video grounding on TACoS dataset

## 5. Conclusion

In this work, we evaluated the effectiveness of using asymmetric co-attention layers and video text contrastive learning in the context of video grounding. Specifically, we observed that while the addition of an asymmetric co-attention block or the contrastive loss alone does not bring any performance gain, the combined use of both modules improves over the baselines in TACoS and performs comparably in ActivityNet. Additionally, our approach requires considerably fewer learnable parameters and captures more robust multi-modal interactions across both modalities compared to the baseline architecture.

## References

[1] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2

[2] Yi-Wen Chen, Yi-Hsuan Tsai, and Ming-Hsuan Yang. End-to-end multi-modal video temporal grounding. *Advances in*

*Neural Information Processing Systems*, 34:28442–28453, 2021. 1, 2

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[4] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 2, 3

[5] Sunan He, Taian Guo, Tao Dai, Ruizhi Qiao, Chen Wu, Xiujun Shu, and Bo Ren. Vlmae: Vision-language masked autoencoder. *arXiv preprint arXiv:2208.09374*, 2022. 1

[6] Zixi Jia, Minglin Dong, Jingyu Ru, Lele Xue, Sikai Yang, and Chunbo Li. Stcm-net: A symmetrical one-stage network for temporal language localization in videos. *Neurocomputing*, 471:194–207, 2022. 5

[7] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715, 2017. 4

[8] Xiaohan Lan, Yitian Yuan, Xin Wang, Zhi Wang, and Wenwu Zhu. A survey on temporal sentence grounding in videos. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(2):1–33, 2023. 1, 2

[9] Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, Hehong Chen, Guohai Xu, Zheng Cao, et al. mplug: Effective and efficient vision-language learning by cross-modal skip-connections. *arXiv preprint arXiv:2205.12005*, 2022. 1, 2

[10] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 1, 2, 3

[11] Kun Li, Dan Guo, and Meng Wang. Proposal-free video grounding with contextual pyramid network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1902–1910, 2021. 2

[12] Daizong Liu, Xiaoye Qu, Xiao-Yang Liu, Jianfeng Dong, Pan Zhou, and Zichuan Xu. Jointly cross-and self-modal graph attention network for query-based moment localization. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4070–4078, 2020. 5

[13] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 4

[14] Jonghwan Mun, Minsu Cho, and Bohyung Han. Local-global video-text interactions for temporal grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10810–10819, 2020. 2

[15] Guoshun Nan, Rui Qiao, Yao Xiao, Jun Liu, Sicong Leng, Hao Zhang, and Wei Lu. Interventional video grounding with dual contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2765–2775, 2021. 2, 5

[16] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 2

[17] Xiaoye Qu, Pengwei Tang, Zhikang Zou, Yu Cheng, Jianfeng Dong, Pan Zhou, and Zichuan Xu. Fine-grained iterative attention network for temporal language localization in videos. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4280–4288, 2020. 5

[18] Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1:25–36, 2013. 4

[19] Marcus Rohrbach, Michaela Regneri, Mykhaylo Andriluka, Sikandar Amin, Manfred Pinkal, and Bernt Schiele. Script data for attribute-based recognition of composite activities. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part I 12*, pages 144–157. Springer, 2012. 4

[20] Rui Su, Qian Yu, and Dong Xu. Stvgbert: A visual-linguistic transformer based framework for spatio-temporal video grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1533–1542, 2021. 2

[21] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 2

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1

[23] Hao Wang, Zheng-Jun Zha, Xuejin Chen, Zhiwei Xiong, and Jiebo Luo. Dual path interaction network for video moment localization. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4116–4124, 2020. 5

[24] Hao Wang, Zheng-Jun Zha, Liang Li, Dong Liu, and Jiebo Luo. Structured multi-level interaction network for video moment localization via language query. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7026–7035, 2021. 5

[25] Sangmin Woo, Jinyoung Park, Inyong Koo, Sumin Lee, Minki Jeong, and Changick Kim. Explore and match: End-to-end video grounding with transformer. *arXiv preprint arXiv:2201.10168*, 2022. 1

[26] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Tubedetr: Spatio-temporal video grounding with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16442–16453, 2022. 2

[27] Yulan Yang, Zhaohui Li, and Gangyan Zeng. A survey of temporal activity localization via language in untrimmed videos. In *2020 International Conference on Culture-oriented Science & Technology (ICCST)*, pages 596–601. IEEE, 2020. 2

[28] Xinli Yu, Mohsen Malmir, Xin He, Jiangning Chen, Tong Wang, Yue Wu, Yue Liu, and Yang Liu. Cross interaction

network for natural language guided video moment retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1860–1864, 2021. 1, 2

[29] 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. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 685–695, 2021. 1, 2

[30] Hao Zhang, Aixin Sun, Wei Jing, Liangli Zhen, Joey Tianyi Zhou, and Rick Siow Mong Goh. Natural language video localization: A revisit in span-based question answering framework. *IEEE transactions on pattern analysis and machine intelligence*, 44(8):4252–4266, 2021. 2

[31] Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Temporal sentence grounding in videos: A survey and future directions. *arXiv preprint arXiv:2201.08071*, 2022. 2

[32] Mingxing Zhang, Yang Yang, Xinghan Chen, Yanli Ji, Xing Xu, Jingjing Li, and Heng Tao Shen. Multi-stage aggregated transformer network for temporal language localization in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12669–12678, 2021. 1, 2, 3, 4, 5, 6

[33] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12870–12877, 2020. 2, 3, 4, 5