VLG-Net: Video-Language Graph Matching Network for Video Grounding Supplementary Material

1. Formulation of Video-Language Graph Matching

In this section, we provide a detailed overview and formulation of the video-language graph matching. This inputs to this layer are the enriched video representation $X_v^{(bv)}$ and query representation $X_l^{(bl)}$ outputs of the single modality stack of computational blocks. The graph matching layer models the cross-modal context and allows for multi-modal fusion. To this purpose the video-language matching graph is constructed and three types of edges are designed: (i) *Ordering Edge* ($\mathcal{O}$), (ii) *Semantic Edge* ($\mathcal{S}$), and (iii) *Matching Edge* ($\mathcal{M}$).

To aggregate the information, we employ relation graph convolution [5] on the constructed video-language matching graph. Eq. 1 shows the high level representation of the convolutions in this layer.

$$\mathbf{X}^{(GM)} = \mathcal{A}_{\mathcal{O}}\mathbf{X}W_{\mathcal{O}} + \mathcal{A}_{\mathcal{S}}\mathcal{B}\mathbf{X}W_{\mathcal{S}} + \mathcal{A}_{\mathcal{M}}\Gamma\mathbf{X}W_{\mathcal{M}} + \mathbf{X} \quad (1)$$

Here, $\mathbf{X} = \{X_{v,1}^{(b_v)}, \ldots, X_{v,n_v}^{(b_v)}, X_{l,1}^{(b_l)}, \ldots, X_{l,n_l}^{(b_l)}\}$ is the feature representation of all the nodes in the video-language matching graph. $\mathcal{A}_r$ and W_r for $r \in \{\mathcal{O}, \mathcal{S}, \mathcal{M}\}$ represent the binary adjacency matrix and learnable weights for each set of edges. Specifically, $\mathcal{B}$ and Γ scale the adjacency matrices $\mathcal{A}_{\mathcal{S}}$ and $\mathcal{A}_{\mathcal{M}}$. Both $\beta_{i,j} \in \mathcal{B}$ and $\gamma_{i,j} \in \Gamma$ are proportional to $\mathbf{x}_i^\top \mathbf{x}_j$,

$$\beta_{i,j} = \frac{\exp\left[\mathbf{x}_i^\top \mathbf{x}_j\right]}{\sum_{\mathcal{A}_{\mathcal{S}}(k,j)=1} \exp\left[\mathbf{x}_k^\top \mathbf{x}_j\right]}, \quad (2)$$

$$\gamma_{i,j} = \frac{\exp\left[\mathbf{x}_i^\top \mathbf{x}_j\right]}{\sum_{\mathcal{A}_{\mathcal{M}}(k,j)=1} \exp\left[\mathbf{x}_k^\top \mathbf{x}_j\right]}. \quad (3)$$

In practise, to implement GPU-memory efficient graph convolution operation, we replace the time-consuming matrix multiplication by indexing operation of tensors. Thus, the semantic and matching edge convolution can be present as

$$\mathcal{A}_{\mathcal{S}}\mathcal{B}\mathbf{X}W_{\mathcal{S}} = \sum_{j \in \mathcal{N}_i^{\mathcal{S}}} (\hat{W}_{\mathcal{S}}^T[\beta_j \mathbf{x}_j || \mathbf{x}_i]), \quad (4)$$

$$\mathcal{A}_{\mathcal{M}}\Gamma\mathbf{X}W_{\mathcal{M}} = \sum_{j \in \mathcal{N}_i^{\mathcal{M}}} (\hat{W}_{\mathcal{M}}^T[\gamma_j \mathbf{x}_j || \mathbf{x}_i]), \quad (5)$$

where $\mathcal{N}_i^*$ is the neighbourhood of node i connected by edge with type $*$, $* \in \{\mathcal{S}, \mathcal{M}\}$. The $||$ sign means concatenation of features. $\hat{W}_{\mathcal{S}}, \hat{W}_{\mathcal{M}}$ are learnable weights.

Moreover, as shown by A.2 of G-TAD[6], our ordering edge convolution, can be efficiently computed as a 1D convolution with kernel size 3.

$$\mathcal{A}_{\mathcal{O}}\mathbf{X}W_{\mathcal{O}} = Conv1D[X] \quad (6)$$

Therefore, we can equivalently formulate Eq. 1 as:

$$\begin{aligned}\mathbf{X}^{(GM)} = & \, Conv1D[X] \\ & + \sum_{j \in \mathcal{N}_i^{\mathcal{S}}} (\hat{W}_{\mathcal{S}}^T[\beta_j \mathbf{x}_j || \mathbf{x}_i]) \\ & + \sum_{j \in \mathcal{N}_i^{\mathcal{M}}} (\hat{W}_{\mathcal{M}}^T[\gamma_j \mathbf{x}_j || \mathbf{x}_i]) \\ & + \mathbf{X}\end{aligned} \quad (7)$$

2. Graph matching edges ablation

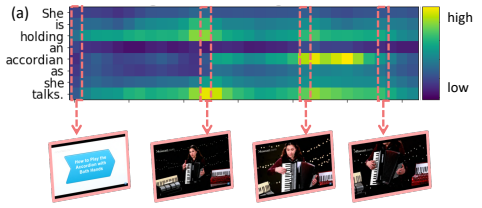
We ablate the contribution of the three different types of edges designed for the graph matching module. We report in Table 1 the performance of VLG-Net for the TACoS dataset when each edge is removed from the architecture. As previously stated, the *Ordering Edges* or *Semantic Edges* are responsible for aggregating contextual information within the graph matching module. When removed, they lead to noticeable degradation of the performance of 2.15% and 3.77%, respectively. Conversely, as expected, when the *Matching Edges* are removed, the performances are severely impaired. We assist in a drop of 27.34%, showcasing the high relevance of the matching operation. Note that, the removal of the *Matching Edges* prevents the fusion between the modalities. Nonetheless, the two modalities still interact in the Masked Attention Pooling module through the learnable cross-attention pooling method. However, this limited interaction cannot bridge the complex semantic information between modalities. The ablation showcases the importance of designing effective operation for multi-modal fusion to achieve high performance on the grounding task. Nonetheless, we can conclude that all edges are relevant and necessary to obtain state-of-the-art performance.

Dataset	Edge Types			R@1 IoU0.5
	Ordering	Semantic	Matching	
TACoS	✓	✓	✓	**34.19**
	✗	✓	✓	32.04
	✓	✗	✓	30.42
	✓	✓	✗	6.85

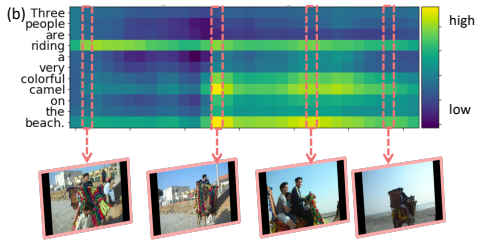
Table 1: **Ablation of different edges.** We investigate the impact of edges within the graph matching layer. We report the performance of our VLG-Net when specific edges are removed, as well as our best performance for TACoS datasets.

3. Visualization graph matching attention

In Fig. 1, we plot the *Matching Edge* weights (before SoftMax) for two video-query pairs, where the *Matching Edge* weights are used to measure the similarity between video snippets and language tokens. In graph convolutions, a *Matching Edge* propagates more information if its weight is high, and vice versa.



(a) Query: "She is holding an accordian as she talks."



(b) Query: "Three people are riding a very colorful camel on the beach."

Figure 1: **Visualization of graph matching attention.** We visualize the *Matching Edge* of the graph matching layer. Correspondence between video snippets and query tokens can be evaluated through the heat-map.

In Fig. 1a, we show the grounding result for a 2 minutes accordian tutorial, with associated query: "She is holding an accordian as she talks". It can be observed from the blue-yellow heat-map that high scores are assigned to the words "holding", "accordian", and "talks", which are the most discriminative tokens for the query localization. Below the heat-map, we visualize the snippets of the video. The unrelated snippets (first and last) are associated with low scores. Conversely, more relevant snippets (central ones) have higher *Matching Edge* weights. This entails that the algorithm is successfully correlating important language cues with relevant video cues when performing the graph matching operation.

Similarly, Fig. 1c shows the result for a 22 second camel riding video, for which the associated query is: "Three people are riding a very colorful camel on the beach." The heat-map highlights the keywords: "riding", "colorful camel", and "beach", which are relatively more informative in the query sentence. Interestingly, the word "riding" is always associated with high attention weights, and a visual inspection confirms that the action happens throughout the whole video. This showcases that our VLG-Net can successfully learn semantic video-language matching. If we focus on the first two snippets of Fig. 1c, we can see that both have associated high scores with the word "riding". However, given the smaller field of view of the first frame, only the second frame contains a more distinguishable camel. In fact, for this particular frame, we observe a high weight score for the words "colorful" and "camel". Moreover, the context of "beach" can be learned from all the last three snippets.

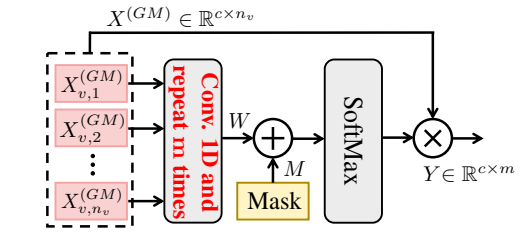
4. Ablation of Masked Attention Pooling

As presented in the main paper, three different implementations of attention for moment pooling operation have been tested. They differ for inputs and operations to achieve the attention scores. Learnable self-attention (Fig. 2a), only relies on the fused features of video and language modalities, which are the output of the graph matching layer, while the cross-attention and learnable cross-attention configurations (Fig. 2b and 2c) also involve a global sentence representation $X_l^{(att)}$ in the process. (See Sec. 3.5 of the paper for more details.) We compare the performances of the three different implementations in Tab 2.

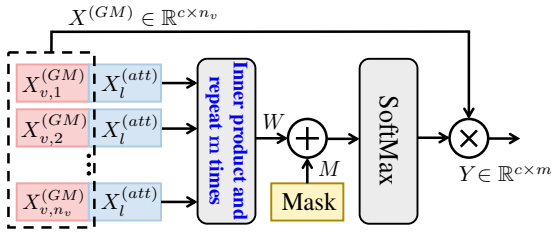
Following the ablation settings in our main paper, we focus on R@1 IoU0.5 and R@5 IoU0.5 for TACoS dataset. We find that the cross-attention setup leads to the lowest performance. Conversely the learnable cross-attention con-

	Learnable self-attention	Cross-attention	Learnable cross-attention
R@1 IoU0.5	29.87	16.62	**34.19**
R@5 IoU0.5	50.24	40.14	**56.56**

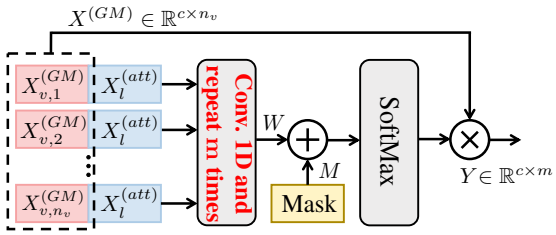
Table 2: **Ablation of masked attention pooling implementations.** The experimental results show that the cross-attention setup leads to sub-optimal performance. Instead, the learnable cross-attention configuration obtains the best performance.



(a) Learnable self-attention



(b) Cross-attention.



(c) Learnable cross-attention.

Figure 2: **Masked attention pooling.** Inputs are video nodes $X_v^{(GM)}$ from the graph matching layer and the query embedding $X_l^{(att)}$ computed through self-attention pooling atop the graph matching output. The output Y represents all moment candidates.

figuration instead, obtains the best performance. This is the reason that motivated our choice of adopting the learnable cross-attention approach to moment pooling in the main paper.

Interestingly we notice that the learnable self-attention setup can achieve relatively high performance. This can be motivated by the intuition that our graph matching layer can effectively fuse the video and language modalities, and by relying on those enriched features only, can we obtain a good representation of the moment's feature. However, involving a global language representation for guiding the moment creation from the enriched snippets features has been shown to yield the best results.

5. Charades-STA

Based on the results obtained from Activitynet-Caption, TACoS, and DiDeMo, our method can theoretically achieve state-of-the-art performance in the Charades-STA dataset. However, we choose not to evaluate VLG-Net on this dataset because of the following observations.

Dataset	Num. Videos	Video-Sentence pairs train	val	test	Vocab. Size
ActivityNet Captions [3]	14926	37421	17505	17031	15406
TACoS [4]	127	10146	4589	4083	2255
DiDeMo [1]	10642	33005	4180	4021	7523
Charades-STA [2]	6670	12404	–	3720	1289

Table 3: **Datasets statistics**. Same as Table 1 in main paper, reported in Supplementary Material for completeness.

(1) This dataset is characterized by the **smallest vocabulary size** and **shortest language annotation** with respect all others datasets (see Tab. 3 and Tab. 4) For example, its vocabulary contains 43% less unique words with respect to TACoS [4], 83% with respect to DiDeMo [1], and 92% with respect to Activity-Captions [3]. This fact can potentially hamper the development of successful methods and reduce the applicability to a real-world scenario where **users might use a richer vocabulary** when querying for moments. Given the great importance of the language for the task at hand, it's diversity in terms of unique tokens' number, and sentence lengths are important factors. This suggests that Charades-STA is less favourable for evaluating the video-language grounding task.

(2) Charades-STA has the **smallest number of video-query pairs** (16124) with respect to all other datasets (See Tab 3). As deep learning methods benefit from a large amount of annotated data, the reduced number of training/testing samples makes the dataset less suited for deep-learning approaches.

(3) Most importantly, Charades-STA **lacks an official validation split.** In machine learning applications, the validation set is mandatory for hyper-parameters search, while the test set is adopted for evaluating the generalization capabilities of a given method to previously unseen data. Given the absence of a validation set, nor a widely accepted procedure for selecting the best models during the development phase, some might use the test set for tuning the hyper-parameters, therefore, harming the measurement of generalization performance. The goal of research is to develop tailor-made solutions for specific problems rather than find-

Dataset	Sentence's lengths Avg.	Std.
Activitynet-Captions [3]	14.4	6.5
TACoS [4]	9.4	5.4
DiDeMo [1]	8.0	3.4
Charades-STA [2]	7.2	1.9

Table 4: **Language annotations statistics**. We report average length (measured in number of tokens) and standard deviation for queries in each dataset. Statistics are computed considering every split for each dataset.

ing the hyper-parameters that can fit the test set best. A conservative researcher could attempt at using the training set (or part of it) as a synthetic validation split. However, this could lead the model to overfit on the specific set of samples. Other methods could be potentially applied (*e.g.* cross-validation), yet no previous work mentioned the adoption of such techniques.

For all these reasons we can conclude that, despite the popularity of Charades-STA as benchmark for the language grounding in video task, we decide not to evaluate our method on it.

References

[1] 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 (ICCV)*, 2017.

[2] Gao Jiyang, Sun Chen, Yang Zhenheng, Nevatia, Ram. TALL: Temporal Activity Localization via Language Query. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[3] 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 (ICCV)*, 2017.

[4] 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 (ACL)*, 2013.

[5] Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. Modeling Relational Data with Graph Convolutional Networks, 2017.

[6] Mengmeng Xu, Chen Zhao, David S. Rojas, Ali Thabet, and Bernard Ghanem. G-TAD: Sub-Graph Localization for Temporal Action Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.