Supplementary Material: Towards Generalisable Video Moment Retrieval: Visual-Dynamic Injection to Image-Text Pre-Training

Dezhao Luo[1], Jiabo Huang[1], Shaogang Gong[1], Hailin Jin[2], and Yang Liu[3*]
[1]Queen Mary University of London
{dezhao.luo, jiabo.huang, s.gong}@qmul.ac.uk
[2]Adobe Research, [3]WICT, Peking University
hljin@adobe.com, yangliu@pku.edu.cn

Scores	Overall	Sub.	Verb.	Obj.
Pos.	0.79	0.82	0.76	**0.86**
Neg.	0.21	0.18	**0.24**	0.14
Dif.	0.59	0.64	0.52	**0.72**

Table 1. Preliminary study on CLIP's [6] understanding of different aspects of the language. The number indicates the average similarity score. The column indicates the negative sample type: "Sub." denotes the negative sample does not correspond to the sentence by subjects; "Obj." denotes objects.

In this material, we present more analysis and experiments for our method (VDI). Sec. 1 provides additional statistical analysis of the image-text pre-training models. Sec. 2 provides more ablations on implementation details.

1. Preliminary Study on CLIP

A key challenge for video moment retrieval (VMR) is to understand the actions from both the visual and the textual modality. To study if image-text pre-training methods, such as CLIP [6], can encode action information, we design an experiment to determine if their pre-training models can distinguish different types of verbs or if they rely solely on nouns in a given sentence. To do so, we design experiments with the dataset [3] which provides image-text triplets with each including a sentence, a positive image and a negative image. Specifically, the positive image is correctly described by the sentence, and the negative image does not match the sentence from one specific aspect (*i.e.* subject, verb or object). In Table 1, each triplet is categorised by its negative type (Sub., Verb. or Obj.), and we report their scores according to the category type to study the model's understanding of different aspects of the language.

Given an image-text triplet (a sentence, a positive image, and a negative image), we take the pre-trained CLIP [6] to calculate the similarity score between the image with the positive sentence (Pos.) and that between the image with the negative sentence (Neg.). The scores are normalised in each triplet, and the difference (Dif.) between the positive and negative scores is calculated. Table 1 shows the averaged score for the whole dataset. The difference between the positive and negative scores indicates how well the model can distinguish between the matched and unmatched images, with smaller differences indicating a more confused model. From the results, it can be seen that CLIP is more likely to assign higher scores to negative samples of the verb type, and the difference between positive and negative is less pronounced than for the other types. For the data samples of the verb type, as the negative image does not match the sentence only from the verb part, a smaller difference between the negative and positive indicates that the verb is less discriminative and more challenging than the subject and object.

In our method, we investigate the problem that it is hard for CLIP to capture video changes, and propose to inject visual context and spatial dynamic information into the words describing video changes.

λ_{vc}	λ_{sd}	R@1, IoU=0.5	R@1, IoU=0.7	mIoU
0	0	43.85	24.17	39.50
0.1	0	44.04	28.35	41.42
0.5		45.47	**29.35**	40.61
1		42.45	24.17	38.70
0	0.1	44.70	25.61	39.70
	0.01	44.60	26.06	40 .09
	0.001	43.60	25.32	39.19
0.5	0.1	45.18	27.91	41.05
	0.01	**46.47**	28.63	**41.60**
	0.001	45.47	27.31	41.28

Table 2. Ablation study on the loss weight λ_{vc} and λ_{sd}.

*Corresponding authors

$\mathcal{L}^{vc}$	$\mathcal{L}^{sd}$	R@1, IoU=0.5	R@1, IoU=0.7	mIoU
Con.	Rel.	**46.47**	**28.63**	**41.60**
Con.	Con.	45.32	28.35	41.08
Rel.	Rel.	38.56	23.31	36.20

Table 3. Ablation study on the loss implementation for $\mathcal{L}^{vc}$ and $\mathcal{L}^{sd}$. "Con." denotes the consistent implementation, "Rel." denotes the relational implementation. "Con."/"Rel." for $\mathcal{L}^{vc}$/$\mathcal{L}^{sd}$ is our default setting.

Method	R@1, IoU=0.5	R@1, IoU=0.7	mIoU
Q.Inj.	44.89	26.04	39.52
Q^d.Inj.	**46.47**	**28.63**	**41.60**

Table 4. Ablation study on the dynamic query emphasis. "Q.Inj." denotes injecting information into the sentence query Q. "Q^d.Inj." is our default setting which injects information into the dynamic query.

2. Ablation Study

In this section, we provide more ablation studies on model design choices. We report the performance under Charades-STA[1] with OOD novel-word split [5] by default.

Loss Weight. Table 2 shows the performance with the loss weight λ_{vc} for $\mathcal{L}^{vc}$ and λ_{sd} for $\mathcal{L}^{sd}$. $\lambda_{vc}/\lambda_{sd}$ = 0/0 denotes the baseline without our design. As one can see from the result, we take the 0.5/0.01 for $\lambda_{vc}/\lambda_{sd}$ as the default setting.

Loss Implementaion. In our experiments, we use a consistent implementation for $\mathcal{L}^{vc}$ to encourage the consistency of the text embedding to the visual context. And we use a relational implementation for $\mathcal{L}^{sd}$ to enforce the correlations between the spatial dynamic features of different videos with their corresponding descriptions. In Table 3, we ablate the different loss implementations (Consistent (Con.) vs Relational (Rel.)) for $\mathcal{L}^{vc}$ and $\mathcal{L}^{sd}$. For visual context injection, relational implementation will force different sentences sharing similar visual context to be similar, which will confuse the learning process as they may describe unrelated actions. For spatial dynamic injection, relational implementation will encourage different sentences sharing similar motion patterns to focus on the common descriptions of such spatial dynamic information.

Dynamic Query Emphasis. In our model design, we inject the visual context and spatial dynamic information into the text with a focus on words describing the video changes (dynamic query Q^d). In this ablation, we demonstrate the necessity to emphasise the dynamic query Q^d with a comparison with injecting information into the sentence query Q. As one can see from the Table 4, Q^d injection (Q^d.Inj.) can yield better performance than Q injection (Q.Inj.). To understand this result, we calculate the weight map of the last Transformer [9] layer in the text encoder and average them by their speech parts (Nouns, Verbs and Others). The weight of each word indicates the contribution to the sentence feature. As shown in Fig. 1, we observe that the CLIP pre-trained model intends to focus more on nouns as the sentence feature, thus less on the learning of verbs. This explains the necessity to highlight the dynamic query.

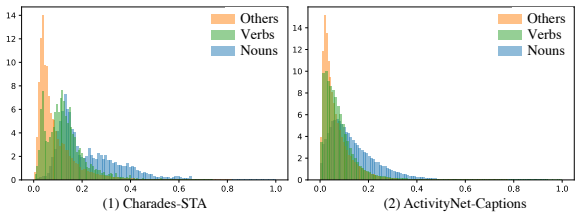

Figure 1. The visualisation of the weight of each word for Charades-STA [1] and ActivityNet-Captions [4]. The x-axis represents the weight from the last Transformer[9] layer in the text encoder. The y-axis represents the number in percentage.

Method	Pre-train	R@1, IoU=0.5	R@1, IoU=0.7	mIoU
MMN[10]	Separated	41.59	23.60	37.90
VDI(Q^s)		41.29	24.46	38.21
MMN[10]	Joint	43.85	24.17	39.50
VDI(Q^h)		46.01	27.05	40.94
VDI(Q^s)		**46.47**	**28.63**	**41.60**

Table 5. Ablation study on visual context generation. The "Pre-train" column indicates how the feature extractors are pre-trained. "Separated" denotes the features are pre-trained separately and no pre-learned visual-textual correlations are provided.

Method	Feature	R@1, IoU=0.5	R@1, IoU=0.7
MMN[10]	VGG-Separated	47.31	27.28
MMN[10]	ViT-Joint	50.48	29.65
VDI		52.32	31.37
MMN[10]	ResNet-Joint	52.88	32.02
VDI		**53.98**	**33.20**

Table 6. Evaluation on the original testing split. The "Feature" column denotes the method using separated or joint pre-training backbones.

Visual Context Generation. For visual context injection, we apply a static query Q^s guided visual context genera-

tion, taking advantage of the pre-learned visual and textual correlations from CLIP [6]. In Table 5, we prove that the pre-learned visual-textual correlation is important by a comparison with different visual context generations.

In Table 5 (Row 2), we apply our method with separated pre-training models and follow the setting as our baseline [10] to use the VGG model [8] as the visual encoder and the DistilBERT [7] as the textual encoder. As one can see from the result, the separated pre-training feature is sub-optimal to the VMR task. This is partially because it is challenging to generate visual context information without visual-textual correlations.

In Table 5 (Row 4), we randomly select half of the words in the sentence as the query (Q^h) for visual context extraction, and the rest of the sentence is injected with visual-dynamic information. One can see from the result that such a random combination of object and dynamic descriptions in Q^h can not fully take advantage of the pre-training visual-textual correlations.

Feature Backbone. To further validate the generalisation of our method, we take the ResNet [2] backbone provided by CLIP [6] for feature extraction. We report the performance under the original split in Table 6, which shows that our method is also effective under ResNet backbones.

References

[1] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *ICCV*, pages 5267–5275, 2017. 2

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 3

[3] Lisa Anne Hendricks and Aida Nematzadeh. Probing image-language transformers for verb understanding. In *ACL-IJCNLP*, pages 3635–3644, 2021. 1

[4] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *ICCV*, pages 706–715, 2017. 2

[5] Juncheng Li, Junlin Xie, Long Qian, Linchao Zhu, Siliang Tang, Fei Wu, Yi Yang, Yueting Zhuang, and Xin Eric Wang. Compositional temporal grounding with structured variational cross-graph correspondence learning. In *CVPR*, pages 3032–3041, 2022. 2

[6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 1, 3

[7] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. 3

[8] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 3

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017. 2

[10] Zhenzhi Wang, Limin Wang, Tao Wu, Tianhao Li, and Gangshan Wu. Negative sample matters: A renaissance of metric learning for temporal grounding. In *AAAI*, volume 36, pages 2613–2623, 2022. 2, 3