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

Abstract

*The correlation between the vision and text is essential for video moment retrieval (VMR), however, existing methods heavily rely on separate pre-training feature extractors for visual and textual understanding. Without sufficient temporal boundary annotations, it is non-trivial to learn universal video-text alignments. In this work, we explore multi-modal correlations derived from large-scale image-text data to facilitate generalisable VMR. To address the limitations of image-text pre-training models on capturing the video changes, we propose a generic method, referred to as Visual-Dynamic Injection (VDI), to empower the model's understanding of video moments. Whilst existing VMR methods are focusing on building temporal-aware video features, being aware of the text descriptions about the temporal changes is also critical but originally overlooked in pre-training by matching static images with sentences. Therefore, we extract visual context and spatial dynamic information from video frames and explicitly enforce their alignments with the phrases describing video changes (*e.g.* verb). By doing so, the potentially relevant visual and motion patterns in videos are encoded in the corresponding text embeddings (injected) so to enable more accurate video-text alignments. We conduct extensive experiments on two VMR benchmark datasets (Charades-STA and ActivityNet-Captions) and achieve state-of-the-art performances. Especially, VDI yields notable advantages when being tested on the out-of-distribution splits where the testing samples involve novel scenes and vocabulary.*

1. Introduction

Video moment retrieval (VMR) aims at locating a video moment by its temporal boundary in a long and untrimmed video according to a natural language sentence [3, 13]. It is a critical task which has been extensively studied in a variety of real-world applications including human-computer interaction [5], and intelligent surveillance [9]. In practice, raw videos are usually unscripted and unstructured, while the words being chosen for describing the same video moments can be varied from person to person [45, 63]. To be generalisable to different scenes, VMR is fundamentally challenging as it requires the comprehension of arbitrary complex visual and motion patterns in videos and an unbounded vocabulary with their intricate relationships.

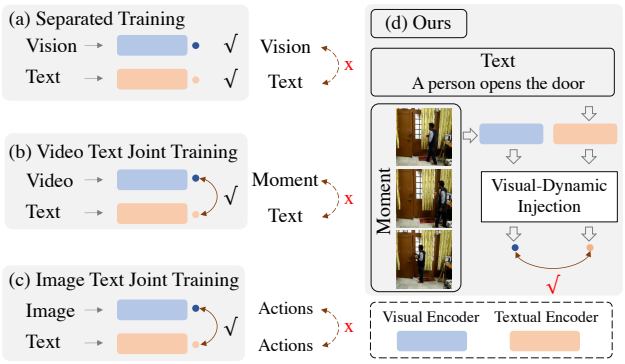

Figure 1. Contemporary methods lack moment-text correlations. Our method takes the advantage of image-text pre-trained models and learns moment-text correlations by visual-dynamic injection.

For the fine-grained retrieval objective of VMR, the precise segment-wise temporal boundary labels are intuitively harder to be collected than conventional image/video-level annotations. In this case, rather than training from scratch with a limited number of temporally labelled videos, existing VMR solutions [3, 13, 14, 62] heavily rely on single-modal pre-training [8, 48] for visual and textual understanding (Fig. 1 (a)). By doing so, they focus on modelling the correlations between the pre-learned features of videos and sentences. Nonetheless, without sufficient training data, it is non-trivial to derive universal video-text alignments so to

*Corresponding authors

generalise to novel scenes and vocabulary.

Separately, the recent successes achieved by joint vision-language pre-training in zero-shot learning [21, 42] demonstrate the potential of adapting the multi-modal correlations derived from large-scale visual-textual data to facilitate generalisable VMR. Whilst it is intuitive to adopt the video-text pre-learned features [34, 38, 50] for moment retrieval (Fig. 1 (b)), it has been shown that the models pre-trained with coarse-grained video-level labels can not transfer well to localisation-based tasks like VMR due to their unawareness of fine-grained alignments between text and frames or clips [2]. Such a misalignment problem is less likely to exist in pre-training by image-text matching. However, image-based pre-training models [21, 42] are less sensitive to the changes in videos and the words describing such dynamics in text [17]. This is inherent in matching sentences and images with static content but is significant in understanding video actions and activities (Fig. 1 (c)). It is suboptimal to directly apply image-text pre-learned features on VMR.

In this work, we propose a generic method for exploiting large-scale image-text pre-training models to benefit generalisable VMR by the universal visual-textual correlations derived in pre-training, dubbed as Visual-Dynamic Injection (VDI). The key idea is to explore the visual context and spatial dynamic information from videos and inject that into text embeddings to explicitly emphasise the phrases describing video changes (*e.g.* verb) in sentences (Fig. 1 (d)). Such visual and dynamic information in text is critical for locating video moments composed of arbitrary evolving events but unavailable or overlooked in image-text pre-training. Specifically, we consider it essential for VMR models to answer two questions: "what are the objects" and "how do the objects change". The visual context information indicates the content in the frames, *e.g.* backgrounds (scenes), appearances of objects, poses of subjects, *etc*. Meanwhile, the spatial dynamic is about the location changes of different salient entities in a video, which potentially implies the development of their interactions. VDI is a generic formulation, which can be integrated into any existing VMR model. The only refinement is to adapt the text encoder by visual-dynamic information injection during training. Hence, no additional computation costs are introduced in inference.

Our contributions are three-folded: **(1)** To our best knowledge, this is the first attempt on injecting visual and dynamic information to image-text pre-training models to enable generalisable VMR. **(2)** We propose a novel method for VMR called Visual-Dynamic Injection (VDI). The VDI method is a generic formulation that can be integrated into existing VMR models and benefits them from the universal visual-textual alignments derived from large-scale image-text data. **(3)** The VDI achieves the state-of-the-art performances on two standard VMR benchmark datasets. More importantly, it yields notable performance advantages when being tested on the out-of-distribution splits where the testing samples involve novel scenes and vocabulary. VDI's superior generalisation ability demonstrates its potential for adapting image-text pre-training to video understanding tasks requiring fine-grained visual-textual comprehensions.

2. Related Work

2.1. Video Moment Retrieval

To tackle the VMR task and predict accurate moment boundaries, existing methods [19, 20, 32, 55, 61, 62] first generated visual features and textual features from pre-training encoders [44, 51], then they designed correlation models to align the two modalities. Ghosh et al. [14], Liu et al. [31] and Zeng et al. [60] proposed to select the starting and ending frames by leveraging cross-modal interactions between text and video. He et al. [15] and Wang et al. [54] proposed reinforcement learning methods for VMR. Gao et al. [13], Wang et al. [55], Zhang et al. [62] took a two-stage pipeline by generating proposals and ranking them relying on the similarity between proposal and query. Li et al. [27], Liu et al. [33], Yang et al. [58] focused on debias problems including the temporal location bias [33, 58], or the word-composition bias [27].

Even though existing methods have demonstrated promising performance for VMR, we argue that models that take separate pre-training visual and textual feature extractors are suboptimal as they need to learn the alignment of the two modalities from scratch. It is demanding to learn from large-scale image-text datasets due to a lack of well-annotated moment-text datasets [38] resulting in poor generalisation [2].

2.2. Vision-Language Pre-Training

Vision-language models have demonstrated great potential in learning generic visual representations and allowing transferring to a variety of downstream tasks [25, 43]. Previously, Frome et al. [12], Mori et al. [40], Weston et al. [56] had explored the connection between images and words using paired text documents. As more and more data accessible from the Internet, image-text pre-training models including CLIP [42], ALIGN [21], ALBEF [28] and Florence [59] proposed to pre-train vision–language models with a contrastive loss. Benefiting from large-scale web data (400M for CLIP, 1.8B for ALIGN and 900M for Florence), image-text pre-training methods can learn powerful visual representation as well as their correlations. Similar ideas can be seen in video-text pre-training [2] with a large-scale video-text dataset Howto100M [38].

Even though the pre-training image-text models can capture the object appearance from visual embeddings and their corresponding description (*e.g.* nouns) in the text embed-

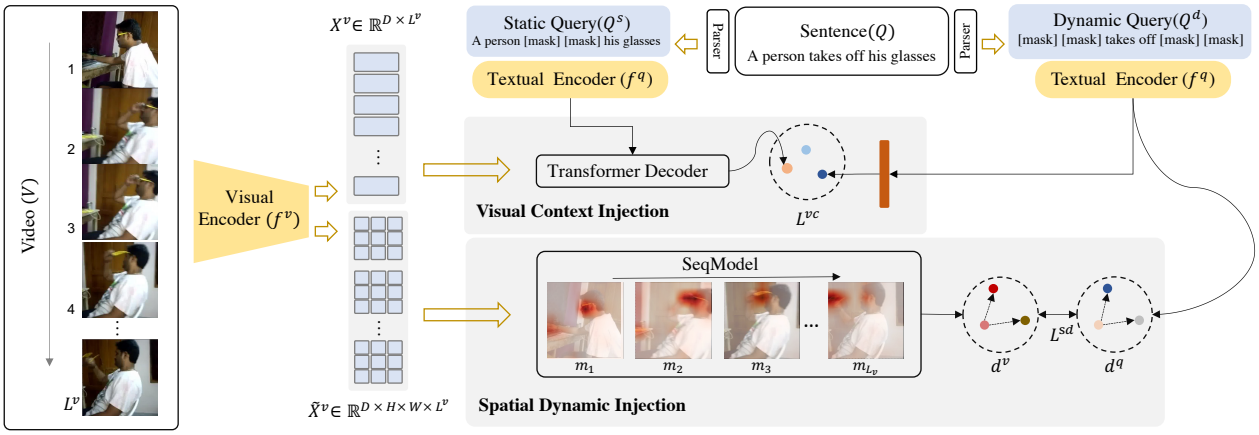

Figure 2. The framework of VDI, in which the video V is fed into a visual encoder to generate image global features X^v and image patch features $\tilde{X}^v$. The sentence Q is parsed into static query Q^s and dynamic query Q^d. Visual Context Injection (L^{vc}) aligns the Q^d with Q^s guided visual context information. Spatial Dynamic Injection (L^{sd}) empowers the Q^d with the awareness of the spatial dynamics.

dings [49], we consider they are suboptimal to capture the change between frames as well as their corresponding descriptions (*e.g.* verbs) in the text.

2.3. Correlation Transfer Learning

Cao et al. [2] explored the video-text pre-training to learn the alignment between moment and text. However, it is suboptimal with a lack of fine-grained well-annotated video-text alignment samples (50% samples are not aligned in dataset Howto100M [38]). Meanwhile, image-text pre-training methods have shown promising generalisable ability with multiple downstream tasks, including image classification [6, 35], action recognition [53], video retrieval [37]. Ju et al. [22] transferred CLIP models [42] to multiple video tasks by textual prompting and visual temporal modelling. Wang et al. [53] utilized CLIP models for video recognition by traditional end-to-end fine-tuning. Lin et al. [29] proposed to build temporal dependencies between images. Cheng et al. [4], Luo et al. [37] proposed to use correlation modelling ability from CLIP models and build their temporal relations by transformers[52] for video retrieval task.

Even though existing methods [4, 22, 29, 37] demonstrated promising results in transferring CLIP to video understanding task, a problem still remains: image-text pre-training models are less sensitive to actions[17]. In this work, we explore multi-modal correlations derived from large-scale image-text data for generalisable VMR. To enable the action understanding for image-text pre-training models, we extract visual context information and spatial dynamic information and enable the text encoder to understand the entities in video frames and their movements.

3. VMR by Image-Text Pre-Training

Given an untrimmed video $V = \{\boldsymbol{I}_i\}_{i=1}^{L^v}$ composed of L^v frames and a natural language sentence Q, the objective of video moment retrieval is to predict when the target video moment starts $t^s \in [1, L^v]$ and ends $t^e \in [t^s, L^v]$ in the video V. For generalisable VMR, not only what happens in V and what is described in Q but also their alignments are supposed to be modelled. This is intrinsically challenging considering the free-form nature of both the unscripted videos and natural language sentences as well as their complex correlations.

To retrieve a target video moment by a text sentence, the untrimmed video V and query sentence Q are first fed into a pre-training visual and a textual encoder in respective to obtain the video $X^v = \{\boldsymbol{x}_i^v\}_{i=1}^{L^v} = \{f^v(\boldsymbol{I}_i)\}_{i=1}^{L^v} \in \mathbb{R}^{D \times L^v}$ and sentence $\boldsymbol{x}^q = f^q(Q) \in \mathbb{R}^D$ features. Both the features will then be taken as the inputs to a video moment retrieval model $f^g(X^v, \boldsymbol{x}^q)$ to predict the temporal boundary of the target moment $(\tilde{t}^s, \tilde{t}^e)$. As the visual and textual encoders are pre-trained on image-text data, they are unaware of temporal changes in videos or the words/phrases describing them in the text. Therefore, we additionally adapt the two encoders (Fig. 2) to model the visual context and spatial dynamic in both modalities. Together with the pre-learned visual-textual correlations which are prone to be universal, the video change-sensitive features will enable f^g to predict accurate temporal boundary even for video moments filmed in novel scenes or described by unseen vocabulary.

3.1. Visual-Dynamic Injection

We start with using a language parsing tool[1] to extract all the noun chunks (a noun plus the words describing the

[1] spaCy: https://spacy.io/

noun) in the query sentence Q, which are supposed to describe the entities of interest in the target video moment. We mask all the other words in the sentence Q and consider such a masked query can be matched with the corresponding video content without knowing the temporal dependencies among frames. In this case, we call the masked sentence about the static content in videos as *static query* Q^s. In contrast, we construct another *dynamic query* Q^d with all the noun chunks in Q being masked, which is critical for understanding complex video moments composed of arbitrary evolving events but missing in the image-text pretraining. Whilst the video change information can be captured by additional sequence analysis in the video moment retrieval model f^g, it is impractical for f^g to understand the phrases describing video changes in the text which are originally overlooked in $\boldsymbol{x}^q$. Therefore, we model video changes with visual context injection and spatial dynamic injection and enforce the text encoder to match the dynamic queries Q^d with them. By doing so, the adapted text encoder is able to yield visual-dynamic sensitive representations for query sentences to ensure more accurate VMR by both visual and dynamic matching.

Visual Context Injection. The visual context we discover in the videos is about how the frames display the entities related to the video changes. Such visual information is likely to encoding the presence of scenes (*e.g.* outdoor or indoor), the status of entities (*e.g.* boiling or cold water) and *etc*. It is important for recognising and locating video moments which involve specific objects and scenes. To that end, we apply a transformer decoder [52] and encode the visual context information into the static query Q^s:

$$\begin{aligned} \boldsymbol{x}^{qs} &= f^q(Q^s) \in \mathbb{R}^D, \\ \tilde{x}^{qs} &= \text{TransDecoder}(\boldsymbol{x}^{qs}, X^v, X^v) \in \mathbb{R}^D. \end{aligned} \tag{1}$$

In Eq. (1), $\boldsymbol{x}^{qs}$ is the D-dimensional textual feature of the static query Q^s obtained by the text encoder f^q, X^v is the video frame features, and $\tilde{x}^{qs}$ is the visual context-aware feature of Q^s computed by a transformer decoder whose three inputs correspond to the query, key and value, respectively. To inject such visual context information to the text encoder with a focus on the words describing the changes in videos, we compute the dynamic query feature Q^d by the text encoder and encourage its consistency with $\tilde{\boldsymbol{x}}^{qs}$:

$$\begin{aligned} \boldsymbol{x}^{qd} &= f^q(Q^d) \in \mathbb{R}^D \\ \mathcal{L}^{vc}(V, Q) &= \|\text{FC}(\boldsymbol{x}^{qd}) - \tilde{\boldsymbol{x}}^{qs}\|_2^2, \end{aligned} \tag{2}$$

where $\text{FC}(\cdot)$ is a linear projection layer. The rationale behind this design is to probe the video frames by the static query in order to select the visual context engaging the entities potentially existing in the target moments. By doing so, the text encoder is updated to align the dynamic query with its visual context and avoid distractions from irrelevant video content.

Spatial Dynamic Injection. Besides the visual context demonstrating the entity of interests in frames, another crucial information for video change is about how the spatial locations of different entities change in time order. However, the motion patterns encoding such dynamic information is hidden in the complex visual patterns in videos. It is non-trivial to discover and use them to raise the text encoder's attention on the corresponding descriptions. Therefore, we propose to extract the location changes of salient entities in videos and explicitly inject such spatial dynamics into the text encoder. To that end, we first obtain the per-frame spatial features $\tilde{X}_i^v \in \mathbb{R}^{D\times H\times W}\ \forall\ i \in [1, L^v]$ as the last feature maps produced by the convolutional neural networks [16, 26, 47] or the patch-wise features in Visual Transformers [10]. The H and W denote the height and width resolutions of concerns. We then adopt a transformer-like formulation to compute a heatmap for each video frame:

$$M_i = \text{FC}(\boldsymbol{x}_i^v) \cdot \text{FC}(\tilde{X}_i^v)/\sqrt{D} \in \mathbb{R}^{H\times W}. \tag{3}$$

The frame-wise heatmap is computed by the correlations between every spatial feature and the global image representation. Therefore, the salient entities whose visual information is encoded in the image feature will result in corresponding salient regions in the heatmap. We then flatten the heatmap and feed it into a linear projection layer to compute a D-dimensional vector as the holistic representation of the spatial feature for each frame. The spatial dynamics in the video can be given by any sequence analysis model:

$$\begin{aligned} \boldsymbol{m}_i &= \text{FC}(\text{Flatten}(M_i)) \in \mathbb{R}^D \\ \tilde{\boldsymbol{m}} &= \text{SeqModel}(\{\boldsymbol{m}_i\}_{i=1}^{L^v}) \in \mathbb{R}^D. \end{aligned} \tag{4}$$

As the visual information is deprecated in the spatial features, we cannot probe them by the static query Q^s. Hence, we choose a transformer encoder [52] to build their dependencies and take the averaged outputs as the spatial dynamic feature $\tilde{\boldsymbol{m}}$ of the video.

To inject such dynamic information into the text encoder, we then enforce consistent correlations between the spatial dynamic features of different videos (V and V') and the corresponding descriptions in text (Q^d and $Q^{d'}$) by

$$\begin{aligned} \epsilon^v = \cos(\tilde{\boldsymbol{m}}, \tilde{\boldsymbol{m}}') \quad \epsilon^q = \cos(\text{FC}(\boldsymbol{x}^{qd}), \text{FC}(\boldsymbol{x}^{qd'})) \\ \mathcal{L}^{sd}(V, Q, V', Q') = (\epsilon^v - \epsilon^q)^2. \end{aligned} \tag{5}$$

In Eq. (5), the notation ϵ^v stands for the cosine similarity between the spatial dynamic features of two videos while ϵ^q is that of the two corresponding dynamic queries Q^d and

$Q^{d'}$. In contrast to the visual context injection, since the spatial features used here lost all the visual cues in videos, we optimise their correlations consistency with that of the dynamic queries rather than directly pushing them closer to the matched textual features. By doing so, different sentences matched with the videos sharing similar motion patterns will be encouraged to focus on the common descriptions of such dynamic information in the text.

Video Moment Retrieval. With the visual and textual features pre-learned from large-scale image-text datasets as well as our adaptation of textual features to be aware of temporal changes in videos, our VDI model is ready to benefit existing VMR models. Here, we take the state-of-the-art Mutual Matching Network (MMN) as an example. Specifically, given the frame-wise video features X^v and the sentence feature $\boldsymbol{x}^q$, we first enumerate all the start-end frame pairs to generate $L^m = L^v \times L^v$ video segments as the candidates of the target moment. We then construct a 2D feature map $X^m = \text{Conv2D}(\{\boldsymbol{x}^m_{i,j}\}^{L^v}_{i,j=1}) \in \mathbb{R}^{D\times L^v\times L^v}$ where $\boldsymbol{x}^m_{i,j}$ is the feature of the segment starting from the i-th frame and ending at the j-th frame. After that, both the features of video segments X^m and query sentences $\boldsymbol{x}^q$ will be linearly projected into a common space and their alignments are then measured by cosine similarities $\cos(\cdot,\cdot)$:

$$\tilde{Y}^{\text{iou}} = \cos(\text{FC}(\boldsymbol{x}^q), \text{Conv1D}(X^m)) \in \mathbb{R}^{L^v\times L^v}. \quad (6)$$

The predicted alignment scores $\tilde{Y}^{\text{iou}}$ between every segment and the query sentence will be supervised by the temporal IoU between it and the manually labelled temporal boundary (t^s, t^e) of the target moment:

$$\begin{aligned} y^{\text{iou}}_{i,j} &= \text{IoU}((t^s,t^e),(i,j)),\\ Y^{\text{iou}} &= \{y^{\text{iou}}_{i,j}\}^{L^v}_{i,j=1} \in \mathbb{R}^{L^v\times L^v},\\ \mathcal{L}^{iou}(V,Q) &= \text{BCE}(Y^{\text{iou}}, \tilde{Y}^{\text{iou}}). \end{aligned} \quad (7)$$

Besides learning to locate video moments by aligning positive segment-text pairs, we follow MMN to conduct mutual-modal contrastive learning among negative sample pairs. In particular, for a moment $\boldsymbol{x}^m_{t^s,t^e}$ in video V and its description $\boldsymbol{x}^q$, we construct a negative video set $\mathcal{X}^{m-}$ and a negative sentence set $\mathcal{X}^{q-}$. We then map the segments and queries features to another shared feature space by linear projections and conduct contrastive learning by:

$$\begin{aligned} \tilde{X}^m = \{\tilde{\boldsymbol{x}}^m_{i,j}\}^{L^v}_{i,j=1} &= \text{Conv1D}(X^m) \in \mathbb{R}^{D\times L^v\times L^v},\\ \tilde{\boldsymbol{x}}^q &= \text{FC}(\boldsymbol{x}^q) \in \mathbb{R}^D,\\ p^m &= \frac{\exp(\cos(\tilde{\boldsymbol{x}}^m_{t^s,t^e}, \tilde{\boldsymbol{x}}^q)/\tau)}{\sum_{\boldsymbol{x}\in\{\boldsymbol{x}^m_{t^s,t^e}\}\cup\mathcal{X}^{m-}}\exp(\cos(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{x}}^q)/\tau)}\\ p^q &= \frac{\exp(\cos(\tilde{\boldsymbol{x}}^m_{t^s,t^e}, \tilde{\boldsymbol{x}}^q)/\tau)}{\sum_{\boldsymbol{x}\in\{\boldsymbol{x}^q\}\cup\mathcal{X}^{q-}}\exp(\cos(\tilde{\boldsymbol{x}}^m_{t^e,t^e}, \tilde{\boldsymbol{x}})/\tau)}\\ \mathcal{L}^{cl}(V,Q) &= -\log p^m - \log p^q. \end{aligned} \quad (8)$$

In Eq. (8), the tilde on top of all the segments and queries features $\{\boldsymbol{x}^m_{t^s,t^e}, \boldsymbol{x}^q, \boldsymbol{x}\}$ denotes their linear projected counterparts. The variables p^m and p^q measure how likely the model identifies the target moment $\boldsymbol{x}^m_{t^s,t^e}$ and the query sentence $\boldsymbol{x}^q$ according to each other from the respective negative sets. In this multi-modal common space, we can compute another alignment scores between every candidate segment and the query sentence:

$$\tilde{Y}^{cl} = \cos(\tilde{\boldsymbol{x}}^q, \tilde{X}^m) \in \mathbb{R}^{L^v, L^v}. \quad (9)$$

The two video-text alignment predictions will then be fused by the hadamard product and the temporal boundary predicted for the target moment can then be computed in a maximum likelihood manner:

$$\begin{aligned} \tilde{Y} &= \tilde{Y}^{iou} \odot \tilde{Y}^{cl}\\ \tilde{t}^s = \arg\max(\text{cmax}(\tilde{Y})), &\quad \tilde{t}^e = \arg\max(\tilde{\boldsymbol{y}}_{t^s}). \end{aligned} \quad (10)$$

By learning from $\mathcal{L}^{iou}$ and $\mathcal{L}^{cl}$ jointly, the model is trained to be aware of both the matched and unmatched video-text information.

Algorithm 1 Visual-Dynamic Injection (VDI)

Input: Untrimmed videos V, Query sentences Q, Temporal boundary labels (t^s, t^e), A visual f^v and a textual encoder f^q from image-text pre-training.
Output: An updated video moment retrieval model.
Generates the static Q^s and dynamic Q^d query sentences;
Computes the features of Q^s and Q^d by f^q;
Computes the features of videos V by f^v;
Computes the visual context $\mathcal{L}^{vc}$ (Eq. (2)) and spatial dynamic $\mathcal{L}^{sd}$ (Eq. (5)) losses;
Adapts the textual encoder f^q by minimising $\mathcal{L}^{vc}$ and $\mathcal{L}^{sd}$;
Computes the features of the query Q by f^q;
Feeds the features of video V and query Q to f^g;
Computes the losses $\mathcal{L}^{iou}$ (Eq. (7)) and $\mathcal{L}^{cl}$ (Eq. (8));
Optimises the VMR model f^g by minimising $\mathcal{L}^{iou}$ and $\mathcal{L}^{cl}$.

3.2. Model Training

The VDI model is trained in a conventional batch-wise manner. A mini-batch is composed of n randomly sampled video-text pair (V, Q) as well as the temporal boundary labels (t^s, t^e) of the target moments. The overall loss functions are computed by:

$$\begin{aligned} \mathcal{L} = \frac{1}{n}\sum_{i=1}^{n}(&\lambda^{iou}\mathcal{L}^{iou}(V_i,Q_i) + \lambda^{cl}\mathcal{L}^{cl}(V_i,Q_i) +\\ &\lambda^{vc}\mathcal{L}^{vc}(V_i,Q_i) +\\ &\lambda^{sd}\frac{1}{n}\sum_{j=1}^{n}\mathcal{L}^{sd}(V_i,Q_i,V_j,Q_j)). \end{aligned} \quad (11)$$

The training objective function $\mathcal{L}$ in Eq. (11) is then be used to optimise both the VMR model f^g and the text encoder f^q from the pre-training model by any stochastic gradient descent algorithms. The overall training process of the VDI model is summarised in Alg. 1.

4. Experiments

To evaluate the importance of generalisable correlations between the visual and textual space, we conduct experiments on video moment retrieval (VMR) and compare with the SOTAs on both out-of-distribution (OOD) and independent and identically distributed (IID) data splits. In this section, we first explain the implementation details and then report our results in the comparison with recent methods. Finally, we evaluate the effectiveness of each component in our methods.

4.1. Experimental Settings

4.1.1 Dataset

Charades-STA [13] is a benchmark dataset for VMR, which is built upon the Charades dataset [46]. The Charades dataset is collected for video action recognition and video captioning. Gao et al. [13] adapt the Charades dataset to VMR by collecting the query annotations. The Charades-STA dataset contains 6670 videos and involves 16124 queries, where 12404 pairs are used for training and 3720 for testing. The average duration of the videos is 30.59 seconds and each video contains 2.41 annotated moments, and the moment has an average duration of 8.09 seconds.

ActivityNet-Captions [24] is collected for the dense video captioning task from ActivityNet [1] where the videos are associated with 200 activity classes, and the content is more diverse compared to Charades-STA. The ActivityNet-Captions dataset consists of 19811 videos with 71957 queries. The average duration of the videos is around 117.75 seconds and each video contains 3.63 annotated moments, and the moment has an average duration of 37.14 seconds. The public split of the ActivityNet-Captions dataset contains a training set and two validation sets val_1 and val_2, including 10009, 4917, 4885 videos separately.

4.1.2 Evaluation Metrics

We adopt "R@n, IoU = μ" and "mIoU" as the evaluation metrics, where "R@n, IoU = μ" denotes the percentage of language queries having at least one result whose intersection over union (IoU) with ground truth is larger than μ in top-n retrieved moments. "mIoU" is the average IoU over all testing samples. We report the results as n $\in \{1, 5\}$ with μ $\in \{0.5, 0.7\}$ for fair IID split comparison following MMN [55] and n$\in \{1\}$ with μ$\in \{0.5, 0.7\}$ and mIoU for fair OOD split comparison with [27].

4.1.3 Implementation Details

We experiment with the MMN [55] as the VMR framework to evaluate our method. Specifically, we apply the pre-training visual extractor of the CLIP (ViT-B/32) [42] as the backbone, and directly feed to the VMR framework to generate proposals. Our VMR framework is similar to MMN, where we only delete the linear layer in the pooling module to maintain the feature structure of CLIP. During training, we only update the parameters of the text encoder to empower the understanding of video change, no additional computation cost is introduced in inference.

We use AdamW [36] optimizer with a learning rate of 1×10^{-4} and a batch size 48 for Charades-STA, a learning rate of 8×10^{-4} and a batch size 48 for ActivityNet-Captions. Following MMN, we early stop the training when we observe the performance on the validation set start to drop. The learning rate of text encoder is always 1/10 of the main model. λ_{vc} and λ_{sd} is set to 0.5 and 0.01.

4.2. Comparison with the SOTAs

In this section, we compare the results of our methods under the VMR task with the baseline MMN [55] and existing SOTAs. To evaluate the importance of generalisability between the vision and text, we report the results under OOD testing and IID testing.

4.2.1 Novel-Word OOD Testing

To validate the generalisation ability of our method to capture unseen words and scenes, we conduct experiments on novel-word OOD testing. Specifically, the novel-word OOD split is recently released by Li et al. [27] where the testing split contains novel words which are not seen in the training split, and the corresponding scenes are not seen as well. We follow the settings in Li et al. [27] to report the performance of Charades-STA[13] and ActivityNet-Captions [24] under R@1.

Novel-word OOD testing results are shown in Table 1. We collect the performance reported by Li et al. [27] and reproduce the baseline model [55] with CLIP features under the same settings. One can see that we outperform the SOTA method by a significant margin. Especially on Charades-STA dataset, we outperform the baseline model MMN [55] with a margin of 2.62%/4.46% under the IoU = 0.5/0.7. We outperform VISA [27] by a margin of 4.12%/7.75%. One can see that on ActivityNet-Captions, the margin is less obvious than Charades-STA (2.21% vs 4.12%), it is partially because ActivityNet-Captions display longer moments than Charades-STA (averaged 37.14s vs 8.09s), so it is more challenging to capture the video change. Also, compared with the 37 long moments ($L_{mom}/L_{vid} \geq 0.5$) in Charades-STA, there are over

Method	Year	Pre-train	Charades-STA			ActivityNet-Captions		
			IoU=0.5	IoU=0.7	mIoU	IoU=0.5	IoU=0.7	mIoU
WSSL [11]	2018	Video&Text Separated	2.79	0.73	7.92	3.09	1.13	7.10
TMN [30]	2018		9.43	4.96	11.23	9.93	5.12	11.38
TSP-PRL [57]	2020		14.83	2.61	14.03	18.05	3.15	14.34
2D-TAN [62]	2020		29.36	13.12	28.47	23.86	10.37	28.88
LGI [41]	2020		26.48	12.47	27.62	23.10	9.03	26.95
VSLNet [61]	2020		25.60	10.07	30.21	21.68	9.94	29.58
VISA [27]	2022		42.35	20.88	40.18	30.14	15.90	**35.13**
MMN [55]	2022	Image-Text Joint	43.85	24.17	39.50	31.05	15.48	33.16
VDI (Ours)	2023		**46.47**	**28.63**	**41.60**	**32.35**	**16.02**	34.32

Table 1. Novel-word testing comparison between our method with other state-of-the-art methods on Charades-STA [13] and ActivityNet-Captions [24]. The "Pre-train" column indicates the types of pre-trained models adopted.

15k in ActivityNet-Captions, which makes it trivial by predicting long predictions instead of capturing the semantics.

Obtaining superior OOD performance over video-based (Kinetics[23]) pre-training models demonstrates that our method can take advantage of the image-text pre-training feature and obtain generalisable correlations to unseen words and scenes.

4.2.2 Original Split Testing

To further evaluate the effectiveness of our method, we also conduct extensive experiments on the standard testing, where the training and testing splits share independent and identical distribution.

In Table 2, we include the recently reported results of SOTA methods as well as their visual pre-training data (Vis.P.T). One can see that when replacing the MMN from separated pre-training with joint pre-training features, the performance increases from 47.31% to 50.48%, indicating that the correlation between the vision and text is essential. With our proposed method to empower the model's understanding of actions, the performance further improves to 52.32%, outperforming the baseline model with a large margin of 5.01%. From the results of ActivityNet-Captions, we can see that even though there is a performance drop from video-based pre-training to image-based pre-training (48.59% vs 46.89%), our method can fill the gap by injecting the video change understanding into the model.

To evaluate the learning of correlations, we compare with existing methods with a specific focus on image-based pre-training datasets[7]. As one can see from Table 2, not only can we outperform SOTAs by a large margin, and we can narrow down the gap between the image-based pre-training feature and video-based pre-training feature.

4.2.3 Ablation Study

In this section, we perform in-depth ablations to evaluate the effectiveness of each component in VDI on Charades-STA dataset [13] with novel-word splits [27]. We report the performance under R@1 for IoU $\in \{0.5, 0.7\}$ and mIoU.

Visual Context Injection. To evaluate that it is essential to inject visual context information into the text embeddings, we study different types of visual context generations as shown in Table 3. We take the MMN with CLIP pre-training features as our baseline. The superior performances yielded by the models with visual context injection over the baseline demonstrate the effectiveness of the design. Moreover, we observe that using static query to probe the videos ($\mathcal{L}^{vc}$ w/ Q^s) produces better results than probing by the complete sentence ($\mathcal{L}^{vc}$ w/ Q) and the pure visual context generation ($\mathcal{L}^{vc}$ w/o Q) without text. This implies the importance to avoid correlating the text descriptions of video changes with irrelevant visual context.

Spatial Dynamic Injection. To evaluate that spatial dynamic information is essential for the text encoder, we study two types of dynamic modelling, including LSTM [18] and Transformer Encoder [52]. As shown in Table 3, by introducing spatial dynamic information, both a recurrent network ($\mathcal{L}^{sd}$ w/ LSTM) or a transformer encoder ($\mathcal{L}^{sd}$ w/ Trans) is ready to contribute to a more precise VMR.

With a combination of $\mathcal{L}^{vc}$ and $\mathcal{L}^{sd}$, the performance improves to 46.47%, which further demonstrates the effectiveness of VDI.

Dynamics Awareness. We further evaluate our baseline and VDI models by using either the complete sentences Q, the static Q^s or the dynamic Q^d queries to retrieve video moments on Charades-STA. As shown in Fig. 3, the baseline MMN model yields better results when retrieving video

Method	Charades-STA					ActivityNet-Captions				
	Vis.P.T	R@1, IoU=0.5	R@1, IoU=0.7	R@5, IoU=0.5	R@5, IoU=0.7	Vis.P.T	R@1, IoU=0.5	R@1, IoU=0.7	R@5, IoU=0.5	R@5, IoU=0.7
VideoBert [50]	Video-Text	32.70	19.50	68.10	46.20	Video-Text	37.20	21.00	66.70	53.60
MIL-NCE [39]		37.00	21.20	74.30	50.40		41.80	24.50	73.50	57.70
LocVTP [2]		43.60	26.30	81.90	55.30		48.20	30.50	80.10	64.70
MGSL [32]	Video	63.98	41.03	93.21	63.85	Video	51.87	31.42	82.60	66.71
D-TSG [33]		65.05	42.77	94.42	65.16		54.29	33.64	86.58	69.36
2D-TAN [62]	Image	39.70	23.31	80.32	51.26		44.51	26.54	77.13	61.96
VSL-Net [61]		39.20	20.80	-	-		43.22	26.16	-	-
CBLN [31]		43.67	24.44	**88.39**	56.49		48.12	27.60	79.32	63.41
DCM [58]		47.80	28.00	-	-		44.90	27.70	-	-
DRN [60]		42.90	23.68	87.80	54.87		45.45	24.36	77.97	50.30
MMN [55]		47.31	27.28	83.74	58.41		48.59	29.26	79.50	64.76
MMN [55]	Image-Text	50.48	29.65	85.27	60.67	Image-Text	46.89	27.26	78.32	63.47
VDI (Ours)		**52.32**	**31.37**	87.03	**62.30**		**48.09**	**28.76**	**79.69**	**64.88**

Table 2. Comparisons to the state-of-the-art methods on the standard splits of VMR benchmark datasets. The "Vis.P.T" column indicates that the video encoders adopted are pre-trained by only videos [23] ("Video"), only images [7] ("Image"), video-text pairs [38] ("Video-Text") and Image-text pairs [42] ("Image-Text"). The best performances among image-based pre-training methods are highlighted in bold.

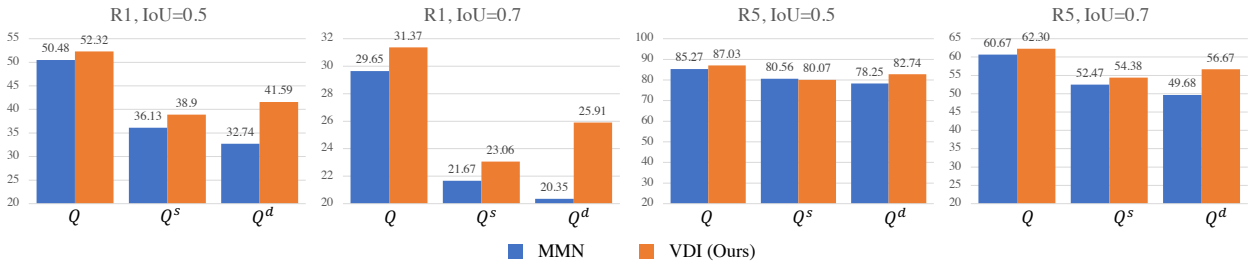

Figure 3. Video Moment Retrieval by the complete Q, the static Q^s or the dynamic Q^d sentence descriptions.

Method	R@1, IoU=0.5	R@1, IoU=0.7	mIoU
Baseline	43.85	24.17	39.50
$\mathcal{L}^{vc}$ w/o Q	43.12	25.32	39.43
$\mathcal{L}^{vc}$ w/ Q	45.02	27.63	40.40
$\mathcal{L}^{vc}$ w/ Q^s	45.47	**29.35**	40.61
$\mathcal{L}^{sd}$ w/ LSTM	45.32	25.76	40.07
$\mathcal{L}^{sd}$ w/ Trans	44.60	26.06	40.09
$\mathcal{L}^{vc} + \mathcal{L}^{sd}$	**46.47**	28.63	**41.60**

Table 3. Ablation studies on visual context injection $\mathcal{L}^{vc}$ with different text queries and spatial dynamic injection $\mathcal{L}^{sd}$ with different sequence analysis models.

moments by static queries than the dynamic ones while our VDI model is in opposed. This verifies the sensitivity of VDI to video changes by correlating visual context and spatial dynamics with text and explains its improvements to the baseline on all VMR tasks.

5. Conclusion

In this paper, we propose to learn universal visual-textual correlations for video moment retrieval (VMR). To address the limitation that the image-text pre-training methods are less sensitive to video changes, we design visual context and spatial dynamic injection to the text encoder with an emphasis on the words describing video changes. By doing so, the potentially relevant visual and motion patterns in videos are encoded in the corresponding text embeddings, enabling more accurate video-text alignments. Experiments on two important datasets (Charades-STA and ActivityNet-Captions) prove that VDI can learn both effective and generic visual-text correlations. Moreover, the comparison between the before and after visual-dynamic injection demonstrate the sensitivity of VDI to video changes.

Acknowledgements

This work was supported by the China Scholarship Council, the Alan Turing Institute Turing Fellowship, Veritone, Adobe Research and Zhejiang Lab (NO. 2022NB0AB05).

References

[1] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, pages 961–970, 2015. 6

[2] Meng Cao, Tianyu Yang, Junwu Weng, Can Zhang, Jue Wang, and Yuexian Zou. Locvtp: Video-text pre-training for temporal localization. In *ECCV*, pages 38–56, 2022. 2, 3, 8

[3] Jingyuan Chen, Xinpeng Chen, Lin Ma, Zequn Jie, and Tat-Seng Chua. Temporally grounding natural sentence in video. In *EMNLP*, pages 162–171, 2018. 1

[4] Xing Cheng, Hezheng Lin, Xiangyu Wu, Fan Yang, and Dong Shen. Improving video-text retrieval by multi-stream corpus alignment and dual softmax loss. *arXiv preprint arXiv:2109.04290*, 2021. 3

[5] Tawanda Blessing Chiyangwa, Judy Van Biljon, and Karen Renaud. Natural language processing techniques to reveal human-computer interaction for development research topics. In *ICARTI*, pages 1–7, 2021. 1

[6] Marcos V Conde and Kerem Turgutlu. Clip-art: contrastive pre-training for fine-grained art classification. In *CVPR*, pages 3956–3960, 2021. 3

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 7, 8

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186, 2019. 1

[9] Aniqa Dilawari, Muhammad Usman Ghani Khan, Yasser D Al-Otaibi, Zahoor-ur Rehman, Atta-ur Rahman, and Yunyoung Nam. Natural language description of videos for smart surveillance. *Applied Sciences*, 11(9):3730, 2021. 1

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 4

[11] Xuguang Duan, Wenbing Huang, Chuang Gan, Jingdong Wang, Wenwu Zhu, and Junzhou Huang. Weakly supervised dense event captioning in videos. *NeurIPS*, 31, 2018. 7

[12] A Frome, GS Corrado, J Shlens, et al. A deep visual-semantic embedding model. *NeurIPS*, pages 2121–2129. 2

[13] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *ICCV*, pages 5267–5275, 2017. 1, 2, 6, 7

[14] Soham Ghosh, Anuva Agarwal, Zarana Parekh, and Alexander G Hauptmann. Excl: Extractive clip localization using natural language descriptions. In *NAACL*, pages 1984–1990, 2019. 1, 2

[15] Dongliang He, Xiang Zhao, Jizhou Huang, Fu Li, Xiao Liu, and Shilei Wen. Read, watch, and move: Reinforcement learning for temporally grounding natural language descriptions in videos. In *AAAI*, volume 33, pages 8393–8400, 2019. 2

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 4

[17] Lisa Anne Hendricks and Aida Nematzadeh. Probing image-language transformers for verb understanding. In *ACL-IJCNLP*, pages 3635–3644, 2021. 2, 3

[18] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 7

[19] Jiabo Huang, Yang Liu, Shaogang Gong, and Hailin Jin. Cross-sentence temporal and semantic relations in video activity localisation. In *ICCV*, pages 7199–7208, 2021. 2

[20] Jiabo Huang, Hailin Jin, Shaogang Gong, and Yang Liu. Video activity localisation with uncertainties in temporal boundary. In *ECCV*, pages 724–740, 2022. 2

[21] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, pages 4904–4916, 2021. 2

[22] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. In *ECCV*, pages 105–124, 2022. 3

[23] 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. 7, 8

[24] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *ICCV*, pages 706–715, 2017. 6, 7

[25] Gihyun Kwon and Jong Chul Ye. Clipstyler: Image style transfer with a single text condition. In *CVPR*, pages 18062–18071, 2022. 2

[26] Yann LeCun, Bernhard Boser, John Denker, Donnie Henderson, Richard Howard, Wayne Hubbard, and Lawrence Jackel. Handwritten digit recognition with a back-propagation network. *NeurIPS*, 2, 1989. 4

[27] 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, 7

[28] 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. *NeurIPS*, 34:9694–9705, 2021. 2

[29] Ziyi Lin, Shijie Geng, Renrui Zhang, Peng Gao, Gerard de Melo, Xiaogang Wang, Jifeng Dai, Yu Qiao, and Hongsheng Li. Frozen clip models are efficient video learners. In *ECCV*, pages 388–404, 2022. 3

[30] Bingbin Liu, Serena Yeung, Edward Chou, De-An Huang, Li Fei-Fei, and Juan Carlos Niebles. Temporal modular networks for retrieving complex compositional activities in videos. In *ECCV*, pages 552–568, 2018. 7

[31] Daizong Liu, Xiaoye Qu, Jianfeng Dong, Pan Zhou, Yu Cheng, Wei Wei, Zichuan Xu, and Yulai Xie. Context-aware biaffine localizing network for temporal sentence grounding. In *CVPR*, pages 11235–11244, 2021. 2, 8

[32] Daizong Liu, Xiaoye Qu, Xing Di, Yu Cheng, Zichuan Xu, and Pan Zhou. Memory-guided semantic learning network for temporal sentence grounding. In *AAAI*, volume 36, pages 1665–1673, 2022. 2, 8

[33] Daizong Liu, Xiaoye Qu, and Wei Hu. Reducing the vision and language bias for temporal sentence grounding. In *ACM MM*, pages 4092–4101, 2022. 2, 8

[34] Song Liu, Haoqi Fan, Shengsheng Qian, Yiru Chen, Wenkui Ding, and Zhongyuan Wang. Hit: Hierarchical transformer with momentum contrast for video-text retrieval. In *ICCV*, pages 11915–11925, 2021. 2

[35] Alexander Long, Wei Yin, Thalaiyasingam Ajanthan, Vu Nguyen, Pulak Purkait, Ravi Garg, Alan Blair, Chunhua Shen, and Anton van den Hengel. Retrieval augmented classification for long-tail visual recognition. In *CVPR*, pages 6959–6969, 2022. 3

[36] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[37] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304, 2022. 3

[38] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*, pages 2630–2640, 2019. 2, 3, 8

[39] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *CVPR*, pages 9879–9889, 2020. 8

[40] Yasuhide Mori, Hironobu Takahashi, and Ryuichi Oka. Image-to-word transformation based on dividing and vector quantizing images with words. In *First international workshop on multimedia intelligent storage and retrieval management*, pages 1–9, 1999. 2

[41] Jonghwan Mun, Minsu Cho, and Bohyung Han. Local-global video-text interactions for temporal grounding. In *CVPR*, pages 10810–10819, 2020. 7

[42] 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. 2, 3, 6, 8

[43] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *CVPR*, pages 18082–18091, 2022. 2

[44] 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. 2

[45] Edward Sapir. Speech as a personality trait. *American journal of sociology*, 32(6):892–905, 1927. 1

[46] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*, pages 510–526, 2016. 6

[47] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 4

[48] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 1

[49] Sanjay Subramanian, William Merrill, Trevor Darrell, Matt Gardner, Sameer Singh, and Anna Rohrbach. Reclip: A strong zero-shot baseline for referring expression comprehension. In *ACL*, pages 5198–5215, 2022. 3

[50] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *ICCV*, pages 7464–7473, 2019. 2, 8

[51] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, pages 4489–4497, 2015. 2

[52] 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. 3, 4, 7

[53] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021. 3

[54] Weining Wang, Yan Huang, and Liang Wang. Language-driven temporal activity localization: A semantic matching reinforcement learning model. In *CVPR*, pages 334–343, 2019. 2

[55] 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, 6, 7, 8

[56] Jason Weston, Samy Bengio, and Nicolas Usunier. Wsabie: Scaling up to large vocabulary image annotation. In *IJCAI*, 2011. 2

[57] Jie Wu, Guanbin Li, Si Liu, and Liang Lin. Tree-structured policy based progressive reinforcement learning for temporally language grounding in video. In *AAAI*, volume 34, pages 12386–12393, 2020. 7

[58] Xun Yang, Fuli Feng, Wei Ji, Meng Wang, and Tat-Seng Chua. Deconfounded video moment retrieval with causal intervention. In *SIGIR*, pages 1–10, 2021. 2, 8

[59] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 2

[60] Runhao Zeng, Haoming Xu, Wenbing Huang, Peihao Chen, Mingkui Tan, and Chuang Gan. Dense regression network for video grounding. In *CVPR*, pages 10287–10296, 2020. 2, 8

[61] Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Span-based localizing network for natural language video localization. In *ACL*, pages 6543–6554, July 2020. 2, 7, 8

[62] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks for moment localization with natural language. In *AAAI*, volume 34, pages 12870–12877, 2020. 1, 2, 7, 8

[63] Hao Zhou, Chongyang Zhang, Yan Luo, Yanjun Chen, and Chuanping Hu. Embracing uncertainty: Decoupling and debias for robust temporal grounding. In *CVPR*, pages 8445–8454, 2021. 1