Video-Text Pre-training with Learned Regions for Retrieval

Rui Yan[1], **Mike Zheng Shou**[2], **Yixiao Ge**[3], **Jinpeng Wang**[2], **Xudong Lin**[4],
Guanyu Cai[5], **Jinhui Tang**[1*]

[1] Nanjing University of Science and Technology, Jiangsu, China
[2] Show Lab, National University of Singapore, Singapore
[3] Tencent PCG, Beijing, China
[4] Columbia University, New York, USA
[5] Tongji University, Shanghai, China
{ruiyan, jinhuitang}@njust.edu.cn, {mike.zheng.shou, geyixiao831}@gmail.com, wangjp23@mail2.sysu.edu.cn,
xudong.lin@columbia.edu, caiguanyu@tongji.edu.cn

Abstract

Video-Text pre-training aims at learning transferable representations from large-scale video-text pairs via aligning the semantics between visual and textual information. State-of-the-art approaches extract visual features from raw pixels in an end-to-end fashion. However, these methods operate at frame-level directly and thus overlook the spatio-temporal structure of objects in video, which yet has a strong synergy with nouns in textual descriptions. In this work, we propose a simple yet effective module for video-text representation learning, namely **RegionLearner**, which can take into account the structure of objects during pre-training on large-scale video-text pairs. Given a video, our module (1) first quantizes continuous visual features via clustering patch-features into the same cluster according to content similarity, then (2) generates learnable masks to aggregate fragmentary features into regions with complete semantics, and finally (3) models the spatio-temporal dependencies between different semantic regions. In contrast to using off-the-shelf object detectors, our proposed module does not require explicit supervision and is much more computationally efficient. We pre-train the proposed approach on the public WebVid2M and CC3M datasets. Extensive evaluations on four downstream video-text retrieval benchmarks clearly demonstrate the effectiveness of our RegionLearner.

Introduction

Video-Text pre-training (Lei et al. 2021; Bain et al. 2021), which aims to learn transferable representations by aligning the semantics of video and text, has attracted researchers' attention in recent years. It enables a series of downstream video-text tasks, such as video-text retrieval (Rohrbach et al. 2015; Xu et al. 2016), video question answering (Jang et al. 2017), and video captioning (Rohrbach et al. 2015; Xu et al. 2016). The conventional pipeline (Bain et al. 2021; Lei et al. 2021) of video-language pre-training is encoding video and text into the shared feature space followed by the cross-modality modeling. The visual input used in existing methods can be categorized as *whole frames* and *whole frames + explicit object boxes*, as shown in Figure 1.

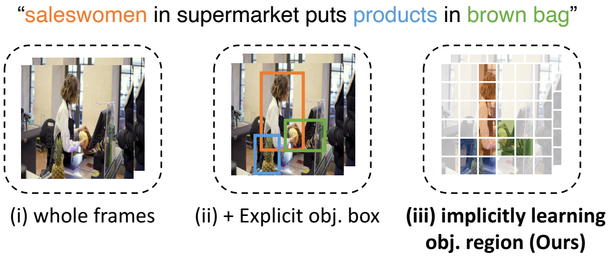

"saleswomen in supermarket puts products in brown bag"
(i) whole frames
(ii) + Explicit obj. box
(iii) implicitly learning obj. region (Ours)

Figure 1: Previous works for video-text pre-training usually extract visual features from the whole frames of video. Inspired by the success of region features in image-text representation (Li et al. 2020b; Chen et al. 2020), some video-based works also extract semantic features (Zhu and Yang 2020; Liu et al. 2019) from explicit object regions. Our motivation is to implicitly learn object regions from raw pixels without any supervision.

i), whole frames (Sun et al. 2019; Lei et al. 2021; Bain et al. 2021): are directly encoded as the video features through the pre-trained 2D or 3D visual backbone. Limited to computing resources, early works (Liu et al. 2019; Sun et al. 2019) extract such video features in an offline way, but recent methods have managed to train the visual backbone on the raw frames in an end-to-end manner (Bain et al. 2021; Lei et al. 2021). Whereas, these methods evenly encode each frame as a number of patch-features, which inevitably destroys the inherent spatio-temporal structure of the visual entities. **ii), whole frames + explicit object box** (Zhu and Yang 2020; Liu et al. 2019; Wang et al. 2022): extracts semantic region features from frames supervised by explicit object boxes detected by the off-the-shelf algorithms for better performance. Intuitively, the visual information of local objects is more effective for semantic alignment. However, existing methods adopt offline region features which are very computationally expensive and not flexible. Beyond that, region features used in these methods also heavily rely on the quality of the off-the-shelf detectors.

These observations motivate us to design a lightweight approach to **implicitly learning object region (as shown in Figure 1 (iii))** without position supervision. We propose

[*] Corresponding author.

a simple yet effective plug-and-play **RegionLearner** module for video-text representation learning. Intuitively, aligning continuous visual content with discrete textual descriptions directly is difficult for models without explicit supervision. Inspired by SOHO (Huang et al. 2021), we first quantize each frame of the video by grouping raw patch features into the same cluster according to visual similarity for better cross-modal alignment. However, patches that make up an object are not necessarily visually similar at the low-level. Therefore, vector quantization based on content similarity tends to assign these dissimilar patches of one object into different clusters, which may destroy the semantic integrity of objects and their dynamic dependencies over time. To this end, based on the quantized visual features, this work further mines object regions with complete spatial semantics and reason the spatio-temporal dependencies among them as follows. `i)`: aggregate fragmentary quantized patch-features to construct integrated semantic regions through multiple learnable region masks; `ii)` : instead of heavy spatio-temporal modeling among dense visual patches from raw features, a lightweight spatio-temporal graph is built on limited learned region features to explore their latent dependencies over time.

Our contributions can be summarized as three-fold. i) To our best knowledge, we are the first to take into account the spatio-temporal structure of objects in the video during video-text pre-training in an end-to-end fashion without supervision. ii) We propose a novel module, namely RegionLearner, to implicitly learn discriminative regions from patch-features. It does not require any explicit supervision and is also computationally efficient, which is friendly for democratizing video-language pre-training technology. Beyond that, the module will be removed for downstream tasks, thus it does not bring any additional parameters or computational overhead. iii) Extensive results on four video-text retrieval downstream benchmarks demonstrate the effectiveness of our approach. As a bonus, the proposed approach also benefits video representations used for visual question answering. We will release relevant code and pre-trained model weights to facilitate the research community.

Related Work

Vision-Language Pre-training Learning visual representation from large-scale video-text pair collections is an emerging research topic. Early methods (Sun et al. 2019; Li et al. 2020a) extract offline visual features from pre-trained video backbones for pre-training. Some recent methods (Lei et al. 2021; Bain et al. 2021) directly extract visual features from raw pixels in an end-to-end fashion. Besides, some works (Liu et al. 2019; Zhu and Yang 2020) attempt to extract regional features from videos as supplementary with the help of off-the-shelf detectors (Anderson et al. 2018) pre-trained on Visual Genome (Krishna et al. 2017). However, frame feature (Lei et al. 2021; Bain et al. 2021) used in existing video-language pre-training methods ignore the complete semantics of visual objects, meanwhile region features heavily rely on the quality of detectors. In this work, we implicitly learn regions from raw pixels without object boxes for video-language pre-training in an end-to-end fashion.

Some recent works for image-text pre-training attempt to get rid of the regional feature (Tan and Bansal 2019; Chen et al. 2020; Kim, Jun, and Zhang 2018) which has been dominant in image-text representations. They either randomly sample some patch-features (Huang et al. 2020) or construct compact discrete representations through visual dictionary (clustering) (Huang et al. 2021) to achieve promising performance. Because semantics involved in language descriptions are visually intertwined, simply clustering (Huang et al. 2021) or randomly sampling (Huang et al. 2020) patch-features will inevitably lead to a large number of fragmented and incomplete areas. In this work, we not only quantizes visual feature into semantic clusters inspired by (Huang et al. 2021), but also further aggregate discrete representation belonging to the same semantic region followed by interactions.

Region-centric Video Representation In the past decade, a large number of deep 2D (Lin, Gan, and Han 2019; Wang et al. 2016) and 3D (Tran et al. 2015; Wang et al. 2018) models have been proposed to extract efficient spatial and temporal representations for videos. Recently, inspired by the success of Transformer in NLP field (Vaswani et al. 2017; Devlin et al. 2018), visual Transformers (Liu et al. 2022; Cheng et al. 2021) are sprung up for video representation. However, these pre-trained video backbones focus more on temporal cues of defined action categories (Kay et al. 2017; Sigurdsson et al. 2016) from the whole frames, and cannot cover rich spatial semantics involved in language descriptions. In this work, we adopt the video backbone pre-trained from ImageNet (Deng et al. 2009), and then aim to learn more fine-grained clues (Tang et al. 2022) of local regions corresponding to the semantics in captions.

In the field of action recognition, to model the dynamic motion of objects, some recent works (Yan et al. 2020; Materzynska et al. 2020) extract the region-centric features from videos according to tracklets of human body or objects. Meanwhile, region-centric features also facilitate semantic alignment between two different modalities in video-language tasks (Lei et al. 2018; Zhu and Yang 2020). However, all these methods rely on the ground truth or detected tracklets of each region (objects or humans). Different from these region-centric works which need explicit positional supervision of regions, we aim at implicitly learning object regions from the raw frames directly.

Approach

Video and Text Encoder

In this work, we extract both visual and textual features in a trainable way, similar to (Bain et al. 2021; Huang et al. 2021). Formally, supposing the input of our approach is a video $\boldsymbol{V} \in \mathbb{R}^{T\times 3\times H\times W}$ which contains T frames of resolution $H \times W$, and the associated tokenized textual description $\boldsymbol{C}$. We obtain their features as,

$$\boldsymbol{F} = E_{\mathrm{V}}(\boldsymbol{V}), \boldsymbol{Y} = E_{\mathrm{C}}(\boldsymbol{C}). \quad (1)$$

Here, $E_{\mathrm{V}}(\cdot)$ and $E_{\mathrm{C}}(\cdot)$ are video and text encoder, respectively. For the text encoder, we choose the current most popular transformer-based architecture and treat the [CLS] token of the last hidden layer as the text feature $\boldsymbol{Y}$. Either

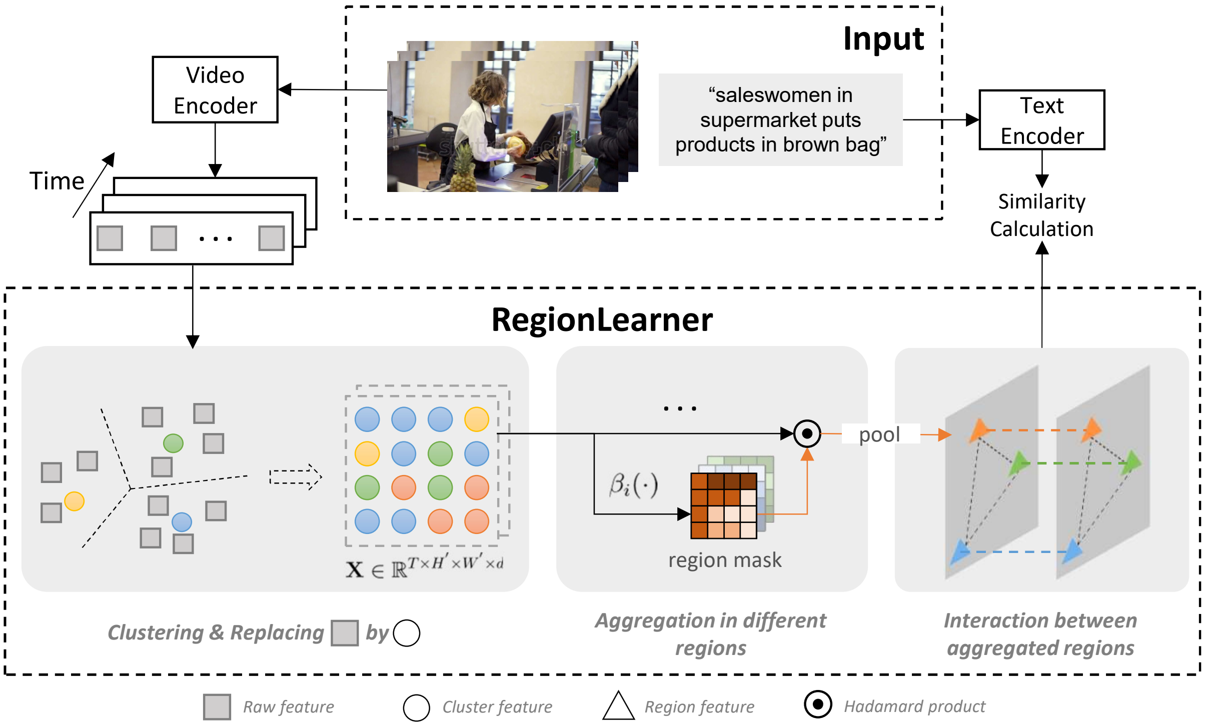

Input
Video Encoder
"saleswomen in supermarket puts products in brown bag"
Text Encoder
Similarity Calculation
Time
RegionLearner
$\mathbf{X} \in \mathbb{R}^{T \times H' \times W' \times d}$
$\beta_i(\cdot)$
pool
region mask
Clustering & Replacing ■ by ○
Aggregation in different regions
Interaction between aggregated regions
■ Raw feature
○ Cluster feature
△ Region feature
⊙ Hadamard product

Figure 2: Overview of the proposed approach. Given a video-text pair, we encode them via the video and text encoder respectively. RegionLearner is proposed to identify and leverage the implicit semantic structure of objects/regions in the video, and it is performed in three steps. **i)** clustering the raw features into semantic clusters and replacing raw features with cluster features; **ii)** multiple region masks are designed to capture the corresponding regions with completed semantics from the quantized feature map and generate a few delicate region features; **iii)** spatio-temporal dependencies between each region feature can be easily dug via self-attention thanks to the limited number of regions. Finally, the video representation generated from Region Learner is fused to compute the similarity with textual representation.

convolution or transformer can be used as a video encoder. Here, we extract visual representations from a video $\boldsymbol{V}$ via ViT (Dosovitskiy et al. 2020; Bertasius, Wang, and Torresani 2021) with the patch size of P, thus we obtain video feature $\boldsymbol{F} \in \mathbb{R}^{T \times L \times d}$ where $L = HW/P^2$.

RegionLearner

Visual Quantization While language used by humans is inherently discrete, the visual information from video is continuous and diverse. Trying to align concepts from these two modalities is already difficult, let alone without any auxiliary supervision (e.g., semantic object annotation). Inspired by (Oord, Vinyals, and Kavukcuoglu 2017; Huang et al. 2021), we believe that quantifying raw continuous visual features to discrete representations will make video-language understanding easier. In this work, we cluster the visual features extracted from patches of each frame based on content similarity in the global space and then replace the raw visual feature with the nearest cluster.

Formally, we first define M learnable clusters as $\{\boldsymbol{c}_0, \boldsymbol{c}_1, \cdots, \boldsymbol{c}_M\}$ in which $\boldsymbol{c}_m \in \mathbb{R}^d$. Our main idea is to aggregate similar visual tokens into the shared cluster in the global space (over the entire dataset). Given a f_s^t at the s-th spatial position and t-th time-step, we update it with the most similar cluster as,

$$f_s^{t'} = \boldsymbol{c}_{m^*}, \quad \text{where} \quad m^* = \text{argmin}_m \, \text{dis}(f_s^t, \boldsymbol{c}_m). \quad (2)$$

Here $\text{dis}(a, b)$ is used to compute the similarity distance between two input features, and it can be implemented by different methods (*e.g.*, euclidean distance, and cosine similarity). In this work, we adopt euclidean distance and update these clusters with momentum learning and stop the gradient on the operation of argmin following (Oord, Vinyals, and Kavukcuoglu 2017; Huang et al. 2021). After that, we can achieve a more compact representation $\boldsymbol{F}' \in \mathbb{R}^{T \times L \times d}$ with the same shape of input feature $\boldsymbol{F}$.

Mining Semantic Region According to content similarity, raw continuous visual features of each frame are quantized through a limited number of cluster representations. However, it is inevitable for the model to represent one visual entity (object/background/people, etc.) via several different cluster features, which will destroy the semantic integrity of the visual entity. Intuitively, each word or phrase in textual descriptions usually refers to a visual instance or region with complete semantic. Therefore, it is necessary to further abstract several visual representations with complete semantics from each quantized feature map. Specifically, inspired by (Ryoo et al. 2021), we extract K region features with complete semantics in space from the sequential features $\boldsymbol{F}^{'}$.

Formally, for each video, we can obtain the semantic representation $\boldsymbol{F}^{'} \in \mathbb{R}^{T\times L\times d}$ in which patch features are arranged in sequence. To find semantic regions, we reshape it back to the original spatial resolution $\boldsymbol{X} \in \mathbb{R}^{T\times H^{'}\times W^{'}\times d}$ where $H^{'} = H/P$ and $W^{'} = W/P$. For each frame t, we aim at learning K region representations $\boldsymbol{S}_t = [\boldsymbol{s}_i]_{i=1}^{K}$ from input frame feature via:

$$\boldsymbol{s}_i = \mathcal{R}(\boldsymbol{X}_t), \tag{3}$$

where $X_t^{'} \in \mathbb{R}^{H^{'}\times W^{'}\times d}$ and $\boldsymbol{s}_i \in \mathbb{R}^{d}$. In this way, region features aggregate informative pixels (*i.e.*, small patches from a video frame) adaptively.

Notably, $\mathcal{R}(\cdot)$ can be implemented via different choices. In this work, we instantiate this function as multiplying the input feature map $\boldsymbol{X}_t$ by a learned spatial attention map and pooling it to a single vector,

$$\boldsymbol{s}_i = \mathcal{R}(\boldsymbol{X}_t) = \texttt{Pool2D}(\boldsymbol{\beta}_i(\boldsymbol{X}_t) \circ \boldsymbol{X}_t). \tag{4}$$

Here, Pool2D and $\circ$ denote the spatial pooling and Hadamard product respectively. $\boldsymbol{\beta} = [\boldsymbol{\beta}_i]_{i=1}^{K}$ is implemented by a 3×3 convolution layer with K channels. After that, for each video, sparse region features $\boldsymbol{S} \in \mathbb{R}^{K\times T\times d}$ are obtained, which allows us to further model the spatio-temporal clues in video.

Video descriptions will inevitably contain dynamic motions, such as "put sth." and "pull sth.", which involves the temporal and spatial dynamic relationships between visual entities rather than only static appearance. Therefore, we further build spatio-temporal dependencies among the region features, which is more efficient and flexible than modeling among raw dense patch-features. The small number of region features allows us to directly build space-time attention as,

$$\boldsymbol{z}_{k,t} = \sum_{k^{'},t^{'}} \boldsymbol{\alpha}(\boldsymbol{\phi}(\boldsymbol{s}_{k,t}), \boldsymbol{\phi}(\boldsymbol{s}_{k^{'},t^{'}}))\boldsymbol{s}_{k^{'},t^{'}}, \tag{5}$$

where $\boldsymbol{\phi}(\cdot)$ is linear embedding, $\boldsymbol{\alpha}(\boldsymbol{a}, \boldsymbol{b})$ computes the attention weights between input via dot product followed by a softmax (Bertasius, Wang, and Torresani 2021). $\boldsymbol{Z} \in \mathbb{R}^{K\times T\times d}$ is the video feature output from RegionLearner plugged on the top of the video encoder.

Objective Function

Following (Bain et al. 2021), in each training batch, paired video-text samples are treated as positives and others are negatives. Formally, we minimise the sum of video-to-text loss ($\mathcal{L}_{\mathrm{v2t}}$) and text-to-video loss ($\mathcal{L}_{\mathrm{t2v}}$) as follows:

$$\mathcal{L}_{\mathrm{v2t}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\boldsymbol{x}_i^{\top}\boldsymbol{y}_j/\tau)}{\sum_j^N \exp(\boldsymbol{x}_i^{\top}\boldsymbol{y}_j/\tau)}, \tag{6}$$

$$\mathcal{L}_{\mathrm{t2v}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\boldsymbol{y}_i^{\top}\boldsymbol{x}_j/\tau)}{\sum_j^N \exp(\boldsymbol{y}_i^{\top}\boldsymbol{x}_j/\tau)}, \tag{7}$$

where $\boldsymbol{x}_i \in \mathbb{R}^{\hat{d}}$ and $\boldsymbol{y}_i \in \mathbb{R}^{\hat{d}}$ are normalized video and text features, respectively. Here, we linear project the [CLS] token of the last layer of video and text encoder into a common dimension as video and text features, respectively. N is the batch-size, and temperature variable τ is used to scale logits.

Experiments

Following the recent work (Bain et al. 2021), we pre-train our model on an affordably large-scale video-text benchmark (WebVid-2M (Bain et al. 2021)) and an image-text benchmark (Google Conceptual Captions (Sharma et al. 2018)). In this work, we evaluate the effectiveness of our pre-trained model on two downstream tasks with nine benchmarks as follows. i) Text-to-Video Retrieval: MSR-VTT (Xu et al. 2016), DiDeMo (Anne Hendricks et al. 2017), LSMDC (Rohrbach et al. 2015, 2017), and MSVD (Chen and Dolan 2011); ii) Video Question Answering: MSRVTT Multiple Choice (Yu, Kim, and Kim 2018), MSRVTT-QA (Xu et al. 2017), and MSVD-QA (Xu et al. 2017).

Comparisons with State-of-the-art

Text-to-Video Retrieval `MSR-VTT.` Our approach exceeds the best no-pretrained method, Support Set (Patrick et al. 2021), by 8.9%. In addition, we are superior to all previous methods pre-trained on HowTo100M (Miech et al. 2019) that is an order of magnitude larger than WebVid2M (Bain et al. 2021) + CC3M (Sharma et al. 2018). Though some of these methods adopt expert features including object (Zhu and Yang 2020), sound (Gabeur et al. 2020; Rouditchenko et al. 2020) and speech (Gabeur et al. 2020) information. Compared with the most related work, Frozen (Bain et al. 2021), our RegionLearner brings significant improvements in text-to-video retrieval. We also provide the results of the zero-shot setting which requires models not to fine-tune on the downstream benchmark. Our approach boosts Frozen by 3.5 on R@1 in Text-Video retrieval and achieves state-of-the-art results against other methods. Beyond that, our approach outperforms OA-Trans which uses explicit object boxes and features extracted from Faster RCNN, in terms of R@1 and R@5 for text-video retrieval. This shows that the learned objects are as good or better than the detected ones, yet there are significant cost savings

`DiDeMo.` As expected, our approach outperforms all existing results on this benchmark. Compared with the most related work (Frozen (Bain et al. 2021)), our approach gains a boost of 1.5% on R@1, which is not very significant. Because this benchmark is collected for Moments Localization (Anne Hendricks et al. 2017), the provided text caption describes only part of the video. Our RegionLearner will be disturbed by many irrelevant frames, which may be an interesting problem for future research.

`LSMDC.` We also report the text-to-video retrieval result on LSMDC in Table 2b. Our approach surpassed all existing methods reported on this benchmark and improves the existing state-of-the-art Frozen by approximately 2.0% in terms of R@1 and MedR.

`MSVD.`Despite this benchmark is relatively small, our

Method	PT dataset	Text ⟹ Video		Video ⟹ Text	
		R@1/@5/@10	MedR	R@1/@5/@10	MedR
JSFusion (Yu, Kim, and Kim 2018)	–	10.2/31.2/43.2	13.0	–/–/–	–
CE (Liu et al. 2019)		20.9/48.8/62.4	6.0	20.6/50.3/64.0	5.3
Support Set (Patrick et al. 2021)		27.4/56.3/67.7	3.0	26.6/55.1/67.5	3.0
HT MIL-NCE (Miech et al. 2019)	HT	14.9/40.2/52.8	9.0	–/–/–	–
ActBERT (Zhu and Yang 2020)		16.3/42.8/56.9	10.0	–/–/–	–
HERO (Li et al. 2020a)		16.8/43.4/57.7	–	–/–/–	–
UniVL (Luo et al. 2020)		21.2/49.6/63.1	6.0	–/–/–	–
MMT (Gabeur et al. 2020)		26.6/57.1/69.6	4.0	27.0/57.5/69.7	3.7
Support Set (Patrick et al. 2021)		30.1/58.5/69.3	3.0	28.5/58.6/71.6	3.0
AVLnet (Rouditchenko et al. 2020)		27.1/55.6/66.6	4.0	–/–/–	–
VidTranslate (Korbar et al. 2020)		14.7/–/52.8	–	–/–/–	–
Noise-Estimation (Amrani et al. 2021)		17.4/41.6/53.6	8.0	–/–/–	–
HIT (Liu et al. 2021)		30.7/60.9/73.2	**2.6**	32.1/62.7/74.1	3.0
DECEMBERT (Tang, Lei, and Bansal 2021)		17.5/44.3/58.6	9.0	–/–/–	–
ClipBERT (Lei et al. 2021)	COCO, VG	22.0/46.8/59.9	6.0	–/–/–	–
Frozen (Bain et al. 2021)	CC, WV	31.0/59.5/70.5	3.0	–/–/–	–
OA-Trans (Wang et al. 2022)	CC, WV	35.8/63.4/76.5	3.0	–/–/–	–
Ours	CC, WV	**36.3/63.9/72.5**	3.0	**35.3/63.5/73.2**	3.0
Zero-shot					
HT MIL-NCE (Miech et al. 2019)	HT	7.5/21.2/29.6	38.0	–/–/–	–
Support Set (Patrick et al. 2021)	HT	12.7/27.5/36.2	24.0	8.7/23.0/31.1	31.0
Frozen (Bain et al. 2021)	CC, WV	18.7/39.5/51.6	10.0	–/–/–	–
Ours	CC, WV	**22.2/43.3/52.9**	**8.0**	**15.3/32.4/42.1**	**17.0**

Table 1: Comparisons with state-of-the-art results on MSR-VTT 1K-A for text-to-video and video-to-text retrieval. COCO, VG, WV2M, CC3M, HT are the abbreviations of COCO Caption (Chen et al. 2015), Visual Genome (Krishna et al. 2017), WebVid2M (Bain et al. 2021), Google Conceptual Captions (Sharma et al. 2018), and HowTo100M (Miech et al. 2019), respectively.

method is still effective and achieves new state-of-the-art results, as shown in Table 2c. In particular, compared with the existing best result from Frozen, our approach improves R@1 by 10.3%. It suggests that our method is still effective on the small-scale downstream dataset, which is very important for pre-training methods to be employed to deal with various real-world tasks.

Video Question Answering We also evaluate our approach on video question answering tasks and reported the results in Table 3. RegionLearner surpasses the prior state-of-the-art VideoCLIP with a performance gain of $+1.4\%$. On MSRVTT-QA and MSVD-QA, our method achieves the accuracy of 38.6% and 39.3%. It's worth noting that DualVGR uses 16 frames twice as many as we use.

Ablation Study

In this section, we study the effectiveness of each component of the proposed approach and the effect of different parameters used in model architecture. All the following experiments are pre-trained on WebVid-2M (Bain et al. 2021) and fine-tuned on the MSR-VTT (Xu et al. 2016) and the results of the 1K-A test set are reported.

Effectiveness of Each Component. To demonstrate the effectiveness of the component of the proposed RegionLearner, we gradually drop each step used in the module and report the results in Table 4a. In general, each component brings improvements. Among them, the improvement of "visual quantization" (step 1) and "Aggregation in different regions" (step 2) is the most obvious, up to $2+\%$. But the gain of "Interaction between aggregated regions" (step 3) is relatively limited, which may be due to the limited spatial-temporal semantics in pre-training data. Through observation, we find that temporal clues contained in the pre-trained dataset (WebVid2M (Bain et al. 2021)) are extremely weak, and it presents almost static visual information.

Different Strategies for Mining Regions. In this work, we compare different strategies used to mine regions with complete semantics in RegionLearner, such as Random Sampling, Empirically Selection, Naive Attention, and the proposed Region Mask. *i), Random Sampling*: samples part of patch features from the feature map X randomly; *Empirically Selection*: selects some high-frequent cluster representations from the map via a threshold of 0.8; *Naive Attention*: performs a simple attention mechanism directly on the feature map X. As reported in Table 4b, all these three methods bring limited improvements or even become worse, but the proposed region mask improves the base method by 2% on R@1. The baseline only quantizes the raw pixels into discrete representations.

Effect of the Number of Regions To determine how many regions the model needs to learn, we set the range of K from 2^0 to 2^6, and the results are shown in Figure 4a. We can see

Method	Text $\Longrightarrow$ Video	
	R@1/@5/@10	**MedR**
S2VT	11.9/33.6/–	13.0
FSE	13.9/36.0/–	11.0
CE (Liu et al. 2019)	16.1/44.1/–	8.3
ClipBERT (Lei et al. 2021)	20.4/44.5/56.7	7.0
Frozen (Bain et al. 2021)	31.0/59.8/**72.4**	3.0
Ours	**32.5**/**60.8**/72.3	**3.0**

(a) DiDeMo.

Method	Text $\Longrightarrow$ Video	
	R@1/@5/@10	**MedR**
JSFusion	9.1/21.2/34.1	36.0
MEE	9.3/25.1/33.4	27.0
CE (Liu et al. 2019)	11.2/26.9/34.8	25.3
MMT (Gabeur et al. 2020)	12.9/29.2/38.8	19.3
Frozen (Bain et al. 2021)	15.0/30.8/39.8	20.0
Ours	**17.1/32.5/41.5**	**18.0**

(b) LSMDC

Method	Text $\Longrightarrow$ Video	
	R@1/@5/@10	**MedR**
VSE	12.3/30.1/42.3	14.0
VSE++ (Faghri et al. 2017)	15.4/39.6/53.0	9.0
Multi. Cues	20.3/47.8/61.1	6.0
CE (Liu et al. 2019)	19.8/49.0/63.8	6.0
Support Set	23.0/52.8/65.8	5.0
Support Set†	28.4/60.0/72.9	4.0
Frozen (Bain et al. 2021)	33.7/64.7/76.3	3.0
Ours	**44.0/74.9/84.3**	**2.0**

(c) MSVD

Table 2: Comparisons with state-of-the-art methods on DiDeMo, LSMDC, and MSVD for text-to-video retrieval. The method pre-trained on HowTo100M is marked '†'. Limited to the space of the table, some references are missing, such as S2VT (Venugopalan et al. 2014), FSE (Zhang, Hu, and Sha 2018), JSFusion (Yu, Kim, and Kim 2018), MEE (Miech, Laptev, and Sivic 2018), VSE (Kiros, Salakhutdinov, and Zemel 2014), Multi. Cues (Mithun et al. 2018), and Support Set (Patrick et al. 2021).

that if K is too large, it may be difficult for the model to find discriminative regions because the module tends to reserve the whole feature map. On the contrary, if K is too small many fragmentary and weak semantics will be discarded in large quantities, leading to poor results. Our approach achieves the best results with $K = 8$ regions.

Effect of the Depth of Spatio-temporal Interaction. We tried to build multiple layers of spatial-temporal dependencies on top of the region features and found that too many layers are not good, as shown in Figure 4b. Probably because the highly abstract regional features are sufficient, and too much spatio-temporal attention will cause the deep model to over-fit on the pre-training dataset. Thus, we use only single-layer attention.

Method	MSRVTT		MSVD
	MC	**QA**	**QA**
JSFusion	83.4	–	–
ActBERT (Zhu and Yang 2020)	85.7	–	–
VideoCLIP (Xu et al. 2021)	92.1	–	–
SSML (Amrani et al. 2021)	–	35.0	35.1
HCRN (Le et al. 2020)	–	35.6	36.1
DualVGR	–	35.5	39.0
ClipBERT (Lei et al. 2021)	88.2	37.4	–
Ours	**93.5**	**38.6**	**39.3**

Table 3: Comparisons with state-of-the-art methods on video question answering. Limited to the space of the table, some references are missing, such as JSFusion (Yu, Kim, and Kim 2018) and DualVGR (Wang, Bao, and Xu 2021).

Method	Text $\Longrightarrow$ Video	
	R@1/@5/@10	**MedR**
RegionLearner	**34.3**/60.2/72.0	**3.0**
w/o step 3	33.5/**60.4**/**72.5**	**3.0**
w/o step 2, 3	31.5/60.2/70.9	**3.0**
w/o step 1,2,3	29.4/56.9/69.0	4.0

(a)

Method	Text $\Longrightarrow$ Video	
	R@1/@5/@10	**MedR**
Baseline	31.5/60.2/70.9	**3.0**
+ Random Sampling	31.2/60.3/72.2	**3.0**
+ Empirically Selection	32.8/59.7/**72.4**	**3.0**
+ Naive Attention	32.8/59.6/69.8	4.0
+ Region Mask	**33.5**/**60.4**/72.5	**3.0**

(b)

Table 4: (a): Effect of each component of RegionLearner. (step 1-3 represents *visual quantization*, *aggregation in different regions* and *interaction between aggregated regions*, respectively.) (b) Effect of different strategies for Aggregation.

Visualization

We also provide some qualitative results in Figure 3 to show the regions learned via the proposed RegionLearner from the pretraining dataset, WebVid-2M (Bain et al. 2021). The second column of each group is the indices map generated by *visual quantization*, and it represents similar visual patches with the same index. It is worth noting that the different colors in this map only represent different index values. As we can see, *visual quantization* is adept at assigning those patches belonging to a large area (such as the backgrounds) into one cluster. However, it seems cannot focus on specific small visual entities in the foregrounds, which are usually associated with local semantics in the text descriptions. After *mining semantic regions*, we can achieve several learned region masks, and here only select two masks for illustration in the last two columns of each group. Bright yellow pixels indicate that the corresponding visual information is more important and the possible regions are annotated by color-

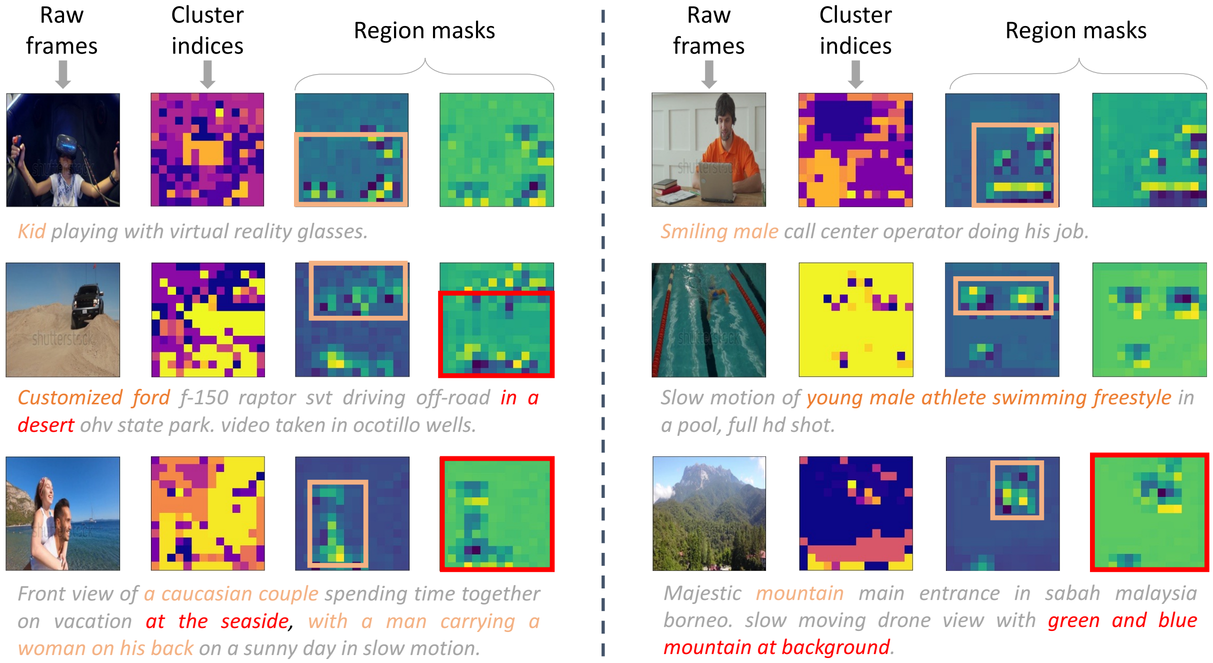
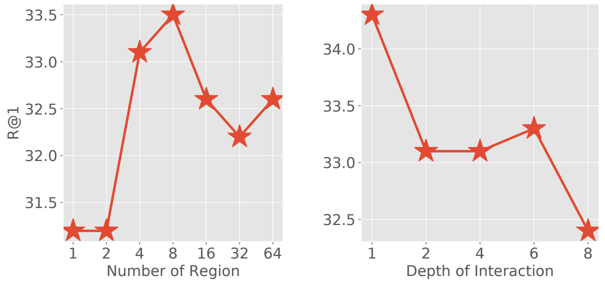

Raw frames
Cluster indices
Region masks
Raw frames
Cluster indices
Region masks
Kid playing with virtual reality glasses.
Smiling male call center operator doing his job.
Customized ford f-150 raptor svt driving off-road in a desert ohv state park. video taken in ocotillo wells.
Slow motion of young male athlete swimming freestyle in a pool, full hd shot.
Front view of a caucasian couple spending time together on vacation at the seaside, with a man carrying a woman on his back on a sunny day in slow motion.
Majestic mountain main entrance in sabah malaysia borneo. slow moving drone view with green and blue mountain at background.

Figure 3: Visualization of the regions learned via the proposed RegionLearner. Each group has a raw frame, the corresponding textual description, a learned map of cluster indices, and two selected learned region masks. We annotate the visual entities in colorful boxes for better understanding. The resolution of these learned maps is 14 × 14. (Best viewed in color.)



(a) (b)

Figure 4: Effect of the number of regions and the depth of spatio-temporal interaction. R@1 performances on the MSR-VTT 1K-A test set are reported.

ful boxes. As we can see, these region masks not only can capture the visual entities in the foregrounds but also reserve discriminative background information for alignment.

Conclusion

To conclude, this work proposes a simple yet effective RegionLearner to mine semantic features from visual objects/entities from pixels for better video-language alignment without any explicit supervision. It first quantizes raw pixels into discrete latent embeddings and then aggregates them into several regions with complete semantics via learnable masks, followed by the spatio-temporal dependencies modeling among these regions. The experimental results of four downstream benchmarks and some visualization results prove the effectiveness and interpretability of our method. We hope that RegionLearner can inspire future work for learning video-text representations in a more fine-grained way.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (Grant No. 2018AAA0102002), and the National Science Foundation for Distinguished Young Scholars of China (Grant No. 61925204). Rui Yan is supported by the China Scholarship Council under Grant 202006840101. Mike Zheng Shou is only supported by the National Research Foundation, Singapore under its NRFF Award NRF-NRFF13-2021-0008.

References

Amrani, E.; Ben-Ari, R.; Rotman, D.; and Bronstein, A. 2021. Noise Estimation Using Density Estimation for Self-Supervised Multimodal Learning. In *AAAI Conference on Artificial Intelligence*, 6644–6652.

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition*, 6077–6086.

Anne Hendricks, L.; Wang, O.; Shechtman, E.; Sivic, J.; Darrell, T.; and Russell, B. 2017. Localizing moments in video with natural language. In *International Conference on Computer Vision*, 5803–5812.

Bain, M.; Nagrani, A.; Varol, G.; and Zisserman, A. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. *International Conference on Computer Vision*.

Bertasius, G.; Wang, H.; and Torresani, L. 2021. Is Space-Time Attention All You Need for Video Understanding? *arXiv preprint arXiv:2102.05095*.

Chen, D.; and Dolan, W. B. 2011. Collecting highly parallel data for paraphrase evaluation. In *Annual Meeting of the Association for Computational Linguistics*, 190–200.

Chen, X.; Fang, H.; Lin, T.-Y.; Vedantam, R.; Gupta, S.; Dollár, P.; and Zitnick, C. L. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Chen, Y.-C.; Li, L.; Yu, L.; Kholy, A. E.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2020. UNITER: UNiversal Image-TExt Representation Learning. arXiv:1909.11740.

Cheng, Y.-B.; Chen, X.; Zhang, D.; and Lin, L. 2021. Motion-transformer: self-supervised pre-training for skeleton-based action recognition. In *ACM International Conference on Multimedia in Asia*, 1–6.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Faghri, F.; Fleet, D. J.; Kiros, J. R.; and Fidler, S. 2017. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*.

Gabeur, V.; Sun, C.; Alahari, K.; and Schmid, C. 2020. Multi-modal transformer for video retrieval. In *European Conference on Computer Vision*, 214–229.

Huang, Z.; Zeng, Z.; Huang, Y.; Liu, B.; Fu, D.; and Fu, J. 2021. Seeing Out of tHe bOx: End-to-End Pre-training for Vision-Language Representation Learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 12976–12985.

Huang, Z.; Zeng, Z.; Liu, B.; Fu, D.; and Fu, J. 2020. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*.

Jang, Y.; Song, Y.; Yu, Y.; Kim, Y.; and Kim, G. 2017. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2758–2766.

Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Kim, J.-H.; Jun, J.; and Zhang, B.-T. 2018. Bilinear Attention Networks. In *Annual Conference on Neural Information Processing Systems*, 1564–1574.

Kiros, R.; Salakhutdinov, R.; and Zemel, R. S. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*.

Korbar, B.; Petroni, F.; Girdhar, R.; and Torresani, L. 2020. Video understanding as machine translation. *arXiv preprint arXiv:2006.07203*.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1): 32–73.

Le, T. M.; Le, V.; Venkatesh, S.; and Tran, T. 2020. Hierarchical conditional relation networks for video question answering. In *IEEE Conference on Computer Vision and Pattern Recognition*, 9972–9981.

Lei, J.; Li, L.; Zhou, L.; Gan, Z.; Berg, T. L.; Bansal, M.; and Liu, J. 2021. Less is more: Clipbert for video-and-language learning via sparse sampling. In *IEEE Conference on Computer Vision and Pattern Recognition*, 7331–7341.

Lei, J.; Yu, L.; Bansal, M.; and Berg, T. 2018. TVQA: Localized, Compositional Video Question Answering. In *Conference on Empirical Methods in Natural Language Processing*, 1369–1379.

Li, L.; Chen, Y.-C.; Cheng, Y.; Gan, Z.; Yu, L.; and Liu, J. 2020a. Hero: Hierarchical encoder for video+ language omni-representation pre-training. In *Conference on Empirical Methods in Natural Language Processing*, 2046–2065.

Li, X.; Yin, X.; Li, C.; Zhang, P.; Hu, X.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; et al. 2020b. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, 121–137.

Lin, J.; Gan, C.; and Han, S. 2019. Tsm: Temporal shift module for efficient video understanding. In *International Conference on Computer Vision*, 7083–7093.

Liu, S.; Fan, H.; Qian, S.; Chen, Y.; Ding, W.; and Wang, Z. 2021. Hit: Hierarchical transformer with momentum contrast for video-text retrieval. In *International Conference on Computer Vision*, 11915–11925.

Liu, Y.; Albanie, S.; Nagrani, A.; and Zisserman, A. 2019. Use what you have: Video retrieval using representations from collaborative experts. In *British Machine Vision Conference*.

Liu, Z.; Ning, J.; Cao, Y.; Wei, Y.; Zhang, Z.; Lin, S.; and Hu, H. 2022. Video swin transformer. In *IEEE Conference on Computer Vision and Pattern Recognition*, 3202–3211.

Luo, H.; Ji, L.; Shi, B.; Huang, H.; Duan, N.; Li, T.; Li, J.; Bharti, T.; and Zhou, M. 2020. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*.

Materzynska, J.; Xiao, T.; Herzig, R.; Xu, H.; Wang, X.; and Darrell, T. 2020. Something-else: Compositional action recognition with spatial-temporal interaction networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1049–1059.

Miech, A.; Laptev, I.; and Sivic, J. 2018. Learning a text-video embedding from incomplete and heterogeneous data. *arXiv preprint arXiv:1804.02516*.

Miech, A.; Zhukov, D.; Alayrac, J.-B.; Tapaswi, M.; Laptev, I.; and Sivic, J. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *International Conference on Computer Vision*, 2630–2640.

Mithun, N. C.; Li, J.; Metze, F.; and Roy-Chowdhury, A. K. 2018. Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In *ACM SIGMM International Conference on Multimedia Retrieval*, 19–27.

Oord, A. v. d.; Vinyals, O.; and Kavukcuoglu, K. 2017. Neural discrete representation learning. In *Annual Conference on Neural Information Processing Systems*, 6309–6318.

Patrick, M.; Huang, P.-Y.; Asano, Y.; Metze, F.; Hauptmann, A.; Henriques, J.; and Vedaldi, A. 2021. Support-set bottlenecks for video-text representation learning. *International Conference on Learning Representations*.

Rohrbach, A.; Rohrbach, M.; Tandon, N.; and Schiele, B. 2015. A dataset for movie description. In *IEEE Conference on Computer Vision and Pattern Recognition*, 3202–3212.

Rohrbach, A.; Torabi, A.; Rohrbach, M.; Tandon, N.; Pal, C.; Larochelle, H.; Courville, A.; and Schiele, B. 2017. Movie description. *International Journal of Computer Vision*, 123(1): 94–120.

Rouditchenko, A.; Boggust, A.; Harwath, D.; Chen, B.; Joshi, D.; Thomas, S.; Audhkhasi, K.; Kuehne, H.; Panda, R.; Feris, R.; et al. 2020. Avlnet: Learning audio-visual language representations from instructional videos. *arXiv preprint arXiv:2006.09199*.

Ryoo, M.; Piergiovanni, A.; Arnab, A.; Dehghani, M.; and Angelova, A. 2021. Tokenlearner: Adaptive space-time tokenization for videos. In *Annual Conference on Neural Information Processing Systems*, volume 34, 12786–12797.

Sharma, P.; Ding, N.; Goodman, S.; and Soricut, R. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Annual Meeting of the Association for Computational Linguistics*, 2556–2565.

Sigurdsson, G. A.; Varol, G.; Wang, X.; Farhadi, A.; Laptev, I.; and Gupta, A. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*, 510–526.

Sun, C.; Myers, A.; Vondrick, C.; Murphy, K.; and Schmid, C. 2019. Videobert: A joint model for video and language representation learning. In *International Conference on Computer Vision*, 7464–7473.

Tan, H.; and Bansal, M. 2019. Lxmert: Learning cross-modality encoder representations from transformers. In *Conference on Empirical Methods in Natural Language Processing*, 5103–5114.

Tang, H.; Yuan, C.; Li, Z.; and Tang, J. 2022. Learning Attention-Guided Pyramidal Features for Few-shot Fine-grained Recognition. *Pattern Recognition*, 108792.

Tang, Z.; Lei, J.; and Bansal, M. 2021. Decembert: Learning from noisy instructional videos via dense captions and entropy minimization. In *The Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2415–2426.

Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; and Paluri, M. 2015. Learning spatiotemporal features with 3d convolutional networks. In *International Conference on Computer Vision*, 4489–4497.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Annual Conference on Neural Information Processing Systems*, 5998–6008.

Venugopalan, S.; Xu, H.; Donahue, J.; Rohrbach, M.; Mooney, R.; and Saenko, K. 2014. Translating videos to natural language using deep recurrent neural networks. *arXiv preprint arXiv:1412.4729*.

Wang, J.; Bao, B.-K.; and Xu, C. 2021. DualVGR: A dual-visual graph reasoning unit for video question answering. *IEEE Transactions on Multimedia*, 24: 3369–3380.

Wang, J.; Ge, Y.; Cai, G.; Yan, R.; Lin, X.; Shan, Y.; Qie, X.; and Shou, M. Z. 2022. Object-aware Video-language Pre-training for Retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition*, 3313–3322.

Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; and Van Gool, L. 2016. Temporal segment networks: Towards good practices for deep action recognition. In *European Conference on Computer Vision*, 20–36.

Wang, X.; Girshick, R.; Gupta, A.; and He, K. 2018. Non-local neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 7794–7803.

Xu, D.; Zhao, Z.; Xiao, J.; Wu, F.; Zhang, H.; He, X.; and Zhuang, Y. 2017. Video question answering via gradually refined attention over appearance and motion. In *ACM International Conference on Multimedia*, 1645–1653.

Xu, H.; Ghosh, G.; Huang, P.-Y.; Okhonko, D.; Aghajanyan, A.; Metze, F.; Zettlemoyer, L.; and Feichtenhofer, C. 2021. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*.

Xu, J.; Mei, T.; Yao, T.; and Rui, Y. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *IEEE Conference on Computer Vision and Pattern Recognition*, 5288–5296.

Yan, R.; Xie, L.; Shu, X.; and Tang, J. 2020. Interactive Fusion of Multi-level Features for Compositional Activity Recognition. *arXiv preprint arXiv:2012.05689*.

Yu, Y.; Kim, J.; and Kim, G. 2018. A joint sequence fusion model for video question answering and retrieval. In *European Conference on Computer Vision*, 471–487.

Zhang, B.; Hu, H.; and Sha, F. 2018. Cross-modal and hierarchical modeling of video and text. In *European Conference on Computer Vision*, 374–390.

Zhu, L.; and Yang, Y. 2020. Actbert: Learning global-local video-text representations. In *IEEE Conference on Computer Vision and Pattern Recognition*, 8746–8755.