EA-VTR: Event-Aware Video-Text Retrieval

Zongyang Ma[1,2,3*], Ziqi Zhang[1*†], Yuxin Chen[1,2,3*], Zhongang Qi[2], Chunfeng Yuan[1], Bing Li[1], Yingmin Luo[2], Xu Li[2], Xiaojuan Qi[5], Ying Shan[2], and Weiming Hu[1,3,4]

[1] MAIS, Institute of Automation, Chinese Academy of Sciences
{mazongyang2020@,chenyuxin2019}ia.ac.cn
{ziqi.zhang,cfyuan,bli,wmhu}@nlpr.ia.ac.cn
[2] ARC Lab, Tencent PCG
{zhongangqi,yingminluo,nelsonxli,yingsshan}@tencent.com
[3] School of Artificial Intelligence, University of Chinese Academy of Sciences
[4] School of Information Science and Technology, ShanghaiTech University
[5] The University of Hong Kong {xjqi@eee.hku.hk}

Abstract. Understanding the content of events occurring in the video and their inherent temporal logic is crucial for video-text retrieval. However, web-crawled pre-training datasets often lack sufficient event information, and the widely adopted video-level cross-modal contrastive learning also struggles to capture detailed and complex video-text event alignment. To address these challenges, we make improvements from both data and model perspectives. In terms of pre-training data, we focus on supplementing the missing specific event content and event temporal transitions with the proposed event augmentation strategies. Based on the event-augmented data, we construct a novel Event-Aware Video-Text Retrieval model, *i.e.*, EA-VTR, which achieves powerful video-text retrieval ability through superior video event awareness. EA-VTR can efficiently encode frame-level and video-level visual representations simultaneously, enabling detailed event content and complex event temporal cross-modal alignment, ultimately enhancing the comprehensive understanding of video events. Our method not only significantly outperforms existing approaches on multiple datasets for Text-to-Video Retrieval and Video Action Recognition tasks, but also demonstrates superior event content perceive ability on Multi-event Video-Text Retrieval and Video Moment Retrieval tasks, as well as outstanding event temporal logic understanding ability on Test of Time task.

1 Introduction

The emergence of web-crawled pre-training datasets [4, 37, 47] has spurred rapid advancements in the field of video-text retrieval. Among the two types of retrieval models, *i.e.*, joint-encoder [15, 27, 30, 68] and dual-encoder [4, 17, 18, 51] models,

* Equal contribution. † Corresponding author.

the latter can achieve real-time retrieval, thus attracting more attention and providing wider practical applications. Enhancing the ability to associate fine-grained video-text information is an effective way to improve the performance of dual-encoder models. While existing works have explored various details such as objects [17, 51], actions [17] and regions [60], the study of video events has been overlooked. A video event refers to the visual activity or scene that occurs within a specific time interval in the video, which is a fundamental component of video and a concept widely used in video understanding works [46, 64]. A video consists of one or more events, as illustrated in Figure 1 (a), where the events *"... street has cars ..."*, *"... subway passes..."*, and *"... people are walking ..."* transpire respectively in the Frames 1, 2, and N of the video.

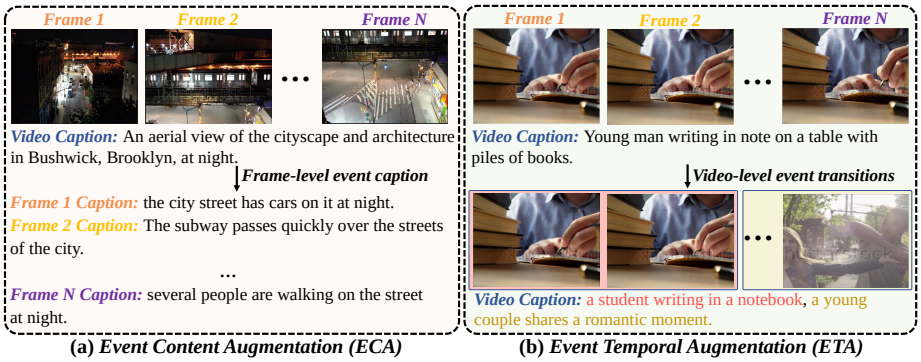

Fig. 1: Examples of missing (a) event content and (b) temporal transitions and corresponding augmentation results. First, the web-crawled video caption in (a) does not contain specific event content. Second, in a video-text pair like (b), the video either lacks event temporal transitions or the caption does not reflect these transitions. Therefore, we propose ECA and ETA to supplement the missing information in both aspects.

Both the current pre-training datasets and elaborated models restrict the ultimate event understanding ability. In terms of data, mainstream pre-training datasets, such as WebVid [4], lack captions about specific event content, *e.g.*, the video caption "An aerial view ..." in Figure 1 (a) does not cover the event *"... subway passes ..."* in Frame 2. SViTT [32] also notes that many WebVid captions are associated with static backgrounds rather than foreground events. Furthermore, a considerable portion of video-text pairs do not exhibit obvious event temporal transitions, as the video at the top of Figure 1 (b) only contains a single event. To support this point, we have counted that 28% of sampled 10,000 WebVid videos have only one event, by calculating whether any video frame has a similarity lower than the pre-defined threshold with the first frame (details in supplementary material). Similarly, SViTT [32] claims that some WebVid videos consist of only simple motions without obviously dynamic changes. In terms of the models, the commonly used video-level contrastive learning directly aligns

global video and text representations, thus falls short in capturing detailed event content and complex event temporal alignment simultaneously, as evidenced by the comparisons in Table 9 and the relevant analysis.

To achieve a video-text retriever with robust event comprehension, we focus on enriching the event information in the pre-training data and improving the model's ability to perform event cross-modal alignment. To obtain the content of multiple events, we first divide the video into several clips uniformly (consistent with the video encoder in video-text retrieval models), and then extract one frame from each clip and get the frame caption through an image captioner to capture the event content of the clip. This approach is based on a validated assumption through our manual verification, which posits that short videos inherently exhibit limited dynamic changes, and thus the divided clips present even less obvious dynamic changes, allowing a single frame to represent the clip adequately. We refer to this process as Event Content Augmentation (ECA) and the demonstration is shown in Figure 1 (a) for the captions *"the city ..."*, *"the subway ..."* and *"several people ..."* corresponding to Frames 1, 2, and *N*. To construct video-text pairs with explicit event temporal transitions to train the model's event temporal understanding ability, we randomly intercept consecutive frames from two videos to synthesize into a new video, and the intercepted frame captions are also concatenated as the video caption, resulting in a synthesized video-text pair with clear event transitions. We refer to the above process as Event Temporal Augmentation (ETA) and give an example in Figure 1 (b).

Based on the event-augmented pre-training data obtained by applying ECA and ETA, we further introduce a novel Event-Aware Video-Text Retrieval model, *i.e.*, EA-VTR, which emphasizes Event Content Learning (ECL) and Event Temporal Learning (ETL) in addition to the common video-level contrastive learning. EA-VTR starts with adding several Frame [CLS] tokens to the video encoder, which interact with the visual patches of the corresponding frames to aggregate intra-frame information, allowing EA-VTR to compute both video-level and frame-level visual features with a negligible increase in computational overhead. ECL then aligns frame features with the text features of the frame captions obtained through ECA to enhance the understanding of event content, and ETL aligns video-text features of the temporally rich synthesized video-text pairs obtained through ETA to improve event temporal logic understanding. During the test, only video-level representation is taken to calculate similarity with the text query, thus ensuring efficient retrieval.

We first evaluate our method on text-to-video retrieval task, and the results demonstrate that our method not only significantly outperforms the best dual-encoder methods on various datasets, such as achieving a 4.7% and 5.3% improvement in R@10 on MSRVTT and DiDeMo under zero-shot settings, but also exhibit comparable performance to the SOTA joint-encoder methods with over 30x speed advantage. Furthermore, our method generalizes well on video action recognition tasks, with a 7.9% increase in mean accuracy on UCF101 compared to the best method. Moreover, we thoroughly evaluate the event understanding ability of our method, including Multi-event Video-Text Retrieval [64] and

Video Moment Retrieval tasks that involve perceiving all events and specific events within a video, as well as Test of Time task [2] that focus on understanding the temporal logic between multiple events. The results demonstrate that our method outperforms previous approaches in all three tasks, encompassing both content and temporal logic aspects of event understanding. The contributions of this work are listed as follows:

- We propose Event Content Augmentation and Event Temporal Augmentation to complement the event content and temporal transitions lacking in the pre-training data.
- We introduce a novel Event-Aware Video-Text Retrieval model, which achieves superior cross-modal retrieval and event understanding abilities by performing Event Content Learning and Event Temporal Learning.
- The results of Text-to-Video Retrieval and Video Action Recognition tasks show our effectiveness in video-text retrieval, while the results of Multi-event Video-Text Retrieval, Video Moment Retrieval, and Test of Time tasks further validate our superiority in video event understanding.

2 Related Work

Pre-training for Video-Text Retrieval. Previous works for video-text retrieval can be divided into joint-encoder [15, 22, 27, 30, 31, 52, 55, 62, 63, 68] and dual-encoder [3, 4, 16–19, 33, 36, 39, 44, 48, 51, 56, 60, 61] models. The consensus within the cross-modal retrieval community [9, 17, 35] is that joint-encoder models perform cross-modal interactions at the expense of speed to achieve higher accuracy, whereas dual-encoder models adopt two individual encoders to extract video [4, 5, 34] and text [12, 45] representations separately and then calculate their cross-modal cosine similarities, enabling efficient real-time retrieval and gaining wide practical application. To improve the generalization of efficient dual-encoder models on downstream tasks, some works [17, 51, 60] have explored improving their ability to associate fine-grained video-text information. MCQ [17] constructs multiple-choice questions about objects and actions in the video and forces the model to select the correct option. OA-Trans [51] aligns object regions, obtained through the object detector [42], with their corresponding labels. RegionLearner [60] clusters similar visual patches to approximate objects to align with text representations. However, current works overlook the study of video events, which are fundamental components of videos. Understanding events not only directly benefits tasks that require event perception [64] and logical judgment [2], but also holds the potential to improve cross-modal retrieval performance by enhancing the grasp of video details. To unleash the potential of dual-encoder models, we are committed to enhancing the event understanding ability of the retriever from both data and model perspectives in this work.

Data Augmentation for Video-Text Pre-Training. There are some works [54, 58, 59, 66] devoted to augmenting the video or text information in the training data to obtain improved video-text pre-training models. In terms of text, CLIP-ViP claims the language domain gap between text in pre-training and down-

stream datasets will reduce the generalization ability, and thus use an image-text pre-training model [53] to generate single-frame captions for HowTo100M [37]. Cap4Video utilizes ZeroCap [50], which combines CLIP [40] and GPT2 [41], to create a caption for each video, simulating the real-world scenario where videos are accompanied by related textual information (such as titles and tags) for retrieval. LAVILA [66] generates additional narrations for long videos from the Ego4D dataset [10, 20] through a fine-tuned large language model [41]. On the video front, BSP [58] employs a strategy of fusing two videos to synthesize a new video for temporal localization pre-training. In contrast to the unimodal data augmentation in the aforementioned works, we perform data augmentation on both video and text modalities to comprehensively supplement event content and temporal transition information in pre-training data.

3 Method

In this section, we first introduce the proposed event augmentation strategy in section 3.1. Then we present the novel Event-Aware Video-Text Retrieval model in section 3.2. Finally, we give the model training and inference in section 3.3.

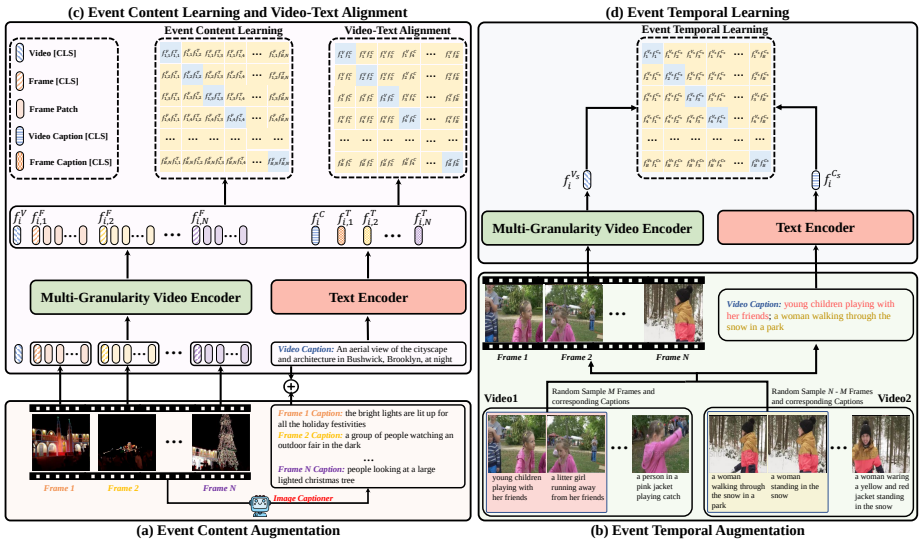

Fig. 2: Overview of the proposed Event Content Augmentation (a) and Event Temporal Augmentation (b) to augment the event information in the pre-training dataset, and EA-VTR model using Event Content Learning (c) and Event Temporal Learning (d) to learn from the augmented data.

3.1 Event Augmentation for Pre-Training

Event augmentation aims to provide data with both rich event content and temporal transitions for model pre-training, which includes Event Content Augmentation and Event Temporal Augmentation.

3.1.1 Event Content Augmentation As shown in Figure 2 (a), Event Content Augmentation (ECA) captures the content of events occurring in different clips (or different segments, different durations) of the video V_i with web-crawled caption C_i. We first uniformly divide the video V_i into N non-overlapping clips, and then sample a single frame from each clip to form the frame set $F_i = \{F_{i,j}\}_{j=1}^{N}$, which is the same input strategy as the video encoder in the retrieval models [4, 17, 18]. The sampled frames F_i are then fed into an off-the-shelf image captioner to obtain the corresponding frame captions $T_i = \{T_{i,j}\}_{j=1}^{N}$, which are used to capture the event content within each clip. The feasibility of using a single frame to represent a clip of short video has been validated through our manual verification. To generate more diverse frame captions for training a model with strong generalization ability, we employ a non-deterministic Top-p (Nucleus) sampling [21] generation strategy, ultimately equipping the video V_i with N frame-level image-text pairs.

3.1.2 Event Temporal Augmentation The process of Event Temporal Augmentation (ETA) is shown in Figure 2 (b). In a batch, two randomly selected videos, named Video1 and Video2, have M_1 and M_2 ($M_1, M_2 \leq N$) frames randomly extracted from multiple consecutive video clips, denoted as $F_1 = \{F_{1,j}\}_{j=1}^{M_1}$ and $F_2 = \{F_{2,j}\}_{j=1}^{M_2}$. The corresponding frame captions, $T_1 = \{T_{1,j}\}_{j=1}^{M_1}$ and $T_2 = \{T_{2,j}\}_{j=1}^{M_2}$, are also selected. Then we perform Video Concatenation for video synthesizing.

Video Concatenation. In this scheme, the frame numbers M_1 and M_2 are supposed to satisfy $M_1 + M_2 = N$, and a newly synthesized video V_i^s is obtained by concatenating all frames of F_1 and F_2:

$$V_i^s = \text{Concat}(F_1, F_2). \tag{1}$$

Caption for the synthesized video. The accompanying caption C_i^s of synthesized video V_i^s is construct by concatenating two frame captions sampled from T_1 and T_2 respectively.

Since the synthesized video contains evident event transition changes, the video events exhibit distinct separability in the temporal dimension, which requires the model to understand the sequential logic between events to effectively comprehend the synthesized video.

3.2 Event-Aware Video-Text Retrieval

With the foundation of the aforementioned augmented pre-training data, the novel Event-Aware Video-Text Retrieval model can be unfolded according to the following outline.

3.2.1 Multi-Granularity Video Encoder Since the pre-training data augmented by ECA and ETA contains both frame-level image-text pairs as well as web-crawled and synthesized video-level video-text pairs, this naturally requires the video encoder to output both video-level and frame-level visual representations to enable subsequent cross-modal alignment at multi-granularity. However, the video encoders of most current video-text retrieval models are TimeSformer [5] or its variants [4, 18], which only support providing video-level visual representations.

To address this issue, we make a simple but generic modification to TimeSformer or its variants [4, 18], enabling it to provide video and frame representations simultaneously. As shown in Figure 2 (c), in contrast to the previous video encoder, we concatenate a new Frame [CLS] token embedding $e^F_{i,j}$ with random initialization before the frame patch embeddings $e^P_{i,j}$ of the video frame $F_{i,j}$. After incorporating the spatio-temporal position information into the aforementioned embeddings, a video [CLS] embedding e^V_i is finally concatenated at the beginning, and the resulting embeddings $e_i = \text{Concat}(e^V_i; \{e^F_{i,j}; e^P_{i,j}\}^N_{j=1})$ are then fed into the video encoder E_V to be converted into representations f_i:

$$f_i = E_V(e_i), \tag{2}$$

where $f_i = \text{Concat}(f^V_i; \{f^F_{i,j}; f^P_{i,j}\}^N_{j=1})$, thus we can directly extract video representation f^V_i and frame representations $\{f^F_{i,j}\}^N_{j=1}$. E_V adopts the similar architecture as [4]. On the one hand, this enables the new Frame [CLS] tokens to primarily interact with the visual patches within the respective frames to form frame representations. On the other hand, the additional computational cost introduced by the N new Frame [CLS] tokens (since N is much smaller than the number of visual patches) is negligible.

3.2.2 Event Content Learning We first perform Event Content Learning (ECL) by aligning video frames with the corresponding frame captions to capture the detailed event content of the video. As shown in Figure 2 (c), each frame caption in $\{T_{i,j}\}^N_{j=1}$ is individually fed into the text encoder E_T, yielding the corresponding text features $\{f^T_{i,j}\}^N_{j=1}$. Subsequently, we employ contrastive learning [23,38] to align the frame visual features with the frame caption features, which is defined as follows:

$$\begin{aligned}
\mathcal{L}^{F2T}_{i,j} &= -log\frac{\exp(f^F_{i,j}, f^T_{i,j}, \tau_c)}{\sum^B_{i'=1}\sum^N_{j'=1}\exp(f^F_{i,j}, f^T_{i',j'}, \tau_c)}, \\
\mathcal{L}^{T2F}_{i,j} &= -log\frac{\exp(f^T_{i,j}, f^F_{i,j}, \tau_c)}{\sum^B_{i'=1}\sum^N_{j'=1}\exp(f^T_{i,j}, f^F_{i',j'}, \tau_c)},
\end{aligned} \tag{3}$$

where $\exp(x, y, \tau) = e^{x^\top y/\tau}$, τ_c is a temperature hyper-parameter, and B is the batch size. The total loss for ECL is defined as:

$$\mathcal{L}_{ECL} = \frac{1}{B \cdot N}\sum^B_{i=1}\sum^N_{j=1}(\mathcal{L}^{F2T}_{i,j} + \mathcal{L}^{T2F}_{i,j})/2. \tag{4}$$

3.2.3 Event Temporal Learning We perform Event Temporal Learning (ETL) by aligning the temporally rich synthesized video and the paired caption. From the process depicted in Figure 2 (d), we can see that the synthesized video V_i^S and its corresponding caption C_i^S are separately passed through the video encoder E_V and text encoder E_T, resulting in their respective features $f_i^{V^s}$ and $f_i^{C^s}$. We then still use contrastive learning to enable cross-modal alignment, which is defined as follows:

$$\begin{aligned}\mathcal{L}_i^{V^s 2C^s} &= -log\frac{\exp(f_i^{V^s}, f_i^{C^s}, \tau_t)}{\sum_{i'=1}^{B}\exp(f_i^{V^s}, f_{i'}^{C^s}, \tau_t)},\\ \mathcal{L}_i^{C^s 2V^s} &= -log\frac{\exp(f_i^{C^s}, f_i^{V^s}, \tau_t)}{\sum_{i'=1}^{B}\exp(f_i^{C^s}, f_{i'}^{V^s}, \tau_t)},\end{aligned} \tag{5}$$

the total loss for ETL is defined as:

$$\mathcal{L}_{ETL} = \frac{1}{B}\sum_{i=1}^{B}(\mathcal{L}_i^{V^s 2C^s} + \mathcal{L}_i^{C^s 2V^s})/2. \tag{6}$$

3.3 Training and Inference

Training. In addition to ECL and ETL, video-level contrastive learning is also employed to align the original video-text pair in training data, which we refer to as Video-Text Alignment (VTA), as shown in Figure 2 (c) and consistent with previous work. The model is optimized through the joint constraints of ECL, ETL, and VTA losses. However, we find that directly mixing original and synthesized video-text pairs within a batch can interfere with model training, as the high distinguishability (due to the different information density) between the two results in minimal contrastive gains between them. To address this, we adopt an Alternating Iteration Training scheme to separately utilize the two sets of data for training, with details in Algorithm 1 in the supplemental material.
Inference. During inference, we only take the video-level representation from the video encoder to respond to text queries for efficient retrieval, just like a typical dual-encoder-based retriever [4].

4 Experiments

4.1 Datasets and Evaluation Metrics

Pre-Training Datasets. We follow recent works and use two datasets for pre-training: Conceptual Captions 3M (CC3M) [47], containing 3 million image-text pairs, and WebVid-2M [4], containing 2.5 million video-text pairs.
Downstream Datasets and Evaluation Metrics. We conduct downstream **Text-to-Video Retrieval** evaluation on four widely used datasets: MSRVTT [57], DiDeMo [1], LSMDC [43] and MSVD [8]. Recall and Median Rank are used as retrieval evaluation metrics. Furthermore, we validate the generalization

ability of our method on two **Video Action Recognition** datasets, UCF101 [49] and HMDB51 [24], as in previous works. Top-1 classification accuracy is used as action recognition evaluation metric. To demonstrate the advantage of our method in understanding video event content, we conducted experiments on the newly proposed **Multi-event Video-Text Retrieval** [64] and **Video Moment Retrieval** tasks. To showcase the superiority of our method in clarifying event temporal logic, we perform experiments on the newly proposed **Test of Time** [2] task. The above three event understanding tasks all involve making certain adjustments to the original ActivityNet [6] annotations to serve as task data. Multi-event Video-Text Retrieval employs Recall@k-Average, Recall@k-One-Hit and Recall@k-All-Hit as evaluation metrics; video moment retrieval uses $R@_n^\theta$ as the evaluation metric; and Test of Time adopts time-order consistency as the evaluation metric. The details of these downstream datasets and evaluation metrics can be found in the supplemental material.

4.2 Implementation Details

EA-VTR adopts the BERTbase [12] model as the text encoder and a TimeSformer-variant model [4] initialized with ViT [13] weights pre-trained on ImageNet-21k [11] as the video encoder. BLIP [29] is utilized as the image captioner with a Top-p of 0.9 for sampling during generation. Pre-training is divided into two stages following previous works. In the first stage, we warm up EA-VTR by pre-training it for 10 epochs on CC3M and WebVid-2M (random sample 1 frame), with a batch size of 2048 and a peak learning rate of 1×10^{-4}. In the second stage, we perform 5 epoch pre-training constrained by ETA, ECL and ETL jointly on WebVid-2M with 4 random sampled frames from evenly divided clips, with a batch size of 1024 and a peak learning rate of 5×10^{-5}. More implementation details are presented in the supplementary material.

4.3 Comparison with the State-of-the-Art

Text-to-Video Retrieval. We separately compare the results of joint-encoder and dual-encoder models as previous works [9, 35]. The zero-shot text-to-video retrieval results on four mainstream downstream datasets are presented in Table 1, we can draw the following conclusions: (1) EA-VTR significantly surpasses previous dual-encoder methods by a large margin. Specifically, EA-VTR improves upon the previous best method TCP by 1.2%, 4.8%, and 4.7% on R@1, R@5, and R@10 of MSRVTT. The advantage is even more pronounced on the temporally-rich DiDeMo, with R@1, R@5, and R@10 being 5.5%, 8.6%, and 5.3% higher than the previous SOTA method Miles. Even when compared to the dual-encoder method CLIP-ViP, which utilizes $\times100$ more training data than ours, our method still maintains competitiveness on MSRVTT and exhibits superior performance on DiDeMo and LSMDC. (2) Compared to the best joint-encoder methods on various datasets, EA-VTR consistently demonstrates comparable or even superior performance. Simultaneously, it has obvious speed advantages, such as being more than $\times30$ faster than Singularity [25]. This is noteworthy,

Table 1: Zero-shot text-to-video retrieval results on MSRVTT, DiDeMo, LSMDC, and MSVD. CLIP-ViP uses ×100 our training data.

Method	MSRVTT				DiDeMo				LSMDC				MSVD			
	R@1↑	R@5 ↑	R@10↑	MedR↓	R@1↑	R@5 ↑	R@10↑	MedR↓	R@1↑	R@5↑	R@10↑	MedR↓	R@1↑	R@5↑	R@10↑	MedR↓
Joint-Encoder																
TACO [61]	9.8	25.0	33.4	29.0	–	–	–	–	–	–	–	–	–	–	–	–
VIOLET [15]	25.9	49.5	59.7	–	23.5	49.8	59.8	–	–	–	–	–	–	–	–	–
ALPRO [27]	24.1	44.7	55.4	8.0	23.8	47.3	57.9	6.0	–	–	–	–	–	–	–	–
Rap [55]	28.9	47.5	56.8	7.0	29.5	55.7	65.6	4.0	12.8	26.6	33.4	37.0	**35.9**	**64.3**	**73.7**	**3.0**
Clover [22]	26.4	49.5	**60.0**	6.0	29.5	55.2	66.3	4.0	**14.7**	29.2	**38.2**	**24.0**	–	–	–	–
TW-BERT [62]	26.4	50.1	59.6	**5.0**	28.4	52.9	64.5	**4.0**	14.2	**30.4**	36.0	28.0	–	–	–	–
Singularity [25]	**28.4**	**50.2**	59.5	–	**36.9**	**52.9**	**64.5**	–	–	–	–	–	–	–	–	–
Dual-Encoder																
Frozen [4]	18.7	39.5	51.6	10.0	21.1	46.0	56.2	7.0	9.3	22.0	30.1	51.0	38.7	70.1	80.1	2.0
LaT [3]	23.4	44.1	53.3	8.0	22.6	45.9	58.9	7.0	-	-	-	-	36.9	68.6	81.0	2.0
RegionL. [60]	22.2	43.3	52.9	8.0	–	–	–	–	–	–	–	–	–	–	–	–
OA-Trans [51]	23.4	47.5	55.6	8.0	23.5	50.4	59.8	6.0	-	-	-	-	-	-	-	-
MCQ [17]	26.0	46.4	56.4	7.0	25.6	50.6	61.1	5.0	12.2	25.9	32.2	42.0	43.6	74.9	84.9	2.0
Miles [18]	26.1	47.2	56.9	7.0	27.2	50.3	63.6	5.0	11.1	24.7	30.6	50.7	44.4	76.2	**87.0**	2.0
TCP [65]	26.8	48.3	57.6	7.0	–	–	–	–	–	–	–	–	–	–	–	–
CLIP-ViP [59]	31.7	51.2	63.2	4.0	24.6	50.7	59.7	5.0	12.5	26.1	33.3	39.0	–	–	–	–
EA-VTR	**28.0**	**53.1**	**62.3**	**5.0**	**32.7**	**58.9**	**68.9**	**3.0**	**15.7**	**29.6**	**36.0**	**30.0**	**46.6**	**78.9**	86.5	**2.0**

as joint-encoder methods are typically considered to have superior performance but at the cost of lower efficiency. The text-to-video retrieval results after fine-tuning are shown in Table 2. It can be observed that the conclusions drawn from zero-shot retrieval still hold true after fine-tuning.

Table 2: Fine-tuned text-to-video retrieval results on MSRVTT, DiDeMo, LSMDC, and MSVD.

Method	MSRVTT				DiDeMo				LSMDC				MSVD			
	R@1↑	R@5 ↑	R@10↑	MedR↓	R@1↑	R@5 ↑	R@10↑	MedR↓	R@1↑	R@5↑	R@10↑	MedR↓	R@1↑	R@5↑	R@10↑	MedR↓
Frozen [4]	31.0	59.5	70.5	3.0	31.0	59.8	72.4	3.0	15.0	30.8	39.8	20.0	45.6	79.8	88.2	2.0
LaT [3]	35.3	61.3	72.9	3.0	32.6	61.3	71.6	3.0	-	-	-	-	40.0	74.6	84.2	2.0
RegionL [60]	36.3	63.9	72.5	3.0	32.5	60.8	72.3	3.0	17.1	32.5	41.5	18.0	44.0	74.9	84.3	2.0
OA-Trans [51]	35.8	63.4	76.5	3.0	34.8	64.4	75.1	3.0	18.2	34.3	43.7	18.5	39.1	68.4	80.3	2.0
MCQ [17]	37.6	64.8	75.1	3.0	37.0	62.2	73.9	3.0	17.9	35.4	44.5	15.0	52.0	82.8	90.0	1.0
Miles [18]	37.7	63.6	73.8	3.0	36.6	63.9	74.0	3.0	17.8	35.6	44.1	15.5	**53.9**	83.5	90.2	1.0
TCP [65]	38.0	65.5	76.4	7.0	–	–	–	–	–	–	–	–	–	–	–	–
EA-VTR	**39.5**	**67.2**	**77.0**	**2.0**	**43.7**	**73.3**	**81.7**	**2.0**	**22.0**	**40.8**	**51.0**	**10.0**	52.7	**83.5**	**90.7**	**1.0**

Video Action Recognition. To verify the generalization ability of our method, we conduct zero-shot evaluation on video action recognition datasets UCF101 and HMDB51, by converting the video categories to captions with text prompts and uniformly sampling 16 frames for videos as input, as in previous works. The results are presented in Table 3, it can be seen that EA-VTR improves the average top-1 classification accuracy (Column "Mean" in the table) on UCF101 and HMDB51 by 8.3% and 0.5% compared to Miles [18]. This can be attributed to the fact that actions are inherently present in video events, and thus the event augmentation and learning method proposed in our work naturally enhances the understanding of actions.

Table 3: Zero-shot video action recognition results on UCF101 and HMDB51.

Method	UCF101				HMDB51			
	S1	S2	S3	Mean	S1	S2	S3	Mean
ClipBert [26]	27.5	27.0	28.8	27.8	20.0	22.0	22.3	21.4
Frozen [4]	45.4	44.7	47.7	45.9	27.5	28.3	27.7	27.8
MCQ [17]	51.1	54.3	53.8	53.1	38.0	36.1	**39.1**	37.7
Miles [18]	51.8	53.4	52.8	52.7	38.4	38.6	37.8	38.3
EA-VTR	**60.0**	**62.4**	**60.5**	**61.0**	**38.6**	**39.8**	37.9	**38.8**

Table 4: Zero-shot Multi-event Video-Text Retrieval results on ActivityNet, with Recall@k-Average/One-Hit/All-Hit as metrics for each k = 1, 5, 10, 50 in Video-to-Text Retrieval and commonly used Recall@k as metric in Text-to-Video Retrieval.

Method	Video-To-Text			
	k=1	k=5	k=10	k=50
Frozen [4]	2.1/6.8/–	7.2/19.5/0.0	11.9/29.8/1.5	30.3/60.7/7.9
MCQ [17]	2.8/8.8/–	9.7/25.2/1.0	15.0/36.8/2.2	35.9/68.9/10.4
Miles [18]	2.9/9.3/–	10.1/26.5/1.0	15.4/37.2/2.2	36.3/69.7/10.7
EA-VTR	**4.4**/**13.9**/–	**13.2**/**33.6**/**1.4**	**19.8**/**45.8**/**3.5**	**42.9**/**77.0**/**14.9**
	Text-To-Video			
	k=1	k=5	k=10	k=50
Frozen [4]	6.9	19.4	28.4	53.7
MCQ [17]	8.5	22.8	32.5	58.6
Miles [18]	8.6	22.9	32.4	58.1
EA-VTR	**10.7**	**27.0**	**37.7**	**63.1**

Table 5: Fine-tuned Video Moment Retrieval results on ActivityNet.

Method	$R@1^{0.5}$	$R@1^{0.7}$	$R@5^{0.5}$	$R@5^{0.7}$
Frozen [4]	43.3	25.8	75.8	59.3
LocVTP [7]	46.1	27.6	78.9	63.7
TCP [65]	46.5	28.4	78.2	64.0
EA-VTR	**48.0**	**29.6**	**80.1**	**65.0**

Table 6: Zero-shot Test of Time results on ActivityNet.

Method	ActivityNet		
	Before	After	All
Frozen [4]	48.9	49.6	49.3
MCQ [17]	50.5	50.1	50.3
Miles [18]	50.2	49.8	50.0
EA-VTR	**62.6**	**60.5**	**61.6**

Multi-event Video-Text Retrieval. The goal of Multi-event Video-Text Retrieval task is to recall all events related to a given video, which effectively reflects the comprehensive perception ability of the model for all video events. We perform zero-shot evaluation without any post-training on the ActivityNet dataset of the task by uniformly sampling 16 video frames as input, and the results are presented in Table 4. On the video-to-text retrieval setting, it is evident that our method significantly outperforms others in all three metrics: Recall@k-Average/One-Hit/All-Hit, and the results for the text-to-video retrieval setting exhibit a similar trend. This validates that the ECL proposed in the work indeed benefits the ability of the model to capture all events within the video.

Video Moment Retrieval. The purpose of video moment retrieval is to find the corresponding clip in the video given a text query, which can reflect the ability of the model to perceive a specific event. We adapt our method for the video moment retrieval task on the ActivityNet dataset following the strategy in LocVTP. The results in Table 5 show that although our method is not specifically designed for this task, it still outperforms the customized SOTA methods LocVTP and TCP, achieving improvements of 1.9% and 1.5% compared to the two methods on $R@1^{0.5}$ respectively.

Test of Time. Each sample in the Test of Time task contains a video and two video captions with reversed event orders, requiring the model to select the correct caption for the video. This task aims to test the temporal understanding

ability of the model for video events, and we performed zero-shot evaluation on the ActivityNet dataset provided by the authors. Results in Table 6 indicate that it is a rather challenging task, as the performance of previous dual-encoder models is close to random guessing (50.0%). It can be seen that our method has a significant advantage in determining the temporal relationships of events, exhibiting a 61.6% time-order consistency across all samples, leading the second-place method by 11.3%. This demonstrates that ETL can effectively enhance the ability of the model to understand event temporal logic.

4.4 Ablation Studies

In this section, we conduct ablations to verify the effectiveness of our design choices through evaluating models for zero-shot text-to-video retrieval on MSRVTT and DiDeMo. Due to the limited resources, we only use 1 million video-text pairs randomly selected from WebVid-2M for pre-training with a batch size of 512.

Table 7: Ablation studies on the effectiveness and compatibility of the components in our method, including VTA, ECL and ETL.

	VTA	ETL	ECL	MSRVTT			DiDeMo		
				R@1↑	R@5↑	R@10↑	R@1↑	R@5↑	R@10↑
A	✓			19.2	39.2	49.7	20.3	45.2	55.2
B	✓	✓		22.0	42.9	53.3	25.3	50.1	59.9
C	✓		✓	21.5	42.1	52.7	25.3	51.2	61.3
D	✓	✓	✓	**23.6**	**44.4**	**54.6**	**27.3**	**54.2**	**63.0**

Are ECL and ETL effective and compatible? Yes. From the results in Table 7, we can make the following observations: (1) Models B and C both outperform baseline model A trained only with VTA. Specifically, model B, incorporating ECL, demonstrates a greater improvement on MSRVTT, while model C, integrating ETL, exhibits a more significant gain on DiDeMo. This suggests that the benefits of ECL and ETL are more prominent in event content and temporal learning. (2) Furthermore, model D, incorporates ECL and ETL, achieves further improvements over models B and C on both datasets, suggesting that ECL and ECL are mutually compatible and beneficial.

Table 8: Ablation studies on the effectiveness of Alternating Iteration Training.

Method	MSRVTT			DiDeMo		
	R@1↑	R@5↑	R@10↑	R@1↑	R@5↑	R@10↑
Baseline	19.2	39.2	49.7	20.3	45.2	55.2
w/o AIT.	20.9	41.6	51.5	23.3	48.6	58.4
with AIT.	**23.6**	**44.4**	**54.6**	**27.3**	**54.2**	**63.0**

Is Alternating Iteration Training effective? Yes. Removing the Alternating Iterative Training (AIT) involves mixing the web-crawled and ETA-synthesized

video-text pairs in the same batch for training. As can be seen from the comparisons in Table 8, the improvement obtained by removing AIT is significantly lower than that of our model, compared to the baseline model. As previously mentioned in section 3.3, web-crawled and synthesized video-text pairs have different information densities, making them easily distinguishable. The contrast between them may not yield much gains.

Table 9: Event learning *vs*. textual augmentation. "VC" and "FC" are abbreviations for "Video Caption" and "Frame Caption".

	Method	MSRVTT			DiDeMo		
		R@1↑	R@5↑	R@10↑	R@1↑	R@5↑	R@10↑
A	VTA + VC	19.2	39.2	49.7	20.3	45.2	55.2
B	VTA + Sample(VC, Concat(all FC))	21.1	41.8	51.9	23.8	48.8	58.9
C	VTA + Concat(VC, FC)	20.9	41.5	51.6	23.1	47.6	57.6
D	VTA + Sample(VC, FC)	21.4	42.0	52.2	24.1	48.6	58.6
E	**EA-VTR**	**23.6**	**44.4**	**54.6**	**27.3**	**54.2**	**63.0**

Does the benefit come from the proposed event learning rather than merely textual augmentation? Yes. We reorganize the original web-crawled video captions and generated frame captions to train three text-augmented baselines with only video-level contrastive learning, *i.e*., VTA, including models B (sampling from video captions and concatenated frame captions), C (concatenating video captions with frame captions) and model D (sampling from video captions and frame captions). As shown in Table 9, text-augmented baselines exhibit better performance than the simple baseline A trained only with web-crawled video captions, but are still notably inferior to our EA-VTR with event learning, especially on DiDeMo with complex events.

Table 10: Ablation studies on the event augmentation strategy.

Captioner	Decoding	MSRVTT			DiDeMo		
		R@1↑	R@5↑	R@10↑	R@1↑	R@5↑	R@10↑
BLIP	BeamSearch	21.7	42.1	52.3	23.9	51.7	60.2
BLIP	Top-p	23.6	**44.4**	54.6	27.3	**54.2**	63.0
BLIP2	Top-p	**23.9**	44.0	**54.7**	**27.8**	53.2	**63.3**

Can more diverse text improve downstream task generalization? Yes. The diversity of frame captions can be manipulated by altering the decoding strategy of the captioner. Specifically, Table 10 shows that the model trained with less diverse frame captions obtained using deterministic BeamSearch [14] decoding underperforms the model trained with non-deterministic Top-p decoded captions. Non-deterministic decoding strategy provides different vocabulary or phrases to describe similar events, which aids in model generalization.
Is a stronger image captioner like MLLM needed for event augmentation? The recent exciting emergence of MLLM (Multi-Modal Large Language

Model) [28,67] raises this question, we thus answer it by replacing the captioner with the most commonly used MLLM, BLIP2. The results in Table 10 show that there is no significant improvement after the replacement, indicating that MLLM does not have a clear advantage in our method but requires longer computation time and memory. It also demonstrates that although our method relies on a captioner, it possesses good robustness to the ability of the captioner.

4.5 Qualitative Analysis

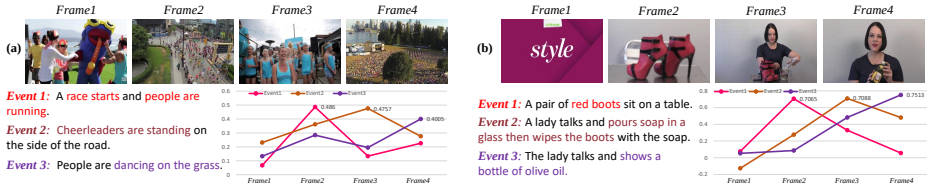

Fig. 3: Examples of frame-event alignment: Above are the extracted video frames, while the bottom left shows multiple events occurring at different times in the video, distinguished by different colors (with key event information in the text also colored). The bottom right displays the similarity score curves between the text features of these events and the visual features of video frames.

Frame-Event Alignment. To demonstrate that the frame-level visual representations optimized by ECL are meaningful, we visualize the frame-event alignment relationships in Figure 3, with samples sourced from ActivityNet. Observing Figure 3 (a), we can find that the events and video frames exhibit the correct corresponding relationships, such as event 1 "A race starts and people are running" with frame 2. Additionally, some challenging frame-event alignment relationships can also be accurately captured by the model. As shown in Figure 3 (b), where the event subjects of events 2 and 3 are the same woman, with the difference being the items she is presenting. Nonetheless, the model correctly associates event 2 with frame 3 and event 3 with frame 4.

5 Conclusion

In this work, we are dedicated to enhancing the ability of video-text retrieval model to comprehensively understand the video events, thereby enabling a event-aware retriever. To achieve this goal, we have made efforts to supplement missing event information in the pre-training data and improve the model's ability to capture detailed and complex video-text event alignment. On this basis, we validate the proposed method by thorough experimentation from multiple perspectives. In the future, we will continue to explore the help of event learning for other video-text tasks such as video question answering and video captioning.

Acknowledgments This work is supported by the Key Research and Development Program of Xinjiang Urumqi Autonomous Region under Grant No. 2023B01005, the Natural Science Foundation of China (Grants 62302501, 62036011, 62122086, 62192782, 61721004, U2033210 and 62372451), Beijing Natural Science Foundation (JQ21017, JQ24022, L243015), CCF-Tencent Rhino-Bird Open Research Fund.

References

1. Anne Hendricks, L., Wang, O., Shechtman, E., Sivic, J., Darrell, T., Russell, B.: Localizing moments in video with natural language. In: Proceedings of the IEEE international conference on computer vision. pp. 5803–5812 (2017)
2. Bagad, P., Tapaswi, M., Snoek, C.G.: Test of time: Instilling video-language models with a sense of time. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2503–2516 (2023)
3. Bai, J., Liu, C., Ni, F., Wang, H., Hu, M., Guo, X., Cheng, L.: Lat: latent translation with cycle-consistency for video-text retrieval. arXiv preprint arXiv:2207.04858 (2022)
4. Bain, M., Nagrani, A., Varol, G., Zisserman, A.: Frozen in time: A joint video and image encoder for end-to-end retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1728–1738 (2021)
5. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? In: ICML. vol. 2, p. 4 (2021)
6. Caba Heilbron, F., Escorcia, V., Ghanem, B., Carlos Niebles, J.: Activitynet: A large-scale video benchmark for human activity understanding. In: Proceedings of the ieee conference on computer vision and pattern recognition. pp. 961–970 (2015)
7. Cao, M., Yang, T., Weng, J., Zhang, C., Wang, J., Zou, Y.: Locvtp: Video-text pre-training for temporal localization. In: European Conference on Computer Vision. pp. 38–56. Springer (2022)
8. Chen, D., Dolan, W.B.: Collecting highly parallel data for paraphrase evaluation. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies. pp. 190–200 (2011)
9. Chen, Y., Ma, Z., Zhang, Z., Qi, Z., Yuan, C., Shan, Y., Li, B., Hu, W., Qie, X., Wu, J.: Vilem: Visual-language error modeling for image-text retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11018–11027 (June 2023)
10. Consortium, E., et al.: Egocentric live 4d perception (ego4d) database: A large-scale first-person video database, supporting research in multi-modal machine perception for daily life activity
11. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
12. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
13. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

14. Freitag, M., Al-Onaizan, Y.: Beam search strategies for neural machine translation. arXiv preprint arXiv:1702.01806 (2017)
15. Fu, T.J., Li, L., Gan, Z., Lin, K., Wang, W.Y., Wang, L., Liu, Z.: Violet: End-to-end video-language transformers with masked visual-token modeling. arXiv preprint arXiv:2111.12681 (2021)
16. Gabeur, V., Sun, C., Alahari, K., Schmid, C.: Multi-modal transformer for video retrieval. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16. pp. 214–229. Springer (2020)
17. Ge, Y., Ge, Y., Liu, X., Li, D., Shan, Y., Qie, X., Luo, P.: Bridging video-text retrieval with multiple choice questions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16167–16176 (2022)
18. Ge, Y., Ge, Y., Liu, X., Wang, J., Wu, J., Shan, Y., Qie, X., Luo, P.: Miles: Visual bert pre-training with injected language semantics for video-text retrieval. In: European Conference on Computer Vision. pp. 691–708. Springer (2022)
19. Ging, S., Zolfaghari, M., Pirsiavash, H., Brox, T.: Coot: Cooperative hierarchical transformer for video-text representation learning. Advances in neural information processing systems **33**, 22605–22618 (2020)
20. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al.: Ego4d: Around the world in 3,000 hours of egocentric video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18995–19012 (2022)
21. Holtzman, A., Buys, J., Du, L., Forbes, M., Choi, Y.: The curious case of neural text degeneration. arXiv preprint arXiv:1904.09751 (2019)
22. Huang, J., Li, Y., Feng, J., Wu, X., Sun, X., Ji, R.: Clover: Towards a unified video-language alignment and fusion model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14856–14866 (2023)
23. Jozefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., Wu, Y.: Exploring the limits of language modeling. arXiv preprint arXiv:1602.02410 (2016)
24. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: Hmdb: a large video database for human motion recognition. In: 2011 International conference on computer vision. pp. 2556–2563. IEEE (2011)
25. Lei, J., Berg, T.L., Bansal, M.: Revealing single frame bias for video-and-language learning. arXiv preprint arXiv:2206.03428 (2022)
26. Lei, J., Li, L., Zhou, L., Gan, Z., Berg, T.L., Bansal, M., Liu, J.: Less is more: Clipbert for video-and-language learning via sparse sampling. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7331–7341 (2021)
27. Li, D., Li, J., Li, H., Niebles, J.C., Hoi, S.C.: Align and prompt: Video-and-language pre-training with entity prompts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4953–4963 (2022)
28. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023)
29. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning. pp. 12888–12900. PMLR (2022)
30. Li, L., Chen, Y.C., Cheng, Y., Gan, Z., Yu, L., Liu, J.: Hero: Hierarchical encoder for video+ language omni-representation pre-training. arXiv preprint arXiv:2005.00200 (2020)

31. Li, L., Gan, Z., Lin, K., Lin, C.C., Liu, Z., Liu, C., Wang, L.: Lavender: Unifying video-language understanding as masked language modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23119–23129 (2023)
32. Li, Y., Min, K., Tripathi, S., Vasconcelos, N.: Svitt: Temporal learning of sparse video-text transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18919–18929 (2023)
33. Liu, Y., Albanie, S., Nagrani, A., Zisserman, A.: Use what you have: Video retrieval using representations from collaborative experts. arXiv preprint arXiv:1907.13487 (2019)
34. Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H.: Video swin transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3202–3211 (2022)
35. Lu, H., Fei, N., Huo, Y., Gao, Y., Lu, Z., Wen, J.R.: Cots: Collaborative two-stream vision-language pre-training model for cross-modal retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15692–15701 (June 2022)
36. Miech, A., Alayrac, J.B., Smaira, L., Laptev, I., Sivic, J., Zisserman, A.: End-to-end learning of visual representations from uncurated instructional videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9879–9889 (2020)
37. Miech, A., Zhukov, D., Alayrac, J.B., Tapaswi, M., Laptev, I., Sivic, J.: Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2630–2640 (2019)
38. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
39. Patrick, M., Huang, P.Y., Asano, Y., Metze, F., Hauptmann, A., Henriques, J., Vedaldi, A.: Support-set bottlenecks for video-text representation learning. arXiv preprint arXiv:2010.02824 (2020)
40. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
41. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)
42. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015)
43. Rohrbach, A., Rohrbach, M., Tandon, N., Schiele, B.: A dataset for movie description. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3202–3212 (2015)
44. Rouditchenko, A., Boggust, A., Harwath, D., Chen, B., Joshi, D., Thomas, S., Audhkhasi, K., Kuehne, H., Panda, R., Feris, R., et al.: Avlnet: Learning audio-visual language representations from instructional videos. arXiv preprint arXiv:2006.09199 (2020)
45. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 (2019)
46. Shao, D., Xiong, Y., Zhao, Y., Huang, Q., Qiao, Y., Lin, D.: Find and focus: Retrieve and localize video events with natural language queries. In: Proceedings of the European Conference on Computer Vision (ECCV) (September 2018)

47. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2556–2565 (2018)
48. Shi, Y., Liu, H., Xu, H., Ma, Z., Ye, Q., Hu, A., Yan, M., Zhang, J., Huang, F., Yuan, C., et al.: Learning semantics-grounded vocabulary representation for video-text retrieval. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 4460–4470 (2023)
49. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
50. Tewel, Y., Shalev, Y., Schwartz, I., Wolf, L.: Zerocap: Zero-shot image-to-text generation for visual-semantic arithmetic. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17918–17928 (2022)
51. Wang, J., Ge, Y., Cai, G., Yan, R., Lin, X., Shan, Y., Qie, X., Shou, M.Z.: Object-aware video-language pre-training for retrieval. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3313–3322 (2022)
52. Wang, J., Ge, Y., Yan, R., Ge, Y., Lin, K.Q., Tsutsui, S., Lin, X., Cai, G., Wu, J., Shan, Y., et al.: All in one: Exploring unified video-language pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6598–6608 (2023)
53. Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., Ma, J., Zhou, C., Zhou, J., Yang, H.: Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In: International Conference on Machine Learning. pp. 23318–23340. PMLR (2022)
54. Wu, W., Luo, H., Fang, B., Wang, J., Ouyang, W.: Cap4video: What can auxiliary captions do for text-video retrieval? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10704–10713 (2023)
55. Wu, X., Gao, C., Lin, Z., Wang, Z., Han, J., Hu, S.: Rap: Redundancy-aware video-language pre-training for text-video retrieval. arXiv preprint arXiv:2210.06881 (2022)
56. Xu, H., Ghosh, G., Huang, P.Y., Okhonko, D., Aghajanyan, A., Metze, F., Zettlemoyer, L., Feichtenhofer, C.: Videoclip: Contrastive pre-training for zero-shot video-text understanding. arXiv preprint arXiv:2109.14084 (2021)
57. Xu, J., Mei, T., Yao, T., Rui, Y.: Msr-vtt: A large video description dataset for bridging video and language. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5288–5296 (2016)
58. Xu, M., Pérez-Rúa, J.M., Escorcia, V., Martinez, B., Zhu, X., Zhang, L., Ghanem, B., Xiang, T.: Boundary-sensitive pre-training for temporal localization in videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7220–7230 (2021)
59. Xue, H., Sun, Y., Liu, B., Fu, J., Song, R., Li, H., Luo, J.: Clip-vip: Adapting pre-trained image-text model to video-language representation alignment. arXiv preprint arXiv:2209.06430 (2022)
60. Yan, R., Shou, M.Z., Ge, Y., Wang, J., Lin, X., Cai, G., Tang, J.: Video-text pre-training with learned regions for retrieval. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 3100–3108 (2023)
61. Yang, J., Bisk, Y., Gao, J.: Taco: Token-aware cascade contrastive learning for video-text alignment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11562–11572 (2021)

62. Yang, X., Li, Z., Xu, H., Zhang, H., Ye, Q., Li, C., Yan, M., Zhang, Y., Huang, F., Huang, S.: Learning trajectory-word alignments for video-language tasks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2504–2514 (October 2023)
63. Ye, Q., Xu, G., Yan, M., Xu, H., Qian, Q., Zhang, J., Huang, F.: Hitea: Hierarchical temporal-aware video-language pre-training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15405–15416 (2023)
64. Zhang, G., Ren, J., Gu, J., Tresp, V.: Multi-event video-text retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22113–22123 (2023)
65. Zhang, H., Liu, D., Lv, Z., Su, B., Tao, D.: Exploring temporal concurrency for video-language representation learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 15568–15578 (October 2023)
66. Zhao, Y., Misra, I., Krähenbühl, P., Girdhar, R.: Learning video representations from large language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6586–6597 (2023)
67. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)
68. Zhu, L., Yang, Y.: Actbert: Learning global-local video-text representations. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8746–8755 (2020)