Exploring Temporal Concurrency for Video-Language Representation Learning

Heng Zhang[1,2†] Daqing Liu[3] Zezhong Lv[1,2] Bing Su[1,2‡] Dacheng Tao[4]
[1] Gaoling School of Artificial Intelligence, Renmin University of China
[2]Beijing Key Laboratory of Big Data Management and Analysis Methods
[3] JD Explore Academy, JD.com [4] The University of Sydney
zhangheng@ruc.edu.cn, {liudq.ustc,zezhonglv0306,subingats,dacheng.tao}@gmail.com

Abstract

Paired video and language data is naturally temporal concurrency, which requires the modeling of the temporal dynamics within each modality and the temporal alignment across modalities simultaneously. However, most existing video-language representation learning methods only focus on discrete semantic alignment that encourages aligned semantics to be close in the latent space, or temporal context dependency that captures short-range coherence, failing in building the temporal concurrency. In this paper, we propose to learn video-language representations by modeling video-language pairs as Temporal Concurrent Processes (TCP) via a process-wised distance metric learning framework. Specifically, we employ the soft Dynamic Time Warping (DTW) to measure the distance between two processes across modalities and then optimize the DTW costs. Meanwhile, we further introduce a regularization term that enforces the embeddings of each modality approximating a stochastic process to guarantee the inherent dynamics. Experimental results on three benchmarks demonstrate that TCP stands as a state-of-the-art method for various video-language understanding tasks, including paragraph-to-video retrieval, video moment retrieval, and video question-answering. Code is available at https://github.com/hengRUC/TCP.

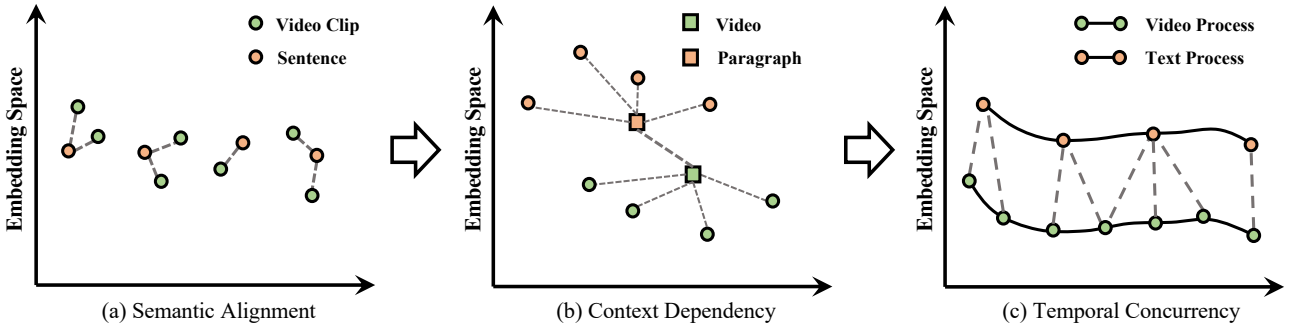
1. Introduction

Video-Language Representation Learning [33, 42, 26, 48, 46] is a fundamental problem of multimodal intelligence, which has demonstrated great practical value in various real-world applications such as video captioning [29, 40], video question answering [46, 26, 57], and video retrieval [7, 17, 16]. Essentially, a video-language pair can be seen as two temporal sequences where each sequence is coherent and change smoothly, and the two sequences are concurrently aligned with each other over time. Therefore, different from single-modality representation learning, *e.g.*, video or text, that requires capturing the temporal dynamics along time [42, 26, 48, 13], multi-modality learning further appeals to the temporal alignment across two concurrent modalities. We refer to the property of requiring modeling of both temporal dynamics and temporal alignment in video-language learning as *temporal concurrency*.

Pioneer works [32, 1, 28, 54, 50] for video-language representation learning typically concentrate on the semantic alignment by discrete cross-modal matching (Figure 1(a)) that encourages paired video clips and sentence [32] to be close in the common latent space [1, 28, 54, 50]. However, the imposition of pulling discrete semantics together without accounting for the temporal dynamics of each individual modality disrupts the inherent temporal coherence of unimodal representations and temporal concurrency between two modalities. Thereby leads to a sub-optimal representation with limited generalization [18]. Recently, an emerging line of endowing temporal dynamics is to capture the temporal context dependency (Figure 1(b)), by global representation alignment across modalities [48], long-form video encoding with video Transformers [45], or temporal order modeling by shuffle discriminations [25, 10, 21]. Despite incorporating temporal constraints to capture temporal context, these methods have limited efficacy in capturing subtle and long-range dependencies due to the neglect of temporal coherence over the entire sequence of data. The incomplete modeling of temporal dynamics undermines temporal concurrency and leads to unsatisfactory performance.

In a nutshell, video-language representation learning entails the *temporal concurrency* with two intrinsic patterns: 1) intra-modal temporal dynamics that indicate the sequential representations should adhere to coherent constraints, and 2) cross-modal temporal alignment that requires both global (video-paragraph) and local (*i.e.*, video clip-sentence) semantic alignments over time. Based on the above insights, we propose to model the video-language

[†]Research Intern at JD Explore Academy.
[‡]Corresponding author.



(a) Semantic Alignment (b) Context Dependency (c) Temporal Concurrency

Figure 1. Compare to mainstream video-language representation learning methods. **(a)** *Semantic Alignment* (*e.g.*, HERO [28], Frozen [1]) enforces video-clip sentence pairs to be close in the embedding space, disrupting the inherent temporal dynamics of each modality. **(b)** *Context Dependency* (*e.g.* HD-VILA [50], MERLOT [54]) endows short-range temporal context dependency within each modality, limited on capturing long-range dependencies. **(c)** The proposed *Temporal Concurrency* models video-language pairs as temporal concurrency processes, therefore capturing temporal alignments while maintaining the coherence of each modality.

pairs as Temporal Concurrent Processes (TCP, Figure 1(c)) where each modality is represented as a goal-oriented stochastic process [35, 4] and the two modalities are further aligned as concurrent processes.

Here the basic assumptions are that 1) a temporal sequence is coherent and smoothly changes from the start to the end, which is essentially a stochastic process, and 2) two concurrent processes can be temporally aligned by minimizing the distance between them.

The implementation of TCP follows a process-wised distance metric learning framework, where the distance is measured by an optimal match between two processes via dynamic programming. Specifically, we employ the differentiable soft Dynamic Time Warping [12] (soft-DTW) to calculate the overall cost. However, directly optimizing networks by soft-DTW costs usually encounter trivial solutions owing to the lack of constraints for each sequence, we further introduce a regularization term that models each temporal sequence as a goal-oriented stochastic process to agree with a time-variant Gaussian distribution, *i.e.*, the Brownian bridge [37, 3], wherein the representations in the process are expected to be the line combination of the bridge head and tail. The overall training objective is the combination of the soft-DTW cost with the regularization term.

Thanks to the global optimization between two self-consistent temporal processes by modeling video-language pairs as Temporal Concurrency Processes, the proposed TCP successfully captures the implicit correlation between the fine-grained semantics of each modality and models the temporal concurrency across modalities. Compared with existing methods that rely on discrete cross-modal alignment or temporal context dependency, TCP empirically produces richer representations of visual contents and language for various downstream video-language tasks, including paragraph-to-video retrieval [53, 1, 17, 16], video moment retrieval [34, 15], and video question-answering [27].

Our contributions are summarized as follows:

- To the best of our knowledge, we are the first work that models video-language pairs as Temporal Concurrent Processes (TCP) to capture the inherent temporal concurrency of video-language representation learning.
- We implement TCP with a novel process-wised distance metric learning framework and optimize it by the soft-DTW cost of two processes to construct temporal alignment, accompanying a regularization term for each process to preserve temporal dynamics.
- The proposed TCP achieves state-of-the-art performance on various video-language understanding tasks across three widely-used datasets.

2. Related Work

Video-Language Pre-training. Early video-language pre-training works are mostly designed for short-form video tasks [1, 50]. VideoBERT [42] trains the encoder by predicting the masked token of video and text representation in a single-stream manner. Two-stream methods [31, 44] are proposed to alleviate the interference when encoding each of the two modalities and interact representations in different latent spaces between video and language. Due to its fundamental role in real applications, video-language pre-training for long-form videos is getting more attention. By applying cross-modality interaction on different scales of video, [28] aggregates the contextualized segment-level representation as the input for further task-specified modules. Moreover, aligning the video segment and its text description is generally introduced as a typical proxy task during video-language pre-training [50, 54]. [50] propose a two-branch scheme to process visual content in different resolutions and perform a divided spatial-temporal operation to reduce the computational cost. [54] takes one step further in modeling fine-grained temporal information by

forcing the model to predict the correct order of frames during a video segment. However, the alignment conducted at the segment level is not sufficient because the natural temporal consistency implied between different segments of long-form video is ignored.

Most recently, sequence alignment for video-language representation learning is explored in TempCLR [52], which discovers the temporal dynamics between different video segments by a contrastive learning framework based on the shuffled segment sequence. Different from TempCLR which uses the unstable DTW for global sequence alignment and learn temporal dynamics by distinguishing normal video sequence from shuffled ones on a single video modality, we leverage a differentiable version of DTW and learn temporal dynamics by modeling sequences as Brownian bridge processes in both video and language modalities.

Long-form Video Temporal Modeling. Recently, a series of more challenging tasks like video moment retrieval [21], video visual grounding [41], and action localization [9], which conducts a contextual alignment between language and visual contents in long-form videos, is becoming a hot research topic because it is more in line with the requirement of the real-world application scenarios. Different from the tasks that rely on only the local dependencies in videos, to semantically match entities and segments in a long-form video with their corresponding description, we need to model the temporal dependencies from aspects at both the segment level and global level. Existing methods mainly focus on exploring the temporal context based on designing multi-scale modeling architecture, while ignoring the correlation between different video segments.

Video-to-Transcript. Prior transcription-based methods can be roughly divided into two categories: 1) Single-modal works leverage transcript (action order) [8] or cluster algorithm [24] for video Temporal Order, ignoring Cross-modal Semantic Alignment. 2) In multimodal works, the frame-sentence similarity matrix is incorporated to determine frame-sentence pseudo labels for denoising [20] or Cross-modal Semantic Alignment [52, 20], while still failing to construct the Inherent Dynamics of sequential input. Although [43] proposes a multimodal temporal contrastive loss for temporal inherent relation, it is process-unperceived and neglects Temporal Order. Comparing to video-to-transcript methods, our TCP achieves 1) Temporal Order and Cross-modal Semantic Alignment via soft-DTW, 2) Inherent Dynamics of each modality with a Brownian Bridge constraint.

3. Method

3.1. Model Architecture

Video Encoder. First of all, given a video containing N clips, we randomly select K frames from each clip to represent them, with each frame divided into $H \times W$ patches. Then patchs are mapped to video tokens by a learnable MLP projection head and taken as the input of the video encoder by adding the position embedding indicating the timestamp. We take Swin-Transformer [30] as the video encoder. Following the last layer, an average pooling over each frames that belong to the same video clip is applied to represent the clip. We initialize the video encoder with the parameters of Swin-Transformer pre-trained on ImageNet-21K.

Text Encoder. Similarly, we first obtain text tokens embedding via lookup and add position and segment embeddings as in BERT [13]. The text encoder is a multi-layer bidirectional Transformer with 12 layers, 768 hidden size, and 12 self-attention heads. Following the last layer, an average pooling over each sentence is applied to obtain the hidden states of each sentence. We initialize our text encoder weights with the $\text{BERT}_{\text{BASE}}$ [13] model.

Notations. We denote the video encoder as f_{θ_v}, the text encoder as f_{θ_p}. Given a long video-paragraph pair $x_c = \left\{c_1^1, ..., c_i^j, ..., c_N^M\right\}$ where c_i^j indicates the i-th video clip belong to j-th segment, $x_s = \{s_1, ..., s_j, ..., s_M\}$ where s_j indicates the j-th sentence in the paragraph. We obtain the final representations with,

$$z_v = f_{\theta_v}(x_c), z_p = f_{\theta_p}(x_s) \tag{1}$$

3.2. Cross-modal Sequence Alignment

Preliminary of DTW. The dense vectors output from the dual-encoder can be taken as temporal sequences of different modalities. We employ Dynamic Time Wrapping (DTW) discrepancy [2] for cross-modal sequence alignment. It is a classical algorithm for finding the minimum cost path in the distance matrix of two sequences wherein the minimum cost can be taken as the distance of two sequences. Different from previous work that uses DTW for video-to-transcript [8] or video-to-video alignment [6, 22], we explore the application of DTW on video-to-language alignment by minimizing the sequence distance.

Given the representations output from dual-encoder, $\boldsymbol{z}_v = \left\{\boldsymbol{v}_1^1, ..., \boldsymbol{v}_i^j, ..., \boldsymbol{v}_N^M\right\}$ where $\boldsymbol{v}_i^j$ indicates the representation of i-th video clip belong to j-th segment, $\boldsymbol{z}_p = \{\boldsymbol{p}_1, ..., \boldsymbol{p}_j, ..., \boldsymbol{p}_M\}$ where $\boldsymbol{p}_j$ indicates the representation of j-th sentence in the paragraph. The distance between i-th video clip belong to k-th ($k \in [1, M]$) segment and j-th sentence is defined as:

$$D_{i,j} = \left\|\boldsymbol{v}_i^k - \boldsymbol{p}_j\right\|_2^2, D \in \mathbb{R}^{N \times M} \tag{2}$$

We first calculate the initial states (*i.e.*, the first column and row) for the distance matrix D:

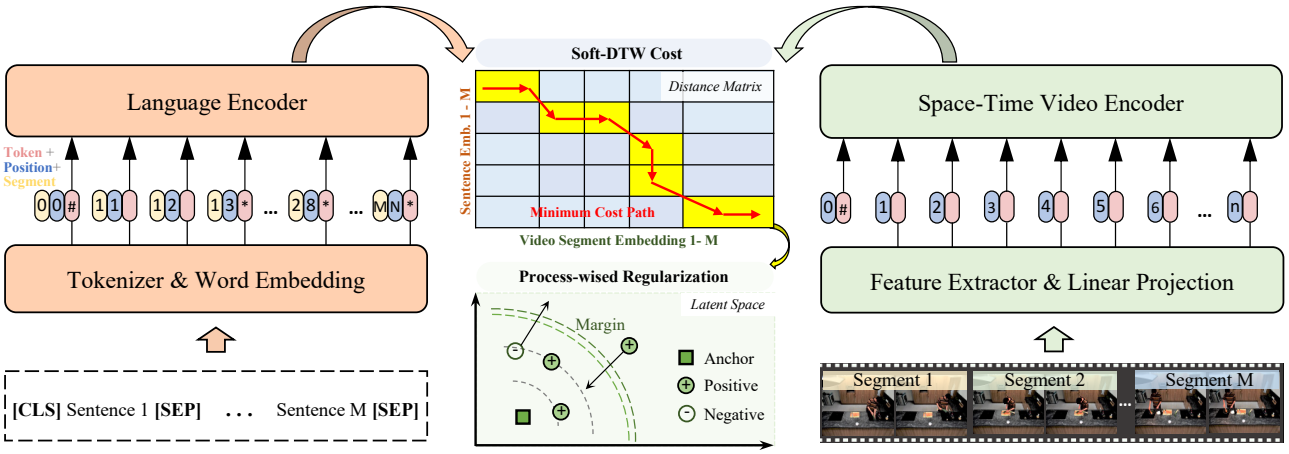


Figure 2. An overview of the proposed TCP. We first obtain video and text embeddings using a feature extractor and tokenizer, respectively. To prepare the inputs for the encoders, we add position and segment embeddings (Section 3.1). We leverage the cost of soft-DTW as the optimization objective for cross-modal sequence alignment (Section 3.2). For the sequences of each modality, we enforce the elements of the sequence to be embedded in the temporal location of the corresponding Brownian bridge process in the latent space via regularization term for the soft-DTW cost (Section 3.3).

$$\begin{aligned} d(1,1) &= D_{1,1}, \\ d(i,1) &= D_{i,1} + d(i-1,1), \\ d(1,j) &= D_{1,j} + d(1,j-1), \end{aligned} \tag{3}$$

where $i \in [2, N]$, $j \in [2, M]$. Then the distance matrix D can be calculated with the dynamic programming:

$$d(i,j) = D_{i,j} + min\left\{d(i,j-1), d(i-1,j), d(i-1,j-1)\right\} \tag{4}$$

Soft-DTW Cost. We can find the minimum cost path from D via the backtracking algorithm and calculate the cost value. However, the minimum cost of soft-DTW can not be directly used as an optimization object due to its undifferentiable min operator. Therefore, we select a smooth and continuous version of DTW, the soft-DTW [12], which replaces the min operator in a smooth way:

$$min^s(d_1, d_2, ..., d_n) = -\lambda \log \sum_{i=1}^{n} e^{\frac{-d_i}{\lambda}}, \tag{5}$$

where $0 < \lambda$ is a parameter for smoothness. A larger λ value leads to smoother results. soft-DTW will degenerate into DTW when $\lim(\lambda \to 0)$.

Replacing min operator with min^s in the Equation (4), we get the soft-distance matrix $\hat{D}$. For concise calculation and description, we use a binary matrix S to record the minimum cost path with the same shape as $\hat{D}$. In S, the cells that belong to the minimum cost path are 1, other cells are 0. Finally, the soft-DTW alignment cost is,

$$\mathcal{L}_{V2P} = \langle S, \hat{D} \rangle \tag{6}$$

Soft-DTW can serve as a better optimization objective with smooth gradients in backpropagation even though no convex optimization function is specifically provided, which makes the training process stable.

3.3. Intra-modal Sequence Modeling

Brownian Bridge Process. Sequence alignment focus on cross-modal temporal connection while missing the temporal dynamic evolution in each modal (*e.g.*, smooth change between continuous clips of a video, context relevance of sentences in a paragraph). In addition, optimizing the network only with the alignment cost may result in trivial solutions. That is to say, all the representations of clip and sense collapse to a small cluster so that all the entries of the video-paragraph distance matrix are 0. Based on these findings and inspired by [47, 55], we model the sequence of video clips or sentences as a Brownian bridge process to capture the dynamic evolution along the temporal dimension. A process indicates a segment of a video or a paragraph. The transition density of the process conforms to a time-variant Gaussian distribution:

$$\begin{aligned} p(\boldsymbol{z}_t | \boldsymbol{z}_A, \boldsymbol{z}_T) &= \mathcal{N}\left((1-\alpha)\boldsymbol{z}_A + \alpha \boldsymbol{z}_T, \alpha(T-t)\right), \\ &\text{where } \alpha = \frac{t-A}{T-A}. \end{aligned} \tag{7}$$

where $\boldsymbol{z}_A$, $\boldsymbol{z}_T$ are the start and end points of the Brownian bridge process respectively, and $\boldsymbol{z}_t$ is an arbitrary point in the process. **1)** The mean value of this distribution requires $\boldsymbol{z}_t$ to be the linear combination of the start and end points of the trajectory according to their relative temporal distance. $\boldsymbol{z}_t$ should be more similar to $\boldsymbol{z}_A$ if $\boldsymbol{z}_t$ is near the start point. Otherwise $\boldsymbol{z}_t$ should be more similar to $\boldsymbol{z}_T$. **2)**

The variance of this distribution requires the uncertainty of $\boldsymbol{z}_t$ to increase and then decrease over time, which conforms to a normal distribution. **To sum up**, this distribution indicates that the elements with closer temporal distance are more similar with a smaller change range, which is very consistent with the dynamic changes of video or language sequences.

Process-wised Regularization. To map the sequence into the latent space of the Brownian bridge, we leverage contrastive learning to build the process. We first define the anchor, positive and negative samples in contrastive learning. Given a Brownian bridge $Z = \{\boldsymbol{z}_A, \cdots, \boldsymbol{z}_t, \cdots, \boldsymbol{z}_T\}$ where $\boldsymbol{z}_t$ is the point at the timestamp t. The linear combination (ref. Equation (7)) of the start and end points at the timestamp t is the anchor, $\boldsymbol{z}_t$ serves as positive while the points of other Brownian bridges serves as negatives. We next define the distance between the anchor and positive or negative points as:

$$d(\boldsymbol{z}_A, \boldsymbol{z}_t, \boldsymbol{z}_T) = \frac{1}{2\sigma^2} \left\| \boldsymbol{z}_t - (1-\alpha)\boldsymbol{z}_A - \alpha \boldsymbol{z}_T \right\|_2^2, \quad \text{where } \alpha = \frac{t-A}{T-A}. \tag{8}$$

where σ^2 is $\alpha(T-t)$, the variance of the Brownian bridge transition density in Equation (7). We finally use the following triplet margin loss for process-wised metric learning, dubbed PRT (Process-wised Regularization Term):

$$\mathcal{L}_{PRT} = [d(\boldsymbol{z}_A, \boldsymbol{z}_t, \boldsymbol{z}_T) - d(\boldsymbol{z}_A, \hat{\boldsymbol{z}}_t, \boldsymbol{z}_T) + \beta]_+ \tag{9}$$

where β is a margin parameter, '+' indicates the result keeps if the value in the brackets is greater than 0 otherwise set to 0. N represents the valid triplets (*e.g.*, greater than 0).

Application of PRT. The sequence can be the representation of video or language. We next discuss the application of PRT on video clips z_v and paragraph z_p respectively.

Given a video representation $\boldsymbol{z}_v = \left\{\boldsymbol{v}_1^1, ..., \boldsymbol{v}_i^j, ..., \boldsymbol{v}_N^M\right\}$ that contains M segments. Each segment is taken as a Brownian bridge process. The j-th segment $\boldsymbol{z}_v^j = \left\{\boldsymbol{v}_A^j, ..., \boldsymbol{v}_t^j, ..., \boldsymbol{v}_T^j\right\}$ where $\boldsymbol{v}_A^j$, $\boldsymbol{v}_T^j$ are the both end of the Brownian bridge. The positive in the timestamp t is $\boldsymbol{v}_t^j$ while the negatives are randomly selected from other segments, denoted as $\hat{\boldsymbol{v}}_t$. The process-wised regularization for video clips sequence modeling is:

$$\mathcal{L}(V) = \sum_{j=1}^{M} \sum_{t=A+1}^{T-1} [d(\boldsymbol{v}_A, \boldsymbol{v}_t, \boldsymbol{v}_T) - d(\boldsymbol{v}_A, \hat{\boldsymbol{v}}_t, \boldsymbol{v}_T) + \beta]_+^j \tag{10}$$

As for text paragraphs, conditions and definitions are somewhat different. Given the paragraph $\boldsymbol{z}_p = \{\boldsymbol{p}_1, ..., \boldsymbol{p}_j, ..., \boldsymbol{p}_M\}$ where each sentence corresponding to a video segment, the whole paragraph is taken as a Brownian bridge process. The positive in the timestamp j is p_j while the negatives $\hat{p_j}$ are randomly selected from another sentence in this paragraph. The process-wised regularization in paragraph sequence modeling is:

$$\mathcal{L}(P) = \sum_{j=2}^{M-1} [d(\boldsymbol{p}_1, \boldsymbol{p}_j, \boldsymbol{p}_M) - d(\boldsymbol{p}_1, \hat{\boldsymbol{p}}_j, \boldsymbol{p}_M) + \beta]_+ \tag{11}$$

3.4. Optimization Objective

The overall training objective used in our method is to minimize the combination of the cross-modal sequence alignment cost in Equation (6) and the intra-modal sequence modeling regularization term in Equation (10, 11):

$$\mathcal{L}(V, P) = \mathcal{L}_{V2P} + \eta(\mathcal{L}(V) + \mathcal{L}(P)), \tag{12}$$

where η is the weight of sequence modeling. The final optimization objective encourages video-paragraph pairs to have minimum alignment costs and fine-grained cross-modal matching (*i.e.*, clip-to-sentence) by modeling the dynamic evolutions along the temporal dimension in each respective modality.

4. Experiments

4.1. Experimental Settings

Pre-training Datasets. Our model is pre-trained on LF-VILA-8M [43] dataset, which combines multiple video-language pairs presented in HD-VILA-100M [50] to obtain long-form video with its text descriptions in temporal sequence. There are 8.5 million video-language pairs in LF-VILA-8M with 100.2 seconds duration on average, which is significantly longer than that in HD-VILA-100M. The language paragraph contains 307.9 words on average. The available large-scale video-language datasets, such as HowTo100M [33], are noisy, and thus competitive performance requires a significant amount of computation resources. Compared to HowTo100M, LF-VILA-8M is much cleaner and smaller in size while longer in average length, which is more suitable for modeling and training long sequences in our TCP. Unless otherwise stated, all the video-text pairs in the dataset are used for pre-training.

Implementation Details. The number of frames that represent each clip is set as $K = 8$ wherein the resolution of the frame is resized to 192×320. The size of patch in the frame is set to 8×8. In the pre-training process, all parameters of the two encoders are optimized by Adam with a learning rate of 1e-5 and a weight decay of 0.05. We implement TCP using PyTorch and pre-train the model on 8 NVIDIA A100 GPUs for 12 epochs with batch size 128. We fix the margin parameter $\beta = 0.2$ in Equation (9) and set the parameter for smoothness $\lambda = 0.5$ in Equation (5), the weight of sequence modeling $\eta = 1$ in Equation (12) as default value. The ablation study on parameters λ, η is illustrated in Appendix.

Method	Date	Video Input	PT Dataset	#Pairs PT	R@1	R@5	R@10	MedR↓
Fine-tuning:								
AVLnet [39]	2021	ResNeXt-101	HowTo100M	120M	27.1	55.6	66.6	4.0
TACo [51]	2021	I3D, S3D	HowTo100M	120M	28.4	57.8	71.2	4.0
Support Set [36]	2021	R(2+1)D-34	HowTo100M	120M	30.1	58.5	69.3	3.0
VideoCLIP [48]	2021	S3D	HowTo100M	110M	30.9	55.4	66.8	-
Frozen [1]	2021	Raw Videos	CC3M, WebVid-2M	5.5M	31.0	59.5	70.5	3.0
HD-VILA [50]	2022	ResNet-50	HD-VILA-100M	103M	35.6	65.3	78.0	3.0
LocVTP [7]	2022	Raw Videos	CC3M, WebVid-2M	5.5M	36.5	64.3	76.8	3.0
MILES [17]	2022	Raw Videos	CC3M, WebVid-2M	5.5M	37.7	63.6	73.8	3.0
Bridging [16]	2022	Raw Videos	CC3M, WebVid-2M	5.5M	37.6	64.8	75.1	3.0
Ours	2023	Raw Videos	LF-VILA-8M*	5.5M	**38.0**	**65.5**	**76.4**	**3.0**
Zero-shot:								
TACo [51]	2021	I3D, S3D	HowTo100M	120M	9.8	25.0	33.4	29.0
VideoCLIP [48]	2021	S3D	HowTo100M	110M	10.4	22.2	30.0	-
Support Set [36]	2021	R(2+1)D-34	HowTo100M	120M	12.7	27.5	36.2	24.0
HD-VILA [50]	2022	ResNet-50	HD-VILA-100M	103M	14.6	34.4	44.1	15.0
Frozen [1]	2021	Raw Videos	CC3M, WebVid-2M	5.5M	18.7	39.5	51.6	10.0
AVLnet [39]	2021	ResNeXt-101	HowTo100M	120M	19.6	40.8	50.7	9.0
LocVTP [7]	2022	Raw Videos	CC3M, WebVid-2M	5.5M	22.1	48.0	55.3	8.0
Bridging [16]	2022	Raw Videos	CC3M, WebVid-2M	5.5M	26.0	46.4	56.4	7.0
MILES [17]	2022	Raw Videos	CC3M, WebVid-2M	5.5M	26.1	47.2	56.9	7.0
Ours	2023	Raw Videos	LF-VILA-8M*	5.5M	**26.8**	**48.3**	**57.6**	**7.0**

Table 1. Comparison with state-of-the-art methods for paragraph-to-video retrieval on MSR-VTT 1K test set under two evaluation settings (zero-shot and fine-tuning). **Video Input:** the type of video embeddings as the input of the video encoder. I3D, S3D, R(2+1)D, and ResNeXt-101 are all pre-extracted features while Raw Videos means raw video frame pixels. **PT Dataset:** the pre-training dataset in each method wherein HowTo100M is the instruction domain and the others are the open domain. **#Pair PT:** the number of video-text pairs in the pre-training dataset. * indicates that we selecte 5.5M pre-training pairs in LF-VILA-8M for fair comparison.

Method	Date	Video Input	PT Dataset	#Pairs PT	R@1	R@5	R@50	MedR↓
ClipBERT [26]	2021	ResNet-50	COCO, Visual Genome	5.6M	21.3	49.0	-	6.0
HD-VILA [50]	2022	ResNet-50	HD-VILA-100M	103M	28.5	57.4	94.0	4.0
Frozen [1]	2021	Raw Videos	CC3M, WebVid-2M	5.5M	28.8	60.9	-	3.0
Support Set [36]	2021	R(2+1)D-34	HowTo100M	120M	29.2	61.6	94.7	3.0
TACo [51]	2021	I3D, S3D	HowTo100M	120M	30.4	61.2	93.4	3.0
LF-VILA [43]	2022	Raw Videos	LF-VILA-8M	8.5M	35.3	65.4	95.0	3.0
Ours	2023	Raw Videos	LF-VILA-8M	8.5M	**36.0**	**66.7**	**95.8**	**3.0**

Table 2. Comparison with state-of-the-art methods for paragraph-to-video retrieval on ActivityNet 1K val set. We additionally report the recall accuracy under a large candidate range, *i.e.*, R@50. The proposed TCP achieves state-of-the-art performance on all metrics.

4.2. Paragraph-to-Video Retrieval

Setup and Evaluation Metrics. In this task, soft-DTW is directly used to match the whole video and the paragraph. We evaluate our method under two settings: zero-shot and fine-tuning. In zero-shot, pre-trained models are directly used for downstream datasets and no fine-tuning is allowed. In fine-tuning, we finetune the downstream training set before evaluation. The performance evaluation metric is measured as the proportion of the queries with correct results in the Top-k retrievals as the performance evaluation metric, which is denoted as $R@k$ in this paper. $MedR$ represents the Median Rank, which measures the median position of the right option in the sequence. Lower $MedR$ indicates better performance.

Datasets. ActivityNet-Captions is constructed based on ActivityNet [5], which has 19,209 videos with an average duration of 117.6 seconds. For each of the videos, there are 3.65 sentences with 13.8 words on average. MSR-VTT [49] contains a total of 10,000 videos in 20 categories, and each video has 20 text descriptions.

Method	Dataset	
	LSMDC	MRSVTT
Fine-tuning:		
JSFusion [53]	73.5	83.4
ActBERT [57]	-	85.7
ClipBERT [26]	-	88.2
MERLOT [54]	81.7	-
VIOLET [14]	82.9	-
All-in-one-B [46]	83.1	91.4
Ours	**83.7**	**92.8**
Zero-shot:		
All-in-one-B [46]	56.3	80.3
Ours	**57.0**	**80.8**

Table 3. Comparison with state-of-the-art methods for video question answering on LSMDC and MRSVTT multiple-choice test set. The proposed TCP achieves state-of-the-art performance on both settings of both datasets.

Comparison with State-of-the-Art. Table 1 and 2 show the performance comparison between state-of-the-art methods and ours on MSR-VTT and ActivityNet datasets respectively. Our method achieves better retrieval accuracy according to all of the metrics on the MSR-VTT 1K test set in both fine-tuning and zero-shot evaluation settings, as presented in Table 1. Our model outperforms state-of-the-art methods with either significantly fewer training video-language data pairs or with the same number of training pairs. For ActivityNet 1K val set, as denoted in Table 2, compared with the latest advanced method [43], our model yields significant performance advantages with the same size of the training dataset, while the pre-extracted feature we utilized to train implies a faster than raw video data that [43] uses. These experimental results validate the effectiveness of our model for understanding long-form videos comprehensively.

4.3. Video Moment Retrieval

Setup and Evaluation Metrics. Video Moment Retrieval (VMR) aims to find the corresponding segment in a video based on a given language query. To transfer to this task, we re-train the widely used temporal grounding method 2D-TAN [56] by only replacing the input with our features of the pre-trained model. Following [21], we choose $R@_n^\theta$ as the evaluation metric for video moment retrieval, which indicates the percentage of the queries with at least one retrieved segment from top-k ones whose IoU with the target segment is larger than θ. In this paper, $n \in \{1, 5\}$ and $\theta \in \{0.5, 0.7\}$ are selected for evaluation.

Datasets. The temporal boundary indexes of the video segment are accessible in the video-sentence pairs given in ActivityNet-Captions [23], and thus we evaluate the video moment retrieval performance of our method in the ActivityNet-Captions dataset.

Comparison with State-of-the-Art. Our model achieves overall better performance than state-of-the-art methods, as illustrated in Table 4. For the model trained based on the pre-extracted feature, our model outperforms Support Set [36] and ClipBERT [26] by 4.6% and 3.9% relatively according to $R@_1^{0.5}$ metric. Compared to the latest [7] trained on raw videos, our pre-extracted feature-based model, which advances in training speed, surpasses it according to $R@_5^{0.7}$ and both of the $R@1$ metrics, and has competitive results according to $R@_5^{0.5}$ metric. This result illustrates the promising ability of our model to capture the fine-grained cross-modality alignment.

4.4. Video Question-Answering

Setup and Evaluation Metrics. Multiple-choice Video Question-Answering is another video-language task that aims to choose the correct answer among multiple choices based on a given video query. To fit on this task, we tune the model with cross-entropy loss to maximize the scores on positive pairs. We use accuracy as the evaluation metric for multiple-choice Video Question-Answering.

Datasets. We use MSR-VTT and LSMDC [38] for evaluation following [46]. LSMDC [38] contains 118,081 video moments collected from 201 movies.

Comparison with State-of-the-Art. In Table 3 we present the performance comparison of video question answering on MRS-VTT and LSMDC datasets respectively. Our model achieves state-of-the-art accuracies in both datasets. Specifically, compared with the latest method All-in-one-B [46], accuracy improvements of 0.6% and 1.4% have been achieved in MRS-VTT [49] and LSMDC [38] respectively. Furthermore, in the zero-shot evaluation setting, our model also surpasses All-in-one-B [46] by 0.7% and 0.5% in terms of accuracy. These results demonstrate that our model can effectively understand the content of the long-form video and match it with the textual descriptions.

4.5. Ablation Study

Regularization Term. We take the none regularization term as the baseline, and employ another two regularization terms as the contrast to PRT (Process-wised Regularization Term): Inverse Difference Moment (IDM) [11] and Noise Contrastive Estimation (NCE) [19].

IDM. The Inverse Difference Moment is defined as:

$$I(\mathbf{x}) = \sum_{i=1}^{n}\sum_{j=1}^{n} W(i,j)\Delta_{\mathbf{x}}(i,j), \tag{13}$$

where $W(i,j) = \frac{1}{(i-j)^2+1}$ serves as the temporal weight, $\Delta_{\mathbf{x}}$ is the self-similarity matrix of sequence $\mathbf{x}$, $\mathbf{x}$ can be any sequence of multiple modalities. Maximizing this objective encourages temporally close elements to be closer in the

Method	Date	Video Input	PT Dataset	#Pairs PT	$R@_1^{0.5}$	$R@_1^{0.7}$	$R@_5^{0.5}$	$R@_5^{0.7}$
Support Set [39]	2021	R(2+1)D-34	HowTo100M	120M	41.9	25.2	74.7	58.3
ClipBERT [26]	2021	ResNet-50	COCO, Visual Genome	5.6M	42.6	24.6	75.3	59.7
Frozen [1]	2021	Raw Videos	CC3M, WebVid-2M	5.5M	43.3	25.8	75.8	59.3
LocVTP [7]	2022	Raw Videos	CC3M, WebVid-2M	5.5M	46.1	27.6	**78.9**	63.7
Ours	2023	Raw Videos	LF-VILA-8M*	5.5M	**46.5**	**28.4**	78.2	**64.0**

Table 4. Comparison with state-of-the-art methods for video moment retrieval on ActivityNet 1K val set. The proposed TCP achieves state-of-the-art performance on most metrics.

Reg. term	MRS-VTT		ActivityNet
	PVR (R@1)	VQA (Accuracy)	VMR ($R@_1^{0.7}$)
None	16.4	51.2	18.6
IDM	24.2	76.3	25.8
NCE	26.4	**81.2**	27.8
PRT	**26.8**	80.8	**28.4**

Table 5. Ablation studies of regularization term.

latent space. To be used as a regularization term for optimization objective, the IDM maximization function can be converted to the following minimization:

$$\bar{I}(\mathbf{x}) = -\sum_{i=1}^{n}\sum_{j=1}^{n} \bar{W}(i,j)\Delta_{\mathbf{x}}(i,j), \tag{14}$$

where $\bar{W}(i,j) = (i-j)^2 + 1$. As Table 5 shows, IDM can alleviate the problem of the trivial solution to a certain extent while failing to achieve optimal performance.

NCE. Different from PRT which learns from triplet contrast, NCE learns from the contrast with multiple negative samples of a batch in the training, all samples except the $\boldsymbol{z}_t^i$ are taken as negative samples denoted as $\mathcal{M}$. We define the NCE-based regularization term as:

$$C(\mathbf{x}) = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(d(\boldsymbol{z}_A^i, \boldsymbol{z}_t^i, \boldsymbol{z}_T^i))}{\sum_{\boldsymbol{z}_t' \in \mathcal{M}} \exp(d(\boldsymbol{z}_A^i, \boldsymbol{z}_t', \boldsymbol{z}_T^i))}. \tag{15}$$

As shown in Table 5, NCE achieves comparable performance with PRT on paragraph-to-video retrieval and video moment retrieval tasks. In addition, NCE outperforms other regularization terms on the video question-answering task. These findings demonstrate that incorporating additional negative samples is beneficial for cross-modal semantic matching without significantly affecting the learning of intra-modal temporal dynamics.

Scalability Study. To figure out how our method scales with the size of pre-training dataset, we pre-train the model with 1.5M, 3.5M, 5.5M, and 8.5M data pairs successively. We report the text-to-video retrieval results on MSR-VTT and ActivityNet in Figure 3, TCP achieves ~4% improvement on MSR-VTT with 7M dataset expansion, ~2.5% higher than Frozen [1], which demonstrates its good scalability. It may be attributed to TCP's construction of inherent dynamics of sequential video/text.

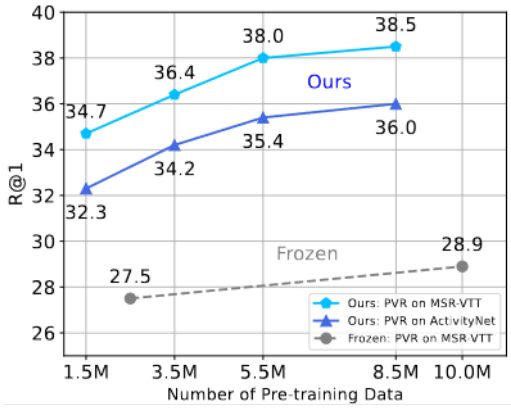


Figure 3. Scalability study with different number of pre-training data pairs. The top two lines indicate our paragraph-to-video retrieval results on MSR-VTT and ActivityNet in order.

5. Conclusion

In this paper, we propose to model video-language pairs as Temporal Concurrent Processes (TCP), which captures the inherent patterns of temporal concurrency, including intra-modal temporal dynamics and cross-modal temporal alignment. Specifically, TCP represents each modality as a goal-oriented stochastic process and aligns them as concurrent processes via dynamic programming and soft-DTW. Meanwhile, we introduce a regularization term that enforces each temporal sequence to agree with a time-variant Gaussian distribution within the Brownian bridge. Extensive experimental results show that TCP empirically produces well-generalized representations that are suitable for various downstream video-language understanding tasks.

Acknowledgment This work was supported in part by the National Natural Science Foundation of China No. 61976206 and No. 61832017, Beijing Outstanding Young Scientist Program NO. BJJWZYJH012019100020098, the Fundamental Research Funds for the Central Universities, the Research Funds of Renmin University of China 21XNLG05.

References

[1] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. *ICCV*, 2021. 1, 2, 6, 8

[2] Donald J. Berndt and James Clifford. Using dynamic time warping to find patterns in time series. In *KDD Workshop*, 1994. 3

[3] Rabi N. Bhattacharya and Edward C. Waymire. The brownian bridge. *Graduate Texts in Mathematics*, 2021. 2

[4] Lothar Breuer. Introduction to stochastic processes. *Statistical Methods for Climate Scientists*, 2022. 2

[5] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015. 6

[6] Kaidi Cao, Jingwei Ji, Zhangjie Cao, C. Chang, and Juan Carlos Niebles. Few-shot video classification via temporal alignment. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10615–10624, 2019. 3

[7] Meng Cao, Tianyu Yang, Junwu Weng, Can Zhang, Jue Wang, and Yuexian Zou. Locvtp: Video-text pre-training for temporal localization. In *European Conference on Computer Vision*, 2022. 1, 6, 7, 8

[8] C. Chang, De-An Huang, Yanan Sui, Li Fei-Fei, and Juan Carlos Niebles. D3tw: Discriminative differentiable dynamic time warping for weakly supervised action alignment and segmentation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3541–3550, 2019. 3

[9] Guo Chen, Yin-Dong Zheng, Limin Wang, and Tong Lu. Dcan: Improving temporal action detection via dual context aggregation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 248–257, 2022. 3

[10] Jinwoo Choi, Gaurav Sharma, Samuel Schulter, and Jia-Bin Huang. Shuffle and attend: Video domain adaptation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 678–695. Springer, 2020. 1

[11] Richard W. Conners and Charles A. Harlow. A theoretical comparison of texture algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1980. 7

[12] Marco Cuturi and Mathieu Blondel. Soft-dtw: a differentiable loss function for time-series. In *International Conference on Machine Learning*, 2017. 2, 4

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *ACL*, 2019. 1, 3

[14] Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. Violet : End-to-end video-language transformers with masked visual-token modeling. *ArXiv*, abs/2111.12681, 2021. 7

[15] Junyu Gao and Changsheng Xu. Fast video moment retrieval. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1503–1512, 2021. 2

[16] Yuying Ge, Yixiao Ge, Xihui Liu, Dian Li, Ying Shan, Xiaohu Qie, and Ping Luo. Bridging video-text retrieval with multiple choice questions. *CVPR*, 2022. 1, 2, 6

[17] Yuying Ge, Yixiao Ge, Xihui Liu, Alex Wang, Jianping Wu, Ying Shan, Xiaohu Qie, and Ping Luo. Miles: Visual bert pre-training with injected language semantics for video-text retrieval. In *ECCV*, 2022. 1, 2, 6

[18] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017. 1

[19] Michael U Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *International Conference on Artificial Intelligence and Statistics*, 2010. 7

[20] Tengda Han, Weidi Xie, and Andrew Zisserman. Temporal alignment networks for long-term video. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2896–2906, 2022. 3

[21] Jiachang Hao, Haifeng Sun, Pengfei Ren, Jingyu Wang, Qi Qi, and Jianxin Liao. Can shuffling video benefit temporal bias problem: A novel training framework for temporal grounding. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVI*, pages 130–147. Springer, 2022. 1, 3, 7

[22] Sanjay Haresh, Sateesh Kumar, Huseyin Coskun, Shahram Najam Syed, Andrey Konin, M. Zeeshan Zia, and Quoc-Huy Tran. Learning by aligning videos in time. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5544–5554, 2021. 3

[23] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715, 2017. 7

[24] Sateesh Kumar, Sanjay Haresh, Awais Ahmed, Andrey Konin, M. Zeeshan Zia, and Quoc-Huy Tran. Unsupervised action segmentation by joint representation learning and online clustering. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20142–20153, 2021. 3

[25] Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Unsupervised representation learning by sorting sequences. In *Proceedings of the IEEE international conference on computer vision*, pages 667–676, 2017. 1

[26] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L. Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. *CVPR*, 2021. 1, 6, 7, 8

[27] Dongxu Li, Junnan Li, Hongdong Li, Juan Carlos Niebles, and Steven C. H. Hoi. Align and prompt: Video-and-language pre-training with entity prompts. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4943–4953, 2021. 2

[28] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+

language omni-representation pre-training. *arXiv preprint arXiv:2005.00200*, 2020. 1, 2

[29] Kevin Lin, Linjie Li, Chung-Ching Lin, Faisal Ahmed, Zhe Gan, Zicheng Liu, Yumao Lu, and Lijuan Wang. Swinbert: End-to-end transformers with sparse attention for video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17949–17958, 2022. 1

[30] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, 2021. 3

[31] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 2

[32] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9879–9889, 2020. 1

[33] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2630–2640, 2019. 1, 5

[34] Guoshun Nan, Rui Qiao, Yao Xiao, Jun Liu, Sicong Leng, Hao Howard Zhang, and Wei Lu. Interventional video grounding with dual contrastive learning. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2764–2774, 2021. 2

[35] Athanasios Papoulis. Probability, random variables and stochastic processes. 1965. 2

[36] Mandela Patrick, Po-Yao Huang, Yuki M. Asano, Florian Metze, Alexander Hauptmann, João F. Henriques, and Andrea Vedaldi. Support-set bottlenecks for video-text representation learning. *ICLR*, 2021. 6, 7

[37] Daniel Revuz and Marc Yor. Continuous martingales and brownian motion. *Springer-Verlag, Berlin*, 1991. 2

[38] Anna Rohrbach, Marcus Rohrbach, and Bernt Schiele. The long-short story of movie description. In *Pattern Recognition: 37th German Conference, GCPR 2015, Aachen, Germany, October 7-10, 2015, Proceedings 37*, pages 209–221. Springer, 2015. 7

[39] Andrew Rouditchenko, Angie Boggust, David F. Harwath, Dhiraj Joshi, Samuel Thomas, Kartik Audhkhasi, Rogério Schmidt Feris, Brian Kingsbury, Michael Picheny, Antonio Torralba, and James R. Glass. Avlnet: Learning audio-visual language representations from instructional videos. In *Interspeech*, 2020. 6, 8

[40] Yaya Shi, Haiyang Xu, Chunfeng Yuan, Bing Li, Weiming Hu, and Zheng-Jun Zha. Learning video-text aligned representations for video captioning. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(2):1–21, 2023. 1

[41] Mykhailo Shvets, Wei Liu, and Alexander C Berg. Leveraging long-range temporal relationships between proposals for video object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9756–9764, 2019. 3

[42] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7464–7473, 2019. 1, 2

[43] Yuchong Sun, Hongwei Xue, Ruihua Song, Bei Liu, Huan Yang, and Jianlong Fu. Long-form video-language pre-training with multimodal temporal contrastive learning. *NIPS*, 2022. 3, 5, 6, 7

[44] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, 2019. 2

[45] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 1

[46] Alex Jinpeng Wang, Yixiao Ge, Rui Yan, Ge Yuying, Xudong Lin, Guanyu Cai, Jianping Wu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. All in one: Exploring unified video-language pre-training. *arXiv preprint arXiv:2203.07303*, 2022. 1, 7

[47] Rose E. Wang, Esin Durmus, Noah D. Goodman, and Tatsunori Hashimoto. Language modeling via stochastic processes. In *ICLR*, 2022. 4

[48] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, and Florian Metze Luke Zettlemoyer Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *ArXiv*, abs/2109.14084, 2021. 1, 6

[49] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016. 6, 7

[50] Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. Advancing high-resolution video-language representation with large-scale video transcriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5036–5045, 2022. 1, 2, 5, 6

[51] Jianwei Yang, Yonatan Bisk, and Jianfeng Gao. Taco: Token-aware cascade contrastive learning for video-text alignment. *ICCV*, 2021. 6

[52] Yuncong Yang, Jiawei Ma, Shiyuan Huang, Long Chen, Xudong Lin, Guangxing Han, and Shih-Fu Chang. Tempclr: Temporal alignment representation with contrastive learning. *ICLR*, 2023. 3

[53] Youngjae Yu, Jongseok Kim, and Gunhee Kim. A joint sequence fusion model for video question answering and retrieval. *ECCV*, 2018. 2, 7

[54] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. *Advances in Neural Information Processing Systems*, 34:23634–23651, 2021. 1, 2, 7

[55] Heng Zhang, Daqing Liu, Qi Zheng, and Bing Su. Modeling video as stochastic processes for fine-grained video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2225–2234, June 2023. 4

[56] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks for moment localization with natural language. In *AAAI Conference on Artificial Intelligence*, 2019. 7

[57] Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8743–8752, 2020. 1, 7