Supplementary Material

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

In this supplementary material, we first provide the preliminaries of soft-DTW and Brownian bridge process in Appendix A. Then, we give more ablation studies on the optimization objective in Appendix C. Next, we present the visualization of cross-modal sequence alignment in Appendix B. Last, we present more qualitative results on the downstream tasks in Appendix D.

A. Preliminaries

A.1. Soft-DTW

Soft Minimum. The soft minimum operation using a log-sum-exp technique:

$$\begin{aligned}\min(d_1, d_2, ..., d_n) &= -\max(d_1, d_2, ..., d_n) \\ &= -\log \sum_{i=1}^{n} e^{-d_i} \\ &= -\log \sum_{i=1}^{n} (e^{\frac{-d_i}{\lambda}})^{\lambda} \\ &\approx -\lambda \log \sum_{i=1}^{n} e^{\frac{-d_i}{\lambda}}.\end{aligned} \tag{1}$$

Then we ge the soft minimum equation for DTW:

$$\min^{\mathrm{s}}(\mathrm{d}_1, \mathrm{d}_2, ..., \mathrm{d}_\mathrm{n}) = -\lambda \log \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{e}^{\frac{-\mathrm{d}_\mathrm{i}}{\lambda}}, \tag{2}$$

where $0 < \lambda$ is a parameter for smoothness.

Differentiation. Given two sequences $\mathbf{x}, \mathbf{y}$, their distance matrix is denoted as $\Delta(\mathbf{x}, \mathbf{y}) := [\delta(x_i, y_j)_{i,j}] \in \mathbb{R}^{n \times m}$, the binary matrix is $\mathcal{S}_{n,m} \subset \{0, 1\}^{n \times m}$, a typical alignment matrix is $S \in \mathcal{S}_{n,m}$. The objective of soft-DTW can be written as:

$$\begin{aligned}\mathbf{dtw}_\lambda(\mathbf{x}, \mathbf{y}) &= \min{}^{\lambda} \{\langle \mathrm{S}, \Delta(\mathbf{x}, \mathbf{y}) \rangle, \mathrm{S} \in \mathcal{S}_{\mathrm{n,m}}\} \\ &= -\lambda \log \sum_{S \in \mathcal{S}_{n,m}} e^{-\langle S, \Delta(\mathbf{x}, \mathbf{y}) \rangle / \lambda}\end{aligned} \tag{3}$$

Differentiation of soft-DTW can be derived by chain rule:

$$\nabla_{\mathbf{x}} \mathbf{dtw}_\lambda(\mathbf{x}, \mathbf{y}) = (\frac{\partial \Delta(\mathbf{x}, \mathbf{y})}{\partial \mathbf{x}})^{\mathrm{T}} \mathbb{E}_\lambda[\mathrm{S}], \tag{4}$$

where $\mathbb{E}_\lambda[S] := \frac{1}{k^{\lambda}_{\mathrm{GA}}(\mathbf{x}, \mathbf{y})} \sum_{S \in \mathcal{S}_{n,m}} e^{-\langle S, \Delta(\mathbf{x}, \mathbf{y}) \rangle / \lambda} S$, is the average binary matrix S, $k^{\lambda}_{\mathrm{GA}}(\mathbf{x}, \mathbf{y})$ is the kernel interpreted by the normalization constant of $p_\lambda \propto e^{-\langle S, \Delta(\mathbf{x}, \mathbf{y}) \rangle / \lambda}$.

A.2. Brownian Bridge Process

A Brownian bridge process is a special form of standard Brownian motion. We first introduce the two important properties of the standard Brownian motion $\{B(t) : t \geq 0\}$:

$$\begin{aligned}&\{B(t + \tau) - B(\tau) : t \geq 0\}, \ \ \tau > 0 \\ &\{\frac{1}{c} B(c^2 t) : t \geq 0\}, \ \ c \neq 0\end{aligned} \tag{5}$$

Equation one indicates the Markov property and equation two is the property of self-similarity. Given a standard Brownian motion $\{B_t : t \geq 0\}$, the probability distribution $X_t = B_t - tB_1$, $X_t \sim N(0, t(1-t))$. The Brownian bridge process can be defined as $\{X_t : 0 \leq t \leq 1\}$. We next prove that it is a conditional stochastic process $\{B_t : 0 \leq t \leq 1 | B_1 = 0\}$:

$$\begin{aligned}P(B_t \leq x | B_1 = 0) =& P(t\hat{B}_{1/t} \leq x | \hat{B}_1 = 0) \\ & P(t(\hat{B}_{1/t} - \hat{B}_1) \leq x)\end{aligned} \tag{6}$$

where $t(\hat{B}_{1/t} - \hat{B}_1) \sim N(0, t(1-t))$. Based on this,

$$\begin{aligned}X_t &\overset{\mathrm{d}}{=} t(\hat{B}_{1/t} - \hat{B}_1) \\ &(B_t | B_1 = 0)\end{aligned} \tag{7}$$

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

We can draw a conclusion that the Brownian bridge is a conditional stochastic process.

B. Cross-modal Sequence Alignment

We provide two visualization examples of cross-modal sequence alignment via soft-DTW in Figure 1, which are randomly selected from the ActivityNet val set. For convenient observation, we have normalized the similarity to [0, 1]. As shown in the figures, the value of the main diagonal of the similarity matrix is the largest in its respective rows and columns. That is to say, semantic alignments are achieved via the minimum cost path of soft-DTW. More importantly, the similarity decreases gradually along the time direction, which demonstrates that the learned model can grab the temporal dynamics within the sequence.

C. More Ablation Studies

The following ablation studies are evaluated by Paragraph-Video Retrieval on MSR-VTT denoted as **PVR**(R@1), Video Question-Answering on MSR-VTT denoted as **VQA (Acc.)**, Video Moment Retrieval on ActivityNet denoted as **VMR (R@$^{0.7}_{1}$)**. We use all data pairs of LF-VILA-8M for pre-training in those ablation studies.

To demonstrate the advantages of soft-DTW for cross-modal sequence alignment, we have used InfoNCE to pre-train our model, which forces representations of video-text pairs to be close. Our TCP outperforms the InfoNCE baseline on all tracks with similar computing resources, as shown in Table 1.

Objective	#PT Time	**PVR**(R@1)	**VQA (Acc.)**	**VMR (R@$^{0.7}_{1}$)**
InfoNCE	75h	34.8	61.4	24.9
TCP	76h	**38.5**	**80.8**	**28.4**

Table 1. Comparison with InfoNCE baseline.

We further explore the effect of each component in the proposed TCP in the Table 2. As we have mentioned before, solely training with the soft-DTW cost encounters trivial solutions, thus resulting in poor performance. Here we further conclude that focusing only on inherited dynamics (*e.g.*, Pw Reg.) without semantic alignment also fails to achieve satisfactory results.

soft-DTW	Pw Reg.	**PVR**(R@1)	**VQA (Acc.)**	**VMR (R@$^{0.7}_{1}$)**
✓	✗	28.7	51.2	18.6
✗	✓	32.1	54.3	22.7
✓	✓	**38.5**	**80.8**	**28.4**

Table 2. Ablation studies on components of pre-training objective.

Hyper-parameters. We conduct ablation studies on the smoothness parameter λ in soft-DTW, and the weight of sequence modeling η in the optimization objective. The results are shown in Table 3. We evaluate on MRS-VTT test set with two downstream tasks: paragraph-to-video retrieval denoted as PVR and video question answering denoted as VQA. The results show that $\lambda = 0.5$ and $\eta = 1$ are optimal. TCP is sensitive to λ. Too low values lead to significant performance degradation both on PVR and VQA. By comparison, TCP is less sensitive to η. Nevertheless, an appropriate value is also necessary.

λ	η	**PVR (R@1)**	**VQA (Accuracy)**
0.1	1.0	24.1	68.7
0.5		**27.8**	**80.8**
0.9		25.6	74.2
0.5	0.5	25.4	77.6
	1.0	**27.8**	**80.8**
	2.0	26.7	78.2

Table 3. Ablation studies of smoothness parameter λ and the weight of sequence modeling η. PVR represents the paragraph-to-video retrieval task.

D. More Qualitative Results

Video Retrieval. A visualization example of video retrieval via soft-DTW is shown in Figure 2, which is randomly selected from the MRS-VTT test set. As shown in the figure, the Top-5 retrieved videos share similar content: 'sing'. The most confident matching video contains all the key information in the text description.

Video Moment Retrieval. Two visualization examples of video retrieval are given in Figure 3, which are randomly selected from the ActivityNet val set. As shown in the figure, the similarity curve is significantly higher in the ground truth than in other places, which shows that our method can approach the results obtained by the temporal grounding method 2D-TAN only via similarity calculation.

Video Question-Answering. Two visualization examples of video question-answering are given in Figure 4, which are randomly selected from the MRS-VTT test set. The first example is the correct answer. We can see that the text description selected by the model is very consistent with the video content as a query. The model can exclude the interference of other options, even if some of them match the video content. The second is an example of the wrong answer. It can be seen that the answer chosen by the model is similar to the video query to a certain extent. The answer is relatively broad. Although it cannot completely express the video content, there are no obvious errors. This also shows that our model has strong semantic generalization ability from the side.

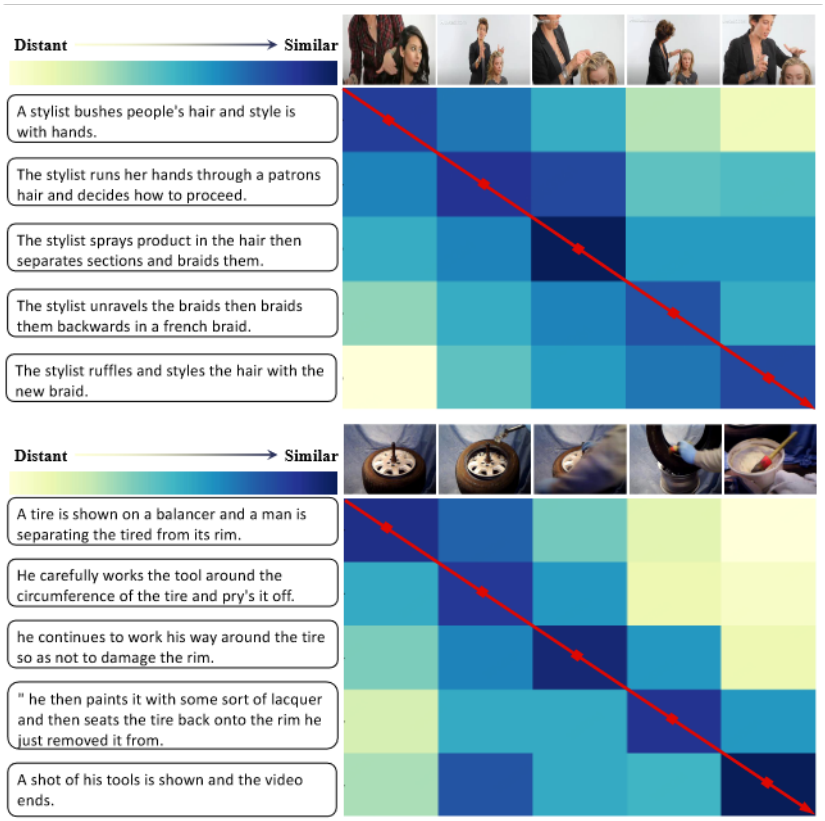

Figure 1. Visualization of cross-modal sequence alignment. The red line indicates the minimum cost path of soft-DTW.

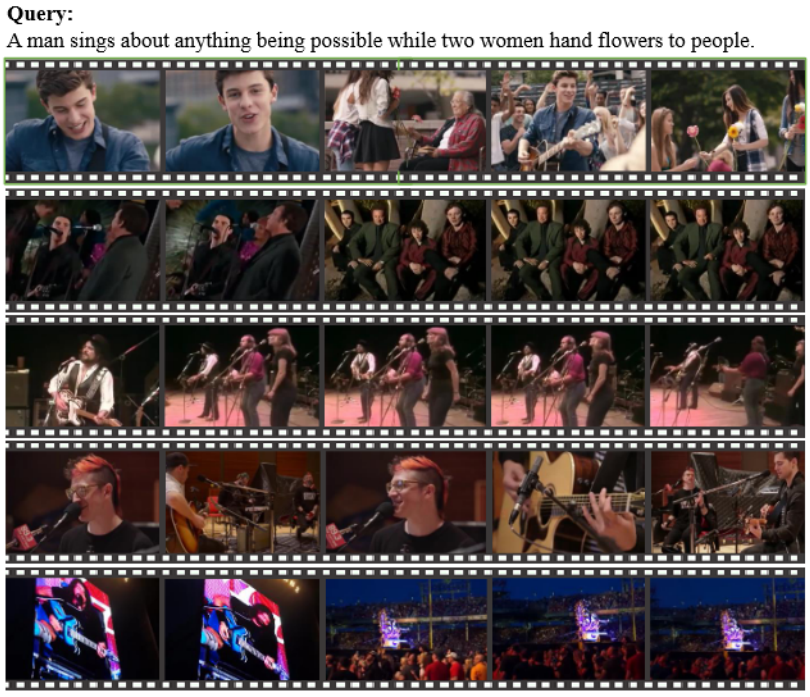

Figure 2. Visualization of video retrieval. The query on the top is a text description of a video. The five rows represent the Top-5 retrieved videos via soft-DTW cost directly. Ground truth is highlighted with green borderlines.

Figure 3. Two visualization examples of video moment retrieval. The query on the top is a text description of a video. The last block records the video clips similarity curve between video clips and the sentence query.

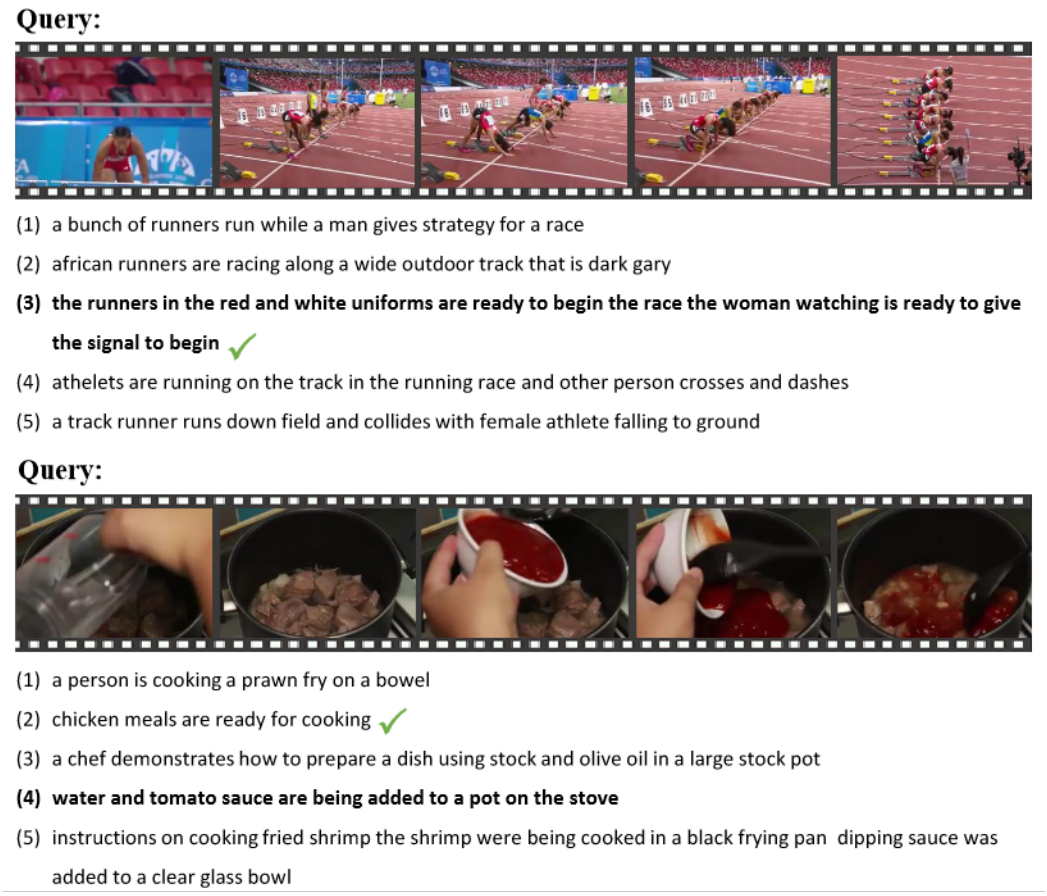

Figure 4. Two visualization examples of video question-answering. The query on the top is a video. The five rows represent the given candidates. **Ground-truth** is highlighted in bold and the prediction is denoted with ✓.