TS2-Net: Token Shift and Selection Transformer for Text-Video Retrieval

Yuqi Liu[1,2*], Pengfei Xiong[2], Luhui Xu[2],
Shengming Cao[2], and Qin Jin[1] (✉)

[1] School of Information, Renmin University of China
[2] Tencent
{yuqi657,qjin}@ruc.edu.cn,
xiongpengfei2019@gmail.com, {lukenxu,devancao}@tencent.com

Abstract. Text-Video retrieval is a task of great practical value and has received increasing attention, among which learning spatial-temporal video representation is one of the research hotspots. The video encoders in the state-of-the-art video retrieval models usually directly adopt the pre-trained vision backbones with the network structure fixed, they therefore can not be further improved to produce the fine-grained spatial-temporal video representation. In this paper, we propose Token Shift and Selection Network (TS2-Net), a novel token shift and selection transformer architecture, which dynamically adjusts the token sequence and selects informative tokens in both temporal and spatial dimensions from input video samples. The token shift module temporally shifts the whole token features back-and-forth across adjacent frames, to preserve the complete token representation and capture subtle movements. Then the token selection module selects tokens that contribute most to local spatial semantics. Based on thorough experiments, the proposed TS2-Net achieves state-of-the-art performance on major text-video retrieval benchmarks, including new records on MSRVTT, VATEX, LSMDC, ActivityNet, and DiDeMo. Code is available at https://github.com/yuqi657/ts2_net.

Keywords: text-video retrieval, token shift, token selection

1 Introduction

With advanced digital technologies, massive amount of videos are generated and uploaded online everyday. Searching for target videos based on users' text queries is a task of great practical value and has attracted increasing research attention. Over the past years, different text-video benchmarks have been established [2,46,10,40,44,25] and various text-video retrieval approaches have been

[0] ✉ Corresponding author
[*] This work is done when Yuqi is an intern at Tencent

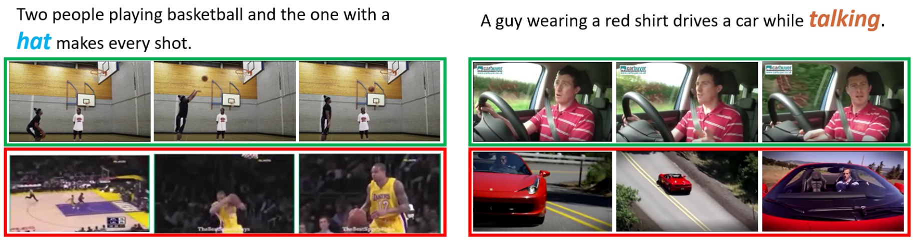
Two people playing basketball and the one with a hat makes every shot.
A guy wearing a red shirt drives a car while talking.

Fig. 1. The text-video retrieval examples that require fine-grained video representation. Left: the small object 'hat' is important for correctly retrieving the target video. Right: the subtle movement of 'talking' is crucial for the correct retrieval of the target video. Green boxes depict the positive video result, while red boxes are negative candidates

proposed [11,17,21,31,30,33], which usually formulate the task as a learning and matching task based on a similarity function between the text query and candidate videos in the corpus. With the success of deep neural networks [9,20,45], deep learned features have replaced manually-designed features. A text-video retrieval engine is generally composed of a text encoder and a video encoder, which maps the text query and the video candidate to the same embedding space, where the similarity can be easily computed using a distance metric.

Building a powerful video encoder to produce spatial-temporal feature encoding for videos, that can simultaneously capture motion between video frames, as well as entities in video frames, has been one of the research focuses for text-video retrieval in recent years [29,3,32]. Lately, Transformer has become the dominant visual encoder architecture, and it enables the training of video-language models with raw video and text data [4,34,19,12]. Various video transformers [3,32,5,8], considering both spatial and temporal representations, have achieved superior performance on major benchmarks. However, these models still lack fine-grained representation capacity in either spatial or temporal dimension. For example, the video encoder in models [34,19,12] normally consists of a single-frame feature extraction module followed by a global feature aggregation module, which lacks fine-grained interaction between adjacent frames and only aggregates the frame-level semantic information. Although the video encoder in Frozen [4] employs divided space-time attention, it uses only one [CLS] token as the video representation, failing to capture the find-grained spatial-temporal details. In general, all these models can effectively represent obvious motions and categorical spatial semantics in the video, but still lack the capacity for subtle movement and small objects. They will fail in cases such as illustrated in Fig.1, where the video encoder needs to capture the small object ('hat') and subtle movement ('talking') in order to retrieve the correct target videos.

Based on the structure of video transformer, video sequence is spatially and temporally divided into consecutive patches. To enhance modeling of small objects and subtle movements, patch enhancement is an intuitive and straightforward solution. This motivates us to find a feasible way to incorporate spatial-temporal patch contexts into encoded features. The shift operation is introduced in TSM[29], which shifts parts of the channel along temporal dimension. Shift Transformer[50] applies shift in visual transformer to enhance temporal modeling. However, the architecture of transformer is different from CNN, such partial shift operation damages the completeness of each token representation.

Therefore, in this paper, we propose TS2-Net, a novel token shift and selection transformer network, to realize local patch feature enhancement. Specifically, we first adopt the token shift module in TS2-Net, which shifts the whole spatial token features back-and-forth across adjacent frames, in order to capture local movement between frames. We then design a token selection module to select top-K informative tokens to enhance the salient semantic feature modeling capability. Our token shift module treats the features of each token as a whole, and iteratively swaps token features at the same location with neighbor frames, to preserve the complete local token representation and capture local temporal semantics at the same time. The token selection module estimates the importance of each token feature of patches with a selection network, which relies on the correlation between all spatial-temporal patch features and [CLS] tokens. It then selects tokens which contributes most to local spatial semantics. Finally, we align cross-modal representation in a fine-grained manner, where we calculate the similarity between text and each frame-wise video embedding and aggregate them together. TS2-Net is optimized with video-language contrastive learning.

We conduct extensive experiments on several text-video retrieval benchmarks to evaluate our model, including MSRVTT, VATEX, LSMDC, ActivityNet, and DiDeMo. Our proposed TS2-Net achieves the state-of-the-art performance on most of the benchmarks. The ablation experiments demonstrate that the proposed token shift and token selection modules both improve the fine-grained text-video retrieval accuracy. The main contributions of this work are as follows:

- We propose a new perspective of video-language learning with local patch enhancements to improve the text-video retrieval.
- We introduce two modules, token shift transformer and token selection transformer, to better model video representation temporally and spatially.
- We report new records of retrieval accuracy on several text-video retrieval benchmarks. Thorough ablation studies demonstrate the merits of our patch enhancement concept.

2 Related Work

2.1 Video Retrieval

Various approaches have been proposed to deal with text-video retrieval task, which usually consist of off-line feature extractors and feature fusion module [48,31,21,17,11,30,14,43]. MMT[21] uses a cross-modal encoder to aggregate feature extracted by different experts. MDMMT[17] further utilizes knowledge learned from multi-domain datasets. Recent works [26,4,34,19,12] attempt to train text-video model in an end-to-end manner. ClipBERT[26] is the pioneering end-to-end text-video pretrain model. Its promising results show that jointly train high-level semantic alignment network with low-level feature extractor is beneficial. CLIP4Clip[34] and CLIP2Video[19] transfer knowledge from pretrained CLIP[37] to video retrieval task. However, these models still lack fine-grained representation capacity in either spatial or temporal dimension. Different from previous works, we aim to model fine-grained spatial and temporal information to enhance text-video retrieval.

2.2 Visual-Language Pre-training

Viusal-language pre-training models has shown promising results in visual-and-language tasks such as image retrieval, image caption and video retrieval. In works such as Unicoder-VL[27], VL-BERT[41] and VLP[51], text and visual sequence are input into a shared transformer encoder. In Hero[28], ClipBERT[26] and Univl[33], text and visual sequence are encoded independently, then a cross-encoder is used to fuse different modality. While in Frozen[4], CLIP[37], text and visual sequence are encoded independently and a contrastive loss is used to align text and visual embedding. Our work use the two-stream structure, where text feature and video feature are encoded independently, then a cross-modal contrastive loss is used to align them.

2.3 Video Representation Learning

Early works use 2D or 3D-CNN to encode video feature [9,20,20,29]. Recently, Visual Transformer(ViT)[16] has shown great potential in image modeling. Many works attempt to transfer ViT into video domain [3,5,8,32]. TimeSformer[5] and ViViT[3] propose variants of spatial-temporal video transformer. There are several works exploring shift operation to enable 2D network learn temporal information, including TSM[29] and Shift Transformer[50]. They shift parts of the channel along the temporal dimension. Different from previous work, we consider token shift operation, which we shift all channels of selected visual tokens to the temporal dimension rather than partial shift (i.e. shift some channels). Token selection has been used to reduce redundancy problem in transformer based visual model. Dynamic ViT[39] and STTS[42] use token selection for efficiency. Perturbed maximum is proposed in [6] to make top-K differentiable. Based on differential top-K[13], our work designs a light-weight token selection module to select informative tokens for effective temporal-spatial modeling.

3 Method

The goal of text-video retrieval is to find the best matching videos based on the text query. Fig.2 illustrates the overall structure of the proposed TS2-Net model for the text-video retrieval task, which consists of three key components:

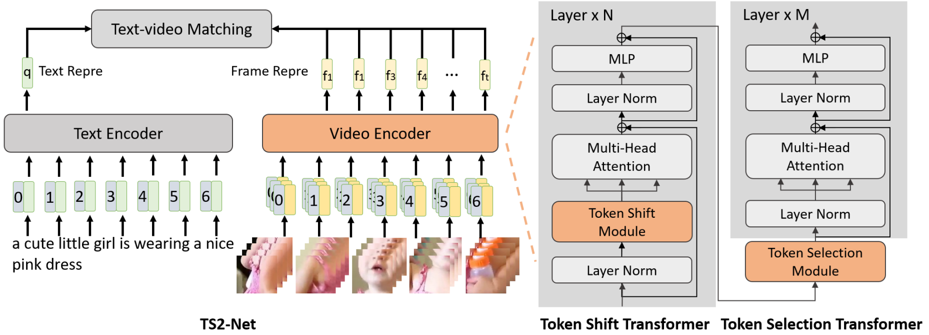

Fig. 2. Overview of the proposed TS2-Net model for text-video retrieval, which consists of three key components: the text encoder, the video encoder, and the text-video matching. The video encoder is composed of the Token Shift Transformer and Token Selection Transformer. ('Repre' is short for 'Representation')

the text encoder, the video encoder, and the text-video matching. The text encoder encodes the sequence of query words into a query representation q. In this paper, we use GPT [38] model as the text encoder. By adding a special token [EOS] at the end of query word sequence, we employ the encoding of [EOS] by the GPT encoder as the query representation q. The video encoder encodes the sequence of video frames into a sequence of frame-wise video representation $v = \{f_1, f_2, \ldots, f_t\}$. Based on the query and video representation, q and v, the text-video matching computes the cross-modal similarity between the query and video candidate. In following sections, we first elaborate the core ingredients of our video encoder, namely the token shift transformer (Sec.3.1) and the token selection transformer (Sec.3.2), and finally present our text-video matching strategy in details (Sec.3.3).

3.1 Token Shift Transformer

Token shift transformer is based on Vision Transformer (ViT) [16]. It inserts a token shift module in the transformer block. Let's review ViT model first, and then describe our modification to ViT. Given an image $\boldsymbol{I}$, ViT first splits $\boldsymbol{I}$ into N patches $\{p_0, p_1, \ldots, p_{n-1}\}$. To eliminate ambiguity, we use *token* to represent *patch* below. After adding a [CLS] token p_{cls}, the token sequence $\{p_{cls}, p_0, p_1, \ldots, p_{n-1}\}$ is fed into a stack of transformer blocks. Then the image embedding is generated by either averaging all the visual tokens or using the [CLS] token p_{cls}. In this work, we use p_{cls} as the image embedding. Token shift transformer aims to effectively model subtle movements in a video. The proposed token shift operation is a parameter-free operation, as illustrated in Fig.3. Suppose we have a video $\boldsymbol{V} \in \mathbb{R}^{T \times N \times C}$, where T represents the number of frames, N refers to the number of tokens per frame, and C represents the feature dimension. We feed T frames into ViT to encode frame feature. In certain ViT layer, we shift some tokens from adjacent frames to the current frame to

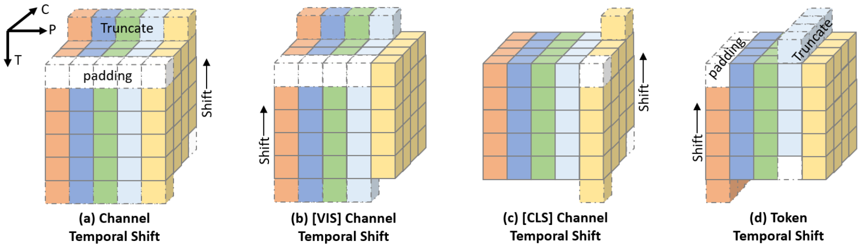
C
P
T
Truncate
padding
Shift
(a) Channel Temporal Shift
Shift
(b) [VIS] Channel Temporal Shift
Shift
(c) [CLS] Channel Temporal Shift
padding
Truncate
Shift
(d) Token Temporal Shift

Fig. 3. Illustration of different types of Shift operation and our proposed Token Temporal Shift. 'T, P, C' refer to video temporal dimension, video token, and feature channel respectively. Each vertical cube group represents a spatial-temporal video token. Cubes with dash line represent tensor truncated, and white cubes represent tensor padding. In Shift-Transformer [50], tokens are shifted along the channel dimension, while our proposed Token Shift Module does not compromise the integrity of a video token

exchange information of adjacent frames. Note that we use a bi-directional token shift in our implementation. By token shift operation across adjacent frames, our model is able to capture subtle movements in the local temporal interval.

Shift-Transformer [50] has also explored several shift variants on the visual transformer architecture. Fig.3 visualizes the difference between these shift variants and our proposed token shift. A naive channel temporal shift swaps part of channels of a frame tensor along temporal dimension, as shown in Fig. 3(a). Shift-Transformer [50] also presents [VIS] channel temporal shift and [CLS] channel temporal shift, as shown in Fig.3(b)(c). They fix tensor in token dimension and shift parts of channels for chosen token along the temporal dimension. Different from these works, our token shift transformer emphasizes the token dimension, where we shift whole channels of a token back-and-forth across adjacent frames, as shown in Fig.3(d). We believe our token shift is better for ViT architecture, because different from the CNN architecture, each token in ViT is independent and contains unique spatial information with respect to its location. Thus shifting parts of channels destroys the integrity of the information contained in a token. On the contrast, shifting a whole token with all channels can preserve complete information contained in a token and enable cross-frame interaction.

However, if we shift most of the tokens in every ViT layer, it damages the spatial modeling ability, and the information contained in these tokens is no longer accessible in the current frame. We therefore use a residual connection between original feature and token shift feature, as illustrated in Fig.2. In addition, we assume that shallow layers are more important to model spatial features, so shifting in shallow layers could harm spatial modeling. We thus choose to apply token shift operation only in deeper layers in our implementation.

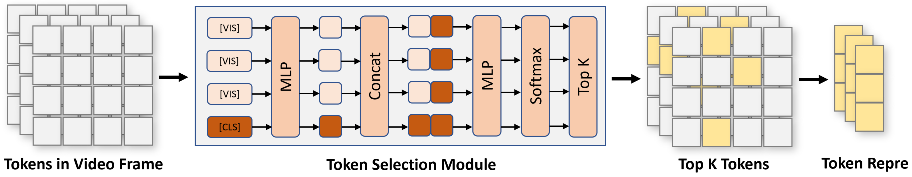

Fig. 4. Illustration of Token Selection Module. Top-K informative tokens are selected per frame from original spatial-temporal tokens for following feature aggregation

3.2 Token Selection Transformer

Aggregating information from each frame is a necessary step in building the video representation. A naive solution to aggregate per-frame information is by adding some temporal transformer layers, or by mean pooling as CLIP4Clip[34]. We argue that aggregation with only the [CLS] token leads to missing important spatial information (i.e. some objects). An alternative way is using all tokens from all frames to aggregate information, but this introduces redundancy problem, leading to the pitfall of some background tokens with irrelevant information dominating the final video representation.

In this work, we propose the token selection transformer by inserting a token selection module, which aims to select informative tokens per frame, especially those tokens containing salient semantics of objects, for video feature aggregation. As shown in Fig.4, top-K informative tokens are selected via the trainable token selection module every frame.

The input of the token selection module is a sequence of tokens of each frame $\boldsymbol{I} = \{p_{cls}, p_0, p_1, \dots, p_{n-1}\} \in \mathbb{R}^{(N+1)\times C}$. We first apply an MLP over $\boldsymbol{I}$ for channel dimension reduction and output $\boldsymbol{I}' = \{p'_{cls}, p'_0, p'_1, \dots, p'_{n-1}\} \in \mathbb{R}^{(N+1)\times \frac{C}{2}}$. We then use $p^{'}_{cls}$ as a global frame feature and concatenate it with each local token p'_i, $\hat{p}_i = [p'_{cls}, p'_i]\,, 0 \leq i < N$. We finally feed all the concatenated token features to another MLP followed by a Softmax layer to predict the importance scores, which can be formulated as:

$$\boldsymbol{S} = \mathrm{Softmax}(\mathrm{MLP}(\hat{p})) \in \mathbb{R}^{(N+1)}. \tag{1}$$

We select indices of K most informative tokens based on $\boldsymbol{S}$, denoting as $\mathbf{M} \in \{0,1\}^{(N+1)\times K}$, where each column in $\mathbf{M}$ is a one-hot $(N+1)$ dimensional indicator. We extract top-K most informative tokens by:

$$\hat{\mathbf{I}} = \mathbf{M}^T \mathbf{I}, \tag{2}$$

After top-K token select on every frame, we input the selected tokens from all frames to a joint spatial-temporal transformer, to learn global spatial-temporal video representation. We also pick the most informative token from each frame as the frame-wise video encoding.

Differentiable TopK. Until now, both top-K operation and one-hot operation are non-differentiable. To make token selection module differentiable, we employ the perturbed maximum method proposed in [6]. Specifically, a discrete optimization problem with input $\boldsymbol{S} \in \mathbb{R}^{(N+1)}$ ($\boldsymbol{S}$ is the importance score matrix in Eq.1) and optimization variable $\mathbf{M} \in \mathbb{R}^{(N+1)\times K}$ ($\mathbf{M}$ is the index indicator matrix in Eq. 2) can be formulated as:

$$F(\boldsymbol{S}) = \max_{\mathbf{M}\in\mathcal{C}} \langle \mathbf{M}, \boldsymbol{S} \rangle, \mathbf{M}^*(\boldsymbol{S}) = \arg\max_{\mathbf{M}\in\mathcal{C}} \langle \mathbf{M}, \boldsymbol{S} \rangle, \tag{3}$$

where $F(\boldsymbol{S})$ represents the top-K selection operation, $\mathbf{M}^*(\boldsymbol{S})$ represents the optimal value. Based on Eq.3, we can select top-K informative tokens by $F(\boldsymbol{S})$. We calculate forward and backward pass following [1,13].

3.3 Text-Video Matching

The similarity between the text query and video candidate is computed by integrating the similarity between the query and each video frame. To be specific, given the query representation q and a sequence of frame-wise video representation $v = \{f_1, f_2, ..., f_t\}$, we compute the frame-level similarity as follows:

$$s_i = \frac{q \cdot f_i}{\|q\| \, \|f_i\|}. \tag{4}$$

The final text-video matching similarity is defined as the weighted combination of frame-level similarities:

$$s = \sum_{i=1}^{n} \alpha_i s_i, \tag{5}$$

where $\alpha_i = \frac{\exp(\lambda s_i)}{\sum_{i=1}^{n} \exp(\lambda s_i)}$ and λ is a temperature parameter. We set λ as 4 empirically in our experiments.

Symmetric cross-entropy loss is adopted as our training objective function. For each training step with B text-video pairs, we calculate symmetric cross-entropy loss as follows:

$$\mathcal{L}_t^{t2v} = -\frac{1}{B} \sum_i^B \log \frac{\exp\left(\tau \cdot \operatorname{sim}\left(q_i, v_i\right)\right)}{\sum_{j=1}^B \exp\left(\tau \cdot \operatorname{sim}\left(q_i, v_j\right)\right)}, \tag{6}$$

$$\mathcal{L}_t^{v2t} = -\frac{1}{B} \sum_i^B \log \frac{\exp\left(\tau \cdot \operatorname{sim}\left(q_i, v_i\right)\right)}{\sum_{j=1}^B \exp\left(\tau \cdot \operatorname{sim}\left(q_j, v_i\right)\right)}, \tag{7}$$

$$\mathcal{L} = \frac{1}{2}\left(\mathcal{L}_{t2v} + \mathcal{L}_{v2t}\right), \tag{8}$$

where τ is a trainable scaling parameter and $\operatorname{sim}(q, v)$ is calculated using Eq.5. During inference, we calculate the matching score between each text and video based on Eq.5, and return videos with the highest ranking.

4 Experiment

In this section, we carry out text-video retrieval evaluations on multiple benchmark datasets to validate our proposed model TS2-Net. We first ablate the core ingredients of our video encoder, the token shift transformer and the token selection transformer, on the dominant MSR-VTT dataset. We then compare our model with other state-of-the-art models on multiple benchmark datasets quantitatively and qualitatively.

4.1 Experimental Settings

Datasets. To demonstrate the effectiveness and generalization ability of our model, we conduct evaluations on five popular text-video benchmarks, including MSR-VTT[46], VATEX[44], LSMDC[40], ActivityNet-Caption[18,25], DiDeMo[2]. All these datasets are collected from different scenarios with various amounts of captions. Videos in different datasets also have different content styles and different lengths.

- **MSR-VTT**[46] contains 10,000 video clips with 20 captions per video. Our experiments follow 1k-A split protocol used in [21,31,35], where the training set has 9,000 videos with its corresponding captions and test set has 1,000 text-video pairs.
- **VATEX**[44] contains 34,991 video clips with several captions per video. We follow HGR[11] split protocol. There are 25,991 videos in the training set, 1,500 videos in the validation set and 1,500 videos in the test set.
- **LSMDC**[40] contains 118,081 video clips, which are extracted from 202 movies. Each video clip has one caption. There are about 100k videos in the training set, 7,408 videos in the validation set and 1,000 videos in the test set. Especially, videos in the test set are from movies disjoint with the training and validation set.
- **ActivityNet-Caption**[18,25] contains 20,000 YouTube videos. Following the same setting as in [34,49,21], we regard it as a paragraph-video retrieval by concatenate all descriptions of a video. We train our model on *train* split and test our model on *val1* split.
- **DiDeMo**[2] contains over 10k videos. There are 8,395 videos in the training set, 1,065 videos in the validation set and 1,004 videos in the test set. Following the same setting as in [34,31,26], we concatenate all descriptions of a video to retrieval videos with paragraphs.

Evaluation Metrics. We measure the retrieval performance using standard text-video retrieval metrics: Recall at K (R@K, higher is better), Median Rank (MdR, lower is better) and Mean Rank (MnR, lower is better). R@K calculates the fraction of correct videos among the top K retrieved videos. Similar to previous works [34,31,12], we use K=1,5,10 for different datasets. We also sum up all the R@K results as rsum to reflect the overall retrieval performance. MedR calculates the median rank of correct results in the retrieval ranking list and MeanR calculates the mean rank of correct results in the retrieval ranking list.

Table 1. Performance comparison with different parameter settings of the Token Shift Transformer on MSR-VTT-1k-A test split

			Text $\Longrightarrow$ Video				Video $\Longrightarrow$ Text				
Method	Layers	Ratio	R@1	R@5	R@10	MnR	R@1	R@5	R@10	MnR	rsum
Baseline	-	-	45.4	74.3	82.7	13.6	44.5	72.3	82.3	9.8	401.5
w/ Token Shift	1-12	25%	42.8	71.2	80.9	14.4	43.2	70.3	80.4	11.3	388.8
w/ Token Shift	3-12	25%	44.1	71.0	81.8	14.5	43.5	71.2	81.8	10.8	393.4
w/ Token Shift	5-12	25%	44.4	71.9	81.6	14.6	44.8	72.0	80.6	11.3	395.3
w/ Token Shift	7-12	25%	44.1	72.3	82.9	13.6	43.8	72.3	82.1	10.3	397.5
w/ Token Shift	9-12	25%	45.2	73.8	83.1	13.4	45.3	72.1	82.5	9.5	402
w/ Token Shift	11-12	12.5%	46.0	73.3	82.2	13.8	**45.8**	72.9	83.0	9.5	403.2
w/ Token Shift	11-12	50%	46.1	**74.5**	83.3	13.3	45.6	72.9	82.2	9.5	404.6
w/ Token Shift	**11-12**	**25%**	**46.2**	73.9	**83.8**	**13.0**	45.6	**73.5**	**83.2**	**9.3**	**406.2**

Table 2. Performance comparison between other shift operation variants and our proposed token shift module on MSR-VTT-1k-A test split

	Text $\Longrightarrow$ Video				Video $\Longrightarrow$ Text				
Method	R@1	R@5	R@10	MnR	R@1	R@5	R@10	MnR	rsum
Baseline	45.4	74.3	82.7	13.6	44.5	72.3	82.3	9.8	401.5
Channel Shift[50]	45.6	73.6	83.1	13.7	45.0	73.2	82.7	9.7	403.2
[VIS] Channel Shift[50]	45.1	73.8	83.5	13.9	44.7	73.3	82.2	9.8	402.6
[CLS] Channel Shift[50]	45.8	**74.3**	83.0	13.6	44.7	72.9	82.5	9.8	403.2
Token Shift	**46.2**	73.9	**83.8**	**13.0**	**45.6**	**73.5**	**83.2**	**9.3**	**406.2**

Implementation Details. The layer of GPT, token shift transformer and token selection transformer is 12, 12 and 4, respectively. The dimension of text embedding and frame embedding is 512. We initialize transformer layers in GPT, token shift transformer and token selection transformer with pre-trained weight from CLIP(ViT-B/32)[37], using parameters with similar dimension, while other modules are initialized randomly. We choose 4 most informative tokens in MSR-VTT, VATEX, ActivityNet-Caption, DiDeMo, and 1 in LSMDC. We set the max query text length as 32 and max video frame length as 12 in MSR-VTT, VATEX, LSMDC. For ActivityNet-Caption and DiDeMo, we set the max query text length and max video frame length as 64. We train our model with Adam[24] optimizer and adopt a warmup[23] setting. We choose a batch size of 128. The learning rate of GPT and token shift transformer is 1e-7 and the learning rate of token selection transformer is 1e-4.

4.2 Ablation Experiments

In this section, we evaluate the proposed token shift transformer and token selection transformer under different settings to validate their effectiveness. We conduct ablation experiments with the 1k-A test split on MSR-VTT[46]. We set our baseline model as the degraded TS2-Net model which removes the token shift and token selection modules from TS2-Net.

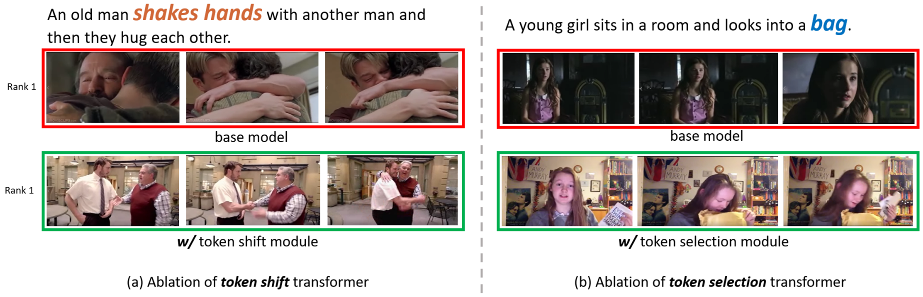
An old man shakes hands with another man and then they hug each other.
Rank 1
base model
Rank 1
w/ token shift module
(a) Ablation of token shift transformer
A young girl sits in a room and looks into a bag.
base model
w/ token selection module
(b) Ablation of token selection transformer

Fig. 5. The text-video retrieval results of different network architecture. Left: with *token shift transformer*, our model is able to distinguish 'shake hands', while the baseline model retrieves an incorrect video. Right: with *token selection transformer*, our model retrieves the correct video, although 'bag' is only shown in small part of video frames. Green boxes: correct target video; red boxes: incorrect target video.

Ablation of Token Shift Transformer. We first analyze the impact of some factors on the token shift module in Tab.1, including shift layer and shift ratio. Shift layer (in which layers should we insert token shift) and shift ratio (how many tokens should we shift) are two main factors that affect the final retrieval performance. The backbone of our token shift transformer is the 12-layer ViT. We thus experiment to insert the token shift module in different layers. As shown in Tab.1, shift operation in deeper layers (i.e. 11-12 layers) brings retrieval performance improvement. But if we shift more layers (i.e. 9-12 layers), it hurts the retrieval performance, and it hurts more if we operate shift in shallower layers (i.e. 1-12 layers). We think that shallow layers in ViT are more important in modeling spatial information, so shift in shallow layers damages spatial modeling ability. We thus choose to insert the token shift module in the 11-12 layers in the following experiments. In terms of shift ratio, we find that shifting 25% tokens back-and-forth across frames achieves the best retrieval performance. Despite some slight fluctuations, token shift with different ratios achieves better results than the baseline model. The improvement is more obvious especially for R@1.

We further conduct experiments to compare our proposed token shift module with other shift operation variants proposed in Shift-ViT[50]. As shown in Tab.2, our proposed token shift module outperforms all other shift operation variants. This is because our token shift operation can preserve the integrity of the token feature, posing minor impact on the spatial modeling ability. We visualize the retrieval results from the baseline model and the model with token shift transformer in Fig.5(a). With token shift transformer, the model is able to capture subtle movement such as 'shake hand'.

Ablation of Token Selection Transformer.

The token selection transformer follows the token shift transformer to select the most informative tokens for the next transformer propagation. We conduct experiment to verify what proportion of tokens is beneficial to the final retrieval in Tab.3. As can be observed, selecting fewer tokens per frame tends to achieve

Table 3. Comparison results with different settings of Token Selection Transformer

		Text $\Longrightarrow$ Video				Video $\Longrightarrow$ Text				
Method	top-K	R@1	R@5	R@10	MnR	R@1	R@5	R@10	MnR	rsum
Token Shift	1	46.2	73.9	83.8	13.0	45.6	73.5	83.2	9.3	406.2
w/ all token	50	45.8	73.5	83.4	13.5	44.7	73.1	82.4	9.4	402.9
w/ Random select	4	46.4	73.9	83.5	13.1	45.1	73.5	82.1	9.5	404.5
w/ Select token	2	47.0	74.2	83.6	13.1	**45.6**	74.0	83.5	9.3	407.9
w/ Select token	6	46.6	74.4	**84.3**	13.2	44.5	73.8	83.2	9.2	406.8
w/ Select token	8	46.4	73.9	83.5	13.2	45.0	74.1	**83.9**	9.2	406.8
TS2-Net	4	**47.0**	**74.5**	83.8	**13.0**	45.3	**74.1**	83.7	**9.2**	**408.4**

Table 4. Retrieval results on MSR-VTT-1kA. Other SOTA methods are adopted as comparisons. Note that CLIP2TV uses patch size of 16×16, so we use TS2-Net(ViT16) for fair comparison. All results in this table do not use inverted softmax

	Text $\Longrightarrow$ Video					Video $\Longrightarrow$ Text				
Method	R@1	R@5	R@10	MdR	MnR	R@1	R@5	R@10	MdR	MnR
CE[31]	20.9	48.8	62.4	6.0	28.2	20.6	50.3	64.0	5.3	25.1
TACo[47]	26.7	54.5	68.2	4.0	-	-	-	-	-	-
MMT[21]	26.6	57.1	69.6	4.0	24.0	27.0	57.5	69.7	3.7	21.3
SUPPORT-SET[36]	27.4	56.3	67.7	3.0	-	26.6	55.1	67.5	3.0	-
TT-CE[14]	29.6	61.6	74.2	3.0	-	-	-	-	-	-
T2VLAD[43]	29.5	59.0	70.1	4.0	-	31.8	60.0	71.1	3.0	-
HIT-pretrained[30]	30.7	60.9	73.2	2.6	-	32.1	62.7	74.1	3.0	-
Frozen[4]	31.0	59.5	70.5	3.0	-	-	-	-	-	-
MDMMT[17]	38.9	69.0	79.7	2.0	16.5	-	-	-	-	-
CLIP[37]	39.7	72.3	82.2	2.0	12.8	11.3	22.7	29.2	5.0	-
CLIP4Clip[34]	44.5	71.4	81.6	2.0	15.3	42.7	70.9	80.6	2.0	11.6
CAMoE[12]	44.6	72.6	81.8	2.0	13.3	45.1	72.4	83.1	2.0	10.0
CLIP2Video[19]	45.6	72.6	81.7	2.0	14.6	43.5	72.3	82.1	2.0	10.2
TS2-Net	**47.0**	**74.5**	**83.8**	**2.0**	**13.0**	**45.3**	**74.1**	**83.7**	**2.0**	**9.2**
CLIP2TV[22]	48.3	74.6	82.8	2.0	14.9	46.5	75.4	84.9	2.0	10.2
TS2-Net(ViT16)	**49.4**	**75.6**	**85.3**	**2.0**	**13.5**	**46.6**	**75.9**	**84.9**	**2.0**	**8.9**

better performance than selecting more. For example, the R@1 performance decreases from 47.0 to 45.8 while the number of selected tokens increases from 2 to 50. We consider that fewer informative tokens are sufficient to preserve the salient spatial information, while adding more tokens may bring redundancy problem. Although random selection also improves the performance slightly, it can not beat the proposed learnable token selection module. In Fig.5(b), we show a retrieval case from the baseline model and the model with token selection transformer. With token selection transformer, the model is able to capture the small object 'bag' in video frames.

4.3 Comparisons with State-of-the-art Models

MSR-VTT-1kA. We compare our proposed TS2-Net with other state-of-the-art methods on five benchmarks. Tab.4 presents the results on MSR-VTT-1kA

Table 5. Text-to-Video retrieval results on VATEX, LSMDC, ActivityNet and DiDeMo. QB-Norm uses dynamic inverted softmax during inference, while other methods report results without inverted softmax

VATEX					
Method	R@1	R@5	R@10	MdR	MeanR
Dual Enc.[15]	31.1	67.5	78.9	3.0	-
HGR[11]	35.1	73.5	83.5	2.0	-
CLIP[37]	39.7	72.3	82.2	2.0	12.8
CLIP4Clip[34]	55.9	89.2	95.0	1.0	3.9
QB-Norm*[7]	58.8	88.3	93.8	1.0	-
CLIP2Video[19]	57.3	90.0	**95.5**	1.0	3.6
TS2-Net	**59.1**	**90.0**	95.2	**1.0**	**3.5**

LSMDC					
Method	R@1	R@5	R@10	MdR	MeanR
JSFusion[48]	9.1	21.2	34.1	36.0	-
CE[31]	11.2	26.9	34.9	25.3	-
Frozen[4]	15.0	30.8	39.8	20.0	-
CLIP4Clip[34]	22.6	41.0	49.1	11.0	61.0
QB-Norm*[7]	22.4	40.1	49.5	11.0	-
CAMoE[12]	22.5	**42.6**	50.9	-	**56.5**
TS2-Net	**23.4**	42.3	**50.9**	**9.0**	56.9

ActivityNet					
Method	R@1	R@5	R@10	MdR	MeanR
CE[31]	20.5	47.7	63.9	6.0	23.1
ClipBERT[26]	21.3	49.0	63.5	6.0	-
MMT-Pretrained[21]	28.7	61.4	-	3.3	16.0
CLIP4Clip[34]	40.5	73.4	-	2.0	**7.5**
TS2-Net	**41.0**	**73.6**	**84.5**	**2.0**	8.4

DiDeMo					
Method	R@1	R@5	R@10	MdR	MeanR
ClipBERT[26]	20.4	48.0	60.8	6.0	-
TT-CE[14]	21.1	47.3	61.1	6.3	-
Frozen[4]	31.0	59.8	72.4	3.0	-
CLIP4Clip[34]	**42.5**	70.2	80.6	2.0	17.5
TS2-Net	41.8	**71.6**	**82.0**	**2.0**	**14.8**

test set. Our model outperforms previous methods across different evaluation metrics. With token shift transformer and token selection transformer, our model is able to capture subtle motion and salient objects, and thus our final video representation contains rich semantics. Compared with video-to-text retrieval, the gain on text-to-video retrieval is more significant. We consider it is because the proposed token shift and token selection modules enhance the video encoder, while a relative simple text encoder is adopted.

Other Benchmarks. Tab.5 presents text-to-video retrieval results on VATEX, LSMDC, ActivityNet-Caption and DiDeMo. Results on these datasets demonstrate the generalization and robustness of our proposed model. Our model achieves consistent improvements across different datasets, which demonstrates that it is beneficial to encode spatial and temporal features simultaneously by our token shift and token selection. Note that our performance surpasses QB-Norm[7] on LSMDC and VATEX even without inverted softmax, as shown in Tab.5. More detailed analysis can be found in supplementary materials.

4.4 Qualitative Results

We visualize some retrieval examples from the MSR-VTT testing set for text-to-video retrieval in Fig.6. In the top left example, our model is able to distinguish ‘hand rubbing’ (in the middle picture) during a guitar-playing scene. The bottom right example shows our model can distinguish ‘computer battery’ from ‘computer’. In the bottom left example, our model retrieves the correct video which contains all actions and objects expressed in the text query, especially the small object ‘microphone’ and tiny movement ‘talking’. In the bottom right

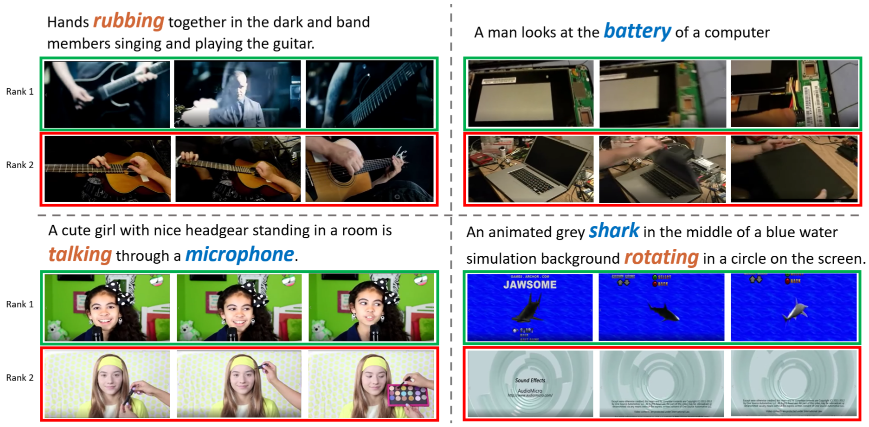
Hands rubbing together in the dark and band members singing and playing the guitar.
A man looks at the battery of a computer
Rank 1
Rank 2
A cute girl with nice headgear standing in a room is talking through a microphone.
An animated grey shark in the middle of a blue water simulation background rotating in a circle on the screen.
Rank 1
Rank 2
JAWSOME
Sound Effects
AudioMicro

Fig. 6. Visualization of text-video retrieval examples. We sorted results based on its similarity scores. Green: ground truth; Red: incorrect

example, our model retrievals the correct result although 'rotating' is a periodic movement and is hard to spot.

We also select a subset from the MSR-VTT-1kA test set. Queries in this subset are selected based on their corresponding video's visual appearance, where objects mentioned in query are shown in a small part of video and movements mentioned in query is slight. Such as '*little pet shop cat* getting a bath and washed with *little brush*', 'a golf player is trying to hit the ball into the *pit*'. Since such cases account for a small proportion, so the total number of this subset is 103. During inference, we calculate similarity between queries in subset with videos in whole test set. We compare our model with another strong baseline on this subset. Our model achieves 79.6 on R@1 metric, while CLIP4Clip[34] only achieves 39.8. There is a significant margin and this verifies the effectiveness of TS2-Net in handling local subtle movements and local small entities.

5 Conclusion

In this work, we propose Token Shift and Selection Network (TS2-Net), a noval transformer architecture with token shift and selection modules, which aims to further improve the video encoder for better video representation. A token shift transformer is used to capture subtle movements, followed by a token selection transformer to enhance salient object modeling ability. Superior experimental results show our proposed TS2-Net outperforms start-of-the-art methods on five text-video retrieval benchmarks, including MSR-VTT, VATEX, LSMDC, ActivityNet-Caption and DiDeMo.

Acknowledgement. This work was partially supported by National Key R&D Program of China (No. 2020AAA0108600) and National Natural Science Foundation of China (No. 62072462).

References

1. Abernethy, J., Lee, C., Tewari, A.: Perturbation techniques in online learning and optimization. Perturbations, Optimization, and Statistics p. 223 (2016)
2. 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)
3. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C.: Vivit: A video vision transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6836–6846 (2021)
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. arXiv preprint arXiv:2102.05095 (2021)
6. Berthet, Q., Blondel, M., Teboul, O., Cuturi, M., Vert, J.P., Bach, F.: Learning with differentiable pertubed optimizers. Advances in neural information processing systems pp. 9508–9519 (2020)
7. Bogolin, S.V., Croitoru, I., Jin, H., Liu, Y., Albanie, S.: Cross modal retrieval with querybank normalisation. arXiv preprint arXiv:2112.12777 (2021)
8. Bulat, A., Perez Rua, J.M., Sudhakaran, S., Martinez, B., Tzimiropoulos, G.: Space-time mixing attention for video transformer. Advances in Neural Information Processing Systems (2021)
9. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
10. 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)
11. Chen, S., Zhao, Y., Jin, Q., Wu, Q.: Fine-grained video-text retrieval with hierarchical graph reasoning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10638–10647 (2020)
12. Cheng, X., Lin, H., Wu, X., Yang, F., Shen, D.: Improving video-text retrieval by multi-stream corpus alignment and dual softmax loss. arXiv preprint arXiv:2109.04290 (2021)
13. Cordonnier, J.B., Mahendran, A., Dosovitskiy, A., Weissenborn, D., Uszkoreit, J., Unterthiner, T.: Differentiable patch selection for image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2351–2360 (2021)
14. Croitoru, I., Bogolin, S.V., Leordeanu, M., Jin, H., Zisserman, A., Albanie, S., Liu, Y.: Teachtext: Crossmodal generalized distillation for text-video retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11583–11593 (2021)
15. Dong, J., Li, X., Xu, C., Ji, S., He, Y., Yang, G., Wang, X.: Dual encoding for zero-example video retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9346–9355 (2019)
16. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. ICLR (2021)

17. Dzabraev, M., Kalashnikov, M., Komkov, S., Petiushko, A.: Mdmmt: Multidomain multimodal transformer for video retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3354–3363 (2021)
18. Fabian Caba Heilbron, Victor Escorcia, B.G., Niebles, J.C.: 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)
19. Fang, H., Xiong, P., Xu, L., Chen, Y.: Clip2video: Mastering video-text retrieval via image clip. arXiv preprint arXiv:2106.11097 (2021)
20. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6202–6211 (2019)
21. Gabeur, V., Sun, C., Alahari, K., Schmid, C.: Multi-modal transformer for video retrieval. In: European Conference on Computer Vision. pp. 214–229 (2020)
22. Gao, Z., Liu, J., Chen, S., Chang, D., Zhang, H., Yuan, J.: Clip2tv: An empirical study on transformer-based methods for video-text retrieval. arXiv preprint arXiv:2111.05610 (2021)
23. Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., He, K.: Accurate, large minibatch sgd: Training imagenet in 1 hour. arXiv preprint arXiv:1706.02677 (2017)
24. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization 3rd int. In: Conf. for Learning Representations, San (2014)
25. Krishna, R., Hata, K., Ren, F., Fei-Fei, L., Carlos Niebles, J.: Dense-captioning events in videos. In: Proceedings of the IEEE international conference on computer vision. pp. 706–715 (2017)
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, G., Duan, N., Fang, Y., Gong, M., Jiang, D.: Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 11336–11344. No. 07 (2020)
28. 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)
29. Lin, J., Gan, C., Han, S.: Tsm: Temporal shift module for efficient video understanding. In: Proceedings of the IEEE International Conference on Computer Vision (2019)
30. Liu, S., Fan, H., Qian, S., Chen, Y., Ding, W., Wang, Z.: Hit: Hierarchical transformer with momentum contrast for video-text retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11915–11925 (2021)
31. 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)
32. Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H.: Video swin transformer. arXiv preprint arXiv:2106.13230 (2021)
33. Luo, H., Ji, L., Shi, B., Huang, H., Duan, N., Li, T., Li, J., Bharti, T., Zhou, M.: Univl: A unified video and language pre-training model for multimodal understanding and generation. arXiv preprint arXiv:2002.06353 (2020)
34. Luo, H., Ji, L., Zhong, M., Chen, Y., Lei, W., Duan, N., Li, T.: Clip4clip: An empirical study of clip for end to end video clip retrieval. arXiv preprint arXiv:2104.08860 (2021)

35. 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)
36. 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)
37. 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 (2021)
38. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog p. 9 (2019)
39. Rao, Y., Zhao, W., Liu, B., Lu, J., Zhou, J., Hsieh, C.J.: Dynamicvit: Efficient vision transformers with dynamic token sparsification. Advances in neural information processing systems (2021)
40. Rohrbach, A., Torabi, A., Rohrbach, M., Tandon, N., Pal, C., Larochelle, H., Courville, A., Schiele, B.: Movie description. International Journal of Computer Vision pp. 94–120 (2017)
41. Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., Dai, J.: Vl-bert: Pre-training of generic visual-linguistic representations. ICLR (2020)
42. Wang, J., Yang, X., Li, H., Wu, Z., Jiang, Y.G.: Efficient video transformers with spatial-temporal token selection. arXiv preprint arXiv:2111.11591 (2021)
43. Wang, X., Zhu, L., Yang, Y.: T2vlad: global-local sequence alignment for text-video retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5079–5088 (2021)
44. Wang, X., Wu, J., Chen, J., Li, L., Wang, Y.F., Wang, W.Y.: Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4581–4591 (2019)
45. Xie, S., Sun, C., Huang, J., Tu, Z., Murphy, K.: Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In: Proceedings of the European conference on computer vision (ECCV). pp. 305–321 (2018)
46. 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)
47. 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)
48. Yu, Y., Kim, J., Kim, G.: A joint sequence fusion model for video question answering and retrieval. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 471–487 (2018)
49. Zhang, B., Hu, H., Sha, F.: Cross-modal and hierarchical modeling of video and text. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 374–390 (2018)
50. Zhang, H., Hao, Y., Ngo, C.W.: Token shift transformer for video classification. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 917–925 (2021)
51. Zhou, L., Palangi, H., Zhang, L., Hu, H., Corso, J., Gao, J.: Unified vision-language pre-training for image captioning and vqa. In: Proceedings of the AAAI Conference on Artificial Intelligence (2020)