SRTube: Video-Language Pre-Training with Action-Centric Video Tube Features and Semantic Role Labeling

Ju-Hee Lee and Je-Won Kang*
Dept. of Electronic and Electrical Engineering and Graduate Program in Smart Factory,
Ewha W. University, Seoul, South Korea
juhee69@ewhain.net, jewonk@ewha.ac.kr

Abstract

In recent years, large-scale video-language pre-training (VidLP) has received considerable attention for its effectiveness in relevant tasks. In this paper, we propose a novel action-centric VidLP framework that employs video tube features for temporal modeling and language features based on semantic role labeling (SRL). Our video encoder generates multiple tube features along object trajectories, identifying action-related regions within videos, to overcome the limitations of existing temporal attention mechanisms. Additionally, our text encoder incorporates high-level, action-related language knowledge, previously underutilized in current VidLP models. The SRL captures action-verbs and related semantics among objects in sentences and enhances the ability to perform instance-level text matching, thus enriching the cross-modal (CM) alignment process. We also introduce two novel pre-training objectives and a self-supervision strategy to produce a more faithful CM representation. Experimental results demonstrate that our method outperforms existing VidLP frameworks in various downstream tasks and datasets, establishing our model a baseline in the modern VidLP framework.

1. Introduction

Video-language pre-training (VidLP) models [7, 9, 16, 35, 43, 45] have been actively used for multi-modal learning. Their effectiveness is acknowledged across various downstream tasks, including video question answering [21, 25], video retrieval [17, 38], and video captioning [32, 47]. High-performance VidLP frameworks include key modules such as temporal modeling and cross-modal (CM) alignment [12, 15, 50, 54]. In recent studies, video transformers [34, 53] are used to establish baselines due to their ability to generate rich and joint representations of video and language [2, 7, 9, 16, 43]. In [12], a baseline VidLP architecture was built through the dissection and reassembly of the crucial components, resulting in superior performance. This approach seems to effectively construct the VidLP framework by empirically optimizing specific modules. However, when further expanding to a temporal direction, current video transformers [7, 9, 34] would deteriorate computational efficiency and overlook essential spatio-temporal contexts within videos due to the conventional patch-wise cross-attention. In [8, 27], visual features extracted from single or just few frames often served as substitutes for temporal features, since the straightforward cross-attention exhibited limited abilities to capture temporal contexts.

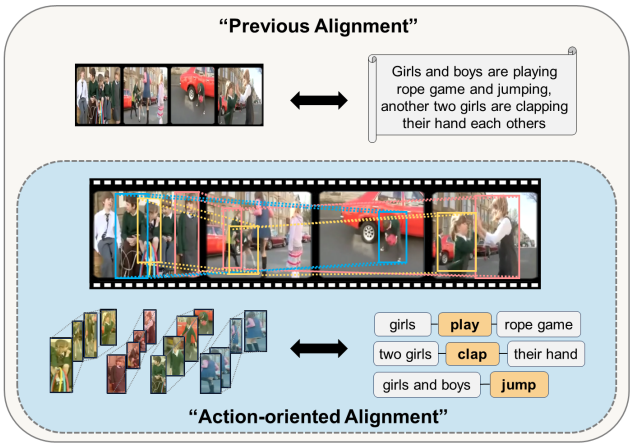


Figure 1. Our motivation is to bridge semantic gaps between video and text using more organized action-related features and their semantic alignments in CM.

To address this problem, recent VidLP studies have introduced object-centric video transformers [3, 20, 28, 48, 57] to manage more relevant objects in adjacent frames. Existing object-centric image-language pre-training approaches have used off-the-shelf object detection models to align objects in images with words [28, 31]. However, these schemes are infeasible to be applied to video, because they require pre-extracted features of all the objects. To mi-

*Je-Won Kang is a corresponding author.

grate this object-centric approach from image to video domain, in [44], an anchor frame was selected to extract visual features using an object-aware video transformer. Despite the reduced computational complexity, since an anchor frame is uniformly sampled, temporal contexts are not sufficiently reflected for CM alignments. Because videos exhibit dynamic changes of objects over time, accurate temporal modeling has been emphasized in the majority of studies [2, 3, 7, 34]. HiTeA [50] proposed a method to capture temporal contexts, by extracting features from long and short clips at various time steps. Although various features obtained from diverse video frames are reflected to the CM alignment, the plane features from the standard video transformer cannot encode the temporal semantics. The previous studies using visual tokens [3, 54], linguistic tokens [45], and region tokens [2] have relied on the query attention to capture objects but ignored semantic relationships in temporal, which remains further challenges in VidLP.

In this paper, we propose an action-centric VidLP framework using structured and action-related video and language features to enhance the CM alignment. Specifically, as shown in Fig. 1, we enrich our framework with a video tube feature for temporal modeling and a language feature grounded in semantic role labeling (SRL) to identify semantically relevant regions of videos, thereby facilitating a more robust CM alignment. Temporal tube along object trajectories offers action information as an inherent advantage for video representation. While the previous video transformers incorporating temporal modeling [3] calculated the attention only to correlated regions in video, our method tries to build the semantic gaps, by generating multiple tube features with precise coordinates of each object and considering their relations, which is a challenge with query-driven attention. Moreover, The proposed method also exploits high-level action-related knowledge from language, which was underutilized in existing VidLP models. The SRL aids in instance-level text matching, by exploring action-verbs and the related semantics among objects within sentences [42]. These enhanced features are combined with a CM fusion. Our study is the pioneering effort that integrates structural insights from video and linguistic features to synergize both, whereas few recent studies used temporal information from either video or language individually [44, 50].

Further, a novel pre-training objective and a self-supervision strategy are proposed to produce a faithful CM representation through the proposed features. It has been pointed out that the pre-training schemes combined with temporal modeling and CM fusion were crucial to improve the performance in recent VidLP studies [12, 17, 28, 44, 50]. However, the existing methods are hardly applied to the proposed video and language features for rich semantic alignments. Therefore, we introduce new action-centric proxy tasks, i.e., masked action modeling (MAM) and action numbering modeling (ANM). Through a new proxy task, we enable the effective integration of features with distinct temporal semantics.

Our work has several primary contributions as follows:

- We propose an action-centric VidLP framework (dubbed SRTube), explicitly combining structured video and linguistic features with CM semantic alignments. This pioneering approach utilizes video tube features and semantic phrase features for improved video representation and is the first to holistically exploit action-related knowledge from both video and language.
- We introduce a novel pre-training objective and a self-supervision strategy to craft a reliable CM representation using our proposed features. Our new proxy task facilitates an effective integration of the multi-modal features.
- We present experimental results on various datasets and downstream tasks to demonstrate the efficacy of our method. The superior performance justifies the effectiveness of our method, which can be considered as a new baseline for modern VidLP frameworks.

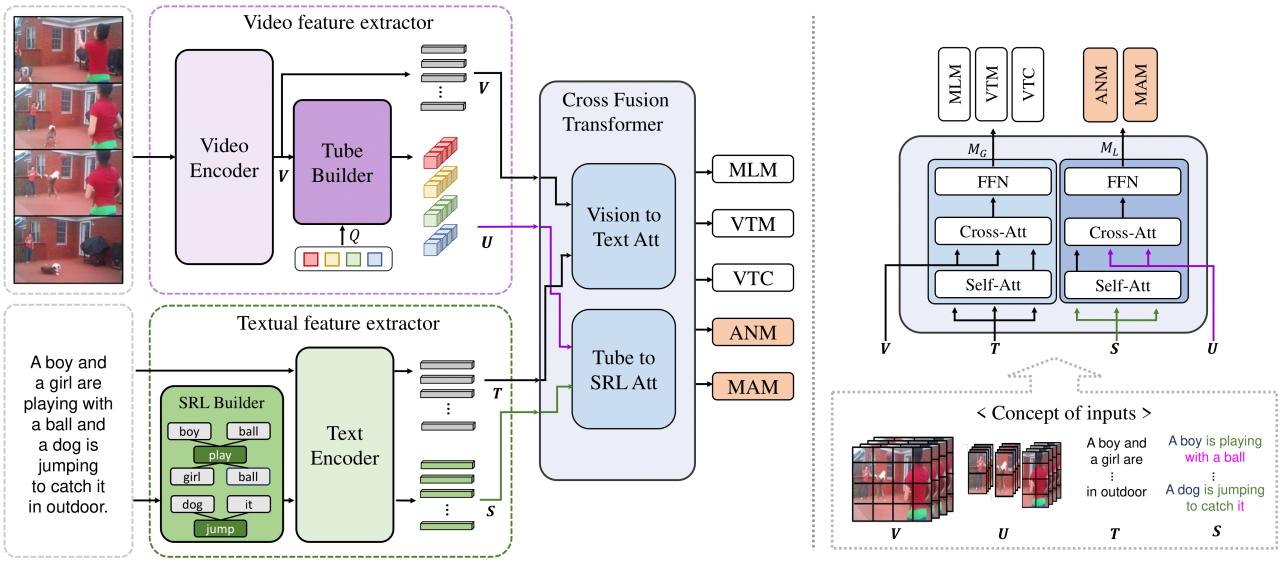
2. Related work

2.1. Video-language Pre-training

Owing to large-scale video text benchmarks [5, 37], pre-trained models have demonstrated impressive performance on diverse multi-modal downstream tasks [5, 16, 45, 49].

Current VidLP methods commonly used multi-modal fusion and temporal modeling to achieve highly improved performance [12, 15, 50, 54]. First, they tried to produce an effective CM representation with contextualized video and language features [10, 12, 17, 28, 50]. Existing CM alignment and fusion methods in VidLP can be broadly categorized into two groups. One strategy used a unified multi-modal fusion module to combine visual and language features, thus creating a CM representation [16, 43]. The other approach used separate visual and language encoders and leveraged a cross-attention mechanism of a transformer for inter-modal interactions [12, 17, 28].

Temporal modeling has been emphasized due to its impact [2, 4, 9, 10, 26, 50], as the current self-attention in transformers tended to exacerbate computational complexity and neglect essential temporal contexts within videos [8]. Video transformer architectures have adopted temporal attention mechanism [7, 9, 34]. TimeSformer [7] decoupled space and time attention to alleviate computational complexity. Also, X-Vit[9] proposed to integrate two lightweight global temporal attention mechanisms. Mformer-L [3] proposed a trajectory attention that tracked relevant objects between adjacent frames. [57] used pre-extracted object trajectory features, and [48] proposed a Patch-2-Word attention applied to related objects in video. However, those studies struggle with the lack of inductive



(a) Overall framework

(b) Cross Fusion Transformer

Figure 2. (a): An overall framework of SRTube is presented. The video feature extractor produces global visual features $\mathbf{V}$ and tube features $\mathbf{U}$. Simultaneously, the text extractor generates global text features $\mathbf{T}$ and semantic phrase features $\mathbf{S}$. $\mathbf{U}$ and $\mathbf{S}$ are combined to produce action-centric features via the cross fusion transformer (CFT) using two types of attention. (b): we show the detailed structure of the CFT, where CM alignments are performed using the global features and action-centric features through the proposed proxy tasks.

biases for videos, which is a challenge with query-driven attention. Our method tries to overcome the drawbacks, by focusing on more relevant spatio-temporal regions identified with tube features. Our tube builder is pre-trained for action recognition, thus providing well-characterized features for this purpose.

2.2. Linguistic Features in Multi-modal Analysis

While language features were often used as supplementary, they were actively applied to recent VidLP tasks [13, 31, 51, 54]. Certain words or phrases in a sentence hold rich semantic information. OSCAR [31] extracted text tags from images and used these tags as anchor points to align with words. In [39], adverbs were employed to enhance video understanding by predicting changes in actions through measuring textual relationships. [17] distinguished verbs and nouns using a part-of-speech (PoS) tag and constructed related questions. This approach enabled the model to synchronize local objects.

Scene graphs have been employed to extract structural knowledge from sentences [13, 51]. ERNIE-ViL [51] proposed a method to build semantic connections across vision and language using scene graphs. ROSITA [13] acquired object region features from images in scene graphs from text. However, while scene graphs concentrated on all words, they were hardly applied to motion representation and temporal modeling. Compared to scene graphs, SRL [42] in NLP was used to discern relationships between verbs and adjacent words in sentences, clarifying the role of each element. This enhanced sentence structure comprehension [40, 42] by designating a semantic role to each word. In our study, the texture features aligned with verbs are used to perform CM alignments with tube features in videos to create action-oriented features.

3. Proposed Method

3.1. Overall Architecture

Fig. 2 presents an overall network architecture of SRTube. Our model integrates action-related video and language features and their explicit semantic cues into the framework.

The video and text pairs are fed into a transformer-based video feature extractor and a text feature extractor. Our architecture employs the same video encoder and text encoder as VindLU [12], which are respectively expressed with light purple and green colors in Fig. 2, as a baseline. In the proposed architecture, a tube builder and an SRL builder as expressed as deep purple and green are incorporated into the feature extractors. On top of the baseline cross-fusion model [12], using visual to text attention (VTA), we further develop a tube to SRL attention (TSA) for the fusion. Those feature builders and TSA module are newly introduced in our architecture.

3.2. SRTube Feature Extraction

3.2.1 Video feature extractor

Visual features. Following VindLU [12], we employ BEiT [6] vision transformer and adapt it by adding a 2-layer temporal attention layer [7] before self-attention layer as our video encoder. Fig. 2 presents the generation of a sequence of visual context embedding, denoted as $\mathbf{V} = \{v_i\}_{i=0}^{L}, v_i \in \mathbb{R}^C$. L and C are a sequence length and an embedding dimension, which is set to 196 and 768, respectively. v_0 is assigned to v_{cls} as [CLS] token embedding.

Tube features. We use a tube feature, representing the trajectory of individual objects over time, to effectively capture the temporal dynamics of objects within the video. Fig. 2 (a) presents the proposed tube builder incorporated into the video feature extractor. The tube builder produces a set of tube features $\mathbf{U} \in \mathbb{R}^{N\times K\times D} = \{u_n^k\}$, in which N is the number of tube features in each frame, which also represents the maximum number of objects to be captured in the frame. D is a feature dimension, which is set to 256, and K is the number of frames. $\mathbf{U}$ is obtained with $\mathbf{V}$ and learned queries, $\mathbf{Q} \in \mathbb{R}^{N\times K\times C} = \{q_n^k\}$. q_n^k is the tube query corresponding to u_n^k in the k-th frame. q_n^k is learnable for a transformer to predict the class or location of objects by attention operations as in [11]. When $\{u_n^k\}$ is obtained from all the frames, the tube features go through a temporal pooling to generate a set of features $\{u_n\}_{n=1}^{N}$, which represent N object trajectories in a video sequence.

We explain a detailed architecture as shown in Fig. 3. Motivated by [11, 56], the tube builder comprises self-attention layers, cross-attention layers, and classification heads with feed-forward neural network (FFN). In our implementation, we develop a spatial-temporal self-attention (STSA), using the query. Specifically, we apply a spatial self-attention, in which q_n^k within the same frame goes through self-attention, and then calculate a temporal self-attention over time. We produce an output embedding z_n^k using cross-attention between $\mathbf{V}$ and the output of the STSA. z_n^k is used to predict a coordinate of a bounding box b_n^k. z_n^k and b_n^k are concatenated to produce u_n^k. We use a multi-layer perceptron (MLP) head, when projecting u_n^k into a cross-fusion transformer.

Tube builder training. We pretrain our tube builder using an action-recognition task that predicts both bounding boxes and their corresponding action labels as in [56] with AVA v2 datasets [18]. While the previous method [56] focused on only the recognition of foreground objects, our method is designed to include background labels to reflect overall temporal dynamics. These strategies lead to richer semantic expressions of both the static and dynamic video scenes. For this, we add several bounding boxes for backgrounds that are not overlapped with action regions and their binary results indicating if the regions are foreground or background to the training dataset. The number of labels is determined by the difference between N and the number of GT bounding boxes.

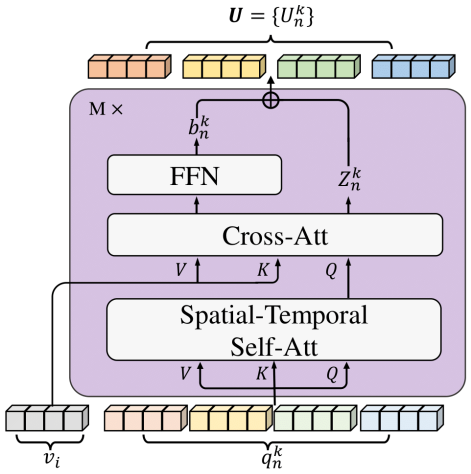


Figure 3. The architecture of the tube builder that produces a tube feature representing the trajectories of several objects in video.

3.2.2 Text feature extractor

Text features. As in the previous studies [5, 12, 16], an input sentence is tokenized and fed into a pre-trained encoder to obtain a text embedding sequence $\mathbf{T} = \{t_i\}_{i=0}^{L}, t_i \in \mathbb{R}^C$, where t_0 is assigned to t_{cls} as the [CLS] token embedding. L and C are the same as in the visual encoder. We adopt BERT [14] as the text encoder.

Semantic phrase features. $\mathbf{S} = \{s_i\}_{i=0}^{L}, s_i \in \mathbb{R}^C$ denotes a set of semantic phrase features generated from an SRL builder. s_0 is assigned to a [CLS] token. These features are used to facilitate the alignments with tube features. The SRL builder identifies verbs and verb-related phrases in a sentence and tags their semantic role labels as `VERB`. Then, the other noun phrases related to the verbs are labeled as `ARG`. `ARG` is assigned numbers such as `ARG0` and `ARG1` in the order of their importance to provide specific semantic roles. Specifically, `ARG0` and `ARG1` represent the agent of the verb and the corresponding object, respectively. Because `VERB` presents an action in a sentence, `ARG` expresses the corresponding actor in video. In this manner, semantic phrase are highly correlated with tube feature in video.

We explain our implementation in detail. We use a BERT-based model to predict semantic labels from training sentences [42]. Then, we remove noisy elements, e.g. `ARGM` label which including time and place, and aggregate relevant `ARG` and `VERB` elements to construct seman-

tic phrases based on predicted labels in the pre-processing. For instance, with the sentence "A girl passionately singing song at the stage", we extract "a girl singing song" as a semantic phrase consisting of ARG0, VERB, and ARG1. Additionally, to address instances involving non-VERB sentence, such as basic descriptions (e.g., "4k panorama video"), we adopted the word 'background' in place of a semantic phrase. More detailed examples of semantic phrase processing are described in the supplementary material.

3.3. Cross Fusion Transformer

Fig. 2 presents a cross fusion transformer (CFT) designed to achieve alignments between the proposed video and language features. The fusion model is composed of a series of self-attention, cross-attention, and FFNs. We use a visual-to-text attention (VTA) and a tube-to-SRL attention (TSA).

Visual-to-Text attention (VTA). As shown in Fig. 2 (b), we use a pair of $\mathbf{V}$ and $\mathbf{T}$ to generate a fused embedding sequence $\mathbf{M_G}$ through cross-attention. $\mathbf{M_G}$ is used to express global contexts, because $\mathbf{V}$ is pooled over an entire video sequence and $\mathbf{T}$ contains all the words in a sentence. The key $\mathcal{K}$, query $\mathcal{Q}$, and value $\mathcal{V}$ are determined as follows:

$$\mathbf{M_G} = MHA(\mathcal{Q} = \mathbf{V}, \mathcal{V} = \mathbf{T}, \mathcal{V} = \mathbf{T}), \quad (1)$$

where MHA is multi-head attention.

Tube-to-SRL attention (TSA). We propose TSA attention in CFT to facilitate action-oriented alignments. $\mathbf{M_L}$ is generated as a fused embedding sequence using a pair of $\mathbf{U}$ and $\mathbf{S}$ through TSA, as shown in Fig. 2. M_L represents specific semantic embedding as compared with M_G, since $\mathbf{U}$ and $\mathbf{S}$ are the action-related features. It is generated by the multi-head attention via a key, query, and value, as follows:

$$\mathbf{M_L} = MHA(\mathcal{Q} = \mathbf{U}, \mathcal{V} = \mathbf{S}, \mathcal{V} = \mathbf{S}). \quad (2)$$

3.4. Pre-training objective

We use five proxy tasks to train SRTube, including two novel proxy tasks that are introduced to incorporate action-oriented elements in the training. We also use three general proxy tasks of a masked language modeling (MLM), a video-text matching (VTM), and video-text contrastive (VTC). Despite improved performance by MLM, VTM, and VTC, they have posed challenges with dynamic temporal events in video. To address this issue, the proposed proxy tasks use the tube features and semantic phrases.

Masked action modeling (MAM). MAM uses a VERB token to enhance a semantic alignment between a set of tube features and semantic phrase features, since the token reflects a temporal attribution in a sentence. Specifically, we generate a masked semantic embedding S_m by masking some words, which tagged as VERB from semantic phrases. As following BERT [14], the masking is done by replacing the word with a special [MASK] token, as shown in Fig. 4.

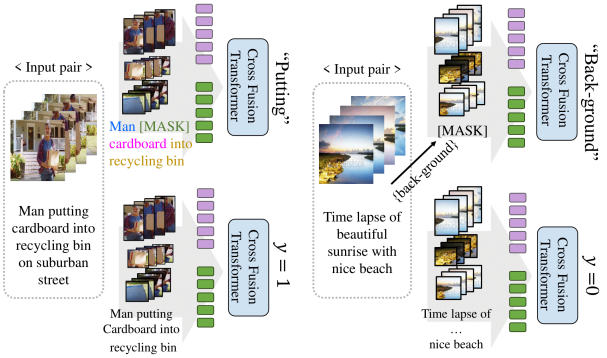


Figure 4. Illustrations of our proxy tasks MAM and ANM. In a sentence "Man putting cardboard into recycling bin on suburban street," MAM masks "putting" tagged as VERB, and ANM counts the verb semantic phrase to 1.

Our objective is to predict the masked text tokens (e.g. "putting" in Fig. 4), motivated by the MLM. U and S_m are provided to TSA in the CFT. The output embedding is fed into the prediction head, and, as a result, the predicted probability distribution of a masked token p_m is obtained. We apply cross-entropy loss to predict masked text tokens and the objective function is defined as,

$$\mathcal{L}_{MAM} = \mathbb{E}_{(U,S_m)\sim D} H(y_v | p_m(U, S_m)), \quad (3)$$

where H is cross-entropy loss and y_v is the masked verb. D is the distribution.

Action numbering modeling (ANM). We propose an ANM task as a self-supervised learning scheme to enhance an ability to discriminate actions from video. It is inspired by our observation that video-text pairs often exhibit more dynamic attributions, when the corresponding sentences contain more verbs. Our model counts the number of verbs and predicts the action-related attribution, as shown in Fig. 4, whereas previous methods relied on only simple linguistic analysis such as pos-tagging [13, 17].

We use the SRL builder to count the number, served as a pseudo label y_n for the task. Our model generates an embedding with the output [CLS] token of $\mathbf{M_L}$. We obtain a predicted probability distribution p_a by feeding $\mathbf{M_L}$ into a prediction network including fully connected and softmax layers. We predict y_n, ranging from 0 to 10, and utilize a mean squared error (MSE) loss function, as follows:

$$\mathcal{L}_{ANM} = MSE(y_n, p_a), \quad (4)$$

where MSE denotes a mean squared error.

The ANM enhances the ability to discriminate temporal attributions of video, including dynamic motions and static scenes (when y_n=0). ANM often acts as a regulator to avoid a motion bias. When relying on only the features pre-trained with action recognition datasets, the static

scenes would not be appropriately distinguished.

3.5. Total Loss Function in Training

The total loss L_{all} is defined as follow :

$$\mathcal{L}_{all} = \mathcal{L}_{MLM} + \mathcal{L}_{VTM} + \mathcal{L}_{VTC} + \mathcal{L}_{MAM} + \mathcal{L}_{ANM}, \quad (5)$$

where $\mathcal{L}_{MLM}$, $\mathcal{L}_{VTM}$ and $\mathcal{L}_{VTC}$ are the general losses in VidLP methods [12].

4. Experiment

4.1. Experimental Setting

Pre-training Datasets and Details. We use AVA v2 dataset [18] to train a tube builder. As in the recent works [5, 12, 28, 50], our model is trained on the large-scaled video-text pair datasets such as WebVid2M [5] and CC3m [41].

We train a tube builder and SRTube for 20 epochs on 24 NVIDIA V100 GPUs. We use AdamW [22] optimizer with an initial learning rate of 5×10^{-5}. We split a video sequence into uniform clips and sample 16 frames from randomly selected clips. We resize frame to 224×224.

Fine-tuning. For TR tasks, we fine-tune only the CFT with a VTC loss. For VQA task, we add an MLP to take an input [CLS] embedding for classification and optimize the model with cross-entropy loss. For VC task, to generate a sentence, the model predicts [MASK] until [END] token appears. All fine-tuning experiments are performed on the same GPU setting with pre-training.

Downstream Datasets. Our method has been extensively evaluated with various downstream tasks and datasets, including text-to-video retrieval (TR), video question and answering (VQA), video captioning (VC), and zero shot retrieval (ZR). We summarize the datasets as follows:

- **TR:** MSRVTT [47], DideMo [19], LSMDC [1], SSv2-Lable, SSv2-Template [27], ActivityNet Caption[23]
- **VQA:** MSRVTT-QA [47], TVQA [25], MSVD-QA [46].
- **VC:** MSRVTT [47], MSVD [46]
- **ZR:** MSRVTT [47], DideMo [19], LSMDC [1]

4.2. Experimental Results

We evaluate the performance of the proposed method in comparison to the state-of-the-art VidLP studies [12, 17, 44, 45, 50] in the following tasks.

Text-to-Video Retrieval. Table 1 summarizes the results in TR on MSR-VTT [47], LSMDC [1], and DiDeMo [19] benchmarks, including the results of the fine-tuning and zero-shot setting. We use R@1, R@5, and R@10 as the evaluation metrics. For clear comparisons, we present the number of pre-training data in the table.

On MSRVTT and LSMDC datasets, our method achieves the highest scores among the tested methods with all the metrics. In the table, we remark several methods [28, 44, 49] with $*$ and †, sharing similar motivations to ours. [28, 44] and [49] use pre-trained object detectors and trajectories, respectively. Our method improves the R@1 score by approximately 10.5% and 7.9% in comparison to [44] and [49]. Moreover, our model surpasses the performance of HiTeA [50] which utilizes features from diverse temporal duration, demonstrating the superiority of the proposed temporal modeling. Our results on DiDeMo dataset are second-ranked with R@5 and R@10 metrics, slightly worse than VindLU [12]. DiDeMo includes video sequences that exceed one minute, contain diverse scene changes. Our tube features assume continuous scenes to hold reliable actions, when sampling short-clips. Future research will consider adaptive frame selection to avoid noisy tube features and irrelevant clips.

We compare the results on the Something-to-Something dataset (SSv2) and ActivityNet caption dataset, including more motion dynamics, in Table 2 and 3. Our model improves 1.2% accuracy with R@1 over X-CLIP [36], which utilized a large pre-training dataset, and exhibits superior performance compared to the other recent models.

Zero-shot Retrieval. We show zero-shot retrieval results without fine-tuning to present the reliable performance of the proposed model. Table 1 presents the zero-shot performance of tested methods. Our model achieves the best scores in all the datasets and metrics. In comparison to a recent method [17], our approach provides an improved score approximately by 3.2%, 3.2%, and 5.2% on the MSR-VTT, LSMDC, and DiDeMo datasets, respectively.

Video Question and Answering. Table 4 presents the results on MSR-VTT, MSVD, and TVQA. Our method shows improved scores on MSR-VTT and MSVD. The TVQA dataset, consisting of drama content, differs from the MSVD dataset, by having more continuous scenes, which enhances the prediction of tube trajectories [55]. Due to the characteristic, our model surpasses [12] by a margin of 0.6. Our action-oriented approach leads to significant differences. Furthermore, as shown in Table 7, our method significantly improves the accuracy of "What type" questions, which typically involve action-related answer.

Video Captioning. We conduct fine-tuning and evaluation tests on VC tasks on MSR-VTT and MSVD datasets. The CIDEr metric is employed to perform an evaluation of captioning performance. Table 5 presents that SRTube significantly outperforms the state-of-the-art [28], which is instant-level matching method, by a 1.6% margin on MSR-VTT dataset. In MSVD dataset, our method is slightly worse than [30], using a larger size of PT pairs and second-ranked.

4.3. Ablation Study

Tube feature and SRL phrase. We demonstrate the effectiveness of our proposed visual and textual semantic feature

Method	Pre-training dataset	#PT pairs	MSR-VTT			LSMDC			DiDeMo		
			R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
UniVL [35]	H100M	136M	-	49.6	63.1	-	-	-	-	-	-
ClipBERT [26]	COCO, VG	0.2M	22.0	46.8	59.9	-	-	-	20.4	48.0	60.8
Frozen [5]	W2M, CC3M	5.5M	31.0	59.5	70.5	15.0	30.8	39.8	31.0	59.8	72.4
VIOLET [16]	YT, H100M	180M	34.5	63.0	73.4	16.1	36.6	41.2	32.6	62.8	74.7
OA-Trans* [44]	W2M, CC3M	5.5M	35.8	63.4	76.5	18.2	34.3	43.7	34.8	64.4	75.1
All-in-one [45]	H100M	138M	37.9	68.1	77.1	-	-	-	37.9	68.1	77.1
BridgeFormer [17]	W2M, CC3M	5.5M	37.6	64.8	751	17.9	35.4	44.5	37.0	62.2	73.9
TW-BERT†[49]	W2M, CC3M	5.5M	38.4	65.1	76.6	21.0	38.8	49.2	41.8	71.1	81.2
VindLU [12]	W2M, CC3M	5.5M	43.8	70.3	75.5	-	-	-	54.6	**82.3**	**88.2**
SRTube(ours)	W2M, CC3M	5.5M	**46.3**	**71.1**	**81.9**	**27.7**	**46.9**	**55.7**	**54.9**	82.1	87.2
Zero shot text to video retrieval											
Frozen [5]	W2M, CC3M	5.5M	18.7	39.6	51.6	9.3	22.0	30.1	21.1	46.0	56.2
VIOLET [16]	W2M, CC3M	5.5M	25.9	**49.5**	59.7	-	-	-	23.5	49.8	59.8
ALPRO* [28]	W2M, CC3M	5.5M	24.1	44.7	55.4	-	-	-	23.8	47.3	57.9
OA-Trans* [44]	W2M, CC3M	5.5M	23.4	47.5	55.6	-	-	-	-	-	-
BridgeFormer [17]	W2M, CC3M	5.5M	26.0	46.4	56.4	12.2	**25.9**	32.2	25.6	50.6	61.1
SRTube(ours)	W2M, CC3M	5.5M	**29.1**	49.4	**60.0**	**16.4**	25.8	**37.7**	**34.4**	**55.1**	**63.4**

Table 1. Performance comparison of our model and SOTA models in TR on MSR-VTT, LSMDC, and DiDeMo datasets under fine-tuning settings (top) and ZR (bottom). **#PT pairs** is the number of video-text pairs for pre-training. **H100M**: HowTo100M, **CC3M**: Conceptual Caption, **YT**: YT-Temporal [52], **W2M**: WebVid-2M, **COCO**: microsoft coco [33], **VG**: Visual Genome [24]

Method	SSv2-Lable			SSv2-Template		
	R@1	R@5	R@10	R@1	R@5	R@10
Frozen [5]	-	-	-	52.9	94.8	99.4
Singularity [27]	44.1	73.5	82.2	77.0	98.9	99.4
VindLU [12]	51.2	78.8	-	82.2	98.9	-
SRTube(ours)	**52.9**	**79.5**	**88.8**	**83.5**	**100.0**	**100.0**

Table 2. Performance comparison of TR on SSv2 dataset. For a fair comparison, we compare the methods pre-trained on W2M and CC3M. SSv2-Label task involves using ground truth sentences for video retrieval. SSv2-Template task employs object-masked sentences to retrieve videos [27].

Method	ActivityNet Caption		
	#PT Pairs	R@1	R@5
All-in-one [45]	138M	22.4	53.7
Singularity [27]	5.5M	43.0	70.6
TW-BERT [49]	5.5M	31.7	62.3
X-CLIP [36]	400M	44.3	**74.1**
HiTeA [50]	5.5M	45.1	73.5
SRTube(ours)	5.5M	**46.5**	**74.1**

Table 3. Performance comparison in TR on ActivityNet Caption.

in Table 6. For fair evaluations, we limit the training objectives to only MLM, VTC, and VTM. We assess the impact of each semantic feature in ZR task on the MSR-VTT dataset. Table 6 presents that **U** and **S** lead to performance increase of 1.6% and 1.4% with R@1 metric, respectively.

Methods	#PT pairs	MSR-VTT	MSVD	TVQA
HERO [29]	136M	-	45.9	74.2
MERLOT [52]	180M	43.1	-	78.7
ALPRO [28]	5.5M	42.1	45.9	-
Singularity [27]	17M	43.5	-	-
VindLU [12]	5.5M	43.6	-	79.0
SRTube(ours)	5.5M	**44.1**	**46.0**	**79.6**

Table 4. Performance comparison with existing methods in VQA.

Method	#PT pairs	MSRVTT	MSVD
UniVL [35]	180M	49.9	-
LAVENDER [30]	5.5M	58.0	**142.9**
STOA-VLP [57]	2.5M	60.2	131.8
SRTube(ours)	5.5M	**61.8**	142.5

Table 5. Performance comparison in VC with state-of-arts method on MSRVTT and MSVD with a CIDEr score.

With R@10, there is a slight loss with **S**, which could occur when the SRL over-recognizes verbs. However, this can be compensated for by the tube features and the proposed objectives. When combining all the features, the features produce 3.8%, 4.3%, and 2.4% improvements on the R@1, R@5, and R@10 metric at final.

Pre-training objectives. We conduct an ablation study of pre-training objectives in Table 7 in VQA task. All the pre-training objectives improve the accuracy on top of L_{Base}. Our analysis includes a detailed breakdown of accuracy in

Method	MSR-VTT		
	R@1	R@5	R@10
Baseline(**V**,**T**)	24.7	42.1	52.1
Baseline + Tube (**V**,**U**,**T**)	26.3	45.0	53.4
Baseline + SRL (**V**,**T**,**S**)	26.1	44.2	51.8
Baseline + Tube + SRL (**V**, **U**, **T**, **S**)	28.5	46.4	54.5

Table 6. Ablation tests of the tube and semantic phrase features.

Method	ALL (13,157)	What (8,149)	Who (4,552)	Others (456)
L_{Base}	44.7	37.6	54.8	73.2
$L_{Base} + L_{MAM}$	44.9	37.8	54.9	73.2
$L_{Base} + L_{ANM}$	45.4	38.5	55.5	73.9
$L_{Base} + L_{MAM} + L_{ANM}$	46.0	39.2	55.7	74.3

Table 7. Ablation tests on the proposed objectives in VQA on MSVD dataset. L_{Base} is sum of the MLM, VTM, and VTC losses.

# Frame	mAP.
4	24.1
8	25.4
16	**26.1**
32	24.9

# Tube queries	mAP.
4	23.0
8	**26.2**
12	24.4
16	24.5

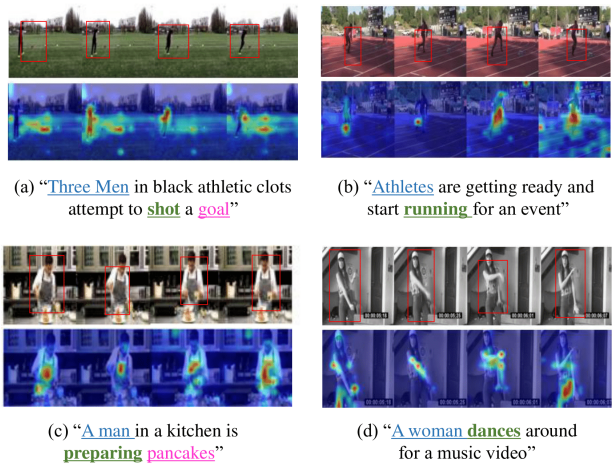
Table 8. Ablation study on frame counts and tube query numbers with the mAP values and AVA v2 dataset.

different question categories, "what", "who", and "other". To predict answers to "what" questions, the model has to focus on the action of objects, while "who" questions are related to appearance information. Our experimental results show that the integration of MAM and ANM lead to observable increase in performance: 1.6% in "what" questions. This result indicates that our model can catch the appearance of objects both action information in temporal domain.

Parameters in tube feature. We conduct an ablation study to examine the effect of the number of frames and queries on the performance of the tube builder. We fix either the number of frames or the number of queries during the ablation tests. Table 8 presents the impact on performance when the number of frames is varied with a constant query count as 8. The right section examines the effects of the number of tube queries with a fixed frame count as 16.

4.4. Qualitative Visualization

Fig. 5 presents qualitative examples of STRTube. We show visualization of a cross-attention map, using a sample from the MSR-VTT test set. In Fig. 5, we show examples of semantic phrases, the predicted bounding boxes, and the attention maps for words corresponding to VERB, highlighted in bold in the figure. For instance, in (a), the attention map



(a) "Three Men in black athletic clots attempt to **shot** a goal"

(b) "Athletes are getting ready and start **running** for an event"

(c) "A man in a kitchen is **preparing** pancakes"

(d) "A woman **dances** around for a music video"

Figure 5. Visualization of the attention map generated by the CFT, using samples from MSR-VTT dataset. Our model attends to the patches associated with moving objects by tracking the action trajectory. SRL label is expressed with blue, pink for ARG, and green for VERB.

for the VERB "shot" focuses on objects related arguments. This indicates that our method understands the motion of moving objects, while generating video tube features. Furthermore, in (c) and (d), where only the subject of the action is present, our method more distinctly focuses on the features necessary to differentiate the actions.

5. Conclusion

We proposed a new VidLP framework that used video tube features for temporal analysis and SRL for language processing, aiming to identify action-related regions in videos more effectively. The video encoder generated multiple tube features along object trajectories to identify action-related regions within videos, addressing limitations of existing attention mechanisms. Moreover, the text encoder incorporated high-level, action-related language knowledge using SRL, enhancing instance-level text matching and the CM alignment. This approach, enhanced by two novel pre-training objectives and a self-supervision strategy, outperformed existing VidLP models in various tasks and datasets, establishing a new baseline in the VidLP domain.

Acknowledgment

This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2021-0-02068, Artificial Intelligence Innovation Hub) and was partly supported by the NRF grant funded by MSIT (No.NRF-2022R1A2C4002052).

References

[1] Rohrbach Anna, Rohrbach Marcus, Tandon Niket, and Schiele Bernt. A dataset for movie description. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3202–3212, 2015.

[2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Int. Conf. Comput. Vis.*, pages 6836–6846, 2021.

[3] Y. Asano, D. Campbell, C. Feichtenhofer, J. Henriques, F. Metze, I. Misra, M. Patrick, A. Vedaldi, et al. Keeping your eye on the ball: Trajectory attention in video transformers. *Adv. Neural Inform. Process. Syst.*, 34:12493–12506, 2021.

[4] Piyush Bagad, Makarand Tapaswi, and Snoek Cees GM. Test of time: Instilling video-language models with a sense of time. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2503–2516, 2023.

[5] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Int. Conf. Comput. Vis.*, pages 1728–1738, 2021.

[6] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

[7] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, page 4, 2021.

[8] Shyamal Buch, Cristóbal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Niebles Juan Carlos. Revisiting the" video" in video-language understanding. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2917–2927, 2022.

[9] Adrian Bulat, Perez Rua Juan Manuel, Swathikiran Sudhakaran, Brais Martinez, and Georgios Tzimiropoulos. Space-time mixing attention for video transformer. *Adv. Neural Inform. Process. Syst.*, 34:19594–19607, 2021.

[10] Meng Cao, Tianyu Yang, Junwu Weng, Can Zhang, Jue Wang, and Yuexian Zou. Locvtp: Video-text pre-training for temporal localization. In *Eur. Conf. Comput. Vis.*, pages 38–56. Springer, 2022.

[11] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Eur. Conf. Comput. Vis.*, pages 213–229. Springer, 2020.

[12] Feng Cheng, Xizi Wang, Jie Lei, David Crandall, Mohit Bansal, and Gedas Bertasius. Vindlu: A recipe for effective video-and-language pretraining. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10739–10750, 2023.

[13] Yuhao Cui, Zhou Yu, Chunqi Wang, Zhongzhou Zhao, Ji Zhang, Meng Wang, and Jun Yu. Rosita: Enhancing vision-and-language semantic alignments via cross-and intra-modal knowledge integration. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 797–806, 2021.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[15] Z. Dou, Y. Xu, Z. Gan, J. Wang, S. Wang, L. Wang, C. Zhu, P. Zhang, L. Yuan, and N. Peng. An empirical study of training end-to-end vision-and-language transformers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 18166–18176, 2022.

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

[17] Yuying Ge, Yixiao Ge, Xihui Liu, Dian Li, Ying Shan, Xiaohu Qie, and Ping Luo. Bridging video-text retrieval with multiple choice questions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16167–16176, 2022.

[18] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6047–6056, 2018.

[19] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Int. Conf. Comput. Vis.*, pages 5803–5812, 2017.

[20] Roei Herzig, Elad Ben-Avraham, Karttikeya Mangalam, Amir Bar, Gal Chechik, Anna Rohrbach, Trevor Darrell, and Amir Globerson. Object-region video transformers. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3148–3159, 2022.

[21] Nayoung Kim, Seong Jong Ha, and Je-Won Kang. Video question answering using language-guided deep compressed-domain video feature. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1708–1717, 2021.

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[23] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Niebles Carlos Juan. Dense-captioning events in videos. In *Int. Conf. Comput. Vis.*, pages 706–715, 2017.

[24] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.

[25] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*, 2018.

[26] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Berg Tamara L, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7331–7341, 2021.

[27] Jie Lei, Tamara L Berg, and Mohit Bansal. Revealing single frame bias for video-and-language learning. *arXiv preprint arXiv:2206.03428*, 2022.

[28] Dongxu Li, Junnan Li, Hongdong Li, Juan Carlos Niebles, and Steven CH Hoi. Align and prompt: Video-and-language pre-training with entity prompts. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4953–4963, 2022.

[29] Linjieb 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.

[30] Linjie Li, Zhe Gan, Kevin Lin, Chung-Ching Lin, Zicheng Liu, Ce Liu, and Lijuan Wang. Lavender: Unifying video-language understanding as masked language modeling. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 23119–23129, 2023.

[31] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Eur. Conf. Comput. Vis.*, pages 121–137. Springer, 2020.

[32] 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.

[33] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[34] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3202–3211, 2022.

[35] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*, 2020.

[36] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 638–647, 2022.

[37] 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. In *Int. Conf. Comput. Vis.*, pages 2630–2640, 2019.

[38] Niluthpol Chowdhury Mithun, Juncheng Li, Florian Metze, and Amit K Roy-Chowdhury. Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, pages 19–27, 2018.

[39] Davide Moltisanti, Frank Keller, Hakan Bilen, and Laura Sevilla-Lara. Learning action changes by measuring verb-adverb textual relationships. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 23110–23118, 2023.

[40] Arka Sadhu, Kan Chen, and Ram Nevatia. Video object grounding using semantic roles in language description. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10417–10427, 2020.

[41] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018.

[42] Peng Shi and Jimmy Lin. Simple bert models for relation extraction and semantic role labeling. *arXiv preprint arXiv:1904.05255*, 2019.

[43] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Int. Conf. Comput. Vis.*, pages 7464–7473, 2019.

[44] Jinpeng Wang, Yixiao Ge, Guanyu Cai, Rui Yan, Xudong Lin, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Object-aware video-language pre-training for retrieval. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3313–3322, 2022.

[45] Jinpeng Wang, Yixiao Ge, Rui Yan, Yuying Ge, Lin Kevin Qinghong, Satoshi Tsutsui, Xudong Lin, Guanyu Cai, Jianping Wu, Ying Shan, et al. All in one: Exploring unified video-language pre-training. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6598–6608, 2023.

[46] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653, 2017.

[47] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5288–5296, 2016.

[48] Xu Yang, Zhangzikang Li, Haiyang Xu, Hanwang Zhang, Qinghao Ye, Chenliang Li, Ming Yan, Yu Zhang, Fei Huang, and Songfang Huang. Learning trajectory-word alignments for video-language tasks. *arXiv preprint arXiv:2301.01953*, 2023.

[49] Xu Yang, Zhangzikang Li, Haiyang Xu, Hanwang Zhang, Qinghao Ye, Chenliang Li, Ming Yan, Yu Zhang, Fei Huang, and Songfang Huang. Learning trajectory-word alignments for video-language tasks. *arXiv preprint arXiv:2301.01953*, 2023.

[50] Qinghao Ye, Guohai Xu, Ming Yan, Haiyang Xu, Qi Qian, Ji Zhang, and Fei Huang. Hitea: Hierarchical temporal-aware video-language pre-training. *arXiv preprint arXiv:2212.14546*, 2022.

[51] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. In *AAAI*, pages 3208–3216, 2021.

[52] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. *Adv. Neural Inform. Process. Syst.*, 34:23634–23651, 2021.

[53] H. Zhang, J. Duan, M. Xue, J. Song, L. Sun, and M. Song. Bootstrapping vits: Towards liberating vision transformers from pre-training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8944–8953, 2022.

[54] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5579–5588, 2021.

[55] Shun Zhang, Jia-Bin Huang, Jongwoo Lim, Yihong Gong, Jinjun Wang, Narendra Ahuja, and Ming-Hsuan Yang. Tracking persons-of-interest via unsupervised representation adaptation. *International Journal of Computer Vision*, 128: 96–120, 2020.

[56] Jiaojiao Zhao, Yanyi Zhang, Xinyu Li, Hao Chen, Bing Shuai, Mingze Xu, Chunhui Liu, Kaustav Kundu, Yuanjun Xiong, Davide Modolo, et al. Tuber: Tubelet transformer for video action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13598–13607, 2022.

[57] Weihong Zhong, Mao Zheng, Duyu Tang, Xuan Luo, Heng Gong, Xiaocheng Feng, and Bing Qin. Stoa-vlp: Spatial-temporal modeling of object and action for video-language pre-training. *arXiv preprint arXiv:2302.09736*, 2023.