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

Supplementary Material

1. Implementation Details

1.1. Model Architecture.

Video and Text Encoders. In our SRTube, we adopt the BEiT [2] architecture as our video encoder which comprises 12 layers with 768 hidden units. Following VindLU [5], two temporal attention layers [3] are added before self-attention layers in a video encoder. Initially, the video encoder was trained for action recognition and trajectory prediction using the AVA [8] dataset. In this stage, visual encoder is initialized using the BEiT [2] model, which is pre-trained on ImageNet[15]. The weights obtained from this stage are then used to further initialize the visual encoder in the SRTube pre-training stage to enhance its capabilities for advanced action-centric video analysis. The Tube builder consists of a 6-layer transformer that follows the DETR [4] decoder architecture and is initialized with DETR. For text encoding, we use the BERT [6] architecture and the cross fusion encoder consists of the last three layers of the BERT model, as suggested in [5].

1.2. Downstream Implementation Details

Text to video retrieval. For retrieval tasks, we jointly optimize the VTC loss for video-text alignment during fine-tuning. During inference, we select top-k candidates by computing the dot-product similarity between the output embedding of VTA and TSA. Then TSA scores are added to VTA scores and then rerank the selected candidates based on their total scores.

Video QA. For video question and answering tasks, we jointly optimize the VTC, VTM, MAM, and ANM loss for fine-tuning. he text decoder receives input from the output embeddings of both VTA and TSA to generate answers. During inference, a high-confidence answer is selected from the output of the text decoder.

Video Captioning. We concatenate the video context feature with the video tube feature and directly input this combination into a text decoder for caption generation. The caption is generated from a [MASK] embedding and continues until an [END] token is reached. To optimize the model, we utilize language modeling loss. The hyperparameters used for fine-tuning are detailed in the Table 3.

1.3. Datasest descriptions.

In this section, we describe all of the video-text pairs used for evaluation. The details of the datasets are described in table 4.

2. Additional Experimental Results.

In this section, we present additional experimental results for text-to-video retrieval, video question and answering for additional ablation studies.

Tube feature and SRL phrase. We present the effectiveness of tube features and SRL features in a fine-tuned model for ablation studies in Table 1. We present experimental results on text-to-video retrieval on MSR-VTT and LSMDC datasets, achieving R@1 metrics. These experiments are conducted using the same settings as in the main paper.

Method	MSR-VTT	LSMDC
	R@1	R@1
Baseline(**V**,**T**)	40.6	21.3
Baseline + Tube (**V**,**U**,**T**)	43.1	24.5
Baseline + SRL (**V**,**T**,**S**)	42.7	24.3
Baseline + Tube + SRL (**V**, **U**, **T**, **S**)	43.9	25.7

Table 1. Ablation tests of the tube and semantic phrase features.

Video encoder. We compare the performance of models with different video encoders, including both traditional (C3D [16]) and conventional method (VideoSwin [14], BEiT [2]) in Table 2. We compare results using the TVQA dataset for visual question answering. The results demonstrate that our model, initialized with weights derived from the tube builder training stage, consistently outperforms alternative approaches.

Method	TVQA
C3D [16]	75.1
VideoSwin [14]	76.7
BEiT [2]	78.2
Ours	78.6

Table 2. Ablation tests of video encoder initialization.

3. Semantic phrase

In Fig. 1, We show the results of applying semantic role labeling to the original video descriptions to get semantic phrases. Each phrase corresponds to a specific semantic label highlighted in different colors. The final extracted semantic phrases are connected using "and" to generate our proposed semantic phrase which is shown as "filtered description" in Fig.1.

CVPR #06358

Datasets	Optimize & Fine-tuning Configs	# Query	Epoch	BS $\times$ GPUs
MSR-VTT-Ret		8	20	24 X 2
MSR-VTT-QA		8	20	24 X 2
MSR-VTT-Cap	optimizer : AdamW [10]	8	20	24 X 2
LSMDC-Ret	weight decay : 0.02	8	25	24 X 2
DiDeMo-Ret	learning rate scheduler : Cosine Decay	8	20	24 X 2
MSVD-QA	learning rate : 10^{-5}	8	15	24 X 2
MSVD-Cap	frame resolution : 224 X 224	8	15	24 X 2
TVQA-QA	number of frame : 16	8	20	24 X 2
Activity net-Ret		8	20	48 X 1
SSv2-Template-Ret		4	20	48 X 1
SSv2-Lable-Ret		4	20	48 X 1

Table 3. Fine-tuning configurations for video-language downstream tasks.**BS**: batch size, **Ret** : retrieval task, **QA** : Question and answering task, **Cap** : captioning task. # denotes the number of.

Dataset	Source	# Clip	# Sentence
MSR-VTT [17]	YouTube	10K	200K
MSVD [17]	YouTube	2K	10K
LSMDC [1]	Movie	118K	118K
DiDeMo [9]	Flickr	10K	40K
TVQA [12]	TV show	22K	153K
SSv2-Template [13]	SSv2 [7]	220K	174
SSv2-Label [13]	SSv2	220K	174
ActivityNet-Captions [11]	YouTube	20K	100K

Table 4. Comparison of video-language benchmarks. We describe each benchmark in terms of the source of the original videos, as well as the number of clips and text. # denotes the number of.

4. Visualization.

To illustrate the functionality of our proposed model, we display examples of predicted object trajectories generated by the tube builder, alongside results from various downstream tasks. We indicate the predicted object trajectory with red bounding boxes in the image. For example, in Fig. 2, we show input video and predicted trajectories of an object and video captioning results compared with ground truth. In cases where videos lack target objects, our model adeptly captures the background region as a video feature, as also depicted in Fig. 2. Furthermore, we present inference results for Visual Question Answering (VQA) in Figs. 3 and 4. In addition to successful outcomes, we analyze error cases in trajectory prediction. We note that the model tends to struggle with complex video inputs, such as those involving scene changes or frames of poor quality, often leading to incorrect trajectory predictions. Lastly, we demonstrate zero-shot retrieval results for MSR-VTT video data through a demo video, with a sample showcased in Fig. 5. The demo video linked is below:

https://youtu.be/g0RkfWlMhcI

Video description	Man putting cardboard into recycling bin on suburban street	
	Semantic Role Labeling	Semantic Groups
	[ARG0: Man] [V: putting] [ARG1: cardboard] [ARG2: into recycling bin on suburban street]	Man putting cardboard into recycling bin on suburban street recycling bin
Filtered description	Man putting cardboard into recycling bin on suburban street recycling bin	

Video description	Tired bearded man in a casual look sits on a couch, uses remote control, switches channels, nothing interests him, stands up and goes away. cozy living-room stuff on the background. having break	
	Semantic Role Labeling	Semantic Groups
	[ARG0: Tired bearded man in a casual look] [V: sits] [ARG1: on a couch] ,uses remote control , switches channels , nothing interests him , stands up and goes away . cozy living - room stuff on the background .	Tired bearded man in a casual look sits on a couch
	[ARG0: Tired bearded man in a casual look] sits on a couch , [V: uses] [ARG1: remote control] , switches channels , nothing interests him , stands up and goes away . cozy living-room stuff on the background. having break	Tired bearded man in a casual look uses remote control ⋮
Filtered description	Tired bearded man in a casual look sits on a couch and Tired bearded man in a casual look uses remote control ⋮ and Tired bearded man in a casual look having break	

Figure 1. Example of SRL pre-processing on WebVid-2M datasets.

CVPR
#06358

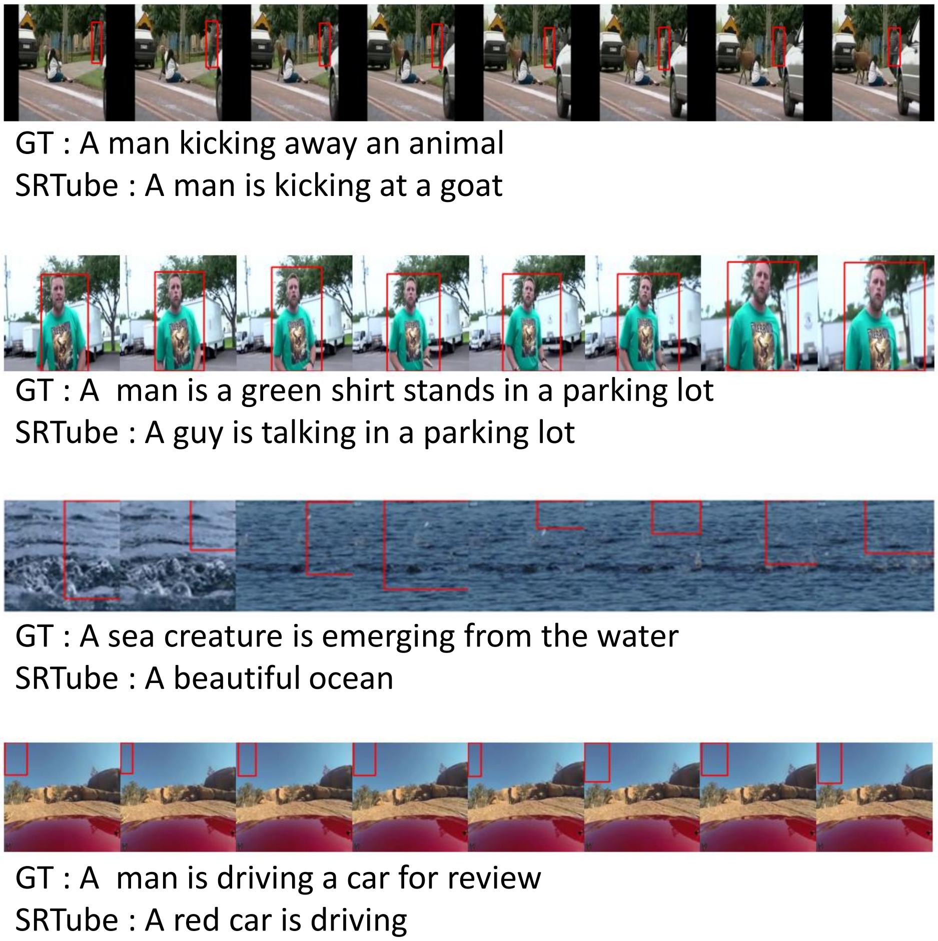

Figure 2. Visualization of the MSR-VTT video captioning results. We also show tube trajectory with red bounding box for i_th tube query.

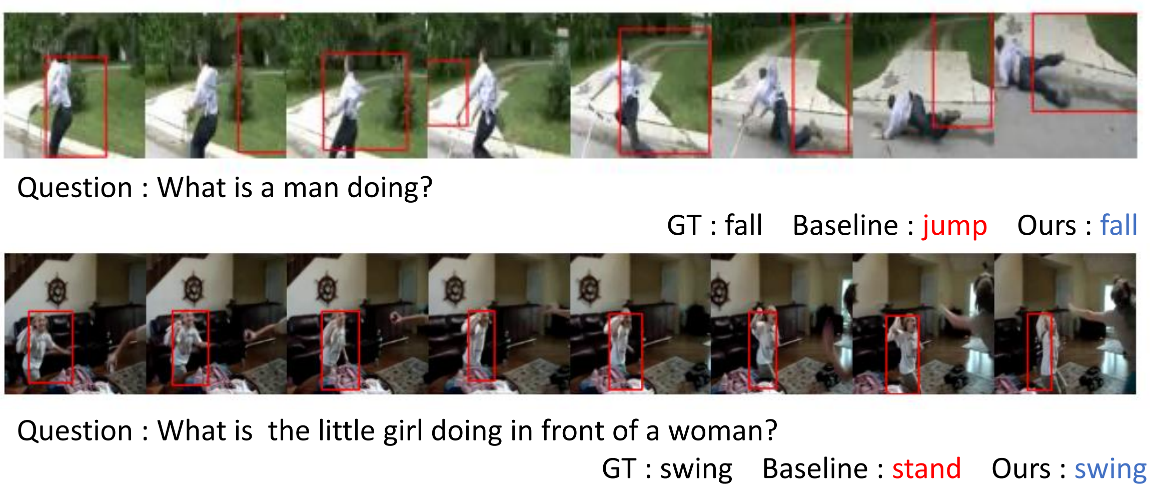

Figure 3. Visualization of the VQA results on MSVD. We show tube trajectory with red bounding boxes for tube query. The baseline is the same setting as the ablation study on features.

CVPR
#06358

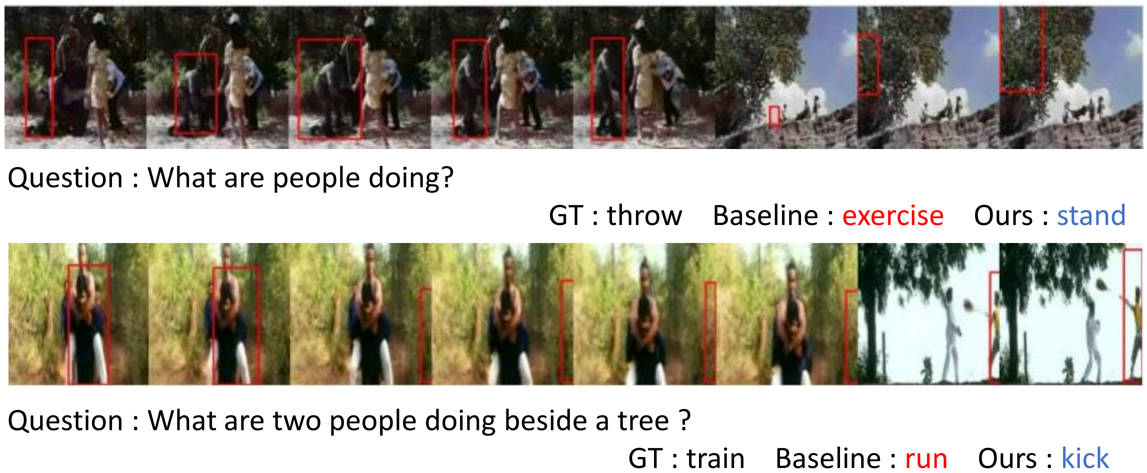

Figure 4. Error cases of VQA on MSVD dataset. We show tube trajectory with red bounding boxes for tube query. The baseline is the same setting as the ablation study on features.

Figure 5. Zero-shot text-to-video retrieval results on MSR-VTT. We type text and find most related video. We present demo video and this is sample of demo.

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

[2] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 1

[3] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, page 4, 2021. 1

[4] 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. 1

[5] 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. 1

[6] 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. 1

[7] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska andSusanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017. 2

[8] 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. 1

[9] 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. 2

[10] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 2

[11] 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. 2

[12] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*, 2018. 2

[13] Jie Lei, Tamara L Berg, and Mohit Bansal. Revealing single frame bias for video-and-language learning. *arXiv preprint arXiv:2206.03428*, 2022. 2

[14] 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. 1

[15] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 1

[16] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Int. Conf. Comput. Vis.*, pages 4489–4497, 2015. 1

[17] 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. 2