Supplementary Material for
SWINBERT: End-to-End Transformers with Sparse Attention for Video Captioning

Kevin Lin*, Linjie Li*, Chung-Ching Lin*, Faisal Ahmed, Zhe Gan, Zicheng Liu, Yumao Lu, Lijuan Wang
Microsoft
{keli, lindsey.li, chungching.lin, fiahmed, zhe.gan, zliu, yumaolu, lijuanw}@microsoft.com

1. Analysis of Different Video Backbones

Table 1 shows our proposed method is generalizable to different video backbones. Note that the SOTA methods are trained with pre-extracted 2D and 3D CNN features. Our end-to-end trained model with only video backbone (TimeSformer [3] or Video Swin Transformer [8]), can often outperform the recent SOTA. Adding sparse attention mask consistently improves model performance across the video backbones considered. Further, a stronger backbone yields better captioning performance.

It is worth noting that, TimeSformer generates longer video tokens compared to that of Video Swin Transformer (VidSwin). This introduces extra memory cost for the language model (due to quadratic complexity), making TimeSformer difficult to scale to longer sequences. Due to GPU memory constraints, during rebuttal period, we can only train TimeSfomer on 8 frames per clip. From another perspective, this shows VidSwin offers a favorable memory-accuracy trade-off for video captioning.

2. Influence of Pre-Training on Backbone

The top rows of Table 2 give a fair comparison where both approaches use the same SlowFast [5] as the backbone. Our method achieves better performance than VALUE [7].

The bottom rows of Table 2 show the best results obtained by the two methods with different pre-training datasets. VALUE uses both CLIP-ViT [11] and SlowFast [5] as backbones, which are pre-trained on 400M image-text pairs [11] and Kinetics-400 (K400) [6]. In contrast, our video backbone is pre-trained on ImageNet [13] and K400/600. Although our video backbone uses less pre-training data than VALUE, we achieve better caption performance. We show that end-to-end training (from video patches to textual outputs) is crucial to the performance of video captioning. Compared with K400, pre-training backbone with K600 slightly improves CIDEr.

* Equal contribution.

3. Choice of Hyperparameter λ

Since we use a regularization hyperparameter λ in our sparsity constraint (see Eq. 1 in our main manuscript), we provide further experiments with difference choices of λ. Table 3 shows that our model gives consistent improvements over different choices of λ.

4. Additional Qualitative Results

We present additional qualitative results in Figure 1, 2, 3, and 4. For each video, we show our prediction and the corresponding ground-truth captions.

In Figure 1, SWINBERT generates semantically correct captions for the considered cooking videos. For example, as presented in the top row, our model predicts "*Place the basil on the pizza,*" while the ground truth is "*Place basil leaves on top of the pizza*." Although the word sequences are not exactly the same, both can be considered semantically correct with respect to the given video.

Figure 2 shows our qualitative results on MSRVTT. We observe that SWINBERT works well for open-domain videos. For example, our model is capable of recognizing different actions, such as *giving a speech*, *applying makeup*, and *playing golf*. In addition, some of our predictions are similar to the ground truths, as presented in the second, third, and fourth rows.

In Figure 3, we show our results on VATEX, where the ground-truth sentences are more descriptive and challenging. SWINBERT recognizes fine-grained objects (*e.g.*, drum set, paper airplane, high chair, and curling iron) in various viewpoints, and generates semantically reasonable captions for the input videos.

Figure 4 shows the results on MSVD. SWINBERT recognizes the video events correctly. As presented in the first row, SWINBERT recognizes "*A woman is dancing on a stage*" by seeing detailed movements of the posture in multiple frames. In the second row, SWINBERT correctly describes "*A man is playing a flute*."

Backbone	#frames (#tokens)	Attn. Mask	MSRVTT	MSVD	VATEX
SOTA	-	-	52.9 [16]	95.2 [17]	58.1 [7]
TimeSformer	8 (1568)	Full	49.9	123.4	57.9
TimeSformer	8 (1568)	Sparse	51.9	127.6	63.0
VidSwin	32 (784)	Full	52.3	127.9	71.1
VidSwin	32 (784)	Sparse	55.1	147.6	71.6

Table 1. Analysis of our method with different video backbones. All backbones are pretrained on Kinetics-600 [4]. We report CIDEr score [14] in this analysis.

Method	Backbone	Pretraining data for backbone	CIDEr
VALUE [7]	SlowFast	K400	51.2
Ours	SlowFast	K400	53.6
VALUE [7]	CLIP-ViT + SlowFast	400M image-text pairs + K400	58.1
Ours	VidSwin	ImageNet + K400	68.1
Ours	VidSwin	ImageNet + K600	71.1

Table 2. Breakdown of pre-training data, evaluated on VATEX.

5. Additional Training Details

We implement our models based on PyTorch [10]. We also adopt mixed-precision training. To be specific, we use DeepSpeed [12] for the majority of our experiments. Additionally, we use Nvidia Apex [2] for the experiments of longer video sequences, which empirically leads to more stable training. All experiments are conducted on Microsoft Azure [1] with multiple Nvidia V100 GPUs (32GB).

Our Video Swin Transformer (VidSwin) is a `Swin-base` model initialized with Kinetics-600 pre-trained weights [8]. Our multimodal transformer has 12 layers, and the hidden size is 512. Our multimodal transformer is randomly initialized. Both VidSwin and the multimodal transformer are trained in an end-to-end manner.

We resize the shorter side of all the video frames to 224. During training, we random crop (224×224) at the same location for all the frames in a given video. During inference, we center crop (224×224) for all the frames.

Since the considered datasets have different data scales and domains, we use task-specific training epochs and learning rates based on the performance of validation sets.

6. Broader Impact and Ethical Concerns

Video captioning offers the possibility to make videos more accessible and inclusive to all users, including low-vision and blind users [9]. In this paper, we aim to improve the accuracy of video captioning with better video representations. While our method outperforms the previous state-of-the-arts, the model does not always guarantee a perfect prediction. As a data-driven system, our model is sensitive to the distribution of training data, therefore may fail when encountering videos in the wild. To avoid any undesirable predictions that could lead to ethical concerns in real-world applications (*e.g.*, incorrect semantics, wrong identity), the generated caption should be considered as a draft that requires further editing.

7. Discussion

Computational Cost: In this work, we primarily focus on improving caption accuracy (CIDEr score), and the sparse attention mask is used as a regularizer for improving training. Since we implement the sparse attention mask via an additional learnable embedding, it does not have a real speed-up. In the future, we plan to investigate CUDA implementations to construct a binary attention mask to reduce computational cost. In our current implementation, our model is computational memory intensive since both VidSwin and BERT require sufficient GPU memory during training. We use mix-precision and checkpointing to remedy the memory issues.

How many frames are sufficient for video captioning: Our experimental results in Table 4(a) of the main text suggest that more frames would benefit captioning performance. However, due to GPU memory constraints, with 128-frame inputs, we are restricted to use batch size=1, making the training inefficient. Hence, we can only empirically conclude that 64 frames give the best performance. Please note that 128-frame is a significant departure from current SOTA, which are typically 8-32 frames.

Token selection: Recently, researchers [15] are exploring

	0	0.1	0.5	1	2	5	10
MSRVTT (32frm)	52.3	53.4	53.8	53.9	54.9	**55.1**	53.4

Table 3. Our model (with sparse attention) gives consistent improvements over baseline (without sparse attention) across different choices of λ.

dynamic token selection to reduce the computation complexity of the transformer. While dynamic token selection is useful for vision or NLP transformers, it needs to be studied further when integrated with multimodal transformers for video captioning. Unlike previous efforts that attempted to reduce the number of tokens, we keep video tokens intact and improve caption accuracy by regularizing attention over time.

Observation in VATEX and MSVD: We observe the two datasets have different characteristics. The groundtruth captions in VATEX include detailed actions, and the caption model requires more temporal features to have a correct generation. For MSVD, the groundtruth captions are more about the scenes and objects, and thus spatial features play a critical role to captioning.

References

[1] Microsoft Azure. https://azure.microsoft.com/. 2

[2] NVIDIA Apex. https://github.com/NVIDIA/apex. 2

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

[4] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018. 2

[5] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, 2019. 1

[6] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 1

[7] Linjie Li, Jie Lei, Zhe Gan, Licheng Yu, Yen-Chun Chen, Rohit Pillai, Yu Cheng, Luowei Zhou, Xin Eric Wang, William Yang Wang, et al. Value: A multi-task benchmark for video-and-language understanding evaluation. In *NeurIPS*, 2021. 1, 2

[8] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. *arXiv preprint arXiv:2106.13230*, 2021. 1, 2

[9] Meredith Ringel Morris. Ai and accessibility. *Communications of the ACM*, 63(6):35–37, 2020. 2

[10] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, 2019. 2

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1

[12] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *KDD*, 2020. 2

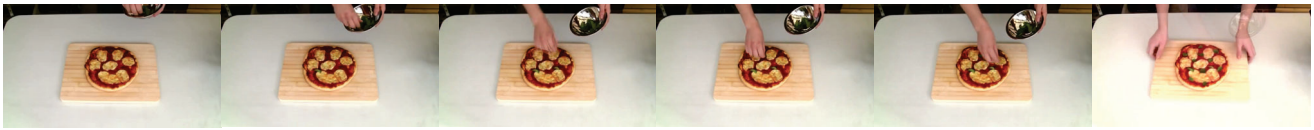
[13] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 2015. 1

[14] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, 2015. 2

[15] Yifan Xu, Zhijie Zhang, Mengdan Zhang, Kekai Sheng, Ke Li, Weiming Dong, Liqing Zhang, Changsheng Xu, and Xing Sun. Evo-vit: Slow-fast token evolution for dynamic vision transformer. *arXiv preprint arXiv:2108.01390*, 2021. 2

[16] Ziqi Zhang, Zhongang Qi, Chunfeng Yuan, Ying Shan, Bing Li, Ying Deng, and Weiming Hu. Open-book video captioning with retrieve-copy-generate network. In *CVPR*, 2021. 2

[17] Ziqi Zhang, Yaya Shi, Chunfeng Yuan, Bing Li, Peijin Wang, Weiming Hu, and Zheng-Jun Zha. Object relational graph with teacher-recommended learning for video captioning. In *CVPR*, 2020. 2



VideoID: efnHOsT7k9s_6
Generated caption: Place the basil on the pizza
GT: Place basil leaves on top of the pizza

VideoID: tYg3lQ5aZv8_2
Generated caption: Chop the green onion
GT: Finely chop green onions

VideoID: E9O9-6TQUw0_2
Generated caption: Season the meat with salt and pepper
GT: Sprinkle salt and pepper on top of the meat

VideoID: 0uaKitJaqmI_6
Generated caption: Boil the potatoes in water
GT: Boil the potatoes

VideoID: 0uaKitJaqmI_7
Generated caption: Mash the potatoes with salt
GT: Mash the potatoes

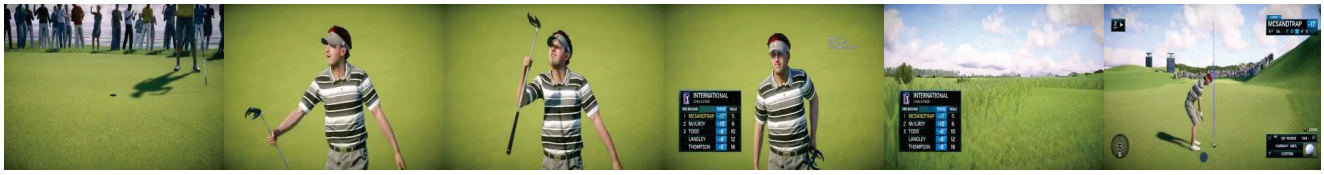
VideoID: 0uaKitJaqmI_9
Generated caption: Spread the mashed potatoes on top of the meat
GT: Cover the meat mixture with the mashed potatoes

Figure 1. Qualitative examples generated by SWINBERT on YouCook2 dataset.

VideoID: video6674
Generated caption: A chef in a white apron is cooking a dish
GT1: A chef shows how to prepare a dish
GT2: A man is cooking and describe his process
GT3: A chef is giving instructions in the kitchen

VideoID: video6741
Generated caption: A woman is giving a speech
GT1: A woman is giving a speech
GT2: A lady gives a speech at a podium
GT3: Hilary Clinton is giving a speech at Columbia University

VideoID: video6909
Generated caption: A girl is applying makeup to her face
GT1: A girl applying makeup to her face
GT2: A woman applies makeup to her face
GT3: A woman is doing makeup and showing eye blush



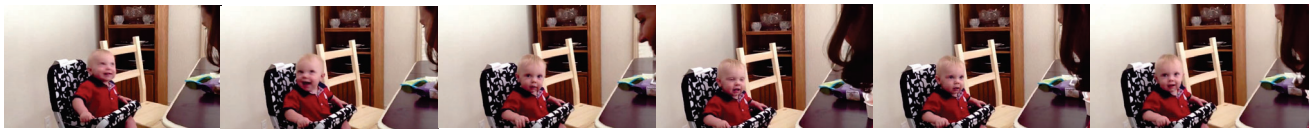
VideoID: video7001
Generated caption: A man is playing golf
GT1: A man is playing golf
GT2: A man playing golf
GT3: A man is having golf bat

VideoID: video6963
Generated caption: A football player is running
GT1: A football player is running to the end zone
GT2: A man is playing football
GT3: A football player makes a touch down

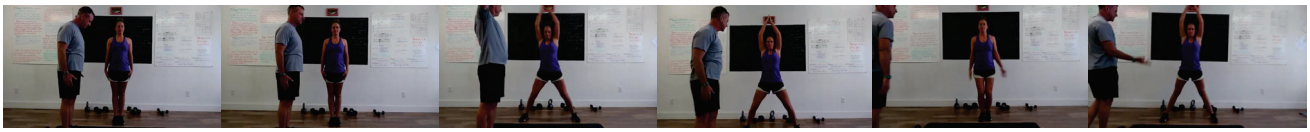
Figure 2. Qualitative examples generated by SWINBERT on MSRVTT dataset.

VideoID: EA3HCx0yTIY_000281_000291
Generated caption: A man is sitting at a drum set and playing the drums
GT1: A man sits and plays music on a set of drums
GT2: A man is seated at a drum set striking the cymbal several times
GT3: A man sits at his drum set and slowly hits the symbols continuously

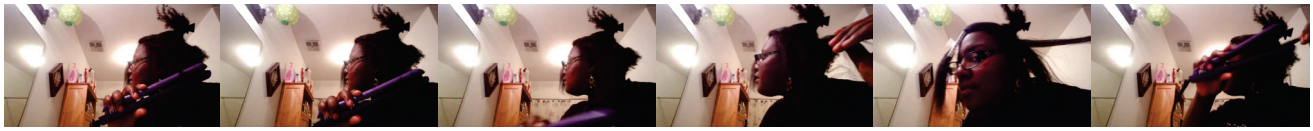
VideoID: G0mjFqytJt4_000152_000162
Generated caption: A young boy is showing how to make a paper airplane
GT1: A young boy in his bathroom as he explains how to make a paper airplane
GT2: A boy gives close up instructions on how to make a paper plane
GT3: A little boy is showing how to stuff paper to make a paper airplane



VideoID: IczD9OzKvco_000102_000112
Generated caption: A baby is sitting in a high chair and shaking his head back and forth
GT1: A baby is sitting in a chair and is shaking his head
GT2: A baby sitting in a high-chair shakes his head at a woman
GT3: A baby in a high chair shakes head back and forth and then looks at a woman



VideoID: Pj_070vBUeQ_000010_000020
Generated caption: A man and a woman are doing jumping jacks in a gym
GT1: A man and a woman are making an instructional video on the proper way to do jumping jacks
GT2: A man is at a gym teaching a woman how to properly do a jumping jack
GT3: A man instructs a woman on how to do star jumps in a gym

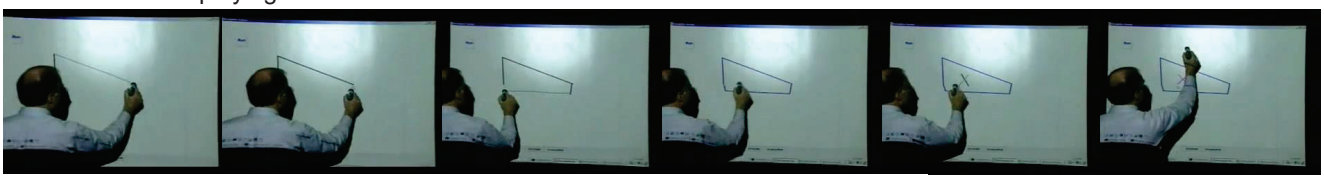


VideoID: ypsPcmnMIg8_000278_000288
Generated caption: A woman is using a curling iron to straighten her hair
GT1: A teenage girl uses a hair iron to straighten her hair
GT2: A young woman is using a straightening iron on her hair
GT3: A lady is looking in the mirror and curling her hair with a curling iron

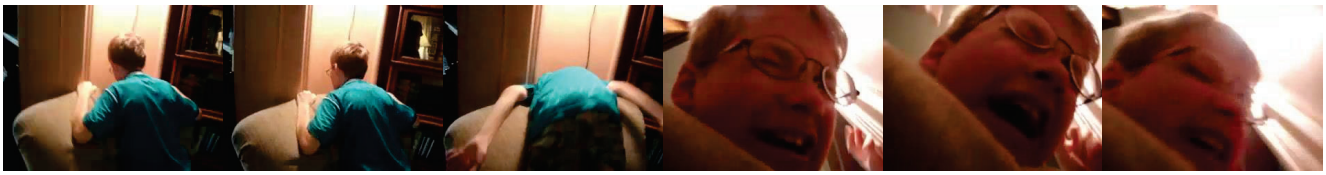
Figure 3. Qualitative examples generated by SWINBERT on VATEX dataset.

VideoID: d7Gs0uGFLh0_5_13
Generated caption: A woman is dancing on a stage
GT1: A woman is dancing
GT2: The little girl is dancing
GT3: A girl is dancing on stage

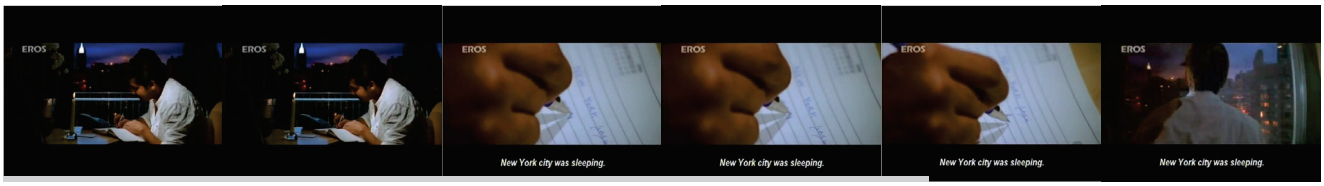
VideoID: crfrKqFp0Zg_15_25
Generated caption: A man is playing a flute
GT1: A man is playing a flute
GT2: A man is playing the flute
GT3: The man is playing flute



VideoID: d7eGypGOlOc_13_22
Generated caption: A man is writing on a white board
GT1: A man is drawing on a white board
GT2: A man is writing on the board
GT3: The man is drawing a geometric shape on a white board



VideoID: dhxE9CNeVeY_0_12
Generated caption: A boy is screaming and shouting
GT1: A boy is screaming
GT2: A kid is screaming behind a chair
GT3: A boy is screaming and crying



VideoID: ejgwQqCHN1E_7_12
Generated caption: A man is writing on a piece of paper
GT1: A man is writing a note
GT2: A man is writing
GT3: A man is writing a letter

Figure 4. Qualitative examples generated by SWINBERT on MSVD dataset.