# Re-examination and improvements to SWINBERT from multiple perspectives

**Coauthor**
Huang He
Peking University
2201111731@stu.pku.edu.cn

**Coauthor**
Sha Guo
Peking University
sandykwokcs@stu.pku.edu.cn

**Coauthor**
Zehong Ma
Peking University
zehongma@stu.pku.edu.cn

## Abstract

Video captioning is one of the hottest topics in multi-modal research today. Among the methods proposed in recent years, SWINBERT achieves the state of the art performance in video caption. In this paper, we further examine the effectiveness of the original SWINBERT and propose the following approaches to improve it: 1)Unlike static images, the additional temporal dimension of video delivers another important cue for semantic understanding, namely motion. In addition to the original input modality of RGB, Our method investigates another modality optical flow as the source of motion representation, and is proven to be effective. 2)High computational complexity is the main obstacle for the practical application of the video transformer. We will investigate sparse masking strategies and propose neighborhood masked attention to reduce the redundancy in attention mechanism and improve the efficiency of the SWINBERT. 3)The attention mask is sparse enough, but the model still take as input as all dense sampled frames. So we introduce an adaptive frame sampling module for selecting sparse frames from dense samplied videos. 4)There exists huge modal heterogeneity since the model is random initialized and trained from scratch. So it's necessary to take advantage of rich multi-modal information in vision-language pretraining model to narrow modal heterogeneity. And in practice, we introduce the popular CLIP model as the multimodal representation extractor and proposed the Video Caption CLIP(VC-CLIP) as a strong baseline. We prove the effectiveness of our methods through concrete experiments.

## 1 Introduction

We argue that the SWINBERT is impeded by three major obstacles. First, although motion information has been proven crucial for understanding the dynamics in traditional video action recognition methods, it has not been considered as a critical factor in SWINBERT framework. Swin focus on perceptual appearance(i.e., color, contour, details) of non-local spatial and temporal, thus lacking the capacity to explicitly incorporate short-term and long-term motion structure.

Thirdly, the attention mask is sparse enough, which means only a few tokens in some specific frames are used during training or inference, but the model still take as input as all dense sampled frames.

And to be honest, the SwinBERT has a large model size and will take a few seconds to generate caption for video clips, which severely restricts application of SwinBERT.

Finally, the parameters in SwinBERT are initialized with uni-modal pretraining weights or random value, which means it need to learn the multi-modal representation and interaction from scratch for a long time and the modal heterogeneity may still exist and hugely influence the model's performance if the size of train data is not large enough.

These challenges motivate us to study the video caption problem in this project from the following three aspects: *1) how to effectively learn motion representation that capture long term and short-term structure; 2) how to design a more efficient masking strategy to improve SwinBert's performance during inference ; 3) how to adaptively select the sparse frames from dense videos. 4) Take advantage of rich multi-modal information in vision-language pretraining model to narrow modal heterogeneity.*

To capture short-term and long-term motion structure, we develop a optical flow modality module, which uses a SOTA pre-trained deep learning based optical flow model PWC-Net(71) to calculate dense motion optical flow between sampled frames. We then extend RGB input modality to RGB and optical flow field by simply adding their embedding together before sending them to Swin architecture. By combing these two inputs, we build the better video caption system, which has numerous potential applications in real-world problems.

Besides, to adaptively select the sparse frames from videos, we propose a frame sampling module, which derives from MG-Sampler(72). The sampling module calculates the difference between two adjacent frames which are preprocessed by an $7\times7$ average kernel as motion representation and then leverage a motion-uniform sampling strategy based on the cumulative motion distribution to ensure the sampled frames evenly cover all the important segments with high motion salience.

To take full advantage of the great multimodal representations in visual-language pretraining, we replace the video transformer with CLIP(16) visual encoder and only keep the multimodal transformer as multimodal interaction module. Each selected frame of video is encoded into a visual token by CLIP visual encoder. And each word is separately encoded into a non-contextual textual token by textual encoder in CLIP. The multimodal transformer take as input the visual and textual tokens encoded by CLIP, which makes the multimodal transformer mainly focus on the interaction of multimodal representation instead of multimodal representations' extraction or alignment.

## 2 Related Work

### 2.1 Video caption

Recent researches (4; 5; 6; 7; 8) mainly focus on modeling the relationship between fixed video representations and the output textual descriptions via an encoder-decoder framework for video captioning. Specifically, these methods(9; 10; 11; 5; 12) employ an encoder to refine video representations from a set of fixed video frame features, and a language decoder operates on top of these refined video representa- tions to learn visual-textual alignment for caption generation. Researchers(4; 10; 6) have focused on exploring different 2D/3D video representations, including IncepRes-NetV2(13), ResNet(14), CLIP-ViT(15; 16), SlowFast(17), C3D(18) and S3D(19; 20), for improving video captioning. In addition, object-level representations(21; 22; 23) have been explored to enrich captions with fine-grained objects and actions. Prior works(24) also studied frame selection schemes to capture informative visual inputs. Unlike previous studies that learn from multiple offline-extracted 2D/3D features with a fixed sampling rate, we introduce Video Swin Transformer(25) as the video encoder in our framework to encode spatial-temporal representations from raw video frames. Benefiting from the flexibility of the transformer architecture, our model can learn with variable number of video tokens and can be trained end-to-end.

### 2.2 Video transformers

Dosovitskiy et al.(15) demonstrate that a pure-transformer based architecture can outperform its convolutional counterparts in ImageNet classification task(26). Since then, there has been a growing interest in applying vision transformer (ViT) to the video domain. For example, ViViT(27) and TimeSformer(28) propose a new transformer architecture that can leverage spatial-temporal attention for improving representation learning. Video Swin Transformer (VidSwin)(29) further introduces locality inductive bias into the transformer self-attention, and achieves state-of-the-art performance on action recognition benchmark(30). While recent studies(27; 28; 29) mainly focus on developing video transformer architecture for action recognition, video captioning has not been explored along this research direction, which is the focus of this work.

### 2.3 Video and language

Recent studies(31; 32; 33; 34; 35; 36) have shown great success on multimodal representation learning for video-and-language understanding. Popular downstream tasks include video question answering(37), text-video retrieval (38; 39)and video captioning(40). Among the literature, Frozen-in-time(41) is a relevant study that explores pure transformer-based model design, but they focus on text-video retrieval. Specifically, they employ two independent transformer encoders for visual and textual inputs, respectively. Retrieval is conducted by estimating the similarity between the outputs of their visual and textual encoders. With a similar spirit, CLIP4Clip(34) studied using the pre-trained CLIP(16) as a feature extractor for video retrieval. While existing architectures(41; 33) are effective for video retrieval, it cannot be directly applied to video captioning, which is the focus of this work.

### 2.4 Optical flow

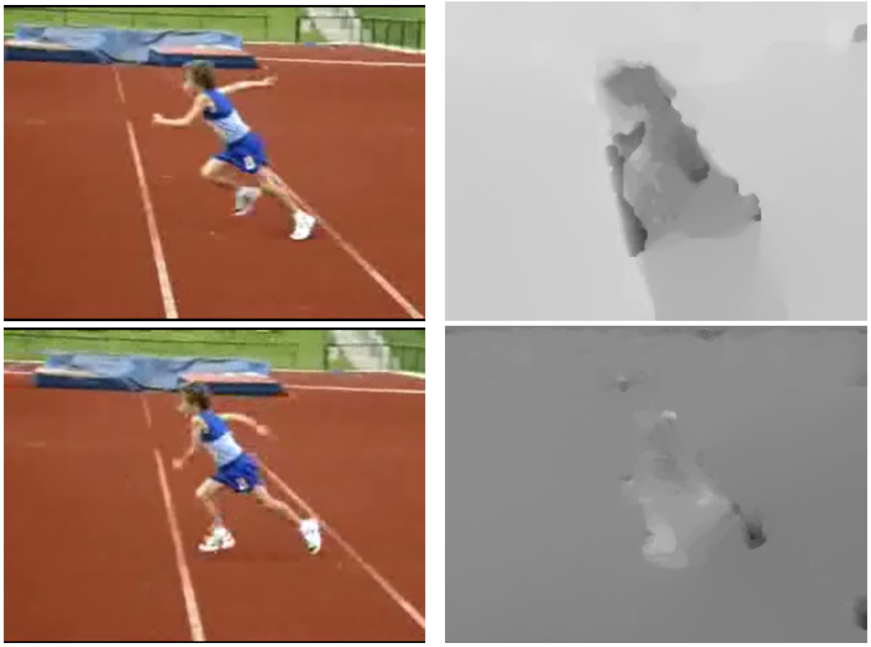
Optical flow estimation is a core computer vision problem and has many applications, e.g., action recognition(42), autonomous driving(43) , and video editing(44). Decades of research efforts have led to impressive performances on challenging benchmarks(45; 46; 47). Most top-performing methods adopt the energy minimization approach introduced by Horn and Schunck(48). However, optimizing a complex energy function is usually computationally expensive for real-time applications.

Horn and Schunck(48) pioneer the variational approach to optical flow by coupling the brightness constancy and spatial smoothness assumptions using an energy function. Black and Anandan(49) introduce a robust framework to deal with outliers, i.e., brightness inconstancy and spatial discontinuities. As it is computationally impractical to perform a full search, a coarse-to-fine, warping-based approach is often adopted(50). Brox et al.(51) theoretically justify the warping-based estimation process. Sun et al.(52) review the models, optimization, and implementation details for methods derived from Horn and Schunck and propose a non-local term to recover motion details. The coarse-to-fine, variational approach is the most popular framework for optical flow. However, it requires solving complex optimization problems and is computationally expensive for real-time applications.

Inspired by the success of CNNs on high-level vision tasks(60), Dosovitskiy et al.(61) construct two CNN networks, FlowNetS and FlowNetC, for estimating optical flow based on the U-Net denoising autoencoder(62). The networks are pre-trained on a large synthetic FlyingChairs dataset but can surprisingly capture the motion of fast moving objects on the Sintel dataset. The raw output of the network, however, contains large errors in smooth background regions and requires variational refinement(63). Mayer et al.(64) apply the FlowNet architecture to disparity and scene flow estimation. Ilg et al.(65) stack several basic FlowNet models into a large one, i.e., FlowNet2, which performs on par with state of the art on the Sintel benchmark. Ranjan and Black(66) develop a compact spatial pyramid network, called SpyNet. SpyNet achieves similar performance as the FlowNetC model on the Sintel benchmark, which is good but not state-of-the-art. Sun et al. present a compact but effective CNN model for optical flow, called PWC-Net(71). PWC-Net has been designed according to simple and well-established principles: pyramidal processing, warping, and the use of a cost volume. Cast in a learnable feature pyramid, PWC-Net uses the current optical flow estimate to warp the CNN features of the second image. It then uses the warped features and features of the first image to construct a cost volume, which is processed by a CNN to estimate the optical flow. PWC-Net is 17 times smaller in size and easier to train than the recent FlowNet2 model. Moreover, it outperforms all published optical flow methods on the MPI Sintel final pass and KITTI 2015 benchmarks, running at about 35 fps on Sintel resolution (1024×436) images. PWC-Net models are available on https://github.com/NVlabs/PWC-Net.

# 3 Our approach

## 3.1 Multi-Modality Swin Module



(a) RGB images.

(b) optical flow fields.

Figure 1: Examples of two types of input modality: RGB images and optical flow fields (x,y directions).

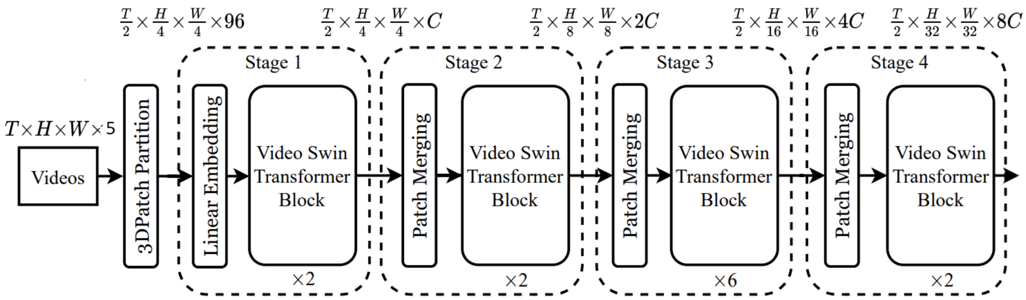

$\frac{T}{2}\times\frac{H}{4}\times\frac{W}{4}\times 96$
$\frac{T}{2}\times\frac{H}{4}\times\frac{W}{4}\times C$
$\frac{T}{2}\times\frac{H}{8}\times\frac{W}{8}\times 2C$
$\frac{T}{2}\times\frac{H}{16}\times\frac{W}{16}\times 4C$
$\frac{T}{2}\times\frac{H}{32}\times\frac{W}{32}\times 8C$
$T\times H\times W\times 5$
Videos
3DPatch Partition
Stage 1
Linear Embedding
Video Swin Transformer Block
×2
Stage 2
Patch Merging
Video Swin Transformer Block
×2
Stage 3
Patch Merging
Video Swin Transformer Block
×6
Stage 4
Patch Merging
Video Swin Transformer Block
×2

Figure 2: Architecture of our Multi-modality Video Swin Transformer.

The architecture of our proposed Multi-Modality Swin Transformer is shown in Fig. 2. The input modality is a little bit different from original Video Swin Transformer. As shown in Fig. 1, the input is defined to be of size $T \times H \times W \times 5$, consisting of T frames which each contain $H \times W \times 3$ RGB pixels and $H \times W \times 2$ optical flow(x and y directions), which is inferenced by pre-trained PWC-Net(71). In Multi-Modality Swin Transformer, we treat each 3D patch of size $2 \times 4 \times 4 \times 3$ and $2 \times 4 \times 4 \times 2$ as RGB token and optical flow token respectively. Thus, the 3D patch partitioning layer obtains $\frac{T}{2} \times \frac{H}{4} \times \frac{W}{4}$ 3D RGB tokens and optical flow tokens, with each patch/token consisting of a 96-dimensional feature. A linear embedding layer is then applied to project the features of each token to an arbitrary dimension denoted by C. After linear embedding layer, We then simply add RGB tokens and optical flow tokens together before sending them into Video Swin transformer Block.

Following the prior art, we then strictly follow the hierarchical architecture of the original Video Swin Transformer, which consists of four stages and performs 2× spatial downsampling in the patch merging layer of each stage. The patching merging layer concatenates the features of each group of $2 \times 2$ spatially neighboring patches and applied a linear layer to project the concatenated features to half of their dimension. For example, the linear layer in the second stage projects 4C-dimensional features of each token to 2C dimensions. The Video Swin Transformer block also follow the original work, which is built by replacing the multi-head self-attention module in the standard Transformer layer with the 3D shifted window based multi-head self-attention module and keeping the other components unchanged. Specifically, a video transformer block consists of a 3D shifted window based MSA module followed by a feed-forward network, specifically a 2-layer MLP, with GELU no-linearity in between. Layer Normalization is applied before each MSA module and FFN, and a residual connection is applied after each module.

### 3.2 Attention Masking Strategy

In this section, we will look into the attention masking strategy that the SWINBERT employs and further propose new masking methods.

#### 3.2.1 Sparse Attention Mask

The motivation of the Sparse Attention Mask is to solve the Calculation power consumption problems the transformer models always encounter in applications, especially in the video processing scenes. the computational demand of attention are proportional to input length, which limits the number of input frames.On the other hand, considering the essence of the video properties, the dense sampling scheme with consecutive video frames contains redundant and perhaps irrelevant information, which may compromise performance. The problem is addressed by introducing a learnable Sparse Attention Mask as a regularizer to the transformer encoder. As is shown in Figure 3, the attention among Language tokens adopt GPT mod, SWINBERT model generates one word token at a time during inference. Also, the Word tokens have full attention mask with video tokens. Sparse Attention mask is applied on the attention between video tokens. The value of the mask is generated from a group of learnable variable corresponding to each position of the attention mask. The value of the sparse is normalized by the softmax function, then the position where value of mask is larger 0.5 is set to be 1. The learnable varibles are trained by the loss sparse is $Lsparse = \lambda \times \sum_{i=1}^{M} \sum_{j=1}^{M} |Vi, j|$ , M represents the length of the tokens the pretrained 3D swin transformer extracts, V is the value of the learnable mask for each position in the attention map. This method is applied as a global mask

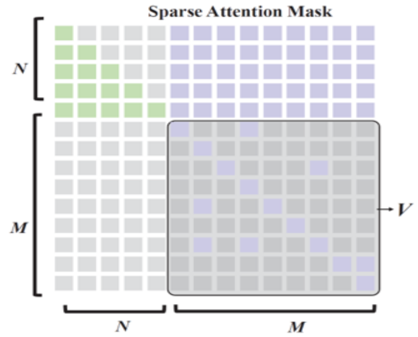


Figure 3: The overview of the sparse masking: a learnable Sparse Attention Mask as a regularizer for multimodal transformer encoder to reduce redundancy among the video tokens.

(same for each layer of the transformer) strengthen the most important relationships among different tokens by reducing the likelihood of meaningless connections, while focusing more on the active video tokens that contain rich spatial-temporal information.

### 3.2.2 The neighbor masked attention

We looked into the attention mask the author offers: we find that in practice, the value of the attention is not sparse enough (there are no >0.5 elements apart from diagonal in 32 frames. This result means that although the sparse masking is carefully designed to find tokens with most useful information for the captioning task and reduce irrelevant calculations, it might end up in failure especially in long videos. The resulting mask might only have elements on diagonal that can be used for attention map calculation. This will bring damage to the final performance.

In order to solve such problem,we introduce the neighbor masked attention, it is originally used in 2 dimension, but it can be extend to an additional t dimension.The idea comes from the observation that the full attention in the transformer contributes to the “identical shortcut”. In full attention, one token is permitted to see itself, so it will be easy to reconstruct by simply copying. This bring unnecessary calculation and may lead the network to learn the easy and plain features. Thus we mask the neighbor tokens when calculating the attention map, called Neighbor Masked Attention (NMA). In the 3D heat map, the spatial and temporal neighbor of the tokens are unseeable in the attention map, for they share similar features, forcing the network to find more global relationships and reduce redundancy. The NMA masking is shown in figure 4.

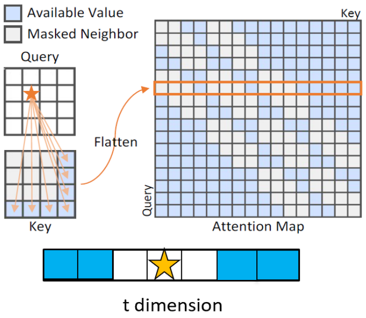


Figure 4: The overview of the NMA: for a token in the input of 3 dimension,we mask a percentage of neighborhood tokens in spatial and temporal dimension

### 3.2.3 Other tried masking strategies

we also apply some other masking strategies to the SwinBert. For example, we tried randomly activation: randomly select a small portion of tokens in the row of token i in the attention map to activation in the attention calculation. We also tried Input-feature Based Attention Generation Network(IBAGN), which takes the extracted tokens as the input then process them with convolutional network,upsample and output the attention mask directly. In this way,the network produce the attention mask individually instead of using the same mask for every input like sparse masking or NMA masking strategies. However, such network may hard to tune for long token sequences.

## 3.3 Adaptive Frame Sampling

There is a lot of sparsity between tokens in SwinBERT, so obviously there also exists sparsity between input frames. So we introduced an adaptive frame sampling module to reduce the number of input frames, and at the same time hope to keep the comparable performance with the dense sampling method. Firstly, following MGSampler(72), we introduce a motion representation $Diff_t$ for each frame as shown in Figure 5 which is extracted by subtraction between adjacent feature-level

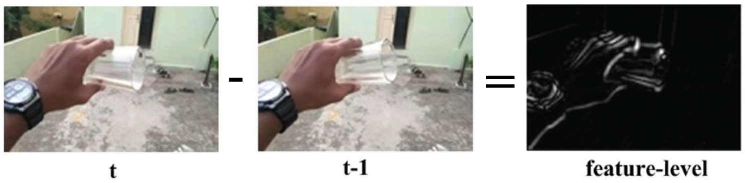


Figure 5: Feature-level motion representation for adaptive frame sampling

frame representation which are processed by a shallow convolution before subtraction. The shallow convolution only contains one convolution layer whose kernel shape is $3\times1\times7\times7$, and the value of kernel is initialized by 1e-3 for average pooling. It is noteworthy that the shallow convolution layer cannot be trained and the kernel is kept constant because the sampling operation is discrete and not differentiable. Then we can get the motion representation $Diff_t$ by subtracting the previous frame's feature-level representation $F_{t-1}$ from the current frame's feature-level representation $F_t$ as:

$$Diff_t = F_t - F_{t-1}. \tag{1}$$

Next, after getting the difference score $Diff_t$ of adjacent frames, we can get a distribution curve of accumulative motion by summing the difference score across the time dimension as shown in Figure 6. Then we uniformly sample in the distribution curve vertically, and calculate the closest frame number corresponding to the intersection point as the selected frame. Now, we can adaptively sample the frames with more variation and more information in the video to help the video captioning.

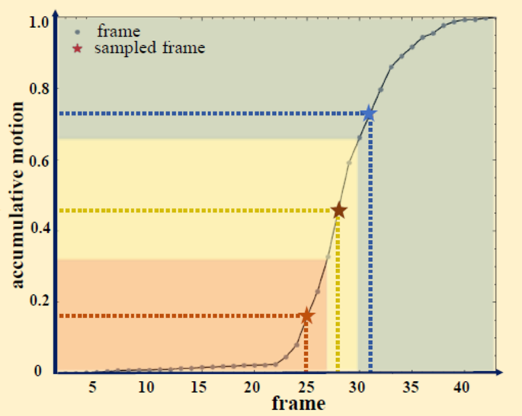


Figure 6: The distribution curve of accumulative motion(72)

### 3.4 VC-CLIP based on Vision-Language Pretraining

The SwinBERT is used to generate caption for video, but it doesn't take use of any vision-language pretraining, which means the model need to extract single-modal feature well and then project them into a shared space or capture the complicated relationship between vision and language. Meanwhile, the model size is so large and the model's visual input is also too large.

To overcome the above two disadvantages, we replace the video transformer with CLIP(16) visual encoder and only keep the multimodal transformer as multimodal interaction module. Each selected frame of video is encoded into a visual token by CLIP visual encoder. And each word is separately encoded into a non-contextual textual token by textual encoder in CLIP. The multimodal transformer take as input the visual and textual tokens encoded by CLIP, which makes the multimodal transformer mainly focus on the interaction of multimodal representation instead of multimodal representations' extraction or alignment. It is worthy that the visual and textual encoder in CLIP are kept fixed during training and only the BERT-based multimodal transformer is trained, which much reduce the training

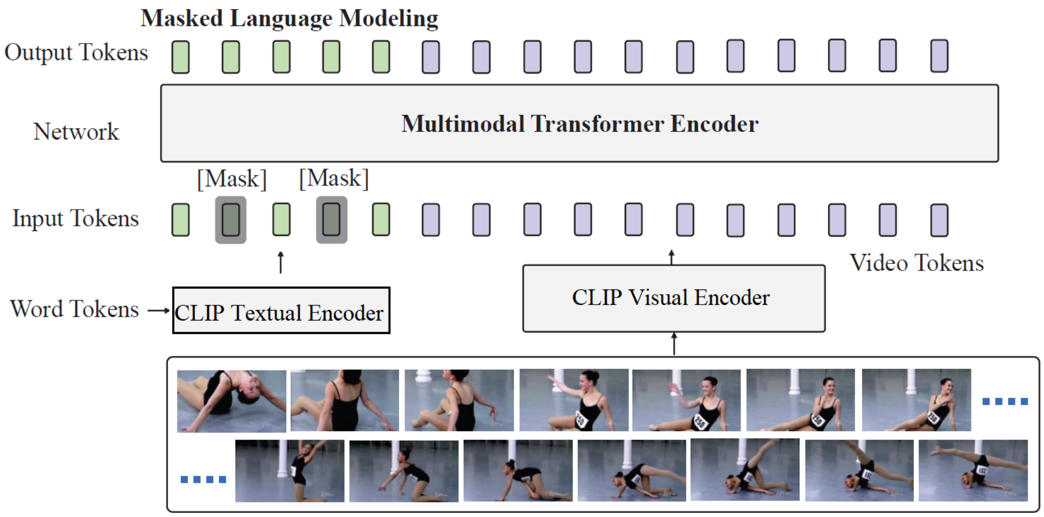


Figure 7: The architecture of VC-CLIP

Table 1: Comparison of multi-modality methods on the test split of MSVD.

| Method | B4 | R | M | C |
|---|---|---|---|---|
| RGB+SwinBert | 0.6016 | 0.7804 | 0.4075 | 128.9 |
| RGB+Optical flow + SwinBert | 0.6426 | 0.8035 | 0.4216 | 138.2 |

time and resources. The new model called Video Caption CLIP which is dubbied as VC-CLIP is showed in Figure 7.

## 4 Experiments

In this section, we focus on investigating the effect of the good practices described in Sec. 3, including the input modalities and the adaptive masking strategies.

### 4.1 Input modalities

We propose new type of modalities in Section 3.1: RGB and optical flow fields. We test different modalities and report the results in Table3. We outperform multiple experiments with good practices verified in original SwinBert to investigate the performance of different modalities. We observe that RGB and optical flow works well with SwinBert, yielding B4 score of 0.6426, R score of 0.8035, M score of 0.4216, and CIDEr score of 138.2.

### 4.2 Effectiveness of The Masking strategies

In this section,we will discuss the effectiveness of proposed making strategies. We test the methods on the MSVD dataset. MSVD is a collection of 2K open-domain video clips downloaded from YouTube. Each video clip has 35 ground-truth captions written by human. We use the standard split which contains 1.2K training videos and 670 test videos. During the experiment, we sample the video in 32 frames and test the CIDEr score. The result is in the Table 2. As is shown in the Table 2, the proposed NMA masking with window size 5 brings a small improvement of 4.1 points to the model,

Table 2: Comparison of different masking strategies

| Method | Sparse Mask | Random mask | NMA(window-5) | IBAGN |
|---|---|---|---|---|
| CIDEr | 145.0 | 143.5 | 149.1 | 145.3 |

while the random mask failed in comparison with the original sparse masking. This results means that the NMA manages to reduce the information redundancy and keep the useful global knowledge that is helpful in final captioning tasks. Meanwhile, the effectiveness of IBAGN is not remarkable, the improvement is only 0.3 point. The learnable network producing masks for each input might not be better than just learning the global masking like the sparse masking.

**Window size's influence:** On the other hand, we also test the influence of the window size of the NMA masking. The result is in Table 3. The result shows that different window size affects the performance of the NMA masking. Clearly, if the window is too small, the redundancy still remains. While if the window size is too large, the CIDER score drops, for the useful information in the attention is over covered. In general, the scores are all above 148 compared with 145.0 from the original method. This proves the effectiveness of the NMA masking.

Table 3: Comparison of different window size in NMA

| Window size | Sparse Mask | window-3 | window-5 | window-9 |
|---|---|---|---|---|
| CIDEr | 145.0 | 148.7 | 149.1 | 148.4 |

### 4.3 Effectiveness of Adaptive Frame Sampling

Table 4: Comparison of different adaptive sampling ratio on MSVD

| Method | CIDEr |
|---|---|
| 8frame(base) | 117.8 |
| 16->8frame adaptive | 129.2 |
| **32->8frame adaptive** | **137.5** |
| 64->8frame adaptive | 136.7 |
| 32frame(target) | 148.3 |

Because the model size of SwinBERT is so large, we set the model trained on the video consisting of uniformly sampled 8 frames as our base model. From the Table 4, we can see that, when we evaluate on the adaptively sampled frames, the performance is improved by a large margin compared with the base model. And when we sample 8 frames from 32 frames adaptively, the performance is improved most. But compared to model trained on 32 frames, there is still a performance gap.

As shown in Table 5, in order to improve the performance of adaptive frame sampling, we firstly finetune the SwinBERT on selected 8 frames which are adaptively chosen from 32 uniformly sampled frames. However, it's disappointing that the performance drops from 137.5 to 122.5. Compared with training on selected 8 frames which contain more motion information and variation, training on uniformly sampled 8 frames directly may require the model to have a stronger reasoning ability or capability and be able to extract useful information from video frames that contain less information. In contrast, training on selected frames may harm the model's reasoning ability because more information usually means learning more easily. So when we directly leverage the base model to generate caption from selected frames including more information, the model can get a better performance. And when we finetune SwinBERT on selected 8 frames, the performance(122.5) is improved a little compared with base model's performance 117.8 which trained on uniformly sampled 8 frames while drops by a large margin compared with no finetuning(137.5).

For the reason that the shallow convolution is discrete and not differentiablewe take the feature-level representation as additional feature and add it to the frame's RGB channel as the fourth channel so that the parameters in shallow convolution can be trained with the model together. However, as shown in the fourth row of Table 5, the performance is only 120.1, which is a little poor and means the constant average kernel is a good feature extractor for adaptive frame sampling.

Table 5: Influence of shallow convolution on MSVD

| Method | CIDEr |
|---|---|
| 8frame(base) | 117.8 |
| 32->8frame adaptive(no futher finetuning) | 137.5 |
| finetune SwinBERT on selected 8 frames | 122.5 |
| finetune SwinBERT+Shallow-Conv on selected 8 frames | 120.1 |

### 4.4 Results of VC-CLIP

We propose a new model VC-CLIP for taking full use of the good multimodal representation learning from vision-language pretraining and reducing the model's complexity. And the VC-CLIP get a promising results without modifying any hyperparameter which means there still a lot of potential in VC-CLIP waiting to be exploited.

Table 6: Result of VC-CLIP on MSVD

| Method | CIDEr | Training Time |
|---|---|---|
| SwinBERT | 117.8 | 5.2h |
| VC-CLIP(RN50) | 85.6 | 1.5h |
| VC-CLIP(RN101) | 91.3 | 2.8h |
| VC-CLIP(RN50×4) | 104.6 | 4 h |
| VC-CLIP(ViT-B/16) | **127.6** | 2.5h |

We freeze the CLIP model in VC-CLIP and only train the parameter in multimodal transformer for multimodal interaction. As shown in the last row of Table 6, when using the CLIP's ViT-B/16 pretraining model, the CIDEr is improved by a large margin compared with SwinBERT while the training time is only a half of SwinBERT. And from the second row to the fourth row of Table 6, we can see that the model's performance is improved with the increasing of model's size and training time. In conclusion, our VC-CLIP is a strong and promising baseline for end-to-end Video caption which have a good comprehension of vision and language information.

## 5 Conclusions

In this paper, we examine the theory of SWINBERT, test its effectiveness and propose methods to improve the performance of the original approach. We come to the conclusion after a variety of experiments that: 1) In addition to the original input modality of RGB, our method investigate another modality optical flow as the source of motion representation, and is proven to be effective. 2) The sparse masking strategies like neighborhood masked attention(NMA) are proven to be helpful to raise the CIDEr score of the sparse attention mechanism.3) The adaptive frame sampling can dynamically select the sparse frames with more information and variation, which outperforms the base model at the frame rate of 8 by a largin margin and achieves comparable performance with the base model at the frame rate of 32. 4) Our VC-CLIP is a strong and promising baseline for end-to-end Video caption which have a better comprehension of vision and language information with lower computational cost and shorter training time.

## References

[1] Alexander, J.A. & Mozer, M.C. (1995) Template-based algorithms for connectionist rule extraction. In G. Tesauro, D.S. Touretzky and T.K. Leen (eds.), *Advances in Neural Information Processing Systems 7*, pp. 609–616. Cambridge, MA: MIT Press.

[2] Bower, J.M. & Beeman, D. (1995) *The Book of GENESIS: Exploring Realistic Neural Models with the GEneral NEural SImulation System.* New York: TELOS/Springer–Verlag.

[3] Hasselmo, M.E., Schnell, E. & Barkai, E. (1995) Dynamics of learning and recall at excitatory recurrent synapses and cholinergic modulation in rat hippocampal region CA3. *Journal of Neuroscience* **15**(7):5249-5262.

[4] Nayyer Aafaq, Naveed Akhtar, Wei Liu, Syed Zulqarnain Gilani, and Ajmal Mian. Spatio-temporal dynamics and semantic attribute enriched visual encoding for video captioning. In CVPR, 2019. 1, 2, 5

[5] 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. 1, 2

[6] Boxiao Pan, Haoye Cai, De An Huang, Kuan-Hui Lee, Adrien Gaidon, Ehsan Adeli, and Juan Carlos Niebles. Spatio-temporal graph for video captioning with knowledge distillation. In CVPR, 2020. 1, 2, 5

[7] Wenjie Pei, Jiyuan Zhang, Xiangrong Wang, Lei Ke, Xiaoyong Shen, and Yu-Wing Tai. Memory-attended recurrent network for video captioning. In CVPR, 2019. 1, 2

[8] Botian Shi, Lei Ji, Zhendong Niu, Nan Duan, Ming Zhou, and Xilin Chen. Learning semantic concepts and temporal alignment for narrated video procedural captioning. In ACM MM, 2020. 1, 2

[9] Shaoxiang Chen, Ting Yao, and Yu-Gang Jiang. Deep learning for video captioning: A review. In IJCAI, 2019. 1, 2

[10] 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, 3, 5, 6, 12, 13

[11] Sheng Liu, Zhou Ren, and Junsong Yuan. Sibnet: Sibling convolutional encoder for video captioning. IEEE TPAMI, 2020. 1, 2, 5

[12] ZiqiZhang,ZhongangQi,ChunfengYuan,YingShan,Bing Li, Ying Deng, and Weiming Hu. Open-book video captioning with retrieve-copy-generate network. In CVPR, 2021. 1, 2, 5, 13

[13] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In AAAI, 2017. 2

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In CVPR, 2016. 2

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In ICLR, 2020. 2

[16] 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. 2,3,12

[17] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In ICCV, 2019. 2, 12

[18] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In CVPR, 2018. 2

[19] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In CVPR, 2020. 2, 3

[20] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In ECCV, 2018. 2

[21] Yaosi Hu, Zhenzhong Chen, Zheng-Jun Zha, and Feng Wu. Hierarchical global-local temporal modeling for video captioning. In ACM MM, 2019.2

[22] Junchao Zhang and Yuxin Peng. Object-aware aggregation with bidirectional temporal graph for video captioning. In CVPR, 2019.2, 5

[23] ZiqiZhang,YayaShi,ChunfengYuan,BingLi,PeijinWang, Weiming Hu, and Zheng Jun Zha. Object relational graph with teacher-recommended learning for video captioning. In CVPR, 2020. 2, 5, 13

[24] Yangyu Chen, Shuhui Wang, Weigang Zhang, and Qingming Huang. Less is more: Picking informative frames for video captioning. In ECCV, 2018. 2, 5

[25] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. arXiv preprint arXiv:2106.13230, 2021. 2, 3, 5, 12

[26] 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. 2, 12

[27] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lucic, and Cordelia Schmid. Vivit: A video vision transformer. In ICCV, 2021. 2, 3

[28] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is spacetime attention all you need for video understanding? In ICML, 2021. 2, 3, 12

[29] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. arXiv preprint arXiv:2106.13230, 2021. 2, 3, 5, 12

[30] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In CVPR, 2017. 2, 3

[31] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In CVPR, 2021. 1, 2, 3

[32] 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, 3, 5, 6, 12, 13

[33] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval. arXiv preprint arXiv:2104.08860, 2021. 3

[34] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In CVPR, 2020. 2, 3

[35] 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 ICCV, 2019. 3

[36] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. In NeurIPS, 2021. 3

[37] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. EMNLP, 2018. 2, 3

[38] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvr: A large-scale dataset for video-subtitle moment retrieval. In ECCV, 2020. 2, 3, 5, 6

[39] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In CVPR, 2016. 2, 3, 5

[40] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In ICCV, 2019. 3, 5

[41] Max Bain, Arsha Nagrani, Gu l Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In ICCV, 2021.3

[42] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In Advances in Neural Information Processing Systems (NIPS), 2014. 1

[43] J. Janai, F. Guney, A. Behl, and A. Geiger. Computer vision for autonomous vehicles: Problems, datasets and state-of-the-art. arXiv preprint arXiv:1704.05519, 2017. 1

[44] N.Bonneel, J.Tompkin, K.Sunkavalli,D.Sun,S.Paris,and H. Pfister. Blind video temporal consistency. ACM SIGGRAPH, 34(6):196, 2015. 1

[45] S.Baker,D.Scharstein,J.P.Lewis,S.Roth,M.J.Black,and R. Szeliski. A database and evaluation methodology for optical flow. International Journal of Computer Vision (IJCV), 2011. 1, 2, 3

[46] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In European Conference on Computer Vision (ECCV), 2012. 1,3

[47] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012. 1, 3

[48] B.HornandB.Schunck. Determining opticalflow. Artificial Intelligence, 1981. 1, 2, 3, 4

[49] M. J. Black and P. Anandan. The robust estimation of multiple motions: Parametric and piecewise smooth flow fields. Computer Vision and Image Understanding (CVIU), 1996. 2, 3, 4

[50] A. Bruhn, J. Weickert, and C. Schnorr. Lucas/Kanade meets Horn/Schunck: combining local and global optic flow methods. International Journal of Computer Vision (IJCV), 2005. 2

[51] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. High accuracy optical flow estimation based on a theory for warping. In European Conference on Computer Vision (ECCV), 2004. 2, 3, 4

[52] D. Sun, S. Roth, and M. J. Black. A quantitative analysis of current practices in optical flow estimation and the principles behind them. International Journal of Computer Vision (IJCV), 2014. 2, 3, 4

[53] E. P. Simoncelli, E. H. Adelson, and D. J. Heeger. Probability distributions of optical flow. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1991. 2

[54] W. T. Freeman, E. C. Pasztor, and O. T. Carmichael. Learning low-level vision. International Journal of Computer Vision (IJCV), 2000. 2

[55] S. Roth and M. J. Black. On the spatial statistics of optical flow. International Journal of Computer Vision (IJCV), 2007. 2

[56] D. Sun, S. Roth, J. P. Lewis, and M. J. Black. Learning optical flow. In European Conference on Computer Vision (ECCV), 2008. 2

[57] Y. Li and D. P. Huttenlocher. Learning for optical flow using stochastic optimization. In European Conference on Computer Vision (ECCV), 2008. 2

[58] J. Wulff and M. J. Black. Efficient sparse-to-dense optical flow estimation using a learned basis and layers. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 120–130, 2015. 2

[59] M. Werlberger, W. Trobin, T. Pock, A. Wedel, D. Cremers, and H. Bischof. Anisotropic Huber-L1 optical flow. In British Machine Vision Conference (BMVC), 2009. 2

[60] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems (NIPS), 2012. 2

[61] A. Dosovitskiy, P. Fischery, E. Ilg, C. Hazirbas, V. Golkov, P. van der Smagt, D. Cremers, T. Brox, et al. FlowNet: Learning optical flow with convolutional networks. In IEEE International Conference on Computer Vision (ICCV), 2015. 1,2,3,4,5

[62] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional networks for biomedical image segmentation. In International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), 2015. 1, 2, 3

[63] T. Brox and J. Malik. Large displacement optical flow: Descriptor matching in variational motion estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2011. 2

[64] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. 2

[65] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox. FlowNet 2.0: Evolution of optical flow estimation with deep networks. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017. 1, 3, 4, 5, 7

[66] A. Ranjan and M. J. Black. Optical flow estimation using a spatial pyramid network. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017. 1, 3, 7

[67] R. Memisevic and G. Hinton. Unsupervised learning of image transformations. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2007. 3

[68] G. Long, L. Kneip, J. M. Alvarez, H. Li, X. Zhang, and Q. Yu. Learning image matching by simply watching video. In European Conference on Computer Vision (ECCV), 2016. 3

[69] J. J. Yu, A. W. Harley, and K. G. Derpanis. Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness. In CoRR. 2016. 3

[70] W.-S. Lai, J.-B. Huang, and M.-H. Yang. Semi-supervised learning for optical flow with generative adversarial networks. In Advances in Neural Information Processing Systems (NIPS), 2017. 3

[71] Sun D, Yang X, Liu M Y, et al. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 8934-8943.

[72] Yuan Zhi, Zhan Tong, Limin Wang, and Gangshan Wu. Mgsampler: An explainable sampling strategy for video action recognition. CoRR, abs/2104.09952, 2021. 1