Streaming Video Model
Supplementary Materials

Yucheng Zhao[1]*, Chong Luo[2], Chuanxin Tang[2], Dongdong Chen[3], Noel Codella[3], Zheng-Jun Zha[1]†
[1]University of Science and Technology of China [2]Microsoft Research Asia [3]Microsoft Cloud + AI
{lnc@mail., zhazj}@ustc.edu.cn {cluo,chutan,dochen,ncodella}@microsoft.com

Table S1. Hyper-parameters used in the action recognition.

Dataset	K400	SSv2
Batch size	256	64
Epochs	30	
Warmup epochs	5	
Learning rate	1e-5	5e-5
Learning rate schedule	cosine	
Optimizer	AdamW	
Weight decay	1e-3	5e-2
RandomFlip	0.5	-
ColorJitter	0.8	-
GrayScale	0.2	-
RandomAugment	-	✓
Random erasing	-	0.25
Repeated augmentation	-	2
Label smoothing	0.1	
Mixup	0.8	
CutMix	1.0	

A. Hyperparameter Details

We present the training hyper-parameters used in different datasets in Tab.S1 and Tab.S2. The hyper-parameters for action recognition are adapted from T2D [2]. And the hyper-parameters for multiple object tracking are adapted from ByteTrack [1]. The learning rates shown in Tables are for CLIP pre-trained parameters. For randomly initialized parameters, we use a 100× learning rate for K400 and a 10× learning rate for SSv2 and MOT17. In MOT training, we apply sequence-level data augmentation, which means the random parameters are kept the same for frames from the same training sequence. Each training sample is sampled from a video sequence with a frame length of 2 and a random frame interval sampled from $\{1, 2, 3, 4\}$.

Table S2. Hyper-parameters used in the multiple object tracking.

Dataset	MOT17
Batch size	16
Epochs	10
Warmup epochs	1
Learning rate	2.5e-5
Learning rate schedule	cosine
Optimizer	AdamW
Weight decay	0.05
RandomFlip	0.5
Mosaic	✓
RandomAffine	✓
RandomHSV	✓
Mixup	✓

B. Additional Experiments

we perform a number of ablation experiments on the MOT17 half-validation set to substantiate the rationale for certain design decisions

In order to investigate the necessity of a ResNet block in S-ViT for cross-window spatial information propagation, we first conduct an examination. The results in Tab. S3 demonstrate that when the ResNet blocks in S-ViT are removed, a significant reduction in tracking performance is observed across all scores, providing evidence for the effectiveness of ResNet blocks in this context.

First, we investigate the necessity of the ResNet block in S-ViT for cross-window spatial information propagation. As shown in Tab. S3, when we remove the ResNet blocks in S-ViT, a significant performance loss across all tracking scores is observed, demonstrating the effectiveness of using ResNet blocks.

Second, we evaluate S-ViT models with various pre-training strategies. Our default model uses CLIP pre-training. Additionally, we test ImageNet 21K pre-training and ImageNet 1K pre-training, as well as training from scratch. The comparison results are shown in Tab.S4. When

[*] This work was done during the internship of Yucheng at MSRA.
[†] Corresponding author.

Table S3. Analysis of the S-ViT architecture. Introducing the ResNet block is crucial to acquiring good performance on MOT. The results are from the MOT17 half-validation set.

Method	MOTA ↑	IDF1 ↑	HOTA ↑
w/ ResNet block	79.6	80.9	68.3
w/o ResNet block	78.6	77.9	66.4

using no pre-training or the weak ImageNet 1K pre-training, our model does not achieve high tracking scores. While a strong pre-trained model like CLIP or ImageNet 21K is utilized, our model could achieve top performance. Another possible pre-training strategy is to use a detection-trained model, which is a more commonly adopted approach by other MOT methods. However, as there is no available out-of-the-box detection checkpoint for our model, we plan to explore this direction in future work.

Table S4. Analysis of the pre-trained model. We use the CLIP pre-trained model in the default setting, which is proven to perform best. The results are from the MOT17 half-validation set.

Pre-trained Model	MOTA ↑	IDF1 ↑	HOTA ↑
CLIP (Default)	79.6	80.9	68.3
ImageNet 21K	79.4	79.7	67.7
ImageNet 1K	76.1	75.3	63.7
No pre-train	76.1	76.9	65.0

Third, we investigate an important hyper-parameter, namely the detection score threshold. As the association method we used in S-ViT does not require any training procedure, the selection of an appropriate hyper-parameter is essential for achieving high performance. Through experiments (see Tab.S5), we observed that a relatively high detection score threshold is necessary, as adopting the same value of 0.6 as ByteTrack results in a significant loss in performance. This observation can be attributed to the distinct detection frameworks utilized by ByteTrack and S-ViT, which tend to generate varying levels of confidence for detection boxes.

Table S5. Analysis of the detection score threshold for S-ViT. Due to the usage of different detection frameworks, S-ViT needs to tune a separate detection threshold to work well. The results are from the MOT17 half-validation set.

Detection Threshold	MOTA ↑	IDF1 ↑	HOTA ↑
0.75	78.9	79.6	67.6
0.7 (S-ViT)	79.6	80.9	68.3
0.65	79.6	79.3	67.2
0.6 (ByteTrack)	78.7	78.0	66.2

References

[1] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *ECCV (22)*, volume 13682 of *Lecture Notes in Computer Science*, pages 1–21. Springer, 2022. 1

[2] Yucheng Zhao, Chong Luo, Chuanxin Tang, Dongdong Chen, Noel C Codella, Lu Yuan, and Zheng-Jun Zha. T2d: Spatiotemporal feature learning based on triple 2d decomposition, 2023. 1