Supplementary Material: Long Movie Clip Classification with State-Space Video Models

Md Mohaiminul Islam, Gedas Bertasius

UNC Chapel Hill

Our supplementary material consists of:

1. Implementation Details.
2. Detailed Quantitative Results.
3. Additional Qualitative Results.

1 Implementation Details

1.1 ViS4mer Architecture

Our ViS4mer model consists of (i) a Transformer Encoder, and (ii) a Multi-scale Temporal S4 decoder. For our Transformer Encoder design, we follow ViT-L [2]. In particular, we use a transformer architecture consisting of 24 blocks with hidden dimension 1024. Each block contains a multi-headed self-attention layer with 16 heads and a Multi-Layer Perceptron.

For the Multi-scale Temporal S4 Decoder, we use a 3 block architecture. Each block contains an S4 layer, where the first S4 layer operates on a hidden dimension of 1024, and each subsequent block reduces the hidden dimension by a factor of 2 using a Multi-Layer Perceptron. Moreover, each block contains a Pooling layer with kernel size and stride of dimension $1 \times 2 \times 2$ which reduces the height and width dimension by a factor of 2. We do not change the temporal dimension throughout the network. Table 1 shows different layers and the input-output dimensions of each layer of the proposed ViS4mer model.

1.2 Training and Inference

We resize each video frame to a spatial resolution of 224×224 and divide the frames using a patch size of 16×16. We train our model for 100 epochs with an initial learning rate of 10^{-3}, which is divided by a factor of 5 when validation loss begins to plateau. We use Adam optimizer [4] with a weight decay of 0.01. We initialize the Transformer Encoder with ImageNet [1] pretraining weights and freeze the encoder network. For the LVU dataset [8], we use a frame rate of 1 fps and clip length of 60 seconds. During inference, we sample multiple clips and aggregate the predictions, following the exact sampling strategy of [8]. For the Breakfast [5] and the COIN [6] datasets, we use a frame rate of 15 fps and the standard splits [3, 7]. Following [3], we uniformly sample 64 segments to construct an input video, where each segment consists of 8 frames. During inference, we sample 64 equidistant segments from the whole video to construct a video and make predictions only using one video clip.

Table 1: The proposed architecture of the ViS4mer model.

(a) Transformer Encoder.

Stages	Operator	Output sizes
Input		$T \times H \times W \times 3$
Patchify	Patch Linear	$T \times \frac{H}{16} \times \frac{W}{16} \times D$
Blocks 1-L	$\begin{bmatrix} \text{SA} \\ \text{MLP} \end{bmatrix} \times L$	$T \times \frac{H}{16} \times \frac{W}{16} \times D$

(b) Mult-scale Temporal S4 Decoder.

Stages	Operator	Output sizes
Input		$T \times H' \times W' \times D$
Block 1	S4 Pooling MLP	$\frac{T}{s_T} \times \frac{H'}{2} \times \frac{W'}{2} \times \frac{D}{2}$
Block 2	S4 Pooling MLP	$\frac{T}{s_T} \times \frac{H'}{2} \times \frac{W'}{4} \times \frac{D}{4}$
Block 3	S4 Pooling MLP	$\frac{T}{s_T} \times \frac{H'}{2} \times \frac{W'}{8} \times \frac{D}{8}$

2 Detailed Quantitative Results

2.1 Ablating the Number of Spatiotemporal Tokens

We compare ViS4mer and the Long Sequence Transformer (LST) models when varying the number of input tokens on the LVU benchmark. We use video clips of 60 seconds and a frame rate of 1. In our main methods, we use a patch size of 16×16 and an image size of 224×224, which produces $60 \times 14 \times 14 = 11760$ tokens for an input video clip. We also experiment with a smaller number of input tokens by pooling the neighboring tokens before feeding through our model. Particularly, we experimented with 2940, and 1500 tokens. Finally, we use only the CLS token from each video frame which yields 60 input tokens for each video clip. In summary, in all of these experiments, we vary the number of input tokens to be 60, 1500, 2940, and 11760.

Table 2 shows the results of such analysis. We observe that increasing the number of spatial tokens increases the performance for both LST and the ViS4mer models. This indicates the necessity of using fine-grained spatiotemporal tokens for these complex movie understanding tasks. However, ViS4mer outperforms the LST by a significant margin while operating on a large number of tokens. This shows the effectiveness of ViS4mer over LST for the LVU benchmark. Note that this experimental setup is similar to Section 5.2 Figure 3 of the main paper, where we present the average performance of the content understanding, metadata prediction, and user engagement tasks. On the contrary, here we present detailed results of all tasks.

2.2 Ablating Temporal Support

In Table 3, we compare the performance of the Long Sequence Transformer (LST) and the ViS4mer while operating on video clips of different duration.

Table 2: Performace of the Object Transformer, Long Sequence Transformer, and the ViS4mer model on the LVU benchmark while operating on a different number of spatiotemporal tokens. Increasing the number of spatiotemporal tokens increases performance, and the ViS4mer outperforms both the Object transformer and the LST models.

	Number of	Content (↑)			Metadata (↑)				User (↓)	
Model	Tokens	Relation	Speak	Scene	Director	Genre	Writer	Year	Like	Views
Obj. Trans.	512	53.10	39.40	49.80	51.20	54.60	34.50	39.10	**0.23**	**3.55**
LST	60	47.61	34.32	47.08	49.53	49.45	40.47	39.16	0.33	4.09
	1500	47.61	35.32	62.34	51.40	49.12	41.66	39.86	0.33	4.01
	2940	50.00	36.31	62.57	53.27	48.64	41.66	40.55	0.32	3.94
	11760	52.38	37.31	62.79	56.07	52.7	42.26	39.16	0.31	3.83
ViS4mer	60	52.38	34.32	61.62	52.33	51.93	42.26	39.86	0.31	3.83
	1500	54.76	35.82	65.11	54.20	52.77	43.45	41.25	0.31	3.77
	2940	54.76	39.30	66.27	55.14	**54.71**	45.83	41.95	0.28	3.78
	11760	**57.14**	**40.79**	**67.44**	**62.61**	**54.71**	**48.80**	**44.75**	0.26	3.63

Particularly, we experiment with video durations of 1s, 20s, 40s, and 60s. We observe that increasing the temporal extent increases the performance for most of the tasks except for the Scene Prediction tasks. We hypothesize that the scene prediction task does not require very long-range temporal reasoning. LST achieves better performance than the ViS4mer when applied on a shorter clip (e.g., 1s), however, ViS4mer archives much better performance when we increase the clip duration. This experiment suggests that ViS4mer has a better long-term modeling ability than the LST baseline. This experiment is similar to Section 5.2 Figure 4 in the main paper, where we plot the performance on the Writer Prediction, Year Prediction, and Speaking Style Prediction tasks as a function of the temporal extent. Here we present the performance of the LST, and the ViS4mer models on all tasks of the LVU benchmark in Table 3 using different input video lengths.

3 Additional Qualitative Results

We present additional qualitative results on the LVU benchmark in Figure 1 and Figure 2. Specifically, we present several instances of correct and incorrect predictions made by our ViS4mer model. Furthermore, we also illustrate our qualitative results on the Breakfast (Figure 3) and the COIN (Figure 4) datasets.

3.1 Qualitative Results on the LVU Benchmark

Figure 1 shows some examples predictions of the ViS4mer on the content understanding tasks. We see that ViS4mer successfully captures the relationship among the characters, way of speaking, and the scene/place in Figures 1(a), (c), (e). We also observe that ViS4mer produces incorrect predictions on several difficult examples illustrated in Figures 1(b), (d), (f). Specifically, in Figure (b),

Table 3: Comparison between Object Transformer, Long Sequence Transformer, and ViS4mer on the LVU benchmark while varying the duration of the input video. Though LST achieves better performance on short video clips (e.g., 1s), ViS4mer excels on much longer clips indicating its effectiveness at long-range temporal reasoning. Both of these methods outperform the Object Transformer.

	Temporal	Content (↑)			Metadata (↑)				User (↓)	
Model	Extent	Relation	Speak	Scene	Director	Genre	Writer	Year	Like	Views
Obj. Trans.	60s	53.10	39.40	49.80	51.20	54.60	34.50	39.10	**0.23**	**3.55**
LST	1s	47.61	28.35	61.62	44.85	47.91	30.35	33.56	0.36	3.95
	20s	47.61	32.33	63.95	46.72	48.37	33.83	36.36	0.32	3.84
	40s	50.00	33.83	62.79	51.40	51.15	36.90	39.86	0.32	3.87
	60s	52.38	37.31	62.79	56.07	52.70	42.26	39.16	0.31	3.83
ViS4mer	1s	45.23	25.37	63.95	42.05	44.82	22.61	29.37	0.41	4.03
	20s	47.61	31.34	**69.76**	47.66	46.67	40.47	36.36	0.35	3.82
	40s	52.38	36.81	68.60	55.14	50.54	45.83	41.25	0.29	3.78
	60s	**57.14**	**40.79**	67.44	**62.61**	**54.71**	**48.80**	**44.75**	0.26	3.63

ViS4mer predicts the relationship among the character to be 'boyfriend_girlfirned', which is difficult to distinguish from the ground truth label 'husband_wife'. Moreover, in figure 1(d), it is very difficult to determine whether the characters are discussing something or explaining something. Similarly, in Figure 1(f), we see a police officer and man holding a gun which our model interprets as a scene in 'prison', whereas the ground truth label is 'office'. These examples suggest that the LVU benchmark is very challenging and ViS4mer yields qualitatively reasonable predictions in most cases.

We show several examples of correct and incorrect predictions made by the ViS4mer on metadata prediction tasks in Figure 2. ViS4mer can successfully recognize the genre, director, writer, and year of the movie in Figure 2(a), (c), (e), (g). Moreover, we present some incorrect predictions in Figures 2(b), (d), (f), (h). We observe that it is quite difficult to predict the director or the writer just by looking at a video clip even for a human which illustrates the challenges of the LVU benchmark. Moreover, ViS4mer predicts the genre of the movie of Figure (b) to be 'Romance' whereas the ground truth label is 'Comedy'. This is a reasonable prediction considering that the video clip is from the movie named 'Nick and Norah's Infinite Playlist, which is a romantic comedy movie.[1] Furthermore, ViS4mer predicts the year of the movie of Figure 2(f) to be '1940', which is close to the ground truth label '1930'.

3.2 Qualitative Results on the Breakfast Dataset

In Figure 3, we visualize several samples of the Breakfast dataset and the corresponding predictions made by the ViS4mer model. ViS4mer correctly identifies the procedural activity of 'making salad', 'making juice', and 'making sandwich' in Figures 3(a), (b), and (c). However, in some cases, we observe that ViS4mer

[1] `https://en.wikipedia.org/wiki/Nick_%26_Norah%27s_Infinite_Playlist`

fails to capture subtle differences among cooking activities like 'making coffee' *vs.* 'making tea' (Figure 3(d)), and 'making scrambled egg' *vs.* 'making fried egg' (Figure 3(e)).

3.3 Qualitative Results on the COIN Dataset

Figure 4 illustrates several examples of the long-range procedural activity of the COIN dataset and the corresponding predictions made by the ViS4mer model. The proposed ViS4mer effectively recognizes a diverse set of procedural activities of the COIN dataset, for example, 'Change Bike Chain', 'Practice Karate', and 'Plant Tree' (Figure 4(a), (b), (c)). Furthermore, we show two incorrect predictions made by the ViS4mer in Figure 4(d) and (e). Note that in many of these cases, the predicted class labels are similar to the ground truth labels (e.g., Figure 4(d) and (e)).

References

1. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
2. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
3. Hussein, N., Gavves, E., Smeulders, A.W.: Videograph: Recognizing minutes-long human activities in videos. arXiv preprint arXiv:1905.05143 (2019)
4. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
5. Kuehne, H., Arslan, A., Serre, T.: The language of actions: Recovering the syntax and semantics of goal-directed human activities. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 780–787 (2014)
6. Tang, Y., Ding, D., Rao, Y., Zheng, Y., Zhang, D., Zhao, L., Lu, J., Zhou, J.: Coin: A large-scale dataset for comprehensive instructional video analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1207–1216 (2019)
7. Tang, Y., Lu, J., Zhou, J.: Comprehensive instructional video analysis: The coin dataset and performance evaluation. IEEE transactions on pattern analysis and machine intelligence **43**(9), 3138–3153 (2020)
8. Wu, C.Y., Krahenbuhl, P.: Towards long-form video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1884–1894 (2021)

Fig. 1: Qualitative results of the ViS4mer model on the content understanding tasks. ViS4mer can successfully recognize the relationship, way of speaking, and scene/place in Figure (a), (c), (e). In Figures (b) and (d), ViS4mer makes incorrect predictions; however, predictions are semantically close to the ground truth labels. Finally, we see a police officer and a man holding a gun in Figure (f), which makes our model predict the place to be 'prison' rather than the 'office'.

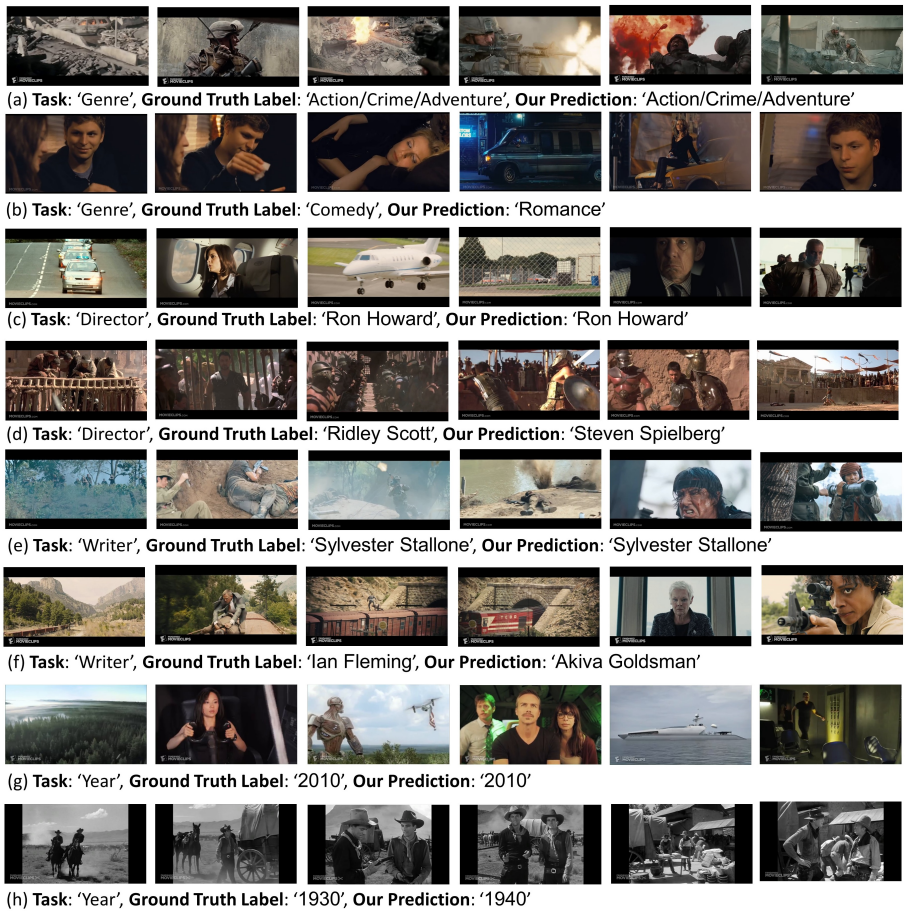

Fig. 2: Qualitative results of the ViS4mer model on the metadata prediction tasks. ViS4mer can successfully classify the genre, director, writer, and year of the movie in Figure 2(a), (c), (e), (g). In Figure (b) ViS4mer predicts the genre of the movie to be 'Romance' which is reasonable considering the movie is a romantic-comedy movie. Predicting the director and the writer of the movie from a video clip might be very difficult, and ViS4mer makes wrong predictions in Figures (d) and (f). Lastly, the predicted year '1940' of the movie is close to the actual label '1930' in Figure (h).

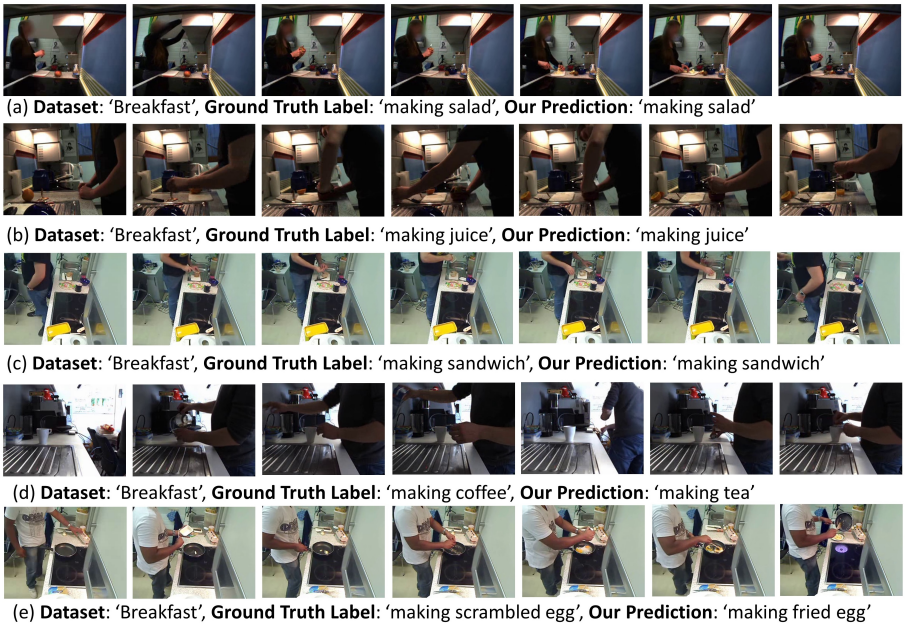

Fig. 3: Qualitative results of ViS4mer on the Breakfast dataset. ViS4mer can correctly identify the procedural activity in Figures (a), (b), (c). However, we also note that ViS4mer produces incorrect predictions in Figures (d) and (e), although the predicted classes are quite similar to the ground truth labels.

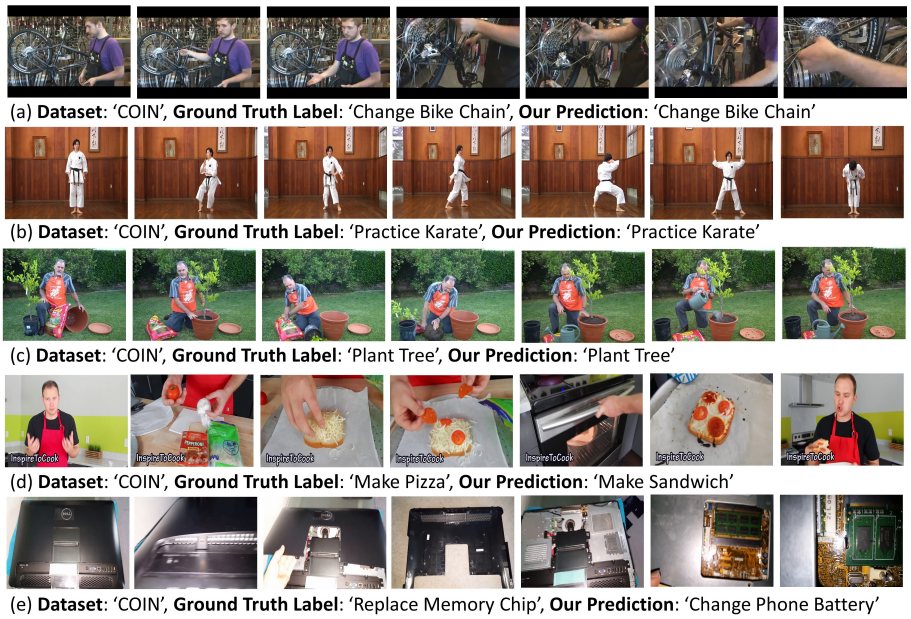

Fig. 4: Example predictions of ViS4mer on the COIN dataset. COIN dataset contains a diverse range of procedural activities, and ViS4mer can effectively recognize the activities in Figures (a), (b), (c). However, it fails in Figures (d) and (e).