Text-Guided Video Masked Autoencoder Supplementary Material

David Fan[1], Jue Wang[1], Shuai Liao[2], Zhikang Zhang[1], Vimal Bhat[1], and Xinyu Li[1]

[1] Amazon Prime Video
[2] Amazon Fulfillment Technology

1 Additional Results

1.1 More Epochs

In Table 1 we display the results with more epochs of pretraining on unlabeled Kinetics-400. We see that both finetuning and linear performance both continue to improve with more epochs of pretraining, with linear evaluation achieving higher delta of improvement with longer pretraining. Performance does not saturate with longer pretraining.

2 Additional Visualizations

In Figure 1 we provide additional visualizations. The model both masks and attends to the most salient regions of video that correspond to the provided natural language description. Again, we emphasize that visualizations are not intended to provide a formal explanation for model behavior. Our intention is to provide additional insights into the model to complement our quantitative results.

3 Additional Hyperparameters

We mostly follow the same hyperparameters as [4]. Table 2 and Table 3 show the configurations for pretraining and finetuning.

		Top-1 Acc	
Mask	Epochs	FT	Linear
$\text{TGM}_{\text{p=0.6}}$	200	79.6	59.4
	400	80.2	62.4
	800	80.3	63.4

Table 1: Results with more epochs of pretraining on unlabeled K400. Results are finetuning (FT) and linear evaluation on K400.

config	SSv2	K400
optimizer	AdamW	
base learning rate†	1.5e-4	
weight decay	0.05	
optimizer momentum	$\beta_1, \beta_2 = 0.9, 0.95$	
batch size	512	
learning rate schedule	cosine decay [3]	
warmup epochs	40	
flip augmentation	no	yes
augmentation	MultiScaleCrop	

Table 2: Pretraining hyperparameters. † : we follow the linear LR scaling rule. $lr = base\_lr \times batch\_size/256$.

config	SSv2	K400
optimizer	AdamW	
base learning rate	5e-4	1e-3
weight decay	0.05	
optimizer momentum	$\beta_1, \beta_2 = 0.9, 0.999$	
layer-wise lr decay	0.75 [1]	0.75
batch size	384	
learning rate schedule	cosine decay	
repeated augmentation	2 [2]	2
warmup epochs	5	5
total epochs	30	75
flip augmentation	no	yes
drop path	0.1	0.1

Table 3: Finetuning hyperparameters.

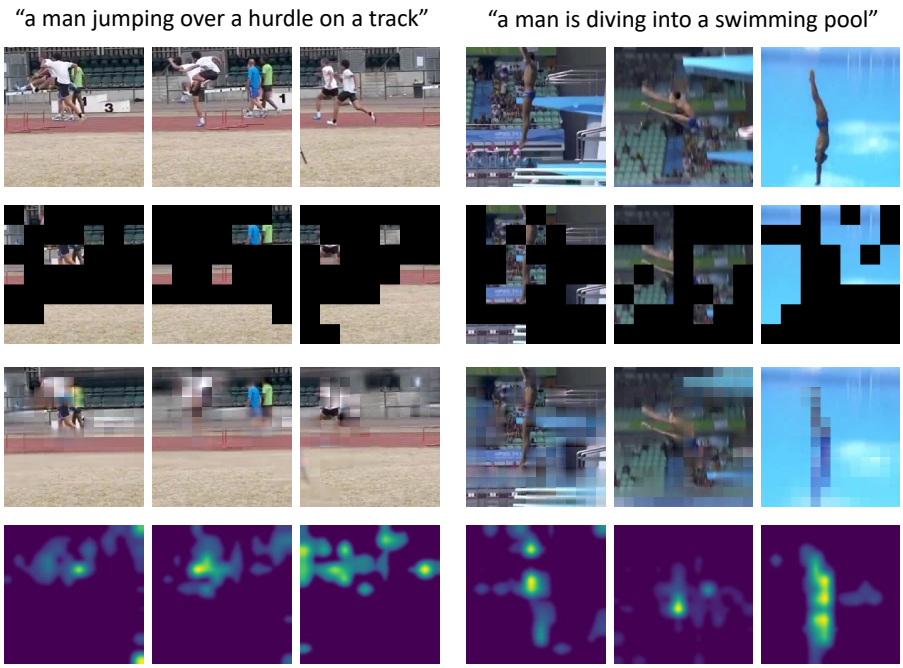

Fig. 1: Additional visualizations from three perspectives: the visualized mask, reconstructed RGB output, and encoder attention map. We see that our TGM solves the reconstruction task reasonably well and learns to attend to the salient regions of the video.

References

1. Bao, H., Dong, L., Piao, S., Wei, F.: Beit: Bert pre-training of image transformers. In: International Conference on Learning Representations (2021)
2. Hoffer, E., Ben-Nun, T., Hubara, I., Giladi, N., Hoefler, T., Soudry, D.: Augment your batch: Improving generalization through instance repetition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8129–8138 (2020)
3. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)
4. Tong, Z., Song, Y., Wang, J., Wang, L.: Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. ArXiv **abs/2203.12602** (2022)