Supplementary: Motion-Guided Masking for Spatiotemporal Representation Learning

David Fan[1] Jue Wang[1] Shuai Liao[2] Yi Zhu[3] Vimal Bhat[1]
Hector Santos-Villalobos[1] Rohith MV[1] Xinyu Li[1]
[1] Amazon Prime Video [2] Amazon Fulfillment Technology [3] AWS AI
{fandavi, juewangn, uliaoshu, yzaws, vimalb, hsantosv, kurohith, xxnl}@amazon.com

1. Additional Hyperparameters

We mostly follow the same hyperparameters as [4]. Suppl. Table 1 and Suppl. Table 2 show the configurations for pretraining and finetuning.

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

config	SSv2	K400	AVA	UCF101	HMDB51	Diving48
optimizer	AdamW					
base learning rate	5e-4	1e-3	2.5e-4	5e-4	5e-4	5e-4
weight decay	0.05					
optimizer momentum	$\beta_1, \beta_2 = 0.9, 0.999$					
layer-wise lr decay	0.75 [1]	0.75	0.7	0.75	0.7	0.7
batch size	128					
learning rate schedule	cosine decay					
repeated augmentation	2 [2]	2	1	2	2	2
warmup epochs	5	5	5	0	0	0
total epochs	30	75	30	100	100	100
flip augmentation	no	yes	yes	yes	yes	yes
drop path	0.1	0.1	0.2	0.2	0.2	0.2

Table 2: Finetuning hyperparameters.

2. Additional Visualizations

In Suppl. Figure 1 we show more attention visualizations for our MGM overlaid on top of the RGB frames. MGM seems to attend mostly to the salient video regions. In Suppl. Figure 2 we visualize more masks and RGB reconstructions for MGM and VideoMAE. Despite MGM being forced to solve a more challenging reconstruction task, it is able to achieve similar if not better reconstruction quality than VideoMAE. 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.

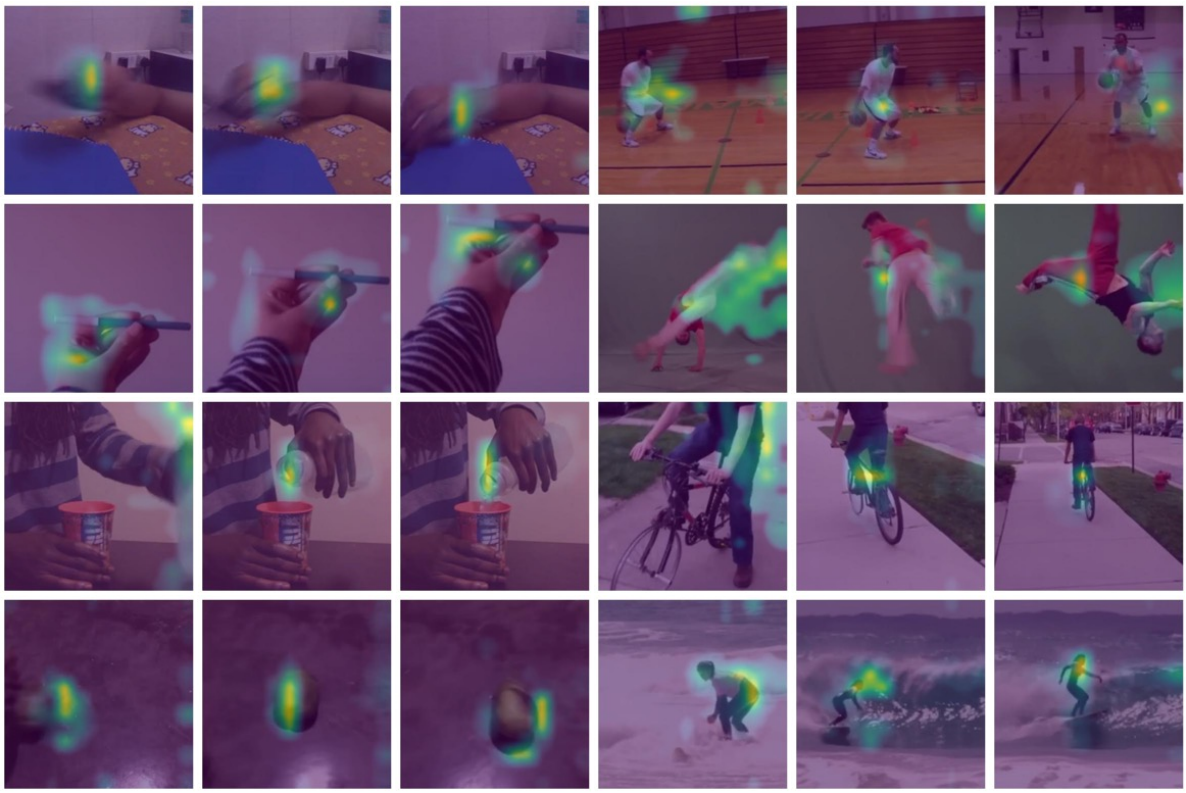

Figure 1: Encoder attention visualization overlaid to RGB frames where the query is the center patch of the center frame. Our MGM attends mostly to the salient regions of motion across frames.

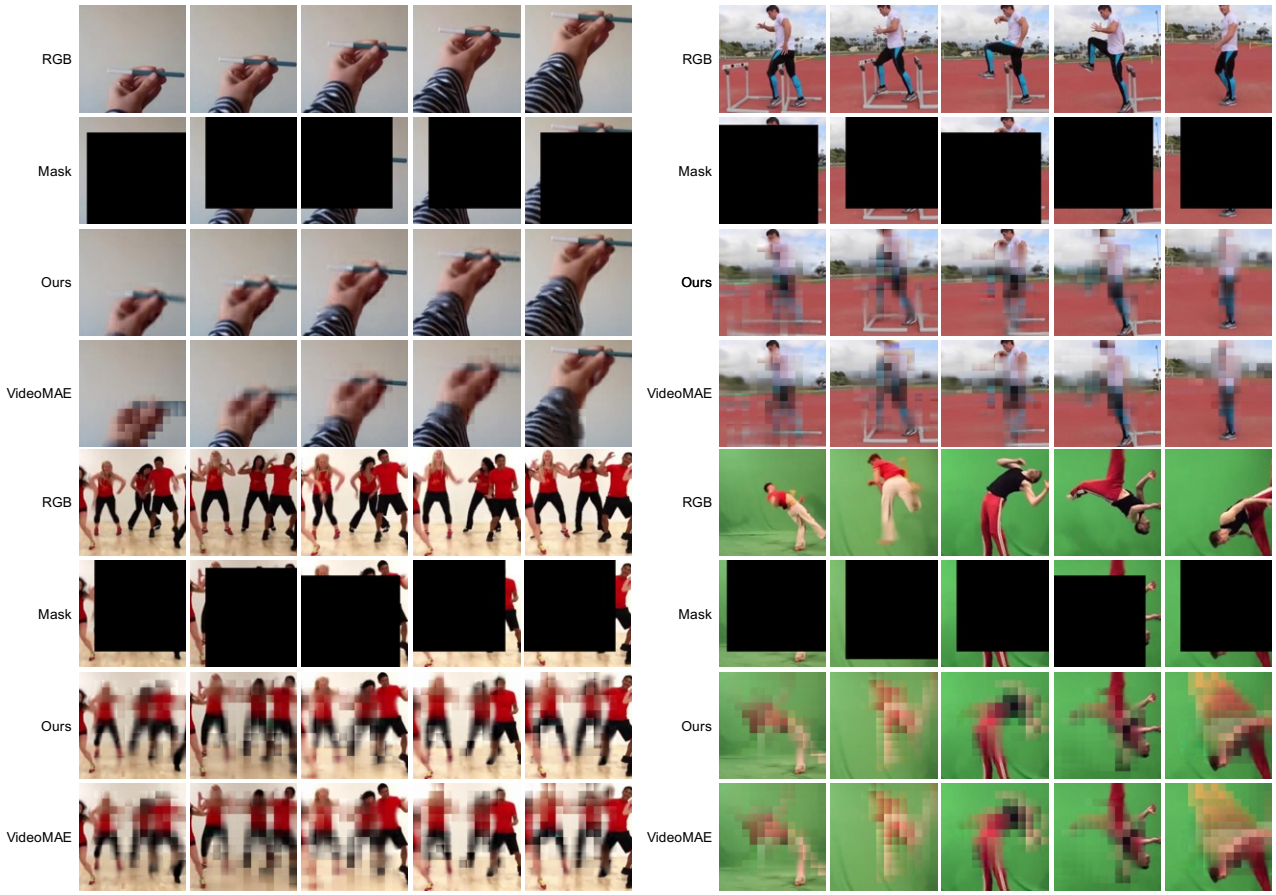

Figure 2: Our MGM achieves similar if not better reconstruction quality as VideoMAE despite using motion-guided masking which makes the reconstruction task more difficult. The masked regions of videos are most informative and therefore cannot be easily inferred from non-masked regions. MGM is forced to learn spatiotemporal semantics throughout the video to reconstruct the spatiotemporally continuous motion-guided masked regions.

References

[1] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. In *International Conference on Learning Representations*, 2021. 1

[2] Elad Hoffer, Tal Ben-Nun, Itay Hubara, Niv Giladi, Torsten Hoefler, and Daniel Soudry. Augment your batch: Improving generalization through instance repetition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8129–8138, 2020. 1

[3] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 1

[4] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *ArXiv*, abs/2203.12602, 2022. 1