A Implementation Details

Training. Our training settings follow [24] and we build on the open-source implementation of MAEs (https://github.com/facebookresearch/mae) for all our experiments. We use the parameters specified in the original implementation unless specified otherwise in Table 4a. All our experiments are performed on 4 Nvidia Titan RTX GPUs for ViT-S/16 models, and on 8 Nvidia Titan RTX GPUs for ViT-S/8 models.

Evaluation methodology. Our evaluation methodology follows prior work [14–16] and in Table 1 we report results previously reported in these studies. For recent self-supervised learning approaches like DINO, MAEs, MAE-ST and VideoMAE, we carry out a comprehensive grid search on the evaluation hyperparameters listed in Table 4b, and report the optimal results obtained. The evaluation parameters for SiamMAE can be found in Table 4b.

config	value
optimizer	AdamW [100]
optimizer momentum	$\beta_1, \beta_2{=}0.9, 0.95$ [103]
weight decay	0.05
learning rate	1.5e-4
learning rate schedule	cosine decay [104]
warmup epochs [105]	40
epochs	2000 (ablations 400)
repeated sampling [99]	2
augmentation	hflip, crop $[0.5, 1]$
batch size	2048
frame sampling gap	$[4, 48]$

(a) **Kinetics pre-training setting.**

config	DAVIS	VIP	JHMDB
top-k	7	10	7
queue length	20	20	20
neighborhood size	20	8	20

(b) **Evaluation setting.**

Table 4: Training and evaluation hyperparameters.