A Implementation Details

Kinetics-400. Our Kinetics-400 dataset contains 240,436 training videos and 19,787 validation videos. We use 224 spatial input size in all experiments. We sample evenly strided frames for Kinetics-400, and use a stride of 16, 16, 8 for the 8-, 16-, 32-frame model variants, respectively. We use *RandomResizedCrop*, *RandomHorizontalFlip* and *RandAugment* (as implemented in `https://github.com/facebookresearch/SlowFast`) for data augmentation and apply a 0.5 dropout rate in each trainable MLP block and before the final classification head. All models are trained using a batch size of 256 for 50,000 steps with AdamW optimizer. We use a half-period cosine learning rate schedule with initial value of 4×10^{-4} and constant weight decay of 0.05. For testing, we resize the short size of videos to 224 and use 3 temporal crops and the center spatial crop.

Something-Something-v2. Training on Something-something v2 is similar to Kinetics-400, except for the following differences. We use TSN-style sampling for Something-Something-v2, i.e., we divide the video evenly into n segments and select one frame from each – A random frame from each segment is sampled during training and the center frame is used during evaluation. 3 spatial crops are used for testing. We also train for a shorter 30,000 steps on Something-something v2. We *do not* use Kinetics-400 pretraining to initialize models for Something-Something-v2, as we have found the accuracy difference negligible.

Model Details. By default we use ViT-B/16 with CLIP pretraining as the image backbone. We use decoder blocks with the same configuration as backbone encoder blocks. Unless otherwise specified, for Kinetics-400, we use 4 Transformer decoder blocks taking information from the last 4 blocks of the backbone as key and value. For Something-something v2, as we have found using deeper decoders helps model motion information, we use 12 (for ViT-B) or 24 (for ViT-L) Transformer decoder blocks, taking information from all Transformer encoder blocks in the CLIP backbone.

Full-finetuning Details. For TimeSformer experiments, we use a training configuration similar to their original implementation, except that training epochs are set to 20 and a 100x learning rate reduction on backbone weights is applied for CLIP-related experiments, as we found these changes lead to higher accuracy. For full-finetuning with our own architecture, we also use a 100x learning rate reduction on backbone weights, and all other training configuration remains the same.