Model	Pre-training Supervision	Pre-training Dataset	Acc (%)
TSN (RGB+Flow) [26]	Supervised: action labels	Kinetics	36.5*
S3D [16]	Unsupervised: MIL-NCE on ASR	HT100M	37.5*
ClipBERT [12]	Supervised: captions	COCO + Visual Genome	30.8
VideoCLIP [28]	Unsupervised: NCE on ASR	HT100M	39.4
SlowFast [10]	Supervised: action labels	Kinetics	32.9
TimeSformer [4]	Supervised: action labels	Kinetics	48.3
LwDS: TimeSformer [4]	Unsupervised: k-means on ASR	HT100M	46.5
LwDS: TimeSformer	Distant supervision	HT100M	54.1
VideoTF (SC)	Unsupervised: NN on ASR	HT100M	47.0
VideoTF (DM)	**Distant supervision**	HT100M	**54.8**
VideoTF (SC)	**Distant supervision**	HT100M	**56.5**

Table 1: **Step classification.** We compare to the accuracy scores for all baselines. VideoTF (SC) pre-trained with step classification loss on distant supervision from WikiHow achieves state-of-the-art performance on the downstream step classification task. We report baseline results from [13]. * indicates results by fine-tuning on COIN

Downstream Model	Base Model	Pre-training Supervision	Pre-training Dataset	Acc (%)
TSN (RGB+Flow) [26]	Inception [24]	Supervised: action labels	Kinetics	73.4*
Transformer	S3D [16]	Unsupervised: MIL-NCE on ASR	HT100M	70.2*
Transformer	ClipBERT [12]	Supervised: captions	COCO + Visual Genome	65.4
Transformer	VideoCLIP [28]	Unsupervised: NCE on ASR	HT100M	72.5
Transformer	SlowFast [10]	Supervised: action labels	Kinetics	71.6
Transformer	TimeSformer [4]	Supervised: action labels	Kinetics	83.5
LwDS: Transformer	TimeSformer [4]	Unsupervised: k-means on ASR	HT100M	85.3
LwDS: Transformer	TimeSformer	Distant supervision	HT100M	88.9
LwDS: Transformer w/ KB Transfer	TimeSformer	Distant supervision	HT100M	90.0
VideoTF (SC; fine-tuning) w/ KB Transfer	TimeSformer	Unsupervised: NN on ASR	HT100M	81.2
VideoTF (SC; linear-probe) w/ KB Transfer	TimeSformer	Distant supervision	HT100M	83.1
VideoTF (DM; linear-probe) w/ KB Transfer	TimeSformer	Distant supervision	HT100M	85.7
VideoTF (SC) w/ KB Transfer	TimeSformer	Distant supervision	HT100M	90.5
VideoTF (DM) w/ KB Transfer	**TimeSformer**	Distant supervision	**HT100M**	**91.0**

Table 2: Accuracy of different methods on the **procedural activity recognition** dataset.

Downstream Model	Base Model	Pre-training Supervision	Pre-training Dataset	Acc (%)
Transformer	S3D [16]	Unsupervised: MIL-NCE on ASR	HT100M	28.1
Transformer	SlowFast [10]	Supervised: action labels	Kinetics	25.6
Transformer	TimeSformer [4]	Supervised: action labels	Kinetics	34.7
LwDS: Transformer	TimeSformer [4]	Unsupervised: k-means on ASR	HT100M	34.0
LwDS: Transformer w/ KB Transfer	TimeSformer	Distant supervision	HT100M	39.4
VideoTF (SC; fine-tuned) w/ KB Transfer	TimeSformer	Unsupervised: NN on ASR	HT100M	35.1
VideoTF (SC; linear-probe) w/ KB Transfer	TimeSformer	Distant supervision	HT100M	39.2
VideoTF (DM; linear-probe) w/ KB Transfer	TimeSformer	Distant supervision	HT100M	40.1
VideoTF (SC) w/ KB Transfer	TimeSformer	Distant supervision	HT100M	41.5
VideoTF (DM) w/ KB Transfer	**TimeSformer**	Distant supervision	**HT100M**	**42.4**

Table 3: Accuracy of different methods on the **short-term step forecasting** dataset.

Downstream Model	Base Model	Pre-training Supervision	Pre-training Dataset	Acc (%)
Transformer (ASR text) w/ Task label	MPNet			39.0
Transformer	SlowFast [10]	Supervised: action labels	Kinetics	15.2
Transformer	TimeSformer [4]	Supervised: action labels	HT100M	17.0
Transformer w/ Task label	TimeSformer [4]	Supervised: action labels	HT100M	40.1
LwDS: Transformer w/ Task label	TimeSformer	Distant supervision	HT100M	41.3
VideoTF (DM)	TimeSformer	Distant supervision	HT100M	40.2
VideoTF (DM) w/ Task label	**TimeSformer**	Distant supervision	**HT100M**	**46.4**

Table 4: Accuracy of different methods on the **long-term step forecasting** dataset.