Downstream Model	Base Model	Pre-training Supervision	Pre-training Dataset	Mistake Detection	
				Step	**Order**
Transformer (ASR text) w/ Task label	MPNet [22]			34.2	33.4
Transformer w/ Task Label	SlowFast [10]	Supervised: action labels	Kinetics	28.6	26.1
Transformer w/ Task label	TimeSformer [4]	Supervised: action labels	HT100M	36.0	34.7
LwDS: Transformer	TimeSformer	Distant supervision	HT100M	17.1	11.2
LwDS: Transformer w/ Task Label	TimeSformer	Distant supervision	HT100M	37.6	31.8
VideoTF (SC)	TimeSformer	Distant supervision	HT100M	20.1	15.4
VideoTF (DM) w/ Task label	TimeSformer	Distant supervision	HT100M	40.8	34.0
VideoTF (SC; fine-tuned) w/ Task label	**TimeSformer**	Distant supervision	**HT100M**	**41.7**	**35.4**

Table 5: Accuracy of different methods on the **mistake step detection** test dataset.

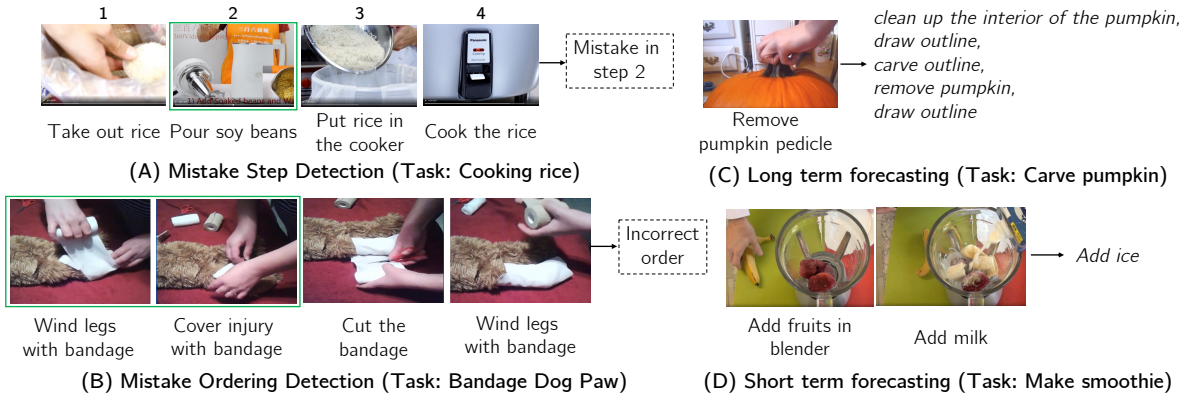

Figure 3: **Qualitative results.** We show qualitative results of our method on 4 tasks. The step labels are not used during training and are only shown here for illustrative purposes.

Figure 4: **Qualitative comparison.** We compare results from our method VideoTF to the baseline LwDS on the short-term forecasting task. Step labels are not passed to the model as input and are only for reference.

LwDS and our unsupervised pre-training using NN with ASR outperforms previous unsupervised methods. We also note that linear-probe performance is competitive in Tab. 2 and outperforms baselines in Tab. 3. VideoTF with achieves a strong improvement of 5% over LwDS on the long-term forecasting task, 4% on mistake step detection, and 4% on mistake ordering detection. Adding task labels improves performance on all three tasks.

Additionally, we evaluate our approach on the activity recognition task in EPIC Kitchens-100 and include results in the Supplemental. We also report our models performance on the step localization task in COIN.

Qualitative Results. Fig. 3 shows qualitative results of our model VideoTF on the mistake detection tasks. Fig. 3 (A) shows a result on mistake step detection, where our model's input is the sequence of video clips on the left and it correctly predicts the index of the mistake step "2" as the output. In (B), the order of the first two steps is swapped and our model classifies the sequence as incorrectly ordered. In (C), for the long-term forecasting task, the next 5 steps predicted by our model match the ground truth and in (D), for the short-term forecasting task, the model predicts the next step correctly given the past 2 steps. In Fig. 4 we show an example result of our method compared to the baseline LwDS on the short-term forecasting task. Our method correctly predicts the next step as "remove air nozzle" since it has acquired knowledge of task structure whereas the baseline predicts the next step incorrectly as "install valve cap."

6. Conclusion

In this work, we introduce a new video model, VideoTaskformer, for learning contextualized step representations through masked modeling of steps in instructional videos. We also introduce 3 new benchmarks: mistake step detection, mistake order detection, and long term forecasting. We demonstrate that VideoTaskformer improves performance on 6 downstream tasks, with particularly strong improve-