Procedure Planning in Instructional Videos
Supplementary Material

Chien-Yi Chang[0000−0001−6508−1161], De-An Huang[0000−0002−6945−7768], Danfei Xu[0000−0002−8744−3861], Ehsan Adeli[0000−0002−0579−7763], Li Fei-Fei[0000−0002−7481−0810], and Juan Carlos Niebles[0000−0001−8225−9793]

Stanford University

1 Dataset Curation Details

To train and test our model for the procedure planning problem, we curate a dataset of size N that takes the form

$$\{(o_t^i, a_t^i, ..., a_{t+T-1}^i, o_{t+T}^i)\}_{i=1}^N. \tag{1}$$

In the following we will drop the superscript i for simplicity. Each example is a sequence of observations o and actions a. Here we note $o \in \mathbb{R}^{M \times K}$ where M is the number of frames sampled for each timestep and K is the dimension of features, and actions a can be represented as one-hot vectors. Each sequence can be parsed to a list of triplets (o_t, a_t, o_{t+1}). We follow the tradition in [1, 3] where we view an action a_t as a transformation which changes the state of the environment before the action happens (precondition o_t) to the state after the action (effect o_{t+1}). For example, the precondition of whisking eggs is a bowl containing unmixed eggs and the effect is the eggs being whisked.

We build our procedure planning dataset from temporally annotated instructional videos. The videos have manually annotated temporal segmentation boundaries and action labels. Please note that since our method requires full supervision, we will leave how to utilize unlabeled instructional videos [2, 4] as future work. Specifically, we use the 2750 labeled videos in CrossTask [4], averaging 4.57 minutes in duration, for a total of 212 hours. Each video depicts one of the 18 complex long-horizon tasks like *Grill Steak*, *Make Pancakes*, or *Change a Tire*.

Concretely, consider a video V_i where i is the index of frames. It has been annotated with a sequence of action labels a_j and each action starts at frame index s_j and ends at frame index e_j. We can write the aforementioned precondition-action-effect triplet (o_t, a_t, o_{t+1}) as

$$o_t = V_{s_t - M/2 : s_t + M/2}, \tag{2}$$

$$a_t = a_t, \tag{3}$$

$$o_{t+1} = V_{e_t - M/2 : e_t + M/2}. \tag{4}$$

This process is illustrated in Figure 1. We can repeat this process for all a_j to obtain a sequence of observations and actions as described above. For a video

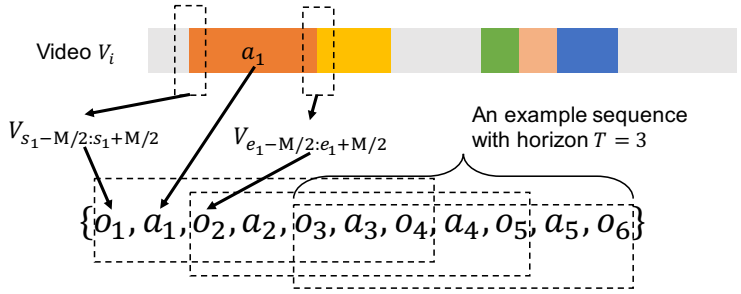

Fig. 1: Consider a video V_i where i is the index of frames. It has been annotated with a sequence of action labels a_j and each action starts at frame index s_j and ends at frame index e_j. We can curate example sequences with horizon T by using sliding window.

with n_{actions} actions, the length of resulting sequence will be $2n_{\text{actions}} + 1$. After processing all 2759 videos, we can derive (1) using sliding windows technique to sample sequences of fixed length $2T + 1$. For all data we have used in our paper, we set $M = 2$ and $K = 3200$.

We sample training and testing examples individually for different choices of planning horizon T. In Figure 2 we show the the number of examples by task for planning horizon $T = 3$ (top) and $T = 4$ (bottom). As the planning horizon increases, the total number of example decreases, making it difficult for our model to learn useful information from limited training data. This also explains why our model's performance decreases when T is large.

2 Additional Visualizations

In Figure 3 we show some examples where our model fails to plan actions that are exactly the same as the ground truth. In the changing tire example (top), our model is not able to capture the subtle visual cues that the tool is already in the man's hand, so there is no need to get the tools out. In the making French strawberry cake example (bottom), our model is not able to predict steps such as "Cut Strawberries", due to the partial observability of instructional videos, where the key instrument for cutting (knife) is not seen in neither the start nor the goal.

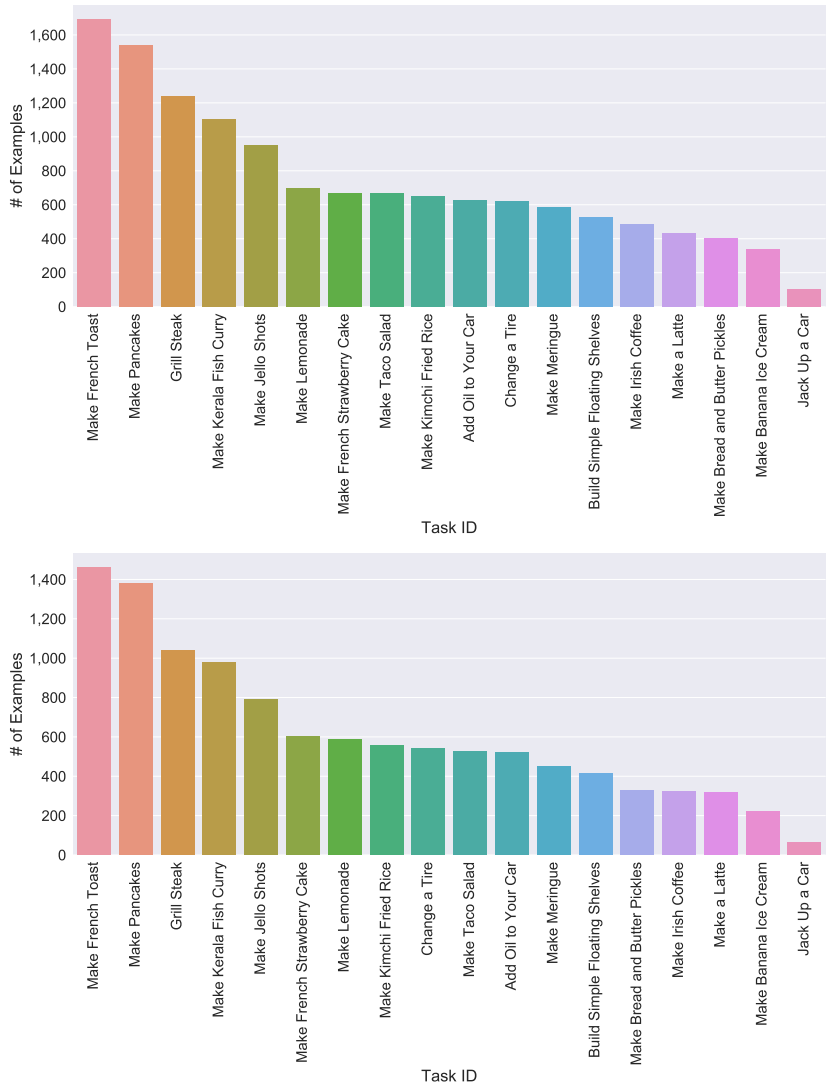

Fig. 2: The number of curated examples in each high-level task for $T = 3$ (top) and $T = 4$ (bottom). The number of examples decreases as the planning horizon T increases.

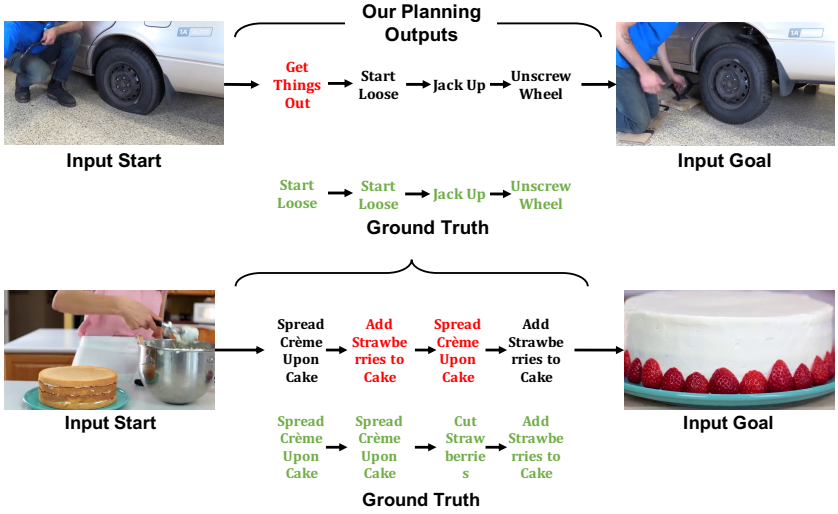

Fig. 3: Examples where our model fails to plan a sequence of actions that is exactly the same as the ground truth. Wrong step is in red. Ground truth step is in green. The results are nevertheless semantically reasonable.

References

1. Bellman, R.: A markovian decision process. Journal of mathematics and mechanics pp. 679–684 (1957)
2. Miech, A., Zhukov, D., Alayrac, J.B., Tapaswi, M., Laptev, I., Sivic, J.: Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2630–2640 (2019)
3. Wang, X., Farhadi, A., Gupta, A.: Actions˜ transformations. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 2658–2667 (2016)
4. Zhukov, D., Alayrac, J.B., Cinbis, R.G., Fouhey, D., Laptev, I., Sivic, J.: Cross-task weakly supervised learning from instructional videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3537–3545 (2019)