Supplementary: Hierarchical Video Prediction using Relational Layouts for Human-Object Interactions

Navaneeth Bodla[1] Gaurav Shrivastava[1] Rama Chellappa[2] Abhinav Shrivastava[1]
[1]University of Maryland, College Park [2]Johns Hopkins University

Contents

1. Additional Training Details

Creating object maps. The output of the relational layout generation phase (i.e, stage 1) is a sequence of objects and poses. The pose outputs are 2D maps of shape $128 \times 128 \times N_{kp}$ for each timestep. Whereas the object outputs are 1D boxes of shape 1×4 for each box at each timestep. After the first stage, the objects are converted into 2D maps before using them as inputs in the second stage for video generation. We do this by first initializing a tensor of shape $128 \times 128 \times d_o$ with zeros. Next, each object is mapped into this tensor such that the channel corresponding to its class is set to 1 in the region occupied by the bounding box. This operation is performed for all the objects to obtain the 2D mapping of the objects per time step.

Additional architecture details. The pose encoder is used to encode the pose before using it as an input to the RNN. It is a convolutional encoder with 8 initial filters. Filter size is doubled after every convolutional layer. The final layer is a fully connected layer with 64 output dimensions. Similarly, the box is encoded before using it as an input to the RNN. The box encoder is a two layer multilayer perceptron with 8 and 16 as output dimensions. The pose decoder is a convolutional decoder with 256 filters. The number of filters gets halved after every convolutional layer. For the second stage of video generation, we use a pix2pixHD architecture with 24 filters. For all the discriminators, we use spectral normalization and 64 filters in the first convolutional layer.

2. Datasets and Pre-processing

We used two datasets to evaluate our approach: 1) UMD-HOI [1] dataset and 2) Bimanual dataset [2]. Additional details about dataset collection and preprocessing are presented below.

UMD-HOI dataset. UMD-HOI dataset has a total of 64 videos with actions performed by ten subjects performing six interactions with four objects. We fully annotate this dataset by labeling the boxes around the objects in every frame. Note that we only annotate the object of interaction since the other objects are stationary in an action. To obtain the keypoints for every frame, we used posenet [3]. Since the dataset size is small, in order to avoid overfitting we do data augmentation. Firstly, the video is resized spatially to 140×140 and then randomly cropped to 128×128. Next, it is randomly flipped from left to right. Finally, the first frame of the video sample is randomly chosen from the first 12 frames (i.e, from the first 0.5 sec) and then the subsequent frames are uniformly sampled to a fixed sequence length of 16. Even though the original dataset size is small, we observe that these data augmentation techniques greatly help in training the video generator. Specifically, the random flipping from left to right helps in making it more robust to the hand (left or right) used for performing the action and the random crop helps in learning a better pose sequence generator. The dataset is divided into train and val splits. The train split contains 50 videos and the val split contains 14 videos. In order to show generalization to new subjects, one particular subject is held out of training. The training and validation videos are randomly chosen.

Bimanual dataset. The Bimanual dataset contains a total of 540 videos with actions performed by 6 subjects where 4 of them are male and 2 of them are female. It has an overall 12 objects and 9 tasks. The tasks in this dataset involve interactions with multiple objects over time. For example, the task of cooking with bowls involves picking up a whisk, mixing in the bowl, poring from another bowl, etc. The authors of the dataset have provided with the extracted bounding boxes using YOLO-v3 [4]. We selected 6 tasks and computed the tracklets for each object in the video us-

ing two heuristics: 1) IoU and 2) similarity score based on the color histogram. Data augmentation is done similarly to the UMD-HOI dataset. The dataset is divided into training and validation splits. The training has 300 videos and validation has 120. The data is split in a way that two random actions performed by every subject are held out from training and are included in the validation set.

3. Qualitative Results

3.1. Object sequence visualization

In Figure 1 we visualize the outputs of our object RNN on test samples with the model trained on Bimanual dataset. Under this setting, for every performer, the actions performed in the training set have no overlap with the actions performed in the test set. That is, a user will never perform the same action in train and test splits. We do this to show that our model can generalize well to new actions. For every example in Figure 1, the first row is the ground truth sequence of objects and the second row is the predict sequence.

We observe various properties from Figure 1. Firstly we see that the relative positions of various objects are very similar to that of ground truth sequences. Note that in all the examples multiple objects evolve over time such as in example 4 the green and blue objects come together and then the blue object goes up and comes down. This is an example of pouring from a bottle and drinking. Our generated sequence is faithfully able to predict the relative configurations of the boxes per frame and overtime. Also, note that our model does not simply mimic the ground truth. It learns to predict the sequence at different speeds. For example, in 1 and 2 we see that the speeds are more or less the same but in examples 3, 4 and 5 our model predicts actions at a slightly slower speed than the ground truth and hence the box sequence is shifted in time.

3.2. Failure Modes

Figure 2 shows a few failure modes of our model. The preprocessing method used to generate pose and object sequences can sometimes be noisy as shown in Figure 2(a). Our layout generator is successfully able to learn to recover from noisy labels and predict smooth pose sequences. Another well-known issue with video prediction problems is missing object permanence sometimes. In Figure 2(b), the predicted video contains two cups, one that stays stationary on the table and the other that is used to perform the action. This is a side effect of the optimization algorithm that tries to propagate most of the information from the guidance frame to all the frames and at the same time tries to perform the desired action with the object. Human beings are known to learn object permanence at a very young age, at about 4 to 7 months, and is a very important property to model. This is an interesting direction for future research.

3.3. Additional qualitative results on Bimanual and HoI datasets

We present additional qualitative results on Bimanual dataset in Figures 3, 4, 5, 6 and UMD-HOI dataset in Figures 7, 8. We also show a few example videos generated by our model in Figure 9.

Acknowledgements. This work was supported by DARPA SAIL-ON program via ARO contract no. W911NF2020009 and IARPA via contract no. D17PC00345. The views and conclusions are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

References

[1] Abhinav Gupta, Aniruddha Kembhavi, and Larry S Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(10):1775–1789, 2009. 1

[2] Christian R. G. Dreher, Mirko Wächter, and Tamim Asfour. Learning object-action relations from bimanual human demonstration using graph networks. *IEEE Robotics and Automation Letters (RA-L)*, 5(1):187–194, 2020. 1

[3] George Papandreou, Tyler Zhu, Liang-Chieh Chen, Spyros Gidaris, Jonathan Tompson, and Kevin Murphy. Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 269–286, 2018. 1

[4] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 1

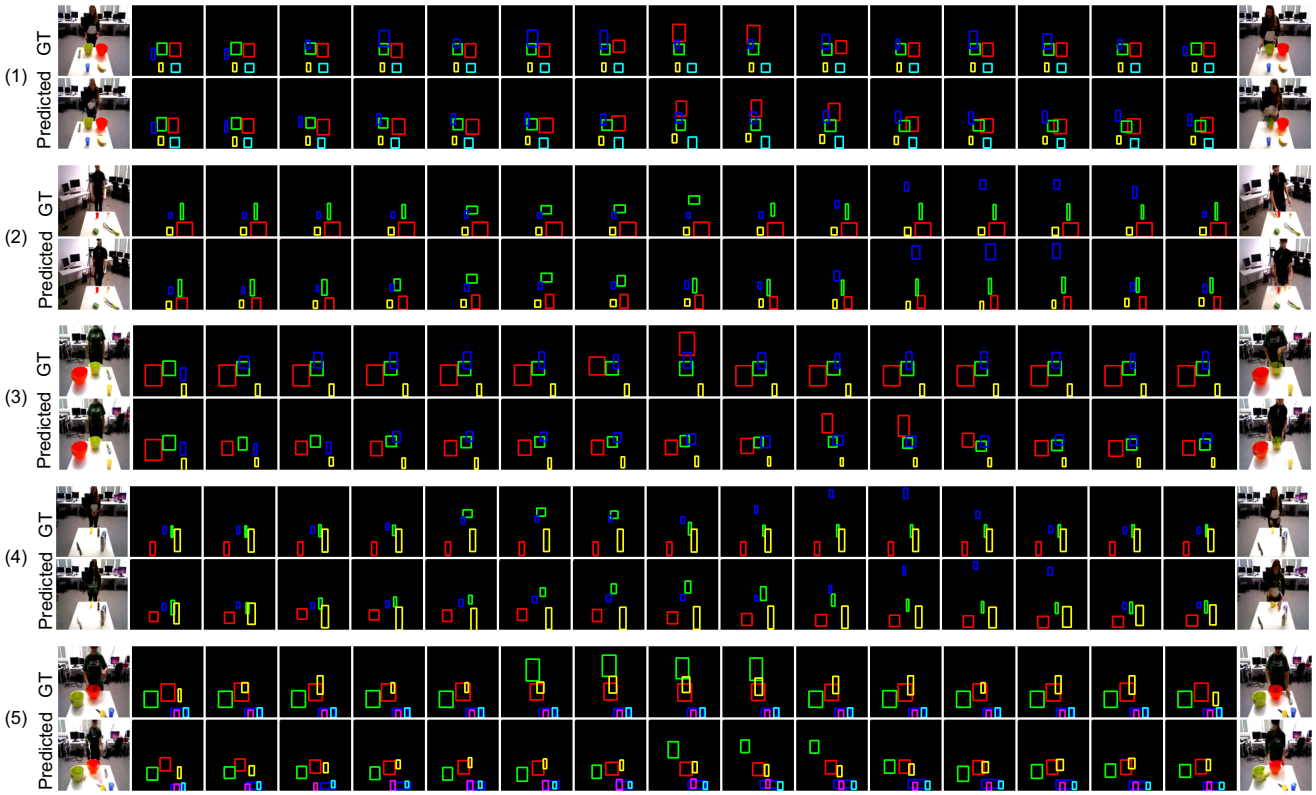

Figure 1: Output of object RNN as compared to ground truth sequences for tasks : cooking with bowls (1,3,5) and pouring and drinking (2,4)

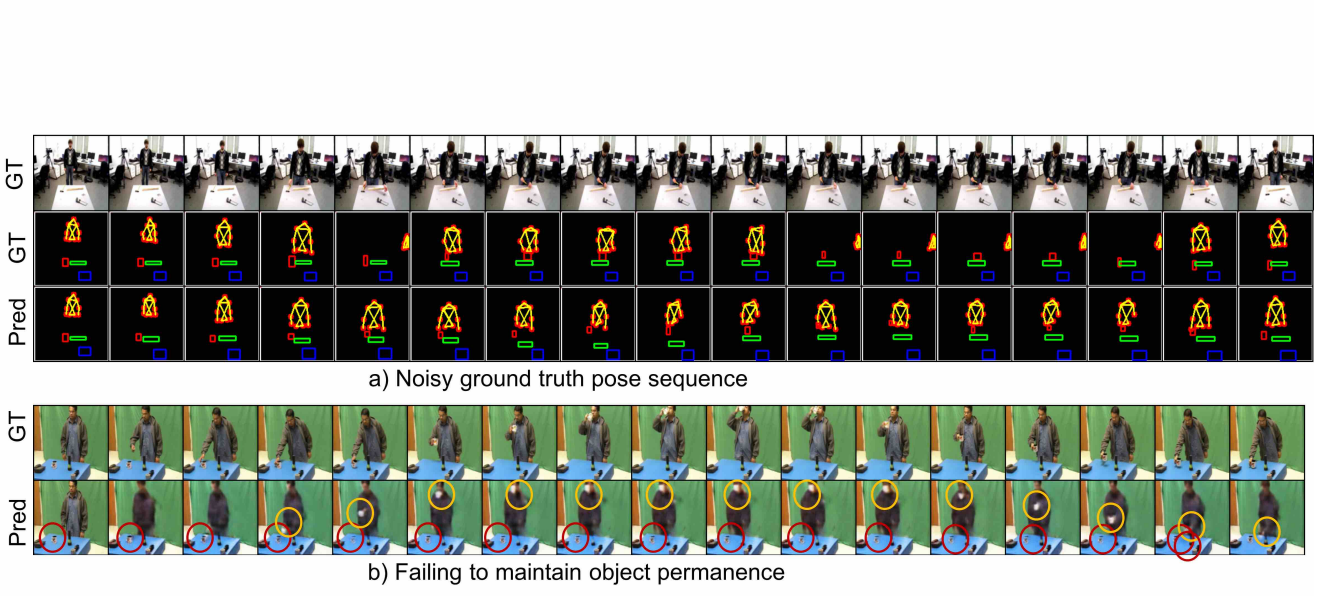

Figure 2: Failure Modes

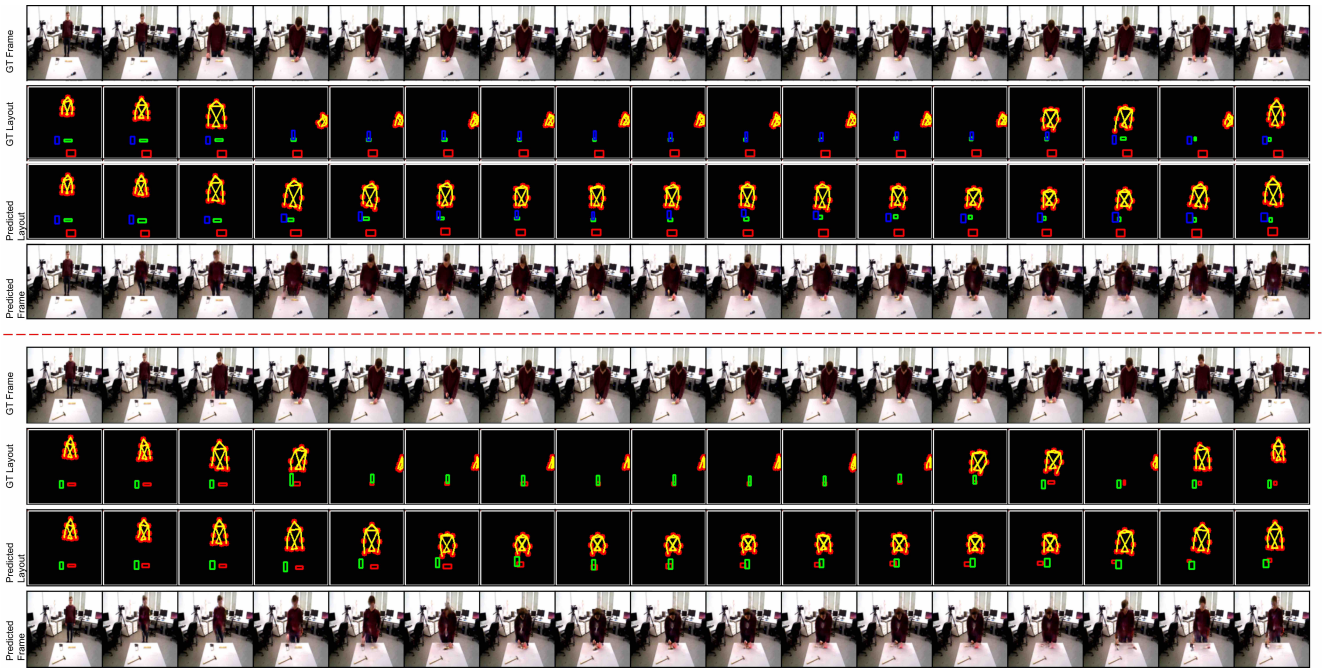

Figure 3: Example visualizations of the task “sawing” on Bimanual dataset.

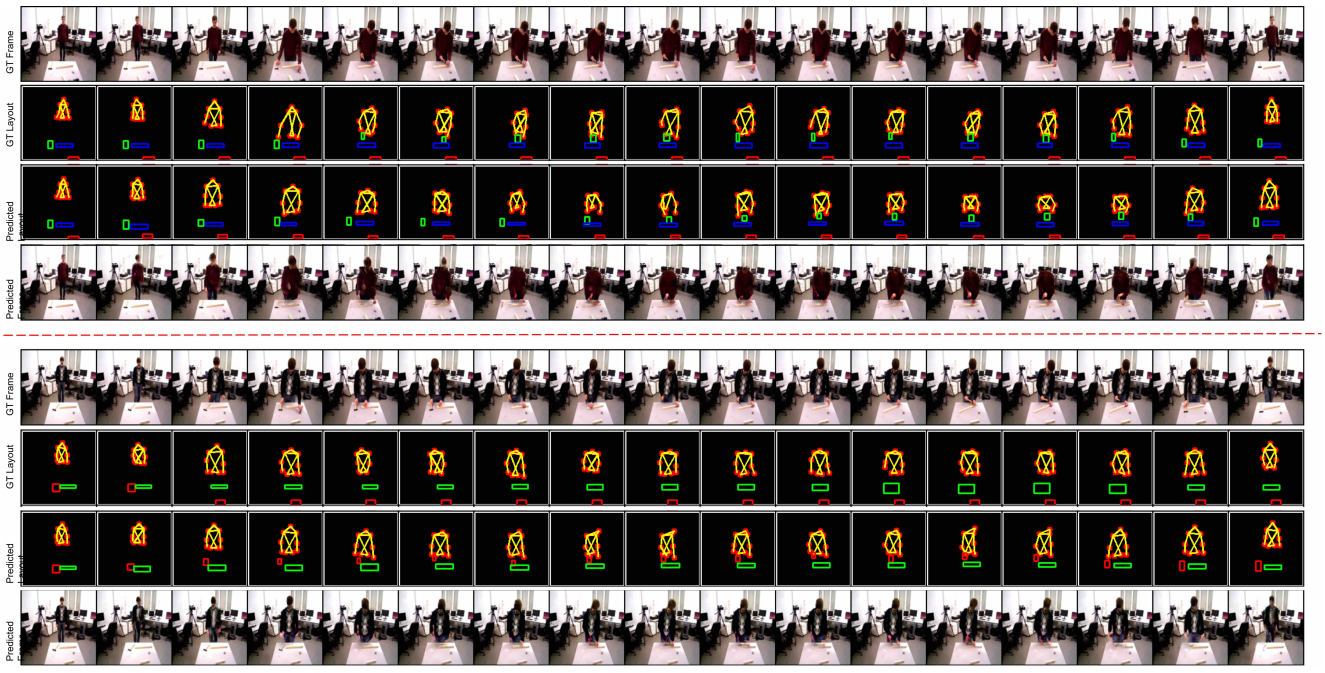

Figure 4: Example visualizations of the task “hammering” on Bimanual dataset.

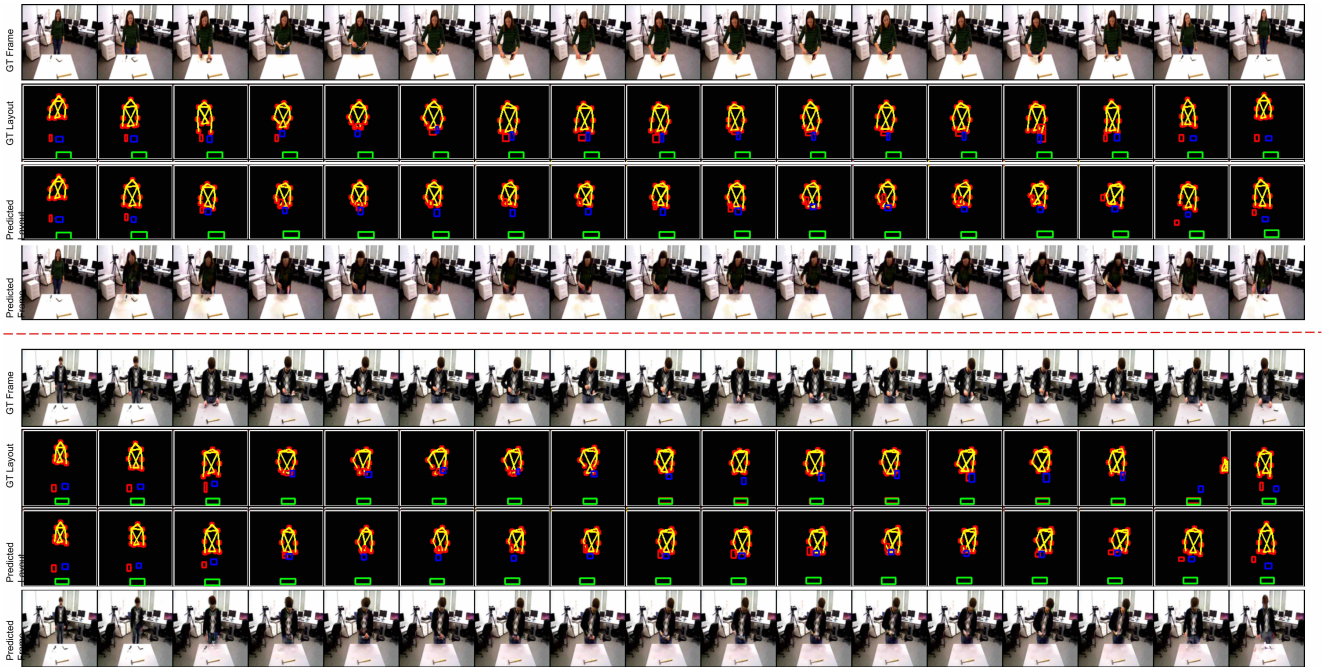

Figure 5: Example visualizations of the task “screwing a hard drive” on Bimanual dataset.

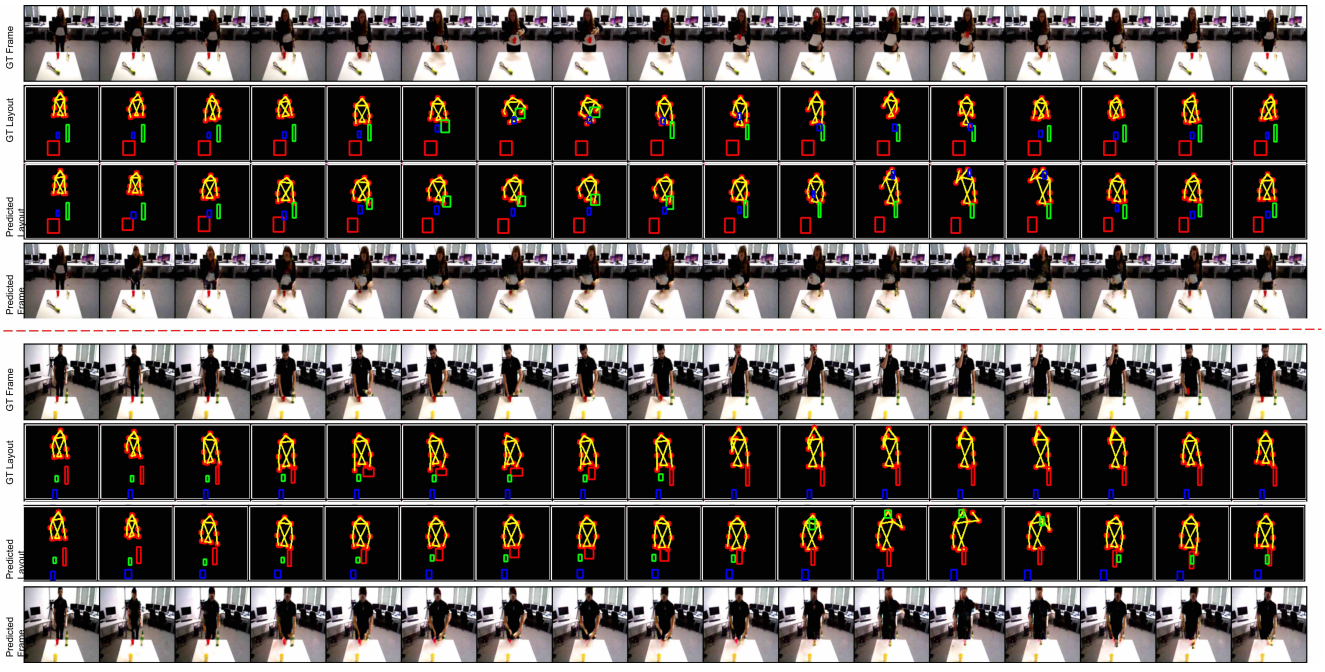

Figure 6: Example visualizations of the task “drinking” on Bimanual dataset.

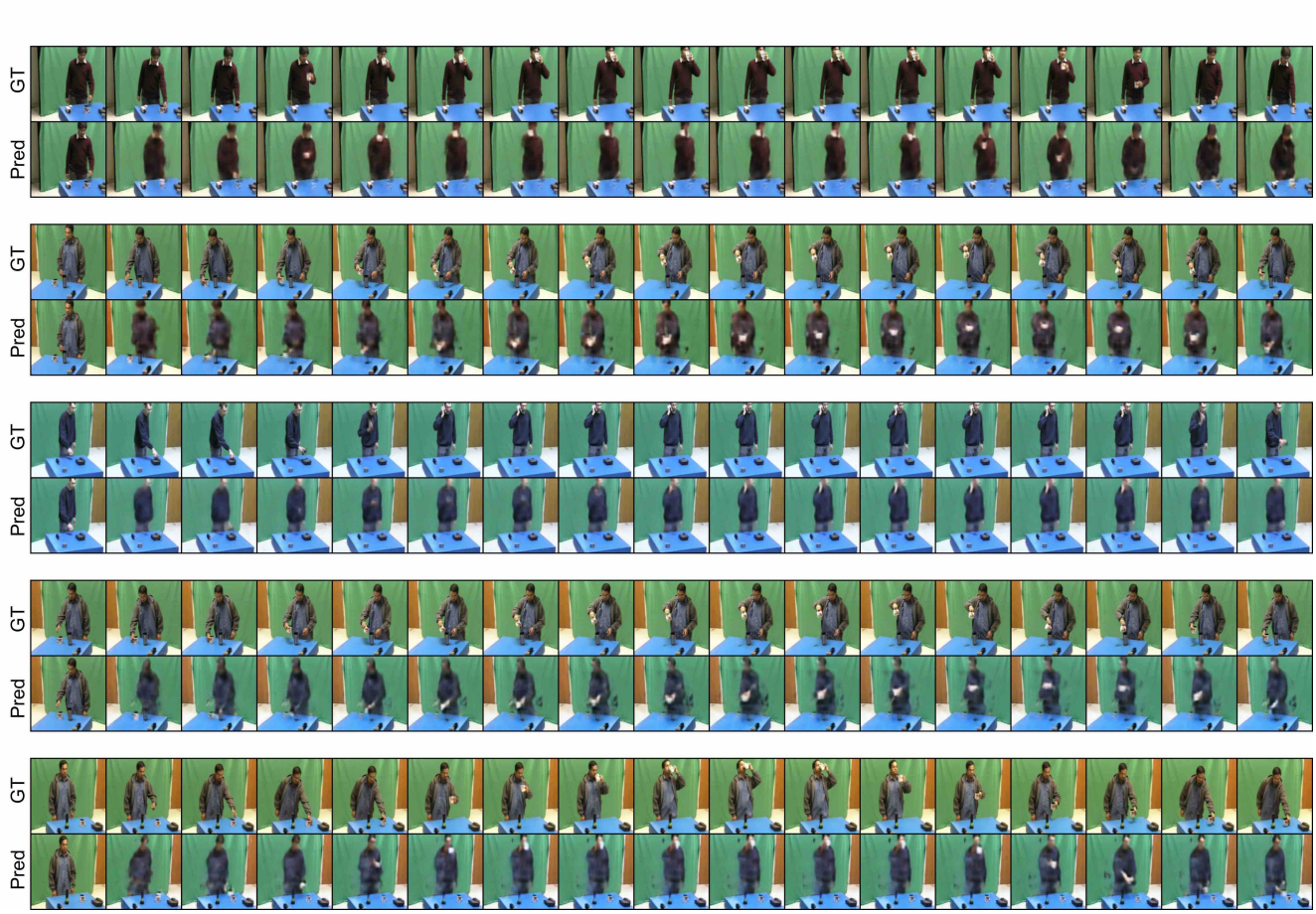

Figure 7: Example visualizations of various tasks on UMD-HOI Dataset.

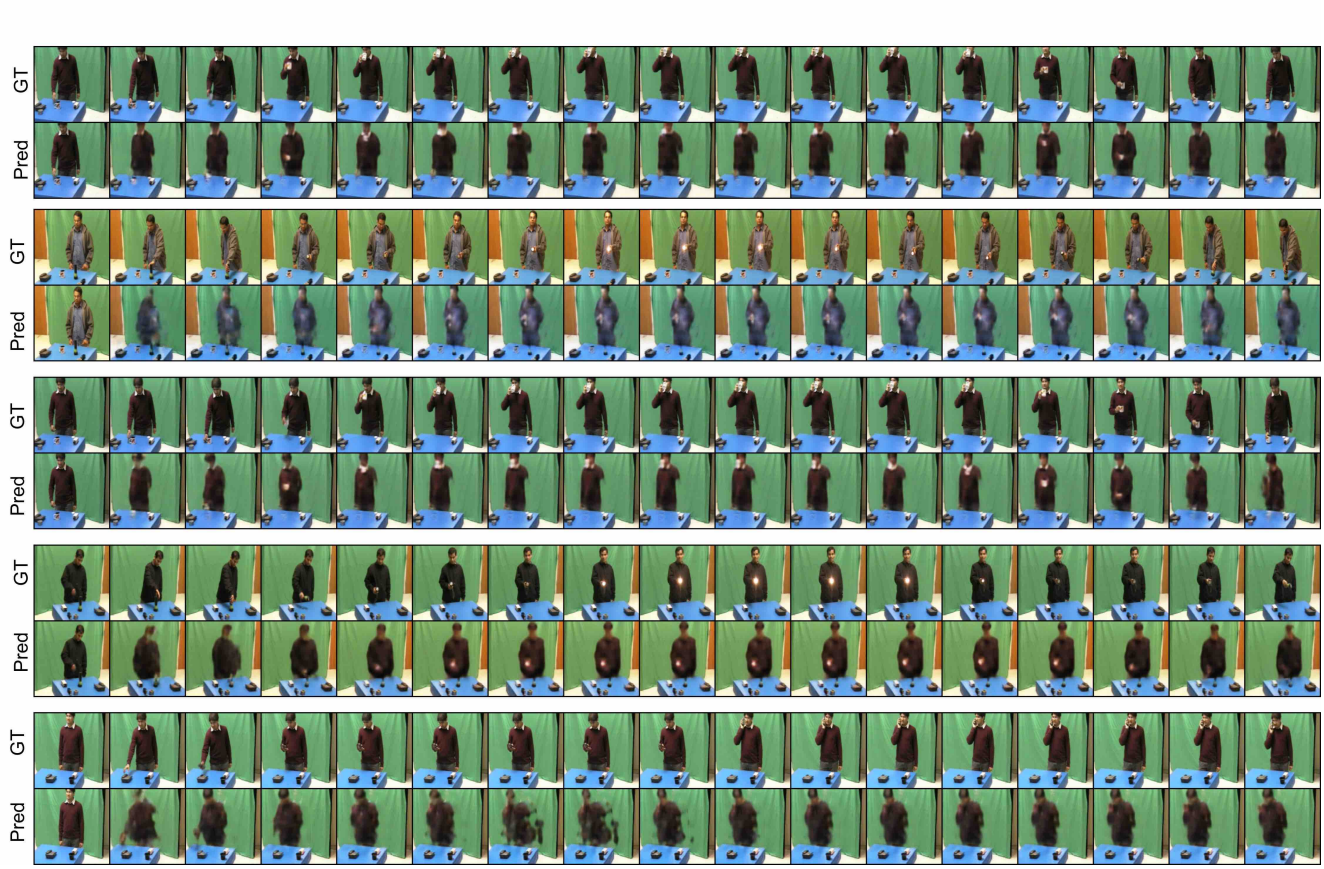

Figure 8: Example visualizations of various tasks on UMD-HOI Dataset.

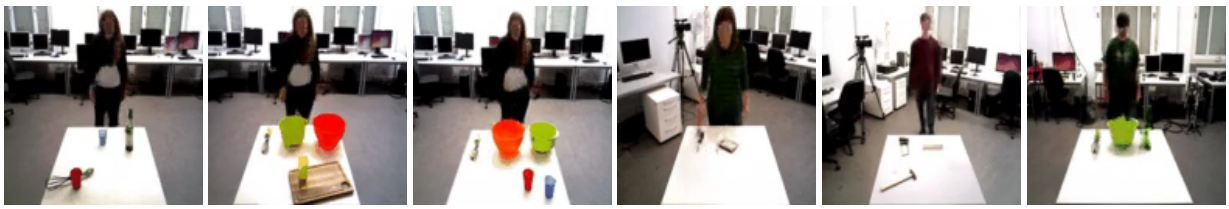

Figure 9: Examples of videos generated by our model.