Appendix: Imitation Learning for Human Pose Prediction

Borui Wang, Ehsan Adeli, Hsu-kuang Chiu, De-An Huang, Juan Carlos Niebles
Stanford University
{wbr, eadeli, hkchiu, dahuang, jniebles}@cs.stanford.edu

A.1 Behavioral Cloning for Human Pose Prediction

Our behavioral cloning algorithm for human pose prediction that was described in Section 4.6 of the main paper is listed in detail below in Algorithm 2:

Algorithm 2 Behavioral Cloning for Human Pose Prediction

Initialize
Randomly initialize the parameter θ of the policy generator network π_θ
for iteration = 1, 2, ..., T **do**
  Randomly sample a batch of N length-$(t+l)$ trajectories of human pose vectors from the training dataset $\mathcal{E}$
  Initialize the loss $L = 0$
  for $j = 1, 2, ..., N$ **do**
    Take the j-th sampled trajectory $\{x_{j,1}, x_{j,2}, ..., x_{j,t+l}\}$
    for $i = 1, 2, ..., K$ **do**
      $s_{j,i} = \{x_{j,1}, \ ..., x_{j,t+(i-1)m}\}$, $a_{j,i} = \{x_{j,t+(i-1)m+1}, \ ..., x_{j,t+im}\}$
      $L = L + \|\pi_\theta(s_{j,i}) - a_{j,i}\|_1$
    end for
  end for
  Take an Adam step on θ to minimize the loss L:
$\theta \leftarrow$ Adam($\nabla_\theta L$) = Adam($\nabla_\theta \frac{1}{N \times K} \sum_{j,i} \|\pi_\theta(s_{j,i}) - a_{j,i}\|_1$)
end for

A.2 Analysis on Varying K

Under our proposed reinforcement learning formulation of the human pose prediction problem, the number of prediction steps K is an important hyperparameter in our imitation learning system that we need to tune during cross-validation. Here we present a detailed analysis on the effects of varying the value of K. We test the prediction performance of our proposed imitation learning algorithm using 10 different K values, $K = \{1, 2, 3, 4, 5, 6, 7, 10, 15, 30\}$, and plot the results for a representative human activity category 'walking' in Figure 1 below. From Figure 1, we can see that varying K does affect the prediction performance of our proposed imitation learning system. For shorter-term predictions, larger K values would generally yield better results. But for longer-term predictions, mid-range values of K would produce the best performance. According to cross-validation results, we chose to set $K = 5$ when evaluating on the test set in order to achieve a good balance between short-term and long-term prediction performance.

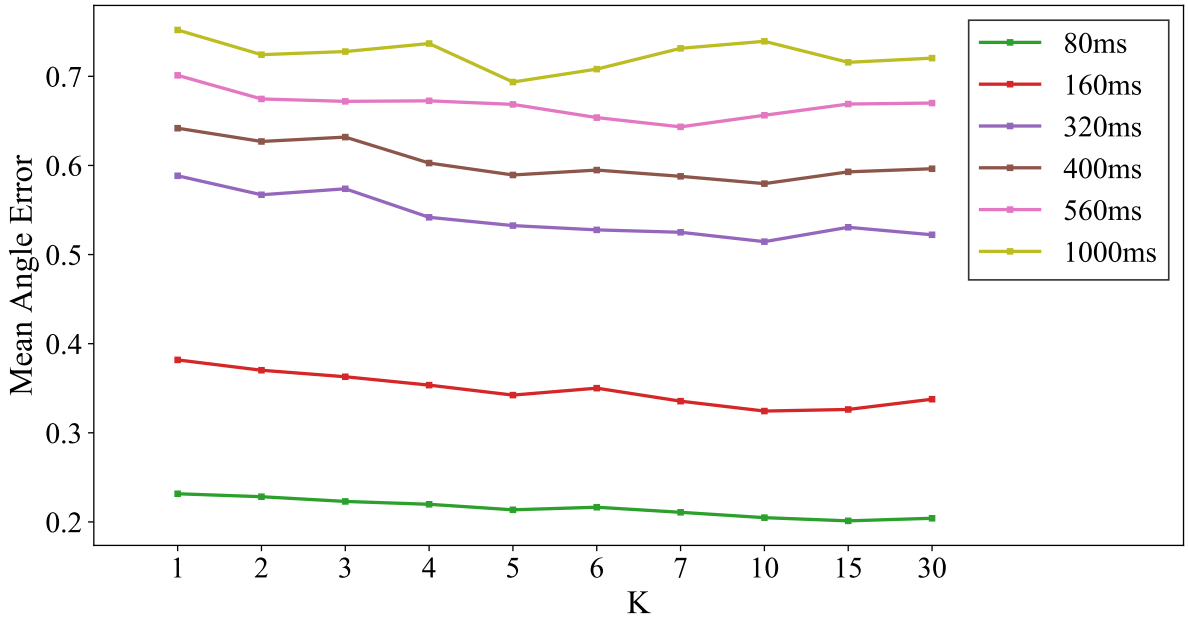

Figure 1: Effects of varying K on the prediction accuracy of our proposed imitation learning system for the activity 'walking'.

A.3 Visualization of Human Pose Prediction Results

In this section, we visualize the results of human pose prediction obtained by our proposed imitation learning method on the Human 3.6M dataset, and compare it with both the ground truth and the results obtained by the benchmark method *Residual* proposed by Martinez et al in [1].

Here we plot visualizations of human pose prediction results for 6 representative human activity categories in the Human 3.6M dataset: taking photo, walking, sitting down, sitting, phoning and eating. For each activity category we generate pose predictions over 1200ms (30 frames) into the future, and plot visualizations of poses every 80ms, resulting in a total of 15 future poses.

Taking Photo:

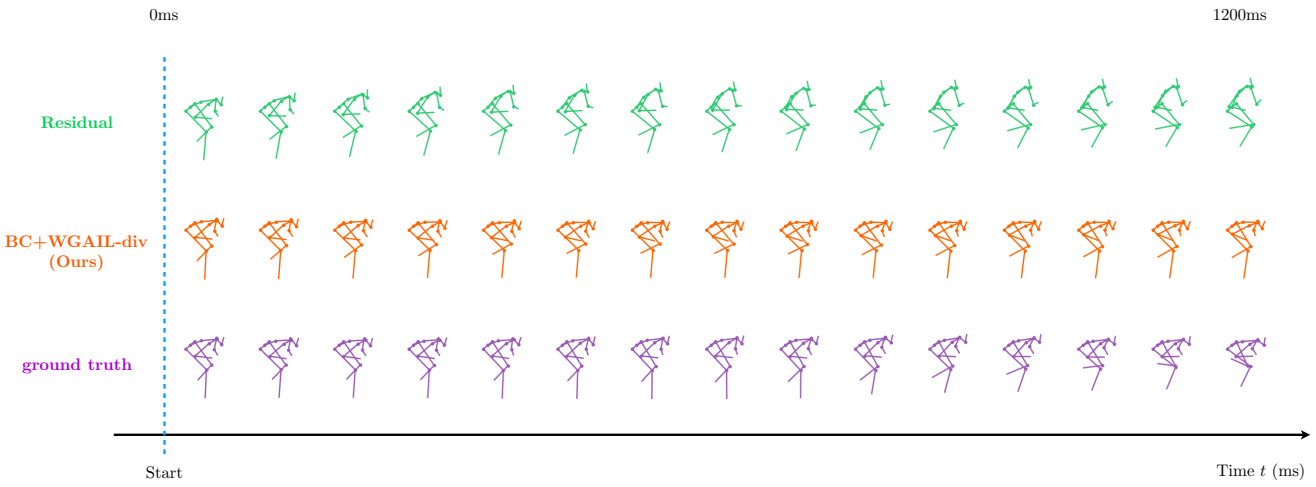

Figure 2: Visualization of human pose prediction results across 1200ms into the future for the activity *taking photo*. Top row: prediction results obtained by the benchmark *Residual* model [1], plotted in green. Middle row: prediction results obtained by our proposed imitation learning method: *behavior cloning* + *WGAIL-div*, plotted in orange. Bottom row: ground truth poses at the corresponding time points, plotted in purple.

Walking:

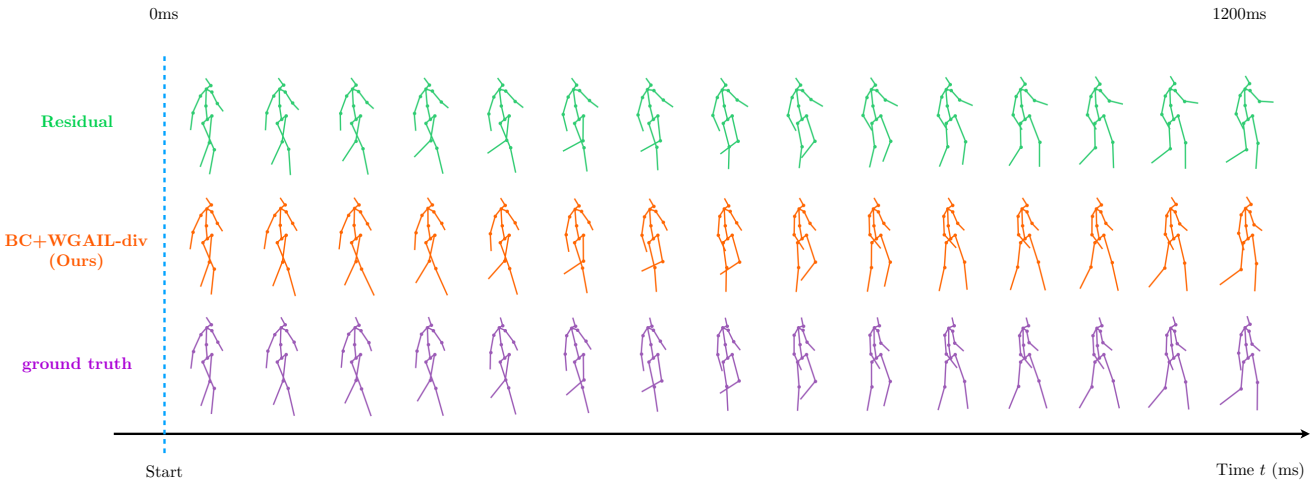

Figure 3: Visualization of human pose prediction results across 1200ms into the future for the activity *walking*. Top row: prediction results obtained by the benchmark *Residual* model [1], plotted in green. Middle row: prediction results obtained by our proposed imitation learning method: *behavior cloning* + *WGAIL-div*, plotted in orange. Bottom row: ground truth poses at the corresponding time points, plotted in purple.

Sitting Down:

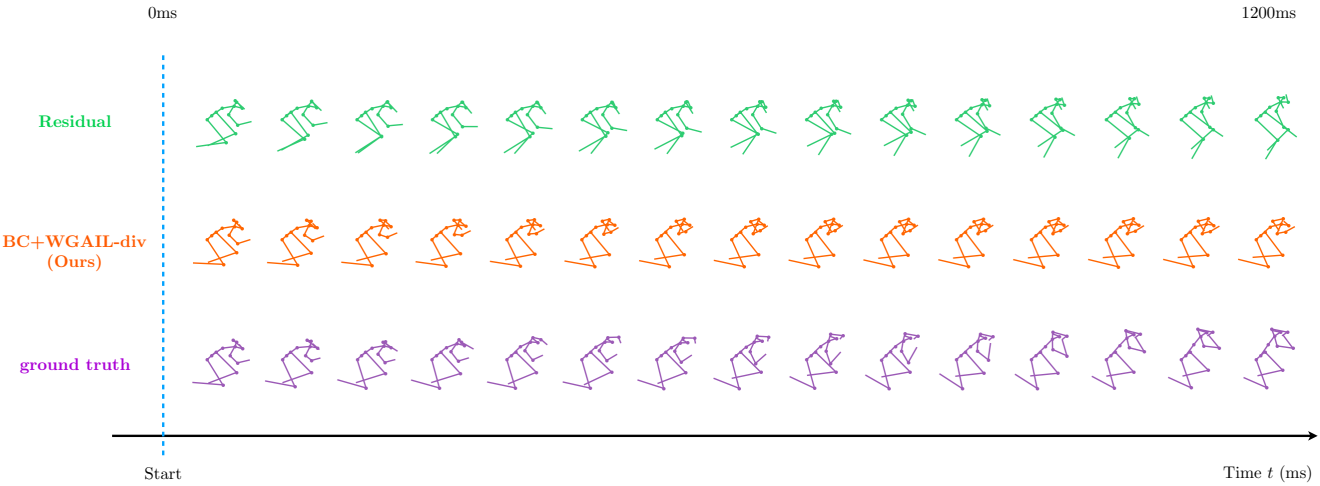

Figure 4: Visualization of human pose prediction results across 1200ms into the future for the activity *sitting down*. Top row: prediction results obtained by the benchmark *Residual* model [1], plotted in green. Middle row: prediction results obtained by our proposed imitation learning method: *behavior cloning* + *WGAIL-div*, plotted in orange. Bottom row: ground truth poses at the corresponding time points, plotted in purple.

Sitting:

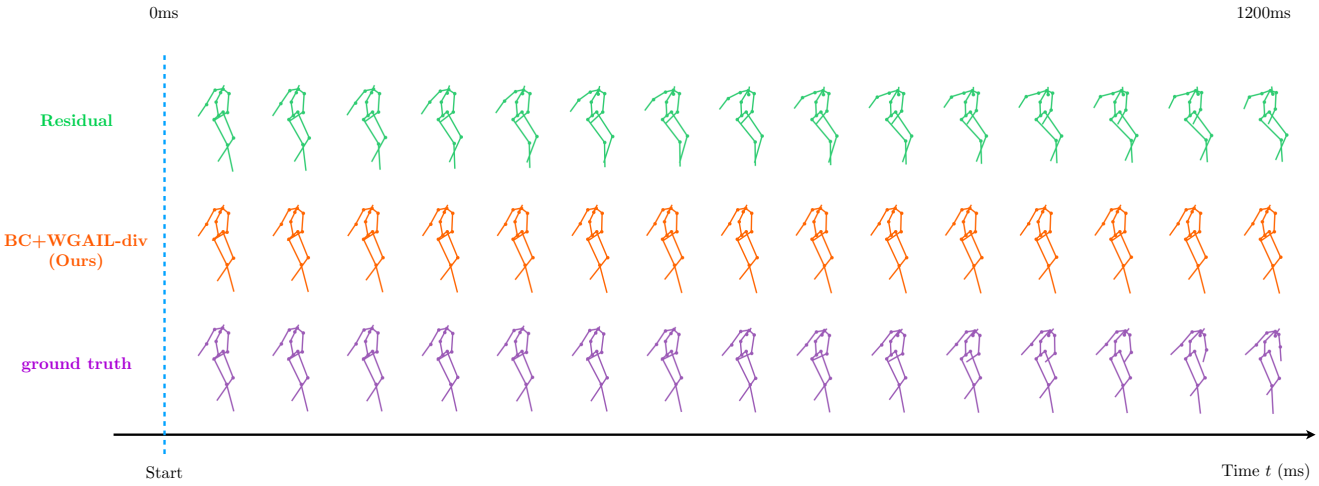

Figure 5: Visualization of human pose prediction results across 1200ms into the future for the activity *sitting*. Top row: prediction results obtained by the benchmark *Residual* model [1], plotted in green. Middle row: prediction results obtained by our proposed imitation learning method: *behavior cloning* + *WGAIL-div*, plotted in orange. Bottom row: ground truth poses at the corresponding time points, plotted in purple.

Phoning:

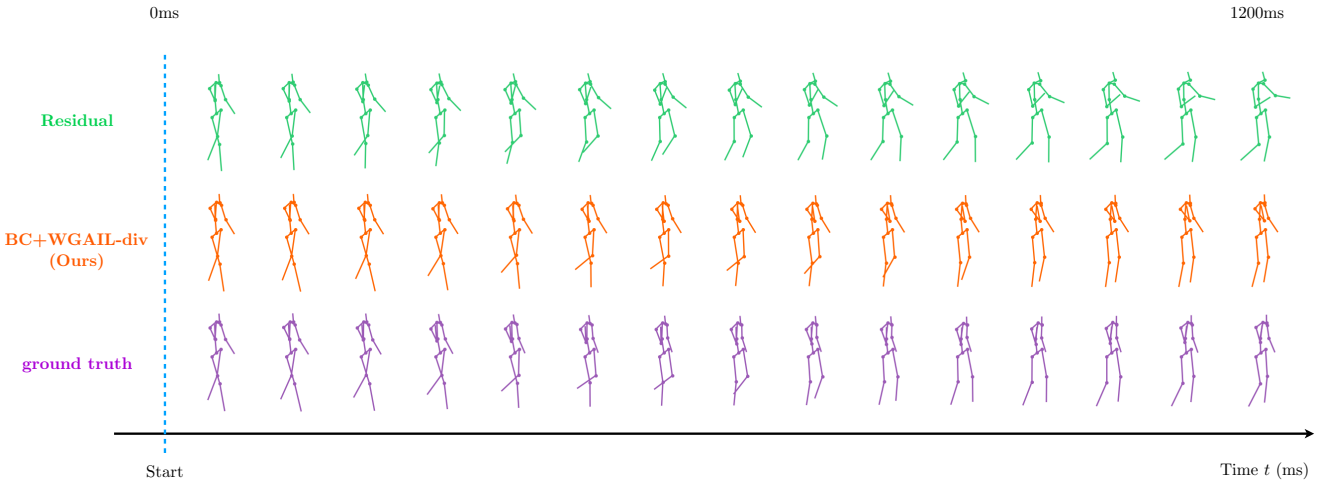

Figure 6: Visualization of human pose prediction results across 1200ms into the future for the activity *phoning*. Top row: prediction results obtained by the benchmark *Residual* model [1], plotted in green. Middle row: prediction results obtained by our proposed imitation learning method: *behavior cloning* + *WGAIL-div*, plotted in orange. Bottom row: ground truth poses at the corresponding time points, plotted in purple.

Eating:

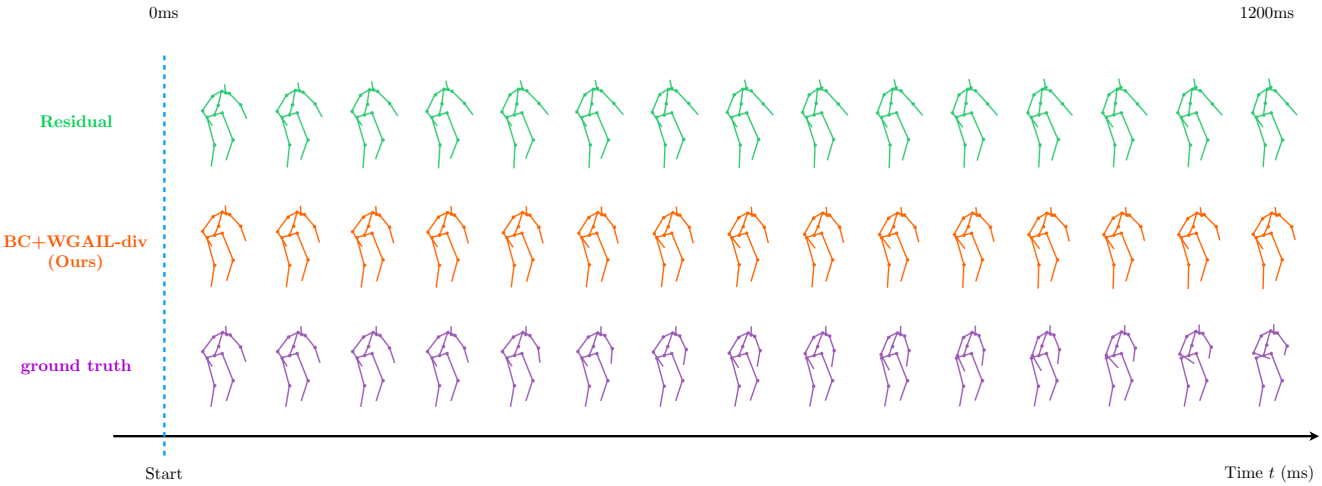

Figure 7: Visualization of human pose prediction results across 1200ms into the future for the activity *eating*. Top row: prediction results obtained by the benchmark *Residual* model [1], plotted in green. Middle row: prediction results obtained by our proposed imitation learning method: *behavior cloning + WGAIL-div*, plotted in orange. Bottom row: ground truth poses at the corresponding time points, plotted in purple.

From the above visualizations of our human pose prediction results, we clearly see that the prediction results of our proposed imitation learning method are significantly better than the results obtained by the benchmark method *Residual*. In these visualizations, we see that our predicted poses look very similar to the ground truth poses, and in many cases are almost indistinguishable from the ground truth poses by human eyes. These visualization results further confirm that our proposed imitation learning method surpasses previous methods by large margins and sets the new state-of-the-art performance for the task of human pose prediction.

References

[1] Julieta Martinez, Michael J Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2891–2900, 2017. 3, 4, 5, 6