Appendix

A. Motion-Based Pixel-Level Evaluation, Analysis, and Control Experiments

In this section, we evaluate the predictions by deciles of motion similar to Villegas et al. (2017) using Peak Signal-to-Noise Ratio (PSNR) measure, where the 10th decile contains videos with the most overall motion. We add a modification to our hierarchical method based on a simple heuristic by which we copy the background pixels from the last observed frame using the predicted pose heat-maps as foreground/background masks (`Ours BG`). Additionally, we perform experiments based on an *oracle* that provides our image generator the exact future pose trajectories (`Ours GT-pose`*) and we also apply the previously mentioned heuristics (`Ours GT-pose BG`*). We put * marks to clarify that these are *hypothetical* methods as they require ground-truth future pose trajectories.

In our method, the future frames are strictly dictated by the future structure. Therefore, the prediction based on the future pose oracle sheds light on how much predicting a different future structure affects PSNR scores. (Note: many future trajectories are possible given a single past trajectory.) Further, we show that our conditional image generator given the perfect knowledge of the future pose trajectory (e.g., `Ours GT-pose`*) produces high-quality video prediction that both matches the ground-truth video closely and achieves much higher PNSRs. These results suggest that our hierarchical approach is a step in the right direction towards solving the problem of long-term pixel-level video prediction.

A.1. Penn Action

In Figures 6, and 7, we show evaluation on each decile of motion. The plots show that our method outperforms the baselines for long-term frame prediction. In addition, by using the future pose determined by the oracle as input to our conditional image generator, our method can achieve even higher PSNR scores. We hypothesize that predicting future frames that reflect similar action semantics as the ground-truth, but with possibly different pose trajectories, causes lower PSNR scores. Figure 8 supports this hypothesis by showing that higher MSE in predicted pose tends to correspond to lower PSNR score.

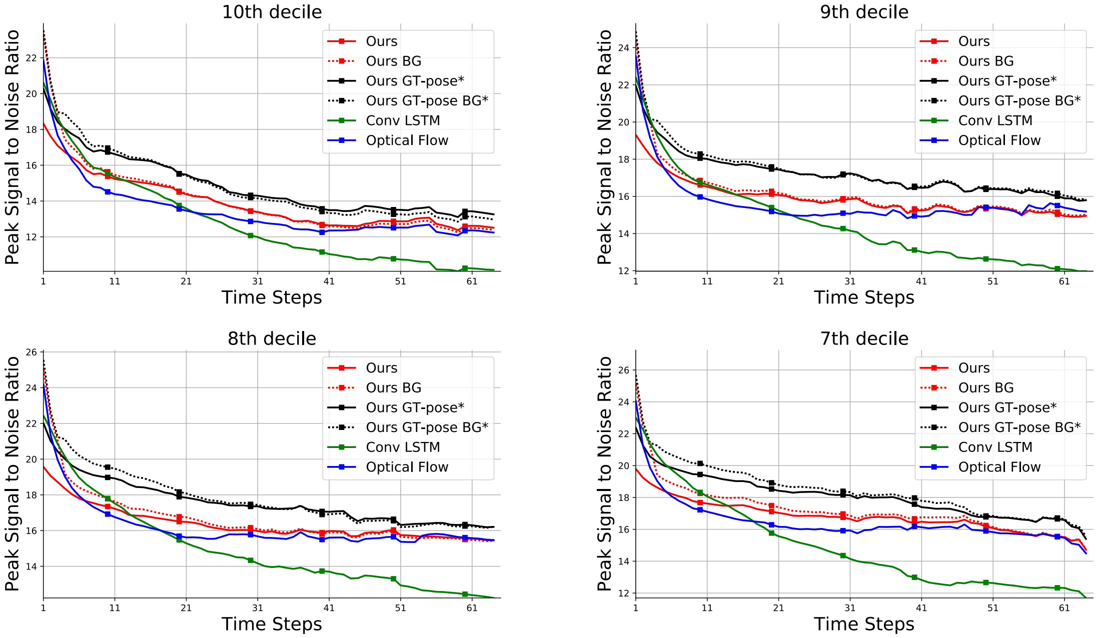

Figure 6. Quantitative comparison on Penn Action separated by motion decile.

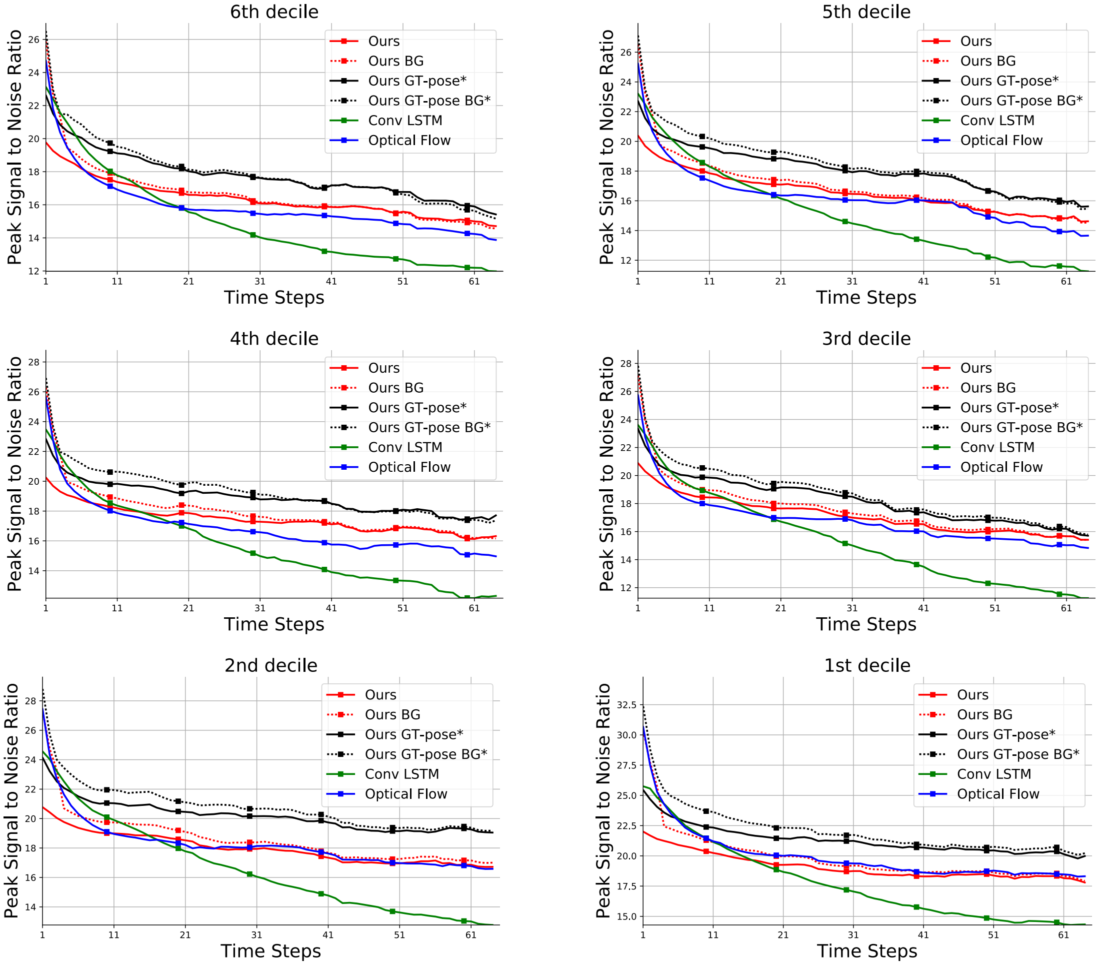

Figure 7. (Continued from Figure 6.) Quantitative comparison on Penn Action separated by motion decile.

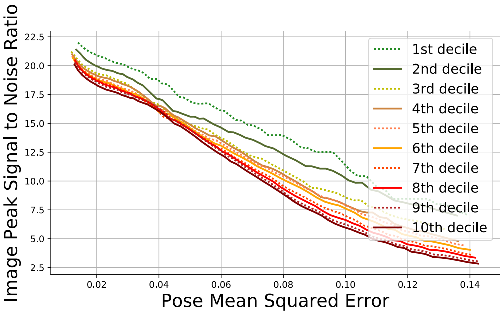

Figure 8. Predicted frames PSNR vs. Mean Squared Error on the predicted pose for each motion decile in Penn Action.

The fact that PSNR can be low even if the predicted future is one of the many plausible futures suggest that PSNR may not be the best way to evaluate long-term video prediction when only a single future trajectory is predicted. This issue might be alleviated when a model can predict multiple possible future trajectories, but this investigation using our hierarchical decomposition is left as future work. In Figures 9 and 10, we show videos where PSNR is low when a different future (from the ground-truth) is predicted (left), and video where PSNR is high because the predicted future is close to the ground-true future (right).

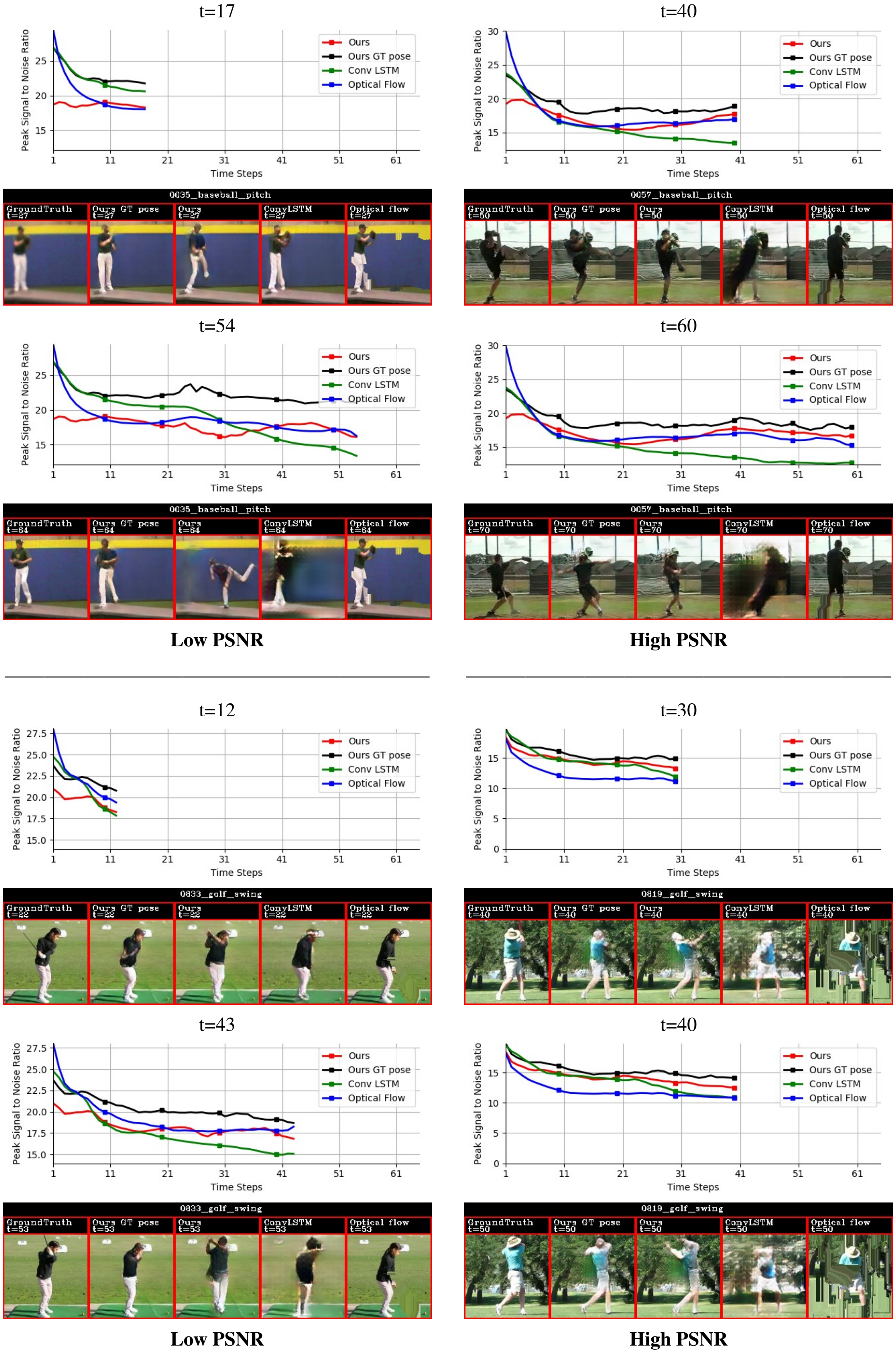

Figure 9. Quantitative and visual comparison on Penn Action for selected time-steps for the action of `baseball pitch` (top) and `golf swing` (bottom). Side by side video comparison can be found in our project website

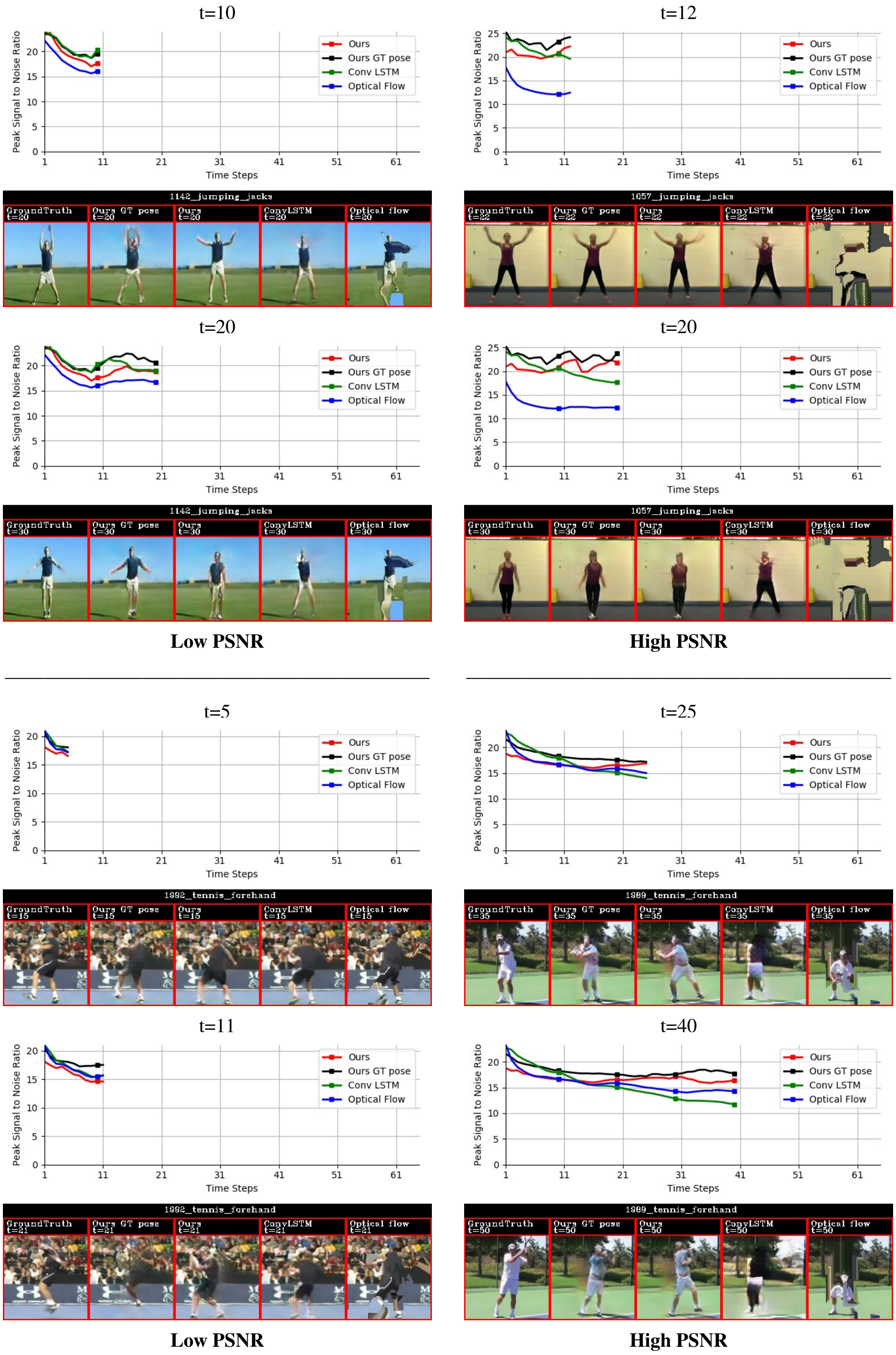

Figure 10. Quantitative and visual comparison on Penn Action for selected time-steps for the actions of `jumping jacks` (top) and `tennis forehand` (bottom). Side by side video comparison can be found in our project website

To directly compare our image generator using the predicted future pose (`Ours`) and the ground-truth future pose given by the oracle (`Ours GT-pose*`), we present qualitative experiments in Figure 11 and Figure 12. We can see that the both predicted videos contain the action in the video. The oracle based video prediction reflects the exact future very well.

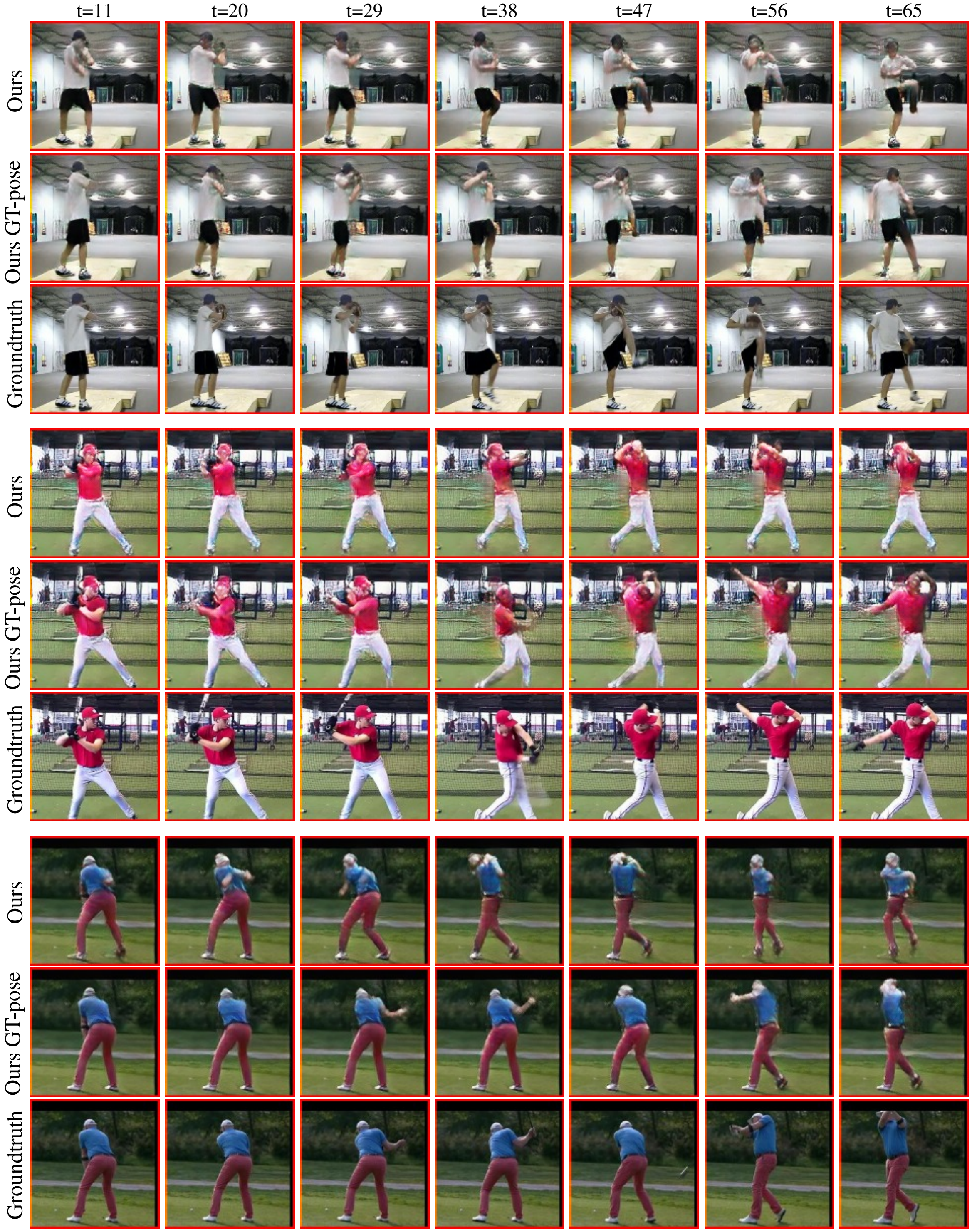

Figure 11. Qualitative evaluation of our network for long-term pixel-level generation. We show the actions of `baseball pitch` (top row), `baseball swing` (middle row), and `gold swing` (bottom row). Side by side video comparison can be found in our project website.

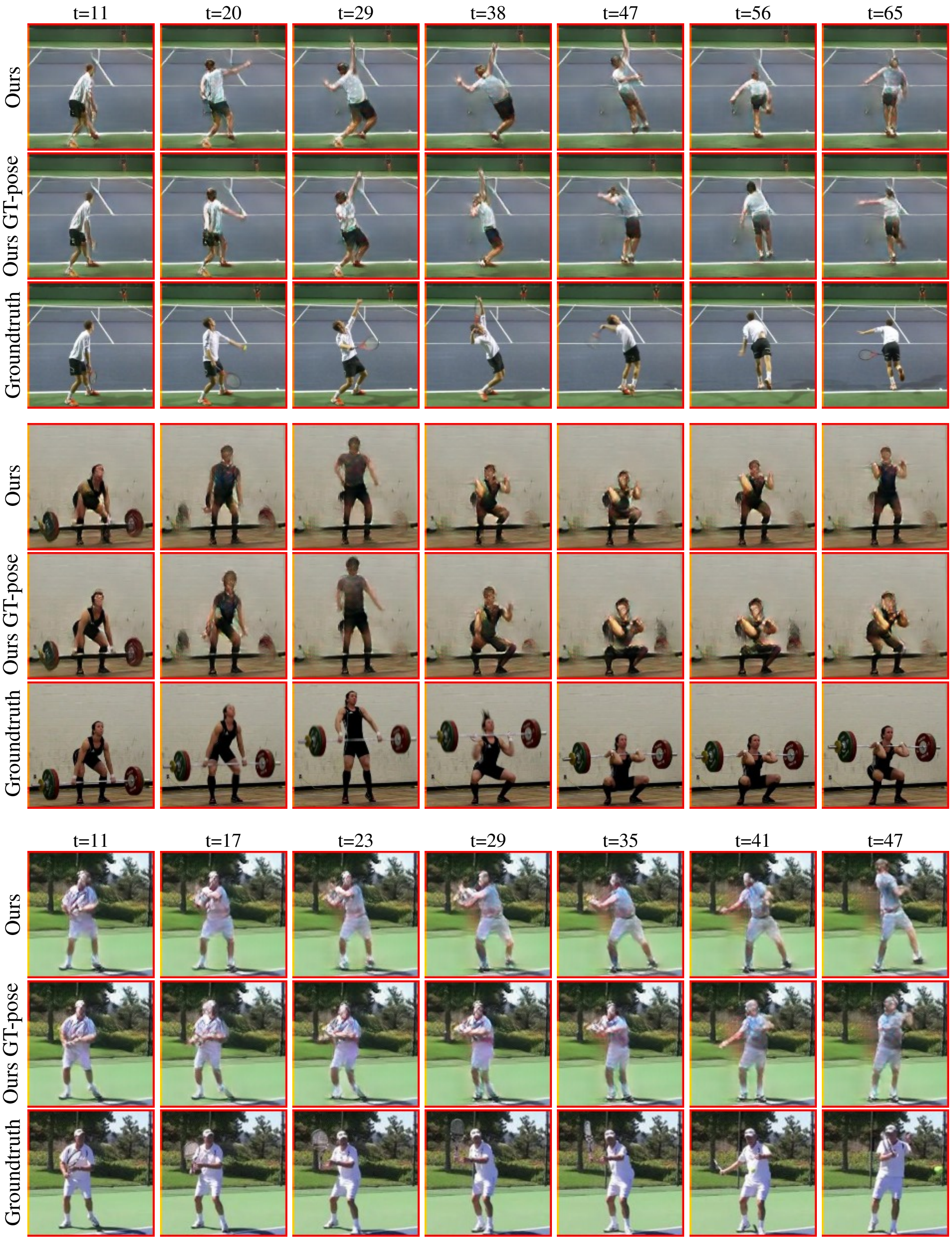

Figure 12. Qualitative evaluation of our network for long-term pixel-level generation. We show the actions of `tennis serve` (top row), `clean and jerk` (middle row), and `tennis forehand` (bottom row). We show a different timescale for `tennis forehand` because the ground-truth action sequence does not reach time step 65. Side by side video comparison can be found in our project website.

A.2. Human3.6M

In Figure 13, we show evaluation (PSNRs over time) of different methods on each decile of motion.

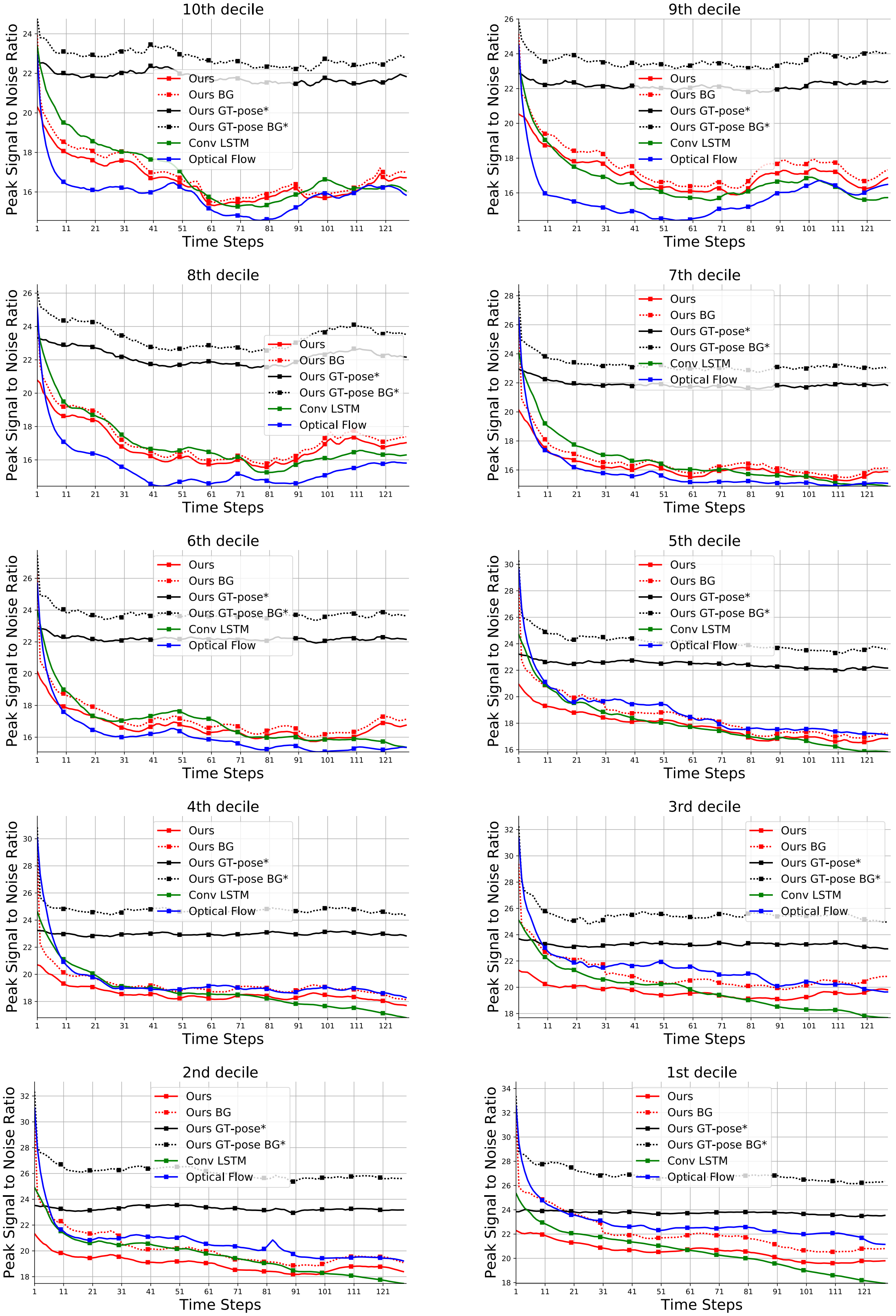

Figure 13. Quantitative comparison on Human3.6M separated by motion decile.

As shown in Figure 13, our hierarchical approach (e.g., `Ours BG`) tends to achieve PSNR performance that is better than optical flow based method and comparable to convolutional LSTM. In addition, when using the oracle future pose predictor as input to our image generator, the PSNR scores get a larger boost compared to Section A.1. This is because there is higher uncertainty of the actions being performed in the Human 3.6M dataset compared to Penn Action dataset. Therefore, even plausible future predictions can still deviate significantly from the ground-truth future trajectory, which can penalize PSNRs.

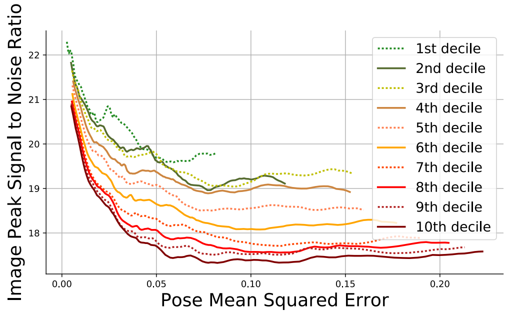

Figure 14. Predicted frames PSNR vs. Mean Squared Error on the predicted pose for each motion decile in Human3.6M.

To gain further insight on this problem, we provide two additional analysis. First, we compute how the average PSNR changes as the future pose MSE increases in Figure 14. The figure clearly shows the negative correlation between the predicted pose MSE and frame PSNR, meaning that larger deviation of the predicted future pose from the ground future pose tend to cause lower PSNRs.

Second, we show snapshots of video prediction from different methods along with the PNSRs that change over time (Figures 15 and 16). Our method tend to make plausible future pose trajectory but it can deviate from the ground-truth future pose trajectory; in such case, our method tend to achieve low PSNRs. However, when the future pose prediction from our method matches well with the ground-truth, the PSNR is much higher and the generated image frame is perceptually very similar to the ground-truth frame. In contrast, optical flow and convolutional LSTM make prediction that often loses the structure of the foreground (e.g., human) over time, and eventually their predicted videos tend to become *static*. It is interesting to note that our method is comparable to convolutional LSTM in terms of PSNR, but that our method still strongly outperforms convolutional LSTM in terms of human evaluation, as described in Section 6.2.

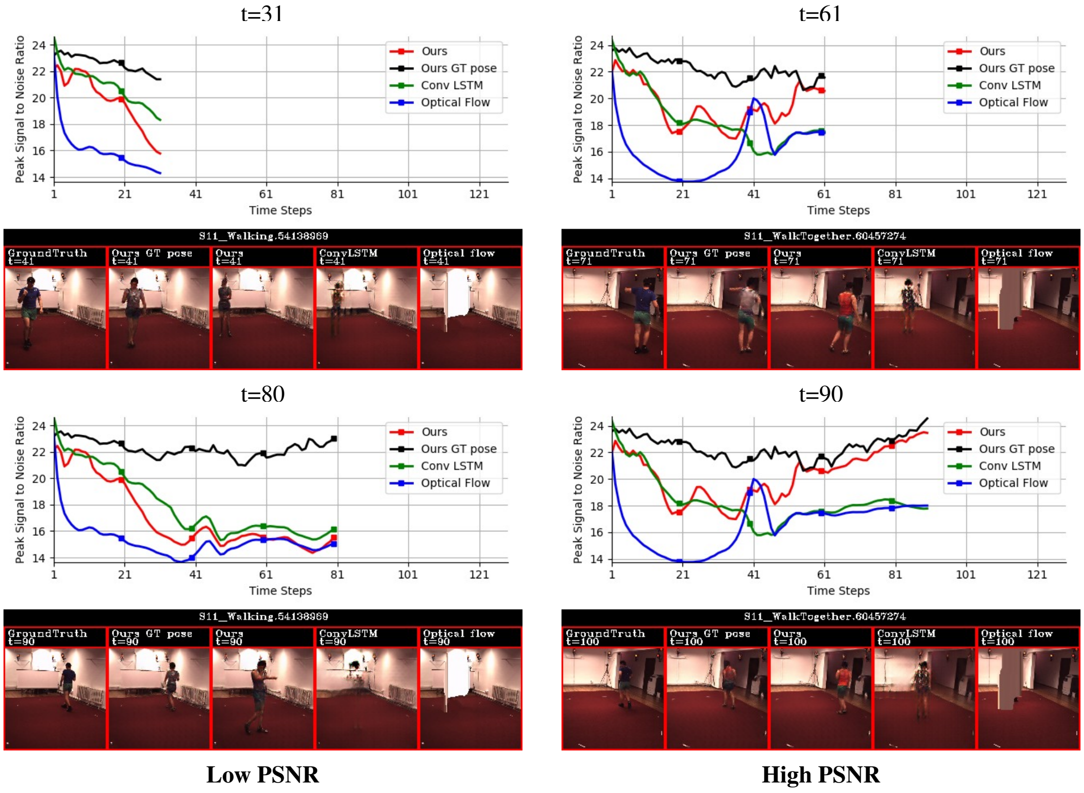

Figure 15. Quantitative and visual comparison on Human 3.6M for selected time-steps for the action of `walking` (left) and `walk together` (right). Side by side video comparison can be found in our project website.

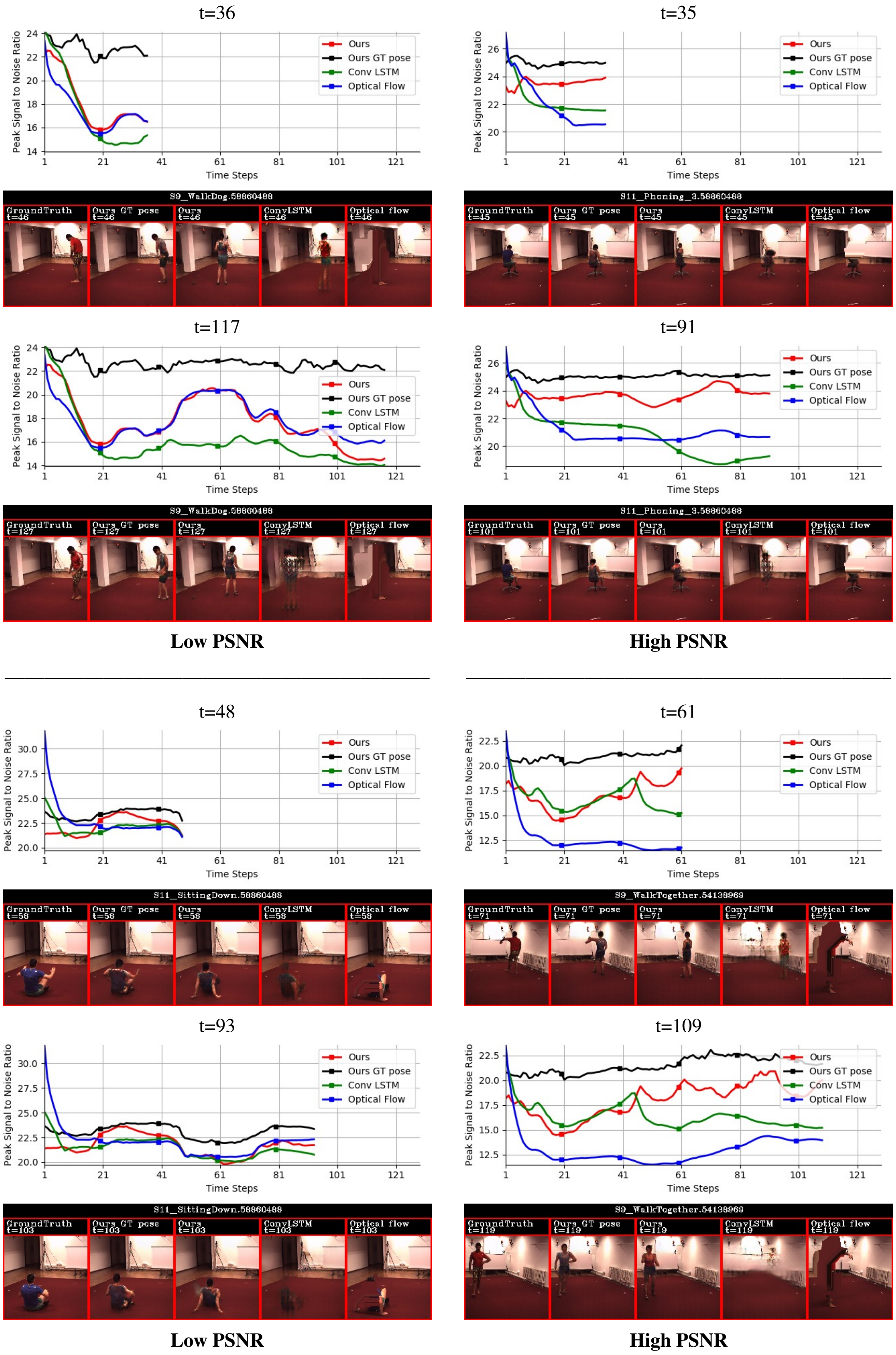

Figure 16. Quantitative and visual comparison on Human 3.6M for selected time-steps for the actions of `walk dog` (top left), `phoning` (top right), `sitting down` (bottom left), and `walk together` (bottom right). Side by side video comparison can be found in our project website.

To directly compare our image generator using the predicted future pose (Ours) and the ground-truth future pose given by the oracle (Ours GT-pose*), we present qualitative experiments in Figure 17 and Figure 18. We can see that the both predicted videos contain the action in the video. However, the oracle based video reflects the exact future very well.

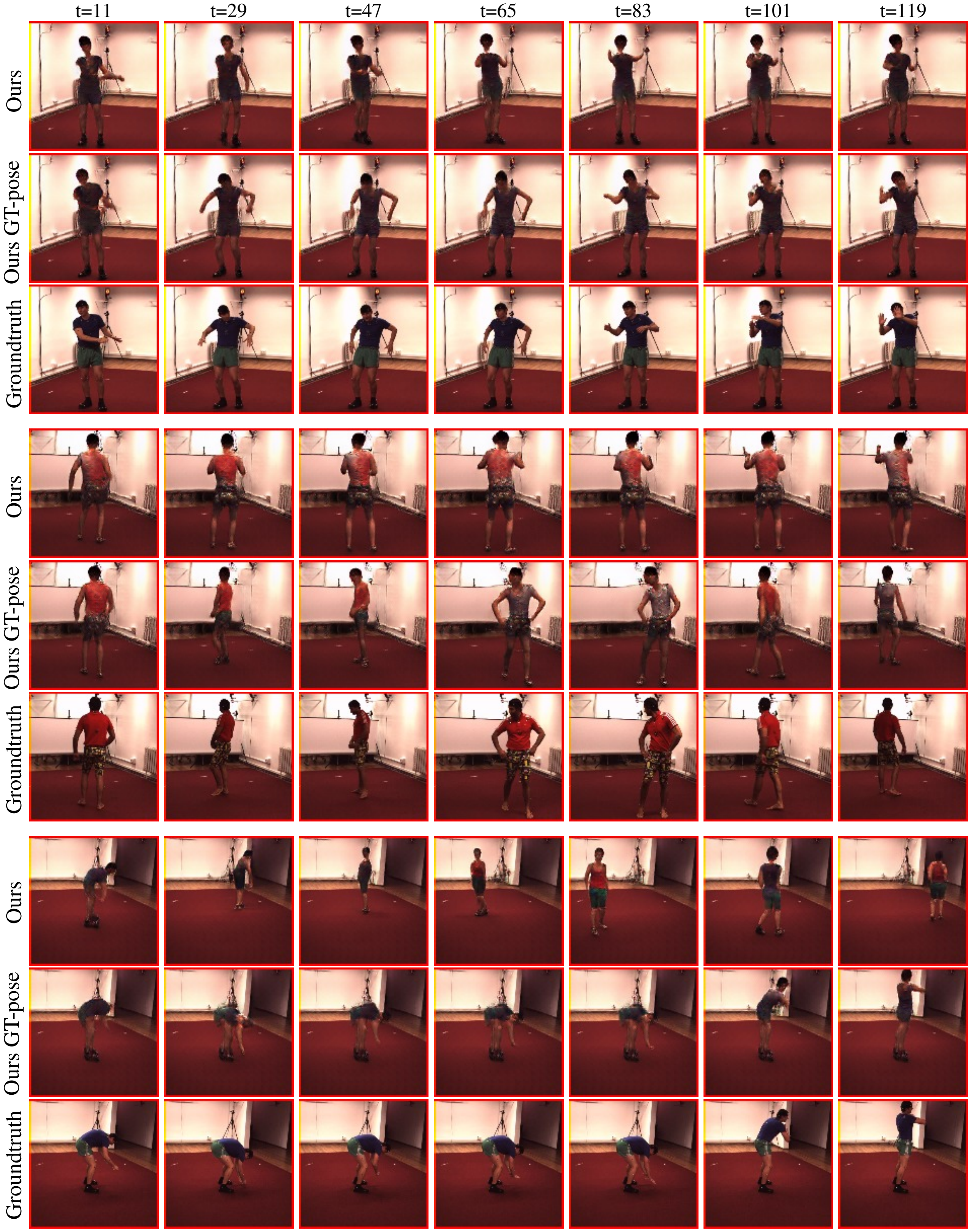

Figure 17. Qualitative evaluation of our network for long-term pixel-level generation. We show the actions of giving directions (top three rows), posing (middle three rows), and walk dog (bottom three rows). Side by side video comparison can be found in our project website.

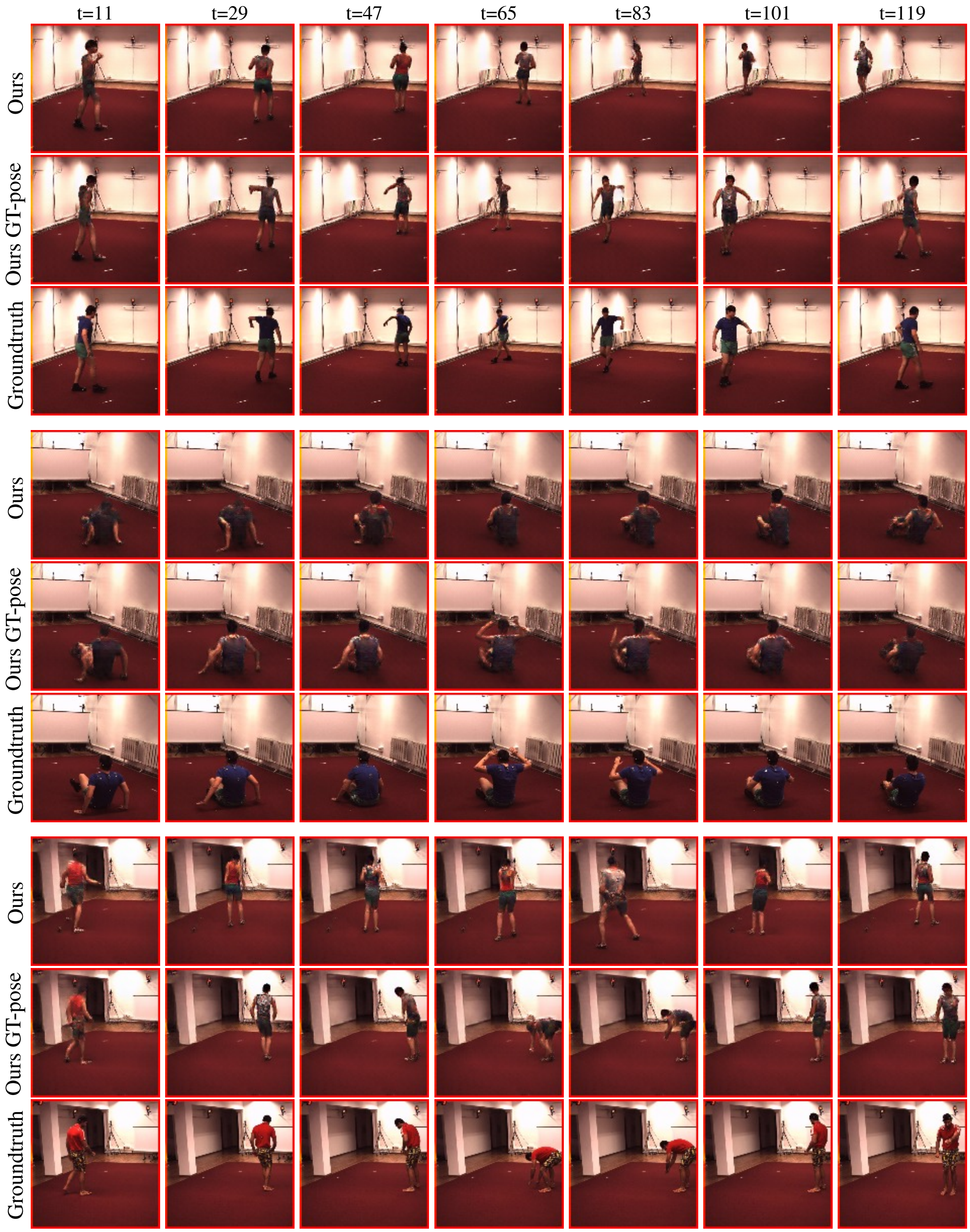

Figure 18. Qualitative evaluation of our network for long-term pixel-level generation. We show the actions of `walk together` (top three rows), `sitting down` (middle three rows), and `walk dog` (bottom three rows). Side by side video comparison can be found in our project website.