Predicting 3D Human Dynamics from Video
Supplemental Material

In this section, we provide:

- Discussion of the implementation details with limited sequence length in Section 0.1.
- A random sample of discovered "Statue" poses from Penn Action in Figure 2.
- An example of Dynamic Time Warping in Figure 3.
- Per-action evaluation of Human3.6M (Table 1) and Penn Action (Table 2) **with** Dynamic Time Warping.
- Per-action evaluation of Human3.6M (Table 3) and Penn Action (Table 4) **without** Dynamic Time Warping.
- A comparison of our method with Constant and Nearest Neighbor baseline **without** Dynamic Time Warping in Table 5.
- A visualization of Nearest Neighbor Predictions in Human3.6m (Figure 4) and Penn Action (Figure 5).
- A comparison of autoregressive predictions in the latent space versus pose space in Figures 6 and 7.
- Discussion of failure modes such as ambiguity of 2D keypoints (Figure 8), poor conditioning (Figure 9), little motion in conditioning (Figure 10), and drifting (Figure 11).

0.1. Implementation Details of Sequence Length

As discussed in the main paper, while our approach can be conditioned on a larger past context by using dilated convolutions, our setting is bottlenecked by the length of the training videos. Here we describe some implementation details for predicting long range future with short video tracks.

The length of consistent tracklets of human detections is limited given that people often walk out of the frame or get occluded. In Penn Action, for instance, the median video length is 56 frames. Thus, we chose to train on videos with at least 40 frames. Recall that to avoid drifting, we train our f_{AR} on its own predictions [**?**]. Since f_{AR} has a receptive field of 13, our model must predict 14 timesteps into the future before it is fully conditioned on its own predicted movie strips. This is further complicated by the fact that each movie strip is also causal and has its own receptive field, again pushing back when f_{AR} can begin its first future prediction. In principle, the maximum number of ground truth images that f_{AR} could be conditioned on would be one less than the sum of the receptive field of f_{AR} and f_{movie}. For a receptive field of 13, this would be $13 + 13 - 1 = 25$ images. However, with tracklets that have a minimum length of 40 frames, this would leave just $40 - 25 = 15$ timesteps for future prediction. This means that just 2 predictions would be fully conditioned on previously predicted movie strips. To support future prediction of 25 frames with a sequence length of 40, we edge pad the first image such that f_{AR} is only conditioned on 15 images. This allows us to compute losses for 25 predictions into the future, leaving enough training samples in which the past input includes previous predictions. See the illustration in Figure 1.

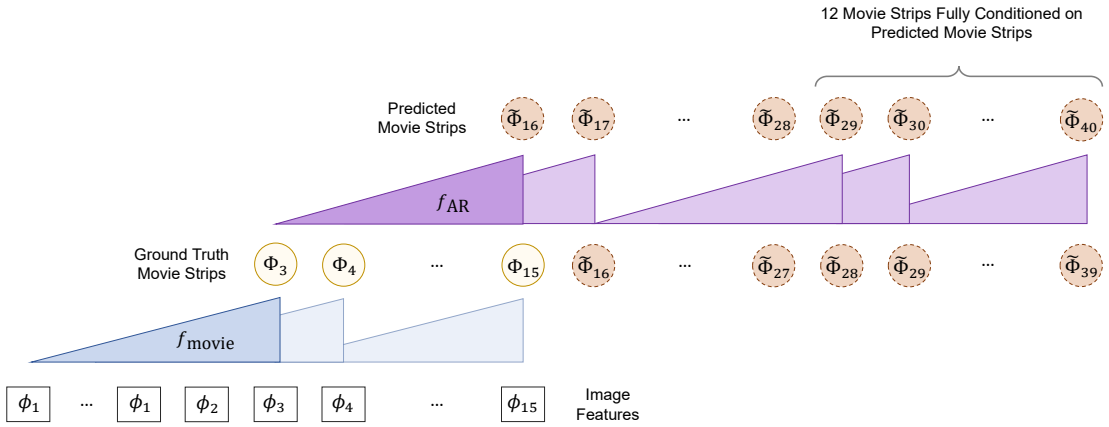

Figure 1: **Illustration of the full field of view of the proposed architecture.** Since f_{AR} and f_{movie} each have a receptive field of 13, it is theoretically possible for f_{AR} to be conditioned on 25 ground truth images. However, we train on videos that have a minimum length of 40 frames. In order to predict 25 frames into the future, we reduce the number of conditioned images to 15 by edge padding the first set of image features. See Section 0.1 for more details.

Figure 2: **Random Samples of Discovered "Statues."** We show 5 *random* samples of the statue discovery on 6 action categories from Penn Action that have fast, acyclic motion. For each sequence, we discovered the statue pose by finding the conditioning window when the prediction accuracy improves the most. Here, we visualize the first frame after the best input conditioning.

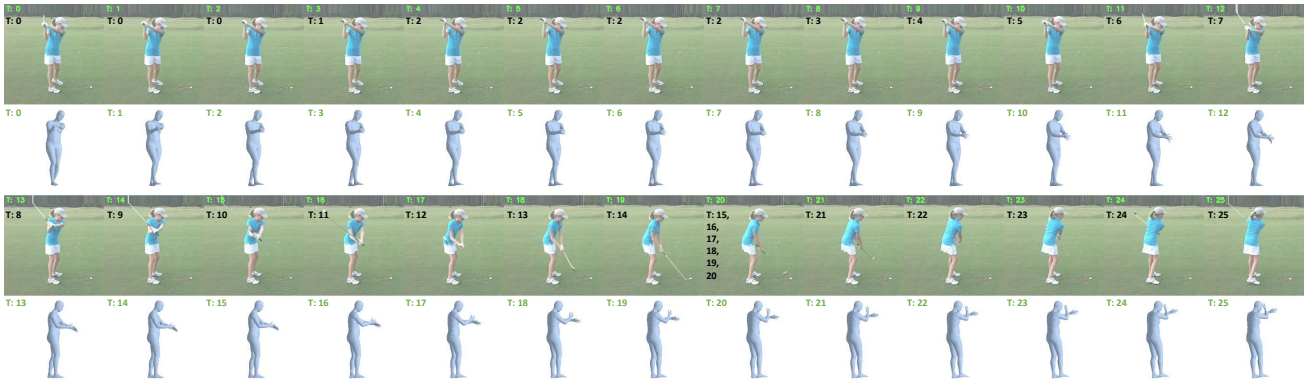

Figure 3: **Motion sequence after alignment via Dynamic Time Warping.** Dynamic Time Warping is often used to compute similarity between sequences that may have different speeds. Here, we show a *Golf Swing* sequence in which the prediction model produces the correct action type but starts the swing too soon. The optimal match produced using Dynamic Time Warping reduces the penalty of mistiming the motion. Each frame of the ground truth sequence is matched with at least one frame of the predicted sequence and vice-versa. **Green:** Index of ground truth image and predicted mesh before applying Dynamic Time Warping. **Black:** Index of predicted mesh that corresponds to the ground truth image after applying Dynamic Time Warping.

	Human3.6M	Reconst. ↓			
Action	1	5	10	20	30
Directions	54.5	57.1	59.2	60.6	63.3
Discussion	58.3	60.0	61.1	61.8	64.8
Eating	50.1	51.7	53.7	54.7	56.9
Greeting	60.4	63.9	68.0	68.1	71.4
Phoning	60.6	61.7	62.6	63.7	66.7
Posing	52.8	55.3	57.1	58.5	61.7
Purchases	52.5	54.0	55.8	56.3	60.0
Sitting	61.1	62.0	62.7	63.3	66.2
Sitting Down	68.3	69.3	70.0	71.3	75.0
Smoking	57.1	58.3	59.5	60.3	63.1
Taking Photo	73.3	74.7	75.9	77.6	81.9
Waiting	53.7	55.4	57.3	58.3	61.0
Walk Together	53.2	55.0	56.1	57.3	59.7
Walking	44.5	47.3	48.9	48.2	50.6
Walking Dog	62.0	63.8	65.8	67.7	70.3
All	57.7	59.5	61.1	62.1	65.1

Table 1: **Per-action evaluation of autoregressive future prediction in the latent space on Human3.6M with Dynamic Time Warping (DTW).** For each action, we evaluate the mean reconstruction error in *mm* after applying Dynamic Time Warping for the predictions. Each column corresponds to a different number of frames into the future. We find that our model most accurately predicts periodic motion (*e.g*. *Walking*), performs reasonably on actions with small motion (*e.g*. *Eating*, *Posing*, *Waiting*), but less so with intricate poses (*e.g*. *Sitting Down*, *Taking Photo*).

	Penn Action	PCK ↑			
Action	1	5	10	20	30
Baseball Pitch	83.9	81.2	78.2	75.5	72.1
Baseball Swing	93.6	92.3	90.2	92.0	90.8
Bench Press	63.0	62.9	62.6	62.6	62.5
Bowl	74.4	69.6	69.1	69.5	70.3
Clean And Jerk	90.8	89.4	88.9	88.9	87.9
Golf Swing	90.6	90.8	88.8	87.5	87.0
Jump Rope	93.0	92.8	92.7	93.0	92.9
Pullup	87.1	86.8	87.0	87.9	87.3
Pushup	71.4	71.0	70.6	71.5	72.2
Situp	67.4	66.6	65.8	66.5	65.4
Squat	80.8	80.7	80.4	78.9	79.1
Strum Guitar	77.8	78.4	79.2	78.8	78.7
Tennis Forehand	92.4	89.9	87.2	86.1	81.4
Tennis Serve	87.0	85.5	82.9	79.1	74.3
All	81.2	80.0	79.0	78.2	77.2

Table 2: **Per-action evaluation of autoregressive future prediction in the latent space on Penn Action with Dynamic Time Warping (DTW).** For each action, we evaluate the Percentage of Correct Keypoints after applying Dynamic Time Warping for the predictions. Each column corresponds to a different number of frames into the future. Note the *Jumping Jacks* action category is omitted because the corresponding video sequences are too short to evaluate. On the fast sequences, our model performs more accurately on linear motion (*e.g*. *Baseball Swing*, *Tennis Forehand*) than sequences that require changes in direction (*e.g*. windups in *Baseball Pitch* and *Bowl*). For actions in which motion is slow, our model performance is dependent on the viewpoint quality. For instance, *Jump Rope* and *Clean and Jerk* tend to have frontal angles whereas *Bench Press* and *Situp* are often viewed from side angles that have self-occlusions.

	H3.6M	Reconst. ↓			
Action	1	5	10	20	30
Directions	54.5	59.7	62.0	64.6	77.4
Discussion	58.3	61.4	63.0	66.4	77.5
Eating	50.1	53.3	57.5	57.5	69.5
Greeting	60.4	67.7	74.4	69.1	94.4
Phoning	60.6	62.7	64.5	67.8	80.2
Posing	52.8	57.2	60.7	63.2	80.7
Purchases	52.5	55.5	58.4	64.4	77.5
Sitting	61.1	62.6	64.6	66.4	79.0
Sitting Down	68.3	69.9	71.9	76.1	90.9
Smoking	57.1	59.2	61.7	64.4	76.8
Taking Photo	73.3	76.1	78.3	83.3	99.2
Waiting	53.7	57.0	61.0	61.9	74.8
Walk Together	53.2	57.0	60.0	65.7	78.7
Walking	44.5	50.5	55.1	58.5	75.9
Walking Dog	62.0	65.4	70.0	74.5	81.9
All	57.7	61.2	64.4	67.1	81.1

Table 3: **Per-action evaluation of autoregressive future prediction in the latent space on Human3.6M without Dynamic Time Warping (DTW).** Without DTW, the reconstruction errors accumulate quickly as the sequence goes on. As with DTW, sequences with less motion (*e.g*. *Eating*, *Posing*, *Waiting*) are easier to predict, and sequences with intricate poses (*e.g*. *Sitting Down*, *Taking Photo*) are challenging. Note that periodic motions (*e.g*. *Walking*) are much better analyzed with DTW, which accounts for uncertainties in speed such as stride frequency. This helps account for the gap in performance without DTW.

	Penn Action	PCK ↑			
Action	1	5	10	20	30
Baseball Pitch	83.9	73.6	62.7	53.8	39.2
Baseball Swing	93.6	88.2	77.2	78.4	68.3
Bench Press	63.0	62.1	61.0	61.4	56.9
Bowl	74.4	63.9	61.3	59.0	55.5
Clean And Jerk	90.8	88.2	86.8	85.6	78.1
Golf Swing	90.6	86.4	77.3	61.8	64.6
Jump Rope	93.0	90.7	91.2	89.4	88.3
Pullup	87.1	85.5	84.4	85.4	77.5
Pushup	71.4	69.8	67.9	67.0	60.9
Situp	67.4	63.6	57.2	53.4	42.8
Squat	80.8	81.5	80.0	77.1	72.7
Strum Guitar	77.8	78.4	79.8	78.7	75.8
Tennis Forehand	92.4	85.7	76.9	75.1	50.2
Tennis Serve	87.0	80.9	71.3	57.7	40.1
All	81.2	77.2	72.4	67.9	60.1

Table 4: **Per-action evaluation of autoregressive future prediction in the latent space on Penn Action without Dynamic Time Warping (DTW).** The prediction accuracy of actions with fast motion (*e.g*. *Baseball Pitch*, *Golf Swing*, *Tennis Serve*, etc.) deteriorates quickly since the speed is challenging to predict. In addition, these sequences often begin with little motion as the player prepares to begin the action or is waiting for the ball to come to them. In such cases, mis-timing the start of the action results in a large quantitative penalty. As with DTW, for the slower sequences, we observe that the actions that tend to have clearer viewpoints (*e.g*. *Jump Rope*, *Pullup*) outperform those that tend to be recorded from the side (*e.g*. *Bench Press*, *Pushup*, *Situp*).

	H3.6M	Reconst. ↓				Penn Action	PCK ↑			
Method	1	5	10	20	30	1	5	10	20	30
AR on Φ	57.7	**61.2**	**64.4**	67.1	**81.1**	**81.2**	**77.2**	**72.4**	**67.9**	**60.1**
AR on Φ, no $L_{\text{movie strip}}$	**56.9**	**61.2**	64.9	**66.8**	83.6	80.4	75.4	70.2	65.6	59.0
AR on Θ	57.8	65.9	75.9	91.9	105.2	79.9	67.8	56.2	43.4	35.1
Constant	59.7	71.4	85.9	101.4	102.8	78.3	65.5	54.6	42.3	32.7
Nearest Neighbor	90.3	99.8	110.3	124.7	133.3	62.5	57.6	53.7	44.6	41.1

Table 5: **Comparison of autoregressive predictions with various baselines without Dynamic Time Warping.** We evaluate our model with autoregressive prediction in the movie strip latent space Φ (AR on Φ), an ablation in the latent space without the distillation loss (AR on Φ, No $L_{\text{movie strip}}$), and predictions in the pose space Θ (AR on Θ). We also show the results of the no-motion baseline (Constant) and Nearest Neighbors (NN). The performance of all methods deteriorates more quickly without Dynamic Time Warping. Our method using autoregressive predictions in the latent space still significantly outperforms the baselines.

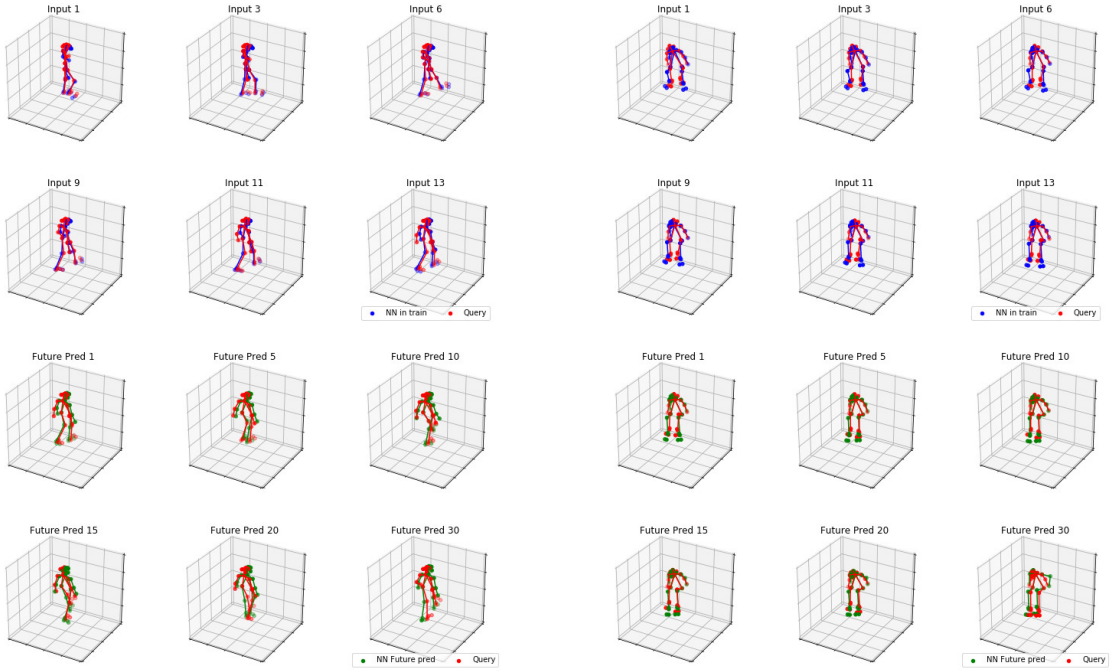

Figure 4: **Nearest Neighbor Future Predictions on Human3.6M.** For each sequence in the test set (red), we search for the best matching sequence in the training set (blue) and use the succeeding poses as the future prediction (green). **Left:** NN for a *Walking* sequence. While the query has good fit, the NN prediction drifts further from the ground truth over time. **Right:** NN for a *Sitting Down* sequence.

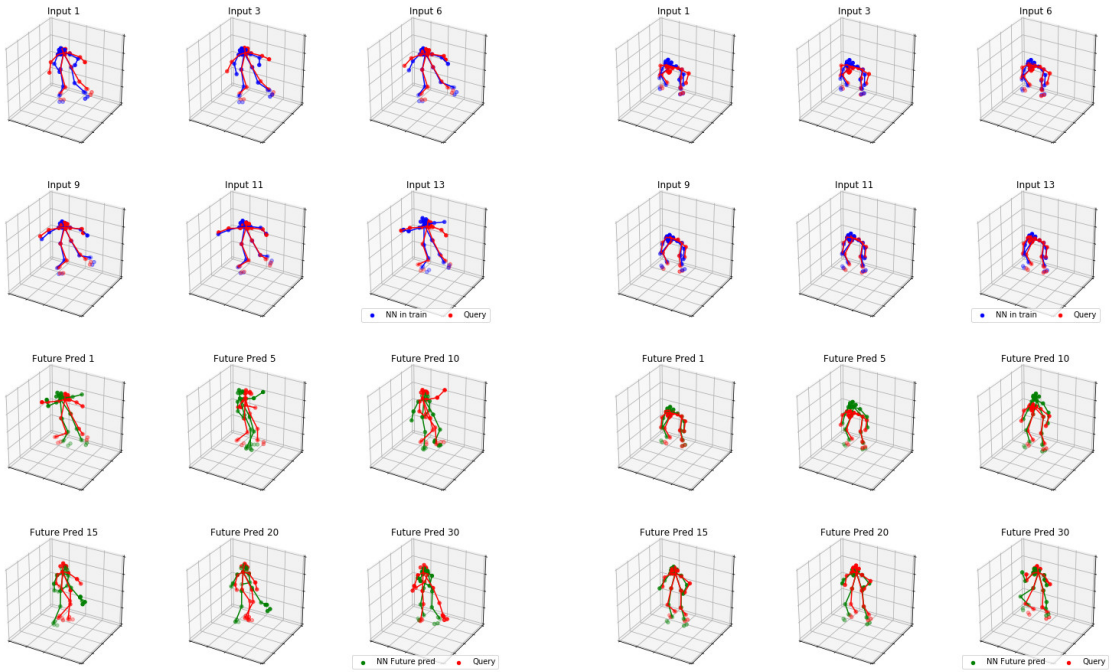

Figure 5: **Nearest Neighbor Future Predictions on Penn Action.** For each sequence in the test set (red), we search for the best matching sequence in the training set (blue) and use the succeeding poses as the future prediction (green). **Left:** NN for a *Baseball Pitch* sequence. The predicted motion is faster than the ground truth motion. **Right:** NN for a *Clean and Jerk*. The NN aligns well with the ground truth motion.

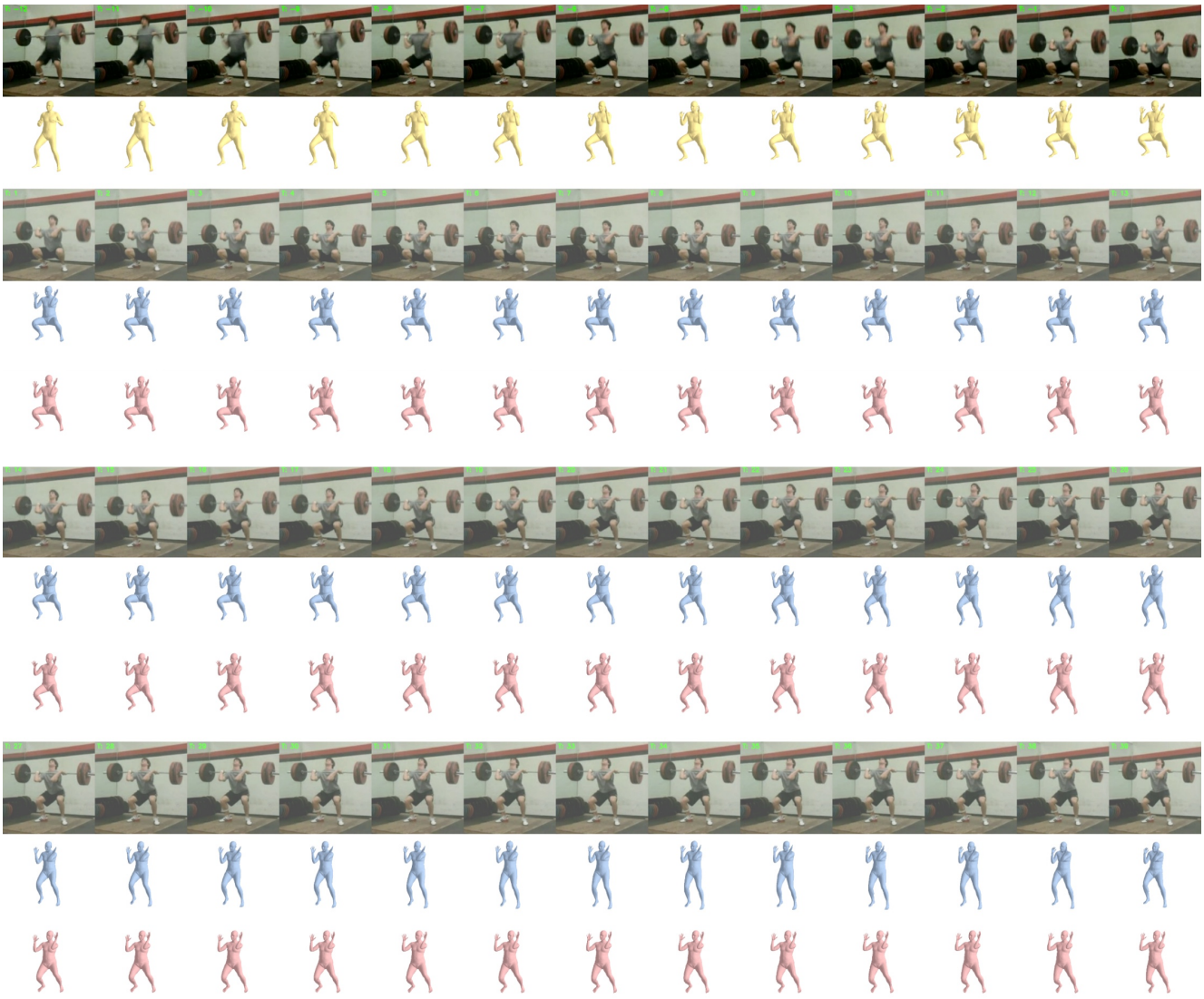

Figure 6: **Comparison of autoregressive models on performing a clean.** For simple motions, predictions in both the latent space and pose space perform reasonably. The first row of images shows the input sequence, and the rest of the images are ground truth for reference. We illustrate the conditioning with yellow meshes which are read out from the ground truth movie strips. The blue meshes show predictions in the latent space while the pink meshes show predictions in the pose space.

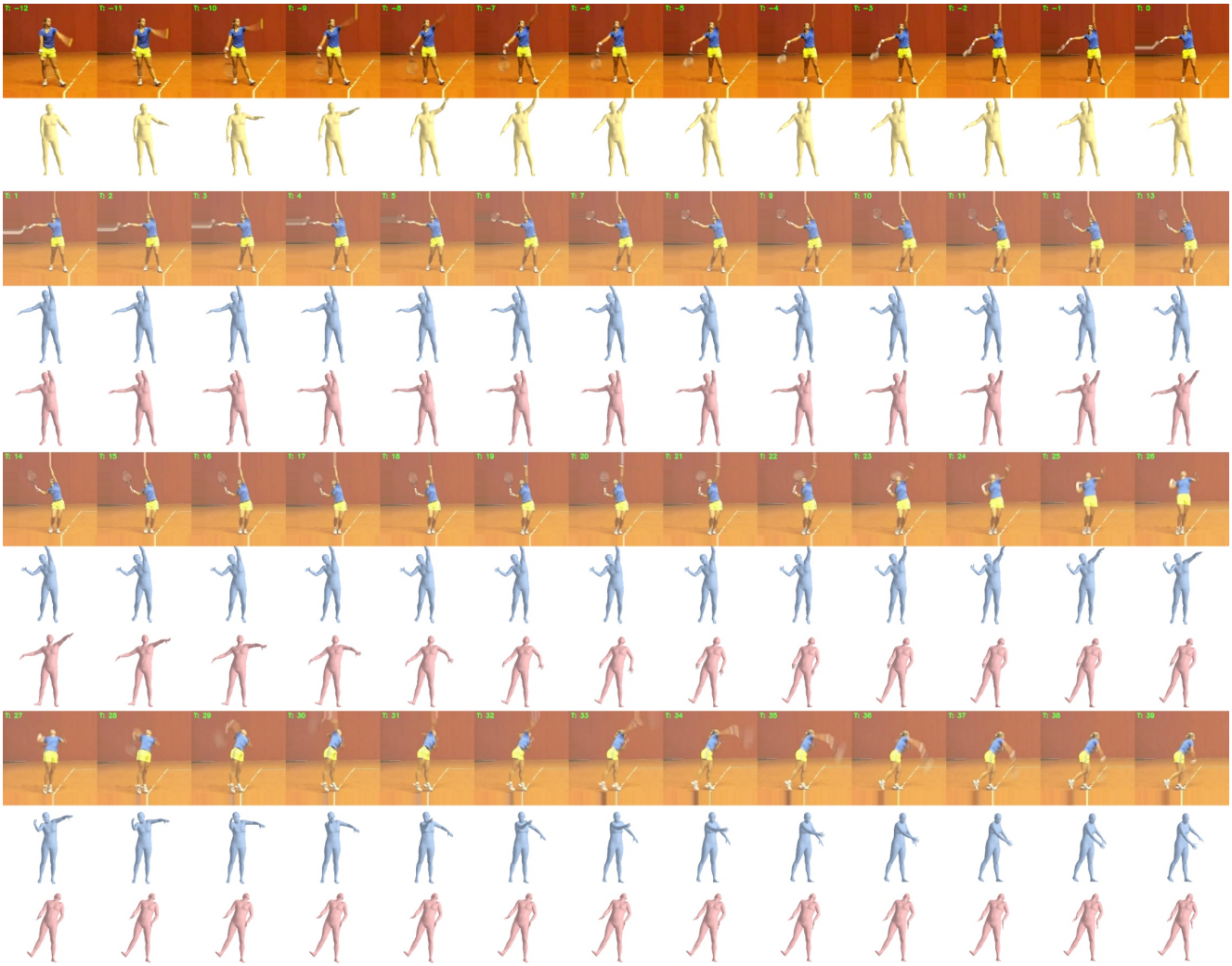

Figure 7: **Comparison of autoregressive models on a tennis serve.** For complex motions, predictions in latent space work reasonably well while predictions in the pose space struggle with identifying the action and motion. The first row of images shows the input sequence, and the rest of the images are ground truth for reference. We illustrate the conditioning with yellow meshes which are read out from the ground truth movie strips. The blue meshes show predictions in the latent space while the pink meshes show predictions in the pose space.

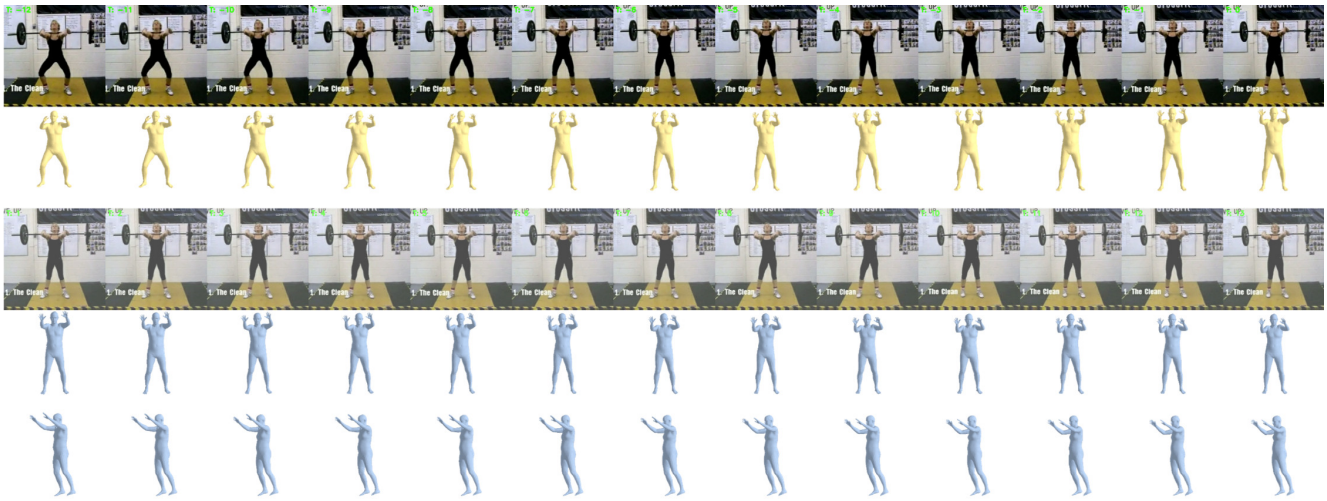

Figure 8: **Failure Mode: Ambiguity of 2D keypoints.** In-the-wild data is generally labeled only with 2D keypoints, which can have multiple 3D interpretations. We rely on an adversarial prior to produce realistic poses. Here, our model predicts motion that incorrectly extends the subject's arms, but it is still anatomically plausible and projects to the correct 2D keypoints. The first row of images shows the input sequence, and the rest of the images are ground truth for reference. We illustrate the conditioning with yellow meshes which are read out from the ground truth movie strips. The blue meshes show predictions in the latent space from two different viewpoints.

Figure 9: **Failure mode: Poor quality conditioning.** Our auto-regressive model is conditioned on the input movie strips from our temporal encoder. Mistakes made by the temporal encoder due to unusual viewpoints thus carry over to our prediction model. Here, the benching sequence is recorded from a top-down view, which is rarely encountered in the training data. The first row of images shows the input sequence, and the rest of the images are ground truth for reference. We illustrate the conditioning with yellow meshes which are read out from the ground truth movie strips. The blue meshes show predictions in the latent space from two different viewpoints.

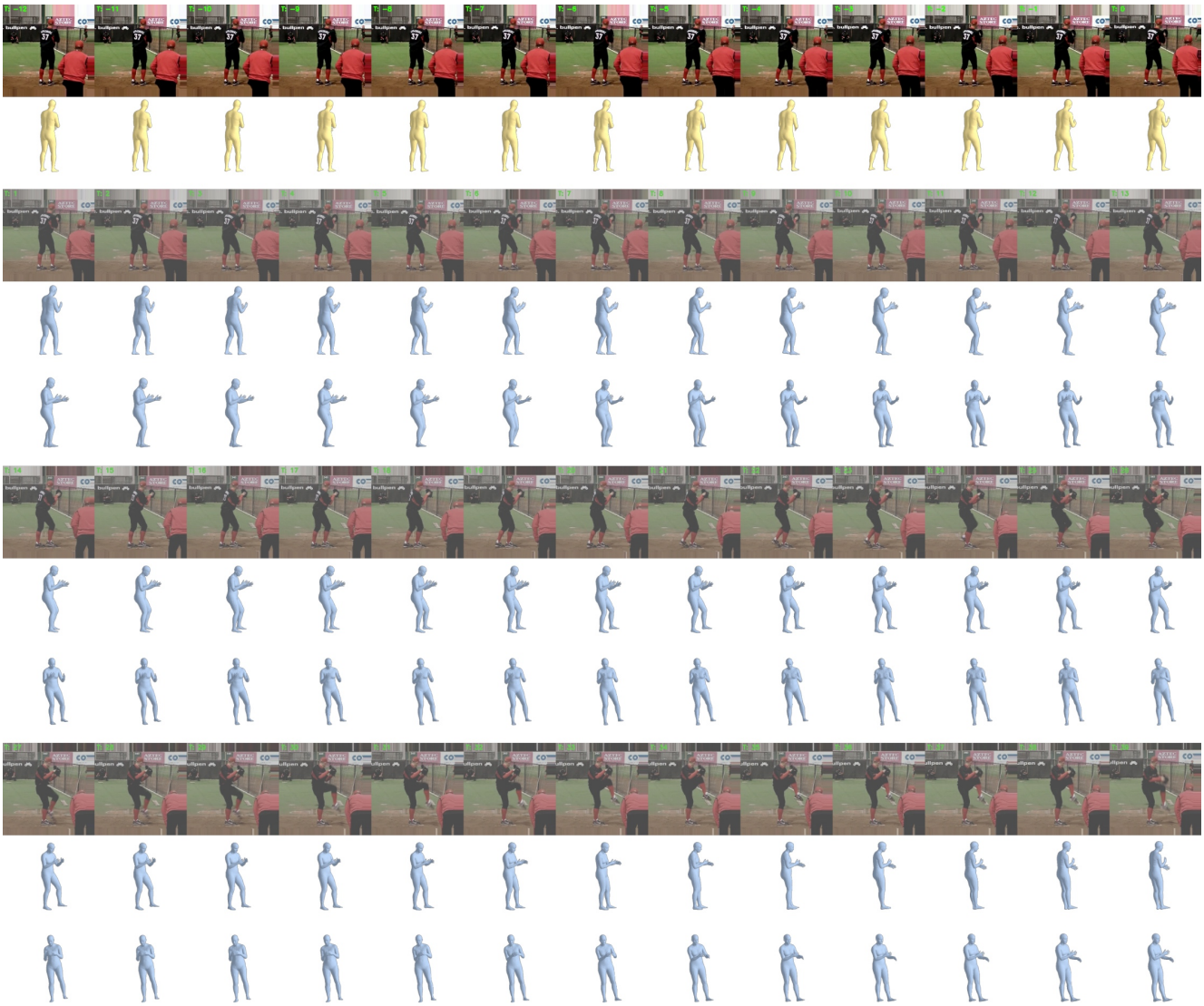

Figure 10: **Failure Mode: Conditioned on little motion.** Most sports actions in the Penn Action dataset begin with a short period with no motion as the player gets ready to pitch a ball, waits to bat, or prepares to golf. Thus, it is challenging to predict when the motion should begin when conditioned on frames corresponding to little motion. Here, the input frames show the pitcher barely moving, so our model predicts no motion while the athlete does begin to pitch later in the sequence. The first row of images shows the input sequence, and the rest of the images are ground truth for reference. We illustrate the conditioning with yellow meshes which are read out from the ground truth movie strips. The blue meshes show predictions in the latent space from two different viewpoints.

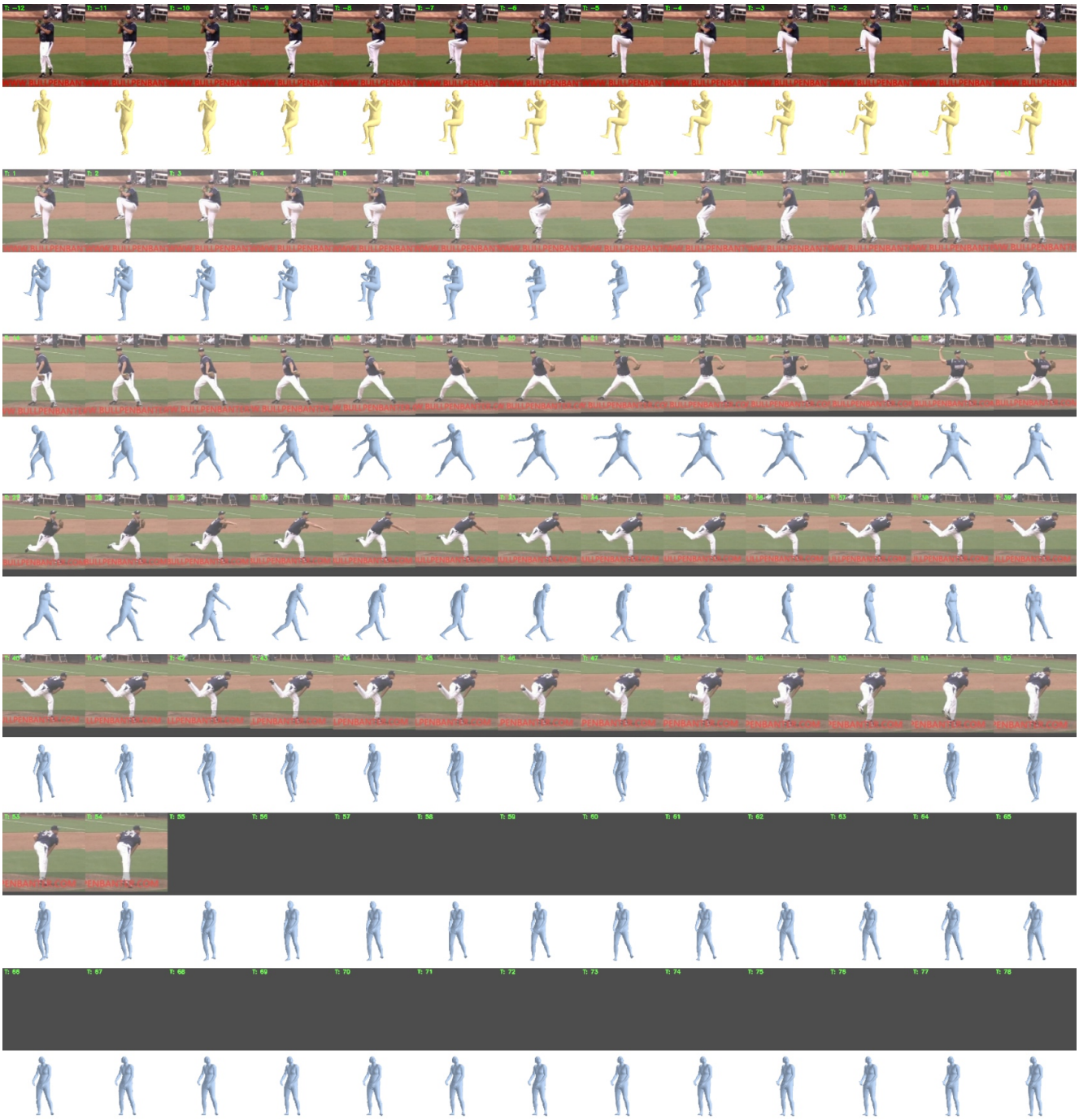

Figure 11: **Failure Mode: Drifting past 35 frames.** Due to the limited length of sequences in our training data, we train with future predictions up to 25 frames into the future. We observe that our model is capable of predicting outputs that look reasonable qualitatively until around 35 frames into the future. Training with longer sequences should alleviate this issue. The first row of images shows the input sequence, and the rest of the images are ground truth for reference. We illustrate the conditioning with yellow meshes which are read out from the ground truth movie strips. The blue meshes show predictions in the latent space.