Supplementary Material
Continual Predictive Learning from Videos

Anonymous CVPR submission

Paper ID 6337

0. Summary of The Supplementary Material

1. Model details of the proposed CPL approach.
2. More experimental configurations and training details.
3. Full quantitative comparisons on RoboNet.
4. Robustness analyses to the training order on RoboNet.
5. Further qualitative results on RobotNet in both action-free and action-conditioned setups.
6. Qualitative results for all previous tasks on KTH, which show significant improvements over the prior art.

1. Model Architecture Details

Fig. 1 and Fig. 2 provide the detailed model architectures of our Mixture World Model. Fig. 3 shows the details of the generative model that generates the initial frames for Predictive Experience Replay.

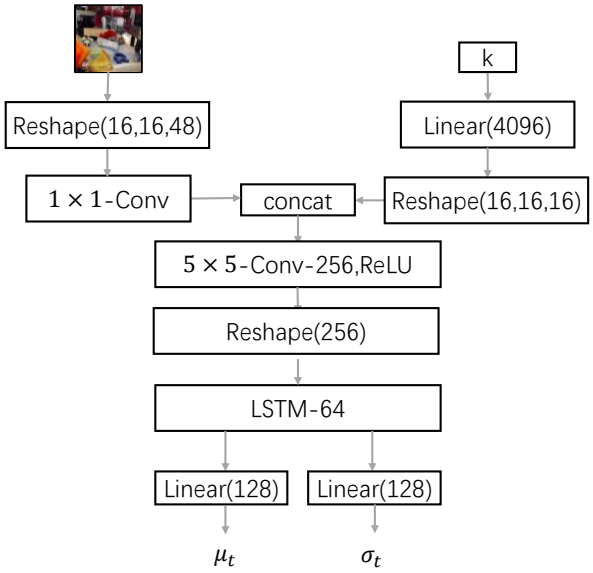

Figure 1. Architecture details of the encoding module and the representation module in our Mixture World Model.

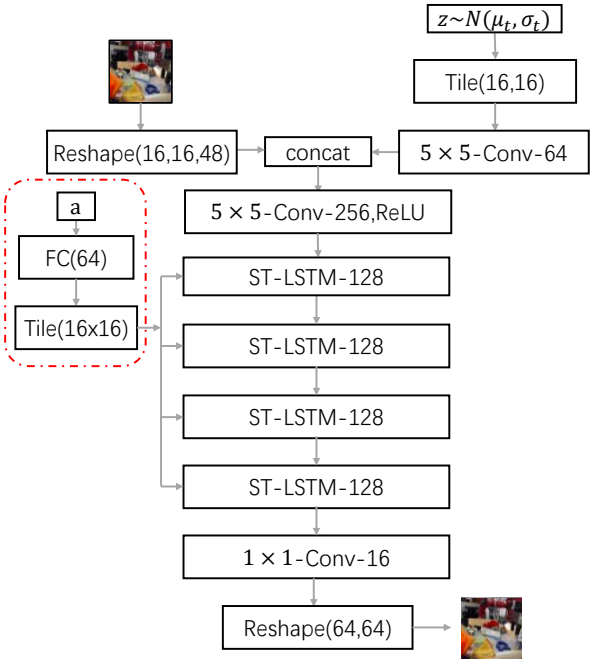

Figure 2. Architecture details of the dynamic module in the proposed Mixture World Model. Modules in the red dashed box are only used when actions are provided.

2. Experimental Configurations

We here provide the training details of CPL. In the predictive experience replay scheme, the number of rehearsal video sequences from all previous tasks is about one-third of that used in the current task. All models are trained using the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and the learning rate is set to 0.0005 for the KTH benchmark and 0.0001 for RoboNet. Besides, the mini-batch size is set to 32 for KTH and 16 for RoboNet. The input frames are pre-resized to 64×64 for both benchmarks. We optimize the entire model by $30{,}000$ iterations for each task in the continual learning process. We train all compared models on a GTX 3090 GPU.

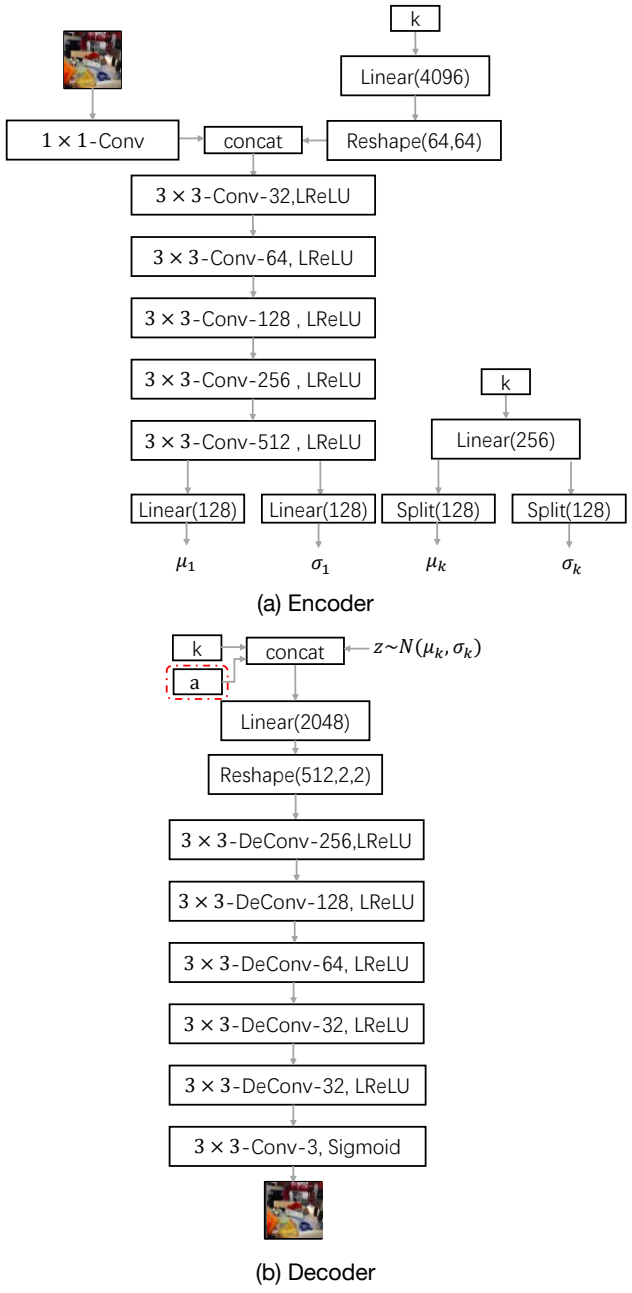

Figure 3. Architecture of the generative model in the proposed Predictive Experience Replay scheme, which learns to generate the initial frames of previous tasks. Modules in the red dashed box are only used when actions are provided.

3. Further Quantitative Results on RoboNet

Fig. 4 shows the full quantitative comparisons on particular tasks after individual training periods on the action-free RoboNet benchmark. Fig. 5 shows the corresponding results under the action-conditioned setup.

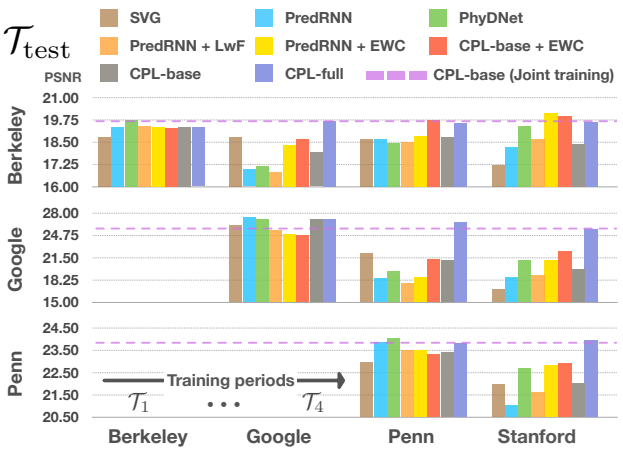

Figure 4. Results on the action-free RoboNet benchmark.

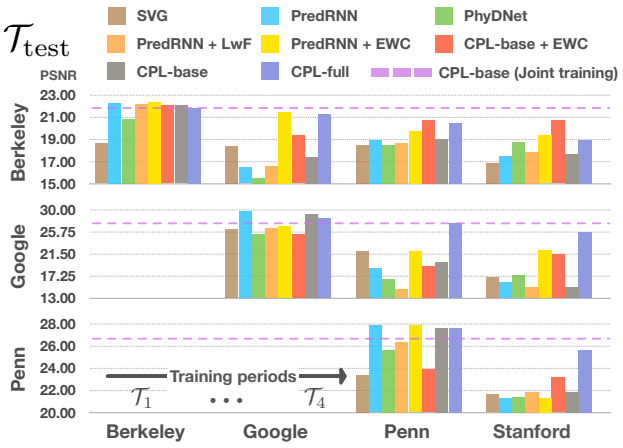

Figure 5. Results on the action-conditioned RoboNet benchmark.

4. Robustness of CPL to RoboNet Task Order

As shown in Table 1 and Table 2, we further conduct experiments on RoboNet to analyze that whether CPL can effectively alleviate catastrophic forgetting regardless of the task order. Specifically, we train CPL with a task order of *Penn* → *Google* → *Berkeley* → *Stanford*, which is different from what has been used in Table 1 in the manuscript. From the results, we find that the proposed techniques, *i.e.*, (i) the mixture world model, (ii) the predictive experience replay, and (iii) the non-parametric task inference, are still effective despite the change of training order under both action-conditioned and action-free setups.

5. Further Qualitative Results on RoboNet

5.1. Action-Free Video Prediction

Fig. 6 gives examples for predicted frames on RoboNet under the action-free setup. We here follow the task or-

Method	PSNR	SSIM ($\times 10^{-2}$)
CPL-base	19.71 $\pm$0.01	68.37 $\pm$0.04
CPL-full	**22.07$\pm$0.04**	**77.08 $\pm$0.14**

Table 1. Results on action-free RoboNet by models trained with a different task order.

Method	PSNR	SSIM ($\times 10^{-2}$)
CPL-base	19.07 $\pm$0.00	62.56 $\pm$0.02
CPL-full	**23.22$\pm$0.02**	**71.32 $\pm$0.15**

Table 2. Results on action-conditioned RoboNet by models trained with a different task order.

der described in the above section, *i.e.*, *Penn* $\rightarrow$ *Google* $\rightarrow$ *Berkeley* $\rightarrow$ *Stanford*. The input sequence is randomly sampled from the test set of the first environment (*Penn*). The prediction results are made by models that have finished the entire training procedure (*i.e.*, after the last training period on *Stanford*).

As we can see from this figure, all existing video prediction models, including SVG, PredRNN, and PhyDNet, do not have an accurate prediction of the motion of the robot arm. Even the Joint-Training baseline (see the bottom line) tends to produce rather static images across multiple time steps. Compared with the models based on LwF and EWC, our approach (CPL-full) makes less blurry predictions around the robot arm in future frames.

In Fig. 7, we also provide results on the other two previous tasks, *i.e.*, *Google* and *Berkeley*. As above, the predicted frames are generated by models that have finished the last training period on *Stanford*.

5.2. Action-Conditioned Video Prediction

For the action-conditioned setup, besides the results shown Fig. 4 in the manuscript, we here provide results on other two tasks (*i.e.*, *Google* and *Penn*) in Fig. 8. To be consistent with the results in the manuscript, we still follow the training order of *Berkeley* $\rightarrow$ *Google* $\rightarrow$ *Penn* $\rightarrow$ *Stanford*, and use the final models after the last training period on *Stanford*.

As we can see, our approach (CPL-full) shows sharper and more accurate prediction results than the state-of-the-art video prediction model (*i.e.*, PhyDNet), as well as the continual learning methods (*i.e.*, LwF and EWC).

6. Further Qualitative Results on KTH

On the KTH benchmark, we set the the training order as *Boxing* $\rightarrow$ *Clapping* $\rightarrow$ *Waving* $\rightarrow$ *Walking* $\rightarrow$ *Jogging* $\rightarrow$ *Running*. In Fig. 6 in the manuscript, we have provided the prediction showcases from the test set of the first task (*i.e.*, *Boxing*) by models trained on the last task (*i.e.*, *Running*). Here, we show corresponding results on the other four previous tasks in Fig. 9.

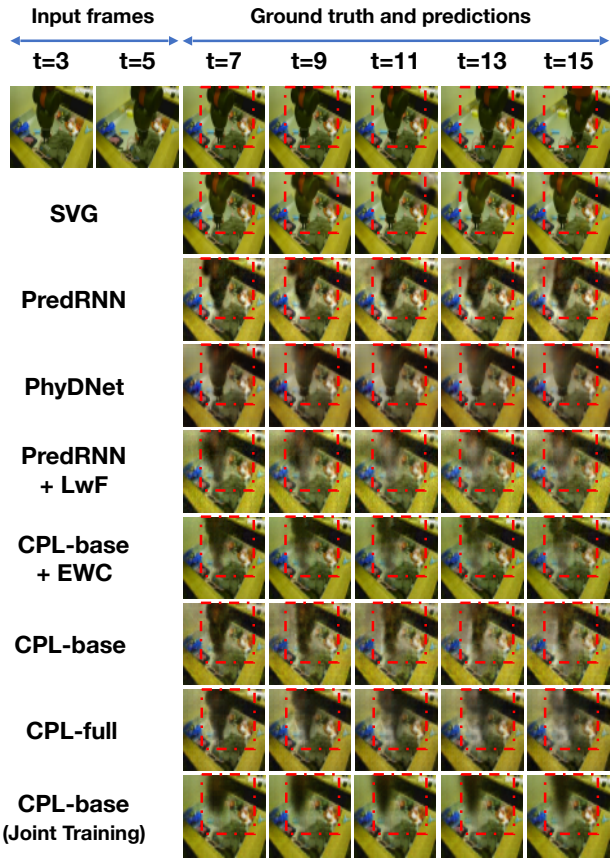

Figure 6. Showcases of predicted frames on the action-free RoboNet benchmark. The example is from the ***Penn*** environment, which is the **first** task in the process of continual learning. We use models that have finished the last training period on *Stanford*.

Our approach shows remarkable improvements over the state-of-the-art video prediction model (*i.e.*, PhyDNet), as well as the continual learning methods (*i.e.*, LwF and EWC). Notably, none of the compared models can "remember" the motion on previous tasks, especially for the *Clapping* and *Waving* tasks that have long gone. While our CPL approach is the only one that shows the ability to effectively mitigate catastrophic forgetting and generate **CORRECT** motions from corresponding observation frames.

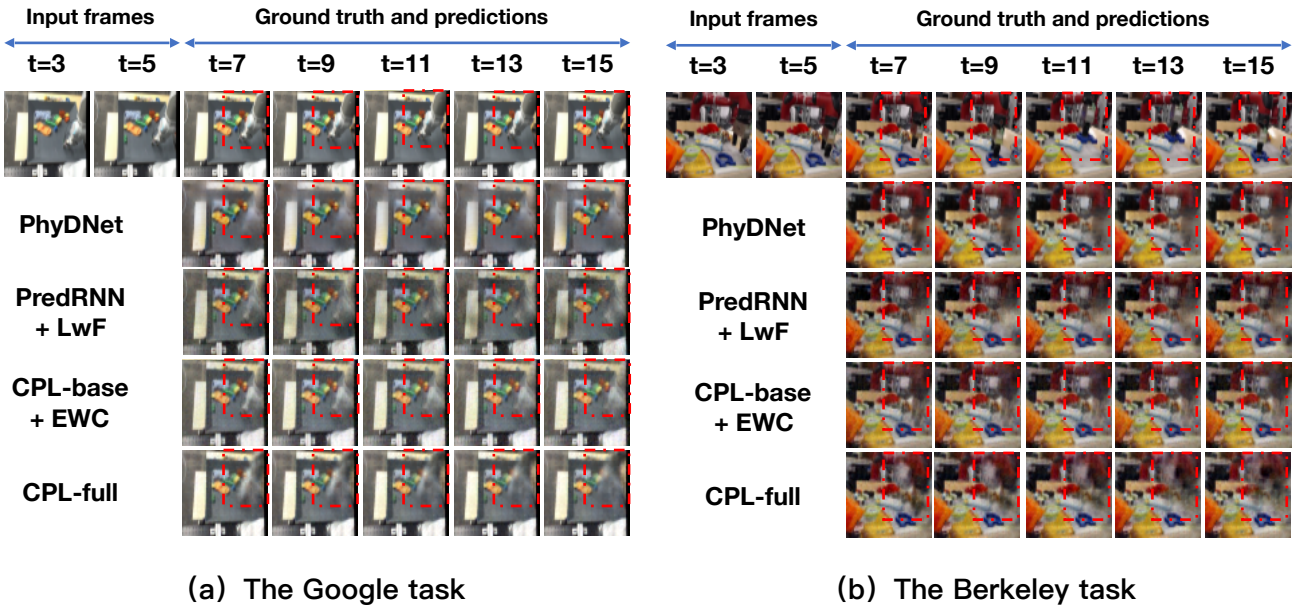

Figure 7. Action-free prediction results from (a) the ***Google*** environment and (b) the ***Berkeley*** environment, which are respectively the **second** and the **third** task in continual learning setup on RoboNet. The task order at training time is *Penn* → *Google* → *Berkeley* → *Stanford*. We use models that have finished the last training period on *Stanford*. For the first task of *Penn*, please refer to Fig. 6.

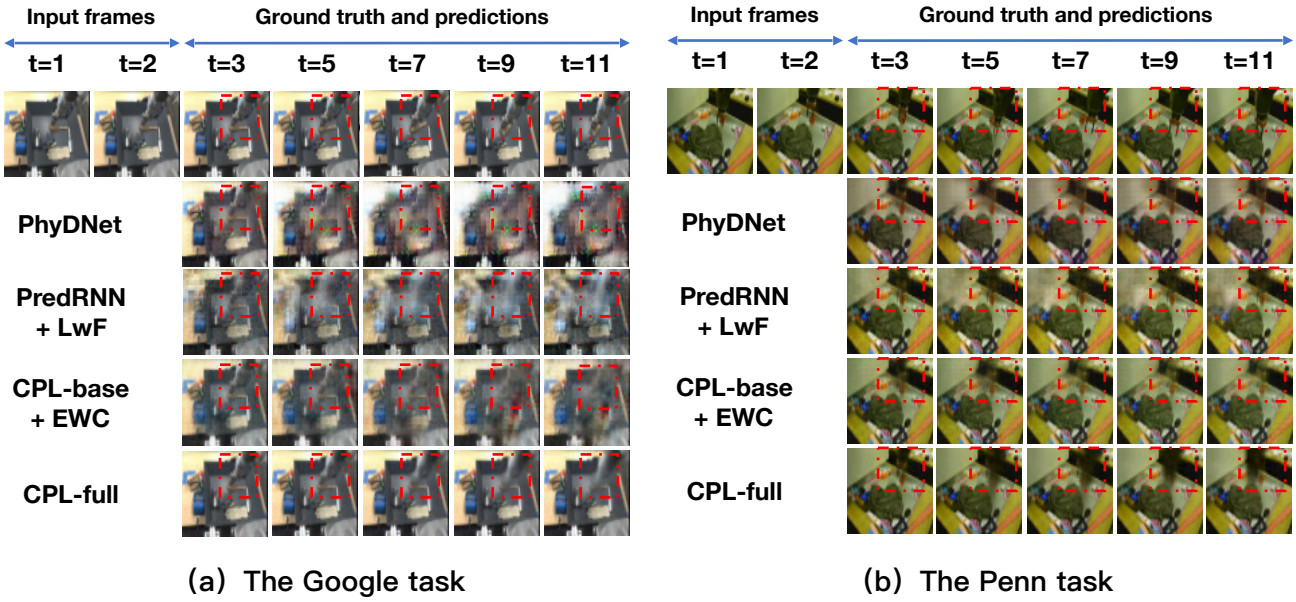

Figure 8. Action-conditioned prediction results from (a) the ***Google*** environment and (b) the ***Penn*** environment, which are respectively the **second** and the **third** task in continual learning setup on RoboNet. The task order at training time is *Berkeley* → *Google* → *Penn* → *Stanford*. We use models that have finished the last training period on *Stanford*. For the first task of *Berkeley*, please refer to Fig. 4 in the manuscript.

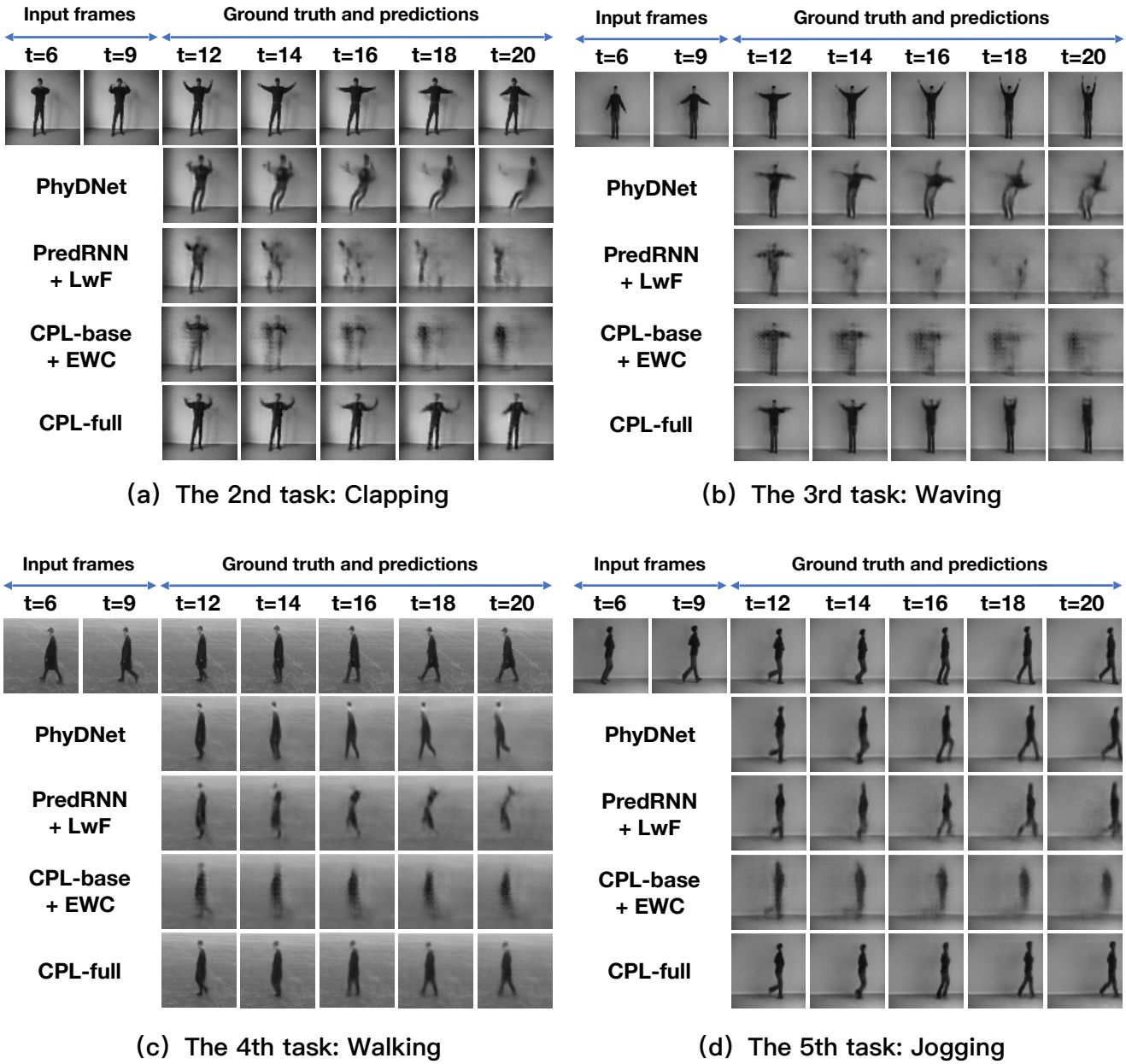

Figure 9. Video prediction results on the previous tasks before *Running*. We use models that have finished the last training period on *Running*. For the first task of *Boxing*, please refer to Fig. 6 in the manuscript. Note that our CPL approach is the only one that shows the ability to effectively mitigate catastrophic forgetting and generate **CORRECT** motions from corresponding observation frames.