RoboScape-R: Unified Reward-Observation World Models for Generalizable Robotics Training via RL

Supplementary Material

6. Broader Impacts

Our world model, as a scalable environment framework, offers positive value for training robotic policies. The world model environment can interact with actions output by the policy while providing observations and rewards for the next frame. Such rewards are unified, which enhances the generalization capability of the policy. This type of reward is "endogenous", which is derived from the world model's understanding of diverse tasks—to facilitate multi-task generalization learning of the policy.

7. Limitations

While our framework enables the world model to act as an environment for training generalizable policies, we acknowledge several limitations:(1) Our current framework lacks robust support for policy learning in long-horizon and complex tasks. As our world model adopts an autoregressive architecture, it can only achieve stable rollout within 300 frames when the window size is set to 48 frames. Exceeding this limit may lead to deteriorated quality and controllability of generated videos. This restricts our tasks to short-duration scenarios, meaning we cannot yet accommodate long-range, complex tasks such as folding clothes. (2) Our framework relies on the empirical assumption that the world model has fully learned the dynamic transitions of the real world. However, this assumption hinges on the fundamental performance of the world model itself.

8. Supplemented Evaluation Results

8.1. Task Setting for In-domain and Out-of-domain Evaluation

In the evaluation part, we conduct both the in-domain and out-of-domain evaluation. We display the task setting in Fig. 6. For in-domain evaluation, we train the policy in one environment and evaluate it in the same environment but with different initial states. For out-of-domain evaluation, we train the policy in several environments and evaluate it in different environments with seen objects and containers, but with different combinations. For example, we train the policy in "pick up the lemon and place it in the plate" and "pick up the peach and place it in the bowl", and we evaluate the policy in "pick up the lemon and place it in the bowl" and "pick up the peach and place it in the plate".

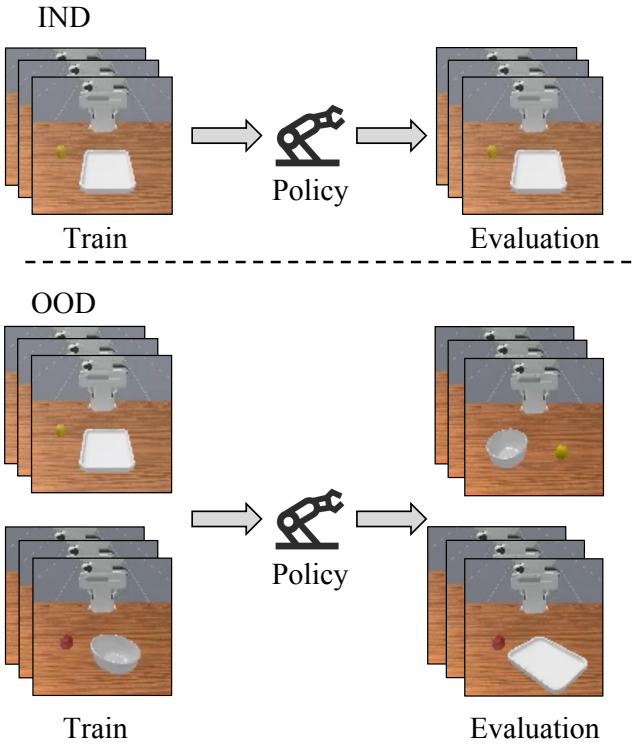


Figure 6. Task setting for in-domain and out-of-domain evaluation. For in-domain evaluation, training and evaluation are in the same environment, while for out-of-domain evaluation are in different environments.

8.2. Supplemented Description for the Self-Collected Dataset

In our experiment, we have collected a dataset from ManiSkill [38]. Specifically, we select 4 tasks, including pick and place, push, pull, and move to aim. For each task, we select 2 tables, 2 containers, and 20 objects to collect the data. A schematic diagram of data collection is presented in Fig. 7. In order to enable a more comprehensive learning of the action space for the world model to learning the dynamics, we also collect both optimal and suboptimal trajectories. The detail can be found in Sec. 4.1, and we display the representative trajectories for each task in Fig. 8.

8.3. Evaluation for the World Model Controlability

Utilizing the world model as an RL environment also poses a challenge to the observation generation quality for the world model, mainly about the action controllability and the robustness to out-of-domain actions. This is due to that the world model is trained in a collected dataset, which indi-

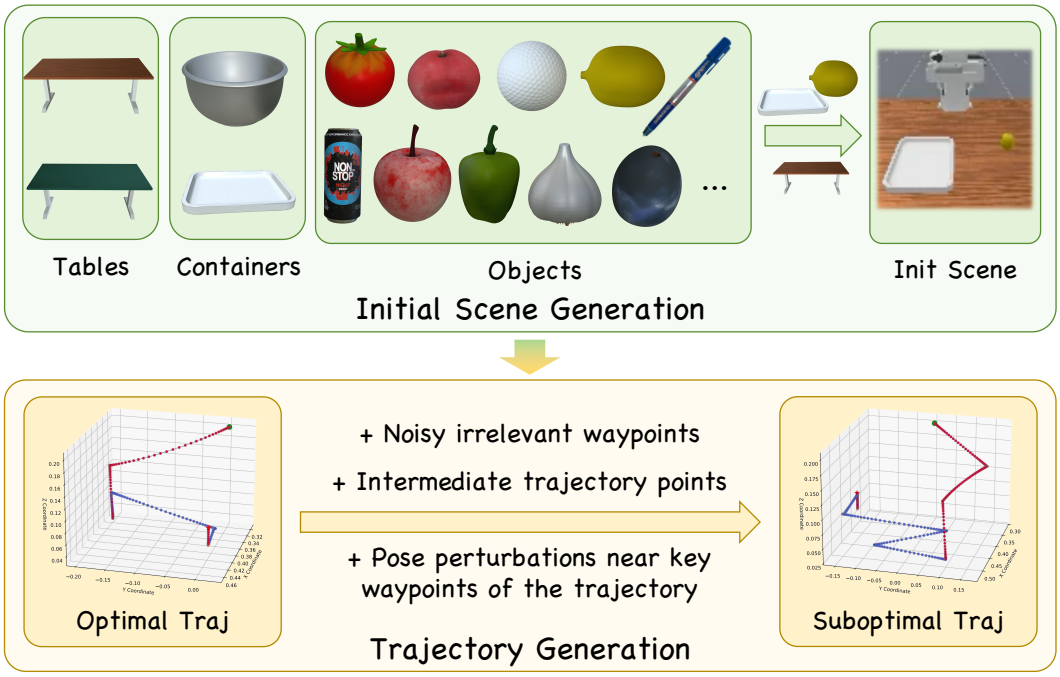
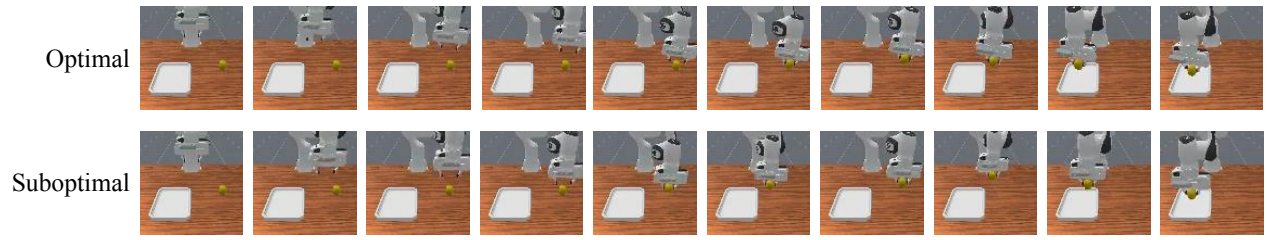
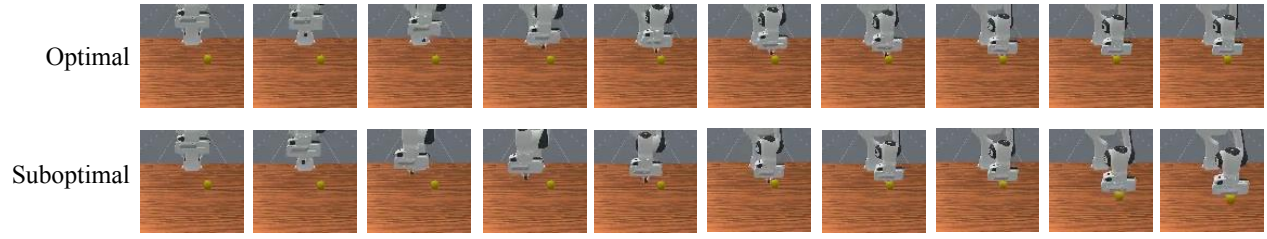


Figure 7. Diagram for the data collection pipeline. We first select tables, containers, and objects to create the initial scene, then we use motion planning to generate the optimal trajectory. We then modify the key waypoints and add the perturbation, and use motion planning to generate the suboptimal trajectories.

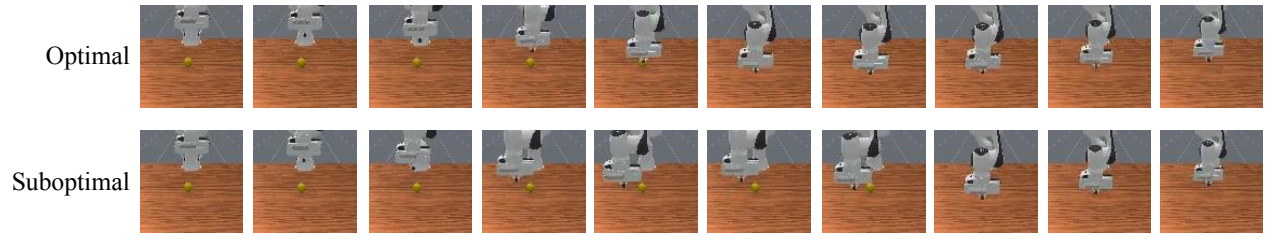
cates a discrete and limited action space, while the policy may generate an extreme action, especially at the beginning stage. As shown in Fig. 9, our world model is able to respond to extreme actions due to the promoted cross-attention-based action injection and the comprehensive pre-training data, while other world models may suffer from meaningless observation generation due to the extreme action sequence.



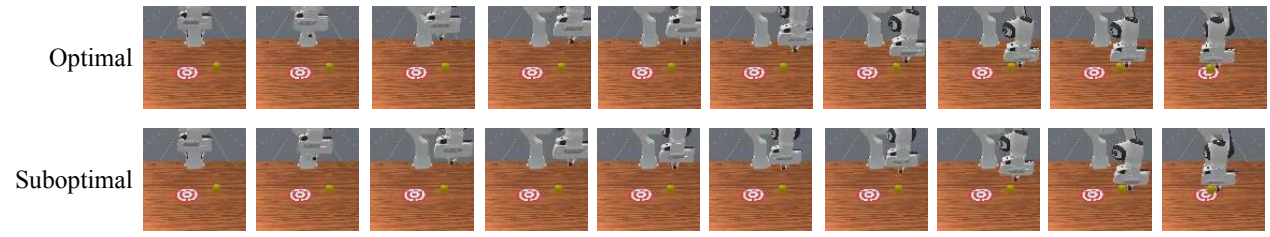
(a) Pick and Place (P.&P.)



(b) Push



(c) Pull



(d) Move to Aim (M.A.)

Figure 8. Visualization for optimal and suboptimal trajectories for different tasks.

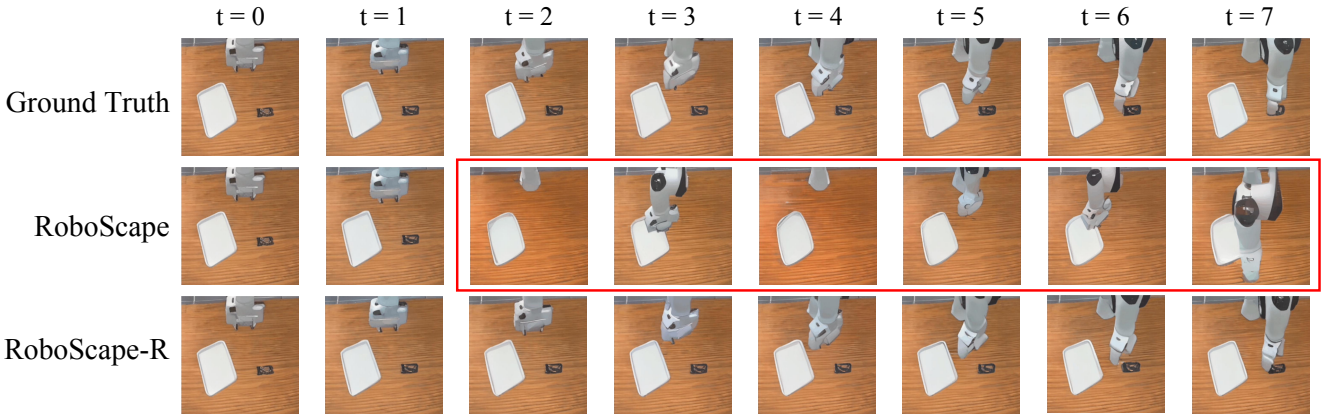


Figure 9. Visualization when the world model responds to an unseen trajectory in the out-of-domain environments.