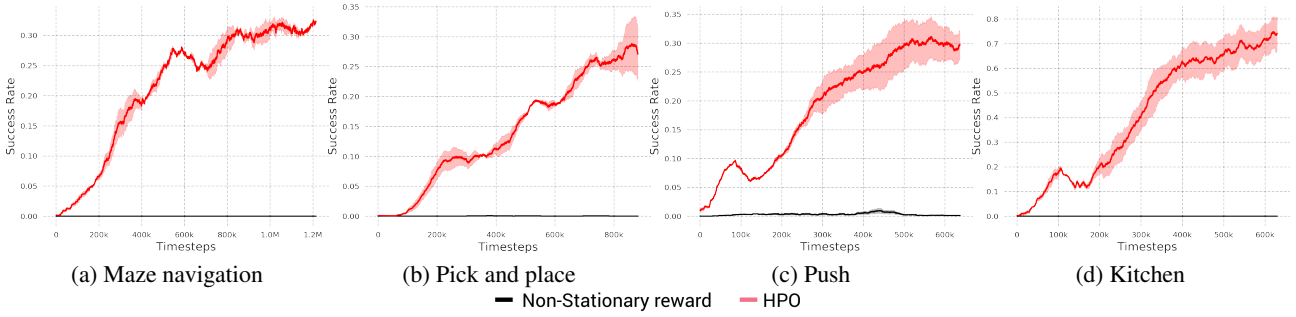
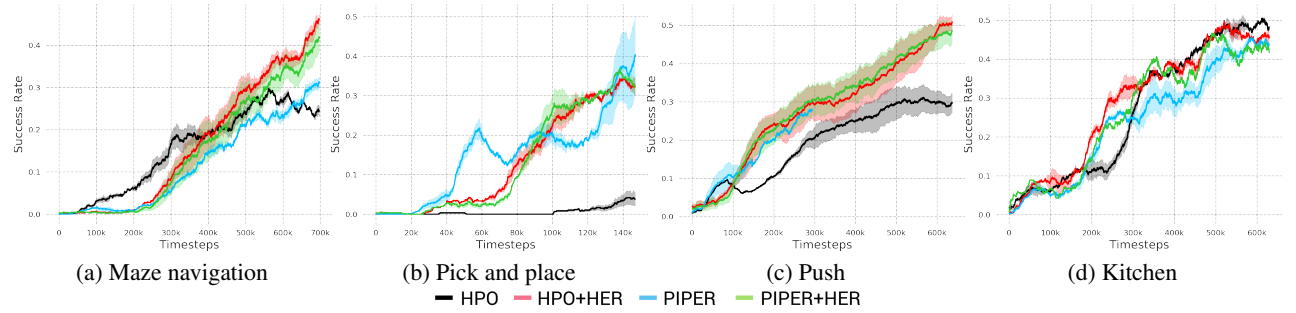
Additional Experiments



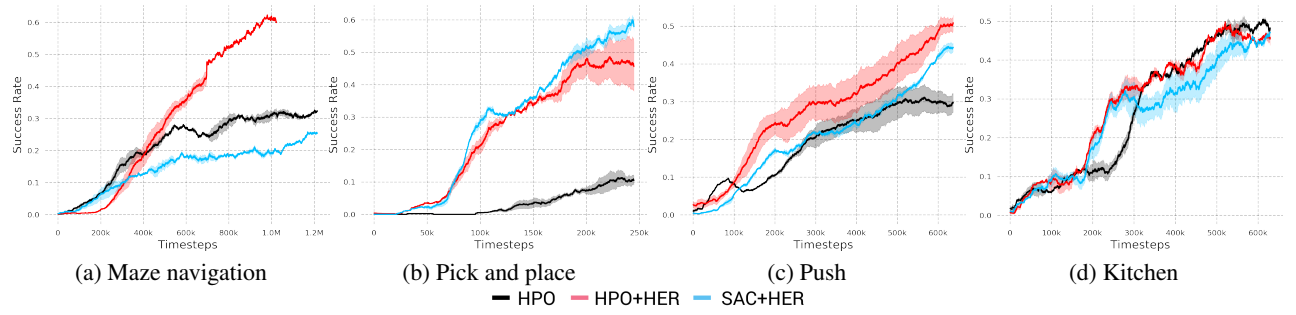
(a) Maze navigation (b) Pick and place (c) Push (d) Kitchen

Figure 1: **Comparison with non-stationary reward function.** This figure illustrates the success rate comparison across four sparse-reward maze navigation and robotic manipulation tasks. We evaluate HPO against the hierarchical approach trained with non-stationary reward function. As seen from figure, HPO demonstrates strong performance and significantly outperforms the baseline.



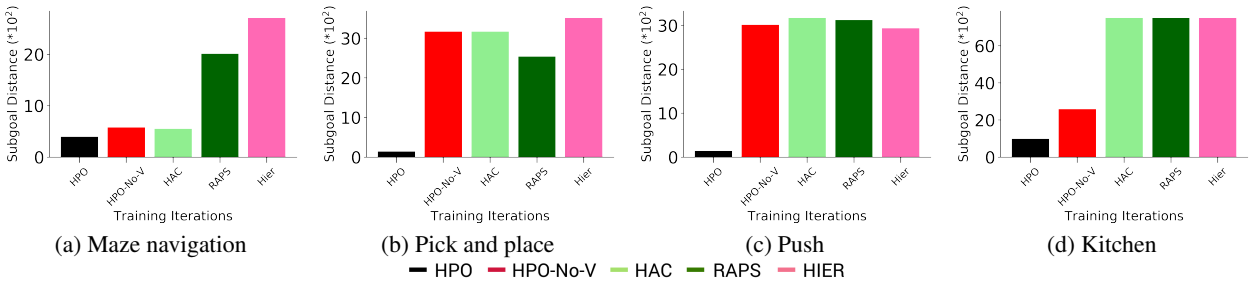
(a) Maze navigation (b) Pick and place (c) Push (d) Kitchen

Figure 2: **Comparison with PIPER.** This figure illustrates the success rate comparison across four sparse-reward maze navigation and robotic manipulation tasks. We evaluate HPO and HPO+HER (HPO with Hindsight Experience Replay) approach against PIPER and PIPER+HER baseline. As seen from figure, although PIPER outperforms HPO in maze, pick and place and push tasks, HPO+HER demonstrates impressive performance and outperforms the PIPER PIPER+HER baselines on all tasks.



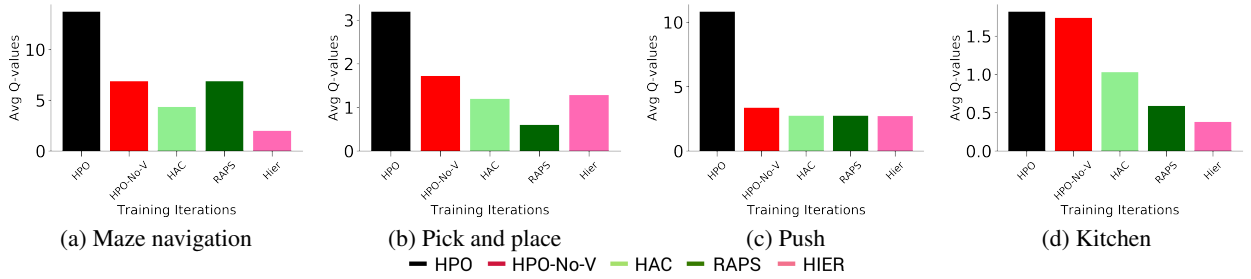
(a) Maze navigation (b) Pick and place (c) Push (d) Kitchen

Figure 3: **Comparison with SAC+HER.** This figure illustrates the success rate comparison across four sparse-reward maze navigation and robotic manipulation tasks. We evaluate HPO and HPO+HER (HPO with Hindsight Experience Replay) approach against SAC+HER. As seen from figure, although SAC+HER outperforms HPO in pick and place and push tasks, HPO+HER demonstrates impressive performance and outperforms the SAC+HER baseline on maze, push and kitchen tasks.



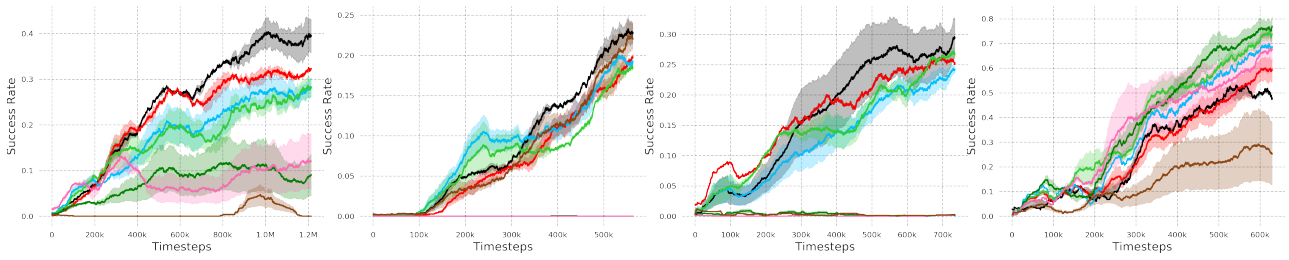
(a) Maze navigation (b) Pick and place (c) Push (d) Kitchen

Figure 4: **Non-stationarity ablation.** This figure compares `HPO` with `HPO-No-V`, `HAC`, `RAPS`, `HIER` baselines, based on average distance between subgoals predicted by the higher level policy and subgoals achieved by the lower level primitive. `HPO` consistently generates low average distance values, which implies that in `HPO`, the higher level policy generates achievable subgoals that induce optimal lower primitive goal reaching behavior. This shows that HPO is able to address non-stationary in HRL and generate feasible subgoals.



(a) Maze navigation (b) Pick and place (c) Push (d) Kitchen

Figure 5: **Q-value ablation.** This figure compares `HPO` with `HPO-No-V`, `HAC`, `RAPS`, `HIER` baselines, based on average lower level Q function values for the subgoals predicted by the higher level policy. `HPO` consistently large Q-function values, which implies that in `HPO`, the higher level policy generates achievable subgoals. This shows that HPO is able to address non-stationary in HRL and generate feasible subgoals.

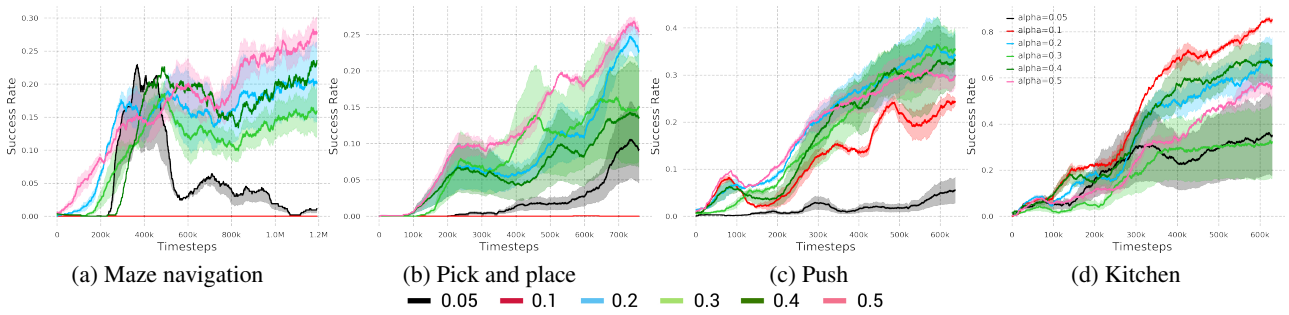


(a) Maze navigation (b) Pick and place (c) Push (d) Kitchen

Figure 6: **Regularization weight ablation.**This figure depicts the success rate performance for varying values of the primitive regularization weight λ. When λ is too small, we loose the benefits of primitive-informed regularization resulting in poor performance, whereas too large λ values can lead to degenerate solutions.

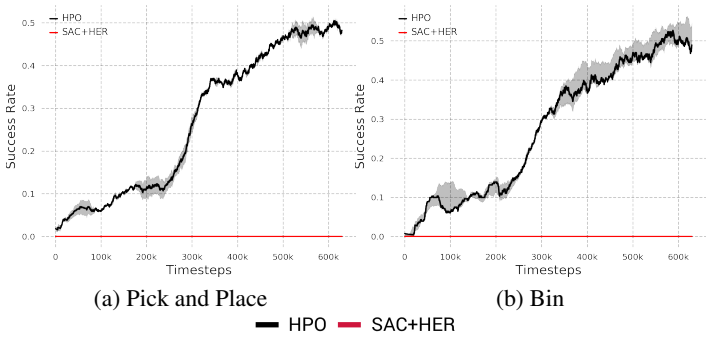
Table 1: Real-world success performance comparison

Method	Pick & Place	Bin Task
HPO	**0.5**	**0.4**
SAC+HER	0.2	0.2



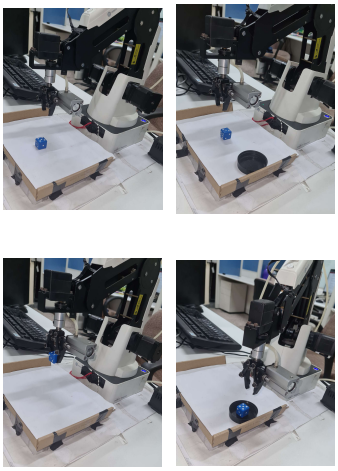
(a) Maze navigation (b) Pick and place (c) Push (d) Kitchen

Figure 7: **Max-ent parameter ablation.** This figure illustrates the success rate performance for different values of the max-ent parameter β hyper-parameter. This parameter controls the exploration in maximum-entropy formulation. If β is too large, the higher-level policy may perform extensive exploration but stay away from optimal subgoal prediction, whereas if β is too small, the higher-level might not explore and predict sub-optimal subgoals.



(a) Pick and Place (b) Bin

Figure 8: **Comparison with `SAC+HER` on hard kitchen task.** This figure illustrates the success rate comparison for franka kitchen tasks. We evaluate `HPO` vs `SAC+HER`. Column 1 is the four step task "open microwave door, put kettle on burner, turn on the burner, slide cabinet", and Column 2 is the four step task "open microwave door, put kettle on burner, turn on the light, slide cabinet". As seen from figure, `HPO` outperforms `SAC+HER` on complex long-horizon tasks, which displays the advantage of our hierarchical preference optimization formulation over single-level SAC+HER approach.



(a) Pick (b) Bin

Figure 9: **Real world tasks** This figure depicts two real world environments in our experiments: (a) pick and place and (b) bin tasks. In pick and place, the robotic gripper has to pick the block and bring it to goal position. In bin task, the gripper has to pick the block and place it in the bin. Row 1 depicts initial state and Row 2 depicts final goal configurations.