Phasic Self-Imitative Reduction for Sparse-Reward Goal-Conditioned Reinforcement Learning

Yunfei Li [*1] **Tian Gao** [*1] **Jiaqi Yang** [2] **Huazhe Xu** [3] **Yi Wu** [1 4]

Abstract

It has been a recent trend to leverage the power of supervised learning (SL) towards more effective reinforcement learning (RL) methods. We propose a novel phasic approach by alternating online RL and offline SL for tackling sparse-reward goal-conditioned problems. In the online phase, we perform RL training and collect rollout data while in the offline phase, we perform SL on those successful trajectories from the dataset. To further improve sample efficiency, we adopt additional techniques in the online phase including task reduction to generate more feasible trajectories and a value- difference-based intrinsic reward to alleviate the sparse-reward issue. We call this overall algorithm, *PhAsic self-Imitative Reduction* (PAIR). PAIR substantially outperforms both non-phasic RL and phasic SL baselines on sparse-reward goal-conditioned robotic control problems, including a challenging stacking task. PAIR is the first RL method that learns to stack *6 cubes* with only *0/1 success rewards* from scratch.

1. Introduction

Despite great advances achieved by deep reinforcement learning (RL) in a wide range of application domains such as playing games (Mnih et al., 2015; Schrittwieser et al., 2020), controlling robots (Lillicrap et al., 2016; Hwangbo et al., 2019; Akkaya et al., 2019), and solving scientific problems (Jeon & Kim, 2020), deep RL methods have been empirically shown to be brittle and extremely sensitive to hyper-parameter tuning (Tucker et al., 2018; Ilyas et al., 2020; Engstrom et al., 2020; Yu et al., 2021a; Andrychowicz et al., 2021), which largely limits the practice use of deep RL in many real-world scenarios. On the contrary, supervised learning (SL) provides another learning paradigm by imitating given demonstrations, which is much simpler for tuning and typically results in a much more steady optimization process (Lynch et al., 2020; Ghosh et al., 2020). Inspired by the success of training powerful fundamental models by SL (Brown et al., 2020; Dosovitskiy et al., 2020; Jumper et al., 2021), it has also been a recent trend in RL to leverage the power of SL to develop more powerful and stable deep RL algorithms (Levine, 2021).

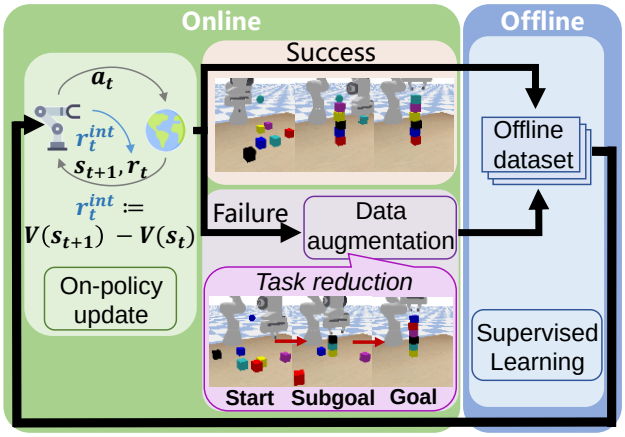


Figure 1. Overall workflow of PAIR. PAIR iteratively alternates between online RL and offline SL phases. During online phases, the agent is trained using both environment reward and a value-difference-based intrinsic reward. Meantime, it collects trajectories and augments them with task reduction into successful demonstrations for offline training. In the offline phases, the agent runs supervised learning to improve its policy using the generated dataset.

One representative line of research that incorporates SL into RL is offline RL, which assumes that a large offline dataset of transition data is available and solely performs learning on the dataset without interacting with the environment (Lange et al., 2012; Wu et al., 2019; Levine et al., 2020; Kumar et al., 2020; Fujimoto & Gu, 2021). However, both empirical (Yu et al., 2021b) and theoretical (Rashidinejad et al., 2021) evidence suggests that the success of recent offline RL methods rely on the quality of the dataset. Accord-

[*]Equal contribution [1]Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China [2]Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA, USA [3]Stanford University, CA, USA [4]Shanghai Qi Zhi Institute, Shanghai, China. Correspondence to: Yunfei Li <liyf20@mails.tsinghua.edu.cn>, Yi Wu <jxwuyi@gmail.com>.

Proceedings of the 39th International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

ingly, popular offline RL datasets are typically constructed by human experts, which may not be feasible for many real-world problems. Besides, offline RL suffers from a final performance gap compared with online RL algorithms.

Another representative line of research is self-imitation learning (SIL) (Oh et al., 2018), which combines SL and RL in a completely online fashion. SIL directly performs SL over selected self-generated rollout trajectories by treating the SL objective as an auxiliary loss and optimizing it jointly with the standard RL objective. SIL does not require a dataset in advance. However, optimizing a mixed objective consisting of both RL and SL makes the optimization process even more brittle and requires substantial efforts of parameter tuning to achieve best empirical performance.

In this paper, we propose a simple phasic approach, *PhAsic self-Imitative Reduction* (PAIR), to effectively balance both RL and SL training for goal-conditioned sparse-reward problems. The main principle of PAIR is to alternate between RL and SL: in the RL phase, we solely perform standard online RL training for optimization stability and collect rollout trajectories as a dataset for the offline SL phase; while in the SL phase, we pick out successful trajectories as SL signals and run imitation learning to improve policy. To improve sample efficiency, PAIR also includes two additional techniques in the RL phase: 1) a value-difference-based intrinsic reward that alleviates the sparse-reward issue, and 2) *task reduction*, a data augmentation technique that largely increases the total number of successful trajectories, especially for hard compositional tasks. Our theoretical analysis suggests that task reduction can converge *exponentially* faster than the vanilla phasic approach that does not use task reduction.

We implement PAIR with Proximal Policy Optimization (PPO) for RL and behavior cloning for SL and we conduct experiments on a variety of goal-conditioned control problems, including relatively simple benchmarks such as pushing and ant-maze navigation, and a challenging sparse-reward cube-stacking task. PAIR substantially outperforms all the baselines including non-phasic online methods, which jointly perform self-imitation and RL training, and phasic SL methods, which only perform supervised learning on self-generated trajectories (Ghosh et al., 2020). We highlight that PAIR successfully learns to stack 6 cubes with 0/1 rewards from scratch. To our knowledge, PAIR is the *first* deep RL method that could solve this challenging sparse reward task.

2. Related Work

Goal-conditioned RL: We study goal-conditioned reinforcement learning (Kaelbling, 1993) with sparse reward in this work. Goal-conditioned RL enables one agent to solve a variety of tasks by predicting actions given both observations and goals, and is studied in a number of works (Schaul et al., 2015; Nair et al., 2018b; Pong et al., 2018; Veeriah et al., 2018; Zhao et al., 2019; Eysenbach et al., 2020). Although some techniques like relabeling (Andrychowicz et al., 2017) are proposed to address the sparse reward issue when learning goal-conditioned policies, there are still challenges in long-horizon problems (Nasiriany et al., 2019).

Offline reinforcement learning: Offline RL (Lange et al., 2012; Levine et al., 2020) is a popular line of research that incorporates SL into RL, which studies extracting a good policy from a fixed transition dataset. A large portion of offline RL methods focus on regularized dynamic programming (e.g., Q-learning) (Kumar et al., 2020; Fujimoto & Gu, 2021), with the constraint that the resulting policy does not deviate too much from the behavior policy in the dataset (Wu et al., 2019; Peng et al., 2019; Kostrikov et al., 2021). Some other works directly treat policy learning as a supervised learning problem, and learn the policy in a conditioned behavior cloning manner (Chen et al., 2021; Janner et al., 2021; Furuta et al., 2021; Emmons et al., 2021), which can be considered as special cases of upside down RL and reward-conditioned policies (Schmidhuber, 2019; Kumar et al., 2019). There are also methods that learn transition dynamics from offline data before extracting policies in a model-based way (Matsushima et al., 2020; Kidambi et al., 2020). Some works also consider a single online fine-tuning phase after offline learning (Nair et al., 2020; Lu et al., 2021; Mao et al., 2022; Uchendu et al., 2022) while we repeatedly alternate between offline and online training.

Imitation learning in RL: Imitation learning is a framework for learning policies from demonstrations, which has been shown to largely improve the sample complexity of RL methods (Hester et al., 2018; Rajeswaran et al., 2018) and help overcome exploration issues (Nair et al., 2018a). Self-imitation learning (SIL) (Oh et al., 2018), which imitates good data rolled out by the RL agent itself, does not require any external dataset and has been shown to help exploration in sparse reward tasks. Our method also performs imitation learning over self-generated data. Self-imitation objective is optimized jointly with the RL objective, while we propose to perform SL and RL separately in two phases. The idea of substituting joint optimization with iterative training for minimal interference between different objectives is similar to phasic policy gradient (Cobbe et al., 2021). Goal-conditioned supervised learning (GCSL) (Ghosh et al., 2020) is perhaps the most related work to ours. GCSL repeatedly performs imitation learning on self collected relabeled data without any RL updates while we leverage the power from both RL and SL and adopt further task reduction for enhanced data augmentation.

Sparse reward problems: There are orthogonal interests in solving long horizon sparse reward with hierarchical

modeling (Stolle & Precup, 2002; Kulkarni et al., 2016; Bacon et al., 2017; Nachum et al., 2018) or by encouraging exploration (Singh et al., 2005; Burda et al., 2019; Badia et al., 2020; Ecoffet et al., 2021). Our framework can be also viewed as an effective solution to tackle challenging sparse-reward compositional problems.

3. Preliminary

We consider the setting of goal-conditioned Markov decision process with 0/1 sparse rewards defined by $(\mathcal{S}, \mathcal{A}, P(s'|s,a), \mathcal{G}, r(s,a,g), \rho_0, \gamma)$. $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{G}$ is the goal space and γ is the discounted factor. $P(s'|s,a)$ denotes transition probability from state s to s' after taking action a. The reward function $r(s,a,g)$ is 1 only if the goal g is reached at the current state s within some precision threshold and otherwise 0. In the beginning of each episode, the initial state s_0 is sampled from a distribution ρ_0 and a goal g is sampled from the goal space. An episode terminates when the goal is achieved or it reaches a maximum number of steps.

The agent is represented as a goal-conditioned stochastic policy $\pi_\theta(a|s,g)$ parametrized by θ. The optimal policy $\pi_{\theta^\star}$ should maximize the objective $J(\theta)$ defined by the expected discounted cumulative reward over all the goals, i.e.,

$$J(\theta) = J(\pi_\theta) = \mathbb{E}_{g\in\mathcal{G}, a_t\sim\pi_\theta}\left[\sum_t \gamma^t r(s_t, a_t, g)\right]. \quad (1)$$

Policy gradient optimizes $J(\theta)$ via the gradient computation

$$\nabla J(\theta) = \mathbb{E}_{g,a_t}\left[\sum_t (R_t - V_\psi(s_t,g))\nabla \log \pi(a_t|s_t,g)\right],$$

where $V_\psi(s_t,g)$ is the goal-conditioned value function parameterized by ψ and R_t denotes the discounted return starting from time t on the current trajectory. Note that in our goal-conditioned setting with 0/1 rewards, R_t will be either[1] 0 or 1 and the value function $V_\psi(s_t,g)$ can be approximately interpreted as the discounted success rate from state s_t towards goal g.

Goal-conditioned imitation learning optimizes a policy π_θ by running supervised learning over a given demonstration dataset $\mathcal{D} = \{\tau : (g; s_0, a_0, s_1, a_1, \ldots)\}$, where τ denotes a single trajectory. The supervised learning loss $L(\theta)$ is typically defined by

$$L(\theta) = -\mathbb{E}_{(g;s,a)\in\mathcal{D}}\left[w(s,a,g)\log\pi(a|s,g)\right], \quad (2)$$

where $w(s,a,g)$ is some sample weight. Behavior cloning (BC) simply sets $w(s,a,g)$ as 1 while more advanced methods may have other choices. Note that self-imitation learning (SIL) is a paradigm that jointly optimizes the RL objective $J(\theta)$ and the SL objective $L(\theta)$ in an online fashion.

[1] More precisely, the return R_t will be 0 or close to 1 due to the discount factor γ.

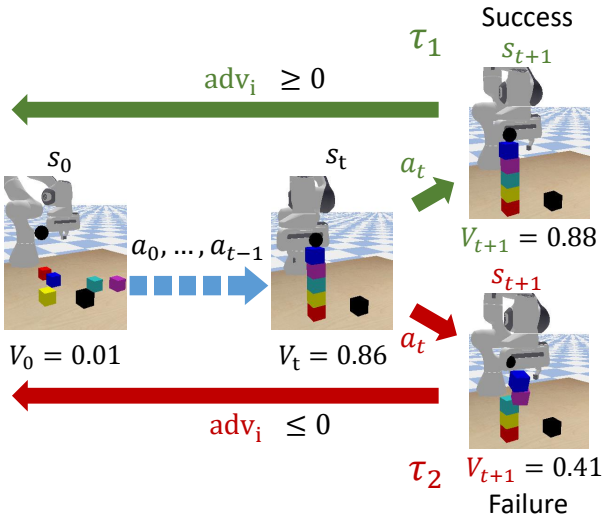


Figure 2. Motivation of value-difference intrinsic reward. Trajectories τ_1 and τ_2 only differ at a single action a_t. τ_1 succeeds and will have positive advantage. τ_2 fails since a_t knocks down the cubes, then all the transitions in τ_2 will have negative advantages due to 0/1 goal-conditioned reward, even if early actions are perfect.

4. Method

Our phasic solution PAIR consists of 3 components, including the online RL phase (Sec. 4.1), which also collects rollout trajectories, task reduction (Sec. 4.2) as a mean for data augmentation, and the offline phase (Sec. 4.3), which performs SL on self-generated demonstrations. We summarize the overall algorithm in Sec. 4.4.

4.1. RL Phase with Intrinsic Rewards

The RL phase follows any standard online RL training. Specifically in our work, we adopt PPO (Schulman et al., 2017) as our RL algorithm, which trains both policy π_θ and value function V_ψ using rollout trajectories. When a trajectory $\tau = (g; s_t, a_t)$ is successful, i.e., goal g is reached, we keep this trajectory τ as a positive demonstration towards goal g in the dataset $\mathcal{D}$ for the offline phase.

Value-difference-based intrinsic rewards:

The sparse-reward issue is a significant challenge for online RL training, which can yield substantially high variance in gradients. In particular, let's assume $\gamma = 1$ for simplicity and consider a single trajectory $\tau = (g; s_t, a_t)$. In our goal-conditioned setting, the return R_t on τ will be binary and the value function $V_\psi(s_t,g)$ will be approximately between 0 and 1. Hence, the advantage function for state-action pair

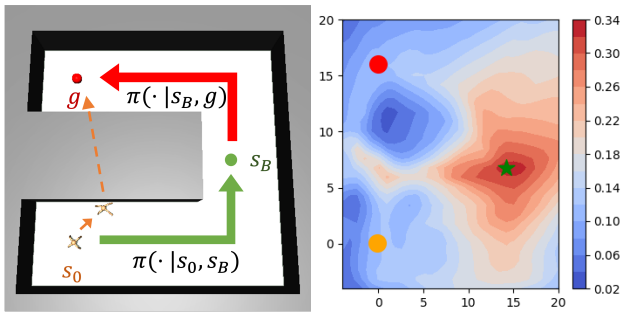


Figure 3. Illustration of task reduction. *(Left)* Given a trajectory that fails to reach g from s_0 (orange arrows), task reduction searches for an intermediate s_B, then executes the resulting two sub-tasks (green and red arrows) to generate a successful demonstration. *(Right)* Heatmap of composed value in Eq. 4 w.r.t. s_B. The green star denotes the position of optimal s_B.

$(s_t, a_t) \in \tau$, i.e., $R_t - V(s_t)$, will be either positive or negative for the entire trajectory τ. Imagine a concrete example as illustrated in Fig. 2, where we have two trajectories, a successful trajectory τ_1 and a failed trajectory τ_2. These two trajectories only differ at the final timestep s_{t+1}, which may be due to some random exploration noise at action a_t. However, only due to a single mistake, all the state-action pairs from τ_2 will have negative advantages, even though most of the actions in τ_2 are indeed approaching the desired target. Likewise, in a successful trajectory, it is also possible that some single action is poor but eventually the subsequent actions fix this early mistake.

Besides the trajectory-based advantage computation, it will be beneficial to have some effective signal of whether a transition (s_t, a_t, s_{t+1}) is properly "*approaching*" the desired goal g or not. Our suggestion is to use $V_\psi(s, g)$ as an empirical measure: if a_t is a good action, it should lead to a higher state value, i.e., $V_\psi(s_{t+1}, g) - V_\psi(s_t, g) > 0$, which indicates that a_t moves to a state with a higher success rate; similarly, a poor action will result in a value drop, i.e., $V_\psi(s_{t+1}, g) - V_\psi(s_t, g) < 0$. Accordingly, we propose to adopt value difference as an intrinsic reward r^{int} for stabilizing goal-conditioned RL training as follows

$$r^{\text{int}}(s_t, a_t, g) := V_\psi(s_{t+1}, g) - V_\psi(s_t, g).$$

We remark that r^{int} relies on an accurately learned value function $V_\psi(s, g)$. Therefore, we suggest to only train $V_\psi(s, g)$ over the sparse goal-conditioned rewards while using another value head to fit the intrinsic rewards for critic learning. Similar techniques have been previously explored in (Burda et al., 2019) as well.

4.2. Task Reduction as Data Augmentation

In order to perform effective supervised learning in the offline phase, it is critical that the online phase can generate as much successful trajectories as possible. In our framework, we consider two data augmentation techniques to boost the positive samples in the dataset $\mathcal{D}$, i.e., *goal relabeling* and *task reduction*. We will first describe the simpler one, goal relabeling, before moving to a much more powerful technique, task reduction.

Goal relabeling was originally proposed by Andrychowicz et al. (2017). In our goal-conditioned learning setting, for each failed trajectory $\tau = (g; s_t, a_t)$ originally targeted at goal g, we can create an artificial goal g' by setting $g' = s_j$ for some reached state $s_j \in \tau$, which naturally yields a successful trajectory τ' as follows:

$$\tau' \leftarrow (g' = s_j; s_{t=0:j}, a_{t=0:j}) \quad \text{where} \quad s_j \in \tau. \tag{3}$$

Therefore, despite its simplicity, goal relabeling can convert every failed trajectory τ to a positive demonstration without any further interactions with the environment.

Task reduction was originally proposed by Li et al. (2020b). The main idea is to decompose a challenging task into a composition of two simpler sub-tasks so that both sub-tasks can be solved by the current policy. As illustrated in the left part of Fig. 3, given a goal g from state s_0, task reduction searches for the best sub-goal $s_B^\star$ through a 1-step planning process over the universal value function $V_\psi(s, g)$ as follows

$$s_B^\star = \arg\max_{s_B} V_\psi(s_0, s_B) \oplus V_\psi(s_B, g), \tag{4}$$

where $\oplus$ is a composition operator typically implemented as multiplication. Since the value function is learned, such a search process can be accomplished without any further environment interactions by gradient descent or cross-entropy method. A heatmap of the composed value for sub-goal search (Eq.(4)) is visualized in the right part of Fig. 3. Notably, after $s_B^\star$ is obtained, we would still need to execute the policy in the environment following the sub-goal $s_B^\star$ and then the final goal g in order to obtain a valid demonstration. Compared to goal relabeling, task reduction consumes additional samples and may even fail to produce a successful trajectory when either of the two sub-tasks fails. Hence, task reduction can be expensive in the early stage of training when the policy and value function have not yet been well trained. However, we will show both theoretically (Sec. 5) and empirically (Sec. 6.2) that task reduction can lead to an exponentially faster convergence compared to using goal relabeling solely in long-horizon problems.

By default, PAIR utilizes both goal relabeling and task reduction for data augmentation unless otherwise stated.

4.3. Offline SL Phase

After a dataset $\mathcal{D}$ of successful demonstrations, including augmented trajectories, is collected, we switch from RL training to offline SL by performing advantage weighted

Algorithm 1 Phasic Self-Imitative Reduction

Initialize: goal-conditioned policy $\pi_\theta(\cdot|s,g)$, value function $V_\psi(s,g)$.
for $k \leftarrow 1, 2, \cdots$ **do**
  Sample a batch of trajectories $\mathcal{B} \leftarrow \{\tau : (g; s_t, a_t)\}$ w.r.t. the current policy π_θ
  Update π_θ, V_ψ by RL on $\mathcal{B}$ (Sec. 4.1)
  Set dataset $\mathcal{D} \leftarrow \emptyset$
  for $\tau \in \mathcal{B}$ **do**
    if not `is_success`(τ) **then**
      $\tau \leftarrow$ `data_augment`$(\tau, \pi_\theta, V_\psi)$ (Sec. 4.2)
    end if
    $\mathcal{D} \leftarrow \mathcal{D} \cup \{\tau\}$
  end for
  Train π_θ with SL over $\mathcal{D}$ (Sec. 4.3)
end for

behavior cloning (BC). In particular, we set the weight $w(s,a,g)$ in Eq. (2) to $\exp\left(\frac{1}{\beta}(R - V_\phi(s,g))\right)$ following (Peng et al., 2019). We remark that we only train policy in the offline phase while keeping the value function V_ψ unchanged for algorithmic simplicity. We also empirically find that training value function during the offline phase does not improve the overall performance.

It is feasible to adopt more advanced methods in the offline phase, such as offline RL methods, which typically assume a pre-constructed dataset but conceptually compatible with our phasic learning process. We conduct experiments by substituting BC with two popular offline RL methods, decision transformer (Chen et al., 2021) and AWAC (Nair et al., 2020), in Sec. 6.1. Empirical results show that these alternatives are much more brittle than BC and perform poorly without a high-quality warm-start dataset. Thus, we simply use BC as our SL method in this paper.

4.4. PAIR: Phasic Self-Imitative Reduction

By repeatedly performing the online RL phase with task reduction and the offline SL phase, we derive our final algorithm, PAIR. The pseudo-code is summarized in Alg. 1. More implementation details can be found in Appendix B.3. More analysis on the phasic training scheme vs. joint RL and SL optimization is in Appendix B.4.

5. Theoretical Analysis

First, we establish the correctness of our framework. The following theorems follows directly from the GCSL framework in (Ghosh et al., 2020), because both GCSL and PAIR are built upon SL.

Theorem 5.1 (Ghosh et al. (2020), Theorem 3.1)**.** *Let $J(\pi)$ be defined in Eq. (1) and $J_{\mathrm{PAIR}}(\pi) = -L(\theta)$ as defined in Eq. (2). Let $\tilde{\pi}$ be the data collection policy induced by data augmentation. Then*

$$J(\pi) \geq J_{\mathrm{PAIR}}(\pi) - 4T(T-1)\alpha^2 + C, \tag{5}$$

where $\alpha = \max_{s,g} D_{\mathrm{TV}}(\pi(\cdot|s,g) \| \tilde{\pi}(\cdot|s,g))$ and C is a constant independent of π.[2]

Theorem 5.2 (Ghosh et al. (2020), Theorem 3.2)**.** *Assume deterministic transition and that $\tilde{\pi}$ has full support. Define*

$$\epsilon := \max_{s,g} D_{\mathrm{TV}}(\pi(\cdot|s,g) \| \hat{\pi}^\star(\cdot|s,g)). \tag{6}$$

Then $J(\pi^\star) - J(\pi) \leq \epsilon \cdot T$, where $\hat{\pi}^\star$ minimizes $L(\theta)$ defined in Eq. (2) and $\pi^\star$ is the optimal policy.

Here, we present theoretical justification to illustrate why our algorithm is efficient on sparse-reward composite combinatorial tasks.

Theorem 5.3. *Under mild assumptions, the PAIR framework could use exponentially less number of iterations compared to the phasic framework that does not use task reduction, e.g., GCSL (Ghosh et al., 2020).*

We defer the exact statement of Theorem 5.3 and its proof to Appendix A.

6. Experiment

We aim to answer the following questions in this section:

- *Does phasic training of RL and SL objectives perform better than non-phasic joint optimization?*
- *Is PAIR compatible with offline RL methods other than BC?*
- *Does PAIR achieve exponential improvement over baselines* without *task reduction?*
- *Are all the algorithmic components of PAIR necessary for good performance?*
- *Can PAIR be used to solve challenging long-horizon sparse-reward problems, e.g., cube stacking?*

We consider 3 goal-conditioned control problems with increasing difficulty: (i) short-horizon robotic pushing (adopted from (Nair et al., 2018b)), (ii) ant navigation in a U-shaped maze, and (iii) robotic stacking with up to 6 cubes. All the tasks are with only 0-1 sparse reward.

We compare PAIR with non-phasic RL baslines that jointly perform RL and SL as well as phasic SL baselines that

[2] $D_{\mathrm{TV}}(\mu\|\nu) := \frac{1}{2}\int_{\mathcal{A}} |\mu(a) - \nu(a)| \, \mathrm{d}a$ is the total variation between distribution μ and ν over the action set $\mathcal{A}$.

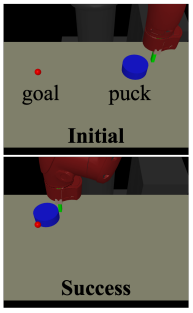


Figure 4. An initial state and a successful state in "*Push*" environment.

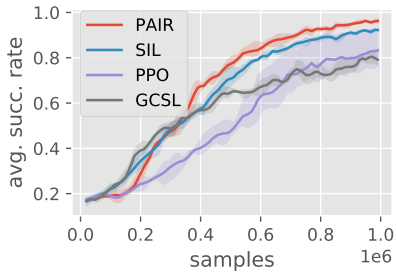


Figure 5. Average success rate vs. number of samples in "*Push*". PAIR achieves the best performance compared to non-phasic, pure RL and pure SL baselines.

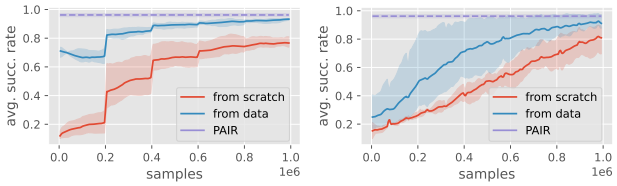


Figure 6. Combining PAIR with offline algorithms AWAC (left) and DT (right) in "*Push*" domain. We train them both from scratch (red) and from a demonstration dataset (blue). The performances of PAIR with PPO/BC are in dashed purple line.

only perform SL over self-generated data. For non-phasic RL baselines, we consider naive PPO, plain self-imitation learning with goal relabeling (SIL) (Oh et al., 2018) and self-imitation learning with task reduction (SIR) (Li et al., 2020b). For phasic SL baselines, we consider goal-conditioned supervised learning (GCSL) (Ghosh et al., 2020). We emphasize that for a fair comparison, all the RL-based baselines leverage our proposed intrinsic rewards. All the experiments are repeated over 3 random seeds on a single desktop machine with a GTX3090 GPU. More implementation and experiment details can be found in appendix.

6.1. Sawyer Push

We first answer whether our phasic framework can outperform non-phasic training algorithms on a simple "*Push*" task. As illustrated in Fig. 4, a Sawyer robot is tasked to push the puck to the goal within 50 steps. The initial position of the robot hand, the puck and the goal are randomly initialized on the table. Since it only requires very few steps to reach the goal – even the largest distance between puck and goal is within 20 steps of actions using a well trained policy, task reduction would not make its best use and may hurt sample efficiency due to extra sample consumption. In this particularly simple domain, we only use goal relabeling as a single data augmentation technique in PAIR.

Effectiveness of phasic training: We compare PAIR with non-phasic RL method, i.e., SIL, which utilizes the successful relabeled demonstrations by jointly optimizing SL and RL objectives. As shown in Fig. 5, PAIR (red) gets higher success rate than SIL (blue) using fewer samples while both methods perform better than naive PPO, which is a pure RL baseline. We also compare with phasic SL method, GCSL (Ghosh et al., 2020), which only performs iterative SL on relabeled data without RL training. Following the original implementation of GCSL, every trajectory is relabeled as a successful one targeting at an achieved state. GCSL learns fast in the beginning, but achieves a substantially lower final success rate than other methods.

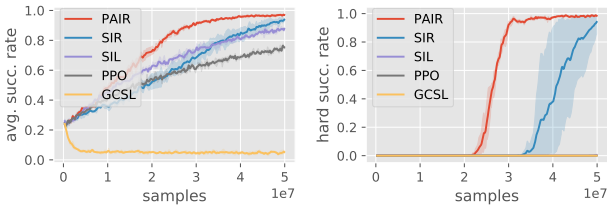


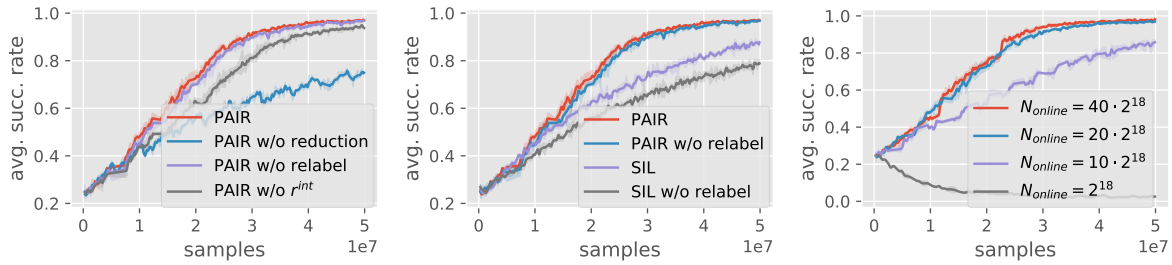
Figure 7. (*Left*) Average success rate over uniformly sampled tasks, where ant and goal positions are uniformly sampled. (*Right*) Success rate evaluated on one particularly difficult task configuration with ant and goal initialized at two ends of the maze.

Combining with offline RL methods: Here we provide the results with initial attempts to combine our framework with representative offline RL methods, AWAC (Nair et al., 2020) and decision transformer (DT) (Chen et al., 2021).

Original AWAC performs a single fine-tuning phase after offline pretraining (Nair et al., 2020; Lu et al., 2021). Following the phasic framework of PAIR, we alternate between offline and online AWAC updates. We examine both training from scratch without any dataset prepared in advance, or starting from the offline phase with a warm-start dataset of successful trajectories. The results are shown on the left of Fig. 6. Phasic-AWAC from scratch (red) continues making progress as it switches from offline to online phase, but finally converges to a much worse policy compared with the variant initialized with a warm-start dataset (blue).

DT predicts actions conditioning on a sequence of desired returns, past states and actions via a transformer model (Vaswani et al., 2017). Since DT is proposed only for a single offline phase, we similarly adopt PPO for RL training in the online phase. We use a context length of 5 for sequence conditioning and train DT both from scratch and from prepared warm-start dataset with successful demonstrations. As shown in the right plot of Fig. 6, the performance progressively improves within each online phase and SL phase. Phasic-DT with warm-start (blue) outperforms the variant from scratch and gets similar final performance as PAIR but the variance is much higher.

We generally observe that offline RL algorithms within our



(a) Ablate PAIR components. (b) The effect of goal-relabeling on SIL. (c) Sensitivity analysis of N_{online}.

Figure 8. Ablation studies on algorithmic components of PAIR in "*Ant Maze*". (a) Performances after removing different components of PAIR. Using reduction data and intrinsic reward help the most for PAIR. (b) Validate the effectiveness of goal relabeling on SIL. (c) Comparison of different N_{online}, i.e., samples collected in the RL phase. PAIR works well with less frequent phase switches.

phasic framework would require a good dataset to initialize and do not perform as robustly as the simple PPO and BC combination when starting from scratch. Therefore, we confirm our use of BC/PPO for the remaining experiments and leave a more competitive combination with offline RL algorithms as future work.

6.2. Ant Maze

We then consider a harder task "*Ant Maze*", which requires a locomotion policy with continuous actions to control the ant robot and plan over an extended horizon to navigate inside a 2-D U-shape maze. The observations include the center of mass and the joint position and velocities of the ant robot. The goal is a 2-D vector denoting the xy position of that the robot should navigate to. For each episode, the initial position of the ant's center and the goal position are uniformly sampled over the empty space inside the maze. We define the *hard task* configuration as when the ant and the goal are initialized at two different ends of the maze (see Fig. 3). In this domain, we leverage both task reduction and goal relabeling as data augmentation techniques in PAIR.

Effectiveness of PAIR and exponential improvement on hard goals: As shown in Fig. 7, we evaluate the performances of different algorithms with the average success rate over the entire task space (left) and on hard tasks only (right). We compare PAIR with: SIR, which runs SL and RL jointly using both reduction and relabeling data; SIL, a plain non-phasic RL method without task reduction; GCSL and vanilla PPO. We observe that all the methods that utilize task reduction (i.e., PAIR, SIR) can finally converge to a much higher average success rate (left plot) and outperform SIL and vanilla PPO. The gap becomes substantially larger on the performance on hard situations (right plot), where only methods with task reduction are able to produce non-zero success rate within the given sample budget. This empirical observation is consistent with our theoretical analysis in Sec. 5 on the effectiveness of task reduction. We also remark that PAIR achieves a significantly higher sample

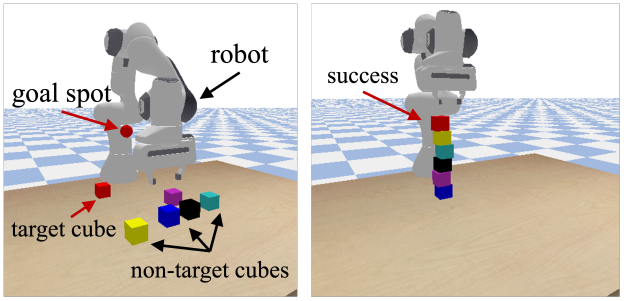


Figure 9. Illustration of the stacking environment. The left figure shows an initial state, where the red ball denotes the identity and the desired position of a target cube. On the right we show a successful state for this task, where the red cube is close to the goal and the robot hand does not touch the tower.

efficiency compared with the non-phasic method SIR on the hard cases. Notably, GCSL completely fails in this problem. We empirically observe that the ant robot may frequently get stuck in the early stage of training (e.g., it may accidentally fall over and cannot recover anymore). We hypothesis that the failure of GCSL is largely due to a tremendous amount of supervision from such corrupted trajectories. This suggests that online RL training would be necessary compared with running SL only (e.g., SIL performs well on this task).

Ablation studies: We then study the effectiveness of different components of PAIR. From Figure 8a, we can find that removing value-difference-based intrinsic reward makes PAIR converge slower while the policy is still able to achieve a high success rate after convergence. After removing task reduction, the performance becomes significantly worse. By contrast, when removing goal relabeling, the performance drop is negligible. These observations suggest that task reduction is the most critical component in PAIR. In Figure 8b, we additionally examine the effectiveness of goal relabeling within a non-phasic RL framework, i.e., SIL, where we can clearly observe a performance drop of SIL with goal relabeling turned off. This indicates that goal relabeling can be still beneficial for methods that do not utilize task

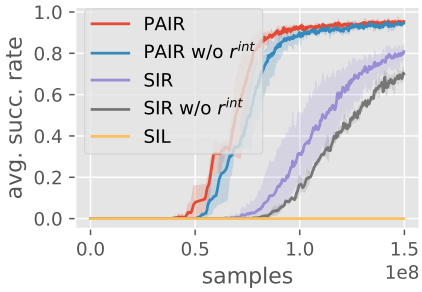


Figure 10. Average success rate in "stack-6-cube" with sparse reward. PAIR solves the task with high success rate most efficiently.

Algorithm	SR @ 7.5e7 steps	SR @ 1.5e8 steps
PAIR	**0.693 ± 0.176**	**0.955 ± 0.005**
PAIR w/o r^{int}	0.493 ± 0.156	0.952 ± 0.016
SIR	0.026 ± 0.043	0.815 ± 0.039
SIR w/o r^{int}	0.003 ± 0.003	0.688 ± 0.011
SIL	0.000 ± 0.000	0.000 ± 0.000

Table 1. Mean and standard deviation of success rates for stacking 6 boxes with different algorithms over 3 seeds. The policies are evaluated at 7.5e7 and 1.5e8 environment samples.

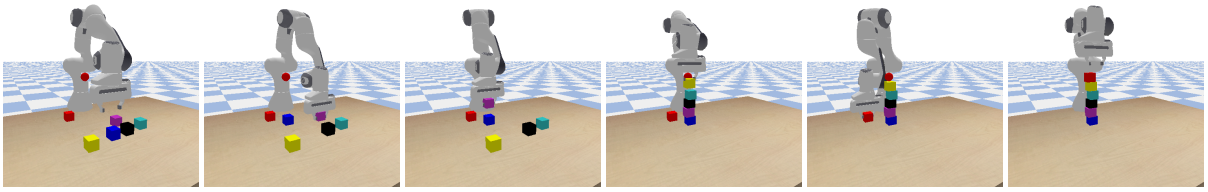


Figure 11. Learned policy of PAIR agent for stacking 6 randomly initialized cubes.

reduction. Finally, in Figure 8c, we also perform sensitivity analysis on the frequency of alternating between the offline and online phase. We perform one offline SL phase after collecting N_{online} samples in the online RL phase. We can observe that a relatively low alternating frequency is critical to the success of PAIR, which suggests that a large dataset for SL training and a long RL fine-tuning period help accelerate learning. We notice that when the phase changes too frequently, PAIR can be unstable or even fail to learn.

6.3. Stacking

Finally, we test whether PAIR can solve an extremely challenging long-horizon sparse-reward robotic manipulation problem. We build a robotic control environment that features stacking multiple cubes given only final success reward. The simulated scene is shown in Fig. 9. A Franka Panda arm is mounted on the side of a table. On top of the table, there are a total of N cubes with random initial positions. A goal spot specifies a target cube using its color and also its desired position. The robot's task is to manipulate the cubes on the table so that the specified target cube can remain stable within 3cm distance to the goal position, while its hand is at least 10cm apart from that position at the same time. In order to accomplish this task, the robot must build a "base" tower using other non-target cubes. During the whole construction process, *the robot cannot receive any external reward for intermediate manipulations unless when the target cube is in place*. We perform curriculum on the number of cubes to stack, and use the similar task reduction process described in (Li et al., 2020b). More details of training specifications can be found in Appendix B.1.3.

Performance on stacking with 6 cubes: We report the success rate on the most challenging "stack-6-cube" scenario using PAIR, SIR and SIL algorithms in Fig. 10 and Table 1. PAIR achieves an impressive 95.5% success rate in this challenging task. SIR is making considerable progress thanks to task reduction, but it learns significantly less sample efficiently than PAIR. SIL baseline without task reduction fails completely, getting 0 success rate throughout the training process. We also show the effectiveness of intrinsic reward in online training: similar to the findings in other domains, r^{int} can significantly speed up the training process.

Learned strategies: The learned policy using PAIR is visualized in Fig. 11. In the initial frame, all 6 cubes are scattered randomly on the table. The robot then picks up non-target cubes, transports them to accurate positions aligned with the goal spot one by one. Even when picking up the red cube which locates close to the half-built tower, the robot is cautious enough to avoid knocking down the tower.

7. Conclusion

We propose a phasic training method PAIR that efficiently combines offline supervised learning with online reinforcement learning for sparse-reward goal-conditioned problems, such as robotic stacking of multiple cubes. PAIR repeatedly alternates between SL on self-generated datasets and RL fine-tuning and leverages value difference as intrinsic rewards and task reduction as data augmentation. We validate the effectiveness of PAIR both theoretically and empirically on a variety of domains. We remark that PAIR provides a general learning paradigm that has the potential to be com-

bined with more advanced offline RL methods, even though our initial attempts are not satisfactory. We hope PAIR can be a promising step to take advantage of both supervised learning and reinforcement learning and help make RL a more scalable tool for complex real-world challenges.

Acknowledgements

We thank Jingzhao Zhang for valuable discussion on phasic optimization. Yi Wu is supported by 2030 Innovation Megaprojects of China (Programme on New Generation Artificial Intelligence) Grant No. 2021AAA0150000.

References

Akkaya, I., Andrychowicz, M., Chociej, M., Litwin, M., McGrew, B., Petron, A., Paino, A., Plappert, M., Powell, G., Ribas, R., et al. Solving rubik's cube with a robot hand. *arXiv preprint arXiv:1910.07113*, 2019.

Andrychowicz, M., Wolski, F., Ray, A., Schneider, J., Fong, R., Welinder, P., McGrew, B., Tobin, J., Abbeel, P., and Zaremba, W. Hindsight experience replay. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 5055–5065, 2017.

Andrychowicz, M., Raichuk, A., Stańczyk, P., Orsini, M., Girgin, S., Marinier, R., Hussenot, L., Geist, M., Pietquin, O., Michalski, M., Gelly, S., and Bachem, O. What matters for on-policy deep actor-critic methods? a large-scale study. In *International Conference on Learning Representations*, 2021.

Azuma, K. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal, Second Series*, 19(3):357–367, 1967.

Bacon, P.-L., Harb, J., and Precup, D. The option-critic architecture. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pp. 1726–1734, 2017.

Badia, A. P., Sprechmann, P., Vitvitskyi, A., Guo, D., Piot, B., Kapturowski, S., Tieleman, O., Arjovsky, M., Pritzel, A., Bolt, A., and Blundell, C. Never give up: Learning directed exploration strategies. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=Sye57xStvB.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020.

Bubeck, S. Convex optimization: Algorithms and complexity. *Found. Trends Mach. Learn.*, 8(3-4):231–357, 2015. doi: 10.1561/2200000050.

Burda, Y., Edwards, H., Storkey, A., and Klimov, O. Exploration by random network distillation. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=H1lJJnR5Ym.

Chen, L., Lu, K., Rajeswaran, A., Lee, K., Grover, A., Laskin, M., Abbeel, P., Srinivas, A., and Mordatch, I. Decision transformer: Reinforcement learning via sequence modeling. *arXiv preprint arXiv:2106.01345*, 2021.

Cobbe, K., Hilton, J., Klimov, O., and Schulman, J. Phasic policy gradient. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2020–2027. PMLR, 2021.

Coumans, E. and Bai, Y. Pybullet, a python module for physics simulation for games, robotics and machine learning. 2016.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

Ecoffet, A., Huizinga, J., Lehman, J., Stanley, K. O., and Clune, J. First return, then explore. *Nature*, 590(7847): 580–586, 2021.

Emmons, S., Eysenbach, B., Kostrikov, I., and Levine, S. Rvs: What is essential for offline rl via supervised learning? *arXiv preprint arXiv:2112.10751*, 2021.

Engstrom, L., Ilyas, A., Santurkar, S., Tsipras, D., Janoos, F., Rudolph, L., and Madry, A. Implementation matters in deep rl: A case study on ppo and trpo. In *International Conference on Learning Representations*, 2020.

Eysenbach, B., Salakhutdinov, R., and Levine, S. C-learning: Learning to achieve goals via recursive classification. In *International Conference on Learning Representations*, 2020.

Fujimoto, S. and Gu, S. S. A minimalist approach to offline reinforcement learning. *arXiv preprint arXiv:2106.06860*, 2021.

Furuta, H., Matsuo, Y., and Gu, S. S. Generalized decision transformer for offline hindsight information matching. In *Deep RL Workshop NeurIPS 2021*, 2021.

Ghosh, D., Gupta, A., Reddy, A., Fu, J., Devin, C. M., Eysenbach, B., and Levine, S. Learning to reach goals via iterated supervised learning. In *International Conference on Learning Representations*, 2020.

Hester, T., Vecerik, M., Pietquin, O., Lanctot, M., Schaul, T., Piot, B., Horgan, D., Quan, J., Sendonaris, A., Osband, I., et al. Deep q-learning from demonstrations. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

Hwangbo, J., Lee, J., Dosovitskiy, A., Bellicoso, D., Tsounis, V., Koltun, V., and Hutter, M. Learning agile and dynamic motor skills for legged robots. *Science Robotics*, 4(26), 2019.

Ilyas, A., Engstrom, L., Santurkar, S., Tsipras, D., Janoos, F., Rudolph, L., and Madry, A. A closer look at deep policy gradients. In *International Conference on Learning Representations*, 2020.

Janner, M., Li, Q., and Levine, S. Offline reinforcement learning as one big sequence modeling problem. *Advances in Neural Information Processing Systems*, 34, 2021.

Jeon, W. and Kim, D. Autonomous molecule generation using reinforcement learning and docking to develop potential novel inhibitors. *Scientific reports*, 10(1):1–11, 2020.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Kaelbling, L. Learning to achieve goals. In *Proc. of IJCAI-93*, pp. 1094–1098, 1993.

Kidambi, R., Rajeswaran, A., Netrapalli, P., and Joachims, T. Morel: Model-based offline reinforcement learning. In *NeurIPS*, 2020.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *ICLR (Poster)*, 2015.

Kostrikov, I., Nair, A., and Levine, S. Offline reinforcement learning with implicit q-learning. *arXiv preprint arXiv:2110.06169*, 2021.

Kulkarni, T. D., Narasimhan, K., Saeedi, A., and Tenenbaum, J. Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation. *Advances in neural information processing systems*, 29: 3675–3683, 2016.

Kumar, A., Peng, X. B., and Levine, S. Reward-conditioned policies. *arXiv preprint arXiv:1912.13465*, 2019.

Kumar, A., Zhou, A., Tucker, G., and Levine, S. Conservative q-learning for offline reinforcement learning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Lange, S., Gabel, T., and Riedmiller, M. Batch reinforcement learning. In *Reinforcement learning*, pp. 45–73. Springer, 2012.

Levine, S. Understanding the world through action. In Faust, A., Hsu, D., and Neumann, G. (eds.), *Conference on Robot Learning, 8-11 November 2021, London, UK*, volume 164 of *Proceedings of Machine Learning Research*, pp. 1752–1757. PMLR, 2021.

Levine, S., Kumar, A., Tucker, G., and Fu, J. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Li, R., Jabri, A., Darrell, T., and Agrawal, P. Towards practical multi-object manipulation using relational reinforcement learning. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4051–4058. IEEE, 2020a.

Li, Y., Wu, Y., Xu, H., Wang, X., and Wu, Y. Solving compositional reinforcement learning problems via task reduction. In *International Conference on Learning Representations*, 2020b.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. Continuous control with deep reinforcement learning. In *ICLR (Poster)*, 2016.

Lu, Y., Hausman, K., Chebotar, Y., Yan, M., Jang, E., Herzog, A., Xiao, T., Irpan, A., Khansari, M., Kalashnikov, D., and Levine, S. AW-opt: Learning robotic skills with imitation andreinforcement at scale. In *5th Annual Conference on Robot Learning*, 2021. URL https://openreview.net/forum?id=xwEaXgFa0MR.

Lynch, C., Khansari, M., Xiao, T., Kumar, V., Tompson, J., Levine, S., and Sermanet, P. Learning latent plans from play. In *Conference on Robot Learning*, pp. 1113–1132. PMLR, 2020.

Mao, Y., Wang, C., Wang, B., and Zhang, C. Moore: Model-based offline-to-online reinforcement learning, 2022.

Matsushima, T., Furuta, H., Matsuo, Y., Nachum, O., and Gu, S. Deployment-efficient reinforcement learning via model-based offline optimization. In *International Conference on Learning Representations*, 2020.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. Human-level control through deep reinforcement learning. *nature*, 518(7540): 529–533, 2015.

Nachum, O., Gu, S. S., Lee, H., and Levine, S. Data-efficient hierarchical reinforcement learning. *Advances in Neural Information Processing Systems*, 31:3303–3313, 2018.

Nair, A., McGrew, B., Andrychowicz, M., Zaremba, W., and Abbeel, P. Overcoming exploration in reinforcement learning with demonstrations. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6292–6299. IEEE, 2018a.

Nair, A., Dalal, M., Gupta, A., and Levine, S. Awac: Accelerating online reinforcement learning with offline datasets. 2020.

Nair, A. V., Pong, V., Dalal, M., Bahl, S., Lin, S., and Levine, S. Visual reinforcement learning with imagined goals. *Advances in Neural Information Processing Systems*, 31: 9191–9200, 2018b.

Nasiriany, S., Pong, V., Lin, S., and Levine, S. Planning with goal-conditioned policies. *Advances in Neural Information Processing Systems*, 32:14843–14854, 2019.

Oh, J., Guo, Y., Singh, S., and Lee, H. Self-imitation learning. In *International Conference on Machine Learning*, pp. 3878–3887. PMLR, 2018.

Peng, X. B., Kumar, A., Zhang, G., and Levine, S. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. *arXiv preprint arXiv:1910.00177*, 2019.

Pong, V., Gu, S., Dalal, M., and Levine, S. Temporal difference models: Model-free deep rl for model-based control. In *International Conference on Learning Representations*, 2018.

Rajeswaran, A., Kumar, V., Gupta, A., Vezzani, G., Schulman, J., Todorov, E., and Levine, S. Learning complex dexterous manipulation with deep reinforcement learning and demonstrations. In Kress-Gazit, H., Srinivasa, S. S., Howard, T., and Atanasov, N. (eds.), *Robotics: Science and Systems XIV, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA, June 26-30, 2018*, 2018. doi: 10.15607/RSS.2018.XIV.049. URL http://www.roboticsproceedings.org/rss14/p49.html.

Rashidinejad, P., Zhu, B., Ma, C., Jiao, J., and Russell, S. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *arXiv preprint arXiv:2103.12021*, 2021.

Schaul, T., Horgan, D., Gregor, K., and Silver, D. Universal value function approximators. In *International conference on machine learning*, pp. 1312–1320. PMLR, 2015.

Schmidhuber, J. Reinforcement learning upside down: Don't predict rewards–just map them to actions. *arXiv preprint arXiv:1912.02875*, 2019.

Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., Guez, A., Lockhart, E., Hassabis, D., Graepel, T., et al. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839): 604–609, 2020.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017. URL http://arxiv.org/abs/1707.06347.

Singh, S., Barto, A. G., and Chentanez, N. Intrinsically motivated reinforcement learning. Technical report, MASSACHUSETTS UNIV AMHERST DEPT OF COMPUTER SCIENCE, 2005.

Stolle, M. and Precup, D. Learning options in reinforcement learning. In *International Symposium on abstraction, reformulation, and approximation*, pp. 212–223. Springer, 2002.

Todorov, E., Erez, T., and Tassa, Y. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033. IEEE, 2012.

Tucker, G., Bhupatiraju, S., Gu, S., Turner, R., Ghahramani, Z., and Levine, S. The mirage of action-dependent baselines in reinforcement learning. In *International conference on machine learning*, pp. 5015–5024. PMLR, 2018.

Uchendu, I., Xiao, T., Lu, Y., Zhu, B., Yan, M., Simon, J., Bennice, M., Fu, C., Ma, C., Jiao, J., et al. Jump-start reinforcement learning. *arXiv preprint arXiv:2204.02372*, 2022.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Veeriah, V., Oh, J., and Singh, S. Many-goals reinforcement learning. *arXiv preprint arXiv:1806.09605*, 2018.

Wu, Y., Tucker, G., and Nachum, O. Behavior regularized offline reinforcement learning. *arXiv preprint arXiv:1911.11361*, 2019.

Yu, C., Velu, A., Vinitsky, E., Wang, Y., Bayen, A., and Wu, Y. The surprising effectiveness of mappo in cooperative, multi-agent games. *arXiv preprint arXiv:2103.01955*, 2021a.

Yu, T., Kumar, A., Chebotar, Y., Hausman, K., Levine, S., and Finn, C. Conservative data sharing for multi-task offline reinforcement learning. *Advances in Neural Information Processing Systems*, 34, 2021b.

Zhao, R., Sun, X., and Tresp, V. Maximum entropy-regularized multi-goal reinforcement learning. In *International Conference on Machine Learning*, pp. 7553–7562. PMLR, 2019.

The project webpage is at https://sites.google.com/view/pair-gcrl.

A. Missing Proofs in Section 5

In this section, we prove Theorem 5.3.

Here, we make the following assumptions to facilitate our analysis. First, we consider a goal-conditioned MDP with deterministic transition. We assume that the goal set is the same as the state space, i.e., $\mathcal{G} = \mathcal{S}$. Furthermore, we consider sparse-reward environment where the reward function is defined by $r(s, a, g) = \mathbb{I}\{s = g\}$.

We note that our framework draw on-policy trajectories with random state-goal pairs. To further simplify our theoretical analysis, we assume that each state-goal pair is sampled at least once in each iteration. Here, we may meet two practical issues. First, the state space may be continuous, which makes it impossible to sample each state-goal pair once. This may fit into our theoretical analysis by discretizing the state space. Second, N_{online} may not be large enough. This fits into our analysis by merging several consecutive iterations.

We use $\pi^{(k)}$ to denote the policy by the end of the k-th iteration, and $\pi^{(0)}$ denotes the initial policy. We assume that the initial policy $\pi^{(0)}$ can reach any one-step goal. Specifically, if $P(s'|s, a) = 1$ then $\pi^{(0)}(a|s, g = s') = 1$.

For a state s, a goal g, and a (deterministic) policy π, we define

$$s \xrightarrow{\pi} g := \begin{cases} \text{true}, & \Pr[\exists t \geq 0 : s_t = g \mid s_0 = s, \forall i \geq 0, a_i \sim \pi(\cdot|s_i, g), s_{i+1} \sim P(\cdot|s_i, a_i)] = 1, \\ \text{false}, & \text{otherwise.} \end{cases} \tag{7}$$

and we define

$$d(s, s') := \min\{t : \exists \text{policy} \pi \text{ such that } \Pr[s_t = s' \mid s_0 = s, \forall i \geq 0, a_i \sim \pi(\cdot|s_i, g), s_{i+1} \sim P(\cdot|s_i, a_i)] = 1\}, \tag{8}$$

$$\ell_k := \sup_{\ell} \{\forall s, g : d(s, g) \leq \ell : s \xrightarrow{\pi^{(k)}} g \text{ is true}\}. \tag{9}$$

Finally, we let

$$D = \max_{s, s' \in \mathcal{S}} \{d(s, s') : d(s, s') < +\infty\} \tag{10}$$

be the maximum possible length of trajectory from state s to goal s' such that s' is reachable from s.

Lemma A.1. *For the PAIR framework, we have $\ell_k \geq 2\ell_{k-1}$ for every iteration k.*

Proof. We prove by induction. For every state-goal pair s, g such that $d(s, g) \leq 2\ell_{k-1}$, we can find some state s' such that $d(s, s') \leq \ell_{k-1}$ and $d(s', g) \leq \ell_{k-1}$. By the inductive hypothesis, we have that $s \xrightarrow{\pi^{(k)}} s'$ and $s' \xrightarrow{\pi^{(k)}} g$ are true.

Now we claim that every $(s, g) \in \mathcal{S} \times \mathcal{G}$ would have a success trajectory after the k-th iteration. Note that by our assumption, the framework would roll out a trajectory from s to g using $\pi(\cdot|s, g)$. If it succeeds, then we get a trajectory, else task reduction (cf. Section 4.2) would find the state s' such that $s \xrightarrow{\pi^{(k-1)}} s'$ and $s' \xrightarrow{\pi^{(k-1)}} g$ are true, because $V(s, s') = V(s', g) = 1$. Therefore, task reduction would run $\pi^{(k-1)}(\cdot|s, g = s')$ followed by $\pi^{(k-1)}(\cdot|s', g)$, and get a successful trajectory from s to g. Finally, the SL step would learn the policy $\pi^{(k)}$ such that $s \xrightarrow{\pi^{(k)}} g$ is true. □

Lemma A.2. *The PAIR framework uses at most $O(|\mathcal{S}|^2 \log D)$ samples to learn a policy that could go from any state to any reachable goal.*

Proof. By the definition of ℓ_k, we know that $\pi^{(k)}$ could go from any state to any reachable goal if $\ell_k \geq D$. By our assumption, we have $\ell_0 = 1$. Therefore, by Lemma A.1, we have $\ell_k \geq D$ after $k = O(\log D)$, i.e., PAIR uses at most $O(\log D)$ iterations.

By our assumption, PAIR uses $O(|\mathcal{S}||\mathcal{G}|) = O(|\mathcal{S}|^2)$ samples, because each state-goal pair is sampled for constant number of times. Thus we conclude that PAIR uses at most $O(|\mathcal{S}|^2 \log D)$ samples in total. □

Lemma A.3. *Without PAIR, under assumptions described in Appendix A, for hard tasks, with high probability, the algorithm needs $\Omega(|\mathcal{S}|^2 D)$ samples to learn a policy that could go from any state to any reachable goal.*

Proof. We consider the following hard task. Suppose $s_0, a_0, \cdots, s_D$ is a trajectory of length D that maximizes Eq. (10). We assume that the initial policy $\pi^{(0)}$ is a uniformly random policy and we assume $\Pr[\pi^{(0)}(\cdot|s_i, g = s_D)]] \propto |\mathcal{A}|^{-1}$ for $i \le D-2$.[3] We observe that for the SL framework, if $s_i \xrightarrow{\pi^{(k)}} s_D$ is true then the dataset must contain a trajectory from s_i to s_D, and thus the policy $\pi^{(k)}$ learned by SL would satisfy $s_j \xrightarrow{\pi^{(k)}} s_D$ for every $j \ge i$. For this, we analyze the sample complexity by analyzing how many iterations would be enough for the algorithm to learn a policy from state s_0 to goal s_D. To facilitate analysis, we define $i^{(k)}$ to be the smallest number such that $s_{i^{(k)}} \xrightarrow{\pi^{(k)}} s_D$. Then the algorithm learns the desired policy after k iterations only if $i^(k) = 0$.

Next, we prove that $\mathbb{E}[i^{(k-1)} - i^{(k)} | i^{(k-1)}] \le O(1)$. To change $i^{(k-1)}$, the algorithm must roll out a success trajectory with goal s_D from some s_j with $j \le i$. Note that

$$\begin{aligned}
\Pr[i^{(k-1)} - i^{(k)} = j | i^{(k-1)}] &\le \Pr[\text{algorithm rolls out a success trajectory from } s_{i^{(k-1)}-j} \text{ to } s_D] && (11)\\
&\le \Pr[\pi^{(k-1)} \text{ could go from } s_{i^{(k-1)}-j} \text{ to } s_{i^{(k-1)}}] && (12)\\
&\le O(|\mathcal{A}|^{-j}), && (13)
\end{aligned}$$

so

$$\begin{aligned}
\mathbb{E}[i^{(k-1)} - i^{(k)} | i^{(k-1)}] &= \sum_{j=0}^{i^{(k-1)}} j \cdot \Pr[i^{(k-1)} - i^{(k)} = j | i^{(k-1)}] && (14)\\
&\le O(\sum_{j=0}^{\infty} j \cdot |\mathcal{A}|^{-j}) && (15)\\
&\le O(|\mathcal{A}|^{-1}) && (16)\\
&\le O(1). && (17)
\end{aligned}$$

Therefore, by the Azuma-Hoeffding inequality (Azuma, 1967), we conclude that with high probability $1-\delta$, we have $i^{(k)} < D$ for $k = \Theta(D - \sqrt{D \log \delta^{-1}}) \ge \Omega(D)$, i.e., $k \ge \Omega(D)$ with high probability. Together with that $|\mathcal{S}|^2$ trajectories are rolled out in each iteration, we conclude that the sample complexity is with high probability at least $\Omega(|\mathcal{S}|^2 D)$. □

Finally, we conclude Theorem 5.3 by collecting Lemmas A.2 and A.3.

B. Experiment Details

B.1. Environment Description

In this section, we describe all the environment-specific details regarding MDP definitions.

B.1.1. SAWYER PUSH

"*Push*" is a robotic pushing environment adopted from (Nair et al., 2018b) simulated with MuJoCo (Todorov et al., 2012) engine. Each episode lasts for at most 50 steps.

Observation: The agent can observe 2-D hand position and 2-D puck position in each step.

Action space: An action is a 2-D vector denoting the relative movement of the robot hand. The height of the hand and the state of the gripper are fixed. Each dimension ranges from -1 to 1, and is categorized into 3 bins (therefore, there are a total of 9 different actions).

[3] Actually, we can construct such an MDP. Assume we have $(D+2)$ states $\mathcal{S} = \{s_0, \cdots, s_D, s_{D+1}\}$ and 2 actions $\mathcal{A} = \{0, 1\}$. The transition is given by $P(s_{D+1}|s_i, a=0) = 1$ and $P(s_{i+1}|s_i, a=1) = 1$ for $0 \le i \le D$. Such an MDP can similarly be constructed for $|\mathcal{A}| \ge 3$.

Goal and reward: The puck goal is a 2-D vector denoting the target position on the table. The agent can only get reward when the $l2$ distance between puck and goal is smaller than 5cm.

Initial state: Each episode starts with the hand initialized in a 0.03m $\times$ 0.03m region on the right side of the table, the puck initialized in a 0.07m $\times$ 0.07m square, and the goal is uniformly sampled in a 0.4m $\times$ 0.4m square on the table.

B.1.2. ANT MAZE

This environment is adopted from (Nachum et al., 2018). The scene is a 24m $\times$ 24m maze simulated with MuJoCo (Todorov et al., 2012). The maximum number of steps for each episode is 500 steps.

Observation: Joint positions and joint velocities of the ant robot. The center of mass of the robot is included in the joint positions.

Action: 8-D real-valued vector controlling the motors on the robot. The actions output from the policy network are clipped to -30 to 30 before sent to the simulator.

Goal and reward: The goals are 2-D vectors denoting the desired xy position of the ant. The agent gets reward when the distance between the ant's center of mass and the goal is smaller than 2m.

Initial state: Both the ant and the goal are uniformly sampled in the empty space (not collided with walls) in the maze. For hard tasks, the ant is initialized at coordinate (0m, 0m), and the goal is at coordinate (0m, 16m).

B.1.3. STACKING

The stacking environment is built with PyBullet (Coumans & Bai, 2016).

Observation: The observation is a concatenation of robot state and objects states. The robot state contains 6-D end effector pose, 3-D end effector linear velocity, 1-D finger position, and 1-D finger velocity. Each object state consists of its 6-D pose, 3-D relative position to the robot's end effector, 3-D linear velocity, 3-D angular velocity and 1-D 0/1 indicator denoting the identity of the target cube.

Action: A concatenation of 3-D relative movement of the end effector and 1-D finger control signal. Each dimension is discretized into 20 bins.

Goal and reward: A goal specifies a 3-D desired position and an identity denoting which cube must be close to the goal spot. A non-zero reward is only given when the target cube is within 3cm distance to the goal, and when the end effector is at least 10cm away from the goal.

Adaptive curriculum: Since directly training on stacking tasks with large number of objects pose severe exploration challenges, we adopt a simple curriculum on the number of cubes to stack, similar to the one in (Li et al., 2020a). In the beginning, 70% of the tasks only need to stack one cube, and the other tasks are uniformly sampled to stack 2 to N cubes. Whenever the average success rate reaches 0.6, the curriculum proceeds from sampling "stack-n-cubes" with 70% probability to "stack-$(n+1)$-cubes" tasks. We switch to offline SL training when the curriculum proceeds and the success rate for stacking $n+1$ cubes is lower than 0.5.

B.2. Combination with Offline RL

B.2.1. OFFLINE RL COMBINATION IN SAWYER PUSH

For the warm-start dataset, we collect rollout trajectories generated by random policy, where we keep successful samples and do goal relabelling on failed trajectories. Therefore, the dataset only consists of successful demonstrations. The dataset size for phasic-DT is $55K$ steps, and $15k$ for phasic-AWAC. Phasic-AWAC switches from online to offline training after collecting $200k$ online samples to match the total number of phasic switches in original PAIR. The online training of Phasic-DT is based on PPO, therefore we adopt the same number of PPO updates as PAIR for phasic-DT before switching to its offline phase.

B.2.2. ADDITIONAL RESULTS IN STACK SCENARIO

We also try to run phasic-DT in the stacking domain. We train phasic-DT starting from the offline phase with a warm-up dataset. The warm-up dataset consists of successful trajectories in "stack-1-cube" scenario generated by a pretrained PPO

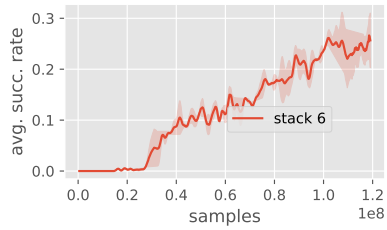


Figure 12. Average success rate of phasic-DT in "stack-6-cube" domain.

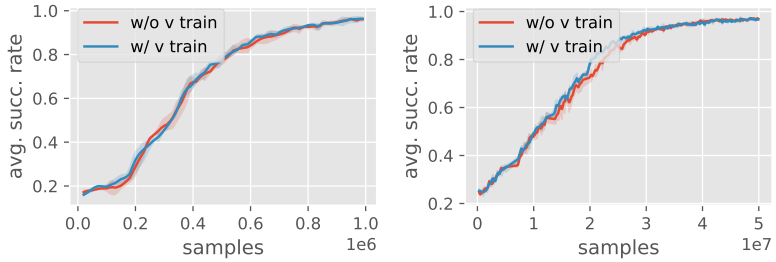


Figure 13. Comparison of training value network or not in PAIR offline phase. Left plot is in "Push" domain, right plot is in "Ant Maze".

model. We similarly adopt adaptive curriculum to alleviate exploration difficulty as when training PAIR. The results are shown in Figure 12. The success rate of phasic-DT for stacking 6 cubes grows slowly, and does not converge to a value as high as the original PAIR.

B.3. Algorithm Implementations

Network architecture: We use separate policy and value networks with the same architecture for PPO-based algorithms. The specific network architectures for different domains are as follows. For "Push", we use MLPs of hidden size 400 and 300. For "Ant Maze", the networks are MLPs of hidden size 256 and 256. For "Stacking", we use a transformer (Vaswani et al., 2017)-based architecture, which stacks 2 self-attention blocks with one head and 64 hidden units, and then goes though linear heads to output action distributions or values. Since there are exponential number of possible actions as we discretize each action dimension into several bins, we assume different action dimensions are independent, and use separate linear heads to predict distributions of different dimensions.

PAIR implementation: We use PPO (Schulman et al., 2017) for online training and advantage weighted imitation learning for offline training by default. For PPO training, we use N_{worker} parallel workers to collect transitions (s, a, r, s') from the environments synchronously. When applying the value-difference-based intrinsic reward, r is replaced by the sum of environment reward and the intrinsic reward calculated with the current value network. After each worker gets N_{steps} data points, we run N_{epoch} epochs of PPO training by splitting all the collected on-policy data into N_{batch} batches. The successful trajectories from these on-policy batch are directly stored into the offline data $\mathcal{D}$, and the failed trajectories are cached into a failure buffer $\mathcal{B}_{\text{fail}}$. We repeat the data collection - PPO training phase until the total amount of on-policy data reaches N_{online}. Then we perform data augmentation (task reduction, goal relabeling) over $\mathcal{B}_{\text{fail}}$ so as to generate more successful data, and insert the resulting demonstrations into $\mathcal{D}$ as well. After the dataset $\mathcal{D}$ is constructed, we run M_{epoch} epochs of advantage weighted behavior cloning with batch size m. We use GAE advantage calculated with the value network learned after the online phase as $(R - V)$ for each data point. We use Adam optimizer (Kingma & Ba, 2015) for PPO and supervised learning. All the hyper-parameters are listed in Table 2.

We do not additionally train the value network using the offline dataset in offline phase since we empirically find that training V does not help the overall performance (see Figure 13).

GCSL baseline implementation: There is a slight difference between our implementation and the original version from (Ghosh et al., 2020): we collect and relabel data in a more phasic fashion, i.e., we perform imitation learning on the data batch collected from only the previous online collection phase. We keep the number of online steps before offline imitation learning the same as PAIR. The original GCSL maintains a data buffer throughout the training process similar to the setting

Domain	Push	Ant Maze	Stacking
N_{worker}	4	64	64
N_{online}	$10 \cdot 2^{14}$	$20 \cdot 2^{18}$	adaptive
α		0.5	
N_{steps}		4096	
N_{epoch}		10	
N_{batch}		32	
β		1	
M_{epoch}		10	
m		64	
lr		2.5e-4	

Table 2. Hyperparameters of PAIR.

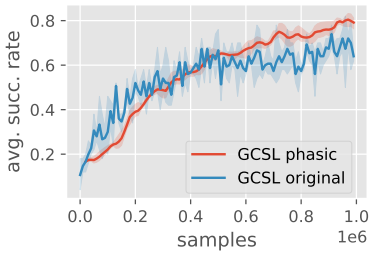


Figure 14. Comparison between phasic GCSL and the original version.

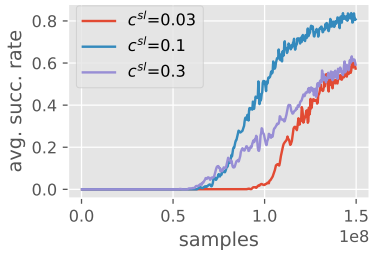


Figure 15. Success rate of joint RL+SL with different c^{SL} in "stack-6-cube" task. We adopt $c^{SL} = 0.1$ for SIR in the main paper.

in off-policy RL. We find our phasic GCSL can get better performance than the original version (see Figure 14), so we present the results of phasic GCSL in the main paper.

DT baseline implementation: The policy and value networks adopt the same GPT-2 architecture similar to the original version in (Chen et al., 2021). They output actions and values conditioning on desired return, past states, and past actions. To avoid gradient explosion, only the last token of the predicted action or value sequence is used for updating the model. To stabilize training, we remove all dropout layers from the transformer model. We use a context length of 5 for sequence conditioning. The specific feature extractors for processing single-step observations in different domains are as follows. For "Push", we use MLPs of hidden size 300 and 300 for representation learning and we train separate policy and value networks. For "Stacking", we use a transformer-based architecture similar to PAIR for representation learning except that the size of hidden layer is 128. We find that using shared or separate feature extractors for policy and value networks leads to similar performances. Therefore we share parameters for them to save computational resources.

B.4. Analysis on Phasic vs. Joint RL and SL Optimization

PAIR decouples RL and SL objectives in two phases instead of optimizing them jointly ($L^{RL} + c^{SL}L^{SL}$), since the two objectives can largely interfere with each other and the choice of c^{SL} is empirically brittle to tune. The RL objective is policy gradient over rollout data (Eqn. 1), which requires (primarily) on-policy samples (both success and failures) to make policy improvement. The SL objective (Eqn. 2) is performed over the successful dataset with both success-only rollout samples and off-policy augmentation trajectories. These two objectives operate on very distinct data distributions. If c^{SL} is too large, the gradient will be pulled away from the policy improvement direction, which makes policy learning unstable or even breaks training. If c^{SL} is too small, the objective may not sufficiently leverage the augmented successful data learning to slow convergence. We report sensitivity analysis in Fig. 15 by trying different c^{SL} in the joint objective. With phasic training, RL and SL are decoupled and the interference is largely reduced. We can also intuitively interpret the benefit of decoupling RL and SL into two phases by echoing one result in stochastic optimization: smaller optimization step size should be taken when the gradient noise is large (Theorem 6.2 in (Bubeck, 2015)). RL objective is with large gradient noise, while SL offers clear supervision signal. When jointly optimizing RL and SL, only small step size is allowed which leads to slow convergence; when optimizing RL and SL separately, different step sizes can be chosen for different objectives, thus converges faster.