# Coherent Off-Policy Improvement of Large Behaviour Models with Learned Rewards

Christian Scherer[1,2], Joe Watson[3], Daniel Palenicek[1,4], Theo Gruner[1], Ingmar Posner[3], Jan Peters[1,4,5,6]
[1]Technical University of Darmstadt, [2]Zuse School ELIZA, [3]University of Oxford,
[4]hessian.AI, [5]German Research Center for AI (DFKI), [6]Robotics Institute Germany (RIG)

***Abstract*—Distilling expert demonstration data into large generative models using behavioural cloning is a scalable approach to learning capable policies for robotic control, particularly for dexterous manipulation. Reinforcement learning (RL) can be used as a means to fine-tune these policies further using additional experience. An open question is whether RL is more sample-efficient than collecting more human demonstrations. Prior work has fine-tuned large pre-trained policies in a scalable fashion by applying RL to a smaller residual policy that corrects the pre-trained model. However, for the typical sparse reward tasks, RL algorithms can struggle to optimize the behaviour in a sample-efficient manner. We look at inverse reinforcement learning, where a dense reward function is learned from the expert demonstrations, potentially reducing the challenge of RL fine-tuning. We specifically consider coherent imitation learning, an IRL method that facilitates improvement of the BC policy through using a specific reward formulation with theoretical guarantees. We show that our IRL method improves the performance of $\pi_0.5$ on all 6 sparse manipulation tasks and achieves a $\geq$ 90% success rate on 5 out of 6 complex manipulation tasks, outperforming RL-based baselines using sparse rewards. By ensuring our initial pretrained finetuning policy is optimal for our initial reward and critic, our method circumvents the initial drop commonly seen in RL fine-tuning and enables faster improvement.**

## I. Introduction

Distilling massive expert demonstration datasets into large generative models (behavioural cloning, BC) has seen large success in robotic control, yielding highly capable policies for complex, long-horizon manipulation [1, 2]. Despite these advances, covariate shift due to accumulating errors remains a fundamental limitation of any BC policy. Solutions such as collecting supplementary recovery demonstrations are fundamentally limited because they rely on teleoperators to anticipate out-of-distribution states. Since online agents will inevitably face novel states, online adaptation becomes indispensable.

Residual reinforcement learning (RL) is an attractive paradigm for improving BC policies via online experience: freezing the base policy and learning a correction enables parameter-efficient fine-tuning [3, 4, 5]. However, two major challenges remain. First, designing dense, task-specific rewards for open-world tasks is difficult and prone to misspecification [6]. Second, even parameter-efficient residual RL frameworks [7] remain vulnerable to initial performance degradation (see Figure 1). Crucially, this unlearning directly degrades sample efficiency.

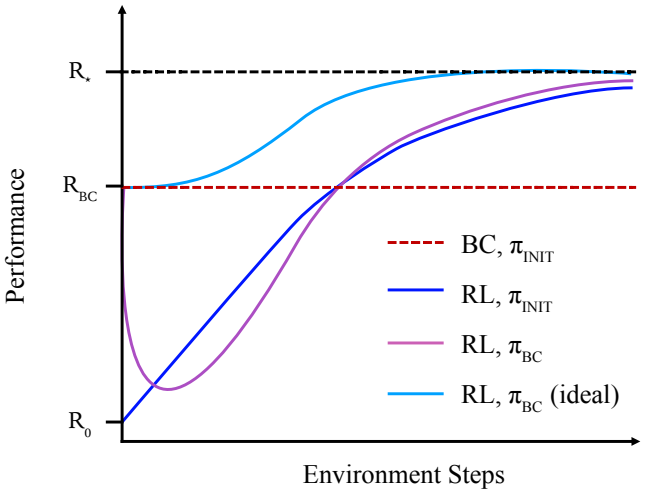


Fig. 1. An illustration of RL fine-tuning of strong BC policies using residuals. Despite strong initial performance by a BC policy, due to function approximation in actor-critic methods, this performance is rapidly unlearned and relearned, which means in practice the BC initialization provides little benefit when running the RL method from scratch. We investigate a means for fine-tuning the BC policy without unlearning using coherent soft imitation learning (CSIL) for (inverse) RL fine-tuning.

To achieve online adaptation without degrading the foundation model (see Figure 1), we build on coherent soft imitation learning (CSIL) [8], an entropy-regularized IRL algorithm. In contrast to other (deep) IRL algorithms, CSIL is designed to fine-tune a pre-trained behaviour policy using IRL due to its specially designed reward structure, as IRL methods with black-box reward models tend to unlearn the BC behaviour rather than improve it [8].

Our main contribution is a novel off-policy inverse RL method for fine-tuning large behaviour models (LBMs). Instead of relying on typically sparse environment rewards, we learn a dense reward function from demonstrations using CSIL. Secondly, we enhance the base CSIL implementation [8] with recent sample-efficient off-policy RL advancements. This integration boosts image-based success on `Square` from 50% to 90% and significantly accelerates convergence in state-based settings.

## II. Related Work

Achieving reliable performance with LBMs remains a critical bottleneck for their widespread deployment, as current pre-trained models typically fail to achieve sufficiently high

success rates without further fine-tuning. While simple BC provides a baseline, a large body of recent literature focuses on improving LBMs through RL. Several works directly apply online RL algorithms, particularly proximal policy optimization (PPO), to fine-tune these models [9]. However, the inherent sample inefficiency of these methods makes them difficult to scale and extend to real-world applications.

To overcome the cost of vast online exploration, offline RL has emerged as a powerful alternative for fine-tuning LBMs [10]. These approaches typically train an offline value function and subsequently condition the LBM's actions upon it. While demonstrating impressive results, such methods are highly compute-intensive and generally out of scope for standard fine-tuning tasks. Seeking to mitigate both sample inefficiency and high computational demands, prior work has investigated off-policy residual RL [3, 11]. In this paradigm, RL is used to train a secondary policy that predicts corrections to the base actions of a pre-trained diffusion policy. Residual RL has recently been successfully adopted to LBMs [7]. Notably, the authors report a severe "initial performance drop" during early fine-tuning [7], which is also observed in [11], driven by the agent's exploration into out-of-distribution states. The observed degradation is a well-documented phenomenon in the offline-to-online RL literature [12, 8, 13]. To prevent unlearning, these methods typically restrict action evaluation to the offline dataset or explicitly align their online objectives with the BC prior to avoid destructive, high-variance gradients.

Because the naive residual approach often suffers from poor exploration, recent works have introduced targeted improvements to this framework. For instance, constraining the residual predictions to a smaller, bounded action range has been shown to guide exploration more effectively [4]. Additionally, to improve and stabilize the value function's learning process, works have proposed filling the initial replay buffer exclusively with the base policy's actions [5].

Our approach contrasts itself in a variety of ways. Instead of training a residual, we ensemble the base policy's and the second policy's actions, as constraining the model to predict small residual corrections led to a collapse of uncertainty estimation in preliminary experiments. Initializing our policy using BC enables us to fill the initial replay buffer with on-policy actions, instead of relying on off-policy actions for this step [5]. Most importantly, we learn a dense reward function in a principled manner in contrast to works that only use a sparse reward [3, 11, 4, 5] or approximate a reward signal using hand-crafted features [9]. Further, we address sample-efficiency by incorporating advances from state-of-the-art off-policy RL such as categorical critics with batch normalization and weight normalization [14].

## III. Coherent Inverse Reinforcement Learning

In this section, we will introduce the foundations of our method that are applicable independent of refining LBMs and, in particular, we will explain CSIL [8] and implementation details that improve its sample efficiency in the deep IRL setting.

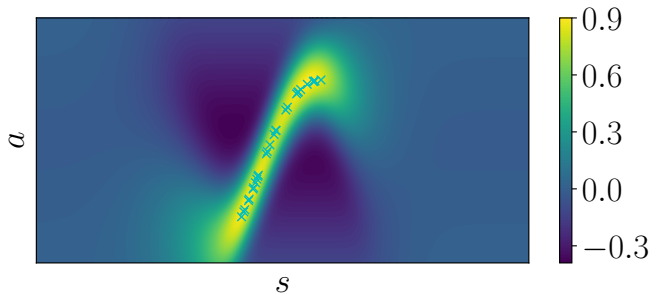


Fig. 2. Given demonstration data (X), the coherent reward provides positive rewards for correct actions for observed states, negative rewards for incorrect actions for seen states, and zero rewards for any action under unseen states. This encourages the agent to act like the expert, and also rewards the agent for returning to the demonstration distribution if the agent deviates far from the expert trajectories due to compounding errors. Taken from Watson et al. [8].

### A. Coherent Soft Imitation Learning

Imitation learning is an umbrella term that covers behavioural cloning and inverse reinforcement learning. Inverse reinforcement learning jointly performs RL and reward inference from demonstrations to learn expert behaviour. For sample efficiency, one would ideally combine these methods to accelerate learning, but in the deep RL setting, Watson et al. observed that most IRL algorithms cannot fine-tune BC policies and instead rapidly unlearn them [8]. This pathology arises from function approximation. Since the BC policy is not optimal for the randomly initialized reward function and critic, a random behaviour is initially learned instead. To alleviate this issue, Watson et al. propose *coherency* in IRL, which requires that the BC policy be optimal for the learned reward function. In the entropy-regularized RL setting, this can be achieved with the following 'coherent' reward function,

$$\tilde{r}_\alpha(\mathbf{s}, \mathbf{a}) = \alpha \ (\log \pi_{\mathrm{BC}}(\mathbf{a}|\mathbf{s}) - \log \pi_{\mathcal{U}}(\mathbf{a}|\mathbf{s})), \qquad (1)$$

where $\pi_{\mathrm{BC}}$ is the pre-trained BC policy and $\pi_{\mathcal{U}}$ is uniform prior policy used for maximum entropy regularization. The ratio is further scaled by a parameter $\alpha > 0$, which is set to be scale invariant w.r.t. the action dimension. A notable property of this coherent reward is that it is a 'shaped' variant of the true unknown reward function, and therefore has the same optimal policy [15, 8]. More intuitively, the reward function specified in (1) should reward state-action pairs for which the state and action are in-distribution, should penalise state-action pairs where the state is in-distribution and the action is not and should return a reward of 0 if the state is out-of-distribution, independent of the action. To achieve the desired reward function behaviour, it is required that $\pi_\alpha$ models the data $\mathcal{D}$ when in-distribution and models the uniform prior otherwise. In practice, $\pi_{\mathrm{BC}}$ is usually a function approximator. As function approximators represent correlations between states, updating their approximation for a subset of states usually changes the approximation for a large set of other states as well. As the reward function's desired behaviour requires $\pi_\alpha$ to match the prior for out-of-distribution states, the authors introduce a MLP architecture that is 'stationary' that approximates this quality. Meronen et

al. showed that a wide final layer with a periodic activation function induces the stationary property into the MLP [16]. Watson et al. specify a stationary policy $a = \tanh(z(\mathbf{s}))$, where $z(\mathbf{s}) = \mathbf{W}\phi(\mathbf{s})$, for which the weights are factorized row-wise $\mathbf{W} = [\mathbf{w}_1, ..., \mathbf{w}_{d_\alpha}]^\mathsf{T}, \mathbf{w}_i = \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ to define a Gaussian Process with independent actions [8]. By using change-of-variables as in SAC [17], the authors express the policy per-action as

$$\phi(\mathbf{s}) = f_{per}(\tilde{\mathbf{W}}\phi_{mlp}(\mathbf{s})),$$
$$q(\mathbf{a}_i|\mathbf{s}) = \mathcal{N}\left(z_i; \boldsymbol{\mu}_i^\mathsf{T}\phi(\mathbf{s})\phi(\mathbf{s})^\mathsf{T}\boldsymbol{\Sigma}_i\phi(\mathbf{s})\right) \cdot \left|\det\left(\frac{\mathrm{d}a_i}{\mathrm{d}z_i}\right)\right|^{-1},$$

where $f_{per}$ is a periodic activation function [16], the weights $\tilde{\mathbf{W}}$ are drawn from a distribution, e.g., a Gaussian that characterises the stochastic process, and $\phi_{mlp}$ is an arbitrary MLP [8]. They refer to this heteroscedastic stationary model as HETSTAT. The critic used for SAC is then initialised via

$$\tilde{Q}_1(\mathbf{s}, \mathbf{a}) = \tilde{r}_{\boldsymbol{\theta}_1}(\mathbf{s}, \mathbf{a}) + \gamma\big(\tilde{Q}_1(\mathbf{s}', \mathbf{a}') - \alpha(\log q_{\boldsymbol{\theta}_1}(\mathbf{a}' \mid \mathbf{s}') - \log \pi(\mathbf{a}' \mid \mathbf{s}'))\big),$$

where $\mathbf{s}, \mathbf{a}, \mathbf{s}', \mathbf{a}' \sim \mathcal{D}$ and $(\mathbf{s}', \mathbf{a}')$ is the successor state-action pair to $(\mathbf{s}, \mathbf{a})$. The reward's weights are then separated from the policy's weights such that both can be updated independently, and the policy and critic are updated using soft actor-critic (SAC) [17].

In MDPs with success-based absorbing states, positive rewards can incentivize agents to simply 'stay alive' and accumulate reward rather than complete the task. To prevent this, the coherent reward is bounded and shifted by its maximum value to remain strictly negative. By adding a small bias term $\sigma^2_{\min}$ to the predictive variance and bounding the $\tanh$ transformation, an upper bound for the reward $r(\mathbf{s}, \mathbf{a})$ is derived. Given action dimension $d_a$, a uniform prior over $[-1, 1]^{d_a}$, and $\alpha = d_a^{-1}$ [8]:

$$r(\mathbf{s}, \mathbf{a}) \leq -0.5 \cdot \log \pi\sigma^2_{min}/2 + \tilde{c},$$

where $\tilde{c}$ depends on the $\tanh$ clipping term [8].

### B. Accelerating Coherent Learning

While Watson et al. showed CSIL demonstrated strong empirical results, the implementation struggled to solve more complex tasks in the pixel-based setting [8]. To improve on CSIL's base performance, we integrated recent advances from the current state-of-the-art sample-efficient off-policy RL methods, which incorporate architectural regularization to facilitate more graceful scaling. In particular, following the results from Palenicek et al. [14], we adapt the CSIL critic and policy to use batch normalization (BN) [18] and weight normalization (WN) [19] instead of layer normalization [20]. Combining batch normalization and weight normalization improves the landscape of the optimization and helps maintain a steady 'effective' learning rate [19, 14]. We observed that Watson et al.'s CSIL implementation, which is based on SAC, suffered from exploding parameter values during training, which impedes learning as the 'effective' learning rate drops.

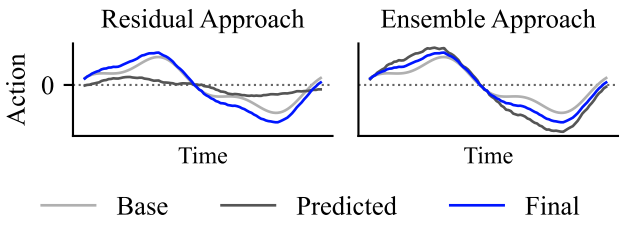


Fig. 3. Visualization of the different action modalities. The light gray line is the action generated by the base policy, the dark gray is the one predicted by the trainable policy, and the blue line is the final action used in the environment. The final action for the residual approach is calculated via Eq. 2 while the total action using the ensemble approach is calculated via Eq. 3.

Incorporating BN and WN into CSIL allowed us to mitigate these numerical issues and improve optimization and scaling. Incorporating this regularization also enables us to increase the update-to-data (UTD) ratio, where the critic is updated several times per environment step to accelerate learning. In accordance with the results from [21], we used policy delay and only updated our policy once per environment step. Additionally, while Watson et al. use a Monte Carlo KL estimator for the BC regularization term in the actor objective, we identify that this KL can be computed exactly between the latent Gaussian actions due to the shared bijection between policies [22]. To improve CSIL's performance in image-based settings, we adapted CSIL's original impala-style encoder [23] by increasing its size, training an encoder per camera instead of one for all cameras and utilizing spatial softmax pooling [24] at the end of the network to extract structural information efficiently. We refer to this improved implementation of CSIL as CSIL++ in the subsequent text. With CSIL++ as our policy improvement operator, we now consider how to improve LBMs effectively.

## IV. Coherent Fine-Tuning of Pre-Trained Policies

This section briefly outlines how CSIL is adapted to improve pre-trained action-chunking policies, in particular LBMs.

One important design decision is the RL action. This action could be the environment action, residual correction term to the LBM (Eq. 2), or an ensembled policy, averaging the base action and the predicted action (Eq. 3),

$$\boldsymbol{a}_{\text{residual}} = \boldsymbol{a}_{\text{base}} + \boldsymbol{a}_\pi, \tag{2}$$
$$\boldsymbol{a}_{\text{ensemble}} = \tfrac{1}{2}(\boldsymbol{a}_{\text{base}} + \boldsymbol{a}_\pi). \tag{3}$$

See Figure 3 for a visualization of both. Directly predicting environment actions, i.e., adapting the base policy directly, would require optimizing billions of parameters with RL, which is typically a prohibitive amount of compute even when using LoRA [25]. It would also require RL over action chunks rather than single actions. To learn an additional, smaller corrective policy, we use the LBM to predict an action chunk every $N$ steps (in our case $N = 10$) and a smaller model to predict a full action every step. Using a residual action, i.e., predicting a correction [26] on top of the base LBM policy, performed worse. We believe that this is because the residual prediction problem is challenging to

(a) Square (b) Coffee (c) Mug Cleanup

(d) Threading (e) Nut Assembly (f) Hammer Cleanup

Fig. 4. A visualization of the 6 manipulation tasks. In Square, the robot is tasked to place the wooden ring on the Square. Nut Assembly requires doing the same for the wooden ring and the silver ring. Coffee requires the capsule to be placed inside the machine and the lid to be closed. Mug Cleanup requires opening the drawer, putting the mug inside, and closing it, while Hammer Cleanup requires doing the same with the hammer. Threading requires picking up the needle and threading it through the tiny hole on top of the stick.

learn as it is similar to learning random noise, rather than a more 'meaningful' mapping from observation to expert action, especially as we are jointly learning internal feature representations from vision. To incorporate 'meaning' back into the action prediction problem, we propose an ensemble approach where the smaller RL policy is averaged with the LBM action. Moreover, rather than using the LBM's internal representation for RL, we found it better and faster to let the RL policy learn its own observation encoder, which is smaller than the LBM's and therefore easier to fine-tune.

## V. Experimental Setup

We test our approach on tasks from Robomimic [27] and Mimicgen [28], which vary in terms of task horizons, precision requirements and number of subtasks to be completed. Both Robomimic and Mimicgen rely on the Robosuite environment framework [29], which is built on MuJoCo [30]. All tasks are run at a control frequency of 20 Hz. The main section of our experiments relies on an image-based setting, where the policy has access to a static camera and wrist-mounted camera, as well as the proprioceptive state of the end-effector. For our CSIL++ ablation, we also used a state-based setting, where the robot has access to ground-truth information about the environment, such as the current pose of the relevant object, as well as the end-effector's proprioceptive state. Our non-CSIL-based approaches use a constant negative reward of $-1$, and all our experiments utilize the successful termination for bootstrapping. We train all methods across 3 independent seeds for 500k steps and evaluate the policy every 50k steps for 50 episodes. To compare the maximum performance of the approaches, we store the parameters with the highest evaluation success rate and evaluate them 5 times, with 50 episodes each, at the end of training for each run. We use $\pi_{0.5}$ [31] as our base model with 10 actions per action chunk.

**Tasks.** Our task selection covers three categories: simple short-horizon manipulation tasks, dexterous, short-horizon tasks and mid-range horizon tasks. For simple short-horizon tasks, we use Coffee from Mimicgen (see Figure 4 for environment visualizations). For dexterous tasks, we use Square from Robomimic and Nut Assembly and Threading from Mimicgen. For mid-range horizon tasks, we use Hammer Cleanup and Mug Cleanup. All tasks use a Franka robot with a parallel-jaw gripper and have an action space dimension of 7 — 6 for the delta end-effector pose, and 1 for the gripper. For Square we use the released expert demonstration dataset with 200 episodes and for all Mimicgen tasks, we use the Mimicgen pipeline to construct a dataset of 200 episodes from a source dataset of 10 episodes.

**Pre-training.** We fine-tune $\pi_{0.5}$ for 30k steps with a batch size of 128, combining all single-task datasets into one dataset, for 13 hours on 8 A100 GPUs.

**Baselines and Ablations.** First, we examine how CSIL++ on top of an action-chunked base policy compares to applying off-policy RL with sparse rewards while including expert demonstrations into the training loop. For this, we chose the state-of-the-art algorithm XQC [32], an off-policy approach utilizing a categorical critic with batch normalization between the hidden layers and furthermore normalizes the weights of the hidden layers to the unit sphere after every gradient step. To incorporate expert demonstrations into the training loop, we modify XQC to concatenate a batch samples from the expert dataset to the batch sampled from the online replay buffer. Additionally we strengthen the XQC baseline, by incorporating BC pre-training and the same HETSTAT architecture used in our approach. The main differences of our approach are thus its dense reward and replacing the entropy regularization with a KL-Divergence based regularization. To ensure a fair comparison to current literature, we evaluate both using the policy trained via XQC + offline data (XQC+OD) as a residual as well as averaging the actions from the frozen base policy and the action predicted by the policy (XQC+OD Ensemble). Since all evaluated methods' policies are initialized via BC, our replay buffer is well initialized during training, as in [5]. Additionally, we quantify the combined impact of our modifications on CSIL by comparing CSIL++ to vanilla CSIL on Square in both image-based and state-based settings.

**Implementation Details and Hyperparameters.** We summarize the specific architectural choices and hyperparameters used across our experiments in Table II.

## VI. Experimental Results

We have introduced both the components of our method that are independent of fine-tuning pre-trained policies and those that are specific to fine-tuning pre-trained models and will now evaluate the performance of our approach. In particular, we will consider two questions:

1) How does CSIL++ compare to CSIL?
2) Can CSIL++ improve the performance of an LBMs?

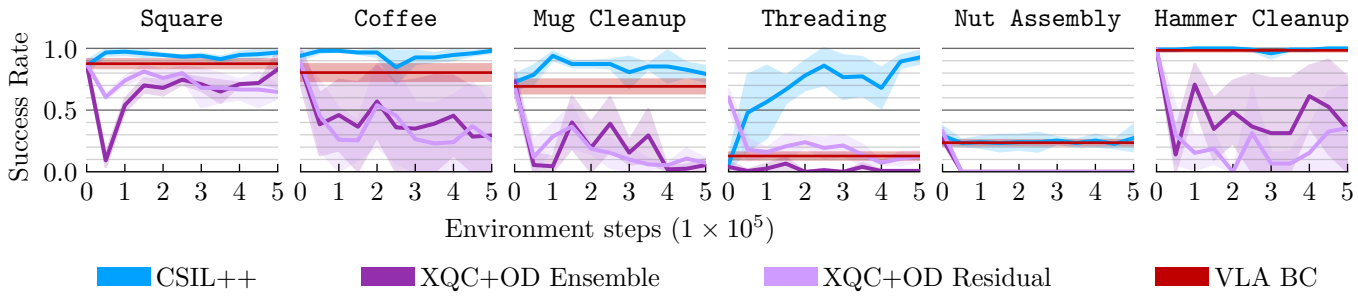


Fig. 5. Performance of CSIL++ and other baselines on 6 challenging simulation environments across three seeds. While CSIL++ combines the coherent reward with the XQC actor-critic algorithm, XQC+OD uses the demonstration data and sparse environment reward. Success rate is computed over 50 evaluations.

### A. *Coherent fine-tuning of Large Behaviour Models*

To evaluate our method, we conduct experiments on six environments from robomimic [27] and mimicgen [28]. Examining Fig 5, we first observe that CSIL++ converges to a very strong policy for almost all tasks except `Nut Assembly`.

When analyzing the training progression, we find that both baselines exhibit a detrimental loss of performance at the beginning of each run. While both the ensemble baseline and the residual baseline recover some performance for simpler tasks like `Square` and `Coffee`, they do not stably reclaim the performance achieved by $\pi_{0.5}$ on its own. In particular, this can be observed for tasks with a longer horizon, such as `Nut Assembly`, `Hammer Cleanup`, and `Mug Cleanup`, where the models are not able to recover their initial performance. We attribute this strong decline to the lack of regularization against moving away from the BC policy. On the other hand, CSIL++ is able to circumvent this initial drop through its main benefits: The *coherence* of policy, reward, and critic detailed in III prevents an immediate and severe degradation of performance commonly seen in RL and inverse RL. The dense reward formulation allows the critic to have a more accurate estimate of the value, and the closed-form KL-Divergence regularization against the BC policy allows CSIL to bias the search space towards searching for solutions close to the approximated expert policy learned via BC.

These benefits also allow CSIL++ to keep the near-perfect performance on `Hammer Cleanup` and significantly improve on the strong $\pi_{0.5}$ base performance for `Square`, `Coffee`, and `Mug Cleanup`. For `Square` and `Coffee`, our approach achieves a near $100\%$ success rate within 50k steps, while for `Mug Cleanup` it requires around 100k steps. Furthermore, for `Threading`, our approach improves the base policy's performance from roughly $13\%$ to $89\%$, demonstrating its potential for improving the performance and robustness of base policy even in the low success rate range. For `Square` and `Mug Cleanup`, a slight drop after the initial performance peak is observable. We believe this is caused by the categorical critic being unable to discern between rewards for expert and online states due to the discretization caused by the binning once performance becomes near-perfect leading to the policy deviating too far from the BC policy. For `Nut Assembly`, we observe that our approach only preserves $\pi_{0.5}$'s performance but does not improve it. Likely this is caused by the sequential nature of the task, requiring to first solve the task of `Square` and then solve a similar task again, making it much harder to gather valuable samples for the second part of the task where it's most crucial to correct compounding errors from the base model.

Despite this drop, the success rate of the best-performing parameters for each seed, reported in Table VI, demonstrates that our approach is able to outperform the base policy's performance as well as the other baselines for all tasks except `Nut Assembly`. Notably, the strongest performing parameters for each baseline run are the initial BC parameters, showing the detrimental effect of the missing *coherence* in policy and critic and the lack of regularization against moving too far away from the BC policy.

### B. *Scaling Coherent Imitation Learning*

To evaluate our modifications, we compare CSIL against our enhanced version, CSIL++, over 500k steps ($\sim 2,000$ rollouts) on the robomimic Square task in the state-based setting, where

TABLE I
SUCCESS RATES REPORTED AS MEDIAN, $[25^{th}, 75^{th}]$ PERCENTILES.

| **Method** | `Square` | `Coffee` | `Mug Cleanup` | `Threading` | `Nut Assembly` | `Hammer Cleanup` |
|---|---|---|---|---|---|---|
| CSIL++ Ensemble | **1.00 [0.99, 1.00]** | **0.98 [0.97, 0.98]** | **0.90 [0.81, 0.90]** | **0.90 [0.88, 0.91]** | 0.32 [0.27, 0.35] | **1.00 [1.00, 1.00]** |
| XQC+OD Ensemble | 0.85 [0.84, 0.85] | **0.98 [0.83, 0.99]** | 0.74 [0.60, 0.82] | 0.07 [0.04, 0.11] | 0.27 [0.26, 0.28] | 0.99 [0.98, 0.99] |
| XQC+OD Residual | 0.94 [0.94, 0.94] | 0.88 [0.87, 0.89] | 0.80 [0.80, 0.80] | 0.62 [0.62, 0.62] | **0.40 [0.40, 0.40]** | **1.00 [1.00, 1.00]** |
| VLA Base | 0.84 [0.82, 0.86] | 0.78 [0.76, 0.86] | 0.68 [0.66, 0.72] | 0.14 [0.12, 0.14] | 0.22 [0.22, 0.26] | 0.98 [0.98, 0.98] |

Success rates of best performing parameters in the evaluation during RL on 6 challenging manipulation tasks in simulation. **Bold** indicates highest median. Success rates are computed from 5 evaluations with 50 episodes each, and aggregate results across 3 seeds per environment and method are concatenated.

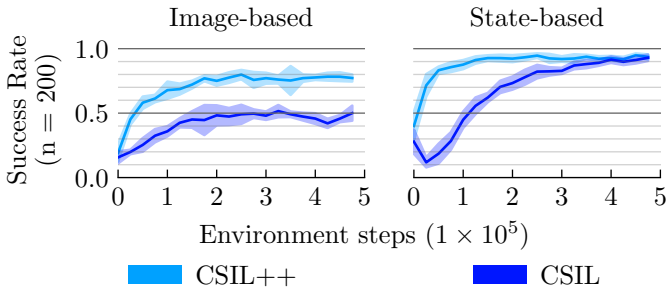


Fig. 6. Performance of CSIL and our improved version CSIL++ on the `Square` task. The 'Image-based' run takes images and the robot's proprioceptive state as observations. The 'State-based' operates on full state information comprising proprioception and ground-truth information of the object. In both settings, CSIL++ is more sample-efficient and improves performance significantly in the 'Image-based' setting.

the actor has access to the robot's proprioceptive state as well as ground truth information about the object, as well as in the image-based setting, where the actor has access to a static camera, a wrist mounted camera and the robot's propriocepive state.

In Figure 6, we are able to see that CSIL++ reaches $85\%$ within roughly 50k steps ($\sim$ 200 rollouts) in the state-based setting while vanilla CSIL takes around 250k steps, showing a 5x improvement in sample efficiency. The image-based performance gains are even more pronounced: CSIL++ reaches $70\%$ success in 50k steps and converges to $90\%$ by 150k steps, whereas vanilla CSIL reaches only $50\%$ after 500k steps. In summary, our modifications lead to strictly superior performance in terms of sample efficiency and in the image-based setting, even in absolute performance.

## VII. Discussion

We demonstrate that utilizing inverse reinforcement learning problem to produce a coherent, dense reward resolves the severe sample inefficiency and instability typically associated with refining LBMs using RL. By replacing sparse environment signals with a dense reward derived from expert demonstrations we enable significant policy improvement. A key insight from our experiments is that the standard residual objective creates an overly complex optimization landscape for BC-initialized policies. Our ensembling approach sidesteps this by predicting full single-step actions and averaging them with the LBM's chunked actions while allowing the policy to learn a residual action by predicting actions deviating from the optimal action to compensate for the LBM's errors.

The primary limitation of our approach is its reliance on the smaller policy's ability to adequately capture the state distribution during the initial BC phase. For highly complex or sequential tasks with large initial state distributions, the smaller model struggles to learn a sufficient prior. This limits the ensemble's ability to correct compounding errors in later stages of the trajectory, as the model cannot gather enough valuable online samples to learn from. Additionally, our method currently only implicitly integrates the vast prior knowledge of the LBM at the action level, without leveraging the base model's rich semantic understanding as input for its predictions.

Looking ahead, a promising direction to overcome these limitations is to share knowledge more explicitly between the models. Routing the LBM's intermediate embeddings directly into the smaller policy could dramatically improve the smaller model's expressiveness and contextual awareness. Furthermore, incorporating inductive biases, such as SE(3) equivariance, into the smaller policy's architecture could help it generalize across much broader initial state distributions, allowing this coherent fine-tuning framework to scale to more complex, multi-stage manipulation tasks.

TABLE II
Core Hyperparameters for CSIL++

| Hyperparameter | Value |
|---|---|
| *Architecture & Optimization* | |
| Batch Size | 128 |
| Base Network Sizes | Same as CSIL [8] |
| Learning Rates | Same as CSIL [8] |
| ResNet Channels | $[16, 32, 64, 128]$ |
| Visual Pooling | Spatial Softmax |
| Critic Normalization | Batch Normalization |
| Policy Normalization | Layer Normalization |
| Categorical Critic Bins | 101 |
| Categorical Critic Range | $[-8, 0]$ |
| Distribution of Bins | By Quantile |
| *Training Dynamics* | |
| Warmup Steps | 10,000 |
| Total Environment Steps | 510,000 |
| Update-To-Data (UTD) Ratio | 4 |
| Entropy Coefficient | 0.3 |
| Replay Buffer Size | 250,000 |
| *Pretraining* | |
| Demonstrations per Env. | 200 |
| Policy Pretraining Steps (Image) | 25,000 |
| Policy Pretraining Steps (State) | 50,000 |
| Critic Pretraining Steps | 20,000 |
| Environment Horizon | Standard (Task-dependent) |

## Acknowledgements

Christian Scherer is supported by the Konrad Zuse School of Excellence in Learning and Intelligent Systems (ELIZA) through the DAAD programme Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Education and Research. This research was funded by the research cluster "Third Wave of AI", funded by the excellence program of the Hessian Ministry of Higher Education, Science, Research and the Arts, hessian.AI and by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy (EXC-3057/1 "Reasonable Artificial Intelligence", Project No. 533677015). It was further supported by a UKRI/EPSRC Programme Grant [EP/V000748/1] and partially supported by the German Federal Ministry of Research, Technology and Space (BMFTR) under the Robotics Institute Germany (RIG). Ingmar Posner holds concurrent appointments as a Professor of Applied AI at the University of Oxford and as an Amazon Scholar. This paper describes work performed at the University of Oxford and is not associated with Amazon.

## References


[1] D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu *et al.*, "PaLM-E: An embodied multimodal language model," in *International Conference on Machine Learning (ICML)*, 2023.

[2] J. Barreiros, A. Beaulieu, A. Bhat, R. Cory, E. Cousineau, H. Dai, C.-H. Fang, K. Hashimoto, M. Z. Irshad, M. Itkina *et al.*, "A careful examination of large behavior models for multitask dexterous manipulation," *arXiv preprint arXiv:2507.05331*, 2025.

[3] L. Ankile, A. Simeonov, I. Shenfeld, M. Torne, and P. Agrawal, "From imitation to refinement-residual RL for precise assembly," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2025.

[4] X. Yuan, T. Mu, S. Tao, Y. Fang, M. Zhang, and H. Su, "Policy decorator: Model-agnostic online refinement for large policy model," in *International Conference on Learning Representations (ICLR)*, 2025.

[5] G. Ma, L. Li, H. Wang, Z. Liu, P.-L. Bacon, and D. Tao, "What makes value learning efficient in residual reinforcement learning?" *arXiv preprint arXiv:2602.10539*, 2026.

[6] J. Eschmann, "Reward function design in reinforcement learning," *Reinforcement learning algorithms: Analysis and Applications*, 2021.

[7] W. Xiao, H. Lin, A. Peng, H. Xue, T. He, Y. Xie, F. Hu, J. Wu, Z. Luo, L. Fan *et al.*, "Self-improving vision-language-action models with data generation via residual rl," in *International Conference of Learning Representation (ICLR)*, 2026.

[8] J. Watson, S. H. Huang, and N. Heess, "Coherent soft imitation learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[9] G. Lu, W. Guo, C. Zhang, Y. Zhou, H. Jiang, Z. Gao, Y. Tang, and Z. Wang, "VLA-RL: Towards masterful and general robotic manipulation with scalable reinforcement learning," 2025.

[10] K. Chen, Z. Liu, T. Zhang, Z. Guo, S. Xu, H. Lin, H. Zang, X. Li, Q. Zhang, Z. Yu, G. Fan, T. Huang, Y. Wang, and C. Yu, "$\pi_{\texttt{RL}}$: Online RL fine-tuning for flow-based vision-language-action models," 2026.

[11] L. Ankile, Z. Jiang, R. Duan, G. Shi, P. Abbeel, and A. Nagabandi, "Residual off-policy RL for finetuning behavior cloning policies," *arXiv preprint arXiv:2509.19301*, 2025.

[12] I. Kostrikov, A. Nair, and S. Levine, "Offline reinforcement learning with implicit q-learning," in *International Conference of Learning Representation (ICLR)*, 2022.

[13] S. Yue, X. Hua, J. Ren, S. Lin, J. Zhang, and Y. Zhang, "OLLIE: imitation learning from offline pretraining to online finetuning," in *International Conference on Machine Learning (ICML)*, 2024.

[14] D. Palenicek, F. Vogt, J. Watson, and J. Peters, "Scaling off-policy reinforcement learning with batch and weight normalization," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[15] A. Y. Ng, D. Harada, and S. J. Russell, "Policy invariance under reward transformations: Theory and application to reward shaping," in *International Conference on Machine Learning (ICML)*, 1999.

[16] L. Meronen, M. Trapp, and A. Solin, "Periodic activation functions induce stationarity," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[17] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *International Conference on Machine Learning (ICML)*, 2018.

[18] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015.

[19] C. Lyle, Z. Zheng, K. Khetarpal, J. Martens, H. van Hasselt, R. Pascanu, and W. Dabney, "Normalization and effective learning rates in reinforcement learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[20] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016.

[21] S. Fujimoto, H. van Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," 2018.

[22] Y. Polyanskiy, "Information theoretic methods in statistics and computer science: Lecture 1 — $f$-divergences," 2020.

[23] L. Espeholt, H. Soyer, R. Munos, K. Simonyan, V. Mnih, T. Ward, Y. Doron, V. Firoiu, T. Harley, I. Dunning, S. Legg, and K. Kavukcuoglu, "IMPALA: Scalable distributed deep-RL with importance weighted actor-learner architectures," in *International Conference on Machine Learning (ICML)*, 2018.

[24] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-end training of deep visuomotor policies," *Journal of Machine Learning Research (JMLR)*, 2016.

[25] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *International Conference on Learning Representations (ICLR)*, 2022.

[26] T. Silver, K. Allen, J. Tenenbaum, and L. Kaelbling, "Residual policy learning," *arXiv preprint arXiv:1812.06298*, 2018.

[27] A. Mandlekar, D. Xu, J. Wong, S. Nasiriany, C. Wang, R. Kulkarni, L. Fei-Fei, S. Savarese, Y. Zhu, and R. Martín-Martín, "What matters in learning from offline human demonstrations for robot manipulation," in *Conference on Robot Learning (CoRL)*, 2021.

[28] A. Mandlekar, S. Nasiriany, B. Wen, I. Akinola, Y. Narang, L. Fan, Y. Zhu, and D. Fox, "MimicGen: A data generation system for scalable robot learning using human demonstrations," in *Conference on Robot Learning (CoRL)*, 2023.

[29] Y. Zhu, J. Wong, A. Mandlekar, R. Martín-Martín,

A. Joshi, S. Nasiriany, Y. Zhu, and K. Lin, "robosuite: A modular simulation framework and benchmark for robot learning," in *arXiv preprint arXiv:2009.12293*, 2020.

[30] E. Todorov, T. Erez, and Y. Tassa, "MuJoCo: A physics engine for model-based control," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2012.

[31] P. Intelligence, K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, M. Y. Galliker, D. Ghosh, L. Groom, K. Hausman, B. Ichter, S. Jakubczak, T. Jones, L. Ke, D. LeBlanc, S. Levine, A. Li-Bell, M. Mothukuri, S. Nair, K. Pertsch, A. Z. Ren, L. X. Shi, L. Smith, J. T. Springenberg, K. Stachowicz, J. Tanner, Q. Vuong, H. Walke, A. Walling, H. Wang, L. Yu, and U. Zhilinsky, "$\pi_{0.5}$: a vision-language-action model with open-world generalization," 2025.

[32] D. Palenicek, F. Vogt, J. Watson, I. Posner, and J. Peters, "XQC: Well-conditioned optimization accelerates deep reinforcement learning," in *International Conference of Learning Representation (ICLR)*, 2026.