REMEMBERING HOW TO ACT THROUGH GREEDY ACTION CONSOLIDATION IN CONTINUAL REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

For value-based continual control, retaining a learned decision need not entail preserving the action values that originally supported it, since greedy action selection depends on their relative ordering rather than their absolute magnitudes. We propose Greedy Action Consolidation, or GAC, which targets the gaps that support historical greedy actions to mitigate catastrophic forgetting while allowing action value estimates to evolve. At the core of GAC lies the continual action gap, which measures the action value difference, under the updated Q function, between a historical greedy action and its best competing action, and is regularized to remain at least a prescribed fraction of its historical action gap. Theoretically, we derive an upper bound on return discrepancy in terms of the GAC regularizer and the distribution of historical action gaps, and show that GAC strictly enlarges the pointwise admissible action value set relative to exact value preservation as instantiated by Value Cloning, affording greater freedom for value adaptation. Experiments on MinAtar and Forager demonstrate the effectiveness of GAC across semi-continual and continual reinforcement learning settings, suggesting that the action gaps provide a parsimonious substrate for remembering how to act in continual reinforcement learning.

1 INTRODUCTION

For all the striking progress that reinforcement learning (RL) has achieved in complex yet stationary task settings (Mirhoseini et al., 2021; Fawzi et al., 2022; Mankowitz et al., 2023; Kaufmann et al., 2023; Seo et al., 2024; Guo et al., 2025; Hubert et al., 2025; Lei et al., 2026), current agents still lack the capacities for memory and continual adaptation that characterize human intelligence (Kudithipudi et al., 2022; Flesch et al., 2023), exhibiting a form of static intelligence at odds with the aspiration of learning that never stops (Abel et al., 2023; Elelimy et al., 2025). Intelligence is distinguished not only by the sophistication of behavior within any particular task, but also by the breadth and diversity of tasks that can be learned, retained, and integrated over a lifetime (Legg & Hutter, 2007). Continual task streams expose this deficiency sharply, as agents may lose plasticity or catastrophically forget prior knowledge, reflecting the classical *stability-plasticity dilemma* (Carpenter & Grossberg, 1987; Mermillod et al., 2013).

Among these intertwined failures, catastrophic forgetting (McCloskey & Cohen, 1989; French, 1999; Kirkpatrick et al., 2017) is particularly inimical to cumulative intelligence, since an agent unable to retain prior competence can only exchange one capability for another rather than expand its repertoire. Efforts to mitigate catastrophic forgetting in RL have accordingly focused on various carriers of prior knowledge, ranging from parameters (Kirkpatrick et al., 2017; Elsayed & Mahmood, 2024) and stored experience (Riemer et al., 2019; Rolnick et al., 2019) to representations (Javed & White, 2019; Meyer et al., 2024; Lan et al., 2025), structured parameterizations of policies and value functions (Kaplanis et al., 2018; 2019; Gaya et al., 2023; Anand & Precup, 2023), and predictive world models (Kessler et al., 2023; Kauvar et al., 2023; Fu et al., 2025; Alyahya et al., 2026). Despite their methodological diversity, these approaches converge on a common strategy of operationalizing retention by stabilizing a chosen carrier of prior knowledge, yet the resulting plurality of preservation targets, whose drift need not imply a change in the agent's decisions (Liu et al.,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2023), leaves a more fundamental question unresolved: *which decision-relevant property offers a more direct target for retention in continual reinforcement learning?*

For a value-based RL agent, the induced greedy behavior at state s depends on the ordering of action values, not on their absolute magnitudes. This observation suggests a decision-centric view of catastrophic forgetting in value-based control, under which what must be retained is not necessarily the numerical action value landscape itself, but the greedy decisions that the landscape induces. Guided by this principle, we define the *continual action gap* as the action value difference, under the updated Q function, between a historical greedy action and its best competing action, thus distinguishing it from the conventional action gap (Farahmand, 2011) defined around the current greedy action. We then propose *Greedy Action Consolidation* (GAC), a hinge-based regularizer that uses an auxiliary memory of historical states and their associated greedy actions and action gaps to penalize the continual action gap whenever it falls below a prescribed fraction of the historical action gap, thereby targeting decision stability while allowing the underlying value estimates to evolve.

Contributions. Our study advances continual value-based control in three respects:

- We formalize behavioral forgetting as the mismatch rate between historical and current greedy actions and introduce the continual action gap, establishing a theoretical connection between the two. Building on this connection, we develop GAC, a principled regularizer that constrains the continual action gap to remain above a prescribed fraction of the historical action gap.
- We connect the GAC objective to control performance by deriving, under a discounted state visitation matching condition, an upper bound on return discrepancy governed jointly by the regularizer and the distribution of historical action gaps, while proving that the pointwise admissible action value set of GAC strictly contains that of exact Value Cloning and thus affords greater flexibility for subsequent value adaptation.
- We evaluate GAC across both semi-continual and continual reinforcement learning settings, showing that low behavioral forgetting can coexist with substantial action value drift, while tighter value retention need not yield better continual adaptation. These results provide empirical support for a decision-centric perspective on continual value-based control grounded in preserving action gaps across tasks.

2 Related Work

Sustained plasticity alone does not ensure that successive experience accumulates into lasting knowledge, which motivates our focus on catastrophic forgetting. One line of work mitigates forgetting by shaping learning through updates informed by taskwise gradient statistics (Malviya et al., 2022), gradients and perturbations scaled by parameter utility (Elsayed & Mahmood, 2024), representations learned to resist interference (Javed & White, 2019), sparse categorical encodings (Meyer et al., 2024), and activation functions with sparse outputs and gradients (Lan et al., 2025). Meta-experience replay (Riemer et al., 2019) combines a reservoir of past experience with Reptile-style meta-learning (Nichol et al., 2018) to encourage gradient alignment between current and replayed samples. Together, these approaches shape learning dynamics or representations, offering potential complements to our explicit consolidation of historical greedy actions.

Parameter consolidation enforces retention through the weights that encode prior knowledge. EWC penalizes parameter displacement using diagonal Fisher information as a local measure of parameter importance (Kirkpatrick et al., 2017), while online EWC maintains a discounted aggregate Fisher and recenters the penalty at the latest parameter solution (Schwarz et al., 2018). Kaplanis et al. (2018) introduce the Benna-Fusi synaptic model into DQN, coupling each weight to interacting variables that consolidate its update history across multiple timescales. Unlike these weight-level mechanisms, our method imposes its retention constraint on action value gaps in function space, drawing on the broader idea of regulating learning through functional changes (Asadi et al., 2024; Sharifnassab et al., 2026). Within this function-space perspective, Rolnick et al. (2019) combine on-policy learning from new experience with off-policy replay, adding KL regularization toward stored policies and regression toward stored state values, while Lan et al. (2023) regularize current action-value vectors toward those of a target network at sampled states to support memory-efficient learning. These value cloning terms penalize deviations from reference predictions, whereas our

method specifically consolidates the action-gap structure supporting historical greedy actions, allowing both parameters and numerical value estimates to evolve without requiring pointwise value agreement. Other methods organize adaptation and retention through distinct learners. PT-DQN (Anand & Precup, 2023) maintains shared knowledge across tasks in a permanent value function that provides a jump-start baseline for rapid adaptation through a transient component. Nevertheless, retaining this baseline alone does not necessarily preserve greedy actions across previously encountered tasks, which are induced by the sum of both the permanent and transient components. Paralleling PT-DQN's separation of adaptation and retention, FAME (Sun et al., 2026) pairs a meta Q-learner for knowledge integration, updated by incremental maximum likelihood derived from an MDP distance-based forward KL objective, with a fast Q-learner for task adaptation that can be warmed up through behavior cloning of the meta learner. FAME measures forgetting through value and policy discrepancies, whereas our measure tracks the mismatch rate between historical and current greedy actions. Unlike FAME, our method penalizes violations of the historical actions' gap constraints with a hinge objective, without fitting action probabilities. Further discussion of related work is provided in Appendix A.

3 Problem Setting

Task-agnosticity. We study task-agnostic continual RL, in which an agent encounters a sequence of tasks without observing their identities (Caccia et al., 2023). Let $(\mathcal{S}, \Sigma)$ be a measurable state representation space. Each task $k \in \{1, \ldots, K\}$ is modeled as a discounted Markov decision process $\mathcal{M}_k = \langle \mathcal{S}, \mathcal{A}, \mathcal{P}_k, \mathcal{R}_k, \nu_k, \gamma \rangle$ where $\mathcal{P}_k$, $\mathcal{R}_k$, and ν_k denote the transition kernel, reward function, and initial state distribution, while all tasks share a common state representation space $\mathcal{S}$, the finite action space $\mathcal{A}$, and the discount factor γ. States reachable under different tasks may lie in distinct subsets $\mathcal{S}_k \subseteq \mathcal{S}$, so sharing $\mathcal{S}$ ensures compatibility with a single function approximator without presupposing semantic correspondence across tasks. We assume uniformly bounded rewards satisfying $|r_k(s,a)| \leq R^\star$ for all k and $(s,a) \in \mathcal{S} \times \mathcal{A}$. Throughout, $\hat{a}^\star$ denotes a learned greedy action, not necessarily an optimal one. We define the expected discounted return of greedy policy π as $J_k(\pi) = \mathbb{E}_{\nu_k, \mathcal{P}_k, \pi}\left[\sum_{t=0}^{\infty} \gamma^t r_k(s_t, a_t)\right]$ and the discounted state visitation distribution as $d_k^\pi(B) = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \Pr_{\nu_k, \mathcal{P}_k, \pi}(s_t \in B)$ for every measurable $B \in \Sigma$.

Boundary Observability. Following Anand & Precup (2023), we consider semi-continual and continual RL. In our semi-continual setting, task switches are signaled, although the signal is neither supplied to the value function nor used to condition action selection, and its known periodicity is used only to set schedule-dependent hyperparameters. In the continual setting, neither task identities nor boundaries are revealed, so all hyperparameters and update schedules are specified independently of underlying task changes.

4 Greedy Action Consolidation

4.1 Behavioral Forgetting and Continual Action Gap

Under greedy evaluation, a value-based agent's decisions depend on which actions have the highest estimated values, not on the values' absolute magnitudes. Historical greedy actions may thus remain unchanged even as their value estimates evolve. We measure forgetting on states representative of a previously encountered task by asking whether the current Q function still induces the same greedy actions as its historical snapshot.

At the completion of task j, let θ_j denote the frozen reference parameters of task j, while θ_k denotes the current parameters after subsequent learning through task $k \geq j$. Under a fixed deterministic tie-breaking rule shared across learning stages, the historical and current greedy actions are defined as $\hat{a}_j^\star(s) := \arg\max_{a \in \mathcal{A}} Q_{\theta_j}(s,a)$ and $\hat{a}_k^\star(s) = \arg\max_{a \in \mathcal{A}} Q_{\theta_k}(s,a)$. Let μ_j denote a reference state distribution on $\mathcal{S}$, formally a probability measure on $(\mathcal{S}, \Sigma)$, fixed with θ_j and independent of θ_k. It may be instantiated by the historical discounted state visitation distribution or by a fixed empirical approximation retained from task j.

Definition 1 (Behavioral Forgetting)**.** *For a historical task j, a learning stage $k \geq j$, and a fixed reference distribution μ_j, behavioral forgetting is defined as*

$$F_{j \to k}(\mu_j) = \mathbb{E}_{s \sim \mu_j}\left[\mathbb{I}\left\{\hat{a}_k^\star(s) \neq \hat{a}_j^\star(s)\right\}\right], \tag{1}$$

where $\mathbb{I}$ denotes the indicator function.

Thus $F_{j\to k}(\mu_j) \in [0,1]$ is the probability that the current greedy policy selects an action different from the historical reference, and it vanishes precisely when the two greedy policies agree μ_j-almost surely. A larger behavioral forgetting indicates more frequent departures from historical decisions after adaptation, potentially changing the return on task j and reducing it when the historical greedy policy is optimal.

At each reference state, behavioral forgetting depends on how the current Q function ranks the historical greedy action against its highest-valued competing action. The conventional action gap measures the difference between the current largest and second-largest action values (Farahmand, 2011). To trace this difference after a task has passed, we retain its historical greedy action as the comparison reference.

Definition 2 (Continual Action Gap)**.** *Let $\hat{a}_j^\star(s)$ denote the greedy action induced by the reference action value function Q_{θ_j} after learning task j. For any subsequent learning stage $k \geq j$, its continual action gap at state s is*

$$M_{j\to k}(s) := Q_{\theta_k}\big(s, \hat{a}_j^\star(s)\big) - \max_{a\in\mathcal{A}\setminus\{\hat{a}_j^\star(s)\}} Q_{\theta_k}(s,a). \tag{2}$$

At $k = j$, the continual action gap reduces to the conventional action gap of the historical snapshot and is therefore nonnegative, while replacing the historical reference throughout Equation 2 with $\hat{a}_k^\star(s)$ yields the conventional gap of the current value function. Once learning proceeds beyond task j, however, $\hat{a}_j^\star(s)$ need no longer maximize Q_{θ_k}, making $M_{j\to k}(s)$ a signed measure of whether the historical decision remains separated from its current competing actions. Under the fixed tie-breaking rule, a positive continual action gap certifies retention of the historical greedy action, a negative gap certifies its displacement, and a zero gap leaves the outcome to tie breaking. Consequently, positivity μ_j-almost surely is sufficient for zero behavioral forgetting and becomes equivalent whenever μ_j assigns no mass to ties. By recasting discrete policy disagreement as a signed decision margin, the continual action gap supplies the surrogate structure on which greedy action consolidation is built.

4.2 Greedy Action Value Retention Is Not Enough

A natural hypothesis is that preserving the values attained by greedy actions may suffice to preserve control. To examine this hypothesis, we quantify greedy value forgetting by replacing the indicator in the behavioral forgetting measure defined in Equation 1 with the squared error between $Q_{\theta_j}\big(s, \hat{a}_j^\star(s)\big)$ and $Q_{\theta_k}\big(s, \hat{a}_j^\star(s)\big)$, thereby measuring the numerical drift through stage k in the value assigned to task j's historical greedy action.

Adapting the MinAtar experimental protocol of Anand & Precup (2023), we compare PT-DQN with DQN on a stream cycling through Breakout, Space Invaders, and Freeway, with the environment switching every one million steps, and track the lifetime return, behavioral forgetting, and greedy value forgetting. Although the upper panel in the leftmost column of Figure 1 shows that PT-DQN exhibits markedly less greedy value forgetting than DQN, the lower panel in the same column reveals no commensurate reduction in behavioral forgetting, while the three rightmost panel columns reveal that PT-DQN loses performance on previously visited tasks after each switch in much the same way as DQN. This result exposes the insufficiency of numerical value retention, since PT-DQN can substantially stabilize the value assigned to a historical greedy action while non-stationarity shifts its competing actions enough to reverse their ordering, displacing that action from the maximum of the current task's Q-function and leaving decision-level forgetting unabated.

4.3 Retaining Decisions by Maintaining Continual Action Gaps

The empirical evidence in Section 4.2, together with the characterization in Section 4.1, shows that behavioral retention depends on keeping a historical greedy action above its competing actions rather than preserving its numerical value. We accordingly introduce **Greedy Action Consolidation** (GAC). Since a positive continual action gap preserves the historical greedy action, GAC converts each historical action gap into a soft constraint on the current Q-value landscape, penalizing shortfalls of the continual action gap below a scaled historical action gap that could otherwise leave the historical decision vulnerable to forgetting.

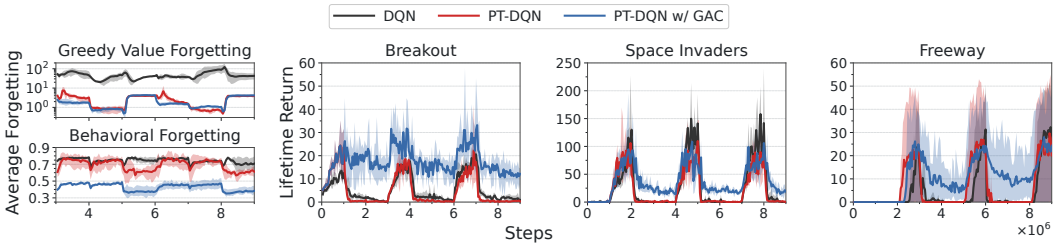


Figure 1: Forgetting and lifetime performance in the semi-continual MinAtar experiment, where the agent cycles through Breakout, Space Invaders, and Freeway in one-million-step phases. The leftmost column reports average forgetting, with greedy value forgetting shown in the upper panel and behavioral forgetting shown in the lower panel. The three rightmost panels show lifetime evaluation return on each task over the full training stream, with higher returns and milder post-switch declines indicating better performance.

Let $M_j(s) := M_{j\to j}(s) = Q_{\theta_j}\big(s, \hat{a}_j^\star(s)\big) - \max_{a\in\mathcal{A}\setminus\{\hat{a}_j^\star(s)\}} Q_{\theta_j}(s,a)$ denote the action gap at state s for task j. Given a reference distribution μ_j and a retention coefficient $\rho > 0$, the GAC regularizer at stage $k \geq j$ is

$$\begin{aligned}\mathcal{R}_{j\to k}(\theta_k) &= \mathbb{E}_{s\sim\mu_j}\Big[\big[\rho M_j(s) - M_{j\to k}(s)\big]_+\Big] \\ &= \mathbb{E}_{s\sim\mu_j}\left[\max_{a\in\mathcal{A}\setminus\{\hat{a}_j^\star(s)\}}\Big[Q_{\theta_k}(s,a) - Q_{\theta_k}\big(s,\hat{a}_j^\star(s)\big) + \rho M_j(s)\Big]_+\right],\end{aligned} \tag{3}$$

where $[x]_+ = \max\{x, 0\}$. A vanishing regularizer implies $M_{j\to k}(s) \geq \rho M_j(s) \geq 0$ for μ_j almost every state, ensuring that the historical greedy action $\hat{a}_j^\star$ remains a maximizer under Q_{θ_k} after learning task k, while the action values themselves remain free to evolve. As shown in Figure 1, equipping PT-DQN with GAC substantially preserves post-switch performance, offering initial empirical support for our core intuition that continually maintaining the action gaps supporting historical greedy actions protects behavior more effectively than retaining greedy action values alone.

Practical Implementation. GAC maintains an auxiliary memory of triples $\big\{s, \hat{a}_j^\star(s), M_j(s)\big\}$ to retain historical greedy actions and their associated action gaps. We populate this memory through greedy action alignment, admitting a record only when the online network and its lagged target network agree on the greedy action at the corresponding state.

4.4 Theoretical Analysis

We first bound behavioral forgetting, as well as the return discrepancy when μ_j matches the historical policy's normalized discounted occupancy, in terms of the GAC regularizer and the probability mass of states with small historical action gaps, and then characterize the additional pointwise flexibility afforded by GAC.

Theorem 1 (**Upper Bound on Return Discrepancy**). *Let $\rho > 0$, and fix a single deterministic tie-breaking rule to be used throughout the learning process. Suppose that the GAC regularizer is defined with $\mu_j = d_j^{\pi_j}$. Then,*

$$\begin{aligned}|J_j(\pi_k) - J_j(\pi_j)| &\leq \frac{2R^\star}{(1-\gamma)^2} F_{j\to k}\left(d_j^{\pi_j}\right) \\ &\leq \frac{2R^\star}{(1-\gamma)^2} \inf_{\delta>0}\left\{\Pr_{s\sim d_j^{\pi_j}}\left(M_j(s) \leq \delta\right) + \frac{\mathcal{R}_{j\to k}(\theta_k)}{\rho\delta}\right\}.\end{aligned} \tag{4}$$

Theorem 1 shows how behavioral forgetting measured with respect to the historical policy's normalized discounted state occupancy translates into performance retention on the corresponding task. Setting aside the infimum, increasing δ attenuates the regularizer term with $\delta \to \infty$ recovering the vacuous bound $F_{j\to k}(\mu_j) \leq 1$. When the historical action gap distribution instead assigns negligible mass to $M_j(s) \leq \delta$ for some positive δ, reducing the GAC regularizer directly tightens the return discrepancy guarantee. The proof is deferred to Appendix D.1.

Theorem 2 (**Pointwise Relaxation of Exact Value Cloning**). *Fix a finite action space $\mathcal{A}$ with $|\mathcal{A}| \geq 2$ and a retention coefficient $0 < \rho \leq 1$. For any state s, let $\hat{a}_j^\star(s)$ be the historical greedy action under a fixed tie-breaking rule, and write $\boldsymbol{Q}_{\theta_j}(s) = [Q_{\theta_j}(s,a)]_{a\in\mathcal{A}} \in \mathbb{R}^{|\mathcal{A}|}$. The exact value-cloning set $\mathcal{Q}_j(s) = \{\boldsymbol{Q}_{\theta_j}(s)\}$ and the GAC-feasible set $\mathcal{Q}_j(s;\rho) = \left\{\boldsymbol{Q} \in \mathbb{R}^{|\mathcal{A}|} \,\middle|\, Q\big(s,\hat{a}_j^\star(s)\big) - \max_{a\in\mathcal{A}\setminus\{\hat{a}_j^\star(s)\}} Q(s,a) \geq \rho M_j(s)\right\}$ satisfy*

$$\mathcal{Q}_j(s) \subsetneq \mathcal{Q}_j(s;\rho)$$

The GAC set is therefore a closed, convex, full-dimensional, unbounded polyhedron, whereas exact value cloning admits only the single historical vector $\boldsymbol{Q}_{\theta_j}(s)$.

Theorem 2 formalizes the additional flexibility afforded by GAC, as the full dimensionality of $\mathcal{Q}_j(s;\rho)$ prevents the retention constraint from confining the action value vector to a lower dimensional manifold, allowing strictly feasible configurations to accommodate subsequent learning without relinquishing the historical greedy action. The proof and an extension showing that sufficiently accurate approximate Value Cloning remains contained within the GAC feasible set are provided in Appendix D.2.

5 Experiments

We evaluate our method on MinAtar (Anand & Precup, 2023) and Forager (Tang et al., 2026). MinAtar primarily serves to evaluate semi-continual RL, in which task switches are signaled, whereas Forager places the agent in a continual partially observable grid world where a restricted local field of view and unannounced reward changes must be resolved from a single stream of experience, thereby furnishing a testbed for fully continual RL.

Baselines. On MinAtar, we benchmark against four continual RL baselines comprising EWC (Kirkpatrick et al., 2017), PT-DQN (Anand & Precup, 2023), and Value Cloning, a value-based adaptation of CLEAR (Rolnick et al., 2019) that omits policy cloning, and FAME (Sun et al., 2026). We additionally evaluate Online EWC (Schwarz et al., 2018) on Forager but omit FAME because its task-boundary-dependent updates require nontrivial redesign for boundary-free continual RL. Complete hyperparameter specifications and further experimental details are deferred to Appendix B.

Evaluation Metrics. The *Average Performance* and *Forgetting* metrics used in prior work respectively evaluate only the final policy and compare task-end performance with sequence-end performance, compressing the dynamics of continual learning into a few endpoint measurements (Ahn et al., 2025; Sun et al., 2026). Since little acquisition leaves little to forget, low forgetting alone does not certify retention. We instead evaluate performance over the learning lifetime, with the concrete metric matched to the benchmark protocol. In MinAtar, where tasks have identifiable phases and can be evaluated separately, we define *Lifetime Return* as taskwise evaluation return over the entire learning stream, irrespective of the task currently being learned. The corresponding normalized AUC increases when forward transfer accelerates acquisition and resistance to forgetting preserves performance after task switches. In Forager, the agent learns and acts in a continuing environment along a single uninterrupted stream of experience. Because separate taskwise evaluation rollouts are not part of this fully online protocol, we report the moving average reward obtained during training and its normalized AUC as the online counterparts of lifetime performance.

5.1 Semi-Continual Reinforcement Learning on MinAtar

Adapting prior MinAtar protocols (Anand & Precup, 2023; Sun et al., 2026), we construct three task sequences by cyclically shifting the base ordering of Breakout, Space Invaders, and Freeway. Each sequence is evaluated over three random seeds and spans three cycles with 1 million steps per phase, yielding nine phases and 9 million interactions to more thoroughly evaluate catastrophic forgetting. Under the semi-continual RL setting, GAC, Value Cloning and FAME update their auxiliary memories at every task transition, while EWC stores the Fisher information and corresponding parameter snapshot from each phase. Complete experimental details are provided in Appendix B.1.

Main Results. Since lifetime return is intended to capture both online adaptation and accumulated knowledge, neither FAME component alone spans both roles (Sun et al., 2026), and we thus report the curves and normalized AUC of both learners. The results are reported in Table 1 and Figure 2.

Table 1: Average normalized AUC of lifetime return on semi-continual MinAtar environments.

Method	Avg. Norm. AUC ↑		
	Breakout	Space Inv.	Freeway
DQN	5.530±0.096	25.486±2.948	6.187±1.415
DQN w/ EWC (Kirkpatrick et al., 2017)	7.403±0.274	29.235±3.787	5.024±1.351
DQN w/ Value Cloning (Rolnick et al., 2019)	13.974±0.582	**93.731±6.680**	16.775±3.150
DQN w/ GAC (ours)	**14.858±0.365**	88.282±19.797	**26.121±2.988**
DQN w/ FAME (meta learner) (Sun et al., 2026)	10.170±0.271	16.798±4.114	11.542±3.351
DQN w/ FAME (fast learner) (Sun et al., 2026)	4.737±0.770	26.338±1.002	11.077±0.095
PT-DQN (Anand & Precup, 2023)	4.914±0.064	21.529±1.176	6.327±0.283
PT-DQN w/ GAC (ours)	14.764±2.096	36.539±4.503	12.303±1.573

Our method GAC delivers the strongest overall lifetime performance, when incorporated into PT-DQN, produces markedly higher and more stable lifetime returns. The two rightmost panels of Figure 2 reveal a clear dissociation between value and action retention. Value Cloning exhibits the least greedy value forgetting, whereas GAC achieves the lowest behavioral forgetting and a pronounced performance advantage on Freeway, suggesting that, remembering how to act by maintaining the action gaps that support historical greedy actions can retain behavior more directly and effectively than remembering the values once assigned to those actions. We further evaluate GAC and Value Cloning in an online MinAtar setting with unobservable task boundaries, where preliminary results indicate that GAC continues to exhibit superior overall performance to Value Cloning. Further details are provided in Appendix C.

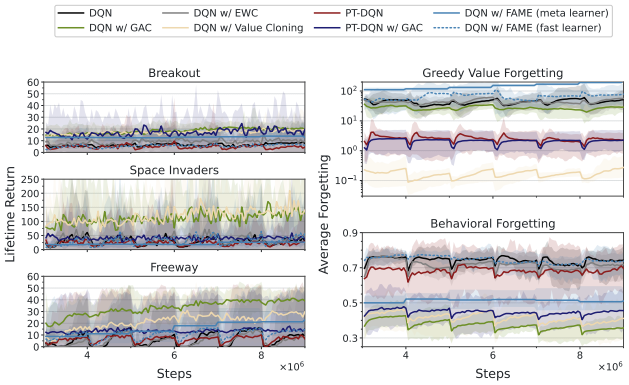


Figure 2: **Semi-continual reinforcement learning on MinAtar**. The three left panels report the average lifetime return of each task across the cyclically shifted sequences from 3 million environment steps onward, after every task has been encountered once. For any fixed task, the cyclic shifts cause this average to pool three task-exposure histories corresponding to the current, immediately preceding, or second preceding phase across the three sequences, thereby characterizing performance under distinct task-recency conditions.

Hyperparameter Sensitivity. Motivated by Mesbahi et al. (2025), we limit hyperparameter exposure across task sequences by tuning on a single sequence and deploying the selected configuration unchanged on the remaining sequences. For GAC, the retention coefficient ρ determines how tightly the continual action gap is constrained relative to its historical counterpart, thereby modulating the stringency of behavioral retention. While our theory covers $0 < \rho \leq 1$, we include $\rho = 0$ as a boundary ablation and evaluate $\rho \in \{0, 0.5, 1\}$. Figure 4a shows that GAC remains above the baseline throughout this candidate set, indicating that its gains do not depend on finely tuning ρ within the tested range. Figure 4b reports a shared regularization coefficient sweep over $\{0.01, 0.1, 1, 3\}$, across which GAC exhibits a stronger overall performance profile than Value Cloning. The corresponding sensitivity analysis for EWC is reported in Appendix B.1.

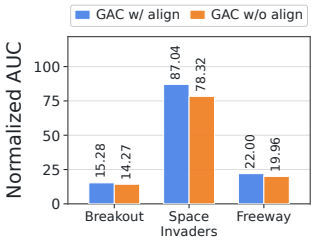
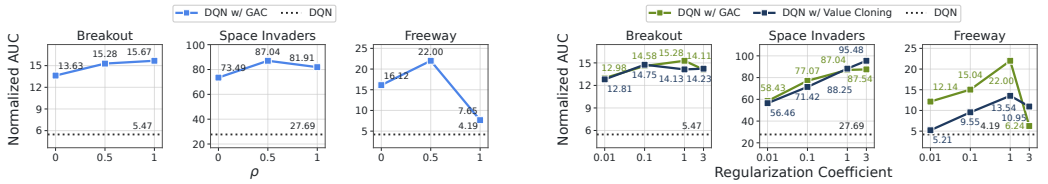


Figure 3: Ablation on greedy action alignment.

Ablation. Before the action value estimates on a sampled state have stabilized, their ordering can remain unreliable, so the action selected by greedy policy may reflect a transient preference rather than the greedy action that would prevail after convergence. Greedy action alignment uses a lightweight consistency check to reduce the risk of turning estimation instability into persistent behavioral constraints. As shown in Figure 3, incorporating alignment into DQN with GAC produces consistent yet modest gains.



(a) (b)

Figure 4: **Hyperparameter sensitivity on semi-continual MinAtar.** (a) Effect of the retention coefficient ρ on GAC. (b) Comparison of GAC and Value Cloning across regularization coefficients.

5.2 Continual Reinforcement Learning on Forager

Forager (Tang et al., 2026) is a lightweight continual grid world that combines perceptual aliasing with latent nonstationarity while avoiding the computational burden of visually elaborate benchmarks. We adopt a configuration that closely follows the two-biome never-ending relearning experiment of Tang et al. (2026): a square-wave task in a 24×15 world with a 9×9 local field of view, as depicted in Figure 5. Mushroom rewards reverse every 250,000 interactions and complete a square wave cycle every 500,000 steps, although neither the switch nor its periodicity is revealed to the agent. Because colors are shared across biomes and no visual cue announces a reward reversal, the same local evidence can warrant incompatible behavior under different latent regimes. Each agent undergoes 10 million uninterrupted interactions without episodic termination, while task identity, phase, clock, and boundary indicators are withheld from both policy evaluation and optimization. The resulting protocol realizes task-agnostic continual control in which adaptation must be elicited from experience rather than synchronized to privileged task boundaries. The full environment configuration and hyperparameter settings for each method are provided in the Appendix B.2.

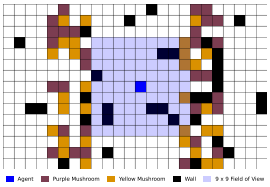


Figure 5: Illustration of the two-biome Forager environment.

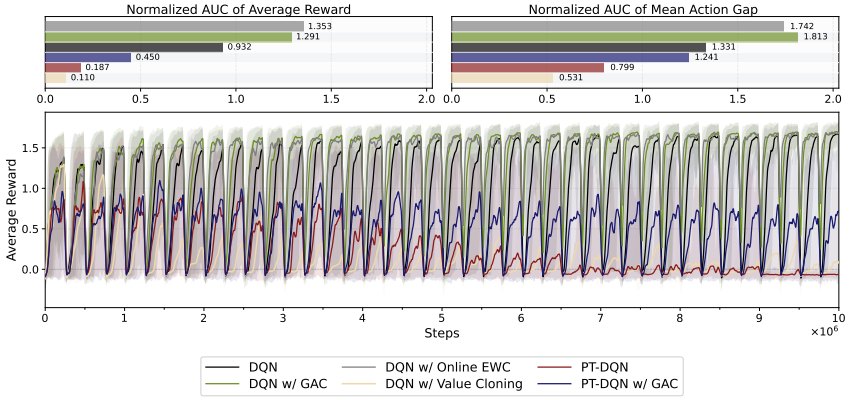
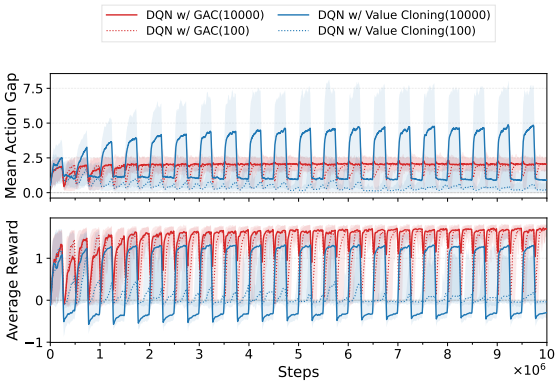
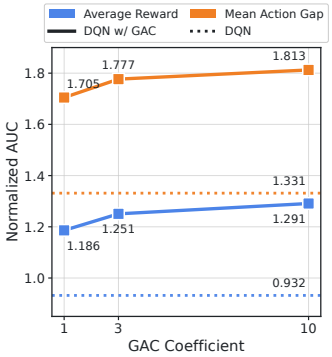


Figure 6: **Continual reinforcement learning on the two-biome Forager**. The upper-left and upper-right panels report the normalized AUCs of average reward and mean action gap over the full training horizon, respectively, while the lower panel presents the corresponding learning curves averaged over 15 independent runs.

Main Results. As illustrated in Figure 6, neither PT-DQN nor Value Cloning achieves effective and sustained learning in Forager, with both undergoing an acute performance collapse late in training. This collapse is unlikely to be attributable solely to a general loss of plasticity, since the baseline DQN continues to learn effectively throughout training. Instead, a plausible explanation is that, because PT-DQN and Value Cloning both retain previously learned values, state aliasing induced by partial observability and periodic reward reversals repeatedly drive their action value functions toward mutually incompatible regression targets, allowing interference to accumulate and eventually cause collapse. Conversely, GAC affords TD learning the latitude to readjust action values as the reward mapping changes. GAC not only improves the performance of DQN but also alleviates the



(a) **Scaling results on the two-biome Forager environment**. Results are averaged over 15 independent runs.



(b) Sensitivity of greedy action consolidation to the regularization coefficient.

Figure 7: Scaling results and effect of the regularization coefficient on Forager.

late collapse of PT-DQN. Although DQN with Online EWC achieves a 0.062 advantage over GAC in the normalized AUC of average reward, Online EWC must additionally retain the consolidated Fisher information and the most recent parameter vector, whose combined auxiliary storage is on the order of $2P$ for a model with P parameters. So the auxiliary memory of Online EWC grows as $\mathcal{O}(P)$ and remains intrinsically tied to model size. GAC also incurs a storage cost, but its principal overhead arises from a configurable auxiliary memory whose capacity can be chosen independently of model size. As the following paragraph subsequently demonstrates, scaling this auxiliary memory yields further performance improvements. Among all evaluated methods, GAC attains the largest mean action gap, a result consistent with more stable action selection under its constraint.

Scaling the Memory Size. With the DQN replay buffer size fixed at 1,000, we increase the auxiliary memory size of GAC and Value Cloning from 100 to 10,000. Figure 7a shows that a larger memory mitigates the deterioration observed with Value Cloning at the smaller size, yet leaves pronounced oscillations between positive and negative rewards across successive task phases. This pattern is consistent with unresolved conflicts between historical value targets under state aliasing and reward reversals, suggesting that greater memory size alone does not resolve the underlying interference. GAC, by contrast, benefits substantially from the increased memory size, exhibiting a highly stable mean action gap curve and markedly smaller reward fluctuations during training, suggesting reduced catastrophic forgetting at the level of decision.

Hyperparameter Sensitivity. Figure 7b shows that DQN with GAC improves the normalized AUC of both average reward and mean action gap over DQN at all three tested regularization coefficient values of 1, 3, and 10. Both metrics increase across these values, supporting the robustness of GAC's gains to coefficient selection in the evaluated Forager setting.

6 Conclusion

In this work, we introduced the continual action gap, establishing its connection to behavioral forgetting, and proposed Greedy Action Consolidation, which penalizes continual action gaps that fall below a prescribed fraction of their historical values. Our analysis bounds historical return discrepancy in terms of the GAC regularizer and the distribution of historical action gaps, while showing that GAC admits a strictly larger pointwise set of action values than exact value cloning. Across both semi-continual and continual RL settings, our experiments demonstrate the effectiveness of GAC, which can be integrated into a broad class of action value based agents. Together, these results support a complementary decision-centric perspective on continual value-based control. The current formulation of GAC nevertheless remains specific to Q-learning agents and finite discrete action spaces. Extending GAC to continuous control is therefore a natural direction for future work, potentially through sampling-based approximation of Q-maximization as in QMLE (Tavakoli et al., 2025), together with action value growth as a continuous-action analogue of the discrete action gap (Wang et al., 2026).

AI USE STATEMENT

In this work, we used generative AI tools to assist with LaTeX formatting, polishing and improving the clarity of the manuscript, writing plotting code for scientific figures, and checking mathematical derivations for correctness and consistency. All AI-assisted content was reviewed by the authors, who take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide the implementation details, experimental configurations, and complete algorithmic hyperparameter settings in Appendix B. These details specify the settings used to produce the experimental results reported in this work.

REFERENCES

Zaheer Abbas, Rosie Zhao, Joseph Modayil, Adam White, and Marlos C Machado. Loss of plasticity in continual deep reinforcement learning. In *Conference on lifelong learning agents*, pp. 620–636. PMLR, 2023.

David Abel, Andre Barreto, Benjamin Van Roy, Doina Precup, Hado van Hasselt, and Satinder Singh. A definition of continual reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=ZZS9WEWYbD.

Armaan A Abraham, Lucy Xiaoyang Shi, and Chelsea Finn. Long-horizon q-learning: Accurate value learning via n-step inequalities. *arXiv preprint arXiv:2605.05812*, 2026.

Hongjoon Ahn, Jinu Hyeon, Youngmin Oh, Bosun Hwang, and Taesup Moon. Prevalence of negative transfer in continual reinforcement learning: Analyses and a simple baseline. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=KAIqwkB3dT.

Abdulaziz Alyahya, Abdallah Al Siyabi, Markus R. Ernst, Luke Yang, Levin Kuhlmann, and Gideon Kowadlo. ARROW: Augmented replay for RObust world models. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=3FK2tFwNwK. Reproducibility Certification, J2C Certification.

Nishanth Anand and Doina Precup. Prediction and control in continual reinforcement learning. *Advances in neural information processing systems*, 36:63779–63817, 2023.

Kavosh Asadi, Yao Liu, Shoham Sabach, Ming Yin, and Rasool Fakoor. Learning the target network in function space. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=jP1zeEqHli.

Marc G. Bellemare, Georg Ostrovski, Arthur Guez, Philip S. Thomas, and Rémi Munos. Increasing the action gap: New operators for reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016. doi: 10.1609/aaai.v30i1.10303.

Massimo Caccia, Jonas Mueller, Taesup Kim, Laurent Charlin, and Rasool Fakoor. Task-agnostic continual reinforcement learning: Gaining insights and overcoming challenges. In Sarath Chandar, Razvan Pascanu, Hanie Sedghi, and Doina Precup (eds.), *Proceedings of The 2nd Conference on Lifelong Learning Agents*, volume 232 of *Proceedings of Machine Learning Research*, pp. 89–119. PMLR, 22–25 Aug 2023. URL https://proceedings.mlr.press/v232/caccia23a.html.

Gail A. Carpenter and Stephen Grossberg. A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer Vision, Graphics, and Image Processing*, 37(1): 54–115, 1987. doi: 10.1016/S0734-189X(87)80014-2.

Wesley Chung, Lynn Cherif, David Meger, and Doina Precup. Parseval regularization for continual reinforcement learning. *Advances in Neural Information Processing Systems*, 37:127937–127967, 2024.

Quentin Delfosse, Patrick Schramowski, Martin Mundt, Alejandro Molina, and Kristian Kersting. Adaptive rational activations to boost deep reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=g90ysX1sVs.

Shibhansh Dohare, J Fernando Hernandez-Garcia, Qingfeng Lan, Parash Rahman, A Rupam Mahmood, and Richard S Sutton. Loss of plasticity in deep continual learning. *Nature*, 632(8026): 768–774, 2024.

Esraa Elelimy, David Szepesvari, Martha White, and Michael Bowling. Rethinking the foundations for continual reinforcement learning. In *Reinforcement Learning Conference*, 2025. URL https://openreview.net/forum?id=cNRVG7y63A.

Benjamin Ellis, Matthew Thomas Jackson, Andrei Lupu, Alexander D. Goldie, Mattie Fellows, Shimon Whiteson, and Jakob Nicolaus Foerster. Adam on local time: Addressing nonstationarity in RL with relative adam timesteps. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=biAqUbAuG7.

Mohamed Elsayed and A Rupam Mahmood. Addressing loss of plasticity and catastrophic forgetting in continual learning. In *International Conference on Learning Representations*, volume 2024, pp. 33399–33430, 2024.

Mohamed Elsayed, Qingfeng Lan, Clare Lyle, and A. Rupam Mahmood. Weight clipping for deep continual and reinforcement learning. In *Reinforcement Learning Conference*, 2024. URL https://openreview.net/forum?id=9Vagp8vnPa.

Amir-Massoud Farahmand. Action-gap phenomenon in reinforcement learning. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 24, pp. 172–180. Curran Associates, Inc., 2011. URL https://papers.nips.cc/paper/4485-action-gap-phenomenon-in-reinforcement-learning.

Alhussein Fawzi, Matej Balog, Aja Huang, Thomas Hubert, Bernardino Romera-Paredes, Mohammad Barekatain, Alexander Novikov, Francisco J. R. Ruiz, Julian Schrittwieser, Grzegorz Swirszcz, David Silver, and Demis Hassabis. Discovering faster matrix multiplication algorithms with reinforcement learning. *Nature*, 610(7930):47–53, 2022.

Timo Flesch, Andrew Saxe, and Christopher Summerfield. Continual task learning in natural and artificial agents. *Trends in Neurosciences*, 46(3):199–210, 2023. ISSN 0166-2236. doi: https://doi.org/10.1016/j.tins.2022.12.006. URL https://www.sciencedirect.com/science/article/pii/S0166223622002600.

Robert M French. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135, 1999.

Haotian Fu, Yixiang Sun, Michael Littman, and George Konidaris. Knowledge retention in continual model-based reinforcement learning. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=DiqeZY27XK.

Jean-Baptiste Gaya, Thang Doan, Lucas Caccia, Laure Soulier, Ludovic Denoyer, and Roberta Raileanu. Building a subspace of policies for scalable continual learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=UKr0MwZM6fL.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025. doi: 10.1038/s41586-025-09422-z. URL https://doi.org/10.1038/s41586-025-09422-z.

Isaac Han, Sangyeon Park, Seungwon Oh, Donghu Kim, Hojoon Lee, and KyungJoong Kim. FIRE: Frobenius-isometry reinitialization for balancing the stability–plasticity tradeoff. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=CfZLxT3zIZ.

Thomas Hubert, Rishi Mehta, Laurent Sartran, Miklós Z. Horváth, Goran Žužić, Eric Wieser, Aja Huang, Julian Schrittwieser, Yannick Schroecker, et al. Olympiad-level formal mathematical reasoning with reinforcement learning. *Nature*, 651(8106):607–613, 2025.

Khurram Javed and Martha White. Meta-learning representations for continual learning. *Advances in neural information processing systems*, 32, 2019.

Christos Kaplanis, Murray Shanahan, and Claudia Clopath. Continual reinforcement learning with complex synapses. In *International Conference on Machine Learning*, pp. 2497–2506. PMLR, 2018.

Christos Kaplanis, Murray Shanahan, and Claudia Clopath. Policy consolidation for continual reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3242–3251. PMLR, 2019. URL https://proceedings.mlr.press/v97/kaplanis19a.html.

Tyler Kastner, Nimrod De La Vega, and Amir-massoud Farahmand. Singularclip: Preventing spectral collapse to maintain plasticity in continual and reinforcement learning. *arXiv preprint arXiv:2608.18319*, 2026.

Elia Kaufmann, Leonard Bauersfeld, Antonio Loquercio, Matthias Müller, Vladlen Koltun, and Davide Scaramuzza. Champion-level drone racing using deep reinforcement learning. *Nature*, 620(7976):982–987, 2023. doi: 10.1038/s41586-023-06419-4.

Isaac Kauvar, Chris Doyle, Linqi Zhou, and Nick Haber. Curious replay for model-based adaptation. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 16018–16048. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/kauvar23a.html.

Samuel Kessler, Mateusz Ostaszewski, Michał PawełBortkiewicz, Mateusz Żarski, Maciej Wolczyk, Jack Parker-Holder, Stephen J. Roberts, and Piotr Miś. The effectiveness of world models for continual reinforcement learning. In Sarath Chandar, Razvan Pascanu, Hanie Sedghi, and Doina Precup (eds.), *Proceedings of The 2nd Conference on Lifelong Learning Agents*, volume 232 of *Proceedings of Machine Learning Research*, pp. 184–204. PMLR, 22–25 Aug 2023. URL https://proceedings.mlr.press/v232/kessler23a.html.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

Dhireesha Kudithipudi, Mario Aguilar-Simon, Jonathan Babb, Maxim Bazhenov, Douglas Blackiston, Josh Bongard, Andrew P. Brna, Suraj Chakravarthi Raja, Nick Cheney, Jeff Clune, Sandeep Madireddy, and Angel Yanguas-Gil. Biological underpinnings for lifelong learning machines. *Nature Machine Intelligence*, 4(3):196–210, 2022. doi: 10.1038/s42256-022-00452-0.

Qingfeng Lan, Yangchen Pan, Jun Luo, and A. Rupam Mahmood. Memory-efficient reinforcement learning with value-based knowledge consolidation. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=zSDCvlaVBn.

Qingfeng Lan, Gautham Vasan, and A Rupam Mahmood. Efficient reinforcement learning by reducing forgetting with elephant activation functions. *arXiv preprint arXiv:2509.19159*, 2025.

Shane Legg and Marcus Hutter. Universal intelligence: A definition of machine intelligence. *Minds and Machines*, 17(4):391–444, 2007. doi: 10.1007/s11023-007-9079-x.

Kun Lei, Huanyu Li, Dongjie Yu, Zhenyu Wei, Lingxiao Guo, Zhennan Jiang, Ziyu Wang, Shiyu Liang, and Huazhe Xu. Rl-100: Performant robotic manipulation with real-world reinforcement learning. *Science Robotics*, 11(116), 2026. doi: 10.1126/scirobotics.aed6267. URL https://www.science.org/doi/10.1126/scirobotics.aed6267.

Alex Lewandowski, Michał Bortkiewicz, Saurabh Kumar, Andras Gyorgy, Dale Schuurmans, Mateusz Ostaszewski, and Marlos C Machado. Learning continually by spectral regularization. In *International Conference on Learning Representations*, volume 2025, pp. 34504–34530, 2025a.

Alex Lewandowski, Dale Schuurmans, and Marlos C. Machado. Plastic learning with deep fourier features. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=NIkfix2eDQ.

Lute Lillo and Nick Cheney. Activation function design sustains plasticity in continual learning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=XZf6wObHX4.

Vincent Liu, Han Wang, Ruo Yu Tao, Khurram Javed, Adam White, and Martha White. Measuring and mitigating interference in reinforcement learning. In *Conference on Lifelong Learning Agents*, pp. 781–795. PMLR, 2023.

Pranshu Malviya, Balaraman Ravindran, and Sarath Chandar. Tag: Task-based accumulated gradients for lifelong learning. In *Conference on Lifelong Learning Agents*, pp. 366–389. PMLR, 2022.

Daniel J. Mankowitz, Andrea Michi, Anton Zhernov, et al. Faster sorting algorithms discovered using deep reinforcement learning. *Nature*, 618:257–263, 2023. doi: 10.1038/s41586-023-06004-9.

Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pp. 109–165. Elsevier, 1989.

Martial Mermillod, Aurélia Bugaiska, and Patrick Bonin. The stability-plasticity dilemma: Investigating the continuum from catastrophic forgetting to age-limited learning effects. *Frontiers in Psychology*, 4:504, 2013. doi: 10.3389/fpsyg.2013.00504.

Golnaz Mesbahi, Parham Mohammad Panahi, Olya Mastikhina, Steven Tang, Martha White, and Adam White. Position: Lifetime tuning is incompatible with continual reinforcement learning. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025. URL https://openreview.net/forum?id=JMoWFkwnvv.

Edan Jacob Meyer, Adam White, and Marlos C. Machado. Harnessing discrete representations for continual reinforcement learning. In *Reinforcement Learning Conference*, 2024. URL https://openreview.net/forum?id=tCXURNlAZ3.

Azalia Mirhoseini, Anna Goldie, Mustafa Yazgan, Joe Wenjie Jiang, Ebrahim Songhori, Shen Wang, Young-Joon Lee, Eric Johnson, Omkar Pathak, Azade Nova, Jiwoo Pak, Andy Tong, Kavya Srinivasa, William Hang, Emre Tuncer, Quoc V. Le, James Laudon, Richard Ho, Roger Carpenter, and Jeff Dean. A graph placement methodology for fast chip design. *Nature*, 594(7862): 207–212, 2021. doi: 10.1038/s41586-021-03544-w. URL https://doi.org/10.1038/s41586-021-03544-w.

Aneesh Muppidi, Zhiyu Zhang, and Heng Yang. Fast trac: A parameter-free optimizer for lifelong reinforcement learning. *Advances in Neural Information Processing Systems*, 37:51169–51195, 2024.

Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018.

Johan S Obando Ceron, Walter Mayor, Samuel Lavoie, Scott Fujimoto, Aaron Courville, and Pablo Samuel Castro. Simplicial embeddings improve sample efficiency in actor–critic agents. In *International Conference on Learning Representations*, volume 2026, pp. 118212–118256, 2026.

Jongchan Park. Scalable reinforcement learning via adaptive batch scaling. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=YI9wSqfz7W.

Sangyeon Park, Isaac Han, Seungwon Oh, and KyungJoong Kim. Activation by interval-wise dropout: A simple way to prevent neural networks from plasticity loss. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=Y0hjl4L1ve.

Ali Saheb Pasand, Johan Obando-Ceron, Aaron Courville, Pouya Bashivan, and Pablo Samuel Castro. Stable deep reinforcement learning via isotropic gaussian representations. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=gc7Gg18ejz.

Matthew Riemer, Ignacio Cases, Robert Ajemian, Miao Liu, Irina Rish, Yuhai Tu, and Gerald Tesauro. Learning to learn without forgetting by maximizing transfer and minimizing interference. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=B1gTShAct7.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. *Advances in neural information processing systems*, 32, 2019.

Andries Rosseau, Robert Müller, and Ann Nowe. Preserving plasticity in continual learning via dynamical isometry. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=vJCOWSkMuq.

Tom Schaul, Andre Barreto, John Quan, and Georg Ostrovski. The phenomenon of policy churn. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=qTCiw1frE_l.

Jonathan Schwarz, Wojciech Czarnecki, Jelena Luketina, Agnieszka Grabska-Barwinska, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. Progress & compress: A scalable framework for continual learning. In *International conference on machine learning*, pp. 4528–4537. PMLR, 2018.

Jaemin Seo, SangKyeun Kim, Azarakhsh Jalalvand, et al. Avoiding fusion plasma tearing instability with deep reinforcement learning. *Nature*, 626:746–751, 2024. doi: 10.1038/s41586-024-07024-9.

Arsalan Sharifnassab, Mohamed Elsayed, Kris De Asis, A. Rupam Mahmood, and Richard S Sutton. Intentional updates for streaming reinforcement learning. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=K9lUSMMI0h.

Ke Sun, Hongming Zhang, Jun Jin, Chao Gao, Xi Chen, Wulong Liu, and Linglong Kong. Principled fast and meta knowledge learners for continual reinforcement learning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=loNTDX3wTn.

Hongyao Tang, Johan Obando-Ceron, Pablo Samuel Castro, Aaron Courville, and Glen Berseth. Mitigating plasticity loss in continual reinforcement learning by reducing churn. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=EkoFXfSauv.

Steven Tang, Xinze Xiong, Anna Hakhverdyan, Andrew Patterson, Jacob Adkins, Jiamin He, Esraa Elelimy, Parham Mohammad Panahi, Martha White, and Adam White. Forager: a lightweight testbed for continual learning with partial observability in rl. *Reinforcement Learning Journal*, 7, 2026.

Arash Tavakoli, Sina Ghiassian, and Nemanja Rakicevic. Learning in complex action spaces without policy gradients. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=nOL9M6D4oM.

Nino Vieillard, Olivier Pietquin, and Matthieu Geist. Munchausen reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. URL https://arxiv.org/abs/2007.14430.

Shengbo Wang, Jose Blanchet, and Peter Glynn. Fast convergence of policy regret in learning stochastic optimal control. *arXiv preprint arXiv:2605.26361*, 2026.

Zihao Wu, Hongyao Tang, Yi Ma, Jiashun Liu, YAN ZHENG, and Jianye HAO. The rank and gradient lost in non-stationarity: Sample weight decay for mitigating plasticity loss in reinforcement learning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=5DpzzTPnJZ.

Hoda Yamani, Yuning Xing, Koen van Rijnsoever, Bruce A. MacDonald, and Henry Williams. Repetition as reinforcement: Enhancing sample efficiency via instant episode repetition in reinforcement learning. *Reinforcement Learning Journal*, 7, 2026.

Hongming Zhang, Chenjun Xiao, Chao Gao, Han Wang, bo xu, and Martin Müller. Exploiting the replay memory before exploring the environment: Enhancing reinforcement learning through empirical MDP iteration. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=lsd27JUJ8v.

Huaicheng Zhou, Zifeng Zhuang, and Donglin Wang. Stay hungry, keep learning: Sustainable plasticity for deep reinforcement learning. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=hTrSxX3kiV.

A Additional Related Work

Loss of Plasticity. Continual reinforcement learning depends as much on whether an agent can still learn as on what it can still remember, making loss of plasticity a complementary concern to catastrophic forgetting. Efforts to sustain trainability and stabilize learning span regularization (Chung et al., 2024; Lewandowski et al., 2025a; Elsayed et al., 2024; Kastner et al., 2026), optimizer design (Ellis et al., 2024; Elsayed & Mahmood, 2024; Muppidi et al., 2024; Rosseau et al., 2026), and activation design (Abbas et al., 2023; Delfosse et al., 2024; Lewandowski et al., 2025b; Park et al., 2025; Lillo & Cheney, 2026). Other approaches shape representational geometry (Obando Ceron et al., 2026; Pasand et al., 2026), rejuvenate neural networks (Dohare et al., 2024; Zhou et al., 2025; Han et al., 2026), or regulate update dynamics (Tang et al., 2025; Wu et al., 2026; Park, 2026).

Others. The role of action gaps in value-based control is well established. Farahmand (2011) relates their distribution to greedy policy performance under value approximation error, while methods that increase action gaps seek to make greedy decisions less susceptible to estimation error (Bellemare et al., 2016; Vieillard et al., 2020). When action gaps are small, modest value perturbations can reverse greedy decisions, contributing to policy churn, the rapid switching of greedy actions across learning updates (Schaul et al., 2022). Although such switching can provide a form of implicit exploration within a task, it constitutes behavioral forgetting in our setting when subsequent learning displaces a historical greedy action. Zhang et al. (2024) regularize action values toward estimates obtained through in-sample bootstrapping on empirical MDPs and report reduced policy churn within individual tasks, connecting their framework to our emphasis on stabilizing greedy decisions across tasks. A related form of behavioral replay appears in Yamani et al. (2026), who store the action sequence whenever an episode achieves a new highest return and re-execute it under new initial conditions to reinforce successful behavior. While their use of past actions as a reference for subsequent learning resonates with our approach, their repetition shapes exploration, whereas ours stores state-specific action gaps to promote greedy action retention across tasks. More closely related in loss construction, Long-Horizon Q-learning (Abraham et al., 2026) uses hinge penalties derived from multistep optimality inequalities to enforce trajectory-wise value consistency and curb compounding estimation error.

B Experimental Details

B.1 MinAtar Configuration

MinAtar Environment. We construct the semi-continual MinAtar sequences using a shifted cyclic scheme:

- `["Breakout", "Space Invaders", "Freeway"]`,
- `["Freeway", "Breakout", "Space Invaders"]`,
- `["Space Invaders", "Freeway", "Breakout"]`.

Whereas prior protocols switch environments every 0.5 million steps and terminate after seven phases (Anand & Precup, 2023; Sun et al., 2026), we allocate 1 million steps to each phase and traverse every ordering for three complete cycles, producing a longer horizon of nine phases and 9 million interactions that exposes each agent to three visits per environment to better evaluate catastrophic forgetting.

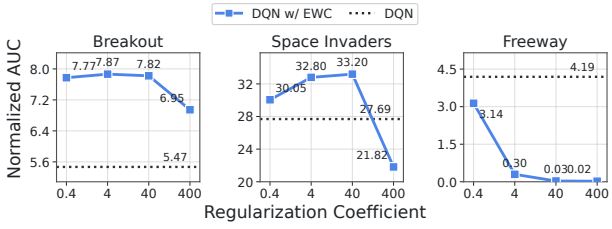


Figure 8: Regularization coefficient tuning results for DQN with EWC.

Table 2: Hyperparameter settings for the semi-continual MinAtar experiments

Hyperparameter	Value
Exploration Strategy	ϵ-greedy, $\epsilon = 0.1$
Minibatch Size	64
Replay Buffer Size	100,000
Buffer Sampling Strategy	Uniform
Target Network Update Frequency	1,000
Discount Factor γ	0.99
Adam β_1	0.9
Adam β_2	0.999
Adam ϵ	10^{-8}
Evaluation ϵ	0.001

Experiment Settings. We provide the experimental configurations used in Section 5.1. Following Anand & Precup (2023), we use a learning rate of 10^{-5} for DQN and all DQN-based methods. PT-DQN trains its permanent value function with SGD at a learning rate of 10^{-8} and its transient value function with Adam at a learning rate of 10^{-4}. DQN with GAC and Value Cloning both use a regularization coefficient of 1, an auxiliary memory size of 10,000 and minibatches of 64 samples to compute their losses, while GAC additionally uses a retention coefficient of 0.5. For PT-DQN with GAC, we compute both the historical and continual action gaps from the sum of the permanent and transient value functions, while allowing gradients from the GAC regularizer to update only the transient component. The transient component thus learns to counter changes that would displace historical greedy actions from the combined value function, and the resulting knowledge is subsequently integrated into the permanent component. PT-DQN with GAC inherits its PT-DQN and GAC hyperparameters from standalone PT-DQN and DQN with GAC, respectively. Based on the tuning results shown in Figure 8, DQN with EWC uses a regularization coefficient of 0.4 and estimates the Fisher information from batches of 100 samples. Following the configuration of Sun et al. (2026), DQN with FAME maintains a meta replay buffer with a capacity of 100,000, to which $N = 12{,}000$ samples are added from each task. Policy evaluation for the one versus all hypothesis test is conducted over $n = 600$ steps, while the warmup period is fixed at $L = 50{,}000$ steps. During knowledge integration, the meta learner is optimized for 200 epochs from an initial learning rate of 10^{-3} under a decaying schedule, whereas the fast learner uses a fixed learning rate of 10^{-5}. At each environment transition, the fast replay buffer is cleared and the parameters of the involved learners are reinitialized in accordance with the original protocol. We use $\tau = 1$ in Q-value based MLE for FAME throughout all tasks. All methods use a CNN architecture consisting of a ReLU-activated convolutional layer with 3×3 filters and stride 1, a ReLU-activated fully connected layer, and a linear action-value output layer. DQN-based methods use 32 filters and 256 hidden units, whereas each of the permanent and transient value functions in PT-DQN uses 16 filters and 128 hidden units. The remaining shared hyperparameters are reported in Table 2.

B.2 Forager Configuration

Forager Environment. We provide details of the environment used for the experiments in Subsection 5.2. The agent inhabits a 24×15 toroidal world, moves one cell using one of four cardinal actions, collects colored mushrooms whose rewards depend on the latent regime, and avoids impassable walls while observing only a local field of view centered on itself, so consequential structure remains beyond its immediate sensory horizon. Its two disjoint 4×15 biomes each contain 15 purple mushrooms, 15 yellow mushrooms, and four walls, while another 24 walls are distributed across the world. Each mushroom is regenerated at a uniformly sampled free location in its original biome after collection or expiry at 500 steps, with regeneration occurring after 9 to 11 steps.

The two reward regimes, each corresponding to a distinct task phase, assign mushroom rewards as specified in Table 3. The regime reverses every 250,000 interactions and completes one square wave cycle every 500,000 steps, while its identity, transitions, and periodicity remain hidden from the agent. Each observation concatenates a flattened 9×9 binary view with separate channels for purple mushrooms, yellow mushrooms, and walls, a four-dimensional one-hot encoding of the previous action, and the previous scalar reward, yielding $9 \times 9 \times 3 + 4 + 1 = 248$ dimensions. Shared colors

Table 3: Mushroom rewards configuration in the Forager environment.

Reward Regime	Western Biome		Eastern Biome	
	Purple	Yellow	Purple	Yellow
Task 1	$+4$	-2	-8	-14
Task 2	-14	-8	-2	$+4$

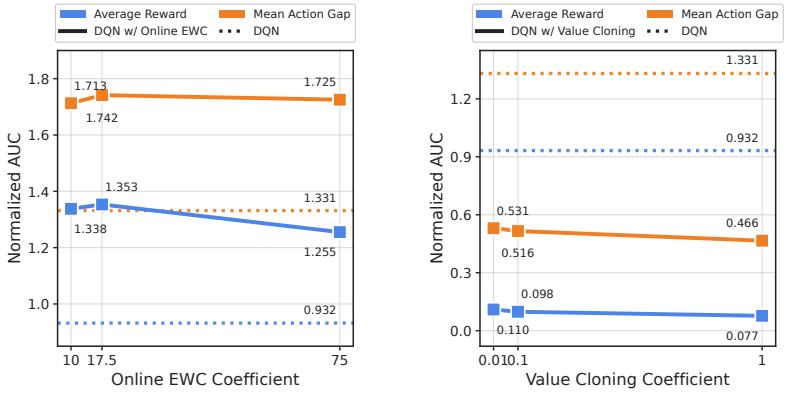


Figure 10: Sensitivity of Online EWC and Value Cloning to the regularization coefficient.

and visually unmarked reward reversals allow identical local evidence to prescribe incompatible behavior under different task phases. Each agent interacts continually for 10 million steps without episodic termination, while task identity, phase, clock, and boundary indicators are excluded from both policy evaluation and optimization.

Experiment Settings. We detail the configurations used in the main and memory size scaling experiments, with DQN and PT-DQN settings primarily guided by the candidates reported in the never-ending relearning experiments of Tang et al. (2026). We allocate the same hyperparameter tuning budget to GAC, Online EWC, and Value Cloning and evaluate three candidate values per method, with the results for Online EWC and Value Cloning reported in Figure 10. All DQN-based methods use a learning rate of 3×10^{-3} except PT-DQN, which trains its permanent value function using SGD at 10^{-8} and its transient value function using Adam at 10^{-3}, with a permanent replay buffer size of 10,000. For the main experiments, GAC and Value Cloning both use an auxiliary memory size of 100, a memory update period of 50,000 steps, and minibatches of 32 samples to compute their losses, with regularization coefficients of 10 and 0.01, respectively. GAC additionally uses a retention coefficient of $\rho = 0.5$. We configure PT-DQN with GAC following the same procedure as in the MinAtar experiments described in Appendix B.1, with its GAC regularization coefficient set to 1 rather than the 10 used for DQN with GAC on Forager. DQN with Online EWC uses a regularization coefficient of 17.5 and a consolidation period of 50,000 steps, and a Fisher information decay factor of 0.95, with Fisher information estimated using batches of 100 samples. For the memory size scaling experiments, we fix the regularization coefficient at 1 for both GAC and Value Cloning and evaluate memory sizes of 100 and 10,000, retaining all remaining settings from the main experiments. All methods use an MLP that flattens the observation and processes it through two fully connected layers, each followed by LayerNorm and a ReLU activation, before a linear action-value output layer. DQN-based methods use 64 hidden units per layer, whereas the permanent and transient value functions of PT-DQN each use 32 hidden units per layer. Other shared hyperparameters for all the methods are listed in Table 4.

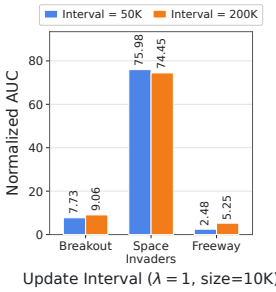


Figure 9: Effect of the auxiliary memory update interval on DQN with Online GAC.

Table 4: Hyperparameter settings for the Forager experiments.

Hyperparameter	**Value**
Exploration Strategy	ϵ-greedy, ϵ=0.1
Update Frequency	4
Minibatch Size	32
Replay Buffer Size	1,000
Minimum Replay History	50
Buffer Sampling Strategy	Uniform
Target Network Update Frequency	128
Discount Factor γ	0.99
Optimizer	Adam
Adam β_1	0.9
Adam β_2	0.999
Adam ϵ	10^{-5}

C MinAtar Experiments in the Online Setting

We provide a preliminary evaluation of GAC and Value Cloning on MinAtar without observable task boundaries in this section. We replace the auxiliary memory updates triggered by task switches used in Section 5.1 with periodic updates every 50,000 steps for both GAC and Value Cloning while leaving the remaining protocol unchanged.

Results. Figure 11 reports results on the sequence comprising Breakout, Space Invaders, and Freeway, showing that GAC retains an overall advantage over Value Cloning, although both methods exhibit lower lifetime performance than their counterparts with observed task boundaries. The degradation is also accompanied by greater behavioral forgetting in Online GAC than in its counterpart with observable task boundaries. This contrast suggests that the timing of auxiliary memory updates affects not only the distribution of retained states but also the reliability of their associated consolidation targets. Theorem 1 relates the GAC regularizer to historical return when $\mu_j = d_j^{\pi_j}$. Updating auxiliary memory after a complete task phase is therefore more likely to record states from a comparatively mature stage of learning, which can better approximate the occupancy measure of the historical policy while yielding more stable greedy actions, action gaps, and value estimates. Fixed periodic updates can instead populate the auxiliary memory with states and targets generated by intermediate policies and may consequently consolidate provisional action and value estimates. The degradation observed in the online setting therefore plausibly reflects the combined effects of occupancy mismatch and target unreliability.

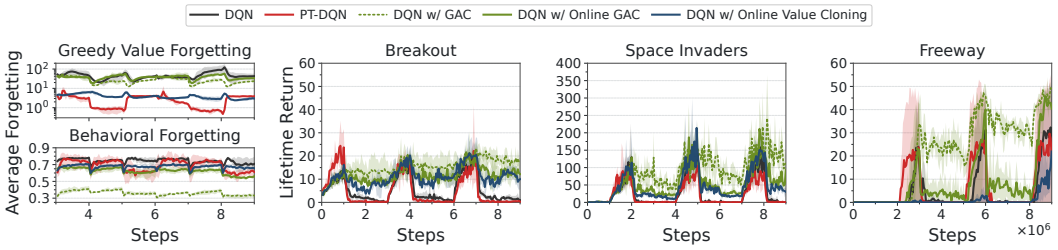


Figure 11: Forgetting and lifetime performance in the task-boundary free continual MinAtar experiment, where the agent cycles through Breakout, Space Invaders, and Freeway in one-million-step phases. The results are averaged across three seeds.

Auxiliary Memory Update Interval. To assess how the timing of auxiliary memory updates affects Online GAC when task boundaries are unobserved, we vary the update interval while fixing the regularization coefficient to 1 and the auxiliary memory size to 10,000. As shown in Figure 9, increasing the interval from 50,000 to 200,000 steps produces only a modest decrease on Space Invaders, where the normalized AUC of lifetime return changes from 75.98 to 74.45, while improving Breakout from 7.73 to 9.06 and Freeway from 2.48 to 5.25, corresponding to relative gains of approximately 17% and 112%. The overall pattern is therefore consistent with our earlier interpretation that auxiliary

memory is more effective when its contents are derived from a sufficiently mature stage of learning, since a longer interval gives the Q function more opportunity to stabilize before its greedy actions, action gaps, and value estimates are retained for subsequent consolidation. The small degradation on Space Invaders further suggests that increasing the update interval does not improve performance uniformly across environments, although the overall pattern suggests that the maturity of the stored targets in auxiliary memory can meaningfully influence GAC. Moreover, when the update interval matches the task duration and the two schedules are aligned, the periodic update rule reproduces the update timing used when task boundaries are observed in the semi-continual RL setting.

The Effect of Regularization Coefficient. We further examine the sensitivity of Online GAC and Online Value Cloning to the regularization strength by varying the coefficient over $\{0.01, 0.05, 0.1, 1\}$ while retaining the same experimental protocol. Figure 12 shows that Online GAC retains an overall advantage over Online Value Cloning throughout the tested range, indicating that its empirical benefit is preserved across substantially different regularization strengths.

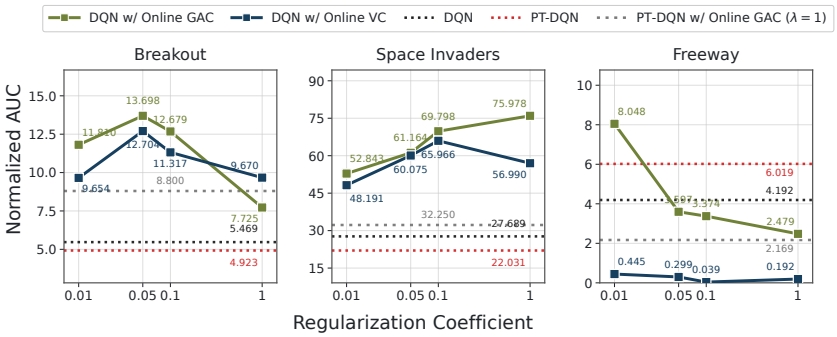


Figure 12: Performance under different regularization coefficients.

D Theoretical Analysis

We theoretically explore the properties of GAC and provide proofs of them in this section.

D.1 Upper Bound on Behavioral Forgetting and Return Discrepancy

Lemma D.1.1 (Upper Bound on Behavioral Forgetting). *Let $\rho > 0$, and assume a single deterministic tie-breaking rule throughout learning. Then,*

$$F_{j\to k}(\mu_j) \le \inf_{\delta>0} \left\{ \Pr_{s\sim\mu_j}\big(M_j(s) \le \delta\big) + \frac{\mathcal{R}_{j\to k}(\theta_k)}{\rho\delta} \right\}. \tag{5}$$

Proof. For brevity, define the pointwise GAC loss

$$\overline{M}_{j\to k}(s;\rho) := \big[\rho M_j(s) - M_{j\to k}(s)\big]_+,$$

then the GAC regularizer can be written as

$$\mathcal{R}_{j\to k}(\theta_k) = \mathbb{E}_{s\sim\mu_j}\Big[\overline{M}_{j\to k}(s;\rho)\Big].$$

Let

$$D_{j\to k}(s) := \mathbb{I}\left\{\hat{a}_k^\star(s) \ne \hat{a}_j^\star(s)\right\} \in \{0,1\},$$

if $D_{j\to k}(s) = 1$, then the historical greedy action $\hat{a}_j^\star(s)$ of task j cannot strictly dominate all of its current competing actions, which implies $M_{j\to k}(s) \le 0$ and hence

$$\overline{M}_{j\to k}(s;\rho) \ge \rho M_j(s).$$

If $D_{j\to k}(s) = 0$, the product $M_j(s)D_{j\to k}(s)$ vanishes. The two cases therefore give the pointwise inequality

$$\overline{M}_{j\to k}(s;\rho) \geq \rho M_j(s)D_{j\to k}(s).$$

Taking expectations under $s \sim \mu_j$ yields

$$\mathbb{E}_{s\sim\mu_j}[M_j(s)D_{j\to k}(s)] \leq \frac{\mathcal{R}_{j\to k}(\theta_k)}{\rho}. \tag{6}$$

Fix any historical action gap threshold $\delta > 0$ and partition the state representation space into

$$\mathcal{S}^\delta := \{s \in \mathcal{S} : M_j(s) \leq \delta\}, \qquad (\mathcal{S}^\delta)^c := \mathcal{S} \setminus \mathcal{S}^\delta = \{s \in \mathcal{S} : M_j(s) > \delta\}.$$

Writing $\Pr_{\mu_j}$ for probability under $s \sim \mu_j$, behavioral forgetting decomposes as

$$\begin{aligned} F_{j\to k}(\mu_j) &= \Pr_{\mu_j}(D_{j\to k} = 1,\ M_j \leq \delta) + \Pr_{\mu_j}(D_{j\to k} = 1,\ M_j > \delta) \\ &\leq \Pr_{\mu_j}(M_j \leq \delta) + \Pr_{\mu_j}(D_{j\to k} = 1,\ M_j > \delta). \end{aligned} \tag{7}$$

Viewing $M_j(s)D_{j\to k}(s)$ as a nonnegative random variable and applying *Markov's inequality* to the second term in Equation 7 gives

$$\begin{aligned} \Pr_{\mu_j}(D_{j\to k} = 1,\ M_j > \delta) &= \Pr_{\mu_j}\big(M_j(s)D_{j\to k}(s) > \delta\big) \\ &\leq \frac{\mathbb{E}_{s\sim\mu_j}\big[M_j(s)D_{j\to k}(s)\big]}{\delta} \\ &\leq \frac{\mathcal{R}_{j\to k}(\theta_k)}{\rho\delta}. \end{aligned} \tag{8}$$

The last inequality above makes use of Equation 6. Substituting Equation 8 into Equation 7 yields

$$F_{j\to k}(\mu_j) \leq \Pr_{s\sim\mu_j}(M_j(s) \leq \delta) + \frac{\mathcal{R}_{j\to k}(\theta_k)}{\rho\delta}.$$

Since this inequality holds for every $\delta > 0$, taking the infimum over δ proves the result. □

We restate Theorem 1 for completeness.

Theorem D.1.1 (Upper Bound on Return Discrepancy). *Let $\rho > 0$, and fix a single deterministic tie-breaking rule to be used throughout the learning process. Suppose that the GAC regularizer is defined with $\mu_j = d_j^{\pi_j}$. Then*

$$\begin{aligned} |J_j(\pi_k) - J_j(\pi_j)| &\leq \frac{2R^\star}{(1-\gamma)^2} F_{j\to k}\left(d_j^{\pi_j}\right) \\ &\leq \frac{2R^\star}{(1-\gamma)^2} \inf_{\delta>0} \left\{ \Pr_{s\sim d_j^{\pi_j}}(M_j(s) \leq \delta) + \frac{\mathcal{R}_{j\to k}(\theta_k)}{\rho\delta} \right\}. \end{aligned} \tag{9}$$

Proof. For every bounded measurable function f,

$$\left(\mathcal{P}_j^\pi f\right)(s) = \int_{\mathcal{S}} f(s')\mathcal{P}_j\left(ds' \mid s, \pi(s)\right),$$

and let its Bellman operator be

$$\left(\mathcal{T}_j^\pi V\right)(s) = r_j\left(s, \pi(s)\right) + \gamma\left(\mathcal{P}_j^\pi V\right)(s).$$

Write

$$V^j := V_j^{\pi_j}, \qquad V^k := V_j^{\pi_k},$$

where both value functions are evaluated on the historical task j, and define

$$\Delta_{j,k} = V^k - \mathcal{T}_j^{\pi_j} V^k.$$

Since $V^k = \mathcal{T}_j^{\pi_k} V^k$ and $V^j = \mathcal{T}_j^{\pi_j} V^j$, we obtain

$$\begin{aligned} V^k - V^j &= V^k - \mathcal{T}_j^{\pi_j} V^k + T_j^{\pi_j} V^k - V^j \\ &= \Delta_{j,k} + \mathcal{T}_j^{\pi_j} V^k - \mathcal{T}_j^{\pi_j} V^j \\ &= \Delta_{j,k} + \gamma \mathcal{P}_j^{\pi_j} \left(V^k - V^j \right). \end{aligned} \tag{10}$$

Because $\gamma < 1$, the inverse of $I - \gamma \mathcal{P}_j^{\pi_j}$ exists on bounded measurable functions and admits its Neumann expansion, which gives

$$V^k - V^j = (I - \gamma \mathcal{P}_j^{\pi_j})^{-1} \Delta_{j,k} \tag{11}$$

$$= \sum_{t=0}^{\infty} \gamma^t \left(\mathcal{P}_j^{\pi_j} \right)^t \Delta_{j,k}. \tag{12}$$

Averaging Equation 12 under ν_j and invoking discounted occupancy yields

$$\begin{aligned} J_j(\pi_k) - J_j(\pi_j) &= \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{\nu_j, \mathcal{P}_j, \pi_j} \left[\Delta_{j,k}(s_t) \right] \\ &= \frac{1}{1-\gamma} \mathbb{E}_{s \sim d_j^{\pi_j}} \left[\Delta_{j,k}(s) \right]. \end{aligned} \tag{13}$$

Since $V^k = \mathcal{T}_j^{\pi_k} V^k$, the residual also satisfies

$$\Delta_{j,k} = \mathcal{T}_j^{\pi_k} V^k - \mathcal{T}_j^{\pi_j} V^k.$$

Its pointwise form is

$$\begin{aligned} \Delta_{j,k}(s) = r_j \left(s, \pi_k(s) \right) + \gamma \mathbb{E}_{s' \sim \mathcal{P}_j(\cdot | s, \pi_k(s))} \left[V^k(s') \right] \\ - r_j \left(s, \pi_j(s) \right) - \gamma \mathbb{E}_{s' \sim \mathcal{P}_j(\cdot | s, \pi_j(s))} \left[V^k(s') \right]. \end{aligned} \tag{14}$$

The true action value function of π_k on task j is

$$Q_j^{\pi_k}(s,a) = r_j(s,a) + \gamma \mathbb{E}_{s' \sim \mathcal{P}_j(\cdot | s,a)} \left[V_j^{\pi_k}(s') \right]. \tag{15}$$

Consequently, Equation 14 becomes

$$\Delta_{j,k}(s) = Q_j^{\pi_k} \left(s, \pi_k(s) \right) - Q_j^{\pi_k} \left(s, \pi_j(s) \right). \tag{16}$$

In general, $Q_j^{\pi_k}$ need not coincide with the learned function Q_{θ_k}. The former is the true action value function of π_k on task j, whereas the latter is the learned function that induces π_k, so the proof requires no assumption that Q_{θ_k} accurately estimates $Q_j^{\pi_k}$.

The bounded reward assumption gives

$$\left\| Q_j^{\pi_k} \right\|_\infty \leq \frac{R^\star}{1-\gamma}.$$

Moreover, Equation 16 vanishes whenever $\pi_k(s) = \pi_j(s)$, and therefore

$$|\Delta_{j,k}(s)| = |Q_j^{\pi_k} \left(s, \pi_k(s) \right) - Q_j^{\pi_k} \left(s, \pi_j(s) \right)| \mathbb{I} \left\{ \pi_k(s) \neq \pi_j(s) \right\} \tag{17}$$

$$\leq \left(|Q_j^{\pi_k} \left(s, \pi_k(s) \right)| + |Q_j^{\pi_k} \left(s, \pi_j(s) \right)| \right) \mathbb{I} \left\{ \pi_k(s) \neq \pi_j(s) \right\} \tag{18}$$

$$\leq \frac{2R^\star}{1-\gamma} \mathbb{I} \left\{ \pi_k(s) \neq \pi_j(s) \right\}. \tag{19}$$

Combining Equations 13 and 19 gives

$$\begin{aligned} |J_j(\pi_k) - J_j(\pi_j)| &\leq \frac{1}{1-\gamma} \mathbb{E}_{s \sim d_j^{\pi_j}} \left[|\Delta_{j,k}(s)| \right] \\ &\leq \frac{2R^\star}{(1-\gamma)^2} \mathbb{E}_{s \sim d_j^{\pi_j}} \left[\mathbb{I} \left\{ \pi_k(s) \neq \pi_j(s) \right\} \right] \\ &= \frac{2R^\star}{(1-\gamma)^2} F_{j \to k} \left(d_j^{\pi_j} \right). \end{aligned} \tag{20}$$

Since $\mu_j = d_j^{\pi_j}$, Lemma D.1.1 further gives

$$F_{j\to k}\left(d_j^{\pi_j}\right) \le \inf_{\delta>0}\left\{ \Pr_{s\sim d_j^{\pi_j}}\left(M_j(s)\le\delta\right) + \frac{\mathcal{R}_{j\to k}(\theta_k)}{\rho\delta}\right\}.$$

Substitution into Equation 20 proves the second inequality. □

If historical action gaps are uniformly bounded away from zero on the reference states, Theorem D.1.1 yields a return discrepancy bound proportional to the GAC regularizer. The following corollary makes this dependence explicit.

Corollary D.1.2. *Under the assumptions of Theorem D.1.1, suppose that the historical action gaps admit a uniform positive lower bound over the reference state distribution, in the sense that there exists $\kappa > 0$ such that $M_j(s) \ge \kappa$ for $d_j^{\pi_j}$-almost every state s. Then*

$$|J_j(\pi_k) - J_j(\pi_j)| \le \frac{2R^\star}{(1-\gamma)^2}\frac{\mathcal{R}_{j\to k}(\theta_k)}{\rho\kappa}. \tag{21}$$

Proof. For every $0 < \delta < \kappa$, the uniform gap assumption gives

$$\Pr_{s\sim d_j^{\pi_j}}\left(M_j(s)\le\delta\right) = 0.$$

Theorem D.1.1 therefore implies

$$|J_j(\pi_k) - J_j(\pi_j)| \le \frac{2R^\star}{(1-\gamma)^2}\frac{\mathcal{R}_{j\to k}(\theta_k)}{\rho\delta} \qquad \text{for every } 0<\delta<\kappa.$$

Because δ can be chosen arbitrarily close to κ from below, the stated bound follows with κ in the denominator. □

D.2 Action Value Capacity

Lemma D.1.1 and Theorem D.1.1 establish that GAC protects historical decisions and, under occupancy matching, bounds the corresponding return discrepancy. Having examined stability against catastrophic forgetting, we now address the complementary requirement of plasticity by asking how much capacity remains after retention is imposed. Value Cloning (Rolnick et al., 2019) preserves historical action values themselves, whereas GAC preserves only the action-gap needed to retain the historical greedy decision. In our theory, GAC strictly enlarges the pointwise admissible action value set, while its induced feasible parameter set contains that of value cloning, ensuring that the infimum of any subsequent learning objective cannot exceed its counterpart under value cloning.

Throughout this section, assume a finite action space with $|\mathcal{A}| \ge 2$ and a retention coefficient $0 < \rho \le 1$. For a fixed reference state s, write

$$\boldsymbol{Q}_{\theta_j}(s) = \left[Q_{\theta_j}(s,a)\right]_{a\in\mathcal{A}} \in \mathbb{R}^{|\mathcal{A}|}.$$

For a candidate action value vector $\boldsymbol{Q} = \left[Q(s,a)\right]_{a\in\mathcal{A}}$, define the exact value cloning admissible set

$$\mathcal{Q}_j(s) = \left\{ \boldsymbol{Q}\in\mathbb{R}^{|\mathcal{A}|} \,\middle|\, \boldsymbol{Q} = \boldsymbol{Q}_{\theta_j}(s)\right\}, \tag{22}$$

and the GAC admissible set

$$\mathcal{Q}_j(s;\rho) = \left\{ \boldsymbol{Q}\in\mathbb{R}^{|\mathcal{A}|} \,\middle|\, Q\big(s,\hat{a}_j^\star(s)\big) - \max_{a\in\mathcal{A}\setminus\{\hat{a}_j^\star(s)\}} Q(s,a) \ge \rho M_j(s)\right\}. \tag{23}$$

Theorem D.2.1. *For every reference state s and every $0 < \rho \le 1$,*

$$\boxed{\mathcal{Q}_j(s) \subsetneq \mathcal{Q}_j(s;\rho)}. \tag{24}$$

Moreover, $\mathcal{Q}_j(s;\rho)$ is a closed convex full dimensional unbounded polyhedron in $\mathbb{R}^{|\mathcal{A}|}$, whereas $\mathcal{Q}_j(s)$ contains only the historical action value vector $\boldsymbol{Q}_{\theta_j}(s)$.

Proof. Fix a reference state s and write $y = \hat{a}_j^\star(s)$. By definition of the historical action gap,

$$M_j(s) = Q_{\theta_j}(s, y) - \max_{a \in \mathcal{A} \setminus \{y\}} Q_{\theta_j}(s, a).$$

Hence, for every $a \neq y$,

$$Q_{\theta_j}(s, y) - Q_{\theta_j}(s, a) \geq M_j(s) \geq \rho M_j(s), \tag{25}$$

where the final inequality follows from $0 < \rho \leq 1$ and $M_j(s) \geq 0$. Therefore

$$\boldsymbol{Q}_{\theta_j}(s) \in \mathcal{Q}_j(s; \rho).$$

Since $\mathcal{Q}_j(s) = \{\boldsymbol{Q}_{\theta_j}(s)\}$, we obtain

$$\mathcal{Q}_j(s) \subseteq \mathcal{Q}_j(s; \rho).$$

It remains to establish strictness. Let $C \neq 0$ be a constant and define the shifted vector $\widetilde{\boldsymbol{Q}}$ by

$$\widetilde{Q}(s, a) = Q_{\theta_j}(s, a) + C \qquad \text{for every } a \in \mathcal{A}.$$

For every competitor $a \neq y$,

$$\begin{aligned} \widetilde{Q}(s, y) - \widetilde{Q}(s, a) &= Q_{\theta_j}(s, y) + C - Q_{\theta_j}(s, a) - C \\ &= Q_{\theta_j}(s, y) - Q_{\theta_j}(s, a) \\ &\geq M_j(s) \geq \rho M_j(s). \end{aligned} \tag{26}$$

Thus $\widetilde{\boldsymbol{Q}} \in \mathcal{Q}_j(s; \rho)$. Since $C \neq 0$, $\widetilde{\boldsymbol{Q}} \neq \boldsymbol{Q}_{\theta_j}(s)$ and hence $\widetilde{\boldsymbol{Q}} \notin \mathcal{Q}_j(s)$, which proves

$$\mathcal{Q}_j(s) \subsetneq \mathcal{Q}_j(s; \rho).$$

The geometric properties follow from the defining inequalities. Writing $\boldsymbol{Q} = \big[Q(s, a)\big]_{a \in \mathcal{A}}$, we have

$$\mathcal{Q}_j(s; \rho) = \bigcap_{a \in \mathcal{A} \setminus \{y\}} \left\{ \boldsymbol{Q} \in \mathbb{R}^{|\mathcal{A}|} \,\middle|\, Q(s, y) - Q(s, a) \geq \rho M_j(s) \right\},$$

which is the intersection of finitely many closed half spaces and is therefore a closed convex polyhedron. Its interior is nonempty because any vector satisfying

$$Q(s, y) - \max_{a \neq y} Q(s, a) > \rho M_j(s)$$

belongs to its interior, so it is full dimensional in $\mathbb{R}^{|\mathcal{A}|}$. Finally, if $\boldsymbol{Q} \in \mathcal{Q}_j(s; \rho)$, then $\boldsymbol{Q} + C\mathbf{1} \in \mathcal{Q}_j(s; \rho)$ for every $C \in \mathbb{R}$, since a common shift cancels from every action difference. The set is therefore unbounded, which completes the proof. □

Theorem D.2.1 characterizes the additional flexibility afforded by GAC, as the full dimensionality of $\mathcal{Q}_j(s; \rho)$ prevents the retention constraint from confining the action value vector to a lower dimensional manifold and instead renders every strictly feasible configuration an interior point, leaving an admissible neighborhood of perturbations through which the action value function can accommodate subsequent learning without relinquishing the historical greedy decision.

To accommodate the residual error arising from imperfect interpolation of historical action value targets, for $\varepsilon \geq 0$ we define the approximate value cloning admissible set

$$\mathcal{Q}_{j,\varepsilon}(s) = \left\{ \boldsymbol{Q} \in \mathbb{R}^{|\mathcal{A}|} \,\middle|\, \max_{a \in \mathcal{A}} \big| Q(s, a) - Q_{\theta_j}(s, a) \big| \leq \varepsilon \right\}. \tag{27}$$

Theorem D.2.2. *For every reference state s, if*

$$\varepsilon \leq \frac{(1 - \rho) M_j(s)}{2}, \tag{28}$$

then

$$\boxed{\mathcal{Q}_{j,\varepsilon}(s) \subseteq \mathcal{Q}_j(s; \rho)}. \tag{29}$$

Proof. Fix s, let $\boldsymbol{Q} \in \mathcal{Q}_{j,\varepsilon}(s)$, and write $y = \hat{a}_j^\star(s)$. By definition of $\mathcal{Q}_{j,\varepsilon}(s)$,

$$Q(s,y) \geq Q_{\theta_j}(s,y) - \varepsilon,$$

while, for every $a \neq y$,

$$Q(s,a) \leq Q_{\theta_j}(s,a) + \varepsilon.$$

Subtracting the two inequalities gives

$$\begin{aligned} Q(s,y) - Q(s,a) &\geq Q_{\theta_j}(s,y) - Q_{\theta_j}(s,a) - 2\varepsilon \\ &\geq M_j(s) - 2\varepsilon. \end{aligned} \tag{30}$$

Under Equation 28,

$$M_j(s) - 2\varepsilon \geq \rho M_j(s).$$

Hence

$$Q(s,y) - Q(s,a) \geq \rho M_j(s) \qquad \text{for every } a \neq y,$$

which proves

$$\mathcal{Q}_{j,\varepsilon}(s) \subseteq \mathcal{Q}_j(s;\rho).$$

□

Theorem D.2.2 shows that sufficiently accurate value cloning preserves the GAC margin, while the converse need not hold.

We next lift this distinction to the parameterized function class. Let Θ denote the parameter space, write

$$\boldsymbol{Q}_\theta(s) = \big[Q_\theta(s,a)\big]_{a\in\mathcal{A}},$$

and define the exact value cloning feasible set

$$\Theta_j = \{\theta \in \Theta \,|\, \boldsymbol{Q}_\theta(s) \in \mathcal{Q}_j(s) \text{ for } \mu_j \text{ almost every } s\}, \tag{31}$$

together with the GAC feasible set

$$\Theta_j(\rho) = \{\theta \in \Theta \,|\, \boldsymbol{Q}_\theta(s) \in \mathcal{Q}_j(s;\rho) \text{ for } \mu_j \text{ almost every } s\}. \tag{32}$$

Let $\mathcal{L}_k(\theta)$ be any extended real valued objective for learning on task k.

Theorem D.2.3. *For every* $0 < \rho \leq 1$*,*

$$\Theta_j \subseteq \Theta_j(\rho), \tag{33}$$

and consequently

$$\boxed{\inf_{\theta\in\Theta_j(\rho)} \mathcal{L}_k(\theta) \leq \inf_{\theta\in\Theta_j} \mathcal{L}_k(\theta)}. \tag{34}$$

If there exists $\widetilde{\theta} \in \Theta_j(\rho) \setminus \Theta_j$ *such that*

$$\mathcal{L}_k(\widetilde{\theta}) < \inf_{\theta\in\Theta_j} \mathcal{L}_k(\theta), \tag{35}$$

then the inequality in Equation 34 is strict.

Proof. Take any $\theta \in \Theta_j$. By Equation 31,

$$\boldsymbol{Q}_\theta(s) = \boldsymbol{Q}_{\theta_j}(s)$$

for μ_j almost every s. For such states,

$$\begin{aligned} &Q_\theta\big(s, \hat{a}_j^\star(s)\big) - \max_{a\in\mathcal{A}\setminus\{\hat{a}_j^\star(s)\}} Q_\theta(s,a) \\ &\quad = Q_{\theta_j}\big(s, \hat{a}_j^\star(s)\big) - \max_{a\in\mathcal{A}\setminus\{\hat{a}_j^\star(s)\}} Q_{\theta_j}(s,a) \\ &\quad = M_j(s) \\ &\quad \geq \rho M_j(s), \end{aligned} \tag{36}$$

where the final inequality follows from $0 < \rho \leq 1$ and $M_j(s) \geq 0$. Thus

$$\boldsymbol{Q}_\theta(s) \in \mathcal{Q}_j(s; \rho)$$

for μ_j almost every s, which proves

$$\Theta_j \subseteq \Theta_j(\rho).$$

Since the infimum of any extended real valued function over a superset cannot exceed its infimum over a subset,

$$\inf_{\theta \in \Theta_j(\rho)} \mathcal{L}_k(\theta) \leq \inf_{\theta \in \Theta_j} \mathcal{L}_k(\theta).$$

This proves Equation 34 without requiring convexity, differentiability, realizability of a subsequent task optimum, or any assumption on the optimization algorithm.

Finally, if $\widetilde{\theta} \in \Theta_j(\rho) \setminus \Theta_j$ satisfies Equation 35, then

$$\inf_{\theta \in \Theta_j(\rho)} \mathcal{L}_k(\theta) \leq \mathcal{L}_k(\widetilde{\theta}) < \inf_{\theta \in \Theta_j} \mathcal{L}_k(\theta).$$

The infimum over the GAC feasible set is therefore strictly smaller under the stated condition, which completes the proof. □

Theorem D.2.3 formalizes plasticity as the capacity of parameter space to minimize subsequent tasks' objective without assumptions on the optimizer, exploration process, or particular form of the subsequent objective. This statement concerns the best objective value permitted by the retention constraint and does not imply that a particular stochastic optimization trajectory must outperform value cloning.