

Adaptive Reward Shaping Helps Only Without Entropy Auto-Tuning: A Boundary Condition for Robotic Reaching

Mugahed Abdulla^{1,2}, Maher Mesto², Francisco Cruz^{2,3}

Abstract—Adaptive Potential Functions (APF) improve convergence when combined with deterministic policy gradient methods such as DDPG, but it is unknown whether these benefits transfer to entropy-regularised algorithms that already manage their own exploration. We evaluate APF integrated with Soft Actor-Critic (SAC) on simulated Kinova Gen3 Lite reaching across 20 random seeds and 3,000 episodes per seed. Under standard SAC with entropy auto-tuning, adaptive shaping was statistically indistinguishable from baseline (mean AUC -43.9 ± 14.2 vs -46.1 ± 18.5 , $p = 0.68$, Cliff’s $\delta = -0.055$). However, fixing the temperature parameter at its default value ($\alpha = 0.2$) revealed a significant medium-sized benefit ($p = 0.016$, Cliff’s $\delta = 0.450$), indicating that external scaffolding helps when internal regulation is absent. Analysis of SAC’s internal training dynamics shows that temperature tuning does not materially change when APF is present, suggesting partial functional redundancy between the two mechanisms. We interpret this as a boundary condition for reward shaping: adaptive scaffolding is redundant when the learner already possesses intrinsic self-regulation, but becomes beneficial when that mechanism is removed.

Index Terms—reinforcement learning, reward shaping, adaptive potential functions, robotic reaching, developmental learning

I. INTRODUCTION

Reinforcement learning (RL) has emerged as a promising approach for learning robotic manipulation policies [1], [2]. A persistent challenge is sample efficiency: training requires many environment interactions, making simulation-based training with subsequent sim-to-real transfer the dominant paradigm [3], [4]. Reward shaping can improve sample efficiency by augmenting environment rewards with additional feedback, and Adaptive Potential Functions (APF) [5] have demonstrated improved convergence when combined with DDPG [1] for robotic reaching.

However, Soft Actor-Critic (SAC) [2] takes a fundamentally different approach to exploration: it maximises an entropy-regularised objective and automatically tunes a temperature parameter α to balance exploration and exploitation. This raises a question that, to our knowledge, has not been investigated: *Does adaptive reward shaping remain beneficial when the learner already possesses an intrinsic mechanism for managing exploration?*

We hypothesise that APF partially overlaps in role with SAC’s entropy auto-tuning. Although the two mechanisms work differently – APF shapes rewards based on trajectory statistics, while SAC penalises overly deterministic policies

through entropy regularisation – both have the downstream effect of encouraging broader state-space coverage during early training. If this partial overlap exists, then APF should be redundant when auto-tuning is active but beneficial when auto-tuning is removed. We test this prediction directly by fixing α at its initial value and comparing performance with and without APF. Beyond reward performance, we analyse SAC’s internal training dynamics (temperature α , policy entropy, and Q-values) to examine whether APF interacts with or operates independently of SAC’s entropy mechanism.

Developmental learning theory provides a lens for this question. In Vygotsky’s zone of proximal development [6], scaffolding supports learners who cannot yet succeed independently. Wood et al. [7] observed that effective scaffolding is contingent on the learner’s current capability. By structural analogy, a deterministic policy learner (DDPG) lacks internal exploration guidance and benefits from external scaffolding, while an entropy-regularised learner (SAC) may already occupy its own zone of proximal development through self-regulated exploration.

The contributions of this paper are:

- 1) Integration of APF with SAC for robotic reaching, extending the original APF-DDPG framework to an entropy-regularised algorithm.
- 2) A 20-seed evaluation showing APF is redundant under auto-tuned SAC ($p = 0.68$, Cliff’s $\delta = -0.055$).
- 3) A fixed- α ablation demonstrating that APF significantly helps when entropy auto-tuning is disabled ($p = 0.016$, Cliff’s $\delta = 0.450$, medium effect), establishing a mechanistic boundary condition for adaptive reward shaping.

II. RELATED WORK

Reward shaping augments an environment’s reward signal with additional feedback to accelerate learning. Ng et al. [8] established that potential-based reward shaping (PBRs) preserves the optimal policy provided the shaping function takes the form $f(s, s') = \gamma\Phi(s') - \Phi(s)$ for a fixed potential Φ . Early work from Randalv and Alstrm [9] demonstrated practical benefits of this approach for learning to ride a bicycle. However, these methods relied on hand-designed potential functions, limiting their applicability to tasks where domain knowledge is readily available.

Chen et al. [5] addressed this limitation by introducing Adaptive Potential Functions (APF), where the potential is learned online from trajectory statistics rather than specified a priori. Combined with the Deep Deterministic Policy Gradient (DDPG) algorithm [1], APF improved convergence

¹Robotic Systems Lab, ETH Zurich, Switzerland

²School of Computer Science and Engineering, UNSW Sydney, Australia

³Escuela de Ingenieria, Universidad Central de Chile, Chile

speed and stability for continuous robotic reaching tasks. Crucially, DDPG is a deterministic policy algorithm that relies entirely on additive Ornstein-Uhlenbeck noise for exploration, with no mechanism for adjusting exploration intensity based on learning progress. The APF framework has not been evaluated with algorithms that possess built-in exploration management.

Soft Actor-Critic (SAC) [2] introduced a fundamentally different approach to exploration by maximising an entropy-regularised objective and automatically tuning a temperature parameter α to maintain a target policy entropy. This closed-loop mechanism enables SAC to self-regulate its exploration throughout training, distinguishing it from deterministic methods. Fujimoto et al. [10] improved actor-critic stability through twin Q-networks and delayed updates (TD3), but retained external noise for exploration. SAC’s entropy auto-tuning raises a question not addressed in prior work: whether external reward shaping remains beneficial when the learner already possesses an intrinsic mechanism for managing exploration. This paper investigates that question directly.

III. METHODOLOGY

A. Problem Formulation

The reaching task is formulated as an MDP $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma)$. A state $s \in \mathcal{S}$ is a 15-dimensional observation vector: end-effector position (3), goal position (3), relative position vector (3), and joint angles (6). An action $a \in \mathcal{A}$ consists of 6 continuous joint position increments, each in $[-\pi/64, \pi/64]$ radians. Episodes terminate upon success, object displacement, or after 100 timesteps.

B. Soft Actor-Critic

We employ SAC [2], which maximises the entropy-regularised objective:

$$J(\pi) = \sum_{t=0}^T \mathbb{E} [r(s_t, a_t) + \alpha \mathcal{H}(\pi(\cdot|s_t))] \quad (1)$$

where α is the temperature parameter and $\mathcal{H}$ denotes entropy. SAC uses twin Q-networks to mitigate overestimation [10] and automatic temperature tuning with target entropy $\bar{\mathcal{H}} = -\dim(\mathcal{A})$. The temperature is adjusted by minimising:

$$J(\alpha) = \mathbb{E}_{a_t \sim \pi_t} [-\alpha (\log \pi_t(a_t|s_t) + \bar{\mathcal{H}})] \quad (2)$$

This mechanism enables SAC to self-regulate its exploration: when policy entropy falls below target, α increases to encourage more stochastic behaviour. When entropy exceeds the target, α decreases to allow more exploitation. This closed-loop regulation is the key distinction from DDPG, which has no mechanism for adapting exploration intensity based on learning progress.

C. Adaptive Potential Function

Following Chen et al. [5], a state mapping $Z : \mathcal{S} \rightarrow \mathcal{S}_p$ projects the state to a 6-dimensional discrete representation of end-effector and goal positions, each coordinate discretised to two decimal places. The APF network $\Phi : \mathcal{S}_p \rightarrow \mathbb{R}$

is a two-layer feedforward network (512–256 units) learned online. After each episode, visited trajectories are stored in a priority buffer of size 2,000. The top 50% of trajectories by episodic return are labelled good (T_g), the rest bad (T_b). The APF is updated by minimising:

$$\mathcal{L}_\phi = \mathbb{E} \left[\left(\Phi(s_p; \phi) - \frac{N_g[s_p] - N_b[s_p]}{N_g[s_p] + N_b[s_p]} \right)^2 \right] \quad (3)$$

where $N_g[s_p]$ and $N_b[s_p]$ count occurrences in good and bad trajectories. Shaping is activated after more than 100 potential network updates to allow the buffer to accumulate sufficient trajectory data. The shaping reward follows the PBRs form: $f(s, s') = \gamma \Phi(Z(s')) - \Phi(Z(s))$, and the shaped reward $r + f$ is stored in the replay buffer. While classical PBRs guarantees policy invariance for a fixed potential, the online adaptation means this guarantee holds only approximately at each update step [5].

D. Fixed-Alpha Ablation

To test whether APF’s redundancy is caused by functional overlap with entropy auto-tuning, we conduct an ablation where the temperature parameter is fixed at its initial value ($\alpha = 0.2$, the standard default from Haarnoja et al. [2]) and gradient updates to $\log \alpha$ are disabled. This keeps the entropy weight constant throughout training, removing SAC’s ability to self-regulate exploration. If our hypothesis is correct, APF should provide a measurable benefit in this setting.

IV. EXPERIMENTAL SETUP

A. Simulation Environment

All experiments use MuJoCo [3] with a Kinova Gen3 Lite 6-DOF arm [11]. The robot model is derived from the official Kinova URDF. Six actuators control the arm joints and four control the gripper, which is kept open throughout. Each control step executes five simulation substeps to ensure stable physics integration. Figs. 1 and 2 show the simulated environment and the physical robot used for a preliminary deployment check (reported in Limitations).

B. Task and Reward

The agent must move the end-effector within 10 cm of a randomly placed target ($x_{obj} \sim \mathcal{U}(0.3, 0.4)$, $y_{obj} \sim \mathcal{U}(0.3, 0.4)$, $z_{obj} = 0.03$ m) within 100 timesteps. The dense reward at each step is $r_t = -d(\mathbf{p}_{ee}, \mathbf{p}_{obj})$, with terminal bonuses of +10 for success and -100 if the object is displaced outside the workspace boundary. For APF-SAC, the shaped reward $r_t + f$ replaces r_t in the replay buffer. All reported performance metrics (learning curves, AUC, final reward) use the raw environment reward r_t to ensure fair comparison. Observations are normalised using Welford’s online algorithm [12], with running statistics frozen after 100,000 steps to prevent distribution shift in late training.

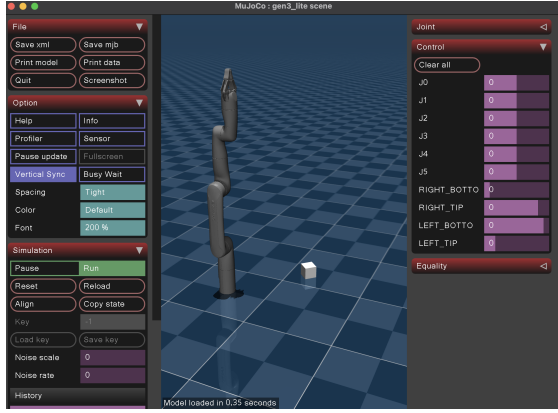


Fig. 1. MuJoCo simulation environment with the Kinova Gen3 Lite arm and target object.

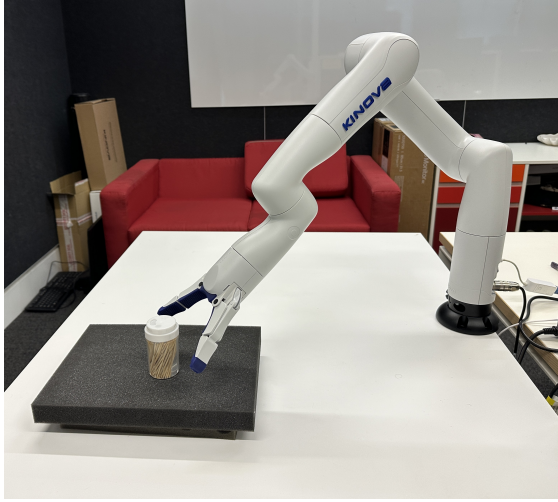


Fig. 2. Physical Kinova Gen3 Lite robot setup used for preliminary real-world deployment.

C. Training Protocol

Each method is trained for 3,000 episodes across 20 independent random seeds. The fixed- α ablation uses identical hyperparameters, seeds, and episode count, yielding four experimental conditions: Auto+SAC and Auto+APF (with entropy auto-tuning), and FixedA+SAC and FixedA+APF (with fixed $\alpha = 0.2$). Table I lists all settings. AUC is defined as the mean raw episodic reward over the full 3,000 episodes for each seed.

To analyse whether APF interacts with SAC’s entropy mechanism, we log four quantities per episode: the temperature $\alpha = \exp(\log \alpha)$, policy entropy $\mathcal{H} = -\mathbb{E}[\log \pi(a|s)]$, and the mean outputs of both Q-networks ($\bar{Q}_1$, $\bar{Q}_2$), each computed over 256 replay buffer samples.

V. RESULTS

A. Standard SAC Shows No Shaping Benefit

Under entropy auto-tuning, APF-SAC and SAC are statistically indistinguishable. Fig. 3 shows overlapping training curves across 20 seeds. Table II summarises the comparison using bootstrap confidence intervals and effect sizes

TABLE I
TRAINING HYPERPARAMETERS.

Parameter	Value
Training episodes	3,000
Max steps per episode	100
Random seeds	20
Actor learning rate	1×10^{-4}
Critic learning rate	2×10^{-3}
Discount factor (γ)	0.99
Soft update (τ)	0.005
Replay buffer size	300,000
Batch size	256
Actor/Critic layers	128–128
Entropy (α)	0.2 (auto or fixed)
APF network layers	512–256
APF learning rate	2×10^{-3}
APF batch size	64
Trajectory buffer size	2,000
Good trajectory ratio	0.5

TABLE II
AUTO-TUNED SAC COMPARISON ACROSS 20 SEEDS WITH BOOTSTRAP
CONFIDENCE INTERVALS AND EFFECT SIZES.

Metric	SAC	APF-SAC
Full AUC (mean $\pm$ sd)	-46.1 ± 18.5	-43.9 ± 14.2
Final reward	-44.5	-44.6
95% Bootstrap CI (final)	$[-55.3, -34.9]$	$[-53.4, -36.1]$
Cliff’s δ (final)	-0.055 (negligible)	
Brown-Forsythe p (final)	0.694	

following Agarwal et al. [13]. Full-training AUC differs by 2.2 points ($p = 0.68$), final reward differs by only 0.08 points (-44.6 vs -44.5), and Cliff’s delta is -0.055 (negligible). Brown-Forsythe testing confirms no significant difference in cross-seed variance ($p = 0.69$). The cross-seed standard deviation ratio (APF-SAC / SAC) ranges from 0.77 in acquisition to 0.87 in late training, suggesting a modest trend toward reduced variability, but this is not statistically significant (Brown-Forsythe $p = 0.37$ for AUC variance).

B. Shaping Does Not Alter Converged Entropy

Figs. 5 and 6 show the temperature α and policy entropy traces across training, plotted as median with interquartile range. Two observations emerge.

Temperature trajectories. Both methods begin at $\alpha = 0.2$ and follow a nearly identical initial descent to approximately 0.02 by episode 500 as the policy becomes more deterministic. After episode 500, temperature rises as SAC’s auto-tuning responds to falling entropy. The median trajectories remain comparable throughout, though SAC exhibits a wider interquartile range indicating more variable temperature adaptation across seeds. In extreme cases, individual SAC seeds develop temperature values orders of magnitude above the median, driven by Q-value overestimation in a subset of seeds. Across training, SAC’s mean α is approximately 938 (vs median $\alpha \approx 0.1$), reflecting the influence of a small number of divergent seeds. At episode 2,999, the mean reaches $2,862 \pm 8,844$ (Table III). APF constrains this variability (training-wide mean $\alpha = 0.43$, final-episode

TABLE III
INTERNAL TRAINING DIAGNOSTICS AT KEY EPISODES (MEAN $\pm$ STD
ACROSS 20 SEEDS).

Episode	Metric	SAC	APF-SAC
100	α	0.075 ± 0.002	0.075 ± 0.002
	Entropy	3.60 ± 0.20	3.60 ± 0.25
	$ Q_1 - Q_2 $	0.052 ± 0.047	0.059 ± 0.043
500	α	0.157 ± 0.323	0.028 ± 0.022
	Entropy	-11.0 ± 11.7	-6.4 ± 0.6
	$ Q_1 - Q_2 $	96.8 ± 313.3	0.4 ± 0.4
2999	α	2862 ± 8844	0.83 ± 2.18
	Entropy	-6.0 ± 0.3	-6.1 ± 0.3
	$ Q_1 - Q_2 $	9388 ± 32554	7.6 ± 16.3

mean 0.83 ± 2.18), but the constraint does not translate into improved reward.

Entropy convergence. Both methods converge to entropy values near the target $\bar{\mathcal{H}} = -6$ (equal to $-\dim(\mathcal{A})$). SAC’s entropy reaches -6.0 ± 0.3 at convergence, while APF-SAC reaches -6.1 ± 0.3 , with no significant difference ($p = 0.12$, Welch t -test). The entropy trajectories suggest that SAC’s automatic temperature mechanism drives policy entropy toward its target regardless of whether APF shaping is present. APF neither accelerates convergence to the entropy target nor alters the converged value.

Table III summarises the internal diagnostics at key training episodes. Note that these means are sensitive to outlier seeds; the figures above use median and interquartile range for a more robust view. At episode 100, both methods are comparable across all metrics. By episode 2,999, both converge to similar entropy, but SAC exhibits substantially higher mean α and Q-values due to outlier seeds with Q-value overestimation. In these seeds, overestimated Q-values create a feedback loop: inflated critic targets cause the policy to become overly deterministic, driving entropy below target and causing α to increase. APF is associated with bounded Q-value estimates in these cases, but since policy entropy converges to the same target regardless, this stabilisation has no measurable effect on task performance in this setting.

C. Disabling Auto-Tuning Reveals a Shaping Benefit

The fixed- α ablation changes the conclusion. When entropy auto-tuning is disabled, APF-SAC significantly outperforms SAC. Fig. 4 shows clear separation between the methods under fixed $\alpha = 0.2$. Table IV reports the ablation statistics.

FixedA+APF achieves a final reward of -78.9 ± 15.0 compared to -90.1 ± 10.7 for FixedA+SAC ($p = 0.016$, Cliff’s $\delta = +0.450$, medium effect). The APF advantage is small initially ($+0.8$ at episode 500) but grows steadily throughout training: $+5.0$ at episode 1,000, $+6.6$ at episode 1,500, and $+10.5$ by episode 3,000. This increasing gap indicates that APF shaping provides cumulative benefit when the learner lacks self-regulation, rather than merely accelerating early learning.

Comparing across conditions reveals the dominance of

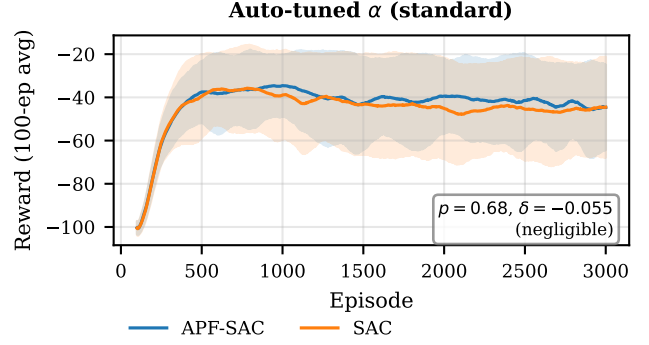


Fig. 3. Training curves under auto-tuned α (mean $\pm$ 1 std across 20 seeds, 100-episode running average). APF-SAC and SAC are statistically indistinguishable.

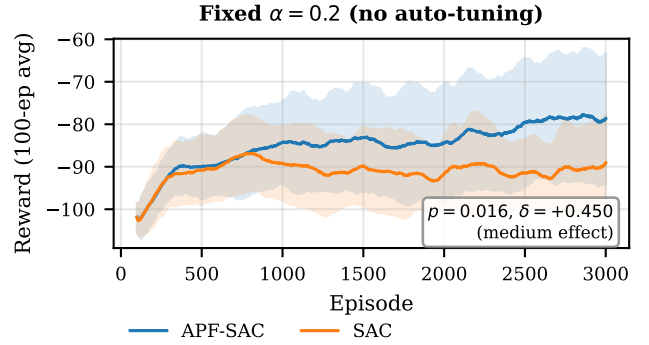


Fig. 4. Training curves under fixed $\alpha = 0.2$ (mean $\pm$ 1 std across 20 seeds, 100-episode running average). APF-SAC significantly outperforms SAC.

auto-tuning. Auto+SAC strongly outperforms FixedA+SAC ($p < 0.001$, Cliff’s $\delta = -0.900$, large effect), demonstrating that entropy auto-tuning is far more valuable than APF shaping. The AUC gap between auto-tuned and fixed conditions is approximately 45 points, compared to the 6-point AUC gap that APF provides under fixed α . Similarly, Auto+APF outperforms FixedA+APF ($p < 0.001$, Cliff’s $\delta = -0.785$, large effect). These results form a consistent pattern: auto-tuning contributes most to performance, and APF provides additional benefit only when auto-tuning is absent.

Note that both fixed- α conditions perform substantially worse than their auto-tuned counterparts, confirming that a constant temperature of $\alpha = 0.2$ provides insufficient exploration regulation across the full training horizon. With auto-tuning, SAC adapts α from 0.2 down to approximately 0.02 during early learning and subsequently adjusts upward as needed. Fixing α prevents this adaptation, leaving the agent with a static exploration-exploitation balance that is suboptimal for most of the training trajectory.

VI. DISCUSSION

A. A Boundary Condition for Reward Shaping

These results suggest a boundary condition for adaptive reward shaping in dense-reward reaching tasks. Chen et

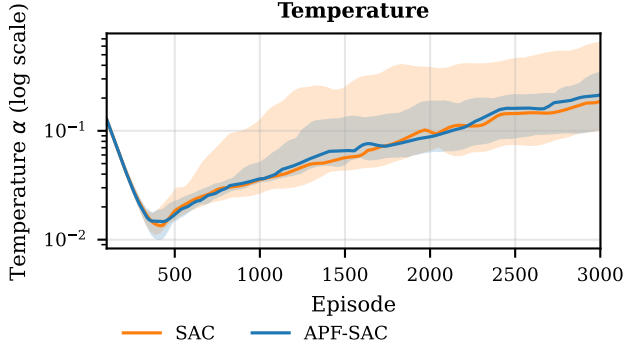


Fig. 5. Temperature α across training (median $\pm$ IQR across 20 seeds, 100-episode smoothing). SAC and APF-SAC follow comparable median temperature trajectories.

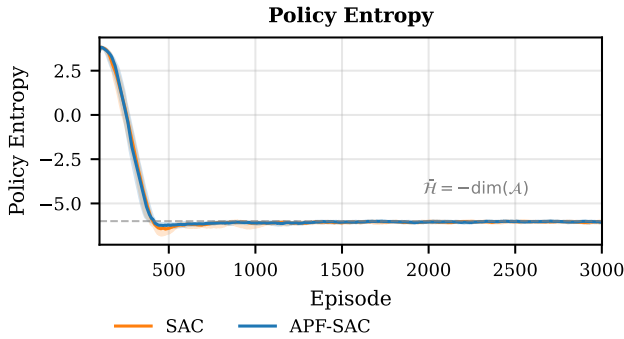


Fig. 6. Policy entropy across training (median $\pm$ IQR across 20 seeds, 100-episode smoothing). Both methods converge to the target entropy $\tilde{H} = -\dim(\mathcal{A})$.

al. [5] demonstrated that APF improves DDPG, a deterministic policy algorithm that relies on additive Ornstein-Uhlenbeck noise for exploration with no mechanism for adjusting exploration intensity based on learning progress. SAC, by contrast, maintains a stochastic policy and adapts its exploration through the learned temperature α . Our training dynamics analysis (Section V-B) shows that APF does not materially alter SAC’s entropy regulation: policy entropy converges to the same target regardless of whether shaping is present, suggesting partial functional redundancy between the two mechanisms at the level of exploration management in this task.

The fixed- α ablation provides direct evidence for this account. By removing SAC’s self-regulation mechanism and demonstrating that APF then becomes effective, we show that the null result under standard SAC is specifically because of partial functional redundancy between APF and entropy auto-tuning, not because APF is inherently ineffective. The results form a 2×2 pattern: $\{\text{auto-tuned, fixed}\} \times \{\text{APF, no-APF}\}$, with a conditional effect where APF helps only without auto-tuning.

This pattern is consistent with broader observations in modern RL, where engineered guidance such as reward shaping, curriculum learning, or auxiliary objectives often

TABLE IV
FIXED-ALPHA ABLATION STATISTICS ACROSS 20 SEEDS.

Comparison	p	Cliff’s δ
FixedA+APF vs FixedA+SAC	0.016	+0.450 (medium)
FixedA+SAC vs Auto+SAC	< 0.001	−0.900 (large)
FixedA+APF vs Auto+APF	< 0.001	−0.785 (large)

shows diminishing returns as the base algorithm’s adaptive capabilities improve. The contribution here is identifying the specific mechanism whose presence renders APF redundant: although entropy auto-tuning and APF operate through different mechanisms (policy stochasticity versus trajectory-based reward shaping), they appear to play partially substitutable roles in exploration management for this task. When the internal mechanism is active, the external one has nothing additional to contribute at the level of task performance.

A secondary observation is that APF substantially reduces critic instability. Table III shows that Q-value discrepancy ($|Q_1 - Q_2|$) at episode 2,999 is three orders of magnitude smaller under APF-SAC (7.6 ± 16.3) than under SAC ($9,388 \pm 32,554$). While this stabilisation does not improve task reward in our dense-reward setting, it may become relevant in longer training horizons, with larger networks, or in tasks where critic divergence poses greater risks to learning stability.

B. Developmental Interpretation

We operationalise self-regulation as entropy auto-tuning and test the prediction that scaffolding should help more when this regulatory mechanism is absent. These results do not show that adaptive shaping is ineffective. They show that its utility depends on whether the learner has an effective mechanism for regulating exploration. In this structural reading, the fixed- α setting resembles a learner without self-regulation, where external support is beneficial. The auto-tuned setting resembles a learner that already regulates its behaviour, where extra scaffolding is redundant. This interpretation is consistent with the zone of proximal development [6] as an analogy, not a causal claim. The key distinction is between a learner that *lacks* internal exploration mechanisms (DDPG, fixed- α SAC) and one that *has* them (auto-tuned SAC). In developmental terms, scaffolding is most effective when calibrated to the gap between a learner’s current and potential capability. When the learner already manages its own exploration, additional external support yields diminishing returns.

C. Limitations

This study evaluates a single dense-reward reaching task with one robot platform. As a preliminary deployment check, separately trained SAC and APF-SAC policies were each evaluated on a physical Kinova Gen3 Lite across 10 fixed goal positions, with both methods achieving 10/10 successes. However, these policies were trained separately from the main simulation study and the limited trial count precludes statistical comparison. The redundancy finding

may not generalise to tasks where SAC’s entropy mechanism is insufficient, such as sparse-reward environments or contact-rich tasks with deceptive local optima. The fixed- α ablation uses a single value ($\alpha = 0.2$), chosen as SAC’s standard default rather than through a sweep of fixed values. Testing APF under multiple fixed temperature values would strengthen the boundary condition claim. Several directions merit investigation: testing APF with SAC on contact-rich tasks such as grasping, where reward landscapes are more complex, could reveal whether the redundancy holds across task difficulty levels. The Q-value stabilisation effect of APF, while not performance-relevant here, may become important in longer training horizons or with larger networks where critic divergence poses greater risks. Finally, the boundary condition we identify generates predictions for other algorithm pairings beyond APF-DDPG and APF-SAC.

VII. CONCLUSION

We integrated Adaptive Potential Functions with Soft Actor-Critic for robotic reaching and found no statistically significant benefit over SAC alone across 20 random seeds. Analysis of SAC’s internal training dynamics showed that entropy regulation is not materially altered by APF’s shaping signal, suggesting partial functional redundancy: APF’s guidance appears substitutable with SAC’s built-in entropy-based exploration in this setting. However, a fixed- α ablation revealed that APF provides a significant medium-sized benefit ($p = 0.016$, Cliff’s $\delta = 0.450$) when entropy auto-tuning is disabled. This establishes a boundary condition for adaptive reward shaping in dense-reward robotic reaching: the benefits observed with deterministic policy methods do not transfer to entropy-regularised algorithms, but the shaping signal retains its value when the learner lacks internal self-regulation. Drawing a structural analogy to developmental learning, these results suggest that external scaffolding is most effective when the learner lacks internal self-regulation, and redundant when the learner already manages its own exploration.

ACKNOWLEDGMENT

The authors thank the Robotic Systems Lab at ETH Zurich and the School of Computer Science and Engineering at

UNSW Sydney for facilities and resources. Maher Mesto is supported by an Australian Government Research Training Program (RTP) Scholarship. This research includes computations using the computational cluster Katana supported by Research Technology Services at UNSW Sydney.

REFERENCES

- [1] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, “Continuous control with deep reinforcement learning,” in *International Conference on Learning Representations (ICLR)*, 2016.
- [2] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, “Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor,” in *International Conference on Machine Learning (ICML)*. PMLR, 2018, pp. 1861–1870.
- [3] E. Todorov, T. Erez, and Y. Tassa, “MuJoCo: A physics engine for model-based control,” in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012, pp. 5026–5033.
- [4] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, “Domain randomization for transferring deep neural networks from simulation to the real world,” in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 23–30.
- [5] Y. Chen, L. Schomaker, and F. Cruz, “Boosting reinforcement learning algorithms in continuous robotic reaching tasks using adaptive potential functions,” 2024, arXiv preprint arXiv:2402.04581.
- [6] L. S. Vygotsky, *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press, 1978.
- [7] D. Wood, J. S. Bruner, and G. Ross, “The role of tutoring in problem solving,” *Journal of Child Psychology and Psychiatry*, vol. 17, no. 2, pp. 89–100, 1976.
- [8] A. Y. Ng, D. Harada, and S. Russell, “Policy invariance under reward transformations: Theory and application to reward shaping,” in *Proceedings of the Sixteenth International Conference on Machine Learning (ICML)*, 1999, pp. 278–287.
- [9] J. Randløv and P. Alstrøm, “Learning to drive a bicycle using reinforcement learning and shaping,” in *Proceedings of the Fifteenth International Conference on Machine Learning (ICML)*, 1998, pp. 463–471.
- [10] S. Fujimoto, H. van Hoof, and D. Meger, “Addressing function approximation error in actor-critic methods,” in *International Conference on Machine Learning (ICML)*. PMLR, 2018, pp. 1587–1596.
- [11] Kinova Inc., “Kinova Gen3 Lite robotic arm,” <https://www.kinovarobotics.com/product/gen3-lite-robots>, 2024, accessed: 2026-03-28.
- [12] B. P. Welford, “Note on a method for calculating corrected sums of squares and products,” *Technometrics*, vol. 4, no. 3, pp. 419–420, 1962.
- [13] R. Agarwal, M. Schwarzer, P. S. Castro, A. C. Courville, and M. G. Bellemare, “Deep reinforcement learning at the edge of the statistical precipice,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.