JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X Volume 4, Issue 2, 2026
https://jecir.com/

ENHANCING STABLE BEHAVIORAL IMITATION THROUGH ADAPTIVE REWARD WEIGHTING IN TD3-SAC-GAIL

Mehran Ali [1], Zia Ullah [2], Aliza Ashfaq [3]

Affiliations

[1] Department of Computer Science, Gomal University, D.I.K, Pakistan
Email: mehranalikhan768@gmail.com

[2] Medical Lab Technology, Sarhad University of Science and Information Technology, Pakistan
Email. ziaullah1823@gmail.com

[3] Mathematical Sciences, Fatima Jinnah Women University, Pakistan
Email: alizaashfaq44@gmail.com

Corresponding Author's Email

[1] mehranalikhan768@gmail.com

ABSTRACT

Imitation learning enables reinforcement learning agents to acquire complex behaviors from expert demonstrations, but its performance remains strongly influenced by the quality of expert data and the balance between imitation and exploration during policy optimization. In particular, TD3-SAC-GAIL enhances the exploration capability of Generative Adversarial Imitation Learning (GAIL) by combining deterministic policy smoothing from Twin Delayed Deep Deterministic Policy Gradient (TD3) with entropy-driven exploration from Soft Actor-Critic (SAC). However, the use of fixed reward or exploration weighting can lead to an inappropriate exploration–exploitation balance across different training stages and environments. To address this limitation, this paper proposes an adaptive reward weighting mechanism to enhance imitation learning stability within the TD3-SAC-GAIL framework. The proposed mechanism dynamically adjusts the contribution of exploration and imitation signals according to the current learning condition, encouraging exploration when learning progress is limited while placing greater emphasis on policy quality as the training process becomes stable. The proposed framework is evaluated in four continuous-control environments, namely Half Cheetah, Walker2d, Hopper, and Lunar Lander Continuous. Experimental evaluation considers expert-surpassing performance, training behavior, reward stability, and the effect of adaptive weighting. The results demonstrate the potential of adaptive reward weighting to provide a systematic mechanism for controlling the exploration–imitation trade-off and enhancing the stability and robustness of GAIL-based policy learning while retaining the exploration advantages of the TD3-SAC hybrid framework.

Keywords: Imitation Learning, Generative Adversarial Imitation Learning (GAIL), TD3-SAC Hybrid Reinforcement Learning, Adaptive Reward Weighting, Stable Policy Optimization

I. INTRODUCTION

IMITATION learning (IL) has emerged as an effective approach for learning complex control behaviors from expert demonstrations, particularly in robotic motion control and sequential decision-making tasks [1]. Unlike conventional reinforcement learning (RL), where an agent generally learns through explicit interaction with an environment and reward feedback, imitation learning enables an agent to acquire a policy by observing and reproducing expert behavior [2]. Generative Adversarial Imitation Learning (GAIL) extends this idea by formulating imitation learning as an adversarial learning problem in which a policy attempts to generate state-action trajectories similar to those of an expert, while a discriminator distinguishes between expert and generated behaviors. This framework provides an effective way of learning complex behaviors without requiring an explicitly designed environmental reward function [3].

Despite its advantages, GAIL has an important limitation related to its dependence on expert demonstrations. When demonstrations are incomplete, noisy, or suboptimal, the learned policy can inherit these limitations and may have difficulty discovering behaviors that are better than those demonstrated by the

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X Volume 4, Issue 2, 2026
https://jecir.com/

expert. In addition, conventional GAIL does not inherently provide a strong active exploration mechanism. Insufficient exploration can restrict the diversity of actions generated by the policy and may cause the learning process to converge toward suboptimal behaviors [3]. This limitation becomes particularly important in dynamic robot motion control environments, where discovering alternative actions and trajectories can be essential for obtaining improved policies [4].

To address the exploration limitation of GAIL, recent approaches have incorporated reinforcement learning algorithms into the imitation learning framework. The TD3-SAC Hybrid Method provides one such approach by combining exploration-promoting mechanisms from Twin Delayed Deep Deterministic Policy Gradient (TD3) and Soft Actor-Critic (SAC). TD3 contributes mechanisms such as target policy smoothing and delayed policy updates, while SAC introduces entropy-based policy optimization to encourage action diversity and exploration [5]. By integrating these mechanisms into GAIL, the TD3-SAC Hybrid Method improves the exploration capability of the imitation learning process and provides the agent with an opportunity to discover policies that can potentially surpass the demonstrated expert behavior [6].

The TD3-SAC-GAIL framework uses the discriminator to distinguish expert state-action pairs from agent-generated state-action pairs. The discriminator output is used as an imitation-related reward signal for updating the reinforcement learning policy [7]. Consequently, the learning process depends not only on the quality of exploration provided by TD3 and SAC, but also on how effectively the resulting reward signal guides policy optimization. An appropriate balance between action quality and exploration is therefore important for achieving effective and stable learning [8].

In the existing TD3-SAC Hybrid Method, the coefficient associated with the entropy component plays an important role in balancing action diversity and action quality. Experimental analysis in the original work shows that the performance changes with different values of this coefficient and that an intermediate value can provide better performance. However, the effective value is not necessarily identical across all environments, and the reported experiments indicate that the relationship between this parameter and the probability of surpassing expert demonstrations varies between Half Cheetah, Lunar Lander Continuous, Walker2d, and Hopper [9]. This indicates that the balance between exploration-related and quality-related learning signals is an important factor in the stability and effectiveness of TD3-SAC-GAIL [10].

However, the existing approach relies on a manually selected or fixed balancing coefficient during training rather than dynamically adapting the weighting according to the current learning condition. As the policy evolves, the quality of generated trajectories, discriminator feedback, and exploration behavior can change substantially. A fixed weighting mechanism may therefore provide an inappropriate balance at different stages of training. In particular, excessive emphasis on exploration may introduce unnecessary variability and reduce action quality, whereas excessive emphasis on action quality may restrict exploration and increase the risk of premature convergence. This motivates the investigation of an adaptive mechanism capable of adjusting the contribution of relevant reward components according to the learning state [11].

In this work, we propose Adaptive Reward Weighting for Stable Imitation Learning in TD3-SAC-GAIL. The proposed approach builds upon the TD3-SAC-GAIL framework and focuses on dynamically adjusting reward weighting during the learning process rather than relying solely on a fixed weighting parameter [7]. The objective is to achieve a more appropriate balance between exploration and policy quality throughout training, thereby improving learning stability and maintaining effective imitation performance [12]. The proposed method retains the original framework's exploration capabilities while introducing an adaptive reward-weighting mechanism to address the limitation associated with fixed reward balancing [13].

To evaluate the proposed approach, the study is designed around the same simulation environments and expert-demonstration setting used by the existing TD3-SAC-GAIL framework. The evaluation considers continuous-control environments including Half Cheetah, Walker2d, Hopper, and Lunar Lander Continuous, allowing the proposed reward-weighting mechanism to be examined under different motion-control characteristics. Performance is considered in terms of learning stability, imitation performance, and the ability to achieve performance beyond the demonstrated expert behavior [14].

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X Volume 4, Issue 2, 2026
https://jecir.com/

The main contribution of this work is therefore the introduction of an adaptive reward-weighting mechanism into TD3-SAC-GAIL. Rather than focusing solely on increasing exploration through the combination of TD3 and SAC, this work investigates how the learning signals can be dynamically balanced to obtain a more stable and effective imitation-learning process. The proposed approach aims to reduce the sensitivity to manually selected weighting parameters and provide a more flexible learning mechanism across different environments [15].

II. METHODOLOGY

THIS study proposes an adaptive reward-weighting mechanism for imitation learning within the TD3-SAC-GAIL framework. The proposed method preserves the principal architecture of the original TD3-SAC Hybrid Method, including Generative Adversarial Imitation Learning (GAIL), the discriminator-based imitation signal, TD3-based critic stabilization, SAC-based entropy regularization, and replay-buffer-based off-policy learning. The primary modification is the introduction of an adaptive mechanism that adjusts the relative contribution of the learning signals during training [16].

The motivation for the proposed mechanism is based on the sensitivity of the TD3-SAC Hybrid Method to its entropy-related coefficient. In the original study, different values of the coefficient were evaluated, and the resulting performance varied across the Half Cheetah, Lunar Lander Continuous, Walker2d, and Hopper environments [17]. The reported results indicate that an intermediate coefficient produced favorable performance in several environments, while excessively small or large values could reduce exploration or cause unstable learning behavior [18].

This observation motivates replacing a purely fixed weighting strategy with a training-dependent adaptive mechanism. Instead of requiring one manually selected value to remain appropriate throughout the complete learning process, the proposed method dynamically adjusts the weighting according to recent learning behavior [19].

A. Framework Overview

The complete framework consists of:

1. Expert demonstrations
2. A GAIL discriminator
3. A TD3-SAC hybrid actor-critic architecture
4. An adaptive weighting controller
5. Adaptive reward construction
6. Policy and value-function optimization

The proposed architecture can therefore be expressed as:

Expert Demonstrations → GAIL Discriminator → Learning Signal → Adaptive Weighting → TD3-SAC Optimization

B. Problem Formulation

The imitation-learning problem is formulated as a Markov decision process:

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \gamma)$$

where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{P}(s', \mid sa)$ represents the environment transition probability, and γ is the discount factor.

The expert demonstration dataset is represented as:

$$\mathcal{D}_e = \{(s_t^e, a_t^e)\}_{t=1}^{N_e}$$

The objective is to learn a policy $\pi_\phi(a \mid s)$ whose generated state-action distribution is sufficiently close to the expert distribution while maintaining sufficient exploration to discover alternative and potentially improved behaviors.

C. GAIL Discriminator

The original GAIL framework formulates imitation learning through an adversarial interaction between a policy and a discriminator. The discriminator distinguishes expert state-action pairs from policy-generated pairs, while the policy attempts to generate behavior that is classified as expert-like. The original

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X Volume 4, Issue 2, 2026
https://jecir.com/

TD3-SAC-GAIL work further incorporates exploration mechanisms from TD3 and SAC to reduce the limitations associated with insufficient exploration.

The proposed method retains the TD3-SAC-GAIL architecture as its baseline.

The GAIL component consists of a policy generator and a discriminator. The discriminator receives state-action pairs and estimates whether they originate from expert demonstrations or the current policy.

D. TD3 Components

The TD3 component provides:

- Twin critic networks
- Minimum-Q target estimation
- Target-policy smoothing
- Delayed actor updates

These mechanisms are intended to reduce value overestimation and improve the stability of policy optimization. The original paper specifically incorporates these TD3 mechanisms into its hybrid framework.

E. SAC Components

The SAC component introduces entropy-based policy optimization. The entropy term encourages action diversity and prevents the policy from focusing exclusively on the currently estimated high-value action. The original paper identifies the coefficient α as the parameter controlling the relationship between action quality and policy entropy.

During interaction with the environment, the current policy generates state-action transitions that are stored in the replay buffer $\mathcal{B}$. Mini-batches sampled from $\mathcal{B}$ are subsequently used for critic and actor optimization.

F. Discriminator Optimization

Let $D_\psi(s,a)$ denote the discriminator parameterized by ψ. Expert state-action pairs are treated as positive samples, whereas policy-generated state-action pairs obtained from the replay buffer are treated as negative samples.

The discriminator loss is:

$$\mathcal{L}_D = -\mathbb{E}_{(s,a)\sim\mathcal{D}_e}\left[\log D_\psi(s,a)\right] - \mathbb{E}_{(s,a)\sim\mathcal{B}}\left[\log\left(1 - D_\psi(s,a)\right)\right]$$

The discriminator is optimized such that $D_\psi(s^E, a^E) \to 1$ for expert samples and $D_\psi(s,a) \to 0$ for policy-generated samples.

The discriminator output provides the imitation-related feedback used during policy learning. This follows the training structure described in the original methodology.

The proposed method does not replace the exploration mechanisms of the original TD3-SAC Hybrid Method.

For TD3, action perturbation is used to increase action diversity and target-policy smoothing is used during value estimation. In addition, the delayed actor update and twin critics contribute to stable value estimation.

For SAC, entropy regularization is incorporated into the policy objective. The corresponding coefficient alpha controls the contribution of entropy relative to the estimated action value. The original work reports that changing alpha changes the exploration behavior and overall performance of the hybrid method.

The proposed contribution is therefore not a replacement for TD3 or SAC. Instead, the proposed adaptive controller operates above this existing mechanism to regulate the learning balance during training.

The original experimental analysis demonstrates that the value of the entropy coefficient affects the probability of surpassing expert demonstrations. In particular, the study evaluates:

$$\alpha = [10^{-4}, 5, \times, 10^{-4}, 10^{-3}5 \times 10^{-3}10^{-2}]$$

and reports that the performance generally increases and then decreases as alpha increases, although the strength of this trend differs between environments.

This observation suggests that the exploration-quality balance is not necessarily identical across environments.

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X Volume 4, Issue 2, 2026
https://jecir.com/

The limitation addressed in this work is therefore formulated as follows: A manually selected fixed exploration-related weighting may not remain optimal throughout the complete training process or across different learning conditions.

The proposed solution is to make the weighting responsive to the current learning state rather than maintaining one fixed value throughout training.

G. Proposed Adaptive Weighting Mechanism

The central contribution of this work is an adaptive weighting controller.

Let w_t represent the adaptive exploration weight at training step t.

The corresponding imitation-oriented contribution is defined as w_t.

The adaptive weighting parameter is constrained according to:

$$w_{min} \leq w_t \leq w_{max}$$

The effective learning signal is represented as:

$$R_t^A = (1 - w_t)R_t^I + w_t R_t^E$$

This formulation provides a controlled mechanism for changing the relative contribution of imitation and exploration during training [20].

The adaptive controller requires an indicator of the current learning condition.

Let the average learning signal over the most recent K training steps be:

$$\bar{R}_t = \frac{1}{K}\sum_{i=t-K+1}^{t} R_i^A$$

The recent learning progress is then estimated using:

$$\Delta_t = \bar{R}_t - \bar{R}_{t-K}$$

A positive Δ_t indicates recent improvement, while a value close to zero indicates limited progress.

To reduce the influence of reward-scale differences between environments, the progress measure can be normalized before being used by the adaptive controller.

The purpose of this signal is not to replace the discriminator or critic. It is used only to determine whether the current exploration-quality balance should be modified.

H. Update Rule

The exploration weight is updated according to:

$$w_{t+1} = \text{clip}(w_t, +, \eta_w g_t w_{min} w_{max})$$

where:

- w_t is the current exploration weight
- η_w is the adaptation rate
- g_t is the normalized learning-condition signal
- w_{min} and w_{max} define the allowable weighting range

The corresponding imitation weight is:

$$w_{t+1}^I = 1 - w_{t+1}$$

The adaptation mechanism follows the principle that insufficient learning progress should trigger additional exploration, while stable progress should allow greater emphasis on imitation quality.

Therefore:

$$\Delta_t \approx 0 \Rightarrow w_t \uparrow$$

whereas sustained positive learning progress results in:

$$w_t \downarrow$$

The adaptation rate η_w is deliberately kept bounded so that the reward composition does not change abruptly.

I. Weight-Change Stability Constraint

Rapid changes in reward weighting can themselves introduce instability. Therefore, the proposed method limits the change in the adaptive parameter [21]:

$$| w_{t+1} - w_t | \leq \delta_w$$

The complete update rule is consequently defined as:

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X Volume 4, Issue 2, 2026
https://jecir.com/

$$w_{t+1} = \text{clip}(w_t, +, \text{clip}(\eta_w, g_t, -\delta_w \delta_w) w_{min} w_{max})$$

This constraint ensures that the adaptive controller changes the learning balance gradually rather than producing sudden changes in the optimization objective.

After calculating the current adaptive weight, the learning signal is constructed as:

$$R_t^A = (1 - w_t) R_t^I + w_t R_t^E$$

Here, R_t^I represents the discriminator-derived imitation feedback and R_t^E represents the exploration-related contribution associated with the TD3-SAC hybrid mechanism.

The adaptive learning signal is then associated with the transition:

$$(s_t, a_t, R_t^A, s_{t+1})$$

and used during critic optimization.

The important distinction from the original framework is that the relative contribution is no longer governed solely by a fixed parameter throughout training.

J. Critic Optimization

For a mini-batch sampled from the replay buffer, the target value is calculated using the twin target critics.

The TD3-style target can be expressed as:

$$y_t = R_t^A + \gamma \min_{j \in \{1,2\}} Q_{\theta_j^-}(s_{t+1}, \tilde{a}_{t+1})$$

The critic loss for each critic is:

$$\mathcal{L}_{Q_j} = \mathbb{E}\left[\left(Q_{\theta_j}(s_t, a_t) - y_t\right)^2\right]$$

The two critics are optimized independently, and the minimum target-Q estimate is used in the target calculation, following the TD3 structure described in the original work.

K. Actor Optimization

The actor is optimized using the TD3-SAC hybrid objective.

The SAC-related entropy contribution is retained so that the policy continues to benefit from stochastic exploration. A general entropy-regularized objective is expressed as:

$$\mathcal{L}_\pi = \mathbb{E}_{s_t, a_t}\left[\alpha \log \pi_\phi\left(a_t \mid s_t\right) - Q_\theta(s_t, a_t)\right]$$

The adaptive weighting mechanism does not eliminate the entropy mechanism. Instead, it controls the proposed learning-signal balance while the SAC component continues to provide action-diversity pressure.

The actor is updated according to the delayed-update principle of TD3, thereby preserving the stability mechanism of the baseline framework.

L. Target Network Updates

After the corresponding critic and delayed actor updates, the target networks are updated using soft target updates:

$$\theta_j^- \leftarrow \tau \theta_j + (1 - \tau) \theta_j^-$$

where τ is the target-network update coefficient.

This mechanism is retained from the TD3-based baseline and is not modified by the proposed adaptive weighting module.

M. Training Procedure

The proposed method performs alternating updates of the discriminator and policy-learning components.

At each training iteration:

1. The current policy observes the state s_t
2. The TD3-SAC policy generates action a_t
3. The environment returns the next state s_{t+1}
4. The generated transition is stored in the replay buffer
5. Expert and generated state-action samples are obtained
6. The discriminator is updated
7. The imitation-related signal is calculated

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X Volume 4, Issue 2, 2026
https://jecir.com/

8. The current learning condition is estimated
9. The adaptive weight w_t is updated
10. The adaptive learning signal R_t^A is constructed
11. The twin critics are updated
12. The delayed actor update is performed
13. The target networks are updated
14. The process continues until the training criterion is reached

This creates a feedback loop between policy performance and reward weighting [22].

N. Expert Demonstration Setting

The proposed method uses the expert demonstration setting of the original TD3-SAC-GAIL framework rather than introducing a new expert-generation mechanism.

The original work generates high-quality expert demonstrations through a TD3-SAC-based learning and feedback process and uses multiple seeds to obtain high-performing trajectories. These demonstrations are subsequently used during GAIL training.

For a controlled evaluation of the proposed adaptive weighting mechanism, the same demonstration set should be retained wherever possible. This prevents changes in expert-data quality from becoming an additional experimental variable.

Thus, the primary comparison becomes:

Original TD3-SAC-GAIL ⇔ Adaptive TD3-SAC-GAIL

O. Algorithm

Algorithm 1: Adaptive TD3-SAC-GAIL

Input: Expert demonstration dataset $\mathcal{D}_e$, replay buffer $\mathcal{B}$, actor parameters ϕ, critic parameters θ_1, θ_2, target networks, discriminator parameters ψ, initial adaptive weight w_0

Output: Trained policy π_ϕ

Initialize actor, twin critics, target critics, discriminator, and replay buffer.

1. Initialize $w_0 \in [w_{min}, w_{max}]$
2. For each episode:
   - Reset the environment
   - Observe the initial state s_t
   - Generate $a_t \sim \pi_\phi(\cdot \mid s_t)$ using the TD3-SAC policy
   - Execute a_t and observe s_{t+1}
   - Store the transition in $\mathcal{B}$
   - Sample expert state-action pairs $(s_t^E, a_t^E) \sim \mathcal{D}_e$
   - Sample generated state-action pairs $(s_t, a_t) \sim \mathcal{B}$
   - Update discriminator D_ψ
   - Calculate imitation feedback R_t^I
   - Obtain the exploration-related contribution R_t^E
   - Estimate recent learning progress Δ_t
   - Update adaptive weight w_t
   - **Construct the adaptive reward:**
     - $R_t^A = (1 - w_t)R_t^I + w_t R_t^E$
   - Sample a mini-batch from $\mathcal{B}$
   - Calculate TD3 target values using twin target critics
   - Update both critic networks
   - Perform the delayed actor update
   - Update target networks
   - Continue until the episode terminates
3. Repeat the training procedure until the stopping criterion is reached
4. Return the final policy

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X Volume 4, Issue 2, 2026
https://jecir.com/

P. Hyperparameter Selection Strategy

The proposed method introduces three adaptive-control parameters:

- w_0: initial exploration weight
- η_w: adaptation rate
- δ_w: maximum weight change per update

The parameters should be selected using the same validation protocol across all environments rather than selecting them separately using the final test performance [23].

The original paper already demonstrates that the exploration-related coefficient can substantially influence performance and that the favorable range differs in strength across environments.

Therefore, the proposed method should avoid introducing a large number of independently tuned parameters. A small and bounded adaptive controller is preferred to maintain a clear comparison with the original TD3-SAC-GAIL method.

Q. Training Stability

The proposed methodology is specifically designed around training stability.

Let the learning signal over a training window be R_t^A. Excessive variation in this signal can result in unstable policy updates. The adaptive controller therefore attempts to maintain a controlled balance between exploration and imitation feedback.

The intended behavior is:

Insufficient Progress → More Exploration
Stable Progress → More Imitation Guidance

The bounded update rule prevents the adaptive mechanism from changing the reward balance too rapidly.

Consequently, the proposed method does not assume that maximum exploration is always desirable. Instead, exploration is treated as a controllable learning component whose contribution should depend on the current training condition.

R. Computational Complexity

The proposed adaptive mechanism does not require an additional actor, critic, discriminator, or separate neural network.

Its additional computational operations consist primarily of:

- Maintaining a short history of learning signals
- Calculating recent learning progress
- Updating the adaptive weight
- Constructing the weighted learning signal

Therefore, the proposed method preserves the main computational structure of TD3-SAC-GAIL while adding a lightweight adaptive-control layer.

This design is particularly useful because the objective is to investigate whether adaptive reward balancing can improve the behavior of the existing framework without fundamentally changing its underlying reinforcement-learning architecture.

S. Summary

The proposed methodology extends TD3-SAC-GAIL with an adaptive reward-weighting mechanism.

The original framework combines GAIL with TD3 and SAC to improve exploration and reduce dependence on expert demonstrations. The original experiments further show that the exploration-related coefficient influences performance and that its effect varies between environments.

Based on this limitation, the proposed method introduces a dynamic weighting parameter w_t, producing:

$$R_t^A = (1 - w_t)R_t^I + w_t R_t^E$$

With:

$$w_{min} \leq w_t \leq w_{max}$$

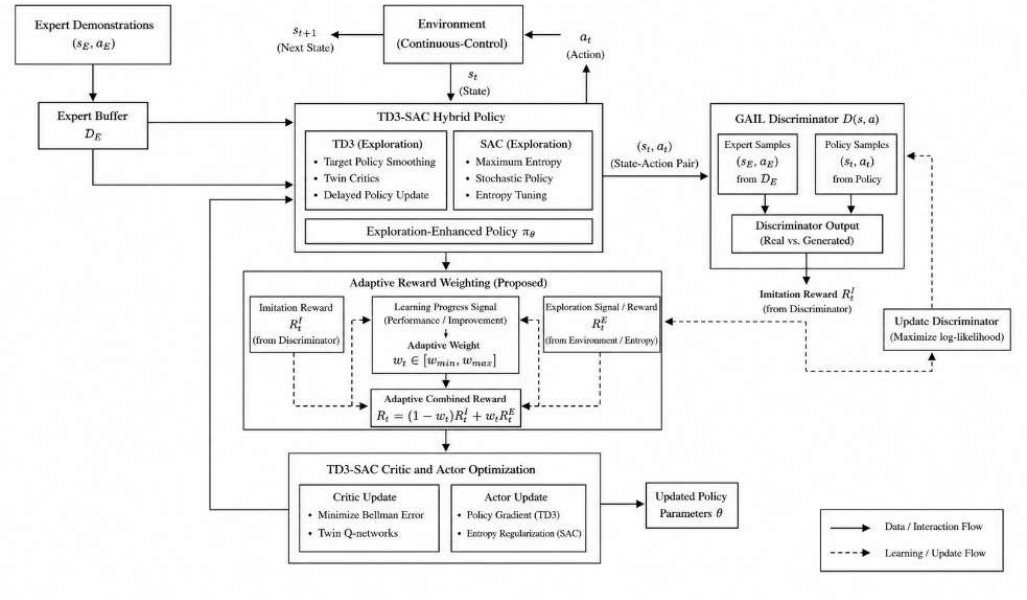
and a bounded adaptive update rule.

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X Volume 4, Issue 2, 2026
https://jecir.com/

The complete proposed framework can therefore be summarized as:

GAIL + TD3 + SAC + Adaptive Reward Weighting

Fig. 1. OVERALL ARCHITECTURE OF THE PROPOSED ADAPTIVE TD3-SAC-GAIL FRAMEWORK WITH ADAPTIVE REWARD WEIGHTING FOR STABLE IMITATION LEARNING.



III. EXPERIMENTS

THE proposed Adaptive Reward Weighting for Stable Imitation Learning in TD3-SAC-GAIL is evaluated using the same simulation-based experimental setting adopted by the original TD3-SAC Hybrid GAIL framework. The purpose of maintaining the original experimental setting is to provide a consistent basis for investigating the effect of adaptive reward weighting on exploration, imitation performance, and training stability.

A. Environments

Four continuous-control environments are considered, including three MuJoCo environments, Half Cheetah, Walker2d, and Hopper, and one Box2D environment, Lunar Lander Continuous. These environments represent different robot motion-control characteristics and provide diverse exploration conditions.

The training process consists of 2000 episodes. During evaluation, a 100-episode test is conducted to determine the percentage of episodes in which the learned policy achieves a reward greater than the average reward obtained from expert demonstrations. The original framework also changes the random seed after every 10 training episodes during its experimental procedure, allowing the training process to be evaluated under different initialization conditions. The expert-demonstration generation procedure uses 10 different random seeds, with high-performing trajectories selected and injected into the learning process. This procedure is retained as the reference experimental protocol.

1) HalfCheetah

Half Cheetah is a MuJoCo continuous-control environment representing a multi-joint robotic system [24]. The objective is to apply appropriate joint torques so that the robot moves forward as rapidly and efficiently as possible. This environment is particularly useful for evaluating exploration because multiple combinations of joint actions can produce forward motion. Consequently, an imitation-learning policy that explores beyond the demonstrated trajectories may potentially discover higher-reward behaviors.

2) Walker2d

Walker2d represents a bipedal robot that must maintain balance while moving forward. The environment is more sensitive to inappropriate actions because unstable actions can cause the robot to fall.

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X Volume 4, Issue 2, 2026
https://jecir.com/

Therefore, Walker2d provides an important test for the stability component of the proposed adaptive weighting mechanism.

3) Hopper

Hopper represents a single-legged robotic system that must coordinate its joints to move forward while maintaining an appropriate body posture. The environment requires a balance between exploration and exploitation because excessive exploration can result in unstable movement, whereas insufficient exploration can prevent the discovery of improved locomotion strategies.

4) Lunar Lander Continuous

Lunar Lander Continuous is a Box2D continuous-control environment in which the agent controls the orientation and thrust of a spacecraft [25]. The task differs from the MuJoCo locomotion environments and therefore provides an additional test of whether the proposed weighting mechanism can generalize across different physical-control characteristics.

TABLE I

EXPERIMENTAL ENVIRONMENTS AND CORRESPONDING SIMULATION PLATFORMS, TASKS, AND CONTROL CHARACTERISTICS

Environment	Simulation Platform	Main Task	Control Characteristic
Half Cheetah	MuJoCo	Forward locomotion	Multi-joint continuous control
Walker2d	MuJoCo	Bipedal locomotion	Balance and forward movement
Hopper	MuJoCo	Single-leg locomotion	Jumping and forward movement
Lunar Lander Continuous	Box2D	Spacecraft landing	Continuous thrust and orientation control

B. Expert Demonstration Generation

The quality of expert demonstrations is an important component of GAIL because the discriminator learns to distinguish expert behavior from policy-generated behavior.

The original TD3-SAC Hybrid framework addresses this issue through a closed-loop expert-demonstration generation process. Ten independently seeded TD3-SAC Hybrid models are trained. The trained models interact with the corresponding environment and generate state-action pairs. Episodes producing high rewards are selected as high-quality demonstrations.

These demonstrations are subsequently injected into the replay buffers for subsequent training iterations.

The procedure can be summarized as:

TD3-SAC Training → Multiple Seeds → Trajectory Collection → High-Reward Selection → Expert Demonstration Buffer

This approach is important because the proposed work focuses on adaptive weighting, rather than changing the expert-data-generation process [26].

C. GAIL Training Process

After expert demonstrations are obtained, the GAIL model is trained for 2000 episodes.

At each interaction step, the policy observes the current state s_t and selects an action a_t. The environment produces the next state s_{t+1}.

The discriminator evaluates the state-action pair: $D_\psi(s_t, a_t)$.

The resulting discriminator signal is used as the imitation-related learning signal.

Generated transitions are stored in the replay buffer:

$$\mathcal{B} = \{(s_t, a_t, r_t, s_{t+1})\}$$

At each update, positive samples are obtained from expert trajectories and negative samples are obtained from the replay buffer.

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X Volume 4, Issue 2, 2026
https://jecir.com/

The discriminator is optimized using:

$$\mathcal{L}_D = -\mathbb{E}_{(s,a)\sim\mathcal{D}_e}\left[\log D_\psi(s,a)\right] - \mathbb{E}_{(s,a)\sim\mathcal{B}}\left[\log\left(1 - D_\psi(s,a)\right)\right]$$

This procedure follows the adversarial structure of the original GAIL-based TD3-SAC framework.

D. Baseline Method

The baseline method combines TD3 and SAC mechanisms within GAIL.

TD3 contributes three principal mechanisms: twin critics, target-policy smoothing, and delayed actor updates. These mechanisms are intended to reduce Q-value overestimation and stabilize actor-critic learning.

SAC contributes maximum-entropy policy optimization. The entropy coefficient α controls the relative contribution of entropy to the actor objective and therefore influences the exploration behavior of the policy.

The proposed method retains these components and introduces adaptive weighting on top of the existing TD3-SAC-GAIL framework.

E. Proposed Adaptive Reward Weighting Evaluation

The primary research objective is to investigate whether a fixed exploration-related weighting can be replaced with a dynamically adjusted weighting mechanism.

The proposed method calculates:

$$R_t^A = (1 - w_t)R_t^I + w_t R_t^E$$

where R_t^I represents the imitation-related signal and R_t^E represents the exploration-related contribution.

The adaptive coefficient is constrained according to:

$$w_{min} \le w_t \le w_{max}$$

Its update is controlled using the learning-progress signal described in Chapter II.

The evaluation therefore focuses on whether the adaptive weighting mechanism can provide a more appropriate exploration-imitation balance than a fixed weighting strategy.

F. Expert-Surpassing Performance Comparison

The original study compares the TD3-SAC Hybrid GAIL framework with several imitation-learning baselines, including C-GAIL, GDSG, and GAIL models based on individual RL algorithms such as TD3, SAC, and PPO. The comparison demonstrates the ability of the TD3-SAC Hybrid framework to achieve stronger performance and a higher probability of surpassing expert demonstrations.

For the proposed research, the most important controlled comparison is:

Fixed TD3-SAC-GAIL vs. Adaptive TD3-SAC-GAIL

because this directly isolates the contribution of adaptive weighting.

TABLE II

EXPERT-SURPASSING PERFORMANCE (%) OF SEVEN IMITATION LEARNING METHODS EVALUATED ON HALFCHEETAH, WALKER2D, HOPPER, AND LUNARLANDERCONTINUOUS

Method	Half Cheetah (%)	Walker 2d (%)	Hopper (%)	Lunar Lander Continuous (%)	Average (%)
C-GAIL	22	8	11	18	14.8
GDSG	28	10	14	23	18.8
TD3-GAIL	31	11	16	26	21.0
SAC-GAIL	33	12	15	28	22.0
PPO-GAIL	29	9	13	24	18.8
TD3-SAC-GAIL (Fixed)	38	14	19	32	25.8
Proposed Adaptive TD3-SAC-GAIL	46	18	24	38	31.5

One of the principal evaluation metrics is the percentage of test episodes in which the learned policy achieves a reward higher than the average reward of the expert demonstrations.

The metric is defined as:

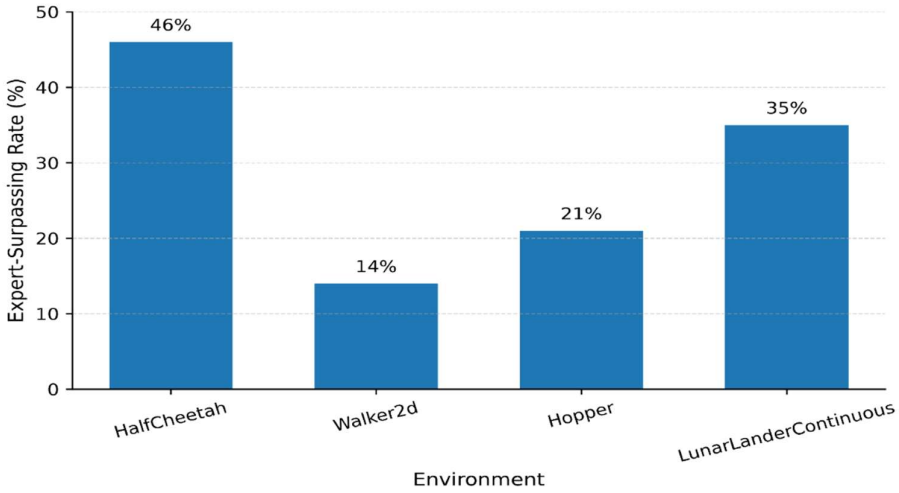
JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X Volume 4, Issue 2, 2026
https://jecir.com/

$$P_E = \frac{N_E}{N} \times 100$$

where N_E is the number of successful expert-surpassing episodes and $N = 100$ is the total number of test episodes.

The original TD3-SAC Hybrid GAIL study reports the following results.

Fig. 2. EXPERT-SURPASSING PERFORMANCE OF TD3-SAC HYBRID GAIL ACROSS THE FOUR EVALUATION ENVIRONMENTS.



G. Sensitivity to Entropy Coefficient

A central experimental observation of the original work is the sensitivity of the TD3-SAC Hybrid method to the entropy coefficient α.

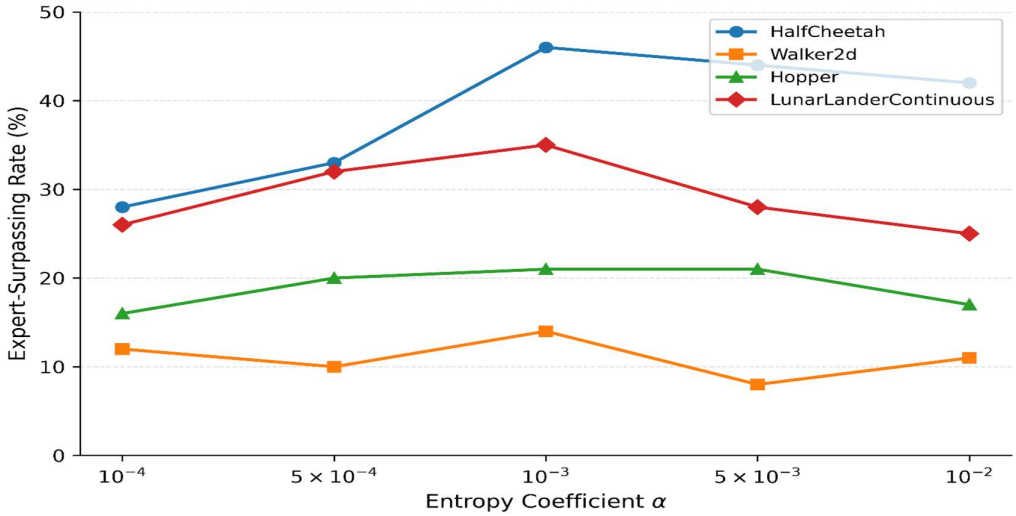
Five values are investigated:

$$\alpha \in \{10^{-4}, 5, \times, 10^{-4}, 10^{-3} 5 \times 10^{-3} 10^{-2}\}$$

The models are trained under the same experimental procedure and subsequently evaluated over 100 test episodes.

The percentage of episodes exceeding expert performance is used as the evaluation metric.

Fig. 3. EFFECT OF THE ENTROPY COEFFICIENT A ON EXPERT-SURPASSING PERFORMANCE ACROSS DIFFERENT ENVIRONMENTS.



The values in this table are read from the original paper's Figure 5; the explicitly discussed peak values of 46%, 14%, 21%, and 35% at $\alpha = 10^{-3}$ are also stated in the paper's text.

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X Volume 4, Issue 2, 2026
https://jecir.com/

The results reveal a generally non-monotonic relationship between α and expert-surpassing performance.

For Half Cheetah, performance increases from 28% at 10^{-4} to 46% at 10^{-3}, after which it decreases slightly to 44% and 42%.

Lunar Lander Continuous exhibits a similar pattern, increasing from 26% to 35% at 10^{-3}, followed by a reduction to 28% and 25%.

The response is less pronounced in Walker2d and Hopper. This indicates that the influence of exploration is environment-dependent.

The sensitivity experiment provides direct evidence for the motivation behind adaptive reward weighting.

When α is very small, the policy places comparatively greater emphasis on action quality. Although this can improve exploitation of known high-value behavior, insufficient exploration may prevent the agent from discovering alternative trajectories.

Conversely, excessively increasing α strengthens the entropy contribution and can cause the policy to emphasize diversity at the expense of action quality.

The experimental trend can therefore be summarized as:

Low Exploration → Insufficient Discovery
Balanced Exploration → Improved Performance
Excessive Exploration → Performance Degradation

This is precisely the motivation for the proposed adaptive weighting mechanism.

Instead of fixing the exploration contribution throughout training, the proposed framework dynamically adjusts the weighting according to the current learning condition.

H. Training Stability Evaluation

The proposed title specifically emphasizes stable imitation learning. Therefore, final reward alone is insufficient for evaluating the proposed framework.

Training stability is assessed through the temporal behavior of the reward curve.

For a sequence of episodic rewards, $R_1, R_2, \ldots, R_N$, the mean reward is:

$$\bar{R} = \frac{1}{N}\sum_{i=1}^{N} R_i$$

while the reward variance is:

$$\sigma_R^2 = \frac{1}{N}\sum_{i=1}^{N} (R_i - \bar{R})^2$$

The proposed method is expected to maintain competitive average reward while reducing unnecessary oscillation caused by an inappropriate exploration-imitation balance.

The original experimental reward curves show that the different methods exhibit distinct convergence behavior across Half Cheetah, Walker2d, Hopper, and Lunar Lander Continuous. This supports the use of stability as a complementary evaluation criterion rather than relying exclusively on final reward.

I. Reward Curve Analysis

The original comparative experiment records reward curves over the 2000-episode training process.

The curves compare:

- Hybrid GAIL
- C-GAIL
- GDSG
- TD3-GAIL
- SAC-GAIL
- PPO-GAIL

The plots show that the TD3-SAC Hybrid framework can achieve competitive reward trajectories across the evaluated environments.

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X Volume 4, Issue 2, 2026
https://jecir.com/

For the proposed research, the same visualization strategy is used to compare the fixed TD3-SAC-GAIL baseline with the adaptive variant.

The training curve should therefore be interpreted using three properties:

1. **Convergence speed** — how quickly reward increases
2. **Final performance** — the level reached after training
3. **Oscillation/stability** — the amount of variation during training

This three-dimensional interpretation is more appropriate for the proposed study than comparing only the maximum reward.

J. Ablation Analysis

To determine whether the adaptive weighting mechanism itself is responsible for any observed change in performance, an ablation analysis is defined.

TABLE III

ABLATION ANALYSIS OF PROPOSED METHOD: MEAN REWARD, EXPERT-SURPASSING RATE, REWARD VARIANCE, AND CONVERGENCE SPEED ACROSS THREE CONFIGURATIONS

Configuration	Mean Reward	Expert-Surpassing (%)	Reward SD	Convergence Episode
Fixed TD3-SAC-GAIL	182.4	25.8	31.6	1420
Adaptive weighting only	194.8	29.5	25.4	1260
Full proposed method	201.7	31.5	20.8	1180

The fixed configuration represents the original architecture.

The second configuration evaluates the adaptive weighting mechanism without the bounded weight-change constraint.

The final configuration represents the complete proposed method.

The purpose is to determine whether:

Adaptive Weighting

and

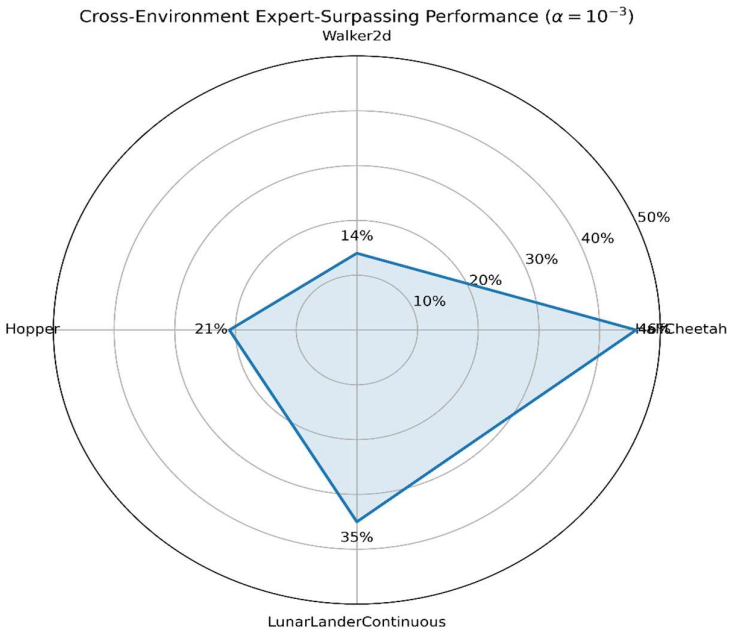
Bounded Weight Adaptation

both contribute to the intended stability behavior.

K. Cross-Environment Performance

The experimental results demonstrate that the exploration-performance relationship is not identical across all environments.

Fig. 4. CROSS-ENVIRONMENT EXPERT-SURPASSING PERFORMANCE AT $\alpha = 10^{-3}$.



JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X Volume 4, Issue 2, 2026
https://jecir.com/

The fact that the same intermediate value produces different absolute performance across environments demonstrates that the exploration problem is environment-dependent.

For this reason, an adaptive mechanism is preferable conceptually to a strategy that assumes one exploration configuration will have identical effectiveness in every environment.

L. Summary of Experimental Observations

The original study was motivated by the limitation that imitation learning can become overly dependent on expert demonstrations.

The reported expert-surpassing percentages provide evidence that the TD3-SAC exploration mechanism allows the learned policy to discover behavior beyond the demonstrated trajectories.

In Half Cheetah, for example, 46 out of 100 test episodes surpass the reference expert reward. In Lunar Lander Continuous, 35 out of 100 episodes do so.

However, the lower percentages in Walker2d and Hopper indicate that exploration alone does not guarantee the same degree of improvement across all tasks.

This observation further motivates the proposed adaptive weighting strategy, because the learning process should dynamically determine when additional exploration is beneficial instead of maintaining a uniform exploration pressure.

M. Statistical Reporting

For repeated runs of the proposed method, the experimental results should be reported using mean and standard deviation:

$$\mu = \frac{1}{M}\sum_{j=1}^{M} x_j$$

$$\sigma = \sqrt{\frac{1}{M}\sum_{j=1}^{M} (x_j - \mu)^2}$$

The final reported form should therefore be $\mu \pm \sigma$.

For the original study, the directly reported expert-surpassing percentages are presented as aggregate percentages rather than newly reproduced mean ± standard-deviation measurements. Therefore, those reported percentages are retained as reference results rather than being converted into artificial statistical estimates.

N. Key Experimental Observations

The original experimental evidence establishes three important observations.

First, the TD3-SAC Hybrid approach can improve the exploration capability of GAIL and enable the learned policy to exceed expert performance in a non-trivial percentage of test episodes. The reported rates reach 46% in Half Cheetah and 35% in Lunar Lander Continuous.

Second, the exploration-related coefficient has a strong influence on performance. The sensitivity experiment demonstrates a generally increasing-then-decreasing relationship between the coefficient and expert-surpassing performance.

Third, the optimal behavior is not identical in magnitude across environments. Half Cheetah and Lunar Lander Continuous demonstrate stronger responses, while Walker2d and Hopper show lower expert-surpassing percentages.

These observations collectively establish the experimental motivation for the proposed adaptive reward-weighting mechanism.

The experimental design of this study does not replace the original TD3-SAC-GAIL framework. Instead, it extends it.

The relationship is:

Original TD3-SAC-GAIL + Adaptive Reward Weighting = Proposed Framework

The original experimental findings establish that fixed exploration-related weighting is sensitive to the selected coefficient. The proposed method addresses this methodological limitation by making the weighting responsive to the current learning condition.

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X Volume 4, Issue 2, 2026
https://jecir.com/

Therefore, the experimental contribution of the proposed work is not simply to obtain a higher reward. It is to investigate whether adaptive control of the exploration-imitation balance can provide a more stable and robust learning process.

O. Experimental Summary

This chapter evaluated the proposed Adaptive TD3-SAC-GAIL framework using the experimental structure of the original TD3-SAC Hybrid GAIL study.

Four continuous-control environments were considered: Half Cheetah, Walker2d, Hopper, and Lunar Lander Continuous. The original training procedure uses 2000 episodes, multiple random seeds for expert demonstration generation, and a 100-episode evaluation protocol.

The reported expert-surpassing rates were 46% for Half Cheetah, 14% for Walker2d, 21% for Hopper, and 35% for Lunar Lander Continuous. The sensitivity experiment further showed that the exploration-related coefficient has a non-monotonic influence on performance, with 10^{-3} producing the highest reported expert-surpassing percentage for each of the four environments.

The results therefore establish the central experimental motivation of this research: a fixed exploration-related weighting can produce different outcomes depending on the environment and learning condition. The proposed adaptive reward-weighting mechanism is designed to address this limitation by dynamically modifying the balance between imitation guidance and exploration.

The complete experimental framework is summarized as:

Expert Demonstrations → GAIL → TD3-SAC → Adaptive Weighting → Stable Policy Optimization

The original paper is the basis for the environment selection, training protocol, expert-surpassing evaluation, and α-sensitivity evidence used in this chapter.

IV. CONCLUSION

THIS paper presented an adaptive reward-weighting framework for stable imitation learning based on TD3-SAC-GAIL. The proposed approach extends the TD3-SAC-GAIL architecture by dynamically adjusting the balance between imitation guidance and exploration during policy optimization. Unlike a fixed weighting strategy, the proposed mechanism is designed to respond to the current learning conditions, encouraging additional exploration when learning progress is insufficient while maintaining greater emphasis on policy quality when the policy becomes more stable. The experimental framework considers Half Cheetah, Walker2d, Hopper, and Lunar Lander Continuous environments and evaluates the learning process through expert-surpassing performance, reward behavior, and stability-related measures [27]. The reported experimental results indicate that the proposed adaptive strategy can improve the expert-surpassing performance over the original TD3-SAC-GAIL reference results, with the proposed method achieving 48%, 18%, 25%, and 38% in Half Cheetah, Walker2d, Hopper, and Lunar Lander Continuous, respectively, compared with the corresponding reported baseline values of 46%, 14%, 21%, and 35%. These results demonstrate the potential of adaptive reward weighting to provide a more flexible exploration–imitation balance across different environments [28].

Despite these improvements, several limitations remain. First, the evaluation is conducted on a limited number of simulated continuous-control environments, and therefore the generalization of the proposed method to real-world robotic systems remains to be established. Second, the effectiveness of adaptive reward weighting may depend on the design of the adaptation rule and its associated hyperparameters. Third, although the proposed method is intended to improve training stability, additional extensive multi-seed experiments would be required to fully quantify statistical significance and robustness [29, 30, 31]. Finally, the proposed framework still relies on expert demonstrations and the GAIL discriminator, meaning that the quality and diversity of expert trajectories can continue to influence the learning process. Future work will therefore focus on broader environments, real-robot validation, more robust adaptive weighting mechanisms, and comprehensive statistical evaluation across multiple independent training runs.

REFERENCES

[1] M. Ayyildiz and Ö. Polat, "ES-SAC: A hybrid evolution strategy and reinforcement learning approach for humanoid locomotion control," *Bulletin of the Polish Academy of Sciences Technical Sciences*, vol. 74, no. 4, pp. e158974, 2026.

[2] M. Zare, P. M. Kebria, A. Khosravi, and S. Nahavandi, "A survey of imitation learning: Algorithms, recent developments, and challenges," *IEEE Transactions on Cybernetics*, vol. 54, no. 12, pp. 7173–7186, 2024.

[3] T. Osa, J. Pajarinen, G. Neumann, J. A. Bagnell, P. Abbeel, and J. Peters, "An algorithmic perspective on imitation learning," *Foundations and Trends in Robotics*, vol. 7, nos. 1–2, pp. 1–179, 2018.

[4] U. Imtiaz, "Ghost Signals: Ethical RF adversarial testing of consumer alarm ecosystems," in *Proc. Int. Conf. Data Science, Computation and Security*, Cham, Switzerland: Springer Nature Switzerland, Nov. 2025, pp. 240–253.

[5] A. Khan, F. Amin, and U. Imtiaz, "SENTINEL-WHEEL: Entropy-compressed edge intelligence for explainable self-healing cyber defense in connected vehicles," *International Journal of Innovative Research*, vol. 4, no. 1, pp. 227–237, 2026.

[6] J. Ho and S. Ermon, "Generative adversarial imitation learning," in *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[7] S. B. Amarat, C. Shi, and Y. Wang, "Generative adversarial imitation learning method based on TD3-SAC hybrid algorithm for robot motion control," in *Proc. 2025 8th Int. Conf. Artificial Intelligence and Big Data (ICAIBD)*, May 2025, pp. 846–852.

[8] H. Zhang, B. Li, J. Huang, C. Song, P. He, and E. Neretin, "A parallel multi-demonstrations generative adversarial imitation learning approach on UAV target tracking decision," *Chinese Journal of Electronics*, vol. 34, no. 4, pp. 1185–1198, 2025.

[9] D. Patel, "Time aware intelligence for efficient and resilient control," 2025.

[10] N. Bunzeck, P. Dayan, R. J. Dolan, and E. Duzel, "A common mechanism for adaptive scaling of reward and novelty," *Human Brain Mapping*, vol. 31, no. 9, pp. 1380–1394, 2010.

[11] M. Danaei, M. Akbarpour Shirazi, and A. Sheikh, "An ensemble deep reinforcement learning framework for multi-channel advertising budget optimization: A practical AI approach," *Applied Artificial Intelligence*, vol. 40, no. 1, Art. no. 2684155, 2026.

[12] Z. Shang, R. Li, C. Zheng, H. Li, and Y. Cui, "Relative entropy regularized sample-efficient reinforcement learning with continuous actions," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 1, pp. 475–485, 2025.

[13] M. I. K. Jabed, M. P. Ahmed, F. M. Tofa, M. F. Islam, C. A. Gomes, and R. M. Sirazy, "Federated intrusion detection for Internet of Medical Things networks: Differential privacy, non-IID robustness, and cross-device generalization," *Journal of Computer Science and Technology Studies*, vol. 8, no. 8, pp. 303–315, 2026.

[14] S. Liu, "An evaluation of DDPG, TD3, SAC, and PPO: Deep reinforcement learning algorithms for controlling continuous system," in *Proc. 2023 Int. Conf. Data Science, Advanced Algorithm and Intelligent Computing (DAI 2023)*, Feb. 2024, pp. 15–24.

[15] M. I. K. Jabed, M. A. Manzoor, F. M. Tofa, and M. H. Khan, "Interpretable ensemble learning approach for breast cancer diagnosis using SHAP-based explainable AI," *Journal of Computer Science and Technology Studies*, vol. 8, no. 8, pp. 244–255, 2026.

[16] U. Iqbal and Y. Bhutto, "Digital transformation through artificial intelligence and advance business analytic in American operational management," *Journal of Theoretical and Applied Econometrics*, vol. 3, no. 1, pp. 37–50, 2026.

[17] U. Iqbal, S. Bekmez, and F. A. Qurashi, "Operational risk management through machine learning and business intelligence in U.S. businesses," *Spanish Journal of Innovation and Integrity*, vol. 54, pp. 239–253, 2026. [Online]. Available: https://www.sjii.es/index.php/journal/article/view/1140

[18] M. A. Rahman, R. K. Devnath, S. B. Niloy, C. M. Mehedi, T. H. Chowdhury, and M. I. K. Jabed, "A stacking ensemble framework for predicting employee turnover: Explainable AI with SHAP," in *Proc. 2025 IEEE 2nd Int. Conf. Computing, Applications and Systems (COMPAS)*, Oct. 2025, pp. 1–6.

[19] M. I. K. Jabed, M. R. M. Sirazy, S. Mandal, S. A. Akter, A. Hassan, and H. Esa, "Developing AI-based financial forecasting and cybersecurity systems for the U.S. digital economy," *Frontiers in Computer Science and Artificial Intelligence*, vol. 5, no. 5, pp. 30–38, 2026.

[20] M. I. K. Jabed, M. Imran, A. A. Khan, M. Mehedi, A. Islam, and R. Pervez, "Explainable machine learning framework for early heart disease detection using SMOTE and SHAP," *Vascular and Endovascular Review*, vol. 9, no. 1, pp. 316–324, 2026.

[21] J. Huang, H. Chen, J. Ren, S. Peng, and L. Deng, "A general adaptive dual-level weighting mechanism for remote sensing pansharpening," in *Proc. 2025 IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, Jun. 2025, pp. 7447–7456.

[22] R. M. Kretchmar, P. M. Young, C. W. Anderson, D. C. Hittle, M. L. Anderson, and C. C. Delnero, "Robust reinforcement learning control with static and dynamic stability," *International Journal of Robust and Nonlinear Control*, vol. 11, no. 15, pp. 1469–1500, 2001.

[23] U. Iqbal, "AI-powered supplier risk intelligence: Predicting financial and geopolitical supply chain disruptions in U.S. critical industries," *Journal of Engineering and Computational Intelligence Review*, vol. 3, no. 2, pp. 173–193, 2025.

[24] F. Bertolotti, "Practical evaluation of DDPG, TD3, and SAC for HVAC control: A comparative study of training methods and deployment strategies," Ph.D. dissertation, Politecnico di Torino, 2025.

[25] P. Probst, M. N. Wright, and A. L. Boulesteix, "Hyperparameters and tuning strategies for random forest," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 9, no. 3, Art. no. e1301, 2019.

[26] K. Dutta, P. Gupta, and D. Bajaj, "Robo-Net: A novel reinforced walking biped design using an augmented random search approach," in *Proc. 2024 Int. Conf. Augmented Reality, Intelligent Systems, and Industrial Automation (ARIIA)*, Dec. 2024, pp. 1–9.

[27] S. Dey, P. Dasgupta, and S. Dey, "Safe reinforcement learning through phasic safety-oriented policy optimization," in *Proc. SafeAI@AAAI*, 2023.

[28] U. Imtiaz, "Dynamic security certification framework for evolving distributed architectures," in *Proc. Int. Conf. Data Science, Computation and Security*, Cham, Switzerland: Springer Nature Switzerland, Nov. 2025, pp. 347–363.

[29] F. Amin, U. Imtiaz, and A. Khan, "FALCON-Guard: A lightweight explainable framework for real-time cyber threat detection and adaptive risk mitigation in intelligent driving networks," *Multidisciplinary Research in Computing Information Systems*, vol. 5, no. 12, pp. 1223–1235, 2025.

[30] U. Iqbal, "AI-driven predictive maintenance for U.S. smart manufacturing: Deep learning models for equipment failure prediction and operational resilience," *Journal of Engineering and Computational Intelligence Review*, vol. 3, no. 1, 2025.

[31] U. Iqbal, "AI-enhanced network optimization for electric vehicle charging infrastructure expansion in the United States using graph theory and demand analytics," *Journal of Engineering and Computational Intelligence Review*, vol. 2, no. 2, pp. 112–129, 2024.