Unlocking Potential in Suboptimal Demonstrations via Progressive Trajectory Valuation

Yiyang Xu[1], Yilin Liu[1], Tao Wang[1], Shiwei Li[1], Xiangfeng Luo[1], and Shaorong Xie[1,*]
[1]School of Computer Engineering and Science, Shanghai University, Shanghai, China
{xuyiyang, liuyilin, wangt96, lishiwei, luoxf, srxie}@shu.edu.cn

***Abstract*—Imitation learning (IL) provides a powerful framework for skill acquisition by leveraging expert demonstrations. However, its real-world application is often hindered by the prevalence of suboptimal demonstrations, which violates the common assumption of optimal expert policies. While existing methods handle imperfect data via static weighting or ranking, they typically fail to capture and utilize the progressive nature of learning—the valuable trends of improvement within trajectories that are evolving towards expert-level performance. To bridge this gap, we propose Progressive Confidence-based Adversarial Imitation Learning (PCAIL). Our framework establishes a dual-assessment paradigm, which dynamically reinterprets imperfect demonstrations throughout training via adaptive target margins that shift the focus from static quality to instructive potential. This enables the model to prioritize demonstrations showing positive learning progression, thereby enhancing both sample efficiency and final performance. Extensive evaluations on six MuJoCo continuous control tasks and a custom unmanned surface vehicle (USV) reconnaissance environment demonstrate that PCAIL achieves consistent and significant improvements over a range of established baselines. Our work offers a practical and principled solution for learning from imperfect data, establishing a paradigm that strategically exploits the progressive potential within imperfect data rather than treating it as homogeneous noise.**

***Index Terms*—reinforcement learning, imitation learning, suboptimal demonstrations, dynamic weighting, continuous control**

I. Introduction

Imitation learning (IL) [1] enables agents to acquire complex skills directly from expert demonstrations, bypassing the need for hand-designed reward functions in reinforcement learning (RL) [2]. This approach circumvents the complexities of reward engineering and has demonstrated effectiveness in diverse applications such as robotic control [3], autonomous navigation [4], and intelligent systems [5]. Among IL methods, Generative Adversarial Imitation Learning (GAIL) [6] has gained prominence for its robustness in framing imitation as a distribution-matching problem, inspiring numerous variants that enhance aspects like data efficiency and stability [7]–[12].

Despite these advances, a fundamental obstacle hinders the real-world deployment of IL: the prevalence of suboptimal or noisy demonstrations [13]–[16]. In practice, data collected from human operators or physical systems is rarely optimal due to cognitive biases, sensor noise, and environmental stochasticity. This directly violates the core assumption in standard IL that demonstrations originate from an optimal policy [17]. Consequently, training on such imperfect data can lead to policy degradation, where the agent converges to suboptimal and brittle behaviors.

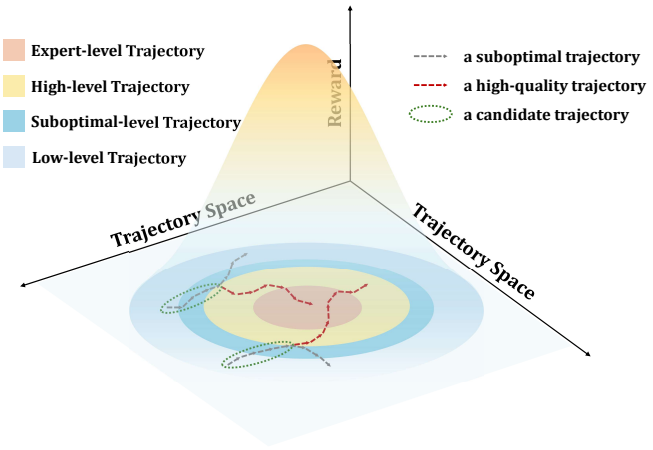


Fig. 1: Conceptual illustration of the progressive reward landscape in trajectory space. Candidate trajectories (green) exhibit improvement from suboptimal regions toward the expert-level center.

To mitigate the impact of imperfect data, existing methods primarily focus on filtering or re-evaluating the provided demonstrations. These approaches can be broadly categorized by their core mechanism. Weighting-based methods [13], [18]–[22] assign static importance scores to trajectories based on estimated quality or confidence. Ranking-based methods [23]–[27], in contrast, infer a reward function from relative quality orderings among demonstrations. A third paradigm first infers a reward function from the imperfect data and then employs offline reinforcement learning for policy extraction [28]–[32].

Despite their contributions, these methods share a fundamental limitation: they treat each demonstration as a *static snapshot*, assessing its global quality in isolation. This perspective overlooks the intrinsic *progressive structure* within real-world demonstrations, where trajectories often contain segments of meaningful improvement. By neglecting this temporal evolution, existing approaches discard a crucial learning signal—the latent pathway from suboptimal to expert performance—and thus fail to fully exploit the potential within imperfect data.

To address this challenge, we propose **Progressive Confidence-based Adversarial Imitation Learning**

*Corresponding author.

(PCAIL). Our key insight is that suboptimal demonstrations contain heterogeneous potential: while many segments are of low quality, others exhibit a promising trend of progressive improvement toward expert behavior, as conceptually visualized in Fig. 1. Building on this, PCAIL introduces a dual-assessment and dynamic-margin mechanism that shifts the focus from static quality assessment to the valuation of both immediate quality and future potential.

Technically, PCAIL extends the adversarial imitation learning framework with two synergistic components: a dual-score assessment module and a phase-aware scheduler. The dual-score assessment module quantifies both immediate confidence and future promise for each demonstration segment. Based on these scores and the current training phase, the phase-aware scheduler dynamically calibrates their relative influence, balancing the influence between high-potential and high-confidence demonstrations. This mechanism induces an adaptive learning strategy that dynamically reinterprets demonstrations via adaptive target margins: it amplifies the contribution of high-potential segments during early training to bootstrap skill acquisition, and progressively shifts the focus toward high-confidence data for policy refinement. This transforms the learning paradigm from indiscriminate imitation to the strategic utilization of latent progression, yielding a more stable and informative learning signal for policy optimization.

Our main contributions are summarized as follows:

- **A Novel Dual-Assessment Paradigm:** We identify and formalize the dual nature of imperfect demonstrations—immediate quality versus future potential—and propose a novel dual-score assessment and dynamic-margin mechanism that adaptively revalues demonstrations throughout training.
- **The PCAIL Framework:** We integrate this mechanism into the adversarial imitation learning paradigm, resulting in the PCAIL framework, which enables robust policy learning directly from suboptimal, noisy data.
- **Comprehensive Empirical Evaluation:** We demonstrate the efficacy of PCAIL through extensive experiments on six MuJoCo continuous control tasks and a photo-realistic Unmanned Surface Vehicle (USV) reconnaissance environment, where it achieves superior performance compared to strong established baselines.

II. Related Work

A. Confidence and Weighting Based Methods

Confidence and weighting based methods address imperfect demonstrations by assigning importance weights to data points based on estimated confidence or quality metrics. Pioneering works such as 2IWIL and IC-GAIL [13] demonstrated the effectiveness of this paradigm. Subsequent research has focused on relaxing the need for manual confidence labels. For instance, WGAIL [18] connects confidence to an exponential advantage function estimated from the discriminator and policy. Other approaches include DWBC [15], which employs an auxiliary discriminator for weight inference, and CAIL [20], which leverages partial rankings to guide confidence estimation. Recently, PN-GAIL [22] innovatively utilizes both positive and negative information from imperfect demonstrations through dual-risk assessment. Despite their effectiveness, these methods predominantly treat demonstrations as static datasets and often rely on assumptions such as the availability of a certain proportion of optimal data. Consequently, they may face convergence challenges and fail to capture the progressive dynamics within trajectories.

B. Ranking and Preference Based Methods

Ranking and preference based methods infer reward functions by leveraging relative quality orderings among trajectories. This line of work was pioneered by T-REX [23], which learns a reward function consistent with provided rankings, enabling generalization to unseen behaviors. D-REX [24] extends this approach by automatically generating ranked trajectories via noise injection, thereby reducing reliance on manual annotations. More recent methods, including SSRR [25] and LERP [27], further refine reward function structures and robustness to noisy preferences. However, these approaches primarily focus on static orderings of complete trajectories. They typically require substantial numbers of ranked pairs to learn robust reward functions and, similar to weighting-based methods, do not fully exploit intra-trajectory dynamics or the progressive nature of skill acquisition within individual demonstrations.

C. Reward Inference and Offline RL Based Methods

Reward inference and offline RL based methods first learn a reward function from demonstrations and subsequently train a policy via offline reinforcement learning. This paradigm explicitly decouples reward learning from policy optimization. Representative works include ORIL [28] and MLIRL [33], which learn discriminative reward functions between expert and non-expert behaviors. While powerful, these methods can exhibit sensitivity to hyperparameters and training instability due to the complexity of adversarial optimization. Moreover, the learned reward functions are typically static and may not generalize beyond the offline data distribution. This limitation can misguide the agent in unseen states and overlook valuable information about how behaviors improve progressively within individual trajectories.

III. Background

A. Reinforcement Learning

Reinforcement learning (RL) addresses sequential decision-making through agent-environment interaction, typically formalized as a Markov Decision Process (MDP) [2] $M = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma, \mu_0)$. Here, $\mathcal{S}$ and $\mathcal{A}$ denote state and action spaces, $\mathcal{P}(s'|s,a)$ is the transition dynamics, $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function, $\gamma \in [0, 1)$ the discount factor, and μ_0 the initial state distribution.

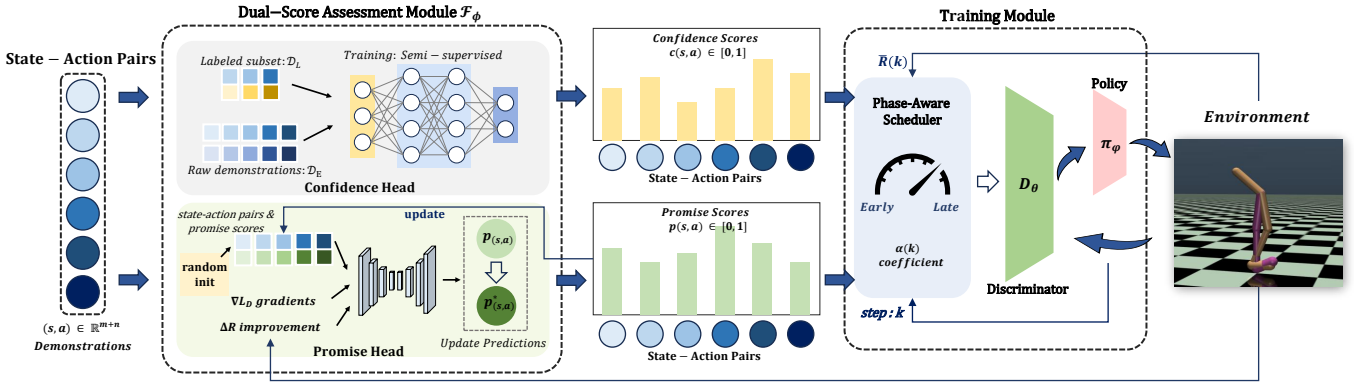


Fig. 2: The PCAIL framework. The dual-score module outputs confidence c and promise p for each demonstration. The phase-aware scheduler computes α from training progress. These produce dynamic targets y for the discriminator D_θ in its adversarial training with policy π_ψ, enabling progressive learning from imperfect data.

The agent's behavior is defined by a policy $\pi(a|s)$. The objective is to find an optimal policy π^* maximizing the expected discounted return:

$$J(\pi) = \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t\mathcal{R}(s_t,a_t)\right], \tag{1}$$

where $\tau = (s_0, a_0, s_1, a_1, \dots)$ denotes a trajectory.

B. Imitation Learning (IL)

Imitation learning (IL) bypasses reward specification by learning directly from expert demonstrations $\mathcal{D}_E = \{(s_i,a_i)\}_{i=1}^N$. The goal is to recover a policy π_θ that approximates the expert policy π_E.

a) Behavioral Cloning (BC): The simplest IL approach treats imitation as supervised learning:

$$\min_\theta \mathbb{E}_{(s,a)\sim\mathcal{D}_E}\left[-\log\pi_\theta(a|s)\right]. \tag{2}$$

While straightforward, BC [34] suffers from compounding error due to distributional shift and lack of interaction with the environment.

b) Generative Adversarial Imitation Learning (GAIL): GAIL [6] formulates imitation as distribution matching via adversarial training. A discriminator $D_\psi : \mathcal{S}\times\mathcal{A}\to[0,1]$ is trained to distinguish expert state-action pairs from those generated by the policy π_θ, while π_θ learns to deceive D_ψ. The objective is:

$$\min_\theta\max_\psi \ \mathbb{E}_{\pi_E}\big[\log D_\psi(s,a)\big] + \mathbb{E}_{\pi_\theta}\big[\log(1-D_\psi(s,a))\big]. \tag{3}$$

Here, $D_\psi(s,a)$ estimates the probability that (s,a) comes from the expert. This adversarial formulation avoids compounding error but assumes near-optimal demonstrations. Performance degrades significantly with suboptimal data, motivating our approach.

IV. Method

In this section, we introduce the Progressive Confidence-based Adversarial Imitation Learning (PCAIL) algorithm. PCAIL reframes the problem of learning from imperfect demonstrations by dynamically quantifying and leveraging the progressive structure—the latent trajectory from error to correction—within sub-optimal data. The framework is built upon three synergistic components: a dual-score assessor, a phase-aware scheduler, and an adversarial trainer with a dynamically shaped objective. At its heart, PCAIL eschews static data valuation. Instead, it employs a time-varying blending mechanism that continuously fuses two complementary data assessments: one of current reliability, and another of future instructive value. This fused signal dynamically recalibrates the discriminator's target, thereby shifting the policy's focus across the training horizon. Consequently, the policy is first guided to distill broadly beneficial behavioral motifs before being finely tuned toward expert precision, achieving robust imitation through structured, self-paced learning. The overall architecture is illustrated in Fig. 2.

A. Dual-Score Assessment Module

Let $\mathcal{D}_E = \mathcal{D}_L \cup \mathcal{D}_U$ be the imperfect demonstration set, where $\mathcal{D}_L = \{(s_i,a_i,y_i)\}$ contains a small labeled subset with binary labels $y_i \in \{0,1\}$ indicating optimal (1) or sub-optimal (0) actions. For each $(s,a)\in\mathcal{D}_E$, PCAIL learns two scores: confidence $c(s,a)$ and promise $p(s,a)$.

1) Immediate Confidence Score: We learn a confidence estimator $g_\psi(s,a)\in[0,1]$ via semi-supervised learning. The overall objective combines supervised and consistency terms:

$$\min_\psi \ \mathcal{L}_{\text{sup}}(\psi) + \lambda\mathcal{L}_{\text{cons}}(\psi), \tag{4}$$

where the supervised loss on labeled data is:

$$\mathcal{L}_{\text{sup}} = -\mathbb{E}_{(s,a,y)\sim\mathcal{D}_L}\big[y\log g_\psi + (1-y)\log(1-g_\psi)\big], \tag{5}$$

and the consistency loss enforces smoothness along trajectories $\tau = \{(s_t,a_t)\}$:

$$\mathcal{L}_{\text{cons}} = \mathbb{E}_{\tau\sim\mathcal{D}_U}\left[\frac{1}{T-1}\sum_{t=1}^{T-1}\big(g_\psi(s_t,a_t) - g_\psi(s_{t+1},a_{t+1})\big)^2\right]. \tag{6}$$

The confidence score is $c(s,a) = g_\psi(s,a)$.

2) Future Promise Score: The promise score $p(s,a) \in [0,1]$ predicts the training utility of emphasizing (s,a). We model it as:

$$p_\phi(s,a) = \sigma(h_\phi(s,a)), \tag{7}$$

where h_ϕ is trained to minimize:

$$\min_\phi \ \mathbb{E}_{(s,a)\sim\mathcal{D}_E}\big[\big(h_\phi(s,a) - \Delta(s,a)\big)^2\big]. \tag{8}$$

Here, $\Delta(s,a)$ measures the actual policy improvement after utilizing (s,a), computed online as the increase in the policy's average discriminator score over a short horizon.

3) Joint Architecture: Both scores use a shared feature encoder $f_\theta(s,a)$. Parameters are updated alternately: (θ,ψ) using $\mathcal{L}_{\text{conf}}$, and (θ,ϕ) using $\mathcal{L}_{\text{promise}}$.

B. Phase-Aware Scheduler

We design a robust scheduler that adapts to learning progress while being resilient to performance fluctuations. The key is to distinguish between meaningful learning plateaus and transient noise.

First, we compute a smoothed performance estimate using exponential moving average:

$$\bar{R}(k) = 0.95 \cdot \bar{R}(k-1) + 0.05 \cdot R_k, \quad \bar{R}(0) = R_0, \tag{9}$$

where R_k is the return at iteration k.

For $k \geq 100$, we compute the average progress rate over a long window:

$$\rho(k) = \frac{\bar{R}(k) - \bar{R}(k-100)}{100}. \tag{10}$$

For $k < 100$, we define $\rho(k) = 0$.

The blending coefficient $\alpha(k)$ is updated adaptively:

$$\alpha(k) = \begin{cases} \min(1.0,\ \alpha(k-1) + 0.01), & \text{if } \rho(k) < -\epsilon, \\ \max(0.0,\ \alpha(k-1) - 0.05), & \text{if } \rho(k) > \epsilon, \\ \alpha(k-1), & \text{otherwise}, \end{cases} \tag{11}$$

with initial value $\alpha(0) = 1.0$ and threshold $\epsilon = 0.005$.

This design embodies three key principles for robust adaptation. First, it employs asymmetric adjustment: decreasing α (shifting focus toward confidence-based refinement) is easier and faster than increasing it, ensuring that once convergence toward expert precision begins, it is not easily disrupted. Second, it incorporates hysteresis by requiring negative progress ($\rho(k) < -\epsilon$) to trigger a shift toward promise-based exploration, which filters out transient performance dips. Third, it ensures bounded and gradual adaptation: α is constrained to $[0,1]$ and changes incrementally, preventing abrupt oscillations in training focus and promoting stable learning dynamics.

C. Dynamic-Margin Adversarial Training

The core of PCAIL is a novel adversarial training mechanism that unifies the dual scores and the scheduling signal into a dynamic objective. Unlike standard GAIL [6] that treats all expert demonstrations equally, we define a *dynamic target margin* for each demonstration sample.

Let $D_\theta(s,a)$ denote the discriminator output, where values near 0 indicate expert-like behavior and values near 1 indicate policy-like behavior. For policy samples $(s,a) \sim \pi$, the discriminator target remains 1 as in standard GAIL. For expert demonstrations $(s,a) \sim \mathcal{D}_E$, we define a dynamic target based on the current blending coefficient $\alpha(k)$ and the dual scores:

$$y(s,a;k) = \alpha(k)(1-p(s,a)) + (1-\alpha(k))(1-c(s,a)), \tag{12}$$

where k is the training iteration. This formulation creates smooth interpolation: when $\alpha(k) \approx 1$, the target is dominated by the promise score $p(s,a)$, encouraging the discriminator to treat high-promise samples as more expert-like; when $\alpha(k) \approx 0$, it is dominated by the confidence score $c(s,a)$, focusing on high-confidence samples.

The discriminator is trained with a least-squares loss to fit these dynamic targets:

$$\mathcal{L}_D(\theta) = \mathbb{E}_\pi[(D_\theta - 1)^2] + \mathbb{E}_{\mathcal{D}_E}[(D_\theta - y)^2]. \tag{13}$$

The policy receives a shaped reward defined as the negative log-probability of being classified as non-expert:

$$r(s,a) = -\log D_\theta(s,a). \tag{14}$$

The policy parameters ψ are updated using Proximal Policy Optimization (PPO) to maximize the expected discounted return:

$$\max_\psi \mathbb{E}_{\tau\sim\pi_\psi}\left[\sum_{t=0}^{T} \gamma^t r(s_t,a_t)\right]. \tag{15}$$

The dynamic target $y(s,a;k)$ bridges the assessment module and adversarial training. When $\alpha(k)$ is high (early training), high-promise samples have targets near 0, making them attractive to imitate. As $\alpha(k)$ decays, high-confidence samples become primary imitation targets, refining the policy toward expert precision. This implements automated curriculum learning where the discriminator's notion of "expertness" coevolves with the policy's capability.

D. Training Procedure

Algorithm 1 presents the complete PCAIL procedure. The key distinction from standard adversarial imitation learning is the integration of dual-score assessment and phase-aware scheduling into the adversarial training loop.

V. EXPERIMENTS

A. Experimental Setup

1) Environments and Datasets: We evaluate PCAIL on two categories of environments to validate its ability to exploit progressive structures within imperfect demonstrations.

MuJoCo Benchmarks. We employ six standard continuous control tasks from MuJoCo: *Pendulum-v1*, *Ant-v2*, *Walker2d-v2*, *Hopper-v2*, *Swimmer-v2*, and *HalfCheetah-v2*. For each environment, we use the publicly available imperfect demonstration datasets from the 2IWIL repository, which include confidence scores. Following prior work, we treat only 20% of each dataset as labeled ($\mathcal{D}_L$), while the remainder serves

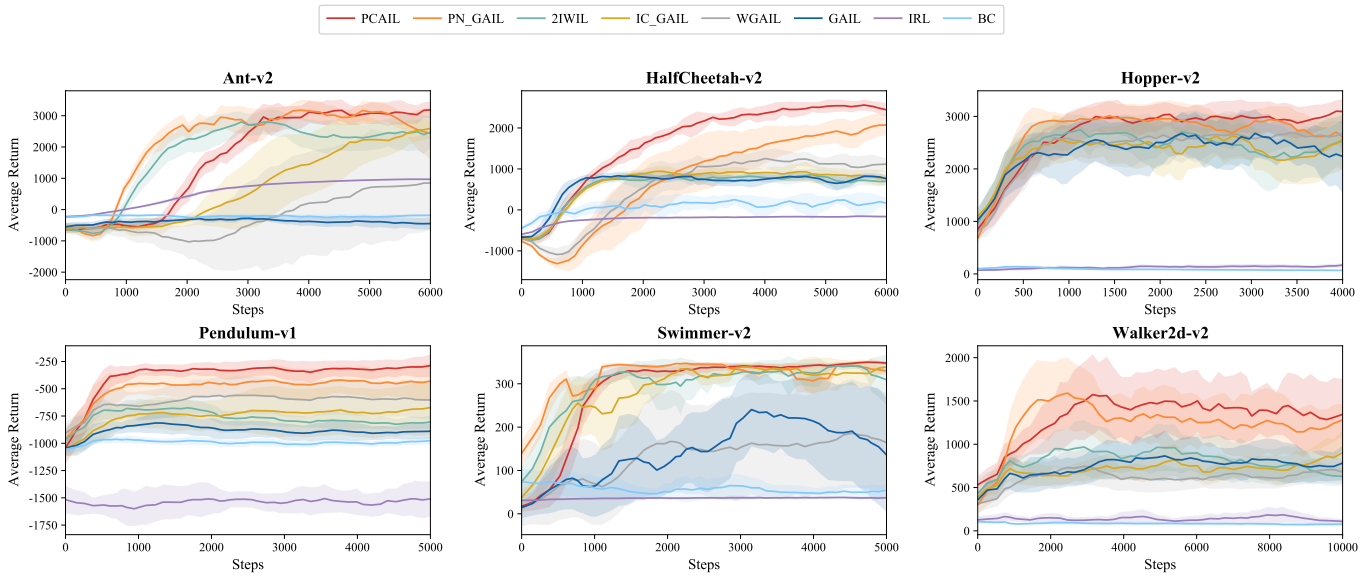


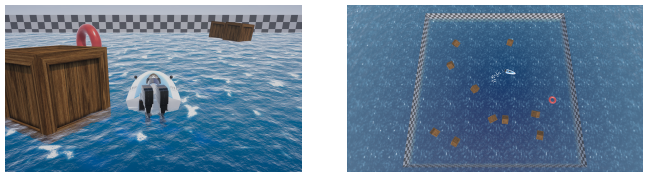
Fig. 3: Learning curves of PCAIL and baseline methods on six MuJoCo environments. PCAIL (red) consistently achieves higher average returns across all environments.

Algorithm 1 Progressive Confidence-based Adversarial Imitation Learning (PCAIL)

Require: Imperfect demonstrations $\mathcal{D}_E$, labeled subset $\mathcal{D}_L$, iterations K
Ensure: Learned policy π_ψ

```
Initialize π_ψ, D_θ, f_φ, α ← 1.0, R̄ ← 0
for k = 1 to K do
    Collect D_π ~ π_ψ
    Evaluate policy to obtain return R_k {Optional: use D_π's average reward}
    Assess: (c, p) ← f_φ(s, a) for (s, a) ∈ D_E
    Schedule: Update α via Eqs. (9, 10, 11)
    Compute targets: y ← α(1 − p) + (1 − α)(1 − c) {Eq. (12)}
    Update discriminator: θ ← θ − ∇_θ L_D {Eq. (13)}
    Update policy: ψ ← ψ + ∇_ψ E[− log D_θ] {PPO}
end for
return π_ψ
```

as unlabeled data ($\mathcal{D}_U$). For Pendulum-v1, demonstrations are generated by mixing trajectories from an optimal and a sub-optimal policy in a 1:4 ratio.



(a) Agent's perspective (b) Bird's-eye view

Fig. 4: The Unity USV navigation environment.

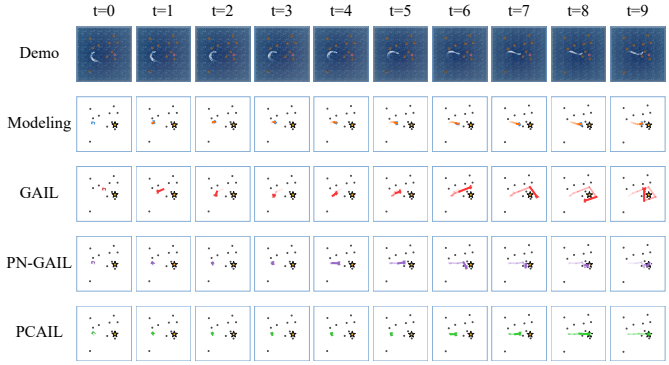


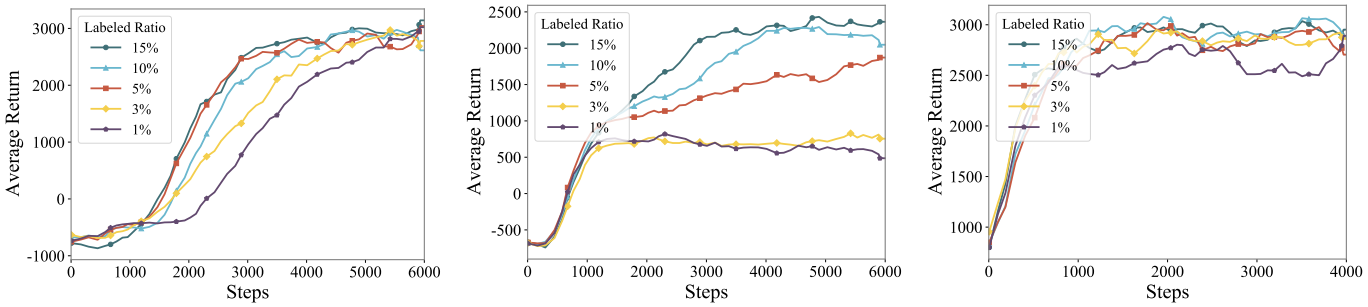
Fig. 5: Qualitative comparison of trajectories on the USV navigation task. PCAIL generates smooth, target-reaching paths, while baselines exhibit erratic or suboptimal behaviors.

Unity USV Navigation Task. To test PCAIL in a more complex, realistic setting, we design a custom navigation environment in Unity (Fig. 4). An Unmanned Surface Vehicle (USV) must navigate a 200m×200m lake to locate a target while avoiding randomly placed obstacles. Imperfect demonstrations are collected from human operators via a game controller. These demonstrations exhibit clear progressive improvement as operators gain experience throughout the collection process. A total of 200 trajectories are collected, with 20% randomly labeled for confidence evaluation.

2) Baselines and Implementation: Our evaluation encompasses three key aspects: comparison against established baselines, ablation studies of our method, and robustness analysis under limited supervision. For baseline comparisons, we include classic imitation learning approaches: Behavioral Cloning (BC) [34] , Inverse Reinforcement

TABLE I: Ablation study of PCAIL components across MuJoCo environments and the USV task.

Method	Ant-v2	Walker2d-v2	Hopper-v2	Swimmer-v2	HalfCheetah-v2	Pendulum-v1	USV
PCAIL\D	2450.3 ± 382.5	1250.8 ± 187.6	2880.7 ± 345.2	290.6 ± 52.1	2150.2 ± 322.8	-295.4 ± 45.3	285.1 ± 42.8
PCAIL\S	2780.6 ± 333.7	1380.2 ± 165.6	2650.4 ± 318.0	325.3 ± 43.9	2380.5 ± 285.7	-335.7 ± 50.4	310.4 ± 37.2
PCAIL (full)	3250.8 ± 260.1	1570.5 ± 141.3	3150.2 ± 252.0	358.9 ± 32.2	2680.7 ± 214.5	-253.6 ± 30.4	356.1 ± 28.5



(a) Ant-v2 (b) HalfCheetah-v2 (c) Hopper-v2

(d) Pendulum-v1 (e) Swimmer-v2 (f) Walker2d-v2

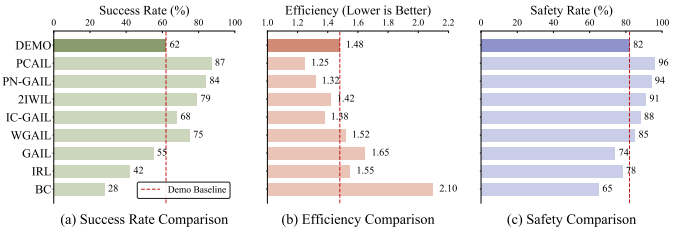
Fig. 6: Performance of PCAIL with varying proportions of labeled data (15%, 10%, 5%, 3%, 1%) across six MuJoCo environments.

Learning (IRL) [35], and Generative Adversarial Imitation Learning (GAIL) [6], alongside state-of-the-art methods designed specifically for imperfect demonstrations: 2IWIL, IC-GAIL [13], WGAIL [18], and PN-GAIL [22]. For ablation studies, we evaluate PCAIL\D (without the dual-score assessment module) and PCAIL\S (without the phase-aware scheduler) to validate the contribution of each component. Finally, to assess robustness, we reduce the proportion of labeled data from 15% to 1% to simulate scenarios with extremely scarce supervision.

All methods share a consistent architecture: three-layer MLPs with 64 units per layer for both policy and discriminator. The dual-score module uses a shared encoder followed by separate output heads for confidence $c(s, a)$ and promise $p(s, a)$. Optimization employs PPO for the policy and Adam for the discriminator and assessment module. Results are averaged over five independent random seeds and reported as mean ± standard deviation.

B. Performance Comparison

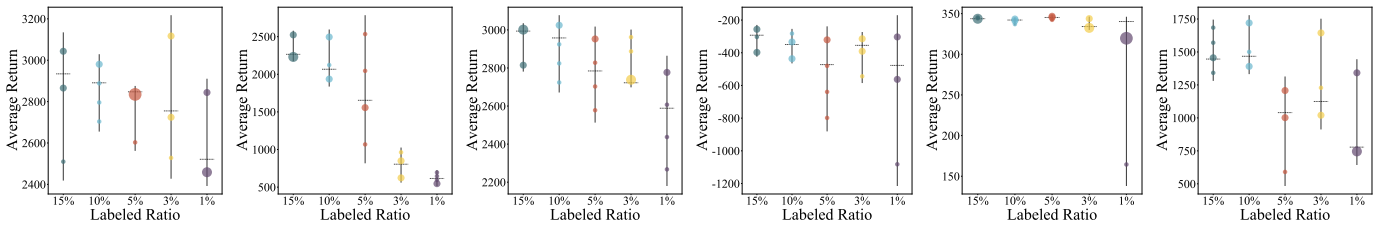
1) MuJoCo Benchmarks: As shown in Fig. 3, PCAIL achieves the highest final returns and most stable learning



(a) Success Rate Comparison (b) Efficiency Comparison (c) Safety Comparison

Fig. 7: Quantitative evaluation on the USV task: (a) success rate, (b) path efficiency (lower is better), (c) safety rate. PCAIL outperforms all baselines and exceeds the performance of the original human demonstrators (red dashed line).

curves across all six MuJoCo environments, validating the effectiveness of our dual-assessment paradigm. The advantage is most pronounced where demonstrations contain systematic biases or complex progressive structures. In Pendulum-v1, PCAIL learns a near-optimal policy despite low-confidence biases that degrade performance for confidence-weighted baselines like 2IWIL. In motor-skills-intensive tasks (HalfCheetah-v2, Walker2d-v2), PCAIL shows a clear margin over other



(a) Ant-v2 (b) HalfCheetah-v2 (c) Hopper-v2 (d) Pendulum-v1 (e) Swimmer-v2 (f) Walker2d-v2

Fig. 8: Performance distribution of PCAIL across labeled data proportions. Vertical lines indicate performance ranges across five random seeds; bubble sizes represent result density.

baselines, indicating that dynamic assessment surpasses static weighting for complex behaviors.

Classic methods show expected limitations: GAIL treats all demonstrations as optimal and performs poorly; BC suffers from compounding error. Among confidence-aware methods, PN-GAIL performs robustly but is consistently surpassed by PCAIL, highlighting the advantage of modeling both immediate quality and future potential.

2) USV Navigation Task: The photo-realistic USV environment provides a challenging testbed with naturally progressive human demonstrations. PCAIL generates smooth, obstacle-avoiding trajectories (Fig. 5), while baselines produce erratic or suboptimal paths. Quantitatively (Fig. 7), PCAIL achieves 87% success rate, 1.25 path efficiency, and 96% safety rate—surpassing all baselines and exceeding the original human demonstrators (62%, 1.48, 82%).

The performance gap widens in this noisy setting. While baselines show limited effectiveness, PCAIL not only outperforms them but also refines policy information beyond the human demonstrations.

Across both standardized benchmarks and the realistic USV task, PCAIL demonstrates consistent superiority through its dynamic integration of dual-score assessment and phase-aware scheduling, enabling adaptive utilization of demonstrations' progressive potential—a capability lacking in prior methods.

C. Ablation Study

To validate the contributions of PCAIL's core components, we conduct ablation experiments across all environments. We evaluate three configurations: (1) PCAIL\D removes the dual-score assessment module, using only immediate confidence scores; (2) PCAIL\S removes the phase-aware scheduler, employing static weighting; (3) PCAIL (full) integrates both components.

As shown in Table I, the full PCAIL achieves the highest performance across all environments. The performance gap is most pronounced in complex tasks: in Ant-v2, PCAIL improves by 32.7% over PCAIL\D and 16.9% over PCAIL\S. This demonstrates that both components contribute significantly to the final performance.

The dual-score module (PCAIL vs. PCAIL\S) provides particular benefit in environments with progressive demonstration structures. In Walker2d-v2 and HalfCheetah-v2, where demonstrations contain meaningful improvement patterns, modeling future promise yields performance gains of 13.8% and 12.6% respectively. The phase-aware scheduler (PCAIL vs. PCAIL\D) shows its value in adapting to different training stages, with the largest improvements observed in early-convergence tasks like Pendulum-v1.

These results confirm that PCAIL's performance advantage stems from the synergistic operation of its dual components: the dual-score module identifies valuable progressive patterns, while the scheduler optimally balances their utilization throughout training.

D. Robustness to Limited Supervision

We evaluate PCAIL's robustness by varying the proportion of labeled demonstrations from the standard 15% down to 1%. This tests the method's ability to operate with minimal supervision—a practical requirement for real-world deployment.

Figures 6 and 8 show that PCAIL maintains strong performance even with severely limited labeled data. The learning curves remain stable, with only gradual degradation as supervision decreases from 15% to 1%. Notably, in complex environments like Ant-v2 and Hopper-v2, PCAIL retains over 85% of its peak performance even with only 1% labeled data.

The distribution analysis reveals consistent behavior across random seeds. Tight vertical ranges and concentrated bubble distributions indicate low variance, confirming PCAIL's stability under varying supervision levels. This robustness stems from PCAIL's semi-supervised design: the dual-score module effectively leverages unlabeled data, while the scheduler adapts to the available supervision quality.

These results demonstrate that PCAIL substantially reduces reliance on expensive human annotations while maintaining competitive performance—a crucial advantage for practical applications where labeled data is scarce.

VI. CONCLUSION

In this paper, we introduced Progressive Confidence-based Adversarial Imitation Learning (PCAIL), a novel framework for learning from suboptimal demonstrations. Our core contribution is a dual-assessment paradigm that dynamically revalues demonstrations throughout training based on both their immediate quality and future potential. This is realized through two key components: a dual-score assessment module

that quantifies confidence and promise scores, and a phase-aware scheduler that adaptively balances their influence. This mechanism enables policies to strategically extract latent instructional value from imperfect data, achieving superior final performance and enhanced decision-making robustness compared to conventional static approaches. Extensive experiments on six MuJoCo tasks and a photo-realistic USV environment demonstrate PCAIL's effectiveness against a range of established imitation learning baselines. These results validate that explicitly modeling the progressive potential within imperfect data enhances learning robustness and sample efficiency. Future work will explore extending the dual-assessment principle to sim-to-real transfer and multi-task learning, broadening the applicability of progressive demonstration valuation in practical unmanned systems.

ACKNOWLEDGMENT

This work was supported by the Foundation for Innovative Research Groups of the National Natural Science Foundation of China (No. 62421004).

REFERENCES

[1] A. Hussein, M. M. Gaber, E. Elyan, and C. Jayne, "Imitation learning: A survey of learning methods," *ACM Computing Surveys (CSUR)*, vol. 50, no. 2, pp. 1–35, 2017.

[2] R. S. Sutton, A. G. Barto *et al.*, *Reinforcement learning: An introduction*. MIT press Cambridge, 1998, vol. 1, no. 1.

[3] B. Fang, S. Jia, D. Guo, M. Xu, S. Wen, and F. Sun, "Survey of imitation learning for robotic manipulation," *International Journal of Intelligent Robotics and Applications*, vol. 3, no. 4, pp. 362–369, 2019.

[4] H. Karnan, G. Warnell, X. Xiao, and P. Stone, "Voila: Visual-observation-only imitation learning for autonomous navigation," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 2497–2503.

[5] B. Yang, C. Ma, and X. Xia, "Drone formation control via belief-correlated imitation learning," in *Proceedings of the 20th international conference on autonomous agents and MultiAgent systems*, 2021, pp. 1407–1415.

[6] J. Ho and S. Ermon, "Generative adversarial imitation learning," *Advances in neural information processing systems*, vol. 29, 2016.

[7] Y. Li, J. Song, and S. Ermon, "Infogail: Interpretable imitation learning from visual demonstrations," *Advances in neural information processing systems*, vol. 30, 2017.

[8] J. Fu, K. Luo, and S. Levine, "Learning robust rewards with adversarial inverse reinforcement learning," *arXiv preprint arXiv:1710.11248*, 2017.

[9] R. Dadashi, L. Hussenot, M. Geist, and O. Pietquin, "Primal wasserstein imitation learning," *arXiv preprint arXiv:2006.04678*, 2020.

[10] K. Zolna, S. Reed, A. Novikov, S. G. Colmenarejo, D. Budden, S. Cabi, M. Denil, N. De Freitas, and Z. Wang, "Task-relevant adversarial imitation learning," in *Conference on Robot Learning*. PMLR, 2021, pp. 247–263.

[11] R. Bhattacharyya, B. Wulfe, D. J. Phillips, A. Kuefler, J. Morton, R. Senanayake, and M. J. Kochenderfer, "Modeling human driving behavior through generative adversarial imitation learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 3, pp. 2874–2887, 2022.

[12] C.-M. Lai, H.-C. Wang, P.-C. Hsieh, F. Wang, M.-H. Chen, and S.-H. Sun, "Diffusion-reward adversarial imitation learning," *Advances in Neural Information Processing Systems*, vol. 37, pp. 95 456–95 487, 2024.

[13] Y.-H. Wu, N. Charoenphakdee, H. Bao, V. Tangkaratt, and M. Sugiyama, "Imitation learning from imperfect demonstration," in *International Conference on Machine Learning*. PMLR, 2019, pp. 6818–6827.

[14] F. Sasaki and R. Yamashina, "Behavioral cloning from noisy demonstrations," in *International Conference on Learning Representations*, 2020.

[15] H. Xu, X. Zhan, H. Yin, and H. Qin, "Discriminator-weighted offline imitation learning from suboptimal demonstrations," in *International Conference on Machine Learning*. PMLR, 2022, pp. 24 725–24 742.

[16] Z. Li, T. Xu, Z. Qin, Y. Yu, and Z.-Q. Luo, "Imitation learning from imperfection: Theoretical justifications and algorithms," *Advances in Neural Information Processing Systems*, vol. 36, pp. 18 404–18 443, 2023.

[17] H. Yang, C. Yu, S. Chen *et al.*, "Hybrid policy optimization from imperfect demonstrations," *Advances in Neural Information Processing Systems*, vol. 36, pp. 4653–4663, 2023.

[18] Y. Wang, C. Xu, B. Du, and H. Lee, "Learning to weight imperfect demonstrations," in *International Conference on Machine Learning*. PMLR, 2021, pp. 10 961–10 970.

[19] V. Tangkaratt, B. Han, M. E. Khan, and M. Sugiyama, "Variational imitation learning with diverse-quality demonstrations," in *International Conference on Machine Learning*. PMLR, 2020, pp. 9407–9417.

[20] S. Zhang, Z. Cao, D. Sadigh, and Y. Sui, "Confidence-aware imitation learning from demonstrations with varying optimality," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12 340–12 350, 2021.

[21] Y. Wang, B. Du, and C. Xu, "Unlabeled imperfect demonstrations in adversarial imitation learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, no. 8, 2023, pp. 10 262–10 270.

[22] Q. Liu, H. Fu, K. Tang, C. Chen, and D. Dong, "PN-GAIL: Leveraging non-optimal information from imperfect demonstrations," in *The Thirteenth International Conference on Learning Representations*, 2025. [Online]. Available: https://openreview.net/forum?id=0e2pcSxQJS

[23] D. Brown, W. Goo, P. Nagarajan, and S. Niekum, "Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations," in *International conference on machine learning*. PMLR, 2019, pp. 783–792.

[24] D. S. Brown, W. Goo, and S. Niekum, "Better-than-demonstrator imitation learning via automatically-ranked demonstrations," in *Conference on robot learning*. PMLR, 2020, pp. 330–359.

[25] L. Chen, R. Paleja, and M. Gombolay, "Learning from suboptimal demonstration via self-supervised reward regression," in *Conference on robot learning*. PMLR, 2021, pp. 1262–1277.

[26] A. Taranovic, A. G. Kupcsik, N. Freymuth, and G. Neumann, "Adversarial imitation learning with preferences," in *The Eleventh International Conference on Learning Representations*, 2022.

[27] L. Huo, Z. Wang, and M. Xu, "Learning noise-induced reward functions for surpassing demonstrations in imitation learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 7, 2023, pp. 7953–7961.

[28] K. Zolna, A. Novikov, K. Konyushkova, C. Gulcehre, Z. Wang, Y. Aytar, M. Denil, N. De Freitas, and S. Reed, "Offline learning from demonstrations and unlabeled experience," *arXiv preprint arXiv:2011.13885*, 2020.

[29] J. Chang, M. Uehara, D. Sreenivas, R. Kidambi, and W. Sun, "Mitigating covariate shift in imitation learning via offline data with partial coverage," *Advances in Neural Information Processing Systems*, vol. 34, pp. 965–979, 2021.

[30] S. Yue, G. Wang, W. Shao, Z. Zhang, S. Lin, J. Ren, and J. Zhang, "Clare: Conservative model-based reward learning for offline inverse reinforcement learning," *arXiv preprint arXiv:2302.04782*, 2023.

[31] S. Zeng, C. Li, A. Garcia, and M. Hong, "When demonstrations meet generative world models: A maximum likelihood framework for offline inverse reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 36, pp. 65 531–65 565, 2023.

[32] G. Cideron, B. Tabanpour, S. Curi, S. Girgin, L. Hussenot, G. Dulac-Arnold, M. Geist, O. Pietquin, and R. Dadashi, "Get back here: Robust imitation by return-to-distribution planning," *arXiv preprint arXiv:2305.01400*, 2023.

[33] S. Zeng, C. Li, A. Garcia, and M. Hong, "Maximum-likelihood inverse reinforcement learning with finite-time guarantees," *Advances in Neural Information Processing Systems*, vol. 35, pp. 10 122–10 135, 2022.

[34] D. A. Pomerleau, "Alvinn: An autonomous land vehicle in a neural network," *Advances in neural information processing systems*, vol. 1, 1988.

[35] A. Y. Ng and S. J. Russell, "Algorithms for inverse reinforcement learning," in *Proceedings of the Seventeenth International Conference on Machine Learning*, ser. ICML '00. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2000, p. 663–670.