Novelty Assessment Report

Paper: Fine-tuning Behavioral Cloning Policies with Preference-Based Reinforcement Learning

PDF URL: https://openreview.net/pdf?id=oIiQZfnSxP

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2026-01-01

Abstract

Deploying reinforcement learning (RL) in robotics, industry, and health care is blocked by two obstacles: the difficulty of specifying accurate rewards and the risk of unsafe, data-hungry exploration. We address this by proposing a two-stage framework that first learns a safe initial policy from a reward-free dataset of expert demonstrations, then fine-tunes it online using preference-based human feedback. We provide the first principled analysis of this offline-to-online approach and introduce BRIDGE, a unified algorithm that integrates both signals via an uncertainty-weighted objective. We derive regret bounds that shrink with the number of offline demonstrations, explicitly connecting the quantity of offline data to online sample efficiency. We validate BRIDGE in discrete and continuous control MuJoCo environments, showing it achieves lower regret than both standalone behavioral cloning and online preference-based RL. Our work establishes a theoretical foundation for designing more sample-efficient interactive agents.

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **Fine-Tuning Imitation Learning Policies with Preference-Based Reinforcement Learning**

A total of **50 papers** were analyzed and organized into a taxonomy with **28 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Offline-to-Online Policy Refinement Frameworks**
- **Preference-Based Reward Learning and Policy Optimization**
- **Imitation Learning with Preference Signals**
- **Large Language Model Alignment via Preference Learning**
- **Domain-Specific Applications of Preference-Based Learning**
- **Interactive and Human-in-the-Loop Learning Systems**
- **Offline Preference Learning and Trajectory Generation**
- **Evaluation Frameworks and Benchmarking**
- **Scalable and Lifelong Learning Paradigms**

Complete Taxonomy Tree

- Fine-Tuning Imitation Learning Policies with Preference-Based Reinforcement Learning Survey Taxonomy
- Offline-to-Online Policy Refinement Frameworks
  - Unified Offline-Online Integration ★ (3 papers)
  - [0] Fine-tuning Behavioral Cloning Policies with Preference-Based Reinforcement Learning (Anon et al., 2026) View paper
  - [8] FDPP: Fine-Tune Diffusion Policy with Human Preference (Yuxin Chen, 2025) View paper
  - [47] TakeAD: Preference-Based Post-Optimization for End-to-End Autonomous Driving With Expert Takeover Data (Deqing Liu, 2025) View paper
  - Reward Model Pre-Training and Fine-Tuning (2 papers)
  - [1] Residual Reward Models for Preference-based Reinforcement Learning (Cao Chenyang, 2025) View paper
  - [21] Meta-reward-net: Implicitly differentiable reward learning for preference-based reinforcement learning (R Liu, 2022) View paper
  - Personalization via Action Representation Learning (1 papers)
  - [4] Personalization in human-robot interaction through preference-based action representation learning (Ruiqi Wang, 2025) View paper
- Preference-Based Reward Learning and Policy Optimization
  - Direct Preference Optimization Without Reward Modeling (3 papers)
  - [16] On a Connection Between Imitation Learning and RLHF (Xiao Teng, 2025) View paper
  - [24] Aligning diffusion behaviors with q-functions for efficient continuous control (Huayu Chen, 2024) View paper
  - [25] Direct preference-based policy optimization without reward modeling (An, 2023) View paper
  - Reward Model Learning from Preferences (3 papers)
  - [18] Reward learning from human preferences and demonstrations in atari (Borja Ibarz, 2018) View paper
  - [19] Exploiting Unlabeled Data for Feedback Efficient Human Preference based Reinforcement Learning (Verma, 2023) View paper
  - [29] Provable Reward-Agnostic Preference-Based Reinforcement Learning (Zhan WenHao, 2023) View paper
  - Preference-Based Exploration and Active Querying (2 papers)
  - [10] Contextual bandits and imitation learning with preference-based active queries (A Sekhari, 2023) View paper
  - [20] Active RLHF via Best Policy Learning from Trajectory Preference Feedback (A Agnihotri, 2025) View paper
  - Uncertainty-Aware and Adaptive Preference Learning (2 papers)
  - [27] Adaptive preference scaling for reinforcement learning with human feedback (Alexander Bukharin, 2024) View paper

  - [31] Uncertainty-aware preference alignment in reinforcement learning from human feedback (S Xu, 2024) View paper
- Imitation Learning with Preference Signals
  - Preference-Guided Demonstration Selection and Weighting (3 papers)
  - [11] Limited preference aided imitation learning from imperfect demonstrations (X Cao, 2024) View paper
  - [14] Learning state importance for preference-based reinforcement learning (Guoxi Zhang, 2024) View paper
  - [23] A ranking game for imitation learning (Sikchi, 2022) View paper
  - Learning from Vague or Implicit Feedback (3 papers)
  - [3] Imitation learning from vague feedback (XQ Cai, 2023) View paper
  - [9] Human Implicit Preference-Based Policy Fine-tuning for Multi-Agent Reinforcement Learning in USV Swarm (Kim Hyeonjun, 2025) View paper
  - [35] Predictive Preference Learning from Human Interventions (Cai Haoyuan, 2025) View paper
  - Adversarial and Ranking-Based Imitation (1 papers)
  - [32] Adversarial Imitation Learning with Preferences (Kupcsik, 2023) View paper
  - Heterogeneous Demonstration Handling (1 papers)
  - [50] Learning to Discern: Imitating Heterogeneous Human Demonstrations with Preference and Representation Learning (Kuhar, 2023) View paper
- Large Language Model Alignment via Preference Learning
  - Inverse RL and Bayesian Approaches for LLM Alignment (2 papers)
  - [5] Imitating Language via Scalable Inverse Reinforcement Learning (Arun Ahuja, 2024) View paper
  - [17] Approximated Variational Bayesian Inverse Reinforcement Learning for Large Language Model Alignment (YuAng Cai, 2024) View paper
  - Multimodal Preference Learning (1 papers)
  - [6] Personalizing reinforcement learning from human feedback with variational preference learning (Abhishek Gupta, 2024) View paper
  - Structured and Stage-Aware Preference Optimization (2 papers)
  - [30] STARE-VLA: Progressive Stage-Aware Reinforcement for Fine-Tuning Vision-Language-Action Models (Feng Xu, 2025) View paper
  - [43] Structured Preference Modeling for Reinforcement Learning-Based Fine-Tuning of Large Models (Zhu Lin, 2025) View paper
- Domain-Specific Applications of Preference-Based Learning
  - Robotic Manipulation and Dexterous Control (2 papers)
  - [13] Deformpam: Data-Efficient Learning for Long-Horizon Deformable Object Manipulation Via Preference-Based Action Alignment (Wendi Chen, 2024) View paper
  - [26] Learning a universal human prior for dexterous manipulation from human preference (Ding, 2023) View paper
  - Autonomous Navigation and Vehicle Control (3 papers)
  - [2] Deep reinforcement learning from human preferences for ROV path tracking (Shilong Niu, 2025) View paper
  - [39] Learning Real-World Acrobatic Flight from Human Preferences (Geles, 2025) View paper
  - [41] Enhancing the Controllability of Visual Navigation Agents with Language-Conditioned Preferences (Putta, 2025) View paper
  - Healthcare and Rehabilitation Systems (2 papers)
  - [28] Adaptive Learning based Upper-Limb Rehabilitation Training System with Collaborative Robot (Jun Hong Lim, 2023) View paper
  - [34] An online trajectory guidance framework via imitation learning and interactive feedback in robot-assisted surgery. (Ziyang Chen, 2025) View paper
  - Recommender Systems and NLP Tasks (2 papers)
  - [15] An Extremely Data-efficient and Generative LLM-based Reinforcement Learning Agent for Recommenders (Feng, 2024) View paper
  - [33] Towards Abstractive Timeline Summarisation using Preference-based Reinforcement Learning (Yuxuan Ye, 2022) View paper
- Interactive and Human-in-the-Loop Learning Systems
  - Multi-Dimensional Feedback Fusion (1 papers)
  - [46] A Human-Machine Reinforcement Learning Framework with Multi-dimensional Human Feedback Fusion (Wei Gao, 2024) View paper
  - Interactive Imitation Learning with Visual Feedback (2 papers)
  - [12] Improving multimodal interactive agents with reinforcement learning from human feedback (Abramson, 2022) View paper
  - [40] VITAL: Interactive Few-Shot Imitation Learning via Visual Human-in-the-Loop Corrections (Kasaei, 2024) View paper
  - Comparative Studies of Human-Guided Learning (3 papers)
  - [44] New approach in human-AI interaction by reinforcement-imitation learning (NÃ©da Navidi, 2021) View paper
  - [48] Unveiling the Role of Expert Guidance: A Comparative Analysis of User-centered Imitation Learning and Traditional Reinforcement Learning (Gomaa, 2024) View paper
  - [49] Leveraging Human Knowledge in Imitation Learning (Das, 2025) View paper
- Offline Preference Learning and Trajectory Generation
  - Offline Trajectory Preference Optimization (1 papers)
  - [22] Flow to better: Offline preference-based reinforcement learning via preferred trajectory generation (Z Zhang, 2023) View paper
  - Safe Reward Inference from Preferences (1 papers)
  - [7] Safe imitation learning via fast bayesian reward inference from preferences (Daniel S. Brown, 2020) View paper
  - Regularized Offline RL with Demonstrations (1 papers)
  - [37] Regularized rl (D Tiapkin, 2023) View paper
- Evaluation Frameworks and Benchmarking
  - Online Evaluation Budget Analysis (1 papers)
  - [38] Showing Your Offline Reinforcement Learning Work: Online Evaluation Budget Matters (Kurenkov, 2021) View paper
  - AI-Based Feedback and Oracle Systems (1 papers)
  - [36] Oracle-RLAIF: An Improved Fine-Tuning Framework for Multi-modal Video Models through Reinforcement Learning from Ranking Feedback (Glatt, 2025) View paper

- Scalable and Lifelong Learning Paradigms
  - Lifelong Imitation Learning for Robot Fleets (1 papers)
  - [45] Scalable Lifelong Imitation Learning for Robot Fleets (Hoque, 2024) View paper
  - Human-in-the-Loop Robotics Frameworks (1 papers)
  - [42] Human-in-the-Loop Robotics: Enhancing Safety and Adaptability through Interactive AI Systems (Ali, 2025) View paper

Narrative

Core task: fine-tuning imitation learning policies with preference-based reinforcement learning. This field addresses how agents can move beyond simple behavioral cloning by incorporating human or oracle preferences to refine learned policies. The taxonomy reveals a rich landscape organized around several complementary themes. Offline-to-Online Policy Refinement Frameworks explore how to transition from static demonstration data to active policy improvement, often blending imitation with online exploration. Preference-Based Reward Learning and Policy Optimization focuses on extracting reward signals from comparative feedback, with methods ranging from classical inverse RL approaches like Scalable Inverse RL[5] and Bayesian Reward Inference[7] to modern direct optimization techniques such as Direct Preference Optimization[25]. Imitation Learning with Preference Signals and Large Language Model Alignment via Preference Learning capture the integration of preference data into both robotic control and language model fine-tuning, while Domain-Specific Applications and Interactive Human-in-the-Loop Learning Systems highlight practical deployments in robotics, autonomous vehicles, and interactive settings. Evaluation Frameworks and Benchmarking provide the infrastructure for measuring progress, and Scalable and Lifelong Learning Paradigms address long-term adaptation challenges.

Several active lines of work reveal key trade-offs and open questions. One central tension is between offline methods that learn purely from logged data—such as Unlabeled Preference Data[19] and Offline Evaluation Budget[38]—and online or interactive approaches like Active RLHF[20] and Human-in-the-Loop Robotics[42], which can query for new preferences but incur higher annotation costs. Another contrast appears between model-based reward learning, where a reward function is explicitly estimated from preferences, and model-free direct policy optimization methods like Direct Preference Optimization[25] that bypass reward modeling. Behavioral Cloning Preference[0] sits within the Unified Offline-Online Integration branch, emphasizing seamless transitions from imitation to preference-driven refinement. It shares this branch with neighbors like Diffusion Human Preference[8] and TakeAD[47], which similarly aim to unify offline demonstration data with preference signals. Compared to purely offline methods or those requiring extensive online interaction, Behavioral Cloning Preference[0] occupies a middle ground, leveraging initial imitation policies while incorporating preference feedback to guide further refinement without fully committing to either extreme.

Related Works in Same Category

The following **2 sibling papers** share the same taxonomy leaf node with the original paper:

1. FDPP: Fine-Tune Diffusion Policy with Human Preference

Authors: Yuxin Chen, Devesh K. Jha, Masayoshi Tomizuka, Diego Romeres | **Year/Venue**: 2025 • IEEE International Conference on Robotics and Automation | **URL**: View paper

Abstract

Imitation learning from human demonstrations enables robots to perform complex manipulation tasks and has recently witnessed huge success. However, these techniques often struggle to adapt behavior to new preferences or changes in the environment. To address these limitations, we propose Fine-tuning Diffusion Policy with Human Preference (FDPP). FDPP learns a reward function through preference-based learning. This reward is then used to fine-tune the pre-trained policy with reinforcement learnin...

Relationship Analysis

Both papers belong to the Unified Offline-Online Integration category, combining demonstration-based pre-training with preference-based online fine-tuning. They overlap in using imitation learning (behavioral cloning/diffusion policies) as initialization followed by preference-based RL refinement. The key difference is that the original paper provides theoretical regret bounds for this paradigm with explicit offline-to-online sample complexity analysis, while FDPP focuses on practical implementation for robotic manipulation tasks using diffusion policies with KL-regularized fine-tuning, without theoretical guarantees.

2. TakeAD: Preference-Based Post-Optimization for End-to-End Autonomous Driving With Expert Takeover Data

Authors: Deqing Liu, Yinfeng Gao, Deheng Qian, Qichao Zhang, Xiaoqing Ye, et al. (12 authors total) | **Year/Venue**: 2025 • IEEE Robotics and Automation Letters | **URL**: View paper

Abstract

Existing end-to-end autonomous driving methods typically rely on imitation learning (IL) but face a key challenge: the misalignment between open-loop training and closed-loop deployment. This misalignment often triggers driver-initiated takeovers and system disengagements during closed-loop execution. How to leverage those expert takeover data from disengagement scenarios and effectively expand the IL policy's capability presents a valuable yet unexplored challenge. In this letter, we propose Ta...

Relationship Analysis

Both papers belong to the Unified Offline-Online Integration category, combining demonstration-based pre-training with preference-based online fine-tuning. While the original paper (BRIDGE) provides a theoretical framework for fine-tuning behavioral cloning policies with preference-based RL using human feedback in general RL settings, TakeAD focuses specifically on autonomous driving applications, using expert takeover data from disengagement scenarios and combining DAgger with Direct Preference Optimization (DPO) rather than the uncertainty-weighted objective approach of BRIDGE. The key difference is that BRIDGE offers theoretical regret bounds and a principled algorithmic framework, whereas TakeAD presents a domain-specific application with a practical data collection pipeline for autonomous driving systems.

Contributions Analysis

Overall novelty summary. The paper proposes a two-stage framework that first learns a safe initial policy from reward-free expert demonstrations, then fine-tunes it online using preference-based human feedback. It sits in the 'Unified Offline-Online Integration' leaf, which contains only three papers total, including this work. This represents a relatively sparse research direction within the broader taxonomy of 50 papers across 36 topics, suggesting the specific combination of principled offline-to-online integration with preference-based refinement remains underexplored compared to adjacent areas like pure preference learning or pure imitation learning.

The taxonomy reveals that neighboring research directions are substantially more populated. The parent branch 'Offline-to-Online Policy Refinement Frameworks' includes related leaves on reward model pre-training and personalization via action representations. Adjacent branches like 'Preference-Based Reward Learning and Policy Optimization' contain multiple subcategories with 2-3 papers each, focusing on direct preference optimization, reward modeling, and uncertainty-aware learning. The paper's position bridges these areas by combining demonstration-based initialization with preference-driven online refinement, distinguishing it from purely offline methods in 'Offline Preference Learning and Trajectory Generation' and fully online approaches in 'Interactive and Human-in-the-Loop Learning Systems'.

Among 20 candidates examined across three contributions, the analysis found limited prior work overlap. The first contribution (theoretical framework for offline-to-online preference learning) examined 9 candidates with none clearly refuting it. The second contribution (BRIDGE algorithm) examined only 1 candidate with no refutation. The third contribution (regret bound connecting offline data to online sample efficiency) examined 10 candidates, with 1 appearing to provide overlapping prior work. This suggests that within the limited search scope, the algorithmic and theoretical framework contributions appear relatively novel, though the regret bound analysis may have more substantial precedent in the examined literature.

Based on the top-20 semantic matches examined, the work appears to occupy a genuinely sparse intersection of offline imitation learning and online preference-based refinement. The limited number of sibling papers in its taxonomy leaf and the low refutation rate across contributions support this impression. However, the analysis explicitly covers only a narrow slice of potentially relevant literature, and the single refutable candidate for the regret bound suggests that deeper theoretical connections may exist in the broader offline-to-online RL literature not captured by this search.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: First theoretical framework for offline-to-online preference learning

Description: The authors establish the first principled theoretical framework for combining offline expert demonstrations with online preference-based reinforcement learning. They formalize this hybrid setting and derive regret bounds that explicitly connect the quantity of offline data to online sample efficiency.

This contribution was assessed against **9 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Efficient Online RL Fine Tuning with Offline Pre-trained Policy Only

URL: View paper

Brief Assessment

Offline Pretrained Finetuning[66] focuses on fine-tuning pre-trained policies with online RL using Q-functions, not preference-based learning. The candidate addresses a different technical problem (avoiding pessimistic Q-functions) rather than combining offline demonstrations with online preference feedback.

2. Accelerating Human Motion Imitation with Interactive Reinforcement Learning

URL: View paper

Brief Assessment

Interactive Motion Imitation[67] focuses on human-in-the-loop RL for humanoid motion control with human demonstrations and labeling, not on combining offline expert demonstrations with online preference-based RL or establishing theoretical regret bounds for this paradigm.

3. Reinforcement learning meets bioprocess control through behaviour cloning: Real-world deployment in an industrial photobioreactor

URL: View paper

Brief Assessment

Bioprocess Behaviour Cloning[62] focuses on pH regulation in photobioreactors using behavior cloning followed by online RL fine-tuning, not on preference-based reinforcement learning or theoretical frameworks for combining offline demonstrations with online preference feedback.

4. Online iterative reinforcement learning from human feedback with general preference model

URL: View paper

Brief Assessment

Online Iterative RLHF[63] focuses on general preference models without reward functions, while the original paper specifically addresses combining offline expert demonstrations with online preference-based RL in a reward-free setting with behavioral cloning initialization. The technical approaches and problem formulations differ substantially.

5. Dialogue Learning with Human Teaching and Feedback in End-to-End Trainable Task-Oriented Dialogue Systems

URL: View paper

Brief Assessment

Dialogue Human Feedback[68] focuses on task-oriented dialogue systems using supervised learning followed by imitation and reinforcement learning from human feedback. This is fundamentally different from the original paper's theoretical framework for combining offline expert demonstrations with online preference-based RL in general MDPs with regret bounds.

6. Offline to Online Learning for Real-Time Bandwidth Estimation

URL: View paper

Brief Assessment

Bandwidth Estimation Online[65] focuses on bandwidth estimation for video applications using imitation learning and behavioral cloning, not on preference-based reinforcement learning or theoretical frameworks for combining offline demonstrations with online preference feedback.

7. Reinforcement Learning in the Era of LLMs: What is Essential? What is needed? An RL Perspective on RLHF, Prompting, and Beyond

URL: View paper

Brief Assessment

RL Era LLMs[64] focuses on RLHF for LLMs, describing it as 'online inverse RL with offline demonstration data' but does not provide a theoretical framework with regret bounds for combining offline expert demonstrations with online preference-based RL in the general setting that the original paper addresses.

8. New approach in human-AI interaction by reinforcement-imitation learning

URL: View paper

Brief Assessment

Reinforcement-Imitation Interaction[44] focuses on combining imitation learning with traditional RL methods (SARSA, A3C) using teacher feedback, not on preference-based reinforcement learning or theoretical frameworks for offline-to-online learning with regret bounds.

9. Improving multimodal interactive agents with reinforcement learning from human feedback

URL: View paper

Brief Assessment

Multimodal RLHF[12] focuses on improving multimodal interactive agents using human feedback in embodied 3D environments, not on establishing theoretical frameworks for combining offline demonstrations with online preference learning or deriving regret bounds.

Contribution 2: BRIDGE algorithm with uncertainty-weighted objective

Description: The authors propose BRIDGE (Bounded Regret with Imitation Data and Guided Exploration), a novel algorithm that combines offline behavioral cloning with online preference-based learning through an uncertainty-weighted objective that constrains exploration to a confidence set constructed from offline data.

This contribution was assessed against **1 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Effective Reinforcement Learning with Information Reuse from Multiple Demonstrators

URL: View paper

Brief Assessment

Multiple Demonstrators Reuse[61] focuses on combining action advice from multiple demonstrations using contextual bandit insights and probabilistic policy reuse (FTSA) and actor-critic networks (TL-AC). This is fundamentally different from BRIDGE's uncertainty-weighted objective that integrates offline behavioral cloning with online preference-based learning through confidence sets constructed from Hellinger distance bounds.

Contribution 3: Regret bound showing offline data reduces online regret

Description: The authors prove that their algorithm achieves optimal square-root-T regret dependence while explicitly showing how the number of offline demonstrations n improves online performance. Their bound formally demonstrates that as offline data increases, online regret approaches zero.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Provably and practically efficient adversarial imitation learning with general function approximation

URL: View paper

Brief Assessment

Adversarial Imitation Function[59] focuses on adversarial imitation learning with general function approximation, not preference-based RL. The paper does not address offline demonstrations reducing online regret in preference learning contexts.

2. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning

URL: View paper

Prior Art Analysis

Policy Finetuning[53] demonstrates prior work establishing regret bounds that explicitly connect offline data quantity to online sample efficiency. The candidate paper proves that with n offline demonstrations, the sample complexity for finding an ε-optimal policy is $\tilde{O}(H^3SC/\varepsilon^2)$, where C is the concentrability coefficient. This bound formally shows how offline data reduces online learning complexity, with the regret vanishing as n→∞. The candidate's theoretical framework predates the original paper and establishes the fundamental relationship between offline demonstration quantity and online regret reduction.

Evidence

Evidence 1 - **Rationale**: Both papers establish regret bounds that explicitly connect offline data quantity to online performance. The candidate paper's bound $\tilde{O}(H^3SC/\varepsilon^2)$ shows how the concentrability coefficient C (which depends on the quality of offline data) directly affects sample complexity, similar to how the original paper's bound shows n (number of offline demonstrations) affects regret. - **Original**: we prove that our algorithm, bridge (algorithm 1) achieves an optimal √ tregret dependence on the online horizon t, while explicitly showing how offline demonstrations improve online performance (theorem 4.1). our bound formally shows that as the number of offline demonstrationsn→ ∞, the online regr... - **Candidate**: we first design a sharp offline reduction algorithm-which simply executes µand runs offline policy optimization on the collected dataset-that finds an εnear-optimal policy within ˜o(h3sc⋆/ε2) episodes, where c⋆ is the single-policy concentrability coefficient between µand π⋆. this offline result is the fir...

Evidence 2 - **Rationale**: Both papers establish theoretical frameworks for understanding how offline data (reference policy in the candidate, demonstrations in the original) improves online learning efficiency. The candidate paper's policy finetuning framework predates and establishes the theoretical foundation for connecting offline and online RL. - **Original**: we derive regret bounds that shrink with the number of offline demonstrations, explicitly connecting the quantity of offline data to online sample efficiency - **Candidate**: this paper initiates the theoretical study of policy finetuning, that is, online rl where the learner has additional access to a 'reference policy'µclose to the optimal policy π⋆ in a certain sense

Evidence 3 - **Rationale**: Both papers provide explicit mathematical relationships between offline data quality/quantity and online learning efficiency. The candidate uses concentrability coefficient C while the original uses radius O(1/√n), but both formalize how offline data reduces online complexity. - Original: our framework (theorem 4.2) uses the hellinger distance between trajectory distributions to construct confidence sets whose radii,o(1/ √n), directly connect the quantity of offline datanto online learning efficiency - Candidate*: we first design a sharp offline reduction algorithm-which simply executes µand runs offline policy optimization on the collected dataset-that finds an εnear-optimal policy within ˜o(h3sc⋆/ε2) episodes, where c⋆ is the single-policy concentrability coefficient

3. Hybrid rl: Using both offline and online data can make rl efficient

URL: View paper

Brief Assessment

Hybrid RL[52] focuses on Q-learning algorithms with bilinear rank assumptions and does not address preference-based RL or demonstrate how offline demonstration quantity relates to online preference learning regret bounds.

4. Leveraging (biased) information: Multi-armed bandits with offline data

URL: View paper

Brief Assessment

Bandits Offline Data[56] addresses multi-armed bandits with offline data, not the preference-based RL with behavioral cloning setting. The technical frameworks differ fundamentally: bandits vs. MDPs with preference feedback.

5. Online Decisions with (Biased) Offline Data

URL: View paper

Brief Assessment

Online Biased Data[55] addresses a different problem setting involving biased offline data in online decision-making, rather than the reward-free expert demonstrations combined with preference-based feedback studied in the original paper.

6. Selective sampling and imitation learning via online regression

URL: View paper

Brief Assessment

Selective Sampling Imitation[58] focuses on online active learning with expert queries in classification/imitation settings, not on offline-to-online RL with preference feedback. The candidate's regret bounds relate offline demonstration quantity to query complexity in a fundamentally different problem setting (noisy expert labels vs. preference-based feedback).

7. Contextual Online Pricing with (Biased) Offline Data

URL: View paper

Brief Assessment

Contextual Pricing Offline[54] addresses contextual pricing with biased offline data, not reward-free expert demonstrations with preference-based feedback. The settings and problem formulations are fundamentally different.

8. Regret minimization in Linear Bandits with offline data via extended D-optimal exploration

URL: View paper

Brief Assessment

D-Optimal Exploration[60] addresses linear bandits with offline data, while the original paper focuses on preference-based RL with behavioral cloning. The technical settings and problem formulations are fundamentally different.

9. Leveraging demonstrations to improve online learning: Quality matters

URL: View paper

Brief Assessment

Demonstration Quality Matters[57] focuses on multi-armed bandits with expert demonstrations, not general RL with preference-based feedback. Their regret bound depends on expert competence parameters (β, λ) rather than demonstration quantity n in the same framework.

10. Balancing optimism and pessimism in offline-to-online learning

URL: View paper

Brief Assessment

Optimism Pessimism Balance[51] addresses multi-armed bandit problems with a focus on balancing pessimistic (LCB) and optimistic (UCB) strategies over time, rather than analyzing how offline demonstration quantity affects online sample efficiency in preference-based RL with behavioral cloning initialization.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] Fine-tuning Behavioral Cloning Policies with Preference-Based Reinforcement Learning View paper
- [1] Residual Reward Models for Preference-based Reinforcement Learning View paper
- [2] Deep reinforcement learning from human preferences for ROV path tracking View paper
- [3] Imitation learning from vague feedback View paper
- [4] Personalization in human-robot interaction through preference-based action representation learning View paper
- [5] Imitating Language via Scalable Inverse Reinforcement Learning View paper
- [6] Personalizing reinforcement learning from human feedback with variational preference learning View paper
- [7] Safe imitation learning via fast bayesian reward inference from preferences View paper
- [8] FDPP: Fine-Tune Diffusion Policy with Human Preference View paper
- [9] Human Implicit Preference-Based Policy Fine-tuning for Multi-Agent Reinforcement Learning in USV Swarm View paper
- [10] Contextual bandits and imitation learning with preference-based active queries View paper
- [11] Limited preference aided imitation learning from imperfect demonstrations View paper
- [12] Improving multimodal interactive agents with reinforcement learning from human feedback View paper
- [13] Deformpam: Data-Efficient Learning for Long-Horizon Deformable Object Manipulation Via Preference-Based Action Alignment View paper
- [14] Learning state importance for preference-based reinforcement learning View paper
- [15] An Extremely Data-efficient and Generative LLM-based Reinforcement Learning Agent for Recommenders View paper
- [16] On a Connection Between Imitation Learning and RLHF View paper
- [17] Approximated Variational Bayesian Inverse Reinforcement Learning for Large Language Model Alignment View paper
- [18] Reward learning from human preferences and demonstrations in atari View paper
- [19] Exploiting Unlabeled Data for Feedback Efficient Human Preference based Reinforcement Learning View paper
- [20] Active RLHF via Best Policy Learning from Trajectory Preference Feedback View paper
- [21] Meta-reward-net: Implicitly differentiable reward learning for preference-based reinforcement learning View paper
- [22] Flow to better: Offline preference-based reinforcement learning via preferred trajectory generation View paper
- [23] A ranking game for imitation learning View paper
- [24] Aligning diffusion behaviors with q-functions for efficient continuous control View paper
- [25] Direct preference-based policy optimization without reward modeling View paper
- [26] Learning a universal human prior for dexterous manipulation from human preference View paper
- [27] Adaptive preference scaling for reinforcement learning with human feedback View paper
- [28] Adaptive Learning based Upper-Limb Rehabilitation Training System with Collaborative Robot View paper
- [29] Provable Reward-Agnostic Preference-Based Reinforcement Learning View paper

- [30] STARE-VLA: Progressive Stage-Aware Reinforcement for Fine-Tuning Vision-Language-Action Models View paper
- [31] Uncertainty-aware preference alignment in reinforcement learning from human feedback View paper
- [32] Adversarial Imitation Learning with Preferences View paper
- [33] Towards Abstractive Timeline Summarisation using Preference-based Reinforcement Learning View paper
- [34] An online trajectory guidance framework via imitation learning and interactive feedback in robot-assisted surgery. View paper
- [35] Predictive Preference Learning from Human Interventions View paper
- [36] Oracle-RLAIF: An Improved Fine-Tuning Framework for Multi-modal Video Models through Reinforcement Learning from Ranking Feedback View paper
- [37] Regularized rl View paper
- [38] Showing Your Offline Reinforcement Learning Work: Online Evaluation Budget Matters View paper
- [39] Learning Real-World Acrobatic Flight from Human Preferences View paper
- [40] VITAL: Interactive Few-Shot Imitation Learning via Visual Human-in-the-Loop Corrections View paper
- [41] Enhancing the Controllability of Visual Navigation Agents with Language-Conditioned Preferences View paper
- [42] Human-in-the-Loop Robotics: Enhancing Safety and Adaptability through Interactive AI Systems View paper
- [43] Structured Preference Modeling for Reinforcement Learning-Based Fine-Tuning of Large Models View paper
- [44] New approach in human-AI interaction by reinforcement-imitation learning View paper
- [45] Scalable Lifelong Imitation Learning for Robot Fleets View paper
- [46] A Human-Machine Reinforcement Learning Framework with Multi-dimensional Human Feedback Fusion View paper
- [47] TakeAD: Preference-Based Post-Optimization for End-to-End Autonomous Driving With Expert Takeover Data View paper
- [48] Unveiling the Role of Expert Guidance: A Comparative Analysis of User-centered Imitation Learning and Traditional Reinforcement Learning View paper
- [49] Leveraging Human Knowledge in Imitation Learning View paper
- [50] Learning to Discern: Imitating Heterogeneous Human Demonstrations with Preference and Representation Learning View paper
- [51] Balancing optimism and pessimism in offline-to-online learning View paper
- [52] Hybrid rl: Using both offline and online data can make rl efficient View paper
- [53] Policy finetuning: Bridging sample-efficient offline and online reinforcement learning View paper
- [54] Contextual Online Pricing with (Biased) Offline Data View paper
- [55] Online Decisions with (Biased) Offline Data View paper
- [56] Leveraging (biased) information: Multi-armed bandits with offline data View paper
- [57] Leveraging demonstrations to improve online learning: Quality matters View paper
- [58] Selective sampling and imitation learning via online regression View paper
- [59] Provably and practically efficient adversarial imitation learning with general function approximation View paper
- [60] Regret minimization in Linear Bandits with offline data via extended D-optimal exploration View paper
- [61] Effective Reinforcement Learning with Information Reuse from Multiple Demonstrators View paper
- [62] Reinforcement learning meets bioprocess control through behaviour cloning: Real-world deployment in an industrial photobioreactor View paper
- [63] Online iterative reinforcement learning from human feedback with general preference model View paper
- [64] Reinforcement Learning in the Era of LLMs: What is Essential? What is needed? An RL Perspective on RLHF, Prompting, and Beyond View paper
- [65] Offline to Online Learning for Real-Time Bandwidth Estimation View paper
- [66] Efficient Online RL Fine Tuning with Offline Pre-trained Policy Only View paper
- [67] Accelerating Human Motion Imitation with Interactive Reinforcement Learning View paper
- [68] Dialogue Learning with Human Teaching and Feedback in End-to-End Trainable Task-Oriented Dialogue Systems View paper