MisoDICE: Multi-Agent Imitation from Unlabeled Mixed-Quality Demonstrations

- **The Viet Bui**, *Singapore Management University*, Singapore
- **Tien Anh Mai**, *Singapore Management University*, Singapore
- **Thanh Hong Nguyen**, *University of Oregon*, USA

NEURAL INFORMATION PROCESSING SYSTEMS

Introduction

- **The Problem:** Obtaining high-quality expert data in multi-agent environments is expensive and impractical due to complex joint state-action spaces. Real-world data is often **unlabeled and mixed-quality**, containing both expert and sub-optimal trajectories.
- **The Challenge:** Existing methods assume access to expert labels or high-quality demonstrations. Learning from unlabeled mixed data requires distinguishing expert behaviors without ground-truth rewards.

Preliminaries

Setting: Cooperative MARL modeled as a POMDP:

$$M = \langle S, A, P, r, Z, O, n, N, \gamma \rangle$$

Data: We operate on an Unlabeled Dataset ($\mathcal{D}_{unlabeled}$) containing a mix of expert and non-expert trajectories.

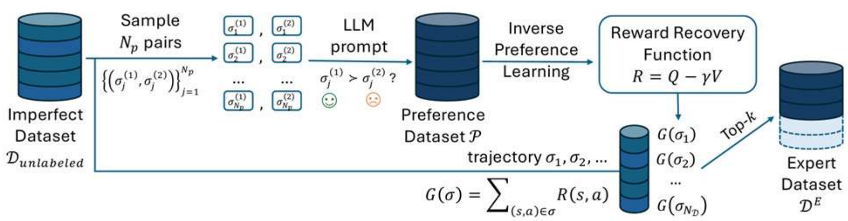
Goal: Recover optimal local policies π_i without access to the ground-truth reward function, maximizing the expected return based on the inferred expert distribution.

Phase 1: Expert Identification (Data Labeling)

Step 1 (LLM Preferences): Sample trajectory pairs and use an LLM (e.g., GPT-4o) to generate preference labels based on semantic game features (health, position).

Step 2 (Reward Recovery): Train O-MAPL (Preference-based MARL) on these labels to learn a soft Q-function. Recover rewards via $R \approx Q - \gamma V$.

Step 3 (Ranking): Rank trajectories by total recovered return. The top-k are selected as the Expert Dataset ($\mathcal{D}^E$); the rest form the suboptimal set ($\mathcal{D}^{Mix}$).



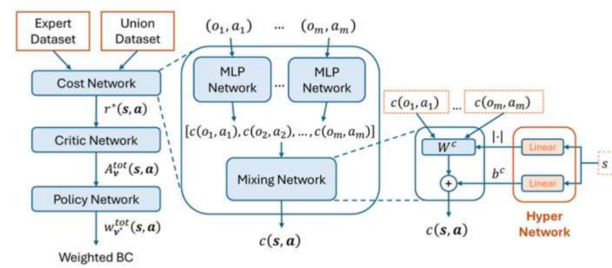
Phase 2: MisoDICE Algorithm

Objective: Minimize the divergence between the learned policy and the expert distribution, regularized by the union (mixed) distribution:

$$\max_{\pi_{tot}} D_{KL}(\rho_{tot}^{\pi} || \rho_{tot}^{E}) + \alpha D_{KL}(\rho_{tot}^{\pi} || \rho_{tot}^{U})$$

- ρ_{tot}^E: Expert distribution (from Phase 1).
- ρ_{tot}^U: Union distribution (Expert + Mixed).
- α: Hyperparameter controlling influence of suboptimal data.

Optimization: We reformulate this as a convex optimization problem over stationary distributions using the DICE framework.



Experiments & Results

	BC ($\beta = 0.0$)	BC ($\beta = 0.5$)	BC ($\beta = 1.0$)	OMAPL	INDD	MARL-SL	VDN	MisoDICE (ours)
2c_vs_64zg	8.5 ± 0.1	9.7 ± 0.3	12.6 ± 0.3	12.2 ± 0.4	14.6 ± 1.0	14.0 ± 1.6	12.7 ± 0.6	$\mathbf{16.4 \pm 1.3}$
5m_vs_6m	5.0 ± 1.1	6.7 ± 0.0	6.1 ± 0.1	5.7 ± 0.2	6.7 ± 0.1	6.8 ± 0.1	6.2 ± 1.4	$\mathbf{7.3 \pm 0.1}$
6h_vs_8z	7.0 ± 0.0	7.4 ± 0.0	7.2 ± 0.1	6.6 ± 0.2	7.5 ± 0.2	7.8 ± 0.1	8.2 ± 0.2	$\mathbf{8.7 \pm 0.2}$
corridor	1.5 ± 0.1	1.5 ± 0.2	4.3 ± 0.7	2.2 ± 1.3	4.4 ± 1.2	1.8 ± 0.2	4.7 ± 0.6	$\mathbf{5.8 \pm 0.8}$
Protoss 5_vs_5	9.2 ± 0.1	11.7 ± 0.5	10.2 ± 0.5	9.6 ± 1.1	10.9 ± 0.1	11.6 ± 0.3	11.5 ± 0.2	$\mathbf{12.4 \pm 0.5}$
Protoss 10_vs_10	10.3 ± 0.6	11.8 ± 0.5	10.6 ± 0.2	10.1 ± 0.9	11.0 ± 0.7	11.9 ± 0.4	12.4 ± 0.2	$\mathbf{12.9 \pm 0.2}$
Protoss 10_vs_11	8.2 ± 0.4	9.6 ± 0.4	8.7 ± 0.3	8.5 ± 1.2	9.4 ± 0.4	9.9 ± 0.3	10.4 ± 0.1	$\mathbf{10.7 \pm 0.4}$
Protoss 20_vs_20	10.1 ± 0.2	10.4 ± 0.5	10.5 ± 0.3	9.4 ± 0.4	11.4 ± 0.5	13.1 ± 0.4	12.1 ± 0.5	$\mathbf{13.5 \pm 0.5}$
Protoss 20_vs_23	8.1 ± 0.2	8.6 ± 0.3	8.3 ± 0.2	7.9 ± 0.3	9.6 ± 0.3	9.6 ± 0.3	10.3 ± 0.4	$\mathbf{10.6 \pm 0.2}$

Conclusion: MisoDICE significantly outperforms baselines by effectively leveraging unlabeled mixed-quality data, confirming the benefits of the two-stage labeling and convex value-decomposition approach.

Value Factorization & consistency

Value Factorization: To handle the combinatorial action space, we decompose the global value function using a linear mixing network to preserve convexity:

$$v^{tot}(s) = \mathcal{M}_\phi[\{v_i(s_i)\}] = \sum \phi_i v_i(s_i) + \phi_0$$

Note: Non-linear mixing (like ReLU networks) destroys the convexity of the DICE objective, leading to instability.

Occupancy Ratio Estimation: We estimate the density ratio $w(s,a)$ using a discriminator trained with a linear mixing network to ensure concavity.

Policy Extraction (Global-Local Consistency)

The optimal local policies are recovered via *Weighted Behavior Cloning (WBC)*, ensuring that maximizing local objectives results in the global optimum:

$$\pi_i^*(a_i|s_i) \propto \exp\left(\frac{\phi_i^*}{1+\alpha} q_i^*(s_i, a_i) + \log \mu_i^U(a_i|s_i)\right)$$