Learning to Trade with Preferences: Interpretable Execution via Mixture-of-Experts Direct Preference Optimization

Haohan Xu[1], Jason Bohne[1,3], Paweł Polak[1], David Byrd[2], David Rosenberg[3], Gary Kazantsev[3]

[1]Stony Brook University [2]Bowdoin College [3]Bloomberg

Broker Execution Strategies & Two-Stage MoE-DPO Framework

Four Canonical Broker Strategies: **TWAP** (uniform execution), **VWAP** (follows market volume), **IS** (early execution), **POV** (adapts to real-time volume).

Key Limitations: Inflexible to changing market microstructure; brokers manually switch between strategies.

Our Solution: Reinforcement learning framework that constructs adaptive, interpretable policies as state-dependent mixtures of these deterministic strategies.

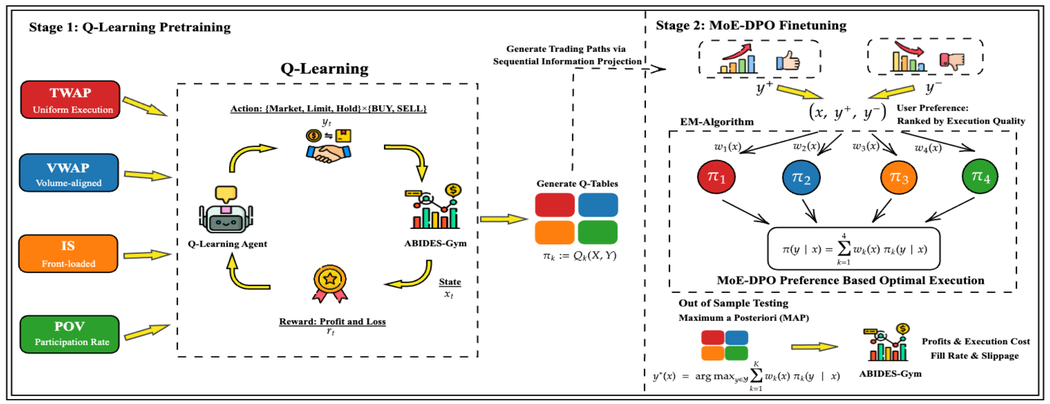

Figure 1: Integration of $\mathcal{Q}$-learning pretraining and MoE-DPO fine-tuning within ABIDES-Gym environment

$\mathcal{Q}$-learned rule-based strategies (TWAP, VWAP, IS, POV) are first integrated via Sequential Information Projection (SIP), which adaptively weights experts based on execution quality. **MoE-DPO** then fine-tunes this SIP mixture through preference-based optimization, producing interpretable and adaptive execution policies.

$\mathcal{Q}$-Learning Pretraining & MoE-DPO Fine-tuning

We conduct experiments in **ABIDES-Gym** with 6.5-hour trading sessions (9:30 AM - 4:00 PM) and disparate **background agents** (market makers, value agents, noise traders, momentum traders, volatility agents).

Stage 1: $\mathcal{Q}$-Learning Pretraining:

We update $\mathcal{Q}$-values using the standard update rule:

$$Q(x_t, y_t) \leftarrow (1-\alpha)Q(x_t, y_t) + \alpha \left[r_{t+1} + \gamma \max_{y' \in Y} Q(x_{t+1}, y') \right]$$

Stage2: MoE-DPO Fine-tuning:

Preference Learning: We construct trajectory-level comparisons (x, y^+, y^-) where $y^+ \succ y^-$ based on execution **quality** or **speed** criteria.

Stochastic EM: Alternates between computing expert responsibilities $q_k(x, y^+, y^-)$ using the **Mixture-of-Bradley-Terry (MBT)** model and updating expert policies and gating weights.

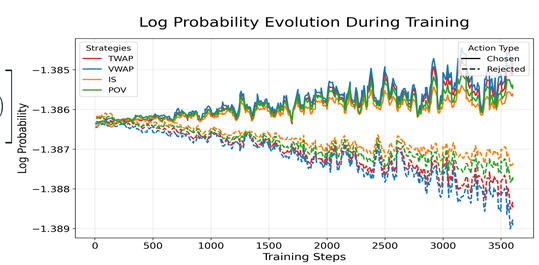

Figure 2: Policy probability dynamics during MoE-DPO training.

The log probabilities of preferred actions rise while rejected actions decline across all experts. This consistent separation, emerging after 500 steps, validates the MBT preference optimization built on SIP-generated trajectories.

Out-of-Sample Performance Across Different Market Regimes

Results compare four $\mathcal{Q}$**-learned baselines**, the **SIP mixture**, and **MoE-DPO** variants. **MoE-DPO Execution Quality** delivers the highest profits and lowest costs, while **Execution Speed** trades for faster completion, and SIP MIX provides a strong adaptive benchmark.

Policy	Baseline	Open/Close	Mid-Day	Bullish	Bearish	Liquidity
Profit ± Standard Error						
TWAP	60.50±1.20	5.53±0.28	23.73±0.23	73.65±1.40	54.69±1.09	21.63±0.21
VWAP	61.45±1.06	5.83±0.29	23.69±0.26	70.53±1.41	52.69±1.09	21.41±0.21
IS	58.64±1.12	6.00±0.31	23.92±0.24	71.61±1.25	53.72±1.01	21.44±0.18
POV	59.94±1.01	5.60±0.27	23.38±0.25	72.49±1.19	53.47±1.09	21.35±0.20
SIP Mix	61.75±1.08	6.15±0.32	24.25±0.23	73.70±1.36	55.42±1.00	21.90±0.18
MoE-DPO: Execution Quality	**67.59**±1.31	**6.21**±0.29	**26.23**±0.29	**76.73**±1.25	**57.86**±1.12	**23.10**±0.18
MoE-DPO: Execution Speed	57.13±1.14	5.82±0.28	22.73±0.28	68.88±1.24	51.61±0.87	20.60±0.19
Execution Cost ± Standard Error						
TWAP	1.490±0.005	0.386±0.005	0.800±0.002	1.507±0.005	1.502±0.005	2.097±0.004
VWAP	1.500±0.005	0.393±0.005	0.804±0.002	1.498±0.005	1.503±0.005	2.088±0.004
IS	1.482±0.006	0.391±0.006	0.805±0.002	1.504±0.005	1.500±0.005	2.089±0.005
POV	1.483±0.006	0.387±0.005	0.800±0.002	1.502±0.005	1.501±0.005	2.080±0.005
SIP Mix	1.453±0.005	0.381±0.005	0.791±0.002	1.457±0.005	1.454±0.005	2.017±0.005
MoE-DPO: Execution Quality	**1.445**±0.006	**0.375**±0.005	**0.784**±0.002	**1.445**±0.004	**1.440**±0.005	**1.989**±0.005
MoE-DPO: Execution Speed	1.519±0.006	0.398±0.005	0.827±0.003	1.515±0.006	1.515±0.006	2.117±0.005

Table 1: Out-of-sample performance across market regimes (100 trajectories)

Fill Rate Dynamics & Volatility Stress Testing

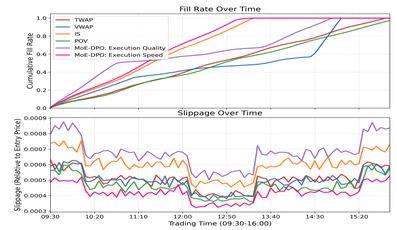

Figure 3: Out-of-sample evaluation of execution strategies.

Figure 4: Out-of-sample evaluation across volatility regimes.

Panel (a) shows cumulative fill rates, where MoE-DPO Execution Quality adapt execution dynamically compared to fixed baselines. Panel (b) reports slippage, with MoE-DPO Execution Quality maintaining competitive or lower levels, improving upon both deterministic and SIP mixtures.

Performance is compared across 100 trajectories under increasing volatility. MoE-DPO Execution Quality consistently outperforms deterministic baselines and the SIP mixture, preserving its profit and cost advantages even in **stressed markets**.

Conclusions

- Extended DPO to multi-expert execution domain with trajectory-level preferences
- Achieved consistent outperformance while maintaining interpretability through expert specialization
- Enabled preference-based customization (quality vs. speed) for institutional trading requirements

[1] Amrouni, Selim, et al. "ABIDES-gym: gym environments for multi-agent discrete event simulation and application to financial markets." Proceedings of the Second ACM International Conference on AI in Finance. 2021.
[2] Rafailov, Rafael, et al. "Direct preference optimization: Your language model is secretly a reward model." Advances in neural information processing systems 36 (2023): 53728-53741.