LEARNING STOCK TRADING POLICIES VIA BARYCENTER-BASED ADVERSARIAL INVERSE REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Designing effective trading strategies using reinforcement learning remains challenging due to delayed and noisy rewards, poor exploration, and the difficulty of enforcing explicit risk constraints. In this work, we propose BRaG, a barycenter-based adversarial inverse reinforcement learning framework for stock trading that learns trading behavior from multiple heterogeneous expert strategies. BRaG aggregates expert demonstrations using a performance-weighted Wasserstein barycenter, yielding a stable pseudo-expert representation that captures shared structure across diverse trading styles. This representation is used to pretrain a trading policy via adversarial imitation learning, which alleviates unstable exploration during reinforcement learning. The pretrained policy is subsequently refined using reinforcement learning with true market rewards. To ensure risk-aware decision-making, BRaG incorporates control barrier functions that constrain action execution and regularize policy learning to satisfy drawdown limits. We evaluate the proposed approach on four major global equity markets, including the US, UK, Indian, and Taiwanese indices. Across all the markets, the proposed approach achieves stronger performance than both classical trading rules and recent deep reinforcement learning methods, while exhibiting more stable risk characteristics.

1 INTRODUCTION

Financial markets play a central role in modern economies by facilitating capital allocation, risk transfer, and price discovery. Within this broad landscape, stock trading has attracted sustained research interest owing to its economic relevance and the availability of large-scale historical data. At the same time, stock prices exhibit high volatility, non-stationarity, and sensitivity to exogenous events, which makes consistent decision-making challenging (Wu et al., 2020). Traditional trading strategies based on fixed rules or handcrafted indicators often fail to adapt to evolving market conditions and may suffer from limited generalization. Supervised learning approaches have been used to predict future price movements or returns. However, their effectiveness is limited by label noise and the difficulty of translating predictions into profitable trading actions. As a result, there has been growing interest in learning frameworks that directly optimize sequential trading decisions through interaction with the market (Zhang et al., 2019; Tsantekidis et al., 2020).

Reinforcement learning (RL) has gained prominence as a natural framework for framing stock trading as a sequential decision-making problem. In a typical setting, an agent observes market states constructed from historical prices and related indicators, executes trading actions such as buy, sell, or hold, and receives feedback based on the resulting returns (Moody et al., 1998). In contrast to supervised approaches, RL focuses on maximizing cumulative performance over time rather than short-term predictive accuracy. With the integration of deep neural networks, deep reinforcement learning (DRL) methods have been successfully applied to a wide range of trading tasks, including single-asset trading, portfolio management, and intraday trading. Existing studies indicate that DRL agents are capable of capturing temporal structure in financial time series and adjusting their behavior as market regimes evolve, which makes this paradigm well-suited for automated trading applications (Sun et al., 2022; Kabbani & Duman, 2022).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Despite these advantages, deploying DRL in real-world stock trading remains non-trivial. One of the central challenges lies in designing reward functions that accurately capture desirable trading behavior over extended horizons. Market feedback is often sparse, delayed, and noisy, and reward signals based purely on short-term profits can lead to unstable learning dynamics. In addition, manually specified rewards typically embed implicit assumptions about risk tolerance and market behavior, which may not transfer well across assets or changing conditions (Rodinos et al., 2023; Orra et al., 2025). While alternative objectives such as risk-adjusted returns or drawdown-based penalties can be incorporated, they require careful tuning and still provide only partial proxies for the complex preferences underlying successful trading strategies. Inverse reinforcement learning (IRL) offers a principled alternative by seeking to recover an underlying reward function that explains observed expert trajectories, under the assumption that the expert behaves optimally or near-optimally (Ng et al., 2000). Recent adversarial formulations, such as Generative Adversarial Imitation Learning (GAIL), reformulate this problem as one of matching the distribution of expert and agent trajectories, enabling scalable policy learning without explicit reward specification (Ho & Ermon, 2016). Beyond reward recovery, this formulation also provides an effective pretraining mechanism, initializing the policy in a region of the policy space aligned with expert behavior and thereby mitigating the instability of exploration under sparse and noisy market rewards. In the context of stock trading, IRL provides a mechanism for capturing latent trading preferences reflected in expert decisions and for training agents whose behavior aligns more closely with expert-level strategies through interaction with the market (Sun et al., 2023; Roa-Vicens et al., 2019; Zhang et al., 2022).

Despite recent advances, important limitations remain in applying IRL to stock trading. Existing methods often rely on demonstrations from a single expert or a narrowly defined strategy, which is restrictive in markets characterized by heterogeneous and evolving trading behaviors. Such approaches are prone to overfitting and may fail to generalize across market regimes. Moreover, adversarial imitation learning methods are commonly applied without explicitly addressing the exploration challenges faced by reinforcement learning in noisy and high-dimensional financial environments. Policies trained from scratch frequently exhibit unstable learning dynamics due to sparse and delayed rewards (Liu et al., 2020; Halperin et al., 2022). In addition, most existing approaches either depend on manually specified reward functions or do not explicitly incorporate risk constraints during policy learning, limiting their ability to capture complex trading preferences and ensure stable performance. These limitations highlight a gap for a unified framework that can leverage diverse expert behaviors, avoid explicit reward specification, and incorporate risk awareness in a principled manner.

To address these challenges, we propose BRaG, a barycenter-based adversarial IRL framework for stock trading. BRaG aggregates multiple heterogeneous expert strategies using a performance-weighted Wasserstein barycenter, yielding a stable pseudo-expert distribution for imitation learning. This aggregated representation is used to pretrain a trading policy via adversarial imitation learning, which mitigates poor exploration by initializing the agent with expert-aligned behavior. The pretrained policy is subsequently refined using reinforcement learning with true market rewards. Explicit risk control is incorporated through control barrier functions that constrain action execution and regularize policy learning to ensure compliance with drawdown limits throughout training and deployment. We evaluate the proposed framework on four major global equity markets, including the US, UK, Indian, and Taiwanese stock indices. Empirical results demonstrate that BRaG consistently outperforms classical trading strategies and state-of-the-art DRL baselines in terms of both return and risk-adjusted performance. Although BRaG builds upon established techniques, its contribution lies in the principled integration of these techniques to address the practical challenges of multi-expert imitation learning for financial trading, including expert aggregation, stable policy initialization, adaptive policy refinement, and risk-aware optimization. The main contributions of this work are summarized as follows:

- We propose BRaG, a barycenter-based adversarial IRL framework for stock trading that aggregates heterogeneous expert strategies to construct a stable pseudo-expert representation for imitation learning.

- We introduce a performance-weighted Wasserstein barycenter to combine multiple expert trajectories, thereby reducing sensitivity to individual expert bias and enhancing robustness across diverse market conditions.

- We demonstrate how adversarial imitation learning can be used as an effective pretraining mechanism to address poor exploration in reinforcement learning for financial markets.
- We integrate explicit risk control through control barrier functions, enforcing drawdown constraints at action execution time and regularizing policy learning during training.
- We conduct extensive empirical evaluation across four global equity markets, showing that the proposed approach consistently outperforms classical trading strategies and state-of-the-art DRL baselines in terms of both return and risk-adjusted performance.

2 PRELIMINARIES

Reinforcement Learning. We model stock trading as a Markov Decision Process (MDP), where an agent observes market states, takes trading actions, and receives rewards based on portfolio performance. The objective is to learn a policy that maximizes expected cumulative return over time (Sutton & Barto, 1998). In this work, we adopt a policy gradient framework with actor–critic methods for stable optimization, specifically using Proximal Policy Optimization (PPO) (Schulman et al., 2017), with implementation details provided in Appendix A.

Inverse Reinforcement Learning. Inverse reinforcement learning (IRL) addresses the problem of inferring an underlying reward function from expert demonstrations, under the assumption that the expert behaves optimally or near-optimally. Rather than relying on manually specified objectives, IRL aims to recover latent preferences that explain observed behavior. This is particularly useful in financial domains, where reward design is challenging due to noisy, delayed, and non-stationary market feedback (Ng et al., 2000; Fu et al., 2017).

Adversarial Imitation Learning. Modern IRL approaches often adopt adversarial formulations that avoid explicit reward recovery. Generative Adversarial Imitation Learning (GAIL) casts policy learning as a distribution-matching problem between expert and agent trajectories. A discriminator is trained to distinguish expert samples from those generated by the policy, while the policy is optimized to produce behavior that is indistinguishable from expert demonstrations. The discriminator output serves as a surrogate reward signal, enabling scalable policy learning without manual reward specification. In our framework, this formulation is used to pretrain a trading policy that aligns with expert behavior before further refinement via reinforcement learning (Ho & Ermon, 2016).

3 PROBLEM FORMULATION

We study algorithmic stock trading as a sequential decision problem. At each trading time, an agent observes the market state, updates its portfolio, and receives feedback through profit and risk. This setting is naturally cast as an MDP, since the objective depends on a sequence of interlinked decisions and the reward is delayed and path-dependent. Formulating the stock trading problem as an MDP enables the use of RL techniques to optimize trading strategies dynamically over time.

Let $\mathcal{I} = \{1, \ldots, n\}$ be the list of tradable stocks and $t = \{0, 1, \ldots, T-1\}$ denote decision times. Then, for our multi-stock trading problem, the key components of the MDP are defined as follows:

State s_t: The state collects market features and portfolio information. For an $n-$asset portfolio, the state comprises the available cash $b_t \in \mathbb{R}_+$, current share holdings $h_t \in \mathbb{Z}^n$, closing price of each stock $p_t \in \mathbb{R}_+^n$, and eight most commonly used technical indicators corresponding to each stock. This formulation assigns a compact dimension of $10n + 1$ to the state s_t for n assets.

Action a_t: The action a_t represents the trading decision across all stocks at time t. It is modeled as an $n-$dimensional vector $a_t = [a_{t,1}, \ldots, a_{t,n}]^\top$, where each component $a_{t,i} \in \{-m, \ldots, 0, \ldots, m\}$ denotes the number of shares of stock i to be bought (positive), sold (negative), or held (zero). The parameter m specifies the upper limit on the number of shares that can be traded in a single period. Consequently, the overall action space has a cardinality of $(2m + 1)^n$.

Reward r_t: After executing the action a_t, the cash updates as

$$b_{t+1} = b_t - p_t^\top a_t - c(a_t, p_t),$$

where c is the fixed commission fee. The reward r_t is defined as the percentage change in the marked-to-market portfolio value, net of trading costs,

$$r_t = \frac{V_{t+1} - V_t}{V_t}, \qquad V_t = b_t + p_t^\top h_t.$$

This reward formulation encourages the agent to pursue profit-maximizing trades while concurrently penalizing excessive transactions, thereby balancing return and cost efficiency.

4 METHODOLOGY

We consider the problem of learning an expert-level stock trading policy in settings where the underlying reward function governing expert behavior is unknown and multiple heterogeneous expert strategies are available. To address this, we propose BRaG (Barycenter-based Risk-aware Adversarial GAIL), a multi-expert adversarial inverse reinforcement learning framework for stock trading. The central idea behind BRaG is to leverage the diversity of multiple expert strategies while avoiding overfitting to any single trading style. This is achieved by aggregating heterogeneous expert demonstrations through a performance-weighted Wasserstein barycenter, which yields a stable and representative pseudo-expert distribution. Policy learning is then carried out using adversarial imitation learning, enabling implicit reward recovery without explicit reward specification. Finally, explicit portfolio risk constraints are enforced through control barrier functions during action execution.

4.1 MULTI-EXPERT DEMONSTRATIONS AND OCCUPANCY MEASURES

We assume access to K expert trading strategies $\{\pi_k\}_{k=1}^K$, each reflecting a distinct trading philosophy. From each expert π_k, we observe a dataset of state-action pairs:

$$\mathcal{D}_k = \{(s_t^{(k)}, a_t^{(k)})\}_{t=1}^{T_k}. \tag{1}$$

Each expert induces a discounted occupancy measure:

$$\rho_{\pi_k}(s, a) = \sum_{t=0}^{\infty} \gamma^t \Pr(s_t = s, a_t = a \mid \pi_k), \tag{2}$$

which captures the long-term behavior of the expert. Rather than selecting a single expert or merging all demonstrations indiscriminately, BRaG seeks to construct a consensus representation that preserves common behavioral structure across experts.

4.2 PERFORMANCE-WEIGHTED BARYCENTER OF EXPERT TRAJECTORIES

To aggregate heterogeneous expert behaviors into a single representative expert distribution, we compute a performance-weighted sliced Wasserstein barycenter over expert state-action trajectories. Unlike uniform mixing or single-expert selection, the Wasserstein barycenter preserves the geometry of expert behaviors and provides a principled consensus representation. This yields robustness to expert bias while capturing shared structure across heterogeneous strategies.

Each expert policy π_k is assigned a non-negative weight α_k proportional to its risk-adjusted performance, measured using the Sharpe ratio:

$$\alpha_k = \frac{\tilde{S}_k}{\sum_l \tilde{S}_l}, \tag{3}$$

where $\tilde{S}_k = \max(\mathrm{Sharpe}(\pi_k), 0)$, the clipping at zero excludes experts with negative risk-adjusted performance from the aggregation. This weighting favors experts who exhibit superior return-to-risk characteristics. This Sharpe ratio-based weighting choice favors experts with high risk-adjusted performance rather than raw returns, so the barycenter is not skewed toward strategies that profited by taking on excessive risk.

Let μ_k denote the empirical distribution of concatenated state-action samples (s, a) generated by expert π_k. The objective is to compute a barycenter distribution μ_{bar} that minimizes the weighted sum of Wasserstein distances to individual expert distributions:

$$\mu_{\text{bar}} = \arg\min_{\mu} \sum_{k=1}^{K} \alpha_k \, W_2^2(\mu, \mu_k), \tag{4}$$

where W_2 denotes the 2-Wasserstein distance.

Computing Wasserstein barycenters directly in high-dimensional spaces is intractable. We therefore adopt a sliced Wasserstein approximation. Specifically, random projection directions $v \in \mathbb{R}^{d_s+d_a}$ with $\|v\|_2 = 1$ are sampled, and expert samples are projected onto one-dimensional subspaces. For each projection, the barycenter is obtained by weighted averaging of the empirical quantile functions:

$$\hat{z}_{\text{bar}}^{(v)}(q) = \sum_{k=1}^{K} \alpha_k \, \hat{z}_k^{(v)}(q), \tag{5}$$

where $\hat{z}_k^{(v)}$ denotes the sorted projected samples of expert k. The barycenter quantile $\hat{z}_{\text{bar}}^{(v)}(q)$ defines the projected barycenter distribution, from which pseudo-samples are obtained by sampling over quantile levels q. Since sliced Wasserstein barycenters do not admit a unique inverse mapping, we construct pseudo-expert samples by aggregating quantiles across multiple random projections.

The final barycenter samples are recovered by averaging over multiple random projections and back-projecting to the original space. The resulting dataset

$$\mathcal{D}_{\text{bar}} = \{(s_t^{\text{bar}}, a_t^{\text{bar}})\} \tag{6}$$

constitutes a set of pseudo-expert demonstrations. Intuitively, the barycenter aggregates expert trajectories in a structured manner rather than simple pooling. Expert datasets are first weighted by performance, and their state-action samples are projected onto multiple one-dimensional directions. For each projection, the barycenter is computed as a weighted average of the sorted projected samples (quantiles) across experts. Repeating this across projections yields pseudo-expert samples $\mathcal{D}_{\text{bar}}$, which capture shared structure while reducing sensitivity to individual strategies. Since the sliced Wasserstein back projection does not admit a unique inverse map, a reconstructed sample may not correspond to a trade that is actually reachable given the agent's cash and holdings at that step. We address this by passing every reconstructed sample through a feasibility check before it enters $\mathcal{D}_{\text{bar}}$. Each candidate action is compared against the available cash and current holdings, and is clipped to the nearest feasible point when a violation occurs.

4.3 Adversarial Inverse Reinforcement Learning via GAIL

BRaG employs Generative Adversarial Imitation Learning (GAIL) to recover an implicit reward function and a corresponding trading policy. GAIL frames imitation learning as a min-max game between a policy π_θ and a discriminator D_ϕ.

The discriminator $D_\phi : \mathcal{S} \times \mathcal{A} \to (0, 1)$ is trained to distinguish pseudo-expert samples from policy-generated samples by maximizing:

$$\mathbb{E}_{(s,a)\sim\rho_{\pi_\theta}}[\log D_\phi(s, a)] + \mathbb{E}_{(s,a)\sim\rho_{\text{bar}}}[\log(1 - D_\phi(s, a))]. \tag{7}$$

From the discriminator, an implicit reward signal is obtained as follows:

$$r_\phi(s, a) = -\log(D_\phi(s, a)). \tag{8}$$

This reward assigns higher values to state-action pairs that resemble barycenter expert behavior. Unlike classical inverse reinforcement learning, this formulation avoids explicit reward parameterization and instead directly matches occupancy measures.

The policy π_θ is optimized to maximize the expected discounted sum of discriminator-derived rewards:

$$\max_{\theta} \; \mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{\infty} \gamma^t r_\phi(s_t, a_t)\right]. \tag{9}$$

In BRaG, policy optimization is performed using Proximal Policy Optimization (PPO) (Schulman et al., 2017), which ensures stable updates through clipped likelihood ratios and value function regularization. This adversarial learning process iteratively aligns the policy distribution with the barycenter expert distribution.

4.4 Risk-Constrained Action Selection using Control Barrier Functions

Financial risk is explicitly controlled using a control barrier function (CBF) that enforces a maximum drawdown constraint. Let V_t denote the portfolio value at time t, and let $V_t^{\max} = \max_{\tau \le t} V_\tau$ denote the running maximum portfolio value. The drawdown at time t is defined as

$$\mathrm{DD}_t = 1 - \frac{V_t}{V_t^{\max}}. \tag{10}$$

We define the barrier function:

$$B(V_t) = \delta_{\max} - \mathrm{DD}_t, \tag{11}$$

where $\delta_{\max} \in (0, 1)$ specifies the maximum allowable drawdown. Safety is ensured when $B(V_t) \ge 0$.

To account for uncertainty, we estimate a conservative lower-bound return vector using a rolling window:

$$r_t^{\mathrm{lb}} = \mu_t - \kappa \sigma_t, \tag{12}$$

where $\mu_t, \sigma_t \in \mathbb{R}^n$ denote the rolling mean and standard deviation of asset returns up to time $t-1$, and $\kappa > 0$ controls risk sensitivity. Using this estimate, the next portfolio value V_{t+1} is predicted conservatively as

$$V_{t+1} = b_t - p_t^\top a_t - c(a_t, p_t) + \hat{p}_{t+1}^\top (h_t + a_t), \tag{13}$$

where

$$\hat{p}_{t+1} = p_t \odot (1 + r_t^{\mathrm{lb}}). \tag{14}$$

The CBF condition is enforced via:

$$B(V_{t+1}) - B(V_t) + \alpha_{\mathbf{CBF}} B(V_t) \ge 0, \tag{15}$$

where $\alpha_{\mathbf{CBF}} > 0$ determines the rate at which the system is driven back towards the safe set.

If the condition is violated, the policy action a_t is projected onto the safe set by solving

$$\begin{aligned} a_t^{\mathrm{safe}} = \; & \arg\min_{a \in \mathcal{A}} \|a - a_t\|_2^2 \\ & \text{s.t.} \;\; B(V_{t+1}) - B(V_t) + \alpha_{\mathbf{CBF}} B(V_t) \ge 0. \end{aligned} \tag{16}$$

In addition to action-level safety enforcement, we also incorporate a soft penalty term in the PPO objective to discourage repeated violations of the barrier condition during training. The combination of hard constraint enforcement at execution time and soft regularization during optimization improves training stability and reduces the need for frequent corrections.

4.5 Training Procedure

Training in BRaG proceeds in a staged manner to ensure stability and effective knowledge transfer from expert demonstrations to the final trading policy.

In the first stage, the policy is initialized through adversarial imitation learning using the barycenter pseudo-expert demonstrations. A discriminator is updated to differentiate between policy-generated state-action pairs and barycenter expert samples. The output of the discriminator is used to compute an imitation-based reward signal, which replaces the environment reward during this phase. Policy updates are performed using PPO, allowing for stable improvement under the adversarial reward. Throughout this stage, every action proposed by the policy is filtered through the CBF before execution, ensuring that drawdown constraints are respected during data collection itself. This pretraining stage enables the policy to internalize expert-like trading behavior while operating within explicit risk limits.

In the second stage, the pretrained policy is further refined using reinforcement learning with the true environment reward. This phase enables the policy to adapt to market-specific dynamics

Algorithm 1 BRaG: Barycenter-based Risk-aware Adversarial GAIL

Require: Expert datasets $\{\mathcal{D}_k\}_{k=1}^{K}$, parameters $\gamma, \delta_{\max}, \kappa, \alpha$
Ensure: Trained policy π_θ
Compute expert weights $\alpha_k \propto \mathrm{Sharpe}(\pi_k)$
Compute sliced Wasserstein barycenter $\mathcal{D}_{\mathrm{bar}}$
Initialize policy π_θ and discriminator D_ϕ
while GAIL not converged **do**
    Collect rollouts τ_π using π_θ
    Apply CBF-based action projection during rollouts
    Update D_ϕ using $\mathcal{D}_{\mathrm{bar}}$ and τ_π
    Set $r_\phi(s,a) = -\log(D_\phi(s,a))$
    Update π_θ using PPO with reward r_ϕ
end while
while fine-tuning not converged **do**
    Roll out π_θ with true environment reward
    Enforce CBF constraints during action execution
    Update π_θ using PPO
end while
return π_θ

and improve realized performance beyond imitation. The optimization objective is now aligned with portfolio returns, while the CBF continues to regulate action execution. By retaining the barrier mechanism during fine-tuning, risk constraints remain active even as the policy explores new behaviors. Model selection is based on validation performance, and the final policy is evaluated on an unseen trading period to assess generalization. The complete training procedure of BRaG is summarized in Algorithm 1.

5 Experimental Evaluation

This section reports the experimental study carried out in this work. We first describe the market datasets used in the experiments and the corresponding experimental settings. The evaluation criteria and baseline methods are then introduced. The section concludes with the empirical results, along with a discussion of the observed performance across datasets.

5.1 Data Description

The proposed methodology is evaluated using daily stock price data from four major global equity indices, representing diverse geographical regions and market structures. These include the Dow Jones Industrial Average (DJI) from the United States, the Financial Times Stock Exchange 100 (FTSE 100) from the United Kingdom, the BSE Sensex from India, and the Taiwan Capitalization Weighted Stock Index (TAIEX or TWII) from Taiwan. To maintain consistency across experiments, a fixed universe of thirty stocks is considered for each index. All constituent stocks of the DJI and Sensex are included. For the FTSE 100 and TAIEX, the thirty stocks with the highest market capitalization are selected. Daily closing price data for all selected stocks is obtained from Yahoo Finance[1] and spans the period from January 1, 2010, to September 30, 2025. These prices are used to compute technical indicators that form the state representation for the trading agent. The dataset for each index is divided into non-overlapping temporal segments to support training, validation, and out-of-sample evaluation. Data spanning January 1, 2010, to December 31, 2020, is used for GAIL-based pretraining, followed by fine-tuning on data from January 1, 2021, to December 31, 2023. The learned policies are then evaluated on an unseen test period ranging from January 1, 2024, to September 30, 2025.

[1] `https://finance.yahoo.com/`

5.2 Experiment settings

All experiments are conducted under consistent environmental settings to ensure fair comparison across models. The trading setup operates at a daily frequency, where decisions are made on each trading day based on daily market features. At the start of each trading period, the agent is initialized with a capital of $1,000,000$. Transaction costs are modeled by applying a fee of 0.1% on both buy and sell operations, and all trades are executed at the daily closing price. To capture heterogeneous expert behavior, five distinct expert trajectories are considered, including Cross-sectional momentum (CSMOM) (Jegadeesh & Titman, 2002), Time-series momentum (TSMOM) (Moskowitz et al., 2012), Moving average (MA) crossover (Brock et al., 1992), Bollinger Bands-based strategy (Kirkpatrick II & Dahlquist, 2010), and a DRL-based PPO trading agent. These strategies are chosen to reflect complementary trading principles, including momentum, mean-reversion, trend-following, and learning-based approaches, thereby ensuring diversity in expert behavior rather than redundancy. A detailed description of the expert strategies is provided in Appendix B. Hyperparameters for the proposed approach and baseline DRL models are tuned using Bayesian optimization (Snoek et al., 2012) implemented through the Hyperopt framework. The search ranges are chosen based on prior empirical studies to strike a balance between performance and computational efficiency (Mnih et al., 2016; Schulman et al., 2017). The complete set of hyperparameters and their search ranges are reported in Appendix E.

5.3 Baselines and Performance Metrics

We compare BRaG against a broad set of baseline strategies spanning classical finance, heuristic trading rules, and deep reinforcement learning methods to ensure a comprehensive evaluation. Non-learning benchmarks include the market index, buy-and-hold, random trading, and the mean–variance optimization (MVO) model (Markowitz, 1990). Rule-based technical strategies consist of CSMOM, TSMOM, MA crossover, and the Bollinger Band strategy. We also evaluate learning-based baselines, including A2C (Mnih et al., 2016), PPO (Schulman et al., 2017), and DDPG (Lillicrap et al., 2015), along with five recent DRL-based trading approaches from the literature, namely, VS-DRL (Zhang et al., 2019), SRRS (Rodinos et al., 2023), RSHF (Orra et al., 2025), Adaptive (Yang et al., 2020), and DREB (Orra et al., 2024). A detailed description of these baseline trading models is provided in Appendix C. All learning-based models are trained and evaluated under identical market conditions and transaction cost assumptions to ensure a fair comparison. The proposed BRaG model is evaluated using standard financial performance metrics, including return-based, risk-adjusted, and downside-risk measures. For completeness, the formal definitions of all evaluation metrics are provided in Appendix D.

5.4 Results and Discussion

Table 1 summarizes the performance of the proposed BRaG framework in comparison with a wide range of baseline strategies across four global stock market indices. The baselines span passive benchmarks, classical optimization methods, technical trading rules, and deep reinforcement learning models, enabling a comprehensive evaluation. Each DRL-based model is trained over five independent runs, and the average values of the evaluation metrics are reported. Due to space constraints, this section reports only the mean results, with the extended results containing both the mean and standard deviation presented in Appendix F (Table 3). The best-performing results are highlighted in bold.

A clear pattern emerges across all datasets. BRaG achieves the strongest overall performance in terms of cumulative and annualized returns on each market considered. On the DJI dataset, the proposed method delivers a markedly higher final portfolio value than both traditional benchmarks and learning-based strategies. Similar behavior is observed across the FTSE, Sensex, and TWII datasets, where BRaG consistently outperforms the strongest competing methods. The fact that this improvement is observed across markets with different levels of volatility and structural characteristics suggests that the learned policy captures general trading principles rather than dataset-specific effects.

Risk-adjusted performance further distinguishes BRaG from competing methods. The proposed model achieves the highest Sharpe ratio across all datasets, indicating a favorable balance between return and volatility. In contrast, several DRL baselines exhibit higher variability in returns despite achieving competitive profitability. BRaG also maintains comparatively lower maximum drawdowns, highlighting the effectiveness of explicit risk control during action execution. The win ratio analysis

Table 1: Performance comparison of the proposed BRaG model against various baselines across the datasets.

Dataset	Metrics	Index	Buy-Hold	MVO	Random	CSMOM	TSMOM	Bollinger	MA Crossover	A2C	DDPG	VS-DRL	SRRS	RSHF	Adaptive	DREB	PPO	BRaG†
DJI	Cumulative Return	22.8053	30.561	19.649	18.678	24.241	25.315	25.625	31.926	37.192	38.268	36.959	12.275	35.708	24.023	42.522	35.594	**53.977**
	Annual Return	12.576	16.522	10.899	10.379	13.334	13.897	14.064	17.323	20.003	20.545	19.885	6.905	19.252	13.219	22.626	17.669	**28.262**
	Sharpe Ratio	0.883	1.117	0.976	0.713	0.661	0.914	0.814	1.062	1.067	1.165	1.193	0.457	1.289	0.947	1.405	1.164	**1.472**
	Max Drawdown	16.369	15.916	**10.663**	17.142	29.982	13.501	15.991	18.443	23.631	19.724	19.351	21.524	15.701	16.401	17.968	16.208	17.958
	Win Ratio	54.13	55.15	53.44	47.91	53.21	49.31	54.36	47.94	55.05	54.19	57.14	50.01	54.59	55.28	57.34	55.73	**58.05**
FTSE	Cumulative Return	20.443	8.604	23.597	-14.446	17.165	32.061	27.605	30.802	33.917	29.585	17.478	19.144	33.788	30.541	43.898	32.294	**64.241**
	Annual Return	11.862	4.818	12.891	-8.523	9.452	17.181	14.914	16.544	18.117	15.923	9.619	10.502	18.052	16.415	23.103	17.298	**32.695**
	Sharpe Ratio	1.018	0.426	1.217	-0.314	0.619	1.232	0.829	0.891	1.034	1.257	0.585	0.709	1.112	1.046	1.376	1.106	**1.632**
	Max Drawdown	13.434	17.293	**9.438**	28.404	16.046	11.006	18.551	17.848	14.475	10.004	16.297	14.907	11.817	14.819	12.413	13.458	11.356
	Win Ratio	55.33	51.03	55.13	26.24	47.61	49.66	49.43	47.38	54.42	51.7	52.38	54.21	53.97	53.06	53.74	54.22	**57.51**
Sensex	Cumulative Return	11.784	16.135	11.098	-14.708	16.994	19.191	12.756	12.027	6.008	5.134	10.113	7.423	23.226	16.217	18.669	21.393	**34.067**
	Annual Return	6.698	9.092	6.316	-8.842	9.565	10.757	7.237	3.833	3.454	2.956	5.767	4.256	12.924	9.148	10.475	11.943	**18.608**
	Sharpe Ratio	0.549	0.698	0.589	-0.533	0.561	0.604	0.536	0.536	0.303	0.263	0.423	0.325	0.848	0.627	0.664	0.784	**1.042**
	Max Drawdown	14.964	15.913	17.747	27.027	17.935	21.311	15.451	**11.971**	17.716	17.833	23.755	15.854	16.235	15.807	17.412	15.354	15.193
	Win Ratio	51.35	53.24	51.19	27.69	47.45	50.46	48.38	43.52	51.16	51.16	53.47	49.39	53.47	52.55	51.39	51.62	**55.07**
TWII	Cumulative Return	43.276	51.428	25.039	31.014	62.625	47.187	31.188	17.875	68.504	95.384	29.394	46.248	71.962	49.536	76.954	66.454	**106.84**
	Annual Return	23.954	28.118	14.275	17.506	33.632	25.963	17.968	10.318	36.559	49.478	16.635	25.483	38.226	27.161	40.608	35.566	**54.343**
	Sharpe Ratio	1.041	1.053	1.125	0.754	1.072	1.565	0.716	0.671	1.436	1.605	0.747	0.983	1.323	1.213	1.501	1.436	**1.732**
	Max Drawdown	28.693	33.761	11.998	36.567	38.745	**10.762**	38.066	18.014	29.551	26.454	27.475	33.783	30.309	26.326	21.789	26.493	28.362
	Win Ratio	55.06	53.68	57.01	32.02	51.31	52.02	49.41	44.42	57.24	56.53	51.78	53.92	53.44	56.06	54.87	55.11	**60.82**

indicates that BRaG achieves a higher proportion of profitable trades across all markets. This suggests that performance improvements are driven by consistent decision-making rather than reliance on a small number of extreme outcomes. Technical trading strategies and random trading display lower win ratios and less stable behavior across datasets.

The corresponding out-of-sample cumulative return trajectories are presented in Appendix F, Figure 1. These plots complement the aggregate results in Table 1 by showing how the portfolio values evolve over the test period. BRaG exhibits a generally upward trajectory across the four markets, with comparatively controlled declines during several volatile intervals.

Overall, BRaG achieves stronger returns while maintaining a more stable portfolio evolution across different markets. The framework benefits from aggregating diverse expert behavior, which provides a structured starting point for learning, and from adversarial pretraining, which reduces unstable exploration. The additional use of explicit risk control further contributes to limiting drawdowns during periods of market volatility. Additional downside-risk metrics and ablation studies further support these findings and are reported in Appendices F and G.

6 Conclusion

In this paper, we presented BRaG, a framework for stock trading that combines multi-expert imitation learning with adversarial training and explicit risk control. By aggregating heterogeneous expert strategies using a barycenter construction, BRaG learns a stable reference behavior that guides policy learning. Adversarial pretraining helps initialize the policy in a meaningful region of the policy space, which reduces unstable exploration during reinforcement learning. The use of control barrier functions further constrains actions during execution and limits excessive drawdowns. The empirical evaluation shows that BRaG performs well across all the considered markets when compared with both traditional trading rules and recent DRL methods. The framework currently relies on predefined rule-based experts and a simplified risk model. Future work may incorporate richer expert sources, more accurate market dynamics, adaptive risk constraints, and extensions to higher-frequency trading.

AI use statement

In this work, we used generative AI tools solely to assist with writing and language editing. The tools were used to improve the clarity, readability, grammar, and organization of parts of the manuscript. Generative AI was not used to develop the research idea, formulate the methodology, generate or process the data, implement the proposed method, conduct experiments, derive mathematical claims, or interpret the experimental results. All AI-assisted text was carefully reviewed, revised, and verified by the authors. We take full responsibility for the final content of the manuscript, including all text, claims, results, and conclusions.

Ethics statement

This study uses historical financial market data and does not involve human participants, personal information, or sensitive data. The proposed framework is evaluated only through retrospective

backtesting and is intended for academic research. The reported results should not be interpreted as financial advice or as evidence of guaranteed performance in live markets.

Reproducibility statement

The main paper and appendices provide the information required to understand and reproduce the proposed framework. Section 5.1 describes the datasets, stock universes, temporal splits, and evaluation period. Section 5.2 reports the trading environment, initial capital, transaction costs, expert strategies, and hyperparameter selection procedure. The complete BRaG training pipeline is summarized in Algorithm 1. Appendix A describes the PPO optimization procedure, while Appendices B and C provide details of the expert trajectories and baseline methods. The evaluation metrics and hyperparameter search spaces are reported in Appendices D and E, respectively. All models are evaluated using the same market conditions, data splits, and transaction-cost assumptions to support a fair and reproducible comparison.

References

William Brock, Josef Lakonishok, and Blake LeBaron. Simple technical trading rules and the stochastic properties of stock returns. *The Journal of Finance*, 47(5):1731–1764, 1992.

Justin Fu, Katie Luo, and Sergey Levine. Learning robust rewards with adversarial inverse reinforcement learning. *arXiv preprint arXiv:1710.11248*, 2017.

Igor Halperin, Jiayu Liu, and Xiao Zhang. Combining reinforcement learning and inverse reinforcement learning for asset allocation recommendations. *arXiv preprint arXiv:2201.01874*, 2022.

Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. *Advances in neural information processing systems*, 29, 2016.

Narasimhan Jegadeesh and Sheridan Titman. Cross-sectional and time-series determinants of momentum returns. *The Review of Financial Studies*, 15(1):143–157, 2002.

Taylan Kabbani and Ekrem Duman. Deep reinforcement learning approach for trading automation in the stock market. *IEEE Access*, 10:93564–93574, 2022.

Charles D Kirkpatrick II and Julie R Dahlquist. *Technical analysis: the complete resource for financial market technicians*. FT press, 2010.

Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

Yang Liu, Qi Liu, Hongke Zhao, Zhen Pan, and Chuanren Liu. Adaptive quantitative trading: An imitative deep reinforcement learning approach. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 2128–2135, 2020.

Harry M Markowitz. Foundations of portfolio theory. *Harry Markowitz: Selected Works, S*, pp. 481–490, 1990.

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International Conference on Machine Learning*, pp. 1928–1937. PMLR, 2016.

John Moody, Lizhong Wu, Yuansong Liao, and Matthew Saffell. Performance functions and reinforcement learning for trading systems and portfolios. *Journal of forecasting*, 17(5-6):441–470, 1998.

Tobias J Moskowitz, Yao Hua Ooi, and Lasse Heje Pedersen. Time series momentum. *Journal of Financial Economics*, 104(2):228–250, 2012.

Andrew Y Ng, Stuart Russell, et al. Algorithms for inverse reinforcement learning. In *Icml*, volume 1, pp. 2, 2000.

Arishi Orra, Aryan Bhambu, Himanshu Choudhary, and Manoj Thakur. Dynamic reinforced ensemble using bayesian optimization for stock trading. In *Proceedings of the 5th ACM International Conference on AI in Finance*, pp. 361–369, 2024.

Arishi Orra, Himanshu Choudhary, Ankit Sharma, and Manoj Thakur. Enhancing deep reinforcement learning for stock trading: a reward shaping approach via expert feedback. *Knowledge and Information Systems*, pp. 1–20, 2025.

Jacobo Roa-Vicens, Cyrine Chtourou, Angelos Filos, Francisco Rullan, Yarin Gal, and Ricardo Silva. Towards inverse reinforcement learning for limit order book dynamics. *arXiv preprint arXiv:1906.04813*, 2019.

Georgios Rodinos, Paraskevi Nousi, Nikolaos Passalis, and Anastasios Tefas. A sharpe ratio based reward scheme in deep reinforcement learning for financial trading. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pp. 15–23. Springer, 2023.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems*, 25, 2012.

Qizhou Sun, Xueyuan Gong, and Yain-Whar Si. Transaction-aware inverse reinforcement learning for trading in stock markets. *Applied Intelligence*, 53(23):28186–28206, 2023.

Shuo Sun, Wanqi Xue, Rundong Wang, Xu He, Junlei Zhu, Jian Li, and Bo An. Deepscalper: A risk-aware reinforcement learning framework to capture fleeting intraday trading opportunities. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pp. 1858–1867, 2022.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

Avraam Tsantekidis, Nikolaos Passalis, Anastasia-Sotiria Toufa, Konstantinos Saitas-Zarkias, Stergios Chairistanidis, and Anastasios Tefas. Price trailing for financial trading using deep reinforcement learning. *IEEE Transactions on neural networks and learning systems*, 32(7):2837–2846, 2020.

Xing Wu, Haolei Chen, Jianjia Wang, Luigi Troiano, Vincenzo Loia, and Hamido Fujita. Adaptive stock trading strategies with deep reinforcement learning methods. *Information Sciences*, 538: 142–158, 2020.

Hongyang Yang, Xiao-Yang Liu, Shan Zhong, and Anwar Walid. Deep reinforcement learning for automated stock trading: An ensemble strategy. In *Proceedings of the first ACM International Conference on AI in Finance*, pp. 1–8, 2020.

Weipeng Zhang, Tao Yin, Yunan Zhao, Bing Han, and Huanxi Liu. Reinforcement learning for stock prediction and high-frequency trading with t+ 1 rules. *IEEE Access*, 11:14115–14127, 2022.

Zihao Zhang, Stefan Zohren, and Stephen Roberts. Deep reinforcement learning for trading. *arXiv preprint arXiv:1911.10107*, 2019.

A Proximal Policy Optimization

Proximal Policy Optimization (PPO) is a first-order policy gradient method designed to provide stable and efficient policy updates in reinforcement learning. PPO belongs to the family of actor-critic algorithms and optimizes a parameterized policy $\pi_\theta(a|s)$ using stochastic gradient ascent while constraining the update to remain close to the previous policy $\pi_{\theta_{\text{old}}}$. This constraint mitigates the instability often observed in vanilla policy gradient methods when large updates are applied. PPO achieves this objective without explicitly solving a constrained optimization problem, which makes it computationally efficient and well-suited for large-scale applications.

The core idea of PPO is to maximize a clipped surrogate objective. Let

$$r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$$

denote the probability ratio between the new and old policies, and let $\hat{A}_t$ be an estimate of the advantage function. The PPO objective is defined as

$$L^{\text{PPO}}(\theta) = \mathbb{E}\Big[\min\big(r_t(\theta)\hat{A}_t,\ \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t\big)\Big], \tag{17}$$

where $\epsilon > 0$ is a hyperparameter that controls the size of the trust region. This clipping operation prevents excessively large policy updates by limiting the contribution of samples where the new policy deviates significantly from the old one (Schulman et al., 2017). In practice, PPO is combined with a learned value function and entropy regularization to reduce variance and encourage exploration.

B Expert Trajectories

To construct diverse and informative expert demonstrations, we employ a set of rule-based technical trading strategies that have been well-studied in the finance literature. These strategies capture distinct market behaviors and provide complementary perspectives on trend persistence and price deviations.

1. **Time-Series Momentum (TSMOM):** TSMOM is a momentum-based strategy that takes positions based on the sign of an asset's own past returns. For an asset i at time t, the trading signal is given by
$$s_{i,t} = \text{sign}\big(r_{i,t-L:t-1}\big),$$
where $r_{i,t-L:t-1}$ denotes the cumulative return over a lookback window of length L. A long position is taken when the signal is positive, and a short position otherwise. TSMOM belongs to the class of time-series momentum strategies. It is beneficial as it captures persistent trends within individual assets and adapts naturally to regime changes where trends strengthen or weaken over time (Moskowitz et al., 2012).
2. **Cross-Sectional Momentum (CSMOM):** CSMOM ranks assets based on their past returns relative to one another and allocates positions accordingly. Let $r_{i,t-L:t-1}$ denote the past return of asset i. Assets are ranked across the universe, and the signal is defined as
$$s_{i,t} = \text{rank}\big(r_{i,t-L:t-1}\big),$$
with long positions assigned to top-ranked assets and short positions to bottom-ranked assets. CSMOM is a momentum-based cross-sectional strategy. It is useful because it exploits relative performance differences across assets and provides diversification benefits through long-short positioning (Jegadeesh & Titman, 2002).
3. **Moving Average Crossover:** The moving average crossover strategy generates signals based on the interaction of short-term and long-term moving averages of price. Let $\text{MA}_f(t)$ and $\text{MA}_s(t)$ denote fast and slow moving averages, respectively. The trading signal is
$$s_t = \begin{cases} +1, & \text{if } \text{MA}_f(t) > \text{MA}_s(t), \\ -1, & \text{otherwise.} \end{cases}$$
This strategy belongs to the trend-following category. It is beneficial due to its simplicity and its ability to filter out short-term noise while responding to sustained price movements (Brock et al., 1992).
4. **Bollinger Bands:** Bollinger Bands are constructed using a moving average and a volatility-based envelope. Let μ_t and σ_t denote the rolling mean and standard deviation of prices. The upper and lower bands are defined as
$$\text{UB}_t = \mu_t + k\sigma_t, \quad \text{LB}_t = \mu_t - k\sigma_t.$$
Trading signals are generated when prices deviate significantly from these bands. This strategy is commonly associated with mean-reversion behavior. It is useful as it captures overbought and oversold conditions, and complements momentum-based experts by focusing on price corrections rather than trend continuation (Kirkpatrick II & Dahlquist, 2010).

C BASELINE TRADING STRATEGIES

The performance of the proposed BRaG framework is benchmarked against a broad spectrum of trading strategies to ensure a comprehensive evaluation. These baselines include:

Rule-Based Technical Experts The individual performance of the CSMOM, TSMOM, MA Crossover, and Bollinger Band strategies.

Standard DRL Agents Vanilla A2C (Mnih et al., 2016), DDPG (Lillicrap et al., 2015), and PPO (Schulman et al., 2017) algorithms.

Traditional Finance Baselines

- **Market Index:** This represents a hypothetical portfolio used to measure the broader macroeconomic performance. Comparing against the index reveals whether the active trading strategy genuinely generated excess alpha over the general market drift.
- **Buy-and-Hold:** The buy-and-hold strategy involves buying and holding the assets over the entire investment horizon without active trading. It assumes no rebalancing, and benefits from long-term market appreciation.
- **Random Trading:** Random trading is a naive baseline where trades are executed randomly without any predictive model or signal. It functions as a statistical floor for performance.
- **Mean-Variance Optimization (MVO):** Mean-Variance Optimization is a foundational portfolio construction technique that aims to balance expected return against risk. It relies on estimates of expected returns and covariances, and assumes normally distributed returns, forming a foundational approach in modern portfolio theory.

State-of-the-Art Literature

- **Adaptive:** An ensemble trading strategy combining three actor-critic algorithms: A2C, DDPG, and PPO, to leverage their complementary strengths (Yang et al., 2020).
- **Volatility-Scaled Reward (VS-DRL):** This study incorporates volatility scaling into the reward function to dynamically adjust position sizes based on market volatility, thereby enhancing robustness and risk management (Zhang et al., 2019).
- **Sharpe Ratio Reward Scheme (SRRS):** This work implements a reward shaping technique by directly integrating an approximation of the Sharpe ratio into the reward function alongside profit and loss (Rodinos et al., 2023).
- **Dynamic Reinforced Ensemble (DREB):** A dynamic ensemble method that uses Bayesian optimization to assign time-varying weights to multiple DRL agents for automated stock trading (Orra et al., 2024).
- **Reward Shaping via Human Feedback (RSHF):** A reward-shaping framework that augments the conventional profit-and-loss objective with trading rules derived from technical indicators as human feedback to improve trading accuracy (Orra et al., 2025).

D PERFORMANCE METRICS

We evaluate all trading strategies using a set of standard financial performance metrics that capture profitability, risk-adjusted return, and downside risk.

1. **Cumulative Return:** The net change in portfolio value between the beginning and end of the trading period.
2. **Annualized Return:** It represents the average yearly growth rate of the portfolio.
3. **Sharpe Ratio:** It evaluates risk-adjusted performance by comparing the mean excess return of the strategy to the standard deviation of returns, thereby quantifying return per unit of risk.

4. **Maximum Drawdown:** Denotes the most severe loss experienced by the portfolio during the trading period, measured as the largest observed decline in portfolio value from a historical peak to a subsequent trough.

5. **Win Ratio:** The percentage of executed trades that yield a positive outcome compared to the total number of trades.

E HYPERPARAMETER SETTING

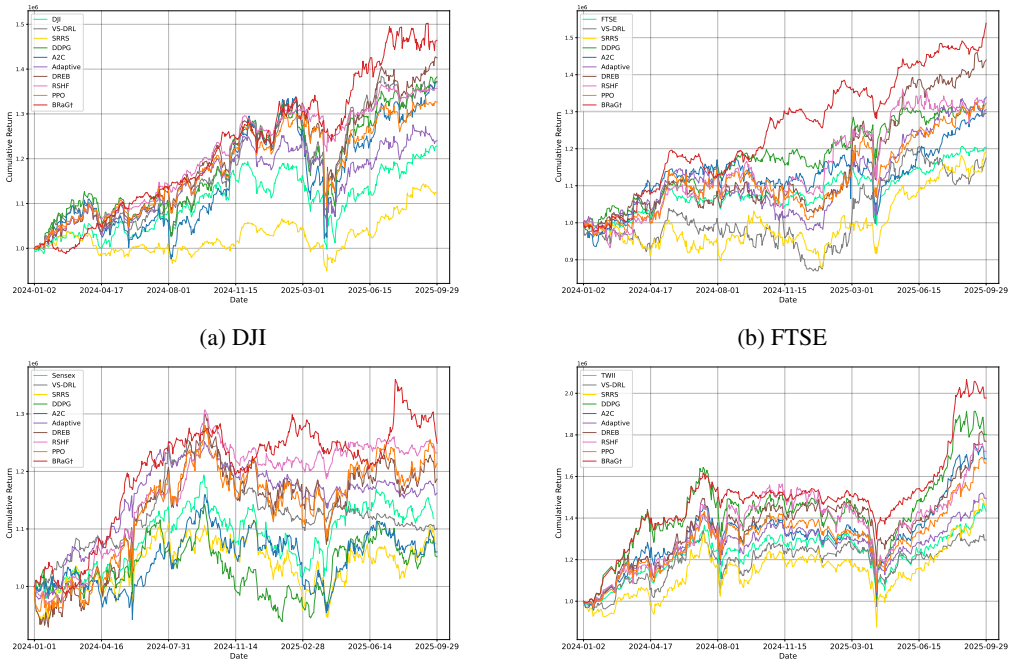
All experiments are conducted under uniform training and evaluation settings to ensure fair comparison across models. Hyperparameters for the proposed BRaG framework are selected using Bayesian optimization, which provides an efficient search strategy in high-dimensional parameter spaces. The tuning process is performed on a held-out validation split derived from the training data. We employ the Hyperopt library, which implements tree-structured Parzen estimators to guide the optimization. The search ranges for policy, discriminator, and optimization-related hyperparameters are chosen based on established empirical practices and preliminary experiments. A complete list of tuned hyperparameters and their corresponding ranges is reported in Table 2.

Table 2: Hyperparameter configuration and search space for the proposed BRaG model.

Model	**Hyperparameter**	**Range**
PPO	Hidden Dimension	[2,512]
	Number of Layers	[1,8]
	Activation Function	[ReLU, Tanh, Sigmoid]
	Learning Rate	$[e^{-8}, e^{-1}]$
	Dropout	[0,0.5]
	γ	[0.9,0.99]
	PPO Epochs	[5,50]
	Value Coefficient	[0.01,0.5]
	Entropy Coefficient	[0.01,0.1]
GAIL	Hidden Dimension	[2,512]
	Number of Layers	[1,8]
	Activation Function	[ReLU, Tanh, Sigmoid]
	Dropout	[0,0.5]
CBF	Rolling Window	20
	κ	1
	α	0.1
	CBF penalty coefficient	[0.01,0.1]

F ADDITIONAL RESULTS

The cumulative return plots in Figure 1 illustrate how the trading strategies behave over time rather than only at the end of the evaluation period. Across all four markets, BRaG shows a relatively steady upward trajectory, with fewer abrupt declines compared to most learning-based baselines. In the DJI and FTSE plots, several DRL methods achieve short phases of rapid growth but suffer noticeable reversals during volatile intervals. BRaG, in contrast, appears to moderate its exposure during these periods and recovers more gradually after downturns. A similar pattern is visible in the Sensex and TWII plots, where technical strategies and some DRL agents display uneven growth and larger fluctuations. Although certain rule-based methods perform competitively in specific sub-periods, their performance is not consistent across the whole horizon. Random trading and passive strategies remain clearly dominated in all markets. The smoother equity curves observed for BRaG suggest that the policy avoids extreme positions and benefits from a more conservative response to unfavorable market movements, which is consistent with the use of expert-guided pretraining and explicit risk constraints.



(a) DJI (b) FTSE

(c) Sensex (d) TWII

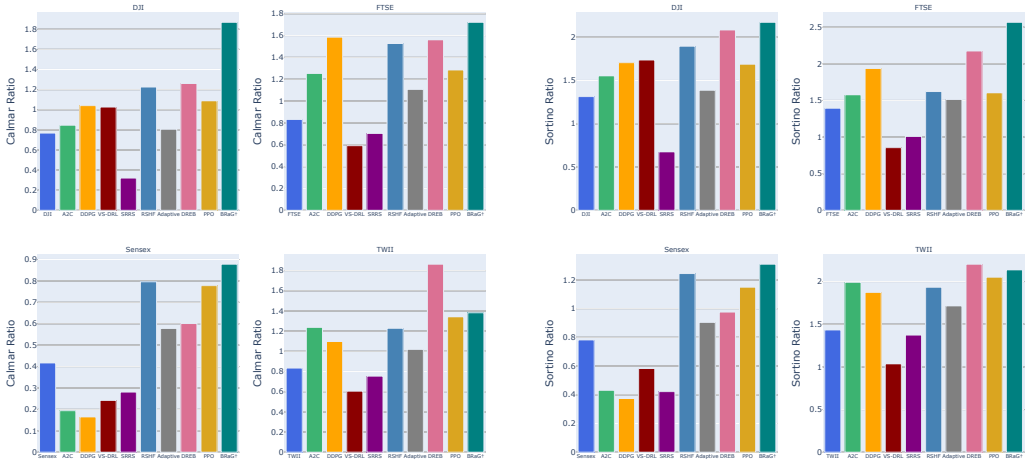
Figure 1: Cumulative return trajectories of the proposed BRaG model and baseline strategies across the DJI, FTSE, Sensex, and TWII datasets during the out-of-sample trading period.

Table 3 presents a comprehensive performance comparison, including the mean and standard deviation across five independent random seeds, for the learning-based methods (deterministic baselines inherently exhibit zero variance). The expanded results further underscore the robustness of the proposed framework. While standard DRL models, such as A2C, DDPG, and PPO, demonstrate moderate to high performance variance across different initializations, BRaG not only achieves the highest mean performance across all datasets but also maintains notably tighter standard deviations. This reduction in variance demonstrates stable behavior across different random initializations.

Table 3: Performance comparison of the proposed BRaG model against various baselines across the datasets. Results are reported as mean ± standard deviation over five independent runs.

Dataset	Metrics	Random	A2C	DDPG	VS-DRL	SRRS	RSHF	Adaptive	DREB	PPO	BRaG†
DJI	Cumulative Return	18.68 ± 5.42	37.19 ± 3.41	38.27 ± 4.12	36.96 ± 3.85	12.28 ± 2.15	35.71 ± 2.94	24.02 ± 2.65	42.52 ± 3.10	35.59 ± 2.87	**53.98 ± 1.95**
	Annual Return	10.38 ± 3.15	20.00 ± 1.84	20.55 ± 2.21	19.89 ± 1.95	6.91 ± 1.12	19.25 ± 1.65	13.22 ± 1.43	22.63 ± 1.75	17.67 ± 1.52	**28.26 ± 1.05**
	Sharpe Ratio	0.71 ± 0.18	1.07 ± 0.12	1.17 ± 0.14	1.19 ± 0.13	0.46 ± 0.16	1.29 ± 0.11	0.95 ± 0.14	1.41 ± 0.09	1.16 ± 0.10	**1.47 ± 0.06**
	Max Drawdown	17.14 ± 3.85	23.63 ± 2.15	19.72 ± 2.84	19.35 ± 2.45	21.52 ± 3.12	**15.70 ± 1.85**	16.40 ± 2.05	17.97 ± 1.75	16.21 ± 1.64	17.96 ± 1.25
	Win Ratio	47.91 ± 2.45	55.05 ± 1.34	54.19 ± 1.65	57.14 ± 1.45	50.01 ± 1.85	54.59 ± 1.25	55.28 ± 1.55	57.34 ± 1.15	55.73 ± 1.15	**58.05 ± 0.85**
FTSE	Cumulative Return	-14.45 ± 6.15	33.92 ± 3.85	29.59 ± 4.15	17.48 ± 3.25	19.14 ± 3.65	33.79 ± 3.15	30.54 ± 2.95	43.90 ± 3.25	32.29 ± 3.14	**64.24 ± 2.41**
	Annual Return	-8.52 ± 4.12	18.12 ± 2.15	15.92 ± 2.45	9.62 ± 1.85	10.50 ± 1.95	18.05 ± 1.75	16.42 ± 1.65	23.10 ± 1.85	17.30 ± 1.75	**32.70 ± 1.25**
	Sharpe Ratio	-0.31 ± 0.22	1.03 ± 0.14	1.26 ± 0.16	0.59 ± 0.18	0.71 ± 0.15	1.11 ± 0.12	1.05 ± 0.14	1.38 ± 0.10	1.11 ± 0.11	**1.63 ± 0.07**
	Max Drawdown	28.40 ± 4.25	14.48 ± 2.25	**10.00 ± 1.85**	16.30 ± 2.45	14.91 ± 2.15	11.82 ± 1.65	14.82 ± 1.95	12.41 ± 1.55	13.46 ± 1.65	11.36 ± 1.15
	Win Ratio	26.24 ± 3.15	54.42 ± 1.45	51.70 ± 1.85	52.38 ± 1.75	54.21 ± 1.65	53.97 ± 1.35	53.06 ± 1.55	53.74 ± 1.25	54.22 ± 1.35	**57.51 ± 0.95**
Sensex	Cumulative Return	-14.71 ± 5.85	6.01 ± 2.85	5.13 ± 3.15	10.11 ± 2.95	7.42 ± 2.45	23.23 ± 2.75	16.22 ± 2.55	18.67 ± 2.15	21.39 ± 2.75	**34.07 ± 1.85**
	Annual Return	-8.84 ± 3.65	3.45 ± 1.65	2.96 ± 1.85	5.77 ± 1.75	4.26 ± 1.45	12.92 ± 1.55	9.15 ± 1.45	10.48 ± 1.35	11.94 ± 1.55	**18.61 ± 1.05**
	Sharpe Ratio	-0.53 ± 0.24	0.30 ± 0.15	0.26 ± 0.18	0.42 ± 0.16	0.33 ± 0.14	0.85 ± 0.12	0.63 ± 0.13	0.66 ± 0.11	0.78 ± 0.12	**1.04 ± 0.08**
	Max Drawdown	27.03 ± 4.15	17.72 ± 2.45	17.83 ± 2.85	23.76 ± 3.15	15.85 ± 2.25	16.24 ± 1.95	15.81 ± 2.15	17.41 ± 1.85	15.35 ± 1.95	**15.19 ± 1.35**
	Win Ratio	27.69 ± 2.95	51.16 ± 1.65	51.16 ± 1.95	53.47 ± 1.85	49.39 ± 1.75	53.47 ± 1.45	52.55 ± 1.55	51.39 ± 1.35	51.62 ± 1.45	**55.07 ± 0.95**
TWII	Cumulative Return	31.01 ± 6.85	68.50 ± 4.85	95.38 ± 6.15	29.39 ± 4.25	46.25 ± 4.65	71.96 ± 4.15	49.54 ± 3.95	76.95 ± 4.35	66.45 ± 4.12	**106.84 ± 3.55**
	Annual Return	17.51 ± 3.95	36.56 ± 2.65	49.48 ± 3.45	16.64 ± 2.45	25.48 ± 2.55	38.23 ± 2.35	27.16 ± 2.25	40.61 ± 2.45	35.57 ± 2.25	**54.34 ± 1.95**
	Sharpe Ratio	0.75 ± 0.19	1.44 ± 0.13	1.61 ± 0.08	0.75 ± 0.17	0.98 ± 0.15	1.32 ± 0.12	1.21 ± 0.14	1.50 ± 0.11	1.44 ± 0.12	**1.73 ± 0.16**
	Max Drawdown	36.57 ± 5.15	29.55 ± 3.45	26.45 ± 4.15	27.48 ± 3.65	33.78 ± 3.85	30.31 ± 3.15	26.33 ± 3.25	**21.79 ± 2.85**	26.49 ± 2.95	28.36 ± 2.15
	Win Ratio	32.02 ± 3.85	57.24 ± 1.85	56.53 ± 2.15	51.78 ± 1.95	53.92 ± 1.85	53.44 ± 1.65	56.06 ± 1.75	54.87 ± 1.55	55.11 ± 1.65	**60.82 ± 1.15**

To further examine the risk characteristics of the proposed method, we report additional results based on the Calmar and Sortino ratios. These metrics are commonly used to assess downside risk and capital preservation, which are not fully captured by return-based measures alone. For this purpose, we plot the distribution of Calmar and Sortino ratios across all models and datasets. The corresponding histograms are shown in Figures 2a and 2b. These plots provide a complementary view of model behavior under adverse market movements.



(a) Calmar Ratio

(b) Sortino ratio

Figure 2: Distribution of Calmar and Sortino ratios across all models and datasets, illustrating comparative downside risk and risk-adjusted performance.

The Calmar ratio compares overall returns with the maximum drawdown experienced during the trading period. BRaG consistently attains higher Calmar values across the DJI, FTSE, Sensex, and TWII datasets. This suggests that the gains achieved by BRaG are not driven by taking excessive drawdown risk. Several DRL baselines achieve reasonable returns but exhibit lower Calmar ratios, indicating larger drawdowns during unfavorable periods. Rule-based strategies and random trading show noticeably weaker performance, particularly on markets with higher volatility. Overall, the Calmar results indicate that BRaG is able to control large losses more effectively than the competing approaches. The Sortino ratio focuses specifically on downside variability and penalizes negative returns rather than total volatility. BRaG again achieves the highest Sortino ratios across all markets. This indicates that negative return fluctuations are comparatively limited for the proposed method. While some learning-based baselines perform well on individual datasets, their Sortino ratios vary substantially across markets. In contrast, BRaG shows more stable behavior, with consistently strong downside-adjusted performance. These observations support the view that combining expert-guided pretraining with explicit risk constraints leads to more controlled trading behavior.

G Ablation Study

To better understand the contribution of the main components of the proposed framework, we conduct a detailed ablation study using the DJI dataset. The first study examines the effect of varying the number of expert trajectories used to construct the barycenter representation. The second study analyzes the impact of the drawdown constraint imposed through the control barrier function. In both cases, we evaluate performance using cumulative return and the Sharpe ratio to capture profitability and risk-adjusted behavior. Figures 3 and 4 present the ablation results, illustrating the effect of varying the number of expert trajectories used for barycenter construction and the drawdown limit in the control barrier function, respectively. In addition to these two studies, we carry out three further ablations to isolate the contribution of individual design choices in BRaG. Table 4 compares the Sharpe-based expert-weighting scheme used in the barycenter against an equal-weighting alternative, to check whether the performance gains of BRaG stem from favoring stronger experts or merely from combining multiple experts. Table 5 presents a component-wise ablation, allowing the contribution of every major module to be assessed individually rather than only in combination. Table 6 compares different expert-aggregation strategies, in order to separate the benefit of using multiple experts at all from the benefit of the specific barycenter formulation adopted in this work.

In the expert ablation study, we vary the number of expert trajectories used during training from one to five. When a single expert is used, it corresponds to the best-performing expert based on validation

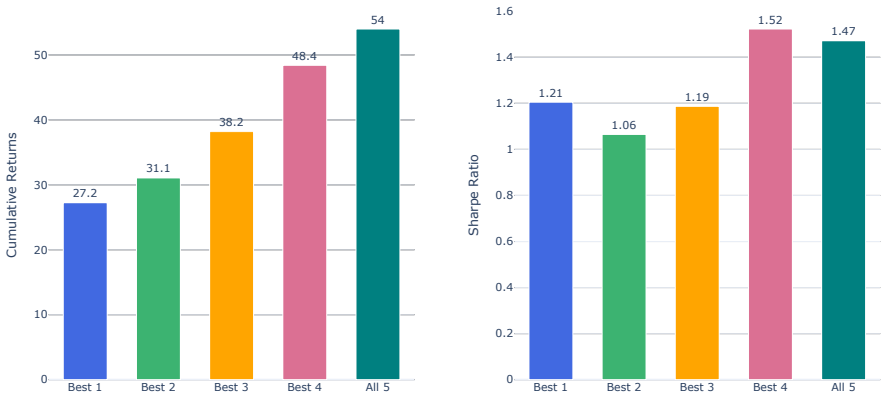


Figure 3: Effect of varying the number of expert trajectories used for barycenter construction on cumulative return and Sharpe ratio. Experts are added incrementally based on validation performance.

performance. For two, three, and four experts, the sets are formed by selecting the top-performing experts in descending order of performance. The final setting uses all five available experts. The results show a clear improvement in cumulative returns as more expert trajectories are incorporated. A similar trend is observed for the Sharpe ratio, which increases as the diversity of expert behavior grows. This suggests that aggregating multiple experts helps the model capture a richer set of trading patterns and reduces over-reliance on a single strategy. The marginal improvement from four to five experts is smaller, indicating that most gains are achieved once sufficient diversity is introduced.

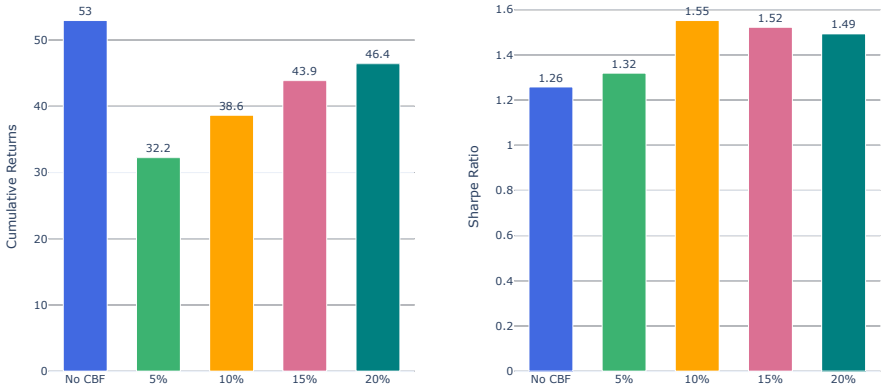


Figure 4: Effect of varying the maximum drawdown limit in the control barrier function on cumulative return and Sharpe ratio.

The drawdown ablation study investigates the role of explicit risk control by varying the maximum allowable drawdown from 5% to 20% and by removing the control barrier function. Without any drawdown constraint, the model achieves high cumulative returns but exhibits weaker risk-adjusted performance. Introducing moderate drawdown limits results in a noticeable improvement in the Sharpe ratio, with the optimal balance observed at intermediate constraint levels. Very tight or very loose constraints result in suboptimal performance, either by limiting profitable opportunities or by allowing excessive risk. These results suggest that the control barrier function plays a crucial role in stabilizing learning and enhancing risk-adjusted returns when appropriately calibrated.

Table 4: Effect of expert-weighting schemes in the Wasserstein barycenter on the DJI dataset. Sharpe-based weighting assigns greater influence to experts with stronger risk-adjusted validation performance, whereas equal weighting assigns identical importance to all experts.

Barycenter weighting scheme	Cumulative Return ↑	Sharpe Ratio ↑	Maximum Drawdown ↓
Sharpe-based weighting	53.977	1.472	17.958
Equal weighting	48.502	1.281	20.133

Table 4 presents the effect of the expert-weighting scheme used while constructing the Wasserstein barycenter. Two settings are compared here. The first is the Sharpe-based weighting used in the proposed BRaG framework, in which each expert is assigned a weight based on its risk-adjusted validation performance. The second is an equal-weighting scheme, in which every expert contributes equally to the barycenter regardless of how well it actually performed. This comparison is included to check whether the performance gains of BRaG come specifically from favoring better experts, or whether simply combining multiple experts is enough on its own. The results in Table 4 show that Sharpe-based weighting gives a clear advantage over equal weighting on all three metrics. Cumulative return drops from 53.977 to 48.502 when equal weighting is used, and the Sharpe ratio falls from 1.472 to 1.281. Maximum drawdown also worsens, rising from 17.958 to 20.133. This pattern suggests that when weaker experts are allowed to influence the pseudo-expert distribution as much as stronger ones, the resulting barycenter becomes a less reliable target for imitation learning. Greater emphasis on reliable experts produces a more useful pseudo-expert representation and improves both profitability and risk control.

Table 5: Component-wise ablation of BRaG on the DJI dataset. Each variant removes one major component from the full framework while retaining the remaining modules.

Component removed	Cumulative Return ↑	Sharpe Ratio ↑	Maximum Drawdown ↓
GAIL Pretraining	33.501	1.245	15.796
PPO Fine-tuning	42.378	1.335	16.953
CBF	52.951	1.258	25.874
None—Full BRaG	53.977	1.472	17.958

Table 5 reports the component-wise ablation of BRaG on the DJI dataset, where each of the three main components, namely GAIL pretraining, PPO fine-tuning, and the control barrier function, is removed one at a time while keeping the remaining components intact. This experiment is carried out to understand how much each individual component contributes to the overall performance of the framework, since the full model combines several mechanisms together, and it is not obvious from the aggregate results alone which part is doing the heavy lifting. The results indicate that removing GAIL pretraining causes the sharpest fall in performance, with cumulative return dropping to 33.501 and the Sharpe ratio to 1.245. This shows that the barycenter-guided pretraining stage plays a central role in giving the policy a useful starting point before it is exposed to the true market reward. Removing PPO fine-tuning also hurts performance, though to a lesser extent, with a cumulative return of 42.378 and a Sharpe ratio of 1.335. This is expected because, without fine-tuning, the policy is limited to the behavior it can imitate from the pseudo-expert and never has the opportunity to improve further using actual portfolio returns. The CBF ablation tells a different kind of story. Removing the CBF actually results in a slightly higher cumulative return of 52.951, close to the full model, but the Sharpe ratio drops to 1.258, and the maximum drawdown increases sharply to 25.874. This shows that the control barrier function does not contribute much to raw profitability, but it does most of the work in keeping the strategy's risk profile in check. Taken together, these three ablations show that the two learning-based components are primarily responsible for BRaG's return advantage, while the CBF is responsible for the risk-adjusted stability that distinguishes BRaG from other DRL baselines.

Table 6 compares different ways of aggregating expert demonstrations, ranging from using just a single best-performing expert to the full barycenter-based approach used in BRaG. Four settings are reported, namely the best single expert, uniform sampling across all experts, Sharpe-weighted sampling without any barycenter construction, and the full BRaG method. This comparison is meant to separate the value of using multiple experts at all from the value of the specific barycenter

Table 6: Comparison of expert-demonstration aggregation strategies on the DJI dataset.

Expert Aggregation Method	Cumulative Return ↑	Sharpe Ratio ↑	Maximum Drawdown ↓
Best Single Expert	27.235	1.205	14.982
Uniform Expert Sampling	46.445	1.285	20.181
Sharpe-weighted Sampling	50.213	1.388	18.875
Full BRaG	53.977	1.472	17.958

construction technique adopted in this work. The best single expert setting performs the weakest among the four, with a cumulative return of only 27.235 and a Sharpe ratio of 1.205, even though it has the lowest maximum drawdown of 14.982. This is not surprising, since relying on a single expert restricts the policy to a single trading philosophy and prevents it from benefiting from complementary strategies from the other experts. Moving to uniform sampling improves the cumulative return considerably to 46.445, confirming that combining multiple experts, even without any weighting, already yields a substantial benefit. Sharpe-weighted sampling further improves results, raising the cumulative return to 50.213 and the Sharpe ratio to 1.388, showing that giving greater weight to better-performing experts during sampling adds value on top of simple pooling. The full BRaG framework, which uses the barycenter construction along with Sharpe-based weighting, achieves the best results overall across all three metrics. This suggests that the barycenter is doing more than just weighting, since it also captures the underlying geometry of the expert distributions in a way that plain weighted sampling cannot fully replicate. Overall, the results in Table 6 indicate that both the multi-expert idea and the specific barycenter formulation contribute meaningfully, with the barycenter providing the final and most useful layer of improvement.