INVESTOR-SPECIFIC MULTI-OBJECTIVE PREFERENCE ALIGNMENT FOR PERSONALIZED PORTFOLIO OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning for portfolio management typically outputs a single strategy for all investors, failing to reflect heterogeneous investor goals. To address this gap, we propose Investor-Type-Wise Multi-Objective Preference Alignment (IMOPA), a framework that accounts for diverse investor profiles and outputs strategies for different investors. Specifically, we integrate preference learning, bandit-based preference weight estimation, and policy optimization into IMOPA. IMOPA first evaluates candidate policies through LLM-generated multi-objective feedback on financial metrics together with an overall evaluation. Utilizing such information, a bandit-based mechanism estimates multi-objective preference weights and selects a candidate policy for each investor type. Finally, IMOPA employs Proximal Policy Optimization (PPO) conditioned on the learned multi-objective preferences to train personalized policies. Experiments on two datasets—S&P 500 and MidCap 100—demonstrate that our estimated preference weights reveal distinct financial trade-offs, leading to customized asset allocations across investor types. Furthermore, ablation studies show that omitting preference alignment narrows or collapses these behavioral differences. Overall, IMOPA enables interpretable, preference-aligned portfolio policies tailored to diverse investor profiles.

1 INTRODUCTION

Classical portfolio methods typically seek to maximize expected return while controlling investment risk under the static setting, often assuming a common objective across all investors (Markowitz, 1952). However, portfolio decisions depend on both time-varying investment and heterogeneity in investors' beliefs and risk preferences in practice (Brandt et al., 2005; Egan et al., 2026). In contrast to static formulations such as mean-variance optimization (Merton, 1972), reinforcement learning (RL) has been used as a sequential decision-making tool for portfolio problems, which help design policies that can respond to evolving market environments and hence handle time-varying opportunities. In such dynamic settings, portfolio based on RL methods also expand the optimization objective based on reward and variance to those based on richer criteria, such as return distributions and downside- or tail-related risks (Abdolmaleki et al., 2020; Wang et al., 2025). Despite accounting for dynamic markets, heterogeneous investor preferences remain underexplored in RL-based portfolio optimization.

In financial applications, RL methods can learn from different forms of feedback, including the scalar reward and preference feedback. Standard RL learns from scalar reward feedback generated through interactions with the environment. For example, Jiang et al. (2017) optimize portfolio policies using the transaction-cost-adjusted log return as a scalar reward signal. Moreover, preference-based RL learns from preference feedback over actions or trajectories, and a recent work by Xu et al. (2025) uses pairwise comparisons of trading trajectories to capture investor preferences over execution strategies. These approaches collect feedback without explicitly distinguishing individual objectives, overlooking the multiple and potentially conflicting financial objectives that may be valued differently from individuals. Therefore, multi-objective feedback provides a more comprehensive evaluation by assessing each policy along distinct dimensions and explicitly retaining the trade-offs among competing objectives (Vamplew et al., 2011; Roijers et al., 2013).

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.

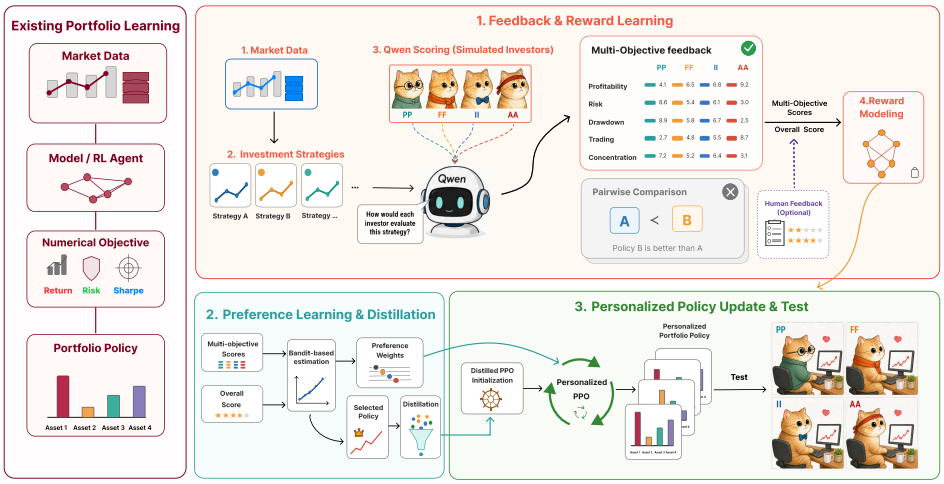

Figure 1: Comparison Between Existing Portfolio Learning and Our Personalized Framework of IMOPA.

Existing multi-objective RL methods can accommodate different objectives, yet their preference vectors are commonly specified in advance or estimated from preference-data statistics (Wang et al., 2024; Xiong & Singh, 2025), limiting their ability to model heterogeneous investor preferences. To fill this gap, we develop personalized portfolio policies for heterogeneous investor types by designing the IMOPA framework. As illustrated in Figure 1, IMOPA utilizes the methods of offline preference elicitation from an LLM evaluator, bandit-based preference estimation, and personalized policy learning. We validate this framework on the S&P 500 and MidCap100 datasets and examine whether the learned preferences induce interpretable behavioral heterogeneity while maintaining competitive investment performance. In particular, the key design in IMOPA is as below:

(i) Within an investor-type framework, we propose IMOPA to model investor-specific preferences, using four investor profiles as representative cases. We use Qwen3.5 to generate multi-dimensional feedback and a overall evaluation for each investor type, then fit surrogate reward models approximate the dimension-level feedback during subsequent policy optimization.

(ii) We formulate investor-specific preference estimation as a bandit-based problem under a linear assumption. The proposed procedure estimates preference weights and selects a policy for each investor type, with theoretical analysis of the confidence set and stopping criterion.

(iii) We develop a personalized PPO framework that incorporates the learned multi-objective reward and estimated preference weights into policy optimization, producing distinct portfolio policies tailored to different investor types.

2 Related Work

The related work can be divided into three strands: portfolio RL and multi-objective optimization, preference-based RL for heterogeneous feedback, and preference weight estimation.

Portfolio RL and Multi-Objective Optimization. Portfolio RL has advanced along two main directions: richer state representations and broader objective functions. Recent methods move beyond price-based market states by modeling temporal dependencies, cross-asset relations, and broader market conditions (Zhang et al., 2022; Wang et al., 2021). On the objective side, Zhang et al. (2022) introduce a cost-sensitive RL framework that accounts for return, variance risk, and transaction costs, while Hwang & Zohren (2026) directly optimizes CVaR for tail-risk control. Multi-objective methods further formulate potentially conflicting objectives jointly, with preference weights characterizing their relative importance (Skalse & Abate, 2023; Zhang et al., 2025; Chen et al., 2026). Taken together, these works advance state representation and objective design, but do not account

for heterogeneous investor preferences when determining the relative importance of different objectives. In contrast, IMOPA estimates investor-specific preference weights from multi-dimensional feedback and incorporates them into portfolio policy optimization.

Preference-based RL for Heterogeneous Feedback. Preference-based RL provides a general mechanism for preference evaluations in policy learning, using preference feedback to learn reward models or directly optimize policies (Christiano et al., 2017; Stiennon et al., 2020). More recent studies examine personalized or group-dependent preferences, suggesting that a single population-level preference model may fail to represent differences across users (Hosking et al., 2024; Poddar et al., 2024; Zhong et al., 2024). Despite this progress, preference-based RL for heterogeneous users remains underexplored in portfolio optimization, where investor heterogeneity is particularly relevant because differences in risk preferences are associated with portfolio diversification and trading behavior (Dorn & Huberman, 2005). Recent work has combined preference elicitation with multi-objective financial decision-making: Dispoto et al. (2026) infer ESG-related preferences from pairwise comparisons of candidate portfolios, while Xu et al. (2025) use trading-trajectory comparisons to learn preferences over execution strategies. Nevertheless, the estimation of investor-specific objective importance from heterogeneous multi-dimensional feedback remains limited.

Preference Weight Estimation. Once multi-objective feedback is available, a remaining issue is estimating objective importance. Linear scalarization represents this through a preference vector over objective-specific rewards. Existing methods obtain these parameters differently: Directional Preference Alignment represents preferences as specified directions in the multi-objective reward space (Wang et al., 2024),whereas Projection Optimization estimates objective weights from statistics of collected preference data (Xiong & Singh, 2025). In contrast, IMOPA estimates investor-specific preference weights from the relationship between multi-dimensional evaluations and a separate overall evaluation. We formulate this estimation as a bandit-based procedure built on linear best-arm identification (BAI)(Soare et al., 2014; Tao et al., 2018), estimating the preference vector and selecting an initialized policy for each investor type.

3 Problem Setting

In this section, we formulate investor-specific preferences through multi-objective feedback, and consider four investor profiles as representative cases. We model portfolio management as a Markov decision process (MDP) defined by (S, A, P, γ). Let $\mathcal{I} = \{i = 1, \ldots, m\}$ denote the set of investor types, with N assets over T decision periods.

3.1 Investors' Preferences Alignment

In the portfolio selection problem, let Π denote the set of admissible portfolio policies. A policy $\pi \in \Pi$ maps a market state s_t to a portfolio action $\mathbf{a}_t \sim \pi(\cdot \mid s_t)$. We use π_t to denote policy π evaluated at time t. We introduce a multi-objective evaluation of portfolio policies and incorporate the preference information into the investment objective.

Investor-type-wise Multi-Objective Reward: Upon receiving portfolio metrics of policies, investors evaluate π_t across multiple dimensions, such as profitability, risk management, and turnover. Investors with different preferences may apply different criteria when evaluating the same policy. Assuming that there are d objectives, the multi-objective feedback for investor i is $\mathbf{r}^{(i)}_{\pi_t} = \left[r^{(i)}_{\pi_t,1}, \ldots, r^{(i)}_{\pi_t,d}\right]^\top \in \mathbb{R}^d$. Here, $r^{(i)}_{\pi_t,j}$ denotes the evaluation of π_t on objective j. Let $\mathbf{w}^{(i)} = \langle w^{(i)}_1, \ldots, w^{(i)}_d \rangle^\top \in \mathbb{R}^d$ be a normalized preference-weight vector for investor type i. The preference utility of policy π_t with $\mathbf{w}^{(i)}$ is defined as the linear combination of d objectives:

$$U^{(i)}_{\pi_t} = \left(\mathbf{w}^{(i)}\right)^\top \mathbf{r}^{(i)}_{\pi_t} = \sum_{j=1}^{d} w^{(i)}_j r^{(i)}_{\pi_t,j}. \tag{1}$$

For $t = 0, \ldots, T-1$, $\mathbf{a}_t =< a^1_t, \ldots, a^N_t > \in A$ is the vector of target portfolio weights at time t. For a deterministic policy, this reduces to $\mathbf{a}_t = \pi(s_t)$. We impose a no-short-selling constraint, following the common long-only portfolio setting (Jiang et al., 2017), and thus the action satisfies

$\mathbf{a}_t \geq 0$ and $\sum_{j=1}^{N} a_t^j = 1$. Let the immediate reward from t to $t+1$ be the logarithmic portfolio return

$$Z_{t+1} = \log \left[\sum_{j=1}^{N} a_t^j \left(c_{t+1}^j / c_t^j \right) \right], \tag{2}$$

where c_t^j is the closing price of the j-th available asset at the stage t.

We define the objective as the sum of the discounted return and an investor-specific preference increment:

$$\max_{\pi \in \Pi} \sum_{t=0}^{T-1} \left[\gamma^t Z_{t+1} + \beta_{\text{pref}} \Delta U_{\pi_{t+1}}^{(i)} \right], \ \forall i \in \mathcal{I}, \tag{3}$$

where $\Delta U_{\pi_{t+1}}^{(i)} = (U_{\pi_{t+1}}^{(i)} - U_{\pi_t}^{(i)})$, and β_{pref} balances portfolio return and investor preference.

Behavioral Investor Profiles: We consider four behavioral investor profiles based on style (passive or active) and risk tolerance following (Pompian, 2012):

- **Passive Preservers (PP):** Low risk tolerance and stronger capital-preservation.
- **Friendly Followers (FF):** Moderate risk tolerance and trend-responsive.
- **Independent Individualists (II):** Moderate risk tolerance and independent.
- **Active Accumulators (AA):** High risk tolerance and more active trading behavior.

The four investor profiles serve only as representative cases for evaluating IMOPA. The IMOPA framework itself is independent of any specific investor classification and applies to a general investor set $\mathcal{I}$.

4 METHODOLOGY

IMOPA consists of three stages: preference elicitation and reward modeling, bandit-based preference estimation and policy selection, and personalized PPO updating. Algorithm 1 summarizes the overall pipeline.

4.1 INVESTOR FEEDBACK ELICITATION AND PERSONALIZED REWARD MODELING

Algorithm 1 Overall IMOPA Pipeline.

Require: Investor set $\mathcal{I}$, learning algorithm set $\mathcal{B}$, time horizon T, number of training policies K
Ensure: Preference vectors $\{\widehat{\mathbf{w}}^{(i)}\}_{i \in \mathcal{I}}$, personalized policies $\{\pi_\theta^{(i)}\}_{i \in \mathcal{I}}$

```
Train candidate policies Π_cand = {π_k}_{k=1}^K using algorithms in B;
Compute m_{π_k,t} and collect {r^(i)_{π_k,t}, R^(i)_{π_k,t}} for π_k ∈ Π_cand, t ∈ T, and i ∈ I;
Let X_t^(i) = { r^(i)_{π_k,t} : π_k ∈ Π_cand };
for all i ∈ I do
    Train and select reward model f^(i) using (m_{π_k,t}, r^(i)_{π_k,t});
end for
({ŵ^(i), π_sel^(i)}_{i∈I}) ← BANDITPREFERENCE({X_t^(i), R^(i)_{π_k,t}}_{i∈I});
Distill each π_sel^(i) into π_{θ_0}^(i) via Eq. (9);
{π_θ^(i)} ← PERSONALIZEDPOLICYUPDATE({f^(i), ŵ^(i), π_{θ_0}^(i)}_{i∈I});
return {ŵ^(i), π_θ^(i)}_{i∈I}
```

In the pipeline of IMOPA, the first stage constructs a representation of investor feedback for subsequent preference estimation and policy optimization. Given a candidate policy set $\Pi_{\text{cand}} = \{\pi_k\}_{k=1}^K$, for each policy $\pi_k \in \Pi_{\text{cand}}$ and evaluation period t. Let $\tau_{\pi_k,t}$ denote the trading trajectory and define the observable metric vector $\mathbf{m}_{\pi_k,t} = \psi(\tau_{\pi_k,t})$, whose components represent

profitability, risk, drawdown, trading activity, and portfolio concentration. For each investor type $i \in \mathcal{I}$, an LLM evaluates the corresponding multi-objective feedback vector

$$\mathbf{r}^{(i)}_{\pi_k,t} = \left[r^{(i)}_{\pi_k,t,\text{profit}}, r^{(i)}_{\pi_k,t,\text{risk}}, r^{(i)}_{\pi_k,t,\text{drawdown}}, r^{(i)}_{\pi_k,t,\text{trading}}, r^{(i)}_{\pi_k,t,\text{concentration}} \right]^\top, \tag{4}$$

together with a separate overall evaluation $R^{(i)}_{\pi_k,t}$. The two outputs serve different purposes: $\mathbf{r}^{(i)}_{\pi_k,t}$ describes policy π_k along individual preference dimensions and is used to train the reward models, whereas $R^{(i)}_{\pi_k,t}$ reflects the joint evaluation of π_k and is used only as the observation in the subsequent bandit-based preference estimation.

To avoid repeated LLM inference during policy optimization, the available surrogates for investor type i are denoted by $f^{(i)}_q : \mathbf{m}_{\pi_k,t} \mapsto \widehat{\mathbf{r}}^{(i)}_{\pi_k,t} \in [1,10]^d, q \in \mathcal{Q}$, where $\mathcal{Q}$ is the set of fitted models. In Algorithm 1, $\mathcal{Q}$ may utilize different regression methods, such as polynomial ridge regression with varying degrees and gradient boosting decision trees (GBDT). We select $f^{(i)}$ from $\mathcal{Q}$ according to the lowest validation MAE.

Algorithm 2 Bandit-Based Preference Estimation and Policy Selection

Require: Feedback sets $\{\mathcal{X}^{(i)}_t\}_{i\in\mathcal{I},t\in T}$, overall evaluations $\{R^{(i)}_{\pi,t}\}$, candidate policies Π_{cand}, ridge parameter λ, maximum steps $N_{\max}$

Ensure: Preference vectors $\{\widehat{\mathbf{w}}^{(i)}\}_{i\in\mathcal{I}}$ and selected policies $\{\pi^{(i)}_{\text{sel}}\}_{i\in\mathcal{I}}$

for all $i \in \mathcal{I}$ **do**
  $\mathcal{X}^{(i)} \leftarrow \bigcup_{t\in T} \mathcal{X}^{(i)}_t$;
  for $n = 1, \ldots, N_{\max}$ **do**
    Select $\mathbf{x}_n \in \mathcal{X}^{(i)}$ by G-allocation and obtain its corresponding $y^{(i)}_n$ from $R^{(i)}_{\pi,t}$;
    Update $\widehat{\mathbf{w}}^{(i)}_n$ using Eq. (6);
    if Eq. (7) holds **then**
      break
    end if
  end for
  $\widehat{\mathbf{w}}^{(i)} \leftarrow \widehat{\mathbf{w}}^{(i)}_n$ and, for each $t \in T$, $\widehat{\pi}^{(i)}_t \leftarrow \arg\max_{\pi\in\Pi_{\text{cand}}} (\mathbf{x}^{(i)}_{\pi,t})^\top \widehat{\mathbf{w}}^{(i)}$;
  $\widehat{\mathbf{x}}^{(i)}_{\text{emp}} \leftarrow 1/T \sum_{t=1}^{T} \mathbf{x}^{(i)}_{\widehat{\pi}^{(i)}_t,t}$;
  Construct alternative policies $\pi^{(i)}_a$ and $\pi^{(i)}_b$, and select $\pi^{(i)}_{\text{sel}}$ by Eq. (8);
end for
return $\{\widehat{\mathbf{w}}^{(i)}, \pi^{(i)}_{\text{sel}}\}_{i\in\mathcal{I}}$.

4.2 Bandit-Based Preference Estimation and Selected Policy Identification

In the real world, investors weigh multiple financial dimensions differently, and hence the aim of the bandit-based method is to estimate the target preference weights for each investor type i within the multi-objective scoring space. It operates on the LLM feedback collected in the preceding stage.

Define the scalar observation of investor type i as

$$y^{(i)}_{\pi,t} = \left(\mathbf{x}^{(i)}_{\pi,t}\right)^\top \mathbf{w}^{(i)} + \xi^{(i)}_{\pi,t}, \tag{5}$$

where $\mathbf{x}^{(i)}_{\pi,t}$ is the normalized dimension-level feedback vector of $r^{(i)}_{\pi_k,t}$ and $\mathbf{w}^{(i)}$ denotes the latent preference weights mentioned in Eq. (1). Assume all candidate arms share the same $\mathbf{w}^{(i)}$, each observation contributes to its estimation. IMOPA adopts the G-allocation strategy of Soare et al. (2014) to prioritize informative arms, and iteratively update $\widehat{\mathbf{w}}^{(i)}_n$ until the stopping criterion in Eq. (7) is satisfied. Because the elements of $\mathbf{w}^{(i)}$ represent the relative importance assigned to the d investment objectives, it can be estimated using constrained ridge regression,

$$\widehat{\mathbf{w}}^{(i)}_n = \arg\min_{\mathbf{w}\in\mathcal{W}} \left\{ \sum_{s=1}^{n} \left(y^{(i)}_s - \mathbf{x}_s^\top \mathbf{w} \right)^2 + \lambda \|\mathbf{w}\|_2^2 \right\}, \tag{6}$$

where $\mathcal{W} = \left\{\mathbf{w} : \mathbf{w} \geq 0,\ \mathbf{1}^\top \mathbf{w} = 1\right\}$. The constraints preserve the economic interpretation of $\widehat{\mathbf{w}}^{(i)}$, while ridge regularization improves estimation stability when the feedback dimensions are correlated. Proposition 1 provides a finite-sample confidence guarantee for identifying the preference-optimal arm and the proof is provided in Appendix B.

Proposition 1. *For investor type $i \in \mathcal{I}$, let $\mathcal{X}^{(i)} \subset \mathbb{R}^d$ be a finite set of feedback arms and $\mathbf{w}^{(i)} \in \mathcal{W}$ the unknown preference vector. Assume that the preference-optimal arm is $\mathbf{x}^{*(i)} = \arg\max_{\mathbf{x}\in\mathcal{X}^{(i)}} \mathbf{x}^\top \mathbf{w}^{(i)}$. After selecting arm $\mathbf{x}_n$ at step n, the scalar observation is modeled as $y_n^{(i)} = \mathbf{x}_n^\top \mathbf{w}^{(i)} + \xi_n^{(i)}$, where $\xi_n^{(i)}$ is zero-mean and σ-sub-Gaussian. Assume that $\|\mathbf{x}\|_2 \leq L$ for all $\mathbf{x} \in \mathcal{X}^{(i)}$ and $\|\mathbf{w}^{(i)}\|_2 \leq S$. Let $\widehat{\mathbf{w}}_n^{(i)}$ be the constrained ridge estimator and $V_n = \lambda I_d + \sum_{s=1}^n \mathbf{x}_s \mathbf{x}_s^\top$. For any $\delta \in (0,1)$, define $\beta_n(\delta) = \sigma\sqrt{d\log\left(\frac{N_{\max}(1+nL^2/\lambda)}{\delta}\right)} + \sqrt{\lambda}S$. Then*

$$\mathbb{P}\left(\forall\, 1 \leq n \leq N_{\max}, \left\|\widehat{\mathbf{w}}_n^{(i)} - \mathbf{w}^{(i)}\right\|_{V_n} \leq \beta_n(\delta)\right) \geq 1 - \delta.$$

Define the empirically optimal arm $\widehat{\mathbf{x}}_n^{(i)} = \arg\max_{\mathbf{x}\in\mathcal{X}^{(i)}} \mathbf{x}^\top \widehat{\mathbf{w}}_n^{(i)}$, if

$$\left(\widehat{\mathbf{x}}_n^{(i)} - \mathbf{x}\right)^\top \widehat{\mathbf{w}}_n^{(i)} > \beta_n(\delta)\left\|\widehat{\mathbf{x}}_n^{(i)} - \mathbf{x}\right\|_{V_n^{-1}}, \qquad \forall\, \mathbf{x} \in \mathcal{X}^{(i)} \setminus \{\widehat{\mathbf{x}}_n^{(i)}\}, \tag{7}$$

then $\widehat{\mathbf{x}}_n^{(i)}$ is the true preference-optimal arm.

Remark 1. *Unlike standard linear bandits with sampled stochastic rewards, each arm in our setting has a fixed LLM-generated overall evaluation $R_{\mathbf{x}}^{(i)}$. We assume $R_{\mathbf{x}}^{(i)} = \mathbf{x}^\top \mathbf{w}^{(i)} + \eta_{\mathbf{x}}^{(i)}$ and $y_n^{(i)} = R_{\mathbf{x}_n}^{(i)} + \epsilon_n^{(i)} = \mathbf{x}_n^\top \mathbf{w}^{(i)} + \xi_n^{(i)}$. Proposition 1 here relies on the assumption that the combined residual $\xi_n^{(i)} = \eta_{\mathbf{x}_n}^{(i)} + \epsilon_n^{(i)}$ is zero-mean σ-sub-Gaussian.*

After estimating $\widehat{\mathbf{w}}^{(i)}$, we construct $\widehat{\mathbf{x}}_{\text{emp}}^{(i)}$ from the period-wise optimal candidate policies $\widehat{\pi}_t^{(i)}$. Since $\widehat{\mathbf{x}}_{\text{emp}}^{(i)}$ does not necessarily correspond to an executable policy, we consider two candidate policies: $\pi_a^{(i)}$, whose average feedback profile is closest to $\widehat{\mathbf{x}}_{\text{emp}}^{(i)}$, and $\pi_b^{(i)}$, which achieves the highest average estimated preference utility. The detailed construction is provided in Appendix C. The final selected policy is determined by

$$\pi_{\text{sel}}^{(i)} = \begin{cases} \pi_b^{(i)}, & \text{if } \overline{Z}_{\pi_b^{(i)}} \geq \overline{Z}_{\pi_a^{(i)}} - \Delta_R \text{ and } \overline{U}_i(\pi_b^{(i)}) \geq \rho\, \overline{U}_i(\pi_a^{(i)}), \\ \pi_a^{(i)}, & \text{otherwise,} \end{cases} \tag{8}$$

where $\overline{Z}_\pi$ and $\overline{U}_i(\pi)$ denote the average realized return and estimated preference utility. Δ_R and ρ are the return-tolerance threshold and utility threshold, respectively.

4.3 Personalized Policy Update via PPO

After obtaining the selected policy $\pi_{\text{sel}}^{(i)}$ for each investor type, we further optimize the portfolio policy using PPO (Schulman et al., 2017). Since the selected policies may originate from base algorithms with incompatible actor architectures, we transfer each selected policy into a common PPO actor through behavior cloning. For investor type i, let $\mathcal{D}_i^{\text{sel}} = \left\{(s_\ell, \mathbf{a}_\ell^{\text{sel}})\right\}_{\ell=1}^{N_i^{\text{sel}}}$ denote the state–action pairs collected by executing $\pi_{\text{sel}}^{(i)}$. The PPO initialization is obtained by

$$\theta_0^{(i)} = \arg\min_\theta \mathbb{E}_{(s,\mathbf{a}^{\text{sel}})\sim\mathcal{D}_i^{\text{sel}}}\left[\left\|\pi_\theta(s) - \mathbf{a}^{\text{sel}}\right\|_2^2\right]. \tag{9}$$

During subsequent PPO training, the reward model $f^{(i)}$ and preference vector $\widehat{\mathbf{w}}^{(i)}$ are fixed. Given the portfolio metrics $\mathbf{m}_t$ computed from the trajectory, the personalized step-wise reward is

$$J_{t+1}^{(i)} = \gamma^t Z_{t+1} + \beta_{\text{pref}} \Delta U_{t+1}^{(i)} - \eta_{\text{ref}} \left\|\pi_\theta(s_t) - \pi_{\theta_0^{(i)}}(s_t)\right\|_2^2, \tag{10}$$

where $\Delta U_{t+1}^{(i)} = \left(\widehat{\mathbf{w}}^{(i)}\right)^\top \left[f^{(i)}(\mathbf{m}_{t+1}) - f^{(i)}(\mathbf{m}_t)\right]$. The last term limits deviation from the initialized policy for each investor type .

Algorithm 3 Personalized Policy Update

Require: Reward models $\{f^{(i)}\}_{i\in\mathcal{I}}$, preference vectors $\{\widehat{\mathbf{w}}^{(i)}\}_{i\in\mathcal{I}}$, initialized policies $\{\pi_{\theta_0^{(i)}}\}_{i\in\mathcal{I}}$, market states $\{s_t\}_{t=0}^{T-1}$, maximum steps $N_{\max}$, parameters $\beta_{\text{pref}}, \gamma, \eta_{\text{ref}}, \epsilon$
Ensure: Personalized policies $\{\pi_{\theta^{(i)}}\}_{i\in\mathcal{I}}$

for all $i \in \mathcal{I}$ **do**
  Initialize $\theta \leftarrow \theta_0^{(i)}$;
  for $n = 1, \ldots, N_{\max}$ **do**
    Set $\theta_{\text{old}} \leftarrow \theta$;
    for $t = 0, \ldots, T-1$ **do**
      Sample $a_t \sim \pi_{\theta_{\text{old}}}(\cdot \mid s_t)$ and obtain portfolio return Z_{t+1} in Eq. (2);
      Compute reward function $J_{t+1}^{(i)}$ using Eq. (10);
    end for
    Estimate advantages $\{A_t^{(i)}\}_{t=1}^{T}$ from $\{J_{t+1}^{(i)}\}_{t=1}^{T}$;
    Update $\theta \leftarrow \arg\max_\theta \mathcal{L}_{\text{PPO}}(\theta)$ using Eq. (11);
  end for
  Set $\theta^{(i)} \leftarrow \theta$;
end for
return $\{\pi_{\theta^{(i)}}\}_{i\in\mathcal{I}}$.

The resulting rewards are used to estimate the advantages $A_t^{(i)}$, θ is updated using the PPO clipped objective (Schulman et al., 2017),

$$\mathcal{L}_{\text{PPO}}(\theta) = \mathbb{E}_t\left[\min\left(\rho_{t,\theta}A_t^{(i)}, \text{clip}\left(\rho_{t,\theta}, 1-\epsilon, 1+\epsilon\right)A_t^{(i)}\right)\right], \tag{11}$$

where $\rho_{t,\theta} = \pi_\theta(a_t \mid s_t)/\pi_{\theta_{\text{old}}}(a_t \mid s_t)$, and different investor types are optimized independently.

5 Experiments

In this experiments, we evaluate IMOPA through three analyses: comparison between bandit-based and direct ridge preference estimation, benchmarking against recent portfolio methods, and ablation studies. The results show that IMOPA provides more stable preference-based policy selection, preserves investor-specific behavioral differences, and remains competitive in out-of-sample performance.

5.1 Experimental Setup

We evaluate IMOPA on the S&P 500 and MidCap100 datasets, with historical price data obtained from Yahoo Finance. After data screening, the two datasets contain 437 and 100 stocks, respectively, and are used for candidate-policy training, preference learning, and out-of-sample testing detailed in Table C.2. Each investment episode contains 63 trading days and starts with an initial wealth normalized to 1. We use a transaction cost of 0.1%. For IMOPA, the number of held assets is restricted to $[2, 15]$, with each asset weight capped at 50%. The candidate pool is constructed from A2C, PPO, DDPG, SAC, and TD3 under five random seeds and five training objectives—return, low risk, low turnover, concentrated active, and downside protection—to induce behavioral diversity. Candidate policies are then evaluated by Qwen 3.5 under the four investor profiles, returning five-objective feedback mentioned in Eq. (4), together with an overall evaluation for bandit-based estimation. Reward models are selected by leave-one-period-out validation from polynomial ridge regression with degrees 1—3 and GBDT. During policy optimization, we impose a maximum daily turnover of 0.2 and use 10^{-3} as the holding threshold. Additional candidate-policy and preference-estimation settings are provided in Appendix C.

5.2 Experimental Results

Preference estimation and validation. Figure 2 presents the bandit-estimated preference weights for the four personas. For PP, greater weights are assigned to risk and drawdown control, whereas AA places relatively greater weight on profitability and trading activity. The preference weights

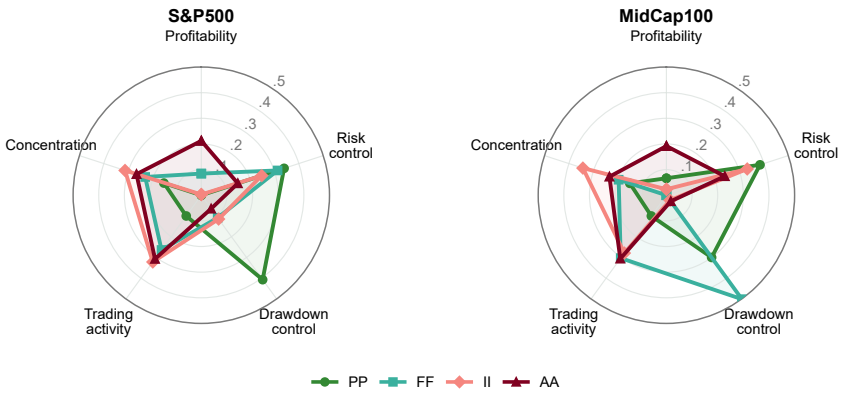

Figure 2: Bandit-estimated preference weights across investor profiles.

are estimated separately on the two datasets, based on different sets of candidate policies and their evaluations. As a result, the estimated values differ across datasets, while the investor-specific patterns for weights remain broadly consistent. Moreover, We compare the bandit-based estimates with those obtained by direct ridge estimation. Figure 3 reports the differences between two estimation methods, and Table F.3 shows that the two methods select the same policy in only one of eight cases. Tables F.4–F.5 further show that policies selected by the bandit-based method achieve better performance across the evaluated cases. Overall, these results indicate that the proposed bandit-based procedure produces preference-consistent estimates and leads to more stable policy allocation than direct ridge estimation.

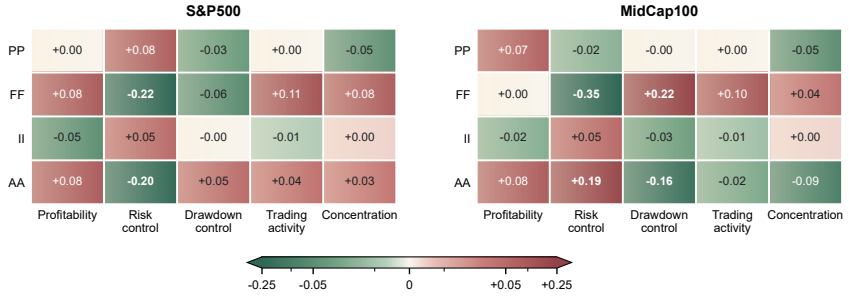

Figure 3: Preference-weight differences between bandit and direct ridge estimation; positive values indicate larger bandit-estimated weights.

We evaluate both validation of reward modeling and preference mapping using leave-one-period-out validation. Reward models are assessed against the five-objective scores, while overall-score mappings are evaluated against the normalized overall evaluations using Constant Mean, Polynomial Ridge, and Linear BAI. As reported in Tables F.1 and F.2, the reward models achieve low MAE and high R^2 on both datasets, while BAI-based estimation achieves the lowest prediction error than the alternative mappings..

Personalized portfolio behavior and performance. The learned preferences translate into distinct portfolio behaviors as depicted in Figure F.1. Passive profiles like PP and FF maintain more diversified allocations, whereas more active profiles of II and AA exhibit greater concentration.

Figure 4 compares the out-of-sample cumulative returns and maximum drawdown of the personalized policies with the fixed-pool equal-weight benchmark. The policies occupy different risk–return positions across both datasets, and their relative financial outcomes vary with the market environment. For example, AA performs strongly on the S&P 500 but less favorably on MidCap100, whereas FF achieves stronger performance on the latter. Table 1 compares our personalized policies with three recent learning-based portfolio methods, including DeepAries (Kim et al., 2025), NCMDP (Nägele et al., 2024), and SIT (Hwang & Zohren, 2026), in terms of financial criteria.

Table 1: Out-of-sample financial and behavioral performance on the S&P 500 and MidCap100 datasets. MaxDD denotes the absolute value of maximum drawdown magnitude.

	S&P 500					MidCap100				
Method	FinWealth ↑	MaxDD ↓	Sharpe ↑	Turnover	MaxWeight	FinWealth ↑	MaxDD ↓	Sharpe ↑	Turnover	MaxWeight
PP	1.205	0.185	0.620	0.075	0.114	1.228	0.420	0.432	0.019	0.077
FF	1.323	0.186	0.750	0.194	0.278	**1.596**	0.478	0.717	0.039	0.074
II	1.556	0.201	1.190	0.177	0.243	1.209	0.462	0.405	0.200	0.380
AA	**1.608**	0.171	**1.398**	0.187	0.356	1.067	0.498	0.273	0.071	0.500
DeepAries	1.298	**0.125**	0.889	0.049	0.072	1.584	0.337	**0.855**	0.048	0.072
SIT	1.317	0.142	0.868	0.005	0.200	1.389	0.416	0.595	0.016	0.331
NCMDP	1.221	0.213	0.450	0.558	0.500	1.489	**0.297**	0.716	0.490	0.500

* Bold values indicate the best result for metrics, Turnover and MaxWeight are reported as behavioral characteristics rather than ranked metriccs. Full results are provided in Tables F.6 and F.7.

Across both datasets, our policies remain competitive while exhibiting investor-specific differences in turnover and portfolio concentration that are consistent with the learned preferences. Comparison baseline, full evaluation metrics and additional results are reported in Appendices D–F.

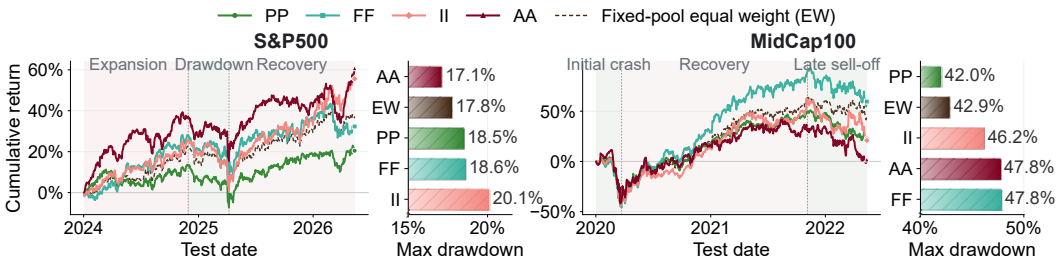

Figure 4: Out-of-sample cumulative returns and absolute value of maximum drawdown of personalized policies and the fixed-pool equal-weight benchmark.

Ablation study. We evaluate three variants of IMOPA: IMOPA-NPR removes the preference reward by setting $\beta_{\text{pref}} = 0$; IMOPA-UW replaces preference weights with uniform weights; and IMOPA-SP uses a shared initialization policy across all personas. As shown in Figure 5, IMOPA-SP produces the most pronounced contraction in behavioral differences, while IMOPA-NPR and IMOPA-UW also alter the resulting portfolio profiles. These results support the roles of both investor-specific initialization and learned preference information in generating personalized portfolio behavior. Full ablation results can be found in Tables F.8 and F.9.

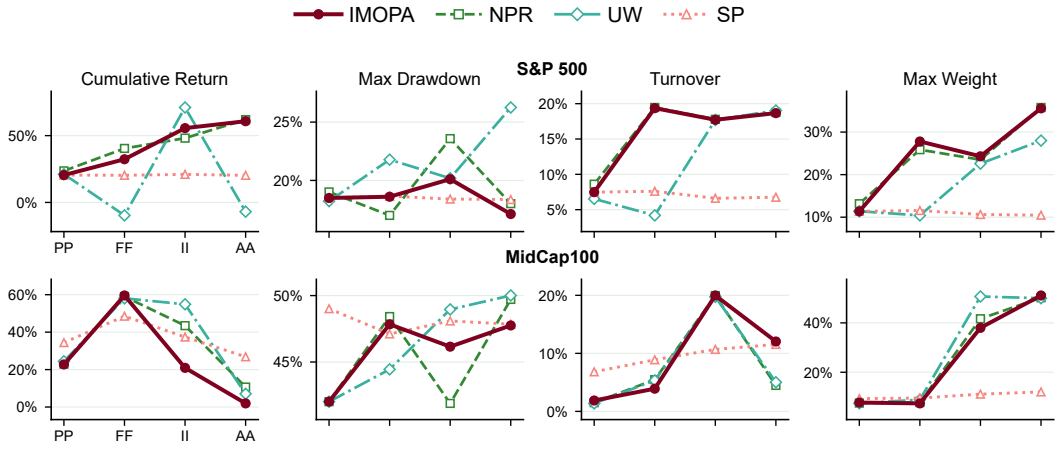

Figure 5: Ablation of preference-related components on datasets. Maximum drawdown is absolute.

AI USE STATEMENT

AI Use Statement. Generative AI tools were used to assist with language editing and migration of the manuscript to the ICLR 2027 LaTeX template. In addition, a Qwen-based language model was used as an experimental evaluator to generate the simulated persona-conditioned feedback described in Section 5. All methodological decisions, analyses, experiments, and claims were reviewed and verified by the authors.

REFERENCES

Abbas Abdolmaleki, Sandy Huang, Leonard Hasenclever, Michael Neunert, Francis Song, Martina Zambelli, Murilo Martins, Nicolas Heess, Raia Hadsell, and Martin Riedmiller. A distributional view on multi-objective policy optimization. In *International Conference on Machine Learning*, pp. 11–22, 2020.

Michael W. Brandt, Amit Goyal, Pedro Santa-Clara, and Jonathan R. Stroud. A simulation approach to dynamic portfolio choice with an application to learning about return predictability. *The Review of Financial Studies*, 18(3):831–873, 2005.

Ying-Tu Chen, Wei Hung, Bing-Shu Wu, Zhang-Wei Hong, and Ping-Chun Hsieh. A reward-free viewpoint on multi-objective reinforcement learning. In *International Conference on Learning Representations*, pp. 1–37, 2026.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, pp. 4299–4307, 2017.

Giovanni Dispoto, Marcello Restelli, and Carmine Ventre. Eliciting esg preferences for reinforcement learning-based portfolio optimization. *arXiv preprint arXiv:2609.02677*, pp. 1–8, 2026.

Daniel Dorn and Gur Huberman. Talk and action: What individual investors say and what they do. *Review of Finance*, 9(4):437–481, 2005.

Mark Egan, Alexander MacKay, and Hanbin Yang. What drives variation in investor portfolios? estimating the roles of beliefs and risk preferences. *The Review of Financial Studies*, 39(10): 3276–3329, 2026.

Tom Hosking, Phil Blunsom, and Max Bartolo. Human feedback is not gold standard. In *International Conference on Learning Representations*, pp. 1–20, 2024.

Yoontae Hwang and Stefan Zohren. Signature-informed transformer for asset allocation. In *International Conference on Machine Learning*, pp. 1–21, 2026.

Zhengyao Jiang, Dixing Xu, and Jinjun Liang. A deep reinforcement learning framework for the financial portfolio management problem. *arXiv preprint arXiv:1706.10059*, pp. 1–31, 2017.

Jinkyu Kim, Hyungjung Yi, Mogan Gim, Donghee Choi, and Jaewoo Kang. DeepAries: Adaptive rebalancing interval selection for enhanced portfolio selection. In *ACM International Conference on Information and Knowledge Management*, pp. 5780–5787, 2025.

Harry Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952.

Robert C. Merton. An analytic derivation of the efficient portfolio frontier. *Journal of Financial and Quantitative Analysis*, 7(4):1851–1872, 1972.

Maximilian Nägele, Jan Olle, Thomas Fösel, Remmy Zen, and Florian Marquardt. Tackling decision processes with non-cumulative objectives using reinforcement learning. *arXiv preprint arXiv:2405.13609*, pp. 1–30, 2024.

Sriyash Poddar, Yanming Wan, Hamish Ivison, Abhishek Gupta, and Natasha Jaques. Personalizing reinforcement learning from human feedback with variational preference learning. In *Advances in Neural Information Processing Systems*, pp. 52516–52544, 2024.

Michael M. Pompian. *Behavioral Finance and Investor Types: Managing Behavior to Make Better Investment Decisions*. John Wiley & Sons, pp. 1–256, 2012.

Diederik M. Roijers, Peter Vamplew, Shimon Whiteson, and Richard Dazeley. A survey of multi-objective sequential decision-making. *Journal of Artificial Intelligence Research*, 48:67–113, 2013.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, pp. 1–12, 2017.

Joar Skalse and Alessandro Abate. On the limitations of Markovian rewards to express multi-objective, risk-sensitive, and modal tasks. In *Conference on Uncertainty in Artificial Intelligence*, pp. 1974–1984, 2023.

Marta Soare, Alessandro Lazaric, and Rémi Munos. Best-arm identification in linear bandits. In *Advances in Neural Information Processing Systems*, pp. 828–836, 2014.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, pp. 3008–3021, 2020.

Chao Tao, Saúl A. Blanco, and Yuan Zhou. Best arm identification in linear bandits with linear dimension dependency. In *International Conference on Machine Learning*, pp. 4877–4886, 2018.

Peter Vamplew, Richard Dazeley, Adam Berry, Rustam Issabekov, and Evan Dekker. Empirical evaluation methods for multiobjective reinforcement learning algorithms. *Machine Learning*, 84 (1–2):51–80, 2011.

Haoxiang Wang, Yong Lin, Wei Xiong, Rui Yang, Shizhe Diao, Shuang Qiu, Han Zhao, and Tong Zhang. Arithmetic control of LLMs for diverse user preferences: Directional preference alignment with multi-objective rewards. In *Annual Meeting of the Association for Computational Linguistics*, pp. 8642–8655, 2024.

Kaiwen Wang, Dawen Liang, Nathan Kallus, and Wen Sun. A reductions approach to risk-sensitive reinforcement learning with optimized certainty equivalents. In *International Conference on Machine Learning*, pp. 63636–63661, 2025.

Zhicheng Wang, Biwei Huang, Shikui Tu, Kun Zhang, and Lei Xu. DeepTrader: A deep reinforcement learning approach for risk-return balanced portfolio management with market conditions embedding. In *AAAI Conference on Artificial Intelligence*, pp. 643–650, 2021.

Nuoya Xiong and Aarti Singh. Projection optimization: A general framework for multi-objective and multi-group RLHF. In *International Conference on Machine Learning*, pp. 69017–69058, 2025.

Haohan Xu, Jason Bohne, Pawel Polak, David Byrd, David Rosenberg, and Gary Kazantsev. Learning to trade with preferences: Interpretable execution via mixture-of-experts. In *ACM International Conference on AI in Finance*, pp. 762–770, 2025.

Wenyi Zhang, Renjun Jia, Yanhao Wang, Dawei Cheng, Minghao Zhao, and Cen Chen. Enhancing portfolio optimization via heuristic-guided inverse reinforcement learning with multi-objective reward and graph-based policy learning. In *International Joint Conference on Artificial Intelligence*, pp. 9483–9491, 2025.

Yifan Zhang, Peilin Zhao, Qingyao Wu, Bin Li, Junzhou Huang, and Mingkui Tan. Cost-sensitive portfolio selection via deep reinforcement learning. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):236–248, 2022.

Yifan Zhong, Chengdong Ma, Xiaoyuan Zhang, Ziran Yang, Haojun Chen, Qingfu Zhang, Siyuan Qi, and Yaodong Yang. Panacea: Pareto alignment via preference adaptation for LLMs. In *Advances in Neural Information Processing Systems*, pp. 75522–75558, 2024.

The appendix is organized as follows. Appendix A presents the conclusion, and Appendix B provides the proof of Proposition 1. Appendix C describes the experimental settings and implementation details, while Appendix D introduces the three recent learning-based portfolio methods used as comparison benchmarks. Appendix E details the computation of the financial criteria used for out-of-sample evaluation, and Appendix F provides the additional figures and tables.

A Conclusion

We introduced IMOPA to learn heterogeneous investor preferences from multi-objective feedback and generate personalized portfolio policies. The framework integrates LLM-based preference elicitation, bandit-based preference estimation and policy selection, and personalized PPO optimization. Experiments show that different investor profiles lead to distinct preference estimates, selected policies, and portfolio allocations. Ablation results further indicate that both learned preference weights and investor-specific policy initialization contribute to this differentiation. Overall, IMOPA produces preference-consistent rather than universally dominating policies, with portfolio behavior reflecting investor priorities across market environments. Future work may extend the framework to richer forms of investor feedback and broader market settings.

B Proof of Proposition 1

Proof. For each $n \le N_{\max}$, let

$$\widetilde{\mathbf{w}}_n^{(i)} = V_n^{-1} \sum_{s=1}^{n} \mathbf{x}_s y_s^{(i)}$$

denote the unconstrained ridge estimator. Set $\delta' = \delta / N_{\max}$. By the regularized least-squares confidence bound, for each $n \le N_{\max}$,

$$\mathbb{P}\left(\left\| \widetilde{\mathbf{w}}_n^{(i)} - \mathbf{w}^{(i)} \right\|_{V_n} \le \beta_n(\delta) \right) \ge 1 - \frac{\delta}{N_{\max}},$$

where

$$\beta_n(\delta) = \sigma \sqrt{d \log\left(\frac{N_{\max}(1 + nL^2/\lambda)}{\delta} \right)} + \sqrt{\lambda} S.$$

A union bound over $n = 1, \dots, N_{\max}$ therefore gives

$$\mathbb{P}\left(\forall\, 1 \le n \le N_{\max}, \left\| \widetilde{\mathbf{w}}_n^{(i)} - \mathbf{w}^{(i)} \right\|_{V_n} \le \beta_n(\delta) \right) \ge 1 - \delta.$$

The constrained ridge estimator $\widehat{\mathbf{w}}_n^{(i)}$ is the V_n-norm projection of $\widetilde{\mathbf{w}}_n^{(i)}$ onto the feasible set $\mathcal{W}$. Since $\mathbf{w}^{(i)} \in \mathcal{W}$, the projection property implies

$$\left\| \widehat{\mathbf{w}}_n^{(i)} - \mathbf{w}^{(i)} \right\|_{V_n} \le \left\| \widetilde{\mathbf{w}}_n^{(i)} - \mathbf{w}^{(i)} \right\|_{V_n}.$$

Hence, with probability at least $1 - \delta$,

$$\left\| \widehat{\mathbf{w}}_n^{(i)} - \mathbf{w}^{(i)} \right\|_{V_n} \le \beta_n(\delta), \qquad \forall\, 1 \le n \le N_{\max}.$$

On this confidence event, for any competing arm $\mathbf{x} \ne \widehat{\mathbf{x}}_n^{(i)}$, the generalized Cauchy–Schwarz inequality gives

$$\left| \left(\widehat{\mathbf{x}}_n^{(i)} - \mathbf{x} \right)^\top \left(\widehat{\mathbf{w}}_n^{(i)} - \mathbf{w}^{(i)} \right) \right| \le \beta_n(\delta) \left\| \widehat{\mathbf{x}}_n^{(i)} - \mathbf{x} \right\|_{V_n^{-1}}.$$

Therefore,

$$\begin{aligned} \left(\widehat{\mathbf{x}}_n^{(i)} - \mathbf{x} \right)^\top \mathbf{w}^{(i)} &= \left(\widehat{\mathbf{x}}_n^{(i)} - \mathbf{x} \right)^\top \widehat{\mathbf{w}}_n^{(i)} \\ &\quad - \left(\widehat{\mathbf{x}}_n^{(i)} - \mathbf{x} \right)^\top \left(\widehat{\mathbf{w}}_n^{(i)} - \mathbf{w}^{(i)} \right) \\ &\ge \left(\widehat{\mathbf{x}}_n^{(i)} - \mathbf{x} \right)^\top \widehat{\mathbf{w}}_n^{(i)} - \beta_n(\delta) \left\| \widehat{\mathbf{x}}_n^{(i)} - \mathbf{x} \right\|_{V_n^{-1}}. \end{aligned}$$

Table C.2: Dataset information and temporal partitions used in the experiments.

Dataset	Stocks	Policy Training	Preference Learning	Testing
S&P 500	437	01/01/2015–31/12/2021	01/01/2022–31/12/2023	01/01/2024–13/05/2026
MidCap100	100	03/01/2011–29/12/2017	01/01/2018–31/12/2019	02/01/2020–13/05/2022

□

C Additional Material for Experimental Setting

Dataset and temporal settings. The temporal partitions are summarized in Table C.2.

Construction of Candidate Policies Given the estimated preference vector $\widehat{\mathbf{w}}^{(i)}$, the preference-optimal candidate policy in period t is

$$\widehat{\pi}_t^{(i)} = \arg\max_{\pi \in \Pi_{\text{cand}}} (\mathbf{x}_{\pi,t}^{(i)})^\top \widehat{\mathbf{w}}^{(i)}. \tag{C.1}$$

The corresponding feedback vectors define the cross-period target

$$\widehat{\mathbf{x}}_{\text{emp}}^{(i)} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{x}_{\widehat{\pi}_t^{(i)},t}^{(i)}. \tag{C.2}$$

For each executable candidate policy, let

$$\bar{\mathbf{x}}_\pi^{(i)} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{x}_{\pi,t}^{(i)}.$$

We then construct two teacher candidates:

$$\pi_a^{(i)} = \arg\min_{\pi \in \Pi_{\text{cand}}} \left\| (\bar{\mathbf{x}}_\pi^{(i)} - \widehat{\mathbf{x}}_{\text{emp}}^{(i)}) \odot \sqrt{\widehat{\mathbf{w}}^{(i)}} \right\|_2, \tag{C.3}$$

and

$$\pi_b^{(i)} = \arg\max_{\pi \in \Pi_{\text{cand}}} (\bar{\mathbf{x}}_\pi^{(i)})^\top \widehat{\mathbf{w}}^{(i)}. \tag{C.4}$$

Thus, $\pi_a^{(i)}$ approximates the synthetic multi-objective target, whereas $\pi_b^{(i)}$ maximizes the estimated preference utility.

Candidate Policy Training Objectives. To construct a behaviorally diverse candidate policy pool, we train A2C, PPO, DDPG, SAC, and TD3 under five portfolio objectives:

$$\mathcal{O} = \{\text{return}, \text{low risk}, \text{low turnover}, \text{concentrated active}, \text{downside protection}\}.$$

Each algorithm–objective configuration is trained independently under five random seeds with a learning rate of 3×10^{-4}. For trading day t, let

$$z_t = \log \frac{W_{t+1}}{W_t},$$

denote the transaction-cost-adjusted log portfolio return, where W_t is the portfolio wealth at the beginning of day t. Let

$$\mathbf{a}_t = (a_{t,1}, \ldots, a_{t,J})^\top,$$

denote the portfolio weights. The daily turnover and portfolio concentration are defined as

$$\tau_t = \sum_{j=1}^{J} |a_{t+1,j} - a_{t,j}|, \qquad \text{HHI}_t = \sum_{j=1}^{J} a_{t,j}^2.$$

Let DD_t denote the cumulative portfolio drawdown observed up to day t, and let $[x]_+ = \max(x, 0)$. The five training objectives are

$$J_{\text{return},t} = z_t, \tag{C.5}$$

$$J_{\text{low risk},t} = z_t - \lambda_r z_t^2, \tag{C.6}$$

$$J_{\text{low turnover},t} = z_t - \lambda_\tau \tau_t, \tag{C.7}$$

$$J_{\text{concentrated active},t} = z_t + \lambda_H \mathrm{HHI}_t + \lambda_A \tau_t, \tag{C.8}$$

$$J_{\text{downside protection},t} = z_t - \lambda_1 [-z_t]_+ - \lambda_2 [-z_t]_+^2 - \lambda_D [\mathrm{DD}_t - \overline{D}]_+. \tag{C.9}$$

Here, all λ parameters are non-negative.

LLM evaluation settings. The four behavioral investor profiles are instantiated through investor-specific LLM evaluation prompts. To avoid revealing candidate-policy identity during evaluation, policy identifiers are anonymized and Qwen 3.5 is not informed of the underlying learning algorithm, random seed, or training objective.

Preference-estimation settings. For each investor profile, the five-objective feedback scores and the overall evaluation are normalized to $[0, 1]$. The scalar observation used for preference estimation is perturbed by $\epsilon \sim \mathcal{N}(0, 0.05^2)$. We set the maximum number of estimation steps to $N_{\max} = 500$ and the ridge parameter to $\lambda = 10^{-6}$. For the gate rule in Eq. (8), we use $\Delta_R = 0.02$ and $\rho = 0.85$ throughout the experiments.

D Comparison Baselines

We compare IMOPA with three recent learning-based portfolio allocation methods.

DeepAries (Kim et al., 2025) combines a Transformer-based market encoder with PPO to jointly determine portfolio allocations and adaptive rebalancing decisions.

NCMDP (Nägele et al., 2024) provides a framework for optimizing non-cumulative objectives with standard reinforcement learning algorithms. Its portfolio application enables direct optimization of risk-adjusted criteria such as the Sharpe ratio.

SIT (Hwang & Zohren, 2026) is an end-to-end asset-allocation framework that incorporates path signatures into a Transformer architecture and directly optimizes a risk-aware portfolio objective.

For each baseline, we randomly construct 10 asset subsets containing 15 stocks each. Each model is independently trained and evaluated on every subset using the same temporal partition and evaluation protocol as in the main experiments. For each performance metric, the reported baseline result is the average over the 10 subsets.

An equally weighted portfolio over the corresponding stock pool is also used as a reference policy over the full testing period.

E Evaluation Metric Definitions

For a policy evaluated over the T trading periods, denote its wealth at period t by W_t, with $W_0 = w_0$. Given the log portfolio return Z_{t+1} in Eq. (2), the corresponding simple return is

$$q_{t+1} = \frac{W_{t+1}}{W_t} - 1 = e^{Z_{t+1}} - 1. \tag{E.1}$$

The reported performance measures are then defined as follows, using the same portfolio weights $\mathbf{a}_t$ and returns Z_{t+1} as in the experiments:

$$W_t = w_0 \prod_{s=0}^{t-1}(1 + q_{s+1}), \tag{E.2}$$

$$\text{FinWealth} = W_T, \tag{E.3}$$

$$\text{Vol} = \left[\frac{252}{T-1}\sum_{t=0}^{T-1}(q_{t+1} - \bar{q})^2\right]^{1/2}, \tag{E.4}$$

$$|\text{MaxDD}| = \max_{0\le t\le T}\left(1 - \frac{W_t}{\max_{0\le s\le t} W_s}\right), \tag{E.5}$$

$$\text{Sharpe} = \frac{\sqrt{252}\,\bar{q}}{\left[\frac{1}{T-1}\sum_{t=0}^{T-1}(q_{t+1} - \bar{q})^2\right]^{1/2}}, \tag{E.6}$$

$$\text{Turnover} = \frac{1}{T-1}\sum_{t=1}^{T-1}\sum_{j=1}^{N}\left|a_{t+1}^j - a_t^j\right|, \tag{E.7}$$

$$\text{EffHold} = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{\sum_{j=1}^{N}(a_t^j)^2}, \tag{E.8}$$

$$\text{MaxWeight} = \max_{1\le t\le T,\, 1\le j\le N} a_t^j, \tag{E.9}$$

where $\bar{q} = T^{-1}\sum_{t=0}^{T-1} q_{t+1}$ is the average daily simple return and 252 is the annualization factor. The turnover calculation excludes the initial deployment from the ongoing rebalancing statistic, consistent with the experimental implementation. MaxDD is reported as a non-negative drawdown magnitude, so smaller values indicate better drawdown control; the Sharpe ratio uses a zero risk-free rate.

F Additional Tables and Figures

Table F.1: Held-out Reward Model performance across the two datasets. MAE and R^2 are computed against the five Qwen dimension scores on the original 1–10 scale.

	S&P 500		MidCap100	
Persona	MAE	R^2	MAE	R^2
PP	0.607	0.885	0.603	0.881
FF	0.628	0.876	0.758	0.800
II	0.749	0.789	0.864	0.705
AA	0.781	0.774	0.937	0.687

Table F.2: Held-out performance of overall-score mappings containing Constant Mean, linear BAI (our method) and Polynomial Ridge mapping across the two datasets. MAE is computed against the normalized Qwen overall evaluation in $[0, 1]$.

	S&P 500			MidCap100		
Persona	Constant Mean	Linear	Polynomial Ridge	Constant Mean	Linear	Polynomial Ridge
PP	0.174	0.095	0.103	0.166	0.105	0.116
FF	0.162	0.101	0.114	0.181	0.109	0.095
II	0.165	0.136	0.155	0.151	0.129	0.136
AA	0.204	0.170	0.190	0.217	0.153	0.170

Table F.3: Selected strategies from bandit-based and direct ridge estimations on the S&P 500 and MidCap100 datasets.

Dataset	Persona	Bandit-Bbased	Ridge
S&P 500	PP	`ddpg return seed44`	`ddpg return seed44`
	FF	`a2c return seed44`	`ddpg low turnover seed43`
	II	`sac concentrated active seed42`	`ppo return seed45`
	AA	`td3 concentrated active seed44`	`ppo concentrated active seed46`
MidCap100	PP	`ddpg low risk seed42`	`ddpg downside protection seed45`
	FF	`ddpg return seed44`	`ddpg low risk seed43`
	II	`sac downside protection seed43`	`sac concentrated active seed42`
	AA	`ddpg downside protection seed42`	`sac low risk seed45`

Table F.4: Out-of-sample performance comparison between bandit-based and direct ridge preference estimation on the S&P 500 dataset.

Method	FinWealth ↑	Vol↓	MaxDD↓	Sharpe↑	Turnover	EffHold	MaxWeight
Bandit–PP	1.236	**0.145**	0.190	0.694	0.086	14.183	0.132
Bandit–FF	1.404	0.181	**0.170**	0.890	0.195	10.063	0.259
Bandit–II	1.481	0.170	0.236	1.070	0.178	11.579	0.235
Bandit–AA	**1.619**	0.153	0.180	**1.417**	0.188	10.094	0.358
Ridge–PP	1.214	**0.145**	0.185	0.642	0.068	14.566	0.109
Ridge–FF	0.902	0.156	0.217	-0.203	0.046	14.843	0.100
Ridge–II	1.507	0.174	0.235	1.092	0.177	11.525	0.249
Ridge–AA	1.031	0.183	0.237	0.162	0.195	6.747	0.497

Table F.5: Out-of-sample performance comparison between bandit-based and direct ridge preference estimation on the MidCap100 dataset.

Method	FinWealth ↑	Vol↓	MaxDD↓	Sharpe↑	Turnover	EffHold	MaxWeight
Bandit–PP	1.228	**0.324**	**0.420**	0.432	0.019	14.977	0.077
Bandit–FF	**1.596**	0.374	0.478	**0.717**	0.039	14.805	0.074
Bandit–II	1.209	0.364	0.462	0.405	0.200	12.236	0.380
Bandit–AA	1.019	0.390	0.478	0.219	0.200	8.039	0.511
Ridge–PP	1.342	0.341	0.483	0.537	0.058	14.451	0.097
Ridge–FF	1.478	0.309	0.433	0.690	0.052	14.158	0.093
Ridge–II	1.076	0.376	0.551	0.273	0.197	12.236	0.197
Ridge–AA	0.937	0.326	0.472	0.082	0.200	8.039	0.413



Figure F.1: Portfolio allocations of personas on the S&P 500 and MidCap100 datasets.

Table F.6: Out-of-sample performance comparison on the S&P 500 dataset. MaxDD is reported as its absolute drawdown magnitude.

Method	FinWealth ↑	Vol↓	MaxDD ↓	Sharpe↑	Turnover	EffHold	MaxWeight
PP	1.205	0.145	0.185	0.620	0.075	14.487	0.114
FF	1.323	0.181	0.186	0.750	0.194	9.665	0.278
II	1.556	0.171	0.201	1.190	0.177	11.577	0.243
AA	**1.608**	0.153	0.171	**1.398**	0.187	10.262	0.356
DeepAries	1.298	0.143	**0.125**	0.889	0.049	14.975	0.072
SIT	1.317	**0.140**	0.142	0.868	0.005	10.564	0.200
NCMDP	1.221	0.169	0.213	0.450	0.558	2.824	0.500

* ∗ ■, ■, and ■ denote first, second, and third place, respectively.
** ∗∗ The evaluation metrics are final wealth, volatility, maximum drawdown, Sharpe ratio, turnover, effective number of holdings, and maximum single-asset weight. ↑ (↓) indicates that a larger (smaller) value is preferred.

Table F.7: Out-of-sample performance comparison on the MidCap100 dataset. MaxDD is reported as its absolute drawdown magnitude.

Method	FinWealth ↑	Vol↓	MaxDD ↓	Sharpe↑	Turnover	EffHold	MaxWeight
PP	1.228	0.324	0.420	0.432	0.019	14.976	0.077
FF	**1.596**	0.374	0.478	0.717	0.039	14.805	0.074
II	1.209	0.364	0.462	0.405	0.200	8.934	0.380
AA	1.067	0.405	0.498	0.273	0.071	2.181	0.500
DeepAries	1.584	0.305	0.337	**0.855**	0.048	14.979	0.072
SIT	1.389	0.315	0.416	0.595	0.016	8.292	0.331
NCMDP	1.489	**0.290**	**0.297**	0.716	0.490	2.603	0.500

Table F.8: Out-of-sample ablation results on the S&P 500 dataset.

Condition	Persona	FinWealth ↑	Vol ↓	MaxDD ↓	Sharpe ↑	Turnover	EffHold	MaxWeight
IMOPA-NPR	PP	1.236	0.145	0.190	0.694	0.086	14.183	0.132
	FF	1.404	0.181	**0.170**	0.890	0.195	10.063	0.259
	II	1.481	0.170	0.236	1.070	0.178	11.579	0.235
	AA	1.619	0.153	0.180	**1.417**	0.188	10.094	0.358
IMOPA-UW	PP	1.211	0.145	0.182	0.637	0.065	14.592	0.114
	FF	0.903	0.156	0.218	-0.201	0.042	14.875	0.104
	II	**1.711**	0.173	0.202	1.410	0.177	11.664	0.226
	AA	0.931	0.158	0.263	-0.115	0.190	10.906	0.280
IMOPA-SP	PP	1.205	0.145	0.185	0.620	0.075	14.487	0.114
	FF	1.203	0.145	0.187	0.616	0.076	14.496	0.116
	II	1.210	0.145	0.184	0.634	0.066	14.609	0.106
	AA	1.203	**0.144**	0.184	0.617	0.068	14.652	0.105

Table F.9: Out-of-sample ablation results on the MidCap100 dataset.

Condition	Persona	FinWealth ↑	Vol ↓	MaxDD ↓	Sharpe ↑	Turnover	EffHold	MaxWeight
IMOPA-NPR	PP	1.229	**0.324**	0.420	0.434	0.015	14.982	0.075
	FF	**1.590**	0.382	0.484	0.707	0.055	14.793	0.077
	II	1.435	0.336	**0.419**	0.624	0.200	8.965	0.417
	AA	1.107	0.406	0.497	0.312	0.045	2.063	0.502
IMOPA-UW	PP	1.244	**0.324**	0.420	0.449	0.014	14.992	0.074
	FF	1.581	0.336	0.444	**0.746**	0.053	14.385	0.089
	II	1.549	0.366	0.490	0.692	0.197	5.992	0.507
	AA	1.070	0.405	0.500	0.277	0.050	2.089	0.500
IMOPA-SP	PP	1.344	0.375	0.490	0.523	0.068	14.668	0.093
	FF	1.485	0.376	0.471	0.635	0.089	14.417	0.095
	II	1.374	0.373	0.481	0.548	0.107	13.029	0.111
	AA	1.268	0.371	0.478	0.458	0.115	12.161	0.120