A Generative Probabilistic Approach for Goal-Based Portfolio Optimization

Tessa Bauman
University of Zagreb, FER
Zagreb, Croatia
tessa.bauman@fer.unizg.hr

Sven Goluža
University of Zagreb, FER
Zagreb, Croatia
sven.goluza@fer.unizg.hr

Lovre Mrčela
University of Zagreb, FER
Zagreb, Croatia
lovre.mrcela@fer.unizg.hr

Stjepan Begušić
University of Zagreb, FER
Zagreb, Croatia
stjepan.begusic@fer.unizg.hr

Zvonko Kostanjčar
University of Zagreb, FER
Zagreb, Croatia
zvonko.kostanjcar@fer.unizg.hr

Abstract

Goal-based portfolio optimization seeks to design investment strategies that maximize the likelihood of achieving specific financial objectives. A major challenge in this domain is data scarcity and non-stationary market dynamics, which undermine the effectiveness of traditional approaches. To address this, we propose a generative modeling framework that integrates probabilistic regression with deep reinforcement learning. The probabilistic model estimates evolving market return distributions for state representation and generates realistic synthetic market trajectories, enabling the agent to train efficiently on diverse market scenarios and adapt to dynamic environments. Experiments on multi-asset historical data demonstrate that our approach achieves superior goal-attainment probabilities compared to established benchmarks, highlighting the value of synthetic market generation for robust goal-based portfolio optimization.

1 Introduction

Goal-based portfolio optimization (GBPO) is an investment approach focused on achieving specific financial objectives within a predefined time frame. Most approaches are centered around maximizing the likelihood of reaching personal financial goals, such as saving for retirement or purchasing a home. While popular, conventional strategies like target date funds often rely on pre-set glide paths that can be suboptimal in non-stationary market conditions [1]. This is where deep reinforcement learning (DRL) shows promise, offering a framework for learning the adaptive strategies that GBPO requires. However, DRL models are known to be sample inefficient, which presents a significant challenge given the scarcity of historical data in long-term financial planning.

To address these challenges, we propose a framework that integrates DRL with probabilistic regression (PR) that serves two functions: first, as a state estimator that provides the DRL agent with estimated market return distributions, and second, as a generative tool to create synthetic market data,

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Generative AI in Finance.

augmenting the historical dataset. Our approach builds upon several research streams. Analytical GBPO methods, including deterministic glide paths [1], utility maximization frameworks [2], and dynamic programming [3], often overlook market non-stationarity. DRL has proven effective across a range of financial tasks such as portfolio optimization [4, 5, 6, 7], trade execution [8, 9, 10], and market making [11, 12]. However, previous DRL applications to GBPO have been limited by assuming stationarity [13] or focusing on simplified two-asset portfolios [14]. Our contribution is a scalable DRL solution that adapts to non-stationary environments through this unified generative and state-estimation framework.

2 Methodology

The objective is to find the portfolio weights $\boldsymbol{w}_1^*, \boldsymbol{w}_2^*, \ldots, \boldsymbol{w}_T^*$ that maximize the probability that the portfolio value meets or exceeds a goal G over time T [15]. Formally, GBPO can be defined as: $\boldsymbol{w}_1^*, \ldots, \boldsymbol{w}_T^* = \arg\max_{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_T} P\left(W_T \geq G\right)$, subject to $\sum_{i=1}^n w_{i,t} = 1$, and $\boldsymbol{w}_t \geq \mathbf{0}$, where n is the number of assets, $\boldsymbol{w}_t = (w_{1,t}, \ldots, w_{n,t})^\top$ are the portfolio weights at time t, W_T is the wealth at T, T is the investment horizon, and G is the target wealth.

2.1 Probabilistic regression

In deep learning, a *deterministic* regression predicts a single value $\hat{y} = \mathbb{E}_\theta\left[y \mid \mathbf{x}\right]$ for the target y given features $\mathbf{x}$, and learns θ by minimizing a pointwise loss (typically squared error). A *probabilistic* regression instead predicts a conditional distribution $P_\theta(y \mid \mathbf{x})$. In our setting we use probabilistic regression to model returns $\boldsymbol{r}_t = (r_{1,t}, \ldots, r_{n,t})^\top$, with $t \in \{1, 2, \ldots, T\}$. Using the last return as input ($k = 1$), the network predicts the distribution of the next n-dimensional return vector:

$$P_\theta(\mathbf{r}_t \mid \mathbf{r}_{t-1}) = \mathcal{N}(\mathbf{r}_t; \boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t).$$

While we adopt a Gaussian specification for tractability, this choice is not essential and heavy-tailed return distributions can be incorporated within the same framework. Given $\mathbf{r}_{t-1}$, the neural network produces: an n-dimensional mean vector $\boldsymbol{\mu}_t$; an n-dimensional vector $\log \boldsymbol{\sigma}_t$ (and we set $\boldsymbol{\sigma}_t = \exp(\log \boldsymbol{\sigma}_t)$); a vector $\boldsymbol{\ell}_t'$ of length $n(n-1)/2$ for the values of the lower triangle. We form a lower-triangular matrix $\mathbf{L}_t$ by placing $\boldsymbol{\sigma}_t$ on the diagonal and $\boldsymbol{\ell}_t'$ below the diagonal, and then define $\boldsymbol{\Sigma}_t = \mathbf{L}_t \mathbf{L}_t^\top$, which guarantees a positive-definite covariance matrix [16]. The loss is the Gaussian negative log-likelihood, $\mathcal{L}_\theta(\mathbf{r}_t) = -\log \mathcal{N}(\mathbf{r}_t; \boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$, augmented with a regularization term. We use this network for two purposes: (i) to simulate trajectories that augment training data for the RL stage, and (ii) to provide part of the RL state (the nonzero entries of $\mathbf{L}_t$ together with $\boldsymbol{\mu}_t$) which will be discussed in 2.2.

2.2 Markov decision process

Deep reinforcement learning addresses sequential decision making by learning through interaction to maximize expected cumulative reward. The standard way to model it is by using a Markov Decision Process (MDP) $\langle S, A, P, R, \gamma \rangle$ with states, actions, transition dynamics (usually unknown), rewards, and a discount factor. We construct the MDP to suit GBPO and to capture market non-stationarity. The state space is therefore constructed with two types of features: goal-based and market-based ones. At time t, goal-based features are represented with: the elapsed-time ratio t/T and the achieved-goal ratio W_t/G (with W_t the current wealth and G the target). For the market-based features, the agent receives the outputs of the probabilistic regression model from Section 2.1. From that model, we include the expected mean returns $\mu_t^{a_1}, \ldots, \mu_t^{a_n}$ and the elements l_t^{jk} of the lower triangular matrix $\mathbf{L}_t$ that induce the covariance $\boldsymbol{\Sigma}_t$. Formally, the state S_t is defined with:

$$S_t = \left(t/T,\ W_t/G,\ \mu_t^{a_1}, \ldots, \mu_t^{a_n},\ l_t^{11}, l_t^{21}, \ldots, l_t^{nn}\right).$$

We note that as the number of assets increases, the state-space dimension grows quadratically; this growth should be taken into account, potentially via factor representations. The action is the portfolio weight vector over the n assets: $A_t = (w_t^{a_1}, w_t^{a_2}, \ldots, w_t^{a_n})$, where each weight $w_t^{a_i}$ lies in $[0, 1]$ (no short selling) and the weights sum to one at each time step. In the context of GBPO, the primary reward is binary in nature: the financial goal is either achieved, or it is not. Therefore, a terminal reward is granted only if the goal is reached: $R_T = \mathbf{1}_{\{W_T \geq G\}}$, and $R_t = 0$ for all $t < T$.

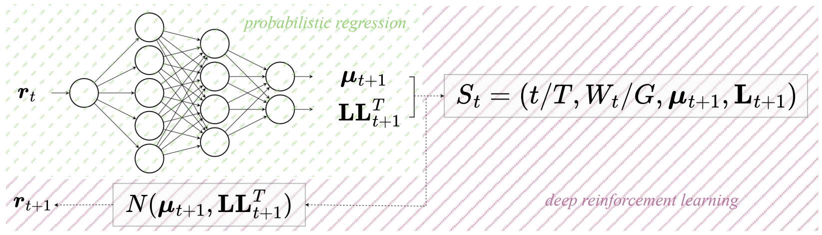

Figure 1: The data generation process intertwined with the definition of a state. The output of the network is used both for the state space representation and as parameters for sampling the next synthetic return.

2.3 Synthetic data generation

In this work, synthetic data was created using the generative model described in 2.1. Each synthetic return was sampled from a multivariate Gaussian distribution. Each trajectory was constructed using an underlying historical return (selected from the historical train dataset) according to the process presented with Algorithm 1. This procedure enables the construction of various return series needed for training the DRL agent. We note that the generative and training processes are intertwined, as displayed in Figure 1. The neural network outputs $\boldsymbol{\mu}$ and $\boldsymbol{L}$ used for sampling are also used as feature variables in the state space. Specifically, if the current state in the DRL learning process is $S_t = (\frac{t}{T}, \frac{W_t}{G}, \mu_t^{a_1}, \mu_t^{a_2}, ..., \mu_t^{a_n}, l_t^{11}, l_t^{21}, ..., l_t^{nn})$, the next synthetic return will be generated from $\mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{L}\boldsymbol{L}_t^\top)$, where $\boldsymbol{\mu}_t = [\mu_t^{a_1}, \mu_t^{a_2}, ..., \mu_t^{a_n}]$ and $\boldsymbol{L}_t = (l_t^{jk})^{j \geq k}$. This way, we merge two processes needed to train the agent: the generation of new data and the market-based features estimations.

Algorithm 1 Data generation

1: Randomly select a historical return $\boldsymbol{r}_0$ from the available train dataset
2: Define an empty list $synthetic\_episode$
3: **for** time step = 1, 2, . . . , T **do**
4: Pass return $\boldsymbol{r}_0$ to neural network and get $\boldsymbol{\mu}$ and $\boldsymbol{L}$
5: Sample $\boldsymbol{r}_{next}$ from the multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{L}\boldsymbol{L}^\top)$
6: Append $\boldsymbol{r}_{next}$ to $synthetic\_episode$ and set $\boldsymbol{r}_0 \leftarrow \boldsymbol{r}_{next}$
7: **end for**
8: **return** $synthetic\_episode$

3 Experimental study

3.1 Data, hyperparameter optimization and training

For evaluating the proposed method, we consider a scenario where the investor starts with an initial investment of W_0 at time $t = 0$. The target time for achieveing the goal G is 10 years and the agent reweights the portfolio monthly, e.g. $T = 120$ months. We use monthly returns from 1973 to 2022 to build a diversified seven-asset portfolio spanning stocks, bonds, and commodities. The stock assets are the MSCI Europe, MSCI North America, MSCI Pacific ex Japan, and MSCI Japan indices; the bond assets are the ICE BofA US Corporate Index and the US 10-Year Treasury Bond; and the commodities are represented by Gold Futures. The data were split into train (01/1973 – 12/1999), validation (01/2000 – 12/2001) and test (01/2002 – 07/2022) sets. For the DRL agent, the train set was only used as an underlying set needed for synthetic data generation For evaluating the agent, we use the historical test set comprised of 247 data points. This amounts to 127 distinct trajectories, considering that sequences of 120 months are required to form the investment scenario of 10 years.

The hyperparameter optimization on the validation set for the PR model resulted in 6 hidden layers, each with 150 neurons and the window size of returns used for estimation was $k = 1$. As for the DRL algorithm, we used the PPO, developed by Schulman et al. [17], and as implemented in Stable Baselines3 [18]. For PPO, we used the learning rate of 0.0001, batch size of 2048 and the discount factor γ equal to 1. DRL agent training was done on the data generated by the PR model. The synthetic episodes' initial historical returns were randomly selected from the historical train set. We used 2 million episodes, each with 120 steps corresponding to 120 months of investing. The agent was initialized with a starting wealth W_0 of 100,000, and the goal G was sampled from a set of {160,000, 170,000, ..., 240,000} at the beginning of each episode. Training ran locally on a personal computer, using the integrated Apple M1 GPU, with seeds parallelized via multiprocessing.

3.2 Results

We present several benchmark methods for evaluation: equally weighted portfolio (`EW`), dynamic programming for goal-based wealth management (`DP`) presented in [15], and deep reinforcement learning for goal-based investing (`DRL-gh`) as presented in [14]. We also include two variants of our model as control benchmarks to isolate the impact of the PR model on the generation of synthetic data and the estimation of the state. Firstly, `DRL-c` employs the same MDP, however, the data is generated using sample estimates of parameters instead of the PR model. Looking at Figure 1, $\boldsymbol{\mu}_{t+1}$ and $\boldsymbol{\Sigma}_{t+1}$ needed for sampling from $\mathcal{N}(\boldsymbol{\mu}_{t+1}, \boldsymbol{\Sigma}_{t+1})$ were calculated as rolling sample estimations of $\boldsymbol{r}_{t-d-1}, \ldots, \boldsymbol{r}_{t-1}, \boldsymbol{r}_t$, where $d = 106$ was chosen based on the maximum likelihood criterion in the test set. In this version, the state space is unchanged and uses the PR model as input for the state. In the other variant, `DRL-se`, the agent uses sample estimates for both the generation of synthetic data and the state space.

Table 1: Average goal achievement (%) on the test set. Values for DRL methods are seed-averaged.

Goal ($\times 10^3$)	`EW`	`DP`	`DRL-gh`	`DRL-c`	`DRL-se`	`DRL-pr`
160	91.3%	73.8%	83.0%	76.3%	88.9%	**95.4%**
180	65.1%	46.8%	56.4%	62.4%	77.0%	**89.7%***
200	36.5%	27.0%	29.0%	49.2%	63.5%	**81.3%***
220	23.0%	11.9%	15.8%	30.6%	45.2%	**71.1%***
240	8.7%	6.3%	7.6%	18.9%	30.2%	**57.4%***

Table 1 displays the proportion of test episodes in which the goal wealth was successfully reached, across multiple levels of goal wealth. For all DRL methods, the reported value is the average across 7 different runs. This approach is recommended due to the inherent instability of DRL models, as highlighted by recent studies [19], which show that performance can vary significantly across different runs. To assess the significance of our results, we performed a two-sample proportions z-test, comparing our approach with the best benchmark. The hypotheses are formulated as: $H_0 : p_1 = p_2$, and $H_1 : p_1 > p_2$, where p_1 is the proportion of test episodes in which the best method achieved the goal for each goal wealth. p_2 represents the proportion of test episodes in which the best benchmark achieved the goal for the same goal wealth. The symbol $*$ in Table 1 indicates statistically significant differences corresponding to the significance level of 1%.

The proposed DRL agent `DRL-pr` outperforms benchmarks on all levels of goal wealth. Moreover, the superior performance of `DRL-pr` over `DRL-se`, which relies solely on rolling parameter estimation, highlights the value of the probabilistic regression model when used for both estimation accuracy and data generation. Interestingly, the control variant `DRL-c` underperforms both the proposed model and `DRL-se`. This result may arise from the learning instability caused by a mismatch between the agent's state information (sample estimates) and the actual environment outcomes (PR model).

4 Conclusion

This study addresses GBPO by integrating DRL and PR into a unified framework. The PR model estimates return distributions—making the approach adaptive to changing market conditions—and also generates synthetic data to increase DRL sample efficiency. We evaluate the proposed DRL agent using historical market returns as test data, ensuring realistic assessment of the proposed method and the proposed method outperforms various benchmarks. Future work could extend beyond the single-goal case to a multi-goal framework. Furthermore, dimensionality reduction in the state space should be considered to improve scalability and enable larger portfolios, thereby enhancing generalization.

Acknowledgments

This research has been supported by the European Regional Development Fund under grant agreement PK.1.1.10.0007 (DATACROSS)

References

[1] Peter A Forsyth, Yuying Li, and Kenneth R Vetzal. Are target date funds dinosaurs? Failure to adapt can lead to extinction. 2017.

[2] Robert C. Merton. Lifetime Portfolio Selection under Uncertainty: The Continuous-Time Case. *The Review of Economics and Statistics*, 51(3):247, 1969.

[3] Sanjiv R Das, Daniel Ostrov, Anand Radhakrishnan, and Deep Srivastav. Dynamic portfolio allocation in goals-based wealth management. *Computational Management Science*, 17:613–640, 2020.

[4] Xin Du and Jinjian Zhai. Algorithm trading using q-learning and recurrent reinforcement learning. 2017.

[5] Parag C Pendharkar and Patrick Cusatis. Trading financial indices with reinforcement learning agents. *Expert Systems with Applications*, 103:1–13, 2018.

[6] Hyungjun Park, Min Kyu Sim, and Dong Gu Choi. An intelligent financial portfolio trading strategy using deep Q-learning. *Expert Systems with Applications*, 158:113573, 2020.

[7] Lin William Cong, Ke Tang, Jingyuan Wang, and Yang Zhang. Alphaportfolio: Direct construction through deep reinforcement learning and interpretable ai. *Capital Markets: Asset Pricing & Valuation eJournal*, 2020.

[8] Yue Deng, Feng Bao, Youyong Kong, Zhiquan Ren, and Qionghai Dai. Deep direct reinforcement learning for financial signal representation and trading. *IEEE Transactions on Neural Networks and Learning Systems*, 28(3):653–664, 2017.

[9] Zekun Ye, Weijie Deng, Shuigeng Zhou, Yi Xu, and Jihong Guan. Optimal trade execution based on deep deterministic policy gradient. In *Database Systems for Advanced Applications: 25th International Conference, DASFAA 2020, Jeju, South Korea, September 24–27, 2020, Proceedings, Part I*, page 638–654, Berlin, Heidelberg, 2020. Springer-Verlag.

[10] Ben Hambly, Renyuan Xu, and Huining Yang. Policy gradient methods for the noisy linear quadratic regulator over a finite horizon. *SIAM Journal on Control and Optimization*, 59(5):3359–3391, 2021.

[11] Bruno Gašperov and Zvonko Kostanjčar. Market making with signals through deep reinforcement learning. *IEEE Access*, 9:61611–61622, 2021.

[12] Thomas Spooner and Rahul Savani. Robust Market Making via Adversarial Reinforcement Learning. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, {IJCAI-20}*, pages 4590–4596. International Joint Conferences on Artificial Intelligence Organization, 2020.

[13] Sanjiv R Das and Subir Varma. Dynamic goals-based wealth management using reinforcement learning. *Journal Of Investment Management*, 18(2):1–20, 2020.

[14] Tessa Bauman, Sven Goluža, Bruno Gasperov, and Zvonko Kostanjcar. Deep reinforcement learning for goal-based investing under regime-switching. In *Northern Lights Deep Learning Conference*, pages 13–19. PMLR, 2024.

[15] Sanjiv Ranjan Das, Daniel N Ostrov, Anand Radhakrishnan, and Deep Srivastav. A New Approach to Goals-Based Wealth Management. *SSRN Electronic Journal*, pages 1–34, 2018.

[16] Diederik P. Kingma and Max Welling. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019.

[17] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov Openai. Proximal Policy Optimization Algorithms. Technical report, 2017.

[18] A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus, and N. Dormann. Stable-baselines3: Reliable reinforcement learning implementations. Journal of Machine Learning Research, 2021. [**22**(1), 12348–12355].

[19] Andrew Patterson, Samuel Neumann, Martha White, and Adam White. Empirical design in reinforcement learning, 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the objective of the paper and the contributions made. The claims made in the abstract and introduction are supported by the results given in the Section 3, and a discussion is provided on their statistical and practical significance.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations are discussed as future work in section 2.2 Markov decision process and 4 Conclusion. We note that the state space grows quadratically which presents a problem if using a large number of assets for portfolio construction. Furthermore, we note that the limitation is a singe-goal setting.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: The paper does not include any theoretical results, only experimental results of the proposed model.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper includes all of the methodology and the hyperparameter selection (in section 3.1 Data, hyperparameter optimization and training). If the hyperparameter is not in the paper it is because the default one was used (namely for PPO from Stable Baselines3). The data set can be find online with free download.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The paper does not include code. However, the architecture used in the paper is simple for implementation, and the DRL algorithm is used from Stable Baselines3 (already implemented).

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper includes all the specifications necessary to understand the results. The data and the training process are to be found in 3.1. Data, hyperparameter optimization and training. The process of testing is explained through 3 Simulation results, and mostly in 3.2 Benchmark methods and evaluation.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: The paper includes statistical testing of the experimental results, section 3.2 Benchmark methods and evaluation. We used a proportion z-test and reported the significant results through Table 1.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The experiments were done on a personal computer, as is written in section 3.1 Data, hyperparameter optimization and training. No large memory or high-end hardware was needed.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conforms in every respect with the NeurIPS Code of Ethics. The research does not involve human subjects or participants. The data used does not contain any private information, and the dataset is used according to the terms of the licence under which it was obtained. There are no potential harmful consequences of the research.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: The paper presents a novel use of generative models for enhancing a procedure for training machine learning models, and is not tied to a deployment with potential societal impact of the work.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The models used have no risk for misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: PPO algorithm was used for the work done in the paper. We cited the original paper along with the citation for Stable Baselines3. This can be found in section 3.1 Data, hyperparameter optimization and training.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing or research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper presents a new method for investing and does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The method presented in the paper was done without any usage of LLMs. LLMs were used only for editing and formatting purposes.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.