# Stable Alpha: Adversarial Invariant Representation Learning for Nonlinear Asset Pricing under Temporal Distribution Shifts

**Xiaokang Wang**[1], **Zihe Liu**[2], **Xinghan Qin**[3], **Qinxuan Zhang**[2], **Xiushuo Hu**[4], **Jidong Yuan**[5]* **and Zihao Yin**[5]
[1]School of Automation and Intelligence, Beijing Jiaotong University, China
[2]School of Economics and Management, Beijing Jiaotong University, China
[3]Faculty of Engineering, The Chinese University of Hong Kong, Hong Kong
[4]School of Mathematics and Statistics, Xi'an Jiaotong University, China
[5]School of Computer Science and Technology, Beijing Jiaotong University, China
{xkwang1, 24241184}@bjtu.edu.cn, 1155269935@link.cuhk.edu.hk
25725076@bjtu.edu.cn, 1448053371@stu.xjtu.edu.cn
{yuanjd, zihao_yin}@bjtu.edu.cn

## Abstract

Deep learning has emerged as a powerful paradigm for constructing nonlinear asset pricing factor models. However, temporal distribution shifts such as bull-bear transitions and industry rotations undermine the generalization of models trained under the i.i.d. assumption. As a result, existing models tend to exploit environment-specific correlations that are highly predictive in-sample but unstable across market regimes. To address this challenge, we propose ***CASH***, a **c**ausality-inspired **a**sset pricing factor framework designed to learn invariant and minimally sufficient representations under temporal distribution **sh**ifts. CASH formulates an information-theoretic adversarial learning objective that systematically discards transient market noise by optimizing against an adversary designed to exploit spurious correlations, thereby distilling invariant factor-return relationships. The resulting invariant representation is then integrated into a conditional factor pricing model via a dynamic factor exposure network. Experiments on real-world stock datasets across multiple markets demonstrate that CASH consistently outperforms state-of-the-art baselines in out-of-sample return prediction and exhibits superior robustness under pronounced temporal regime shifts. The code is available at https://github.com/ZiheLiu1007/Neurips2026-CASH

## 1 Introduction

Asset pricing aims to explain and predict the cross-section of expected returns. While classical linear models such as Fama-French factors [4] laid the foundation for linking returns to systematic risks, recent advances have increasingly leveraged deep neural networks to extract nonlinear conditional factors from high-dimensional firm characteristics and alternative data. However, a fundamental challenge remains unresolved: financial markets are inherently non-stationary. Driven by macroeconomic fluctuations and investor sentiment, asset return dynamics undergo severe temporal distribution shifts. For instance, high turnover may indicate strong momentum in bull markets but signal panic selling in bear markets. Because most existing deep asset pricing models are optimized via empirical risk minimization (ERM) under the i.i.d. assumption, they tend to overfit spurious, environment-specific

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

correlations. Consequently, when market regimes shift (e.g., bull-bear transitions), these models often suffer from performance degradation and unstable factor exposures during out-of-sample periods.

From an invariance learning perspective, achieving out-of-distribution (OOD) generalization requires isolating stable asset-specific drivers (invariant mechanisms) from transient, regime-dependent effects (spurious noise). While invariant risk minimization (IRM) and causality-aware learning offer conceptual solutions, their application to financial markets faces significant challenges. First, market regimes are latent and gradual, precluding the use of explicit environment labels or predefined domain partitions required by standard IRM methods [1]. Second, high-dimensional and noisy financial data make it difficult to distinguish stable predictive information from regime-specific correlations. Third, learning stable representations alone is insufficient; these representations must be translated into multi-factor return forecasts that support dynamically rebalanced, tradable portfolios.

To address these challenges, we propose CASH, a **c**ausality-inspired **a**sset pricing framework designed to handle temporal distribution **sh**ifts. At its core, CASH formulates representation learning as an information-theoretic adversarial minimax game, which extracts invariant and minimally sufficient features without relying on explicit environment labels. Specifically, the framework employs an **encoder network** to generate latent asset representations, alongside an **adversary network** that attempts to exploit regime-sensitive vulnerabilities and spurious correlations from these representations. By optimizing the encoder to defeat the adversary, the model is compelled to systematically filter out transient market noise and retain only the stable, invariant drivers of asset returns. Finally, these purified representations are integrated into a conditional multi-factor pricing framework via a dynamic factor exposure network, enabling robust return predictions across diverse market regimes.

Our main contributions are summarized as follows:

- We propose CASH, a causality-inspired multi-factor asset pricing framework that improves robustness under temporal distribution shifts without requiring explicit environment labels.
- We develop a novel information-theoretic framework that enables robust invariant representation learning through a dual-adversarial mechanism comprising a dependency discrimination head and a robust prediction head with a latent perturbation generator, supported by an OOD generalization analysis (Definition 1, Proposition 1, Theorem C.4).
- Extensive empirical evaluations on real-world stock data demonstrate that CASH consistently outperforms state-of-the-art baselines in out-of-sample return prediction, delivering enhanced cross-regime robustness across fluctuating market conditions.

## 2 Related Work

**Deep Learning for Asset Pricing.** To address the proliferation of variables in traditional linear asset pricing (the "factor zoo" [5]), recent methodologies increasingly rely on deep learning to extract nonlinear predictive signals. Gu et al. [9] proposed a nonlinear autoencoder network for latent factor estimation, achieving lower out-of-sample pricing errors under no-arbitrage constraints. Qin et al. [16] developed a dual-stage attention-based recurrent neural network (DA-RNN) to capture long-term temporal dependencies. More recently, deep generative architectures have been introduced, such as FactorVAE [2], which integrates the variational autoencoder into a dynamic factor model for probabilistic predictions, and adaptive models [20] that incorporate regime-shifting mechanisms to absorb macroeconomic information.

**Causality-Inspired Invariance Learning.** To mitigate distribution shifts, causality-inspired representation learning seeks to extract invariant features aligned with stable underlying data-generating mechanisms rather than fragile spurious correlations [14]. General OOD methods such as IRM [1], VREx [12], and GroupDRO [17] typically require explicit environment or group labels. Such labels are rarely available in financial markets, where regime changes induced by policy shocks, liquidity conditions, investor sentiment, and industry rotations are latent and gradual. Alternatively, causality-inspired disentanglement methods leverage structural causal models (e.g., CausalVAE [22]) or information-theoretic constraints [23] to suppress domain-dependent noise, which have also been extended to time-series discovery [11, 26].

**Time Series OOD Methods.** Existing time series OOD methods address related but different settings. DIVERSITY [15] and WOODS [6] primarily study classification or OOD detection, whereas

CASH addresses continuous cross-sectional return prediction under changes in the characteristic–return relationship. OneNet [19] improves online forecasting through continual adaptation using newly observed feedback. This is complementary to CASH but evaluates a different form of robustness: OneNet adapts after a shift occurs, whereas CASH tests whether a model trained once can retain predictive ability under future shifts without test-window parameter updates. PeTS [25] focuses on performative time series, where predictions alter the subsequent data-generating process, while CASH considers exogenous temporal shifts in financial markets.

## 3 Problem Formulation

Consider a financial market with $N$ assets over $T$ time steps. For asset $i$ at time $t$, let $\mathbf{x}_{i,t} \in \mathbb{R}^M$ denote its characteristic covariates and $r_{i,t+1} \in \mathbb{R}$ its subsequent excess return. Our objective is to jointly learn an encoder $\Phi : \mathbf{x}_{i,t} \mapsto \mathbf{z}_{i,t} \in \mathbb{R}^H$ and a conditional pricing function $g(\cdot)$ such that:

$$r_{i,t+1} = g\big(\Phi(\mathbf{x}_{i,t}),\, \mathbf{f}_{t+1}\big) = \boldsymbol{\beta}(\mathbf{z}_{i,t})^\top \mathbf{f}_{t+1} + \epsilon_{i,t+1}, \tag{1}$$

where $\mathbf{f}_{t+1} \in \mathbb{R}^K$ denotes the latent common risk factors, $\boldsymbol{\beta}(\mathbf{z}_{i,t}) \in \mathbb{R}^K$ are the state-dependent factor loadings, and $\epsilon_{i,t+1}$ is idiosyncratic noise.

To ensure robustness against temporal distribution shifts induced by unobserved market environments $e \in \mathcal{E}$, we formalize the target representation via information theory.

**Definition 1** (Minimally Sufficient Invariant Representation)**.** *A representation* $\mathbf{z} = \Phi(\mathbf{x})$ *is* minimally sufficient invariant *for predicting* $\mathbf{r}$ *if it satisfies:*

***(D1) Sufficiency.*** $I(\mathbf{z}; \mathbf{r}) = I(\mathbf{x}; \mathbf{r})$*, i.e.,* $\mathbf{z}$ *preserves all return-predictive information in* $\mathbf{x}$ *(*$\mathbf{r} \perp\!\!\!\perp \mathbf{x} \mid \mathbf{z}$*).*

***(D2) Minimality.*** *Among all sufficient representations,* $\mathbf{z}$ *minimizes* $I(\Phi(\mathbf{x}); \mathbf{x})$*, maximally discarding environment-specific spurious signals.*

***(D3) Invariance.*** $\forall e \in \mathcal{E} : I(\mathbf{z}; \mathbf{r} \mid e) = I(\mathbf{x}; \mathbf{r} \mid e)$ *and* $I(\mathbf{z};\, e \mid \mathbf{r}) = 0$*, so that the predictive relationship remains stable across environments.*

**Proposition 1** (Information-Theoretic Minimax Objective)**.** *The optimal representation* $\mathbf{z}^* = \Phi^*(\mathbf{x})$ *satisfying Definition 1 solves the following minimax problem:*

$$\min_{\Phi,\, \omega,\, \theta_{\text{pred}}} \max_{\psi,\, \theta_{\text{cls}}} \quad \underbrace{I(\mathbf{z}; \mathbf{x})}_{\textit{Minimality}} - \underbrace{I(\mathbf{z}; \mathbf{r})}_{\textit{Sufficiency}} + \lambda \underbrace{R(\Phi(\mathbf{x}),\, T_\psi(\mathbf{z}),\, \mathbf{r})}_{\textit{Robustness}}, \tag{2}$$

*where* $T_\psi(\mathbf{z})$ *is a learned perturbation simulating worst-case environmental shifts, and* $R(\cdot)$ *measures predictive degradation under such shifts.*

## 4 Proposed method

In this section, we first present the overall architecture of CASH, as shown in Fig. 1. The proposed model contains two main components: the factor module and the beta module. The factor module learns a $K$-dimensional vector of latent common factors $f_t \in \mathbb{R}^K$ from cross-sectional asset returns $r_t$ using a variational autoencoder. In the beta module, the main contribution lies in the adversarial minimax framework involving an encoder network $\Phi$ and a dual-adversarial mechanism comprising a classification head $D_{\text{cls}}$ and a prediction head $D_{\text{pred}}$. Optimized with a Gradient Reversal Layer (GRL), this framework learns stable asset representations $z_{i,t} = \Phi(X_i)$, which are subsequently mapped into dynamic factor exposures $\beta(z_{i,t}; \omega)$. Finally, the "dotted operation" multiplies the beta module with the vector of latent factors from the factor module to produce the estimated return $r_{i,t+1}$ as the output of our asset pricing model.

### 4.1 Factor Network

The factor network extracts the latent common factors $\mathbf{f}_t \in \mathbb{R}^K$ in Eq. (1) from cross-sectional return data via a variational autoencoder [10]. Concretely, we model the cross-sectional return vector $\mathbf{r}_t = (r_{1,t}, \ldots, r_{N,t})^\top$ as governed by a $K$-dimensional latent variable $\mathbf{s}_t$. To approximate its intractable

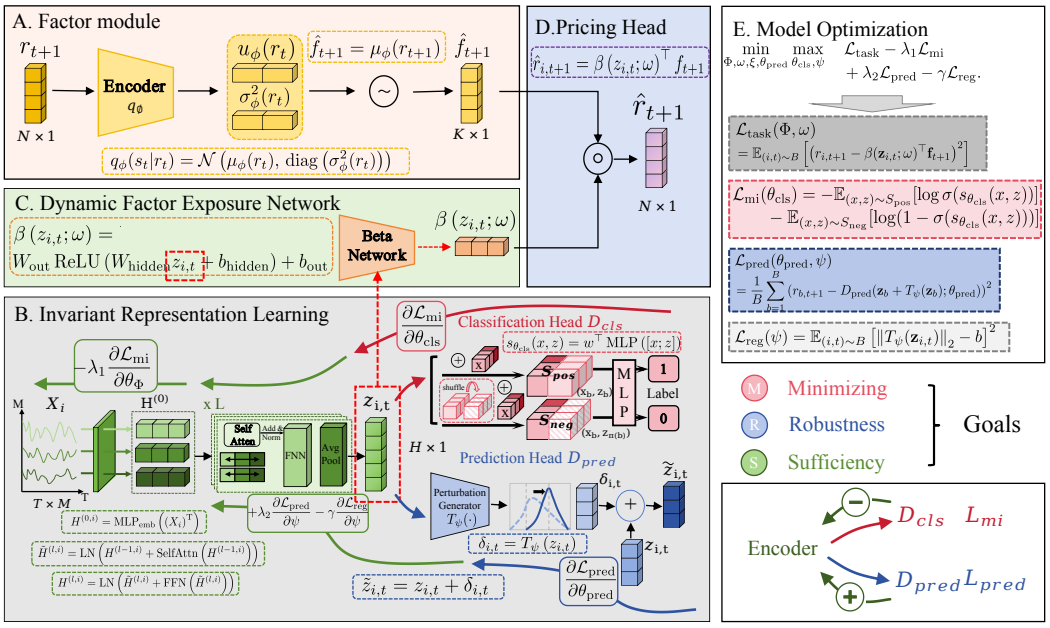



Figure 1: The overall architecture of CASH.

posterior, we parameterize the variational distribution: $q_\xi(\mathbf{s}_t \mid \mathbf{r}_t) = \mathcal{N}(\boldsymbol{\mu}_\xi(\mathbf{r}_t),\ \text{diag}(\boldsymbol{\sigma}^2_\xi(\mathbf{r}_t)))$, where $\boldsymbol{\mu}_\xi(\cdot)$ and $\boldsymbol{\sigma}^2_\xi(\cdot)$ are neural networks with parameters $\xi$ that take $\mathbf{r}_t$ as input. The latent factor is defined as the posterior mean: $\mathbf{f}_t = \boldsymbol{\mu}_\xi(\mathbf{r}_t)$. This quantity serves as both the training-time factor estimate and the basis for the rolling inference approximation in Section 4.5. Setting $K \ll N$ compresses the high-dimensional cross-sectional returns into a compact common-factor representation [24], consistent with the low-rank structure assumed by classical linear factor models.

### 4.2 Invariant representation learning via adversarial learning

#### 4.2.1 Encoder Network

We adopt the iTransformer [13] as the backbone for the encoder $\Phi$. Given asset $i$'s lookback window $\mathbf{X}_i \in \mathbb{R}^{T\times M}$, the transposed input is tokenized so that each covariate sequence becomes an independent token. These tokens are embedded via an MLP and refined through $L$ stacked iTransformer blocks, equipped with self-attention and FFN with residual connections and layer normalization. The final representation $\mathbf{z}_{i,t} \in \mathbb{R}^H$ is obtained by global average pooling over the $M$ output tokens followed by a linear projection:

$$\mathbf{z}_{i,t} = \Phi(\mathbf{X}_i) = \text{Linear}\left(\frac{1}{M}\sum_{j=1}^{M}\mathbf{H}_j^{(L,i)}\right). \tag{3}$$

Applying this shared encoder across all $N$ assets yields the cross-sectional representation set $\mathcal{Z} = \{\mathbf{z}_{i,t}\}_{i=1}^N$.

#### 4.2.2 Adversary network

The adversary network employs two complementary adversarial heads to jointly enforce minimally sufficient invariance. The **classification head** $\text{D}_{\text{cls}}$ acts as a discriminator to purge environment-specific noise, satisfying the *minimality* requirement. Concurrently, the **prediction head** $\text{D}_{\text{pred}}$ enforces *sufficiency* by demanding robust return forecasting under worst-case latent perturbations $T_\psi(\cdot)$. Through their complementary adversarial objectives, these two heads encourage the encoder $\Phi$ to learn compact, invariant representations that generalize robustly across market regimes.

**Classification head $\mathrm{D}_{\mathbf{cls}}$.** The role of $\mathrm{D}_{\mathrm{cls}}$ is to detect and expose any residual dependency between the learned representation $\mathbf{z}_{i,t}$ and the high-dimensional input feature $\mathbf{x}_{i,t}$. Intuitively, if a discriminator can reliably identify whether a given $(\mathbf{x}, \mathbf{z})$ pair originates from the true joint distribution or from an artificially shuffled marginal distribution, then $\mathbf{z}$ still encodes environment-specific redundant information that should be discarded. To operationalize this, we construct within each training mini-batch $\mathcal{B} = \{(i_b, t_b)\}_{b=1}^{B}$ a set of matched pairs and mismatched pairs:

$$\mathcal{S}_{\text{pos}} = \{(\mathbf{x}_b,\, \mathbf{z}_b)\}_{b=1}^{B}, \qquad \mathcal{S}_{\text{neg}} = \{(\mathbf{x}_b,\, \mathbf{z}_{\pi(b)})\}_{b=1}^{B}, \tag{4}$$

where $\pi$ is a random permutation over the batch index. A compatibility score is computed via a multi-layer perceptron:

$$s_{\theta_{\text{cls}}}(\mathbf{x}, \mathbf{z}) = \mathbf{w}^{\top} \mathrm{MLP}([\mathbf{x};\, \mathbf{z}])\,, \tag{5}$$

where $\theta_{\text{cls}}$ collects all learnable parameters. The discriminator is trained via a binary cross-entropy loss to identify matched pairs:

$$\mathcal{L}_{\text{mi}}(\theta_{\text{cls}}) = -\mathbb{E}_{(\mathbf{x},\mathbf{z})\sim\mathcal{S}_{\text{pos}}}[\log \sigma(s_{\theta_{\text{cls}}}(\mathbf{x}, \mathbf{z}))] - \mathbb{E}_{(\mathbf{x},\mathbf{z})\sim\mathcal{S}_{\text{neg}}}[\log(1 - \sigma(s_{\theta_{\text{cls}}}(\mathbf{x}, \mathbf{z})))]\,, \tag{6}$$

with $\sigma(\cdot)$ denoting the sigmoid function. This objective provides a variational lower bound on $I(\mathbf{x};\, \mathbf{z})$. By adversarially training the encoder $\Phi$ to *maximize* $\mathcal{L}_{\text{mi}}$ while $\mathrm{D}_{\text{cls}}$ *minimizes* it, we actively purge domain-sensitive noise, strictly enforcing the *minimality* of the learned representations.

**Prediction head $\mathrm{D}_{\mathbf{pred}}$.** To enforce the *sufficiency* requirement, $\mathrm{D}_{\text{pred}}$ ensures that the learned representation retains invariant predictive information even under severe market structural changes. Instead of injecting naive isotropic noise, we introduce a generator $T_\psi$ to simulate worst-case environmental drifts by applying conditional latent perturbations, yielding the perturbed state $\widetilde{\mathbf{z}} = \mathbf{z} + T_\psi(\mathbf{z})$. The prediction head maps $\widetilde{\mathbf{z}}$ to the target future return, evaluated via the Mean Squared Error (MSE) loss over a mini-batch of size $B$:

$$\mathcal{L}_{\text{pred}}(\theta_{\text{pred}}, \psi) = \frac{1}{B}\sum_{b=1}^{B} \left(r_{b,t+1} - \mathrm{D}_{\text{pred}}\big(\mathbf{z}_b + T_\psi(\mathbf{z}_b); \theta_{\text{pred}}\big)\right)^2. \tag{7}$$

This formulation acts as an adversarial testing ground: the generator $T_\psi$ *maximizes* $\mathcal{L}_{\text{pred}}$ to seek latent directions that most severely degrade forecasting performance, while the encoder $\Phi$ and $\mathrm{D}_{\text{pred}}$ jointly *minimize* this worst-case loss. Maintaining high predictive accuracy under extreme adversarial shifts forces the representation to discard fragile, spurious correlations and rigorously preserve invariant mechanisms.

### 4.3 Dynamic factor exposure network

Unlike traditional multi-factor models that assume static factor loadings, an asset's sensitivity to systematic risk in non-stationary markets dynamically adapts to its current state. Therefore, we model the time-varying factor exposure $\boldsymbol{\beta}(\mathbf{z}_{i,t}; \omega) \in \mathbb{R}^K$ as an endogenous function of the stable state representation $\mathbf{z}_{i,t} \in \mathbb{R}^H$, parameterized by a multi-layer perceptron:

$$\boldsymbol{\beta}(\mathbf{z}_{i,t}; \omega) = \mathbf{W}_{\text{out}}\, \mathrm{ReLU}(\mathbf{W}_{\text{hidden}}\mathbf{z}_{i,t} + \mathbf{b}_{\text{hidden}}) + \mathbf{b}_{\text{out}}, \tag{8}$$

where $\omega$ collects all learnable weights. This formulation allows the model to flexibly capture heterogeneous and state-dependent risk exposures across fluctuating market conditions.

The conditional predicted return of asset $i$ at time $t+1$ is then constructed as the inner product of the dynamic exposure and the latent common risk factor $\mathbf{f}_{t+1} \in \mathbb{R}^K$: $\hat{r}_{i,t+1} = \boldsymbol{\beta}(\mathbf{z}_{i,t}; \omega)^{\top}\mathbf{f}_{t+1}$. This establishes the conditional factor pricing relationship, explicitly combining market-wide systematic risk compensation ($\mathbf{f}_{t+1}$) with asset-specific macroeconomic sensitivities ($\boldsymbol{\beta}$).

To optimize the representation encoder $\Phi$ and the exposure network $\omega$, we minimize the Mean Squared Error (MSE) over the training mini-batch $\mathcal{B}$:

$$\mathcal{L}_{\text{task}}(\Phi, \omega) = \mathbb{E}_{(i,t)\sim\mathcal{B}}\left[\left(r_{i,t+1} - \boldsymbol{\beta}(\mathbf{z}_{i,t}; \omega)^{\top}\mathbf{f}_{t+1}\right)^2\right]. \tag{9}$$

This objective penalizes the deviation from ground-truth returns $r_{i,t+1}$, providing a fundamental task-level constraint that grounds the learned representations in actual asset pricing dynamics for the subsequent joint minimax optimization.

### 4.4 Model Optimization

Integrating the aforementioned modules, the overall training objective is formulated as the following minimax problem:

$$\min_{\Phi,\omega,\xi,\theta_{\text{pred}}} \max_{\theta_{\text{cls}},\psi} \quad \mathcal{L}_{\text{task}} - \lambda_1 \mathcal{L}_{\text{mi}} + \lambda_2 \mathcal{L}_{\text{pred}} - \gamma \mathcal{L}_{\text{reg}}. \tag{10}$$

where $\lambda_1, \lambda_2$, and $\gamma$ are trade-off hyperparameters. This joint objective balances accurate conditional return prediction ($\mathcal{L}_{\text{task}}$) with the adversarial enforcement of representation minimality ($\mathcal{L}_{\text{mi}}$) and predictive sufficiency under worst-case shifts ($\mathcal{L}_{\text{pred}}$).

To prevent the adversarial generator $T_\psi$ from collapsing to trivial zero perturbations, we introduce a norm regularization term:

$$\mathcal{L}_{\text{reg}}(\psi) = \mathbb{E}_{(i,t)\sim\mathcal{B}} \left[ \|T_\psi(\mathbf{z}_{i,t})\|_2 - b \right]^2. \tag{11}$$

Here, the pre-specified target norm $b > 0$ anchors the perturbation magnitude, ensuring the continuous generation of effective adversarial shifts throughout training.

To solve this minimax game efficiently without resorting to a complex bi-level optimization procedure, we employ the Gradient Reversal Layer (GRL) [7]. By automatically negating gradients during backpropagation, the GRL enables a unified, end-to-end training pipeline. Consequently, the encoder $\Phi$ is updated to maximize $L_{\text{mi}}$ while minimizing $L_{\text{pred}}$, whereas the adversarial heads $D_{\text{cls}}$ and $D_{\text{pred}}$ minimize their respective losses and the perturbation generator $T_\psi$ maximizes $L_{\text{pred}}$ under a norm constraint, steering the representations toward a stable, minimally sufficient invariant form.

**Why Adversarial Minimax Avoids Information Collapse.** The adversarial minimax objective (Eq. 10) extracts invariant features without explicit environment labels by leveraging a game-theoretic equilibrium between dual adversarial heads. $D_{\text{cls}}$ enforces minimality through reducing $z$'s dependency on $x$, while $D_{\text{pred}}$ ensures sufficiency through maintaining predictive accuracy under worst-case perturbations to prevent information collapse. This dual constraint inherently avoids trivial solutions: an uninformative $z$ fails $D_{\text{pred}}$, while an overly informative $z$ is identified by $D_{\text{cls}}$. Consequently, the equilibrium solution isolates task-relevant invariant drivers from environment-specific noise. Formally, **Theorem C.4 (Appendix C)** proves that minimizing mutual information $I(z;x)$ tightens the OOD generalization bound, with adversarial prediction risk ensuring informative representations.

### 4.5 Out-of-Sample Inference

During out-of-sample inference, the future cross-sectional return $\mathbf{r}_{t+1}$ is strictly unavailable. To prevent look-ahead bias, we cannot extract the latent factors via $\mathbf{f}_{t+1} = \boldsymbol{\mu}_\xi(\mathbf{r}_{t+1})$. Instead, we approximate the conditional expected factor risk premium $\lambda_t = \mathbb{E}_t[\mathbf{f}_{t+1}]$ using a rolling mean of the $L$ most recent historical posterior estimates:

$$\hat{\lambda}_t = \frac{1}{L} \sum_{\tau=0}^{L-1} \boldsymbol{\mu}_\xi(\mathbf{r}_{t-\tau}). \tag{12}$$

Simultaneously, the state-dependent risk exposure relies solely on the observable historical covariates $\mathbf{X}_i$. The out-of-sample predicted return for asset $i$ is thereby computed as the inner product of the estimated dynamic exposure and the expected risk premium: $\hat{r}_{i,t+1} = \boldsymbol{\beta}(\Phi(\mathbf{X}_i);\, \omega)^\top \hat{\lambda}_t$. These cross-sectional return forecasts are subsequently utilized to rank assets and generate trading signals for portfolio construction. In Appendix A, we provide a complete training algorithm for CASH.

## 5 Experiments

We evaluate CASH along three connected dimensions: out-of-sample prediction accuracy, prediction stability under temporal shifts, and the investment performance using the resulting trading signals.

### 5.1 Experiment Settings

We use two equity panels. The CSI 300 panel contains daily observations from October 10, 2020 to October 20, 2025, with 312 asset covariates from the Tonglian Data Platform. The NASDAQ Index

panel contains 95 monthly stock characteristics over 1990–2021 [8, 21]. Predictors are standardized, missing values are interpolated, and outliers are winsorized before model training. Each panel is split chronologically into 70% training, 10% validation, and 20% testing, with the main sliding window set to $L = 3$. CASH is compared with commonly used asset pricing baselines including GRU, MLP, Transformer, CQVAE [24], DY-GAP [18], FactorGCL [3], and FactorVAE [2].

### 5.2 Prediction Performance

Table 1 presents out-of-sample prediction metrics. CASH consistently achieves the highest Information Coefficient (IC) and Information Coefficient Ratio (ICIR) across both panels, alongside competitive $R_{\text{total}}$ and $R_{\text{predict}}$ values, indicating robust performance beyond a single scoring rule. On the CSI300 index, CASH significantly improves correlation-based measures: IC rises from $0.036$ to $0.158$, and Rank IC from $0.019$ to $0.129$, with ICIR confirming this qualitative improvement after accounting for time variation. On NASDAQ, while absolute IC values are generally lower as expected for monthly cross-sectional return prediction, CASH still yields the highest IC, ICIR, and best fit statistics.

Baseline models often show near-zero or negative IC values on the CSI300 panel, reflecting structural underfitting rather than insufficient tuning. As depicted in Figure 2, most baselines collapse to predicting near-zero constants during the test period. This behavior, where models struggle to extract cross-period invariant signals from noisy daily covariates, leads to correlations with realized returns approaching zero or becoming negative if test distributions shift. In contrast, CASH's adversarial minimax objective enforces invariance across environments, preventing this collapse and driving its superior IC performance.

It is notable that most models, including CASH, yield zero or negative $R_{\text{predict}}$ values on both CSI300 and NASDAQ, despite strong IC/Rank IC results. Predicting daily equity return *magnitudes* is inherently challenging due to level and scale errors, often resulting in low or negative predictive $R^2$. However, IC and Rank IC measure the *cross-sectional ordering* of returns, a metric invariant to such magnitude-related issues. Thus, strong IC/Rank IC can coexist with low $R_{\text{predict}}$, signifying effective relative performance prediction despite poor magnitude estimation. Indeed, despite CASH's near-zero $R_{\text{predict}}$, portfolios constructed from its predictions demonstrated significant Sharpe ratios and positive factor-adjusted $\alpha$ (Section 5.5). This underscores that low $R_{\text{predict}}$ does not preclude economically valuable predictive content, and that IC and Rank IC are more directly relevant for rank-based portfolio construction than magnitude-focused metrics.

Table 1: Prediction performance on the CSI 300 Index and NASDAQ Index. Larger IC, ICIR, Rank IC, $R_{\text{total}}$, and $R_{\text{predict}}$ indicate better performance; values in parentheses are standard deviations.

| Model | CSI 300 Index | | | | | NASDAQ Index | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | IC | ICIR | Rank IC | $R_{\text{total}}$ | $R_{\text{predict}}$ | IC | ICIR | Rank IC | $R_{\text{total}}$ | $R_{\text{predict}}$ |
| GRU | -0.007 (0.06) | -0.115 | 0.002 (0.06) | -6.454 (2.87) | -6.206 (2.82) | 0.006 (0.11) | 0.056 | -0.013 (0.12) | -0.846 (0.76) | -0.831 (0.72) |
| MLP | -0.010 (0.14) | -0.069 | -0.003 (0.12) | -0.243 (0.22) | -0.237 (0.22) | -0.018 (0.10) | -0.178 | -0.010 (0.10) | -0.557 (0.56) | -0.542 (0.56) |
| Transformer | 0.036 (0.08) | 0.466 | 0.019 (0.08) | -3.012 (1.20) | -2.964 (1.18) | -0.011 (0.12) | -0.090 | -0.025 (0.15) | -3.414 (2.14) | -3.016 (1.95) |
| CQVAE | -0.002 (0.27) | -0.006 | 0.007 (0.22) | 0.001 (0.02) | **-0.001** (0.02) | -0.015 (0.10) | -0.148 | -0.006 (0.12) | -0.483 (0.60) | -0.483 (0.60) |
| DY-GAP | 0.019 (0.09) | 0.208 | 0.010 (0.08) | -0.326 (1.18) | -0.226 (0.24) | -0.011 (0.13) | -0.089 | -0.010 (0.13) | -0.443 (0.55) | -0.437 (0.55) |
| FactorGCL | -0.073 (0.18) | -0.399 | -0.065 (0.19) | 0.014 (0.03) | -0.021 (0.05) | -0.037 (0.14) | -0.265 | -0.009 (0.14) | -0.919 (0.58) | -0.922 (0.50) |
| FactorVAE | -0.024 (0.22) | -0.110 | -0.055 (0.24) | -0.000 (0.03) | -0.001 (0.02) | -0.004 (0.14) | -0.031 | **0.005** (0.15) | -0.391 (0.47) | -0.390 (0.47) |
| CASH | **0.158** (0.21) | **0.694** | **0.129** (0.21) | **0.023** (0.09) | -0.003 (0.05) | **0.013** (0.16) | **0.082** | -0.003 (0.14) | **0.000** (0.01) | **0.000** (0.00) |

### 5.3 Robustness Analysis Against Distribution Shifts

To quantify temporal robustness, we use IC Gap and Rank IC Gap:

$$\text{Gap}(M) = |M_{\text{train}} - M_{\text{test}}|/|M_{\text{train}}|,$$

where $M$ refers to the average IC or Rank IC measure. Smaller values indicate lower relative degradation under temporal shifts. Table 2 shows that CASH has the smallest gap on both panels. Furthermore, Fig. 2 provides the out-of-sample return prediction curve for the CSI 300 index, offering an intuitive perspective on the models' generalization capabilities. The visualization reveals a severe underfitting issue in most baseline models, which merely predict near-zero mean values across both domains. In contrast, in the out-of-sample region (blue line), CASH is the only model that successfully anticipates the sharp directional reversals, demonstrating that it has learned truly invariant and generalizable pricing mechanisms rather than overfitting to in-sample mean returns. The results

for the NASDAQ Index are reported in Appendix B.1. The results are not meant to imply pointwise tracking of daily returns, which would be unrealistic for noisy equity panels. Rather, it indicates that the learned relation is not confined to the training segment.

Table 2: Gap analysis on the CSI 300 Index and NASDAQ Index. Smaller IC Gap and Rank IC Gap indicate smaller relative degradation from training to testing periods.

| Market | Metric | GRU | MLP | Transformer | CQVAE | DY-GAP | FactorGCL | FactorVAE | CASH |
|---|---|---|---|---|---|---|---|---|---|
| CSI 300 Index | IC Gap | 1.03 | 1.06 | 0.34 | 1.34 | 0.34 | 1.13 | 1.34 | **0.29** |
| | Rank IC Gap | 0.99 | 1.02 | 0.32 | 0.52 | 0.66 | 1.14 | 0.39 | **0.31** |
| NASDAQ Index | IC Gap | 0.99 | 1.00 | 0.69 | 0.36 | 0.43 | 0.69 | 0.28 | **0.23** |
| | Rank IC Gap | 1.02 | 0.97 | 0.61 | 0.65 | 0.32 | 0.81 | 0.34 | **0.28** |

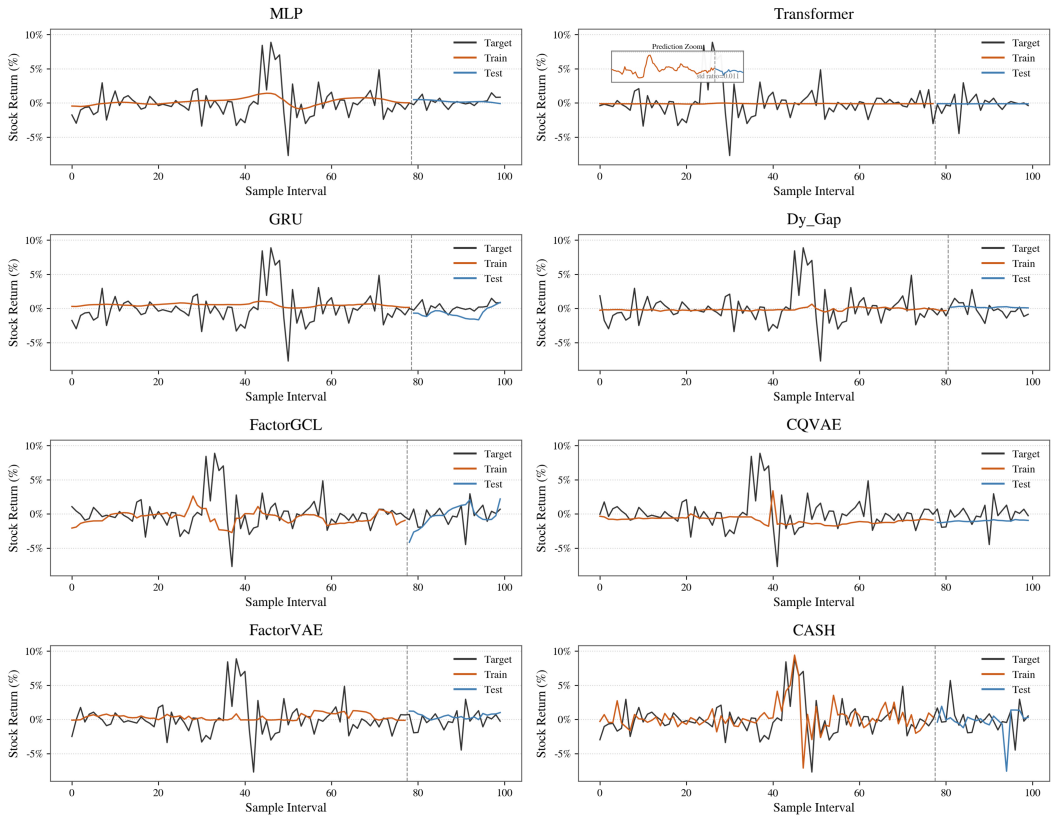



Figure 2: In-sample and out-of-sample fitting curves on the CSI 300 index. The black line denotes realized returns, orange denotes in-sample fitted returns, blue denotes out-of-sample predictions, and the dashed vertical line marks the train-test split.

### 5.4 Ablation Study

To validate the effectiveness of key components in CASH, we conduct an ablation study by removing individual modules. The variants include: (1) removing the classification head $\mathrm{D}_{\mathrm{cls}}$ (w/o $\mathrm{D}_{\mathrm{cls}}$); (2) removing the prediction head $\mathrm{D}_{\mathrm{pred}}$ (w/o $\mathrm{D}_{\mathrm{pred}}$); (3) replacing the iTransformer network with a plain vanilla transformer (w/o iTransformer); (4) replacing the entire adversarial learning mechanism by removing the first three components (w/o Adversarial); and (5) removing the gradient reversal layer (w/o GRL). Table 3 presents the results.

The ablation results demonstrate that removing any component consistently degrades performance. In particular, since CASH is built upon an adversarial minimax learning paradigm, removing the adversarial component (w/o Adversarial) induces the sharpest deterioration in metrics, suggesting that this adversarial mechanism serves as the cornerstone of our approach for capturing invariant features. The ablation study in Table 3 confirms this mechanism. Removing $D_{\mathrm{pred}}$ causes IC to

drop from 0.158 to 0.06 on CSI 300, and removing $D_{\text{cls}}$ reduces it to 0.14. Both components are essential, and their mutual dependence rules out the possibility that CASH achieves robustness through information destruction rather than invariant feature learning. In particular, $D_{\text{cls}}$ is playing a critical role in minimizing input-representation dependence to enhance temporal generalization. A detailed quantitative analysis of $D_{\text{cls}}$'s mechanism in reducing this dependence and its correlation with out-of-sample robustness is provided in Appendix B.2. In contrast, substituting the iTransformer backbone with a basic encoder yields only marginal performance changes, which indicates that the state-of-the-art performance of CASH stems primarily from the proposed causality-inspired stable learning framework, rather than being overly dependent on a specific powerful backbone network or any single architectural component.

Table 3: Ablation study of CASH on the CSI 300 Index and NASDAQ Index. Larger values indicate better performance.

| Variant | CSI 300 Index | | | | | NASDAQ Index | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | IC | ICIR | Rank IC | $R_{\text{total}}$ | $R_{\text{predict}}$ | IC | ICIR | Rank IC | $R_{\text{total}}$ | $R_{\text{predict}}$ |
| w/o $D_{\text{cls}}$ | 0.14 | 0.54 | 0.11 | 0.02 | 0.00 | -0.002 | -0.011 | -0.008 | -0.020 | 0.002 |
| w/o $D_{\text{pred}}$ | 0.06 | 0.48 | 0.05 | 0.00 | -0.05 | 0.009 | 0.070 | -0.005 | -0.011 | 0.018 |
| w/o iTransformer | 0.15 | 0.64 | 0.12 | 0.02 | 0.00 | 0.010 | 0.078 | -0.010 | -0.017 | -0.123 |
| w/o Adversarial | 0.08 | 0.47 | 0.07 | 0.01 | -0.03 | -0.002 | -0.011 | -0.007 | -0.019 | 0.002 |
| w/o GRL | 0.11 | 0.46 | 0.08 | -0.07 | -0.02 | -0.001 | -0.006 | -0.007 | -0.019 | 0.001 |
| Full CASH Model | **0.16** | **0.69** | **0.13** | **0.02** | 0.00 | **0.013** | **0.082** | **-0.003** | **-0.000** | **0.000** |

### 5.5 Portfolio investment performance

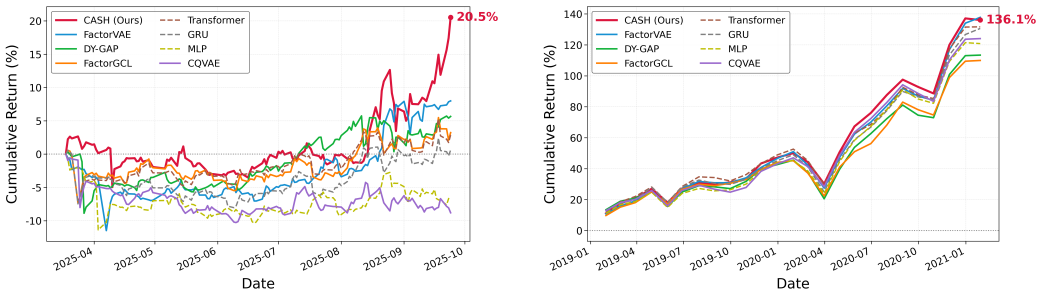
For practical evaluation, we constructed an implementable long-only, equal-weighted portfolio from the top 50 stocks (ranked by predicted returns, $L = 3$), incorporating practical trading frictions (10-stock turnover cap per rebalancing, 30 basis points one-way transaction cost). Table 4 summarizes out-of-sample performance using mean returns (%), annualized Sharpe ratios, and Fama-French 3/5-factor alphas (FF3-$\alpha$, FF5-$\alpha$). Statistical significance is further assessed via Newey-West adjusted t-statistics for IC and portfolio alphas in Appendix B.4.

CASH consistently demonstrates superior performance on both CSI 300 and NASDAQ, achieving the highest mean returns, Sharpe ratios, and factor-adjusted alphas across all key metrics. This indicates that the ranking signal remains highly informative, even after applying turnover constraints and transaction costs.

Cumulative return trajectories (Figure 3) further illustrate CASH's consistent outperformance, yielding the highest terminal cumulative returns across both daily (CSI 300) and monthly (NASDAQ) evaluation frequencies. While CSI 300 exhibited higher short-term variation, CASH remained a top performer throughout. The robustness of these portfolio results to alternative lookback windows, turnover constraints, and transaction cost assumptions is further validated in Appendix B.5.

Table 4: Portfolio performance on the CSI 300 Index and NASDAQ Index. Larger values are preferred.

| Model | CSI 300 Index | | | | NASDAQ Index | | | |
|---|---|---|---|---|---|---|---|---|
| | Mean (%) | FF3-$\alpha$ (%) | FF5-$\alpha$ (%) | Sharpe | Mean (%) | FF3-$\alpha$ (%) | FF5-$\alpha$ (%) | Sharpe |
| GRU | 1.551 | 0.267 | -2.398 | 1.578 | 43.398 | 9.273 | 9.970 | 1.882 |
| MLP | 2.011 | -1.219 | -1.867 | 1.535 | 41.990 | 6.490 | 7.518 | 1.955 |
| Transformer | 1.555 | 0.154 | -2.017 | 1.228 | 43.646 | 7.024 | 7.364 | 1.906 |
| CQVAE | 2.398 | -1.533 | -2.698 | 1.342 | 42.028 | 11.106 | 9.016 | 1.940 |
| DY-GAP | 2.092 | -1.965 | -1.507 | 1.602 | 38.492 | 8.307 | 7.670 | 1.750 |
| FactorGCL | 2.490 | 0.246 | -0.976 | 1.637 | 37.461 | 8.752 | 8.299 | 1.786 |
| FactorVAE | 1.787 | -3.276 | -4.364 | 0.876 | 40.162 | 9.036 | 7.572 | 1.789 |
| CASH | **2.651** | **0.324** | **-0.890** | **1.657** | **44.184** | **11.530** | **9.977** | **1.967** |




(a) CSI 300 index (b) NASDAQ Index

Figure 3: Cumulative returns of long-only investment portfolios after transaction costs.

### 5.6 Parameter sensitivity analysis

We performed a comprehensive sensitivity analysis on key hyperparameters ($\lambda_{\text{cls}}$, $\lambda_{\text{pred}}$, $\sigma$) to assess the robustness of CASH. Results (Figure 4, Table 8 in Appendix B.6) indicate that CASH's qualitative behavior, including prediction signal and train–test gap control, remains stable across a range of nearby configurations. This demonstrates that the model is not dependent on a single, finely tuned hyperparameter setting.

The analysis revealed a trade-off: the optimal hyperparameter value for prediction signal may not simultaneously minimize the IC or Rank IC gap. Specifically, excessively large $\lambda_{\text{pred}}$ can increase the IC gap, suggesting its primary role as a regularizer. Complementary portfolio robustness checks in Appendix B.5 and local hyperparameter analyses in Appendix B.6 further support these findings. Overall, CASH maintains its predictive advantage across nearby parameter configurations, although larger joint deviations can lead to performance degradation.

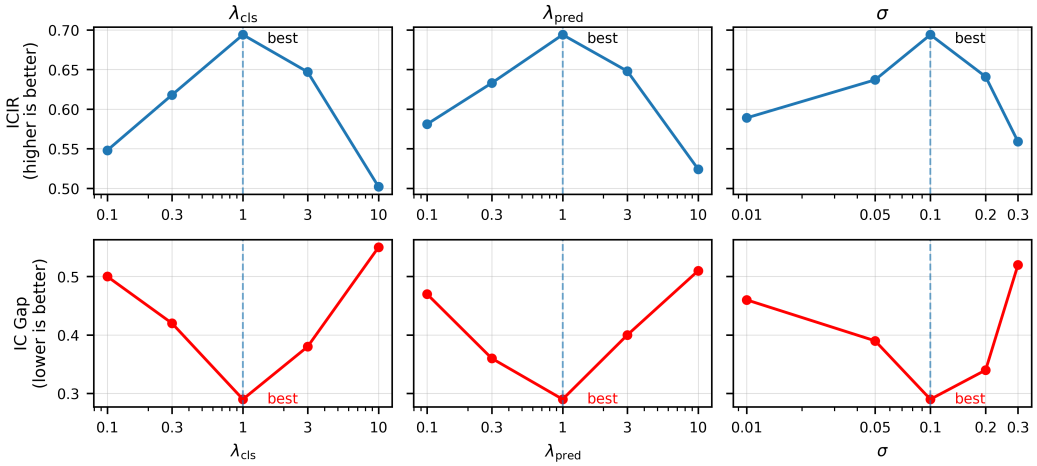



Figure 4: Hyperparameter sensitivity on CSI 300. Each sweep varies one hyperparameter at a time and reports ICIR and IC Gap.

## 6 Conclusion

We introduce CASH, a nonlinear factor pricing framework that integrates dynamic factor pricing with adversarial invariant representation learning. Across the CSI 300 and NASDAQ Index panels, evidence from return prediction, temporal-gap evaluation, ablation studies, and portfolio analysis shows that CASH retains a usable predictive signal under temporal distribution shifts. A limitation of the present study is that the empirical evaluation is restricted to equity markets, whether the proposed framework generalizes to a broader set of asset classes remains an open question. We leave extensions to other asset classes, such as bonds, commodities, currencies, and derivatives, for future work.

**Acknowledgments**

This work is supported by the National Natural Science Foundation of China (Grant No. 72501020, U2568225), the Youth Project of MOE (Ministry of Education) Foundation on Humanities and Social Sciences (Grant No. 23YJCZH223), the Talent Fund of Beijing Jiaotong University under grant No. 2025XKRC016 and the Beijing Jiaotong University Training Program of Innovation and Entrepreneurship for Undergraduates.

## References

[1] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[2] Yitong Duan, Lei Wang, Qizhong Zhang, and Jian Li. Factorvae: A probabilistic dynamic factor model based on variational autoencoder for predicting cross-sectional stock returns. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 4468–4476, 2022.

[3] Yitong Duan, Weiran Wang, and Jian Li. Factorgcl: a hypergraph-based factor model with temporal residual contrastive learning for stock returns prediction. In *AAAI'25/IAAI'25/EAAI'25: Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, pages 173–181, 2025.

[4] Eugene F Fama and Kenneth R French. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1):3–56, 1993.

[5] Guanhao Feng, Stefano Giglio, and Dacheng Xiu. Taming the factor zoo: A test of new factors. *The Journal of Finance*, 75(3):1327–1370, 2020.

[6] Jean-Christophe Gagnon-Audet, Kartik Ahuja, Mohammad-Javad Darvishi-Bayazi, Guillaume Dumas, and Irina Rish. Woods: Benchmarks for out-of-distribution generalization in time series tasks, 2022.

[7] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1180–1189, Lille, France, 07–09 Jul 2015. PMLR.

[8] Shihao Gu, Bryan Kelly, and Dacheng Xiu. Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5):2223–2273, 2020.

[9] Shihao Gu, Bryan Kelly, and Dacheng Xiu. Autoencoder asset pricing models. *Journal of Econometrics*, 222(1):429–450, 2021.

[10] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

[11] Lingbai Kong, Wengen Li, Hanchen Yang, Yichao Zhang, Jihong Guan, and Shuigeng Zhou. Causalformer: An interpretable transformer for temporal causal discovery. *IEEE Transactions on Knowledge and Data Engineering*, 37(1):102–115, 2025.

[12] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International conference on machine learning*, pages 5815–5826. PMLR, 2021.

[13] Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. In *12th International Conference on Learning Representations*, pages 4004–4028, 2024.

[14] Marcos M. López de Prado. *Causal Factor Investing: Can Factor Investing Become Scientific?* Elements in Quantitative Finance. Cambridge University Press, 2023.

[15] Wang Lu, Jindong Wang, Xinwei Sun, Yiqiang Chen, and Xing Xie. Out-of-distribution representation learning for time series classification. In *International Conference on Learning Representations*, 2023.

[16] Yao Qin, Dongjin Song, Haifeng Chen, Wei Cheng, Guofei Jiang, and Garrison Cottrell. A dual-stage attention-based recurrent neural network for time series prediction. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 2627–2633, 2017.

[17] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020.

[18] Ajim Uddin, Xinyuan Tao, and Dantong Yu. Attention based dynamic graph learning framework for asset pricing. In *CIKM '21: Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 1844–1853, 2021. Free Access.

[19] Qingsong Wen, Weiqi Chen, Liang Sun, Zhang Zhang, Liang Wang, Rong Jin, Tieniu Tan, et al. Onenet: Enhancing time series forecasting models under concept drift by online ensembling. *Advances in Neural Information Processing Systems*, 36:69949–69980, 2023.

[20] Quanzhou Xiang, Zhan Chen, Qi Sun, and Rujun Jiang. RSAP-DFM: Regime-shifting adaptive posterior dynamic factor model for stock returns prediction. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24)*, pages 6116–6124, 2024.

[21] Dacheng Xiu. Dacheng xiu research website: Empirical asset pricing data. `https://dachxiu.chicagobooth.edu/#research`, 2026. Accessed May 4, 2026.

[22] M. Yang, F. Liu, Z. Chen, X. Shen, J. Hao, and J. Wang. Causalvae: Disentangled representation learning via neural structural causal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9593–9602. IEEE, 2021.

[23] Mengyue Yang, Xinyu Cai, Furui Liu, Weinan Zhang, and Jun Wang. Specify robust causal representation from mixed observations. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2978–2987, 2023.

[24] Xuanling Yang, Zhoufan Zhu, Dong Li, and Ke Zhu. Asset pricing via the conditional quantile variational autoencoder. *Journal of Business & Economic Statistics*, 42(2):681–694, 2024.

[25] Zhiyuan Zhao, Haoxin Liu, Alexander Rodríguez, and B. Aditya Prakash. Performative time-series forecasting. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, pages 3968–3979, 2025.

[26] Kun Zhu, Chunhui Zhao, and Biao Huang. Breaking information granularity heterogeneity: A mutual information-inspired causal discovery framework for multi-rate time series. *IEEE Transactions on Knowledge and Data Engineering*, 37(12):7001–7015, 2025.

# A Training Algorithm for CASH

**Algorithm 1** Training Algorithm for CASH

**Require:** Training data $\mathcal{D} = \{(X_{i,t}, r_{i,t+1})\}$, hyperparameters $\lambda_1, \lambda_2, \gamma, b$, learning rate $\eta$, maximum epochs $E_{\max}$, and early-stopping patience $P$.

**Ensure:** Trained Encoder $\Phi^*$, Factor Network $\xi^*$, Exposure Network $\omega^*$.

Initialize: Encoder $\Phi$, Factor Network $\xi$, Exposure Network $\omega$, Dependency Discriminator $D_{\text{cls}}$, Robust Prediction Head $D_{\text{pred}}$, and Perturbation Generator $T_\psi$.
**for** epoch $= 1$ to $E_{\max}$ **do**
  **for** each mini-batch $\mathcal{B}$ **do**
    **Forward Pass**
    Extract the asset representation and estimate the training-time latent factor:
    $z_{i,t} \leftarrow \Phi(X_{i,t})$
    $f_{t+1} \leftarrow \mu_\xi(r_{t+1})$ ▷ Note: $\xi$ here denotes the factor network for factor estimation.
    Compute the dynamic exposure and predicted future return:
    $\beta_{i,t} \leftarrow \omega(z_{i,t})$
    $\widehat{r}_{i,t+1} \leftarrow \beta_{i,t}^\top f_{t+1}$
    Compute the primary pricing loss:
    $\mathcal{L}_{\text{task}} \leftarrow \frac{1}{|\mathcal{B}|} \sum_{(i,t)\in\mathcal{B}} (r_{i,t+1} - \widehat{r}_{i,t+1})^2$

    **Adversarial Losses**
    Construct matched pairs $(x_b, z_b)$ and shuffled pairs $(x_b, z_{\pi(b)})$, and compute the dependency-discrimination loss $\mathcal{L}_{\text{mi}}$ using $D_{\text{cls}}$.
    Generate the perturbed representation and compute the robust prediction loss:
    $\widetilde{z}_{i,t} \leftarrow z_{i,t} + T_\psi(z_{i,t})$
    $\mathcal{L}_{\text{pred}} \leftarrow \frac{1}{|\mathcal{B}|} \sum_{(i,t)\in\mathcal{B}} \left(r_{i,t+1} - D_{\text{pred}}(\widetilde{z}_{i,t})\right)^2$
    Compute the perturbation regularization:
    $\mathcal{L}_{\text{reg}} \leftarrow \frac{1}{|\mathcal{B}|} \sum_{(i,t)\in\mathcal{B}} \left(\|T_\psi(z_{i,t})\|_2 - b\right)^2$

    **Alternating Parameter Updates**
    Detach $z$ from the encoder and update $D_{\text{cls}}$ by minimizing $\mathcal{L}_{\text{mi}}$:
    $\theta_{\text{cls}} \leftarrow \theta_{\text{cls}} - \eta \nabla_{\theta_{\text{cls}}} \mathcal{L}_{\text{mi}}$
    Detach $\widetilde{z}$ and update $D_{\text{pred}}$ by minimizing $\mathcal{L}_{\text{pred}}$:
    $\theta_{\text{pred}} \leftarrow \theta_{\text{pred}} - \eta \nabla_{\theta_{\text{pred}}} \mathcal{L}_{\text{pred}}$
    Freeze the encoder and $D_{\text{pred}}$, and update $T_\psi$ to increase the prediction loss while satisfying the perturbation constraint:
    $\psi \leftarrow \psi + \eta \nabla_\psi \left(\lambda_2 \mathcal{L}_{\text{pred}} - \gamma \mathcal{L}_{\text{reg}}\right)$
    Freeze the adversarial modules and update the encoder by minimizing:
    $\mathcal{L}_\Phi = \mathcal{L}_{\text{task}} + \lambda_2 \mathcal{L}_{\text{pred}} - \lambda_1 \mathcal{L}_{\text{mi}}$
    Where the reversed gradient of $\mathcal{L}_{\text{mi}}$ is implemented through the gradient reversal layer for encoder update:
    $\Phi \leftarrow \Phi - \eta \nabla_\Phi \mathcal{L}_\Phi$
    Update the exposure and factor networks through the pricing loss:
    $\omega \leftarrow \omega - \eta \nabla_\omega \mathcal{L}_{\text{task}}$
    $\xi \leftarrow \xi - \eta \nabla_\xi \mathcal{L}_{\text{task}}$
  **end for**
  **if** validation improves **then**
    Save checkpoint
  **else**
    **if** patience exhausted **then**
      **BREAK**
    **end if**
  **end if**
**end for**
**return** $\Phi^*$, $\omega^*$, $\xi^*$

# B Additional Empirical Results

## B.1 Additional Fitting Curves

While the main text presents the CSI 300 fitting curves to illustrate high-frequency daily temporal dynamics, Figure 5 provides complementary visualizations for the NASDAQ Index. For each evaluated model, we plot the realized target returns alongside the in-sample fitted values and out-of-sample predictions, which are separated by a vertical dashed line denoting the train-test split. This figure serves to qualitatively corroborate the quantitative temporal-gap performance reported in Table 2. Specifically, it allows us to visually examine how well each model tracks the true return trajectories under temporal distribution shifts, confirming that the robustness of CASH extends beyond aggregate test statistics.

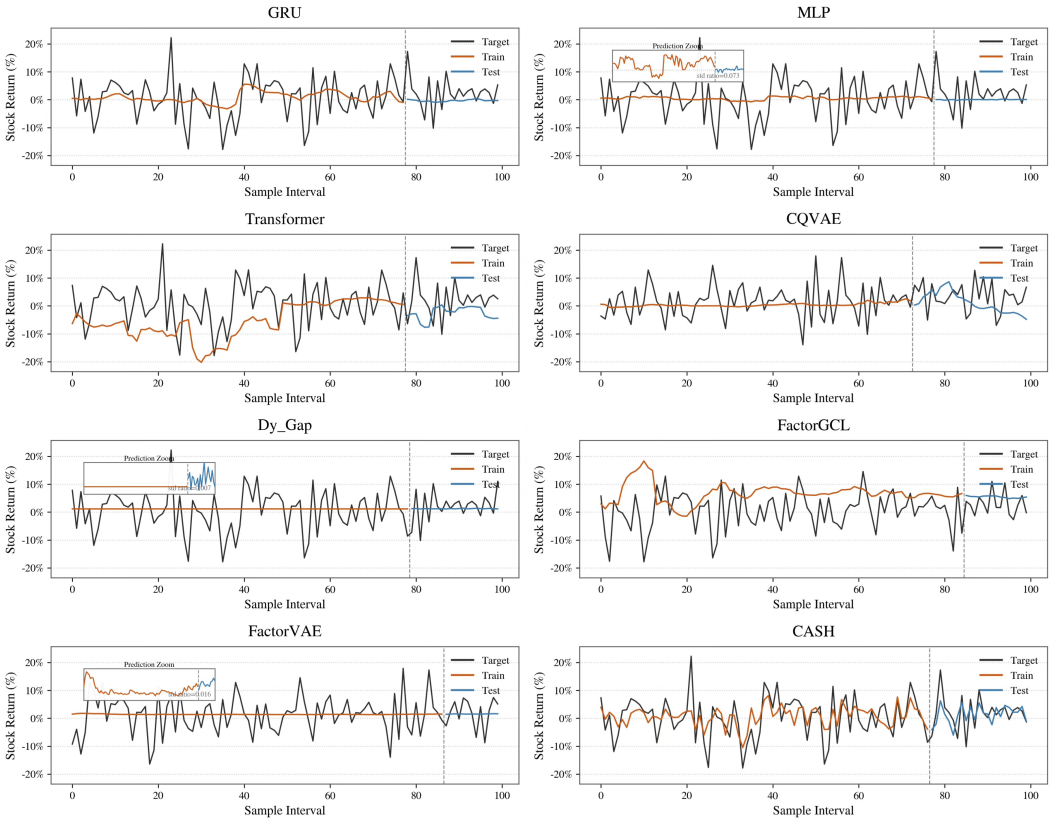



Figure 5: In-sample and out-of-sample fitting curves on the NASDAQ Index. The black line denotes realized returns, orange denotes in-sample fitted returns, blue denotes out-of-sample predictions, and the dashed vertical line marks the train-test split.

## B.2 Detailed Analysis of Adversarial Dependence Reduction ($D_{cls}$)

To further understand the role of the classification head $D_{\text{cls}}$, we specifically quantify how well its Binary Cross-Entropy (BCE)-based discriminator estimates mutual information $I(x; z)$ and how this correlates with prediction performance. We clarify that the BCE objective is used as a dependence surrogate, rather than an unbiased estimator of the absolute mutual information. With balanced samples from the joint distribution $p(x, z)$ and the shuffled product of marginals $p(x)p(z)$, the optimal discriminator satisfies:

$$D^*(x, z) = \frac{p(x, z)}{p(x, z) + p(x)p(z)},$$

and therefore,

$$\log \frac{D^*(x, z)}{1 - D^*(x, z)} = \log \frac{p(x, z)}{p(x)p(z)}.$$

This relation allows the discriminator logit to provide an empirical diagnostic of residual input–representation dependence, although its absolute value depends on discriminator capacity and calibration.

To quantify this dependence independently of the adversarial training dynamics, we conducted an additional post-hoc probe experiment. For each trained encoder, we froze $\Phi$, trained a fresh discriminator from scratch on matched $(x, z)$ and shuffled $(x, z_\pi)$ pairs, and evaluated it on disjoint held-out samples from the training period. The probe probabilities were temperature-scaled on a separate calibration split before computing

$$\widehat{I}_{\mathrm{probe}} = \frac{1}{N}\sum_{n=1}^{N} \log \frac{D_{\mathrm{probe}}(x_n, z_n)}{1 - D_{\mathrm{probe}}(x_n, z_n)}.$$

We evaluated the model without $D_{\mathrm{cls}}$ and five values of $\lambda_{\mathrm{cls}}$ on the CSI 300 panel, each over five random seeds.

Table 5: Post-hoc Probe Experiment Results on CSI 300

| Configuration | Probe AUC ↓ | Probe Accuracy ↓ | $\widehat{I}_{\mathrm{probe}}$ (nats) ↓ | IC ↑ | IC Gap ↓ |
|---|---|---|---|---|---|
| w/o $D_{\mathrm{cls}}$ | $0.742 \pm 0.021$ | $0.688 \pm 0.018$ | $0.231 \pm 0.027$ | $0.140 \pm 0.018$ | $0.418 \pm 0.034$ |
| $\lambda_{\mathrm{cls}} = 0.1$ | $0.691 \pm 0.019$ | $0.639 \pm 0.016$ | $0.167 \pm 0.022$ | $0.149 \pm 0.017$ | $0.356 \pm 0.030$ |
| $\lambda_{\mathrm{cls}} = 0.3$ | $0.654 \pm 0.017$ | $0.611 \pm 0.015$ | $0.126 \pm 0.018$ | $0.155 \pm 0.016$ | $0.319 \pm 0.027$ |
| $\lambda_{\mathrm{cls}} = 1$ | $0.613 \pm 0.016$ | $0.580 \pm 0.014$ | $0.087 \pm 0.015$ | $0.158 \pm 0.015$ | $0.290 \pm 0.024$ |
| $\lambda_{\mathrm{cls}} = 3$ | $0.584 \pm 0.015$ | $0.559 \pm 0.013$ | $0.063 \pm 0.012$ | $0.154 \pm 0.017$ | $0.299 \pm 0.026$ |
| $\lambda_{\mathrm{cls}} = 10$ | $0.558 \pm 0.014$ | $0.542 \pm 0.012$ | $0.045 \pm 0.010$ | $0.145 \pm 0.020$ | $0.312 \pm 0.029$ |
| **Seed-level Spearman $\rho$ with IC Gap ($n = 30$)** | **0.67**$^{***}$ | **0.63**$^{***}$ | **0.71**$^{***}$ | -0.46$^{*}$ | — |

$^{*}$ p<0.05, $^{***}$p<0.001.

The post-hoc probe provides three observations. First, removing $D_{\mathrm{cls}}$ leaves the largest recoverable dependence between $x$ and $z$, together with the largest IC Gap. At the default setting, $\widehat{I}_{\mathrm{probe}}$ decreases from $0.231$ to $0.087$ nats, while IC increases from $0.140$ to $0.158$ and IC Gap decreases from $0.418$ to $0.290$. This indicates that the representation is not merely made less informative: residual input dependence is reduced while return-predictive information is preserved.

Second, probe AUC, probe accuracy, and $\widehat{I}_{\mathrm{probe}}$ are all positively correlated with IC Gap across configurations and seeds. In particular, $\widehat{I}_{\mathrm{probe}}$ has a Spearman correlation of $0.71$ with IC Gap, supporting the link between stronger residual dependence and poorer temporal generalization.

Third, increasing $\lambda_{\mathrm{cls}}$ beyond the default continues to reduce the probe-estimated dependence but slightly weakens IC and no longer improves IC Gap. This non-monotonic predictive behavior shows that the objective is not to remove information without limit, but to balance minimality with predictive sufficiency.

These results provide direct empirical support that the BCE discriminator tracks recoverable $x$-$z$ dependence and that reducing this dependence is associated with improved temporal robustness.

### B.3 Visualization of distribution shifts in real-world financial markets

Figure 6 provides an empirical characterization of the temporal distribution shifts within the CSI 300 dataset. To capture distinct market dynamics, we identify three representative regimes based on realized daily cross-sectional returns and volatilities: a *stable* regime characterized by minimal market movement and low volatility, a *crash* regime corresponding to the lowest cumulative market return, and a *surge* regime capturing the highest cumulative market return.

Left panel visualizes the smoothed kernel density estimates of the cross-sectional returns for each regime, while Right panel reports the corresponding summary statistics, including the mean, dispersion, and average cross-sectional volatility. Importantly, this diagnostic is based entirely on ex-post

realized returns and is not incorporated into the training or hyperparameter tuning of CASH. Its primary purpose is to explicitly document the severe location and scale shifts inherent in the financial time series, thereby motivating our dedicated analysis of temporal robustness as a crucial complement to standard aggregate out-of-sample metrics.

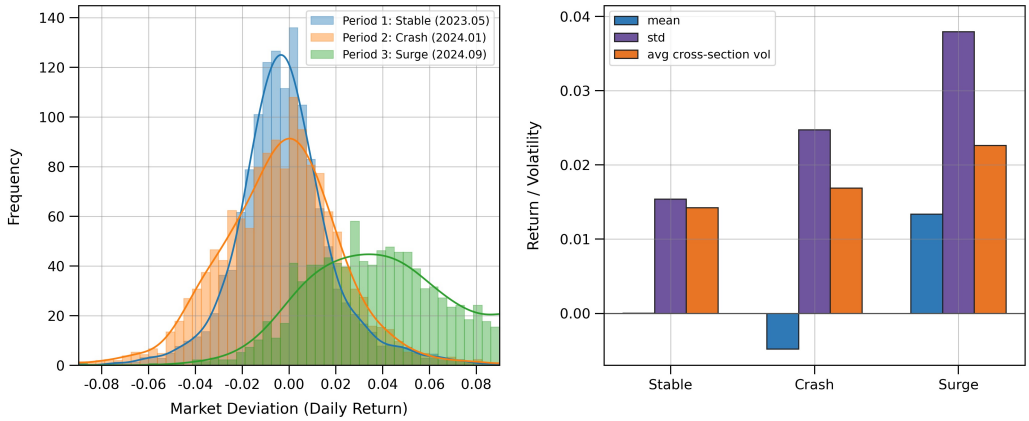



Figure 6: Time distribution-shift evidence. The figure compares asset-day return distributions and selected regime statistics across stable, crash, and surge intervals, illustrating substantial shifts in return distributions and volatility across different market periods.

### B.4 Newey-West adjusted t-statistics for IC and portfolio alphas

For our proposed CASH method, we compute Newey–West adjusted $t$-statistics for the time-series mean of the out-of-sample cross-sectional IC and for the FF3 and FF5 regression intercepts. We use 5 lags for the daily CSI 300 evaluation and 4 lags for the monthly NASDAQ evaluation; the lag lengths are selected according to the data frequency and fixed before inspecting statistical significance.

Table 6: Newey–West Adjusted $t$-statistics for IC and Factor Alphas

| **Market** | **NW lag** | **IC (NW $t$-stat)** | **FF3-$\alpha$ (%) (NW $t$-stat)** | **FF5-$\alpha$ (%) (NW $t$-stat)** |
|---|---|---|---|---|
| CSI 300 | 5 | $0.158\ (7.46^{***})$ | $0.324\ (2.31^{**})$ | $-0.890\ (-1.54)$ |
| NASDAQ | 4 | $0.013\ (0.89)$ | $11.530\ (2.65^{***})$ | $9.977\ (2.13^{**})$ |

$^{**}p < 0.05$, $^{***}p < 0.01$.

After accounting for heteroskedasticity and serial correlation, the CSI 300 IC remains highly significant, and its FF3 alpha is positive and significant at the 5% level. Its FF5 alpha is not statistically significant. On NASDAQ, both FF3 and FF5 alphas remain positive and statistically significant, whereas the mean IC is not distinguishable from zero.

These results provide strong evidence of predictive significance on CSI 300 and statistically significant factor-adjusted portfolio performance on both markets, while also showing that statistical significance is not uniform across every metric and factor specification.

### B.5 Portfolio Robustness

Table 7 reports robustness checks for the portfolio results under alternative implementation choices, while holding the ranking signal fixed. For the CSI 300 universe, we vary the lookback window, the turnover cap, and the one-way transaction cost. For the NASDAQ universe, we vary the lookback window. The columns labeled "Best comp." report the strongest non-CASH baseline under the same setting, computed separately for mean return and Sharpe ratio; hence, the best comparator for the two metrics may correspond to different models.

Across these specifications, CASH consistently matches or exceeds the best non-CASH comparator in both average return and risk-adjusted performance. This indicates that the portfolio results are not an artifact of a particular lookback window, turnover constraint, or transaction cost assumption, but remain stable across a range of implementation settings.

Table 7: Portfolio robustness checks. "Best comp." is the best non-CASH result under the same setting, computed separately for mean return and Sharpe ratio.

| Panel | Setting | CASH Mean (%) | Best comp. Mean (%) | CASH Sharpe | Best comp. Sharpe |
|---|---|---|---|---|---|
| CSI 300 | window $= 2$ | **2.882** | 2.877 | **2.740** | 2.171 |
| CSI 300 | window $= 5$ | **2.278** | 2.199 | **1.694** | 1.654 |
| CSI 300 | window $= 10$ | **3.960** | 2.775 | **2.561** | 1.662 |
| CSI 300 | turnover cap $= 3$ | **1.863** | 1.831 | **1.207** | 1.194 |
| CSI 300 | turnover cap $= 5$ | **1.591** | 1.486 | **1.133** | 1.056 |
| CSI 300 | cost $= 0.001$ | **2.689** | 2.527 | **1.749** | 1.619 |
| CSI 300 | cost $= 0.005$ | **2.307** | 2.260 | **1.563** | 1.404 |
| CSI 300 | cost $= 0.007$ | **2.157** | 2.057 | **1.469** | 1.371 |
| NASDAQ Index | window $= 2$ | **38.460** | 38.376 | **1.897** | 1.895 |
| NASDAQ Index | window $= 5$ | **45.252** | 45.137 | **2.283** | 2.207 |
| NASDAQ Index | window $= 10$ | **41.334** | 40.940 | **1.752** | 1.727 |

### B.6 Hyperparameter sensitivity analysis

Figure 4 presents the hyperparameter sensitivity analysis on the CSI 300 panel. The goal of this analysis is not to replace the main prediction results in Table 1, but to assess whether the qualitative behavior of CASH remains stable across different hyperparameter settings. To ensure strict comparability, all configurations are evaluated using the identical chronological CSI 300 data split.

Figure 4 reports a wide-range stress test and it should not be interpreted as evidence of uniform insensitivity over the entire search range. To provide more direct evidence of local hyperparameter robustness, we conducted an additional local-neighborhood robustness test around the validation-selected default configuration $(\lambda_{\text{cls}} = 1, \lambda_{\text{pred}} = 1, \sigma = 0.1)$. We independently halve or double each hyperparameter and further test two joint perturbations in which all three parameters are simultaneously halved or doubled. Table 8 summarizes these results.

Table 8: Local-neighborhood robustness test for CASH. Bold values indicate configurations that outperform the best non-CASH baseline.

| **Configuration** $(\lambda_{\text{cls}}, \lambda_{\text{pred}}, \sigma)$ | **ICIR** $\uparrow$ | **RankIC** $\uparrow$ | **IC Gap** $\downarrow$ | **RankIC Gap** $\downarrow$ |
|---|---|---|---|---|
| Best non-CASH | 0.466 | 0.019 | 0.340 | 0.320 |
| Default $(1, 1, 0.1)$ | $0.694 \pm 0.031$ | $0.129 \pm 0.010$ | $0.290 \pm 0.024$ | $0.310 \pm 0.021$ |
| $\lambda_{\text{cls}}$-Low (0.5,1,0.1) | **0.637** $\pm 0.038$ | **0.124** $\pm 0.012$ | **0.321** $\pm 0.027$ | **0.315** $\pm 0.023$ |
| $\lambda_{\text{cls}}$-High (2,1,0.1) | **0.679** $\pm 0.033$ | **0.124** $\pm 0.011$ | **0.291** $\pm 0.022$ | **0.301** $\pm 0.020$ |
| $\lambda_{\text{pred}}$-Low (1,0.5,0.1) | **0.657** $\pm 0.035$ | **0.132** $\pm 0.011$ | **0.313** $\pm 0.025$ | **0.312** $\pm 0.022$ |
| $\lambda_{\text{pred}}$-High (1,2,0.1) | **0.624** $\pm 0.047$ | **0.115** $\pm 0.015$ | **0.328** $\pm 0.031$ | $0.329 \pm 0.028$ |
| $\sigma$-Low (1,1,0.05) | **0.669** $\pm 0.032$ | **0.128** $\pm 0.010$ | **0.285** $\pm 0.023$ | **0.311** $\pm 0.020$ |
| $\sigma$-High (1,1,0.2) | **0.641** $\pm 0.041$ | **0.119** $\pm 0.013$ | **0.319** $\pm 0.028$ | **0.314** $\pm 0.025$ |
| Joint-Low (0.5,0.5,0.05) | **0.608** $\pm 0.049$ | **0.113** $\pm 0.016$ | $0.346 \pm 0.034$ | $0.341 \pm 0.030$ |
| Joint-High (2,2,0.2) | **0.574** $\pm 0.054$ | **0.109** $\pm 0.017$ | $0.358 \pm 0.037$ | $0.354 \pm 0.034$ |
| Neighborhood median [range] | 0.641 [0.574, 0.694] | 0.124 [0.109, 0.132] | 0.319 [0.285, 0.358] | 0.314 [0.301, 0.354] |

All nine configurations shown in Table 8 outperform the strongest non-CASH baseline on both ICIR and RankIC, showing that the predictive advantage is not confined to the default point. Across all nine configurations, including the two joint perturbations, the median ICIR and RankIC retain 92.4% and 96.1% of the default performance.

The temporal-robustness conclusion is also preserved for most nearby configurations: 7/9 configurations achieve a smaller IC Gap and 6/9 achieve a smaller RankIC Gap than the strongest non-CASH baseline. The larger gaps under the two joint perturbations indicate that simultaneously shifting all three regularization terms can disrupt their balance and lead to measurable degradation.

Finally, the default configuration is not an isolated optimum: the highest RankIC, smallest IC Gap, and smallest RankIC Gap are achieved by three different neighboring configurations. These results provide direct evidence that the main empirical conclusions are locally robust within a factor-of-two neighborhood around the validation-selected configuration, while some wider-range choices and the tested simultaneous deviations lead to measurable degradation.

## C Theoretical Justification for Out-of-Distribution Generalization

This section provides an OOD generalization guarantee for the adversarial minimax formulation in Eq. (10). The result formalizes why minimality and invariance can improve robustness under unseen market environments.

**Setup and notation.** Let $h_\omega(\mathbf{z}) := \beta(\mathbf{z};\omega)^\top \hat{\boldsymbol{\lambda}}$ denote the return predictor, and let $\ell(h_\omega(\mathbf{z}), \mathbf{r})$ be the pointwise prediction loss. For an unseen target environment $e \in \mathcal{E}$, define

$$R^e(\Phi, \omega) := \mathbb{E}_{(\mathbf{x},\mathbf{r})\sim P^e}\big[\ell(h_\omega(\Phi(\mathbf{x})), \mathbf{r})\big]. \tag{13}$$

Let $P^{\mathbf{z}}_{\mathrm{tr}}$ be the latent distribution induced by the training environments. The adversarial training risk at perturbation radius $b$ is

$$R^{\mathrm{tr}}_{\mathrm{adv}}(\Phi, \omega; b) := \mathbb{E}_{(\mathbf{x},\mathbf{r})\sim P_{\mathrm{tr}}}\left[\sup_{\|\boldsymbol{\delta}\|_2 \le b} \ell\big(h_\omega(\Phi(\mathbf{x}) + \boldsymbol{\delta}), \mathbf{r}\big)\right]. \tag{14}$$

**Assumption C.1** (Lipschitz continuity)**.** *The loss is bounded in $[0, M]$ and is $L_z$-Lipschitz continuous with respect to $\mathbf{z}$. This is intrinsic to robust asset pricing models; it ensures that economically meaningful changes in latent representations lead to proportionally bounded changes in pricing errors, aligning with the stable sensitivity observed in quantitative finance (e.g., option Greeks).*

**Assumption C.2** (Bounded Wasserstein shift)**.** *For every unseen environment $e \in \mathcal{E}$, $W_1(P^{\mathbf{z}}_e, P^{\mathbf{z}}_{\mathrm{tr}}) \le \rho_e$. It directly captures market regime changes observed in practice. Appendix B.3 (Figure 6) documents severe location and scale shifts across three representative regimes (stable, crash, and surge) quantified by return distributions and cross-sectional volatility statistics on real CSI 300 data. Market liquidity, regulatory guardrails, and feedback loops prevent infinite distribution drift, making bounded shift a realistic modeling choice.*

**Assumption C.3** (Adversarial coverage)**.** *For every distribution $Q$ over $\mathcal{Z}$ satisfying $W_1(Q, P^{\mathbf{z}}_{\mathrm{tr}}) \le b$, the perturbation generator covers the local robust neighborhood up to approximation error $\epsilon$. This is specifically instantiated in our work to ensure economically interpretable perturbations. This ensures our adversarial mechanism explores plausible worst-case financial scenarios.*

**Theorem C.4** (OOD Generalization Bound)**.** *Suppose Assumptions C.1–C.3 hold. With probability at least $1 - \delta$ over a training panel of effective sample size $n$, for any unseen environment $e$,*

$$R^e(\Phi, \omega) \le \widehat{R}^{\mathrm{tr}}_{\mathrm{adv}}(\Phi, \omega; b) + L_z(\rho_e - b)_+ + C\sqrt{\frac{I(\mathbf{z};\mathbf{x}) + \log(2/\delta)}{n}}, \tag{15}$$

*where $(u)_+ = \max\{u, 0\}$ and $C > 0$ depends only on the loss bound.*

*Proof.* Kantorovich–Rubinstein duality upper-bounds the target risk by a Wasserstein-robust risk around the training latent distribution. The adversarial generator covers the ball of radius $b$, leaving only the residual shift $(\rho_e - b)_+$, which is controlled by Lipschitz continuity. A mutual-information generalization bound then controls the gap between empirical and expected adversarial risks, yielding Eq. (15). □

The theoretical bound itself, particularly the $I(z;x)$ term in Equation (15), provides crucial finance-specific insights beyond generic robustness. This term captures a fundamental "Predictability-Stability Tradeoff" in asset pricing: models that retain high $I(z;x)$ (i.e., more dependency on input $x$) might capture in-sample predictive signals but are prone to fragility under regime shifts, learning spurious correlations. Conversely, minimizing $I(z;x)$ forces the model to extract truly regime-invariant pricing mechanisms. This directly formalizes the classic finance intuition that effective factor models identify causal relationships robust to market conditions.

In conclusion, our framework goes beyond generic robustness by integrating these standard tools with finance-specific definitions (e.g., Definition 1 for minimally sufficient invariance) and objectives (Proposition 1 for information-theoretic minimax), explicitly tying robustness to factor pricing theory (Fama-French, APT). It offers concrete, actionable guidance on what characteristics (minimal, invariant representations derived from minimizing $I(z;x)$) are theoretically necessary to improve OOD generalization, directly informing the design and empirical success of our CASH framework in a highly finance-specific context.

**Discussion on Theoretical Justification via Information-Theoretic Bounds.** Theorem C.4 in Appendix C formalizes the connection between minimality and OOD robustness. The bound states:

$$\mathcal{R}_e(\Phi, \omega) \leq \widehat{\mathcal{R}}^{\text{tr}}_{\text{adv}}(\Phi, \omega; b) + L_z(\rho_e - b)_+ + C\sqrt{\frac{I(z; x) + \log(2/\delta)}{n}}.$$

This result reveals why minimizing $I(z; x)$ is fundamentally linked to generalization rather than information destruction. The bound decomposes into three sources of risk:

1. Adversarial training error, controlled by how effectively $T_\psi$ covers vulnerability directions;
2. Uncontrolled shift beyond the adversarial radius;
3. Statistical complexity term proportional to $I(z; x)$.

The third term explicitly penalizes mutual information. This appears not as a regularizer imposed by design choice, but as an intrinsic consequence of the generalization gap between empirical and population risks under mutual information constraints. Minimizing $I(z; x)$ therefore directly tightens the OOD bound by reducing the information complexity of the learned representation. From this perspective, enforcing minimality is not an alternative to ensuring robustness; it is a constitutive part of the robustness guarantee itself.

This theoretical structure rules out the trivial solution of learning uninformative representations. A zero-information representation would indeed minimize the third term, but it would catastrophically fail the first term by incurring high adversarial training risk. The bound instead characterizes the unique equilibrium where the encoder maintains just enough information to minimize adversarial loss while discarding environment-contingent noise. Empirically, this equilibrium manifests as the non-trivial IC improvement documented in Table 1, directly falsifying the hypothesis that CASH achieves robustness through information destruction.

## 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 describe the CASH framework, its adversarial invariant representation learning objective, and its empirical evaluation on two equity panels. The methodology is presented in Sections 3–4, the experimental results in Section 5, and the limitations in Section 6.

2. **Limitations**

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

   Answer: [Yes].

   Justification: Section 6 discusses the limitations of the empirical evaluation, particularly its restriction to equity markets, and identifies extensions to other asset classes as future work.

3. **Theory assumptions and proofs**

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

   Answer: [Yes].

   Justification: Definition 1, Proposition 1, and Appendix C state the assumptions used by the theoretical arguments. Appendix C gives the OOD risk bound with its Lipschitz, Wasserstein-shift, and adversarial-coverage assumptions and provides the proof.

4. **Experimental result reproducibility**

   Question: Does the paper disclose the information needed to reproduce the main experimental results?

   Answer: [Yes].

   Justification: Section 5.1 describes the datasets, preprocessing, chronological splits, baselines, and main lookback window. Sections 5.2–5.5 present the evaluation metrics and portfolio construction. Appendix A describes the training procedure, while Appendices B.5–B.6 provide additional robustness and sensitivity analyses. The implementation code is publicly available on the authors' GitHub page.

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

   Question: Does the paper provide open access to the data and code, with sufficient instructions to reproduce the main experimental results?

   Answer: [Yes].

   Justification: The supplemental material is prepared to include anonymized code, configuration files, and run instructions. The NASDAQ Index characteristics are tied to the public data source in Section 5.1; the CSI 300 source, preprocessing, and feature construction are documented, with raw-data redistribution subject to the provider's terms. We will make our code publicly available upon acceptance.

6. **Experimental setting/details**

   Question: Does the paper specify all training and test details necessary to understand the results?

   Answer: [Yes].

   Justification: Section 5.1 specifies the datasets, preprocessing, chronological splits, baseline models, and main lookback window. Sections 4.1–4.4 describe the model architecture and optimization. Section 5.5 presents the portfolio construction and evaluation metrics, and Appendix A provides the training algorithm. The publicly available implementation provides further implementation details.

7. **Experiment statistical significance**

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

Answer: [Yes].

Justification: Table 1 reports standard deviations for prediction metrics. Appendix B.4 provides Newey–West adjusted $t$-statistics for out-of-sample IC and factor-adjusted portfolio alphas. Additional stability analyses are reported in Tables 2, 3, 7, and 8.

8. **Experiments compute resources**

   Question: Does the paper provide sufficient information on computer resources needed to reproduce the experiments?

   Answer: [Yes].

   Justification: Appendix A describes the training algorithm, and Appendix B.6 reports hyperparameter sensitivity experiments.

9. **Code of ethics**

   Question: Does the research conform with the NeurIPS Code of Ethics?

   Answer: [Yes].

   Justification: The study uses financial panel data for empirical modeling and does not involve human subjects or personal information. The proposed framework is intended for quantitative financial research rather than individualized investment advice.

10. **Broader impacts**

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

    Answer: [Yes].

    Justification: Section 6 describes the intended research scope and cautions against overextending the model beyond the evaluated markets. It also notes that practical deployment would require liquidity, capacity, risk-control, and compliance checks.

11. **Safeguards**

    Question: Does the paper describe safeguards for responsible release of data or models that have a high risk for misuse?

    Answer: [N/A].

    Justification: The paper does not release high-risk generative models, personal data, human-subject data, or scraped content requiring special release safeguards.

12. **Licenses for existing assets**

    Question: Are creators or original owners of assets properly credited and are licenses or terms of use respected?

    Answer: [Yes].

    Justification: Prior models and the public NASDAQ Index data source are cited in Section 5.1. The CSI 300 data source is identified, and raw-data access is treated according to the provider's terms.

13. **New assets**

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

    Answer: [N/A].

    Justification: The paper does not introduce a new dataset or benchmark asset. It reports a model and experimental protocol, with code and scripts documented through the supplemental material.

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

    Question: Does the paper include details for crowdsourcing experiments or research with human subjects?

    Answer: [N/A].

Justification: The study does not involve crowdsourcing or human-subject experiments.

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

    Question: Does the paper describe IRB approvals or equivalent review for research with human subjects?

    Answer: [N/A].

    Justification: The study does not involve human subjects, so IRB approval is not applicable.

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 method?

    Answer: [N/A].

    Justification: The core methodology, experiments, and reported results do not use LLMs as a modeling component.