From Pixels to Signals: Hierarchical Vision Transformers for Return Prediction*

Jonas Theill Bøjstrup[1,§] Bezirgen Veliyev[1,‡,§,†] Jesper N. Wulff[1,†]

jobo@econ.au.dk bveliyev@econ.au.dk jwulff@econ.au.dk

Abstract

Recent evidence suggests that candlestick-chart image predictability weakens in more recent data and remains hard to interpret. Using CRSP U.S. equities (1993–2024), we revisit this claim by comparing a CNN benchmark with a hierarchical vision transformer (Hiera) trained on candlestick-and-volume images. Hiera produces a stronger and more stable cross-sectional return ordering that persists in a post-2019 holdout and survives implementability checks. We develop two diagnostics—controlled synthetic probes and occlusion maps—paired with spanning regressions to characterize the learned signal. The diagnostics constrain interpretation and localize predictive content to patterns near the end of the lookback window.

*We are grateful to Erik Christian Montes Schütte, Jonas Eriksen, and Tim Bollerslev for helpful comments. We also thank participants at the CoRE Retreat (Sandbjerg Gods, 2025) and the DPGE Workshop (2025) for valuable discussions and suggestions. We gratefully acknowledge funding support from Aarhus University Research Foundation (AUFF-E-2025-9-9). Veliyev and Wulff are affiliated with the Danish Finance Institute (DFI), which financially supports researchers in Finance.

[1]Department of Economics and Business Economics, Aarhus University, Denmark.
[§]Center for Research in Energy: Economics and Markets (CoRE), Aarhus University, Denmark.
[‡]Aarhus Center for Econometrics (ACE), Aarhus University, Denmark.
[†]Research fellow at the Danish Finance Institute (DFI).

1 Introduction

Price charts compress the trading history of a security into a visually structured object—candlesticks and volume charts that practitioners routinely use to summarize trends and turning points in technical analysis. A long-standing question in empirical finance is whether such price-history patterns contain predictive information that survives market evolution and trading frictions. Because any remaining predictability is likely to be nonlinear and to depend on interactions among local and longer-window features, it is natural to revisit chart-based signals through the lens of modern machine learning.

Recent work sharpens the issue in the specific context of image-based forecasting. Jiang et al. (2023) show that convolutional neural network (CNN) models applied to candlestick charts can generate out-of-sample portfolio spreads by learning predictive patterns without pre-specifying technical rules. However, replication evidence suggests that these CNN-based chart signals weaken once the sample is extended beyond 2019 (Kaczmarek and Pukthuanthong, 2024), consistent with broader evidence that return predictors can decay as markets evolve and as signals become widely known (McLean and Pontiff, 2016). This raises two questions that are central for assessing the empirical relevance of chart-image predictors. First, does chart-image predictability remain economically meaningful in more recent data when evaluated using modern computer-vision architectures under a finance-standard portfolio protocol? Second, if performance improves, what features of the chart images appear to drive any incremental gains?

We address these questions by revisiting chart-image forecasting with a hierarchical vision transformer architecture designed to represent images at multiple resolutions. In price charts, short-horizon cues often reside in the most recent days, but their meaning depends on the surrounding path (e.g., whether a move follows a drawdown, a range, or a breakout attempt). A hierarchical transformer is built to integrate such local configurations with broader within-window context. We therefore adopt Hiera (Ryali et al., 2023) as our primary architecture and evaluate it against a CNN benchmark under the same task and

portfolio protocol.

Our empirical analysis uses daily data on U.S. common stocks from CRSP over 1993–2024. We construct candlestick-and-volume images using rolling lookback windows (with a 20-trading-day window as the main specification) and train models to forecast the direction of 5-day-ahead returns. Following a strictly chronological design, we use 1993–2000 for model development (training and validation, split at the firm level) and evaluate out-of-sample performance over 2001–2019, with 2020–2024 treated as a post-2019 holdout regime. Daily model predictions are converted into weekly cross-sectional decile portfolios held for five trading days. Performance is summarized by annualized High–Low (H–L) return spreads and annualized Sharpe ratios, reported both gross and net of turnover-based transaction costs, for a broad universe and a large-cap (top-500) subset.

Our main findings are as follows. Relative to the CNN benchmark, the hierarchical model delivers a stronger and more stable cross-sectional return ordering. In the primary window-20 specification, the combined out-of-sample period (2001–2024) H–L return spread increases from 0.15 (Sharpe 1.43) for the benchmark to 0.24 (Sharpe 2.17), and the improvement is present in both 2001–2019 and the post-2019 holdout (2020–2024). The advantage persists under a large-cap (top-500) restriction and after turnover-based transaction-cost adjustments, although magnitudes compress.

We also provide evidence that disciplines interpretation of the learned signal. In spanning regressions, the Hiera-based chart-image signal overlaps with standard technical predictors but is not subsumed by them. Two complementary diagnostics developed in this paper sharpen what the model appears to use. First, controlled synthetic probes show that predictions respond to structured price patterns rather than to mechanical changes in trend, scale, or volatility alone. Second, occlusion diagnostics localize predictive content within the lookback window and reveal strong terminal concentration: sensitivity on real inputs is disproportionately concentrated near the end of the window. Taken together, the regressions and diagnostics offer a clearer picture of what the model is picking up in charts, without requiring a unique mapping to named technical rules.

This paper contributes to two debates in empirical finance. The first concerns the durability and implementability of chart-based predictors. Using a strictly chronological design, a post-2019 holdout through 2024, and portfolio tests with large-cap and transaction-cost adjustments, we provide a durability reassessment of chart-image predictability and show that modern hierarchical architectures deliver a more stable return ordering than a CNN benchmark in this setting. The second concerns interpretation of machine-learned return signals. We provide a reusable set of diagnostics—spanning regressions, controlled probes, and occlusion-based localization—that bounds what the model appears to learn and clarifies how predictive content is distributed within the lookback window.

1.1 Related Literature

Our paper relates to several strands spanning (i) price-based predictability and technical analysis, (ii) machine learning in empirical asset pricing, and (iii) unstructured inputs for financial prediction.

First, our work connects to research on return predictability from past prices and to the long-standing debate on technical analysis and chart-based signals. A large literature documents momentum and reversal patterns across horizons (De Bondt and Thaler, 1985; Jegadeesh, 1990; Jegadeesh and Titman, 1993; Carhart, 1997; Rouwenhorst, 1998). Evidence on rule-based technical trading strategies is mixed once one accounts for data-snooping concerns and trading frictions. Early empirical support for simple rules (Brock et al., 1992) motivated a large follow-up literature emphasizing data-snooping adjustments and implementation (Sullivan et al., 1999; Bessembinder and Chan, 1998; Bajgrowicz and Scaillet, 2012). Related work also proposes systematic, prespecified pattern-detection approaches (Lo et al., 2000) and finds that some indicators can retain predictive content in certain settings (Han et al., 2013; Neely et al., 2014). We differ from rule-based technical analysis by treating charts as a high-dimensional input and allowing the model to learn which visual configurations, if any, carry incremental predictive content under a finance-standard out-of-sample protocol.

Second, we relate to the rapidly growing machine learning literature in empirical asset pricing, which emphasizes flexible prediction under disciplined out-of-sample evaluation and transparent benchmarking (Kelly et al., 2019; Gu et al., 2020). Representative applications span cross-sectional stock return prediction and regularization-based model selection (Chinco et al., 2019), factor construction and interpretation (Kozak et al., 2018; Feng et al., 2020), and other asset classes and objects of interest (Bianchi et al., 2021; Fuster et al., 2022; Chen et al., 2024). A parallel stream embeds modern deep-learning architectures directly into asset-pricing models and trading problems (Kelly et al., 2025; Guijarro-Ordonez et al., 2025). Our contribution is to bring modern vision-learning tools into the empirical-asset-pricing evaluation framework using chart images as inputs, while maintaining finance-standard portfolio tests, large-cap robustness checks, and turnover-adjusted net performance.

Third, we connect to work that uses unstructured inputs—including images—for financial prediction. Representative examples include image-based sentiment measures (Obaid and Pukthuanthong, 2022) and satellite imagery proxies for economic activity and information frictions (Hao et al., 2024). A related stream tests whether machine learning can extract stable cross-sectional predictability from historical performance measures that mirror what is visible in price plots (Murray et al., 2024). The closest related approach to ours transforms price charts into images and applies deep learning to forecast short-horizon returns (Jiang et al., 2023). Beyond this, existing image-based finance studies are often limited by first-generation convolutional architectures, narrower universes, or shorter sample windows (e.g., Bang and Ryu, 2023; Ren et al., 2024; Ye et al., 2024; Gezici and Sefer, 2024). Motivated by concerns about post-2019 durability (Kaczmarek and Pukthuanthong, 2024), we reassess chart-image forecasting under finance-standard, large-scale out-of-sample portfolio tests and complement performance evidence with interpretability diagnostics designed to clarify what drives any incremental gains. Methodologically, we build on modern computer-vision advances in transformer-based representation learning—including hierarchical vision transformers that provide explicit multi-scale visual representations (e.g., Dosovitskiy et al., 2020; Liu et al., 2021; Ryali et al., 2023)—and evaluate their value

added in a finance-standard prediction-and-portfolio setting.

The remainder of the paper proceeds as follows. Section 2 outlines the model architecture and training methodology. Section 3 describes the data, experimental design, and empirical results, including model comparisons and interpretability analyses. Section 4 concludes.

2 Methodology

In this paper, we study short-horizon return predictability from chart images. For each stock i and trading day t, the input $I_{i,t}$ is a standardized candlestick-and-volume image constructed solely from information available up to day t using a rolling lookback window (with $L = 20$ as the baseline specification). The chart rendering, normalization, and alternative lookback windows are described in Subsection 3.1. The prediction target is the sign of the subsequent 5-day-ahead return: we set $y_{i,t} = 1$ if the 5-day-ahead return is positive and $y_{i,t} = 0$ otherwise. The learning problem is binary classification, and the model output $p_{i,t}$ is interpreted as an estimate of $P(y_{i,t} = 1 \mid I_{i,t})$.

Economic performance is evaluated through weekly cross-sectional portfolio tests. Each day we generate model scores for all stocks, form weekly decile portfolios by sorting on the signal, and hold positions for five trading days to match the prediction horizon. Performance is summarized by the annualized H–L spread and annualized Sharpe ratio, reported both gross and net of turnover-based transaction-cost adjustments, and reported for both a broad universe and a large-cap (top-500) subset. Unless stated otherwise, all models (including the CNN benchmark) use the same image representation and prediction target and differ only in architecture and training.

2.1 Hiera Model and Learning Framework

Figure 2.1 summarizes the learning setup. The model maps a chart image $I_{i,t}$ to a probability forecast $p_{i,t}$ for a positive 5-day-ahead return. The student network is Hiera, a hierarchical vision transformer initialized from ImageNet-pretrained weights[1]. Training uses supervised classification with a teacher–student consistency regularizer (Mean Teacher) (Tarvainen and Valpola, 2017). The student is updated by gradient descent, while the teacher is an exponential moving average (EMA) of student parameters and provides

[1]The publicly available pretrained weights are obtained via large-scale pretraining in the upstream training pipeline, which includes masked-image modeling objectives such as Masked Autoencoder (He et al., 2022).

target predictions for the consistency loss. Appendix A1 reports the update rule and the ramp-up schedules used for the EMA decay and consistency weight.

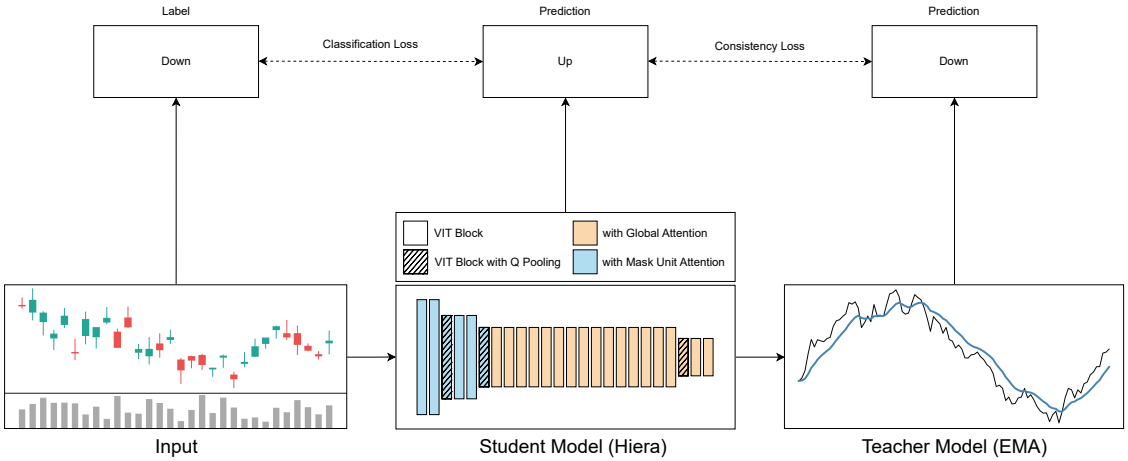


Figure 2.1: Hiera–Mean Teacher framework. The student model (Hiera) maps a candlestick-and-volume image to an Up/Down probability and is trained using supervised classification together with a teacher–student consistency loss. The teacher model is an exponential moving average of the student parameters and provides a stabilized target prediction for the consistency term. Within Hiera, attention blocks alternately mix information broadly across the image (global attention) or within local token groups (mask-unit attention), while Q-pooling reduces the number of tokens between stages to form a progressively coarser, multi-scale representation.

The student network, Hiera (Ryali et al., 2023), is a hierarchical vision transformer that builds multi-scale representations by progressively aggregating information from fine to coarse spatial resolutions. This design permits the model to encode localized price/volume configurations near the end of the lookback window while also representing broader within-window context (e.g., the location of recent movements relative to the preceding path). This is consistent with our later diagnostics, which quantify where within the lookback window the model-implied signal concentrates. Following standard vision-transformer practice, the input image I is a standard RGB image represented as a height-by-width pixel grid with three color channels (red, green, and blue), which is partitioned into non-overlapping $p \times p$ patches and mapped to a sequence of token embeddings.

The token sequence is processed by a sequence of transformer stages. A defining feature of Hiera is the use of *mask units*, which restrict self-attention to local groups of tokens. This encourages early layers to focus on local chart structure while deeper stages

aggregate information at coarser spatial resolutions through hierarchical downsampling. We follow the architecture and block definitions in Ryali et al. (2023); the key object for our application is the resulting multi-scale representation that jointly captures local configurations and broader within-window context.

2.2 Training and Implementation

We summarize the objectives used for supervised fine-tuning with teacher–student consistency regularization. The supervised term is cross-entropy on the Up/Down label. The consistency term penalizes deviations between student and teacher predicted probabilities under stochastic perturbations induced by training-mode randomness (e.g., dropout).

During fine-tuning, following Mean Teacher (Tarvainen and Valpola, 2017), the teacher parameters θ' are maintained as an exponential moving average of the student parameters θ:

$$\theta'_t = \alpha_t \, \theta'_{t-1} + (1 - \alpha_t) \, \theta_t, \tag{2.1}$$

where $\alpha_t \in (0, 1)$ is the EMA decay coefficient. The teacher provides target predictions for a consistency objective that matches student and teacher predicted probabilities under stochastic perturbations.

Let $\mathcal{L}_{\text{class}}\Big(f(x;\theta), y\Big)$ denote the supervised cross-entropy loss, where $f(x;\theta)$ are the student logits for input x. We define the Mean Teacher consistency term as mean squared error between student and teacher softmax probabilities under training-mode randomness,

$$J(\theta) = \mathbb{E}_x\Big[\Big\|\text{softmax}(f(x;\theta)) - \text{softmax}(f(x;\theta'))\Big\|^2\Big].$$

The fine-tuning objective is

$$\mathcal{L} = \mathcal{L}_{\text{class}}\Big(f(x;\theta), y\Big) + \lambda_t \, J(\theta), \tag{2.2}$$

where $\lambda_t \geq 0$ is a (ramped) consistency weight.

We fine-tune the student network on the 1993–2000 development sample and select hyperparameters using the validation split. Unless stated otherwise, we use the same training protocol across lookback windows and universes. Optimization uses AdamW (Loshchilov and Hutter, 2019) with decoupled weight decay and warmup followed by cosine annealing (Loshchilov and Hutter, 2017). We fine-tune the classification head and the last four transformer blocks using separate optimizer parameter groups with block-specific base learning rates. We apply global norm gradient clipping (Pascanu et al., 2013) and use early stopping based on validation loss. For Mean Teacher, we ramp the consistency weight λ_t and the EMA decay α_t as described in Appendix A1.

All optimization and schedule details—including the warmup, cosine learning-rate schedule, weight decay, gradient clipping threshold, and the ramp-up specifications for λ_t and α_t—are provided in Appendix A1, with selected hyperparameter values summarized in Appendix Table A1.1. After selecting hyperparameters in the 1993–2000 development period, we keep the specification fixed and report out-of-sample results over 2001–2019 and 2020–2024 without retraining on later data.

3 Empirical Analysis

3.1 Data

The equity data are obtained from CRSP (v2), covering U.S. common stocks listed on NYSE, AMEX, and NASDAQ from January 1993 through December 2024. To form chart images, we require daily open, high, low, close and volume (OHLCV). As a preprocessing step, we remove any ticker that has at least one missing observation in these OHLCV fields anywhere in the sample to avoid ad hoc imputation and ensure that the image-construction procedure is mechanically well-defined for all windows and dates. This restriction changes the investable universe by design; our main conclusions are qualitatively unchanged in large-cap subsamples where OHLCV coverage is nearly complete. We then work with the resulting sample in which firms enter and exit over time. Across the full sample we have roughly 7,000 unique firms, with the active cross-section varying by date.

We adopt a strictly chronological split into three non-overlapping periods. The years 1993–2000 are reserved for model development (training and validation). The period 2001–2019 is the primary out-of-sample evaluation window. The years 2020–2024 form a separate post-2019 holdout used to assess out-of-sample stability in more recent data.

Candlestick-chart images are generated by a rolling-window procedure at the daily frequency. For each firm and date t with sufficient history, we render a chart using the most recent $L \in \{5, 20, 60\}$ trading days (see Figure 3.1 for example images) ending at t, and we predict the sign of the subsequent five-trading-day return.

Each image is a 224×224 RGB rendering of a candlestick chart with a volume panel and an overlaid simple moving average computed within the L-day window (with $MA = L$). Charts are rendered with axes removed and a fixed style to standardize appearance across firms and time. Images are labeled "Up" or "Down" based on the sign of the close-to-close return over the next five trading days: $y_{i,t} = 1$ if subsequent five-trading-day return is positive and $y_{i,t} = 0$ otherwise, where t denotes the last day in the lookback window.

The image tensor used for model training is obtained by resizing the rendered chart to the input resolution required by the model. For Hiera, we use 224×224 inputs. For the Jiang et al. (2023)-style CNN benchmark, we follow the paper's input geometry by resizing to $(32, 15)$ for $L = 5$, $(64, 60)$ for $L = 20$, and $(96, 180)$ for $L = 60$.[2] Additional rendering, resizing, and normalization details are summarized in Appendix A1.2.

The 1993–2000 development period contains roughly 0.6 million images per window, and the out-of-sample periods contain roughly 1.5 million images per window. To prevent cross-firm leakage within the development period, we partition firms (tickers) into training and validation sets: 70% of firms are assigned to training and 30% to validation, and all images for a given firm belong to the same split.

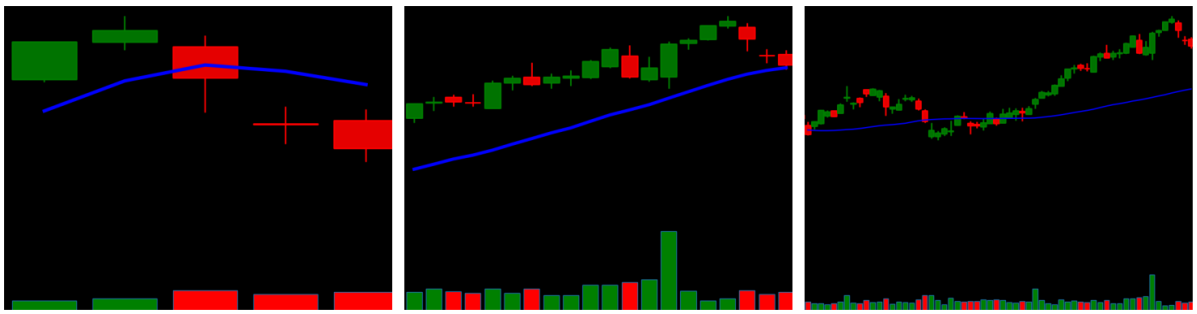


Figure 3.1: Example candlestick renderings for $L \in \{5, 20, 60\}$.

We treat $L = 20$ as the main specification. The $L \in \{5, 60\}$ windows serve as robustness checks for sensitivity to the temporal resolution of the input representation. For benchmarking, we implement the CNN architecture in Jiang et al. (2023) and train it on the same labels, chronological split, and portfolio-evaluation protocol as our main models. We do not seek an exact replication of their full pipeline, as our universe and chart-rendering procedure are constructed independently; the benchmark is intended to isolate the role of model architecture under a common task definition.

[2] Resizing is applied after rendering and does not use forward-looking information.

3.2 Empirical Results

We evaluate economic relevance using cross-sectional portfolio sorts based on each model's predicted probability of a positive 5-day return. At each formation date, we sort stocks into deciles by the model score and form equal-weighted decile portfolios. Portfolios are held for $H = 5$ trading days to match the prediction horizon, and the next rebalance occurs on the trading day immediately after the holding period ends. Table 3.1 reports annualized decile returns and Sharpe ratios for the main $L = 20$ lookback window, shown separately for 2001–2019, 2020–2024, and the combined sample. We then assess whether the ranking advantage persists under large-cap restrictions, and after turnover-based transaction-cost adjustments.

3.2.1 Main Results

Table 3.1: Window 20 Performance Comparison

	Benchmark (2001–2019)		Hiera (2001–2019)		Benchmark (2020–2024)		Hiera (2020–2024)		Benchmark (2001–2024)		Hiera (2001–2024)	
Decile	Ret	SR	Ret	SR	Ret	SR	Ret	SR	Ret	SR	Ret	SR
Low	0.05	0.24	0.01	0.03	0.08	0.32	0.06	0.22	0.06	0.26	0.02	0.08
2	0.09	0.41	0.06	0.29	0.10	0.43	0.08	0.32	0.09	0.41	0.06	0.30
3	0.12	0.58	0.09	0.41	0.12	0.49	0.08	0.36	0.12	0.56	0.08	0.40
4	0.11	0.54	0.12	0.61	0.12	0.51	0.11	0.46	0.11	0.53	0.12	0.58
5	0.12	0.59	0.13	0.65	0.12	0.51	0.13	0.55	0.12	0.57	0.13	0.63
6	0.14	0.68	0.15	0.73	0.14	0.58	0.14	0.59	0.14	0.66	0.15	0.70
7	0.14	0.68	0.16	0.80	0.16	0.66	0.16	0.68	0.14	0.68	0.16	0.77
8	0.14	0.71	0.17	0.80	0.17	0.72	0.19	0.77	0.15	0.71	0.17	0.79
9	0.20	0.97	0.20	0.95	0.17	0.70	0.19	0.77	0.19	0.91	0.20	0.90
High	0.21	1.02	0.26	1.19	0.23	0.88	0.29	1.08	0.22	0.98	0.26	1.16
H–L	0.16***	1.51	0.25***	2.29	0.14**	1.16	0.22***	1.77	0.15***	1.43	0.24***	2.17

Note: Gross performance for decile and H-L portfolios is reported separately for 2001–2019, the post-2019 holdout (2020–2024), and the combined sample (2001–2024). At each formation date, stocks are sorted into deciles based on model scores and equal-weighted decile portfolios are formed. Portfolios are held for a five-trading-day holding period ($H = 5$), and the next rebalance occurs on the trading day immediately after the holding period ends. Returns are annualized geometric means of the realized holding-period returns; Sharpe ratios are computed from the holding-period return series using annualization factor $\sqrt{252/H}$. Test statistics are reported for H-L portfolios and are based on the annualized arithmetic mean of holding-period returns using Newey–West standard errors (4 lags). Statistical significance at the 10%, 5%, and 1% levels is denoted by *, **, and ***, respectively.

Table 3.1 shows a broadly increasing pattern in both returns and Sharpe ratios from the lowest to the highest decile for each model. The ordering is visible in both subperiods and in the combined sample, suggesting that the model scores map into economically relevant cross-sectional variation in short-horizon returns under the portfolio protocol. Relative to the benchmark, Hiera produces a steeper decile profile. The difference is most apparent in the upper part of the ranking, where the high-decile portfolio earns higher returns and

Sharpe ratios under Hiera than under the benchmark in each sample split.

The H-L portfolio summarizes the ranking performance in a single spread. In the combined sample (2001–2024), the benchmark yields an annualized H-L return of 0.15 with a Sharpe ratio of 1.43, whereas Hiera attains 0.24 and 2.17. The same ordering holds within each subsample: Hiera's H-L return exceeds the benchmark in 2001–2019 (0.25 versus 0.16) and in 2020–2024 (0.22 versus 0.14). Beyond the endpoints, several intermediate deciles also exhibit higher Sharpe ratios under Hiera, consistent with a generally sharper cross-sectional ordering rather than a difference driven by a single extreme portfolio. These patterns are not specific to the $L = 20$ chart construction; using shorter ($L = 5$) or longer ($L = 60$) lookback windows yields a similar decile ordering and the same Hiera–benchmark performance ranking (see Appendix A2).

3.2.2 Large-Cap Robustness

To assess sensitivity to large-cap implementability, Table 3.2 restricts the universe at each formation date to the 500 largest firms by market capitalization. This restriction reduces heterogeneity in firm size and trading activity, and it aligns the universe more closely with the subset of securities that institutional investors can trade at scale. The Top-500 set is recomputed at each rebalance date.

Table 3.2: Window 20 Performance Comparison (Top 500 by Market Capitalization)

	Benchmark (2001–2019)		Hiera (2001–2019)		Benchmark (2020–2024)		Hiera (2020–2024)		Benchmark (2001–2024)		Hiera (2001–2024)	
Decile	Ret	SR	Ret	SR	Ret	SR	Ret	SR	Ret	SR	Ret	SR
Low	0.08	0.42	0.05	0.28	0.09	0.44	0.09	0.48	0.08	0.42	0.06	0.32
2	0.08	0.43	0.06	0.30	0.08	0.37	0.09	0.44	0.08	0.42	0.06	0.34
3	0.10	0.54	0.08	0.42	0.14	0.65	0.11	0.52	0.11	0.56	0.08	0.44
4	0.10	0.57	0.09	0.51	0.10	0.47	0.10	0.47	0.10	0.55	0.09	0.50
5	0.09	0.49	0.09	0.49	0.13	0.64	0.13	0.62	0.10	0.52	0.10	0.52
6	0.10	0.55	0.12	0.64	0.14	0.70	0.13	0.61	0.11	0.59	0.12	0.63
7	0.11	0.62	0.14	0.76	0.15	0.71	0.12	0.57	0.12	0.64	0.14	0.71
8	0.11	0.63	0.12	0.64	0.14	0.66	0.19	0.87	0.12	0.63	0.13	0.69
9	0.14	0.75	0.14	0.75	0.17	0.80	0.15	0.68	0.15	0.76	0.15	0.73
High	0.14	0.75	0.19	0.94	0.18	0.79	0.22	0.93	0.15	0.76	0.19	0.94
H–L	0.06***	0.68	0.13***	1.30	0.08	0.66	0.11*	0.89	0.07***	0.67	0.13***	1.20

Note: Gross performance for decile and H-L portfolios is reported separately for 2001–2019, the post-2019 holdout (2020–2024), and the combined sample (2001–2024). At each formation date, the investment universe is restricted to the top 500 stocks by market capitalization and the set is recomputed at each rebalance date. Stocks are sorted into deciles based on model scores and equal-weighted decile portfolios are formed. Portfolios are held for a five-trading-day holding period ($H = 5$), and the next rebalance occurs on the trading day immediately after the holding period ends. Returns are annualized geometric means of the realized holding-period returns; Sharpe ratios are computed from the holding-period return series using annualization factor $\sqrt{252/H}$. Test statistics are reported for H-L portfolios and are based on the annualized arithmetic mean of holding-period returns using Newey–West standard errors (4 lags). Statistical significance at the 10%, 5%, and 1% levels is denoted by *, **, and ***, respectively.

The cross-sectional patterns observed in the broad universe remain visible under this size constraint. Returns and Sharpe ratios generally rise with the model ranking for both methods, and H-L spreads remain positive in the combined sample. Magnitudes compress relative to Table 3.1, which is consistent with lower cross-sectional dispersion among larger firms and with the reduced role of microcap extremes.

Hiera continues to produce a steeper return gradient across deciles than the benchmark, with the largest differences concentrated near the top of the ranking. In the combined sample, Hiera's H-L return is 0.13 versus 0.07 for the benchmark, with Sharpe ratios of 1.20 and 0.67, respectively. The high-decile portfolio earns 0.19 (Sharpe 0.94) under Hiera compared to 0.15 (Sharpe 0.76) under the benchmark. Thus, Hiera's performance edge persists even when the universe is restricted to the largest, most liquid stocks. Appendix A3 formalizes this comparison by reporting Hiera−benchmark return differentials for the H–L series (broad universe and Top 500).

3.2.3 Transaction Costs

A natural question is whether the return spreads documented above survive once we account for trading frictions. Table 3.3 reports combined-sample performance after deducting transaction costs based on realized portfolio turnover, assuming a cost of 5 bps per unit of turnover. Portfolios are rebalanced at the same non-overlapping $H = 5$ schedule used in the gross results, so turnover arises from cross-sectional reordering of firms between rebalances. The cost adjustment therefore provides a direct view of how sensitive the gross spreads are to an implementability friction that scales with trading intensity. Appendix A4 reports robustness to alternative cost assumptions (1, 10, and 25 bps) and summarizes average turnover by decile.

Introducing transaction costs compresses H-L spreads for both approaches, but the model ordering remains the same as in the gross results. In the broad universe, Hiera's net H-L return is 0.17 versus 0.09 for the benchmark, with Sharpe ratios of 1.56 and 0.84, respectively. Net returns remain increasing across deciles for both models, and the larger

Table 3.3: Window 20 Performance with Transaction Costs, Combined Sample (2001–2024)

	Benchmark		Hiera		Benchmark (Top 500)		Hiera (Top 500)	
Decile	Ret	SR	Ret	SR	Ret	SR	Ret	SR
Low	0.03	0.12	-0.01	-0.05	0.05	0.27	0.03	0.16
2	0.06	0.27	0.03	0.16	0.05	0.26	0.03	0.18
3	0.09	0.42	0.05	0.26	0.08	0.40	0.05	0.28
4	0.08	0.39	0.09	0.43	0.07	0.38	0.06	0.33
5	0.09	0.43	0.10	0.48	0.07	0.36	0.07	0.35
6	0.11	0.51	0.12	0.55	0.08	0.42	0.09	0.47
7	0.11	0.53	0.13	0.62	0.09	0.48	0.11	0.55
8	0.12	0.56	0.14	0.64	0.09	0.47	0.10	0.53
9	0.16	0.75	0.16	0.75	0.11	0.59	0.11	0.57
High	0.18	0.83	0.23	1.00	0.12	0.60	0.16	0.78
H–L	0.09***	0.84	0.17***	1.56	0.01	0.08	0.07***	0.62

Note: Net performance for decile and H-L portfolios is reported for the combined period 2001–2024 using a non-overlapping five-trading-day holding period ($H = 5$). At each rebalance date, transaction costs are applied as a proportional deduction from gross portfolio returns equal to $c \times \text{Turnover}_t$, where $c = 0.0005$ (5 bps per unit of turnover). Turnover is measured as firm turnover within each equal-weighted decile portfolio, $\text{Turnover}_t = |S_t \triangle S_{t-1}|/|S_t|$, where S_t denotes the set of stocks in the decile at formation date t and $\triangle$ is the symmetric difference operator. H-L turnover is defined as the sum of turnover for the high and low deciles, and costs are applied to both legs. Returns are annualized geometric means of net holding-period returns; Sharpe ratios are computed from the net holding-period return series using annualization factor $\sqrt{252/H}$. H-L test statistics are computed from the annualized arithmetic mean of net holding-period returns using Newey–West standard errors (4 lags). Statistical significance at the 10%, 5%, and 1% levels is denoted by *, **, and ***, respectively.

H-L spread under Hiera reflects a stronger separation between the high and low ranked portfolios after the turnover-based deduction.

The large-cap restriction sharpens the contrast. In the Top-500 universe, the benchmark's net H-L spread is close to zero (0.01, Sharpe 0.08), whereas Hiera retains a positive net spread (0.07, Sharpe 0.62). Relative to the gross Top-500 results in Table 3.2, both models exhibit lower net performance, reflecting the combined effect of spread compression in a large-cap universe and turnover-based deductions at each rebalance. The remaining net spread under Hiera indicates that its ranking advantage is not eliminated by this cost adjustment in a large-cap setting.

Overall, applying turnover-based costs reduces spreads but does not overturn the ranking of models. Hiera continues to deliver larger net H-L spreads than the benchmark in both the broad and Top-500 universes, indicating that the gross performance differences remain detectable after accounting for realistic implementation frictions.

3.3 Technical-Signal Benchmarks and Spanning Tests

We next benchmark the model-implied spread against a representative set of widely used technical signals—transparent, rule-based transformations of past prices and volume.[3] Because these signals are simple functions of past OHLCV, they provide a scale for the magnitude of the image-based H-L portfolios and help assess whether the learned ranking resembles canonical rule-based behavior. The goal is not to treat individual signals as full competitors, but to quantify how much of the Hiera-implied cross-sectional spread aligns with standard technical rules.

3.3.1 Signal-Sorted H-L Spreads

Table 3.4: Technical-Signal H-L Spread Portfolios

	2001–2019		2020–2024		2001–2024	
Signal	Ret	SR	Ret	SR	Ret	SR
5-Day Momentum (Contrarian)	0.09*	0.43	0.27*	0.95	0.12**	0.55
5-Day Momentum (Standard)	−0.08*	-0.40	−0.21*	-0.76	−0.11**	-0.49
EMA(12) - EMA(26) (MACD Line)	−0.07**	-0.48	0.00	-0.01	−0.06*	-0.36
MACD Histogram	−0.05*	-0.39	-0.06	-0.37	−0.05**	-0.38
SMA(5) - SMA(20)	−0.07**	-0.52	-0.03	-0.18	−0.06**	-0.42
SMA(20) - SMA(60)	-0.05	-0.36	-0.07	-0.41	−0.06*	-0.37
SMA(5) - SMA(60)	−0.08**	-0.51	-0.06	-0.32	−0.07**	-0.47
VWMA - SMA (20-day)	−0.04*	-0.42	-0.01	-0.08	−0.03*	-0.32
OBV Momentum (20-day)	0.02*	0.40	0.01	0.23	0.02*	0.37
RSI(14)†	0.14***	0.73	0.12	0.49	0.13***	0.66
Stochastic K (14)†	0.12***	0.73	0.14*	0.67	0.13***	0.71
Stochastic D (14)†	0.11***	0.61	0.10	0.47	0.11***	0.58
Stoch (K - D, 14)†	0.06**	0.53	0.15***	1.07	0.08***	0.67
Bollinger Z-Score (20)†	0.09***	0.63	0.13*	0.66	0.10***	0.63
ATR-Scaled Return (14)†	0.09***	0.67	0.15**	0.90	0.11***	0.73

Note: Each row reports the H-L spread from sorting stocks into deciles on the corresponding technical signal at each rebalance date and holding for five trading days ($H = 5$). Unless otherwise noted, H-L is Decile 10 minus Decile 1. Signals marked with † are reported in the contrarian orientation (Decile 1 minus Decile 10) so that a positive spread corresponds to the conventional contrarian interpretation. All signals are standardized cross-sectionally on each rebalance date. Signal lookbacks follow conventional parameterizations and are not tied to the image lookback $L \in \{5, 20, 60\}$. Test statistics are based on the annualized arithmetic mean of weekly returns using Newey–West standard errors (4 lags). Statistical significance at the 10%, 5%, and 1% levels is denoted by *, **, and ***, respectively. Mathematical definitions are provided in Table A5.1.

Table 3.4 provides a transparent scale for interpreting the magnitude of the image-based H-L portfolios. For reference, the equal-weighted market portfolio earns an annualized return

[3]The set is not exhaustive; we focus on common signals that appear frequently in practitioner treatments of technical analysis and are widely used as baseline chart-based rules. See, e.g., Murphy (1999).

of about 0.12 over 2001–2024 (Sharpe 0.58) under the same five-trading-day holding-period aggregation.

Many trend-oriented technical signals deliver weak or negative H-L spreads at a five-day horizon in this design. For example, standard 5-day momentum produces a negative combined-sample spread (-0.11; Sharpe -0.49), consistent with short-horizon reversal. Its contrarian orientation yields a positive spread (0.12; Sharpe 0.55), with relatively stronger performance in 2020–2024. MACD-based signals and moving-average differences similarly exhibit small or negative spreads across subsamples.

Oscillator-type contrarian signals are comparatively stronger. RSI, stochastic oscillators, Bollinger-type deviations, and ATR-scaled returns generate positive spreads in the combined sample, typically with Sharpe ratios around 0.6–0.7, and some strengthen in 2020–2024 (e.g., Stoch (K–D) with Sharpe above 1.0). Even so, their magnitudes remain below the baseline window-20 Hiera H-L portfolio reported earlier (0.24; Sharpe 2.17).

These comparisons motivate a spanning test: we ask how much of the week-to-week variation in the Hiera window-20 H-L return is explained by the joint set of technical-signal H-L portfolios.

3.3.2 Multivariate Spanning Regressions

Table 3.5 shows that common technical signals explain a limited share of the week-to-week variation in the Hiera H-L return. The R^2 is 0.22 in the combined period, similar to 0.20 in 2001–2019 and higher in 2020–2024 (0.42). Thus, even in the subsample where comovement is strongest, more than half of the variation in the Hiera spread remains unexplained by this signal set.

Across periods, most slope estimates are small and statistically indistinguishable from zero, and partial correlations are generally close to zero. A few signals load more consistently—for example, VWMA minus SMA(20) enters with a negative and statistically significant coefficient in all three samples—but the associated partial correlations remain modest in

Table 3.5: Multivariate Regressions of Hiera H-L (Window 20) on Technical Signals

	2001–2019 ($R^2 = 0.20,\ N = 942$)		2020–2024 ($R^2 = 0.42,\ N = 238$)		2001–2024 ($R^2 = 0.22,\ N = 1180$)	
Signal	Coeff (t)	Partial Corr	Coeff (t)	Partial Corr	Coeff (t)	Partial Corr
Intercept	0.0035*** (6.99)	–	0.0026*** (2.93)	–	0.0034*** (7.69)	–
5-Day Momentum (Contrarian)	4.32** (2.08)	0.07	−3.02 (−1.54)	−0.10	2.82 (1.50)	0.04
5-Day Momentum (Standard)	4.27** (2.06)	0.07	−3.15 (−1.61)	−0.11	2.75 (1.46)	0.04
EMA(12) - EMA(26) (MACD Line)	0.11 (0.40)	0.01	0.64* (1.90)	0.13	0.19 (0.82)	0.02
MACD Histogram	−0.13 (−1.33)	−0.04	−0.13 (−0.82)	−0.05	−0.07 (−0.82)	−0.02
SMA(5) - SMA(20)	−0.07 (−0.57)	−0.02	−0.34* (−1.78)	−0.12	−0.18 (−1.65)	−0.05
SMA(20) - SMA(60)	−0.18 (−1.14)	−0.04	−0.20 (−1.28)	−0.09	−0.19 (−1.54)	−0.05
SMA(5) - SMA(60)	0.19 (0.56)	0.02	−0.08 (−0.25)	−0.02	0.19 (0.68)	0.02
VWMA - SMA (20-day)	−0.26*** (−3.79)	−0.12	−0.29** (−2.34)	−0.15	−0.28*** (−4.38)	−0.13
OBV Momentum (20-day)	0.01 (0.06)	0.00	−0.10 (−0.63)	−0.04	−0.03 (−0.29)	0.01
RSI(14)†	0.14 (1.62)	0.05	−0.13 (−1.04)	−0.07	0.09 (1.15)	0.03
Stochastic K (14)†	−0.02 (−0.21)	−0.01	0.05 (0.30)	0.02	0.01 (0.15)	0.00
Stochastic D (14)†	−0.06 (−0.61)	−0.02	0.37*** (2.61)	0.17	0.03 (0.31)	0.01
Stoch (K - D, 14)†	−0.10 (−1.18)	−0.04	−0.11 (−1.00)	−0.07	−0.11 (−1.49)	0.00
Bollinger Z-Score (20)†	−0.16 (−1.48)	−0.05	−0.19 (−1.63)	−0.11	−0.19** (−2.30)	−0.07
ATR-Scaled Return (14)†	−0.04 (−0.60)	−0.02	0.06 (0.81)	0.05	−0.01 (−0.27)	−0.01

Note: Ordinary least squares regressions of the weekly Hiera H-L return (window $L = 20$) on the weekly H-L returns of the technical-signal portfolios in Table 3.4, estimated separately for 2001–2019, 2020–2024, and 2001–2024. All returns are in weekly units (not annualized). Newey–West standard errors use four lags. Statistical significance at the 10%, 5%, and 1% levels is denoted by *, **, and ***, respectively. For each regressor, the reported "Partial Corr" is the partial correlation between the Hiera H–L return and that regressor conditional on all others, computed as $r_{\text{partial}} = t/\sqrt{t^2 + \text{df}}$ with $\text{df} = N - k - 1$ (where k is the number of regressors). The two 5-day momentum variables are constructed from the same underlying signal with opposite orientation and are therefore highly correlated; their coefficients should be interpreted jointly. Signals marked with † follow the contrarian orientation defined in Table 3.4 and Table A5.1.

magnitude, implying limited incremental explanatory content conditional on the full set. Overall, the evidence suggests that the Hiera spread is not well approximated by a linear combination of these widely used technical signals. Because multivariate specifications can mask gross comovement when predictors are correlated, we also estimate univariate regressions of the Hiera spread on each signal in isolation, reported in Table 3.6.

3.3.3 Univariate Spanning Regressions

Table 3.6 reports univariate regressions that relate the weekly Hiera window-20 H-L return to the weekly H-L returns of individual technical-signal portfolios. Several signals exhibit statistically detectable comovement with Hiera. Short-horizon momentum is the clearest example: the contrarian 5-day momentum portfolio loads positively in all three samples, with correlations of 0.29 (2001–2019), 0.44 (2020–2024), and 0.34 (2001–2024). As expected, the standard 5-day momentum specification reverses sign one-for-one in both slope and correlation because it is the same underlying signal with opposite orientation.

Two patterns are central. First, even where correlations are strongest, they remain moderate: the largest correlations are around 0.3–0.4, and most are closer to 0.2–0.3 in

magnitude. This implies that any single technical signal explains only a limited share of the week-to-week variation in the Hiera H–L return, consistent with the multivariate fit in Table 3.5. Second, the intercepts are large and stable across specifications. In nearly every row, the estimated intercept is close to 0.004 per week (i.e., per five-trading-day holding period) and is strongly statistically significant. Conditioning on any single signal therefore leaves the average Hiera H-L premium largely unchanged.

Table 3.6: Univariate Regressions of Hiera H-L (Window 20) on Technical Signals

	2001–2019			2020–2024			2001–2024		
Signal	Coeff (t)	Corr	Intercept (t)	Coeff (t)	Corr	Intercept (t)	Coeff (t)	Corr	Intercept (t)
5-Day Momentum (Contrarian)	0.159*** (4.46)	0.29	0.00415*** (6.75)	0.195*** (3.54)	0.44	0.00300*** (2.99)	0.170*** (5.39)	0.34	0.00392*** (7.36)
5-Day Momentum (Standard)	−0.159*** (−4.46)	−0.29	0.00415*** (6.75)	−0.195*** (−3.55)	−0.44	0.00300*** (2.99)	−0.170*** (−5.39)	−0.34	0.00392*** (7.36)
EMA(12) - EMA(26)	−0.062 (−1.58)	−0.09	0.00432*** (6.93)	−0.088 (−0.92)	−0.14	0.00391*** (3.23)	−0.070* (−1.73)	−0.10	0.00423*** (7.65)
MACD Histogram	−0.244*** (−4.21)	−0.26	0.00419*** (6.61)	−0.262*** (−5.10)	−0.37	0.00357*** (3.11)	−0.251*** (−5.95)	−0.29	0.00406*** (7.32)
SMA(5) - SMA(20)	−0.143*** (−3.27)	−0.18	0.00422*** (6.69)	−0.186** (−2.33)	−0.28	0.00381*** (3.24)	−0.157*** (−3.81)	−0.21	0.00412*** (7.42)
SMA(20) - SMA(60)	−0.031 (−0.63)	−0.04	0.00406*** (7.28)	0.030 (0.30)	0.04	0.00402*** (3.82)	−0.016 (−0.35)	−0.02	0.00405*** (8.26)
SMA(5) - SMA(60)	−0.069 (−1.46)	−0.10	0.00398*** (7.21)	−0.002 (−0.02)	−0.00	0.00398*** (3.73)	−0.052 (−1.18)	−0.08	0.00399*** (8.17)
VWMA - SMA (20-day)	−0.271*** (−5.42)	−0.22	0.00423*** (6.85)	−0.344*** (−4.68)	−0.37	0.00386*** (3.38)	−0.298*** (−6.98)	−0.26	0.00414*** (7.64)
OBV Momentum (20-day)	0.042 (0.37)	0.02	0.00441*** (7.05)	−0.181 (−0.62)	−0.07	0.00397*** (3.37)	0.005 (0.04)	0.00	0.00432*** (7.81)
RSI(14)†	−0.115*** (−3.24)	−0.20	0.00408*** (6.80)	−0.152* (−1.92)	−0.30	0.00353*** (3.17)	−0.127*** (−3.48)	−0.23	0.00395*** (7.47)
Stochastic K (14)†	−0.167*** (−4.65)	−0.26	0.00398*** (6.49)	−0.174 (−1.69)	−0.29	0.00340*** (3.11)	−0.169*** (−4.35)	−0.27	0.00386*** (7.19)
Stochastic D (14)†	−0.129*** (−3.55)	−0.21	0.00412*** (6.68)	−0.114 (−1.10)	−0.20	0.00367*** (3.23)	−0.125*** (−3.24)	−0.21	0.00403*** (7.44)
Stoch (K - D, 14)†	−0.227*** (−4.01)	−0.23	0.00415*** (6.58)	−0.302*** (−3.35)	−0.33	0.00298*** (2.79)	−0.247*** (−5.05)	−0.25	0.00392*** (7.12)
Bollinger Z-Score (20)†	−0.227*** (−5.20)	−0.31	0.00398*** (6.47)	−0.265*** (−2.88)	−0.42	0.00321*** (2.98)	−0.238*** (−5.65)	−0.34	0.00382*** (7.10)
ATR-Scaled Return (14)†	−0.160*** (−3.52)	−0.21	0.00410*** (6.82)	−0.227*** (−2.86)	−0.31	0.00317*** (2.86)	−0.178*** (−4.49)	−0.23	0.00391*** (7.37)

Note: Each row reports an ordinary least squares regression of the weekly Hiera window-20 H-L return on the weekly H-L return of a single technical-signal portfolio, estimated separately for 2001–2019, 2020–2024, and 2001–2024. Slope coefficients and intercepts are reported with t-statistics in parentheses; "Corr" is the unconditional correlation between the two weekly return series. Newey–West standard errors use four lags. Statistical significance at the 10%, 5%, and 1% levels is denoted by *, **, and ***, respectively. Signals marked with † follow the contrarian orientation defined in Table 3.4 and Table A5.1.

Overall, the univariate regressions show that the Hiera H-L return has nontrivial comovement with familiar contrarian and oscillator-style signals, but that comovement is not strong enough for any single rule to account for the level of the spread: the intercept remains economically large and precisely estimated across rows. This complements the multivariate evidence in Table 3.5—common signals jointly explain only a limited share of the week-to-week variation—while also clarifying what the spanning regressions cannot tell us. Because these tests operate on portfolio-level time series, they diagnose alignment with known signal families but do not identify which parts of the chart input drive the residual component. The next section therefore turns to interpretability diagnostics that localize the predictive content within the lookback window.

3.4 Interpretability Diagnostics

We discipline interpretation of the chart-image signal using a complementary set of diagnostics. Spanning regressions quantify overlap with standard technical signals but do not control the visual content of the input images. We therefore develop two direct, model-facing exercises. First, controlled synthetic probes vary the strength of stylized chart concepts and measure how predicted up-move probabilities respond. Second, occlusion sensitivity maps provide a model-agnostic "where" statement on real-market images by identifying which regions of the chart drive forecast changes under localized perturbations. Together, these tools narrow the set of interpretations that remain consistent with the performance evidence, without claiming causal identification.

3.4.1 Synthetic Probe Analysis

Synthetic probes provide controlled variation in the visual content of the input charts, offering a direct diagnostic of the image channel that spanning regressions cannot provide. The spanning regressions quantify overlap between the Hiera signal and standard technical signal families on realized market paths, but they cannot isolate how the model responds when a single chart configuration is strengthened in isolation. Synthetic probes address

this limitation by generating charts from explicit stochastic designs in which one stylized concept is the intended driver, indexed by a scalar strength parameter $s \in [0, 1]$, and by averaging predictions across many stochastic realizations at each strength level. Mathematical definitions (including stochastic perturbations) and example images are provided in Table A5.1 and Figure A7.1, respectively.

We focus on a compact set of canonical concepts used by practitioners—breakouts, head-and-shoulders formations, momentum legs, mean-reverting ranges, and volatility clusters—chosen to span distinct classes of visual structure. For each concept and strength level, we generate 5,000 synthetic images, evaluate them with the trained Hiera model, and record the mean predicted probability of an up move, Mean($p_\uparrow$). We also report ΔMean, which measures the deviation of Mean($p_\uparrow$) from a common reference level $\bar{p}_\uparrow$ computed from a pooled baseline set of probes. This common-reference normalization is useful because some probe generators can differ in level even at low strength, while the strength–response within a concept is captured by how Mean($p_\uparrow$) changes as s increases. We report 95% bootstrap confidence intervals based on 1,000 resamples.

Table 3.7 can be read as a set of strength–response curves. Within each concept, the object of interest is how Mean($p_\uparrow$) changes with strength, in either direction. We use these within-concept slopes to assess whether the model responds systematically as a configuration becomes more salient, and whether that response is symmetric when the direction of a trend leg is flipped.

Three patterns summarize the main responses. First, the momentum legs exhibit a pronounced sign asymmetry. As down-momentum strengthens, Mean($p_\uparrow$) increases monotonically from 0.516 to 0.535. The corresponding up-momentum probe shows little change at low strength and declines at high strength (from 0.516 at Strength = 0.00 to 0.507 at Strength = 1.00). This asymmetry emerges under controlled inputs and therefore constrains interpretations in which momentum enters as a direction-symmetric driver of the model's signal.

Second, the head-and-shoulders constructions generate large level differences and sharply

Table 3.7: Core Per-Cell Metrics (Mean Predictions and Deltas, Bootstrapped Confidence Intervals), Window 20

Concept	Strength	Mean($p_\uparrow$) ± CI	Δ Mean ± CI	n
Breakout	0.00	0.519 ± 0.001	0.007 ± 0.001	5000
	0.25	0.522 ± 0.001	0.009 ± 0.001	5000
	0.50	0.522 ± 0.001	0.010 ± 0.001	5000
	0.75	0.522 ± 0.001	0.009 ± 0.001	5000
	1.00	0.522 ± 0.001	0.009 ± 0.001	5000
Down Momentum	0.00	0.516 ± 0.001	0.004 ± 0.001	5000
	0.25	0.521 ± 0.001	0.009 ± 0.001	5000
	0.50	0.526 ± 0.001	0.013 ± 0.001	5000
	0.75	0.530 ± 0.001	0.018 ± 0.001	5000
	1.00	0.535 ± 0.001	0.022 ± 0.001	5000
Head & Shoulders	0.00	0.481 ± 0.001	-0.032 ± 0.001	5000
	0.25	0.485 ± 0.001	-0.028 ± 0.001	5000
	0.50	0.492 ± 0.001	-0.021 ± 0.001	5000
	0.75	0.497 ± 0.001	-0.015 ± 0.001	5000
	1.00	0.500 ± 0.001	-0.013 ± 0.001	5000
Inverse H&S	0.00	0.481 ± 0.001	-0.032 ± 0.001	5000
	0.25	0.507 ± 0.001	-0.005 ± 0.001	5000
	0.50	0.523 ± 0.001	0.011 ± 0.001	5000
	0.75	0.530 ± 0.000	0.017 ± 0.001	5000
	1.00	0.534 ± 0.001	0.021 ± 0.000	5000
Mean Reversion	0.00	0.522 ± 0.001	0.009 ± 0.001	5000
	0.25	0.516 ± 0.001	0.004 ± 0.001	5000
	0.50	0.508 ± 0.001	-0.005 ± 0.001	5000
	0.75	0.496 ± 0.001	-0.017 ± 0.001	5000
	1.00	0.483 ± 0.001	-0.030 ± 0.001	5000
Up Momentum	0.00	0.516 ± 0.001	0.004 ± 0.001	5000
	0.25	0.518 ± 0.001	0.005 ± 0.001	5000
	0.50	0.517 ± 0.001	0.004 ± 0.001	5000
	0.75	0.513 ± 0.001	0.001 ± 0.001	5000
	1.00	0.507 ± 0.001	-0.005 ± 0.001	5000
Volatility Cluster	0.00	0.521 ± 0.001	0.008 ± 0.001	5000
	0.25	0.520 ± 0.001	0.007 ± 0.001	5000
	0.50	0.518 ± 0.001	0.005 ± 0.001	5000
	0.75	0.514 ± 0.001	0.001 ± 0.001	5000
	1.00	0.507 ± 0.001	-0.005 ± 0.001	5000

Note: Each cell is based on $n = 5000$ synthetic probe images for a given concept and strength level. Mean($p_\uparrow$) reports the average predicted probability of an up move. Δ Mean reports the deviation of Mean($p_\uparrow$) from a common reference level $\bar{p}_\uparrow$ computed from a pooled baseline probe set. Confidence intervals are 95% bootstrap intervals based on 1,000 resamples.

different strength responses. The inverse head-and-shoulders response increases strongly with strength (Mean($p_\uparrow$) rises from 0.481 to 0.534), whereas head-and-shoulders remains below 0.5 even at high strength (0.500 at Strength = 1.00). In this diagnostic, the model responds differently to the trough-based and peak-based variants of the same stylized structure, and the distinction is visible in the strength response rather than in a single illustrative chart.

Third, range and regime structure move predicted probabilities in the opposite direction. As mean-reversion strength increases, Mean($p_\uparrow$) declines from 0.522 to 0.483, and volatility clustering also declines with strength (from 0.521 to 0.507). Breakouts produce modest, largely flat changes across strength. Taken together, the probes indicate selectivity: some controlled configurations induce strong and monotone shifts in predicted probabilities, while others induce weak or non-monotone responses.

These controlled sensitivities discipline interpretation of the chart-image signal by isolating which configurations move the model's predictions in a one-factor environment. The next section returns to real-market charts and asks where, within the lookback window, the model draws that predictive content using occlusion sensitivity maps.

3.4.2 Occlusion Sensitivity Maps

Synthetic probes isolate *what* stylized configurations move predictions under controlled inputs. We next ask *where*, within real-market chart images, the model's decision statistic is most sensitive to perturbations. We develop a series of occlusion sensitivity maps by masking a localized region of the candlestick chart, re-evaluating the model, and recording the change in the Up–Down logit margin. Regions whose occlusion lowers the margin contribute positive evidence for an Up forecast; regions whose occlusion raises the margin contribute evidence in the opposite direction. We implement this diagnostic at multiple grid resolutions to reflect the model's multi-scale representation. Appendix A8 details the occlusion operator, the margin construction, and visualization choices; Appendix A8.1 defines the aggregate metrics reported below.

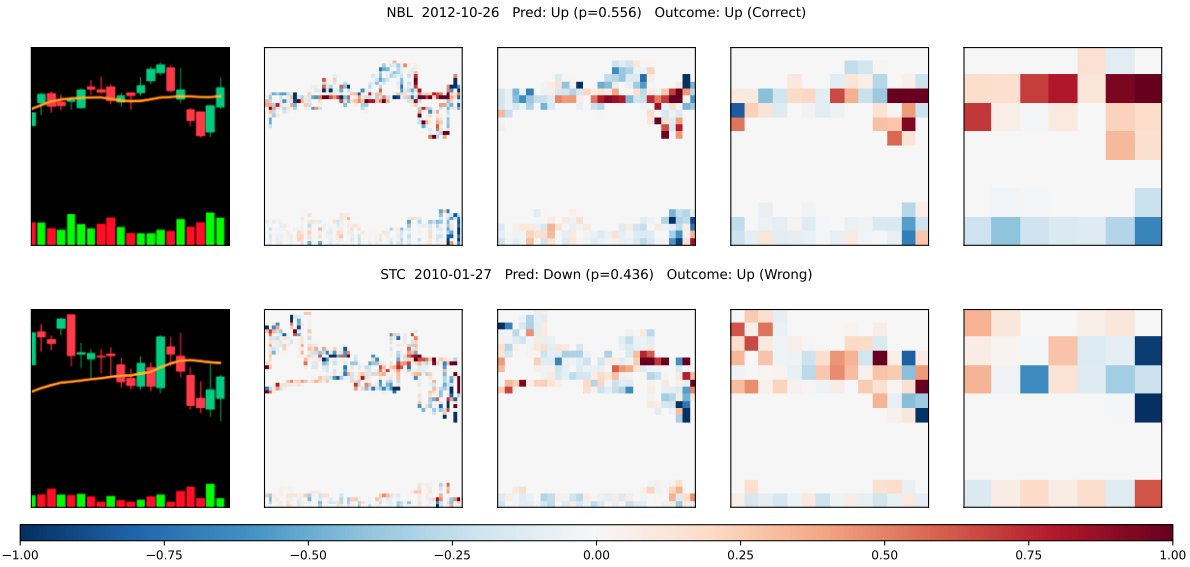


Figure 3.2: Occlusion sensitivity maps for two example forecasts. Each row shows one stock–date observation (20-day lookback window), Hiera's 5-day directional prediction, the predicted probability p_{Up}, and the realized outcome. The first column is the input chart; the remaining columns report signed occlusion effects on the Up–Down margin at grid resolutions 56×56, 28×28, 14×14, and 7×7. Red (blue) indicates that occluding the region lowers (raises) the margin.

To summarize occlusion patterns beyond individual examples, we compute maps for a balanced sample of 1,000 high-confidence out-of-sample forecasts drawn from the two tails of the predictive distribution (Table 3.8 reports the eligibility thresholds and counts). For each resolution $G \in \{56, 28, 14, 7\}$, we aggregate the absolute occlusion evidence by horizontal position within the 20-day window to obtain an evidence profile over time.

Table 3.8: Aggregate Diagnostics from Occlusion Sensitivity Maps

	Sample	Evidence Location Along Window						Evidence Concentration		
Grid G	n	$S_{0.10}$	$S_{0.20}$	$S_{0.33}$	TDR	$\mathbb{E}[t]$	t_{50}	t_{80}	Top 5%	Top 10%
56	1000	0.246	0.368	0.468	0.603	0.567	0.618	0.919	0.289	0.428
28	1000	0.225	0.365	0.495	0.598	0.575	0.640	0.915	0.273	0.408
14	1000	0.205	0.266	0.383	0.377	0.475	0.456	0.868	0.313	0.408
7	1000	0.273	0.409	0.507	0.748	0.552	0.604	0.953	0.297	0.398

Note: The diagnostic sample is drawn from high-confidence forecasts satisfying $p_{\text{Up}}(x) \geq 0.55$ or $p_{\text{Up}}(x) \leq 0.45$, which comprise 239,174 forecasts (15.6%) out of 1,531,408 total forecasts. A balanced subsample is formed from the two tails. For each grid resolution G, n denotes the number of successfully processed images. S_α denotes the share of absolute occlusion evidence located in the last α fraction of the window. TDR is the terminal dominance ratio $S_{0.20}/(1 - S_{0.20})$. $\mathbb{E}[t]$ is the centroid of the evidence profile, while t_{50} and t_{80} denote the relative positions at which cumulative evidence reaches 50% and 80%, respectively. Top 5% and Top 10% report the share of absolute evidence contained in the most influential cells ranked by $|e^{(G)}_{i,j}(x)|$. For more information see Appendix A8.1.

Figure 3.3 and Table 3.8 summarize where sensitivity concentrates within the lookback

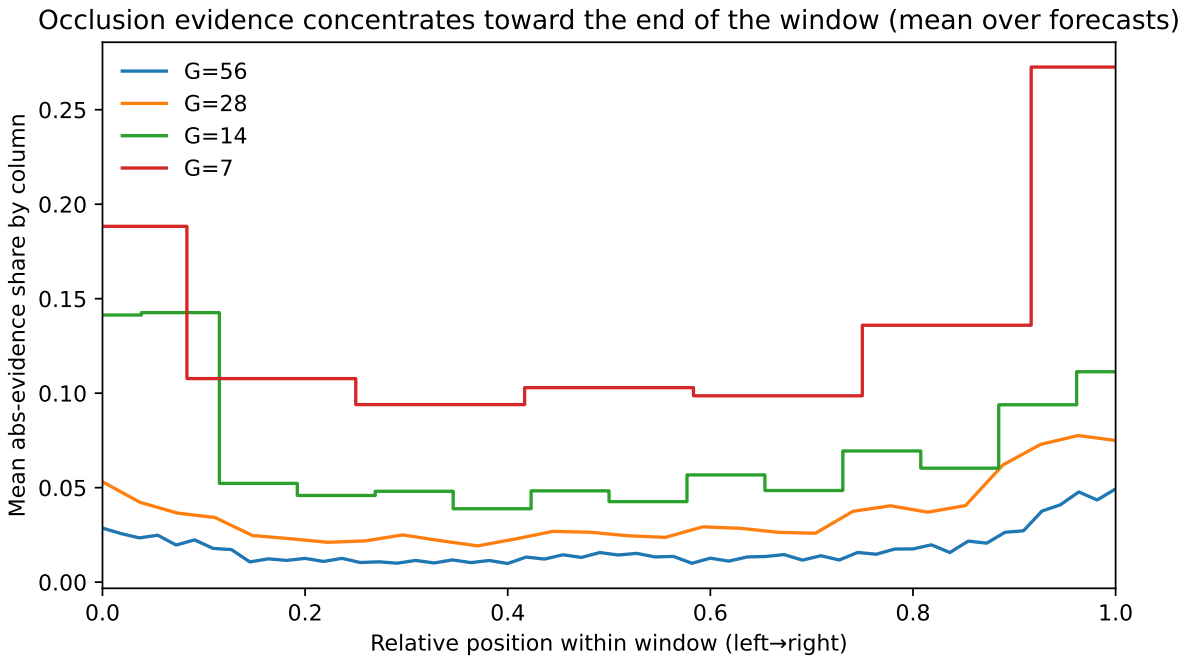


Figure 3.3: Evidence profiles from occlusion maps. The figure plots the mean share of total absolute occlusion evidence attributable to each relative horizontal position in the 20-day lookback window (mean over 1,000 forecasts), for each grid resolution $G \in \{56, 28, 14, 7\}$.

window. Across most resolutions, a substantial share of absolute evidence lies near the end of the 20-day window. For example, at the finer grids ($G = 56$ and $G = 28$), the last fifth of the window accounts for about 37% of total absolute evidence and the last third accounts for about 47–50%. The median evidence location t_{50} is correspondingly late in the window. These patterns provide a disciplined localization result that complements the probe results: the model's decision statistic is most sensitive to perturbations in the most recent portion of the input chart.

The profiles vary with spatial support. Fine grids isolate sensitivity to tightly localized regions, while coarser grids distribute evidence over broader areas of the chart. The 14×14 resolution exhibits weaker terminal concentration than the other grids, which is consistent with scale-dependent aggregation rather than reliance on a single fixed patch size. We interpret the multi-resolution evidence as descriptive: it characterizes how sensitivity shifts with the perturbation scale, without assigning a specific named chart rule.

Occlusion evidence is also spatially concentrated. Across resolutions, a small fraction of

cells accounts for a disproportionate share of total absolute evidence. This concentration strengthens the interpretability takeaway relative to spanning regressions: the latter quantify overlap with known signals on realized market paths, but they do not localize which parts of the chart the model is sensitive to. Occlusion maps provide that localization directly on the model input under an explicit perturbation operator, complementing the controlled probe evidence.

Taken together, the interpretability diagnostics deliver two complementary constraints on the chart-image signal. The synthetic probes identify which stylized configurations shift predicted probabilities in isolation and which do not, providing a controlled benchmark for interpreting model responses. The occlusion maps then localize sensitivity on real-market charts and show that, across perturbation scales, the decision statistic is primarily driven by information concentrated near the end of the lookback window and by a sparse set of influential regions. In combination, these exercises narrow the set of narratives consistent with the performance evidence: the model does not behave like a generic replication of a broad library of chart rules, and its forecasts are most sensitive to localized end-of-window structure on the input chart. The next section concludes by summarizing the empirical findings and their implications under the paper's implementability and robustness design.

4 Conclusion

Recent work shows that market data encoded as candlestick-chart images can be informative for return prediction, but two concerns remain central for empirical asset pricing: durability in recent market regimes and clarity about what information image models use. Building on this agenda (e.g., Jiang et al. (2023)) and motivated by weaker post-2019 performance in replication settings (e.g., Kaczmarek and Pukthuanthong (2024)), we reassess chart-image predictability in U.S. equities using CRSP data through 2024.

Under a shared task definition and portfolio-evaluation protocol, a hierarchical vision transformer (Hiera) delivers a stronger and more stable cross-sectional return ordering than a CNN benchmark out of sample. The advantage remains visible in post-2019 data and under implementability constraints, although spreads attenuate in large-cap universes and after transaction costs. These patterns emphasize both progress and limits: architecture improvements strengthen stability under the same evaluation discipline, but robustness depends on universe and frictions.

We complement performance evidence by developing diagnostics that sharpen interpretation. Spanning regressions show limited explanatory power of standard technical indicators and stable intercepts, indicating that the image signal is not well captured by linear combinations of familiar rules. Controlled synthetic probes isolate how predictions change when a single stylized configuration is strengthened under controlled inputs, and occlusion sensitivity maps localize where the model is most sensitive to perturbations on real-market charts. Together, the diagnostics constrain plausible readings of the learned signal: responses are selective across concepts, and sensitivity on real inputs is concentrated near the end of the lookback window and spatially sparse across perturbation scales.

Overall, the evidence points to two takeaways: modern hierarchical vision models deliver a more stable chart-image return ordering under a shared protocol, and a spanning–probes–occlusion workflow can bound what such models rely on.

References

Bajgrowicz, P. and Scaillet, O. (2012). Technical Trading Revisited: False Discoveries, Persistence Tests, and Transaction Costs. *Journal of Financial Economics*, 106(3):473–491.

Bang, J. and Ryu, D. (2023). CNN-Based Stock Price Forecasting by Stock Chart Images. *Romanian Journal of Economic Forecasting*, 26(3):120–128.

Bessembinder, H. and Chan, K. (1998). Market Efficiency and the Returns to Technical Analysis. *Financial Management*, pages 5–17.

Bianchi, D., Büchner, M., and Tamoni, A. (2021). Bond Risk Premiums with Machine Learning. *Review of Financial Studies*, 34(2):1046–1089.

Brock, W., Lakonishok, J., and LeBaron, B. (1992). Simple Technical Trading Rules and the Stochastic Properties of Stock Returns. *Journal of Finance*, 47(5):1731–1764.

Carhart, M. M. (1997). On Persistence in Mutual Fund Performance. *Journal of Finance*, 52(1):57–82.

Chen, L., Pelger, M., and Zhu, J. (2024). Deep Learning in Asset Pricing. *Management Science*, 70(2):714–750.

Chinco, A., Clark-Joseph, A. D., and Ye, M. (2019). Sparse Signals in the Cross-Section of Returns. *Journal of Finance*, 74(1):449–492.

De Bondt, W. F. and Thaler, R. (1985). Does the Stock Market Overreact? *Journal of Finance*, 40(3):793–805.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv preprint arXiv:2010.11929*.

Feng, G., Giglio, S., and Xiu, D. (2020). Taming the Factor Zoo: A Test of New Factors. *Journal of Finance*, 75(3):1327–1370.

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., and Walther, A. (2022). Predictably Unequal? The Effects of Machine Learning on Credit Markets. *Journal of Finance*, 77(1):5–47.

Gezici, A. H. B. and Sefer, E. (2024). Deep Transformer-Based Asset Price and Direction Prediction. *IEEE Access*, 12:24164–24178.

Gu, S., Kelly, B., and Xiu, D. (2020). Empirical Asset Pricing via Machine Learning. *Review of Financial Studies*, 33(5):2223–2273.

Guijarro-Ordonez, J., Pelger, M., and Zanotti, G. (2025). Deep Learning Statistical Arbitrage. *Management Science*.

Han, Y., Yang, K., and Zhou, G. (2013). A New Anomaly: The Cross-Sectional Profitability of Technical Analysis. *Journal of Financial and Quantitative Analysis*, 48(5):1433–1461.

Hao, X., Wang, S., Wang, Y., and Wu, L. (2024). Is Information Risk Priced? New Evidence from Outer Space. *Management Science*.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. (2022). Masked Autoencoders Are Scalable Vision Learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009.

Jegadeesh, N. (1990). Evidence of Predictable Behavior of Security Returns. *Journal of Finance*, 45(3):881–898.

Jegadeesh, N. and Titman, S. (1993). Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency. *Journal of Finance*, 48(1):65–91.

Jiang, J., Kelly, B., and Xiu, D. (2023). (Re-)Imag(in)ing Price Trends. *Journal of Finance*, 78(6):3193–3249.

Kaczmarek, J. and Pukthuanthong, K. (2024). Eroding and Non-Scalable CNN Signals: A Replication and Extension. SSRN working paper.

Kelly, B. T., Kuznetsov, B., Malamud, S., and Xu, T. A. (2025). Artificial Intelligence Asset Pricing Models. Technical report, National Bureau of Economic Research.

Kelly, B. T., Pruitt, S., and Su, Y. (2019). Characteristics Are Covariances: A Unified Model of Risk and Return. *Journal of Financial Economics*, 134(3):501–524.

Kozak, S., Nagel, S., and Santosh, S. (2018). Interpreting Factor Models. *Journal of Finance*, 73(3):1183–1223.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021). Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022.

Lo, A. W., Mamaysky, H., and Wang, J. (2000). Foundations of Technical Analysis: Computational Algorithms, Statistical Inference, and Empirical Implementation. *Journal of Finance*, 55(4):1705–1765.

Loshchilov, I. and Hutter, F. (2017). SGDR: Stochastic Gradient Descent with Warm Restarts.

Loshchilov, I. and Hutter, F. (2019). Decoupled Weight Decay Regularization.

McLean, R. D. and Pontiff, J. (2016). Does Academic Research Destroy Stock Return Predictability? *Journal of Finance*, 71(1):5–32.

Murphy, J. J. (1999). *Technical Analysis of the Financial Markets: A Comprehensive Guide to Trading Methods and Applications*. Penguin.

Murray, S., Xia, Y., and Xiao, H. (2024). Charting by Machines. *Journal of Financial Economics*, 153:103791.

Neely, C. J., Rapach, D. E., Tu, J., and Zhou, G. (2014). Forecasting the Equity Risk Premium: The Role of Technical Indicators. *Management Science*, 60(7):1772–1791.

Obaid, K. and Pukthuanthong, K. (2022). A Picture Is Worth a Thousand Words: Measuring Investor Sentiment by Combining Machine Learning and Photos from News. *Journal of Financial Economics*, 144(1):273–297.

Pascanu, R., Mikolov, T., and Bengio, Y. (2013). On the Difficulty of Training Recurrent Neural Networks.

Ren, X., Jiang, W., Ji, Q., and Zhai, P. (2024). Seeing Is Believing: Forecasting Crude Oil Price Trend from the Perspective of Images. *Journal of Forecasting*, 43(7):2809–2821.

Rouwenhorst, K. G. (1998). International Momentum Strategies. *Journal of Finance*, 53(1):267–284.

Ryali, C., Hu, Y.-T., Bolya, D., Wei, C., Fan, H., Huang, P.-Y., Aggarwal, V., Chowdhury, A., Poursaeed, O., Hoffman, J., et al. (2023). Hiera: A Hierarchical Vision Transformer Without the Bells-and-Whistles. In *International Conference on Machine Learning*, pages 29441–29454. PMLR.

Sullivan, R., Timmermann, A., and White, H. (1999). Data-Snooping, Technical Trading Rule Performance, and the Bootstrap. *Journal of Finance*, 54(5):1647–1691.

Tarvainen, A. and Valpola, H. (2017). Mean Teachers Are Better Role Models: Weight-Averaged Consistency Targets Improve Semi-Supervised Deep Learning Results. *Advances in Neural Information Processing Systems*, 30.

Ye, W., Yang, J., and Chen, P. (2024). Short-Term Stock Price Trend Prediction with Imaging High Frequency Limit Order Book Data. *International Journal of Forecasting*, 40(3):1189–1205.

Appendix

A1 Additional Optimization and Training Strategies

This appendix records the optimization and scheduling details referenced in the main text. We fine-tune an ImageNet-pretrained Hiera model (timm `hiera_tiny_224`, `pretrained=True`) using supervised classification with Mean Teacher consistency regularization. Training uses AdamW, a warmup+cosine learning-rate schedule, gradient clipping, and ramp-up schedules for the consistency weight and the teacher EMA decay. Unless stated otherwise, optimizer constants not listed here use PyTorch defaults.

We use a warmup schedule followed by cosine annealing. Let t denote the epoch index, and let T_{warmup} and $T_{\max}$ denote the warmup duration and total number of fine-tuning epochs, respectively. The learning-rate multiplier is

$$s(t) = \begin{cases} \frac{t+1}{T_{\text{warmup}}}, & \text{if } t < T_{\text{warmup}}, \\ \frac{1}{2}\left(1 + \cos\left(\pi \frac{t - T_{\text{warmup}}}{T_{\max} - T_{\text{warmup}}}\right)\right), & \text{if } t \geq T_{\text{warmup}}. \end{cases} \tag{A1.1}$$

For each optimizer parameter group g with base learning rate $\ell_{g,\text{base}}$, the effective learning rate is

$$\ell_g(t) = \eta_{\min} + (\ell_{g,\text{base}} - \eta_{\min})\ s(t), \tag{A1.2}$$

where $\eta_{\min}$ is the minimum learning rate.

Optimization uses AdamW with decoupled weight decay. Unless stated otherwise, AdamW uses PyTorch defaults $(\beta_1, \beta_2) = (0.9, 0.999)$ and $\epsilon = 10^{-8}$; weight decay is reported in Appendix Table A1.1.

We apply global norm gradient clipping with threshold $g_{\max} = 5.0$ (PyTorch `clip_grad_norm_`). If the gradient norm exceeds $g_{\max}$, gradients are rescaled so the

post-clipping norm equals $g_{\max}$:

$$\tilde{g} = g \cdot \min\left(1, \frac{g_{\max}}{\|g\|}\right). \tag{A1.3}$$

The teacher parameters θ'_t are updated as an exponential moving average (EMA) of the student parameters θ_t,

$$\theta'_t = \alpha_t\, \theta'_{t-1} + (1 - \alpha_t)\, \theta_t, \tag{A1.4}$$

where $\alpha_t \in (0, 1)$ is the EMA decay coefficient. We ramp α_t from α_{initial} to α_{final} over the first T_{ramp} epochs using

$$r(t) = \exp\left(-2\left(1 - \frac{t}{T_{\text{ramp}}}\right)^2\right), \qquad t < T_{\text{ramp}}, \tag{A1.5}$$

and set $r(t) = 1$ for $t \geq T_{\text{ramp}}$. To avoid near-zero consistency early in training, we cap the ramp-up from below at $r(t) \geq 0.1$ for $t < T_{\text{ramp}}$. The EMA decay is then

$$\alpha_t = \alpha_{\text{initial}} + (\alpha_{\text{final}} - \alpha_{\text{initial}})\; r(t). \tag{A1.6}$$

We use the same ramp-up factor to scale the maximum consistency weight $\lambda_{\max}$:

$$\lambda_t = \lambda_{\max}\, r(t). \tag{A1.7}$$

The supervised loss is cross-entropy on the Up/Down label. The consistency term is mean squared error between student and teacher *softmax* probabilities computed on the same input image:

$$J(\theta) = \mathbb{E}\Big[\|\text{softmax}(f(x;\theta)) - \text{softmax}(f(x;\theta'))\|^2\Big].$$

Stochasticity required for consistency training is induced by network dropout (and other training-mode randomness); we do not apply time-inverting or other label-altering image transformations.

We fine-tune the classification head and the last four transformer blocks, keeping earlier layers frozen. The optimizer uses separate parameter groups for the head and each of the last four blocks, allowing distinct base learning rates per group; the warmup+cosine schedule in (A1.2) is applied to each group.

We monitor validation cross-entropy loss and apply early stopping with patience reported in Appendix Table A1.1. The best-performing teacher checkpoint (lowest validation loss) is retained and used as the final model; we copy these weights to the student for downstream evaluation.

A1.1 Model Hyperparameters

Table A1.1: Hyperparameters for Student–Teacher Fine-Tuning (Hiera).

Common Parameters (all variants)	
Initialization	ImageNet-pretrained weights (`timm hiera_tiny_224`, `pretrained=True`)
Fine-tuned layers	Classification head + last 4 transformer blocks (earlier layers frozen)
Batch Size	256
Loss (supervised)	Cross-entropy on Up/Down label
Loss (consistency)	MSE between student/teacher softmax probabilities
$\lambda_{\max}$ (max consistency weight)	0.5
EMA decay α	initial 0.99, final 0.999
Ramp-up length T_{ramp}	5 epochs; ramp factor capped at $r(t) \geq 0.1$ for $t < T_{\text{ramp}}$
Learning-rate schedule	Warmup + cosine annealing; applied to each parameter group
$\eta_{\min}$ (cosine floor)	1×10^{-8}
Optimizer	AdamW (PyTorch defaults $(\beta_1, \beta_2) = (0.9, 0.999)$, $\epsilon = 10^{-8}$)
Gradient clipping	Global norm clipping, $g_{\max} = 5.0$
Dropout rate (head)	0.10
Fine-tuning epochs ($T_{\max}$)	10
Early stopping	Validation cross-entropy loss; patience = 2 epochs
Variant-Specific Parameters	
1. Window 5	
Base learning rates (per group)	Head: 1×10^{-3} Block -1: 1×10^{-4} Block -2: 1×10^{-4} Block -3: 1×10^{-4} Block -4: 1×10^{-4}
Weight decay	5×10^{-3}
Warmup epochs	3
2. Window 20	
Base learning rates (per group)	Head: 1×10^{-4} Block -1: 1×10^{-4} Block -2: 1×10^{-4} Block -3: 1×10^{-5} Block -4: 1×10^{-5}
Weight decay	5×10^{-3}
Warmup epochs	3
3. Window 60	
Base learning rates (per group)	Head: 5×10^{-4} Block -1: 3×10^{-4} Block -2: 2×10^{-4} Block -3: 1×10^{-4} Block -4: 5×10^{-5}
Weight decay	5×10^{-4}
Warmup epochs	1

A1.2 Chart Rendering and Preprocessing Details

Table A1.2: Chart-image construction, rendering, and model-specific preprocessing. The table reports fixed choices used throughout unless stated otherwise.

Component	Specification
Underlying inputs	Daily OHLCV from CRSP (v2). Charts are constructed from the most recent $L \in \{5, 20, 60\}$ trading days ending at date t.
Label definition	Binary "Up/Down" label based on the sign of the close-to-close horizon return: $y_{i,t} = 1$ if $C_{i,t+5} > C_{i,t}$ and $y_{i,t} = 0$ otherwise, where t is the last day in the lookback window.
Window update frequency	Daily rolling windows (step size = 1 trading day) within each ticker, subject to having at least $L + 5$ observations to form both the window and the horizon label.
Moving average overlay	Simple moving average with period equal to the window length L; when needed, the rendered window is preceded by up to $L - 1$ additional days so the moving average is fully defined over the displayed window. Only the last L days are shown in the final chart.
Chart type and panels	Candlestick chart with a volume panel; an SMA line is overlaid on the price panel.
Rendering library	`mplfinance` (candlestick renderer) with a fixed style and axes removed to standardize appearance across firms and time.
Visual standardization	Axes and ticks removed (`axisoff`); fixed chart style (e.g., `charles`); fixed background color (black); fixed figure aspect (square) before resizing.
Bar-width configuration	Fixed candlestick/volume widths and line widths to reduce variation due to rendering defaults.
Base image resolution	Rendered charts are saved and then resized to a canonical 224×224 RGB image.
Model-specific preprocessing	
Hiera inputs	Uses 224×224 RGB inputs. Images are normalized using ImageNet channel-wise mean and standard deviation, consistent with the pretrained initialization used for Hiera.
CNN benchmark inputs	Uses window-specific input geometries following Jiang et al. (2023): $(32, 15)$ for $L = 5$, $(64, 60)$ for $L = 20$, and $(96, 180)$ for $L = 60$. (Resizing is applied after rendering and does not use forward-looking information.)
CNN initialization	CNN baseline weights are initialized using Xavier/Glorot initialization for convolutional and linear layers.
Data hygiene and leakage prevention	
Missing-data screen	Tickers with at least one missing value in OHLCV anywhere in the sample are removed prior to image generation.
Train/validation split unit	Development-period split is performed at the ticker (firm) level: all images for a ticker belong exclusively to either train or validation.
Chronological holdout	The out-of-sample periods (2001–2019 and 2020–2024) are never used for model selection.

A2 Lookback-Window Robustness

This appendix section reports supplementary evidence on how portfolio performance varies with the lookback window length used to construct the chart images. The main specification in Table 3.1 uses $L = 20$, which summarizes roughly one trading month of recent information. To assess sensitivity to this choice, we repeat the same evaluation with a shorter lookback window ($L = 5$) and a longer lookback window ($L = 60$), holding the forecast horizon fixed at $H = 5$. Throughout, the formation, holding-period, and performance-measurement conventions follow the main analysis.

Table A2.1: Window 5 Performance Comparison

	Benchmark (2001–2019)		Hiera (2001–2019)		Benchmark (2020–2024)		Hiera (2020–2024)		Benchmark (2001–2024)		Hiera (2001–2024)	
Decile	Ret	SR	Ret	SR	Ret	SR	Ret	SR	Ret	SR	Ret	SR
Low	0.08	0.41	-0.00	-0.02	0.04	0.20	0.02	0.10	0.07	0.36	0.00	0.00
2	0.10	0.50	0.09	0.43	0.08	0.35	0.09	0.38	0.09	0.46	0.09	0.42
3	0.10	0.52	0.11	0.54	0.10	0.44	0.10	0.42	0.10	0.50	0.11	0.51
4	0.10	0.52	0.13	0.62	0.10	0.45	0.13	0.55	0.10	0.50	0.13	0.60
5	0.12	0.58	0.13	0.62	0.12	0.49	0.12	0.49	0.12	0.56	0.13	0.59
6	0.14	0.67	0.14	0.68	0.15	0.65	0.16	0.68	0.14	0.67	0.15	0.68
7	0.13	0.63	0.16	0.75	0.16	0.64	0.16	0.64	0.14	0.63	0.16	0.72
8	0.14	0.68	0.16	0.77	0.17	0.65	0.16	0.64	0.15	0.67	0.16	0.74
9	0.18	0.83	0.19	0.91	0.19	0.71	0.19	0.75	0.18	0.80	0.19	0.87
High	0.23	0.96	0.22	0.99	0.32	1.10	0.30	1.09	0.25	0.99	0.24	1.01
H–L	0.14***	1.23	0.23***	2.00	0.27***	1.87	0.27***	1.90	0.17***	1.37	0.24***	1.96

Note: Gross performance for decile and H-L portfolios is reported separately for 2001–2019, the post-2019 holdout (2020–2024), and the combined sample (2001–2024), using a non-overlapping five-trading-day holding period ($H = 5$). Chart images used to generate model predictions are constructed from the most recent $L = 5$ trading days. At each formation date, stocks are sorted into deciles based on model scores and equal-weighted decile portfolios are formed. Returns are annualized geometric means of realized holding-period returns; Sharpe ratios are computed from the holding-period return series using annualization factor $\sqrt{252/H}$. H-L test statistics are based on the annualized arithmetic mean of holding-period returns using Newey–West standard errors (4 lags). Statistical significance at the 10%, 5%, and 1% levels is denoted by *, **, and ***, respectively.

Table A2.1 reports results for the shorter lookback window ($L = 5$). The decile sorts remain economically ordered in the combined out-of-sample period, and both models continue to produce positive H–L spreads. Relative to the $L = 20$ baseline in Table 3.1, the combined-sample H–L spreads are similar in magnitude (benchmark: 0.17 at $L = 5$ versus 0.15 at $L = 20$; Hiera: 0.24 at both $L = 5$ and $L = 20$), while Sharpe ratios are modestly lower at $L = 5$ for both models.

The subsample patterns differ across periods. In 2001–2019, H–L spreads and Sharpe ratios are slightly lower at $L = 5$ than at $L = 20$ for both models. In the post-2019 holdout (2020–2024), H–L spreads are larger at $L = 5$ (0.27 for both models) than at $L = 20$ (0.14 for the benchmark and 0.22 for Hiera). We next consider whether these

patterns persist when the input aggregates information over a longer lookback window.

Table A2.2: Window 60 Performance Comparison

	Benchmark (2001–2019)		Hiera (2001–2019)		Benchmark (2020–2024)		Hiera (2020–2024)		Benchmark (2001–2024)		Hiera (2001–2024)	
Decile	Ret	SR	Ret	SR	Ret	SR	Ret	SR	Ret	SR	Ret	SR
Low	0.09	0.40	0.03	0.15	0.19	0.81	0.13	0.48	0.11	0.49	0.05	0.22
2	0.10	0.45	0.11	0.52	0.12	0.58	0.21	0.90	0.10	0.48	0.13	0.60
3	0.15	0.71	0.14	0.66	0.20	0.98	0.21	0.98	0.16	0.76	0.15	0.73
4	0.13	0.64	0.15	0.76	0.22	1.07	0.24	1.20	0.15	0.72	0.17	0.85
5	0.16	0.76	0.17	0.85	0.21	1.04	0.24	1.22	0.17	0.81	0.19	0.92
6	0.14	0.70	0.14	0.71	0.22	1.09	0.26	1.31	0.16	0.77	0.16	0.82
7	0.16	0.80	0.17	0.86	0.25	1.23	0.25	1.30	0.18	0.88	0.19	0.95
8	0.15	0.76	0.18	0.92	0.24	1.20	0.26	1.36	0.17	0.85	0.20	1.01
9	0.18	0.88	0.24	1.14	0.32	1.52	0.30	1.49	0.21	1.00	0.25	1.21
High	0.22	1.07	0.24	1.14	0.37	1.65	0.30	1.49	0.25	1.19	0.25	1.21
H–L	0.13***	1.15	0.20***	1.62	0.15**	1.15	0.15*	0.95	0.13***	1.14	0.19***	1.44

Note: Gross performance for decile and H-L portfolios is reported separately for 2001–2019, the post-2019 holdout (2020–2024), and the combined sample (2001–2024), using a non-overlapping five-trading-day holding period ($H = 5$). Chart images used to generate model predictions are constructed from the most recent $L = 60$ trading days. At each formation date, stocks are sorted into deciles based on model scores and equal-weighted decile portfolios are formed. Returns are annualized geometric means of realized holding-period returns; Sharpe ratios are computed from the holding-period return series using annualization factor $\sqrt{252/H}$. H-L test statistics are based on the annualized arithmetic mean of holding-period returns using Newey–West standard errors (4 lags). Statistical significance at the 10%, 5%, and 1% levels is denoted by *, **, and ***, respectively.

Table A2.2 reports results for the longer lookback window ($L = 60$). The decile sorts remain economically ordered in the combined out-of-sample period, and both models continue to deliver positive H–L spreads. Relative to $L = 20$, performance is more compressed at $L = 60$ for both methods in the combined sample, and the incremental H–L advantage of Hiera is smaller.

The subsample results again vary across periods. In 2001–2019, Hiera retains a larger H–L spread than the benchmark at $L = 60$. In the post-2019 holdout (2020–2024), the H–L spreads are equal at 0.15 and the benchmark exhibits a higher Sharpe ratio. Taken together, the lookback-window exercise indicates that (i) the economic ordering implied by model scores is not specific to a single L, and (ii) the incremental advantage of Hiera is largest at the baseline $L = 20$ and becomes smaller at longer lookbacks, particularly in 2020–2024. These patterns support the interpretation that the main $L = 20$ results are not an artifact of a particular lookback choice.

A3 Hiera–Benchmark Return Differentials

Table A3.1: Hiera vs. Benchmark: H–L Return Differential

	Broad Universe			Top 500		
Sample (n)	Hiera−Bench	t (NW=4)	p-value	Hiera−Bench	t (NW=4)	p-value
2001–2019 (951)	0.077***	(4.020)	0.0001	0.065***	(2.879)	0.0040
2020–2024 (247)	0.066*	(1.675)	0.0940	0.028	(0.479)	0.6318
2001–2024 (1198)	0.075***	(4.337)	0.0000	0.057***	(2.650)	0.0081

Note: This table reports the performance differential between Hiera and the Benchmark for the H-L portfolio return series. Results are shown separately for 2001–2019, the post-2019 holdout (2020–2024), and the combined sample (2001–2024). The differential is computed at each rebalance date as $d_t = r^{\text{Hiera}}_{t,\text{H–L}} - r^{\text{Bench}}_{t,\text{H–L}}$, using the weekly holding-period return series. Reported point estimates are annualized arithmetic means of d_t using factor $252/H$ with holding period $H = 5$. Test statistics are based on a constant-only regression of d_t with Newey–West standard errors (4 lags). Statistical significance at the 10%, 5%, and 1% levels is denoted by *, **, and ***, respectively.

A4 Transaction Cost Robustness

This appendix section reports net-of-transaction-cost portfolio results for the $L = 20$, $H = 5$ Hiera signal under alternative proportional cost assumptions. We apply a cost deduction of $c \times \text{Turnover}_t$ at each rebalance, where turnover is measured as firm turnover within each equal-weighted decile portfolio, $\text{Turnover}_t = |S_t \triangle S_{t-1}|/|S_t|$. We consider three values of c: 1, 10, and 25 bps per unit of turnover. We also report average turnover by decile for the combined period.

The results show that net performance is sensitive to the assumed cost level, as expected given the weekly rebalancing schedule. At 1 bps, the Hiera signal continues to produce a positive H–L spread in both universes and across subsamples. At 10 bps, the broad-universe H–L spread remains positive in the combined sample, while net performance in the Top 500 universe is close to zero. At 25 bps, net returns are mostly negative across deciles (except for a few of the high deciles), and the H–L spreads become negative in both universes.

Two features are worth noting. First, turnover is substantially lower in the Top 500 universe, which mechanically reduces the cost drag. Second, even at higher cost assumptions, the highest deciles continue to earn relatively larger net returns than the lowest deciles, consistent with a preserved ranking even when the H–L spread is substantially attenuated.

Table A4.1: Window 20 Hiera Net Performance with Transaction Costs and Average Turnover

	Broad Universe (2001–2019)		Broad Universe (2020–2024)		Broad Universe (2001–2024)		Top 500 (2001–2019)		Top 500 (2020–2024)		Top 500 (2001–2024)	
Decile	Ret	SR	Ret	SR	Ret	SR	Ret	SR	Ret	SR	Ret	SR
Panel A: 1 bps per unit of turnover												
Low	-0.02	0.01	0.02	0.20	-0.01	0.05	0.03	0.24	0.07	0.43	0.04	0.28
2	0.03	0.25	0.04	0.29	0.03	0.26	0.03	0.27	0.06	0.39	0.04	0.30
3	0.06	0.37	0.05	0.32	0.05	0.36	0.05	0.37	0.08	0.46	0.06	0.39
4	0.10	0.55	0.07	0.42	0.09	0.52	0.07	0.45	0.07	0.41	0.07	0.44
5	0.10	0.58	0.09	0.50	0.10	0.56	0.06	0.44	0.10	0.55	0.07	0.46
6	0.12	0.65	0.10	0.53	0.12	0.63	0.09	0.58	0.10	0.55	0.09	0.57
7	0.13	0.72	0.12	0.61	0.13	0.69	0.11	0.68	0.09	0.51	0.11	0.64
8	0.13	0.71	0.15	0.69	0.14	0.70	0.10	0.58	0.15	0.77	0.11	0.62
9	0.17	0.84	0.15	0.69	0.16	0.80	0.12	0.68	0.11	0.60	0.12	0.66
High	0.22	1.03	0.24	0.93	0.22	1.01	0.16	0.84	0.18	0.82	0.16	0.83
H–L	0.23	1.94	0.20	1.52	0.22	1.85	0.12	1.12	0.09	0.74	0.11	1.02
Panel B: 10 bps per unit of turnover												
Low	-0.07	-0.23	-0.03	0.01	-0.06	-0.17	-0.02	-0.05	0.01	0.14	-0.02	-0.01
2	-0.02	0.01	-0.01	0.09	-0.02	0.03	-0.02	-0.02	-0.00	0.10	-0.02	0.00
3	0.00	0.12	-0.00	0.11	0.00	0.12	-0.00	0.08	0.01	0.17	0.00	0.10
4	0.04	0.30	0.02	0.21	0.04	0.28	0.01	0.15	0.00	0.12	0.01	0.14
5	0.05	0.33	0.04	0.30	0.05	0.32	0.01	0.14	0.03	0.26	0.01	0.17
6	0.06	0.40	0.05	0.33	0.06	0.39	0.04	0.29	0.03	0.26	0.04	0.28
7	0.08	0.47	0.07	0.41	0.08	0.45	0.06	0.39	0.03	0.23	0.05	0.36
8	0.08	0.46	0.09	0.49	0.08	0.45	0.04	0.30	0.09	0.49	0.05	0.34
9	0.11	0.59	0.10	0.50	0.11	0.57	0.06	0.41	0.05	0.33	0.06	0.39
High	0.16	0.79	0.18	0.75	0.16	0.78	0.10	0.57	0.11	0.57	0.10	0.57
H–L	0.11	0.99	0.09	0.74	0.10	0.93	0.00	0.10	-0.03	-0.16	-0.00	0.04
Panel C: 25 bps per unit of turnover												
Low	-0.15	-0.62	-0.10	-0.30	-0.14	-0.55	-0.11	-0.53	-0.08	-0.33	-0.10	-0.49
2	-0.10	-0.40	-0.08	-0.26	-0.10	-0.37	-0.10	-0.51	-0.10	-0.39	-0.10	-0.48
3	-0.08	-0.30	-0.08	-0.24	-0.08	-0.28	-0.09	-0.41	-0.08	-0.32	-0.09	-0.39
4	-0.05	-0.12	-0.06	-0.14	-0.05	-0.13	-0.07	-0.35	-0.09	-0.36	-0.08	-0.35
5	-0.04	-0.10	-0.04	-0.04	-0.04	-0.08	-0.08	-0.35	-0.07	-0.23	-0.07	-0.33
6	-0.02	-0.02	-0.03	-0.00	-0.03	-0.02	-0.05	-0.20	-0.07	-0.21	-0.05	-0.20
7	-0.01	0.04	-0.01	0.07	-0.01	0.05	-0.03	-0.08	-0.07	-0.25	-0.04	-0.12
8	-0.01	0.05	0.01	0.16	-0.01	0.08	-0.05	-0.17	-0.02	0.03	-0.04	-0.13
9	0.02	0.18	0.01	0.18	0.02	0.18	-0.03	-0.06	-0.05	-0.11	-0.03	-0.07
High	0.06	0.40	0.09	0.45	0.07	0.41	0.01	0.13	0.01	0.16	0.01	0.14
H–L	-0.07	-0.60	-0.07	-0.55	-0.07	-0.59	-0.16	-1.60	-0.20	-1.68	-0.16	-1.62

Panel D: Average Turnover by Decile (2001–2024)

	Broad Universe	Top 500
Low	9.48	5.34
2	11.32	6.79
3	11.91	6.92
4	11.35	6.69
5	11.28	6.55
6	11.91	7.14
7	11.91	6.87
8	11.36	6.74
9	11.87	6.97
High	9.96	5.76
H–L	19.45	11.09

Note: Net performance for decile and H–L portfolios is reported separately for 2001–2019, the post-2019 holdout (2020–2024), and the combined sample (2001–2024), using a non-overlapping five-trading-day holding period ($H = 5$). At each rebalance date, transaction costs are applied as a proportional deduction from gross portfolio returns equal to $c \times \text{Turnover}_t$, where $c \in \{0.0001, 0.0010, 0.0025\}$ corresponds to 1, 10, and 25 bps per unit of turnover. Turnover is measured as firm turnover within each equal-weighted decile portfolio, $\text{Turnover}_t = |S_t \triangle S_{t-1}|/|S_t S_{t-1}|$ where S_t denotes the set of stocks in the decile at formation date t and $\triangle$ is the symmetric difference operator. H–L turnover is defined as the sum of turnover for the high and low deciles, and costs are applied to both legs. Returns are annualized geometric means of net holding-period returns; Sharpe ratios are computed from the net holding-period return series using annualization factor $\sqrt{252/H}$. Panel D reports average turnover by decile for the combined period 2001–2024 in the Broad Universe and Top 500 universes.

A5 Definitions of Technical Indicators

Table A5.1: Mathematical Definitions of Technical Indicators

Indicator	Definition
Simple Moving Average (SMA, window n)	$\mathrm{SMA}_t^{(n)} = \frac{1}{n}\sum_{i=0}^{n-1} P_{t-i}$
Exponential Moving Average (EMA, window n)	$\mathrm{EMA}_t^{(n)} = \alpha P_t + (1-\alpha)\mathrm{EMA}_{t-1}^{(n)}, \quad \alpha = \frac{2}{n+1}$
5-Day Momentum (Standard)	$\mathrm{Mom5}_t = P_t - P_{t-5}$
5-Day Momentum (Contrarian)	Same as above, but the portfolio is formed as Decile 1 minus Decile 10
MACD Line	$\mathrm{MACD}_t = \mathrm{EMA}_t^{(12)} - \mathrm{EMA}_t^{(26)}$
MACD Signal Line	$\mathrm{Signal}_t = \mathrm{EMA}_t^{(9)}(\mathrm{MACD})$
MACD Histogram	$\mathrm{Hist}_t = \mathrm{MACD}_t - \mathrm{Signal}_t$
VWMA (Volume Weighted Moving Average)	$\mathrm{VWMA}_t^{(n)} = \frac{\sum_{i=0}^{n-1} V_{t-i}P_{t-i}}{\sum_{i=0}^{n-1} V_{t-i}}$
On-Balance Volume (OBV)	$\mathrm{OBV}_t = \mathrm{OBV}_{t-1} + \begin{cases} V_t, & P_t > P_{t-1} \\ -V_t, & P_t < P_{t-1} \\ 0, & P_t = P_{t-1} \end{cases}$
OBV Momentum (n-day)	$\mathrm{OBVmom}_t^{(n)} = \mathrm{OBV}_t - \mathrm{OBV}_{t-n}$
Relative Strength Index (RSI, n-day)	$\mathrm{RSI}_t = 100\left[1 - \frac{1}{1+RS_t}\right], \quad RS_t = \frac{\mathrm{AvgGain}_t^{(n)}}{\mathrm{AvgLoss}_t^{(n)}}$
Stochastic K (n)	$K_t = 100 \cdot \frac{P_t - \min(P_{t-n+1},\ldots,P_t)}{\max(P_{t-n+1},\ldots,P_t) - \min(P_{t-n+1},\ldots,P_t)}$
Stochastic D (n)	$D_t = \mathrm{SMA}_t^{(3)}(K)$
Stochastic (K - D)	$(K-D)_t = K_t - D_t$
Bollinger Z-Score (n)	$Z_t = \frac{P_t - \mathrm{SMA}_t^{(n)}}{\sigma_t^{(n)}}, \quad \sigma_t^{(n)} = \sqrt{\frac{1}{n}\sum_{i=0}^{n-1}(P_{t-i} - \mathrm{SMA}_t^{(n)})^2}$
ATR (Average True Range, n)	$\mathrm{ATR}_t^{(n)} = \mathrm{EMA}_t^{(n)}(\mathrm{TR}_t), \quad \mathrm{TR}_t = \max\{H_t - L_t,\ \lvert H_t - C_{t-1}\rvert,\ \lvert L_t - C_{t-1}\rvert\}$
ATR-Scaled Return (n)	$\mathrm{ATRret}_t = \frac{P_t - P_{t-1}}{\mathrm{ATR}_t^{(n)}}$

A6 Definitions of Synthetic Probe Strategies

Table A6.1: Mathematical Definitions of Synthetic Probe Strategies (I): Trend, Range, and Regime Processes

Strategy	Definition
Breakout	Strength $s \in [0,1]$ controls the tightness of the prebreak range, the postbreak drift, the postbreak volatility, and how tightly the break time is concentrated in the visible window. A random direction $\xi \in \{-1,+1\}$ is drawn. Simple returns are $$r_t = \begin{cases} \varepsilon_t, & t < \tau(s), \\ \mu_{\mathrm{br}}(s)\,\xi + \varepsilon_t, & t \geq \tau(s), \end{cases}$$ where $\tau(s)$ is a random break time located near the center of the visible window, $\varepsilon_t \sim \mathcal{N}\big(0, \sigma^2_{\mathrm{pre}}(s)\big)$ before the break and $\varepsilon_t \sim \mathcal{N}\big(0, \sigma^2_{\mathrm{post}}(s)\big)$ after the break. The variance contrast $\sigma_{\mathrm{post}}(s)/\sigma_{\mathrm{pre}}(s)$ and the expected cumulative move over the window increase in s.
Down Momentum	Strength $s \in [0,1]$ determines the total downward drift over the visible window. Let L be the total series length and W the visible window size. Returns are $$r_t = \mu_{\mathrm{mom}}(s)\,(-1) + c_t(s) + \varepsilon_t,$$ where the constant component satisfies $\sum_{t \in \mathrm{window}} \mu_{\mathrm{mom}}(s)\,(-1) \approx -D_{\max}\, s$ for a fixed $D_{\max} > 0$, $c_t(s)$ is a small linear curvature term that varies across the window to avoid perfectly linear drift, and $\varepsilon_t \sim \mathcal{N}\big(0, \sigma^2(s)\big)$ with $\sigma(s)$ decreasing in s.
Up Momentum	Upward momentum uses the same construction as Down Momentum but with positive drift. Strength $s \in [0,1]$ determines the total upward move over the visible window. Returns are $$r_t = \mu_{\mathrm{mom}}(s)\,(+1) + c_t(s) + \varepsilon_t,$$ where $\sum_{t \in \mathrm{window}} \mu_{\mathrm{mom}}(s) \approx D_{\max}\, s$, $c_t(s)$ is a small linear curvature term, and $\varepsilon_t \sim \mathcal{N}\big(0, \sigma^2(s)\big)$ with $\sigma(s)$ decreasing in s. At $s = 0$ the drift vanishes and the process is close to a random walk.
Mean Reversion	A mean-reverting process around a long-run level m (centered near 100). For strength $s \in [0,1]$, $$X_{t+1} = X_t + \kappa(s)\,(m - X_t) + \eta_t,$$ where $\kappa(s)$ increases in s (stronger pull toward m as s increases) and $\eta_t \sim \mathcal{N}\big(0, \sigma^2_{\mathrm{mr}}(s)\big)$ with $\sigma_{\mathrm{mr}}(s)$ decreasing moderately in s. The observed price series is obtained by rescaling X_t such that the first visible bar equals 100.
Volatility Cluster	A two-state volatility regime with hidden state $Z_t \in \{\mathrm{low}, \mathrm{high}\}$ and strength-dependent persistence. Let $$\Pr(Z_t = Z_{t-1}) = p_{\mathrm{pers}}(s),$$ where $p_{\mathrm{pers}}(s)$ and the contrast between regimes both increase in s. For $s > 0$, one or two regime switches are forced near the visible window. Conditional on Z_t, simple returns satisfy $$r_t \sim \mathcal{N}\big(0, \sigma^2_{Z_t}(s)\big)$$ with $\sigma_{\mathrm{high}}(s) > \sigma_{\mathrm{low}}(s)$. The price path is constructed from the cumulative product of $(1 + r_t)$ and rescaled so that the first visible bar equals 100.

Note: All strategies depend on a strength parameter $s \in [0,1]$ and include Gaussian noise terms so that each probe image corresponds to a different stochastic realisation of the same underlying pattern. Simple returns r_t are mapped to prices via $S_t = S_0 \prod_{u=1}^{t}(1 + r_u)$, and the series is rescaled so that the first visible bar is equal to 100 before chart rendering. Volume dynamics are concept specific and adjust the scale or clustering of volume bars in a way that is consistent with each pattern.

Table A6.2: Mathematical Definitions of Synthetic Probe Strategies (II): Shape-Based Price Profiles

Strategy	Definition
Head and Shoulders	A deterministic profile $h_s(u)$, $u \in [0,1]$, encodes two lower shoulders and a higher head. The peak amplitude $A(s)$ is proportional to the strength parameter s, with additional randomisation of shoulder asymmetry and end slope. For the visible window of length W, let $u_t = t/W$ and define a visible segment $$S_t^{\text{vis}} = S_{\text{base}}\Big(1 + h_s(u_t)\Big)\,(1+\varepsilon_t),$$ where S_{base} is a baseline level and ε_t is a small Gaussian noise term that adds local variation. A short stochastic prefix is generated as a random walk and then smoothly joined to S_1^{vis} so that the full series is continuous. The entire trajectory is rescaled so that the first visible bar is normalised to 100.
Inverse Head and Shoulders	Same construction as Head and Shoulders, but the profile is inverted. The visible segment is $$S_t^{\text{vis}} = S_{\text{base}}\Big(1 - h_s(u_t)\Big)\,(1+\varepsilon_t),$$ so that troughs replace peaks and the overall pattern is the inverse of the head and shoulders shape. Amplitude again scales with s, and the path is preceded and normalised in the same way as in the head and shoulders case.

Note: The construction uses a strength parameter $s \in [0,1]$ and includes a Gaussian noise term so that each probe image corresponds to a different stochastic realisation of the same underlying shape. The full series is rescaled so that the first visible bar is equal to 100 before chart rendering.

A7 Synthetic Probe Examples

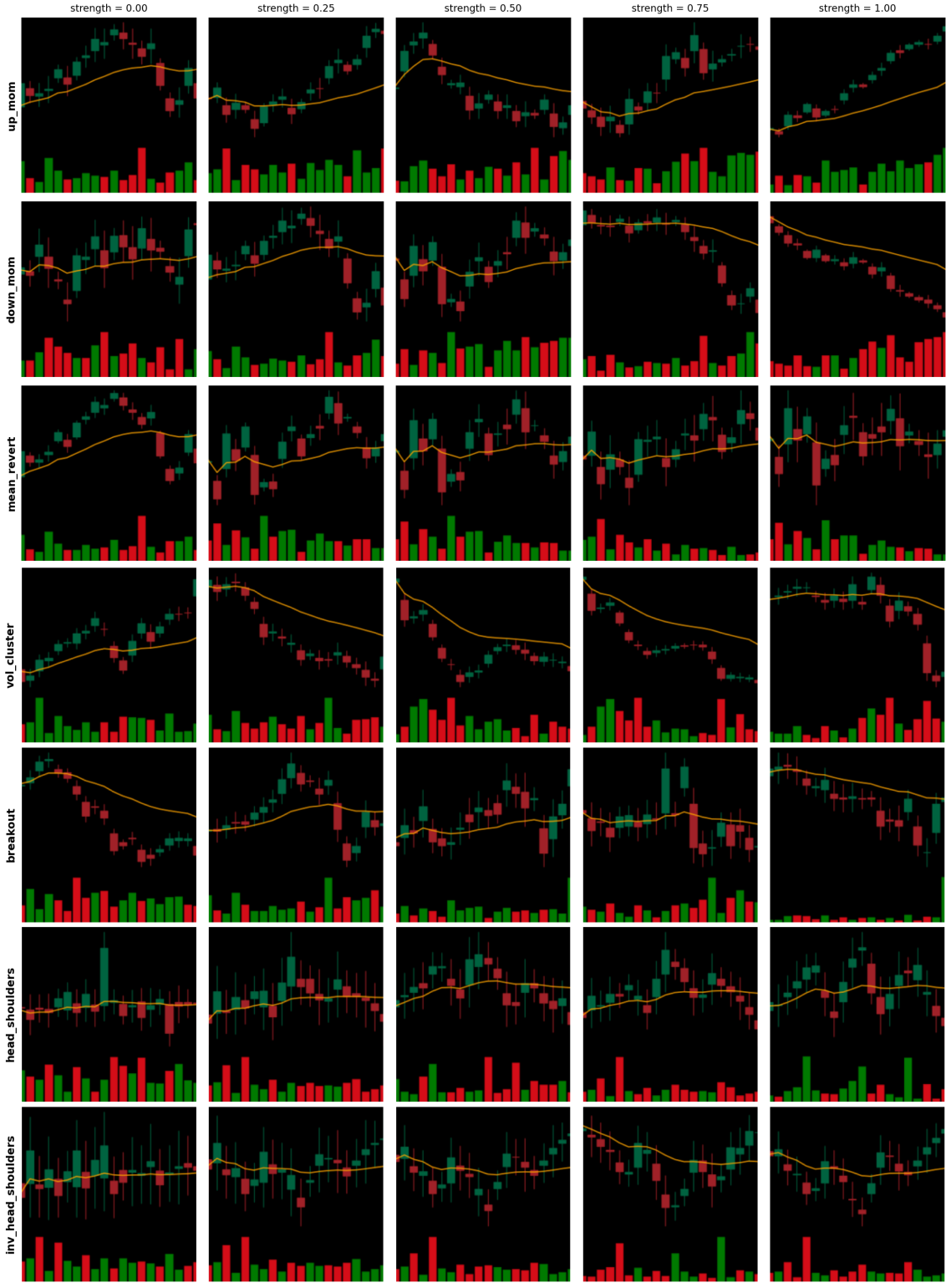


Figure A7.1: Examples for different strength levels accross different trading strategies. See Appendix A6 for details.

A8 Occlusion Sensitivity Maps Details

This appendix describes the procedure used to visualize which regions of a candlestick-chart image the model is most sensitive to when forming directional return forecasts. Because our primary model is a hierarchical vision transformer (Hiera), we use an architecture-agnostic occlusion procedure: we perturb a localized region of the input image, re-evaluate the model, and record how the decision statistic changes. This yields signed evidence maps that indicate whether a region supports an *Up* or *Down* prediction.

Let $x \in \mathbb{R}^{H\times W\times 3}$ denote a candlestick image (resized to $H = W = 224$ and normalized by standard ImageNet channel statistics). The Hiera model maps x into a set of token-level logits

$$\ell(x) \;=\; \Big(\ell_1(x), \ldots, \ell_T(x)\Big), \qquad \ell_t(x) \in \mathbb{R}^2, \qquad T = 49, \tag{A8.1}$$

where T corresponds to the final-stage 7×7 token grid and the two logit coordinates represent the *Down* and *Up* classes. Following the model used to generate the predictions, we aggregate token logits using a log-mean-exp pooling operator:

$$z_c(x) \;=\; \log\left(\frac{1}{T}\sum_{t=1}^{T}\exp\Big(\ell_{t,c}(x)\Big)\right), \qquad c \in \{0, 1\}, \tag{A8.2}$$

and obtain the predicted probability of an *Up* move as

$$p_{\mathrm{Up}}(x) \;=\; \frac{\exp\Big(z_1(x)\Big)}{\exp\Big(z_0(x)\Big) + \exp\Big(z_1(x)\Big)}. \tag{A8.3}$$

For interpretability, we work with the *logit margin*

$$m(x) \;=\; z_1(x) - z_0(x), \tag{A8.4}$$

which is strictly increasing in $p_{\mathrm{Up}}(x)$ and thus preserves the model's ranking of images by "Up" confidence. Using $m(x)$ rather than $p_{\mathrm{Up}}(x)$ is also numerically convenient because it avoids saturation when $p_{\mathrm{Up}}(x)$ is close to 0 or 1.

For a given resolution $G \in \{56, 28, 14, 7\}$, we partition the image into a regular $G \times G$ grid of non-overlapping cells. Since $H = W = 224$, each cell corresponds to a square region of side length

$$s_G \;=\; \frac{224}{G} \in \{4, 8, 16, 32\}. \tag{A8.5}$$

Let $\mathcal{R}_{i,j}^{(G)}$ denote the pixel region corresponding to cell (i, j) in the $G \times G$ grid. We define a perturbed image $x_{i,j}^{(G)}$ by applying a localized occlusion operator $\mathcal{O}$ to $\mathcal{R}_{i,j}^{(G)}$ while leaving the remainder of the image unchanged:

$$x_{i,j}^{(G)} \;=\; \mathcal{O}\Big(x; \mathcal{R}_{i,j}^{(G)}\Big)\,. \tag{A8.6}$$

In our baseline implementation, $\mathcal{O}$ replaces the region by a Gaussian-blurred version of itself (blur-occlusion). This choice preserves low-frequency structure and reduces artifacts introduced by unnatural constant fills, which can be particularly important for sparse line drawings such as financial charts.[4]

We then evaluate how the decision statistic changes under this perturbation:

$$\Delta m_{i,j}^{(G)}(x) \;=\; m\Big(x_{i,j}^{(G)}\Big) - m(x). \tag{A8.7}$$

We convert this into a signed "evidence" measure by reversing the sign:

$$e_{i,j}^{(G)}(x) \;=\; -\Delta m_{i,j}^{(G)}(x) \;=\; m(x) - m\Big(x_{i,j}^{(G)}\Big)\,. \tag{A8.8}$$

This sign convention is intuitive: if occluding a region *reduces* the margin (i.e., $m(x_{i,j}^{(G)}) < m(x)$), then $e_{i,j}^{(G)}(x) > 0$ and the region provides *positive evidence* for an *Up* prediction. Conversely, if occluding a region *increases* the margin, then $e_{i,j}^{(G)}(x) < 0$ and the region provides evidence in favor of *Down*. In the figures, we visualize $e_{i,j}^{(G)}(x)$ using a diverging colormap with red indicating $e_{i,j}^{(G)}(x) > 0$ (supports *Up*) and blue indicating $e_{i,j}^{(G)}(x) < 0$ (supports *Down*).

[4] We also verify that results are qualitatively similar under mean-fill occlusion in normalized pixel space.

The hierarchical nature of Hiera motivates computing occlusion maps at multiple resolutions. We therefore compute the evidence grids

$$E^{(G)}(x) \;=\; \left(e_{i,j}^{(G)}(x)\right)_{i,j=1}^{G} \qquad \text{for } G \in \{56, 28, 14, 7\}. \tag{A8.9}$$

These grids can be interpreted as local sensitivity maps at different spatial scales: $G = 56$ corresponds to fine-grained 4×4 pixel regions, while $G = 7$ corresponds to coarse 32×32 regions that reflect broader chart-level structure. Importantly, we do *not* aggregate these maps across resolutions; rather, we present them side-by-side to illustrate how the model's reliance on image regions evolves with coarser spatial support.

For visualization, each grid $E^{(G)}(x)$ is upsampled to 224×224 pixels using nearest-neighbor interpolation, producing piecewise-constant blocks that correspond exactly to the occluded cells. To ensure that the colormap is not dominated by a small number of outlier cells, we apply a robust normalization separately within each map:

$$\widetilde{e}_{i,j}^{(G)}(x) \;=\; \mathrm{clip}\left(\frac{e_{i,j}^{(G)}(x)}{q^{(G)}(x)}, -1,\, 1\right), \qquad q^{(G)}(x) = \mathrm{Quantile}_{\tau}\left(\left\{|e_{i,j}^{(G)}(x)|\right\}_{i,j}\right), \tag{A8.10}$$

where τ is set to 0.99 (the 99th percentile). We then overlay $\widetilde{E}^{(G)}(x)$ on the original image using transparency proportional to $|\widetilde{e}_{i,j}^{(G)}(x)|$, so that weak evidence does not obscure the underlying chart. Aggregate diagnostics are computed from the unnormalized evidence values $e_{i,j}^{(G)}(x)$; the normalization in (A8.10) is used only for visualization.

The occlusion maps are interventional in the sense that they measure changes in the model's decision statistic induced by localized perturbations of the input. They should be interpreted as local sensitivity diagnostics rather than unique decompositions of the prediction. Because the model is nonlinear, evidence scores can reflect interactions across regions: occluding one region may change how the model uses other regions. Reporting maps at multiple resolutions helps assess the stability of salient regions across perturbation scales. Figure A8.1 reports six additional examples.

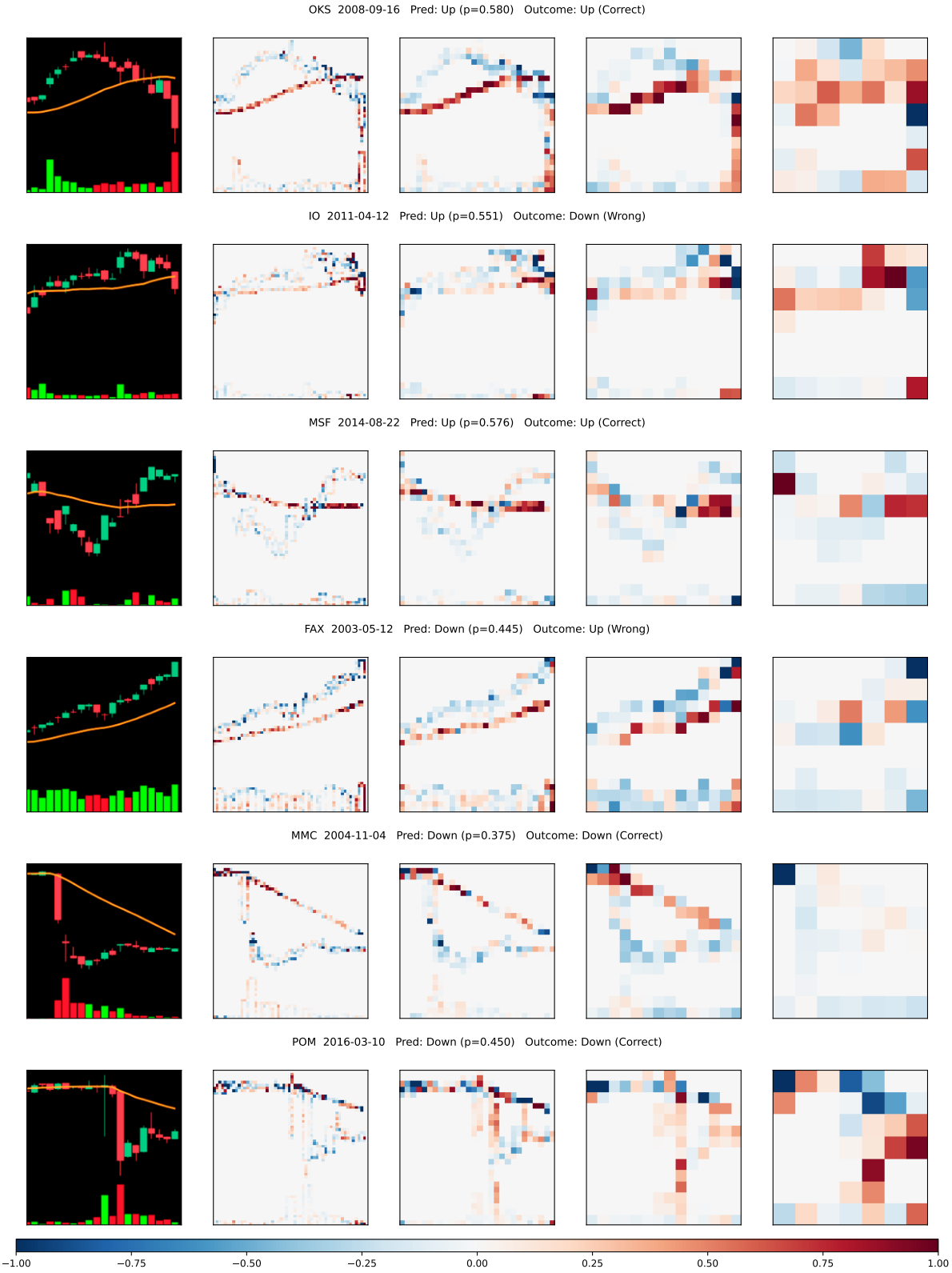


Figure A8.1: Occlusion sensitivity maps for Hiera (six examples). Each row corresponds to one stock–date observation and reports the model's directional forecast (Up/Down), its probability, and whether the forecast is correct. The first column shows the input candlestick image; the remaining columns display signed occlusion effects at four grid resolutions (56×56, 28×28, 14×14, and 7×7). Colors indicate how masking a region changes the model's Up–Down margin: red (blue) denotes positive (negative) evidence for an Up prediction.

A8.1 Aggregate Diagnostics from Occlusion Maps

In addition to the per-observation visualizations, we compute aggregate diagnostics that summarize how occlusion evidence is distributed within the window and how concentrated it is across image regions. Let $e_{i,j}^{(G)}(x)$ denote the signed evidence map defined in (A8.8). For each image x and grid resolution G, define the absolute-evidence mass by column

$$a_j^{(G)}(x) \;=\; \sum_{i=1}^{G} \left| e_{i,j}^{(G)}(x) \right|, \qquad j = 1, \ldots, G, \tag{A8.11}$$

and the corresponding normalized column profile

$$\pi_j^{(G)}(x) \;=\; \frac{a_j^{(G)}(x)}{\sum_{u=1}^{G} a_u^{(G)}(x)}. \tag{A8.12}$$

The profile $\pi^{(G)}(x)$ is a probability distribution over relative horizontal locations in the image and therefore summarizes where occlusion evidence is concentrated along the time axis of the candlestick window. For reporting, we map column indices to the unit interval using

$$t_j \;=\; \frac{j-1}{G-1} \in [0,1], \qquad j = 1, \ldots, G. \tag{A8.13}$$

We aggregate these diagnostics over a balanced sample of model forecasts drawn from the set of high-confidence predictions satisfying $p_{\mathrm{Up}}(x) \geq 0.55$ or $p_{\mathrm{Up}}(x) \leq 0.45$. The resulting sample size and eligibility fraction are reported in Table 3.8. The main plot in the paper reports the cross-sectional mean profile $\mathbb{E}[\pi_j^{(G)}(x)]$ for each G.

To quantify terminal concentration, we report the end-of-window evidence share for the last α fraction of the window,

$$S_\alpha^{(G)}(x) \;=\; \sum_{j=G-\lceil \alpha G \rceil + 1}^{G} \pi_j^{(G)}(x), \qquad \alpha \in \{0.10, 0.20, 0.33\}, \tag{A8.14}$$

and a terminal dominance ratio

$$\mathrm{TDR}^{(G)}(x) \;=\; \frac{S_{0.20}^{(G)}(x)}{1 - S_{0.20}^{(G)}(x)}. \tag{A8.15}$$

We also report distributional location summaries based on the cumulative profile. Define $t_q^{(G)}(x)$ as the smallest t_j such that $\sum_{u=1}^{j} \pi_u^{(G)}(x) \geq q$, with $q \in \{0.50, 0.80\}$. In addition, we report the profile centroid $\sum_{j=1}^{G} t_j \, \pi_j^{(G)}(x)$, which measures the mean time location of absolute evidence.

Finally, to summarize concentration over two-dimensional image regions rather than along the time axis only, let $\{|e_{i,j}^{(G)}(x)|\}_{i,j}$ denote the absolute evidence values across the G^2 cells. For $\kappa \in \{0.01, 0.05, 0.10\}$, define the top-$\kappa$ mass share as

$$\mathrm{TopMass}_{\kappa}^{(G)}(x) \;=\; \frac{\sum_{(i,j)\in\mathcal{I}_\kappa(x)} |e_{i,j}^{(G)}(x)|}{\sum_{i=1}^{G}\sum_{j=1}^{G} |e_{i,j}^{(G)}(x)|},$$

where $\mathcal{I}_\kappa(x)$ indexes the $\lceil \kappa G^2 \rceil$ cells with the largest $|e_{i,j}^{(G)}(x)|$. Table 3.8 reports cross-sectional means of these diagnostics for each grid resolution.