PHBS WORKING PAPER SERIES

Binocular Directional Forecasting: A Cross-Attention Fusion Approach

Yutao Yuan
Peking University

Lingxiao Zhao
Peking University

January 2026

Working Paper 20260102

Abstract

We develop a binocular directional forecasting framework that jointly leverages information from equity and option markets through a multimodal deep learning architecture. Realized price and trading dynamics, together with option-implied volatility information, are encoded as two-dimensional images and processed using convolutional neural networks (CNNs), then integrated through a cross-attention mechanism with an adaptive gating network. This design enables bidirectional information flow across markets and state-dependent weighting of heterogeneous signals. Using U.S. equity and option daily data from 1996 to 2023, we show that the binocular model significantly outperforms stock-only benchmarks in out-of-sample directional prediction at both monthly and quarterly horizons. Trading strategies based on the fused forecasts achieve higher Sharpe ratios and lower turnover than traditional momentum and reversal strategies. Performance gains are most pronounced during periods of elevated market uncertainty and weak historical predictability, highlighting the incremental value of forward-looking option-implied expectations. Overall, our findings demonstrate that cross-attention provides an effective and scalable approach for integrating realized and expectations-based information in financial forecasting.

Keywords: equity return prediction, image-based learning, cross-attention, portfolio performance
JEL Classification: G11, G12, G17, C45, C55

Peking University HSBC Business School
University Town, Nanshan District
Shenzhen 518055, China

Binocular Directional Forecasting: A Cross-Attention Fusion Approach*

Yutao Yuan[a], Lingxiao Zhao[a,*]

[a]*Peking University HSBC Business School University Town, Nanshan District Shenzhen, 518055 P.R. China*

Abstract

We develop a binocular directional forecasting framework that jointly leverages information from equity and option markets through a multimodal deep learning architecture. Realized price and trading dynamics, together with option-implied volatility information, are encoded as two-dimensional images and processed using convolutional neural networks (CNNs), then integrated through a cross-attention mechanism with an adaptive gating network. This design enables bidirectional information flow across markets and state-dependent weighting of heterogeneous signals. Using U.S. equity and option daily data from 1996 to 2023, we show that the binocular model significantly outperforms stock-only benchmarks in out-of-sample directional prediction at both monthly and quarterly horizons. Trading strategies based on the fused forecasts achieve higher Sharpe ratios and lower turnover than traditional momentum and reversal strategies. Performance gains are most pronounced during periods of elevated market uncertainty and weak historical predictability, highlighting the incremental value of forward-looking option-implied expectations. Overall, our findings demonstrate that cross-attention provides an effective and scalable approach for integrating realized and expectations-based information in financial forecasting.

Keywords: Equity return prediction, Image-based learning, Cross-attention, Portfolio performance

JEL: G11, G12, G17, C45, C55

*The project is supported by the National Nature Science Foundation of China (Grant No: 72503015).

*Corresponding author

Email addresses: 2301212286@phbs.pku.edu.cn (Yutao Yuan), lingxiao@phbs.pku.edu.cn (Lingxiao Zhao)

Binocular Directional Forecasting:

A Cross-Attention Fusion Approach

Abstract

We develop a binocular directional forecasting framework that jointly leverages information from equity and option markets through a multimodal deep learning architecture. Realized price and trading dynamics, together with option-implied volatility information, are encoded as two-dimensional images and processed using convolutional neural networks (CNNs), then integrated through a cross-attention mechanism with an adaptive gating network. This design enables bidirectional information flow across markets and state-dependent weighting of heterogeneous signals. Using U.S. equity and option daily data from 1996 to 2023, we show that the binocular model significantly outperforms stock-only benchmarks in out-of-sample directional prediction at both monthly and quarterly horizons. Trading strategies based on the fused forecasts achieve higher Sharpe ratios and lower turnover than traditional momentum and reversal strategies. Performance gains are most pronounced during periods of elevated market uncertainty and weak historical predictability, highlighting the incremental value of forward-looking option-implied expectations. Overall, our findings demonstrate that cross-attention provides an effective and scalable approach for integrating realized and expectations-based information in financial forecasting.

Key Words: Equity return prediction; Image-based learning; Cross-attention; Portfolio performance

JEL Classification: G11, G12, G17, C45, C55

1 Introduction

Financial markets generate multiple distinct informational signals about the same underlying asset. Focusing on a single informational channel may leave complementary signals from related markets unexploited. Yet empirical frameworks that coherently integrate heterogeneous information into a single, economically meaningful predictor remain relatively scarce.

To address this gap, we introduce a novel *binocular directional forecasting* framework that fuses complementary information sources within a unified multimodal deep learning model. Recent work in empirical asset pricing has shown that flexible machine learning methods are well suited for modeling complex and nonlinear relationships in financial data (Gu et al., 2020; Chen et al., 2024). Building on this insight, a growing and influential literature pioneers the use of convolutional neural networks (CNNs) and image-based representations to extract rich predictive signals from financial markets, including historical price and trading patterns (Jiang et al., 2023; Murray et al., 2024) and option-implied volatility surfaces (Kelly et al., 2023). Extending these important contributions, we show that jointly modeling these distinct yet complementary information sources can enhance directional return prediction.

This paper is motivated by the observation that a single asset can be viewed through two complementary informational views. In binocular vision, depth perception arises from the interaction of two eyes rather than from either eye alone. By the same logic, more reliable directional forecasts emerge from jointly modeling stock price dynamics and option-implied information. The equity market provides a backward-looking view, summarizing realized stock prices and historical trading dynamics. The options market offers a forward-looking view that reflects market expectations about future risk and uncertainty. Each view contains relevant but incomplete information for predicting return direction on its own. Treating these signals separately ignores their interaction and limits predictive performance.

Specifically, we develop a unified mixture-of-experts forecasting architecture that integrates heterogeneous signals from the stock and option markets through a cross-attention fusion layer and an adaptive gating mechanism. The cross-attention layer enables bidirectional interaction between stock-based and option-based representations, allowing information from each market to be interpreted in the context of the other. Building on these context-aware representations, the adaptive gating mechanism governs how information from the two sources is combined. Rather than relying on fixed linear aggregation or a prespecified interaction structure, the gating network assigns state-dependent weights to the stock-based and option-based expert branches, which allows their relative importance to adjust across market conditions.

The framework represents stock price histories and option-implied volatility surfaces as two complementary image-based inputs. Following Jiang et al. (2023), we encode realized stock price dynamics as two-dimensional images constructed from Open–High–Low–Close (OHLC) bars, moving average overlays, and trading volume, providing a transparent and well-established benchmark for the historical information channel. In parallel, we introduce option-implied volatility images, a novel forward-looking representation that maps the implied volatility surface at a maturity matched to the forecast horizon into a standardized two-dimensional image parameterized by option delta. While the stock price images summarize recent realized price and trading dynamics, the option images capture market expectations about future risk and asymmetry embedded in option prices.

Our empirical analysis combines daily stock market data from CRSP with option-implied volatility surface data from OptionMetrics over the period January 1996 to December 2023, focusing on stocks for which both price and option data are available. We evaluate three classes of predictive models: a *Stock Only Model* based on image representations of historical price and trading dynamics, an *Option Only Model* based on implied volatility images, and a *Fusion Model* that jointly incorporates both information sources. Across all specifications, stock image

lookback windows and option maturities are aligned with the return prediction horizon to ensure temporal consistency between inputs and targets. All models are trained using data from January 1996 to December 2005, with a random 70/30 split between training and validation, and are evaluated exclusively out of sample over the period January 2006 to December 2023. This design allows us to isolate the incremental predictive value of option-implied information and assess the economic relevance of multimodal fusion.

Our empirical results show that integrating option-implied volatility with historical stock price and trading information delivers substantial and robust improvements in cross-sectional return direction prediction. Across both monthly and quarterly horizons, the *Fusion Model* consistently outperforms stock-only benchmarks in terms of returns, Sharpe ratios, and cumulative performance, while exhibiting equal or lower portfolio turnover. The gains are largest when short-horizon price images are combined with horizon-matched option information, indicating that forward-looking option-implied expectations are especially valuable when historical price signals are noisy or slow to adjust. The *Fusion Model* also dominates traditional momentum and reversal strategies, whose predictive power decays rapidly at longer horizons. Although option-implied volatility alone contains economically meaningful predictive content, its strongest and most persistent value emerges when combined with historical price dynamics. Overall, the evidence demonstrates that multimodal fusion of backward-looking price patterns and forward-looking option information yields more stable, persistent, and economically significant return predictability, particularly during periods of elevated market uncertainty.

Our work builds upon a broad foundation of financial literature regarding market efficiency and information theory. The predictive power of historical price dynamics we leverage is well-documented, spanning from short-term (monthly and weekly) reversals (Jegadeesh, 1990; Lehmann, 1990) to medium-term momentum (Jegadeesh and Titman, 1993). Theoretical work suggests these price sequences allow for the inference of private information (Treynor and

Ferguson, 1985; Brown and Jennings, 1989; Blume et al., 1994), which can be captured via automated pattern recognition (Brock et al., 1992; Lo et al., 2000; Neely et al., 2014; Han et al., 2016). Similarly, our use of the option-implied volatility surface is grounded in the literature showing that the cross-section of option prices identifies the risk-neutral return distributions (Jackwerth and Rubinstein, 1996; Aït-Sahalia and Lo, 1998). Previous studies confirm that a steeper volatility smirk, negative skewness, and even higher-order moments like kurtosis significantly forecast future returns (Bakshi et al., 2003; Xing et al., 2010; Diavatopoulos et al., 2012; Christoffersen et al., 2013). Furthermore, we bridge findings that discrepancies between call and put IVs identify informed sentiment (Ofek et al., 2004; Cremers and Weinbaum, 2010) with recent advancements in deep learning for asset pricing (Chen et al., 2024). By demonstrating that cross-attention mechanisms can bridge the semantic gap between temporal and structural financial data, we provide a scalable blueprint for the next generation of multimodal financial prediction models.

The remainder of the paper is organized as follows. Section 2 describes the image representation and the multimodal deep learning architecture. Section 3 presents the data and empirical results. Section 4 concludes.

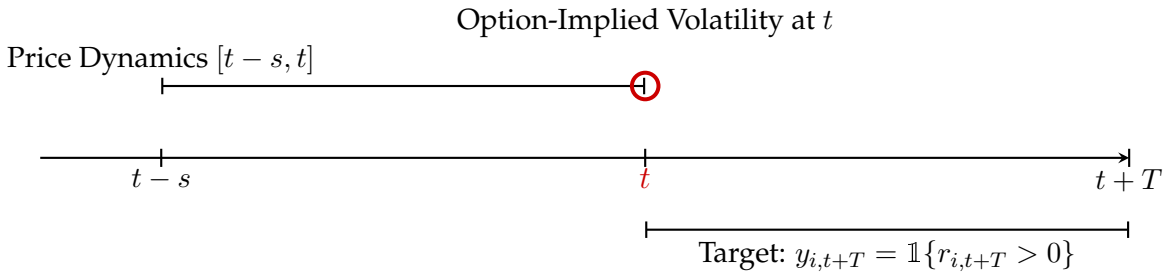
2 Methodology

This section presents the methodological framework for binocular directional forecasting based on the joint use of stock price images and option-implied volatility images. For each stock i at date t, we construct two complementary image-based representations that serve as inputs to a single predictive model: a stock price image summarizing recent historical price dynamics and an option-implied volatility image capturing contemporaneous market expectations derived from options written on the same underlying stock.

The prediction target is defined as the binary directional indicator

$$y_{i,t+T} = \mathbb{1}\{r_{i,t+T} > 0\}, \tag{1}$$

where $r_{i,t+T}$ denotes the cumulative holding-period return from date t to $t+T$, with $T \in \{20, 60\}$ trading days. The stock price image encodes realized price dynamics over the lookback window $[t-s, t]$, with $s \in \{5, 20, 60\}$ trading days, while the option-implied volatility image provides a cross-sectional snapshot of the implied volatility surface for the same stock observed at date t.



The above figure illustrates the temporal alignment between the model inputs and the prediction target. At each date t for each stock, that date serves simultaneously as the final observation in the stock price image and as the date at which the option-implied volatility image is constructed. This alignment ensures that both input modalities reflect the same underlying market information set at time t.

To integrate the two image-based representations, we employ a single deep learning model that uses a cross-attention fusion architecture. Rather than simply combining features from the two sources, cross-attention allows information from the stock-based and option-based views to interact dynamically and in both directions. We further incorporate an adaptive gating mechanism that controls the relative influence of stock-based and option-based features in the final representation, allowing the model to place greater weight on the more informative view under different market conditions.

The remainder of this section first describes the construction of the stock price and option-implied volatility images, and then introduces the cross-attention and gating mechanisms used to fuse these representations for directional forecasting.

2.1 Image Representations

2.1.1 Stock Price Images

As the historical component of our binocular framework, a stock price image captures the realized price dynamics of the underlying asset. To ensure a clean benchmark and to isolate the incremental contribution of option market information, we construct stock price images by closely following the image transformation approach proposed by Jiang et al. (2023). This design choice ensures that any subsequent performance differences can be attributed to the integration of option-based information rather than to differences in stock image construction.

For each trading day, the opening, high, low, and closing prices are represented as a single OHLC bar in a grayscale image. The high and low prices define the endpoints of a vertical line, while the opening and closing prices are indicated by short horizontal ticks on the left and right sides of the vertical line, respectively. Each OHLC bar is rendered in white against a black background and has a fixed width of three pixels: one pixel for the vertical line and one pixel for each horizontal tick. This pixel-level standardization ensures a consistent visual encoding of

daily price movements across assets and time.

By concatenating OHLC bars across consecutive trading days, we obtain a grayscale stock price image whose total width equals $3s$ pixels for an s-trading-day lookback window. We construct stock images using window lengths of $s \in \{5, 20, 60\}$ trading days, corresponding approximately to one week, one month, and one quarter, respectively. For a given window length, the image height is fixed across all stocks and dates to ensure consistent input dimensions. Within each image, the vertical scale is normalized so that the top and bottom pixel rows correspond to the highest and lowest prices observed over the window, allowing relative price levels to be encoded by vertical pixel position.

Consistent with Jiang et al. (2023), we further enrich the stock price images by incorporating moving average price lines and trading volume information. A moving average line, computed using the same window length as the image, is overlaid on the OHLC bars and rendered in white with a thickness of one pixel. Trading volume is displayed in the bottom one-fifth of the image, while the remaining four-fifths are allocated to price information. Each volume bar is rendered in white, centered beneath its corresponding OHLC bar, and scaled proportionally to that day's trading volume relative to the maximum volume observed within the window.

Taken together, these elements form a two-dimensional (2D) stock price image that summarizes recent price dynamics and trading volume over the given lookback window. Figure 1 illustrates an example constructed using a 20-trading-day lookback window.

2.1.2 Option-Implied Volatility Images

As the forward-looking component of our binocular framework, an option-implied volatility image captures market expectations about future price movements and risk embedded in option prices. While the stock price images summarize realized historical price dynamics, the option-implied volatility images provide a contemporaneous snapshot of the market's

assessment of uncertainty, tail risk, and asymmetry in the return distribution. Together, these two representations offer complementary views of the same underlying stock.

From an economic perspective, option data form an implied volatility surface, a three-dimensional (3D) object defined over option moneyness and time to maturity. Although the Black–Scholes–Merton (Black and Scholes (1973); Merton (1973)) framework assumes constant volatility across strikes and maturities, empirical option prices exhibit systematic variation in implied volatility, giving rise to well-documented smile and skew patterns (Rubinstein, 1985, 1994; Bates, 1991). In this study, we parameterize the moneyness dimension using option delta. Following common practice in equity and foreign exchange derivatives markets (Campa et al., 1998; Gatheral, 2011; Mingone, 2023), delta provides a scale-free and economically meaningful ordering of options that facilitates comparison across assets and over time.

Building on recent work that represents the implied volatility (IV) surface as an image (Kelly et al., 2023), we exploit option data in a way that is explicitly aligned with our return directional forecasting horizon. While the option market provides a full three-dimensional IV surface over deltas and maturities at each date, we construct a two-dimensional (2D) representation by fixing maturity at a horizon consistent with the forecast window. This avoids pooling information across maturities, which would dilute horizon-specific signals.

For prediction horizons $T \in \{20, 60\}$ trading days, we use options with maturities closest to 30 and 90 calendar days, respectively. At each stock–date pair, implied volatilities at the selected maturity are collected across delta levels and arranged into a grayscale image, with delta on the horizontal axis and implied volatility on the vertical axis. As with stock price images, the vertical scale is normalized within each image to emphasize relative shape rather than absolute level, allowing the model to focus on option-implied asymmetries relevant for directional forecasting. All option images have a fixed height of 32 pixels and a fixed width of 34 pixels, corresponding to delta levels from -0.90 to 0.90 in increments of 0.05.

Figure 2 illustrates two representative examples of the resulting option-implied volatility images. In the resulting images, the left portion corresponds to negative-delta options (puts), while the right portion corresponds to positive-delta options (calls). The vertical structure of the image captures cross-sectional variation in implied volatility across delta levels, reflecting asymmetries between downside and upside risk as well as the overall shape of the implied volatility surface.

2.2 Deep Learning Model: A Cross-Attention Fusion Approach

We employ a deep learning architecture that integrates stock price images and option-implied volatility images through a cross-attention fusion mechanism. The model consists of three main components: (i) modality-specific convolutional neural networks (CNNs) for feature extraction, (ii) a bidirectional cross-attention module for cross-modal interaction, and (iii) an adaptive gating network that regulates the contribution of each modality in the final prediction.

CNN Feature Extraction. We begin by using two independent CNNs to extract high-level representations from the stock price images and the option-implied volatility images, respectively. CNNs are well suited for image-based financial prediction tasks because weight sharing substantially reduces parameterization, making them easier to train with limited data, while convolution and pooling operations provide robustness to shifts, scale changes, and local deformations in the input. These properties allow CNNs to automatically extract localized and nonlinear spatial patterns from images without manual feature engineering (Jiang et al., 2023).

A CNN processes an input image through stacked convolutional blocks, each comprising a convolutional layer, a nonlinear activation function, and a pooling operation. Convolutional layers slide filters across the image to summarize local pixel patterns, activation functions such as Leaky ReLU introduce nonlinearity, and max-pooling layers reduce dimensionality by retaining the most informative local features. Through this hierarchical structure, the network

transforms raw pixel values into increasingly abstract and informative representations.

For stock price images, we adopt the CNN architecture proposed by Jiang et al. (2023), which uses different network depths for 5-day, 20-day, and 60-day images to capture price patterns at multiple time scales. This specification serves as our *Stock Only Model*, providing a benchmark for evaluating the incremental value of option-based information. For option implied volatility images, which have lower spatial resolution, we design a lighter CNN with two convolutional blocks. This network produces predicted return probabilities and is referred to as the *Option Only Model*. The corresponding architectures are illustrated in Figures 3 and 4.

Cross-Attention Fusion. Our main model replaces the final fully connected layers of the stock and option CNNs with linear projection layers that map each input to a d-dimensional feature vector, where $d = 64$ in our implementation. Let $X \in \mathbb{R}^{N\times d}$ and $Y \in \mathbb{R}^{N\times d}$ denote the resulting feature matrices extracted from the stock and option CNNs, respectively. Here, $N = 128$ denotes the batch size.

To integrate these representations, we employ a cross-attention mechanism that allows information from one modality to selectively attend to relevant features in the other. Cross-attention, originally introduced in Vaswani et al. (2017), enables dynamic and directional interaction between feature sets and is well suited for multimodal fusion. In our setting, it allows the model to adaptively determine when forward-looking option information should dominate and when realized price patterns are more informative.

Focusing on the stock-to-option direction, the query, key, and value matrices are obtained through learned linear projections:

$$Q = XW_Q, \qquad K = YW_K, \qquad V = YW_V, \tag{2}$$

where $W_Q, W_K, W_V \in \mathbb{R}^{d\times d}$ are trainable parameter matrices. The cross-attention weights are

computed using scaled dot-product attention:

$$\alpha = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d}}\right), \tag{3}$$

where $\alpha \in \mathbb{R}^{N\times N}$ is the normalized attention matrix and the scaling factor $\sqrt{d}$ stabilizes gradients. The attended representation is given by $\alpha V \in \mathbb{R}^{N\times d}$.

To increase representational capacity, we implement multi-head attention with $H = 8$ heads, each operating on a subspace of dimension $d_h = d/H = 8$. For head $h = 1, \ldots, H$, the projections are $W_h^Q, W_h^K, W_h^V \in \mathbb{R}^{d\times d_h}$, and the head-specific outputs are

$$\alpha_h = \text{Softmax}\left(\frac{(XW_h^Q)(YW_h^K)^\top}{\sqrt{d_h}}\right), \tag{4}$$

$$z_h = \alpha_h(YW_h^V). \tag{5}$$

The outputs $z_1, \ldots, z_H$ are concatenated and projected back to a d-dimensional space through a linear transformation. Residual connections and layer normalization are applied to stabilize training and preserve representation integrity, following standard practice in attention-based architectures (Vaswani et al., 2017; Ba et al., 2016). An analogous attention module is constructed in the opposite direction by reversing the roles of X and Y. Additional implementation details are provided in Appendix A.2.

Adaptive Gating Network. While cross-attention captures rich interaction patterns between modalities, it does not explicitly regulate their relative importance in the final prediction. To address this, we introduce an adaptive gating network that controls the contribution of each modality at the fusion stage. As emphasized by Arevalo et al. (2017) and Tsai et al. (2019), attention and gating serve distinct roles: attention allocates weights within a feature set, whereas gating allocates weights across modalities.

Consistent with mixture-of-experts models (Jacobs et al., 1991; Shazeer et al., 2017), the gating network functions as a learned expert-weighting mechanism. Let $Z_s \in \mathbb{R}^{N\times d}$ and $Z_o \in \mathbb{R}^{N\times d}$ denote the bidirectionally attended stock and option feature matrices. We concatenate these features as $f = [Z_s, Z_o] \in \mathbb{R}^{N\times 2d}$ and pass them through a two-layer gating network:

$$g = \text{Softmax}(W_2\sigma(W_1 f + b_1) + b_2)\,, \tag{6}$$

where $W_1 \in \mathbb{R}^{h\times 2d}$, $W_2 \in \mathbb{R}^{2\times h}$, and $\sigma(\cdot)$ denotes the ReLU activation function. The resulting weights g_s and g_o correspond to the stock and option modalities, respectively. The final fused representation is computed as

$$Z_f = g_s \odot Z_s + g_o \odot Z_o, \tag{7}$$

where $\odot$ denotes element-wise multiplication. This gating mechanism allows the model to dynamically emphasize the modality that provides more informative signals for each prediction instance. The fused representation Z_f is finally mapped to the predicted probabilities associated with the directional target $y_{t,T}$ via a fully connected softmax layer.

2.3 Training and Validation Procedure

To ensure a fair comparison with the benchmark *Stock Only Model* of Jiang et al. (2023), we closely follow their training protocol. The full sample spans January 1996 through December 2023 and is partitioned into training, validation, and testing sets. The in-sample cover the initial ten-year period from January 1996 to December 2005. Within this in-sample data, we randomly assign 70% of observations to the training set and the remaining 30% to the validation set. Random splitting helps maintain a balanced distribution of positive and negative labels in both sets, which is important for classification tasks. Across all prediction horizons considered, the resulting class distribution is approximately balanced, with positive outcomes accounting for roughly 50–55% of observations. The testing set covers January 2006 to December 2023 and is

reserved exclusively for out-of-sample performance evaluation.

Following Jiang et al. (2023), we formulate stock return direction forecasting as a binary classification problem, with the target variable defined in Equation 1. For notational simplicity, we suppress time and stock subscripts when no confusion arises. Model training minimizes the binary cross-entropy loss,

$$L(y, \hat{y}) = -y \log(\hat{y}) - (1 - y) \log(1 - \hat{y}), \tag{8}$$

where $y \in \{0, 1\}$ denotes the true label and $\hat{y} \in \{0, 1\}$ denotes the predicted class label obtained from the softmax output.

We adopt the regularization and optimization strategy proposed by Gu et al. (2020) and implemented by Jiang et al. (2023), with minor adaptations to accommodate our architecture. All network parameters are initialized using the Xavier initializer (Glorot and Bengio, 2010) to stabilize the variance of activations across layers.

Model training employs the Adam optimizer (Kingma and Ba, 2014) with an initial learning rate of 1×10^{-5} and mini-batches of size $N = 128$. Each convolutional block includes batch normalization (Ioffe and Szegedy, 2015) prior to the nonlinear activation to improve convergence stability. To mitigate overfitting, we apply a dropout rate of 50% to the fully connected layer, consistent with the recommendations of Gu et al. (2020). Training is terminated early when the validation loss fails to improve for two consecutive epochs.

Because parameter initialization and optimization are stochastic, we repeat the training procedure five times and average the resulting forecasts, following Gu et al. (2020).

2.4 Out-of-Sample Performance Evaluation Measures

We evaluate out-of-sample performance using cross-sectional long–short portfolios formed on model-predicted return probabilities. At each rebalancing date, stocks are sorted into deciles

based on their predicted probability of a positive return. The top decile constitutes the long portfolio, while the bottom decile constitutes the short portfolio, yielding an equal-weighted long–short strategy.

The portfolio holding period matches the prediction horizon and begins immediately after the final observation used to construct the stock price image. Portfolios are rebalanced monthly for $T = 20$ trading-day forecasts and quarterly for $T = 60$ trading-day forecasts, ensuring consistency between portfolio formation, rebalancing frequency, and the forecasting task.

Portfolio performance is summarized using annualized returns, Sharpe ratios, and portfolio turnover. Annualized returns are computed by scaling average returns by 12 for monthly portfolios and by 4 for quarterly portfolios. The annualized Sharpe ratio is defined as

$$\text{Sharpe Ratio} = \frac{\text{Annualized Return}}{\text{Std(Annualized Return)}}, \tag{9}$$

where Std(Annualized Return) denotes the standard deviation of annualized portfolio returns.

To assess trading intensity and implementability, we compute portfolio turnover. Let $w_{i,t}$ denote the weight of stock i at the beginning of rebalancing period t, and let $r_{i,t+T}$ denote its return over the subsequent holding period. The normalized portfolio weight prior to rebalancing is

$$\tilde{w}_{i,t+T} = \frac{w_{i,t}(1 + r_{i,t+T})}{1 + \sum_j w_{j,t} r_{j,t+T}}. \tag{10}$$

Turnover in period t is computed as

$$\text{Turnover}_t = \sum_i |w_{i,t+T} - \tilde{w}_{i,t+T}|, \tag{11}$$

and average monthly turnover is defined as

$$\text{Turnover} = \frac{1}{M}\frac{1}{T}\sum_{t=1}^{T}\text{Turnover}_t, \tag{12}$$

where M denotes the number of months per holding period (e.g., $M = 1$ for monthly portfolios and $M = 3$ for quarterly portfolios). Under this definition, a strategy that fully reconstitutes its holdings each period attains a maximum turnover of $200\%/M$, whereas a buy-and-hold strategy has zero turnover.

To illustrate performance dynamics, we plot cumulative log returns of the high–minus–low (H–L) portfolios. At each rebalancing date t, the realized return $R_{p,t}$ is transformed into a log return, $\log(1 + R_{p,t})$, and cumulative log returns up to time τ are computed as

$$\text{CumLogRet}_\tau = \sum_{t\leq\tau}\log(1 + R_{p,t}). \tag{13}$$

Log returns are additive over time and yield numerically stable equity curves for long-horizon comparisons; accordingly, all cumulative performance figures report cumulative log returns rather than compounded simple returns.

Taken together, the methodology integrates image-based representations of historical price dynamics and option-implied volatility through a cross-attention fusion architecture, with model training and validation conducted in a fixed in-sample period and economic relevance assessed through out-of-sample portfolio tests.

3 Empirical Analysis

This section presents the empirical analysis. We begin by describing the data sources, sample construction, and model configurations used in the analysis. We then report the main out-of-sample results comparing the *Fusion Model* with the *Stock Only Model*, which serves as our primary benchmark and isolates the incremental value of incorporating option-implied information. Next, we evaluate the *Fusion Model* against traditional historical return based predictors widely studied in the asset-pricing literature. Finally, we examine the performance of the *Option Only Model* to assess the predictive content of option-implied volatility information in isolation and to highlight the complementary role of multimodal fusion.

3.1 Data

We use daily stock return and price data from the Center for Research in Security Prices (CRSP) and option-implied volatility data from the OptionMetrics Volatility Surface file. The full sample spans January 1996 through December 2023, reflecting the earliest availability of OptionMetrics data in January 1996.

Matching CRSP and OptionMetrics yields a sample of 7,786 unique stocks with linked stock and option data. For each stock, we construct three types of stock price images based on historical windows of 5, 20, and 60 trading days. In parallel, we construct two types of option-implied volatility images using options with maturities closest to 30-calendar-day and 90-calendar-day horizons, respectively. The prediction target is a binary indicator equal to one if the stock's holding-period return over the subsequent 20 or 60 trading days is positive and zero otherwise.

We retain only those observations for which both stock price data and option data are available, ensuring that all image representations can be constructed consistently across modalities.

3.2 Model Configurations and Experimental Design

This subsection describes the set of predictive models evaluated in the empirical analysis and clarifies how input horizons and prediction targets are aligned across model specifications.

The *Stock Only Model* comprises six separately trained versions, corresponding to all combinations of the three stock image lookback windows (5, 20, and 60 trading days) and the two return prediction horizons (20 and 60 trading days). The *Option Only Model* consists of two versions: one using 30-calendar-day maturity option implied volatility images to predict 20-trading-day returns, and the other using 90-calendar-day maturity option-implied volatility images to predict 60-trading-day returns.

The *Fusion Model* mirrors the stock-only setup and also contains six versions, in which option-implied information is aligned with the return prediction horizon. Specifically, a 5-day stock image used to predict 20-trading-day returns is combined with a 30-calendar-day maturity option-implied volatility image, while a 60-trading-day return prediction is paired with a 90-calendar-day maturity option-implied volatility image. This horizon matching ensures that option-implied information reflects market expectations over a time scale comparable to the stock return forecast window.

For notational convenience, we denote model configurations as "Is/RT," where s indicates the stock image lookback window and T denotes the return prediction horizon. Under this notation, "I5/R20" refers to a model that uses 5-trading-day stock images to predict 20-trading-day ahead returns. For the *Fusion Model*, this notation additionally implies that the option image is constructed from options observed on the last day of the stock image window and with a maturity matched to the prediction horizon. For the *Option Only Model*, we use "RT" to denote the prediction horizon, as there is a unique option image type for each horizon.

Our empirical analysis combines daily stock market data from CRSP with option-implied

volatility surface data from OptionMetrics over the period January 1996 to December 2023. We restrict attention to stocks for which both price and option data are available, yielding a large and representative cross section of stocks. The analysis evaluates three classes of predictive models: a *Stock Only Model* benchmark based on image representations of historical price and trading dynamics, an *Option Only Model* based on implied volatility images, and a *Fusion Model* that jointly incorporates both information sources. Across specifications, stock image lookback windows and option maturities are explicitly aligned with the return prediction horizon to ensure temporal consistency between inputs and targets. All models are trained using a fixed in-sample period and evaluated exclusively out of sample, allowing us to isolate the incremental predictive value of option-implied information and to assess the economic relevance of multimodal fusion relative to established historical return–based predictors.

All models are trained once using data from January 1996 to December 2005. Within this period, 70% of observations are randomly assigned to the training set, with the remaining 30% reserved for validation. Out-of-sample performance is then evaluated over the period from January 2006 through December 2023, using the models estimated during the initial training phase.

3.3 Performance of the Fusion Model versus the Stock Only Benchmark

We begin the empirical analysis by comparing the *Fusion Model* with the *Stock Only Model*, which serves as our primary benchmark and isolates the incremental economic value of incorporating option-implied information beyond historical stock price dynamics. From an investment perspective, this comparison asks whether forward-looking information embedded in option markets improves the profitability and risk-adjusted performance of trading strategies formed on image-based stock price signals.

Tables 1 and 2 report the out-of-sample performance of long–short portfolios constructed from the two models' predicted return directions over 20-trading-day and 60-trading-day

horizons, respectively. Performance is evaluated using annualized returns, Sharpe ratios, and portfolio turnover, allowing us to assess not only economic profitability but also risk adjustment and implementability.

One-month-ahead return direction prediction. Table 1 summarizes the comparison results for the 20-trading-day (one-month-ahead) prediction horizon. Across all stock image lookback windows, the *Fusion Model* consistently outperforms the corresponding *Stock Only Model* in terms of both economic profitability and risk-adjusted performance. This finding suggests that forward-looking information embedded in option markets contains incremental predictive content beyond that captured by historical price patterns alone.

Among all specifications, the *Fusion Model* based on 5-day stock images (I5/R20) delivers the strongest performance, generating the largest high-minus-low (H–L) portfolio returns, the highest t-statistics, and the most favorable Sharpe ratios. Quantitatively, the I5/R20 Fusion specification achieves an annualized return of approximately 9% with a Sharpe ratio of 0.85, compared with an annualized return of about 5% and a Sharpe ratio of 0.51 for the corresponding *Stock Only Model*. Similar, albeit slightly smaller, improvements are observed for the 20-day and 60-day stock image horizons, indicating that the economic benefits of incorporating option-implied volatility are robust to the choice of historical price window rather than driven by a single configuration.

In addition to higher returns, the *Fusion Model* exhibits slightly lower portfolio turnover across all stock image horizons. From an economic perspective, this reduction in turnover suggests that option-implied volatility information helps stabilize trading signals by anchoring price-based patterns to market expectations about future risk. As a result, the *Fusion Model* reduces excessive rebalancing while delivering stronger risk-adjusted returns, thereby improving implementability and mitigating concerns related to transaction costs.

One-quarter-ahead return direction prediction. Table 2 reports the results for the 60-trading-day (one-quarter-ahead) prediction horizon. Consistent with the shorter-horizon findings, the *Fusion Model* uniformly outperforms the *Stock Only Model* in terms of annualized returns and Sharpe ratios across all stock image horizons, while maintaining comparable or lower portfolio turnover. This evidence indicates that the incremental information embedded in option-implied volatility remains economically relevant even as the forecast horizon lengthens.

Although the absolute magnitude of returns is lower at the quarterly horizon, the relative performance ranking of the models is unchanged, underscoring the robustness of the fusion framework across prediction horizons. From an economic perspective, this pattern suggests that option-implied volatility captures persistent market expectations about future risk that extend beyond short-term price dynamics, allowing the *Fusion Model* to retain predictive power when purely price-based signals begin to decay.

Cumulative performance and dynamics. Beyond summary statistics, Figures 6 and 7 plot the cumulative log returns of the high–minus–low (H–L) portfolios constructed from the *Fusion Model* and the *Stock Only Model*. These cumulative return profiles provide insight into the temporal stability and economic persistence of the models' predictive signals, beyond what is captured by average returns and Sharpe ratios.

For the 20-trading-day return prediction task (Figure 6), the *Fusion Model* consistently dominates the Stock Only benchmark across all three stock image horizons. The performance gap widens steadily over time, indicating that the economic gains from incorporating option-implied volatility information accumulate gradually rather than arising from a small number of extreme observations. This pattern suggests that option-implied information enhances the reliability of return signals on an ongoing basis, rather than merely improving performance during isolated market episodes. The gains are most pronounced for shorter stock image horizons, with the I5/R20 Fusion specification achieving the highest cumulative return among all

models. Economically, this finding implies that forward-looking option market expectations are particularly valuable when historical price signals are short-term and potentially contaminated by transitory noise.

A similar, though attenuated, pattern emerges for the 60-trading-day horizon (Figure 7). Although cumulative returns are generally lower than those observed at the shorter horizon, Fusion Models continue to outperform their Stock Only counterparts across all stock image horizons, with the performance advantage again being most evident for the I5/R60 specification. Moreover, Fusion-based portfolios exhibit smoother return trajectories over time, consistent with superior risk-adjusted performance and reduced exposure to episodic drawdowns. Taken together, these dynamics suggest that combining backward-looking price patterns with forward-looking option-implied volatility improves both the persistence and the stability of return predictability across forecast horizons.

Performance during market stress. An informative contrast emerges during periods of heightened market stress, most notably around the COVID-19 shock in 2020. While the I20/R60 and I60/R60 portfolios—under both the Stock Only and Fusion specifications—experience pronounced drawdowns, the I5/R60 portfolios remain comparatively resilient. This pattern indicates that short-horizon price images, which place greater weight on recent market information, adjust more rapidly to abrupt changes in economic conditions and therefore offer improved downside protection during turbulent periods.

In contrast, longer-horizon price representations appear slower to adapt to sudden regime shifts. Although incorporating option-implied volatility improves performance on average, option-based information alone does not fully offset the inertia inherent in long-horizon historical price signals during sharp market dislocations. Taken together, these findings underscore the importance of information timeliness—alongside information richness—for achieving robust performance under extreme market conditions.

Overall, the results in this subsection provide strong evidence that integrating option-implied volatility into an image-based forecasting framework yields economically and statistically meaningful improvements over models that rely solely on historical stock prices and trading patterns. The gains are particularly pronounced for short-horizon price images and during periods of elevated market uncertainty, suggesting that option-implied information supplies complementary forward-looking signals rather than redundant noise.

3.4 Performance of the Fusion Model versus Traditional Prior Return-Based Predictors

While the preceding analysis demonstrates that the *Fusion Model* outperforms the Stock Only benchmark, an important remaining question is whether these improvements simply reflect exposure to well-known return predictors embedded in historical price dynamics. To address this concern, we benchmark the *Fusion Model* against a set of traditional price-based trading strategies that have been extensively studied in the asset pricing literature.

Specifically, we compare the out-of-sample performance of the Fusion Model with three widely used return predictors: the 2–12 momentum (MOM) strategy, the one-month short-term reversal (STR) strategy, and the one-week short-term reversal (WSTR) strategy. Because these benchmarks rely exclusively on historical price information, they provide a natural and economically meaningful basis for assessing whether the *Fusion Model* delivers incremental predictive power beyond established price-based return regularities.

Table 3 reports the results for the 20-trading-day return prediction horizon. The *Fusion Model* substantially outperforms all traditional price-based strategies in terms of risk-adjusted performance, with particularly strong gains for specifications based on short-horizon price images. Among all strategies, the I5/R20 Fusion specification achieves the highest Sharpe ratio, exceeding those of the MOM, STR, and WSTR benchmarks by a wide margin. Although the WSTR strategy generates statistically significant excess returns, its performance is accompanied by considerably higher return volatility, resulting in inferior Sharpe ratios. These results

indicate that the *Fusion Model* delivers more stable excess returns, rather than merely amplifying short-term price fluctuations.

Differences in portfolio turnover further illuminate the economic nature of the *Fusion Model*'s performance gains. The I5/R20 Fusion specification exhibits turnover comparable to STR and WSTR, and higher than that of MOM, reflecting its reliance on short-horizon information. Importantly, this higher trading intensity is accompanied by markedly superior Sharpe ratios, suggesting that the improved performance is driven by more informative return signals rather than by excessive trading activity.

Table 4 extends the comparison to the 60-trading-day horizon. In contrast to traditional strategies, whose predictive power largely dissipates as the forecast horizon lengthens, the *Fusion Model* remains both economically and statistically significant. Notably, the I5/R60 Fusion specification is the only strategy that delivers a statistically significant long–short return at this horizon. This pattern highlights a key limitation of purely price-based predictors: their signals decay rapidly over longer horizons. By incorporating option-implied volatility, which reflects forward-looking market expectations about risk, the *Fusion Model* retains predictive power even when historical price information becomes less informative.

Turnover comparisons at the longer horizon reinforce this conclusion. Although the *Fusion Model* exhibits higher turnover than the MOM benchmark, its trading intensity remains comparable to STR and WSTR while delivering substantially higher Sharpe ratios. This trade-off suggests that the *Fusion Model* achieves superior performance through enhanced information content rather than increased trading frequency.

Overall, the evidence demonstrates that the *Fusion Model* does not merely replicate well-known momentum or reversal effects. Instead, it extracts incremental predictive information from the interaction between historical price and trading dynamics and option-implied volatility. By combining backward-looking price patterns with forward-looking market expectations, the

Fusion Model delivers more robust and persistent return predictability than traditional single-modality strategies.

3.5 Performance of the Option Only Model

We conclude the empirical analysis by isolating the predictive content of the options market. Specifically, we evaluate the *Option Only* Model, which uses option-implied volatility (IV) images as the sole input to forecast the direction of future holding-period stock returns. This exercise serves two complementary roles. First, it provides a direct test of whether IV surfaces contain standalone information that is economically meaningful out of sample. Second, it helps interpret the performance of the *Fusion Model* by distinguishing pure option-driven predictability from gains that arise only when option information is combined with historical price dynamics.

Table 5 reports out-of-sample results for $T = 20$ and $T = 60$ trading-day horizons. At the 20-day horizon, portfolios formed on option-based signals exhibit a clear cross-sectional spread: average returns increase across deciles, and the high-minus-low (H–L) portfolio delivers economically meaningful performance with strong risk-adjusted returns. The monotone pattern across deciles indicates that variation in the shape of the IV profile contains systematic information about the cross-section of subsequent stock return direction, rather than reflecting isolated episodes. At the 60-day horizon, the *Option Only* strategy remains profitable, though performance attenuates relative to $T = 20$. The H–L return spread is still statistically meaningful and the decile ordering remains broadly stable, suggesting that option-implied expectations contain predictive content that extends beyond very short horizons, albeit with weaker signal-to-noise.

Turnover patterns indicate that IV-based signals are not purely transitory. Monthly turnover for the *Option Only* strategy is comparable to the other benchmarks, implying that its performance is not mechanically driven by excessive rebalancing. This stability improves implementability and is consistent with the interpretation that option-implied information

reflects persistent shifts in market beliefs about risk and return asymmetry.

Figure 8 plots the cumulative log returns of the high–minus–low (H–L) portfolios for the *Fusion Model* (best-performing specifications at each horizon, I5/R20 and I5/R60) and the *Option Only Model* (R20 and R60). The *Option Only* portfolios exhibit a steadily increasing cumulative return profile over the full out-of-sample period, confirming that option-implied volatility alone supports a persistent long–short return spread. Moreover, the relatively smooth evolution of these equity curves suggests that option-based signals generate stable performance over time, which may be attractive to volatility-sensitive investors.

At the same time, the *Fusion Model* consistently dominates the *Option Only Model* in cumulative performance at both horizons, with the performance gap widening over time. This pattern highlights the incremental value of integrating forward-looking option-implied expectations with backward-looking price and trading dynamics through the fusion architecture. Overall, while option-implied volatility contains economically meaningful standalone directional information, the most reliable and persistent predictability arises when option-based signals are combined with historical price information in a unified multimodal framework.

Taken together, the empirical results demonstrate that incorporating option-implied volatility into an image-based forecasting framework yields economically meaningful and robust improvements in directional return predictability. Among all specifications, the I5/R20 *Fusion Model* consistently delivers the strongest performance across all evaluation criteria, including return magnitude, annualized Sharpe ratio, and cumulative performance. This dominance underscores the importance of combining short-horizon price and trading dynamics with horizon-matched option-implied expectations, and highlights the economic value of multimodal fusion for capturing complementary backward- and forward-looking information in equity return direction forecasting.

4 Conclusion

This paper develops a multimodal deep learning framework for cross-sectional equity return direction prediction that integrates two complementary image-based modalities observed at the same decision date: a backward-looking stock price image summarizing recent realized price and trading volume dynamics, and a forward-looking option-implied volatility image capturing market expectations about future risk. To synthesize these data sources, we employ a cross-attention fusion architecture that enables bidirectional interaction between stock-based and option-based feature representations, enhanced by a mixture-of-experts strategy with an adaptive feature-level gating mechanism. This learned routing system assigns state-dependent weights to stock and option "experts," allowing the relative influence of historical price and trading dynamics and forward-looking volatility expectations to vary across assets and over time in forming the final directional forecast.

Using daily CRSP stock data and OptionMetrics volatility surface data from January 1996 to December 2023, we evaluate economic relevance through out-of-sample portfolio tests over the period from January 2006 to December 2023. Across both 20-trading-day and 60-trading-day prediction horizons and across all stock-image lookback windows, the Fusion Model consistently outperforms the Stock Only Model in terms of annualized returns and Sharpe ratios. These improvements persist across horizons and image constructions, indicating that the incremental value of option-implied information is robust rather than an artifact of model tuning. The strongest performance is achieved by the short-horizon configuration (I5/R20), suggesting that option-implied expectations are especially valuable when short-horizon price and volume signals are most exposed to transitory noise and rapid regime changes.

Benchmarking against traditional prior return-based strategies further clarifies the economic content of the fusion signal. Relative to widely studied return-based predictors, fusion-based portfolios deliver superior risk-adjusted performance and retain predictive power at

longer horizons, where purely stock-market-based signals tend to weaken. While the *Option Only Model* confirms that implied volatility surfaces contain economically meaningful standalone information, the *Fusion Model* typically delivers stronger and more stable cumulative performance by combining forward-looking option signals with backward-looking price information.

Beyond these empirical findings, this paper addresses a broader research question central to modern quantitative asset management: how to synthesize high-dimensional, heterogeneous data streams into a coherent predictive signal. As artificial intelligence enables the extraction of rich technical features from diverse markets and modalities, the primary challenge shifts from information availability to disciplined integration. Traditional approaches—relying on fixed linear combinations, prespecified factor structures, or single-market perspectives—are ill-suited for environments where predictive relevance varies across assets, horizons, and market regimes.

The binocular framework proposed in this paper offers a data-driven response to this challenge. By formalizing signal integration as a mixture-of-experts problem, the model learns to dynamically weight and combine information from different modalities through cross-attention and adaptive gating, avoiding manual calibration of signal importance. While option-implied volatility serves as a concrete and economically meaningful example of forward-looking information, the underlying mixture-of-experts architecture remains agnostic to the specific modalities employed. The empirical results demonstrate that explicitly modeling cross-modal interactions can translate heterogeneous technical information into economically meaningful portfolio performance.

Tables and Figures

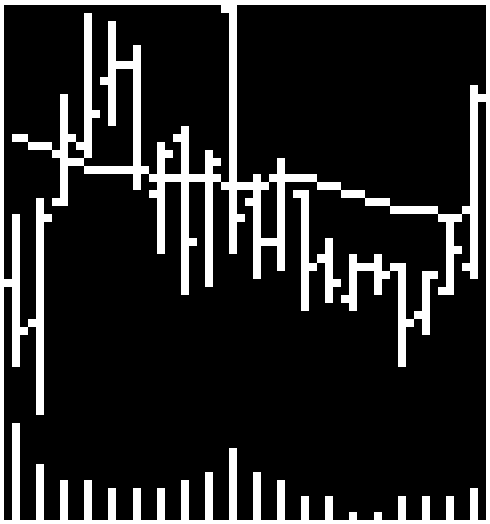
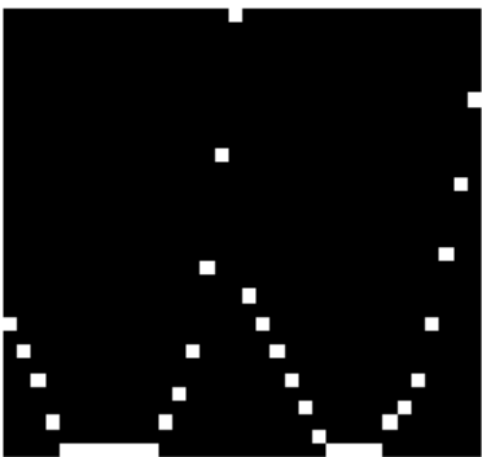
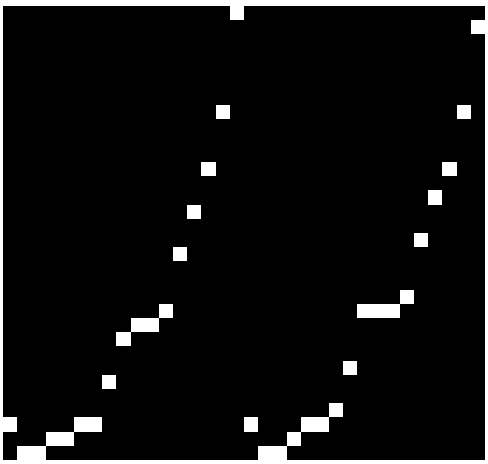


Figure 1 Example of Stock Price Image (20-Trading-Day Lookback Window)



(a) Option Image Type A



(b) Option Image Type B

Figure 2 Example of Option-Implied Volatility Images

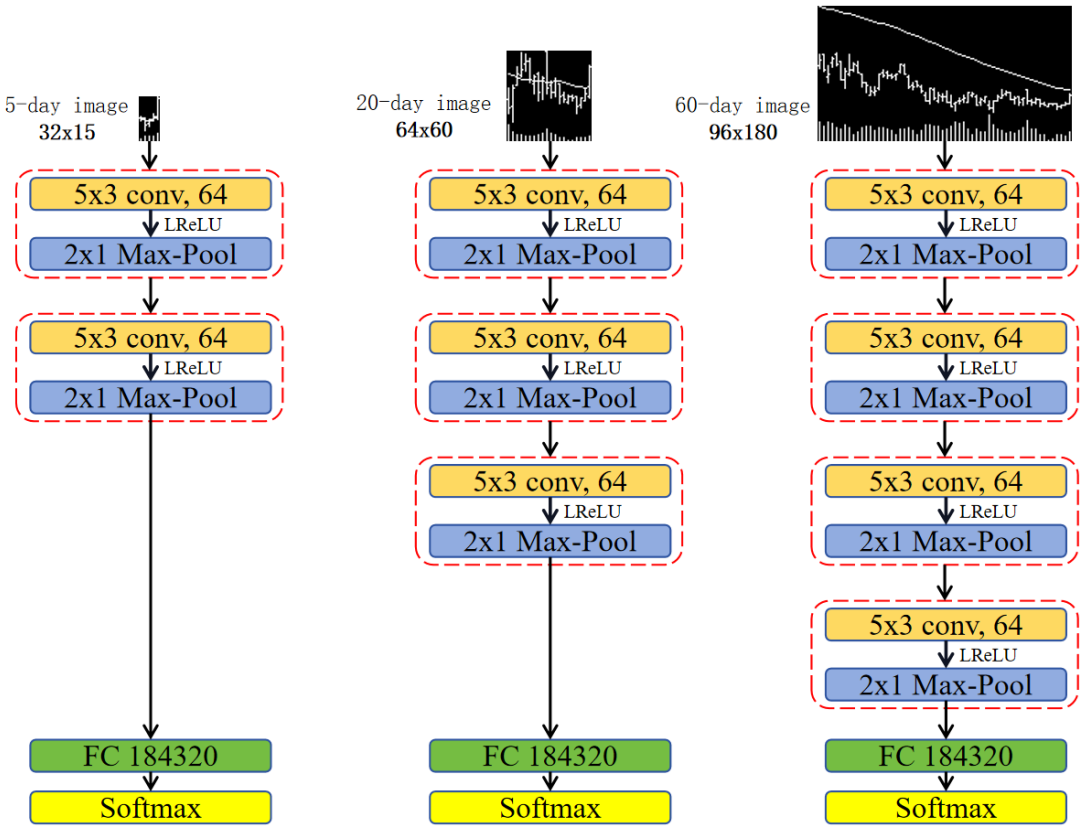


Figure 3 CNN Architecture for Stock Price Image Processing

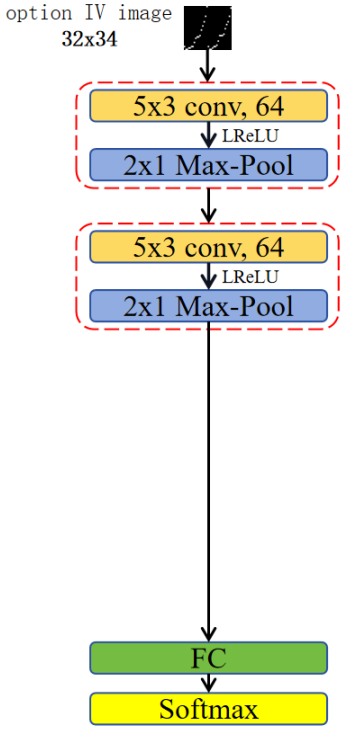


Figure 4 CNN Architecture for Option-Implied Volatility Image Processing

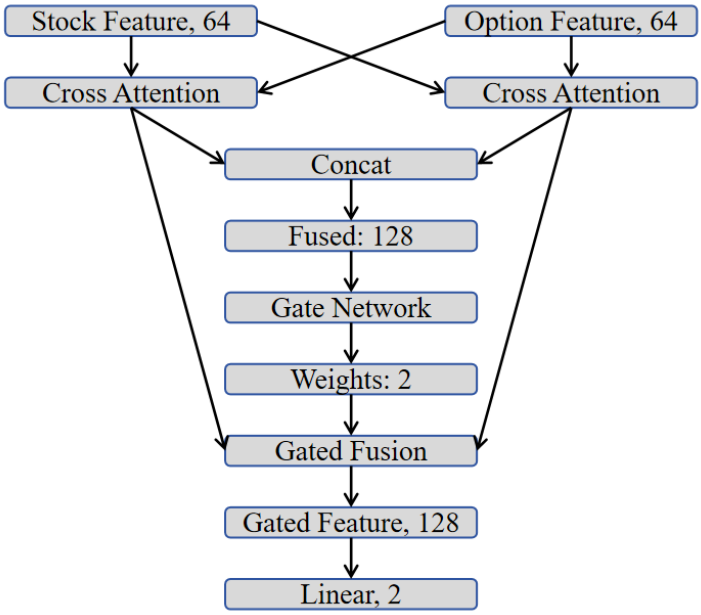


Figure 5 Cross-Attention Fusion Architecture

Table 1 Fusion vs Stock Only Model Performance for R20 Horizon

	Equal-Weight											
	Fusion Model						Stock Only Model					
	I5/R20		I20/R20		I60/R20		I5/R20		I20/R20		I60/R20	
	Ret	SR	Ret	SR	Ret	SR	Ret	SR	Ret	SR	Ret	SR
Low	0.05	0.25	0.07	0.33	0.07	0.33	0.07	0.34	0.08	0.40	0.09	0.40
2	0.07	0.33	0.09	0.44	0.10	0.47	0.09	0.42	0.10	0.46	0.10	0.47
3	0.09	0.44	0.10	0.46	0.10	0.48	0.09	0.44	0.10	0.49	0.11	0.50
4	0.10	0.51	0.09	0.45	0.11	0.54	0.10	0.47	0.10	0.48	0.11	0.51
5	0.10	0.49	0.10	0.46	0.10	0.48	0.09	0.45	0.09	0.44	0.10	0.50
6	0.11	0.51	0.11	0.51	0.10	0.50	0.10	0.50	0.11	0.54	0.10	0.49
7	0.11	0.53	0.12	0.56	0.12	0.57	0.11	0.54	0.11	0.56	0.10	0.49
8	0.11	0.56	0.11	0.53	0.10	0.49	0.11	0.56	0.11	0.55	0.10	0.49
9	0.12	0.58	0.11	0.53	0.11	0.53	0.12	0.56	0.10	0.48	0.11	0.53
High	0.14	0.69	0.12	0.59	0.10	0.50	0.12	0.59	0.10	0.50	0.10	0.51
H-L	**0.09*****	0.85	**0.05****	0.47	0.03	0.27	**0.05****	0.51	0.01	0.17	0.02	0.21
Turnover	1.76		1.79		1.75		1.80		1.81		1.78	

Note: This table presents the out-of-sample portfolio performance results comparing the *Fusion Model* with *Stock Only Model* for R20 horizon returns of equal-weighted portfolios using data from January 2006 through December 2023. Portfolios are formed by sorting stocks into deciles based on the respective model predictions. Ret represents annualized returns, SR denotes the Sharpe ratio. H-L shows the hedge portfolio return (High minus Low) with significance levels: * p<0.10, ** p<0.05, *** p<0.01. Turnover indicates the average monthly portfolio turnover.

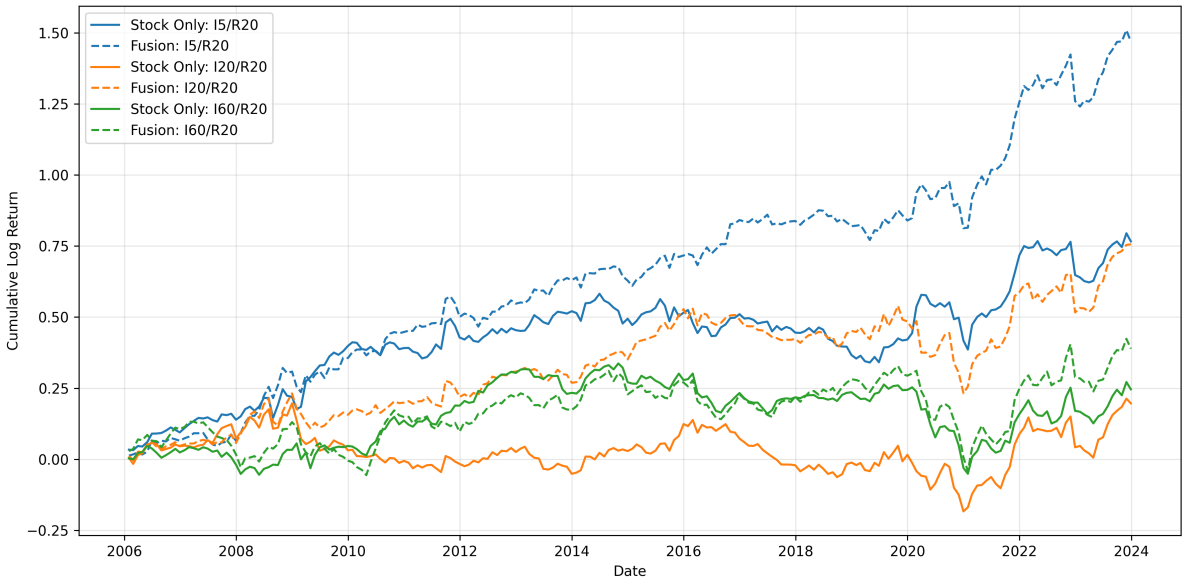


Figure 6 Fusion vs Stock Only Model: H-L Portfolio Returns for R20 Horizon

Note: This figure illustrates the out-of-sample cumulative returns of the High-Low (H-L) portfolios formed based on the *Fusion Model* and *Stock Only Model* predictions for R20 horizon using data from January 2006 through December 2023. The dotted line represents the *Fusion Model*, while the solid line represents the *Stock Only Model*. Different stock image horizons are shown in different colors.

Table 2 Fusion vs Stock Only Model Performance for R60 Horizon

	Equal-Weight											
	Fusion Model						Stock Only Model					
	I5/R60		I20/R60		I60/R60		I5/R60		I20/R60		I60/R60	
	Ret	SR	Ret	SR	Ret	SR	Ret	SR	Ret	SR	Ret	SR
Low	0.07	0.32	0.10	0.39	0.10	0.35	0.07	0.33	0.09	0.38	0.10	0.37
2	0.10	0.44	0.09	0.41	0.11	0.44	0.10	0.41	0.10	0.41	0.13	0.52
3	0.11	0.49	0.11	0.46	0.09	0.40	0.11	0.51	0.11	0.46	0.12	0.48
4	0.10	0.46	0.10	0.44	0.11	0.49	0.11	0.48	0.11	0.46	0.12	0.52
5	0.10	0.47	0.11	0.48	0.10	0.46	0.11	0.49	0.11	0.46	0.10	0.45
6	0.11	0.47	0.10	0.46	0.11	0.49	0.12	0.53	0.12	0.53	0.10	0.46
7	0.12	0.54	0.11	0.50	0.12	0.54	0.11	0.48	0.11	0.51	0.09	0.43
8	0.12	0.52	0.12	0.52	0.11	0.51	0.11	0.50	0.11	0.50	0.11	0.51
9	0.13	0.51	0.11	0.51	0.12	0.55	0.12	0.52	0.12	0.52	0.11	0.51
High	0.13	0.54	0.13	0.61	0.12	0.60	0.12	0.53	0.11	0.55	0.11	0.55
H-L	**0.06*****	0.66	0.03	0.21	0.02	0.19	**0.05****	0.57	0.02	0.20	0.01	0.10
Turnover	0.59		0.61		0.61		0.60		0.61		0.61	

Note: This table presents the out-of-sample portfolio performance results comparing the *Fusion Model* with *Stock Only Model* for R60 horizon returns of equal-weighted portfolios using data from January 2006 through December 2023. Portfolios are formed by sorting stocks into deciles based on the respective model predictions. Ret represents annualized returns, SR denotes the Sharpe ratio. H-L shows the hedge portfolio return (High minus Low) with significance levels: * p<0.10, ** p<0.05, *** p<0.01. Turnover indicates the average monthly portfolio turnover.

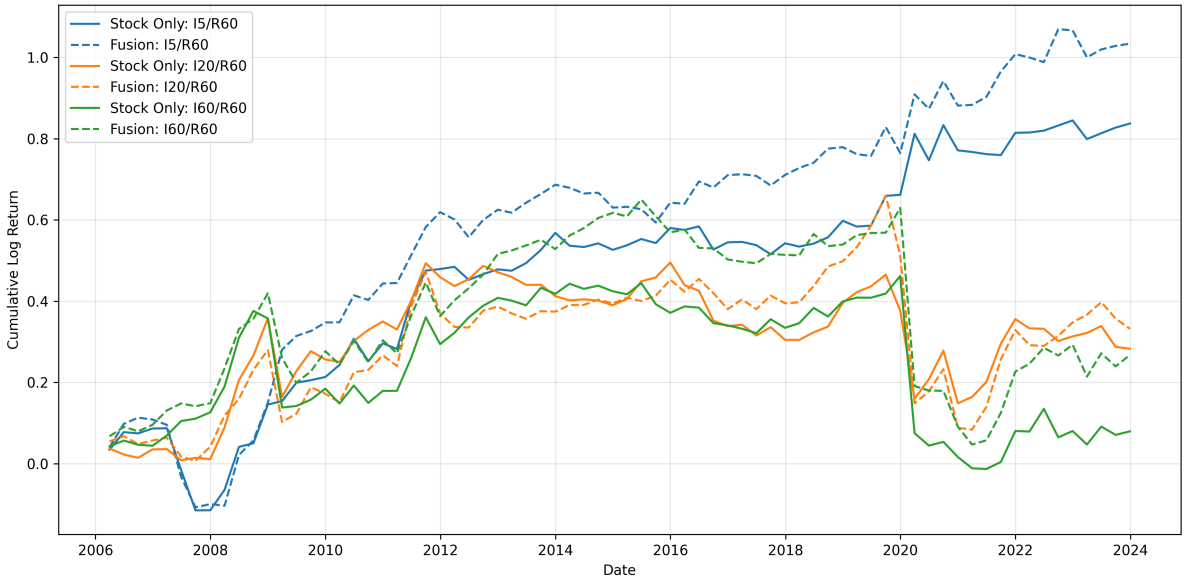


Figure 7 Fusion vs Stock Only Model: H-L Portfolio Returns for R60 Horizon (2023)

Note: This figure illustrates the cumulative returns of the High-Low (H-L) portfolios formed based on the *Fusion Model* and *Stock Only Model* predictions for R60 horizon. The dotted line represents the *Fusion Model*, while the solid line represents the *Stock Only Model*. Different stock image horizons are shown in different colors.

Table 3 Performance Comparison for R20 Horizon of the Fusion Model with Traditional Prior Return-Based Predictors

	Equal-Weight											
	Fusion Model						Traditional Strategy					
	I5/R20		I20/R20		I60/R20		MOM/R20		STR/R20		WSTR/R20	
	Ret	SR	Ret	SR	Ret	SR	Ret	SR	Ret	SR	Ret	SR
Low	0.05	0.25	0.07	0.33	0.07	0.33	0.09	0.27	0.09	0.42	0.08	0.33
2	0.07	0.33	0.09	0.44	0.10	0.47	0.10	0.39	0.10	0.53	0.09	0.44
3	0.09	0.44	0.10	0.46	0.10	0.48	0.11	0.49	0.10	0.54	0.09	0.47
4	0.10	0.51	0.09	0.45	0.11	0.54	0.10	0.53	0.10	0.57	0.09	0.49
5	0.10	0.49	0.10	0.46	0.10	0.48	0.11	0.62	0.10	0.60	0.09	0.52
6	0.11	0.51	0.11	0.51	0.10	0.50	0.11	0.63	0.11	0.60	0.10	0.58
7	0.11	0.53	0.12	0.56	0.12	0.57	0.11	0.63	0.10	0.55	0.12	0.64
8	0.11	0.56	0.11	0.53	0.10	0.49	0.11	0.65	0.11	0.55	0.12	0.60
9	0.12	0.58	0.11	0.53	0.11	0.53	0.12	0.67	0.11	0.46	0.13	0.59
High	0.14	0.69	0.12	0.59	0.10	0.50	0.13	0.61	0.15	0.48	0.18	0.59
H-L	**0.09*** **	0.85	**0.05** **	0.47	0.03	0.27	0.05	0.20	0.06	0.28	**0.10** **	0.58
Turnover	1.76		1.79		1.75		0.66		1.69		1.68	

Note: This table presents the out-of-sample portfolio performance results comparing the *Fusion Model* with Traditional Prior Return-Base Strategies for R20 horizon returns of equal-weighted portfolios using data from January 2006 through December 2023. Portfolios are formed by sorting stocks into deciles based on the respective model predictions. Ret represents annualized returns, SR denotes the Sharpe ratio. H-L shows the hedge portfolio return (High minus Low) with significance levels: * p<0.10, ** p<0.05, *** p<0.01. Turnover indicates the average monthly portfolio turnover.

Table 4 Performance Comparison for R60 Horizon of the Fusion Model with Traditional Prior Return-Based Predictors

	Equal-Weight											
	Fusion Model						Traditional Strategy					
	I5/R60		I20/R60		I60/R60		MOM/R60		STR/R60		WSTR/R60	
	Ret	SR	Ret	SR	Ret	SR	Ret	SR	Ret	SR	Ret	SR
Low	0.07	0.32	0.10	0.39	0.10	0.35	0.13	0.33	0.09	0.32	0.13	0.43
2	0.10	0.44	0.09	0.41	0.11	0.44	0.12	0.41	0.10	0.46	0.11	0.50
3	0.11	0.49	0.11	0.46	0.09	0.40	0.13	0.52	0.11	0.53	0.09	0.47
4	0.10	0.46	0.10	0.44	0.11	0.49	0.12	0.55	0.10	0.54	0.11	0.56
5	0.10	0.47	0.11	0.48	0.10	0.46	0.12	0.61	0.11	0.59	0.11	0.57
6	0.11	0.47	0.10	0.46	0.11	0.49	0.11	0.59	0.12	0.61	0.11	0.59
7	0.12	0.54	0.11	0.50	0.12	0.54	0.11	0.59	0.11	0.56	0.12	0.57
8	0.12	0.52	0.12	0.52	0.11	0.51	0.11	0.58	0.13	0.56	0.10	0.50
9	0.13	0.51	0.11	0.51	0.12	0.55	0.11	0.57	0.12	0.48	0.12	0.48
High	0.13	0.54	0.13	0.61	0.12	0.60	0.12	0.48	0.14	0.42	0.14	0.43
H-L	**0.06*** **	0.66	0.03	0.21	0.02	0.19	-0.01	-0.02	0.05	0.22	0.01	0.10
Turnover	0.59		0.61		0.61		0.38		0.57		0.57	

Note: This table presents the out-of-sample portfolio performance results comparing the Fusion Model with Traditional Prior Return-Base Strategies for R60 horizon returns of equal-weighted portfolios using data from January 2006 through December 2023. Portfolios are formed by sorting stocks into deciles based on the respective model predictions. Ret represents annualized returns, SR denotes the Sharpe ratio. H-L shows the hedge portfolio return (High minus Low) with significance levels: * p<0.10, ** p<0.05, *** p<0.01. Turnover indicates the average monthly portfolio turnover.

Table 5 Option Only Model Performance

	Equal-Weight			
	R20		R60	
	Ret	SR	Ret	SR
Low	0.06	0.30	0.08	0.40
2	0.09	0.41	0.10	0.44
3	0.09	0.42	0.11	0.48
4	0.10	0.49	0.10	0.46
5	0.10	0.48	0.11	0.49
6	0.11	0.53	0.12	0.52
7	0.11	0.56	0.10	0.45
8	0.11	0.53	0.11	0.49
9	0.12	0.57	0.12	0.50
High	0.13	0.64	0.13	0.56
H-L	**0.06*****	0.80	**0.05*****	0.75
Turnover	1.69		0.59	

Note: This table presents the out-of-sample portfolio performance results for the *Option Only Model* using data from January 2006 through December 2023. Portfolios are formed by sorting stocks into deciles based on the model predictions for R20 and R60 horizon returns of equal-weighted portfolios. Ret represents annualized returns, SR denotes the Sharpe ratio. H-L shows the hedge portfolio return (High minus Low) with significance levels: * p<0.10, ** p<0.05, *** p<0.01. Turnover indicates the average monthly portfolio turnover.

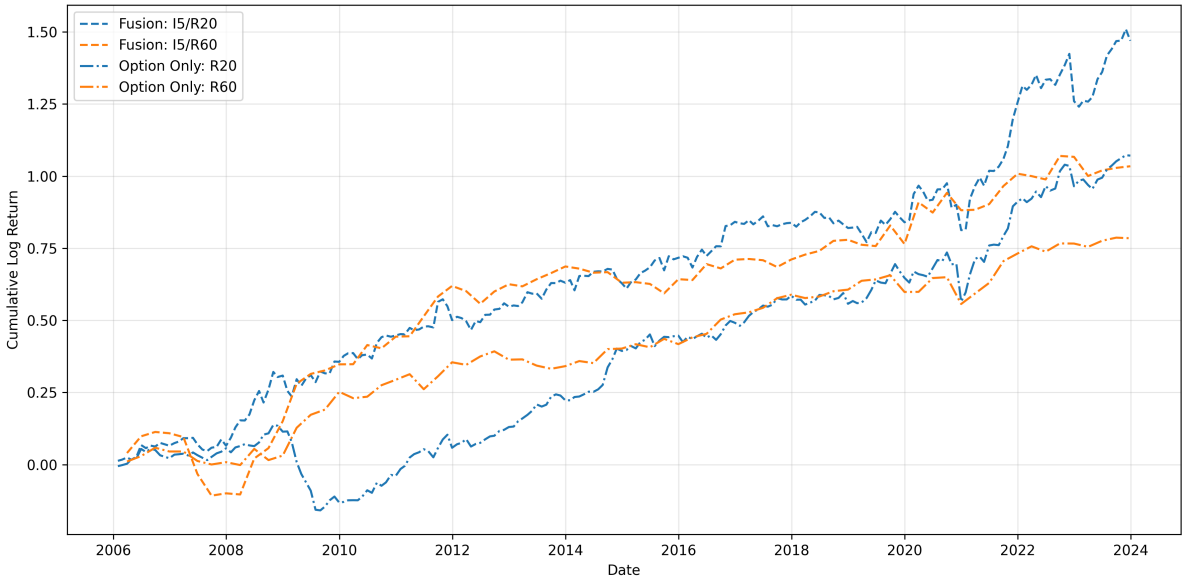


Figure 8 Fusion vs Option Only Model: H-L Portfolio Returns for R20 and R60 Horizons

Note: This figure illustrates the cumulative returns of the High-Low (H-L) portfolios formed based on the Fusion Model (I5/R20 and I5/R60) and *Option Only Model* (R20 and R60) predictions. The dot lines represent the Fusion Model, while the dashdot lines represent the *Option Only Model*. Different prediction horizons are shown in different colors.

References

Aït-Sahalia, Y. and A. W. Lo (1998). Nonparametric estimation of state-price densities implicit in financial asset prices. *The journal of finance 53*(2), 499–547.

Arevalo, J., T. Solorio, M. Montes-y Gómez, and F. A. González (2017). Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992*.

Ba, J. L., J. R. Kiros, and G. E. Hinton (2016). Layer normalization.

Bakshi, G., N. Kapadia, and D. Madan (2003). Stock return characteristics, skew laws, and the differential pricing of individual equity options. *The Review of Financial Studies 16*(1), 101–143.

Bates, D. S. (1991). The crash of '87: was it expected? the evidence from options markets. *The journal of finance 46*(3), 1009–1044.

Black, F. and M. Scholes (1973). The pricing of options and corporate liabilities. *Journal of Political Economy 81*(3), 637–654.

Blume, L., D. Easley, and M. O'hara (1994). Market statistics and technical analysis: The role of volume. *The journal of finance 49*(1), 153–181.

Brock, W., J. Lakonishok, and B. LeBaron (1992). Simple technical trading rules and the stochastic properties of stock returns. *The Journal of finance 47*(5), 1731–1764.

Brown, D. P. and R. H. Jennings (1989). On technical analysis. *The Review of Financial Studies 2*(4), 527–551.

Campa, J. M., P. K. Chang, and R. L. Reider (1998). Implied exchange rate distributions: evidence from otc option markets. *Journal of international Money and Finance 17*(1), 117–160.

Chen, L., M. Pelger, and J. Zhu (2024). Deep learning in asset pricing. *Management Science 70*(2), 714–750.

Christoffersen, P., K. Jacobs, and B. Y. Chang (2013). Forecasting with option-implied information. *Handbook of Economic Forecasting 2*, 581–656.

Cremers, M. and D. Weinbaum (2010). Deviations from put-call parity and stock return predictability. *Journal of Financial and Quantitative Analysis 45*(2), 335–367.

Diavatopoulos, D., J. S. Doran, A. Fodor, and D. R. Peterson (2012). The information content of implied skewness and kurtosis changes prior to earnings announcements for stock and option returns. *Journal of Banking & Finance 36*(3), 786–802.

Gatheral, J. (2011). *The volatility surface: a practitioner's guide*. John Wiley & Sons.

Glorot, X. and Y. Bengio (2010). Understanding the difficulty of training deep feedforward neural networks. *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics 9*, 249–256.

Gu, S., B. Kelly, and D. Xiu (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies 33*(5), 2223–2273.

Han, Y., G. Zhou, and Y. Zhu (2016). A trend factor: Any economic gains from using information over investment horizons? *Journal of Financial Economics 122*(2), 352–375.

Ioffe, S. and C. Szegedy (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning*, Volume 37, pp. 448–456.

Jackwerth, J. C. and M. Rubinstein (1996). Recovering probability distributions from option prices. *The journal of Finance 51*(5), 1611–1631.

Jacobs, R. A., M. I. Jordan, S. J. Nowlan, and G. E. Hinton (1991). Adaptive mixtures of local experts. *Neural Computation 3*(1), 79–87.

Jegadeesh, N. (1990). Evidence of predictable behavior of security returns. *The Journal of finance 45*(3), 881–898.

Jegadeesh, N. and S. Titman (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *Journal of Finance 48*(1), 65–91.

Jiang, J., B. Kelly, and D. Xiu (2023). Re-imag(in)ing price trends. *The Journal of Finance 78*(6), 3193–3249.

Kelly, B., B. Kuznetsov, S. Malamud, and T. A. Xu (2023). Deep learning from implied volatility surfaces. *Swiss Finance Institute Research Paper* (23-60).

Kingma, D. P. and J. Ba (2014). Adam: A method for stochastic optimization.

Lehmann, B. N. (1990). Fads, martingales, and market efficiency. *The Quarterly Journal of Economics 105*(1), 1–28.

Lo, A. W., H. Mamaysky, and J. Wang (2000). Foundations of technical analysis: Computational algorithms, statistical inference, and empirical implementation. *The Journal of Finance 55*(4), 1705–1765.

Merton, R. C. (1973). An intertemporal capital asset pricing model. *Econometrica: Journal of the Econometric Society*, 867–887.

Mingone, A. (2023). Smiles in delta. *Quantitative Finance 23*(12), 1713–1728.

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

Neely, C. J., D. E. Rapach, J. Tu, and G. Zhou (2014). Forecasting the equity risk premium: the role of technical indicators. *Management science 60*(7), 1772–1791.

Ofek, E., M. Richardson, and R. F. Whitelaw (2004). Limited arbitrage and short sales restrictions: Evidence from the options markets. *Journal of Financial Economics 74*(2), 305–342.

Rubinstein, M. (1985). Nonparametric tests of alternative option pricing models using all reported trades and quotes on the 30 most active cboe option classes from august 23, 1976 through august 31, 1978. *The Journal of Finance 40*(2), 455–480.

Rubinstein, M. (1994). Implied binomial trees. *The Journal of Finance 49*(3), 771–818.

Shazeer, N., A. Mirhoseini, K. Maziarz, A. Davis, Q. V. Le, G. E. Hinton, and J. Dean (2017). Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

Treynor, J. L. and R. Ferguson (1985). In defense of technical analysis. *The Journal of Finance 40*(3), 757–773.

Tsai, Y.-H. H., S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov (2019). Multimodal transformer for unaligned multimodal language sequences. *arXiv preprint arXiv:1906.00295*.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin (2017). Attention is all you need. *Advances in neural information processing systems 30*.

Xing, Y., X. Zhang, and R. Zhao (2010). What does individual option volatility smirk tell us about future equity returns? *Journal of Financial and Quantitative Analysis 45*(3), 641–662.

Appendix

A.1 Convolutional Neural Network Architecture

Convolutional Neural Networks (CNNs) are a core methodological tool in this study for extracting nonlinear features from both stock price images and option-implied volatility (IV) images. This section systematically introduces the CNN architecture, progressing from its fundamental building blocks to the overall network structure.

A CNN extracts hierarchical representations from an input image through a sequence of specialized operations, ultimately mapping these representations to predictive outputs. In practice, CNNs are composed of a series of modules that share a common structural design but differ in their hyperparameters, such as filter sizes and channel dimensions. Each module typically consists of three components: a convolutional layer, a nonlinear activation layer, and a pooling layer.

The first component is the convolutional layer, which applies a set of learnable filters that slide across the input image to capture local spatial patterns. These filters enable the network to detect salient features such as edges, shapes, and texture-like structures in early layers, and more abstract patterns in deeper layers. The mathematical principles and operational procedures of the convolutional layer are illustrated in Figure A.1 and explained as follows.

In Figure A.1, the input image is represented as a 6×6 matrix and the convolutional filter as a 3×3 matrix. The filter slides over the input image from the top-left to the bottom-right with a stride of one. At each location, element-wise multiplication is applied between the filter and the corresponding 3×3 subregion of the input image, and the resulting values are summed to produce a single output.

For example, at the initial position, the filter covers the top-left 3×3 region of the input image, yielding an output value of zero after multiplication and summation. As the filter

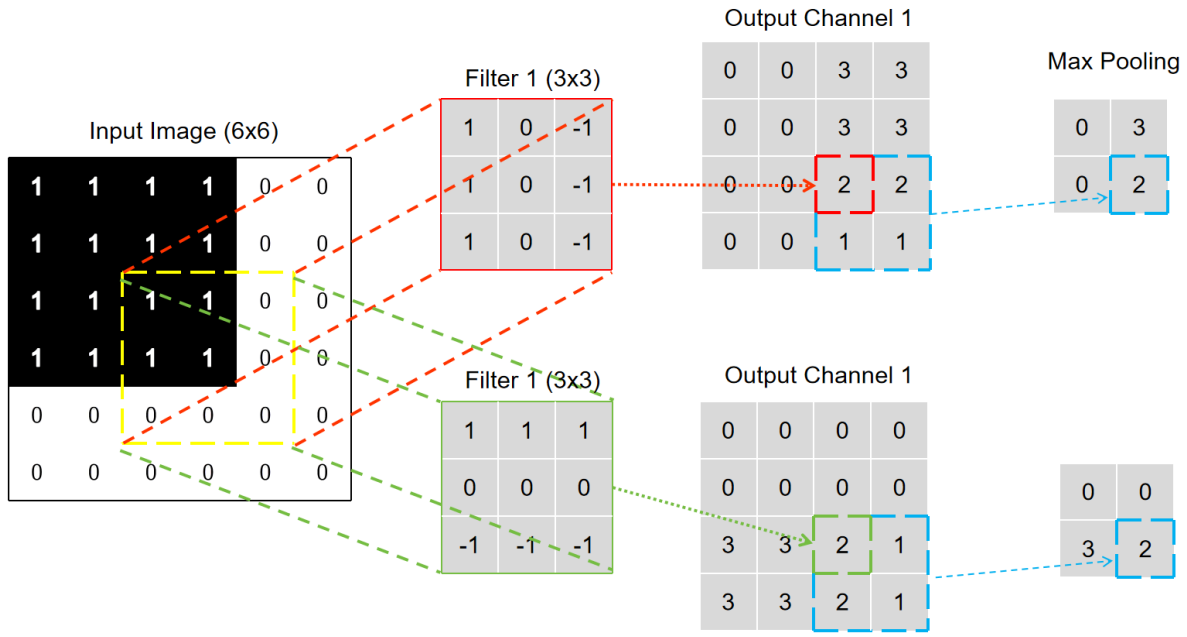


Figure A.1 Convolutional Operation

continues to move across the image, this operation generates a 4×4 output feature map that captures the local spatial patterns detected by the filter.

Also, the channel dimension is important in convolutional layers. For grayscale images, there is only one channel, while RGB images have three channels (red, green, blue). In the example of Figure A.1, the input image has one channel, and the filter also has one channel. If the input image had multiple channels, the filter would have the same number of input channels, and the convolution operation would involve summing over all channels. The output channel dimension is determined by the number of filters used in the convolutional layer and can be adjusted based on the desired feature representation. In this example, two filters are applied, resulting in an output with two channels. Filter 1 captures horizontal edges, while Filter 2 captures vertical edges. In our study, we set the number of output channels to 64 for the first convolutional layer, which allows the model to learn a diverse set of features from the input images.

Convolution operations offer flexibility in adjusting the "stride" of the filter, which refers to the number of pixels the filter shifts horizontally or vertically as it slides across the input matrix.

Specifically, a larger stride results in a coarser-grained convolution output, as the filter covers fewer local regions of the input. Following the setting of Jiang et al. (2023), our model adopts a horizontal stride of 1 and a vertical stride of 3: horizontally, the filter moves one pixel at a time, recalculating the feature value for every position along each row; vertically, the filter skips three rows between consecutive computations, reducing redundant information capture in the vertical dimension while preserving key structural features.

To expand the receptive field of the filter without increasing its size, we occasionally employ "dilated filters" in feature extraction. A dilated filter with a dilation rate k along the vertical or horizontal dimension (or both) extends its effective coverage by inserting $k-1$ zeros between adjacent elements of the original filter along the corresponding dimensions. This design allows the filter to capture multi-scale contextual information from the image without enlarging its parameter count. Consistent with the model of Jiang et al. (2023), our CNN architecture uses a vertical dilation rate of 2 and no horizontal dilation—this configuration enhances the model's ability to detect long-range dependencies in the vertical dimension while maintaining precise capture of sequential information in the horizontal dimension.

The second layer is the activation layer, which introduces nonlinearity into the model. We use LeakyReLU as the activation function, defined as:

$$\text{LeakyReLU}(x) = \begin{cases} x, & x > 0, \\ \alpha x, & x \leq 0, \end{cases}$$

where α is a small constant (typically 0.01) that allows a small, non-zero gradient when the unit is not active. This helps prevent the "dying ReLU" problem, where neurons become inactive and stop learning.

The third layer is the pooling layer, which reduces the spatial dimensions of the feature maps while retaining the most important information. We use max pooling, which takes the

maximum value from each non-overlapping sub-region of the feature map. This downsampling operation helps to reduce computational complexity and exempt small noise in the input image because only the most prominent features are retained.

A key methodological choice that is worth explicitly stating, is that our CNN architecture retains strict consistency with the baseline of Jiang et al. (2023) except for necessary adaptations to option IV images. This design choice ensures that any performance improvement in our multimodal model can be attributed to the addition of option IV information (the core innovation of this study) rather than architectural modifications. Specifically:

- Stock Image CNN: For 5/20/60-day images (32×15, 64×60, 96×180), we use 2-4 convolution-pooling blocks (more blocks for longer horizon images to capture long-term trends), each with 5×3 filters (64 channels), and LeakyReLU activation.
- Option-Implied Volatility Image CNN: For fixed 32×34 images, we use 2 convolution-pooling blocks (matching the volatility surface's simpler grid structure) with identical filter size, stride, and dilation settings.

After passing through several convolutional, activation, and pooling layers, the extracted feature maps are flattened into a one-dimensional vector. This vector is then fed into cross-attention modules for multimodal fusion and finally into fully connected layers for return prediction.

A.2 Cross-Attention

Cross-attention is the core mechanism enabling dynamic interaction between stock price image features and option IV image features in this study. Unlike self-attention, which models dependencies within a single modality, cross-attention facilitates directional information exchange across two distinct modalities, allowing stock historical trend features to "query" forward-looking risk signals from options, and vice versa. This section systematically explains the working principle of cross-attention, starting with scaled dot-product attention and extend-

ing to the multi-head attention design, while linking each step to the study's specific multimodal fusion objectives.

A.2.1 Scaled Dot-Product Attention

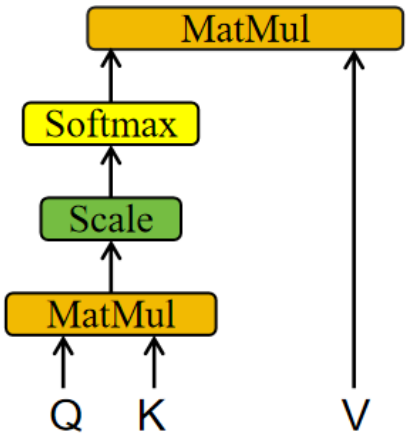


Figure A.2 Scaled Dot Product

Scaled dot-product attention (Figure A.2) is the basic building block of cross-attention, responsible for computing the relevance between "query" features from one modality and "key-value" features from the other modality. As illustrated in Figure A.2, the cirtical input consists of three vectors: Query(Q), Key(K), and Value(V) and the computation follows three squential steps.

The first step is to calculate the similarity between Q and the transpose of K using the dot product operation to measure the revelance between each element in Q and each element in K. For 64-dimensional features, this yields a 64x64 similarity matrix, where entry (i, j) represents the correlation between the i-th dimension of Q and the j-th dimension of K. The second step is to divide the similarity matrix by the square root of the feature dimension (i.e., $\sqrt{64} = 8$) of Q and K, which mitigates the gradient vanishing problem caused by large dot-product values when d is large. The third step is to apply the softmax function to the scaled similarity matrix to obtain normalized attention weights, which sum to 1 per row, then multiply these weights

by V to generate the attention weighted value vector. This vector contains the most revelant information from V aligned with Q. Mathematically, the operation is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V.$$

A.2.2 Multi-Head Attention

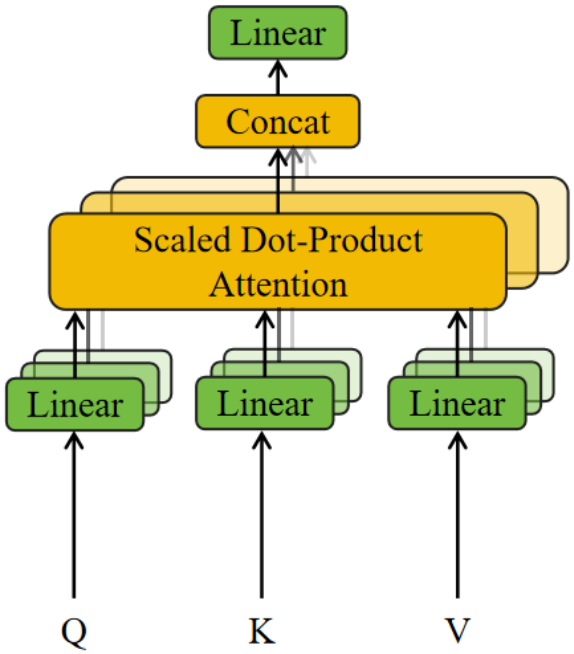


Figure A.3 Multi-Head Attention

While scaled dot-product attention captures basic cross-modal revelance, multi-head attention (Figure A.3) extends this capability by modeling interactions across multiple "subspaces" of the feature vectors, which is essential for financial data, where stock and option features may exhibit complementary patterns in distinct subspaces.

As shown in Figure A.3, the multi-head attention mechanism involves four key steps. First, Q, K, and V are projected into H independent subspaces via learnable linear layers. Second, each of the H head-specific Q_h, K_h, V_h (8-dimensional) undergoes independent scaled dot-product attention.

This parallel computation allows the model to capture 8 distinct types of cross-modal interactions simultaneously. Third, the H attention outputs are concatenated along the dimension axis to reconstruct the vector, merging the diverse cross-modal patterns captured by individual heads. Finally, a final learnable linear layer transforms the concatenated vector into an output to be able to fitted into subsequent network layers. Formally, the multi-head attention operation is expressed as:

$$\text{MultiHead}(Q, K, V) = \text{Linear}\left(\text{Concat}\left(\text{head}_1, \ldots, \text{head}_H\right)\right),$$

$$\text{where} \quad \text{head}_h = \text{Attention}\left(QW_h^Q, KW_h^K, VW_h^V\right),$$

W_h^Q, W_h^K, W_h^V denote the projection weights for the h-th head.

To stabilize training and preserve feature integrity, we add residual connections (Vaswani et al., 2017) and layer normalization (Ba et al., 2016) after the final linear projection.

We now discuss our choice of hyperparameters in the multi-head attention module. We set the number of attention heads H to 8, which balances model capacity and computational efficiency. Each head processes 8-dimensional subspaces, allowing the model to capture diverse cross-modal patterns without excessive complexity.