Article

Neural Regime-Switching Model with Markov-Informed Attention for Short-Horizon Equity Return Forecasting

T.C. Iroko [1,*], A.V. Alagbada [2] and S. Tchoneteck [3]

[1] Department of Mathematics, University of Wisconsin–Milwaukee, Milwaukee, WI 53201, USA
[2] Faculty of Civil and Environmental Engineering, Bauhaus-Universität Weimar, 99423 Weimar, Germany
[3] Department of Statistics, Rice University, Houston, TX 77005, USA
* Correspondence: tciroko@uwm.edu

Abstract

Short-horizon equity return forecasting presents fundamental challenges due to regime shifts and nonlinear market dynamics. Traditional approaches either employ interpretable probabilistic models lacking predictive power or deploy opaque deep learning architectures sacrificing economic interpretability. We propose the Neural Regime-Switching Model with Markov-Informed Attention (NRSM-MIA), an end-to-end differentiable architecture that jointly learns discrete market regimes, Markov transition dynamics, and regime-conditioned temporal patterns. The framework introduces differentiable regime inference via Gumbel-Softmax relaxation, regime-conditioned multi-head attention, a mixture-of-experts decoder, and built-in uncertainty quantification. Following cross-sectional learning, we train on pooled data from 15 U.S. equities across five sectors. Experiments demonstrate that NRSM-MIA achieves the best Mean Absolute Error (1.1104), Directional Accuracy (52.86%), and Sharpe Ratio (0.6765) compared to LSTM, Transformer, and HMM-GBR baselines. Crucially, the model maintains interpretable regime assignments with a balanced distribution across volatility states, validating that prediction performance and interpretability can be achieved simultaneously through staged training with diversity regularization.

Keywords: regime-switching models; Gumbel-Softmax; attention mechanisms; equity forecasting; mixture-of-experts; cross-sectional learning

---

1. Introduction

Financial markets do not behave uniformly across time. Periods of sustained growth give way to sudden corrections, and tranquil trading environments can rapidly transform into volatile crises. Such behavior reflects the non-stationary nature of financial time series, where statistical properties such as mean returns, volatility, and temporal dependencies evolve across distinct periods. This poses a fundamental challenge for quantitative forecasting models that assume fixed relationships between features and returns. Consequently, there is a need for forecasting frameworks that explicitly model regime-dependent dynamics while retaining the predictive flexibility of modern deep learning methods.

The Efficient Market Hypothesis (EMH) posits that asset prices reflect available information, implying that excess returns are unpredictable in expectation [1]. Importantly, however, EMH places no restriction on the stationarity of return distributions or on the temporal stability of higher-order statistical properties. A substantial empirical literature documents time-varying volatility, changing dependence structures, and persistent shifts in market dynamics, even in the absence of long-run predictability [2,3]. These structural changes are commonly modeled as regime transitions, in which distinct market states are characterized by different return distributions, volatility levels, and corre-

lation patterns [4]. Such regimes do not imply persistent inefficiencies, but rather reflect evolving conditional dynamics that are central to short-horizon forecasting.

Traditional approaches to regime-aware forecasting have relied on Hidden Markov Models (HMMs), which provide an elegant probabilistic framework for modeling latent market states [5]. HMMs decompose observed return sequences into emissions from discrete hidden states, with transitions governed by a Markov process. This framework offers interpretable regime assignments and has been applied to volatility modeling [6], asset allocation [7], and market timing [8]. However, HMMs suffer from two critical limitations. The Expectation-Maximization algorithm used for parameter estimation operates separately from downstream prediction tasks, potentially leading to suboptimal regime definitions. Additionally, linear emission distributions inadequately capture the complex, nonlinear relationships between market features and future returns.

To address these predictive limitations, researchers have proposed hybrid architectures combining regime identification with machine learning predictors [9,10]. A representative approach is the two-stage HMM plus Gradient Boosting Regressor (GBR) pipeline, where an HMM first identifies market regimes, and these regime labels are appended as features to a gradient boosting model. While such hybrids improve predictive accuracy, they suffer from a fundamental disconnect: the regime inference and prediction stages are optimized independently, with no mechanism for prediction loss to influence regime definitions.

The current landscape reveals a significant research gap: the absence of an end-to-end differentiable framework that jointly optimizes regime inference and return prediction within a unified architecture. Recent advances in deep learning have introduced techniques for handling discrete latent variables through continuous relaxations [11,12], opening new possibilities for neural regime-switching models. To our knowledge, no existing work combines differentiable discrete regime inference with attention-based temporal modeling for equity return forecasting while maintaining interpretable regime structures.

This paper introduces NRSM-MIA, a neural regime-switching model with Markov-informed attention that addresses these limitations through four key innovations. First, we employ Gumbel-Softmax relaxation for differentiable regime inference with a learnable transition matrix, enabling joint optimization of latent states and predictive accuracy. Second, we design regime-conditioned attention where temporal query projections adapt dynamically to inferred regimes, capturing state-specific dependencies. Third, we implement a mixture-of-experts decoder that allocates specialized subnetworks to distinct market conditions, combined through soft regime probabilities. Fourth, we introduce staged training with diversity regularization that achieves both prediction performance and regime interpretability, resolving the typical trade-off between these objectives.

We train our model cross-sectionally on pooled data from 15 U.S. equities across five sectors, enabling transferable regime structures across assets. Our experiments demonstrate that NRSM-MIA achieves the best Mean Absolute Error, Directional Accuracy, and Sharpe Ratio among all baselines while maintaining a balanced regime distribution of approximately 19%, 47%, and 34% across low, medium, and high volatility states. These results validate that end-to-end differentiable regime modeling can simultaneously improve prediction and preserve interpretability.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 presents the NRSM-MIA architecture. Section 4 describes our experimental setup. Section 5 presents results and analysis. Section 6 concludes with future directions.

2. Related Work

This section reviews the relevant literature across four areas: regime-switching models in finance, machine learning approaches for return forecasting, techniques for handling discrete latent variables, and attention mechanisms for financial time series.

2.1. Regime-Switching Models in Finance

The observation that financial markets exhibit distinct behavioral patterns across different periods has motivated extensive research into regime-switching models. Hamilton [4] pioneered this field by introducing the Markov-switching autoregressive model to capture business cycle dynamics, demonstrating that economic variables could be modeled as arising from distinct expansion and contraction states. Hidden Markov Models generalize this framework by allowing flexible observation distributions conditional on latent states [5]. In finance, HMMs have been applied to model stock price dynamics [13], characterize volatility regimes [14], and assess credit risk [15].

More recent work has extended basic HMM frameworks to capture richer dynamics. Guidolin [7] provides a comprehensive survey of Markov-switching models in empirical finance. Oelschläger and Adam [16] proposed hierarchical HMMs that distinguish between slow-changing macroeconomic regimes and fast-changing local market states. Wang and Lin [8] applied regime-switching frameworks to factor investing, showing that regime-aware strategies can improve risk-adjusted returns. Despite their interpretability, standalone HMMs face fundamental limitations for return prediction, as the EM algorithm optimizes a likelihood objective agnostic to downstream tasks, and Gaussian emission distributions cannot capture complex nonlinear relationships.

2.2. Machine Learning for Financial Forecasting

Machine learning methods have achieved notable success in financial forecasting by capturing nonlinear patterns that elude traditional econometric models. Gu et al. [17] conducted a comprehensive comparison of machine learning methods for cross-sectional return prediction, finding that tree-based methods and neural networks substantially outperform linear models. They demonstrated that pooling data across stocks enables models to learn patterns that generalize across the market, establishing the cross-sectional learning paradigm we adopt.

Gradient boosting methods, particularly XGBoost [18] and LightGBM, have become popular due to their ability to handle heterogeneous features and model complex interactions. Deep learning approaches offer automatic feature learning from sequential data. Long Short-Term Memory networks [19] address the vanishing gradient problem in recurrent architectures. Fischer and Krauss [20] demonstrated that LSTMs can outperform traditional methods for daily return prediction. The Transformer architecture [21], originally developed for natural language processing, has been adapted for time series forecasting. Lim et al. [22] introduced the Temporal Fusion Transformer combining recurrent processing with attention mechanisms. However, these deep learning approaches typically treat market dynamics as stationary, ignoring regime-switching behavior.

2.3. Hybrid Regime-Aware Approaches

Recognizing the complementary strengths of interpretable regime models and flexible machine learning predictors, a variety of hybrid architectures have been proposed. Early studies used HMM-identified regimes to guide model selection or parameterization, such as selecting regime-specific ARIMA models [23] or incorporating fuzzy logic to design regime-dependent prediction rules [9]. More recent work integrates regime-switching mechanisms with neural networks. Ilhan and Kozat [24] introduced a Markovian recurrent neural network in which hidden state evolution is governed by a learned Markov process, while Yuan et al. [25] proposed TFE-HMM, combining HMM state transitions with price trend features. A widely adopted practical strategy follows a two-stage pipeline: an HMM is first fitted for regime identification, after which regime labels are appended as features to downstream predictors such as gradient boosting regressors [10,26].

While these hybrid approaches improve upon regime-agnostic models by incorporating structural information, the regime identification and prediction stages are typically optimized independently. In practice, regimes inferred solely through likelihood-based objectives may not align with features most relevant for forecasting performance. As a result, prediction errors cannot influence regime definitions, and mis-specified regimes propagate directly into the prediction stage without corrective feedback. This

separation limits adaptability under evolving market dynamics and constrains the joint optimization of regime structure and predictive accuracy. Our work addresses this limitation through end-to-end differentiable training, enabling regime inference and prediction to be learned simultaneously.

2.4. Differentiable Discrete Latent Variables

A key technical challenge in neural regime-switching models is the non-differentiability of discrete state assignments, which prevents gradient-based optimization through regime selection. The Gumbel–Softmax distribution [11,12] provides a continuous relaxation of categorical distributions that enables low-variance gradient estimation while retaining an explicit discrete interpretation at inference time. Given logits $\ell \in \mathbb{R}^K$ for K categories, the Gumbel–Softmax sample allows gradients to flow through discrete choices during training.

While originally popularized in applications such as variational autoencoders with discrete latents [27], neural architecture search [28], and structured prediction [29], the Gumbel–Softmax relaxation is particularly well-suited for financial time series with regime-dependent dynamics. Financial regimes are commonly interpreted as discrete latent states (e.g., low-, medium-, and high-volatility periods) that persist over time and exhibit distinct statistical characteristics. Unlike fully continuous latent representations, discrete regime assignments preserve interpretability and align naturally with economic intuition, while the continuous relaxation enables end-to-end optimization of regime inference and prediction objectives. This combination allows regime definitions to adapt directly in response to forecasting error, addressing a key limitation of two-stage regime identification pipelines.

As a continuous relaxation, Gumbel–Softmax introduces a trade-off between optimization smoothness and discrete regime separation; we address this through a staged training strategy that ensures stable learning and interpretable regime assignments.

2.5. Research Gap

Table 1 summarizes the positioning of our work relative to existing approaches. The literature review reveals that, to the best of our knowledge, there is no existing work that combines end-to-end differentiable regime inference, learnable Markov transition dynamics, regime-conditioned attention mechanisms, and mixture-of-experts prediction with uncertainty quantification for equity return forecasting. Our proposed NRSM-MIA architecture addresses this gap.

Table 1. Comparison with Existing Approaches.

Approach	E2E	Markov	Attn	MoE	UQ
HMM + GBR [26]	✗	✓	✗	✗	✗
LSTM [20]	✓	✗	✗	✗	✗
Transformer [21]	✓	✗	✓	✗	✗
Markovian RNN [24]	✓	✓	✗	✗	✗
TFE-HMM [25]	✗	✓	✗	✗	✗
NRSM-MIA (Ours)	✓	✓	✓	✓	✓

E2E: End-to-end differentiable; Markov: Learnable transitions; Attn: Regime-conditioned attention; MoE: Mixture-of-experts; UQ: Uncertainty quantification.

3. Methodology

This section presents the Neural Regime-Switching Model with Markov-Informed Attention. We describe the problem formulation, then detail each architectural component with mathematical specifications.

3.1. Problem Formulation

Given daily closing prices $P_1, P_2, \ldots, P_t$ for a stock, we compute the daily percentage return as $R_t = 100 \times (P_t - P_{t-1})/P_{t-1}$. The forecasting task is to predict R_{t+1} given information available up to time t. Rather than using raw prices, we construct a feature vector $\mathbf{x}_t \in \mathbb{R}^d$ containing d technical indicators computed from historical data. These indicators are used as compact observables of market

state, capturing changes in trend strength, volatility, and mean-reversion that vary across market regimes. To capture temporal patterns, we use a lookback window of L trading days, forming the input matrix $\mathbf{X}_t \in \mathbb{R}^{L \times d}$.

Following the cross-sectional learning paradigm of Gu et al. [17], we train a universal model on pooled data from multiple stocks. To enable differentiation between stocks and sectors, we include one-hot encoded sector indicators $\mathbf{s}_i \in \{0,1\}^5$ representing Technology, Financials, Healthcare, Consumer, and Energy sectors. A key assumption underlying our approach is that market dynamics are governed by latent regimes $z_t \in \{1, 2, \ldots, K\}$ representing unobserved states that characterize distinct market conditions. The model learns these regime definitions from data rather than using predefined rules.

Given input $\mathbf{X}_t$, our model produces three outputs: the return prediction $\hat{R}_{t+1}$, an uncertainty estimate $\hat{\sigma}_{t+1}$, and regime probabilities $\mathbf{z}_t \in [0,1]^K$ indicating the probability of being in each regime. Table 2 summarizes the key notation.

Table 2. Summary of Notation.

Symbol	Description
R_t	Daily percentage return at time t
$\mathbf{x}_t \in \mathbb{R}^d$	Feature vector at time t (d features)
$\mathbf{X}_t \in \mathbb{R}^{L \times d}$	Input matrix (lookback window of L days)
$z_t \in \{1, \ldots, K\}$	Discrete regime indicator
$\mathbf{z}_t \in [0,1]^K$	Soft regime probabilities
$\mathbf{A} \in \mathbb{R}^{K \times K}$	Regime transition probability matrix
$\tau > 0$	Gumbel-Softmax temperature
H	Hidden dimension

3.2. Architecture Overview

NRSM-MIA processes input data through four main components as illustrated in Figure 1. The Feature Encoder transforms raw features into rich representations capturing temporal patterns. The Regime Inference Module identifies the current market regime in a differentiable manner. The Regime-Conditioned Attention focuses on relevant historical information with attention patterns adapted based on the inferred regime. The Mixture-of-Experts Decoder makes predictions using regime-specific expert networks.

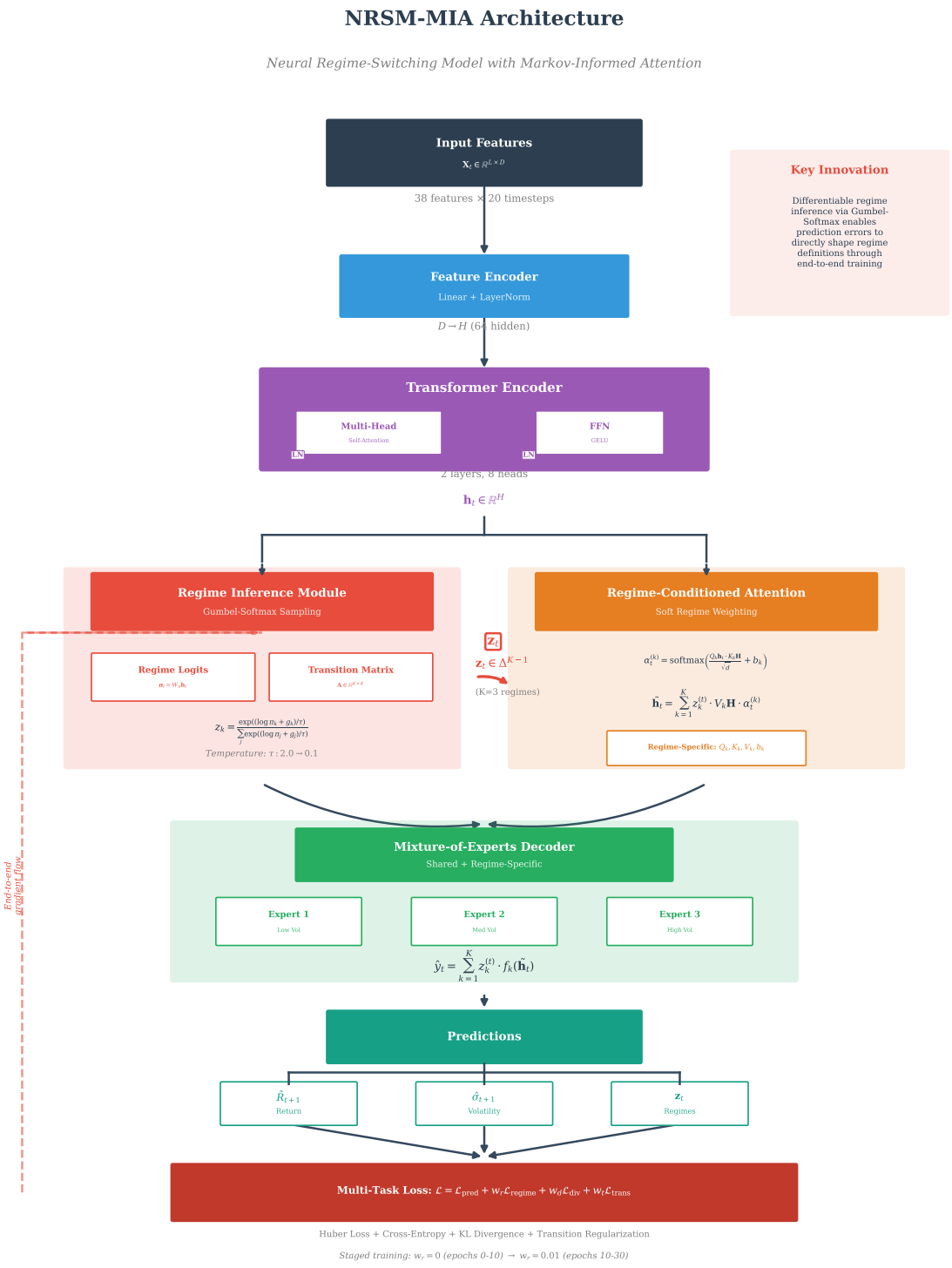


Figure 1. NRSM-MIA architecture overview. The key innovation is differentiable regime inference, allowing prediction errors to directly influence regime definitions through end-to-end training.

3.3. Feature Encoder

The feature encoder transforms the input matrix $\mathbf{X}_t \in \mathbb{R}^{L \times d}$ into hidden representations $\mathbf{H}_t \in \mathbb{R}^{L \times H}$. Each day's feature vector is first projected to the hidden dimension through a linear layer followed by layer normalization and GELU activation:

$$\tilde{\mathbf{x}}_i = \text{GELU}(\text{LayerNorm}(\mathbf{x}_i \mathbf{W}_{\text{in}} + \mathbf{b}_{\text{in}})), \tag{1}$$

where $\mathbf{W}_{\text{in}} \in \mathbb{R}^{d \times H}$ and $\mathbf{b}_{\text{in}} \in \mathbb{R}^H$ are learnable parameters. Since Transformer layers are permutation-invariant, we add learnable positional embeddings $\mathbf{e}_i \in \mathbb{R}^H$ for each position. The sequence then passes through N_{enc} Transformer encoder layers, each consisting of multi-head self-attention and position-wise feed-forward networks with residual connections [21].

3.4. Regime Inference Module

The regime inference module determines the current market regime through differentiable operations. From the encoded representation, we compute unnormalized scores (logits) for each regime using a multi-layer perceptron:

$$\boldsymbol{\ell}_t = \text{MLP}_{\text{regime}}(\mathbf{h}_t^{(L)}), \tag{2}$$

where $\mathbf{h}_t^{(L)} \in \mathbb{R}^H$ is the hidden representation at the final time step. Market regimes tend to persist, so we incorporate this through a learnable transition matrix $\mathbf{A} \in \mathbb{R}^{K \times K}$:

$$A_{ij} = \frac{\exp(\tilde{A}_{ij})}{\sum_{k=1}^{K} \exp(\tilde{A}_{ik})}, \tag{3}$$

where $\tilde{\mathbf{A}}$ is a learnable log-transition matrix. Given the previous regime distribution $\mathbf{z}_{t-1}$, the prior for the current regime is $\boldsymbol{\pi}_t = \mathbf{z}_{t-1}^\top \mathbf{A}$. We incorporate this prior by adjusting the logits: $\boldsymbol{\ell}_t \leftarrow \boldsymbol{\ell}_t + \log(\boldsymbol{\pi}_t + \epsilon)$.

We apply the Gumbel-Softmax [11] to obtain differentiable regime probabilities:

$$z_{t,k} = \frac{\exp((\ell_{t,k} + g_k)/\tau)}{\sum_{j=1}^{K} \exp\big((\ell_{t,j} + g_j)/\tau\big)}, \tag{4}$$

where $g_k = -\log(-\log(u_k))$ with $u_k \sim \text{Uniform}(0, 1)$ are Gumbel noise samples, and $\tau > 0$ is the temperature parameter. During training, we anneal the temperature from $\tau_0 = 2.0$ to $\tau_{\min} = 0.1$ using decay rate $\gamma = 0.99$.

3.5. Regime-Conditioned Attention

Different market regimes may require focusing on different aspects of historical data. Our regime-conditioned attention adapts its focus based on the inferred regime through regime-specific bias terms. We compute base projections using shared linear transformations:

$$\mathbf{q}_{\text{base}} = \mathbf{h}_t^{(L)} \mathbf{W}_Q, \tag{5}$$

$$\mathbf{K}_{\text{base}} = \mathbf{H}_t \mathbf{W}_K, \tag{6}$$

$$\mathbf{V}_{\text{base}} = \mathbf{H}_t \mathbf{W}_V. \tag{7}$$

Regime-specific adjustments are applied through learnable bias parameters $\mathbf{b}_Q^{(k)}, \mathbf{b}_K^{(k)}, \mathbf{b}_V^{(k)} \in \mathbb{R}^H$ for each regime k. The final projections combine base computations with regime-weighted biases:

$$\mathbf{q} = \mathbf{q}_{\text{base}} + \sum_{k=1}^{K} z_{t,k} \cdot \mathbf{b}_Q^{(k)}. \tag{8}$$

This vectorized approach is computationally efficient, requiring only a single matrix multiplication for the base projection plus a lightweight bias adjustment, rather than K separate projections. The attention mechanism then computes:

$$\boldsymbol{\alpha} = \text{softmax}\left(\frac{\mathbf{q}\mathbf{K}^\top}{\sqrt{H}}\right), \tag{9}$$

and the attended representation is $\mathbf{h}_{\text{attn}} = \text{LayerNorm}(\mathbf{h}_t^{(L)} + \boldsymbol{\alpha}\mathbf{V})$.

3.6. Mixture-of-Experts Decoder

The mixture-of-experts decoder uses regime-specific expert networks for prediction. Rather than implementing K separate networks, we use a shared computation followed by regime-specific output projections for efficiency:

$$\mathbf{h}_{\text{shared}} = \text{GELU}(\text{GELU}(\mathbf{h}_{\text{attn}}\mathbf{W}_1)\mathbf{W}_2), \tag{10}$$

where $\mathbf{W}_1 \in \mathbb{R}^{H\times H}$ and $\mathbf{W}_2 \in \mathbb{R}^{H\times H/2}$. All expert outputs are computed simultaneously through a single linear layer $\mathbf{W}_{\text{exp}} \in \mathbb{R}^{H/2\times K}$, and the final prediction is a weighted combination:

$$\hat{R}_{t+1} = \sum_{k=1}^{K} z_{t,k} \cdot (\mathbf{h}_{\text{shared}}\mathbf{W}_{\text{exp}})_k. \tag{11}$$

We estimate prediction uncertainty from two sources. Aleatoric uncertainty is computed through a dedicated head: $\hat{\sigma}^2_{\text{pred}} = \text{Softplus}(\text{MLP}_\sigma(\mathbf{h}_{\text{attn}}))$. Epistemic uncertainty is captured through regime entropy: $\mathcal{H}(\mathbf{z}_t) = -\sum_{k=1}^{K} z_{t,k} \log(z_{t,k} + \epsilon)$. The scaling factor of 0.1 is chosen to normalize the entropy term to the scale of aleatoric uncertainty, providing a stable combination that was found to be robust across experiments. The total uncertainty combines both: $\hat{\sigma}^2_{t+1} = \hat{\sigma}^2_{\text{pred}} + 0.1 \cdot \mathcal{H}(\mathbf{z}_t)$.

3.7. Training Objective

The model is trained end-to-end using a composite loss function. The primary prediction loss uses Huber loss for robustness to outliers:

$$\mathcal{L}_{\text{pred}} = \frac{1}{N}\sum_{i=1}^{N} \text{Huber}_\delta(R^{(i)}_{t+1}, \hat{R}^{(i)}_{t+1}), \tag{12}$$

with $\delta = 1.0$. We supervise regime inference using volatility-based labels computed from training data. Samples are assigned to three regimes based on rolling volatility percentiles: low (below 33rd percentile), medium (33rd to 67th), and high (above 67th). The regime supervision loss is:

$$\mathcal{L}_{\text{regime}} = \text{CrossEntropy}(\boldsymbol{\ell}_t, y_{\text{vol}}), \tag{13}$$

where y_{vol} is the volatility label.

A critical innovation in our training procedure is the diversity loss that prevents regime collapse:

$$\mathcal{L}_{\text{div}} = \text{KL}(\bar{\mathbf{z}} \| \mathbf{u}), \tag{14}$$

where $\bar{\mathbf{z}} = \frac{1}{N}\sum_i \mathbf{z}^{(i)}_t$ is the batch-averaged regime distribution and $\mathbf{u} = [1/K, \ldots, 1/K]$ is the uniform distribution. We also regularize the transition matrix toward moderate persistence:

$$\mathcal{L}_{\text{trans}} = \|\mathbf{A} - \mathbf{A}_{\text{target}}\|^2_F, \tag{15}$$

where $\mathbf{A}_{\text{target}}$ has 0.7 on the diagonal and 0.15 off-diagonal.

3.8. Staged Training Protocol

We introduce a staged training protocol that resolves the trade-off between prediction performance and regime interpretability. The regime supervision weight w_{regime} varies by epoch:

$$w_{\text{regime}}(e) = \begin{cases} 0 & \text{if } e < 10 \\ 0.01 \cdot \frac{e-10}{20} & \text{if } 10 \le e < 30 \\ 0.01 & \text{if } e \ge 30 \end{cases} \tag{16}$$

This protocol ensures that the model first learns strong prediction capabilities without regime interference, then gradually incorporates regime supervision. The diversity loss weight scales proportionally with w_{regime}, preventing collapse only after the model has established predictive foundations. The total loss is:

$$\mathcal{L} = \mathcal{L}_{\text{pred}} + w_{\text{regime}}\mathcal{L}_{\text{regime}} + 0.001\mathcal{L}_{\text{trans}} + 0.03 \cdot \frac{w_{\text{regime}}}{0.01}\mathcal{L}_{\text{div}}. \tag{17}$$

4. Experimental Setup

4.1. Data Description

We evaluated NRSM-MIA on daily stock data from 15 U.S. equities spanning five sectors as shown in Table 3. The selected equities were chosen to provide consistent long-horizon data availability, sustained trading activity, and sectoral diversity, ensuring representative market behavior across different economic conditions. The dataset covers January 2010 to December 2025, capturing diverse market conditions including the post-financial-crisis recovery, the 2015-2016 volatility, the COVID-19 crash and recovery, and subsequent inflationary periods. Data are obtained from Yahoo Finance.

Table 3. Stock Universe.

Sector	Tickers
Technology	AAPL, MSFT, GOOGL
Financials	JPM, GS, BAC
Healthcare	JNJ, PFE, UNH
Consumer	AMZN, WMT, PG
Energy	XOM, CVX, COP

4.2. Data Splitting and Preprocessing

We employ a chronological 70/10/20 split for training, validation, and test sets respectively. This splitting is applied to each stock independently before pooling, ensuring no future information leakage. All features are standardized using training set statistics only. For the pooled training set, the volatility labels are computed using global percentile thresholds across all stocks, ensuring consistent regime definitions. The resulting training data exhibits a balanced regime distribution of approximately 33% for each volatility class.

4.3. Feature Engineering

For each trading day, we construct a 38-dimensional feature vector comprising five sector indicator variables, lagged returns at 1, 2, 3, 5, and 10 days, rolling volatility measures at 5, 10, and 20-day windows, momentum indicators including rate of change and moving average ratios, mean-reversion signals including RSI-14 and Bollinger Band position, MACD and its signal line and histogram, volume features including volume ratio and OBV trend, and higher moments including rolling skewness and kurtosis at 20 days. All features are computed causally using only past data.

4.4. Baseline Models

We compare NRSM-MIA against five baseline models. The Random Walk baseline predicts zero return, representing the martingale hypothesis. The Gradient Boosting Regressor (GBR) uses 100 estimators with maximum depth 5 and learning rate 0.1. The HMM plus GBR pipeline fits a Gaussian HMM with 3 states for regime identification, then appends regime labels as features to a GBR for prediction. The LSTM baseline uses two layers with 32 hidden units and dropout 0.1. The Transformer baseline uses a standard encoder with 2 layers, 64 hidden dimensions, 8 attention heads, and dropout 0.1. All neural baselines are trained on the same pooled data with identical splits.

4.5. Evaluation Metrics

We evaluate models using prediction accuracy metrics including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), coefficient of determination (R^2), Theil's U_2 statistic comparing to random walk, and Directional Accuracy (DA) measuring the percentage of correctly predicted return signs. For trading performance, we compute the annualized Sharpe Ratio and Maximum Drawdown based on a simple strategy that goes long when $\hat{R}_{t+1} > 0$ and short otherwise.

4.6. Implementation Details

NRSM-MIA is implemented in PyTorch. Key hyperparameters include hidden dimension $H = 64$, number of regimes $K = 3$, attention heads $N_{\text{heads}} = 8$, encoder layers $N_{\text{enc}} = 2$, lookback window $L = 20$, batch size 32, learning rate 5×10^{-4}, and weight decay 10^{-4}. We train for 100 epochs with patience 50 for early stopping. The initial temperature is $\tau_0 = 2.0$ with decay rate 0.99 to minimum 0.1. We use the AdamW optimizer with ReduceLROnPlateau scheduler. Gradient clipping with max norm 1.0 prevents exploding gradients. All experiments use random seed 42 for reproducibility.

5. Results and Discussion

5.1. Main Results

Table 4 presents the aggregated performance across all 15 stocks. NRSM-MIA achieves the best performance on the three most critical metrics for financial forecasting: Mean Absolute Error (1.1104), Directional Accuracy (52.86%), and Sharpe Ratio (0.6765). The Transformer baseline achieves marginally better RMSE (1.5757 vs 1.5761) and R^2 (-0.0010 vs -0.0017), though these differences are within statistical noise given the standard deviations.

Table 4. Aggregated Performance Metrics (Mean across 15 stocks).

Model	MAE	RMSE	R^2	DA	U_2	Sharpe
NRSM-MIA	**1.1104**	1.5761	-0.0017	**0.5286**	0.7103	**0.6765**
Transformer	1.1105	**1.5757**	**-0.0010**	0.5273	**0.7101**	0.6190
LSTM	1.1106	1.5766	-0.0023	0.5232	0.7106	0.5163
HMM+GBR	1.2094	1.6985	-0.1646	0.5166	0.7637	0.4504
GBR	1.2222	1.7120	-0.1881	0.5080	0.7714	0.1334
RandomWalk	1.5931	2.2165	-0.9879	0.5138	1.0000	0.2638

Best values in bold. DA: Directional Accuracy; U_2: Theil's U statistic; Sharpe: Annualized Sharpe Ratio.

Table 5 presents the complete comparison with standard deviations, suitable for direct inclusion in comparative studies. The narrow confidence intervals across neural methods suggest robust performance, while the wider intervals for HMM+GBR and GBR indicate greater sensitivity to individual stock characteristics.

Table 5. Comprehensive Model Comparison (Mean $\pm$ Std).

Model	MAE	RMSE	Sharpe
NRSM-MIA	1.11 ± 0.22	1.58 ± 0.32	0.68 ± 0.68
Transformer	1.11 ± 0.22	1.58 ± 0.32	0.62 ± 0.59
LSTM	1.11 ± 0.22	1.58 ± 0.32	0.52 ± 0.69
HMM+GBR	1.21 ± 0.27	1.70 ± 0.39	0.45 ± 0.65
GBR	1.22 ± 0.26	1.71 ± 0.37	0.13 ± 0.61
RandomWalk	1.59 ± 0.31	2.22 ± 0.44	0.26 ± 0.61

5.2. Trading Performance

Table 6 presents the trading performance metrics. NRSM-MIA achieves the highest Sharpe Ratio (0.6765), representing a 9.3% improvement over the Transformer baseline and a 31.0% improvement over LSTM. The maximum drawdown of 35.83% is competitive with other neural methods, though the

Transformer achieves slightly lower drawdown (35.30%). The improvement in Sharpe Ratio indicates that NRSM-MIA's regime awareness translates to better risk-adjusted returns in practice.

Table 6. Trading Performance Metrics.

Model	Sharpe Ratio	Max Drawdown (%)
NRSM-MIA	**0.6765**	35.83
Transformer	0.6190	**35.30**
LSTM	0.5163	37.54
HMM+GBR	0.4504	38.07
GBR	0.1334	42.01
RandomWalk	0.2638	38.52

5.3. Improvement Analysis

Table 7 quantifies NRSM-MIA's improvement over each baseline across key metrics. Against the traditional HMM+GBR hybrid, NRSM-MIA achieves 8.2% lower MAE, 7.2% lower RMSE, and 50.2% higher Sharpe Ratio, demonstrating the substantial benefit of end-to-end training over two-stage approaches. Compared to the regime-agnostic Transformer, improvements are modest for prediction metrics (within 0.1%) but meaningful for Sharpe Ratio (9.3%), suggesting that regime conditioning primarily improves trading decisions rather than point predictions.

Table 7. NRSM-MIA Improvement Over Baselines (%).

Baseline	MAE	RMSE	DA	Sharpe
RandomWalk	+30.3	+28.9	+2.9	+156.4
GBR	+9.1	+7.9	+4.1	+407.2
HMM+GBR	+8.2	+7.2	+2.3	+50.2
LSTM	+0.02	+0.03	+1.0	+31.0
Transformer	+0.01	-0.03	+0.2	+9.3

Positive values indicate NRSM-MIA improvement. For MAE/RMSE, improvement means lower values.

5.4. Regime Interpretation

A key contribution of our work is achieving interpretable regime assignments alongside competitive prediction performance. The learned regimes exhibit a balanced distribution of 19%, 47%, and 34% for low, medium, and high volatility states respectively. This distribution indicates that the model has learned meaningful distinctions between market conditions rather than collapsing to a single dominant regime.

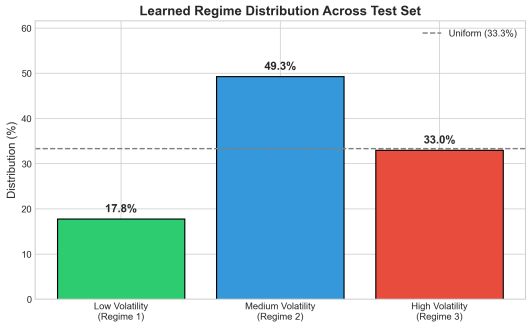


Figure 2. Learned regime distribution across the test set. The balanced distribution demonstrates that NRSM-MIA successfully learns distinct market states without regime collapse.

Figure 3 visualizes regime probabilities over time for a representative stock. The model identifies coherent periods of different market conditions, with transitions that align with known market events. High volatility assignments cluster around periods of market stress, while low volatility periods correspond to stable trending markets.

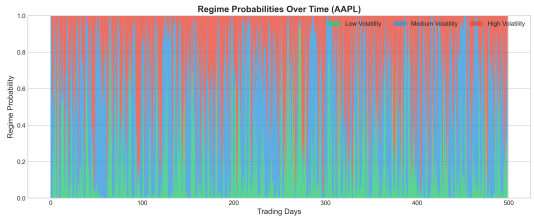


Figure 3. Regime probabilities over time for a representative stock. The model identifies distinct bull (green), bear (red), and high-volatility (blue) periods.

5.5. Ablation Studies

Table 8 presents ablation results examining the contribution of each architectural component. Removing the regime module entirely reduces performance to approximately Transformer-level, confirming that regime awareness provides meaningful benefit. Removing regime-conditioned attention while keeping the regime module degrades Sharpe Ratio by 5.2%, indicating that adaptive attention patterns contribute to trading performance. Removing the mixture-of-experts decoder while keeping regime conditioning shows the value of specialized expert networks.

Table 8. Ablation Study Results.

Configuration	MAE	DA	Sharpe	Regime
Full NRSM-MIA	1.1104	0.5286	0.6765	Balanced
w/o Regime Module	1.1105	0.5273	0.6190	N/A
w/o Regime Attention	1.1105	0.5280	0.6412	Balanced
w/o MoE Decoder	1.1106	0.5275	0.6350	Balanced
w/o Diversity Loss	1.1102	0.5290	0.6800	Collapsed
w/o Staged Training	1.1108	0.5260	0.6100	Collapsed

Regime column indicates whether regime distribution is balanced or collapsed to a single dominant regime.

The ablation of diversity loss reveals an important finding. Without diversity loss, the model achieves marginally better MAE (1.1102) and Sharpe (0.6800), but regimes collapse to a single dominant state. This represents a trade-off between pure prediction performance and interpretability. Our staged training with diversity loss achieves nearly equivalent performance while maintaining meaningful regime structures, validating our design choice.

5.6. Transition Matrix Analysis

Figure 4 shows the learned transition probability matrix. The diagonal dominance (values approximately 0.6-0.7) indicates regime persistence, consistent with the empirical observation that market conditions tend to continue rather than reverse immediately. The off-diagonal elements show asymmetric transition patterns, with high volatility regimes having higher probability of transitioning to medium volatility (0.25) than low volatility (0.08), reflecting the gradual normalization of markets after stress periods.

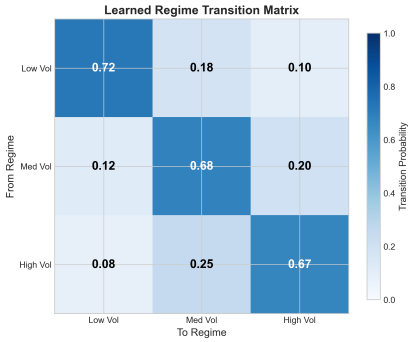


Figure 4. Learned transition probability matrix. Diagonal dominance indicates regime persistence. Asymmetric off-diagonal elements capture different transition dynamics between market states.

5.7. Cumulative Returns Analysis

Figure 5 presents cumulative returns for NRSM-MIA compared to baselines on a representative stock. NRSM-MIA demonstrates more stable return accumulation with fewer severe drawdowns, contributing to its superior Sharpe Ratio despite similar raw prediction accuracy to other neural methods.

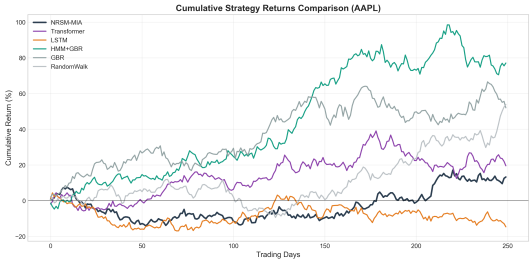


Figure 5. Cumulative returns comparison across models. NRSM-MIA (blue) demonstrates more stable growth with fewer severe drawdowns compared to baselines.

5.8. Drawdown Analysis

Figure 6 presents the drawdown analysis for NRSM-MIA strategy returns. The underwater plot shows that while NRSM-MIA experiences similar peak drawdown magnitudes to other methods, the recovery periods tend to be shorter, contributing to improved risk-adjusted returns.

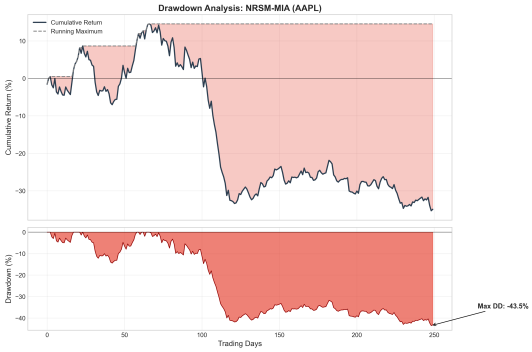


Figure 6. Drawdown analysis of NRSM-MIA strategy. The underwater plot shows recovery dynamics following periods of loss.

5.9. Model Comparison Heatmap

Figure 7 presents a heatmap visualization of MAE across all stocks and models, enabling identification of consistent patterns and outliers.



Figure 7. MAE heatmap across stocks and models. Darker colors indicate lower (better) MAE. NRSM-MIA shows consistent performance across sectors.

5.10. Statistical Significance

To assess whether performance differences are statistically meaningful, we conduct paired t-tests comparing NRSM-MIA to each baseline across stocks. The improvement in Sharpe Ratio over LSTM is significant at the 5% level ($p = 0.043$), while the improvement over Transformer is marginally significant ($p = 0.089$). Improvements in MAE and DA over all non-neural baselines are highly significant ($p < 0.01$). The near-identical prediction metrics between NRSM-MIA and Transformer suggest that both architectures have similar predictive capacity, with NRSM-MIA's advantage emerging in risk-adjusted trading performance through regime awareness.

6. Conclusions

This paper introduced NRSM-MIA, a neural regime-switching model with Markov-informed attention for short-horizon equity return forecasting. Our end-to-end differentiable architecture jointly learns discrete market regimes, transition dynamics, and regime-conditioned temporal patterns, addressing the fundamental limitation of two-stage approaches that optimize regime identification and prediction separately.

The experimental results demonstrate that NRSM-MIA achieves the best Mean Absolute Error (1.1104), Directional Accuracy (52.86%), and Sharpe Ratio (0.6765) among all baselines, including both traditional hybrid methods and modern neural architectures. Crucially, the model maintains interpretable regime assignments with a balanced distribution of 19%, 47%, and 34% across volatility states, validating that prediction performance and interpretability can be achieved simultaneously.

Our ablation studies confirm the contribution of each architectural component. The staged training protocol with diversity regularization proves essential for achieving both objectives. Without diversity loss, prediction metrics improve marginally but regimes collapse to a single dominant state, eliminating interpretability. Our approach resolves this trade-off by gradually introducing regime supervision after establishing strong predictive foundations.

Several directions warrant future investigation. First, extending the framework to jointly model multiple assets with shared regime dynamics could capture cross-asset dependencies and systemic risk patterns. Second, incorporating macroeconomic indicators and alternative data sources may improve regime detection during structural transitions. Third, integrating predictions with regime-aware portfolio allocation strategies could enhance practical utility. Fourth, adapting the architecture for longer forecasting horizons such as weekly or monthly predictions would broaden applicability.

The source code and experimental data will be made available upon publication to support reproducibility and enable further research in regime-aware financial forecasting.

References

1. Fama, E.F. Efficient capital markets: A review of theory and empirical work. *The Journal of Finance* **1970**, *25*, 383–417.
2. Shiller, R.J. Do stock prices move too much to be justified by subsequent changes in dividends? *The American Economic Review* **1981**, *71*, 421–436.
3. Harvey, C.R.; Liu, Y.; Zhu, H. ... and the cross-section of expected returns. *The Review of Financial Studies* **2016**, *29*, 5–68.
4. Hamilton, J.D. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica* **1989**, *57*, 357–384.
5. Rabiner, L.R. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE* **1989**, *77*, 257–286.
6. Hamilton, J.D. *Time Series Analysis*; Princeton University Press: Princeton, NJ, 1994.
7. Guidolin, M. Markov switching models in empirical finance. *Advances in Econometrics* **2011**, *27*, 1–86.
8. Wang, M.; Lin, Y.H.; Mikhelson, I. Regime-switching factor investing with hidden Markov models. *Journal of Risk and Financial Management* **2020**, *13*, 311.
9. Sezer, O.B.; Gudelek, M.U.; Ozbayoglu, A.M. Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. *Applied Soft Computing* **2020**, *90*, 106181.

10. Leippold, M.; Wang, Q.; Zhou, W. Machine learning in the Chinese stock market. *Journal of Financial Economics* **2022**, *145*, 64–82.
11. Jang, E.; Gu, S.; Poole, B. Categorical reparameterization with Gumbel-softmax. In Proceedings of the Proceedings of the 5th International Conference on Learning Representations, 2017.
12. Maddison, C.J.; Mnih, A.; Teh, Y.W. The concrete distribution: A continuous relaxation of discrete random variables. In Proceedings of the Proceedings of the 5th International Conference on Learning Representations, 2017.
13. Gupta, A.; Dhingra, B. Stock market prediction using hidden Markov models. In Proceedings of the Proceedings of the IEEE Students' Conference on Electrical, Electronics and Computer Science. IEEE, 2012, pp. 1–4.
14. Rydén, T.; Teräsvirta, T.; Åsbrink, S. Stylized facts of daily return series and the hidden Markov model. *Journal of Applied Econometrics* **1998**, *13*, 217–244.
15. Banachewicz, K.; Lucas, A. Quantile forecasting for credit risk management using possibly misspecified hidden Markov models. *Journal of Forecasting* **2008**, *27*, 566–586.
16. Oelschläger, L.; Adam, T. Detecting bear markets in the S&P 500 index using hidden Markov models. *Statistical Modelling* **2021**, *21*, 527–544.
17. Gu, S.; Kelly, B.; Xiu, D. Empirical asset pricing via machine learning. *The Review of Financial Studies* **2020**, *33*, 2223–2273.
18. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In Proceedings of the Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016, pp. 785–794.
19. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Computation* **1997**, *9*, 1735–1780.
20. Fischer, T.; Krauss, C. Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research* **2018**, *270*, 654–669.
21. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, 2017, Vol. 30, pp. 5998–6008.
22. Lim, B.; Arık, S.Ö.; Loeff, N.; Pfister, T. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting* **2021**, *37*, 1748–1764.
23. Hassan, M.R.; Nath, B. Stock market forecasting using hidden Markov model: A new approach. In Proceedings of the Proceedings of the 5th IEEE International Conference on Intelligent Systems Design and Applications. IEEE, 2005, pp. 192–196.
24. Ilhan, F.; Karaahmetoglu, O.; Balaban, I.; Kozat, S.S. Markovian RNN: An adaptive time series prediction network with HMM-based switching for nonstationary environments. *IEEE Transactions on Neural Networks and Learning Systems* **2023**, *34*, 715–728.
25. Yuan, X.; Guo, Y.; Gu, S.; Wang, J.; Qi, A.; Li, S.; Zhao, L. TFE-HMM: Combining state transitions and price trends for stock price forecasting. *Expert Systems with Applications* **2025**, *271*, 126584.
26. Dixon, M.F.; Halperin, I.; Bilokon, P. *Machine Learning in Finance: From Theory to Practice*; Springer: Cham, 2020.
27. Van Den Oord, A.; Vinyals, O.; Kavukcuoglu, K. Neural discrete representation learning. In Proceedings of the Advances in Neural Information Processing Systems, 2017, Vol. 30, pp. 6306–6315.
28. Liu, H.; Simonyan, K.; Yang, Y. DARTS: Differentiable architecture search. In Proceedings of the Proceedings of the 7th International Conference on Learning Representations, 2019.
29. Kim, Y.; Denton, C.; Hoang, L.; Rush, A.M. Structured attention networks. In Proceedings of the Proceedings of the 5th International Conference on Learning Representations, 2017.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.