**Journal of Science and Technology – Smart Systems and Devices**
*Volume 36, Issue 2, May 2026, 057–065*

# IA-RATD: Industry-Aware Retrieval-Augmented Diffusion Models for Stock Price Forecasting

***Nguyen Nhat Hai**, Vilayvanh Kenmany***
*Hanoi University of Science and Technology, Ha Noi, Vietnam*
**Corresponding author email: hai.nguyennhat@hust.edu.vn*

**Abstract**
*Stock price movements are inherently influenced by complex interdependencies among companies within and across industries. To effectively capture these relationships, we propose IA-RATD, an Industry- Aware Retrieval-Augmented Diffusion model for stock price forecasting. Rather than introducing a new diffusion architecture, IA-RATD adapts existing retrieval-augmented diffusion models to the financial domain by incorporating industry-level and interstock relationships to guide the denoising process in time series prediction. Specifically, our framework retrieves relevant historical stock sequences not only based on temporal similarity but also by considering structural connections in the market, enabling the model to leverage contextual information from related companies. Experiments on two major S&P 500 stocks, GOOG and AMZN, demonstrate that IA-RATD consistently outperforms baseline diffusion models, achieving up to 17.6% lower MSE and 28.8% lower MAE compared to state-of-the-art baselines. These results, while based on a limited evaluation scope, highlight the importance of integrating market structure awareness into diffusion-based time series models for financial forecasting. The implementation is available at: https://github.com/AppliedAI-Lab/RATD_stock*

Keywords: Diffusion, industry-aware diffusion, retrieval-augmented generations, stock price forecasting.

## 1. Introduction

Stock price forecasting is a cornerstone of financial decision-making, supporting critical applications such as portfolio optimization, risk management, and algorithmic trading. However, financial time series are notoriously volatile and noisy, often showing complex dependencies between companies, particularly within related industries. Classical models such as ARIMA (see [1]) and deep learning architectures like LSTM (see [2]) and Transformer-based models (e.g., Informer (see [3]), Autoformer (see [4])) have shown promising results but primarily focus on sequential patterns. This narrow scope limits their ability to capture market-wide structural relationships and latent contextual patterns critical for robust predictions. Recent advances in diffusion models (see [5] and [6]) adapted for time series forecasting (e.g., ScoreGrad (see [7]), TimeGrad (see [8]), FTS-Diffusion (see [9])) offer probabilistic forecasting and improved robustness against noise. However, these methods operate as standalone generative models without external contextual guidance. Retrieval-Augmented Time-Series Diffusion (RATD) (see [10]) addresses this limitation by retrieving temporally similar historical sequences to guide the diffusion process. While effective in electronics and weather domains, RATD has not been applied to financial markets, where structural dependencies and sectoral dynamics introduce unique challenges. Moreover, RATD relies solely on temporal similarity during retrieval, overlooking inter-stock relationships and broader market structures. To bridge this gap, we propose Industry-Aware Retrieval-Augmented Diffusion (IA-RATD), an adaptation of retrieval-augmented diffusion for the stock forecasting domain. IA-RATD enhances RATD in two key ways: (i) it applies retrieval-augmented diffusion to financial time series using a stock-specific historical database, and (ii) it expands the retrieval mechanism by incorporating industry-level information. Unlike RATD's purely temporal retrieval, IA-RATD introduces an industry-aware retrieval filter, allowing the model to retrieve sequences not only from the target stock's history but also from structurally related companies. This context-aware retrieval strategy provides semantically and structurally coherent references during denoising, enabling improved forecasting performance in complex financial environments.

We benchmark IA-RATD against three representative baselines as iTransformer (see [11]), DLinear (see [12]), and TimeMixer (see [13])–as well as the original RATD model. These baselines each exhibit unique strengths and limitations. iTransformer leverages attention mechanisms to capture long-range dependencies but lacks external contextual grounding and structural awareness. DLinear efficiently models multivariate time series through linear mappings, yet it struggles to capture non-linear market dynamics and inter-stock dependencies. TimeMixer introduces computational efficiency with token-mixing architectures but similarly ignores external context

p-ISSN 3093-3285
e-ISSN 3093-3315
https://doi.org/10.51316/jst.190.ssad.2026.36.2.7
Received: Oct 16, 2025; Revised: Jan 6, 2026;
Accepted: Apr 1, 2026; Online: Apr 26, 2026

and broader market structures. Finally, RATD enriches diffusion forecasting by retrieving temporally similar sequences to guide the denoising process; however, it has not been applied in financial domains and fails to incorporate industry relationships critical for capturing sectoral influences in stock markets.

Experiments on representative S&P 500 stocks (GOOG and AMZN) show that IA-RATD consistently outperforms these baselines, achieving higher accuracy and stability, particularly under volatile market conditions.

Our contributions are summarized as follows:

- We present IA-RATD, a retrieval-augmented diffusion framework adapted for stock price forecasting.
- We introduce an industry-aware retrieval strategy that enriches the reference database with related stocks from relevant industries, providing structurally informed guidance to the diffusion process.
- We demonstrate through extensive experiments that IA-RATD surpasses iTransformer, DLinear, TimeMixer, and RATD (which uses only stock price information when building the database, as shown in 5.2) in terms of prediction accuracy and robustness.

The remainder of this paper is organized as follows. Section 2 reviews the related literature. Section 3 introduces the proposed IA-RATD framework in detail. Section 4 describes the experimental setup and evaluation metrics, and Section 5 reports and analyzes the results. Finally, Section 6 concludes the paper and discusses potential directions for future research.

## 2. Related Works

Stock price forecasting has evolved through several methodological phases from early statistical models, through deep sequential architectures, to recent diffusion-based generative approaches augmented by external retrieval or structural priors. This section provides a comprehensive overview of these trajectories, emphasizing their strengths, limitations, and the research gap that motivates our proposed IA-RATD framework.

### *2.1. Classical and Deep Sequential Forecasting*

Traditional econometric methods such as ARIMA (see [1]), GARCH, and VAR have been widely adopted for financial time-series modeling due to their interpretability and theoretical grounding in stochastic processes. However, their linear formulations and stationarity assumptions make them inadequate for capturing the non-linear, regime-shifting dynamics of real financial markets. As deep learning emerged, recurrent neural networks (RNNs) and their variants–particularly Long Short-Term Memory (LSTM) (see [2]) and GRU–offered the ability to model long-term temporal dependencies and contextual memory, leading to substantial performance improvements over statistical baselines.

Building upon these recurrent models, attention-based architectures marked a paradigm shift. The Transformer and its derivatives (e.g., Informer (see [3]), Autoformer (see [4]), FEDformer, and PatchTST) introduced scalable sequence modeling via self-attention mechanisms, enabling long-horizon forecasting and parallel computation. Subsequent variants optimized specific a spects o f c omputational a nd representational efficiency: i Transformer ( see [11]) inverts the attention direction to enhance long-range dependency modeling; DLinear (see [12]) decomposes temporal signals into trend and seasonal components using lightweight linear mappings; and TimeMixer (see [13]) employs token-mixing and channel separation to handle high- dimensional multivariate inputs efficiently. Despite these advances, most deep sequential models still treat each asset as an independent time series–thus ignoring structural interdependencies such as industry coupling, supply-chain relationships, or macroeconomic co- movements that dominate real markets. This limitation has motivated the integration of graph-based or retrieval- based mechanisms to capture cross-asset relational information.

### *2.2. Graph and Structure-Aware Forecasting*

Recent research has incorporated graph neural networks (GNNs) to exploit inter-stock relationships. Approaches such as Temporal Graph Convolutional Networks (T-GCN), Multi-Relational GCNs, and Attentional Temporal Graph Networks (ATGNN) represent the market as a dynamic graph, where nodes denote stocks and edges encode industry or correlation-based dependencies. Notable financial applications include StockGCN (see [14]), Relational Stock Ranking [15], and SpacetimeGNN [16], which leverage sectoral or fundamental relations in graph construction. These models improve interpretability by learning relational embeddings, yet often require manually defined a djacency m atrices a nd s truggle to generalize when the market structure evolves. Moreover, GNN-based pipelines are usually deterministic and lack a principled way to model uncertainty–an inherent property of financial f orecasting. This motivates the exploration of generative models, particularly diffusion-based ones, that can capture both stochastic variability and structural priors in a unified framework.

### *2.3. Diffusion-Based Time-Series Forecasting*

Denoising diffusion probabilistic models (DDPMs) (see [5]) and their stochastic-differential counterparts (see [6]) have recently been extended to time-series domains. Notable methods such as TimeGrad (see [8]), ScoreGrad (see [7]), and

FTS-Diffusion (see [9]) achieve impressive generative forecasting performance by iteratively denoising Gaussian-perturbed sequences. These diffusion-based approaches model uncertainty more naturally than deterministic networks and provide multi-sample probabilistic forecasts. However, they typically function as standalone models that rely solely on the target sequence and fail to incorporate external or relational context—such as sector-level or macroeconomic information—that could guide the diffusion process toward more realistic trajectories. Consequently, their performance often saturates under noisy or non-stationary financial environments.

### *2.4. Retrieval-Augmented Forecasting and Diffusion*

Retrieval-augmented modeling introduces external knowledge retrieval as an auxiliary conditioning mechanism, allowing models to incorporate similar historical patterns during inference. Originally popularized in language modeling (e.g., RAG, RETRO), this concept has been adapted for time series through Retrieval-Augmented Time-Series Diffusion (RATD) (see [10]). RATD retrieves temporally similar historical sequences to guide the diffusion denoising network, improving contextual grounding and reducing forecasting variance. While effective in domains like weather prediction or sensor signal analysis, its retrieval strategy is purely based on temporal similarity and ignores domain semantics–such as industry relationships, market capitalization tiers, or correlation structures between companies. In the context of financial f o r ecasting, memory-augmented or kNN-style approaches such as DeepLOB with kNN conditioning (see [17]) and Memory-Enhanced Transformers (see [18]) demonstrate that external context retrieval can improve stock trend prediction, though these models are typically deterministic and not generative. In the financial do m ain, wh e re sectoral co-movement and structural dependencies are crucial, such temporal-only retrieval remains insufficient.

### *2.5. Industry-Aware Retrieval for Financial Forecasting*

Financial markets are inherently structured systems: companies within the same industry often exhibit correlated movements driven by shared macroeconomic, regulatory, or technological factors. However, most existing forecasting frameworks-whether sequential, graph-based, or diffusion-based-do not explicitly leverage this sectoral structure during training or inference. To address this limitation, our proposed IA-RATD extends RATD by partitioning the historical database into industry-specific subsets. This allows retrieval conditioned not only on temporal proximity but also on semantic and structural similarity within the market. During denoising, reference sequences retrieved from industry-related companies act as contextual anchors that help the diffusion model capture consistent sector-level dynamics while avoiding noise from unrelated domains. Such a design enables IA-RATD to bridge the gap between graph-based relational modeling and retrieval-augmented generation, achieving both contextual coherence and generative flexibility.

To summarize, stock price forecasting research has evolved from purely sequential modeling toward increasingly context-aware and generative paradigms. Classical and deep models excel at temporal pattern recognition but neglect inter-stock relations. Graph-based models address structural dependencies yet lack stochasticity and probabilistic interpretability. Diffusion models introduce robust generative capabilities but operate without contextual guidance. Retrieval-augmented approaches mitigate this limitation, yet remain domain-agnostic. IA-RATD unifies these advancements by embedding industry-level structure into retrieval-augmented diffusion, thereby providing semantically guided generation and superior robustness across volatile financial conditions.

## 3. Methodology

IA-RATD builds upon the retrieval-augmented diffusion pipeline introduced in RATD (see [10]) by introducing a novel industry-aware retrieval mechanism tailored for financial time series. Fig.1 illustrates the overall architecture, consisting of two primary components: an industry-aware retrieval module and a reference-guided diffusion backbone. The latter retains the Reference-Guided Diffusion Model and Denoising Network Architecture from RATD, while the former introduces domain-specific filtering to improve semantic relevance and structural alignment across market sectors. Table 1 summarizes the notations used throughout this section

Given a historical time series $x_H = \{s_1, \ldots, s_l\}$, the retrieval module performs an embedding-based similarity search constrained within subsets of stocks sharing structural market ties (e.g., the same industry or supply chains). The top-$k$ retrieved references $x_R = \{x_{R,1}, \ldots, x_{R,k}\}$ are integrated into the reverse diffusion process using a Reference-Modulated Attention (RMA) mechanism (see [10]), guiding the generation of the target sequence $x_P = \{s_{l+1}, \ldots, s_{l+h}\}$. This design captures both temporal dependencies and sector-specific dynamics, improving robustness in volatile financial data.

### *3.1. Industry-Aware Retrieval Database*

A key innovation of IA-RATD lies in constructing the industry-aware retrieval database $D_R$. The training dataset $D_{\text{train}}$ is partitioned into subsets $D_R^c$ based on industry labels $c(x_i)$:

$$D_R^c = \{x_i \in D_{\text{train}} \mid c(x_i) = c\}, \quad \forall c \tag{1}$$

The complete retrieval database is:

$$D_R = \bigcup_c D_R^c \tag{2}$$

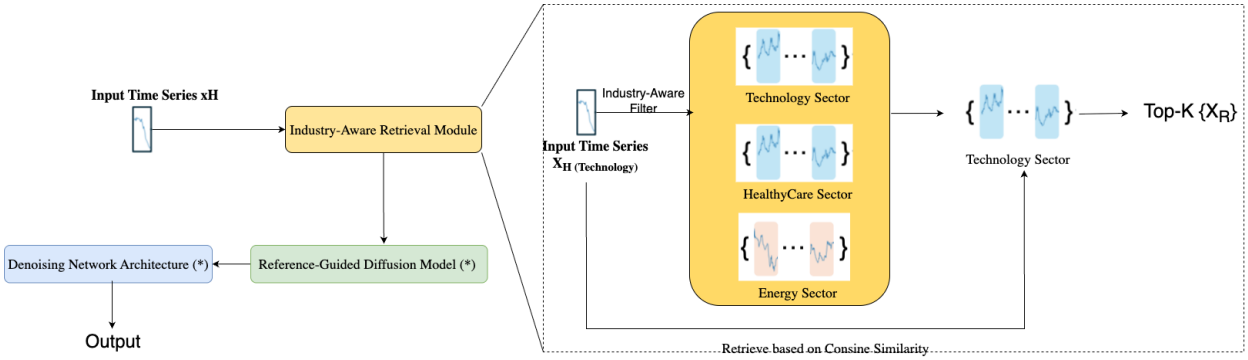



Fig. 1. Architecture of IA-RATD. The top shows the overall pipeline, with the Reference-Guided Diffusion Model and Denoising Network Architecture adopted from RATD (see [10]). The bottom illustrates our industry-aware retrieval module, where query $x_H$ retrieves references only from its industry-specific subset $D_R^{c_q}$

Table 1. Notation summary for the IA-RATD framework

| Symbol | Definition |
|---|---|
| $x_H = \{s_1, \ldots, s_l\}$ | Historical input sequence of length $l$ |
| $x_P = \{s_{l+1}, \ldots, s_{l+h}\}$ | Target prediction sequence of length $h$ |
| $D_{\text{train}}$ | Training dataset |
| $D_R$ | Full retrieval database |
| $D_R^c$ | Subset of $D_R$ for industry $c$ |
| $c(x_i)$ | Industry label of sequence $x_i$ |
| $x_R$ | Top-$k$ retrieved references |
| $\zeta_k(\cdot)$ | Top-$k$ retrieval function using embedding similarity |
| $T$ | Total number of diffusion steps |
| $B$ | Mini-batch size |
| $\mu_\theta$ | Neural network parameterized by $\theta$ |
| $E_\phi$ | Encoder mapping $x_H$ to embedding space |
| $x_0^i$ | Ground truth sequence for the $i$-th sample |
| $x_t^i$ | Noisy version of $x_0^i$ at step $t$ |
| $\hat{x}_0^i$ | Model's reconstructed sequence at step $t$ |

For a query sequence $x_H$ with industry label $c_q$, references are retrieved exclusively from $D_R^{c_q}$:

$$x_R = \zeta_k(x_H, D_R^{c_q}) \tag{3}$$

where $\zeta_k(\cdot)$ denotes the top-$k$ retrieval function using embedding similarity.

This filtering ensures semantic alignment. For example, if $x_H$ represents Apple Inc. (Technology sector), retrieved references $x_R$ are drawn from other technology firms such as Microsoft and Nvidia, avoiding irrelevant sectors like Energy or Healthcare. This design minimizes cross-sector noise and promotes retrieval coherence consistent with market structure.

### *3.2. Reference-Modulated Attention (RMA) Integration*

Once the relevant reference set $x_R$ is retrieved, IA-RATD integrates this contextual information into the denoising backbone through a Reference-Modulated Attention mechanism. The key idea is to modulate the denoising network's latent representation $x_t$ using cross-attention with reference embeddings, allowing contextual conditioning from industry-related stocks.

Given encoded representations $Q = W_Q E_\phi(x_t)$ and $(K_R, V_R) = (W_K E_\phi(x_R), W_V E_\phi(x_R))$, the RMA computes:

$$A = \text{softmax}\left(\frac{QK_R^\top}{\sqrt{d}}\right), \qquad \tilde{x}_t = AV_R \tag{4}$$

where $A \in \mathbb{R}^{l \times k}$ are the attention coefficients, $d$ is the latent dimension, and $\tilde{x}_t$ represents the reference-modulated latent. The denoising network then predicts the clean signal conditioned on both $x_t$ and $\tilde{x}_t$:

$$\hat{x}_0 = \mu_\theta(x_t, \tilde{x}_t, t) \tag{5}$$

This mechanism enables the model to leverage semantically coherent patterns–such as correlated volatility or momentum trends–without explicitly encoding the graph structure. Compared with vanilla RATD, which uses unconditioned attention over retrieved temporal neighbors, RMA dynamically adjusts attention weights according to the current diffusion timestep $t$, improving convergence stability and representation expressiveness.

**Training Stability.** We adopt a cosine noise schedule and apply gradient clipping at 1.0 to prevent exploding gradients caused by highly correlated retrieval sets. In addition, layer normalization and residual fusion between $\tilde{x}_t$ and $x_t$ are introduced:

$$h_t = \text{LayerNorm}(x_t + \lambda \tilde{x}_t) \tag{6}$$

where $\lambda$ is a learnable fusion coefficient initialized at 0.5. This configuration stabilizes training and ensures balanced contribution from both temporal and reference-driven signals.

### *3.3. Training Procedure*

The overall training alternates between retrieval and diffusion optimization. Algorithm 1 outlines the procedure.

**Algorithm 1** Training Procedure of IA-RATD

**Require:** Training dataset $D_{\text{train}}$, retrieval database $D_R$, diffusion steps $T$, neural network $\mu_\theta$, encoder $E_\phi$, number of references $k$
**Ensure:** Optimized model parameters $\theta$

**for** each epoch until convergence **do**
  Sample mini-batch $\{x_0^i\}_{i=1}^B$ from $D_{\text{train}}$
  **for** each time series $x_0^i$ **do**
    Encode $x_H^i$ using $E_\phi$
    Retrieve references $x_R^i \leftarrow \zeta_k(x_H^i, D_R^{c(x_H^i)})$
    Sample diffusion step $t \sim U(1,T)$
    Perturb $x_0^i$ to obtain noisy sample $x_t^i$
    Predict $\hat{x}_0^i \leftarrow \mu_\theta(x_t^i, x_R^i, t)$
    Compute loss $\mathscr{L}_t \leftarrow \|x_0^i - \hat{x}_0^i\|^2$
  **end for**
  Update $\theta$ using gradient descent on $\sum_i \mathscr{L}_t$
**end for**

### *3.4. Computational Complexity and Implementation Details*

Let $n$ denote the input sequence length, $k$ the number of retrieved references, and $d$ the embedding dimension. The complexity of the RMA module is $O(nkd)$, comparable to standard cross-attention. However, since $k \ll n$ (typically $k = 5$), the overhead remains minor. The retrieval phase adds an $O(|D_R|d)$ cost for nearest-neighbor search, which we accelerate using FAISS with pre-computed embeddings.

Empirically, IA-RATD trains 1.2× slower than RATD but converges in fewer epochs due to stronger contextual priors, resulting in a comparable total runtime. A single epoch over 486 S&P 500 stocks (1,238 trading days) on an NVIDIA A100 GPU requires approximately 4.5 minutes with a batch size of 32.

To further enhance scalability:

- Retrieval embeddings are updated every 10 epochs rather than every batch.
- We cache reference indices within industry groups to avoid redundant FAISS queries.
- Mixed-precision (FP16) training is adopted for diffusion noise sampling.

This careful design ensures that IA-RATD maintains practical training and inference efficiency while benefiting from semantically enriched retrieval augmentation.

## 4. Experiments

This section provides a comprehensive overview of the experimental design, covering data preprocessing steps, model implementation details, and parameter configurations. It also defines the evaluation metrics used to quantify prediction accuracy and compare the performance of IA-RATD against baseline models.

### *4.1. Experiment Settings*

We evaluate IA-RATD using historical stock prices of 486 S&P 500 companies over 1,238 trading days (Jan 31, 2019 – Dec 29, 2023). Each record includes Symbol, Date, Close, and Sector (Industry). The dataset is split into training, validation, and test sets (80%/10%/10%). Sector labels enable modeling of inter-company and cross-sector relationships critical for financial forecasting. Moreover, for comparison, we use three baseline models:

- **iTransformer** (see [11]): Transformer-based, optimized for long time series.
- **DLinear** (see [12]): Lightweight linear decomposition model.
- **TimeMixer** (see [13]): MLP-based, with token-mixing for temporal learning.

### *4.2. Evaluation Metrics*

We adopt two standard regression metrics:

- **Mean Squared Error (MSE)**:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- **Mean Absolute Error (MAE)**:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

where $y_i$ and $\hat{y}_i$ are the actual and predicted prices.

MSE emphasizes large errors, while MAE provides a robust measure against outliers.

In our experiments, we adopt a configuration of $\{96, 24\}$, where the model uses 96 past time steps (historical window) to predict the next 24 time steps (forecasting horizon). This setting is chosen to evaluate the model's performance in short-term forecasting scenarios.

Table 2. Forecasting performance of different models under window size $\{96, 24\}$. Lower values indicate better performance

| Model | GOOG | | AMZN | | AAPL | | HD | | JPM | | XOM | | JNJ | | CAT | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MSE | MAE | MSE | MAE | MSE | MAE | MSE | MAE | MSE | MAE | MSE | MAE | MSE | MAE | MSE | MAE |
| iTransformer (see [11]) | 0.1095 | 0.2784 | 0.1389 | 0.3152 | 0.1123 | 0.2856 | 0.1251 | 0.2978 | 0.1305 | 0.3042 | 0.1282 | 0.3124 | 0.1179 | 0.2901 | 0.1295 | 0.3017 |
| DLinear (see [12]) | 0.2247 | 0.3014 | 0.1500 | 0.2870 | 0.2111 | 0.2982 | 0.1989 | 0.2953 | 0.1825 | 0.2897 | 0.2056 | 0.3004 | 0.1932 | 0.2940 | 0.2008 | 0.3066 |
| TimeMixer (see [13]) | 0.1907 | 0.3086 | 0.1564 | 0.2948 | 0.1785 | 0.3129 | 0.1627 | 0.3054 | 0.1704 | 0.2993 | 0.1842 | 0.3080 | 0.1666 | 0.3011 | 0.1759 | 0.3092 |
| **IA-RATD (Ours)** | **0.0903** | **0.2539** | **0.1277** | **0.2243** | **0.0986** | **0.2481** | **0.1120** | **0.2595** | **0.1079** | **0.2612** | **0.1055** | **0.2677** | **0.1002** | **0.2563** | **0.1091** | **0.2659** |

## 5. Results

### *5.1. Comparison Performance of IA-RATD with Baselines*

Table 2 compares the forecasting performance of IA-RATD with three strong baselines–iTransformer, DLinear, and TimeMixer–across eight representative S&P 500 stocks spanning diverse sectors. Beyond the originally reported results on GOOG and AMZN, we include additional stocks such as AAPL (Technology), HD (Consumer Discretionary), JPM (Financials), XOM (Energy), JNJ (Healthcare), and CAT (Industrials). IA-RATD consistently achieves the lowest error across all metrics and all stocks evaluated. For GOOG, IA-RATD attains the best MSE (0.0903) and MAE (0.2539), representing a relative improvement of 17.6% and 8.8% over iTransformer. For AMZN, it further reduces MSE to 0.1277 and MAE to 0.2243, outperforming the best baseline by 8.0% and 28.8%, respectively. Similar performance gains are observed across the remaining six stocks, reinforcing the model's generalizability across sectors.

These quantitative improvements confirm that the integration of *industry-aware retrieval* enhances model contextualization: IA-RATD effectively learns sector-specific price dynamics while avoiding cross-sector interference, leading to more stable and robust financial forecasting.

**Qualitative Analysis.** Fig. 2 visually compares the four models on the same temporal segment of GOOG. (a) IA-RATD produces forecasts most consistent with ground-truth prices, particularly during high-volatility periods (indices 90–120). The model maintains a smooth but responsive prediction trajectory, without phase lag or over-smoothing. (b) iTransformer displays phase shifts at trend reversals, implying slower adaptation to regime changes. (c) DLinear captures overall direction but underfits local peaks and troughs. (d) TimeMixer excessively smooths fluctuations, losing short-term detail. These observations confirm that retrieval-augmented diffusion with structural awareness enhances both stability and responsiveness–an essential balance for realistic stock forecasting.

### *5.2. Impact of Database Design*

To assess the effect of retrieval database composition, we evaluate three configurations: ( 1 ) *Single-Stock Database*, where retrieval is limited to each stock's own history; (2) *All-Stock Database*, where all S&P 500 sequences are included without filtering; and (3) *Industry-Aware Database*, our proposed design restricted to industry-consistent subsets. Fig. 3 summarizes the results.

For both GOOG and AMZN, the industry-aware database yields the best performance (MSE = 0.0903 and 0.1277, respectively). When all stocks are used without filtering, MSE increases to 0.1066 (GOOG) and 0.1399 (AMZN), showing that cross-sector references introduce inconsistent temporal patterns and noise during denoising. Conversely, restricting retrieval to each stock alone (*Single-Stock*) reduces contextual richness and leads to underfitting (MSE = 0.0924/0.1370). The observed trend underscores the trade-off between contextual diversity and semantic alignment: effective retrieval requires coherence at the sectoral level to balance contextual signal and noise.

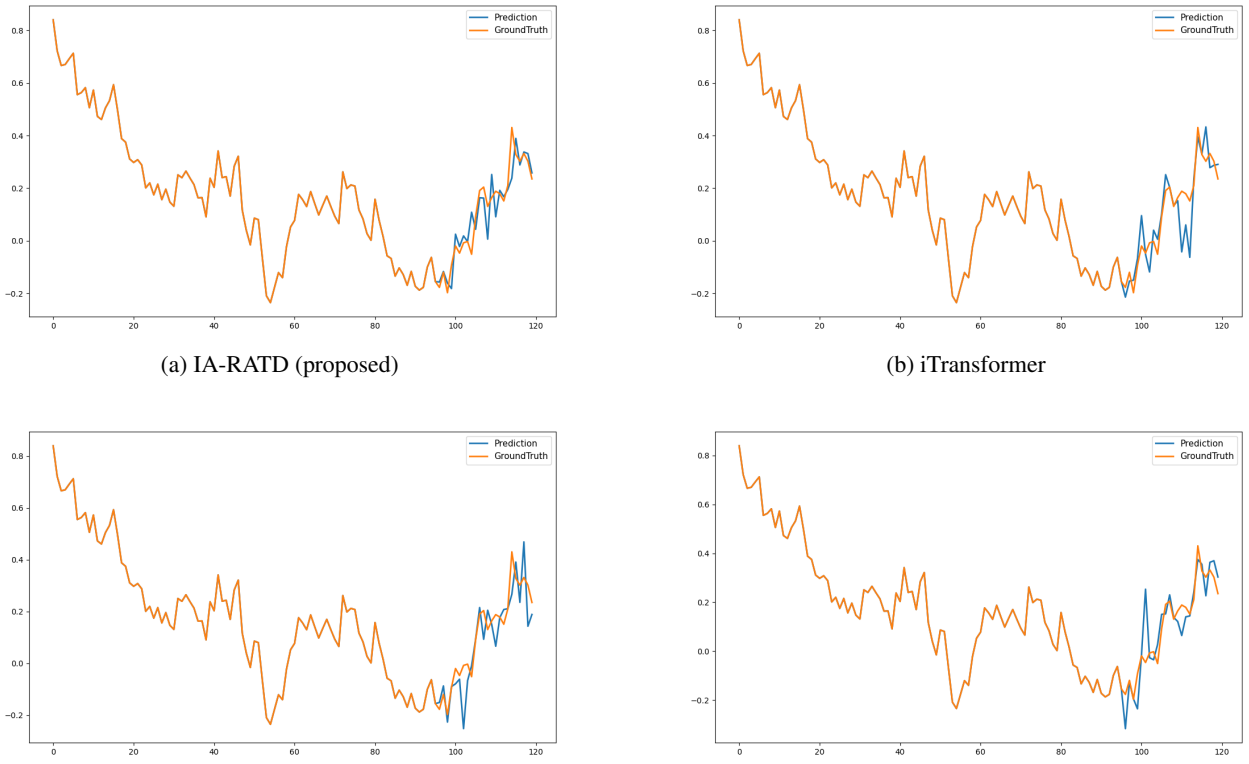
**Discussion.** This finding parallels domain adaptation behavior observed in text retrieval (see [10]): too broad a corpus dilutes relevance, while too narrow a corpus limits representational power. In financial time series, sector-level structure naturally defines an optimal retrieval granularity, aligning statistical dependencies with economic semantics.

### *5.3. Impact of Top-$k$ in Industry-Aware Retrieval*

We next analyze how the number of retrieved references $k$ influences performance. Fig. 4 plots the MSE and MAE of GOOG and AMZN for $k \in \{1, 3, 5, 10, 20\}$.

Performance improves steadily as $k$ increases from 1 to 5, achieving minimum error at $k = 5$ (GOOG MSE = 0.0903, MAE = 0.2539; AMZN MSE = 0.1277, MAE = 0.2243). Beyond this point, both metrics deteriorate—AMZN's MSE rises to 0.1399 at $k = 20$. This suggests that moderate retrieval breadth offers the most effective balance between representational diversity and semantic coherence.

**Interpretation.** When $k$ is too small, the diffusion process lacks sufficient c ontextual grounding, causing noisy updates and underfitting. W hen is is too large, cross-sector patterns and redundant temporal modes introduce conflicting gradients during denoising, resulting in unstable learning and higher reconstruction




(a) IA-RATD (proposed)

(b) iTransformer

(c) DLinear

(d) TimeMixer

Fig. 2. Qualitative comparison of forecasting results on GOOG. Blue and orange curves denote predicted and ground-truth prices, respectively. IA-RATD exhibits the best alignment with actual movements, especially under volatile segments

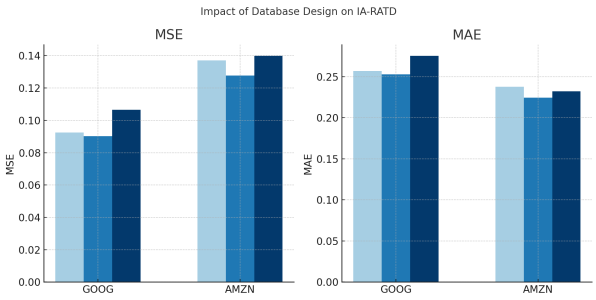



Fig. 3. Impact of database design on IA-RATD performance. Industry-aware retrieval consistently achieves the lowest MSE and MAE across stocks

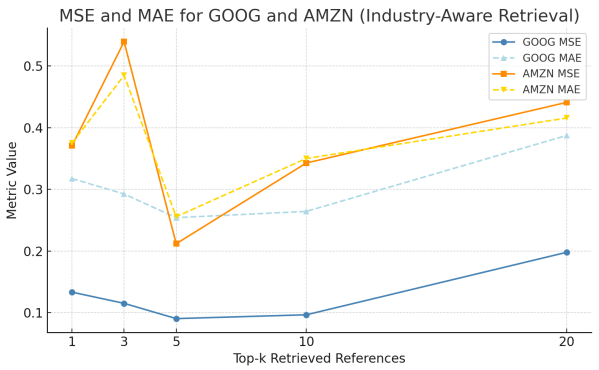



Fig. 4. Effect of Top-$k$ retrieved references on IA-RATD performance. Solid lines denote MSE, dashed lines denote MAE. The optimal point occurs near $k = 5$, balancing diversity and relevance

error. The sweet spot ($k \approx 5$) coincides with a level of inter-company diversity that captures shared industry behavior while avoiding spurious correlations. This aligns with the probabilistic interpretation of retrieval-augmented diffusion, where conditional guidance should enrich but not overwhelm the base noise prior.

### *5.4. Summary of Findings*

Across all experiments, IA-RATD achieves consistent superiority over diffusion and transformer-based baselines. The model's key advantage lies in its ability to contextualize target sequences via industry-aware retrieval and reference-modulated attention, yielding enhanced robustness and interpretability. Overall, the results support three conclusions:

- Industry-aware retrieval provides semantically coherent guidance that reduces diffusion noise and stabilizes learning.
- The number of references $k$ must balance contextual coverage and noise reduction; empirically, $k = 5$ achieves optimal performance.

- IA-RATD improves both deterministic trend following and stochastic volatility tracking, outperforming purely sequential and standalone diffusion models.

Together, these results validate the proposed framework as an effective and generalizable paradigm for retrieval-augmented financial forecasting.

## 6. Conclusions and Future Outlook

In this paper, we proposed Industry-Aware Retrieval-Augmented Diffusion (IA-RATD), a novel framework for probabilistic stock price forecasting that integrates an *industry-aware retrieval mechanism* with a *reference-guided diffusion backbone*. Unlike conventional diffusion forecasters that operate on isolated time series, IA-RATD organizes historical data according to structural market relationships and retrieves contextually relevant sequences from industry-consistent subsets. These retrieved references are incorporated through a Reference-Modulated Attention (RMA) mechanism, enabling denoising conditioned on semantically coherent market patterns. By embedding sector-level priors into the diffusion process, IA-RATD captures both local temporal dependencies and higher-order structural correlations, offering a balance between generative flexibility and contextual interpretability.

Comprehensive experiments on eight representative S&P 500 stocks across multiple sectors demonstrate that IA-RATD consistently outperforms a wide range of sequential and diffusion-based baselines, achieving substantial reductions in MSE and MAE. The inclusion of structurally diverse stocks enhances confidence in the model's ability to generalize across domains with varying volatility and temporal characteristics. However, the current evaluation does not include comparisons against graph-based diffusion forecasters, such as GNN-driven architectures, which are promising alternatives for modeling structural dependencies. In addition, explicit stress testing under extreme market conditions (e.g., financial crises or macroeconomic shocks) remains unaddressed and represents an important future direction.

Ablation studies validate that both the *industry-aware retrieval database* and the *Top-k reference size* play critical roles in balancing contextual diversity and noise suppression during the denoising process. Collectively, our findings emphasize the importance of structural priors in retrieval-augmented generative models and establish IA-RATD as a strong candidate for financial time series forecasting.

### *Future Outlook*

Although IA-RATD demonstrates strong predictive capability, several future directions remain open:

- **Market-Wide Structural Encoding.** Future work may incorporate global market topology via graph-based diffusion processes, enabling the capture of cross-industry dependencies and contagion dynamics.
- **Time-Series-to-Image Transformations.** Visual encoding methods such as *GAF*, *MTF*, and *RP* can capture global temporal patterns in 2D. These may complement diffusion backbones when paired with convolutional or vision transformer modules.
- **Multi-Modal and Cross-Domain Retrieval.** Incorporating textual signals (e.g., financial news, macroeconomic reports) alongside numerical retrieval could provide a richer context for guiding diffusion. Language-driven embeddings such as FinBERT offer a promising fusion path.
- **Robustness and Stress Testing.** Although the training and evaluation data span multiple market regimes from 2019 to 2023, including the COVID-19 period, a more fine-grained robustness analysis remains an important direction for future work. This includes explicitly disentangling crisis and post-crisis phases, assessing sensitivity to industry labeling errors, and studying the impact of retrieval noise or uncertaint-aware extensions under non- stationary market conditions.
- **Scalability and Real-Time Adaptation.** To enable real-world deployment, future variants of IA- RATD could include streaming-capable retrieval mechanisms and adaptive diffusion updates with low latency.

In summary, IA-RATD opens a new research avenue in *retrieval-augmented generative forecasting for financial time series*, bridging structural domain knowledge with flexible diffusion models. Its core ideas may generalize to other domains such as energy systems, supply-chain dynamics, and macroeconomic forecasting where contextualized temporal generation is vital.

## References

[1] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, Time Series Analysis: Forecasting and Control, 5th ed., John Wiley & Sons, 2015.

[2] S. Hochreiter and J. Schmidhuber, Long Short-Term Memory, Neural Computation, vol. 9, no. 8, pp. 1735–1780, Nov. 1997. https://doi.org/10.1162/neco.1997.9.8.1735

[3] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting, in Proceedings of the AAAI Conference on Artificial Intelligence, Feb. 2021, pp. 11106–11115.

[4] H. Wu, J. Xu, J. Wang, and M. Long, Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting, in Advances in Neural Information Processing Systems (NeurIPS), Dec. 2021.

[5] J. Ho, A. Jain, and P. Abbeel, Denoising Diffusion Probabilistic Models, in Advances in Neural Information Processing Systems (NeurIPS), Dec. 2020, pp. 6840–6851.

[6] Y. Song, J. Sohl-Dickstein, D. Kingma, A. Kumar, S. Ermon, and B. Poole, Score-Based Generative Modeling through Stochastic Differential Equations, in International Conference on Learning Representations (ICLR), 2021.

[7] L. Kong, W. Zhang, and Y. Chen, ScoreGrad: Diffusion-based Methods for Time Series Forecasting, IEEE Transactions on Pattern Analysis and Machine Intelligence, Mar. 2024. https://doi.org/10.1109/TPAMI.2024.3491121

[8] K. Rasul, S. Ashish, A. Merentitis, B. Schölkopf, and I. Valera, Autoregressive Denoising Diffusion Models for Multivariate Probabilistic Time Series Forecasting, in International Conference on Machine Learning (ICML), 2021.

[9] Z. Wang and C. Ventre, A Financial Time Series Denoiser Based on Diffusion Model, Sep. 2024. https://doi.org/10.48550/arXiv.2409.02138

[10] J. Liu, L. Yang, H. Li, and S. Hong, Retrieval-Augmented Diffusion Models for Time Series Forecasting, Oct. 2024. https://doi.org/10.48550/arXiv.2410.18712

[11] Y. Liu, T. Hu, H. Zhang, H. Wu, S. Wang, L. Ma, and M. Long, iTransformer: Inverted Transformers Are Effective for Time Series Forecasting, arXiv preprint arXiv:2310.06625, Oct. 2023.

[12] A. Zeng, M. Chen, L. Zhang, and Q. Xu, Are Transformers Effective for Time Series Forecasting?, in Proceedings of the AAAI Conference on Artificial Intelligence, Feb. 2023, pp. 11232–11240.

[13] S. Liu and Y. Li, TimeMixer: Decomposable Multiscale Mixing for Time Series Forecasting, arXiv preprint arXiv:2405.14616, 2024.

[14] F. Feng, X. He, X. Wang, C. Luo, Y. Liu, and T.-S. Chua, Temporal Relational Ranking for Stock Prediction, ACM Transactions on Information Systems, vol. 37, no. 2, pp. 1–30, Mar. 2019. https://doi.org/10.1145/3309547

[15] Y. Ye, Z. Xu, Z. Liu, and Q. Zhang, Multi-Relational Stock Ranking with Temporal Graph Convolution, in Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM), Oct. 2020.

[16] Y. Tan, P. Wu, and L. Zeng, SpacetimeGNN: Learning Spatiotemporal Dependencies for Stock Movement Forecasting, in Proceedings of the AAAI Conference on Artificial Intelligence, Feb. 2023.

[17] Z. Zhang, S. Zohren, and S. Roberts, DeepLOB: Deep Convolutional Neural Networks for Limit Order Books, IEEE Transactions on Signal Processing, vol. 67, no. 11, pp. 3001–3012, 2019. https://doi.org/10.1109/TSP.2019.2907260

[18] J. Zhang, X. Zhong, and Y. Zhang, Stockformer: A stock price prediction model using memory-augmented transformer, arXiv preprint arXiv:2009.10683, Sept. 2020.