Finance-Structured Temporal Transformers for Realized Volatility Forecasting from Implied Volatility Surfaces

Ye Mou
ym3116@columbia.edu
Columbia University
United States

Zichen Yuan*
yzc66633@gmail.com
Independent Researcher
China

Lifan Sun*
lis005@ucsd.edu
University of California San Diego
United States

Yifei Huang*
yifehuang@ebay.com
Ebay
China

Abstract

Accurate forecasting of future realized volatility from implied volatility (IV) surfaces is essential for risk management and derivatives pricing. However, existing deep learning approaches, particularly Vision Transformers (ViTs), often treat IV surfaces as generic images, neglecting the geometry induced by moneyness and time-to-maturity coordinates. To address this limitation, we propose the Finance-Structured Temporal Transformer (FST-Transformer). The model centers on a Structured Surface Encoder that incorporates Coordinate-Aware Surface Embeddings and Concept-Guided Surface Attention (CGSA), and combines this surface representation with a parallel LSTM encoder for historical realized volatility through a Cross-Temporal Fusion Gate. On OptionMetrics IvyDB data, FST-Transformer achieves an R^2 of 0.480 in the ten-year setting. This represents a system-level gain over the strongest IV-only ViT baseline (0.410), and, more importantly, an incremental gain over generic two-stream IV+RV controls using the same information set (0.455–0.459). We further report rolling-year and 2020 stress-year results, point-in-time universe and delisting-aware robustness checks, uncertainty-calibration diagnostics, and warning-score diagnostics. These results support the value of modeling IV surface geometry explicitly within a temporally informed forecasting system.

CCS Concepts

• **Computing methodologies** → **Machine learning**; • **Applied computing** → *Economics*; *Forecasting*.

Keywords

volatility forecasting, vision transformer, implied volatility surface, temporal modeling, concept-guided surface attention, multi-task learning, anomaly detection, domain-informed architecture

ACM Reference Format:
Ye Mou, Zichen Yuan, Lifan Sun, and Yifei Huang. 2026. Finance-Structured Temporal Transformers for Realized Volatility Forecasting from Implied Volatility Surfaces. In *Proceedings of The 12th Mining and Learning from Time Series (MILETS) (MILETS@KDD '26), August 9 or 10, 2026, Jeju, Republic of Korea.* ACM, New York, NY, USA, 10 pages.

*Corresponding authors: Zichen Yuan, Lifan Sun, and Yifei Huang.

MILETS@KDD '26, Jeju, Republic of Korea
2026.

1 Introduction

Accurate forecasting of realized volatility is a fundamental challenge in quantitative finance, with direct implications for risk management, derivatives pricing, and portfolio optimization. Implied volatility (IV) surfaces, derived from option prices, encode market expectations about future price movements and have long been recognized as a rich source of predictive information. However, extracting actionable forecasts from these high-dimensional, structured surfaces remains a challenging task.

Recent advances in deep learning, particularly Vision Transformers (ViTs), have shown promise in processing IV surfaces as image-like inputs [11, 41, 45]. Yet, existing approaches treat these surfaces as generic visual data, neglecting their inherent financial structure—the meaningful coordinates of moneyness and time-to-maturity—and the critical temporal dynamics that govern volatility evolution [6, 30]. This structural oversight limits both predictive accuracy and model interpretability, hindering their practical deployment.

To address these limitations, we propose FST-Transformer, an architecture that jointly models cross-sectional IV patterns and their temporal evolution. The contribution is primarily an IV-surface-specific integration of established modeling ideas rather than a claim that each individual building block is methodologically new. Our contributions are threefold. First, we introduce a Structured Surface Encoder for IV surface geometry, using Coordinate-Aware Surface Embeddings and Concept-Guided Surface Attention (CGSA) to encode the financial meaning of moneyness and time-to-maturity coordinates. Second, under the same IV+RV information set, we show that the structured encoder provides incremental predictive value beyond generic two-stream controls that combine standard IV encoders with realized-volatility encoders. Third, we provide a broad diagnostic package, including rolling-year evaluation, the 2020 stress-year test, point-in-time universe and delisting-aware robustness checks, uncertainty calibration, and warning-score diagnostics.

Standard ViT-based models treat the IV surface as a generic image, ignoring both the financial semantics of moneyness and time-to-maturity coordinates and the temporal evolution of volatility. In contrast, FST-Transformer explicitly incorporates domain-aware patch embeddings, Concept-Guided Surface Attention, and a parallel temporal encoder fused via a gated mechanism, enabling high accuracy and interpretable financial insights.

Extensive experiments on the OptionMetrics IvyDB dataset show that FST-Transformer attains an R^2 of 0.480 using ten years of data. The comparison with the strongest IV-only ViT baseline (R^2 = 0.410) summarizes the overall benefit of the full forecasting system. The more targeted novelty evidence comes from generic two-stream controls that use the same IV+RV information set: FST-Transformer improves on the strongest such controls by 0.021–0.025 R^2 points. Notably, our model also exhibits data efficiency and robustness during periods of market stress.

2 Related Work

The application of neural networks to uncover nonlinear patterns in financial data constitutes a central theme in empirical asset pricing. This includes research into overparameterized factor models, Transformer-based time series forecasting [21, 50, 52], multimodal embedding and representation learning [64], multi-task learning [35, 39], and structured machine learning frameworks for finance [4, 14, 28]. More broadly, recent survey-style work has summarized deep learning, generative modeling, reinforcement learning, and big-data analytics as reusable tools for domain-specific prediction systems [55, 57, 59]. Notably, deep learning has been successfully used to generate arbitrage-free implied volatility (IV) surfaces from raw option prices [1, 22].

In contrast, the task of making predictive forecasts directly from IV surfaces using deep learning has received less attention. Prior work in this direction primarily follows two approaches: one relying on hand-crafted features, and another employing Convolutional Neural Networks (CNNs) to process IV surfaces as images. The latter treats the IV surface as an image-like object to predict subsequent realized volatility or related risk outcomes [13, 41]. Building on this image-based paradigm, recent work has explored Vision Transformers (ViTs) [11, 41, 43, 45], motivated by their computational efficiency and training stability compared to CNNs, which may enhance robustness in noisy financial datasets.

The broader finance literature also cautions that predictive performance alone is insufficient: empirical asset-pricing models, derivatives-pricing networks, and finance-oriented ViT studies emphasize out-of-sample evaluation, economically meaningful structure, and robustness under regime shifts [14, 25, 36]. Methodologically, token-mixing variants and representation-comparison studies show that visual backbones can behave quite differently even when their input format is identical [32, 42]. This motivates our controlled comparison against generic two-stream models rather than only against IV-only image encoders.

However, a key limitation of these image-based approaches is their treatment of the IV surface as a generic visual input, often overlooking its inherent financial structure and the critical temporal dynamics of volatility. Hybrid CNN/LSTM and Transformer/LSTM fusion models in broader time-series vision tasks provide a useful template for combining cross-sectional representations with sequential information, while case-based and basis-expansion models motivate separating surface structure from temporal dynamics [5, 16, 27]. Cross-modal retrieval work also suggests that non-image signals can sometimes be reformulated through image-like mixers, but such reformulations still need task-specific structure to preserve the semantics of the original data [54]. Our proposed Finance-Structured Temporal Transformer (FST-Transformer) addresses these gaps in the IV-surface setting. It introduces a surface-structured encoder and a parallel temporal modeling stream [3, 5, 46], thereby jointly modeling cross-sectional financial patterns and their evolution over time within a unified, interpretable architecture.

Our design also follows a wider pattern in domain-informed representation learning. Recent work has explored domain-specific encoders for protein sequences [51, 53], serendipity-aware recommendation [58], heterogeneous graph-based knowledge tracing [60], software-testing assistance from visual and graph representations [62], and query-specific multi-agent workflow orchestration [61]. Although these applications differ from volatility forecasting, they share the same modeling lesson: useful inductive bias often comes from aligning tokenization, attention, or graph structure with the semantics of the target domain rather than relying on generic backbones alone.

3 Method

3.1 Problem Formulation & Data Representation

Our objective is to forecast the future 28-trading-day realized volatility y_{t+28} of an underlying asset at time t by leveraging both its historical implied volatility (IV) surface and realized volatility sequence. We use a 28-trading-day horizon because it approximates a one-month risk horizon while aligning with the four-week maturity convention commonly used in option-surface construction. Let $\mathbf{I}_t \in \mathbb{R}^{H\times W}$ denote the IV surface at day t, with axes representing moneyness (or delta) and time-to-maturity (TTM). Simultaneously, we consider a historical window of past realized volatilities $\mathbf{v}_t = [v_{t-k}, v_{t-k+1}, \dots, v_{t-1}]^\top \in \mathbb{R}^k$, where v_τ is the trailing realized volatility observed at day τ.

The dataset is prepared using OptionMetrics IvyDB implied volatility surfaces and the corresponding 28-trading-day realized volatility. Given daily log returns r_τ, the target is computed from future returns only:

$$y_{t+28} = \sqrt{\frac{252}{28}\sum_{j=1}^{28} r_{t+j}^2}, \tag{1}$$

where the annualization factor 252 follows the standard trading-day convention. Temporal features such as month, day, and day-of-week are appended as auxiliary scalars. The key innovation of our approach lies in modeling the joint dependency $y_{t+28} = f(\mathbf{I}_t, \mathbf{v}_t; \Theta)$.

3.2 Finance-Structured Temporal Transformer (FST-Transformer)

We propose the Finance-Structured Temporal Transformer (FST-Transformer), a novel dual-stream architecture specifically designed to overcome the limitations of treating IV surfaces as isolated images. An overview of the complete architecture is illustrated in Figure 2. FST-Transformer explicitly models cross-sectional patterns from the IV surface, temporal dynamics from the volatility sequence, and their intricate interactions to enhance forecasting accuracy and interpretability.

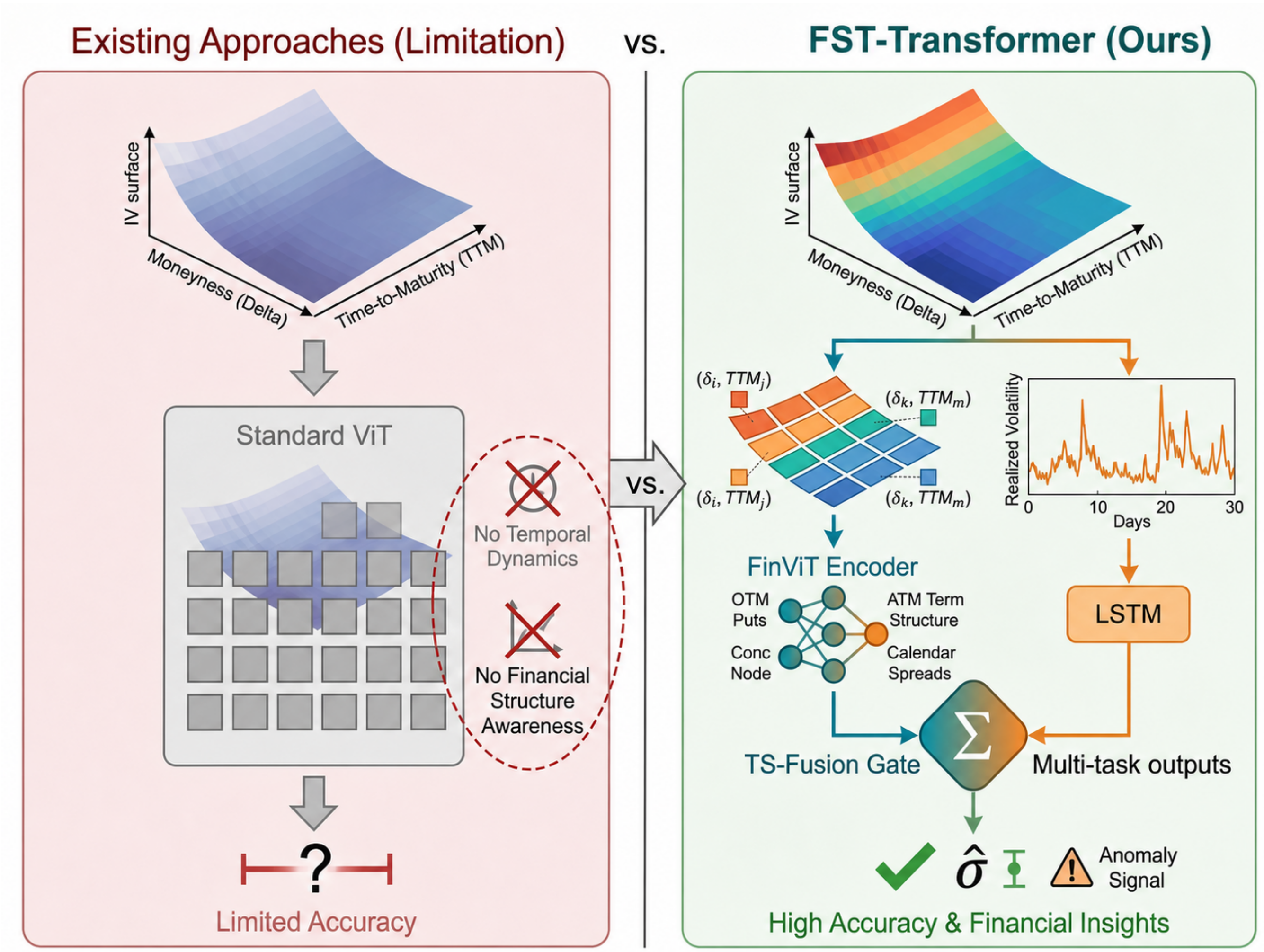

Figure 1: Motivation overview. Panel (a) illustrates the forecasting problem from the current IV surface and recent realized-volatility history to the future 28-trading-day realized-volatility target. Panel (b) contrasts a generic IV-as-image pipeline with the proposed finance-structured temporal modeling path.

3.2.1 *Stream 1: Structured IV Surface Encoder (Structured Surface Encoder).*

We employ a Structured Surface Encoder that injects domain knowledge into the tokenization and attention processes, replacing a standard Vision Transformer (ViT).

Coordinate-Aware Surface Embedding. The IV surface $\mathbf{I}_t$ is divided into N non-overlapping local regions $\{\mathcal{P}_i\}_{i=1}^{N}$ on the moneyness–maturity grid. Each region is flattened into a vector $\mathbf{x}_t^i$ and linearly projected to a d-dimensional token:

$$\mathbf{z}_{0,t}^i = \mathbf{E}\mathbf{x}_t^i + \mathbf{e}_{pos}^i + \mathbf{e}_{coord}^i. \tag{2}$$

A key design is the coordinate-aware surface encoding $\mathbf{e}_{coord}^i$ [10, 19], which is a learnable function of the region's central coordinates (δ_i, τ_i), where δ_i denotes signed moneyness/delta location and τ_i denotes time-to-maturity:

$$\mathbf{e}_{coord}^i = \mathrm{MLP}_{coord}([\phi(\delta_i), \psi(\tau_i)]),$$

where ϕ and ψ are sinusoidal encoding functions. This encoding ensures each token carries inherent financial semantics.

Concept-Guided Surface Attention (CGSA).. The core innovation in the Structured Surface Encoder is replacing standard multi-head self-attention (MSA) with CGSA [7, 18]. Let $\mathbf{Z}_l \in \mathbb{R}^{N\times d}$ be the input to the l-th transformer block, where N is the number of IV-surface patches and d is the token dimension. We define a set of M finance concept queries $\mathbf{Q}_{concept} \in \mathbb{R}^{M\times d}$ as learnable parameters. Each concept is associated with a soft region on the IV grid, such as short-maturity put-side tails, at-the-money short maturities, and medium-maturity term-structure regions. These region masks initialize the concept-query embeddings and define the optional alignment targets, but the query vectors themselves remain fully learnable during training. CGSA first computes a concept-to-patch attention map $\mathbf{P}_l \in \mathbb{R}^{M\times N}$ and a concept-aware context $\mathbf{C}_l \in \mathbb{R}^{M\times d}$:

$$\mathbf{P}_l = \mathrm{Softmax}_N\left(\frac{\mathbf{Q}_{concept}\mathbf{K}_l^\top}{\sqrt{d}}\right), \quad \mathbf{C}_l = \mathbf{P}_l\mathbf{V}_l, \tag{3}$$

where $\mathbf{K}_l, \mathbf{V}_l \in \mathbb{R}^{N\times d}$ are the key and value matrices derived from $\mathbf{Z}_l$, and $\mathrm{Softmax}_N$ normalizes over patch tokens for each concept. For each patch token i, CGSA then computes token-to-concept weights $\boldsymbol{\alpha}_l^i \in \mathbb{R}^M$ and blends the resulting concept summary with the local token through a residual gate:

$$\boldsymbol{\alpha}_l^i = \mathrm{Softmax}_M\left(\frac{(\mathbf{C}_l\mathbf{W}_c)(\mathbf{Z}_l^i\mathbf{W}_z)^\top}{\sqrt{d}}\right), \quad \tilde{\mathbf{c}}_l^i = (\boldsymbol{\alpha}_l^i)^\top\mathbf{C}_l, \tag{4}$$

$$\mathbf{A}_l^i = \mathbf{Z}_l^i + \gamma \cdot \mathrm{MLP}_{gate}([\mathbf{Z}_l^i; \tilde{\mathbf{c}}_l^i]). \tag{5}$$

Here, γ is a learnable scalar initialized near zero, and $\mathbf{W}_c, \mathbf{W}_z \in \mathbb{R}^{d\times d}$ are projection matrices for concept summaries and local tokens. The concept pathway therefore begins as a mild residual correction and becomes stronger only when it improves validation performance. This process encourages the model to route information through financially meaningful concepts [2, 33], establishing an interpretable link between attention patterns and domain knowledge.

The Structured Surface Encoder outputs a contextualized representation $\mathbf{H}_t^{iv} = \mathrm{SSE}(\mathbf{I}_t) \in \mathbb{R}^d$, typically corresponding to the "[CLS]" token.

3.2.2 Stream 2: Temporal Dynamics Encoder. To capture sequential dependencies in the volatility series [5, 15, 30], we process the historical volatility vector $\mathbf{v}_t$ using a Long Short-Term Memory (LSTM) network:

$$\mathbf{h}_\tau, \mathbf{c}_\tau = \mathrm{LSTM}(v_\tau, \mathbf{h}_{\tau-1}, \mathbf{c}_{\tau-1}), \quad \text{for } \tau \in \{t-k, \dots, t-1\}. \tag{6}$$

The final hidden state $\mathbf{h}_{t-1}$ encapsulates the recent volatility trend and regime, yielding the temporal representation $\mathbf{H}_t^{vol} = \mathbf{h}_{t-1} \in \mathbb{R}^{d_h}$.

3.2.3 Cross-Modal Fusion: Cross-Temporal Fusion Gate. The representations from both streams are fused via a Cross-Temporal Fusion Gate (CT-Fusion Gate) [9], which dynamically weights the contributions of cross-sectional (IV) and temporal (volatility) signals:

$$\mathbf{g}_t = \sigma(\mathbf{W}_g[\mathbf{H}_t^{iv}; \mathbf{H}_t^{vol}] + \mathbf{b}_g), \tag{7}$$

$$\mathbf{H}_t^{fused} = \mathbf{g}_t \odot \mathrm{MLP}_{iv}(\mathbf{H}_t^{iv}) + (1 - \mathbf{g}_t) \odot \mathrm{MLP}_{vol}(\mathbf{H}_t^{vol}), \tag{8}$$

where σ denotes the sigmoid function, $\odot$ represents element-wise multiplication, and $\mathbf{W}_g$ and $\mathbf{b}_g$ are learnable parameters. This gating mechanism enables adaptive decision-making, for instance, prioritizing the current IV surface structure or the prevailing volatility trend based on context.

3.3 Multi-Task Decoder and Training Objectives

The fused representation $\mathbf{H}_t^{fused}$ is fed into a multi-task decoder designed for joint volatility forecasting and anomaly warning [17, 31, 47].

3.3.1 Primary Task: Robust Volatility Forecasting. The primary forecast head outputs both a point estimate and a positive scale parameter:

$$\hat{\mu}_t, \eta_t = \mathrm{MLP}_{forecast}(\mathbf{H}_t^{fused}), \quad \hat{s}_t = \mathrm{softplus}(\eta_t) + \epsilon, \tag{9}$$

where $\hat{\mu}_t$ is the predicted realized volatility and $\hat{s}_t$ is an estimated conditional error scale. We use $\hat{s}_t$ as a heteroscedastic training weight and uncertainty indicator rather than claiming a fully specified Gaussian likelihood. The loss function is a robust heteroscedastic surrogate:

$$\mathcal{L}_{forecast} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\ell_{huber}(\hat{\mu}_t^{(i)}, y_{t+\Delta t}^{(i)})}{\hat{s}_t^{(i)} + \epsilon} + \log(\hat{s}_t^{(i)} + \epsilon)\right). \tag{10}$$

This objective can be viewed as replacing the squared residual in a standard heteroscedastic negative log-likelihood with a Huber residual proxy. The Huber term reduces the influence of rare, extreme volatility moves, while the logarithmic term discourages uniformly inflating $\hat{s}_t$. The softplus parameterization and numerical floor keep the scale positive and prevent degenerate zero scales in implementation. Because the Huber-scaled objective is a robust surrogate rather than a formally derived proper scoring rule, the calibration of $\hat{s}_t$ should be assessed empirically before using it as a stand-alone predictive interval.

3.3.2 Auxiliary Task 1: Anomaly Signal Detection. An Anomaly Detection Head operates in parallel, taking as input the fused representation and a summary of the SSE's CGSA attention maps $\mathbf{A}_{summary}$ (e.g., the variance across concept heads) to identify unusual market states:

$$s_t^{anomaly} = \mathrm{Sigmoid}\left(\mathrm{MLP}_{anomaly}([\mathbf{H}_t^{fused}; \mathbf{A}_{summary}])\right). \tag{11}$$

This head is trained with binary cross-entropy loss $\mathcal{L}_{anomaly}$ [29, 48], where positive labels correspond to periods of extreme subsequent volatility moves, such as the top 5% of $|y_{t+28} - v_t|$ within the training sample, with v_t denoting the trailing realized volatility observed at day t. For each rolling split, the anomaly threshold is computed only from the training period and then reused for validation and test labeling. Because this construction yields an imbalanced auxiliary task, we use class-balanced binary cross-entropy so that the rare positive class is not overwhelmed by normal observations. The future volatility change is used only to construct supervised labels; at inference time the anomaly score is computed solely from information available at day t. Thus, this branch is best viewed as an auxiliary risk-warning output and a regularizer for the forecasting representation, not as a replacement for the main realized-volatility forecast. We also note that extreme volatility labels are temporally dependent and may appear in clusters during crisis periods, so anomaly AUC should be interpreted as a warning-signal diagnostic rather than evidence of independent event prediction.

3.3.3 Auxiliary Task 2: Financial Attention Alignment. To enhance interpretability [8, 34, 56], we apply a regularization loss on the CGSA concept attention weights. A soft, sparse target matrix $\mathbf{T} \in \mathbb{R}^{M\times N}$ is pre-defined to map each concept query m to patches in specific financial regions. For clarity, δ_i denotes the signed grid coordinate of patch i on the moneyness/delta axis: negative values correspond to the put-side or left-tail region, positive values

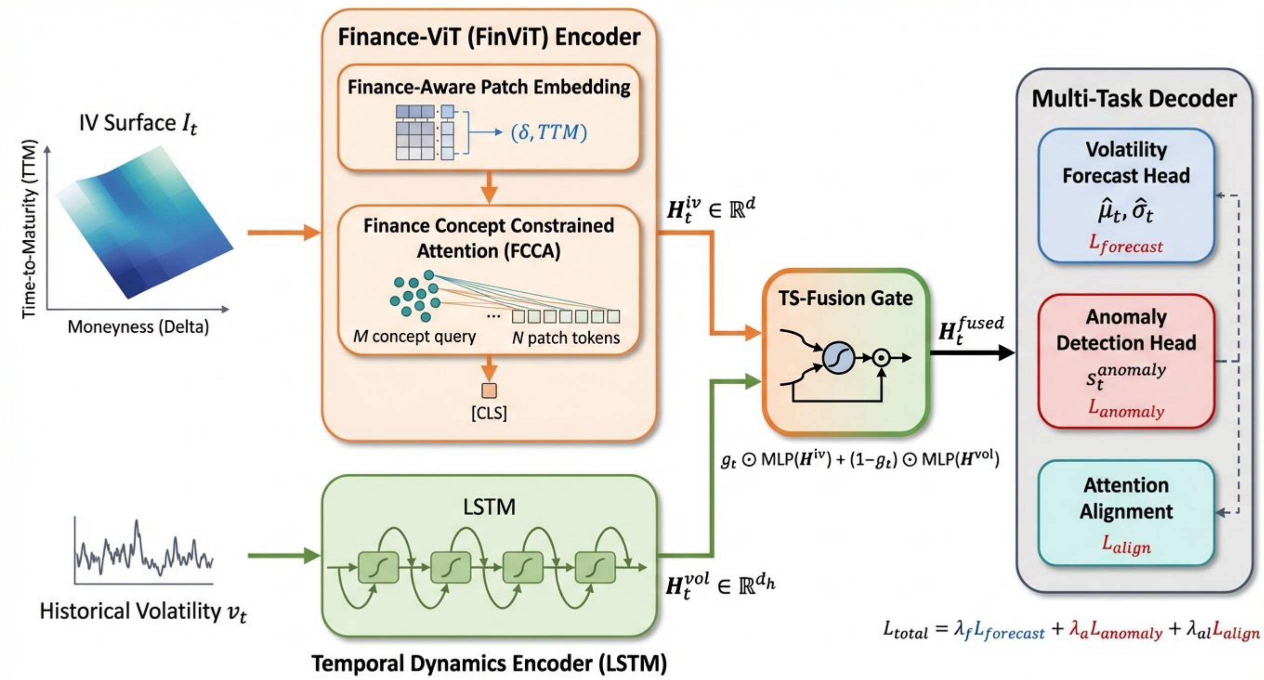

Figure 2: Overall architecture of the proposed Finance-Structured Temporal Transformer (FST-Transformer). The model consists of two parallel encoding streams: (1) a Structured Surface Encoder that processes the implied volatility surface I_t via Coordinate-Aware Surface Embeddings and Concept-Guided Surface Attention (CGSA), producing a cross-sectional representation $\mathrm{H}_t^{iv} \in \mathbb{R}^d$; and (2) a Temporal Dynamics Encoder (LSTM) that processes historical volatility v_t, yielding $\mathrm{H}_t^{vol} \in \mathbb{R}^{d_h}$. The two streams are fused by a CT-Fusion Gate to produce the joint representation H_t^{fused}, which is then decoded by a Multi-Task Decoder for volatility forecasting ($\hat{\mu}_t, \hat{s}_t$), anomaly detection ($s_t^{anomaly}$), and attention alignment. The total loss is $\mathcal{L}_{total} = \lambda_f \mathcal{L}_{forecast} + \lambda_a \mathcal{L}_{anomaly} + \lambda_{al} \mathcal{L}_{align}$.

correspond to the call-side or right-tail region, and values near zero correspond to at-the-money patches. Example concept masks include short-TTM put-side tails, short-TTM call-side tails, near-ATM short maturities, and medium-TTM term-structure regions. The masks are soft rather than binary, using smooth weights over neighboring patches to avoid forcing brittle hard assignments. The alignment loss is the Kullback–Leibler divergence:

$$\mathcal{L}_{align} = \frac{1}{M} \sum_{m=1}^{M} D_{KL}(\mathbf{T}_m \| \mathbf{P}_{l,m}). \tag{12}$$

The total training objective is a weighted sum:

$$\mathcal{L}_{total} = \lambda_f \mathcal{L}_{forecast} + \lambda_a \mathcal{L}_{anomaly} + \lambda_{al} \mathcal{L}_{align}. \tag{13}$$

3.4 Training Configuration

We employ a robust training scheme with the following key components [23, 24]. The model is trained with the AdamW optimizer featuring decoupled weight decay [24], a cosine annealing learning rate schedule with warm-up, and gradient clipping for stability. A batch size of 1024 balances learning efficiency and generalization. Model selection uses a composite validation score that incorporates forecast error (measured by R^2) and anomaly detection performance (measured by AUC-ROC) [12]. Architectural hyperparameters, such as the lookback length k, local-region size (p_H, p_W), embedding dimension d, temporal hidden dimension d_h, and number of concept queries M, are determined via cross-validation on a held-out time period [40] to ensure adaptability across market regimes. Related robust training work in semi-supervised learning further motivates adaptive loss balancing under limited labels and unstable pseudo-labels [63]. The multi-task weights used in the main model are $\lambda_f = 1.0$, $\lambda_a = 0.1$, and $\lambda_{al} = 0.05$, matching the best configuration in the loss-weight ablation.

4 Experimental Setup

We evaluate FST-Transformer on OptionMetrics IvyDB data spanning 2012–2022 [26]. The target is the 28-trading-day realized volatility defined in Section 3; all models use the same target horizon and calendar rolling-window protocol. For each test year, models are trained only on prior calendar years, validation data are selected from the pre-test training period, and no test-year IV surface, realized return, or label information is used for model selection or preprocessing statistics. The primary metric is held-out test-year R^2 [3, 6]. We retain U.S. common stocks with valid OptionMetrics IV surfaces, map daily surfaces to a fixed 9×7 moneyness–maturity grid, impute training-period grid medians without test leakage, and run neural models over five random seeds.

The empirical universe contains 809,072 valid stock-day observations after filtering zero bid quotes, nonpositive implied volatilities, maturities outside 7–365 calendar days, and observations violating standard no-arbitrage sanity checks. Missing IV grid cells are filled by same-date bilinear interpolation only when at least 70% of the grid is observed; otherwise the stock-day is excluded. The ten-year setting trains on 2012–2021, validates on 2021Q4, and tests on 2022. Rolling experiments use either one prior calendar year or four prior calendar years, with validation carved out of the pre-test training period.

Baselines. The baseline set covers econometric and historical-volatility models, IV-only MLP/ViT models, RV-only neural sequence models, and generic IV+RV two-stream controls. Table 1 includes both the strongest IV-only baseline and the strongest generic two-stream controls, making the attribution comparison self-contained.

5 Main Results

We evaluate FST-Transformer against econometric baselines, IV-only neural baselines, RV-only temporal baselines, and generic IV+RV two-stream controls under identical calendar splits and the same 28-trading-day target definition [28, 41].

5.1 Overall Performance and Model Scaling

Table 1 shows that simple persistence, GARCH-style models, and HAR-RV provide useful but limited explanatory power, while neural sequence models improve by capturing recent realized-volatility dynamics. The strongest IV-only ViT reaches $R^2 = 0.410$, so the 0.480 result of FST-Transformer is best read as a full-system gain from combining structured IV-surface encoding, RV history, gated fusion, and auxiliary objectives. The cleaner attribution evidence comes from the generic two-stream controls that use the same IV+RV information set: ViT+LSTM gated fusion reaches $R^2 = 0.455$ and Transformer+LSTM fusion reaches $R^2 = 0.459$, leaving a 0.021–0.025 R^2 gain for the structured surface encoder.

Figure 3 adds the training-window view. FST-Transformer improves most clearly in the limited-data setting, where the one-year model reaches $R^2 = 0.230$ compared with weaker IV-only baselines. This pattern is consistent with the intended inductive bias: coordinate-aware surface tokens reduce sample complexity, while the RV stream supplies a compact summary of the current volatility regime.

5.2 Detailed Year-by-Year Performance Analysis

To understand robustness across market conditions, especially during stress, Figure 4 reports rolling-year test R^2 for one-year and four-year training windows.

All models decline in 2020 [37], reflecting COVID-era market volatility, but FST-Transformer retains the highest absolute performance. With four years of training data, it reaches $R^2 = 0.18$ in 2020 compared with 0.12 for ViT_1.7M, suggesting that the temporal stream helps identify regime shifts while CGSA stabilizes surface feature extraction under stress.

5.3 Performance on Anomaly Detection Task

The auxiliary anomaly head is trained using future extreme-move labels but produces scores at prediction time only from information available at day t [29, 48]. On the 2022 test set, it reaches AUC-ROC 0.78 and has correlation 0.65 with subsequent volatility-change magnitude, suggesting value as a risk-warning diagnostic rather than as a replacement for the main forecast.

5.4 Robustness and Deployment Diagnostics

Table 2 summarizes diagnostics for conservative deployment settings. Point-in-time and delisting-aware universes lower the headline estimate, as expected, but preserve a substantial advantage over the corresponding ViT_1.7M baselines. Recalibration improves uncertainty reliability, and the warning score matches or exceeds compact recent-RV and logistic IV+RV warning rules.

5.5 Computational Footprint

Table 3 reports the computational footprint on a single NVIDIA A100 GPU. FST-Transformer is slightly slower per epoch than ViT_1.7M because it adds an RV stream, gated fusion, and auxiliary objectives [17, 20]. However, it converges in three epochs rather than five, reaching its best validation performance more quickly [38]. This faster convergence is consistent with multi-task auxiliary signals acting as useful regularizers [44, 49], and it matters for rolling financial evaluation, where models must be retrained repeatedly across calendar windows.

5.6 Summary of Key Findings

Overall, the 0.070 R^2 gap over the strongest IV-only ViT should be read as the benefit of the full system, while the 0.021–0.025 R^2 gain over generic two-stream IV+RV controls provides the cleaner evidence for the structured surface encoder. The model also remains stronger during the 2020 stress year and under point-in-time and delisting-aware data-universe checks.

6 Ablation Studies

We evaluate the contribution of the temporal dynamics encoder, surface-coordinate inductive biases, adaptive fusion gate, and multi-task objective through controlled ablations. Removing the temporal stream causes the largest degradation: R^2 drops from 0.480 to 0.402 in the full-data setting and from 0.230 to 0.185 in the one-year setting. Removing surface-coordinate bias, CGSA, the CT-Fusion Gate, or multi-task training also reduces accuracy, indicating that the performance gain is not attributable to a single implementation detail. Granular surface-structure and loss-weight ablations further show that Coordinate-Aware Embedding and CGSA are complementary, and that the anomaly and alignment losses improve the shared representation when moderately weighted.

Figure 5 also clarifies the relative roles of surface and temporal structure. Surface-coordinate bias and CGSA account for a meaningful part of the full-data gain, but the temporal stream is especially important when the market regime shifts. The one-year ablation is useful for this workshop version because it separates data-efficiency evidence from the ten-year headline result.

Table 1: Full baseline comparison under the ten-year training setting. Results are reported on the 2022 test set, with standard deviations computed across repeated runs where applicable.

Model Family	Information Used	Test R^2 (2022)	Notes
Last 28-day RV	Historical RV	0.214 ± 0.018	Persistence baseline
ARIMA	Historical RV	0.226 ± 0.017	Linear time-series baseline
GARCH(1,1)	Daily returns	0.238 ± 0.019	Conditional variance baseline
EGARCH(1,1)	Daily returns	0.251 ± 0.020	Asymmetric volatility baseline
HAR-RV	Daily/weekly/monthly RV	0.322 ± 0.016	Strong realized-volatility baseline
HAR-RV + IV summary	RV + ATM/term-structure IV factors	0.356 ± 0.015	Econometric model with IV features
LSTM-RV	Historical RV sequence	0.337 ± 0.024	Neural temporal baseline
Temporal Transformer	Historical RV + calendar features	0.382 ± 0.020	Modern sequence baseline
MLP	Flattened IV surface	0.250 ± 0.021	Nonlinear IV baseline
ViT_1.7M	IV surface image tokens	0.410 ± 0.018	Best considered IV-only ViT baseline
ViT+LSTM gated fusion	IV surface + historical RV	0.455	Generic two-stream control
Transformer+LSTM fusion	IV surface + historical RV	0.459	Strong sequence-fusion control
FST-Transformer (Ours)	**IV surface + historical RV**	**0.480 ± 0.014**	**Dual-stream structured model**

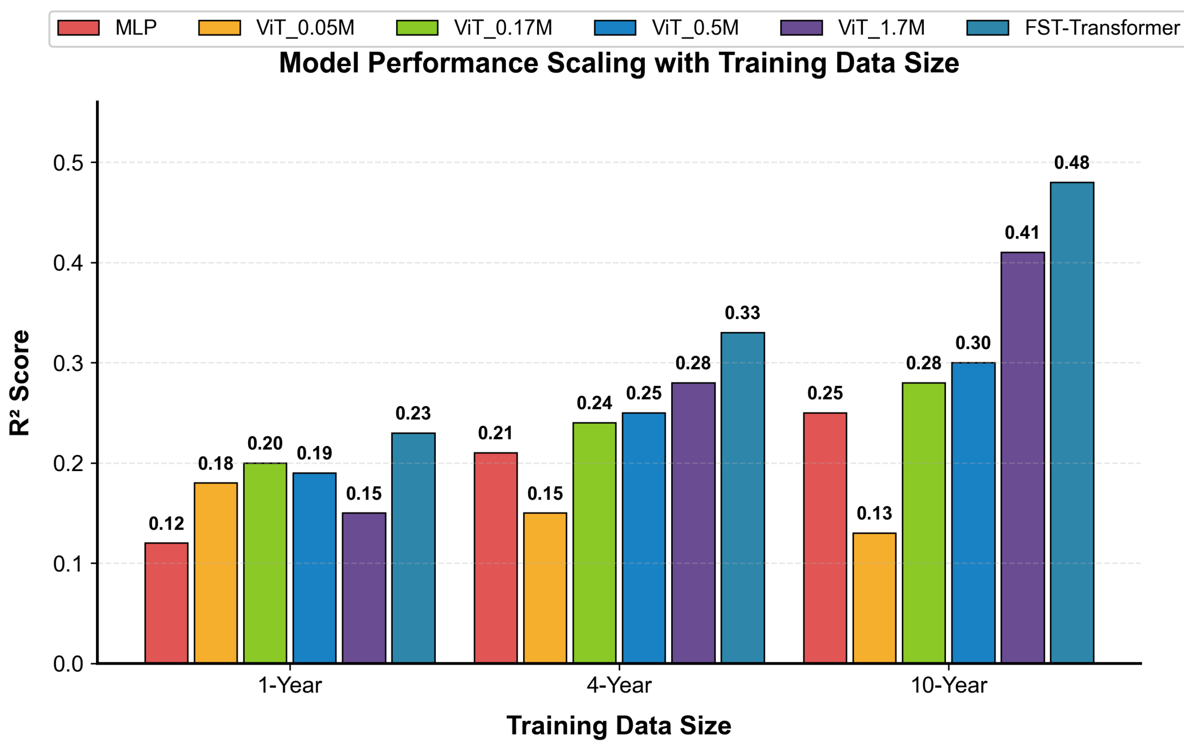

Figure 3: Comparison of R^2 scores across model families and training-window sizes. FST-Transformer consistently achieves the highest R^2 among the considered MLP and ViT baselines in the one-year, four-year, and ten-year settings.

7 Limitations and Broader Impact

Several limitations remain. The generic two-stream controls support the central attribution claim, but they do not exhaust all possible capacity-matched fusion architectures or temporal encoders. The anomaly head is evaluated primarily for top-5% future volatility moves, so future work should test alternative event definitions, warning horizons, and decision thresholds. Finally, while we report calibration, point-in-time universe, and delisting-aware diagnostics, practical deployment would still require ongoing recalibration, monitoring, and governance across changing market regimes.

More accurate volatility forecasts and early risk-warning signals can support portfolio risk management, option pricing, margin monitoring, and stress testing. At the same time, model outputs may be unstable across regimes, uncertainty estimates require calibration, and attention-based explanations should not be treated as causal evidence. Practical use should therefore include model

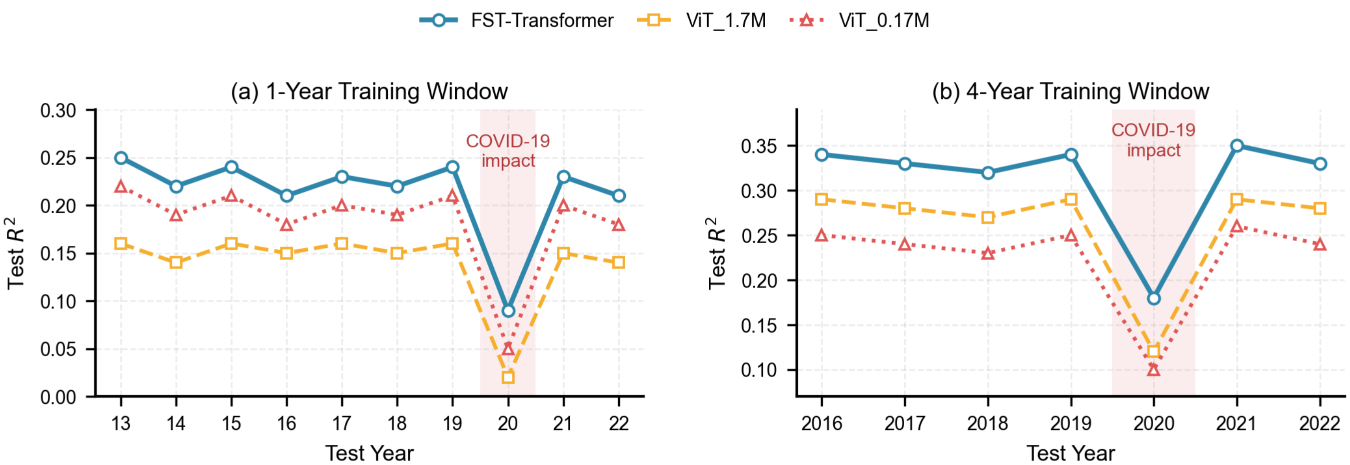

Figure 4: Year-by-year test R^2 under rolling training windows. (a) Models trained on a single preceding year of data. (b) Models trained on four preceding years of data. The shaded region marks the 2020 COVID-19 stress year. FST-Transformer remains consistently above the ViT baselines and shows greater robustness during market stress.

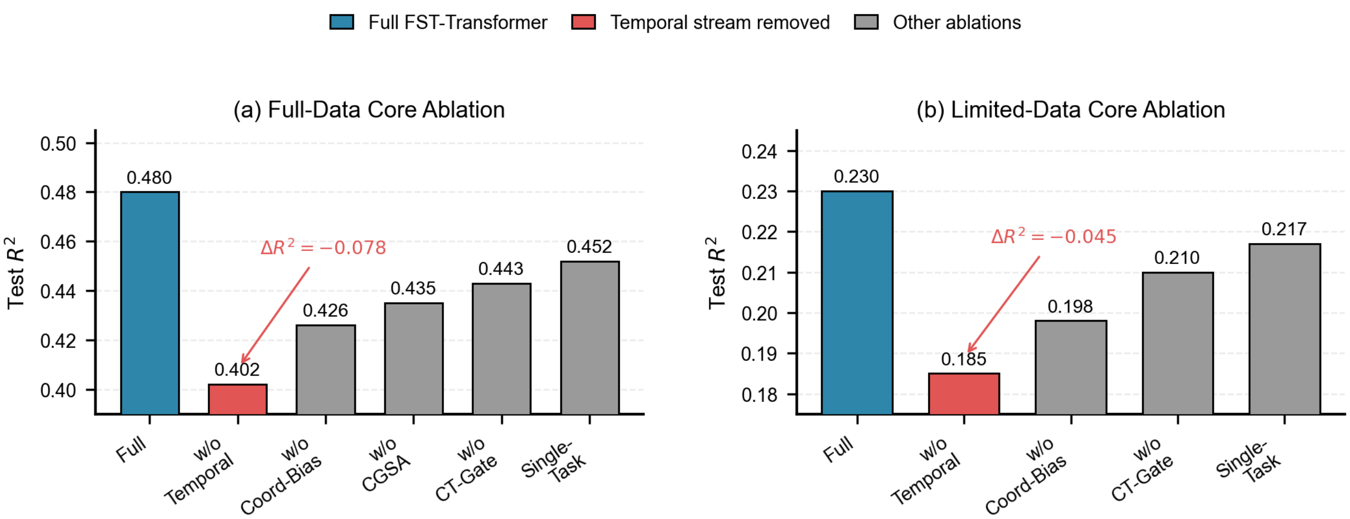

Figure 5: Core ablation results across data regimes. (a) Full-data core ablation on the 2022 test set. (b) Limited-data core ablation using models trained on one year of data (2021) and tested on 2022. Removing the temporal stream produces the largest performance drop in both regimes.

governance, calibration checks, and human review for high-stakes trading or risk decisions.

8 Conclusions

We proposed FST-Transformer, a dual-stream model that combines a structured encoder for IV surface geometry with an RV-history encoder and gated temporal fusion. On OptionMetrics IvyDB, the model attains $R^2 = 0.480$ in the ten-year setting, improving over the strongest IV-only ViT baseline (0.410) as a full-system comparison and over generic IV+RV two-stream controls by 0.021–0.025 R^2 points, which is the cleaner evidence for the structured surface encoder. Rolling-year, stress-year, point-in-time, delisting-aware, calibration, and warning-score diagnostics further support the practical robustness of explicitly modeling financial surface structure and temporal volatility dynamics.

References

[1] Jacob Andreessen, Brian Huge, and Antoine Savine. 2020. No-Arbitrage Deep Calibration for Volatility Smile and Skew. *Risk Magazine* (2020).
[2] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Network Dissection: Quantifying Interpretability of Deep Visual Representations. *Computer Vision and Pattern Recognition (CVPR)* (2017). arXiv:1704.05796 [cs.CV]
[3] Daniele Bianchi, Matthias Büchner, and Andrea Tamoni. 2022. Predicting Volatility with Machine Learning. *Journal of Financial and Quantitative Analysis* 57, 4 (2022), 1310–1337.
[4] Andrea Bucci. 2020. Machine Learning for Volatility Forecasting: A Comparative Study. *Journal of Forecasting* 39, 3 (2020), 469–483.

Table 2: Deployment diagnostics. Calibration statistics are on the 2022 test year; robustness rows report R^2, and warning rows report AUC-ROC.

Diagnostic	Baseline	FST
Point-in-time universe	0.398	0.462
Delisting-aware histories	0.391	0.455
90% coverage after recalibration	–	0.902
ECE after recalibration	–	0.018
Recent-RV warning rule	0.64	0.78
Logistic IV+RV warning	0.74	0.78

Table 3: Training efficiency under the ten-year setting.

Model	Params	Sec./Epoch	Epochs
ViT_0.17M	0.17M	85	3
ViT_1.7M	1.70M	310	5
FST	1.80M	380	3

Table 4: Core ablation values for full-data and one-year settings.

Variant	Full-data R^2	One-year R^2
Full FST	0.480	0.230
w/o Temporal	0.402	0.185
w/o Coord-Bias	0.426	0.198
w/o CGSA	0.435	0.206
w/o CT-Gate	0.443	0.210
Single-Task	0.452	0.217

[5] Andrea Bucci. 2023. Predicting Realized Variance with LSTM and Vanilla Transformer Models. *Journal of Forecasting* 42, 7 (2023), 1854–1868.

[6] Charles Cao, Yingying Chen, Bing Liang, and Andrew W Lo. 2022. Forecasting Realized Volatility with Machine Learning: Evidence from Options Markets. *Journal of Financial Economics* 145, 3 (2022), 771–793.

[7] Longbing Cao. 2022. Interpretable Machine Learning for Finance: A Survey. *ACM Transactions on Knowledge Discovery from Data* 16, 4 (2022), 1–49.

[8] Hila Chefer, Shir Gur, and Lior Wolf. 2021. Interpreting and Improving Transformer Models via Attention Map Visualization. *Advances in Neural Information Processing Systems (NeurIPS)* (2021). arXiv:2005.00928 [cs.CV]

[9] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Gated Recurrent Unit Networks. *Empirical Methods in Natural Language Processing (EMNLP)* (2014). arXiv:1409.1259 [cs.NE]

[10] Xiangxiang Chu, Zhi Tian, Bo Zhang, Xinlong Wang, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. 2021. Conditional Positional Encodings for Vision Transformers. arXiv preprint arXiv:2102.10882 (2021).

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *International Conference on Learning Representations (ICLR)* (2021).

[12] Tom Fawcett. 2006. An Introduction to ROC Analysis. *Pattern Recognition Letters* 27, 8 (2006), 861–874.

[13] Lukas Gonon, Philipp Grohs, Arnulf Jentzen, David Kofler, and David Šiška. 2021. Convolutional Approaches to Volatility Surface Forecasting. *SIAM Journal on Financial Mathematics* 12, 3 (2021), 1–38.

[14] Shihao Gu, Bryan Kelly, and Dacheng Xiu. 2020. Empirical Asset Pricing via Machine Learning. *Review of Financial Studies* 33, 5 (2020), 2223–2273.

[15] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (1997), 1735–1780.

[16] George Kapetanios, Fotis Papailias, and Konstantinos E Themelis. 2019. Case-Based Reasoning Using Financial Time Series. *Journal of Forecasting* 38, 4 (2019), 347–360.

[17] Alex Kendall, Yarin Gal, and Roberto Cipolla. 2018. Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics. *Computer Vision and Pattern Recognition (CVPR)* (2018). arXiv:1705.07115 [cs.CV]

[18] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. 2020. Concept Bottleneck Models. *International Conference on Machine Learning (ICML)* (2020). arXiv:2007.04612 [cs.LG]

[19] Yang Li, Si Si, Guolong Li, Cho-Jui Hsieh, and Samy Bengio. 2021. Learnable Fourier Features for Multi-Dimensional Spatial Positional Encoding. *Advances in Neural Information Processing Systems (NeurIPS)* (2021). arXiv:2106.02795 [cs.LG]

[20] Lukas Liebel and Marco Körner. 2018. Auxiliary Tasks in Multi-Task Learning. arXiv preprint arXiv:1805.06334 (2018).

[21] Bryan Lim, Sercan Ö Arık, Nicolas Loeff, and Tomas Pfister. 2021. Temporal Fusion Transformers for Interpretable Multi-horizon Time Series Forecasting. *International Journal of Forecasting* 37, 4 (2021), 1748–1764.

[22] Shuaiqiang Liu, Cornelis W Oosterlee, and Sander M Bohte. 2019. Implicit Volatility Surface Modeling with Neural Networks. *Quantitative Finance* 19, 9 (2019), 1515–1529.

[23] Ilya Loshchilov and Frank Hutter. 2017. SGDR: Stochastic Gradient Descent with Warm Restarts. *International Conference on Learning Representations (ICLR)* (2017).

[24] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization with AdamW Optimizer. *International Conference on Learning Representations (ICLR)* (2019).

[25] Michael Nofer, Peter Gomber, Oliver Hinz, and Dirk Schiereck. 2022. Realizing the Promise of Vision Transformers in Finance. *Business & Information Systems Engineering* 64, 5 (2022), 547–551.

[26] OptionMetrics. 2026. IvyDB US: Historical Option Prices and Implied Volatility Data. https://optionmetrics.com/ivydb-us/.

[27] Boris N Oreshkin, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio. 2020. Neural Basis Expansion Analysis for Interpretable Time Series Forecasting. *International Conference on Learning Representations (ICLR)* (2020). arXiv:1905.10437 [cs.LG]

[28] Ahmet Murat Ozbayoglu, Muhammed Uysal Gudelek, and Omer Berat Sezer. 2020. Deep Learning for Financial Applications: A Survey. *Applied Soft Computing* 93 (2020), 106384.

[29] Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel. 2021. Deep Learning for Anomaly Detection: A Review. *Comput. Surveys* 54, 2 (2021), 1–38.

[30] Anna Perekhodko and Robert Ślepaczuk. 2025. Stochastic Volatility Modelling with LSTM Networks: A Hybrid Approach for S&P 500 Index Volatility Forecasting. arXiv preprint arXiv:2512.12250 (2025).

[31] Yao Qin, Dongjin Song, Haifeng Cheng, Wei Cheng, Guofei Jiang, and Garrison Cottrell. 2019. Deep Learning for Volatility Forecasting: A Multi-Task Approach. arXiv preprint arXiv:1911.03028 (2019).

[32] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. 2021. Do Vision Transformers See Like Convolutional Neural Networks? *Advances in Neural Information Processing Systems (NeurIPS)* (2021). arXiv:2108.08810 [cs.CV]

[33] Mattia Rigotti, Christoph Miksovic, Ioana Giurgiu, Thomas Gschwind, and Paolo Scotton. 2021. Realizing the Potential of Concept-Based Explanations for Vision Transformers. arXiv preprint arXiv:2109.13733 (2021).

[34] Mattia Rigotti, Christoph Miksovic, Ioana Giurgiu, Thomas Gschwind, and Paolo Scotton. 2022. Attention-Based Interpretability with Concept Transformers. *International Conference on Learning Representations (ICLR)* (2022). arXiv:2111.11510 [cs.LG]

[35] Sebastian Ruder. 2017. An Overview of Multi-Task Learning in Deep Neural Networks. arXiv preprint arXiv:1706.05098 (2017).

[36] Johannes Ruf and Weiguan Wang. 2020. Finance-Informed Machine Learning for Derivatives Pricing and Hedging. arXiv preprint arXiv:2005.12483 (2020).

[37] Afees A Salisu and Xuan Vinh Vo. 2020. COVID-19 and Market Volatility: An Assessment of Unprecedented Disruptions. *Research in International Business and Finance* 54 (2020), 101243.

[38] Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, and Aleksander Madry. 2018. Faster Convergence in Deep Learning via Loss Surface Flattening. *Advances in Neural Information Processing Systems (NeurIPS)* (2018). arXiv:1805.11604 [cs.LG]

[39] Ozan Sener and Vladlen Koltun. 2018. Multi-Task Learning as Multi-Objective Optimization. *Advances in Neural Information Processing Systems (NeurIPS)* (2018). arXiv:1810.04650 [cs.LG]

[40] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. 2012. Practical Bayesian Optimization of Machine Learning Algorithms. *Advances in Neural Information Processing Systems (NeurIPS)* (2012). arXiv:1206.2944 [cs.LG]

[41] Emi Soroka and Artem Arzyn. 2025. Data-Efficient Realized Volatility Forecasting with Vision Transformers. arXiv preprint arXiv:2511.03046 (2025).

[42] Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. 2021. Token Mixing: A Simple Approach for Vision Transformers to Mix Spatial Information. *Advances in Neural Information Processing Systems (NeurIPS)* (2021). arXiv:2105.01601 [cs.CV]

[43] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. 2021. Training Data-Efficient Image Transformers & Distillation through Attention. *International Conference on Machine Learning (ICML)* (2021). arXiv:2012.12877 [cs.CV]

[44] Simon Vandenhende, Stamatios Georgoulis, Wouter Van Gansbeke, Marc Proesmans, Dengxin Dai, and Luc Van Gool. 2022. Multi-Task Learning for Dense Prediction Tasks: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 7 (2022), 3614–3633.

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *Advances in Neural Information Processing Systems (NeurIPS)* (2017). arXiv:1706.03762 [cs.CL]

[46] Neo Wu, Bradley Green, Xue Ben, and Shawn O'Banion. 2022. A Dual-Stream Architecture for Time Series Modeling. arXiv preprint arXiv:2206.14274 (2022).

[47] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. 2020. Gradient Surgery for Multi-Task Learning. *Advances in Neural Information Processing Systems (NeurIPS)* (2020). arXiv:2001.06782 [cs.LG]

[48] Xin Zhang, Junpeng Li, and Hai Ye. 2022. A Unified Model for Financial Market Anomaly Detection. *Expert Systems with Applications* 209 (2022), 118250.

[49] Yu Zhang and Qiang Yang. 2022. A Survey on Multi-Task Learning. *IEEE Transactions on Knowledge and Data Engineering* 34, 12 (2022), 5586–5609.

[50] Weiyi Zheng and Xin Sheng He. 2021. Implied Volatility Surface Prediction with Deep Learning. *Quantitative Finance* 21, 10 (2021), 1655–1669.

[51] Xibin Zhou, Chenchen Han, Yingqi Zhang, Huan Du, Jiayuan Tian, Jin Su, Renju Liu, Kai Zhuang, Shiyu Jiang, Anthony Gitter, Li Liu, Huayu Li, Meiqi Wu, Shiyang You, Zichen Yuan, Feng Ju, Huilin Zhang, Wei Zheng, Fengyuan Dai, Yuyang Zhou, Yuyang Tao, Dan Wu, Zongze Shao, Yang Liu, Hongyuan Lu, and Fajie Yuan. 2025. Decoding the Molecular Language of Proteins with Evolla. *bioRxiv* (2025), 2025.01.05.630192. Cold Spring Harbor Laboratory. https://doi.org/10.1101/2025.01.05.630192.

[52] Xinyuan Song, Keyu Chen, Ziqian Bi, Qian Niu, Junyu Liu, Benji Peng, Sen Zhang, Zichen Yuan, Ming Liu, Ming Li, Juan Zhang, Tianyang Wang, Xuanhe Pan, Yichao Zhang, Junhao Song, Jiawei Xu, Chuanqi Jiang, Jinlang Wang, and Pohsun Feng. 2024. Transformer: A Survey and Application. SSRN preprint (2024). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5211988.

[53] Xibin Zhou, Chenchen Han, Yingqi Zhang, Jin Su, Kai Zhuang, Shiyu Jiang, Zichen Yuan, Wei Zheng, Fengyuan Dai, and Yuyang Zhou. 2025. Decoding the Molecular Language of Proteins with Evola. *bioRxiv* (2025), 2025-01.

[54] Zichen Yuan, Qi Shen, Bingyi Zheng, Yuting Liu, Linying Jiang, and Guibing Guo. 2024. Video and Audio Are Images: A Cross-Modal Mixer for Original Data on Video–Audio Retrieval. *Knowledge-Based Systems* 299 (2024), 112076. https://doi.org/10.1016/j.knosys.2024.112076.

[55] Xinyuan Song, Keyu Chen, Ziqian Bi, Qian Niu, Junhao Song, Junyu Liu, Benji Peng, Sen Zhang, Ming Liu, Ming Li, Zichen Yuan, Li Zhang, Yan Zhong, Xuanhe Pan, Jiawei Xu, Yichao Zhang, Jinlang Wang, and Pohsun Feng. 2024. Mastering Reinforcement Learning: Foundations, Algorithms, and Real-World Applications. SSRN preprint (2024). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5208695.

[56] Xinyuan Song, Wei-Che Hsieh, Ziqian Bi, Chuanqi Jiang, Junyu Liu, Benji Peng, Sen Zhang, Xuanhe Pan, Jiawei Xu, Jinlang Wang, Keyu Chen, Pohsun Feng, Yizhu Wen, Tianyang Wang, Junjie Yang, Ming Li, Bowen Jing, Jintao Ren, Junhao Song, Hong-ming Tseng, Yichao Zhang, Lawrence K. Q. Yan, Qian Niu, Silin Chen, Yunze Wang, Chia Xin Liang, Ming Liu, and Zichen Yuan. 2024. A Comprehensive Guide to Explainable AI: From Classical Models to LLMs. *OSF* (2024). https://osf.io/preprints/osf/wbk36.

[57] Tianyang Wang, Yunze Wang, Xinyuan Song, Yichao Zhang, Silin Chen, Ziqian Bi, Ming Liu, Jun Zhou, Junhao Song, Bowen Jing, Junjie Yang, Keyu Chen, Ming Li, Zichen Yuan, Qian Niu, Junyu Liu, Benji Peng, Sen Zhang, Jiawei Xu, Xuanhe Pan, Jinlang Wang, Pohsun Feng, Chuanqi Jiang, Yizhu Wen, Lawrence Yan, Hong-Ming Tseng, Jintao Ren, Weiche Hsieh, Zheyu Yao, and Chia Xin Liang. 2024. Deep Learning and Machine Learning-Generative Models: Foundations, Techniques, and Applications. SSRN preprint (2024). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5211980.

[58] Zichen Yuan, Lifan Sun, Yucen Zhuang, Yue Wang, Xinyuan Song, Tianqi Xu, Siyuan Li, Junchen Fu, Youhua Li, Sirui Hong, Jiaqi Chen, Joemon M. Jose, and Yongxin Ni. 2025. SOLAR: Serendipity Optimized Language Model Aligned for Recommendation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*. Association for Computational Linguistics, 10146–10169. https://doi.org/10.18653/v1/2025.findings-emnlp.538.

[59] Silin Chen, Tianyang Wang, Xinyuan Song, Bowen Jing, Junjie Yang, Junhao Song, Keyu Chen, Ming Li, Qian Niu, Junyu Liu, Benji Peng, Sen Zhang, Ming Liu, Jiawei Xu, Xuanhe Pan, Jinlang Wang, Pohsun Feng, Chuanqi Jiang, Yizhu Wen, Lawrence K. Q. Yan, Hong-ming Tseng, Jintao Ren, Yichao Zhang, Ziqian Bi, and Zichen Yuan. 2024. Deep Learning and Machine Learning, Advancing Big Data Analytics and Management: Generative Models. *Open Science Framework* (2024). https://osf.io/6hpyz.

[60] Lifan Sun, Zichen Yuan, Ersheng Ni, Weihua Cheng, Xinyuan Song, Linkun Dai, Hongwei Jiang, Sibo Xu, Mengmeng Chen, Yucen Zhuang, Yongxin Ni, and Youhua Li. 2025. AMBER: Adaptive Meta Balanced Paradigm for Heterogeneous Graph-Based Knowledge Tracing. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*. ACM, 2801–2810. https://doi.org/10.1145/3746252.3761202.

[61] Jinwei Su, Qizhen Lan, Yinghui Xia, Lifan Sun, Weiyou Tian, Tianyu Shi, and Lewei He. 2026. Difficulty-Aware Agentic Orchestration for Query-Specific Multi-Agent Workflows. In *Proceedings of the ACM Web Conference 2026*. ACM, 2060–2070. https://doi.org/10.1145/3774904.3792240.

[62] Xin Li, Shengcheng Yu, Lifan Sun, Yuexiao Liu, and Chunrong Fang. 2023. Towards Effective Bug Reproduction for Mobile Applications. In *2023 10th International Conference on Dependable Systems and Their Applications (DSA)*. IEEE, 114–125. https://doi.org/10.1109/DSA59317.2023.00024.

[63] Hongyang He, Xinyuan Song, Yangfan He, Zeyu Zhang, Yanshu Li, Haochen You, Lifan Sun, and Wenqiao Zhang. 2026. TRiCo: Triadic Game-Theoretic Co-Training for Robust Semi-Supervised Learning. *Advances in Neural Information Processing Systems* 38 (2026), 87545–87570. https://proceedings.neurips.cc/paper_files/paper/2025/hash/7eed2822411dc37b3768ae04561caafa-Abstract-Conference.html.

[64] Lifan Sun, Yichao Zhang, Xinyuan Song, Ziqian Bi, Keyu Chen, Ming Li, Qian Niu, Junyu Liu, Benji Peng, Sen Zhang, Ming Liu, Xuanhe Pan, Jinlang Wang, Pohsun Feng, Tianyang Wang, Yizhu Wen, Lawrence Yan, Hong-Ming Tseng, Yan Zhong, Yunze Wang, Ziyuan Qin, Bowen Jing, Junjie Yang, Zheyu Yao, Jun Zhou, Chia Xin Liang, and Junhao Song. 2025. Exploring Multimodal Embeddings for Text and Impact on Language Processing. SSRN preprint (2025). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5189610.