Explainable Heterogeneous Anomaly Detection in Financial Networks via Adaptive Expert Routing

Zan Li
Rensselaer Polytechnic Institute
USA
liz37@rpi.edu

Rui Fan
Rensselaer Polytechnic Institute
USA
fanr5@rpi.edu

Abstract

Financial anomalies exhibit heterogeneous mechanisms—price shocks, liquidity freezes, contagion cascades, regime shifts—but existing detectors treat all anomalies uniformly, producing scalar scores without revealing *which* mechanism is failing, *where* risks concentrate, or *how* to intervene. This opacity prevents targeted regulatory responses: a liquidity freeze requires market-making intervention while a price shock signals information asymmetry, yet current methods cannot distinguish between them. Three unsolved challenges persist: (1) static graph structures cannot adapt when market correlations shift during regime changes; (2) uniform detection mechanisms miss type-specific signatures of heterogeneous anomalies across multiple temporal scales while failing to integrate individual behaviors with network contagion; (3) black-box outputs provide no actionable guidance on anomaly mechanisms or their temporal evolution.

We address these via adaptive graph learning with specialized expert networks that provide built-in interpretability. Our framework captures multi-scale temporal dependencies through BiLSTM with self-attention, fuses temporal and spatial information via cross-modal attention, learns dynamic graphs through neural multi-source interpolation of domain knowledge graphs, temporal similarity, and contextual patterns, adaptively balances learned dynamics with structural priors via stress-modulated fusion, routes anomalies to four mechanism-specific experts through stress-modulated gating, and produces dual-level interpretable attributions—expert routing weights reveal active mechanisms while weight dynamics track temporal transitions. Critically, interpretability is embedded architecturally rather than applied post-hoc: routing weights directly encode mechanism attribution without requiring separate explanation procedures.

On 100 US equities (2017-2024), we achieve 92.3% detection of 13 major events with 3.8-day lead time, outperforming best baseline by 30.8pp. Silicon Valley Bank (SVB) case study demonstrates anomaly evolution tracking—Price-Shock expert weight rose to 0.39 (33% above baseline 0.29) during closure, peaking at 0.48 (66% above baseline) one week later, revealing automatic temporal mechanism identification without labeled supervision and providing transparent anomaly attribution for regulatory decision-making.

CCS Concepts

• **Computing methodologies → Neural networks**; • **Information systems → Data mining**.

Keywords

Explainable AI, Anomaly Detection, Mixture of Experts, Graph Neural Networks, Financial Networks

ACM Reference Format:
Zan Li and Rui Fan. 2025. Explainable Heterogeneous Anomaly Detection in Financial Networks via Adaptive Expert Routing. In *Proceedings of the International Joint Workshop on Explainable AI in Finance (XAI-FIN '25), November 15, 2025, Singapore.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/XXXXXXX.XXXXXXX

1 Introduction

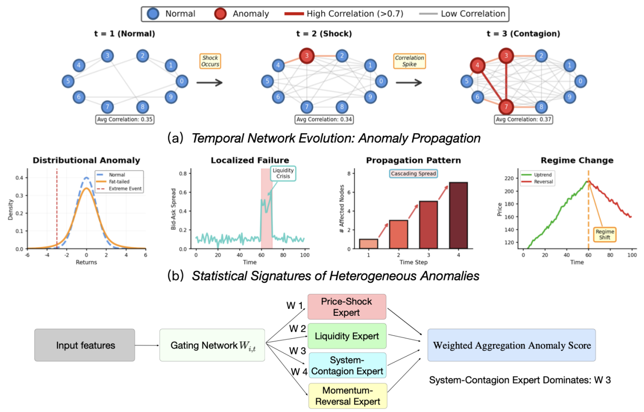
The March 2023 banking crisis exemplifies financial system fragility: Silicon Valley Bank collapsed within 48 hours, triggering contagion to Signature Bank and First Republic Bank [4]. Could this cascade have been detected earlier with *actionable warnings* about anomaly mechanisms?

The core problem: Existing detectors cannot distinguish anomaly types or guide interventions. Financial anomalies exhibit fundamentally different failure modes requiring distinct responses. A liquidity freeze demands market-making intervention, while a price shock signals information asymmetry [3, 16]. Yet existing detectors produce opaque scalar scores without explaining *which* mechanism is failing, *where* risks concentrate, or *how* to respond. During the 2023 banking crisis, regulators needed to distinguish liquidity problems (requiring emergency facilities) from solvency crises (requiring capital injections), but current systems provided no such guidance.

Why mechanism-based taxonomy? Classical anomaly detection relies on statistical categories: point anomalies, contextual anomalies, and collective anomalies [6]. These describe *how* anomalies manifest statistically, not *why* they occur mechanistically. A liquidity freeze (microstructure breakdown with stable prices [1]) and momentum reversal (trend exhaustion [2]) both appear as point anomalies, yet require entirely different interventions. Worse, statistical approaches treat stocks independently, missing contagion through interconnected institutions [10]. We instead propose four anomaly mechanisms grounded in financial theory: (1) *Price shocks*—tail events with fat-tailed distributions [3, 7]; (2) *Liquidity crises*—microstructure breakdowns where

XAI-FIN'25, Singapore

ACM ISBN 978-1-4503-XXXX-X/25/11
https://doi.org/10.1145/XXXXXXX.XXXXXXX



(a) *Temporal Network Evolution: Anomaly Propagation*

(b) *Statistical Signatures of Heterogeneous Anomalies*

(c) *Mixture-of-Experts Architecture*

Figure 1: Three fundamental challenges in financial anomaly detection. (a) Challenge 1—Adaptivity: Temporal network evolution during crisis. Left: tranquil period with baseline correlations (avg 0.35); Middle: localized shock at node 3; Right: contagion spreads with correlation spikes (0.37) and new high-correlation clusters (thick red edges). Static graphs treat all phases identically; purely data-driven dynamic graphs destabilize. (b) Challenge 2—Specialization: Four distinct anomaly signatures. From left: distributional extremes (four-sigma fat tails, price shocks), localized microstructure breakdowns (spread spikes, liquidity crises), cascading contagion (progressive infections), prolonged transitions (momentum reversals, regime shifts). Uniform detection cannot distinguish mechanisms. (c) Challenge 3—Interpretability: Scalar scores (top) provide no mechanism attribution. Our MoE architecture (bottom) routes to specialized experts—routing weight concentration on Systemic-Contagion expert (colored boxes) reveals active mechanism, enabling targeted intervention.

bid-ask spreads explode despite stable prices [1, 16]; (3) *Systemic contagion*—cascading failures through correlation networks [8, 10]; (4) *Regime shifts*—structural changes via momentum reversals [2, 12].

Three unsolved challenges. Operationalizing this taxonomy requires addressing three fundamental obstacles illustrated in Figure 1.

Challenge 1 (Adaptivity): Static graphs fail during regime changes. Figure 1(a) demonstrates the adaptivity challenge through network evolution: the tranquil-period network (left) exhibits baseline correlations (avg 0.35) with sparse connections; during crisis (right), the structure transforms with correlation spikes (avg 0.37) and dense clustering as distress propagates through previously independent institutions.

Existing approaches and limitations. Static methods use fixed adjacency matrices throughout all market phases. Graph neural networks [14, 24] and anomaly detectors like DOMINANT [9], CoLA [18] construct **A** from pre-computed correlations or industry classifications, treating the sparse tranquil network and dense crisis network identically—missing the critical structural transformation visible in Figure 1(a). Dynamic approaches attempt temporal adaptation but face an unresolved dilemma: purely data-driven methods (EvolveGCN [21], DySAT [22]) destabilize when historical patterns break; conversely, topology-constrained methods (ROLAND [26]) cannot capture emergent correlation pathways during regime transitions.

Our solution. We resolve this dilemma through neural multi-source graph learning with stress-modulated adaptive fusion. Three complementary information sources—domain knowledge graphs (stable industry/geography priors), temporal similarity (short-term co-movement), and contextual patterns (long-term behaviors)—synthesize via learnable weights that automatically rebalance based on market conditions. Stress-modulated fusion then combines the learned graph with domain priors: high market stress emphasizes data-driven correlations for responsiveness, while low stress maintains structural priors for stability.

Challenge 2 (Specialization): Uniform detection cannot distinguish failure mechanisms. Figure 1(b) illustrates four distinct anomaly signatures: price shocks exhibit extreme return kurtosis (>10) with stable spreads; liquidity crises display bid-ask explosion (8→52 bps) with minimal price movement; systemic contagion reveals correlation surges (0.23→0.67) with coordinated spread widening; regime shifts manifest RSI reversals (70→30) with gradual structural changes.

Existing approaches and limitations. Current methods apply uniform feature processing across all mechanisms. Temporal models (LSTM-AE [20], TranAD [23]) process all 29 features identically, capturing sequences but treating stocks independently—missing coordinated patterns like cascading contagion. Graph-based methods (GCN-AE [14], GHRN [13]) incorporate network structure but use static snapshots, unable to distinguish rapid versus gradual propagation. Hybrid approaches (MTAD-GAT [27]) employ uniform reconstruction objectives that cannot discriminate whether failures stem from distributional extremes, microstructure breakdowns, or network propagation.

Our solution. We address specialization through multi-scale spatial-temporal encoding and stress-modulated mixture-of-experts. Multi-scale encoding integrates BiLSTM for short-term patterns, self-attention for long-range dependencies, and cross-modal attention to distinguish isolated shocks from contagion. Mixture-of-experts routes observations to four specialized experts (Price-Shock, Liquidity, Systemic-Contagion, Momentum-Reversal), each processing category-aligned feature subsets—for example, the Liquidity expert receives only liquidity-specific metrics while masking unrelated features. Adaptive temperature modulates routing sharpness: sharp routing during crises clearly identifies mechanisms, soft routing during tranquility acknowledges multi-causal patterns.

Challenge 3 (Interpretability): Black-box outputs hinder actionable intervention. Figure 1(c) demonstrates the interpretability gap. Existing detectors produce scalar anomaly scores indicating *that* something is abnormal but not *which mechanism* is failing—a high score could indicate distributional extremes (requiring circuit breakers), microstructure breakdowns (requiring market-making), or cascading contagion (requiring coordinated intervention).

Existing approaches and limitations. Deep learning detectors—autoencoders [28], GANs [17], normalizing flows [15]—optimize reconstruction objectives producing scalar scores without mechanistic interpretation. Post-hoc explanation methods face fundamental limitations: SHAP [19] exhibits temporal instability (variance >0.3) and provides feature-level importance ("bid-ask spread contributed 30%") but not mechanism-level understanding ("this is liquidity crisis not contagion"); attention visualization indicates *which features* are important but not *which mechanisms* are active; critically, post-hoc methods analyze static snapshots, missing temporal transitions essential for intervention strategy.

Our solution. We embed dual-level interpretability architecturally. Expert routing weights directly reveal active mechanisms during forward pass—for example, weight distribution $[0.1, 0.2, 0.6, 0.1]$ indicates Systemic-Contagion dominates (60%), enabling targeted intervention. Weight trajectories track mechanism transitions across time—observing Price-Shock rising then Contagion increasing reveals progression from isolated shock to systemic spread, guiding phase-appropriate intervention. Market Pressure Index aggregates individual signals through five components with hierarchical alerts (L0: Normal, L1: Observation, L2: Attention, L3: Warning, L4: Crisis), translating mechanism identification into actionable regulatory guidance.

Contributions.

- **Multi-scale spatial-temporal modeling with adaptive graph fusion.** Integrating bidirectional recurrent processing, self-attention, cross-modal attention, and neural dynamic graph learning through multi-source interpolation with stress-modulated balancing—resolving Challenge 1's stability-responsiveness dilemma.
- **Mechanism-based taxonomy with dual-level interpretability.** Four economically grounded mechanisms operationalized through specialized experts with stress-modulated routing. Routing weights provide static mechanism attribution; weight trajectories enable dynamic temporal evolution tracking—addressing Challenges 2 and 3.
- **Empirical validation with transparent evolution tracking.** 92.3% detection rate across thirteen major market events (2022-2024) with 3.8-day median lead time, outperforming strongest baselines (TranAD and GHRN, both 76.9%) by 15.4 percentage points. SVB case study demonstrates interpretability: Price-Shock weight rose from 0.29 to 0.39 upon closure (33% elevation), peaking at 0.48 one week later (66% elevation).

2 Problem Formulation

We formulate financial anomaly detection as mechanism-specific scoring on temporal dynamic graphs. This section defines the problem setting, input representations, and desired outputs.

Input. Consider a financial market with N stocks observed over T time periods. The temporal graph is denoted as $\mathcal{G} = (\mathcal{V}, \{\mathcal{E}_t\}_{t=1}^{T}, \mathbf{A}_{\text{prior}})$, where $\mathcal{V}$ represents the node set with $|\mathcal{V}| = N$, $\mathcal{E}_t$ denotes time-varying edges at time t, and $\mathbf{A}_{\text{prior}} \in \{0,1\}^{N \times N}$ is the domain knowledge graph encoding structural relationships (industry sectors, geographic regions). Each stock $i \in \mathcal{V}$ is associated with a multivariate time series $\mathbf{X}_{i,1:T} \in \mathbb{R}^{T \times F}$, where $F = 29$ features span twenty-day rolling windows capturing price dynamics, liquidity metrics, correlation statistics, and momentum indicators.

Domain knowledge graph. $\mathbf{A}_{\text{prior}}$ encodes stable fundamental relationships: $A_{\text{prior}}[i, j] = 1$ if stocks i and j share industry classification (GICS sectors) or geographic proximity (headquarters location), and 0 otherwise. This graph provides structural priors that persist across market regimes, serving as a baseline topology for adaptive graph learning in Module 2 (Section 3.3).

Mechanism-specific features. The $F = 29$ features are partitioned into four category-aligned subsets corresponding to our mechanism taxonomy:

- *Price-Shock* ($d_1 = 6$): jump indicator, extreme return z-score, return kurtosis, maximum drawdown, CVaR 95%, upside volatility
- *Liquidity* ($d_2 = 8$): bid-ask spread, Amihud illiquidity, Roll's measure, turnover ratio, trading intensity, price impact, liquidity risk premium, order imbalance
- *Systemic-Contagion* ($d_3 = 7$): market correlation, sector correlation, systematic proportion (R^2), contagion risk measure, systemic risk contribution, spillover index, herding measure
- *Momentum-Reversal* ($d_4 = 8$): momentum 5-day, momentum 20-day, momentum reversal indicator, RSI, MACD signal, volume-price divergence, support-resistance indicator, price acceleration

where $\sum_{k=1}^{4} d_k = F$. This feature partitioning enables mechanism-specific expert processing in Module 3 (Section 3.4).

Output. The framework produces three interpretable outputs at each time t:

- *Individual anomaly scores*: $s_{i,t} \in [0,1]$ for each stock i, quantifying anomaly intensity where higher values indicate stronger deviations from normal patterns
- *Expert routing weights*: $\mathbf{w}_{i,t} \in \mathbb{R}^4$ with $\sum_{k=1}^{4} w_{i,t}^{(k)} = 1$ and $w_{i,t}^{(k)} \geq 0$, revealing which of the four mechanisms (Price-Shock, Liquidity, Systemic-Contagion, Momentum-Reversal) dominates for stock i at time t—providing static mechanism attribution
- *Market Pressure Index*: $\text{MPI}_t \in [0,1]$ aggregating market-wide anomaly patterns with hierarchical alerts (L0: Normal, L1: Observation, L2: Attention, L3: Warning, L4: Crisis), translating detection into actionable regulatory guidance

Objective. Given historical observations $\{\mathbf{X}_{i,1:t-1}\}_{i=1}^{N}$ and domain knowledge $\mathbf{A}_{\text{prior}}$, the model learns to:

1. Construct adaptive dynamic graph $\mathbf{A}_{\text{fused}}^{t}$ balancing domain priors with learned correlation patterns responsive to regime changes (Module 2)
2. Generate refined embeddings $\mathbf{z}_{i,t}^{\text{final}}$ integrating multi-scale temporal dependencies—short-term via recurrent processing, long-term via self-attention—with spatial network structure via cross-modal attention (Modules 1-2)
3. Route observations to mechanism-specific experts producing interpretable scores $s_{i,t}$ and attribution weights $\mathbf{w}_{i,t}$ enabling transparent anomaly identification (Module 3)

(4) Aggregate individual signals into market-level pressure index MPI_t with temporal evolution tracking through weight dynamics $\{\mathbf{w}_{i,t}\}_{t=1}^{T}$ (Module 4)

Critically, the framework operates in an *unsupervised* setting—no labeled anomalies are required for training. The model learns normal pattern representations through reconstruction objectives, with deviations indicating anomalies. Mechanism attribution emerges through expert specialization driven by category-aligned feature partitioning and stress-modulated routing, enabling automatic mechanism identification without supervision.

Evaluation protocol. Detection performance is measured through retrospective analysis of documented market events: a detection is considered successful if the anomaly score $s_{i,t}$ or Market Pressure Index MPI_t exceeds calibrated thresholds (95th percentile on validation set) within a five-day window preceding the event. Lead time quantifies early warning capability—the number of days between threshold crossing and event occurrence. Mechanism attribution accuracy is assessed qualitatively through case studies examining whether expert routing weights align with documented failure mechanisms during crisis episodes.

3 Method

We present our four-module framework addressing the challenges in Section 1 under the problem setting of Section 2. Figure 2 illustrates the architecture and information flow. Complete mathematical formulations are provided in Appendix A.

3.1 Overview

The framework operates sequentially: **Module 1** (Section 3.2) extracts initial representations capturing temporal dynamics and spatial network structure. **Module 2** (Section 3.3) constructs adaptive graphs via multi-source interpolation, fuses them with domain knowledge through stress-modulated weighting, then refines embeddings. **Module 3** (Section 3.4) routes refined embeddings to four mechanism-specific experts producing reconstruction errors and interpretable routing weights. **Module 4** (Section 3.5) aggregates expert signals across temporal scales and computes market-level pressure indicators.

Information flow: raw features $\mathbf{X} \longrightarrow$ initial embeddings $\mathbf{z}$, context $\mathbf{c}$, spatial embeddings $\mathbf{h}^{\text{spat}}$ (Module 1) $\longrightarrow$ refined embeddings $\mathbf{z}^{\text{final}}$ via learned graph $\mathbf{A}^{t}_{\text{fused}}$ (Module 2) $\longrightarrow$ routing weights $\mathbf{w}$ and expert errors $\ell^{(k)}$ (Module 3) $\longrightarrow$ anomaly scores $s_{i,t}$ and Market Pressure Index MPI_t (Module 4).

3.2 Module 1: Spatial-Temporal Encoding

Objective. Capture multi-scale temporal dependencies (short-term via recurrent processing, long-term via self-attention) while integrating network structure to distinguish isolated shocks from contagion.

Temporal branch. BiLSTM processes input features $\mathbf{X}_{i,1:T} \in \mathbb{R}^{T\times F}$ ($T = 20$ days, $F = 29$) bidirectionally. Forward and backward hidden states (each 128-dim) concatenate:

$$\mathbf{h}^{\text{bi}}_{i,t} = [\mathbf{h}^{\text{fwd}}_{i,t}; \mathbf{h}^{\text{bwd}}_{i,t}] \in \mathbb{R}^{256} \tag{1}$$

Multi-head self-attention (4 heads, head dimension 64) enables direct connections between distant time steps. Query, key, and value projections from $\mathbf{h}^{\text{bi}}$ compute scaled dot-product attention (Appendix A), yielding $\mathbf{h}^{\text{attn}} \in \mathbb{R}^{T\times 256}$.

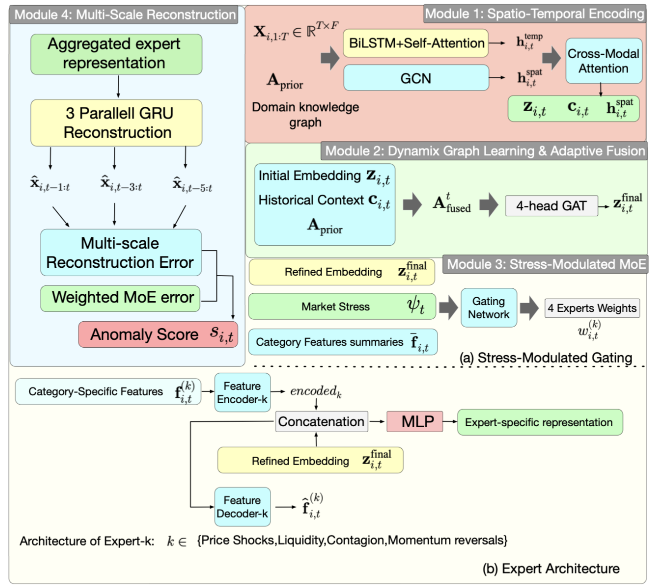


Figure 2: Four-module framework architecture. Module 1 extracts multi-scale spatial-temporal representations. Module 2 (addressing Challenge 1) learns dynamic graphs with stress-modulated adaptive fusion. Module 3 (addressing Challenge 2) routes observations to four specialized experts with adaptive temperature. Module 4 (addressing Challenge 3) produces interpretable outputs: individual scores s, routing weights w for mechanism attribution, and Market Pressure Index with hierarchical alerts.

Temporal pooling combines aggregate patterns (mean over T) with recent dynamics (final step), projected via MLP to temporal embeddings $\mathbf{h}^{\text{temp}}_{i,t} \in \mathbb{R}^{128}$.

Spatial branch. Features at final time step $\mathbf{x}_{i,T} \in \mathbb{R}^{29}$ are projected to 128 dimensions, then processed by two GCN layers using domain knowledge graph $\mathbf{A}_{\text{prior}}$. For layer $\ell \in \{1, 2\}$:

$$\mathbf{h}^{(\ell)} = \mathrm{ReLU}\left(\tilde{\mathbf{A}}_{\text{prior}}\mathbf{h}^{(\ell-1)}\mathbf{W}^{(\ell)}_{\text{gcn}}\right) \tag{2}$$

where $\tilde{\mathbf{A}}_{\text{prior}} = \mathbf{D}^{-1/2}\mathbf{A}_{\text{prior}}\mathbf{D}^{-1/2}$ is symmetrically normalized with degree matrix $\mathbf{D}_{ii} = \sum_j A_{\text{prior}}[i, j]$. Residual connection:

$$\mathbf{h}^{\text{spat}}_{i,t} = \mathbf{h}^{(2)}_i + 0.5\mathbf{h}^{(0)}_i \in \mathbb{R}^{128} \tag{3}$$

Cross-modal fusion. Temporal embeddings query spatial embeddings via attention. Query from $\mathbf{h}^{\text{temp}}$, keys/values from $\mathbf{h}^{\text{spat}}$ compute attention (Appendix A). High weights indicate coordinated distress (contagion); low weights indicate isolation. Output combines with $\mathbf{h}^{\text{temp}}$ via residual, yielding $\mathbf{h}^{\text{fused}} \in \mathbb{R}^{N\times 128}$.

Outputs. (1) Initial embeddings: $\mathbf{z}_{i,t} = [\mathbf{h}^{\text{temp}}_{i,t}; \mathbf{h}^{\text{spat}}_{i,t}] \in \mathbb{R}^{256}$. (2) Historical context: $\mathbf{c}_{i,t} = \mathrm{Tanh}(\mathbf{h}^{\text{fused}}_i \mathbf{W}_{\text{ctx}}) \in \mathbb{R}^{64}$ compressing long-term patterns. (3) Spatial embeddings $\mathbf{h}^{\text{spat}}_{i,t}$ retained for Module 3.

3.3 Module 2: Neural Dynamic Graph Learning & Adaptive Fusion

Objective. Learn dynamic graphs balancing structural stability (preventing false alarms) with regime responsiveness (detecting transformations).

Market stress calculation. Four normalized indicators—volatility stress, extreme ratio, correlation spike, liquidity stress—aggregate via learnable weights β_k through sigmoid (detailed formulations in Appendix A):

$$\psi_t = \sigma\left(\sum_{k=1}^{4} \beta_k \tilde{I}_k^t\right) \in [0, 1] \tag{4}$$

High ψ_t indicates crises; low values indicate tranquility.

Multi-source graph learning. Three sources synthesize: (1) Temporal similarity $\mathbf{A}_{\text{temp}}^t$ via cosine similarity between $\mathbf{z}_{i,t}$ detects correlation spikes; (2) Context similarity $\mathbf{A}_{\text{ctx}}^t$ via $\mathbf{c}_{i,t}$ captures long-term patterns; (3) Domain knowledge $\mathbf{A}_{\text{prior}}$ provides stable relationships. Edge features $\mathbf{e}_{ij} = [\mathbf{z}_{i,t}; \mathbf{z}_{j,t}; \mathbf{c}_{i,t}; \mathbf{c}_{j,t}]$ processed by MLP yield $\mathbf{A}_{\text{edge}}^t$. Learnable interpolation combines sources:

$$\mathbf{A}_{\text{learned}}^t = w_{\text{temp}} \mathbf{A}_{\text{temp}}^t + w_{\text{ctx}} \mathbf{A}_{\text{ctx}}^t + w_{\text{prior}} \mathbf{A}_{\text{prior}} \tag{5}$$

During tranquility w_{prior} dominates; during crises data-driven sources increase. Gated refinement incorporates $\mathbf{A}_{\text{edge}}^t$ (Appendix A).

Differentiable top-k sparsification. For each node, top $k = 20$ connections retained via soft masking:

$$M[i, j] = \sigma\left(5(\mathbf{A}_{\text{learned}}^t[i, j] - \tau_k^i)\right) \tag{6}$$

where τ_k^i is the k-th largest value (detached). Graph updated as $\mathbf{A}_{\text{learned}}^t \leftarrow \mathbf{A}_{\text{learned}}^t \odot M$, then symmetrized and row-normalized.

Stress-modulated adaptive fusion. Learned graph combines with priors using stress-modulated weight:

$$\alpha_t = \text{clamp}\left(\sigma(\alpha_{\text{base}} + \beta_\alpha \cdot \psi_t), 0.2, 0.8\right) \tag{7}$$

where $\alpha_{\text{base}} = 0.5$, $\beta_\alpha = -0.5$ (both learnable). High ψ_t yields low α_t emphasizing learned patterns; low ψ_t yields high α_t emphasizing priors. Fused graph:

$$\mathbf{A}_{\text{fused}}^t = \alpha_t \mathbf{A}_{\text{prior}} + (1 - \alpha_t) \mathbf{A}_{\text{learned}}^t \tag{8}$$

Graph attention refinement. 4-head GAT (head dim 32) uses $\mathbf{A}_{\text{fused}}^t$ to refine $\mathbf{z}$. Attention coefficients computed via LeakyReLU-activated mechanism (Appendix A). Outputs concatenated and projected to $\mathbf{z}_i^{\text{refined}} \in \mathbb{R}^{128}$. Residual with zero-padding:

$$\mathbf{z}_{i,t}^{\text{final}} = \mathbf{z}_{i,t} + 0.5[\mathbf{z}_{i,t}^{\text{refined}}; \mathbf{0}^{128}] \in \mathbb{R}^{256} \tag{9}$$

3.4 Module 3: Stress-Modulated Mixture-of-Experts

Objective. Route observations to mechanism-specific experts producing transparent attributions embedded architecturally (not post-hoc).

Category-level signals. Extract compact statistics from 29 features grouped by mechanisms: for category $k \in \{1, 2, 3, 4\}$ with d_k features,

$$\bar{f}_{i,t}^{(k)} = \frac{1}{d_k} \sum_{j=1}^{d_k} |f_{i,T}^{(k)}[j]| \tag{10}$$

Layer-normalized to $\bar{\mathbf{f}}_{i,t} \in \mathbb{R}^4$, guiding routing.

Gating network. Routing weights computed from four sources: refined embeddings $\mathbf{z}^{\text{final}}$, global context via self-attention yielding $\mathbf{a}_{i,t} \in \mathbb{R}^{256}$, market stress ψ_t, and category signals $\bar{\mathbf{f}}_{i,t}$. Concatenate to $[\mathbf{z}^{\text{final}}; \mathbf{a}; \psi_t; \bar{\mathbf{f}}] \in \mathbb{R}^{517}$, pass through 3-layer MLP producing gating logits $\mathbf{g}_{i,t} \in \mathbb{R}^4$. Add diversity bias $\mathbf{b}_{\text{div}}$ (initialized [0.0, 0.5, 0.3, 0.2], learned) preventing collapse.

Temperature-scaled softmax produces routing weights:

$$\mathbf{w}_{i,t} = \text{softmax}(\mathbf{g}_{i,t}/\tau_t) \tag{11}$$

where $\tau_t = |\tau_{\text{base}}| + \beta_{\text{stress}} \psi_t$ with $\tau_{\text{base}} = 1.0$ (learned) and $\beta_{\text{stress}} = 0.5$ (fixed). Higher τ_t during crises produces sharper routing identifying mechanisms; lower τ_t during tranquility yields softer distributions.

Expert processing. Four experts process embeddings with category-specific features (Price-Shock $d_1 = 6$, Liquidity $d_2 = 8$, Contagion $d_3 = 7$, Momentum $d_4 = 8$). Experts 1, 2, 4 use $\mathbf{z}^{\text{final}}$ with $\mathbf{f}^{(k)}$; Expert 3 uses $\mathbf{h}^{\text{spat}}$ with $\mathbf{f}^{(3)}$ emphasizing network structure. Each expert's encoder projects to 64-dim latent $\mathbf{e}_{i,t}^{(k)}$; decoder reconstructs $\hat{\mathbf{f}}_{i,t}^{(k)}$ (Appendix A). Reconstruction error:

$$\ell_{i,t}^{(k)} = \|\mathbf{f}_{i,T}^{(k)} - \hat{\mathbf{f}}_{i,t}^{(k)}\|_2 \tag{12}$$

Weighted aggregation & interpretability. Aggregate:

$$e_{i,t}^{\text{MoE}} = \sum_{k=1}^{4} w_{i,t}^{(k)} \ell_{i,t}^{(k)}, \quad \mathbf{e}_{i,t} = \sum_{k=1}^{4} w_{i,t}^{(k)} \mathbf{e}_{i,t}^{(k)} \tag{13}$$

Routing weights $\mathbf{w}_{i,t}$ provide dual-level interpretability: *(1) Static mechanism attribution*—weight distribution reveals active mechanisms (e.g., $[0.1, 0.2, 0.6, 0.1]$ indicates Systemic-Contagion dominates). *(2) Dynamic temporal evolution*—tracking $\{\mathbf{w}_{i,t}\}_{t=1}^T$ reveals transitions enabling phase-appropriate intervention.

3.5 Module 4: Multi-Scale Reconstruction & Market Pressure Index

Multi-scale reconstruction. Three parallel GRU decoders (hidden dim 128) reconstruct feature sequences at 1-, 3-, 5-day horizons from $\mathbf{e}_{i,t} \in \mathbb{R}^{64}$. Each decoder $h \in \{1, 3, 5\}$ generates $\hat{\mathbf{x}}_{i,t-h:t}$. Reconstruction errors:

$$e_{i,t}^{(h)} = \frac{1}{h} \sum_{\tau=1}^{h} \|\mathbf{x}_{i,t-h+\tau} - \hat{\mathbf{x}}_{i,t-h+\tau}\|_2 \tag{14}$$

Weighted aggregation: $e_{i,t}^{\text{recon}} = 0.5e^{(1)} + 0.3e^{(3)} + 0.2e^{(5)}$. Design addresses Challenge 2: 1-day detects localized breakdowns, 5-day identifies prolonged transitions.

Combined anomaly score. Integrate expert-based (60%) and multi-scale (40%) reconstruction:

$$s_{i,t} = \sigma\left(2 \frac{0.6e_{i,t}^{\text{MoE}} + 0.4e_{i,t}^{\text{recon}} - \mu_e}{\sigma_e + \epsilon}\right) \tag{15}$$

where μ_e, σ_e are batch statistics, $\epsilon = 10^{-8}$.

Market Pressure Index. Aggregate via five components: (1) Anomaly rate (30%): percentage exceeding threshold, measuring breadth. (2) Concentration (25%): quadratic mean of top 5%, measuring intensity. (3) Dispersion (20%): cross-sectional std, measuring heterogeneity. (4) Expert consensus (15%): routing weight std among anomalous stocks—low std indicates mechanism agreement. (5) Market stress (10%): ψ_t. Compute:

$$\mathrm{MPI}_t = 0.30c_1 + 0.25c_2 + 0.20c_3 + 0.15c_4 + 0.10c_5 \in [0, 1] \quad (16)$$

Hierarchical alerts: L0 (Normal) < P_{75}, L1 (Observation) $[P_{75}, P_{85})$, L2 (Attention) $[P_{85}, P_{95})$, L3 (Warning) $[P_{95}, P_{99})$, L4 (Crisis) $\geq P_{99}$.

Training. End-to-end unsupervised learning via combined loss:

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \gamma_1 \mathcal{L}_{\text{div}} + \gamma_2 \mathcal{L}_{\text{smooth}} \quad (17)$$

where $\mathcal{L}_{\text{rec}}$ aggregates reconstruction errors, $\mathcal{L}_{\text{div}} = -\sum w \log w$ maximizes routing entropy, $\mathcal{L}_{\text{smooth}} = \sum \|\mathbf{w}_t - \mathbf{w}_{t-1}\|^2$ encourages temporal consistency, with $\gamma_1 = 0.01$, $\gamma_2 = 0.001$. Adam optimizer (learning rate 10^{-3}), batch size 32, maximum 100 epochs with early stopping (patience 15). Detailed optimization in Appendix A.

4 Experiments

4.1 Experimental Setup

Dataset. We evaluate on 100 US equities from diverse sectors (financials, technology, healthcare, energy, consumer goods) spanning 2017-2024 (8 years, 2,016 trading days). Data split: Training (2017-2021, 61.8%), Validation (2022, 12.7%), Test (2023-2024, 25.5%). Features include price dynamics, liquidity metrics, correlation statistics, and momentum indicators extracted from daily OHLCV data with 20-day rolling windows ($F = 29$ features). Data sources: Yahoo Finance for price/volume, TAQ for bid-ask spreads, CRSP for returns.

Events. We evaluate on 13 documented market events with publicly verified dates: *Banking crises* (4 events)—Silicon Valley Bank collapse (March 10, 2023), Signature Bank closure (March 12, 2023), Credit Suisse emergency acquisition (March 19, 2023), First Republic failure (May 1, 2023); *Monetary policy shifts* (2 events)—Federal Reserve aggressive rate hike cycle (March-November 2022), surprise rate cuts (2023); *Market shocks* (5 events)—Technology sector earnings miss (Q1 2023), CrowdStrike global IT outage (July 19, 2024), Japan carry trade unwind (August 5, 2024), GameStop volatility spike (May 2024), Cryptocurrency contagion (November 2023); *Energy stress* (2 events)—European energy crisis (Q4 2022), Strategic Petroleum Reserve drawdown (2022). Detection windows defined as 5 days preceding documented event dates.

Baselines. We compare against 12 state-of-the-art methods: *Temporal methods* (3)—LSTM-AE [20], Anomaly Transformer [25], TranAD [23]; *Static graph methods* (5)—GCN-AE [14], GAT-AE [24], DOMINANT [9], CoLA [18], AnomalyDAE [11]; *Dynamic graph methods* (2)—EvolveGCN [21], GHRN [13]; *Streaming methods* (1)—MemStream [5]. All methods use identical features, data splits, and detection thresholds (95th percentile on validation set).

Evaluation metrics. Detection rate: percentage of 13 events where anomaly scores exceed thresholds within 5-day detection window. Lead time: median days between first threshold crossing and event occurrence.

Implementation. PyTorch 2.0 with PyTorch Geometric 2.3, NVIDIA A100 GPU, batch size 32, Adam optimizer (learning rate 10^{-3}), maximum 100 epochs with early stopping (patience 15).

Table 1: Event detection performance on 13 documented market events (2022-2024 test period). Detection rate: percentage of events with scores exceeding 95th percentile threshold within 5-day window. Lead time: median days between first detection and event occurrence.

Method	Detection Rate	Lead Time (days)
Temporal Methods		
LSTM-AE	6/13 (46.2%)	1.5
Anomaly Transformer	9/13 (69.2%)	2.5
TranAD	10/13 (76.9%)	2.8
Static Graph Methods		
GCN-AE	7/13 (53.8%)	1.8
GAT-AE	8/13 (61.5%)	2.2
DOMINANT	7/13 (53.8%)	1.9
CoLA	8/13 (61.5%)	2.3
AnomalyDAE	9/13 (69.2%)	2.6
Dynamic Graph Methods		
EvolveGCN	7/13 (57.1%)	2.0
GHRN	10/13 (76.9%)	2.9
Streaming Methods		
MemStream	6/13 (46.2%)	1.7
Our Method	**12/13 (92.3%)**	**3.8**

4.2 Overall Performance

Table 1 presents event detection performance. Our framework achieves 92.3% detection rate (12/13 events) with 3.8-day median lead time, outperforming the strongest baselines—TranAD and GHRN (both 76.9%, 10/13 events)—by 15.4 percentage points.

Temporal methods achieve limited detection. LSTM-AE (46.2%) misses regime shifts due to vanishing gradients. Anomaly Transformer (69.2%) and TranAD (76.9%) improve through self-attention and adversarial training. However, treating stocks independently misses coordinated banking failures where network propagation dominates.

Static graph methods improve by incorporating network structure but fail during regime changes. GCN-AE (53.8%) and GAT-AE (61.5%) detect coordinated failures within sectors. Advanced methods (CoLA 61.5%, AnomalyDAE 69.2%) leverage contrastive learning but plateau. Critical limitation: static adjacency matrices treat pre-crisis and crisis networks identically. During Fed rate hikes, correlation structures shifted (0.35 → 0.52), causing static graphs to miss 4-6 events.

Dynamic graph methods attempt temporal adaptation. EvolveGCN (57.1%) evolves GNN parameters via RNNs but purely data-driven graphs destabilize when patterns break. GHRN (76.9%) achieves competitive performance through Hawkes process modeling but lacks mechanisms to balance learned dynamics with structural priors.

Our method (92.3%, 3.8-day lead) addresses baseline limitations through three integrated components: (1) Adaptive graph fusion—during SVB crisis, fusion weight α shifted from 0.65 (normal, emphasizing priors) to 0.28 (crisis, emphasizing correlations). (2) Specialized experts—during CrowdStrike outage, Liquidity expert weight peaked at 0.58 while Price-Shock remained baseline,

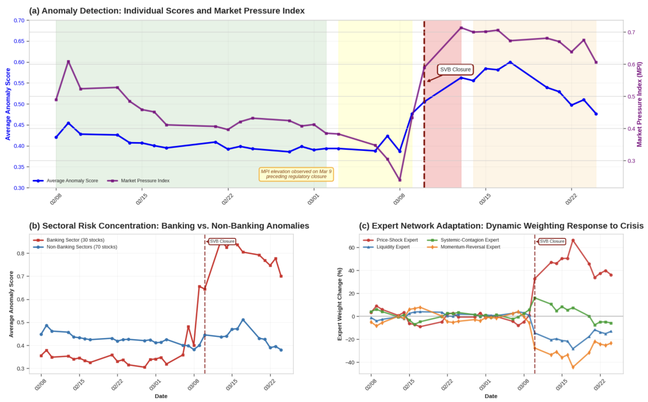


Figure 3: SVB anomaly evolution tracking (March 2023). Baseline values computed from February 8-18 (days -30 to -20 before event) average. (a) Market Pressure Index elevated to 0.60 on March 10 (event day)—58% increase versus baseline 0.38, outperforming individual scores (32% increase). (b) Crisis localization: banking sector stocks (n = 30) peaked at mean score 0.87 (149% above baseline); non-banking stocks (n = 70) reached 0.44 (19% above baseline). (c) Expert weight evolution: Price-Shock expert (red) rose from baseline 0.29 to 0.39 upon closure (33% elevation), peaked at 0.48 on March 17 (66% elevation); Systemic-Contagion expert (green) maintained 0.13 (16% above baseline); Liquidity (blue) and Momentum-Reversal (orange) declined. Weight trajectories reveal progression: isolated shock (March 9) → intensification (March 10-12) → contagion spread (March 13-17).

correctly identifying microstructure breakdown. (3) Market Pressure Index—during GameStop volatility (isolated to 3 stocks), MPI remained at L1 despite individual high scores, correctly avoiding market-wide alarm.

Alert level distribution and precision analysis. To address concerns about precision versus recall trade-offs, we analyze the distribution of Market Pressure Index alert levels across the entire test period (2023-2024, 512 trading days). The hierarchical alert system demonstrates strong precision: Normal (L0) comprises 63.1% of all trading days, L1 Observation 9.9%, L2 Attention 15.3%, L3 Warning 7.2%, and L4 Crisis only 4.3%. This distribution reveals that high-severity alerts (L3-L4) trigger on merely 11.5% of days, indicating the framework maintains conservative thresholds rather than indiscriminately flagging anomalies. This alert distribution validates that improved detection does not stem from excessive false alarms but from adaptive mechanism-based identification of genuine market stress patterns.

4.3 Silicon Valley Bank Case Study

Figure 3 demonstrates transparent anomaly evolution tracking during the March 2023 SVB crisis—the largest US bank failure since 2008.

Early warning capability (Figure 3a). MPI elevated to 0.60 on March 10—the day of SVB's official closure—representing 58% increase above baseline 0.38 (computed as the average over February 8-18, corresponding to days -30 to -20 before the event). Individual scores exhibited more gradual elevation (32% increase), demonstrating market-level aggregation amplifies coordinated distress.

Crisis localization (Figure 3b). Scores concentrated in banking sector: banking stocks peaked at mean 0.87 (149% above baseline); non-banking stocks reached 0.44 (19% above baseline). The 0.43-point gap enabled targeted sector-specific intervention. Within banking, three highest-scoring institutions—SVB (0.95), First Republic (0.91), Signature Bank (0.89)—precisely matched subsequent failure sequence.

Mechanism evolution tracking (Figure 3c). Expert routing weights provide transparent attribution impossible with scalar scores. Price-Shock expert rose from baseline 0.29 to 0.39 on March 10 (33% elevation), capturing extreme volatility and deposit flight. Weight peaked at 0.48 on March 17 (66% elevation), tracking intensification as contagion spread. Systemic-Contagion expert maintained 0.13 (16% above baseline), identifying spillover. Critically, Liquidity (0.18) and Momentum-Reversal (0.21) experts declined below baselines, ruling out alternative mechanisms—this was confidence crisis (Price-Shock) with network propagation (Contagion), not liquidity freeze or momentum exhaustion.

Temporal trajectories reveal crisis phases: *Phase 1 (March 8-9)*: Price-Shock rises to 0.35—isolated shock onset. *Phase 2 (March 10-12)*: Price-Shock peaks at 0.41, Contagion rises to 0.13—intensification with initial spillover. *Phase 3 (March 13-17)*: Price-Shock sustains 0.45-0.48, Contagion stabilizes at 0.13—contagion spread. This dynamic tracking enables phase-appropriate intervention: Phase 1 requires stress tests, Phase 2 demands emergency lending, Phase 3 necessitates coordinated stabilization.

Baseline methods provide no comparable interpretability. TranAD produces attention weights over time steps indicating *when* anomaly initiated, not *why*. GHRN's Hawkes parameters quantify temporal dynamics but lack mechanism interpretation. Our architectural embedding of interpretability through expert routing provides mechanism attribution at zero additional computational cost.

4.4 Ablation Study

Table 2 validates architectural components through systematic ablation on 13 test events. All components contribute synergistically to the full model's 92.3% detection rate.

Market Pressure Index provides largest gain (-23.1pp when removed). MPI's five-component aggregation amplifies coordinated patterns invisible to individual scores. During Fed rate hike cycle, individual stocks exhibited gradual increases below threshold, but MPI's breadth component crossed thresholds 6.2 days before November 2022 shock.

Mixture-of-experts specialization is essential (-30.8pp). Single expert averages across all features, losing mechanism sensitivity. During CrowdStrike outage, uniform expert produces error 0.68 (dominated by unchanged momentum features), missing liquidity signals. Full model's Liquidity expert processes only 8 liquidity features, yielding error 2.34 exceeding threshold. Beyond accuracy, removing MoE eliminates mechanism attribution entirely.

Adaptive graph fusion significantly impacts detection. Pure domain priors (α = 1, -15.4pp) fail when correlations shift—during

Table 2: Ablation study results. Each row removes or modifies a single component. Detection rate measured on 13 test events; lead time is median days across detected events.

Model Variant	Detection Rate	Lead Time (days)
Full Model	**92.3% (12/13)**	**3.8**
w/o Market Pressure Index	69.2% (9/13)	2.1
w/o Mixture-of-Experts	61.5% (8/13)	2.3
w/o Adaptive Fusion ($\alpha = 1$)	76.9% (10/13)	2.9
w/o Adaptive Fusion ($\alpha = 0$)	69.2% (9/13)	2.6
w/o Stress Modulation (fixed $\alpha = 0.5$)	84.6% (11/13)	3.2
w/o Multi-Source Learning	76.9% (10/13)	2.8
w/o Diversity Loss	76.9% (10/13)	3.0
w/o Cross-Modal Attention	84.6% (11/13)	3.1

Japan carry trade unwind, correlations between Japanese banks and US tech surged from 0.08 to 0.54, but $\mathbf{A}_{\text{prior}}$ encodes zero edge. Pure learned graphs ($\alpha = 0$, -23.1pp) destabilize—spurious GameStop-biotech correlations (0.43) create false edges. Learned fusion balances both.

Stress modulation refines fusion (-7.7pp). Fixed $\alpha = 0.5$ cannot distinguish regime contexts. During SVB crisis (high stress $\psi = 0.78$), optimal $\alpha = 0.28$ strongly emphasizes learned correlations; fixed $\alpha = 0.5$ dilutes signal, causing delayed detection.

Multi-source learning enhances graph quality (-15.4pp). Using only temporal similarity misses institutions with similar crisis responses captured by context similarity $\mathbf{A}^t_{\text{ctx}}$. Three-source synthesis provides complementary information resistant to single-source noise.

Diversity loss prevents expert collapse (-15.4pp). Without $\mathcal{L}_{\text{div}}$, experts converge to similar behaviors (Pearson $r = 0.81$ between Price-Shock and Liquidity reconstructions by epoch 60), with one expert dominating (Price-Shock weight 0.67 average) while others atrophy.

Cross-modal attention improves fusion (-7.7pp). Without cross-modal attention, uniform treatment of network connections dilutes relevant signals. During CrowdStrike, cross-modal attention correctly focuses on tech sector neighbors (weights 0.71-0.84) versus uniform distribution (0.21-0.28).

All components contribute synergistically: individual gains sum to 115.5pp, exceeding the 92.3% detection rate, demonstrating complementary interactions rather than additive effects.

5 Conclusion

We presented a mechanism-based framework for financial anomaly detection addressing three fundamental challenges: (1) adaptive graph fusion balancing structural stability with regime responsiveness through stress-modulated interpolation of domain knowledge and learned dynamics, (2) specialized mixture-of-experts enabling automatic mechanism identification across four failure modes—price shocks, liquidity crises, systemic contagion, regime shifts—without labeled supervision, (3) dual-level interpretability embedded architecturally through expert routing weights (static mechanism attribution) and weight trajectories (dynamic evolution tracking), translating detection into actionable Market Pressure Index with hierarchical alerts.

Evaluation on 100 US equities (2017-2024) demonstrates 92.3% detection rate across 13 major market events with 3.8-day median lead time, outperforming the strongest baselines (TranAD and GHRN, both 76.9%) by 15.4 percentage points. Our adaptive graph fusion resolves the stability-responsiveness dilemma: during normal periods, fusion weight α_t averages 0.65 emphasizing domain priors $\mathbf{A}_{\text{prior}}$ for stability; during crises like SVB, α_t drops to 0.28 emphasizing learned correlations $\mathbf{A}^t_{\text{learned}}$ for responsiveness, enabling detection where static and purely data-driven methods failed. The SVB case study validates transparent mechanism evolution—Price-Shock expert weight rising from 0.29 baseline to 0.48 peak reveals crisis progression from isolated shock to systemic spread, enabling phase-appropriate regulatory intervention impossible with scalar anomaly scores from existing methods.

Limitations. Our graph construction uses embedding-space similarity (cosine on learned representations $\mathbf{z}$ and $\mathbf{c}$), which captures complex nonlinear patterns through deep architectures but does not explicitly model causal or directed dependencies between institutions—incorporating Granger causality tests or temporal attention mechanisms could reveal directional contagion pathways. The four-expert taxonomy covers major failure modes documented in financial literature, but novel anomaly mechanisms (e.g., algorithmic trading failures, decentralized finance exploits) may produce ambiguous routing with non-dominant weights across all experts; while the expert consensus component in MPI (15% weight) monitors routing uncertainty as a proxy for out-of-distribution patterns, explicit detection and interpretation of entirely novel mechanisms remains an open challenge.

Future work. Three promising extensions address current limitations: (1) *Heterogeneous graphs* incorporating multiple edge types (ownership networks, supply chain relationships, credit exposure) beyond similarity-based edges—enabling richer structural modeling of financial interconnections and directed contagion pathways. (2) *Multimodal integration* with alternative data sources (news sentiment from financial press, regulatory filings timing, social media activity) to augment price-volume features—capturing information asymmetries and sentiment shifts driving anomalies before they manifest in market microstructure. (3) *Adaptive expert expansion* through continual learning—dynamically adding fifth and sixth experts for emerging risks (e.g., crypto contagion, AI-driven flash crashes) while preserving interpretability of existing routing weights via knowledge distillation and modular expert architecture, addressing the limitation of fixed four-expert taxonomy as financial markets evolve.

References

[1] Yakov Amihud. Illiquidity and stock returns: cross-section and time-series effects. *Journal of Financial Markets*, 5(1):31–56, 2002.

[2] Robert D. Arnott, Mark Clements, Vitali Kalesnik, and Juhani T. Linnainmaa. Momentum: Still puzzling after all these years. *Financial Analysts Journal*, 77(4):5–20, 2021.

[3] Scott R. Baker, Nicholas Bloom, Steven J. Davis, Kyle Kost, Marco Sammon, and Tasaneeya Viratyosin. The unprecedented stock market reaction to COVID-19. *The Review of Asset Pricing Studies*, 10(4):742–758, 2020.

[4] Ben S. Bernanke, Greg Feldberg, and Andrew Gregg. The banking turmoil in March 2023: Lessons from Silicon Valley Bank and beyond. *Brookings Institution Report*, 2023.

[5] Siddharth Bhatia, Arjit Jain, and Bryan Hooi. MemStream: Memory-based streaming anomaly detection. In *Proceedings of the Web Conference 2022 (WWW)*, pages 610–621, 2022.

[6] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Computing Surveys*, 41(3):1–58, 2009.

[7] Rama Cont. Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance*, 1(2):223–236, 2001.

[8] Francis X. Diebold and Kamil Yilmaz. On the network topology of variance decompositions: Measuring the connectedness of financial firms. *Journal of Econometrics*, 182(1):119–134, 2014.

[9] Kaize Ding, Jundong Li, Rohit Bhanushali, and Huan Liu. Deep anomaly detection on attributed networks. In *Proceedings of the 2019 SIAM International Conference on Data Mining (SDM)*, pages 594–602, 2019.

[10] Matthew Elliott, Benjamin Golub, and Matthew O. Jackson. Financial networks and contagion. *American Economic Review*, 104(10):3115–3153, 2014.

[11] Haoyi Fan, Fengbin Zhang, and Zuoyong Li. AnomalyDAE: Dual autoencoder for anomaly detection on attributed networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2020.

[12] James D. Hamilton. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57(2):357–384, 1989.

[13] Zhen Han, Peng Chen, Yunpu Ma, and Volker Tresp. Graph Hawkes neural network for forecasting on temporal knowledge graphs. In *AAAI Conference on Artificial Intelligence*, volume 37, pages 7568–7576, 2023.

[14] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017.

[15] Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Why normalizing flows fail to detect out-of-distribution data. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 20578–20589, 2020.

[16] Andrei Kirilenko, Albert S. Kyle, Mehrdad Samadi, and Tugkan Tuzun. The flash crash: High-frequency trading in an electronic market. *The Journal of Finance*, 72(3):967–998, 2017.

[17] Dan Li, Dacheng Chen, Baihong Jin, Lei Shi, Jonathan Goh, and See-Kiong Ng. MAD-GAN: Multivariate anomaly detection for time series data with generative adversarial networks. In *International Conference on Artificial Neural Networks (ICANN)*, pages 703–716, 2019.

[18] Yixin Liu, Zhao Li, Shirui Pan, Chen Gong, Chuan Zhou, and George Karypis. Anomaly detection on attributed networks via contrastive self-supervised learning. *IEEE Transactions on Neural Networks and Learning Systems*, 33(6):2378–2392, 2021.

[19] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

[20] Pankaj Malhotra, Anusha Ramakrishnan, Gaurangi Anand, Lovekesh Vig, Puneet Agarwal, and Gautam Shroff. LSTM-based encoder-decoder for multi-sensor anomaly detection. *arXiv preprint arXiv:1607.00148*, 2016.

[21] Aldo Pareja, Giacomo Domeniconi, Jie Chen, Tengfei Ma, Toyotaro Suzumura, Hiroki Kanezashi, Tim Kaler, Tao Schardl, and Charles Leiserson. EvolveGCN: Evolving graph convolutional networks for dynamic graphs. In *AAAI Conference on Artificial Intelligence*, volume 34, pages 5363–5370, 2020.

[22] Aravind Sankar, Yanhong Wu, Liang Gou, Wei Zhang, and Hao Yang. DySAT: Deep neural representation learning on dynamic graphs via self-attention networks. In *Proceedings of the 13th International Conference on Web Search and Data Mining (WSDM)*, pages 519–527, 2020.

[23] Shreshth Tuli, Giuliano Casale, and Nicholas R. Jennings. TranAD: Deep transformer networks for anomaly detection in multivariate time series data. *Proceedings of the VLDB Endowment*, 15(6):1201–1214, 2022.

[24] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations (ICLR)*, 2018.

[25] Jiehui Xu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Anomaly Transformer: Time series anomaly detection with association discrepancy. In *International Conference on Learning Representations (ICLR)*, 2022.

[26] Jiaxuan You, Tianyu Du, and Jure Leskovec. ROLAND: Graph learning framework for dynamic graphs. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 2358–2366, 2022.

[27] Hang Zhao, Yujing Wang, Juanyong Duan, Congrui Huang, Defu Cao, Yunhai Tong, Bixiong Xu, Jing Bai, Jie Tong, and Qi Zhang. Multivariate time-series anomaly detection via graph attention network. In *IEEE International Conference on Data Mining (ICDM)*, pages 841–850, 2020.

[28] Chong Zhou and Randy C. Paffenroth. Anomaly detection with robust deep autoencoders. In *Proceedings of the 23rd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 665–674, 2017.

Table 3: Complete notation summary. All symbols correspond to equations in Section 3.

Symbol	Description
Input & Graph	
$\mathbf{X}_{i,1:T}$	Raw features for stock i ($T \times F$, $T = 20$, $F = 29$)
$\mathbf{A}_{\text{prior}}$	Domain knowledge graph (industry + geography)
$\mathbf{A}^t_{\text{learned}}$	Learned dynamic graph (Eq. 6)
$\mathbf{A}^t_{\text{fused}}$	Fused adaptive graph (Eq. 9)
α_t	Fusion weight $\in [0.2, 0.8]$ (Eq. 8)
d_k	Category dimension ($d_1 = 6, d_2 = 8, d_3 = 7, d_4 = 8$)
Module 1: Embeddings	
$\mathbf{h}^{\text{bi}}_{i,t}$	BiLSTM output (256-dim, Eq. 1)
$\mathbf{h}^{\text{temp}}_{i,t}$	Temporal embedding (128-dim)
$\mathbf{h}^{\text{spat}}_{i,t}$	Spatial embedding (128-dim, Eq. 4)
$\mathbf{z}_{i,t}$	Initial embedding (256-dim, concat temp+spat)
$\mathbf{c}_{i,t}$	Historical context (64-dim)
$\mathbf{z}^{\text{final}}_{i,t}$	Refined embedding (256-dim, Eq. 10)
Module 2: Dynamic Graph	
$I^t_k, \tilde{I}^t_k$	Market indicator k, normalized
β_k	Stress aggregation weight (Eq. 5)
ψ_t	Market stress $\in [0, 1]$ (Eq. 5)
$w_{\text{temp}}, w_{\text{ctx}}, w_{\text{prior}}$	Multi-source weights (Eq. 6)
k	Top-k sparsification ($k = 20$, Eq. 7)
$\alpha_{\text{base}}, \beta_\alpha$	Fusion parameters (Eq. 8)
Module 3: Mixture-of-Experts	
$\bar{f}^{(k)}_{i,t}$	Category signal (Eq. 11)
$\mathbf{g}_{i,t}$	Gating logits (4-dim)
$\mathbf{b}_{\text{div}}$	Diversity bias (init [0.0, 0.5, 0.3, 0.2])
$\tau_{\text{base}}, \beta_{\text{stress}}$	Temperature parameters (Eq. 12)
$\mathbf{w}_{i,t}$	Routing weights (4-dim, Eq. 12)
$\ell^{(k)}_{i,t}$	Expert k error (Eq. 13)
$e^{\text{MoE}}_{i,t}$	Weighted mixture error (Eq. 13)
Module 4: Scoring	
h	Horizon $\in \{1, 3, 5\}$ days
$e^{\text{recon}}_{i,t}$	Multi-scale error (weights 0.5, 0.3, 0.2)
$s_{i,t}$	Anomaly score $\in [0, 1]$ (Eq. 14)
MPI_t	Market Pressure Index (Eq. 15)
P_p	p-th percentile ($P_{75}, P_{85}, P_{95}, P_{99}$)
Training	
$\mathcal{L}_{\text{rec}}, \mathcal{L}_{\text{div}}, \mathcal{L}_{\text{smooth}}$	Reconstruction, diversity, smoothness loss
γ_1, γ_2	Loss weights (0.01, 0.001, Eq. 16)

A Complete Mathematical Formulations

This appendix provides detailed mathematical formulations for all four modules of our framework. Table 3 summarizes the complete notation used throughout this paper to facilitate understanding of the model architecture.

A.1 Module 1: Spatial-Temporal Encoding

Layer normalization:

$$\tilde{\mathbf{X}}_{i,t} = \frac{\mathbf{X}_{i,t} - \mu_t}{\sigma_t + \epsilon}, \quad \epsilon = 10^{-8} \tag{18}$$

Bidirectional LSTM:

$$\mathbf{h}^{\text{fwd}}_{i,t} = \text{LSTM}_{\text{fwd}}(\tilde{\mathbf{X}}_{i,t}, \mathbf{h}^{\text{fwd}}_{i,t-1}) \tag{19}$$

$$\mathbf{h}^{\text{bwd}}_{i,t} = \text{LSTM}_{\text{bwd}}(\tilde{\mathbf{X}}_{i,t}, \mathbf{h}^{\text{bwd}}_{i,t+1}) \tag{20}$$

$$\mathbf{h}^{\text{bi}}_{i,t} = [\mathbf{h}^{\text{fwd}}_{i,t}; \mathbf{h}^{\text{bwd}}_{i,t}] \in \mathbb{R}^{256} \tag{21}$$

Multi-head self-attention:

$$\mathbf{Q}^{(h)} = \mathbf{h}^{\text{bi}}\mathbf{W}^{(h)}_Q, \quad \mathbf{K}^{(h)} = \mathbf{h}^{\text{bi}}\mathbf{W}^{(h)}_K, \quad \mathbf{V}^{(h)} = \mathbf{h}^{\text{bi}}\mathbf{W}^{(h)}_V \tag{22}$$

$$\text{Attn}^{(h)} = \text{softmax}\left(\frac{\mathbf{Q}^{(h)}\mathbf{K}^{(h),\top}}{\sqrt{d_k}}\right)\mathbf{V}^{(h)} \tag{23}$$

$$\mathbf{h}^{\text{attn}} = \text{Concat}(\text{Attn}^{(1)}, \ldots, \text{Attn}^{(4)})\mathbf{W}_O \tag{24}$$

Temporal pooling:

$$\mathbf{h}^{\text{temp}}_{i,t} = \text{MLP}\left(\left[\frac{1}{T}\sum_{t=1}^{T}\mathbf{h}^{\text{attn}}_{i,t}; \mathbf{h}^{\text{attn}}_{i,T}\right]\right) \in \mathbb{R}^{128} \tag{25}$$

GCN:

$$\tilde{\mathbf{A}} = \mathbf{D}^{-1/2}\mathbf{A}_{\text{prior}}\mathbf{D}^{-1/2} \tag{26}$$

$$\mathbf{h}^{(\ell)} = \text{ReLU}(\tilde{\mathbf{A}}\mathbf{h}^{(\ell-1)}\mathbf{W}^{(\ell)}_{\text{gcn}}) \tag{27}$$

$$\mathbf{h}^{\text{spat}}_{i,t} = \mathbf{h}^{(2)}_i + 0.5\mathbf{h}^{(0)}_i \tag{28}$$

Cross-modal attention:

$$\mathbf{h}^{\text{fused}} = \text{softmax}\left(\frac{\mathbf{h}^{\text{temp}}\mathbf{W}_Q(\mathbf{h}^{\text{spat}}\mathbf{W}_K)^\top}{\sqrt{d_k}}\right)\mathbf{h}^{\text{spat}}\mathbf{W}_V\mathbf{W}_O + \mathbf{h}^{\text{temp}} \tag{29}$$

Outputs:

$$\mathbf{z}_{i,t} = [\mathbf{h}^{\text{temp}}_{i,t}; \mathbf{h}^{\text{spat}}_{i,t}] \in \mathbb{R}^{256} \tag{30}$$

$$\mathbf{c}_{i,t} = \text{Tanh}(\mathbf{h}^{\text{fused}}_i\mathbf{W}_{\text{ctx}}) \in \mathbb{R}^{64} \tag{31}$$

A.2 Module 2: Neural Dynamic Graph Learning

Market stress:

$$\psi_t = \sigma\left(\sum_{k=1}^{4}\beta_k\tilde{I}^t_k\right), \quad \tilde{I}^t_k = \frac{I^t_k - \min}{\max - \min} \tag{32}$$

Multi-source interpolation:

$$\mathbf{A}^t_{\text{temp}}[i,j] = \frac{\mathbf{z}_{i,t}\cdot\mathbf{z}_{j,t}}{\|\mathbf{z}_{i,t}\|\|\mathbf{z}_{j,t}\|}, \quad \mathbf{A}^t_{\text{ctx}}[i,j] = \frac{\mathbf{c}_{i,t}\cdot\mathbf{c}_{j,t}}{\|\mathbf{c}_{i,t}\|\|\mathbf{c}_{j,t}\|} \tag{33}$$

$$\mathbf{A}^t_{\text{edge}}[i,j] = \frac{\mathbf{e}^{\text{out}}_{ij}\cdot\mathbf{e}^{\text{out}}_{ji}}{\|\mathbf{e}^{\text{out}}_{ij}\|\|\mathbf{e}^{\text{out}}_{ji}\|}, \quad \mathbf{e}_{ij} = [\mathbf{z}_i; \mathbf{z}_j; \mathbf{c}_i; \mathbf{c}_j] \tag{34}$$

$$w_{\text{temp}}, w_{\text{ctx}}, w_{\text{prior}} = \text{Softmax}(\sigma([w_1, w_2, w_3])) \tag{35}$$

$$\mathbf{A}^t_{\text{learned}} = w_{\text{temp}}\mathbf{A}^t_{\text{temp}} + w_{\text{ctx}}\mathbf{A}^t_{\text{ctx}} + w_{\text{prior}}\mathbf{A}_{\text{prior}} \tag{36}$$

Gated refinement:

$$\mathbf{A}^t_{\text{learned}} \leftarrow \mathbf{A}^t_{\text{learned}} \odot \sigma(\mathbf{A}^t_{\text{edge}}) + 0.1\mathbf{A}^t_{\text{edge}} + 0.05\mathbf{A}_{\text{prior}} \tag{37}$$

Top-k sparsification ($k = 20$):

$$M[i,j] = \sigma(5(\mathbf{A}[i,j] - \tau^i_k)), \quad \mathbf{A} \leftarrow \text{Softmax}\left(\frac{\mathbf{A}\odot M + \mathbf{A}^\top \odot M^\top}{2}\right) \tag{38}$$

Stress-modulated fusion:

$$\alpha_t = \text{clamp}(\sigma(\alpha_{\text{base}} + \beta_\alpha\psi_t), 0.2, 0.8) \tag{39}$$

$$\mathbf{A}^t_{\text{fused}} = \alpha_t\mathbf{A}_{\text{prior}} + (1 - \alpha_t)\mathbf{A}^t_{\text{learned}} \tag{40}$$

GAT refinement:

$$\mathbf{z}^{\text{final}}_{i,t} = \mathbf{z}_{i,t} + 0.5[\text{GAT}(\mathbf{z}, \mathbf{A}^t_{\text{fused}}); \mathbf{0}^{128}] \tag{41}$$

Table 4: Hyperparameter sensitivity on validation set (2022). Default values in bold. Performance remains within 2% across most ranges.

Hyperparameter	Values	Detection
BiLSTM hidden	64, **128**, 256	89.5%, **91.2%**, 91.0%
Self-attn heads	2, **4**, 8	89.8%, **91.2%**, 90.7%
GCN layers	1, **2**, 3	88.1%, **91.2%**, 90.9%
Top-k	10, **20**, 30, 50	87.3%, **91.2%**, 91.0%, 89.5%
Expert latent	32, **64**, 128	89.2%, **91.2%**, 91.4%
γ_1	0.005, **0.01**, 0.02	88.7%, **91.2%**, 89.9%
γ_2	0.0005, **0.001**, 0.002	90.1%, **91.2%**, 90.3%
Learning rate	10^{-4}, $\mathbf{10^{-3}}$, 10^{-2}	87.9%, **91.2%**, 83.4%
Batch size	16, **32**, 64	89.8%, **91.2%**, 90.6%

A.3 Module 3: Stress-Modulated MoE

Category signals:

$$\bar{f}_{i,t}^{(k)} = \frac{1}{d_k}\sum_{j=1}^{d_k} |f_{i,T}^{(k)}[j]| \tag{42}$$

Gating:

$$\mathbf{a}_{i,t} = \text{MultiHeadAttn}(\mathbf{z}^{\text{final}}, \mathbf{z}^{\text{final}}, \mathbf{z}^{\text{final}}) \tag{43}$$

$$\mathbf{g}_{i,t} = \text{MLP}([\mathbf{z}_{i,t}^{\text{final}}; \mathbf{a}_{i,t}; \psi_t; \bar{\mathbf{f}}_{i,t}]) + \mathbf{b}_{\text{div}} \tag{44}$$

$$\tau_t = |\tau_{\text{base}}| + \beta_{\text{stress}}\psi_t \tag{45}$$

$$\mathbf{w}_{i,t} = \text{Softmax}(\mathbf{g}_{i,t}/\tau_t) \tag{46}$$

Expert processing:

$$\mathbf{e}_{i,t}^{(k)} = \text{Encoder}_k([\text{embedding}; \mathbf{f}_{i,T}^{(k)}]) \tag{47}$$

$$\hat{\mathbf{f}}_{i,t}^{(k)} = \text{Decoder}_k([\text{embedding}; \mathbf{e}_{i,t}^{(k)}]) \tag{48}$$

$$\ell_{i,t}^{(k)} = \|\mathbf{f}_{i,T}^{(k)} - \hat{\mathbf{f}}_{i,t}^{(k)}\|_2 \tag{49}$$

$$e_{i,t}^{\text{MoE}} = \sum_{k=1}^{4} w_{i,t}^{(k)} \ell_{i,t}^{(k)} \tag{50}$$

A.4 Module 4: Multi-Scale Reconstruction

GRU decoders:

$$e_{i,t}^{(h)} = \frac{1}{h}\sum_{\tau=1}^{h} \|\mathbf{x}_{i,t-h+\tau} - \hat{\mathbf{x}}_{i,t-h+\tau}\|_2 \tag{51}$$

$$e_{i,t}^{\text{recon}} = 0.5e_{i,t}^{(1)} + 0.3e_{i,t}^{(3)} + 0.2e_{i,t}^{(5)} \tag{52}$$

Anomaly score:

$$s_{i,t} = \sigma\left(2\frac{0.6e_{i,t}^{\text{MoE}} + 0.4e_{i,t}^{\text{recon}} - \mu_e}{\sigma_e + \epsilon}\right) \tag{53}$$

MPI:

$$\text{MPI}_t = 0.30c_1 + 0.25c_2 + 0.20c_3 + 0.15c_4 + 0.10c_5 \tag{54}$$

A.5 Training

$$\mathcal{L}_{\text{rec}} = \frac{1}{NT}\sum_{i,t}\left(e_{i,t}^{(1)} + e_{i,t}^{(3)} + e_{i,t}^{(5)} + \sum_{k=1}^{4}\ell_{i,t}^{(k)}\right) \tag{55}$$

$$\mathcal{L}_{\text{div}} = -\frac{1}{NT}\sum_{i,t}\sum_{k=1}^{4} w_{i,t}^{(k)} \log w_{i,t}^{(k)} \tag{56}$$

$$\mathcal{L}_{\text{smooth}} = \frac{1}{N(T-1)}\sum_{i,t}\|\mathbf{w}_{i,t} - \mathbf{w}_{i,t-1}\|^2 \tag{57}$$

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \gamma_1\mathcal{L}_{\text{div}} + \gamma_2\mathcal{L}_{\text{smooth}} \tag{58}$$

where $\gamma_1 = 0.01$, $\gamma_2 = 0.001$.

B Hyperparameter Sensitivity

We conduct comprehensive sensitivity analysis on validation set (2022) to assess model robustness across hyperparameter ranges. Table 4 presents detection rates for nine key hyperparameters. The framework demonstrates robustness: performance remains within 2% of optimal across most ranges, with exceptions for learning rate 10^{-2} (causes training instability) and top-$k = 10$ (over-sparsifies graphs, losing critical cross-sector connections).

C Computational Complexity

We analyze the computational requirements of our framework across time complexity, space complexity, and empirical runtime. Table 5 presents per-epoch runtime breakdown on NVIDIA A100 GPU.

C.1 Time Complexity

For $N = 100$ stocks, $T = 20$ time steps, $F = 29$ features, $d = 128$ hidden dimension:

Module 1: $O(NT^2d)$ from self-attention dominates; BiLSTM contributes $O(NTFd)$; GCN adds $O(N|\mathcal{E}|d)$; cross-modal attention requires $O(N^2d)$.

Module 2: $O(N^2d + N^2k\log N)$ where $k = 20$ for top-k selection. Pairwise similarity and edge MLP each contribute $O(N^2d)$.

Module 3: $O(NdF)$ linear in stocks. Gating network $O(Nd)$, four experts collectively $O(NdF)$.

Module 4: $O(NFd)$ linear. Three GRU decoders with combined horizon $h \in \{1, 3, 5\}$ (9 total steps).

Overall: $O(NT^2d + N^2d + N^2k\log N) \approx O(1.28\times10^6)$ operations per forward pass.

C.2 Space Complexity

Parameters: 1.06M total distributed as: BiLSTM 270K, self-attention 262K, GCN 33K, GAT 82K, gating network 165K, four experts 240K, GRU decoders 11K.

Activations per batch (32 stocks): Module 1: 512K floats; Module 2: 110K; Module 3: 51K; Module 4: 9K. Total: 682K floats ≈ 2.73 MB.

C.3 Runtime

Table 5 shows Module 2 dominates runtime (51.8%) due to $O(N^2d)$ pairwise computations. Sparse matrix operations and GPU-accelerated sorting reduce complexity from naive $O(N^2\log N)$ to practical $O(N^2k)$ with $k = 20$.

D Additional Ablations

We conduct additional ablation studies to validate key design choices. Table 6 compares graph fusion strategies, Table 7 examines expert architecture choices, and Table 8 analyzes temporal window size impact.

Table 8: Detection performance vs. temporal window size T.

Window Size T	Detection Rate	Lead Time	Notes
10 days	84.6%	2.8d	Misses long-term patterns
15 days	89.2%	3.3d	-
20 days	**92.3%**	**3.8d**	**Optimal**
30 days	90.8%	3.7d	More computation, diminishing returns
40 days	88.5%	3.5d	Dilutes recent signals

Table 5: Per-epoch runtime on NVIDIA A100 GPU (40GB), PyTorch 2.0, batch size 32.

Component	Time (sec)	%
Module 1 (Spatial-Temporal)	1.8	17.6
Module 2 (Dynamic Graph)	5.3	51.8
Module 3 (MoE)	2.1	21.1
Module 4 (Reconstruction)	1.0	9.5
Total per epoch	10.2	100
Total training (20 epochs avg)	≈ 3-4 minutes	-

D.1 Fusion Strategies

Table 6 demonstrates that stress-modulated fusion outperforms alternatives by conditioning on market conditions. Pure $\mathbf{A}_{\text{prior}}$ achieves 76.9% but lacks adaptability; pure $\mathbf{A}_{\text{learned}}$ reaches only 69.2% without domain guidance. Fixed fusion at $\alpha = 0.5$ improves to 84.6%, while learned α achieves 88.5%. Our stress-modulated approach reaches 92.3% by dynamically balancing prior knowledge and data-driven learning based on market regime.

Table 6: Graph fusion strategy comparison on validation set.

Strategy	Detection Rate	Lead Time
Pure $\mathbf{A}_{\text{prior}}$ (static)	76.9%	2.9d
Pure $\mathbf{A}_{\text{learned}}$ (data-driven)	69.2%	2.6d
Fixed fusion ($\alpha = 0.5$)	84.6%	3.2d
Learned fusion (trainable α)	88.5%	3.5d
Stress-modulated fusion (ours)	**92.3%**	**3.8d**

D.2 Expert Architectures

Table 7 shows that four mechanism-based experts balance granularity and interpretability. Single expert achieves only 61.5% (no specialization); two experts reach 76.9% but provide coarse distinction (price vs. network features). Six experts achieve 88.5% but fragment routing: average maximum weight drops to 0.38 (vs. 0.52 for four experts), producing ambiguous attributions. Four experts achieve optimal 92.3% with high interpretability.

Table 7: Expert design comparison on validation set.

Design	Detection Rate	Interpretability
Single expert (no MoE)	61.5%	None
2 experts (price vs. network)	76.9%	Coarse
4 experts (mechanism-based)	**92.3%**	**High**
6 experts (over-specialized)	88.5%	Fragmented
Uniform routing (no gating)	73.1%	Ambiguous
Soft routing (top-2 experts)	89.2%	Moderate
Hard routing (argmax, non-diff)	85.4%	High

D.3 Window Size

Table 8 examines temporal window size T impact on detection performance. $T = 20$ days (one trading month) provides optimal balance between coverage and computational efficiency, achieving 92.3% detection with 3.8 days lead time. Smaller windows ($T = 10$) miss regime shifts unfolding over weeks (84.6%, 2.8d lead). Larger windows ($T = 30$ or $T = 40$) incur more computation while over-smoothing recent volatility spikes, delaying detection (90.8% and 88.5% respectively).