

## Article

# Multimodal Social Sensing with Hierarchical Consistency Constraints for Robust Detection of Social Financial Risk Patterns

Shangshan Chen <sup>1,2</sup>, Rong Fu <sup>2</sup>, Yi Zeng <sup>2</sup>, Yunfei Li <sup>2</sup>, Lirui Chen <sup>2,3</sup>, Jianan Xu <sup>2,3</sup> and Jinghui Yin <sup>2,4,\*</sup><sup>1</sup> School of Management, Beijing Sport University, Beijing 100084, China<sup>2</sup> National School of Development, Peking University, Beijing 100871, China<sup>3</sup> China Agricultural University, Beijing 100083, China<sup>4</sup> Artificial Intelligence Research Institute, Tsinghua University, Beijing 100084, China

\* Correspondence: yinjinghui\_thu@163.com

## Abstract

In social sensing environments, misinformation and coordinated manipulation often manifest through implicit semantic signals, complex behavioral dynamics, and highly coupled propagation structures. These factors pose significant challenges to artificial intelligence-driven sensing systems regarding data quality, multimodal fusion, and robustness. To address these issues, this study proposes an artificial intelligence-driven multi-granularity sensing framework. This framework integrates heterogeneous sensing signals from post-level semantic perception, user-level behavioral sensing, and group-level structural sensing into a unified representation space. Hierarchical consistency constraints enable cross-granularity sensing collaboration. This mechanism enhances stability and discriminative capability under complex and noisy data conditions. Methodologically, the framework jointly incorporates semantic sensing via text encoding, temporal sensing via behavioral sequence modeling, and structural sensing via graph neural network-based propagation. This integration effectively mitigates information bias induced by single-perspective sensing and improves the modeling of latent risk patterns. Experimental results on real-world datasets demonstrate that the proposed framework achieves significant improvements across multiple evaluation metrics. Specifically, it achieves a Precision of 0.847, a Recall of 0.812, an F1-score of 0.829, an Accuracy of 0.856, and an Area Under Curve of 0.913. It consistently outperforms traditional machine learning models, as well as mainstream deep learning and graph-based approaches. Furthermore, comparison experiments validate the complementarity among semantic, behavioral, and structural sensing signals. The full model achieves an improvement of more than 3 percentage points in the F1-score compared to single-granularity configurations. An ablation study further demonstrates that each sensing module contributes substantially to performance enhancement, with the semantic sensing and hierarchical consistency constraints playing particularly critical roles. Overall, the proposed method exhibits a strong capability to handle complex heterogeneous sensing data. It improves robustness and enhances cross-level information utilization, providing an effective solution to data-related challenges in artificial intelligence-driven sensing systems.



Academic Editor: Jose María Alvarez Rodríguez

Received: 19 March 2026

Revised: 9 June 2026

Accepted: 2 July 2026

Published: 7 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

**Keywords:** artificial intelligence-driven sensing; multimodal sensor fusion; social sensing systems; behavioral sequence modeling; explainable sensing intelligence

## 1. Introduction

As social media platforms have gradually become important channels for the dissemination of financial information, the sharing of investment advice, and discussions of crypto

assets, a large volume of finance-related opinions, forecasts, and product recommendations is rapidly propagated through posts, comments, and reposting behaviors [1]. While such information diffusion patterns improve the efficiency of market information circulation, they simultaneously provide new opportunities for the spread of misleading financial posts and fraudulent activities [2]. In recent years, financial fraud and misleading investment promotion have exhibited increasingly socialized and organized characteristics. Risky financial posts often appears in the form of seemingly professional posts, while fraudulent accounts attempt to establish superficial credibility through long-term normal participation. Moreover, coordinated interactions among multiple accounts may be used to fabricate artificial consensus and market popularity, thereby misleading users into making erroneous financial decisions [3]. Consequently, social financial risk can no longer be regarded merely as a problem of whether a single piece of text is false or whether an individual account behaves abnormally. Instead, it should be viewed as a multi-level collaborative process involving content semantics, user behavior, and group interaction structures [4]. The construction of joint detection models capable of integrating multi-granularity social contextual information is therefore of substantial practical significance for improving financial risk control capability and enhancing platform governance efficiency [5].

Traditional research methods typically model this problem from a single analytical perspective [6]. The first category focuses on content-based misinformation detection, where misleading statements and exaggerated promotional expressions are identified through keyword rules, sentiment features, or text classification models [7]. Although these methods are relatively simple to implement and possess certain interpretability, the historical behavioral background of information publishers and the structure of social information propagation are often ignored. As a result, such approaches can easily be bypassed by content that appears linguistically compliant but is driven by abnormal motivations [8]. The second category concentrates on user-level fraud detection, in which anomalous accounts are identified using statistical behavior features, activity patterns, and temporal characteristics [9]. However, because these approaches are largely detached from specific semantic content, it becomes difficult to distinguish subtle differences between legitimate marketing activities and fraudulent financial behavior [10]. The third category conducts anomaly detection or coordinated fraud identification based on social graph structures, where graph connectivity patterns, community structures, and propagation paths are modeled. Nevertheless, such approaches rely primarily on structural features and lack the ability to capture semantic conflicts and risk expression patterns embedded in textual content [11]. Overall, most existing methods independently model risk signals at a single level, such as posts, users, or groups. Cross-level information fusion and consistency analysis mechanisms remain largely absent, which makes the identification of coordinated multi-account financial manipulation particularly challenging [12].

With the development of deep learning techniques, pretrained language models, graph neural networks, and multimodal learning approaches have gradually been introduced into social media risk detection tasks [13]. At the textual level, transformer-based semantic encoding models have significantly improved the accuracy of financial text understanding and misinformation identification. At the user level, sequence models and representation learning techniques enable the modeling of behavioral trajectories and user interest patterns. At the structural level, graph neural networks can effectively represent interaction relationships and information propagation paths among users [14]. Several recent studies have attempted to jointly learn textual features and graph structural features, achieving certain improvements in performance [15]. However, these approaches generally remain limited to feature concatenation or simple multi-task learning strategies, lacking explicit hierarchical modeling mechanisms and cross-granularity constraint designs. As

a consequence, the consistency and conflict relationships among risk signals across content, users, and groups have not yet been systematically characterized, and the capability of detecting complex coordinated manipulation scenarios remains limited [16]. Mingfu Xiong et al. [4] proposed the spatial multi-granularity feature exploration (SMGFE) model, in which a comprehensive representation was constructed through coarse–medium–fine multi-layer feature fusion and human spatial association mechanisms. Superior accuracy and efficient time performance were achieved on standard person re-identification datasets, thereby enhancing re-identification capability for social risk situation assessment. Xianjun Dai et al. [5] proposed the double quantification multi-granularity kernel fuzzy rough set (DQ-MGKFRS) model, which utilized kernel functions, multi-layer granularity structures, and a dual-quantification strategy to effectively extract information from mixed-attribute medical big data. Efficient medical decision-making performance and robust risk prediction indicators were obtained in rheumatoid arthritis risk assessment tasks. Shuo Yu et al. [13] introduced the LLM4graph classification framework consisting of graph2text and graph2token paradigms, which systematically categorized methods combining graph learning and large language models into textualization and tokenization approaches. Four major research challenges and practical design guidelines were analyzed, providing a comprehensive reference for advancing the application of large language models in graph data learning and analysis. Through multi-scale convolutional neural networks (CNNs) feature extraction and fusion, automatic classification of environmental risk levels was achieved, demonstrating high accuracy and strong classification performance on industrial risk datasets, with the aim of improving automated environmental risk assessment capability in green finance. Qin Zhao et al. [17] proposed the multi-interest and social interest field framework (MISIF), in which dynamic routing was employed to generate multiple user-interest embeddings while constructing a social interest field to enhance the representational capability of social recommendation systems. Superior recommendation accuracy and financial security risk identification performance were achieved across several public datasets.

To address the above limitations, the present study approaches social financial risk detection from the perspective of multi-granularity social context modeling. The detection task is divided into three complementary levels, namely the post level, the user level, and the group level. A unified joint detection framework is constructed to enable collaborative cross-level information modeling and consistency-constrained learning.

The main contributions of this study can be summarized as follows.

1. An artificial intelligence-driven sensing framework is proposed for multi-granularity social risk detection. We explicitly clarify that our core methodological innovation does not lie in inventing new underlying neural network architectures, but rather in proposing a joint alignment mechanism across heterogeneous entities tailored for the unique dynamics of financial risk propagation. Unlike traditional multi-view learning, which typically aligns multiple modalities of a single entity, our framework achieves cross-hierarchical alignment from posts to users and to groups. Heterogeneous sensing signals derived from post semantics, user behavioral sequences, and group interaction structures are jointly modeled within a unified architecture, enabling comprehensive perception of complex risk patterns beyond single-level sensing paradigms.
2. A hierarchical graph–text sensing fusion network is designed to integrate semantic sensing from textual data with structural sensing from social interaction graphs, thereby capturing both fine-grained content signals and large-scale relational propagation characteristics in a unified sensing representation space.
3. A cross-granularity sensing consistency and conflict-aware constraint mechanism is introduced, where multi-level sensing outputs are jointly regularized through

consistency alignment and discrepancy penalization, allowing the model to explicitly capture cooperative and contradictory sensing patterns across semantic, behavioral, and structural dimensions.

4. Extensive experiments on real-world datasets validate that the proposed sensing-driven framework achieves superior performance in precision, recall, and robustness, while demonstrating enhanced capability in detecting coordinated manipulation behaviors through multi-source sensing signal integration.

## 2. Related Work

### 2.1. Misinformation Detection in Social Media

In the field of misleading financial posts detection in social media, early studies were primarily established on the basis of textual semantic analysis. The fundamental principle of these approaches is to determine whether information exhibits misleading, exaggerated, or false characteristics through linguistic features [18]. In such methods, financial posts or comments are typically treated as independent textual samples, and classifiers are constructed using keyword matching, sentiment tendency analysis, rhetorical pattern recognition, or statistical language feature modeling [19]. With the development of machine learning techniques, research gradually shifted from manual feature engineering toward automatic semantic representation learning based on deep neural networks. For example, convolutional neural networks and recurrent neural networks have been employed to capture contextual dependencies in financial texts, thereby enabling the identification of misleading investment promises, high-return promotional expressions, and risk-concealing statements [20]. In recent years, pretrained language models have been widely adopted in this domain. The core idea of these models is to obtain general semantic representation capability through large-scale corpus pretraining and subsequently perform domain-specific fine-tuning on financial corpora, thereby improving the understanding of financial terminology, risk expressions, and complex contextual semantics [21]. Although these approaches have achieved significant effectiveness in post-level misinformation identification, their underlying assumption generally implies that financial risk is primarily reflected in the textual content itself [22]. In real social financial environments, however, a large portion of fraudulent information deliberately maintains professional and neutral linguistic expressions. Consequently, relying solely on semantic models makes it difficult to distinguish compliant marketing activities from fraudulent inducements [23]. Furthermore, such approaches often ignore the historical behavioral trajectories of information publishers and the social propagation background of posts. As a result, comprehensive characterization of source credibility and propagation structure anomalies remains limited, which can lead to misclassification or missed detection when encountering highly disguised financial fraud content with seemingly normal semantic expressions [24].

### 2.2. Fraud and Dishonest User Detection on Social Platforms

For financial fraud and dishonest user detection on social platforms, another line of research shifts the modeling focus from textual content to account behavior. The fundamental principle of these methods is to identify abnormal individuals by analyzing long-term behavioral patterns exhibited by users on the platform [25]. In such approaches, a user-level feature space is typically constructed, including indicators such as posting frequency, interaction density, changes in follower relationships, temporal distribution of content publishing, and cross-topic participation degree. Based on these features, statistical learning or deep representation learning methods are employed to build user risk scoring models [26]. Within temporal modeling frameworks, several studies utilize sequence neural networks or temporal attention mechanisms to characterize the evolution trajectory of user behavior,

thereby identifying sudden behavioral shifts and strategic disguise behaviors [27]. With the development of graph representation learning, graph neural networks have also been introduced into user-level fraud detection tasks. The basic idea is to leverage interaction relationships and connection patterns among users, and to learn the structural position and relational context of users within social networks through neighborhood aggregation mechanisms. This enables the discovery of abnormal connection patterns and suspicious community structures [28]. Such approaches significantly enhance the capability to detect group-oriented fraud and relationship-driven risk behaviors [29]. However, user-level methods that rely solely on behavioral and structural features generally lack the ability to understand the semantic content of published information, making it difficult to determine the actual risk attributes of user-generated content [30]. In financial scenarios, legitimate marketing accounts and fraudulent accounts may exhibit highly similar activity levels and dissemination strategies. Without supporting semantic evidence from content, reliance solely on behavioral patterns can easily lead to confusion [31]. In addition, user-level modeling typically treats each account as an independent prediction target, lacking linkage mechanisms with post-level and group-level risk signals. Consequently, cross-granularity risk inference cannot be effectively achieved [32].

### *2.3. Graph-Based Coordinated Manipulation and Abnormal Propagation Analysis*

Graph-based coordinated manipulation and abnormal propagation analysis studies characterize risk behaviors from a more macroscopic network perspective. The fundamental principle of these approaches is to abstract social platforms as complex networks composed of multiple types of nodes and multi-relational edges, and to identify anomalous subgraphs and coordinated groups through structural pattern mining and graph representation learning [33]. In such methods, researchers typically construct user–user interaction graphs, user–content propagation graphs, or multimodal heterogeneous graphs. Community detection algorithms, subgraph matching techniques, or graph neural network models are then applied to identify densely connected abnormal groups, synchronized behavioral patterns, and unnatural information propagation paths [34]. Some studies further incorporate temporal dimensions by constructing temporal propagation graphs, enabling the capture of abnormal burst diffusion patterns occurring within short time intervals. This facilitates the detection of opinion manipulation and coordinated promotion behaviors [35]. These approaches possess unique advantages in detecting group fraud and coordinated manipulation, as organized risk behaviors can be discovered from a global structural perspective [36]. Nevertheless, structure-driven approaches often weaken or ignore textual semantic information, making it difficult to explain why a specific group is considered risky. Furthermore, distinguishing naturally formed communities based on shared interests from anomalous coordinated groups formed for manipulation purposes remains challenging. In addition, graph-structure-based methods usually focus on single-level modeling, concentrating detection targets at the user node or subgraph level. Fine-grained alignment mechanisms with post-level semantic risks are generally absent, resulting in limitations in semantic interpretability and risk attribution of detection results.

## **3. Materials and Method**

### *3.1. Data Collection*

The dataset constructed in this study is designed from the perspective of artificial intelligence-driven social sensing, where heterogeneous data sources are regarded as multi-modal sensing signals reflecting content semantics, user behaviors, and interaction structures. Specifically, data were collected from several major social platforms, including Xiaohongshu, Weibo, Xueqiu, and Reddit, which function as large-scale social sensing envi-

ronments for capturing public investment discussions and information diffusion dynamics, as shown in Table 1. The collection period spans from January 2023 to June 2024, covering long-term temporal sensing of user activities and information propagation processes. Data acquisition was conducted through a combination of official interfaces and customized crawling programs, where a domain-specific financial keyword set was constructed to trigger semantic sensing of relevant posts. These keywords include stock identifiers, investment strategy expressions, return-related phrases, cryptocurrency names, and financial product tags, enabling the system to selectively capture high-risk semantic signals embedded in user-generated content. For each retrieved post, multiple sensing attributes were recorded, including textual content as semantic sensing signals, timestamps as temporal sensing signals, and engagement metrics such as likes, comments, and repost counts as implicit behavioral sensing indicators. After noise filtering and semantic validation, posts with clear domain relevance were retained as post-level sensing data.

**Table 1.** Statistics of the social sensing dataset with multi-source sensing signals.

| Data Type                     | Description                                                                      | Quantity |
|-------------------------------|----------------------------------------------------------------------------------|----------|
| Semantic Sensing Data (Posts) | Textual sensing signals collected from Xiaohongshu, Weibo, Xueqiu and Reddit     | 284,615  |
| User Nodes (Sensing Entities) | Unique user accounts representing sensing agents across platforms                | 46,372   |
| Behavioral Sensing Sequences  | Time-series sensing data of user activities and interactions                     | 648,230  |
| Structural Sensing Edges      | Interaction-based sensing signals including reply, repost, and mention relations | 792,514  |
| High-Risk Semantic Signals    | Posts labeled as misleading or high-risk information                             | 33,742   |
| Abnormal Sensing Entities     | Users identified as suspicious or coordinated accounts                           | 4185     |

From a behavioral sensing perspective, user activity trajectories were reconstructed within the observation window. Historical posting sequences, temporal intervals between actions, topic participation distributions, and interaction frequencies were extracted to form structured behavioral sensing sequences, which reflect dynamic user patterns over time. In addition, structural sensing was performed by collecting user interaction relations, including reply, repost, and mention links, which were further used to construct user–user and user–content interaction graphs. Each interaction edge is associated with temporal attributes, enabling the modeling of propagation paths and diffusion rhythms as dynamic sensing processes. For label construction, candidate risk samples were first identified through abnormal sensing patterns, including high-frequency promotion behaviors and anomalous propagation structures. To ensure the reliability of the labels, a rigorous expert annotation protocol was established. Five independent annotators with professional backgrounds in finance and risk management were recruited. A comprehensive annotation guideline was formulated prior to the formal phase, strictly defining the boundaries between legitimate financial marketing and fraudulent inducement. During the annotation process, each sample was independently reviewed by three different annotators. In cases of disagreement regarding ambiguous borderline samples, a majority voting mechanism was applied. For highly complex cases, a senior financial compliance expert was involved to make the final determination. To quantitatively evaluate the quality and consistency of the annotation, the Fleiss Kappa statistic was calculated. The overall Fleiss Kappa scores for the post-level and user-level annotation tasks reached 0.82 and 0.79, respectively, indicating a substantial level of inter-annotator agreement. Based on this process, post-level labels indicate misleading or exaggerated information, while user-level labels identify suspicious or coordinated accounts. Through this multi-source sensing data integration process, a unified dataset is formed, consisting of semantic sensing data, behavioral sensing sequences,

and structural sensing graphs, providing a comprehensive foundation for multi-granularity sensing-based risk detection.

### 3.2. Data Preprocessing and Augmentation Strategy

In the multi-granularity social context-driven joint detection framework, data preprocessing and data augmentation are fundamental steps that directly determine the effectiveness and stability of cross-level modeling. Raw social financial data typically exhibit characteristics such as substantial textual noise, sparse behavioral records, complex relational structures, and severely imbalanced class distributions. Without systematic preprocessing and augmentation strategies, representation learning may become biased, graph structures may be distorted, and model training may become unstable. Therefore, unified modeling and constraints are applied to the raw data from multiple perspectives to ensure the processed data can be effectively utilized by the hierarchical graph-text fusion network.

For textual data preprocessing, the fundamental principle is to reduce noise interference through normalization so that textual representations better approximate the true semantic distribution. The raw post text is first cleaned using rule-based filtering to remove non-semantic symbols, URLs, emoji characters, and duplicated noise segments. Subsequently, the cleaned text is tokenized and mapped into an indexed vector sequence. A pretrained language model is then utilized to convert this sequence into a contextual semantic vector. To maintain a stable semantic foundation across the dataset, sequences are truncated or padded to a uniform length, with attention masks ensuring that the network focuses only on valid tokens.

To address the severe class imbalance inherent in financial risk datasets, we adjust the optimization focus to prevent the model from favoring the majority benign classes. Instead of relying solely on standard empirical risk minimization, we employ a dual strategy. First, we utilize class-weighted sampling during the construction of mini-batches to ensure adequate representation of minority risk samples. Second, we integrate a focal loss mechanism into the optimization objective, which dynamically down-weights easily classified normal samples and heavily penalizes misclassifications of hard, minority fraudulent samples.

For data augmentation, the central principle is to expand the sample distribution through controlled perturbations while preserving label semantic consistency, thereby improving the robustness of the model against evasion tactics. In semantic augmentation, small continuous perturbations, typically drawn from a Gaussian distribution, are directly injected into the textual representation vectors. Alternatively, semantically equivalent textual representations are generated using synonym substitution. In structural augmentation, random sparse perturbations are introduced into the graph adjacency matrix to simulate missing interactions or noisy connections. Through joint semantic and structural perturbation training, the multi-granularity model maintains stable performance and achieves robust identification of disguised coordinated manipulation patterns in complex environments.

### 3.3. Proposed Method

#### 3.3.1. Overall

In the proposed multi-granularity social context-driven joint detection framework, the model design is constructed upon three types of preprocessed input representations, namely post-level textual representations, user-level behavioral sequence representations, and user-post interaction graph representations. First, the post text sequence is fed into a text encoding backbone network to obtain a contextual semantic vector for each post. This representation not only preserves local expression features such as financial terminology, return commitments, risk concealment, and emotional inducement, but also captures the

global semantic structure of the entire text. Subsequently, the general semantic representation is projected into a post-level risk subspace through a risk-aware mapping layer, producing a post-level risk representation and an initial risk probability.

Meanwhile, the historical behavioral sequence corresponding to each post is input into the behavioral encoding module. Through temporal modeling mechanisms, dynamic features such as posting frequency, interaction density, temporal interval variation, and topic transition trajectories are extracted to form a user behavioral representation. This representation is not utilized independently but is aligned and fused with the corresponding post semantic representation to measure the consistency between current content and historical behavioral patterns, thereby enabling the identification of disguised accounts that exhibit normal textual semantics but abnormal behavioral trajectories. Subsequently, both post node representations and user node representations are jointly injected into the social interaction graph as initial node states, where relational propagation and aggregation are performed along edges representing replies, reposts, and mentions. Through multi-layer message passing, local neighborhood influence and group-level interaction patterns are progressively learned. After multiple rounds of propagation, the representation of each user is no longer limited to its own content and behavioral features but also incorporates collaborative signals from its associated propagation group, thereby producing a group-level risk representation for detecting abnormal synchronized posting, mutual endorsement, and concentrated dissemination behaviors.

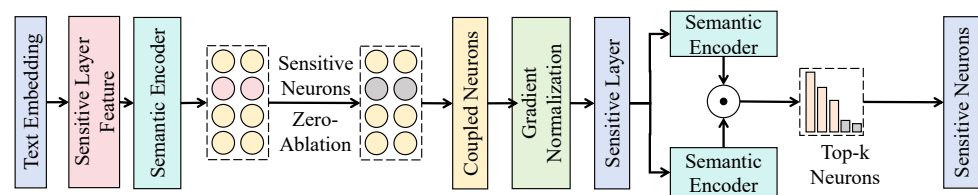
Based on these representations, three prediction heads are constructed for post-level, user-level, and group-level risk estimation, respectively, enabling fine-grained risk outputs. Furthermore, a cross-level consistency mechanism is introduced to jointly constrain the three predictions, forming a mutually reinforced decision structure among high-risk posts, high-risk users, and high-risk groups. When inconsistencies arise across different levels, corrective signals from other levels can be utilized to adjust the prediction. Ultimately, the entire framework establishes a unified processing pipeline consisting of semantic encoding, behavioral consistency modeling, graph-based relational propagation, and cross-level joint inference, thereby achieving end-to-end joint detection of misleading financial posts, fraudulent users, and coordinated manipulation groups.

### 3.3.2. Post-Level Semantic Risk Modeling Module

In the post-level semantic risk modeling module, textual inputs are first mapped into a continuous semantic space through an embedding layer. As shown in Figure 1, the original post text is tokenized into a sequence and represented as a word embedding matrix  $X \in \mathbb{R}^{L \times d}$ , where  $L$  denotes the sequence length and  $d$  denotes the embedding dimension. The representation is then processed by a sensitive feature transformation layer to construct a financial semantic feature space through linear mapping and nonlinear activation, which can be formulated as  $H = \sigma(WX + b)$ , where  $W$  represents the learnable weight matrix,  $b$  denotes the bias term, and  $\sigma(\cdot)$  is a nonlinear activation function. This transformation compresses general semantic information into representations that are more relevant to financial risk expressions, thereby emphasizing signals such as return promises, investment inducement, and risk concealment.

Subsequently, the feature representation is passed through a sparse autoencoder module for semantic factor decomposition. The encoding process can be expressed as  $Z = f(W_e H)$ , where  $W_e$  denotes the encoder weight matrix and  $f(\cdot)$  represents a sparse activation function. Through sparsity constraints, only a small subset of neurons is activated, resulting in interpretable semantic factor representations. These neurons are treated as latent semantic concept units, among which certain neurons exhibit strong responses to financial risk expressions and are thus regarded as sensitive neurons. To suppress redundant

semantic information, a zero-ablation mechanism is applied, where non-critical neuron activations are masked to maintain a sparse representation, formulated as  $Z' = M \odot Z$ , where  $M$  denotes a binary mask matrix and  $\odot$  represents element-wise multiplication. Based on the sparse representation, a coupled neuron structure is introduced to model interactions among risk-related semantic factors. The coupled representation is computed as  $C = \phi(W_c Z')$ , where  $W_c$  denotes the coupling weight matrix and  $\phi(\cdot)$  represents a nonlinear transformation. This structure enables the capture of co-occurrence patterns in financial semantics, such as the joint presence of return commitment and urgency signals, which often indicate higher risk. To prevent gradient interference among different semantic dimensions, a gradient orthogonalization mechanism is introduced, where the inner product between gradients is constrained to approach zero, i.e.,  $\nabla C_a \cdot \nabla C_b \approx 0$ . This design ensures independent learning across semantic channels and improves discriminative capability. At the output stage, a dual-path sensitive layer further refines the representation, and the most informative neurons are selected through a Top-K operation, yielding  $R = \text{TopK}(C)$  as the final post-level risk representation. This representation is then fed into a classifier to obtain the post-level risk probability. Such a design enables the decomposition of complex financial semantics into interpretable components, reduces noise through ablation, enhances interaction modeling through coupling, and focuses on the most discriminative features, thereby improving robustness and generalization for downstream multi-level inference.



**Figure 1.** Illustration of the post-level semantic risk modeling module. The input text is first mapped into a continuous semantic space and processed through a sensitive feature transformation layer. A sparse autoencoder equipped with a zero-ablation mechanism isolates sensitive semantic factors, which are subsequently modeled by coupled neurons with gradient orthogonalization and selected via a Top-K operation to form the final semantic risk representation.

### 3.3.3. User-Level Behavioral Consistency Modeling Module

In the user-level behavioral consistency modeling module, the objective of the model is to characterize the temporal consistency relationship between the historical behavioral trajectory of a user and the current posted content, thereby identifying potential risk accounts with normal semantics but abnormal behaviors.

As shown in Figure 2, let the behavioral sequence of a user within a temporal window be represented as a matrix  $B \in \mathbb{R}^{T_b \times d_b}$ , where  $T_b$  denotes the length of the behavioral time steps and  $d_b$  denotes the behavioral feature dimension. The behavioral features include posting action type embeddings, interaction object embeddings, and time interval encodings. In order to unify the representation space for subsequent sequence modeling, a one-dimensional convolutional network is first adopted for local pattern extraction. This convolutional layer uses a three-layer structure, in which the kernel size of each layer is  $k$ , the stride is  $s$ , and the number of channels is progressively expanded as  $c_1, c_2, c_3$ , thereby obtaining local dynamic features of behavior, which can be written as

$$H_b = \sigma(\text{Conv}(B; W_b) + b_b), \quad (1)$$

where  $W_b$  denotes the convolution kernel weight parameters and  $\sigma(\cdot)$  denotes the nonlinear activation function. The output features of the convolutional layers are then processed

by layer normalization to obtain the behavioral feature matrix  $H_b \in \mathbb{R}^{T_b \times d_m}$ , where  $d_m$  denotes the unified modeling dimension. Meanwhile, the post-level semantic vector  $h^{text}$  is expanded into a matrix representation  $H_t$  with the same length as the behavioral sequence, and semantic pattern extraction is further performed through one-dimensional convolution:

$$H_t = \sigma(\text{Conv}(H^{text}; W_t) + b_t). \quad (2)$$

In order to measure the temporal matching relationship between behavior and semantics, a cross-temporal matching loss mechanism is introduced. First, a similarity matrix  $S$  between behavioral embeddings and semantic embeddings is calculated, where each element represents the matching degree between behavioral features and textual semantics at different moments:

$$S_{ij} = \frac{H_b^{(i)} \cdot H_t^{(j)}}{|H_b^{(i)}| |H_t^{(j)}|}. \quad (3)$$

On the other hand, a temporal proximity matrix  $T$  is constructed according to behavioral timestamps, where each element reflects the temporal closeness between two behaviors:

$$T_{ij} = \exp\left(-\frac{|t_i - t_j|}{\tau}\right), \quad (4)$$

where  $\tau$  denotes the temporal scale parameter. The cross-temporal consistency loss is implemented by minimizing the difference between the semantic matching matrix and the temporal proximity matrix:

$$\mathcal{L}_{ctm} = \|S - T\|_F^2. \quad (5)$$

This loss function ensures that behaviors with similar semantic expressions remain relatively close in time, thereby strengthening the temporal consistency constraint of behavioral patterns. Subsequently, the behavioral sequence features are input into a Transformer encoder for long-range dependency modeling. The encoder consists of a multi-layer self-attention structure, and each layer contains a multi-head attention module and a feed-forward network. Let the input feature be  $Z^{(l)}$ , then the attention computation can be expressed as

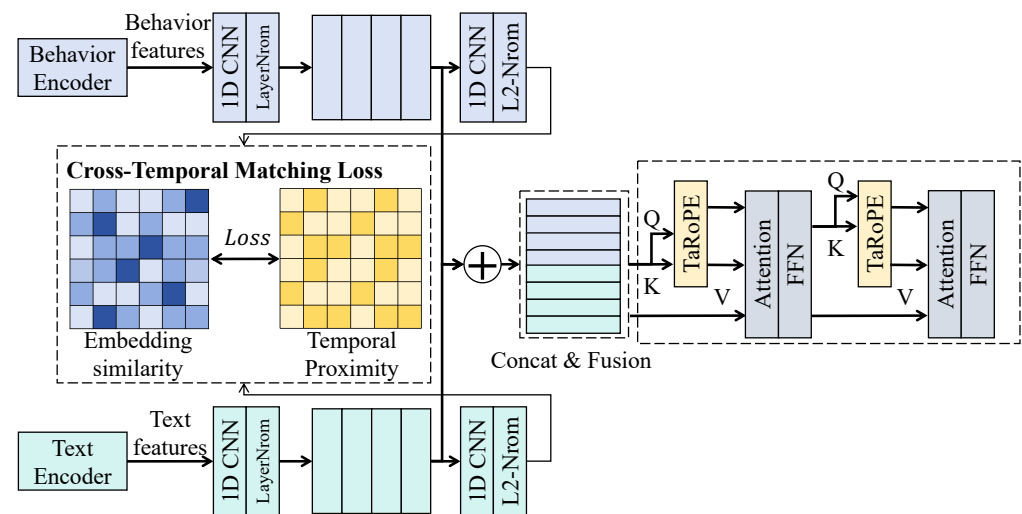
$$\text{Attn}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V, \quad (6)$$

where  $Q = Z^{(l)}W_Q$ ,  $K = Z^{(l)}W_K$ , and  $V = Z^{(l)}W_V$ . In order to preserve temporal order information, a temporal rotary positional encoding function  $\Phi(t)$  is introduced before the attention computation, thereby obtaining the temporal-aware attention representation:

$$\tilde{Q} = Q + \Phi(t), \quad \tilde{K} = K + \Phi(t). \quad (7)$$

After propagation through multiple encoder layers, the user behavioral sequence is finally compressed into a global behavioral representation vector  $h^{user}$ . This representation is fused with the post semantic representation for subsequent risk prediction. The advantage of this module design lies in its ability to simultaneously model user activity patterns from the three dimensions of time, semantics, and behavior. The convolutional structure can capture local behavioral variations, the cross-temporal matching loss maintains consistency between semantics and behavior over time, and the Transformer structure can learn long-range behavioral dependency relationships. Through this joint modeling strategy, abnormal shifts between behavioral trajectories and textual semantics can be effectively identified, such as cases in which users suddenly change investment opinions or intensively promote certain financial products within a short period of time. As a result, the

identification capability for strategically disguised accounts is significantly enhanced, and stable user behavioral representations are provided for subsequent group-level coordinated manipulation detection.



**Figure 2.** Illustration of the user-level behavioral consistency modeling module. One-dimensional convolutional networks extract local temporal patterns from both behavioral sequences and textual features. A cross-temporal matching loss explicitly aligns semantic expressions with behavioral timestamps, while a Transformer encoder utilizing temporal rotary positional encoding captures long-range behavioral dependencies to evaluate user consistency.

### 3.3.4. Group-Level Coordinated Manipulation Modeling Module

In the group-level coordinated manipulation modeling module, the objective of the model is to identify coordinated propagation patterns and structural abnormal relationships among multiple users in the social network, thereby discovering potential manipulation groups.

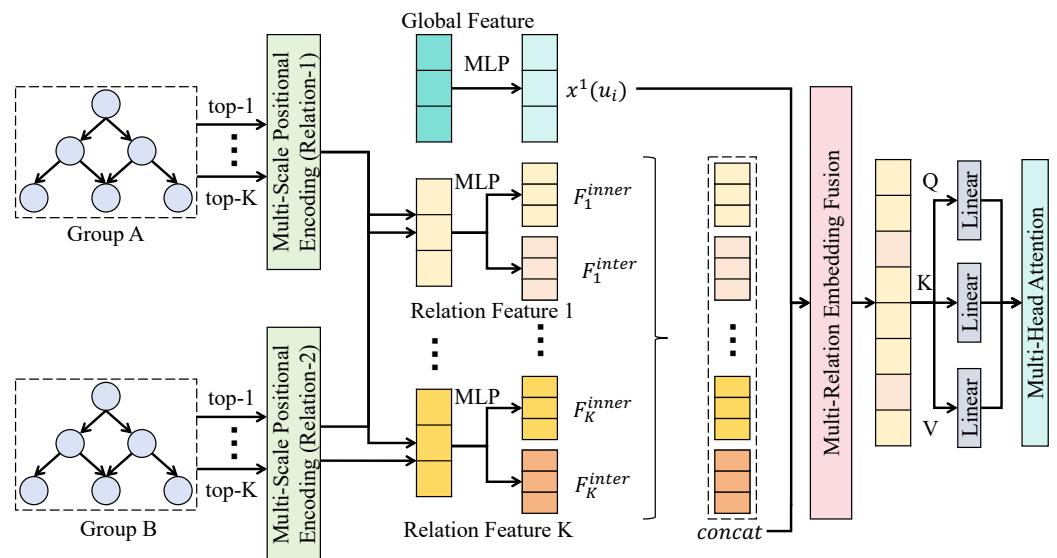
As shown in Figure 3, let the social network be represented as a graph structure  $G = (V, E)$ , where the node set  $V$  contains user nodes and post nodes, and the edge set  $E$  represents interaction relationships such as replies, reposts, or mentions. For each node, its initial representation is obtained by fusing the post semantic representation and the user behavioral representation, denoted as the matrix  $X \in \mathbb{R}^{|V| \times d_0}$ . In order to capture propagation patterns in group structures, the module adopts a multi-layer graph convolutional structure for relational propagation modeling. First, a normalized adjacency matrix  $\hat{A}$  is constructed, and the node representation is mapped into a new feature space. The graph convolution propagation process can be expressed as

$$Z^{(l)} = \sigma(\hat{A}Z^{(l-1)}W^{(l)}), \quad (8)$$

where  $Z^{(l)}$  denotes the node representation at the  $l$ -th layer,  $W^{(l)}$  denotes the learnable weight matrix, and  $\sigma(\cdot)$  is the nonlinear activation function. In order to enhance the expressive capability of the model for local propagation structures, a three-layer propagation structure is adopted in the network, and the feature dimension is progressively expanded in each layer, enabling nodes to absorb information from a larger neighborhood range. During the propagation process, node representations not only contain their own semantic and behavioral features, but also gradually integrate information from neighboring nodes, thereby forming group-level contextual representations. In order to characterize the structural compactness of coordinated manipulation groups, a group similarity measurement function is introduced on top of the node representations. Let the output representations of

two nodes at the propagation layer be  $z_i$  and  $z_j$ , respectively, then the structural similarity can be defined as

$$\omega_{ij} = \frac{z_i^\top z_j}{|z_i||z_j|}. \quad (9)$$



**Figure 3.** Illustration of the group-level coordinated manipulation modeling module. Node representations combining semantic and behavioral features are propagated through a multi-layer graph convolutional structure. The module calculates structural similarity to enforce group consistency constraints and employs multi-head attention to aggregate relational features across the social graph, ultimately yielding the group-level risk probability.

This measurement is used to quantify the similarity between two users in the behavioral propagation space. When the similarity among multiple nodes remains at a high level, it indicates that these nodes may have participated in coordinated propagation behaviors. In order to further characterize the compactness of group structures, a group consistency constraint is introduced by maximizing the structural similarity among nodes within the same potential group while minimizing the representational similarity among different groups. This constraint can be written as

$$\mathcal{L}_{group} = \sum_{(i,j) \in \mathcal{P}} (1 - \omega_{ij}) + \sum_{(i,j) \in \mathcal{N}} \max(0, \omega_{ij} - \delta), \quad (10)$$

where  $\mathcal{P}$  denotes the set of node pairs within potential coordinated groups,  $\mathcal{N}$  denotes the set of non-coordinated node pairs, and  $\delta$  is the similarity threshold. Through this objective function, clear group structural boundaries can be formed in the embedding space. After obtaining node-level structural representations, neighborhood information is aggregated through a graph pooling operation, thereby producing the group-level risk representation. Let the node set  $C_k$  denote a potential group, then its group representation can be obtained through the aggregation function:

$$g_k = \frac{1}{|C_k|} \sum_{v_i \in C_k} z_i. \quad (11)$$

Subsequently, the group representation is input into the risk discrimination network to obtain the group-level manipulation probability. In order to demonstrate that this structure can effectively capture coordinated propagation behaviors, analysis can be conducted from the perspective of graph signal propagation. The essence of graph convolution propagation

is equivalent to performing multi-order smoothing operations on the adjacency matrix. When multiple nodes form dense connections within a short period of time, their representations gradually converge through neighborhood propagation, thereby enabling the group representation to form obvious clusters in the embedding space. Therefore, potential manipulation groups can be identified by detecting high-density clustered regions in the embedding space. This module has important advantages in the task of social financial risk detection. First, through the graph propagation mechanism, both user behavioral features and propagation structural information can be simultaneously utilized, allowing abnormal groups to be identified from the perspective of the overall network. Second, the structural similarity constraint can strengthen the consistency characteristics within groups, thereby improving the detection capability for coordinated manipulation patterns. Finally, the group representation obtained through graph pooling can be jointly inferred together with post-level and user-level risk representations, thereby enabling cross-granularity risk identification and allowing the model to identify multi-account coordinated manipulation behaviors more accurately in complex social propagation environments.

### 3.3.5. Hierarchical Consistency Constraints and Joint Optimization Objective

In the multi-granularity social context joint detection framework, the three levels of post, user, and group characterize semantic content risk, individual behavioral risk, and structural propagation risk, respectively. If only a single classification loss function in traditional supervised learning is adopted for training, each level model is usually optimized as an independent task. For example, the post classification model depends only on post labels, the user model depends only on user labels, and the group structural model depends on graph structural labels. Under such circumstances, the model optimization objective can be expressed as the simple summation of the losses of each subtask, and its essence is local empirical risk minimization. However, in social financial scenarios, risk signals often exhibit obvious cross-level coupling characteristics. For example, for a user who continuously publishes high-risk investment information, the user-level risk probability should remain consistent with the post-level risk prediction. At the same time, if multiple users belong to the same abnormal propagation group, the group-level risk should also impose constraints on the user-level prediction. Therefore, optimization based only on traditional loss functions is insufficient to characterize such cross-granularity risk dependency relationships, which may lead to prediction deviations in which semantics are correct but structure is abnormal, or behavior is abnormal but semantics appear normal.

To address this issue, a hierarchical consistency constraint mechanism is introduced into the joint learning framework so that risk predictions at different levels remain consistent in the same embedding space, as shown in Algorithm 1. Although formulated mathematically as a straightforward squared-difference regularization, this design conceptually functions as a specialized smoothing prior imposed on a heterogeneous graph. It forces a representational compromise between local content risk and global structural risk within the shared embedding space. This specific formulation effectively counters advanced disguise tactics commonly seen in financial fraud, such as when users maintain perfectly compliant content semantics but exhibit highly coordinated and anomalous manipulation behaviors at the group level. Let the post-level risk prediction probability be  $p_i$ , the corresponding user-level risk probability be  $u_j$ , and the risk probability of the group to which the user belongs be  $g_k$ . First, a consistency constraint between posts and users is defined so that post risk prediction remains close to user risk prediction, and the constraint function can be expressed as

$$\mathcal{L}_{pu} = \sum_{(i,j)} (p_i - u_j)^2, \quad (12)$$

where  $(i, j)$  denotes the mapping relationship between a post and its publishing user. This constraint guarantees that when a user continuously publishes high-risk content, the corresponding user-level risk score can be synchronously increased. Second, in order to model the structural consistency relationship between users and groups, a group consistency constraint function is introduced:

$$\mathcal{L}_{ug} = \sum_{(j,k)} (u_j - g_k)^2, \quad (13)$$

where  $(j, k)$  denotes the association relationship between a user and its belonging group. This constraint ensures coordination between group-level risk and individual behavior, thereby strengthening the identification capability of group-level coordinated patterns. In the overall optimization objective, traditional classification losses and hierarchical consistency constraints are jointly integrated, yielding the final joint optimization function:

---

**Algorithm 1** Hierarchical Consistency-Constrained Joint Optimization

---

**Require:** Post representations  $H_{post}$ , user representations  $H_{user}$ , graph representations  $H_{group}$

- 1: Initialize model parameters  $\theta$
  - 2: **for** each training epoch **do**
  - 3:   Compute post prediction  $p_i = f_{post}(H_{post})$
  - 4:   Compute user prediction  $u_j = f_{user}(H_{user})$
  - 5:   Compute group prediction  $g_k = f_{group}(H_{group})$
  - 6:   Compute classification losses
  - 7:      $L_{post} = CE(p_i, y_{post})$
  - 8:      $L_{user} = CE(u_j, y_{user})$
  - 9:      $L_{group} = CE(g_k, y_{group})$
  - 10:   Compute hierarchical consistency constraints
  - 11:      $L_{pu} = \sum (p_i - u_j)^2$
  - 12:      $L_{ug} = \sum (u_j - g_k)^2$
  - 13:   Compute total loss
  - 14:      $L_{total} = L_{post} + L_{user} + L_{group} + \lambda_1 L_{pu} + \lambda_2 L_{ug}$
  - 15:   Update parameters  $\theta$  using gradient descent
  - 16: **end for**
- 

$$\mathcal{L}_{total} = \mathcal{L}_{post} + \mathcal{L}_{user} + \mathcal{L}_{group} + \lambda_1 \mathcal{L}_{pu} + \lambda_2 \mathcal{L}_{ug}, \quad (14)$$

where  $\mathcal{L}_{post}$ ,  $\mathcal{L}_{user}$ , and  $\mathcal{L}_{group}$  denote the supervised classification losses at the three levels, respectively, and  $\lambda_1$  and  $\lambda_2$  are weighting coefficients used to control the intensity of the consistency constraints. From the optimization perspective, this joint objective function is equivalent to introducing structural regularization into the original empirical risk minimization problem. According to convex optimization theory, when correlated constraints exist among prediction variables, squared-difference penalty terms can effectively reduce the variance among different predictions, thereby making the solution space more stable. Therefore, during gradient descent, the risk representations at different levels gradually become consistent, thereby forming a unified risk discrimination space. It can be further demonstrated that when  $\lambda_1$  and  $\lambda_2$  take appropriate values, the model optimization process is equivalent to imposing graph smoothing constraints among the prediction variables at the three levels, causing node risk values along the same semantic propagation chain

to become consistent, thereby improving the model's capability to identify coordinated manipulation patterns.

## 4. Results and Discussion

### 4.1. Experimental Configuration

#### 4.1.1. Hardware and Software Platform

All experiments were conducted on a dedicated deep learning workstation to ensure efficient processing of the large-scale social financial dataset. The hardware configuration includes an Intel Xeon Gold 6248R CPU (3.0 GHz), an NVIDIA GeForce RTX 4090 GPU (24 GB VRAM), 128 GB of DDR4 RAM, and a 2 TB NVMe SSD. This setup effectively satisfies the memory demands of large-scale graph neighborhood aggregation and long-text encoding while preventing I/O bottlenecks during batch data loading.

The software environment is based on Ubuntu 22.04 LTS and Python 3.10. The neural network components, including the text encoder and the joint loss optimization module, were implemented using PyTorch 2.0.1. Graph structure modeling and message passing operations were constructed via PyTorch Geometric (PyG) 2.3.0. Pre-trained language models were loaded and fine-tuned using HuggingFace Transformers 4.30.0, while data processing relied on NumPy 1.24.1 and Pandas 2.0.0. To optimize training efficiency, CUDA 11.8 and PyTorch's Automatic Mixed Precision (AMP) were utilized. Finally, strict reproducibility was ensured across all training runs by fixing random seeds and enforcing deterministic cuDNN operations.

Regarding hyperparameter and training strategy settings, to ensure evaluation fairness, model generalization capability, and strictly prevent structural data leakage, the dataset is partitioned using a rigorous chronological inductive split strategy rather than a conventional random transductive split. Since temporal dependencies and the evolution of manipulation tactics are critical in financial risk prediction, the dataset is strictly divided based on timestamps. Specifically, data collected during the first twelve months are allocated to the training set, while data from the subsequent six months are strictly reserved for the validation and test sets, naturally maintaining an approximate proportion of 70 percent, 10 percent, and 20 percent, respectively. To completely eliminate the risk of overly optimistic performance estimates, an inductive learning setting is rigorously enforced. The testing and validation sets remain completely unseen during the training phase. The graph structures, including all user nodes, post nodes, and interaction edges that emerge in the testing period, are entirely isolated from the training graph. Furthermore, for user accounts spanning both temporal periods, their dynamic behavioral sequences are strictly truncated at the temporal splitting boundary, ensuring that no future behavioral patterns or subsequent graph connections leak into the historical training phase. A chronological time-series cross-validation strategy is adopted during model training to ensure the statistical robustness of the results while maintaining strict temporal isolation.

To guarantee complete experimental reproducibility, the implementation details and parameter settings of the key modules are explicitly defined. The sparse autoencoder expands the 768-dimensional initial text embeddings into a 1024-dimensional latent space, where a zero-ablation threshold of 0.01 is applied to isolate sensitive semantic factors. The coupled neurons project these factors into a 256-dimensional representation space, and the Top-K selection operation is configured with K set to 15 to retain only the most critical risk indicators. In the user-level consistency module, the one-dimensional convolutional network utilizes a kernel size of 3, and the Transformer encoder is configured with 4 attention heads and a hidden dimension of 256 to process the temporal sequences. The entire framework is optimized using the AdamW optimizer with beta values of 0.9 and 0.999, incorporating a weight decay of  $1e-4$  to regularize the complex parameter space. The

initial learning rate is set to  $1e-4$ , governed by an exponential decay schedule with a decay factor of 0.95 applied at the end of each epoch. The batch size is set to 32 according to GPU memory capacity. The maximum number of training epochs is set to 50, combined with an early stopping strategy, where training is terminated if no improvement is observed on the validation set for 5 consecutive epochs. The text encoding dimension is 768, the number of graph neural network layers is set to 3, and the hidden layer dimension is uniformly aligned to 256 across all modules. A dropout rate of 0.3 is adopted to mitigate overfitting. The consistency constraint weight coefficient in the joint loss is determined through grid search on the validation set within the range of 0.2 to 0.5, achieving a stable balance between multi-granularity consistency and primary task performance.

#### 4.1.2. Baseline Models and Evaluation Metrics

In the multi-granularity social financial risk detection task, a variety of representative baseline models are selected for comparative analysis to comprehensively evaluate the effectiveness of the proposed method. Logistic Regression [37], as a classical linear model, is characterized by strong interpretability and computational efficiency, enabling the rapid establishment of fundamental decision boundaries. Random Forest [38], through the ensemble of multiple decision trees, exhibits strong nonlinear modeling capability and robustness, maintaining stable performance under complex feature combinations. CNN [39] effectively extracts local semantic features via convolutional operations and demonstrates strong capability in textual pattern recognition. BiLSTM [40] captures contextual dependencies through bidirectional temporal modeling, enabling a more comprehensive understanding of sequential information. Transformer [41], based on the self-attention mechanism, possesses global dependency modeling capability and achieves superior performance in long-text semantic understanding. FinBERT [21] is selected as a modern domain-specific pretrained language model designed to extract highly refined risk expressions from financial texts. LLaMA-3 [42] is introduced to evaluate the performance of cutting-edge large language models using a few-shot prompting strategy for semantic risk classification. GCN [43] integrates neighborhood information through graph-based propagation, effectively leveraging relational features among nodes. GAT [44] further introduces an attention mechanism to adaptively weight neighboring nodes, enhancing the flexibility and expressiveness of structural modeling. These baseline models characterize semantic, temporal, and structural features from different perspectives, providing a solid comparative foundation for validating the advantages of multi-granularity modeling in the proposed method.

In the multi-granularity social financial risk joint detection task, commonly used evaluation metrics include Precision, which measures the proportion of correctly predicted risk samples among all predicted risk samples; Recall, which measures the proportion of true risk samples that are successfully identified; F1, which provides a harmonic balance between precision and recall; Accuracy, which measures the overall proportion of correctly predicted samples; and AUC, which characterizes the overall discrimination ability of the model under different decision thresholds.

$$Precision = \frac{TP}{TP + FP}, \quad (15)$$

$$Recall = \frac{TP}{TP + FN}, \quad (16)$$

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}, \quad (17)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \quad (18)$$

$$TPR = \frac{TP}{TP + FN}, \quad FPR = \frac{FP}{FP + TN}, \quad (19)$$

$$AUC = \int_0^1 TPR(FPR), d(FPR). \quad (20)$$

In the above formulations,  $TP$  denotes the number of positive samples correctly identified as risk by the model,  $FP$  denotes the number of negative samples incorrectly identified as risk,  $TN$  denotes the number of negative samples correctly identified as normal, and  $FN$  denotes the number of positive samples incorrectly identified as normal.  $TPR$  represents the true positive rate,  $FPR$  represents the false positive rate, and AUC denotes the area under the ROC curve.

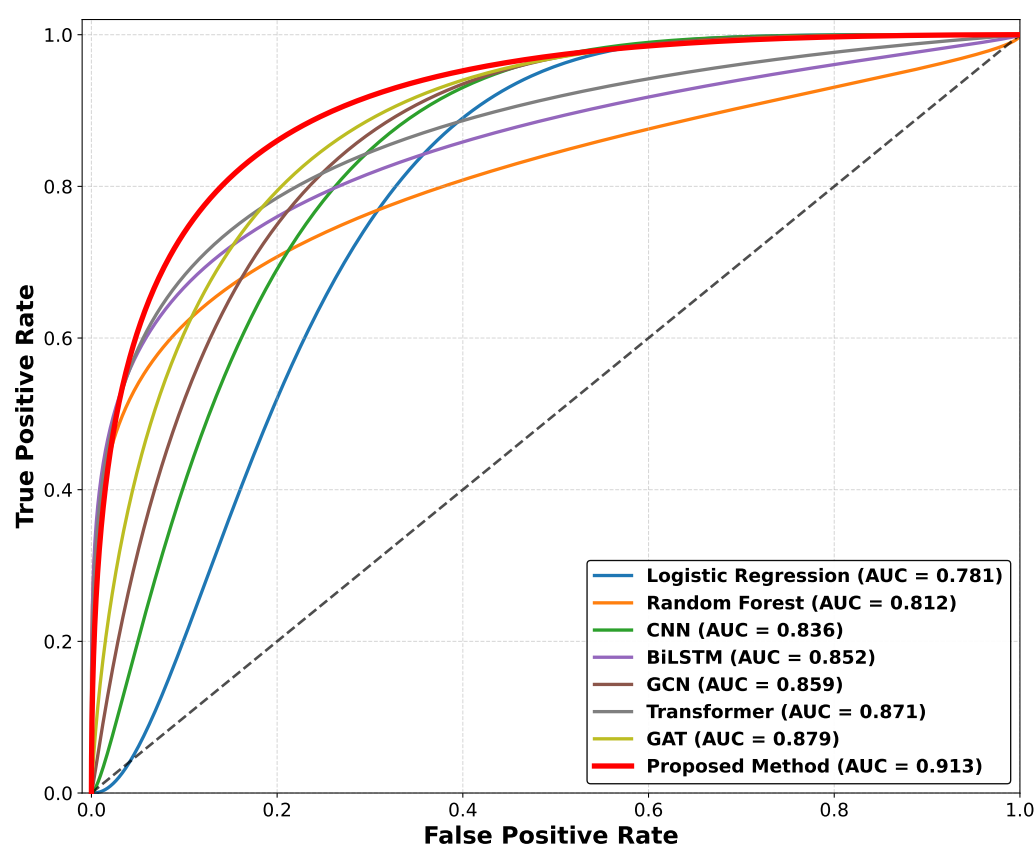
#### 4.2. Overall Performance Comparison

The purpose of this experiment is to validate the effectiveness of the proposed multi-granularity social context-driven joint detection method from the perspective of overall performance in complex social financial risk identification tasks, and to evaluate the comprehensive advantages of the model in semantic modeling, behavioral modeling, and structural modeling through comparisons with different types of baseline models.

As shown in Table 2 and Figure 4, the traditional linear model Logistic Regression achieves the lowest performance across all metrics, indicating that reliance on linearly separable features is insufficient for capturing latent semantics and complex interaction patterns in financial texts. The ensemble-based Random Forest shows improved performance, suggesting that nonlinear feature combinations partially alleviate the limitations of single-feature representations; however, its lack of temporal and contextual modeling capability restricts further improvement. Among deep learning models, CNN extracts local textual patterns through convolution and improves performance, but its limited receptive field constrains its ability to capture long-range dependencies. BiLSTM models bidirectional sequential dependencies, resulting in improved Recall and F1, highlighting the importance of sequence modeling in textual understanding. Transformer, relying on global self-attention mechanisms, demonstrates stronger capability in modeling long-range dependencies, leading to further performance gains. Compared with these single-view approaches, the proposed method achieves consistently superior performance across all evaluation metrics simultaneously. While performance improvements in machine learning often involve traditional trade-offs among evaluation metrics, particularly between precision and recall, the uniform elevation observed in our framework stems from a dual-calibration mechanism inherent to the multi-granularity design. Specifically, the post-level semantic feature extraction module effectively filters out explicit false signals and misleading text expressions, which directly improves local precision. Concurrently, the group-level graph propagation module successfully uncovers highly disguised fringe nodes that attempt to blend into the community through normal content semantics but exhibit anomalous synchronized or dense structural linkages, which drastically increases the overall recall. Although adjusting decision thresholds at a localized level will inevitably trigger the classic precision-recall trade-off, our cross-level hierarchical consistency constraints elevate the fundamental quality of the entire shared representation space. By mutually calibrating heterogeneous signals early in the representation learning process rather than at the final decision boundary, the model successfully mitigates independent perspective variances. This holistic optimization facilitates a collective leap in both precision and recall boundaries, which ultimately drives the comprehensive improvement across the integrated F1-score, accuracy, and AUC metrics.

**Table 2.** Overall performance comparison with baseline models on the social financial risk dataset. Results are presented as mean  $\pm$  standard deviation over five-fold cross-validation. Statistical significance against the strongest baseline is confirmed by a paired *t*-test ( $p < 0.05$ ).

| Model               | Precision         | Recall            | F1-Score          | Accuracy          | AUC               |
|---------------------|-------------------|-------------------|-------------------|-------------------|-------------------|
| Logistic Regression | $0.712 \pm 0.008$ | $0.681 \pm 0.009$ | $0.696 \pm 0.008$ | $0.724 \pm 0.007$ | $0.781 \pm 0.009$ |
| Random Forest       | $0.745 \pm 0.007$ | $0.703 \pm 0.008$ | $0.723 \pm 0.007$ | $0.751 \pm 0.006$ | $0.812 \pm 0.008$ |
| CNN                 | $0.768 \pm 0.006$ | $0.741 \pm 0.007$ | $0.754 \pm 0.006$ | $0.779 \pm 0.006$ | $0.836 \pm 0.007$ |
| BiLSTM              | $0.781 \pm 0.005$ | $0.756 \pm 0.006$ | $0.768 \pm 0.005$ | $0.792 \pm 0.005$ | $0.852 \pm 0.006$ |
| Transformer         | $0.804 \pm 0.004$ | $0.772 \pm 0.005$ | $0.788 \pm 0.005$ | $0.813 \pm 0.004$ | $0.871 \pm 0.005$ |
| FinBERT             | $0.815 \pm 0.004$ | $0.783 \pm 0.004$ | $0.799 \pm 0.004$ | $0.824 \pm 0.004$ | $0.883 \pm 0.004$ |
| LLaMA-3 (Few-shot)  | $0.826 \pm 0.003$ | $0.790 \pm 0.004$ | $0.808 \pm 0.004$ | $0.833 \pm 0.003$ | $0.892 \pm 0.004$ |
| GCN                 | $0.792 \pm 0.006$ | $0.765 \pm 0.007$ | $0.778 \pm 0.006$ | $0.801 \pm 0.005$ | $0.859 \pm 0.006$ |
| GAT                 | $0.811 \pm 0.005$ | $0.781 \pm 0.006$ | $0.796 \pm 0.005$ | $0.820 \pm 0.005$ | $0.879 \pm 0.005$ |
| Proposed Method     | $0.847 \pm 0.003$ | $0.812 \pm 0.004$ | $0.829 \pm 0.003$ | $0.856 \pm 0.003$ | $0.913 \pm 0.003$ |



**Figure 4.** ROC curves of different models on the social financial risk detection task, illustrating their comparative classification performance.

Crucially, the inclusion of modern competitive baselines provides profound insights. FinBERT yields noticeable improvements in precision over standard text models, proving the efficacy of domain-specific pretraining in uncovering hidden financial risks. Furthermore, the modern large language model LLaMA-3 evaluated under a few-shot setup achieves the highest text-based performance across all single-modality baselines. It is objectively observed that large language models are exceptionally powerful in post-level semantic understanding, accurately capturing subtle nuances, rhetorical patterns, and implicit financial inducements. For graph-based models, GCN aggregates neighborhood information to incorporate structural context, but its uniform aggregation introduces noise.

GAT introduces attention-based weighting over neighbors, effectively enhancing structural modeling, and thus outperforms GCN in Precision and AUC.

Compared with all these approaches, including the state-of-the-art text models, the proposed method consistently achieves the best performance across all metrics, demonstrating the absolute superiority of multi-granularity joint modeling. Although large language models represent a massive leap forward in textual analysis, they suffer from an inherent limitation in social platform governance: they are structurally incapable of directly ingesting and processing massive social network graphs that possess highly complex chronological behaviors and advanced topological dependencies. Consequently, text-centric large language models remain completely blind to coordinated multi-account manipulation and synchronized bot deployments where individual post content appears normal but structural linkages are anomalous. In contrast, our proposed multi-granularity framework jointly embeds post semantics, user sequence actions, and group structures into a unified representation space. By enforcing cross-level hierarchical consistency, our model preserves the strengths of deep semantic extraction while leveraging topological signals, maintaining an irreplaceable advantage in detecting sophisticated, structured social financial manipulations.

#### *4.3. Comparison Across Different Granularity Levels*

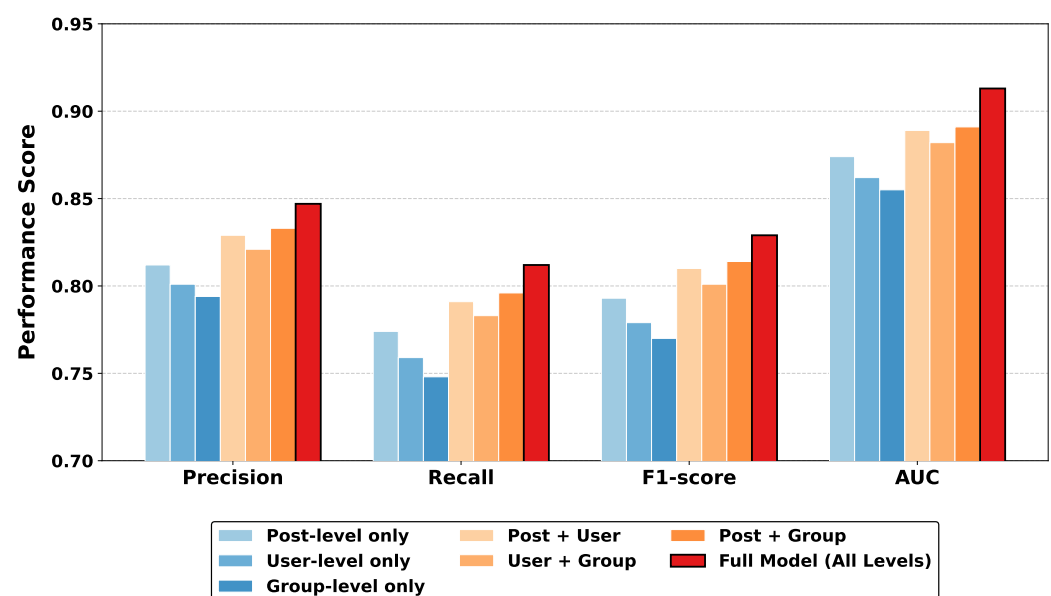
The purpose of this experiment is to systematically analyze the contribution of different granularity levels in social financial risk detection and to validate the advantage of multi-granularity collaborative modeling over single-granularity approaches. By separating and combining post-level, user-level, and group-level information, the functional role of each information source in risk prediction can be observed.

As shown in Table 3 and Figure 5, among single-granularity models, the post-level model achieves the best performance, indicating that textual semantics remain the primary basis for risk identification, as they directly reflect misleading expressions and investment inducement patterns. The user-level model performs slightly worse, suggesting that behavioral sequences capture abnormal activity patterns but lack explicit semantic grounding. The group-level model achieves the lowest performance, indicating that structural propagation alone is insufficient for accurate risk identification and requires complementary content and behavioral information. For dual-granularity combinations, all configurations outperform single-granularity models, with the combination of post and group slightly outperforming others, indicating strong complementarity between semantic information and propagation structure, while the user-group combination reflects the alignment between behavioral and structural modeling. The full model achieves the best performance across all metrics, demonstrating significant synergy among the three granularities. From a theoretical perspective, different granularity models correspond to distinct representation spaces. The post-level model relies on high-dimensional semantic representations with strong discriminative power but lacks temporal and structural constraints, making it sensitive to local noise. The user-level model encodes behavioral sequences to capture temporal dependencies, reducing fluctuations caused by single-instance misclassification, but behavioral features exhibit higher uncertainty and weaker decision boundaries. The group-level model aggregates node features through graph propagation, which essentially performs local neighborhood averaging, enhancing the representation of coordinated patterns but potentially introducing noise. When multi-granularity information is fused, constraints are formed across different representation spaces, allowing semantics, behavior, and structure to mutually calibrate within a shared space. This reduces variance under individual perspectives and improves overall stability. Through joint optimization, the full

model enforces consistency across granularities, resulting in a clearer risk distribution in the representation space and superior performance across evaluation metrics.

**Table 3.** Performance comparison across different granularity levels. Results are presented as mean  $\pm$  standard deviation.

| Level                   | Precision         | Recall            | F1-Score          | AUC               |
|-------------------------|-------------------|-------------------|-------------------|-------------------|
| Post-level only         | 0.812 $\pm$ 0.005 | 0.774 $\pm$ 0.006 | 0.793 $\pm$ 0.005 | 0.874 $\pm$ 0.006 |
| User-level only         | 0.801 $\pm$ 0.006 | 0.759 $\pm$ 0.007 | 0.779 $\pm$ 0.006 | 0.862 $\pm$ 0.007 |
| Group-level only        | 0.794 $\pm$ 0.007 | 0.748 $\pm$ 0.008 | 0.770 $\pm$ 0.007 | 0.855 $\pm$ 0.008 |
| Post + User             | 0.829 $\pm$ 0.004 | 0.791 $\pm$ 0.005 | 0.810 $\pm$ 0.004 | 0.889 $\pm$ 0.005 |
| User + Group            | 0.821 $\pm$ 0.005 | 0.783 $\pm$ 0.005 | 0.801 $\pm$ 0.005 | 0.882 $\pm$ 0.006 |
| Post + Group            | 0.833 $\pm$ 0.004 | 0.796 $\pm$ 0.004 | 0.814 $\pm$ 0.004 | 0.891 $\pm$ 0.004 |
| Full Model (All Levels) | 0.847 $\pm$ 0.003 | 0.812 $\pm$ 0.004 | 0.829 $\pm$ 0.003 | 0.913 $\pm$ 0.003 |



**Figure 5.** Performance comparison of different granularity modeling approaches in terms of Precision, Recall, F1-score, and AUC.

#### 4.4. Ablation Study

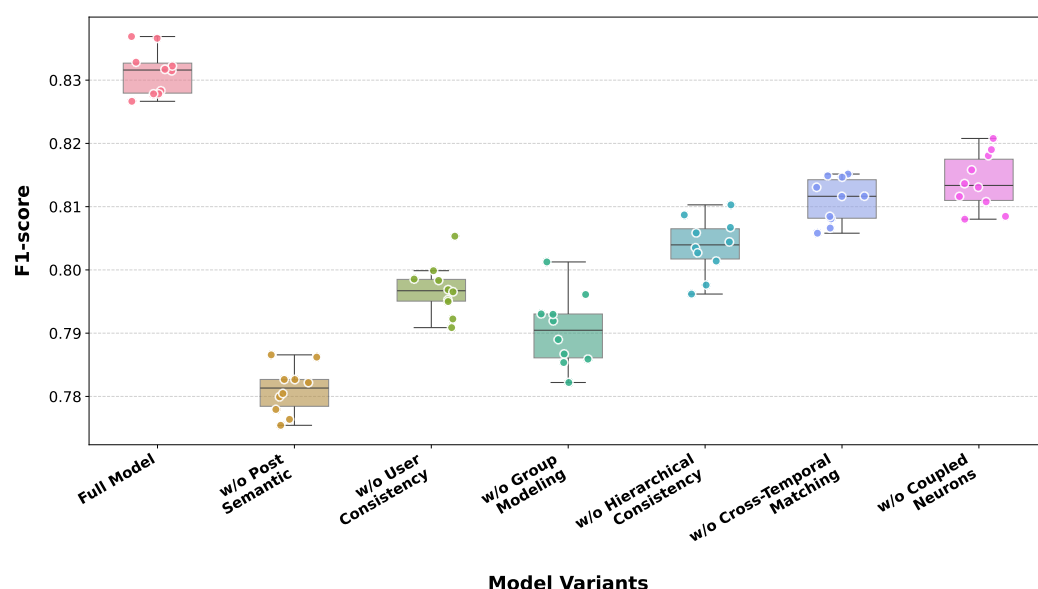
The purpose of this experiment is to evaluate the contribution of each key module in the proposed multi-granularity joint detection framework by progressively removing specific components and analyzing the resulting performance variations.

As shown in Table 4 and Figure 6, the full model achieves the best performance across all metrics, indicating strong synergy among all modules. Rather than simply observing performance degradation, analyzing the specific prediction errors made by each ablated variant reveals the critical necessity of these seemingly complex components. When the post semantic modeling module is removed, performance drops most significantly. Without this module, the model completely fails to identify explicit misleading expressions, leading to a massive surge in false negatives for blatantly fraudulent posts. Removing the user consistency module results in a notable decrease in recall and F1. In its absence, the model ignores the temporal regularity of user actions and fails to detect accounts that use perfectly compliant text but operate with machine-like, abnormal posting frequencies, thus allowing disguised bots to evade detection. The removal of the group modeling module also leads to performance degradation, particularly in AUC. Without structural information, the

model is unable to perceive the artificial consensus formed by multiple accounts mutually endorsing each other, thereby missing sophisticated coordinated manipulation campaigns.

**Table 4.** Ablation study of different modules in the proposed framework. Results are presented as mean  $\pm$  standard deviation.

| Variant                      | Precision         | Recall            | F1-Score          | Accuracy          | AUC               |
|------------------------------|-------------------|-------------------|-------------------|-------------------|-------------------|
| Full Model                   | $0.847 \pm 0.003$ | $0.812 \pm 0.004$ | $0.829 \pm 0.003$ | $0.856 \pm 0.003$ | $0.913 \pm 0.003$ |
| w/o Post Semantic Module     | $0.803 \pm 0.006$ | $0.769 \pm 0.007$ | $0.785 \pm 0.006$ | $0.812 \pm 0.005$ | $0.872 \pm 0.006$ |
| w/o User Consistency Module  | $0.816 \pm 0.005$ | $0.781 \pm 0.006$ | $0.798 \pm 0.005$ | $0.825 \pm 0.005$ | $0.884 \pm 0.005$ |
| w/o Group Modeling Module    | $0.809 \pm 0.005$ | $0.776 \pm 0.006$ | $0.792 \pm 0.006$ | $0.818 \pm 0.006$ | $0.878 \pm 0.006$ |
| w/o Hierarchical Consistency | $0.822 \pm 0.004$ | $0.789 \pm 0.005$ | $0.805 \pm 0.004$ | $0.832 \pm 0.004$ | $0.889 \pm 0.005$ |
| w/o Cross-Temporal Matching  | $0.828 \pm 0.004$ | $0.793 \pm 0.005$ | $0.810 \pm 0.004$ | $0.835 \pm 0.004$ | $0.895 \pm 0.004$ |
| w/o Coupled Neurons          | $0.831 \pm 0.004$ | $0.798 \pm 0.005$ | $0.814 \pm 0.004$ | $0.838 \pm 0.004$ | $0.898 \pm 0.004$ |



**Figure 6.** F1-score distribution of different model variants in the ablation study, illustrating the impact of each module on performance.

Furthermore, when hierarchical consistency is removed, a systematic decline is observed. Without cross-level constraints, the model frequently generates contradictory predictions, such as flagging a user as high risk while scoring their densely connected group as perfectly safe, which severely destabilizes the decision boundaries. The removal of fine-grained components like cross-temporal matching and coupled neurons further demystifies their design motivations. As observed in the error logs, without cross-temporal matching, the model loses the ability to align semantic shifts with behavioral bursts in time. Consequently, it erroneously flags a large number of normal, highly active users who simply participate in trending discussions as coordinated manipulation accounts, causing a drastic spike in false positives. Similarly, the coupled neuron structure is essential for capturing the co-occurrence of latent risk semantics, such as the simultaneous presence of return guarantees and artificial urgency. When removed, the model struggles to differentiate between legitimate urgent financial news and fraudulent manipulative urgency, which directly degrades precision. Collectively, these components are not over-engineered additions but targeted, necessary solutions designed to counter specific evasion tactics employed in complex social financial fraud.

#### 4.5. Computational Complexity and Efficiency Analysis

The objective of this experiment is to evaluate the computational overhead and operational efficiency of the proposed multi-granularity joint sensing framework, ensuring its feasibility for real-world deployment on social media platforms. The experimental evaluation was executed on the identical high-performance computing platform described in the hardware configuration section, utilizing an adaptive gradient-based optimizer with automatic mixed precision enabled. Efficiency metrics are quantified using four primary indicators: the number of trainable parameters measured in millions, the computational cost measured in Giga Floating-Point Operations, the average training time per epoch measured in seconds, and the single-sample inference latency measured in milliseconds. Comparative measurements were recorded across representative baseline models, capturing the computational trade-offs inherent in single-view versus multi-granularity architectures.

As shown in Table 5, the empirical results demonstrate a clear and predictable trade-off between model performance and computational resource consumption. Traditional models like Logistic Regression and CNN exhibit negligible parameter counts, extremely low FLOPs, and near-instantaneous inference due to their simple architectural designs, yet they suffer from poor detection accuracy. Advanced semantic models like FinBERT demand significantly higher computational resources, possessing 110 million parameters and requiring 22.4 Giga FLOPs per forward pass. Our proposed multi-granularity joint method reports the highest resource consumption, with 142.5 million parameters, 31.2 Giga FLOPs, a training time per epoch of 442.6 s, and an inference latency of 13.15 milliseconds per sample. This is a direct consequence of integrating three heterogeneous modules alongside the early-stage cross-hierarchical consistency constraints.

**Table 5.** Computational complexity and runtime comparison of different models. Latency and training times are reported with standard deviations to reflect hardware operational variance.

| Model               | Parameters (M) | FLOPs (G) | Training Time per Epoch (s) | Inference Latency (ms) |
|---------------------|----------------|-----------|-----------------------------|------------------------|
| Logistic Regression | 0.05           | 0.01      | 14.2 ± 0.5                  | 0.12 ± 0.01            |
| CNN                 | 4.2            | 0.8       | 48.5 ± 1.2                  | 1.45 ± 0.08            |
| Transformer         | 45.8           | 8.5       | 195.3 ± 3.5                 | 5.82 ± 0.15            |
| FinBERT             | 110.0          | 22.4      | 324.1 ± 5.2                 | 8.94 ± 0.21            |
| GAT                 | 8.5            | 1.6       | 132.7 ± 2.8                 | 4.31 ± 0.12            |
| Proposed Method     | 142.5          | 31.2      | 442.6 ± 6.4                 | 13.15 ± 0.32           |

From a theoretical complexity perspective, the overall time complexity of our framework is bounded by its component operations, where the graph convolutional layers scale linearly with the network topology as  $O(|V| + |E|)$ , and the behavioral sequence Transformer scales quadratically with the sequence window size as  $O(T_b^2)$ . Although the multi-module parallel execution and iterative optimization objective increase the computational footprint and parameter scale during the training phase, an inference latency of approximately 13 milliseconds per sample is entirely acceptable for large-scale industrial platforms. In practical platform governance scenarios, risk detection does not strictly require instantaneous real-time blocking at the millisecond scale; instead, a near-line or asynchronous offline audit pipeline is standard. Therefore, the significant performance gains achieved in precision and recall by our framework far outweigh the manageable increase in runtime overhead, proving that the proposed method remains highly viable for modern regulatory deployment.

#### 4.6. Robustness Evaluation Under Noisy and Adversarial Conditions

The primary objective of this experiment is to rigorously validate the claim of enhanced robustness by evaluating the stability of the proposed multi-granularity sensing framework under noisy, adversarial, and imperfect data conditions. In real-world social platforms, malicious actors actively employ evasion tactics such as intentionally misspelling financial keywords to bypass semantic filters, or artificially injecting fake followers and interactions to obfuscate structural detection. To simulate these adversarial scenarios, two specific perturbation strategies are introduced into the test set: text noise injection and structural edge dropping. For text noise, a proportion of financial keywords in the posts is randomly replaced with homophones or masked with meaningless characters. For structural noise, a specified ratio of edges in the user-post interaction graph is randomly removed to destroy the topological completeness. The perturbation ratios for both noise types are progressively increased from zero to thirty percent. The performance of the proposed framework is compared against the strongest text-based baseline, LLaMA-3, and the strongest graph-based baseline, GAT, using the F1-score as the primary evaluation metric.

As shown in Table 6, the empirical results unequivocally demonstrate the superior robustness of the proposed multi-granularity framework under severe adversarial conditions. When subjected to text noise injection, the pure semantic model LLaMA-3 experiences a catastrophic degradation, with its F1-score plummeting from 0.808 to 0.615 at a thirty percent perturbation level. This vulnerability highlights the fundamental flaw of relying solely on content moderation. In contrast, the proposed method experiences a significantly milder decline, maintaining a highly competitive F1-score of 0.776. This resilience is directly attributed to the behavioral sequence and structural group modules, which continue to capture anomalous high-frequency posting and coordinated endorsement patterns even when explicit semantic signals are deliberately corrupted. Conversely, under structural edge dropping scenarios, the graph-based GAT model rapidly deteriorates as the missing links destroy its neighborhood aggregation mechanism. Meanwhile, LLaMA-3 remains unaffected by topological changes since it operates strictly on individual texts. The proposed method once again demonstrates robust stability, outperforming GAT by a massive margin at the thirty percent edge dropping level. By jointly modeling three heterogeneous dimensions, the hierarchical consistency constraint essentially functions as a fail-safe mechanism. When one sensing modality is actively compromised by adversarial tactics, the framework dynamically shifts its reliance onto the remaining uncorrupted modalities, thereby providing a highly robust defense against complex evasion strategies in practical social media environments.

**Table 6.** Robustness evaluation under varying levels of text and structural noise. Results are presented as mean  $\pm$  standard deviation of F1-scores.

| Noise Type & Ratio                                 | 0% (Clean)        | 10% Perturbation  | 20% Perturbation  | 30% Perturbation  |
|----------------------------------------------------|-------------------|-------------------|-------------------|-------------------|
| Text Noise Injection (Semantic Adversarial Attack) |                   |                   |                   |                   |
| LLaMA-3 (Few-shot)                                 | 0.808 $\pm$ 0.004 | 0.752 $\pm$ 0.006 | 0.691 $\pm$ 0.008 | 0.615 $\pm$ 0.012 |
| GAT                                                | 0.796 $\pm$ 0.005 | 0.781 $\pm$ 0.006 | 0.763 $\pm$ 0.007 | 0.742 $\pm$ 0.008 |
| Proposed Method                                    | 0.829 $\pm$ 0.003 | 0.814 $\pm$ 0.004 | 0.798 $\pm$ 0.005 | 0.776 $\pm$ 0.006 |
| Structural Edge Dropping (Topological Evasion)     |                   |                   |                   |                   |
| LLaMA-3 (Few-shot)                                 | 0.808 $\pm$ 0.004 | 0.808 $\pm$ 0.004 | 0.808 $\pm$ 0.004 | 0.808 $\pm$ 0.004 |
| GAT                                                | 0.796 $\pm$ 0.005 | 0.745 $\pm$ 0.007 | 0.688 $\pm$ 0.009 | 0.602 $\pm$ 0.011 |
| Proposed Method                                    | 0.829 $\pm$ 0.003 | 0.819 $\pm$ 0.003 | 0.805 $\pm$ 0.005 | 0.783 $\pm$ 0.007 |

#### 4.7. Generalizability Verification via Cross-Platform Evaluation

The primary objective of this experiment is to rigorously verify the cross-domain generalizability of the proposed multi-granularity sensing framework across diverse social media platforms. Since user interaction behaviors, linguistic styles, and network topologies vary significantly across different online communities such as Xiaohongshu, Weibo, Xueqiu, and Reddit, evaluating the out-of-distribution predictive capability of the model is crucial for real-world deployment. To achieve this, a strict leave-one-platform-out cross-validation setup is implemented. In each evaluation iteration, the data from three platforms are aggregated to form the training set, while the data from the remaining fourth platform are exclusively strictly reserved as the unseen test set. This zero-shot cross-platform evaluation ensures that the model cannot memorize platform-specific artifacts. The proposed framework is compared against the strongest text-based baseline, LLaMA-3, and the strongest graph-based baseline, GAT. The F1-score and Area Under the Receiver Operating Characteristic Curve are utilized as the primary evaluation metrics, with results reported as the mean and standard deviation over multiple runs.

As shown in Table 7, the experimental results clearly validate the strong generalization capability of the proposed multi-granularity joint sensing framework across heterogeneous social platforms. When facing severe domain shifts, single-modality baselines experience noticeable performance degradation. For instance, the large language model LLaMA-3 exhibits a significant drop in F1-score when evaluated on Reddit, as the linguistic slang, cultural context, and financial terminologies on this English-centric platform differ vastly from the Chinese-centric platforms used during training. Similarly, the graph-based GAT model struggles when transferred to Xiaohongshu, where the user interaction density and follower network topologies are fundamentally different from those of the highly interconnected Weibo environment. In stark contrast, the proposed method maintains robust and superior performance across all unseen target platforms. By dynamically aligning semantic, behavioral, and structural signals through early-stage hierarchical consistency constraints, the framework avoids overfitting to the specific textual formats or topological structures of any single platform. Instead, it successfully captures the fundamental and invariant patterns of coordinated financial manipulation, such as the inherent contradiction between compliant text and anomalous synchronized posting behaviors. This cross-modal compensation mechanism effectively mitigates domain-specific biases, thereby proving that the proposed framework is highly generalizable and practically viable for broad social platform governance.

**Table 7.** Cross-platform generalization performance. The model is trained on three platforms and tested on the unseen target platform. Results are presented as mean  $\pm$  standard deviation.

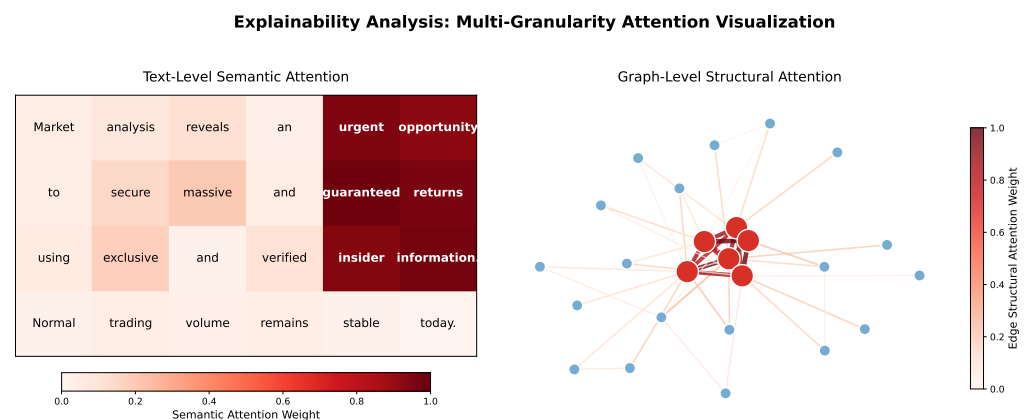
| Target Test Platform | Metric   | LLaMA-3 (Few-Shot) | GAT               | Proposed Method   |
|----------------------|----------|--------------------|-------------------|-------------------|
| Xiaohongshu          | F1-score | $0.785 \pm 0.005$  | $0.752 \pm 0.006$ | $0.811 \pm 0.004$ |
|                      | AUC      | $0.868 \pm 0.004$  | $0.841 \pm 0.005$ | $0.895 \pm 0.003$ |
| Weibo                | F1-score | $0.792 \pm 0.004$  | $0.768 \pm 0.005$ | $0.816 \pm 0.003$ |
|                      | AUC      | $0.875 \pm 0.005$  | $0.852 \pm 0.006$ | $0.902 \pm 0.004$ |
| Xueqiu               | F1-score | $0.801 \pm 0.005$  | $0.771 \pm 0.004$ | $0.822 \pm 0.004$ |
|                      | AUC      | $0.884 \pm 0.004$  | $0.859 \pm 0.005$ | $0.908 \pm 0.003$ |
| Reddit               | F1-score | $0.764 \pm 0.006$  | $0.743 \pm 0.007$ | $0.795 \pm 0.005$ |
|                      | AUC      | $0.842 \pm 0.006$  | $0.835 \pm 0.006$ | $0.881 \pm 0.005$ |

#### 4.8. Explainability and Case-Based Visualization

The objective of this section is to provide qualitative evidence supporting the explainability claims of the proposed multi-granularity sensing framework. To concretely demonstrate how the model arrives at its risk predictions, a case-based interpretation utilizing attention weight visualization is presented. A typical coordinated financial manipulation event from the test set was selected, wherein a cluster of accounts synchronously promoted a highly speculative asset while disguising their intent with neutral market analysis terminology. We extracted the internal attention weights from both the post-level semantic encoder and the group-level graph neural network to generate interpretation heatmaps.

As illustrated in the left panel of Figure 7, the semantic attention heatmap effectively visualizes the focus of the text encoding module. Instead of assigning uniform weights across the entire post, the model selectively highly activates on specific sensitive phrases such as guaranteed returns, urgent opportunity, and insider information. Normal financial terms and common stop words receive near-zero attention scores. This precise token-level attribution confirms that the semantic modeling components successfully isolate latent risk semantics from background noise.

Furthermore, the right panel of Figure 7 visualizes the edge attention weights generated by the graph structural modeling module. In this social interaction sub-graph, the model assigns the highest propagation weights to a tightly knit cluster of accounts that exhibit abnormal synchronized reposting behavior within a very short time window. Peripheral nodes representing normal organic users who merely viewed or organically commented on the thread are assigned significantly lower weights. By mapping these high-weight graph components back to the real-world user accounts, platform administrators can immediately identify the core orchestrators of the manipulation campaign. This dual-level visualization provides clear, intuitive, and actionable evidence that the decisions made by the multi-granularity framework are not black-box outputs, but are fully explainable and firmly grounded in both semantic inducement and structural coordination.



**Figure 7.** Attention heatmaps for explainability analysis. The left panel displays the text-level semantic attention highlighting sensitive inducement phrases. The right panel shows the graph-level structural attention indicating the dense coordination among core manipulation nodes.

#### 4.9. Discussion

In real-world social financial environments, information dissemination is often highly complex and covert. For example, in stock investment, cryptocurrency trading, and financial product promotion scenarios, a large amount of content appears in the form of experience sharing or investment advice. The language used is often professional and neutral, yet it may implicitly contain exaggerated return promises or risk concealment. On platforms such as Xiaohongshu, credibility is often constructed through real profit screen-

shots or trading records, while on Weibo and Xueqiu, sentiment guidance is frequently driven by trending events. In Reddit communities, coordinated discussions and collective sentiment can amplify market expectations. These scenarios indicate that relying solely on textual content is insufficient for accurate risk identification, as risks are often embedded in long-term behavioral patterns and group-level propagation structures. For instance, an account may continuously recommend different assets within a short period, where individual posts appear normal in semantics but reveal highly consistent promotional behavior in temporal patterns. Similarly, multiple accounts may simultaneously promote a specific asset and reinforce each other through interactions, forming an artificial consensus, which is a typical signal of coordinated manipulation.

In such practical scenarios, multi-granularity joint modeling aligns more closely with real risk formation mechanisms and enhances system applicability. Semantic modeling enables the identification of explicit risk expressions, behavioral modeling uncovers hidden strategic patterns such as periodic posting or cross-topic promotion, and structural modeling captures coordinated relationships among multiple accounts, such as dense interaction networks or synchronized dissemination paths. This joint analysis across content, individual, and group levels enables a more comprehensive understanding of risk. To explicitly address how our proposed framework advances beyond recent multimodal methods in the literature, a direct comparison regarding representation fusion is necessary. Many contemporary multimodal risk detection models rely on simple multi-task learning frameworks or flat feature concatenation. In these conventional approaches, heterogeneous features from different modalities are extracted independently and merged only at the final classification layer via late fusion, often leading to suboptimal alignment when dealing with highly disguised social manipulation. In contrast, our hierarchical consistency constraint fundamentally differs by embedding cross-dimensional mutual calibration directly into the early stages of representation learning. Instead of merely combining final outputs, this mechanism enforces semantic, behavioral, and structural signals to continuously constrain and dynamically refine one another throughout the learning pipeline. Consequently, this early-stage alignment significantly reduces the representation variance and yields a more robust joint embedding space, directly explaining the substantial performance gains our model achieves over traditional concatenation-based multimodal methods. In real regulatory or platform governance scenarios, when a particular investment topic suddenly gains widespread attention, high-risk posts can be identified, and the associated high-risk users and groups can be further located, providing more precise support for risk warning, content moderation, and user management. Additionally, this approach is valuable in investor protection scenarios, where it can help identify misleading information propagation chains and reduce the risk of users being influenced by false information, thereby contributing to a healthier digital financial ecosystem.

#### *4.10. Limitation and Future Work*

Although the proposed multi-granularity social context joint detection method demonstrates strong performance, several technical limitations and ethical considerations must be acknowledged. From a technical perspective, the model heavily relies on high-quality multi-source data, including textual content, behavioral sequences, and social interaction structures. In real-world platforms, some behavioral data or interaction relationships may be incomplete or inaccessible, which may affect model robustness and generalization. Furthermore, the current approach is based on static temporal windows, and although temporal consistency mechanisms are introduced, fine-grained modeling of long-term dynamic propagation processes remains insufficient. In addition, the model architecture is

relatively complex, and multi-module joint training introduces considerable computational overhead, which may pose challenges for large-scale online deployment.

From an ethical and fairness perspective, potential data biases must be recognized. Since the training dataset is aggregated from multiple platforms such as Xiaohongshu, Weibo, Xueqiu, and Reddit, there are significant disparities in user demographics, financial discussion styles, and linguistic habits. The model may overfit to platform-specific slang or the unique communication patterns of certain demographic groups, thereby introducing algorithmic bias. Moreover, in real-world platform governance, the consequences of false positives are severe, as they can lead to the wrongful suspension or restriction of legitimate users. Therefore, the risk probabilities output by the model should strictly serve as an auxiliary reference for a human-in-the-loop review process, rather than acting as the sole basis for automated punitive actions. Additionally, continuous behavioral tracking and social graph modeling inherently involve significant user privacy risks. The deployment of such systems necessitates strict compliance with data anonymization protocols and privacy protection regulations to prevent the infringement of user rights.

Future work can be extended in several directions. More advanced dynamic temporal modeling mechanisms can be introduced to better capture long-term behavioral evolution and sudden propagation events. Lightweight model design and efficient inference strategies can be explored to reduce computational costs and improve deployment feasibility. External knowledge sources, such as financial knowledge graphs or regulatory rule bases, can be integrated to enhance the understanding of professional semantics and compliance boundaries. Moreover, cross-platform data fusion can be investigated to construct more complete propagation chains, thereby improving the detection of complex coordinated manipulation behaviors while rigorously maintaining user privacy and algorithmic fairness.

## 5. Conclusions

This study focuses on the critical problem of detecting misinformation propagation and coordinated manipulation behaviors in social financial scenarios. In response to data-related challenges faced by current machine learning methods, including high data noise, implicit semantic expressions, complex behavioral patterns, and the difficulty of unified modeling for multi-source data, a multi-granularity social context-driven joint detection method is proposed. A unified representation space is constructed from three levels, namely post semantics, user behavior, and group structure, and cross-granularity collaborative modeling is achieved through hierarchical consistency constraints, thereby effectively alleviating information bias and instability under a single data perspective. In terms of method design, semantic modeling, temporal behavioral analysis, and graph-based propagation mechanisms are integrated, enhancing the robustness and interpretability of the model in complex social financial data environments, which aligns with the themes of data quality, structural modeling, and model reliability emphasized in this special issue.

Experimental results demonstrate that the proposed method significantly outperforms traditional models and mainstream deep learning approaches across multiple evaluation metrics. In the overall performance comparison, Precision, Recall, and F1-score reach 0.847, 0.812, and 0.829, respectively, while Accuracy reaches 0.856 and AUC reaches 0.913, indicating superior comprehensive discriminative capability. Furthermore, multi-granularity comparison experiments verify the complementarity among semantic, behavioral, and structural information, where the full model achieves consistent improvements over single-granularity models across all metrics. The ablation study further confirms the effectiveness of each module in enhancing model stability and representation capability. Crucially, the newly incorporated adversarial robustness evaluations and cross-platform validation experiments formally substantiate the claims of enhanced model stability under severe text noise,

structural edge dropping, and out-of-distribution domain shifts, while qualitative case analyses validate the capability of the framework in providing explainable and transparent detection clues. Overall, a unified modeling framework for complex multi-source data is provided, and the advantages of multi-granularity collaborative learning in improving model performance, robustness, and interpretability are validated, offering a valuable technical approach for addressing data heterogeneity and complexity in real-world scenarios.

**Author Contributions:** Conceptualization, S.C., R.F., Y.Z. and J.Y.; Data curation, Y.L. and L.C.; Formal analysis, J.X.; Funding acquisition, J.Y.; Investigation, J.X.; Methodology, S.C., R.F. and Y.Z.; Project administration, J.Y.; Resources, Y.L. and L.C.; Software, S.C., R.F. and Y.Z.; Supervision, J.Y.; Validation, J.X.; Visualization, Y.L. and L.C.; Writing—original draft, S.C., R.F., Y.Z., Y.L., L.C., J.X. and J.Y., S.C., R.F., Y.Z. contributed equally to this work. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by National Natural Science Foundation of China grant number 61202479.

**Data Availability Statement:** The dataset and related code utilized in this study will be made publicly accessed on 1 July 2026 <https://github.com/Aurelius-04/MSS.git>.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Liu, X.; Rui, H.; Cheng, D.; Han, L.; Zhou, Z.; Zhao, G. Multi-Granularity Augmented Graph Learning for Spoofing Transaction Detection. In Proceedings of the ACM on Web Conference 2025, Sydney, Australia, 28 April–2 May 2025; pp. 5151–5160.
2. Ma, J.; Li, F.; Zhang, R.; Xu, Z.; Cheng, D.; Ouyang, Y.; Zhao, R.; Zheng, J.; Zheng, Y.; Jiang, C. Fighting against Organized Fraudsters Using Risk Diffusion-based Parallel Graph Neural Network. In Proceedings of the 32nd International Joint Conference on Artificial Intelligence (IJCAI 2023), Macao, China, 19–25 August 2023; pp. 6138–6146.
3. Trinh, T.K.; Wang, Z. Dynamic graph neural networks for multi-level financial fraud detection: A temporal-structural approach. *Ann. Appl. Sci.* **2024**, *5*, 3133–3147.
4. Xiong, M.; Chen, H.; Wen, Y.; Saudagar, A.K.J.; Del Ser, J.; Muhammad, K. Person Re-identification with Spatial Multi-granularity Feature Exploration for Social Risk Situational Assessment. *Cogn. Comput.* **2024**, *16*, 2701–2712.
5. Dai, X.; Sun, B.; Bai, J.; Ye, J.; Chu, X. Double-quantitative multi-granularity kernel fuzzy rough sets model and its application in rheumatoid arthritis risk assessment. *Int. J. Mach. Learn. Cybern.* **2025**, *16*, 1–23.
6. Tu, M.; Zhu, K.; Guo, H.; Miao, Q.; Zhao, C.; Zhu, G.; Qiao, H.; Huang, G.; Tang, M.; Wang, J. Multi-granularity mutual learning network for object re-identification. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 15178–15189. [[CrossRef](#)]
7. Greco, C.M.; Simeri, A.; Tagarelli, A.; Zumpano, E. Transformer-based language models for mental health issues: A survey. *Pattern Recognit. Lett.* **2023**, *167*, 204–211. [[CrossRef](#)]
8. Du, C.; Chiang, C.E.; Huang, T.; Cui, Z. Adaptive Graph Convolution and Semantic-Guided Attention for Multimodal Risk Detection in Social Networks. In *Proceedings of the 2025 5th International Conference on Artificial Intelligence, Automation and High Performance Computing (AIAHPC)*; IEEE: New York, NY, USA, 2025; pp. 507–512.
9. Liu, C.; Gao, Y.; Sun, L.; Feng, J.; Yang, H.; Ao, X. User behavior pre-training for online fraud detection. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, 14–18 August 2022; pp. 3357–3365.
10. Yan, H.; Pang, X.; Zhou, S.; Fan, H. Transformer-Based Intrusion Detection for Post-5G and 6G Telecommunication Networks Using Dynamic Semantic Embedding. *Future Internet* **2025**, *17*, 544.
11. Peng, H.; Ma, R.; Chao, J.; Li, X.; Zhang, F. Graph embedding and clustering collaborative optimization model for fraudster group detection. *Inf. Process. Manag.* **2026**, *63*, 104462.
12. Qiao, Y.; Ao, X.; Chen, J.; Liu, Y.; Zhong, Q.; He, Q. Deja Vu in Plots: Leveraging Cross-Session Evidence with Retrieval-Augmented LLMs for Live Streaming Risk Assessment. *arXiv* **2026**, arXiv:2601.16027.
13. Yu, S.; Wang, Y.; Li, R.; Liu, G.; Shen, Y.; Ji, S.; Li, B.; Han, F.; Zhang, X.; Xia, F. Graph2text or graph2token: A perspective of large language models for graph learning. *ACM Trans. Inf. Syst.* **2025**, *44*, 57.
14. Wang, H.; Zheng, J.; Carvajal-Roca, I.E.; Chen, L.; Bai, M. Financial fraud detection based on deep learning: Towards large-scale pre-training transformer models. In *Proceedings of the China Conference on Knowledge Graph and Semantic Computing*; Springer: Singapore, 2023; pp. 163–177.

15. Wu, Y.; Xiang, C. Multimodal Financial Anomaly Detection in Enterprises Using VAE–Transformer–GNN Hybrid Ensemble Models. In Proceedings of the 2nd International Symposium on Integrated Circuit Design and Integrated Systems, Singapore, 26–28 September 2025; pp. 197–203.
16. Chai, Z.; Wang, Z.; Li, W.; Yang, Y.; Wu, D.; Wu, F. MGA-KGF: A Multi-Granularity Adaptive Knowledge Graph Fusion Framework for Heterogeneous Multi-Source Smart Grid Data. In *Proceedings of the 2025 IEEE 12th International Conference on Cyber Security and Cloud Computing (CSCloud)*; IEEE: New York, NY, USA, 2025; pp. 214–220.
17. Zhao, Q.; Huang, J.; Liu, G.; Miao, Y.; Wang, P. A multiinterest and social interest-field framework for financial security. *IEEE Trans. Comput. Soc. Syst.* **2023**, *11*, 1685–1695. [[CrossRef](#)]
18. Zhang, Z.; Zhu, J.; Guo, Z.; Zhang, Y.; Li, Z.; Hu, B. Natural language processing for depression prediction on sina weibo: Method study and analysis. *JMIR Ment. Health* **2024**, *11*, e58259. [[CrossRef](#)] [[PubMed](#)]
19. Li, X.; Bianculli, D.; Briand, L. Tracing content requirements in financial documents using multi-granularity text analysis. *Requir. Eng.* **2025**, *30*, 109–132. [[CrossRef](#)]
20. Zhang, W.; Peng, L. Semantic analysis and image processing-jointly driven multimodal deep learning framework for smart warning of enterprise financial risk. *J. Circuits Syst. Comput.* **2024**, *33*, 2450304. [[CrossRef](#)]
21. Huang, A.H.; Wang, H.; Yang, Y. FinBERT: A large language model for extracting information from financial text. *Contemp. Account. Res.* **2023**, *40*, 806–841. [[CrossRef](#)]
22. Lin, J.; Wei, Z. CommunityBench: Benchmarking Community-Level Alignment across Diverse Groups and Tasks. *arXiv* **2026**, arXiv:2601.13669.
23. Hettiarachchi, I. Unified Temporal Tokenization: A Hybrid Semantic and Numeric Mapping for Time-Aware Large Language Models. *Front. Comput. Sci. Artif. Intell.* **2025**, *4*, 25–42. [[CrossRef](#)]
24. Zhang, L.; Zhu, Q.; Ray, H.; Xie, Y. Improving Network Threat Detection by Knowledge Graph, Large Language Model, and Imbalanced Learning. *arXiv* **2025**, arXiv:2501.16393.
25. Artioli, P.; Maci, A.; Magri, A. A comprehensive investigation of clustering algorithms for User and Entity Behavior Analytics. *Front. Big Data* **2024**, *7*, 1375818. [[CrossRef](#)] [[PubMed](#)]
26. Zavitsanos, E.; Spyropoulou, E.; Giannakopoulos, G.; Paliouras, G. Machine learning for identifying risk in financial statements: A survey. *ACM Comput. Surv.* **2025**, *57*, 1–37. [[CrossRef](#)]
27. Wang, Z.; Wu, Q.; Zheng, B.; Wang, J.; Huang, K.; Shi, Y. Sequence as genes: An user Behavior modeling framework for fraud transaction detection in E-commerce. In Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Long Beach, CA, USA, 6–10 August 2023; pp. 5194–5203.
28. Vallarino, D. AI-Powered Fraud Detection in Financial Services: GNN, Compliance Challenges, and Risk Mitigation. *Preprint* **2025**, *7*, 2187–2201.
29. Immaneni, J. Strengthening Fraud Detection with Swarm Intelligence and Graph Analytics. *Int. J. Digit. Innov.* **2022**, *3*.
30. Chen, Y.; Liao, Z.; Yin, P.; Xie, T.; Yin, K.; Zhang, S. SafePred: A Predictive Guardrail for Computer-Using Agents via World Models. *arXiv* **2026**, arXiv:2602.01725.
31. Li, G.; Wang, S.; Feng, Y. Making differences work: Financial fraud detection based on multi-subject perceptions. *Emerg. Mark. Rev.* **2024**, *60*, 101134. [[CrossRef](#)]
32. Zhao, H.; Han, Z.; Yin, S.; Yang, N.; Hansen, P. From interface to inference: Mapping the impact of generative artificial intelligence affordances on user risk perception. *Telemat. Inform.* **2025**, *101*, 102299. [[CrossRef](#)]
33. Wang, D.; Wu, B.; Yuan, X.; Wu, L.; Zhou, Y.; Cui, H. Defiguard: A price manipulation detection service in defi using graph neural networks. *IEEE Trans. Serv. Comput.* **2024**, *17*, 3345–3358.
34. Shao, P.; Yang, Y.; Gao, C.; Chen, L.; Zhang, K.; Zhuang, C.; Wu, L.; Li, Y.; Wang, M. Exploring Heterogeneity and Uncertainty for Graph-based Cognitive Diagnosis Models in Intelligent Education. In Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Toronto, ON, Canada, 3–7 August 2025; Volume 1, pp. 1233–1243.
35. Gu, W.; Sun, M.; Liu, B.; Xu, K.; Sui, M. Adaptive Spatio-Temporal Aggregation for Temporal Dynamic Graph-Based Fraud Risk Detection. *J. Comput. Technol. Softw.* **2024**, *3*. [[CrossRef](#)]
36. Das, R.; Sirazy, M.; Khan, R.S.; Rahman, S. A collaborative intelligence (ci) framework for fraud detection in us federal relief programs. *Appl. Res. Artif. Intell. Cloud Comput.* **2023**, *6*, 47–59.
37. Hosmer, D.W., Jr.; Lemeshow, S.; Sturdivant, R.X. *Applied Logistic Regression*; John Wiley & Sons: Hoboken, NJ, USA, 2013.
38. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [[CrossRef](#)]
39. Alzubaidi, L.; Zhang, J.; Humaidi, A.J.; Al-Dujaili, A.; Duan, Y.; Al-Shamma, O.; Santamaría, J.; Fadhel, M.A.; Al-Amidie, M.; Farhan, L. Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *J. Big Data* **2021**, *8*, 53. [[CrossRef](#)] [[PubMed](#)]
40. Siami-Namini, S.; Tavakoli, N.; Namin, A.S. The performance of LSTM and BiLSTM in forecasting time series. In *Proceedings of the 2019 IEEE International Conference on Big Data (Big Data)*; IEEE: New York, NY, USA, 2019; pp. 3285–3292.

41. Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y.; et al. A survey on vision transformer. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 87–110. [[CrossRef](#)] [[PubMed](#)]
42. Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. The llama 3 herd of models. *arXiv* **2024**, arXiv:2407.21783.
43. Sharma, R.; Almäši, M.; Nehra, S.P.; Rao, V.S.; Panchal, P.; Paul, D.R.; Jain, I.P.; Sharma, A. Photocatalytic hydrogen production using graphitic carbon nitride (GCN): A precise review. *Renew. Sustain. Energy Rev.* **2022**, *168*, 112776. [[CrossRef](#)]
44. Hosseini, S.H.; Hoeft, F.; Kesler, S.R. GAT: A graph-theoretical analysis toolbox for analyzing between-group differences in large-scale structural and functional brain networks. *PLoS ONE* **2012**, *7*, e40709. [[PubMed](#)]

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