CSSA: A Cross-Modal Semantic-Structural Alignment Framework via LLMs and Graph Contrastive Learning for Fraud Detection of Online Payment

Zirui Zhao , Keyu Yuan * , Ziyue Wang , Jiaqing Shen , Yirui Huang

Posted Date: 6 February 2026

doi: 10.20944/preprints202602.0543.v1

Keywords: Large Language Models; Graph Neural Networks; cross-modal alignment; contrastive learning; representation learning

 doi:10.20944/preprints202602.0543.v1

Article

CSSA: A Cross-Modal Semantic-Structural Alignment Framework via LLMs and Graph Contrastive Learning for Fraud Detection of Online Payment

Zirui Zhao [1], Keyu Yuan [2], Ziyue Wang [3], Jiaqing Shen [4] and Yirui Huang [5]

[1] Suzhou University of Technology, Changshu, Jiangsu, P.R. China
[2] New York University, NY 11201, USA
[3] Independent Researcher, New York, NY, USA
[4] Rochester Institute of Technology, 1 Lomb Memorial Dr, Rochester, NY 14623, USA
[5] University of Southern California, Los Angeles, CA 90089, USA
* Correspondence: ky2074@nyu.edu

Abstract

Graph Neural Networks (GNNs) have demonstrated exceptional performance in modeling structural dependencies within networked data. However, in complex decision-making environments, structural information alone often fails to capture the latent semantic logic and domain-specific heuristics. While Large Language Models (LLMs) excel in semantic reasoning, their integration with graph-structured data remains loosely coupled in existing literature. This paper proposes CSSA, a novel Cross-modal Semantic-Structural Alignment framework that synergizes the zero-shot reasoning of LLMs with the topological aggregation of GNNs through a contrastive learning objective. Specifically, we treat node attributes as semantic prompts for LLMs to distill high-level "risk indicators," while a GNN branch encodes the local neighborhood topology. A cross-modal alignment layer is then introduced to minimize the representational gap between semantic intent and structural behavior. We evaluate CSSA on a massive dataset of 2.84 million online transaction records. Experimental results demonstrate that CSSA achieves a superior F1-score and AUC compared to state-of-the-art GNNs, particularly in scenarios characterized by extreme class imbalance and covert adversarial patterns.

Keywords: Large Language Models; Graph Neural Networks; cross-modal alignment; contrastive learning; representation learning

I. Introduction

One of the main challenges of modern learning models on relational data lies in reconciling local topological structures and global semantic reasoning, while GNN has emerged as a de facto solution to process unstructured data. Their main strategy is to combine nodes with close neighbors, a method which assumes that proximity implies functionality similarity[1].However, for some crucial applications it is insufficient to only consider the topology structure. One shortcoming of GNNs lies in that they cannot perform complicated analysis and processing operations over nodes' features.reducing complex concepts to simplistic numbers [2].

Meanwhile, recent breakthroughs on large scale LLMs have significantly expanded the horizon of NLP tasks. Large-scale LLMs are capable to perform general reasoning beyond their training data as well as domain-specific expertises,which allows them to understand human intent, which is not trivial for standard networks.Nevertheless, they do not possess any kind of internal structure

 doi:10.20944/preprints202602.0543.v1

perception since they treat the interdependent data like a sequence of elements without considering intricate network patterns and dynamic structural properties that exist in big networks [3].

This restriction is problematic for researchers, as they have to give up either knowledge about the structure of their network and hence lose semantic information, or they have to give up analysis and therefore loose syntactic information[4]. This has serious implications especially when it comes to critical applications such as online bankingwhere adversaries intentionally modify the graph connectivity in order to look benign and have contradictory behavioral properties.

In this work, we propose CSSA (Cross-modal Semantic-Structural Alignment) which overcomes the aforementioned limitation of learning multi-modal latent alignment. We argue that the representation of each node cannot be only determined by its graph structure and features butbut rather from their interplay:The technical foundation of CSSA is a two path architecture and one of the branches includes the Semantic Reasoner with a finetuned large language model through LORA training,while the second one includes a Structural Encoder based on an attention-based graph convolutional neural network.

The main contribution of our method lies in the proposed CCA module. Unlike existing "featurefusion" methods, which simply concatenate feature vectors together to generate a fused representation,contrastive learning, which is used in our CCA model for representation disagreement alignment.This new module aligns LM inference with the topology of a graph,thus facilitating the identification of "semantic-structural inconsistencies"—cases when the node has normal connections, but its logical behavior is inconsistent with that.or even vice versa. We experimentally evaluate our approach using an extensive e-commerce payment dataset comprising of 2.84 million transactions [5].By showcasing how novel aspects of this system enable us to demonstrate that our CSSA framework is a powerful tool for processing text-attributed graphs which need both network structure and semantics-aware inference capability.

II. Related Work

A. Graph Neural Networks in Finance

The research and progress in GNNs have been rapid since GCNs were proposed[6]. Subsequent work, like Graph Attenion Networks(GAT), used attention mechanisms to weight neighbor nodes by their relevance,while GraphSAGE enabled learning on dynamical large scale graph structure.In finance, such graph representations have been widely used for detecting money laundering or credit card fraud by identifying anomalous subgraphs and suspiciously dense communities [37]. One limitation to this class of methods is "feature dilution", where nodes' features with useful information will be lost due to the lack of discriminative power in consecutive aggregation operations.While some recent work has considered integrating different views within a GNN architecture, these approaches mostly remain restricted to the numerical space,overlooking the advantages of combination with explainability in NLP.

Recent state-of-the-art large language models (LLMs) e.g., GPT-4 or Llama can perform many complex tasks far beyond generation[7]. With a series of prompt engineering techniques such as chain-of-thought prompting and task-specific finetuning,these complex models show excellent results on analysis work such as law, medical care, finance etc...For the task of detecting fraud, such models have clear advantages: they are able to combine various factors (e.g., transaction time,geographic information, and vendor's integrity) into a general analysis[8,9].However, such approaches are computationally intensive and have limited scalability for high-order networks—i.e., they do not natively support high-non-linear graph structures as are common in online payments networks.Previous work has attempted to use LM's for converting a topology of networks into text,but is not feasible in practice for a system with many interacting components (our experiments show).

The main source of inspiration for this kind of investigation is a novel idea proposed by Zheng & Lin et.al(2026) that has been the first piece of work combining CNNs used to recognize local

 doi:10.20944/preprints202602.0543.v1

structure characteristics and also LSTMs exploited to extract global temporal dependence between blocks chains [10].They discovered that their hybrid method of combining spatial and temporal information could reduce the danger of finance data considerably in cases where large amounts of class imbalance exist,thus setting up the important experimental ground for our CSSA model. Inspired by their idea to combine time and space,we extend them to coordinate across modalities, linking their lightweight model and advanced linguistic processing together.

III. Methodology

The CSSA framework aims to learn a unified representation $Z_i$for each node i by aligning its semantic representation Z_i^{sem} and structural representation Z_i^{str}[11].

A. Semantic Logic Distillation (LLM Branch)

We model the semantics for each node (i.e., context) as an embedding that consists in a combination of features related to the node itself and past interactions containing such a node. We don't employ standard embeddings,Instead, we turn to an LLM for a Reasoning Vector [12] and with a LoRA-extended version of our model train the embedding module:

$$Z_i^{sem} = \mathrm{MLP}(\mathrm{LLM}(P_i)) \tag{1}$$

The prompt explicitly tells the LLM to determine whether or not the facts about a given node are "logically consistent" with what it knows about that topic[13].

B. Structural Feature Encoding (GNN Branch)

Meanwhile, another GNN is applied to learn on the graph G =(V,E) . To avoid being distracted by useless edges,we propose a dynamic message passing scheme which learns the importance of edges in relation to nodes' importance[14]:

$$h_i^{l+1} = \sigma\Big(\sum_{j\in N(i)\cup\{i\}} \partial_{ij} W^l h_j^l\Big) \tag{2}$$

where B is the mini-batch and τ is a temperature parameter that forces the network to find a joint embedding space in which semantically inconsistent examples are structurally inconsistent[15].

C. Joint Optimization

The output of the classifier is made through an ensemble decision:

$$y_i = Softmax(MLP(Z_i^{sem} \oplus) Z_i^{str}) \tag{3}$$

The final loss function is given by, L = , where is cross entropy loss.

IV. Data

A. Dataset Synthesis and Pre-Processing

In order to evaluate the ability of our CSSA architecture to capture more complicated outliers and abnormalities, we use large scale time series data obtained through one of the largest online marketplaces operating on a worldwide basis; this provides detailed information about the sales transactions made by almost 2.84 million unique customers during a fortnight period.We represent our dataset by a multi-layered dynamic network denoted by G = (N,E,T), where N (~= 2030 nodes) corresponds to agents,consisting of individual users as well as organizations. The transaction relation E represents transactions between these two parties with the following feature matrix for each transaction: Transaction Features : Times when transactions took place to detect fast frauds.Description: Sector code, address history. Quantitative Variables: Price, frequency of fraud

 doi:10.20944/preprints202602.0543.v1

label. Response variable is the fraud label which takes value Y ∈ {0,1}), with the label y = 1 meaning that the transaction was proven to be fraudulent.

The dataset is also contaminated by frauds and presents an imbalance problem since the fraud samples are relatively rare (~0.22% of the data), which requires that the model be able to learn from few positives while being drown into many negatives:which is given in ref.[16].

B Implementation and Training Protocol

We implement our CSSA model with PyTorch Geometric (Fey et al., 2019) and HuggingFace Transformers libraries (Wolf et al., 2019). The overall structure of our model has two main modules:

Semantic Reasoning Layer: We choose ChatGLM3-6B to be our base reasoner[17]. In order to improve the financial background without losing too much ability due to full fine-tuning,we utilize LoRA where r = 16, α = 32; i.e., extracting "semantic risk indicator" information from natural-language descriptions about transactions.

Graph-based Embedding Layer: We utilize a three-layer GCNs with the size of 256 for each layer as graph representation module[18], in which we use batch norm and set drop rate as 0.3 to avoid over-fitting certain popular classes.

We train our joint model for 150 epochs with four V100-80G GPUs on a distributed computer system and we are careful not to over-fit because of scarce data while doing the contrastive learning.

In the latent space, to avoid overfitting of the alignment loss on the majority legitimate class category, we use an oversampling ratio of 5:1 (minority) [19].

V. Experiments

A. Comparative Performance Analysis

To comprehensively verify the superiority of the proposed CSSA model, we compare it with following baselines:

(1) Standard GCN: a standard graph convolution method with only structure feature aggregation.

(2) Graph attention network (GAT). A state-of-the-art structural modeling approach, which is able to learn the weight of a neighbor's feature via self-attention mechanism.

(3) Transformer on Tabular: Sequential model in which we consider the series of transactions as a set of features, learning their complicated interrelationships with multi-head attention.

(4) LLM Zero-shot: The zero-shot LLM is a pure text-reasoning and non-structural-aware model with no finetune.

In Table 1, we summarize the results of analyzing our dataset to understand better how different approaches perform for crossmodal learning. First, it is clear that GCN and GAT have very low recalls:indicating that purely structural features may be insufficient to distinguish between complex fraud cases from legitimate large purchases.While Tabular Transformer performs better than GCN at learning feature correlations, it does not take into account the networkbased context of graphs.

Table 1. Performance comparison of CSSA against others.

Model Architecture	Precision	Recall	Macro-F1	PR-AUC
Vanilla GCN	0.654	0.122	0.206	0.315
GAT (Self-Attention)	0.712	0.185	0.294	0.388
Tabular Transformer	0.685	0.254	0.371	0.402
LLM-only (ChatGLM3)	0.451	0.382	0.414	0.42
CSSA (Ours)	**0.953**	**0.481**	**0.639**	**0.712**

CSSA achieves the highest values in terms of Macro-F1 (0.639) and PR-AUC (0.712), which are considerable improvements over our baseline model based on GAT. It also outperforms it significantly by a large margin in terms of Precision (0.953).suggesting that the semantic-structural

 doi:10.20944/preprints202602.0543.v1

fusion system acts like a very good filter.By rigorous mutual consensus between LLM's reasoning capability and GN's structure,We show that CSSA is able to effectively reduce the number of wrong classifications which often affect attention based methods when large amounts of information are present and there is a high amount of noise in the background.

B Ablation Study

To better understand the impact of different modules in the proposed CSSA architecture and compare the contributions from semantical information with that from modality alignment, we further systematically remove each module for an ablation study, as shown in Table 2.

Table 2. Ablation study results demonstrating the impact of LLM reasoning and Cross-modal Alignment.

Configuration	Precision	Recall	Macro-F1	Δ F1
Full CSSA Framework	0.953	0.481	0.639	-
w/o Cross-modal Alignment (L_align)	0.824	0.312	0.452	-18.70%
w/o LLM Reasoning (Raw Embeddings)	0.745	0.214	0.332	-30.70%
w/o GNN Structure (LLM-only)	0.451	0.382	0.414	-22.50%

When we remove the contrastive alignment loss function () and keep both branches of models, i.e., "Concatenationonly", the performance on Macro-F1 drops by 18.7%. It indicates that simple concatenation is not effective,and the model requires an explicit module for mapping between semantics and structure.When replacing the distilled LR of the LLM by ordinary number embedding (no LLM Reasoning),we see the worst performance degradation (-30.7%) which suggests, for the e-commerce fraud detection task,the context-based reasoning ability of the LLM is the most important feature element, which can be derived as follows:

Removing the GNN module (no GNN Structure), causes a sharp drop in Precision scores [20], which suggests that although LLMs are very good at detecting possibly fake stories,graph neural networks are crucial to test these suspicions with real transaction network data.

C Explainability and Case Study

The CSSA method has a clear advantage in terms of explainability; that is, when predicting results, the explanation path will be provided by the language model part to each abnormal node found.For instance, for "Seller Coordination", we found that the graph component identified a spike in interactions while the NLP component returned "mismatched activity times and location information from the seller." Such evidence provides strong grounds on which one can rely as they manually verify the results.

VI. Conclusions

In this paper, we propose CSSA to bridge the gap between context-sensitive knowledge in large-scale LMs and topology-aware capabilities of GNs by casting AD as a cross-modal retrieval problem.our model avoids limitations of classic feature learning approaches, as well as rigid graph diffusion strategies.

Specifically, we mainly introduce a new concept named Cross-modal Contrastive Alignment (CCA), which aims at aligning semantic "inference routes" encoded in LLMs and topological structures learned through GNNs in terms of mathematical representation.Large-scale empirical evaluation conducted over real-world large e-commerce dataset shows that our proposed CSSA model can achieve very good trade-offbetween precision and recall,especially for the difficult cases with only 0.22% fraud transaction rates, we are able to detect them at an accuracy level of as high as

 doi:10.20944/preprints202602.0543.v1

95.3% which proves that combining the semantic equivalence with relational structure provides much stronger indication about entities' functioning than each of these sources alone.

References

1. Dahiphale, D., Madiraju, N., Lin, J., Karve, R., Agrawal, M., Modwal, A., ... & Merchant, A. (2024, December). Enhancing Trust and Safety in Digital Payments: An LLM-Powered Approach. In 2024 IEEE International Conference on Big Data (BigData) (pp. 4854-4863). IEEE.
2. Ouyang, K., Ke, Z., Fu, S., Liu, L., Zhao, P., & Hu, D. (2024). Learn from global correlations: Enhancing evolutionary algorithm via spectral gnn. arXiv preprint arXiv:2412.17629.
3. Hacini, A. D., Benabdelouahad, M., Abassi, I., Houhou, S., Boulmerka, A., & Farhi, N. (2025). LLM-Assisted Financial Fraud Detection with Reinforcement Learning. Algorithms, 18(12), 792.
4. Kanikanti, V. S. N., Mula, K., Muthukumarasamy, K., Kubam, C. S., Goswami, B., & Gadam, H. (2026). Streaming analytics pipelines for LLM-based financial anomaly detection in Real-Time retail transaction flows. International Journal of Information Technology, 1-6.
5. Luo, R., Wang, N., & Zhu, X. (2025). Fraud detection and risk assessment of online payment transactions on e-commerce platforms based on llm and gcn frameworks. arXiv preprint arXiv:2509.09928.
6. Mingxiu Sui, Yiyun Su, Jiaqing Shen, et al. Intelligent Anti-Money Laundering on Cryptocurrency: A CNN-GNN Fusion Approach. Authorea. January 12, 2026, DOI:10.22541/au.176824645.56752786/v1.
7. Bisht, K. S. (2025). Conversational Finance: LLM-Powered Payment Assistant Architecture. European Journal of Computer Science and Information Technology, 13(27), 116-130.
8. Chen, Y., Liu, L., & Fang, L. (2024). An Enhanced Credit Risk Evaluation by Incorporating Related Party Transaction in Blockchain Firms of China. Mathematics, 12(17), 2673.
9. Lizi Chen, Yue Zou, Pengfei Pan, et al. Cascading Credit Risk Assessment in Multiplex Supply Chain Networks. Authorea. January 16, 2026, DOI: 10.22541/au.176858311.10362606/v1.
10. Zheng, H., Lin, Y., He, Q., Zou, Y., & Wang, H. (2026, January 30). Blockchain Payment Fraud Detection with a Hybrid CNN-GNN-LSTM Model.Https://Www.researchgate.net/Publication/400235797_Blockchain_Payment_Fraud_Detection_with_a_Hybrid_CNN-GNN-LSTM_Model. https://doi.org/10.13140/RG.2.2.26663.20641
11. Malingu, C. J., Kabwama, C. A., Businge, P., Agaba, I. A., Ankunda, I. A., Mugalu, B., ... & Musinguzi, D. (2025). Application of LLMs to Fraud Detection. World J. Adv. Res. Rev, 26, 178-183.
12. O'Neill, O., Ramanayake, R., Mandal, A., Pawar, U., Flanagan, W., Chatbri, H., & Martin, C. A Practical Taxonomy for Finance-Specific LLM Risk Detection and Monitoring. In NeurIPS 2025 Workshop: Generative AI in Finance.
13. Qin Wang, Bolin Huang, and Qianying Liu. 2026. Deep Learning-Based Design Framework for Circular Economy Supply Chain Networks: A Sustainability Perspective. In Proceedings of the 2025 2nd International Conference on Digital Economy and Computer Science (DECS '25). Association for Computing Machinery, New York, NY, USA, 836–840. https://doi.org/10.1145/3785706.3785837.
14. Artola Velasco, A., Tsirtsis, E., Okati, N., & Gomez Rodriguez, M. (2025). Is Your LLM Overcharging You? Tokenization, Transparency, and Incentives. arXiv preprint arXiv:2505.21627.
15. Trozelli, P., & Andersson Holm, A. (2024). Comparing GPT and Traditional Machine Learning in Fraud Detection.
16. Ke, Z., Cao, Y., Chen, Z., Yin, Y., He, S., & Cheng, Y. (2025). Early warning of cryptocurrency reversal risks via multi-source data. Finance Research Letters, 107890.
17. Bhatt, S., & Garg, G. (2025). NLP for Fraud Detection and Security in Financial Documents. In Transformative Natural Language Processing: Bridging Ambiguity in Healthcare, Legal, and Financial Applications (pp. 131-155). Cham: Springer Nature Switzerland.
18. Keyu Yuan, Yuqing Lin, Wenjun Wu, et al. Detection of Blockchain Online Payment Fraud Via CNN-LSTM. Authorea. January 15, 2026,doi:10.22541/au.176851576.63306241/v1.
19. Rollinson, N., & Polatidis, N. (2026). LLM-Generated Samples for Android Malware Detection. Digital, 6(1), 5.

 doi:10.20944/preprints202602.0543.v1

20. Li, Z., & Ke, Z. RepoLLM: A Multi-modal Foundation Model for Drug Repurposing via Alignment of Molecules, EHRs, and Knowledge Graphs.
21. Li, Long, Jiaran Hao, Jason Klein Liu, Zhijian Zhou, Yanting Miao, Wei Pang, Xiaoyu Tan et al. "The choice of divergence: A neglected key to mitigating diversity collapse in reinforcement learning with verifiable reward." arXiv preprint arXiv:2509.07430 (2025).
22. Wang, Chengkai, Di Wu, Yunsheng Liao, Wenyao Zheng, Ziyi Zeng, Xurong Gao, Hemmings Wu et al. "NeuroCLIP: A Multimodal Contrastive Learning Method for rTMS-treated Methamphetamine Addiction Analysis." arXiv preprint arXiv:2507.20189 (2025).
23. Wei, D., Wang, Z., Kang, H., Sha, X., Xie, Y., Dai, A., & Ouyang, K. (2025). A comprehensive analysis of digital inclusive finance's influence on high quality enterprise development through fixed effects and deep learning frameworks. Scientific Reports, 15(1), 30095.
24. Min, M., Duan, J., Liu, M., Wang, N., Zhao, P., Zhang, H., & Han, Z. (2026). Task Offloading with Differential Privacy in Multi-Access Edge Computing: An A3C-Based Approach. IEEE Transactions on Cognitive Communications and Networking.
25. Wang, Chengkai, Yifan Zhang, Chengyu Wu, Jun Liu, Xingliang Huang, Liuxi Wu, Yitong Wang, Xiang Feng, Yiting Lu, and Yaqi Wang. "MMDental-A multimodal dataset of tooth CBCT images with expert medical records." Scientific Data 12, no. 1 (2025): 1172.