BEHAVIOR-GROUNDED SEMANTIC ENRICHMENT FOR FINANCIAL FRAUD MODELING AND REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

In financial fraud detection, rich semantic context can provide important evidence for transaction behavior modeling and fraud reasoning. However, public real-world financial datasets often lack rich semantics due to privacy constraints. Consequently, synthetic datasets incorporate generated semantics, but at the cost of behavioral realism; textual descriptions for contextual reasoning remain scarce. We address this gap through a semantic enrichment framework grounded in original transaction behavior to simulate multimodal financial data. We (1) propose a multi-agent semantic enrichment framework that generates interpretable financial semantics grounded in transaction behavior through role-specialized agents and consistency refinement, and (2) newly contribute a valuable multimodal financial fraud dataset, MS-FFSD, enriched with structured semantics and textual semantics while preserving real-data-grounded transaction behavior. Furthermore, we systematically analyze the quality and utility of semantic enrichment. Results demonstrate statistical fidelity and framework generalizability, while showing that richer semantics benefit fraud modeling and context-aware LLM reasoning. Overall, this work advances multimodal financial fraud research and bridges emerging LLM and multi-agent capabilities with operational anti-fraud practice. The framework and dataset are released at `https://anonymous.4open.science/r/MS-FFSD-D7FD`.

1 INTRODUCTION

Fraud detection is a critical task across diverse real-world domains (Cherif et al., 2023; Cheng et al., 2025), and corresponding public datasets range from social reviews (Ni et al., 2019; Rayana & Akoglu, 2015) to consumer finance (Huang et al., 2022; Vesta Corporation, 2019), with varying modality richness. Among these modalities, semantic and textual context can provide important evidence for understanding fraudulent behavior, and recent advances in large language models (LLMs) have created new opportunities for context-aware fraud analysis (Zhang et al., 2024; Caffagni et al., 2024; Chang et al., 2024). However, financial fraud detection presents a distinct challenge: transaction data constitute a central resource across banking (Pozzolo et al., 2015) and e-commerce systems (Vu, 2018), yet public financial transaction datasets remain semantically sparse (Grover et al., 2023; Ma et al., 2023), thereby limiting support for leveraging semantics in fraud modeling and reasoning.

In real-world financial systems, transaction data are often accompanied by rich semantic information describing participating entities and transaction scenarios, which can support risk modeling and reasoning (Zhang et al., 2026b; Dou et al., 2026; Zhang et al., 2026c). Yet upon public release, privacy constraints necessitate the removal or anonymization of semantically rich information, leaving the resulting datasets dominated by numerical features and coded attributes (Grover et al., 2023; Fan et al., 2025; Li et al., 2025b). This semantic scarcity has motivated the development of synthetic financial datasets that provide richer information, such as user profiles, merchant categories, and geographic attributes (Altman, 2022; Oztas et al., 2023). However, because their transaction behaviors are largely generated through hand-crafted rules or simulators, these datasets inevitably differ from real financial systems in structural distributions and behavioral complexity (Zhao et al., 2025; Altman et al., 2023).

This creates an important gap in existing financial transaction resources: public real-world datasets retain realistic behavioral structure but limited semantic context, whereas rule-driven synthetic datasets provide richer semantics but limited behavioral realism. Moreover, existing datasets rarely provide entity-level textual descriptions, limiting their support for semantic reasoning. Table 1

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Comparison of public financial transaction datasets. Real-Data Grounded denotes real transaction origin or generation from real transaction data; Feature Interpretability and Original Representation indicate retained feature meanings and untransformed values; Transaction, Entity, and Textual indicate available semantic information. MS-FFSD covers all dimensions.

	Data Foundation	Feature Information		Semantic Information		
Dataset	**Real-Data Grounded**	**Feature Interpretability**	**Original Representation**	**Transaction**	**Entity**	**Textual**
Elliptic++ (Elmougy & Liu, 2023)	✓	✓	✓	✓	✓	✗
FraudEcom (Vu, 2018)	✓	✓	✓	✓	✗	✗
IEEE-CIS (Vesta Corporation, 2019)	✓	✗	✓	✗	✗	✗
Credit Card Fraud (Pozzolo et al., 2015)	✓	✗	✗	✗	✗	✗
Sparkov (Shenoy, 2020)	✗	✓	✓	✓	✓	✗
CCT (Altman, 2019)	✗	✓	✓	✓	✓	✗
SAML-D (Oztas et al., 2023)	✗	✓	✓	✓	✓	✗
S-FFSD (Xiang et al., 2023)	✓	✓	✓	✗	✗	✗
MS-FFSD (ours)	✓	✓	✓	✓	✓	✓

summarizes these differences across representative public financial transaction datasets. This motivates behavior-grounded semantic enrichment. To model the multidimensional and interdependent financial semantics involved, we propose a multi-agent framework. Designed for transaction data with user–merchant interactions and known feature meanings, the framework integrates observed transaction behavior with external statistical and domain knowledge to guide structured and textual semantic generation, while preserving the original transaction structure and fraud labels. Specifically, the framework comprises agents for temporal construction, user semantic initialization, merchant semantic assignment, cross-entity consistency refinement, and entity-level textual description generation. The agents share transaction and entity states, with generation guided by external knowledge, global constraints, and fraud-aware preservation. The resulting structured semantics augment the tabular modality, while textual semantics form the textual modality. Together, they yield multimodal financial transaction data that support both fraud modeling and contextual reasoning.

As a semantically rich and publicly accessible data contribution, we construct and release MS-FFSD, a multimodal financial fraud dataset supporting both fraud modeling and semantic reasoning. S-FFSD (Xiang et al., 2023) provides a suitable foundation because it preserves feature meanings and explicit user–merchant interactions, while its transactions were generated by a graph-based model trained on real-world data (Li et al., 2025a), rather than through hand-crafted rules. Applying the framework, we enrich the original transactions with structured and textual semantics while preserving their real-data-grounded behavior. We further apply the framework to two private real-world financial datasets to establish its broader applicability across financial datasets.

Our main contributions are as follows:

- We propose a multi-agent semantic enrichment framework that generates financial semantics grounded in transaction behavior through role-specialized agents and consistency refinement.
- We contribute and release MS-FFSD, a valuable multimodal financial fraud dataset enriched with structured and textual semantics while preserving real-data-grounded transaction behavior.
- We systematically analyze the quality and utility of semantics for financial fraud analysis. Specifically, we assess statistical fidelity and framework generalizability, and provide insights into how semantic enrichment affects financial fraud modeling and semantic reasoning.

2 Related Works

Real-World Financial Transaction Datasets. Publicly available real-world financial transaction datasets preserve transaction records and fraud signals from operational systems, but privacy and regulatory constraints necessitate semantic information to be removed or anonymized before release (Grover et al., 2023; Yoon et al., 2026; Yeh et al., 2026). Elliptic++ (Elmougy & Liu, 2023) preserves real Bitcoin transactions and wallet entities, but the entities remain pseudonymous rather than real-world identities. FraudEcom (Vu, 2018) provides interpretable transaction context but

limited entity semantics. IEEE-CIS (Vesta Corporation, 2019) retains transaction and identity features, but many variables are anonymized into uninterpretable numerical values or categorical codes. Credit Card Fraud (Pozzolo et al., 2015) further replaces original values with transformed representations. Consequently, these datasets support fraud modeling with real transaction behavior but provide limited semantics for LLM-based fraud detection and contextual reasoning.

Synthetic Financial Transaction Datasets. Synthetic datasets alleviate the above limitation by constructing entity and transaction semantics through simulation (Altman et al., 2023; Altman, 2022; Zhao et al., 2025). Sparkov (Shenoy, 2020) generates transactions from predefined user profiles and fraud patterns. CCT (Altman, 2019; Padhi et al., 2021) further models user motivations and merchant dependencies to generate transaction sequences. SAML-D (Oztas et al., 2023) extends this paradigm to anti-money laundering by jointly simulating transaction semantics and laundering patterns. In these datasets, transaction structures are therefore largely determined by predefined semantics and simulation mechanisms, inevitably differ from the behavioral complexity of real financial systems.

Semantic Enrichment Frameworks. The above limitations motivate behavior-grounded semantic enrichment. Sato (Zhang et al., 2020) recovers column semantics from table context; ReTabAD (Yoon et al., 2026) restores dataset, column and label semantics from original documentation. Yet for privacy reasons, financial semantics are irreversibly lost, motivating generation from observable evidence. In general, LogicNLG (Chen et al., 2020) and PLOG (Liu et al., 2022) generate logical text grounded in tabular facts. In the financial domain, semantic enrichment requires generating inferential text grounded in transaction behavior. Recent multi-agent frameworks support this process through agent coordination and iterative refinement (Islam et al., 2024; Chen et al., 2024b; Zhang et al., 2025b; 2026a). Given the multidimensionality and interdependent nature of financial semantics, our framework assigns agents distinct semantic roles with dedicated external knowledge and priors, and enforces consistency both with observed transactions and across related entities.

3 Multi-Agent Semantic Enrichment Framework

3.1 Framework Overview

We propose a multi-agent semantic enrichment framework grounded in transaction behavior, shown in Figure 1. Let $\mathcal{D}$ denote the original transactions and $\mathcal{Z}_S = (\boldsymbol{\tau}, \boldsymbol{r}, \boldsymbol{q})$ the temporal, user, and merchant semantic state. The structured enrichment is summarized by the following unified objective:

$$\mathcal{F}_S(\mathcal{Z}_S) = F_T(\boldsymbol{\tau}) + \lambda_U F_U(\boldsymbol{r}) + \lambda_M F_M(\boldsymbol{q}; \boldsymbol{r}, \boldsymbol{\tau}) + \lambda_C F_C(\boldsymbol{r}, \boldsymbol{q}; \boldsymbol{\tau}, \mathcal{G}_{UM}), \tag{1}$$

The four terms model temporal construction, user initialization, merchant initialization, and cross-entity refinement under the external knowledge base $\mathcal{K}$, feasible region $\mathcal{C}(\mathcal{D}, \mathcal{K})$, and observed user–merchant graph $\mathcal{G}_{UM}$. They are realized through staged agent collaboration while preserving the original transaction structure and fraud labels. The Textual Agent then generates entity-level descriptions from the refined structured state. Detailed formulations are provided in Appendix B.

3.2 Structured Semantic Initialization

Structured semantic enrichment grounds anonymous transactions in interpretable temporal, user, and merchant contexts. The three initialization agents are formulated under a unified objective:

$$F_A = \sum_{j \in \mathcal{J}_A} \lambda_j \mathcal{L}_j^A, \quad A \in \{T, U, M\}, \quad \mathcal{J}_T = \{r, u, b\},\ \mathcal{J}_U = \{g_n, g_a, d\},\ \mathcal{J}_M = \{c, a, s\}. \tag{2}$$

For the Temporal Agent, r, u, and b denote global rhythm matching, user-specific temporal patterns, and fraud-burst preservation, respectively. It reconstructs order-preserving timestamps, using IEEE-CIS[1] as a public reference for coarse temporal rhythms and incorporating user-level patterns when

[1] https://www.kaggle.com/c/ieee-fraud-detection/

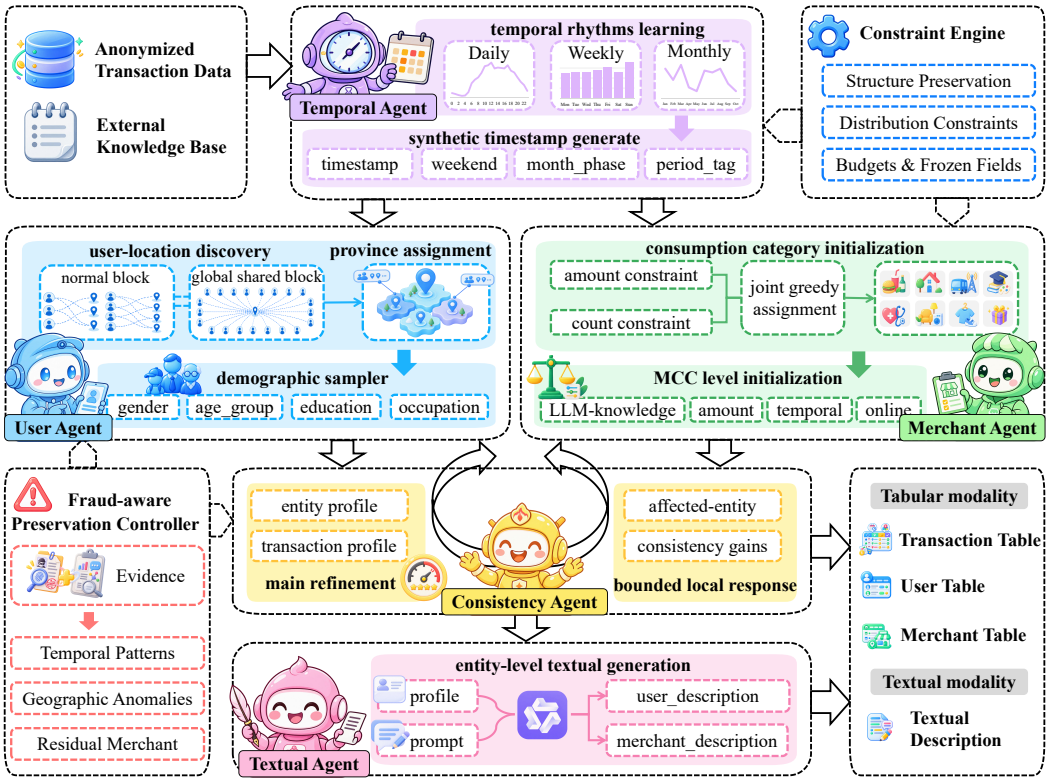


Figure 1: Overview of the proposed multi-agent semantic enrichment framework. The agents sequentially construct synthetic timestamp, user attributes, merchant semantics, cross-entity consistency, and entity-level textual descriptions, while shared knowledge, constraint, and fraud-preservation mechanisms, resulting in multimodal financial transaction data with tabular and textual modalities.

sufficient evidence is available. For the User Agent, g_n and g_a capture geographic matching by user count and transaction amount, while d represents conditional demographic distributions. The agent combines observed user–location interactions with external priors to initialize geographic attributes, gender, age group, education, and occupation. For the Merchant Agent, c, a, and s represent merchant-count, transaction-amount, and semantic-compatibility terms. These objectives guide hierarchical category assignment by balancing distributions and semantic compatibility.

3.3 Cross-Entity Semantic Refinement

Independently initialized user and merchant semantics may remain inconsistent with observed cross-entity interactions. We introduce a Consistency Agent that iteratively refines user and merchant semantics over graph $\mathcal{G}_{UM}$. Its objective combines user and merchant side consistency:

$$F_C = \mathcal{L}_U^C + \mathcal{L}_M^C, \tag{3}$$

where the two terms capture cross-entity consistency. At iteration t, let $Z^{(t)}$ be the current state and $Z_{-e}^{(t)}$ exclude entity e. Each candidate update is evaluated by its net consistency gain:

$$\Delta_e(z'_e; Z^{(t)}) = S_e(z'_e; Z_{-e}^{(t)}) - S_e(z_e^{(t)}; Z_{-e}^{(t)}) - \kappa_e \mathbb{I}[z'_e \neq z_e^{(t)}]. \tag{4}$$

where κ_e penalizes unnecessary semantic changes. Strictly positive-gain updates satisfying distributional and modification constraints are accepted. Changes on one side activate only affected counterparts on the other, enabling bidirectional responses without global reassignment. Refinement terminates when no admissible update remains or the modification budgets are exhausted.

3.4 TEXTUAL SEMANTIC GENERATION

After structured refinement, the Textual Agent converts the refined entity states into natural-language descriptions using Qwen3-32B (Yang et al., 2025). The generation process is summarized as

$$\boldsymbol{d}^\star = A_X(\mathcal{D}, \mathcal{Z}_S^\star, \mathcal{K}), \qquad \mathcal{Z}^\star = (\mathcal{Z}_S^\star, \boldsymbol{d}^\star). \tag{5}$$

where $\boldsymbol{d}^\star$ denotes the generated entity descriptions and $\mathcal{Z}^\star$ the final enriched semantic state. Generation is grounded in the refined structured semantics and aggregated transaction evidence associated with each entity. Single-transaction users are described using deterministic templates due to insufficient evidence for reliable behavioral aggregation, while multi-transaction users use LLM-based generation from aggregated behavioral profiles. Merchant descriptions combine business semantics, transaction behavior, and associated user profiles, with category-level memory reducing repetitive concepts. The Textual Agent does not modify the refined structured state during generation. The complete generation prompts and deterministic template are provided in Appendix C.

3.5 AGENT COLLABORATION AND SHARED CONTROL

The specialized agents operate over shared transaction and entity states, forming a staged workflow in which temporal semantics support entity initialization, user and merchant states are iteratively refined through observed interactions, and the resulting structured states are passed to the Textual Agent. Cross-entity updates are restricted to affected counterparts and bounded by shared constraints, enabling controlled information exchange without unconstrained propagation.

Agent collaboration is governed by three shared components. The External Knowledge Base $\mathcal{K}$ provides statistical and domain knowledge for semantic construction, with its sources, granularity, temporal alignment, and roles detailed in Appendix A. The Constraint Engine defines the feasible region $\mathcal{C}(\mathcal{D}, \mathcal{K})$ through structural preservation, distributional consistency, category restrictions, and modification budgets. The Fraud-aware Preservation Controller further protects evidence-supported fraud patterns from being removed during enrichment. Together, these components coordinate agent updates while preserving the structural, statistical, and fraud-related properties of the original data.

4 MS-FFSD DATASET

4.1 DATASET OVERVIEW

MS-FFSD is the multimodal financial fraud dataset constructed and released in this work. It builds on S-FFSD (Xiang et al., 2023), whose transaction structure was generated by a graph-based generative model trained on real-world data, and enriches it with semantics grounded in observed transaction behavior and informed by real-world statistical distributions and domain knowledge. The enrichment process introduces two forms of semantic information, *Structured Semantics* and *Textual Semantics*, while preserving the original transaction records, interactions between users and merchants, and fraud labels. The resulting dataset integrates the original transaction data with enriched semantic information to support multimodal financial fraud research. Table 2 summarizes its scale, label, and period.

Table 2: Summary statistics of MS-FFSD.

Statistic	Value
Transactions	77,881
Users	30,346
Merchants	886
Normal transactions	31.31%
Fraud transactions	6.75%
Unlabeled transactions	61.94%
Observation period	Jan.–Oct. 2021

4.2 DATA ORGANIZATION AND MODALITIES

MS-FFSD comprises two modalities: tabular and textual, as illustrated in Figure 2. The transaction table preserves all original transaction fields and augments with generated temporal and geography. The user table provides demographic and geographic attributes, including province, gender, age group, education, and occupation, while the merchant table provides hierarchical business categories and

transaction.csv

Datetime	Source	Target	Amount	Location	Type	Transaction_province	Labels
2021/5/27 14:51:38	S24623	T1023	18.62	L100	TP120	Shanghai	2
2021/5/27 14:57:58	S24589	T1013	37.63	L102	TP108	Chongqing	1
2021/5/27 15:06:42	S24589	T1013	37.62	L102	TP108	Xinjiang	0

user.csv

Source	Province	Gender	Age_group	Education	Occupation_industry
S10200	Shanxi	Female	Older Adults	Senior Secondary School	Real Estate
S10201	Chongqing	Male	Middle-aged Adults	Junior College and Above	Manufacturing
S10202	Hainan	Male	Young Adults	Junior Secondary School	Education

merchant.csv

Target	MCC_level1	MCC_level2	Merchant_name
T1004	Apparel and General Shopping	Outlet Shopping	Value Threads Outlet
T1005	Food, Tobacco and Liquor Retail	Specialty Food Retail	Global Pantry Co.
T1006	Education Payments and Training Services	Higher Education	Summit Academic Institute

Tabular modality

user description.csv

The user's consumption focuses on urban public transport, convenience stores, and full-service dining, reflecting localized, routine-based spending. Activity clusters around transit hours and weekends, with recurring early- and late-month peaks. Frequent, modest expenditures occur primarily through offline channels, emphasizing structured yet flexible short-term consumption rhythms.

merchant description.csv

Springboard Youth Academy offers after-school programs and enrichment workshops for children and teenagers, focusing on creative arts, STEM activities, and leadership development. The center provides structured classes in music, robotics, and drama, alongside mentorship sessions tailored to foster personal growth and academic support. Its curriculum emphasizes interactive learning through hands-on projects and collaborative group work.

Textual modality

Figure 2: Data organization and modalities of MS-FFSD. Tabular modality: *transaction* preserves original fields and adds temporal and geographic attributes. *user* provides demographic and geographic attributes. *merchant* provides hierarchical MCC categories and identities. Textual modality: *description* summarizes user behavior and merchant business characteristics in natural language.

merchant names. These tables constitute the *tabular modality*. The corresponding description files contain entity-level natural-language descriptions of users and merchants, representing behavioral characteristics and business activities, forming the *textual modality*.

The enriched semantics can be viewed from three semantic perspectives. *Transaction-level Structured Semantics* characterize individual transactions through temporal and geographic context in addition to the original transaction attributes, including amounts and types. *Entity-level Structured Semantics* describe users and merchants through demographic, geographic, and business-category attributes. *Textual Semantics* further capture user behavioral characteristics and merchant identities and business activities through natural-language descriptions. Single-transaction users are described using deterministic templates, while multi-transaction users receive LLM-generated descriptions based on aggregated behavioral patterns; merchant descriptions are generated from category semantics, transaction behavior, and associated user profiles. Together, these three perspectives provide a comprehensive semantic characterization of financial transactions and entities. Appendix D provides complete field definitions, user semantic attributes, merchant category definitions, and textual description statistics.

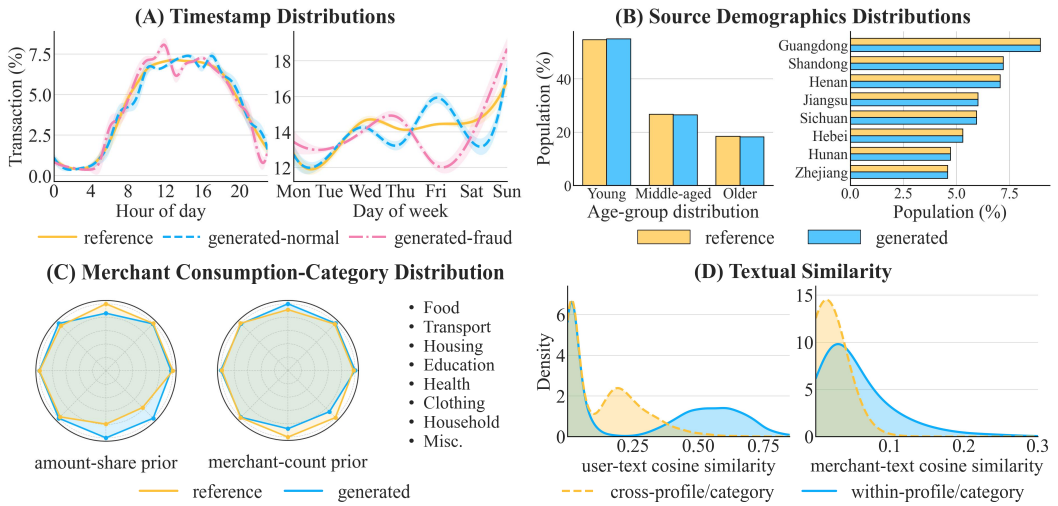


Figure 3: Semantic enrichment quality of MS-FFSD. (A) Generated timestamps follow reference rhythms. (B) Generated user demographics align with reference age-group and provincial distributions. (C) Merchant radar plots show category relative deviations from reference amount and count priors. (D) Textual descriptions show clear within-group coherence and cross-group discriminability.

Table 3: Downstream fraud detection performance on MS-FFSD under progressive semantic enrichment. *Structured* augments *Original* with structured semantic attributes, while *Textual* further adds entity-level textual descriptions. Positive gains over *Original* are shaded, with darker colors indicating larger improvements within each metric. All results are reported in percentage (%).

	Original Features			+ Structured Semantics			+ Textual Semantics		
Model	**AUC**	**AP**	**F1**	**AUC**	**AP**	**F1**	**AUC**	**AP**	**F1**
GCN	85.30	96.29	70.04	86.47	96.31	74.29	89.04	97.14	76.05
GAT	86.92	96.58	73.47	87.83	97.09	74.07	88.24	97.19	73.76
GraphSAGE	89.43	97.43	74.86	89.82	97.27	76.42	90.27	97.59	75.13
CARE-GNN	86.03	96.07	71.02	86.48	96.27	71.61	88.28	96.81	72.81
PC-GNN	87.86	96.82	76.73	87.86	96.80	76.15	88.94	96.97	77.89
H2-FDetector	87.05	96.06	70.04	87.77	96.93	71.69	88.41	97.00	73.12
BWGNN	88.00	96.80	73.97	87.34	96.54	70.95	89.34	97.14	76.08
GAGA	85.75	95.62	75.96	87.54	95.96	76.45	85.73	96.51	76.18
SparseGAD	89.99	97.16	76.30	87.75	96.45	76.12	90.42	97.47	77.26
SplitGNN	87.42	96.51	71.81	86.29	96.31	74.38	88.65	96.99	76.27
ConsisGAD	86.57	96.64	74.27	86.87	96.57	74.59	88.34	96.95	71.97
PMP	89.40	97.23	74.77	89.38	97.32	73.93	89.75	97.44	74.83
DGA-GNN	88.76	97.19	79.17	90.86	97.56	79.06	91.09	97.62	80.37
SpaceGNN	88.77	97.51	77.60	88.59	97.00	74.82	90.33	96.33	75.38
CGADM	90.09	97.46	80.67	89.76	97.24	78.34	90.37	97.33	80.71
GAAP	89.73	97.35	76.09	87.60	96.35	75.16	89.57	97.18	77.58
Average	87.94	96.80	74.80	88.01	96.75	74.88	89.17	97.10	75.96

5 EVALUATION

We first assess whether the generated semantics exhibit distributional fidelity and semantic informativeness, establishing the validity of the enrichment results. We then investigate how progressively enriched semantic information changes the way models exploit and interpret financial transaction behavior. From the predictive perspective, we evaluate its value through graph-based fraud detection and further examine the broader applicability of the enrichment framework on two additional private real-world datasets. From the reasoning perspective, we analyze how progressively enriched semantics support LLM-based contextual interpretation of transaction behavior.

5.1 SEMANTIC ENRICHMENT QUALITY

The validity of the generated semantics is examined from two complementary aspects: distributional fidelity and semantic informativeness, covering temporal, user, merchant, and textual perspectives. As shown in Figure 3(A)–(C), the generated semantics exhibit strong distributional fidelity to their corresponding references. Normal timestamps closely follow the reference hourly and weekly rhythms, while the generated age-group and provincial distributions show little deviation from the demographic priors. Merchant-category assignments also remain broadly aligned with both amount and count priors. Overall, the enrichment process preserves the major statistical characteristics.

Semantic informativeness arises from preserving fraud-relevant variation and discriminative semantic structure. In Figure 3(A), fraud timestamps exhibit distinct temporal patterns from normal transactions, with stronger activity around midday. Across the week, fraud transactions show a pronounced decline before the weekend followed by a strong weekend increase. In Figure 3(D), within-group descriptions show greater density at higher cosine similarities than cross-group descriptions, indicating semantic coherence within profiles and categories while retaining variation and cross-group discriminability.

5.2 DOWNSTREAM FRAUD DETECTION WITH GRAPH MODELS

From the predictive perspective, we examine how progressive semantic enrichment affects downstream graph-based fraud detection on MS-FFSD. Users and merchants are represented as nodes in an entity graph, with fraud detection performed on user nodes. We consider 16 representative graph-based detectors: GCN Kipf & Welling (2017), GAT Veličković et al. (2018), GraphSAGE Hamilton et al. (2017), CARE-GNN Dou et al. (2020), PC-GNN Liu et al. (2021), H2-FDetector Shi et al. (2022), BWGNN Tang et al. (2022), GAGA Wang et al. (2023), SparseGAD Gong et al. (2023), SplitGNN Wu et al. (2023), ConsisGAD Chen et al. (2024a), PMP Zhuo et al. (2024), DGA-GNN Duan et al. (2024), SpaceGNN Dong et al. (2025), CGADM Wei et al. (2025), and GAAP Duan et al.

Table 4: Downstream fraud detection performance on private real-world financial datasets under progressive semantic enrichment. *Structured* augments *Original* with structured semantics, while *Textual* further adds entity textual descriptions. Positive gains over *Original* are shaded, with darker colors indicating larger improvements within each metric. Results are reported in percentage (%).

	Private-1									Private-2								
	Original			+ Structured			+ Textual			Original			+ Structured			+ Textual		
Model	**AUC**	**AP**	**F1**	**AUC**	**AP**	**F1**	**AUC**	**AP**	**F1**	**AUC**	**AP**	**F1**	**AUC**	**AP**	**F1**	**AUC**	**AP**	**F1**
GCN	90.36	98.84	70.09	93.73	99.28	76.82	94.17	99.47	78.39	96.43	98.98	87.21	96.41	98.97	88.69	97.90	99.41	89.88
GAT	93.43	99.26	75.34	94.22	99.40	77.32	94.68	99.41	77.99	96.85	99.04	89.58	97.00	99.08	90.21	97.36	99.20	91.24
GraphSAGE	94.66	99.42	76.08	94.67	99.40	77.61	95.04	99.45	78.16	98.67	99.63	92.95	98.51	99.59	93.16	98.76	99.64	93.18
CARE-GNN	90.10	98.31	64.96	92.05	99.09	71.71	93.99	99.34	77.83	97.75	99.38	91.50	98.02	99.43	90.75	98.06	99.44	92.54
PC-GNN	90.66	97.82	61.85	91.05	98.83	75.74	93.54	99.26	75.72	97.81	99.40	90.34	98.51	99.59	90.04	98.25	99.52	92.13
H2-FDetector	90.78	97.07	71.69	93.82	99.29	73.78	93.66	99.39	73.67	97.56	99.31	86.76	96.71	99.03	86.75	97.02	99.43	88.49
BWGNN	93.47	99.23	75.63	92.65	99.14	71.12	94.10	99.36	76.02	97.75	99.36	91.13	97.75	99.38	90.96	98.31	99.52	92.38
GAGA	92.82	99.18	78.32	92.55	98.95	67.14	92.93	99.24	79.35	97.99	99.46	91.37	98.11	99.48	92.84	98.18	99.50	91.96
SparseGAD	93.98	99.34	73.83	94.40	99.37	73.54	94.87	99.44	78.88	97.95	99.40	89.71	97.87	99.42	90.96	98.39	99.56	91.30
SplitGNN	93.22	99.33	77.91	94.11	99.23	76.16	93.91	99.38	76.63	97.48	99.10	91.28	97.43	99.29	91.48	97.82	99.41	91.19
ConsisGAD	92.40	99.16	65.14	92.58	99.17	66.28	93.55	99.22	80.21	95.36	98.15	85.29	95.58	98.17	86.92	96.56	99.27	87.37
PMP	95.60	99.51	80.03	95.33	99.45	78.65	95.72	99.66	78.16	98.32	99.55	88.19	98.26	99.52	88.48	98.36	99.57	92.81
DGA-GNN	93.92	99.32	78.22	94.38	99.38	78.71	95.98	99.21	79.65	97.79	99.39	91.28	97.83	99.43	91.86	98.40	99.54	92.68
SpaceGNN	94.78	99.42	74.82	94.43	99.35	75.64	95.00	99.45	74.54	98.51	99.59	91.19	98.45	99.60	93.64	98.71	99.65	94.16
CGADM	95.42	99.49	79.80	95.31	99.47	78.87	95.46	99.52	79.28	96.86	97.38	93.04	98.17	98.75	93.14	98.58	99.62	94.25
GAAP	94.74	99.37	80.40	94.98	99.38	81.51	95.31	99.10	82.68	97.83	99.40	92.38	98.19	99.49	93.93	98.77	99.65	94.16
Average	93.15	99.00	74.01	93.77	99.26	75.04	94.49	99.37	77.95	97.56	99.16	90.20	97.68	99.26	90.86	98.09	99.50	91.86

(2025). Each model is evaluated under three progressive settings: *Original Features*, + *Structured Semantics*, and + *Textual Semantics*. The graph structure and experimental configuration remain fixed, with only the available semantics progressively enriched. Structured and textual semantics are encoded with Qwen3-Embedding-8B (Zhang et al., 2025a), while numerical features are projected separately. Detailed configurations are provided in Appendix F, including entity-graph construction and relation design, semantic template construction, feature encoding, and experimental settings.

Table 3 shows that the enriched semantic information in MS-FFSD improves downstream fraud detection across diverse graph architectures, with darker shading indicating larger gains over *Original*. First, structured semantics produce mixed effects, with some detectors benefiting while others show marginal changes or occasional degradation, indicating architecture-dependent integration of multiple semantic signals. Second, textual semantics yield more consistent improvements and often recover performance drops introduced under structured semantics. This suggests that textual semantics enrich the structured representation with additional behavioral and business context aggregated for entity-level description generation. Third, the gains are more evident in AUC and F1-macro than in AP, suggesting that semantic enrichment primarily improves the separation of difficult or borderline samples and classification decisions, while producing smaller changes to the positive-sample ranking.

5.3 General Applicability of the Semantic Enrichment Framework

To establish the broader applicability of the proposed semantic enrichment framework, we further examine whether its predictive benefits can be consistently reproduced across datasets with different characteristics. We apply the enrichment framework to two private real-world financial transaction datasets. Further dataset statistics and descriptions are provided in Appendix E. The generated structured and textual semantics are evaluated following the progressive protocol in Section 5.2.

Table 4 shows that the semantics generated by the framework improve downstream fraud detection on both datasets. Structured semantics exhibit model-dependent effects, whereas incorporating textual semantics leads to broader and more consistent improvements across detectors. Notably, the gains persist even when the original features already provide strong predictive performance, suggesting that the generated semantics add predictive information beyond the existing feature space. The recurring improvement pattern across independently enriched datasets provides empirical evidence for the effectiveness and broader applicability of the proposed framework.

Reasoning Example A: Contextualizing an Apparent Anomaly

Behavioral Evidence. The user shows substantial transaction-amount variability, ranging from small purchases to several-hundred-dollar expenditures.

Textual Evidence. Small payments are associated with quick-service food and convenience purchases, whereas larger transactions occur at an appliance and furniture repair provider offering scheduled maintenance and repair services.

LLM Reasoning. The amount variability partly reflects heterogeneous consumption contexts. Small payments are compatible with quick-service food purchases, while larger expenditures are more plausible for household repair and maintenance services. The cross-category amount range should therefore not be treated as anomalous based on numerical dispersion alone.

Reasoning Example B: Identifying a Residual Anomaly

Behavioral Evidence. The user's transaction amounts span several orders of magnitude, from repeated sub-dollar payments to transactions above $1,000, with a maximum of $2,867.

Textual Evidence. Sub-dollar payments occur at a low-cost urban transit provider, while transactions above $1,000 include large-scale catering services. The $2,867 transaction, however, occurs at an accessible-priced workwear retailer.

LLM Reasoning. The amount variability partly reflects heterogeneous merchant contexts rather than inconsistent spending within a single activity. Sub-dollar payments are compatible with low-cost transit, while larger expenditures are plausible for large-scale catering. In contrast, the $2,867 purchase is difficult to reconcile with the positioning of an accessible workwear retailer, leaving a residual anomaly unexplained by the semantic context.

Figure 4: Illustrative examples of LLM-based semantic reasoning. Example A shows how rich-text semantics support context-aware interpretation of numerical variation, while Example B shows how semantic reasoning distinguishes contextually plausible variation from a residual anomaly.

5.4 LLM-BASED SEMANTIC REASONING

From the reasoning perspective, we conduct zero-shot semantic reasoning with DeepSeek-V3 (DeepSeek-AI et al., 2025) under progressive enrichment settings (Kojima et al., 2022; Tsai et al., 2025). We use a qualitative case-based analysis to examine whether richer semantics support contextual reasoning over transaction behavior. Original features, structured and textual semantics are introduced in sequence. At each level, the LLM and graph models receive the same underlying information but in different representations: graph models use projected numerical features and text embeddings, whereas the LLM directly reads the corresponding numerical values and textual semantics. Setup, prompts, and case selection criteria are detailed in Appendix G.

Figure 4 presents illustrative examples illustrating how progressive semantic enrichment changes the LLM's interpretation of transaction behavior. With original features, the reasoning is primarily driven by numerical patterns and observed transaction variations. Structured semantics ground these observations in explicit user and merchant attributes, while textual semantics further integrate them into contextual behavioral interpretations. In Example A, this richer context helps reconcile numerical variations with plausible behavior. In Example B, part of the variation becomes contextually explainable, while a residual anomaly remains inconsistent with the entity profile. Together, these cases show that textual semantics support contextual interpretation by integrating transaction patterns, entity attributes, and behavioral context, while preserving sensitivity to unexplained abnormalities.

6 CONCLUSION

In this work, we proposed a multi-agent semantic enrichment framework for financial fraud detection and newly contributed a valuable multimodal fraud detection dataset MS-FFSD that enriches with structured and textual semantics grounded in observed transaction behavior. Our evaluation demonstrates the statistical fidelity and generalizability of the enrichment framework, while showing the value of richer semantics for financial fraud modeling and contextual reasoning. These results highlight the potential of behavior-grounded semantic enrichment for advancing multimodal financial fraud research. Furthermore, our work facilitates systematic probing of the capabilities and boundaries of LLMs and multi-agent systems in realistic and complex financial settings. Crucially, it establishes a bridge between academic advances and anti-fraud practice, positioning emerging modeling and reasoning capabilities for operational financial fraud prevention.

AI USE STATEMENT

In this work, we used generative AI tools as part of the proposed semantic enrichment framework to generate textual semantics for the released dataset. Generative AI was also used during research development to assist with methodological discussion, mathematical formulation, experimental analysis, and interpretation of results. In addition, we used generative AI for literature analysis, brainstorming, drafting and editing parts of the manuscript, improving readability, and refining the presentation and structure of the paper. All AI-assisted methodological decisions, generated data, mathematical formulations, experimental interpretations, and manuscript content were reviewed and verified by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work focuses on semantic enrichment for financial fraud research and does not involve direct interaction with human subjects. The released dataset does not contain personally identifiable information, and the enriched demographic, geographic, merchant, and textual semantics are synthetically constructed rather than observations of real individuals. These generated attributes should therefore not be interpreted as factual profiles of specific persons. The released dataset and framework are intended for research purposes.

REPRODUCIBILITY STATEMENT

We provide detailed documentation and open-source resources to support reproducibility of both the semantic enrichment framework and the resulting dataset. An anonymous repository containing the released dataset and the complete implementation of the semantic enrichment framework is linked in the abstract. The main paper describes the overall framework, dataset organization, and evaluation protocols. Appendix B provides the mathematical formulation and agent-level objectives, Appendix A documents the external knowledge sources and their roles, and Appendix D provides detailed definitions and statistics of the enriched semantic attributes. Appendix E describes the two private transaction datasets used for framework evaluation. The prompts and deterministic templates used for textual generation are provided in Appendix C. Detailed settings for graph-based fraud detection and LLM reasoning are reported in Appendices F and G, respectively.

REFERENCES

Erik Altman. Credit Card Transactions. Kaggle Dataset, 2019. URL `https://www.kaggle.com/datasets/ealtman2019/credit-card-transactions`.

Erik Altman. Synthesizing credit card transactions. In *Proceedings of the Second ACM International Conference on AI in Finance*, ICAIF '21, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391481.

Erik Altman, Jovan Blanuša, Luc von Niederhäusern, Beni Egressy, Andreea Anghel, and Kubilay Atasu. Realistic synthetic financial transactions for anti-money laundering models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 29851–29874. Curran Associates, Inc., 2023.

Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. The revolution of multimodal large language models: A survey. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13590–13618, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15 (3), March 2024. ISSN 2157-6904.

Nan Chen, Zemin Liu, Bryan Hooi, Bingsheng He, Rizal Fathony, Jun Hu, and Jia Chen. Consistency training with learnable data augmentation for graph anomaly detection with limited supervision. In *The Twelfth International Conference on Learning Representations*, 2024a.

Nuo Chen, Ning Wu, Jianhui Chang, Linjun Shou, and Jia Li. ControlMath: Controllable data generation promotes math generalist models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 12201–12217, Miami, Florida, USA, November 2024b. Association for Computational Linguistics.

Wenhu Chen, Jianshu Chen, Yu Su, Zhiyu Chen, and William Yang Wang. Logical natural language generation from open-domain tables. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7929–7942, Online, July 2020. Association for Computational Linguistics.

Dawei Cheng, Yao Zou, Sheng Xiang, and Changjun Jiang. Graph neural networks for financial fraud detection: a review. *Frontiers of Computer Science*, 19(9):199609, 2025.

Asma Cherif, Arwa Badhib, Heyfa Ammar, Suhair Alshehri, Manal Kalkatawi, and Abdessamad Imine. Credit card fraud detection in the era of disruptive technologies: A systematic review. *Journal of King Saud University - Computer and Information Sciences*, 35(1):145–174, 2023. ISSN 1319-1578.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. Deepseek-v3 technical report, 2025.

Xiangyu Dong, Xingyi Zhang, Lei Chen, Mingxuan Yuan, and Sibo Wang. SpaceGNN: Multi-space graph neural network for node anomaly detection with extremely limited labels. In *The Thirteenth International Conference on Learning Representations*, 2025.

Yingtong Dou, Zhiwei Liu, Li Sun, Yutong Deng, Hao Peng, and Philip S. Yu. Enhancing graph neural network-based fraud detectors against camouflaged fraudsters. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, pp. 315–324, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450368599.

Yingtong Dou, Zhimeng Jiang, Tianyi Zhang, Mingzhi Hu, Zhichao Xu, Huiyuan Chen, Shubham Jain, Uday Singh Saini, Xiran Fan, Jiarui Sun, Menghai Pan, Junpeng Wang, Chin-Chia Michael Yeh, Xin Dai, and Yuzhong Chen. Transactiongpt: Toward foundational transaction modeling. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '26, pp. 7185–7196, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400722592.

Mingjiang Duan, Tongya Zheng, Yang Gao, Zunlei Feng, and Xinyu Wang. Dga-gnn: Dynamic grouping aggregation gnn for fraud detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

Mingjiang Duan, Da He, Tongya Zheng, Lingxiang Jia, Mingli Song, Xinyu Wang, and Zunlei Feng. Global attribute-association pattern aggregation for graph fraud detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(11):11616–11624, Apr. 2025.

Youssef Elmougy and Ling Liu. Demystifying fraudulent transactions and illicit nodes in the bitcoin network for financial forensics. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, pp. 3979–3990, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701030.

Xiran Fan, Zhimeng Jiang, Chin-Chia Michael Yeh, Yuzhong Chen, Yingtong Dou, Menghai Pan, and Yan Zheng. Enhancing foundation models in transaction understanding with LLM-based sentence embeddings. In Saloni Potdar, Lina Rojas-Barahona, and Sebastien Montella (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 903–911, Suzhou (China), November 2025. Association for Computational Linguistics. ISBN 979-8-89176-333-3.

Zheng Gong, Guifeng Wang, Ying Sun, Qi Liu, Yuting Ning, Hui Xiong, and Jingyu Peng. Beyond homophily: robust graph anomaly detection via neural sparsification. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, IJCAI '23, 2023. ISBN 978-1-956792-03-4.

Prince Grover, Julia Xu, Justin Tittelfitz, Anqi Cheng, Zheng Li, Jakub Zablocki, Jianbo Liu, and Hao Zhou. Fraud dataset benchmark and applications, 2023.

William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 1025–1035, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Li Han, Longxun Wang, Ziyang Cheng, Bo Wang, Guang Yang, Dawei Cheng, and Xuemin Lin. Mitigating the tail effect in fraud detection by community enhanced multi-relation graph neural networks. *IEEE Transactions on Knowledge and Data Engineering*, 37(4):2029–2041, 2025.

Xuanwen Huang, Yang Yang, Yang Wang, Chunping Wang, Zhisheng Zhang, Jiarong Xu, Lei Chen, and Michalis Vazirgiannis. Dgraph: A large-scale financial dataset for graph anomaly detection. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 22765–22777. Curran Associates, Inc., 2022.

Mohammed Saidul Islam, Md Tahmid Rahman Laskar, Md Rizwan Parvez, Enamul Hoque, and Shafiq Joty. DataNarrative: Automated data-driven storytelling with visualizations and texts. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 19253–19286, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

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

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Fan Li, Xiaoyang Wang, Dawei Cheng, Cong Chen, Ying Zhang, and Xuemin Lin. Efficient dynamic attributed graph generation. In *2025 IEEE 41st International Conference on Data Engineering (ICDE)*, pp. 1415–1428. IEEE, 2025a.

Guilin Li, Yun Zhang, Xiuyuan Chen, Chengqi Li, Bo Wang, Linghe Kong, Wenjia Wang, Weiran Huang, and Matthias Hwai Yong Tan. PANTHER: Generative pretraining beyond language for sequential user behavior modeling. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b.

Ao Liu, Haoyu Dong, Naoaki Okazaki, Shi Han, and Dongmei Zhang. PLOG: Table-to-logic pretraining for logical table-to-text generation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 5531–5546, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

Yang Liu, Xiang Ao, Zidi Qin, Jianfeng Chi, Jinghua Feng, Hao Yang, and Qing He. Pick and choose: A gnn-based imbalanced learning approach for fraud detection. In *Proceedings of the Web Conference 2021*, WWW '21, pp. 3168–3177, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383127.

Xiaoxiao Ma, Jia Wu, Shan Xue, Jian Yang, Chuan Zhou, Quan Z. Sheng, Hui Xiong, and Leman Akoglu. A comprehensive survey on graph anomaly detection with deep learning. *IEEE Transactions on Knowledge and Data Engineering*, 35(12):12012–12038, 2023.

Jianmo Ni, Jiacheng Li, and Julian McAuley. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 188–197, Hong Kong, China, November 2019. Association for Computational Linguistics.

Berkan Oztas, Deniz Cetinkaya, Festus Adedoyin, Marcin Budka, Huseyin Dogan, and Gokhan Aksu. Enhancing anti-money laundering: Development of a synthetic transaction monitoring dataset. In *2023 IEEE International Conference on e-Business Engineering (ICEBE)*, pp. 47–54, 2023.

Inkit Padhi, Yair Schiff, Igor Melnyk, Mattia Rigotti, Youssef Mroueh, Pierre Dognin, Jerret Ross, Ravi Nair, and Erik Altman. Tabular transformers for modeling multivariate time series. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3565–3569, 2021.

Andrea Dal Pozzolo, Olivier Caelen, Reid A. Johnson, and Gianluca Bontempi. Calibrating probability with undersampling for unbalanced classification. In *2015 IEEE Symposium Series on Computational Intelligence*, pp. 159–166, 2015.

Shebuti Rayana and Leman Akoglu. Collective opinion spam detection: Bridging review networks and metadata. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pp. 985–994, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450336642.

Kartik Shenoy. Credit Card Transactions Fraud Detection Dataset. Kaggle Dataset, 2020. URL `https://www.kaggle.com/datasets/kartik2112/fraud-detection`.

Fengzhao Shi, Yanan Cao, Yanmin Shang, Yuchen Zhou, Chuan Zhou, and Jia Wu. H2-fdetector: A gnn-based fraud detector with homophilic and heterophilic connections. In *Proceedings of the ACM Web Conference 2022*, WWW '22, pp. 1486–1494, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450390965.

Jianheng Tang, Jiajin Li, Ziqi Gao, and Jia Li. Rethinking graph neural networks for anomaly detection. In *International Conference on Machine Learning*, 2022.

Che-Ping Tsai, Ganyu Teng, Phillip Wallis, and Wei Ding. AnoLLM: Large language models for tabular anomaly detection. In *The Thirteenth International Conference on Learning Representations*, 2025.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.

Vesta Corporation. IEEE-CIS Fraud Detection. Kaggle Competition, 2019. URL https://www.kaggle.com/c/ieee-fraud-detection.

Binh Vu. Fraud Ecommerce. Kaggle Dataset, 2018. URL https://www.kaggle.com/datasets/vbinh002/fraud-ecommerce.

Yuchen Wang, Jinghui Zhang, Zhengjie Huang, Weibin Li, Shikun Feng, Ziheng Ma, Yu Sun, Dianhai Yu, Fang Dong, Jiahui Jin, et al. Label information enhanced fraud detection against low homophily in graphs. In *Proceedings of the ACM Web Conference 2023*, pp. 406–416, 2023.

Chunyu Wei, Haozhe Lin, Yueguo Chen, and Yunhai Wang. Conditional diffusion anomaly modeling on graphs. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Bin Wu, Xinyu Yao, Boyan Zhang, Kuo-Ming Chao, and Yinsheng Li. Splitgnn: Spectral graph neural network for fraud detection against heterophily. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, 2023.

Sheng Xiang, Mingzhi Zhu, Dawei Cheng, Enxia Li, Ruihui Zhao, Yi Ouyang, Ling Chen, and Yefeng Zheng. Semi-supervised credit card fraud detection via attribute-driven graph representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 14557–14565, 2023.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Chin-Chia Michael Yeh, Uday Singh Saini, Xin Dai, Xiran Fan, Shubham Jain, Yujie Fan, Jiarui Sun, Junpeng Wang, Menghai Pan, Yingtong Dou, Yuzhong Chen, Vineeth Rakesh, Liang Wang, Yan Zheng, and Mahashweta Das. Treasure: A transformer-based foundation model for high-volume transaction understanding. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*, KDD '26, pp. 2518–2527, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400722585.

Sanghyu Yoon, Dongmin Kim, Suhee Yoon, Ye Seul Sim, Seungdong Yoa, Hye-Seung Cho, Soonyoung Lee, Hankook Lee, and Woohyung Lim. RetabAD: A benchmark for restoring semantic context in tabular anomaly detection. In *The Fourteenth International Conference on Learning Representations*, 2026.

Dan Zhang, Yoshihiko Suhara, Jinfeng Li, Madelon Hulsebos, Çağatay Demiralp, and Wang-Chiew Tan. Sato: Contextual semantic type detection in tables. *Proc. VLDB Endow.*, 13(12):1835–1848, 2020.

Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. MM-LLMs: Recent advances in MultiModal large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 12401–12430, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

Guangjun Zhang, Hu Zhang, Yazhou Han, Yue Fan, Yuhang Shao, Hongye Tan, and Ru Li. Learning to generate and extract: A multi-agent collaboration framework for zero-shot document-level event arguments extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 34665–34673, 2026a.

Siwei Zhang, Yun Xiong, Xi Chen, Yateng Tang, Zi'an Jia, Xuehao Zheng, and Jiarong Xu. Think-like-lstm: Memory-augmented large language models via dynamic fine-tuning for financial risk assessment. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '26, pp. 8510–8521, New York, NY, USA, 2026b. Association for Computing Machinery. ISBN 9798400722592.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025a.

Zhi Zhang, Yan Liu, Sheng-hua Zhong, Gong Chen, Yu Yang, and Jiannong Cao. Mixture of knowledge minigraph agents for literature review generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 26012–26020, 2025b.

Zhongjian Zhang, Mengmei Zhang, Dehua Xu, Rongjun Shi, Jianfeng Liu, Fuli Meng, Huajian Xu, Xiao Wang, Ruijia Wang, Junze Chen, Minwei Tang, and Chuan Shi. Friskgpt: A generative foundation model for financial risk detection. In *Proceedings of the ACM Web Conference 2026*, WWW '26, pp. 7733–7744, New York, NY, USA, 2026c. Association for Computing Machinery. ISBN 9798400723070.

Zilong Zhao, Robert Birke, and Lydia Y Chen. Tabula: Harnessing language models for tabular data synthesis. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 247–259. Springer, 2025.

Wei Zhuo, Zemin Liu, Bryan Hooi, Bingsheng He, Guang Tan, Rizal Fathony, and Jia Chen. Partitioning message passing for graph fraud detection. In *The Twelfth International Conference on Learning Representations*, 2024.

A External Knowledge and Semantic Priors

The semantic enrichment framework relies on external knowledge to ensure that generated attributes are not only semantically plausible but also statistically grounded and reproducible. We therefore prioritize authoritative statistical resources and domain-specific documentation that provide explicit distributions, category systems, or behavioral reference points relevant to the generated semantics. These resources are selected according to four considerations: coverage of the required semantic attributes, sufficient granularity for regional or category-level assignment, geographic alignment with the original data context of S-FFSD, and temporal alignment with the observation period of MS-FFSD. Geographic alignment ensures that the semantic priors reflect the same underlying population and market context as the transaction data. Since the transaction records span January to October 2021, we prioritize authoritative resources that are temporally close to the observation period and use the nearest available official statistics for each semantic component, rather than simply adopting the most recent data. This reduces temporal mismatch between the external priors and the transaction behavior that grounds semantic enrichment. The selected knowledge is subsequently transformed into structured distributions, constraints, or semantic priors before being incorporated into the corresponding enrichment components.

The selected resources serve distinct roles in enrichment. For temporal semantics, *IEEE-CIS* provides a public reference for coarse temporal rhythms. The *2020 Population Census* provides the primary regional population and demographic statistics for geographic assignment and user attributes such as gender, age, and education. The *Statistical Yearbook 2021* supplements these with consumption expenditure, employment structure, and detailed age statistics for consumption, occupation, and age-related priors. On the merchant side, the *Economic Census Yearbook 2023* informs merchant-count priors, while the Alipay merchant category documentation provides the hierarchical taxonomy for merchant categories. The *2021 Economic and Social Development Statistical Bulletin* and the *PwC 2021 Global Consumer Insights Survey* support category-level merchant profile priors. The online-tendency component is additionally calibrated using aggregate online-retail statistics reported in the *2024 Economic and Social Development Statistical Bulletin* as a supplementary reference. Together, these sources provide complementary statistical, categorical, and behavioral priors at the granularity required by different enrichment components.

The primary external resources used in the framework are listed below:

- *IEEE-CIS*: `https://www.kaggle.com/c/ieee-fraud-detection/`
- *2020 Population Census*: `https://www.stats.gov.cn/sj/pcsj/rkpc/d7c/202111/P020211126523667366751.pdf`
- *Statistical Yearbook 2021*: `https://www.stats.gov.cn/sj/ndsj/2021/indexch.htm`
- *Economic Census Yearbook 2023*: `https://www.stats.gov.cn/sj/pcsj/jjpc/5jp/zk/indexch.htm`
- *Alipay Merchant Category Documentation*: `https://opendocs.alipay.com/solution/0df0ir?pathHash=aea1536c`
- *2021 Economic and Social Development Statistical Bulletin*: `https://www.stats.gov.cn/xxgk/sjfb/zxfb2020/202202/t20220228_1827971.html`
- *PwC 2021 Global Consumer Insights Survey*: `https://runwise.co/wp-content/uploads/2021/12/%E6%B6%88%E8%B4%B9%E8%80%85%E6%B4%9E%E5%AF%9F%E6%8A%A5%E5%91%8A.pdf`

The external resources are used at different levels of granularity according to the semantic component being constructed. For user-related semantics, geographic assignment is constrained by regional population and consumption statistics, while gender, age group, education, and occupation industry are generated from demographic and employment distributions at the corresponding regional or attribute-specific levels. In particular, age-group assignment combines gender-specific fine-grained age statistics with regional broad-age distributions to construct a region-adjusted categorical prior.

For merchant-related semantics, consumption statistics provide the basis for broad consumption-category allocation, while business census information supplies an additional prior on the relative

prevalence of merchant activities. The Alipay merchant category documentation defines the hierarchical candidate space for MCC Level 1 and MCC Level 2 assignment. Behavioral references from statistical bulletins and consumer research are further transformed into category-level merchant profile priors, and more recent online retail statistics are used to calibrate online tendency. Table 5 summarizes the primary external knowledge used in each component.

Table 5: Primary external knowledge sources and their roles in semantic enrichment.

Component	**External Knowledge**	**Granularity**	**Usage**
Temporal Priors			
Temporal Rhythm	IEEE-CIS Fraud Detection	Coarse Temporal Activity Pattern	Temporal rhythm reference
User-related Priors			
Geographic Assignment	2020 Population Census	Region	User-count distribution
Geographic Assignment	Statistical Yearbook 2021	Region	Transaction-amount distribution
Gender	2020 Population Census	Region $\times$ Gender	Region-specific sampling
Age Group	Statistical Yearbook 2021 + 2020 Population Census	Gender $\times$ 5-year Age Group; Region $\times$ Broad Age Group	Base age distribution and regional aging adjustment
Education	2020 Population Census	Region $\times$ Education	Region-specific sampling
Occupation Industry	Statistical Yearbook 2021	Region $\times$ Industry	Region-specific sampling
Merchant-related Priors			
Consumption Category	Statistical Yearbook 2021	Consumption Category	Merchant amount prior
Merchant Distribution	Economic Census Yearbook 2023	Industry	Merchant count prior
Merchant Category	Alipay Merchant Category Documentation	MCC Level 1 $\times$ MCC Level 2	Merchant category space
Merchant Profile	2021 Statistical Bulletin + PwC 2021 Survey	Category $\times$ Behavioral Attribute	MCC Level 2 profile priors

B Mathematical Details of the Semantic Enrichment Framework

This section provides the detailed mathematical formulation of the multi-agent semantic enrichment framework, including transaction representation, agent dependencies, agent-specific objectives, constrained refinement, and textual generation.

B.1 Problem Definition and Agent Workflow

Let the anonymized transaction set be

$$\mathcal{D} = \{x_n\}_{n=1}^{N}, \qquad x_n = (o_n, u_n, m_n, a_n, l_n, c_n, y_n), \tag{6}$$

where o_n denotes the original transaction order, $u_n \in \mathcal{U}$ and $m_n \in \mathcal{M}$ denote the associated user and merchant, a_n is the transaction amount, l_n and c_n are the anonymized location and transaction type, and y_n is the original fraud label. The enrichment process introduces temporal, user, merchant, and textual semantics without overwriting these original fields.

The agent-level workflow is represented as a directed graph

$$\mathcal{G} = (\mathcal{V}, \mathcal{E}), \qquad \mathcal{V} = \{A_T, A_U, A_M, A_C, A_X\}, \tag{7}$$

where A_T, A_U, A_M, A_C, and A_X denote the Temporal, User, Merchant, Consistency, and Textual Agents, respectively. The dependency edges are

$$\begin{aligned}\mathcal{E} = \{&(A_T, A_M), (A_U, A_M), (A_T, A_C), (A_U, A_C),\\ &(A_M, A_C), (A_T, A_X), (A_C, A_X)\}.\end{aligned} \tag{8}$$

An edge $(A_i, A_j) \in \mathcal{E}$ indicates that the state or statistics produced by A_i are used by A_j. Temporal and user states support merchant construction; temporal, user, and merchant states support cross-entity refinement; and the resulting structured states are subsequently passed to the Textual Agent.

To formalize structure preservation, let $\mathcal{D}^+$ denote the enriched dataset and Π_{orig} the projection onto the original transaction fields. Semantic enrichment is required to satisfy

$$\Pi_{\text{orig}}(\mathcal{D}^+) = \mathcal{D}. \tag{9}$$

Thus, the transaction order, participating entities, amounts, original anonymized attributes, and fraud labels remain unchanged throughout enrichment. The components summarized in Eq. 1 are realized through staged construction and constrained refinement following the workflow above, after which the Textual Agent performs the generation mapping defined in Eq. 5.

B.2 Temporal Semantic Construction

The Temporal Agent constructs synthetic timestamps using coarse periodic priors derived from IEEE-CIS while preserving the original transaction order. Since IEEE-CIS provides relative timestamps rather than absolute calendar dates, they are mapped to a fixed pseudo-calendar solely for estimating periodic temporal statistics.

The daily phase is first calibrated by aligning the lowest-activity window with a predefined early-morning interval. For each candidate hourly shift δ, let $C_h^{(\delta)}$ denote the transaction count at hour h after applying the shift, and define

$$s_\delta = \arg\min_s \sum_{j=0}^{L-1} C^{(\delta)}_{(s+j) \bmod 24}. \tag{10}$$

The selected shift is

$$\delta^\star = \arg\min_\delta \left[\text{cdist}(s_\delta, h_0), \sum_{j=0}^{L-1} C^{(\delta)}_{(s_\delta+j) \bmod 24} \right]_{\text{lex}}, \tag{11}$$

where L is the low-activity window length, h_0 its target starting hour, cdist denotes circular distance, and $[\cdot]_{\text{lex}}$ denotes lexicographic minimization. The shifted relative timestamps are then anchored to a fixed reference date to obtain a pseudo-calendar representation.

From this representation, the agent estimates a half-hour-of-week distribution $p^{\text{ref}}_{\text{how}}$, a day-of-month distribution $p^{\text{ref}}_{\text{dom}}$, and a fraud-specific half-hour-of-week distribution $p^{\text{fraud}}_{\text{how}}$. Each distribution is obtained from smoothed empirical counts,

$$p(j) = \frac{C_j + \alpha}{\sum_k C_k + \alpha|\mathcal{B}_{\text{time}}|}, \tag{12}$$

where $\mathcal{B}_{\text{time}}$ denotes the corresponding temporal-bin space and α is the smoothing parameter.

For users with sufficient transaction evidence, source-specific rhythms are constructed by matching their behavioral profiles to reference pseudo users. Let $\mathcal{N}_K(u)$ denote the matched reference users and d_{uj} their profile distance. Their temporal distributions are aggregated as

$$\bar{p}_u^{\mathrm{ref}} = \sum_{j \in \mathcal{N}_K(u)} \beta_{uj} p_j^{\mathrm{ref}}, \qquad \beta_{uj} = \frac{\exp(-d_{uj})}{\sum_{k \in \mathcal{N}_K(u)} \exp(-d_{uk})}, \tag{13}$$

and combined with the global rhythm as

$$p_u^{\mathrm{src}} = \eta_u \, \bar{p}_u^{\mathrm{ref}} + (1 - \eta_u) \, p_{\mathrm{how}}^{\mathrm{ref}}, \tag{14}$$

where $\eta_u \in [0, 1]$ controls the contribution of the matched user-specific rhythm.

The target observation interval is partitioned into calendar bins. For a bin b, its global weight combines the half-hour-of-week and day-of-month priors:

$$w_{\mathrm{global}}(b) \propto p_{\mathrm{how}}^{\mathrm{ref}}(h(b)) \, p_{\mathrm{dom}}^{\mathrm{ref}}(d(b)) \, , \tag{15}$$

where $h(b)$ and $d(b)$ denote its half-hour-of-week and day-of-month indices. For each order-preserving transaction block B, the temporal distribution is defined as

$$p_B^{\mathrm{time}} = \begin{cases} \eta_f p_{\mathrm{how}}^{\mathrm{fraud}} + \eta_s p_u^{\mathrm{src}} + \eta_g p_{\mathrm{how}}^{\mathrm{ref}}, & B \text{ is fraud-related,} \\ \eta_m p_u^{\mathrm{src}} + (1 - \eta_m) p_{\mathrm{how}}^{\mathrm{ref}}, & u \text{ is matched,} \\ p_{\mathrm{how}}^{\mathrm{ref}}, & \text{otherwise,} \end{cases} \tag{16}$$

where $\eta_f, \eta_s, \eta_g \geq 0$, $\eta_f + \eta_s + \eta_g = 1$, and $\eta_m \in [0, 1]$. Candidate start bins are then sampled according to

$$w_B(b) \propto \eta_b \, w_{\mathrm{global}}(b) + (1 - \eta_b) \, p_B^{\mathrm{time}}(h(b)) \, , \tag{17}$$

where $\eta_b \in [0, 1]$ balances global calendar rhythms and block-specific temporal patterns.

Fraud-related blocks additionally reuse relative inter-transaction intervals extracted from reference fraud bursts when a suitable template is available; otherwise compact intervals are generated. The resulting intervals are fitted within the admissible calendar segment while maintaining non-decreasing transaction order. Distributional discrepancies used by subsequent components are measured using total variation distance,

$$D_{\mathrm{TV}}(p, q) = \frac{1}{2} \sum_v |p(v) - q(v)|. \tag{18}$$

Finally, the generated timestamps satisfy

$$\tau_{(1)} \leq \tau_{(2)} \leq \cdots \leq \tau_{(N)}, \qquad \tau_n \in [t_{\min}, t_{\max}], \tag{19}$$

where $(1), \ldots, (N)$ follows the ordering induced by the original order field o_n. Thus, temporal enrichment introduces calendar semantics without altering the original transaction sequence.

B.3 User Semantic Initialization

The User Agent initializes geographic and demographic semantics from observed user–location interactions and external priors. Let $B \in \mathbb{R}^{|\mathcal{U}| \times |\mathcal{L}|}$ denote the user–location interaction matrix, where

$$B_{u\ell} = \sum_{n=1}^{N} \mathbb{I}[u_n = u, \, l_n = \ell]. \tag{20}$$

Locations exhibiting similar user interaction patterns are identified using cosine similarity,

$$s(\ell, \ell') = \frac{B_{:\ell}^{\top} B_{:\ell'}}{\|B_{:\ell}\|_2 \|B_{:\ell'}\|_2}, \tag{21}$$

and thresholded similarities are used to form a set of location blocks $\mathcal{B}$. Let $\rho : \mathcal{B} \to \mathcal{R}$ denote the mapping from location blocks to geographic regions. The mapping jointly matches regional distributions measured by user count and transaction amount:

$$\rho^{\star} = \arg\min_{\rho} \lambda_N D_{\mathrm{TV}}\big(\widehat{p}_{\rho}^{N}, p_{\mathrm{ref}}^{N}\big) + \lambda_A D_{\mathrm{TV}}\big(\widehat{p}_{\rho}^{A}, p_{\mathrm{ref}}^{A}\big), \tag{22}$$

where $\widehat{p}_{\rho}^{N}$ and $\widehat{p}_{\rho}^{A}$ are the resulting geographic distributions by user count and transaction amount, respectively. These two terms correspond to g_n and g_a in Eq. 2.

After assigning the geographic attribute r_u^{prov}, demographic attributes are sampled conditionally as

$$\begin{aligned} r_u^{\mathrm{gender}} &\sim P(G \mid r_u^{\mathrm{prov}}), \\ r_u^{\mathrm{age}} &\sim P(A \mid r_u^{\mathrm{prov}}, r_u^{\mathrm{gender}}), \\ r_u^{\mathrm{edu}} &\sim P(E \mid r_u^{\mathrm{prov}}), \\ r_u^{\mathrm{occ}} &\sim P(O \mid r_u^{\mathrm{prov}}). \end{aligned} \tag{23}$$

Accordingly, the demographic component d is expressed through conditional distribution matching. Expanding the User Agent objective gives

$$\begin{aligned} F_U(\boldsymbol{r}) = \lambda_N D_{\mathrm{TV}}(\widehat{p}_{\mathrm{prov}}^{N}, p_{\mathrm{prov}}^{N,\mathrm{ref}}) + \lambda_A D_{\mathrm{TV}}(\widehat{p}_{\mathrm{prov}}^{A}, p_{\mathrm{prov}}^{A,\mathrm{ref}}) \\ + \sum_{f \in \mathcal{F}_U} \lambda_f \sum_{\xi \in \Xi_f} \pi_{\xi} D_{\mathrm{TV}}\big(\widehat{p}_f(\cdot \mid \xi), p_f^{\mathrm{ref}}(\cdot \mid \xi)\big), \end{aligned} \tag{24}$$

where $\mathcal{F}_U = \{\mathrm{gender}, \mathrm{age}, \mathrm{education}, \mathrm{occupation}\}$. For gender, education, and occupation, Ξ_f represents the geographic region used for conditioning, while age is conditioned jointly on geographic region and gender. π_{ξ} denotes the weight of context ξ. Thus, geographic semantics are grounded in user–location behavior and regional distributions, while demographic semantics preserve the corresponding conditional priors.

B.4 Merchant Semantic Initialization

The Merchant Agent constructs hierarchical business semantics by combining observed transaction behavior with merchant-side priors. For each merchant m, we define a behavioral profile

$$\phi_m = \big(n_m, \bar{a}_m, \mathrm{cv}_m, r_m^{\mathrm{cross}}, H_m, P_m^{\mathrm{demo}}\big), \tag{25}$$

where n_m is the transaction count, $\bar{a}_m$ the mean transaction amount, cv_m the amount coefficient of variation, r_m^{cross} the cross-region transaction ratio, H_m the dominant activity periods, and P_m^{demo} the demographic distribution of associated users.

Let $c_m \in \mathcal{C}_M$ denote the broad consumption category assigned to merchant m. Category initialization targets the reference merchant-count and transaction-amount distributions, summarized by

$$\mathcal{L}_{\mathrm{broad}}^{M}(\boldsymbol{c}) = \lambda_c D_{\mathrm{TV}}\big(\widehat{p}_c^{\mathrm{count}}, p_c^{\mathrm{count}}\big) + \lambda_a D_{\mathrm{TV}}\big(\widehat{p}_c^{\mathrm{amount}}, p_c^{\mathrm{amount}}\big). \tag{26}$$

The assignment is implemented greedily by tracking category-level count and amount quotas. The two terms correspond to the merchant-count and transaction-amount components c and a in Eq. 2. Within each assigned broad category, finer merchant categories are initialized from the admissible hierarchy using category-relative behavioral semantics and soft distributional balancing. Accordingly, letting $q_m = (c_m, k_m)$ denote the structured merchant state, the Merchant Agent objective is summarized as

$$F_M(\boldsymbol{q};\boldsymbol{r},\boldsymbol{\tau}) = \lambda_c D_{\mathrm{TV}}(\widehat{p}_c^{\,\mathrm{count}}, p_c^{\mathrm{count}}) + \lambda_a D_{\mathrm{TV}}(\widehat{p}_c^{\,\mathrm{amount}}, p_c^{\mathrm{amount}}) + \lambda_s \mathcal{L}_s^M(\boldsymbol{q}), \tag{27}$$

where $\mathcal{L}_s^M$ captures category-relative semantic compatibility and within-category balancing during hierarchical initialization, corresponding to the semantic component s in Eq. 2.

For subsequent cross-entity refinement, the compatibility between merchant m and candidate fine-grained category k is measured by

$$\begin{aligned} S_M(m,k) = \alpha_o\big(1 - \big|\widetilde{r}_m^{\mathrm{cross}} - q_k^{\mathrm{online}}\big|\big) + \alpha_a(1 - d_{\mathrm{ord}}(a_m, a_k)) \\ + \alpha_v(1 - d_{\mathrm{ord}}(v_m, v_k)) + \alpha_t J(H_m, H_k) \\ + \frac{\alpha_g}{|\mathcal{F}_D|} \sum_{f \in \mathcal{F}_D} \Big[1 - D_{\mathrm{TV}}(P_m^f, Q_k^f)\Big] + \alpha_s \log \omega_k. \end{aligned} \tag{28}$$

Here, $\widetilde{r}_m^{\mathrm{cross}}$ denotes the normalized cross-region tendency derived from r_m^{cross}, and q_k^{online} is the online tendency associated with category k. The ordinal variables a_m and v_m are derived from $\bar{a}_m$ and cv_m, respectively, while a_k and v_k denote the corresponding category-level references. The normalized ordinal distance is

$$d_{\mathrm{ord}}(x,y) = \frac{|\,\mathrm{rank}(x) - \mathrm{rank}(y)|}{|\mathcal{L}| - 1}, \tag{29}$$

where $\mathcal{L}$ is the ordered set of semantic levels, $J(\cdot,\cdot)$ denotes Jaccard similarity, and $\mathcal{F}_D$ denotes the demographic attributes used for merchant–customer consistency. P_m^f and Q_k^f denote the observed and reference customer distributions for attribute f, respectively, H_k denotes the reference active periods of category k, and ω_k is a balancing weight. This score is subsequently reused by the Consistency Agent to refine fine-grained merchant semantics.

B.5 Cross-Entity Semantic Refinement

The Consistency Agent refines initialized user and merchant semantics according to their observed interactions. The transaction records induce a bipartite interaction graph

$$\mathcal{G}_{UM} = (\mathcal{U} \cup \mathcal{M}, \mathcal{E}_{UM}), \qquad (u,m) \in \mathcal{E}_{UM} \iff \exists n:\ u_n = u,\ m_n = m, \tag{30}$$

where an edge indicates that at least one transaction is observed between user u and merchant m. This graph differs from the agent workflow graph $\mathcal{G}$: $\mathcal{G}_{UM}$ represents entity interactions, whereas $\mathcal{G}$ describes information dependencies among agents.

For a refinable user attribute $f \in \mathcal{F}_R = \{\mathrm{age}, \mathrm{education}, \mathrm{occupation}\}$ and candidate value v, the user-side consistency score is defined as

$$\begin{aligned} S_U(u,f,v) = \gamma_p P_{\mathcal{N}(u)}^f(v) + \gamma_m \left[1 - D_{\mathrm{TV}}(P_u^{\mathrm{MCC}}, Q_v^{\mathrm{MCC}})\right] \\ + \gamma_a C_a(u,v) + \gamma_t C_t(u,v) + \gamma_s \log \omega_{f,v}. \end{aligned} \tag{31}$$

Let n_{um} denote the number of transactions between user u and merchant m, and $P_m^f(v)$ the prevalence of attribute value v among users associated with merchant m. The merchant-mediated prevalence is

$$P_{\mathcal{N}(u)}^f(v) = \frac{\sum_{m \in \mathcal{N}(u)} n_{um} P_m^f(v)}{\sum_{m \in \mathcal{N}(u)} n_{um}}. \tag{32}$$

P_u^{MCC} denotes the merchant-category distribution observed for user u, and Q_v^{MCC} the corresponding reference distribution for candidate value v. Amount and temporal compatibility are defined as

$$C_a(u,v) = 1 - d_{\mathrm{ord}}(a_u, a_v^{\mathrm{ref}}), \qquad C_t(u,v) = J(H_u, H_v^{\mathrm{ref}}), \tag{33}$$

where a_u and H_u denote the user's ordinal amount level and dominant activity periods, and a_v^{ref} and H_v^{ref} are their reference counterparts associated with v. The balancing term $\omega_{f,v}$ discourages excessive concentration on particular semantic values.

Using S_U together with the merchant compatibility score S_M defined in Eq. 28, the two consistency components in Eq. 3 are

$$\mathcal{L}_U^C = -\sum_{u\in\mathcal{U}}\sum_{f\in\mathcal{F}_R} S_U(u,f,r_u^f), \qquad \mathcal{L}_M^C = -\sum_{m\in\mathcal{M}} S_M(m,k_m). \tag{34}$$

Thus, larger compatibility scores reduce the consistency objective, while the change penalty in Eq. 4 discourages unnecessary reassignment. User attributes are refined independently, and merchant candidates remain restricted to fine-grained categories within the parent category assigned during initialization.

At iteration t, the feasible candidate set for entity e is denoted by $\mathcal{A}_e^{(t)}$. Using the net gain in Eq. 4, the update rule is

$$z_e^{(t+1)} = \begin{cases} \arg\max\limits_{z\in\mathcal{A}_e^{(t)}} \Delta_e(z;Z^{(t)}), & \max\limits_{z\in\mathcal{A}_e^{(t)}} \Delta_e(z;Z^{(t)}) > 0, \\ z_e^{(t)}, & \text{otherwise.} \end{cases} \tag{35}$$

For constrained user attributes, the Constraint Engine adjusts candidate weights as

$$\omega_{f,v}^{(t)} = \mathrm{clip}\left(1+\eta\frac{p_{f,v}^{\mathrm{ref}}-\widehat{p}_{f,v}^{(t)}}{b_{f,v}}, \omega_{\min}, \omega_{\max}\right), \qquad b_{f,v} = \max\left(p_{f,v}^{\mathrm{ref}}, \frac{1}{|\mathcal{V}_f|}\right), \tag{36}$$

where $\mathcal{V}_f$ is the value space of attribute f. Underrepresented values therefore receive larger weights and overrepresented values receive smaller ones. Final user distributions satisfy

$$D_{\mathrm{TV}}\left(\widehat{p}_f, p_f^{\mathrm{ref}}\right) \le \epsilon_f, \qquad f\in\mathcal{F}_R. \tag{37}$$

Merchant-side refinement is additionally constrained by category-level structural limits on modification and semantic concentration.

Let $\Delta\mathcal{U}^{(t)}$ and $\Delta\mathcal{M}^{(t)}$ denote the user and merchant sets changed during the main update stage. Only directly affected counterparts are activated for response:

$$\begin{aligned} \mathcal{R}_M^{(t)} &= \{m\in\mathcal{M} : \exists u\in\Delta\mathcal{U}^{(t)}, (u,m)\in\mathcal{E}_{UM}\}, \\ \mathcal{R}_U^{(t)} &= \{u\in\mathcal{U} : \exists m\in\Delta\mathcal{M}^{(t)}, (u,m)\in\mathcal{E}_{UM}\}. \end{aligned} \tag{38}$$

Responses are therefore localized to entities whose connected counterparts have changed, rather than rescanning the full user or merchant population. Response updates are bounded and do not recursively trigger unbounded cascades.

Let $N_U^{(t)}$ and $N_M^{(t)}$ denote the numbers of accepted primary-stage user-field and merchant-category updates at iteration t, respectively, and let N_U and N_M denote the corresponding candidate-pool sizes. The primary-stage modification budgets satisfy

$$\begin{aligned} N_U^{(t)} &\le b_U^{\mathrm{round}} N_U, \quad \sum_t N_U^{(t)} \le b_U^{\mathrm{total}} N_U, \\ N_M^{(t)} &\le b_M^{\mathrm{round}} N_M, \quad \sum_t N_M^{(t)} \le b_M^{\mathrm{total}} N_M. \end{aligned} \tag{39}$$

The refinement terminates when no admissible positive-gain update remains, the modification budgets are exhausted, the accepted gains become sufficiently small for consecutive rounds, or the maximum

number of iterations is reached. If a final hard constraint is violated, eligible low-gain updates associated with the violated attribute or category are reverted, with the constraints re-evaluated after each repair step.

B.6 Textual Semantic Generation

The Textual Agent first constructs an evidence-grounded representation for each entity from its refined structured semantics and observed transaction context. For users and merchants, the corresponding inputs are defined as

$$\begin{aligned} h_u &= \mathrm{Serialize}(r_u, \mathrm{Agg}\{\chi(x_n) : u_n = u\}), \\ h_m &= \mathrm{Serialize}(q_m, \mathrm{Agg}\{\chi(x_n) : m_n = m\}, \mathrm{Agg}\{r_{u_n} : m_n = m\}), \end{aligned} \tag{40}$$

where $\chi(x_n)$ extracts transaction-level temporal, amount, geographic, and interaction statistics, Agg summarizes information associated with the entity, and $\mathrm{Serialize}$ converts the resulting structured state into the input representation used for text generation.

Let $n_u = \sum_{n=1}^{N} \mathbb{I}[u_n = u]$ denote the number of observed transactions associated with user u. User descriptions are generated according to the available transaction evidence:

$$d_u = \begin{cases} \Gamma(h_u), & n_u = 1, \\ \mathrm{LLM}(P_U(h_u; \mathcal{K})), & n_u > 1, \end{cases} \tag{41}$$

where Γ denotes the deterministic template used for single-transaction users, and P_U denotes the structured user prompt. The deterministic template verbalizes the available user attributes together with the geographic, temporal, and merchant-category context of the observed transaction. Single-transaction users are handled deterministically because one observation does not provide sufficient evidence for reliable behavioral aggregation. For multi-transaction users, aggregated behavioral patterns are combined with the refined user semantics and provided to the fixed language model.

Merchant descriptions are generated as

$$d_m = \mathrm{LLM}(P_M(h_m; \mathcal{K}, \mathcal{H}_{c_m})), \tag{42}$$

where P_M denotes the merchant prompt and $\mathcal{H}_{c_m}$ stores previously generated concepts within the same category to reduce repetition. Textual generation is conditioned on the refined semantic state and observed evidence without modifying the structured state. The complete prompts and deterministic templates are provided in Appendix C.

B.7 Shared Constraints and Control

The agents are coordinated through a shared external knowledge base, a global feasible region, and fraud-aware preservation rules. The external knowledge base $\mathcal{K}$ contains the temporal rhythms, demographic distributions, consumption structures, merchant category systems, and behavioral priors used by the corresponding agents. The specific knowledge sources, statistical granularity, temporal alignment, and their roles are detailed in Appendix A.

The Constraint Engine defines the feasible region $\mathcal{C}(\mathcal{D}, \mathcal{K})$ by intersecting several constraint sets:

$$\mathcal{C}(\mathcal{D}, \mathcal{K}) = \mathcal{C}_{\mathrm{orig}} \cap \mathcal{C}_{\mathrm{time}} \cap \mathcal{C}_{\mathrm{dist}} \cap \mathcal{C}_{\mathrm{hier}} \cap \mathcal{C}_{\mathrm{budget}}, \tag{43}$$

where $\mathcal{C}_{\mathrm{orig}}$ preserves the original transaction fields as specified in Eq. 9, $\mathcal{C}_{\mathrm{time}}$ enforces temporal ordering and valid timestamp ranges, $\mathcal{C}_{\mathrm{dist}}$ constrains deviations from reference distributions, $\mathcal{C}_{\mathrm{hier}}$ restricts semantic changes to valid category hierarchies, and $\mathcal{C}_{\mathrm{budget}}$ bounds the extent of semantic modification. Together, these constraint sets provide shared control across enrichment stages.

The Fraud-aware Preservation Controller further protects abnormal patterns supported by transaction evidence, including temporal bursts, geographic anomalies, and residual merchant abnormalities, from being removed solely to improve global semantic consistency. Together, the shared knowledge

Multi-transaction User Description Prompt

You are an assistant for semantic enrichment of financial transaction data. Based on the gender, transaction profile, and spending preferences of an anonymous user, generate a natural-language English description of the user's overall consumption behavior.
Construct a coherent and evidence-grounded description that combines explicit consumption interests with higher-level behavioral semantics. Do not mechanically translate the input profile into natural language.
[Input Information]
{basic_profile}
[Information Usage]
1. Spending Preferences: preferred_categories defines the user's main consumption categories. These categories may be explicitly mentioned and combined into a coherent consumption profile. Stay within the supplied category semantics and do not infer purposes, motivations, lifestyle, or personal roles.
2. Transaction Profile: transaction_profile is used only as reasoning evidence. Combine multiple behavioral signals to derive higher-level structural semantics without restating individual signals or inferring routines, financial conditions, lifestyle characteristics, or other real-life explanations.
3. Gender: gender is weak supplementary context only. Do not explicitly mention gender or use it independently to infer personal or behavioral characteristics.
[Generation Requirements]
1. Generate one English Source_description describing the user's overall consumption profile over full observation period.
2. Combine the supplied consumption categories and their overall composition with higher-level structural behavioral semantics derived from multiple transaction-profile signals.
3. Consumption categories may be stated directly, but transaction-profile signals must remain implicit and must not be restated or paraphrased.
4. State derived characteristics directly without describing the supporting evidence or reasoning. Do not infer purposes, motivations, personality, lifestyle, personal circumstances, financial conditions, or other real-life explanations.
5. Include only information supported by the supplied profile. Do not include demographic information, field names, profile labels, statistics, exact numerical values, risk-related information, or meta-analytical language.
6. Use natural, fluent, neutral, and objective English. Avoid generic, repetitive, analytical, or inferential phrasing while preserving meaningful distinctions among users.
[Output Format]
Output exactly the following line and nothing else:
Source_description: English user description

Figure 5: Prompt used for multi-transaction user description generation.

and constraints coordinate stage-specific updates while maintaining structural, distributional, and fraud-related preservation requirements.

Single-Transaction User Description Template

[Deterministic Template]
The user is {age_gender_description} from {province}, has {education_description}, and works in {occupation_industry}. The user made a {locality} transaction {period_description} with a merchant associated with {MCC_level2}. The transaction occurred {calendar_context}.

[Placeholder Definitions]
age_gender_description: Age-group description combined with gender.
province: User province.
education_description: Natural-language education description.
occupation_industry: User occupation industry.
locality: local / out-of-province.
period_description: Natural-language time-period phrase.
MCC_level2: Merchant Level-2 business category.
calendar_context: Day-of-week information with optional monthly-cycle and bill-payment context.

Figure 6: Deterministic template used for single-transaction user description generation.

Merchant Description Generation Prompt

You are an assistant for semantic enrichment of financial transaction data. Based on the merchant category, transaction profile, associated user profile, and existing merchant concepts of an anonymous merchant, generate an English merchant name and a natural-language merchant description.
Construct a specific, coherent, and restrained merchant instance rather than restating the input fields.
[Input Information]
{basic_profile}
[Existing Merchant Concepts in This MCC Level 2]
{existing_merchant_concepts}
[Information Usage]
1. Merchant Category: consumption_category, MCC_level1, MCC_level2, and fine_grained_candidates form a hierarchical category system. The merchant must remain within the supplied MCC_level2, which defines its business type and product or service scope. When fine_grained_candidates are available, select one coherent direction without combining unrelated candidates or leaving the category hierarchy.
2. Transaction Profile: use transaction_profile only to select a plausible product or service emphasis within the fixed category. Do not convert profile signals into unsupported operating details such as pricing, schedules, memberships, promotions, locations, or service procedures.
3. Associated User Profile: use associated_user_profile only as weak supplementary context for product or service emphasis. Treat demographic attributes as weak signals and occupation_top3 as the lowest-priority signal. Do not explicitly describe customer demographics, occupations, identities, or environments.
4. Existing Merchant Concepts: use existing_merchant_concepts only to reduce naming and semantic repetition. Avoid duplicating existing names or core concepts. When overlap exists, prefer a meaningfully different category-consistent product or service emphasis, combination, or business format. Do not introduce unsupported facts or leave the supplied MCC_level2 merely to create diversity.
Information priority:
merchant_category > transaction_profile > associated_user_profile > existing_merchant_concepts
[Generation Requirements]
1. Generate one English merchant name, Merchant, and one English merchant description, Merchant_description.
2. Merchant should be a natural fictional small- or medium-sized business name consistent with the business type. Do not duplicate or closely imitate existing names, and avoid real entities, specific locations, Chinese pinyin, and overly literary or promotional naming.
3. Merchant_description should directly describe the merchant through representative products or services, business format, service methods, consumption occasions, or consumer needs when appropriate. Do not explain the input evidence or invent unsupported details.
4. Do not include field names, profile labels, statistics, exact numerical values, analytical expressions, or risk- and label-related information. Use natural, fluent, neutral, and objective English.
5. Generate Memory_keywords as 3–5 concise English concept labels separated by semicolons. Capture distinctive products, services, combinations, or business formats for semantic-overlap detection; exclude merchant names, complete sentences, transaction patterns, demographics, locations, and generic category labels.
[Output Format]
Output exactly the following three lines and nothing else:
Merchant: English merchant name
Merchant_description: English merchant description
Memory_keywords: keyword one; keyword two; keyword three

Figure 7: Prompt used for merchant description generation.

C Textual Semantic Generation Prompts

C.1 User Description Generation

For users with multiple observed transactions, we first aggregate their transaction history into a structured behavioral profile containing transaction patterns and spending preferences. The generation input retains only the information required for textual construction, and the prompt instructs the language model to combine explicit consumption categories with higher-level behavioral semantics derived from multiple transaction signals. Individual transaction-profile fields are not directly reproduced in the description, and demographic information is not explicitly expressed. The complete prompt is shown in Figure 5.

Merchant Name Deduplication Prompt

You are a merchant naming assistant. The merchant below must be renamed because its current Merchant name is duplicated. Based on the original merchant name and merchant description, generate a new English merchant name that replaces only the Merchant field.

[Input Information]

Original Merchant Name:
{old_merchant}
Merchant Description:
{merchant_description}

[Requirements]

1. The new merchant name must be consistent with the business type, products or services, and overall semantics described in the merchant description.
2. The new name must be meaningfully different from the original. Do not create variants by merely adding numbers, letters, branch labels, location labels, or similar prefixes or suffixes.
3. Generate a natural English name suitable for a fictional small- or medium-sized merchant and appropriate to its specific business type.
4. Do not use Chinese characters, Chinese pinyin, real brands, platforms, institutions, chain businesses, or specific real-world place names.
5. Avoid overly literary, promotional, exaggerated, or obviously generated names, as well as excessive use of generic endings such as “House,” “Hub,” “Corner,” “Place,” “Club,” or “Lounge.”
6. Generate only the replacement merchant name. Do not modify the merchant description.

[Output Format]

Output exactly the following JSON object and nothing else:

```
{
  "new_merchant": "New English merchant name"
}
```

Figure 8: Prompt used for duplicate merchant name resolution.

For users with only one observed transaction, behavioral aggregation is not sufficiently supported by repeated evidence. We therefore use a deterministic template rather than LLM-based generation. The template verbalizes the transaction’s geographic, temporal, and merchant-category context, producing a concise description directly from the observed and enriched fields. The template and its placeholder definitions are shown in Figure 6.

The template verbalizes the transaction’s geographic, temporal, and merchant-category context without repeating demographic attributes already represented in the structured user profile.

C.2 Merchant Description Generation

Merchant descriptions are generated from the merchant category hierarchy, transaction profile, and aggregated profiles of associated users. The category hierarchy defines the admissible business scope, while transaction and user information provide supporting context for selecting a plausible product or service emphasis. The model is instructed to construct a specific and coherent merchant instance without directly reproducing profile fields or introducing unsupported operating details. The complete generation prompt is shown in Figure 7.

To reduce semantic repetition among merchants within the same MCC Level 2 category, the generation process maintains category-level memory. For each merchant, the model additionally generates a small set of *Memory_keywords* summarizing its distinctive products, services, combinations, or business format. Previously generated merchant names and core concepts are supplied to subsequent generations within the same category, encouraging alternative category-consistent concepts when substantial overlap occurs. This memory mechanism affects only textual construction and does not modify the structured merchant semantics.

After generation, duplicated merchant names are detected separately from semantic description construction. When a duplicate occurs, a dedicated renaming prompt takes the original merchant name and its generated description as input and produces a replacement name consistent with the

existing business semantics. Only the merchant name is changed; the merchant description and structured attributes remain unchanged. The corresponding prompt is shown in Figure 8.

D MS-FFSD DATASET DETAILS

Dataset Fields. MS-FFSD organizes the released data into transaction, user, and merchant components. The transaction component records preserved transaction information together with enriched temporal and geographic semantics. The user and merchant components contain their respective structured semantic attributes and textual descriptions, linked to transactions through anonymized identifiers. Table 6 provides the complete field definitions.

Table 6: Field definitions of the released MS-FFSD dataset.

Transaction	
Datetime	Synthesized physical date and time constructed using IEEE-CIS-based temporal priors.
Source	User identifier and foreign key to the user table.
Target	Merchant identifier and foreign key to the merchant table.
Amount	Original transaction amount preserved during semantic enrichment.
Location	Original transaction-location identifier.
Type	Original transaction-type identifier.
Transaction_province	Assigned physical province of the transaction.
Labels	Transaction label indicating normal, fraudulent, or unlabeled status.
User	
Source	User identifier and primary key.
Province	Assigned home province of the user.
Gender	User gender category.
Age_group	User age-group category.
Education	User education-level category.
Occupation_industry	Industry associated with the user's occupation.
User_description	Textual description of the user's consumption behavior and semantic profile.
Merchant	
Target	Merchant identifier and primary key.
MCC_level1	Broad merchant consumption and business category.
MCC_level2	Fine-grained merchant category within MCC_level1.
Merchant_name	Synthetic English merchant name.
Merchant_description	Textual description of the merchant's business characteristics and behavioral context.

User Semantic Attributes. MS-FFSD enriches users with demographic, geographic, and occupational semantics. The province attribute follows the administrative divisions adopted in the dataset, while gender and education are sampled according to province-specific demographic distributions. Since transaction users are expected to have independent payment capability, individuals below 15 years old are excluded from age assignment. Age is represented as a categorical semantic attribute rather than a numerical value, with users grouped into Young Adults (15–44), Middle-aged Adults (45–59), and Older Adults ($\geq$60). For each gender, national age statistics are first aggregated into the three age groups to obtain the base proportions. The proportion of Older Adults is then adjusted for each province according to its relative 65+ population level, while the remaining probability is divided between Young Adults and Middle-aged Adults according to their gender-specific national proportions. The aging adjustment is left unbounded to retain regional differences. Occupation industry is represented using 19 industry categories and sampled according to province-specific employment distributions. The resulting categorical value spaces are summarized in Table 7.

Merchant Semantic Categories. Merchant semantics are organized into a hierarchical category system to represent business activities at different levels of granularity. Each merchant is first associated with one of eight broad consumption categories and is then assigned an MCC Level 1 category describing its general business function, followed by a more specific MCC Level 2 category. This hierarchical design allows merchants within the same broad consumption domain to be further

Table 7: Categorical definitions of user-level semantic attributes in MS-FFSD.

Attribute	Categories
Gender	Male; Female.
Age Group	Young Adults (15–44); Middle-aged Adults (45–59); Older Adults ($\geq$60).
Education	Junior College and Above; Senior Secondary School; Junior Secondary School; Primary School.
Occupation Industry	Agriculture, Forestry, Animal Husbandry and Fishery; Mining; Manufacturing; Production and Supply of Electricity, Heat, Gas and Water; Construction; Wholesale and Retail Trade; Transport, Storage and Postal Services; Accommodation and Food Services; Information Transmission, Software and Information Technology Services; Financial Services; Real Estate; Leasing and Business Services; Scientific Research and Technical Services; Water Conservancy, Environment and Public Facilities Management; Resident Services, Repair and Other Services; Education; Health and Social Work; Culture, Sports and Entertainment; Public Administration, Social Security and Social Organizations.

distinguished according to their business characteristics while maintaining a consistent category structure. The complete category hierarchy used in MS-FFSD is summarized in Table 8.

Table 8: Hierarchical definitions of merchant-level semantic categories in MS-FFSD.

MCC Level 1	MCC Level 2
Food, Tobacco and Liquor	
Food, Tobacco and Liquor Retail	Convenience Store; Supermarket and Hypermarket; Fresh Produce Retail; Specialty Food Retail; Tobacco, Liquor and Tea Retail.
Food and Beverage Services	Fast Food and Snacks; Beverages, Bakery and Coffee; Cafeteria and Group Dining; Full-Service Restaurant; Nightlife Bar and Dining; Banquet and Large-Scale Catering.
Clothing and Footwear	
Apparel and General Shopping	Apparel, Footwear and Bags; Department Store; Shopping Mall; Outlet Shopping; Commercial Street Retail.
Housing	
Housing Payments and Home Improvement Services	Utility Bill Payments; Property Management; Real Estate and Housing Agency; Hardware and Building Materials; Home Furnishings and Textiles; Building Decoration and Renovation Services; Appliance, Furniture and Other Repair Services.
Household Equipment, Furnishings and Services	
Household and Personal Goods Retail	Beauty and Personal Care; Mother and Baby Products and Children's Toys; Office Supplies; Flowers and Green Plants; Pets and Pet Supplies; Digital Electronics and Home Appliances.
Local Lifestyle and Self-Service Sharing	Hair, Beauty and Nail Services; Bathing, Wellness and Health Services; Housekeeping and Cleaning Services; Pet Hospital and Other Pet Services; Digital and Entertainment Equipment Rental; Self-Service Sharing Services.
Transport and Communications	
Local Commuting and On-Demand Travel	Urban Public Transport; Shared Two-Wheel Mobility Services; On-Demand Ride-Hailing; Designated Driving and Parking Services.
Long-Distance Travel Transport	Long-Distance Ground Travel; Air Travel Services; Water Sightseeing Transport; Car Rental; Expressway Toll and Service Areas.

MCC Level 1	MCC Level 2
Vehicle Energy and Automotive Services	Fuel Supply Services; New Energy Charging and Battery Swap Services; Vehicle Supplies, Repair and Maintenance.
Telecommunications, Logistics and Digital Infrastructure Services	Telecom Payment and Operator Services; Information Search and Online Forums; Internet Data Services; Software Development Services; Express Delivery and Logistics Services; Basic Postal Services.
Education, Culture and Recreation	
Education Payments and Training Services	Preschool Care Services; Primary and Secondary Education; Higher Education; Youth Palace and Development Center; Other Education and Training.
Offline Culture, Sports and Entertainment	KTV and Leisure Clubs; Board Games, Table Games and Gaming Cafes; Fitness, Yoga and Dance; Cinema, Performances and Events; Amusement Parks and Carnivals; Cultural and Sports Venues; Books, Media, Arts and Musical Instruments; Outdoor Sports Equipment.
Online Entertainment and Hobby Spending	Online Social Networking; Online Video, Audio and Reading; Gaming; Live Streaming; Gaming Transactions and Related Services.
Hotels, Scenic Attractions and Tourism Services	Hotels, Inns and Homestays; Scenic Attractions; Travel Agencies and Tourism Services; Direct Tourism-Related Services; Scenic Attraction Touring Services.
Health Care and Medical Services	
Pharmaceutical, Medical Device and Health Retail	Pharmaceutical Sales; Nutrition and Health Products; Medical Device Sales; Health and Assistive Therapy Equipment; Optical Store.
General Clinical Care and Health Management	Public Primary Medical Care; Private Hospital; Clinics and Independent Practitioners; Medical Laboratories and Diagnostic Centers; Health Examinations and Consultations.
Specialized Consumer Medical and Care Services	Medical Aesthetics; Ophthalmic Medical Services; Dental Medical Services; Online Medical Services; Care Institution Services.
Miscellaneous Goods and Services	
High-Value Goods and Professional Services	High-Value Jewelry and Accessories; Wedding and Photography Services; Advertising, Exhibition and Printing Services; Legal Services; Accounting and Financial Consulting Services; Recruitment Services; Online Tools.

Textual Descriptions. The textual modality provides one entity-level description for each user and merchant. Financial transaction activity typically exhibits a long-tailed distribution, where a substantial fraction of users have only limited transaction histories (Han et al., 2025). MS-FFSD exhibits the same characteristic: as shown in Table 9, 69.56% of users have only one observed transaction. This imbalance in behavioral evidence motivates us to construct descriptions separately for single-transaction and multi-transaction users. For single-transaction users, a single observation is insufficient to support reliable aggregation of behavioral patterns, and forcing such aggregation may introduce unstable or weakly supported profile information. Their descriptions are therefore constructed deterministically from the available transaction and entity attributes. For multi-transaction users, repeated observations provide sufficient behavioral evidence to derive aggregated behavioral profiles and generate richer entity-level descriptions. Merchant descriptions are constructed at the entity level from merchant category semantics, transaction behavior, and associated user profiles.

Despite the different construction strategies, the resulting descriptions exhibit highly similar average lengths. Single-transaction and multi-transaction user descriptions contain 46.38 and 44.51 words on average, respectively, while the overall user average is 45.81 words. Merchant descriptions have a comparable average length of 46.94 words. This consistency indicates that neither the user-generation strategy nor the entity type introduces a substantial length discrepancy, reducing the likelihood that description length itself serves as a trivial shortcut or unintended signal in downstream modeling. The corresponding group sizes and description-length statistics are summarized in Table 9.

Table 9: Statistics of textual descriptions in MS-FFSD.

Description Group	Number	Proportion	Avg. Length
User Descriptions			
All Users	30,346	–	45.81 words
Single-transaction Users	21,110	69.56%	46.38 words
Multi-transaction Users	9,236	30.44%	44.51 words
Merchant Descriptions			
All Merchants	886	–	46.94 words

E Private Real-World Dataset Details

Private-1 is a large-scale real-world financial transaction dataset containing diverse transaction patterns and substantial behavioral heterogeneity. Each transaction is associated with a label indicating whether it is normal, fraudulent, or unlabeled. The dataset records transaction order, interacting entities, transaction amount, and anonymized transaction context. Detailed definitions of the transaction attributes are provided in Table 10.

Table 10: Descriptions of transaction fields in the Private-1 dataset.

Field	Description
Time	Global transaction order indicating the relative sequence of transactions
Source	Anonymized identifier of the user
Target	Anonymized identifier of the merchant
Amount	Transaction amount
Location	Anonymized location associated with the transaction
Type	Anonymized transaction type
Label	Transaction label indicating normal, fraudulent, or unlabeled status

Private-2 is collected from a major commercial bank and consists of customer transaction records spanning four consecutive months from January to April. The records are organized into four monthly subsets in chronological order. Each transaction is associated with a label indicating whether it is normal, fraudulent, or unlabeled. Each record contains customer and merchant identifiers together with geographic, temporal, payment, and transaction attributes. Detailed definitions of all transaction fields are provided in Table 11.

Table 11: Descriptions of transaction fields in the Private-2 dataset.

Field	Description
customer_number	Anonymized identifier of the transaction customer
merchant_code	Anonymized identifier of the merchant
receiving_customer_code	Anonymized identifier of the merchant receiving the customer's transaction
card_area	Geographical region associated with the card
pre_trade_result	Outcome of the previous transaction attempt
phone_equal	Whether the transaction phone matches the registered phone
trade_time	Time of the transaction
is_common_ip	Whether the IP address is commonly used
e_pay_single_limit	Limit for a single electronic payment
e_pay_accumulate_limit	Cumulative electronic payment limit over time
trade_amount	Transaction amount
Label	Transaction label indicating normal, fraudulent, or unlabeled status

The two private datasets differ substantially in both temporal structure and user transaction-frequency distribution. First, their original temporal fields encode transaction order differently. In Private-1, the temporal index is globally unique and strictly increasing, providing a complete ordering of all transactions. In Private-2, the temporal index is non-decreasing but may contain duplicate values,

meaning that multiple transactions can share the same index and their relative order cannot be further distinguished. Private-2 additionally provides explicit monthly boundaries. The Temporal Agent therefore adapts its timestamp construction to the available ordering information: Private-1 preserves the unique global order, whereas transactions sharing the same index in Private-2 are handled within the same temporal bucket without imposing an unsupported intra-bucket ordering.

For graph construction and behavioral feature extraction, the dataset-specific schema is mapped to the unified representation used in Appendix F. In Private-2, customer_number and merchant_code are used as the user and merchant identifiers, respectively. Card_area is used as the geographic attribute corresponding to Location, while the dataset-specific transaction-context attributes are encoded into a unified Type representation. The resulting Location and Type representations are then used to construct the corresponding structural relations and the same 12-dimensional behavioral features as in the other datasets.

The two datasets also exhibit markedly different user activity distributions. Private-1 shows a pronounced long-tailed pattern, with most users associated with only a single observed transaction, whereas single-transaction users constitute only a minority in Private-2. This difference reflects substantially different levels of transaction-history sparsity and user behavioral coverage, and may also arise from different data collection or preprocessing regimes. These differences provide distinct conditions for evaluating the semantic enrichment framework. The framework is applied to both datasets without relying on a single temporal representation or user-activity distribution, and the resulting multimodal data improve downstream fraud detection in both settings. The consistent gains across these different data conditions therefore provide evidence for the generalizability of the proposed semantic enrichment framework.

F Graph-based Fraud Detection Details

This section provides implementation details for graph-based fraud detection, covering graph construction, semantic text construction, feature encoding, and experimental settings.

F.1 Graph Construction

We construct an entity graph in which users and merchants are represented as nodes, rather than treating individual transactions as graph nodes. This design is motivated by the granularity of the generated semantics. User descriptions characterize demographic profiles, consumption preferences, and behavioral patterns over the observation period, while merchant descriptions capture business categories, service characteristics, and customer profiles. These semantics therefore correspond naturally to persistent entities rather than individual transactions.

A transaction-level graph would repeatedly attach the same entity descriptions to many transaction nodes and weaken the alignment between node identity and semantic information. More importantly, for transaction-time fraud detection, descriptions summarized from the full observation period would expose future behavioral information to earlier transactions. Avoiding such leakage would require dynamically updating entity descriptions at each transaction time, substantially increasing both generation and encoding costs. Under our user-level fraud identification setting after the complete observation period, the entity graph provides a more consistent alignment among prediction targets, node labels, structured attributes, and textual descriptions.

Nodes and Labels. The graph contains two types of entities: users and merchants. Users serve as the target nodes for fraud detection, while merchants are retained as background nodes that provide transaction connectivity and merchant-side semantic context. Following the use of background nodes in financial graph anomaly detection (Huang et al., 2022), merchant nodes participate in message passing but are excluded from the supervised prediction objective.

User labels are derived from their associated transaction labels using a priority rule. A user is labeled as fraudulent ($y = 1$) if they have initiated at least one fraudulent transaction. Among the remaining users, those associated with unlabeled transactions are assigned $y = 2$, while all others are labeled as normal ($y = 0$). All merchant nodes are assigned the background label $y = 3$. Only labeled users with $y \in \{0, 1\}$ participate in supervised training, validation, and testing, whereas unlabeled users and background merchants remain in the graph to preserve structural connectivity and support

message passing. This node-level distribution differs substantially from the transaction-level label distribution in Table 2, since any user with at least one fraudulent transaction is labeled fraudulent, whereas users containing unlabeled but no fraudulent transactions are excluded from supervised evaluation.

Edge Construction. We construct five types of entity relations from the original transaction records. The original User–Merchant transaction edges are preserved directly, while all induced structural relations are scored using inverse-document-frequency (IDF) weighting and sparsified with Mutual Top-K. IDF weighting reduces the contribution of highly frequent entities or attributes that provide weak discriminative information and assigns larger scores to rarer shared relations. When a node pair is supported by multiple shared entities, attributes, or repeated co-occurrences, their contributions are accumulated into a candidate-edge score. Mutual Top-K then retains an edge only when each endpoint ranks the other among its K highest-scoring candidate neighbors. This procedure filters asymmetric weak associations, limits excessive connectivity around high-degree nodes, and preserves relatively stable bidirectional relations.

The five relation types are constructed as follows:

1. U–M: User–Merchant Transaction Relation. A user and a merchant are connected if at least one transaction occurs between them. These edges directly preserve the observed user–merchant interactions and are not subject to IDF weighting or Mutual Top-K sparsification.
2. U–W–U: Local Temporal Co-occurrence Relation. For each merchant, its associated transactions are ordered by time. A local window of length temporal_window is constructed around each transaction, and distinct users appearing within the same window form candidate user–user relations. These candidates are subsequently scored and sparsified using IDF and Mutual Top-K.
3. U–A–U: Temporal Attribute-Matching Relation. All transactions are globally ordered by time, and only adjacent transactions are compared. When two adjacent transactions originate from different users and share either the same Location or the same Type, their users form a candidate relation. IDF weighting and Mutual Top-K are then applied.
4. U–M–U: Shared-Merchant Relation. Two users form a candidate relation when they have transacted with the same merchant. Unlike the local temporal relation above, this relation captures shared merchants over the full observation period. Candidate edges are filtered using IDF and Mutual Top-K.
5. M–U–M: Shared-User Relation. Merchants connected to the same user form candidate merchant–merchant relations. This relation is structurally symmetric to the shared-merchant relation and captures merchants linked through common users. The same IDF and Mutual Top-K procedure is applied.

After sparsification, all five relation types are merged into a single homogeneous graph. The final graph does not expose relation types or IDF scores to downstream detectors, and all retained edges are treated as unweighted connections. IDF weighting and Mutual Top-K are therefore used only during graph construction to select informative structural relations.

Original Node Features. For each user or merchant node, the associated transaction records are aggregated into a 12-dimensional behavioral representation. The features characterize four aspects of transaction behavior. First, transaction activity is described by the total number of transactions and the accumulated transaction amount. Second, transaction magnitude and dispersion are captured through the maximum amount, standard deviation, three quartiles of the amount distribution, and skewness, which together reflect typical spending levels, variability, and the presence of long-tailed or unusually large transactions. Third, behavioral concentration is measured by the proportions of transactions occurring at the most frequently used location and transaction type. Finally, the diversity of locations and transaction types is quantified using entropy-based measures, with larger values indicating more dispersed geographic activity or more diverse transaction behaviors. These aggregated statistics form the original node representation before projection and semantic fusion.

F.2 Semantic Text Construction

Semantic attributes are converted into natural-language sequences before embedding. Rather than encoding each semantic field independently, we serialize all fields of an entity into a unified sequence before embedding. This design allows the embedding model to capture semantic relationships

across attributes and provides richer context for fields that are less informative in isolation. It also produces a fixed-dimensional representation regardless of the number of fields, avoiding field-specific dimensional allocation and simplifying the incorporation of additional attributes. Moreover, encoding a complete semantic description better matches the natural-language input form of text embedding models than encoding isolated and often semantically sparse field fragments.

User Structured Semantic Template

Customer profile.
Province: {province}.
Gender: {gender}.
Age group: {age_group}.
Education: {education}.
Occupation industry: {occupation_industry}.

(A) User structured semantic template containing user demographic and geographic information.

User Textual Semantic Template

[Single-transaction users]
User description: {User_description}.

[Multi-transaction users]
Customer profile.
Province: {province}.
Gender: {gender}.
Age group: {age_group}.
Education: {education}.
Occupation industry: {occupation_industry}.
User description: {User_description}.

(B) User textual semantic template containing user profile information and entity-level behavioral descriptions.

Figure 9: Semantic templates for user.

For user nodes, the structured semantic template in Figure 9(A) serializes the generated demographic and geographic information into a unified user profile. When using textual semantics, user inputs are serialized according to the description-generation strategy. For single-transaction users, the deterministic description already verbalizes the structured user attributes, and only the description is therefore retained. For multi-transaction users, whose LLM-generated descriptions explicitly exclude demographic information, the structured user profile is prepended to the behavioral description. This ensures that structured profile information is represented once in both cases while avoiding redundant semantic content, as shown in Figure 9(B). This allows explicit profile attributes and higher-level behavioral characteristics inferred from the observed transaction history to be jointly represented within the same semantic sequence.

For merchant nodes, the structured semantic template in Figure 10(A) organizes the generated hierarchical business-category information into a unified merchant representation. When using textual semantics, the same category information is retained while merchant identity and the entity business description are further incorporated, as illustrated in Figure 10(B). The resulting sequence jointly captures business category, merchant identity, and business context.

Importantly, textual semantic inputs are constructed within the same semantic encoding pipeline rather than through an additional feature branch. For user nodes, structured profile information is included only once, avoiding repetition when it is already verbalized in the deterministic description of single-transaction users. Thus, structured and textual semantics share the same semantic encoding pipeline, with their difference arising from the semantic content included in the input sequence.

Merchant Structured Semantic Template

Merchant category level 1: {MCC_level1}.
Merchant category level 2: {MCC_level2}.

(A) Merchant structured semantic template containing hierarchical business category information.

Merchant Textual Semantic Template

Merchant category level 1: {MCC_level1}.
Merchant category level 2: {MCC_level2}.
Merchant name: {Merchant}.
Merchant description: {Merchant_description}.

(B) Merchant textual semantic template containing business category information, merchant name, and entity business descriptions.

Figure 10: Semantic templates for merchant.

F.3 Feature Encoding

The original features and semantic information are processed through separate encoding branches before being provided to the graph-based detectors. For the original branch, the 12-dimensional node features are standardized using statistics computed from the training set and mapped to a 128-dimensional representation through a node-type-specific linear layer followed by LayerNorm and ReLU. User and merchant nodes use separate projection and normalization parameters to accommodate their different feature distributions.

For the semantic branch, the serialized entity text constructed above is encoded using Qwen3-Embedding-8B (Zhang et al., 2025a). Leveraging its Matryoshka representation capability, we directly truncate the original 4096-dimensional output to 128 dimensions during encoding and apply L2 normalization, without additional dimensionality reduction such as PCA. These semantic embeddings are precomputed offline and kept fixed during graph-model training, without updating the parameters of Qwen3-Embedding-8B. The fixed embeddings are subsequently processed through trainable node-type-specific linear projections, followed by LayerNorm and ReLU.

Under the original features setting, the projected 128-dimensional original representation is directly provided to the graph model. When structured or textual semantics are introduced, the two 128-dimensional branch representations are first concatenated and then mapped back to 128 dimensions through a node-type-specific fusion projection followed by LayerNorm and ReLU. Consequently, all three settings provide the graph backbone with the same 128-dimensional node input, keeping the backbone input dimensionality fixed across semantic configurations.

F.4 Experimental Settings

Implementation Details. For all experiments, labeled user nodes are randomly divided into training, validation, and test sets at a ratio of 6:2:2, and identical splits are used across models and semantic settings for fair comparison. Optimization is performed using Adam with a learning rate of 0.001 and a weight decay of 0.001. Training is capped at 300 epochs with an early-stopping patience of 50 epochs based on validation loss, and the checkpoint with the lowest validation loss is used for test evaluation. Model-specific configurations follow the officially recommended settings of the corresponding graph-based detectors. Each experiment is repeated over five random seeds, and the mean performance is reported. All experiments are conducted on a server equipped with four 32GB NVIDIA Tesla V100 GPUs.

The larger training portion provides sufficient supervision for learning how the introduced semantics interact with the original features, while retaining separate validation and test sets for model selection and reliable evaluation. The 300-epoch upper bound provides sufficient optimization time for models

with different convergence behaviors, while early stopping with a patience of 50 epochs prevents unnecessary prolonged training and reduces overfitting.

Evaluation Metrics. We evaluate fraud detection performance using AUC, AP, and F1-macro, which characterize model effectiveness from different aspects. AUC measures the overall ranking ability between fraudulent and normal samples across different decision thresholds. AP focuses on the ranking quality of fraudulent samples and is particularly informative under class imbalance. F1-macro evaluates classification performance by computing F1 scores for both classes and assigning them equal importance, thereby reducing the influence of the majority class. Together, these three metrics provide a comprehensive evaluation of model performance.

G LLM-BASED SEMANTIC REASONING DETAILS

G.1 REASONING SETUP

We conduct zero-shot semantic reasoning using DeepSeek-V3 (DeepSeek-AI et al., 2025). To isolate the effect of semantic enrichment, the same model configuration and fixed reasoning prompt are used across all enrichment levels, with only the information provided in the input changing accordingly. Ground-truth fraud labels are excluded from the input, so the model must form its assessment solely from the observed behavioral and semantic evidence.

The reasoning task examines how enriched semantics change transaction behavior interpretation. The prompt asks the model to identify salient evidence, determine whether available semantics contextualize each observation, and produce an evidence-grounded assessment. This avoids treating semantic consistency as evidence of normality while retaining unexplained abnormalities.

The reasoning input follows the same three progressive information settings used in graph-based fraud detection. Under *Original Features*, the input contains the target user's behavioral features and transaction-derived information associated with its connected merchants, without generated semantic attributes or textual descriptions. Under *Structured Semantics*, the generated user profile and merchant business-category information are additionally provided. Under *Textual Semantics*, the structured information is retained and further enriched with the user behavioral description and merchant identity and business descriptions. Across the three settings, the underlying target user and associated merchants remain unchanged, with only the available semantic information progressively expanded. This enables direct comparison of how additional semantic context affects the interpretation of the same observed behavior. Unlike graph models, which consume projected numerical representations and semantic embeddings, the LLM directly receives the corresponding values and semantic information in textual form.

G.2 REASONING PROMPT AND OUTPUT PROTOCOL

A fixed reasoning prompt is applied across all cases and enrichment levels, as shown in Figure 11. The prompt consists of three components: analysis guidelines, target-user information, and a structured reasoning procedure. The analysis guidelines restrict the model to explicitly provided evidence and require semantic explanations to correspond to the specific behavioral patterns being interpreted. Textual descriptions may provide additional context beyond structured attributes, while abnormalities that remain unexplained are retained in the reasoning process.

For each target user, the model first identifies up to three behavioral observations relevant to the assessment. It then associates each observation with available semantic evidence and classifies their relationship as contextualized, partially contextualized, unresolved, or semantic–behavioral mismatch. The model subsequently summarizes the overall interpretation and returns a final normal or fraud assessment with a concise evidence-grounded rationale.

This structured output enables comparison across enrichment levels at both the behavioral-interpretation and final-assessment levels, allowing us to examine how structured and textual semantics progressively influence contextual reasoning under the same transaction evidence.

Financial Fraud Semantic Reasoning Prompt

[Analysis Guidelines]

You are a financial fraud analysis assistant. Assess the fraud risk of a target user using the information explicitly provided. Analyze behavioral evidence first, and then determine how the available semantic information changes the interpretation of the observed transaction behavior.

Follow these principles:

1. Use only the information provided, without adding unsupported scenarios or relationships between attributes.
2. Use semantic information only when it directly helps interpret the corresponding behavioral pattern; general semantic consistency is not sufficient to explain quantitative or structural abnormalities.
3. Keep each interpretation evidence-grounded and limited to the behavior it supports.
4. When textual descriptions are available, use them to enrich the interpretation beyond structured attributes while retaining attention to abnormalities that remain unexplained.

[Target User]

Behavioral Features: {behavioral_features}

User Semantic Information: {user_semantic_information}

Associated Merchants: {merchant_information}

[Reasoning Procedure]

1. Behavioral Evidence: identify up to three behavioral observations most relevant to the assessment.
2. Semantic Interpretation: for each observation, identify relevant semantic evidence and classify its relationship to the behavior as one of the following:
   - Contextualized: the semantic information directly explains the observed behavioral dimension.
   - Partially Contextualized: part of the behavior is explained, but an important irregularity remains.
   - Unresolved: the semantic information does not explain the observation.
   - Semantic-Behavioral Mismatch: the semantic context makes the observed behavior more difficult to reconcile.
3. Overall Interpretation: Summarize what the available semantic information helps interpret and what behavioral irregularities remain unexplained.
4. Final Assessment
   - Assessment: Normal / Fraud
   - Rationale: Provide a concise assessment based only on the evidence established above.

Figure 11: Prompt used for LLM-based semantic reasoning across all enrichment settings.

G.3 Evaluation Scope and Case Selection

The LLM experiment is designed as a qualitative zero-shot reasoning analysis rather than a predictive benchmark. It examines whether progressively enriched semantic information supports more contextual interpretation of the same transaction behavior, rather than evaluating the LLM as a supervised fraud detector. We therefore do not report classification metrics for this experiment, but focus on how the reasoning process changes as structured and textual semantics are progressively introduced.

Illustrative cases for qualitative analysis are selected according to the following criteria:

- **Cross-setting comparability.** The same target user must have complete reasoning outputs under all three enrichment settings, so that changes in the reasoning process can be attributed to the progressively introduced semantic information rather than differences in the analyzed case. This ensures that the comparison is conducted under a consistent behavioral context.
- **Reasoning contrast across settings.** The selected case should allow meaningful comparison of the reasoning process as semantic information is progressively introduced. Such differences may appear in the identified evidence, behavioral interpretation, or the connection between observed behavior and semantic context.
- **Diversity of semantic effects.** The selected cases should reflect different effects of semantic enrichment, including cases where additional context helps explain ambiguous behavior and cases where important abnormalities remain only partially contextualized or unresolved.