

Dual-Scale Relational Graph Transformers for Ecosystem-Aware Fraud Detection

Mohsen Nayebi Kerdabadi¹ Xinrou Li² Yao Xiao² Zijun Yao¹ Xin Sun²

¹University of Kansas, Lawrence, KS, USA

²AI Foundations, Capital One, McLean, VA, USA

{xinrou.li,yao.xiao,xin.sun}@capitalone.com

{mohsen.nayebi,zyao}@ku.edu

Abstract

Account takeover (ATO) fraud is a growing threat to digital banking, requiring effective detection while minimizing friction for legitimate customers. Production systems predominantly rely on tabular models that score sessions in isolation, discarding the relational structure of the underlying interaction network. Although graph-based models exploit relationships among sessions and network entities, they primarily reason over local neighborhoods and therefore capture only part of the problem: fraud risk depends jointly on the local relational structure surrounding a session and the evolving global state of the fraud ecosystem. We present HERMES (**HE**terogeneous **R**elational **Micro**-macro graph transformer **E**ncoder for high-risk **S**essions), a dual-scale architecture that jointly models these complementary scales of information. *Micro-GT* captures local heterogeneous graph structure through structured, relation-aware attention over a temporally safe session neighborhood. Complementing this local representation, *Macro-GT* models ecosystem-level context using non-anticipative climate tokens that summarize fraud dynamics, platform shifts, and infrastructure reuse, together with adaptive class prototypes that track representative fraud and benign session patterns over time. Evaluated on more than 130 million high-risk transaction sessions from a leading U.S. financial institution, HERMES consistently outperforms production and strong graph-based baselines, achieving a 44.44% relative reduction in customer friction and a 24.66% relative improvement in fraud recall over the production system. Ablation and temporal-stability analyses further demonstrate complementary gains from local relational modeling and global ecosystem context across changing fraud regimes.

1 Introduction

Account takeover (ATO) fraud is a growing threat to digital financial services, causing substantial financial losses and eroding customer trust. In an ATO attack, an adversary gains access to a legitimate user’s account to conduct unauthorized activity or obtain sensitive information. Effective detection therefore presents a fundamental operational tradeoff: increasing fraud recovery often requires challenging more legitimate customers, introducing friction that degrades user experience [Doerfler et al., 2019]. As attack strategies become increasingly sophisticated, practical fraud detection requires models that improve fraud recovery while minimizing unnecessary customer friction.

ATO fraud is inherently a structured and non-stationary prediction problem. Sessions are connected through a heterogeneous interaction graph of shared accounts, devices, and network identifiers, creating relation-specific and temporal dependencies that session-level features alone cannot capture [Kerdabadi et al., 2025]. At the same time, the risk associated with a local graph pattern depends on the broader fraud ecosystem, which evolves as attack volume, platform concentration, and infrastructure reuse shift over time. Reliable detection therefore requires reasoning jointly over the local relational structure surrounding a session and the evolving global ecosystem state in which that structure occurs. This challenge is further compounded by extreme class imbalance and delayed fraud adjudication, requiring models to operate non-anticipatively using only information available at scoring time.

Existing fraud-detection models capture these complementary sources of information only partially. Production tabular systems score sessions largely in isolation, discarding relational dependencies among connected entities. Graph neural networks [Hamilton et al., 2017, Rao et al., 2021, Xiang et al., 2023] improve upon this by aggregating information from local neighborhoods, and temporal label propagation [Kerdabadi et al., 2025] can further incorporate confirmed fraud signals from connected sessions. However, conventional message passing can struggle to preserve informative dependencies across multiple hops because repeated aggregation may induce oversmoothing and oversquashing [Dwivedi et al., 2026]. Generic aggregation can also blur the distinct semantics of account, device, and network relations and inadequately capture higher-order interactions across heterogeneous paths within a multi-hop neighborhood. Moreover, models restricted to local neighborhoods remain largely unaware of ecosystem-level regime changes. Consequently, existing approaches lack a unified mechanism for combining expressive local relational modeling with the evolving global context in which that structure should be interpreted.

We present HERMES: **H**eterogeneous **R**elational **M**icro–**m**acro Graph Transformer **E**ncoder for high-risk transaction **S**essions, a dual-scale architecture that addresses these limitations by jointly modeling local relational structure and global ecosystem context. *Micro-GT* applies a relational graph transformer over a temporally safe sampled neighborhood, using four structured attention masks to preserve relation-specific semantics and capture higher-order interactions across heterogeneous paths. Complementing this local representation, *Macro-GT* attends over non-anticipative ecosystem climate tokens that summarize evolving fraud dynamics, platform shifts, and infrastructure reuse, together with adaptive class prototypes that provide representative references for fraud and benign session patterns as they evolve. The two representations are fused and trained end-to-end for session-level fraud prediction.

Evaluated on more than 130 million high-risk transaction sessions collected over one year at a leading U.S. financial institution, HERMES consistently outperforms both the production gradient-boosted system and strong graph-based baselines. Relative to the production system, HERMES achieves a 44.44% reduction in customer friction at fixed fraud recall and a 24.66% improvement in fraud recall at a fixed flag rate. Ablation studies demonstrate complementary gains from the proposed components, while temporal evaluation across four held-out test months demonstrates robustness to distribution shift. Our main contributions are as follows:

- We present HERMES, a dual-scale architecture that jointly reasons over local graph neighborhood structure and evolving global ecosystem-level distributional context, enabling session-level fraud prediction across two complementary scales within a unified framework.
- We propose Micro-GT, a local relational graph transformer that uses four structured attention masks to preserve heterogeneous relation semantics and capture higher-order interactions across temporally safe multi-hop neighborhoods.
- We propose Macro-GT, an ecosystem-scale transformer that attends over a rich set of non-anticipative daily climate tokens covering multiple fraud signal categories, augmented with adaptive class prototypes that provide representative references for fraud and benign session patterns.
- We conduct a production-scale evaluation on more than 130 million high-risk transaction sessions from a leading U.S. financial institution, with extensive ablations and temporal analyses demonstrating complementary gains from the local and global components and substantial improvements over the deployed production system.

2 Problem Formulation

2.1 Heterogeneous Temporal Session Graph

Let $\mathcal{S}$ denote the set of high-risk digital transaction sessions. Each session $v \in \mathcal{S}$ is associated with a feature vector $\mathbf{x}_v \in \mathbb{R}^d$, an occurrence timestamp $\tau(v)$, and a binary fraud label $y_v \in \{0, 1\}$. Because fraud labels are adjudicated with delay, y_v becomes observable only at $\tau(v) + \ell$, where $\ell > 0$ denotes the *label lag*.

Let $\mathcal{R}$ denote the set of relation types corresponding to account, device, and network identifiers, and let $\text{id}_r(v)$ denote the identifier of type $r \in \mathcal{R}$ associated with session v . We define the relation-specific temporal predecessors of v as

$$\mathcal{N}_r^-(v) = \text{TopK}_\tau(\{u \in \mathcal{S} : \text{id}_r(u) = \text{id}_r(v), \tau(v) - T \leq \tau(u) < \tau(v)\}), \quad (1)$$

where T is the recency window and $\text{TopK}_\tau(\cdot)$ retains up to K candidate sessions with the largest occurrence timestamps.

For each destination session v , relation type r , and associated identifier $a = \text{id}_r(v)$, we instantiate a time-conditioned identifier node $i_{r,a}^{\tau(v)}$, representing the state of identifier a at time $\tau(v)$. Rather than treating identifiers as static entities, this formulation allows their representations to evolve as additional historical activity becomes available. Each shared-identifier relation is represented as

$$u \longrightarrow i_{r,a}^{(\tau(v))} \longrightarrow v, \quad u \in \mathcal{N}_r^-(v), \quad a = \text{id}_r(u) = \text{id}_r(v). \quad (2)$$

Thus, $i_{r,a}^{\tau(v)}$ represents the time-conditioned state of identifier a , summarizing its history available when session v is scored. For a target session v_0 , we recursively apply Eq. (1) for H session hops to construct the local computation graph $\mathcal{G}(v_0)$. We define *session-hop distance* by treating each path $u \rightarrow i_{r,a}^{(\tau(v))} \rightarrow v$ as one relational hop. Accordingly, v_0 is assigned hop 0, its direct session predecessors hop 1, and more distant predecessors increasing hop indices up to H . Because every expansion satisfies $\tau(u) < \tau(v)$, $\mathcal{G}(v_0)$ is temporally ordered and contains no future-session information.

2.2 Temporally-Safe Node Representation

Session nodes carry their session-level features $\mathbf{x}_v$. To incorporate historical fraud information without label leakage, label-dependent features are computed only from predecessors whose labels are observable when v is scored. For relation r , the lag-safe predecessor set is

$$\mathcal{L}_r^-(v) = \{u \in \mathcal{N}_r^-(v) : \tau(u) + \ell \leq \tau(v)\}. \quad (3)$$

Thus, structural inclusion requires $\tau(u) < \tau(v)$, while label-dependent features additionally require $\tau(u) + \ell \leq \tau(v)$. Identifier-node features likewise use only information available at the destination timestamp. Full feature definitions are provided in Appendix A.1.

2.3 Task Definition

For a target session v_0 , let $\mathbf{X}(v_0)$ denote the serve-time features associated with its local computation graph, and let $\mathbf{C}_{\tau(v_0)}$ denote the global ecosystem context available at time $\tau(v_0)$. The session-level fraud detection task is

$$\hat{y}_{v_0} = f_\theta(\mathcal{G}(v_0), \mathbf{X}(v_0), \mathbf{C}_{\tau(v_0)}) \in [0, 1], \quad (4)$$

where $\hat{y}_{v_0}$ estimates the probability that v_0 is fraudulent. All local and global inputs are restricted to information available at $\tau(v_0)$, ensuring consistent temporal constraints during training, evaluation, and production inference.

3 Method

HERMES is a dual-scale architecture that represents each target session from two complementary perspectives. *Micro-GT* models the local relational structure of the target-conditioned temporal graph, while *Macro-GT* captures the evolving ecosystem context available at the target’s scoring time. The two representations are fused through a learned projection and optimized jointly for session-level fraud prediction. Figure 1 provides an overview.

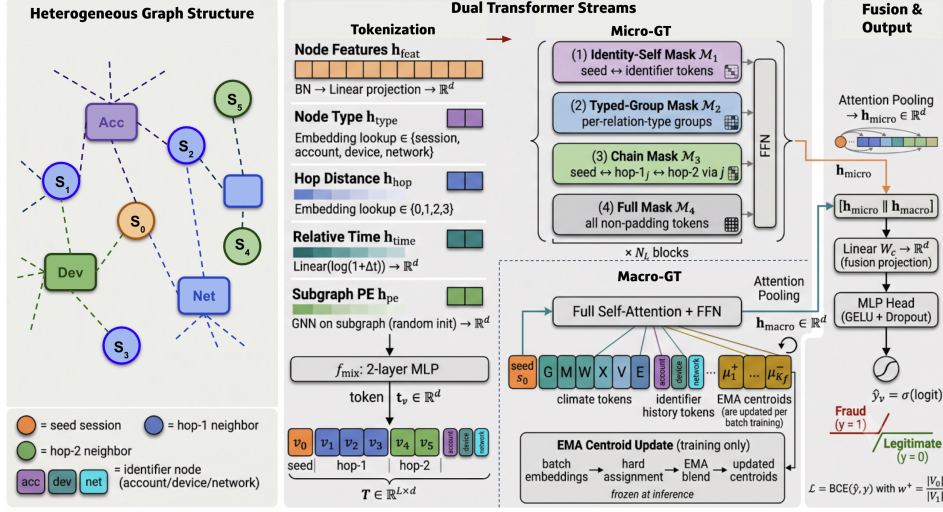


Figure 1: HERMES architecture. Micro-GT encodes the target session’s local heterogeneous neighborhood through structured relational attention, while Macro-GT models non-anticipative ecosystem context together with adaptive fraud and benign reference prototypes.

3.1 Local Graph Tokenization

For a target session v_0 , Micro-GT operates on the target-conditioned computation graph $\mathcal{G}(v_0)$ defined in Section 2. We instantiate $H = 2$ session hops and retain fixed budgets of at most K_1 hop-1 and K_2 hop-2 session predecessors, selected according to the temporal sampling rule in Eq. (1). Missing positions are padded and masked from all subsequent attention operations.

Time-conditioned identifier states mediate relational paths throughout $\mathcal{G}(v_0)$, but only those adjacent to the target are explicitly tokenized. Specifically, for each relation $r \in \mathcal{R}$, we include $i_{r,a_r}^{(\tau(v_0))}$, where $a_r = \text{id}_r(v_0)$. Their serve-time feature construction is detailed in Appendix A.3. Figure 2 illustrates the resulting local context.

Each tokenized session or identifier state q is mapped to $\mathbf{t}_q \in \mathbb{R}^d$ by combining feature, type, hop, temporal, and structural encodings following Dwivedi et al. [2026]:

$$\mathbf{t}_q = f_{\text{mix}}(\mathbf{h}_{\text{feat}}(q), \mathbf{h}_{\text{type}}(q), \mathbf{h}_{\text{hop}}(q), \mathbf{h}_{\text{time}}(q, v_0), \mathbf{h}_{\text{pe}}(q)), \quad (5)$$

where f_{mix} is a two-layer MLP projecting the concatenated components to dimension d . The components are

$$\begin{aligned} \mathbf{h}_{\text{feat}}(q) &= \text{MLP}_{\phi(q)}(\text{BN}_{\phi(q)}(\mathbf{x}_q)) \in \mathbb{R}^d, \\ \mathbf{h}_{\text{type}}(q) &= \mathbf{E}_{\text{type}}[\phi(q)], \\ \mathbf{h}_{\text{hop}}(q) &= \mathbf{E}_{\text{hop}}[p(q, v_0)], \\ \mathbf{h}_{\text{time}}(q, v_0) &= \mathbf{w}_t \log(1 + \Delta\tau(q, v_0)) + b_t. \end{aligned} \quad (6)$$

Here, $\text{MLP}_{\phi(q)}$ and $\text{BN}_{\phi(q)}$ denote type-specific feature projection and normalization, mapping raw node features of potentially different dimensions to d ; $\phi(q)$ denotes the node type and $p(q, v_0)$ the session-hop index. The target and its adjacent identifier states are assigned hop 0, while hop-1 and hop-2 sessions receive indices 1 and 2, respectively. For session tokens, $\Delta\tau(q, v_0) = \tau(v_0) - \tau(q)$; for target-adjacent identifier states, $\Delta\tau(q, v_0) = 0$. Finally, $\mathbf{h}_{\text{pe}}(q)$ is a permutation-equivariant structural positional encoding produced by a lightweight GNN with randomized initialization over $\mathcal{G}(v_0)$, following Dwivedi et al. [2026].

Let $\{v_k^{(1)}\}_{k=1}^{K_1}$ and $\{v_k^{(2)}\}_{k=1}^{K_2}$ denote the padded hop-1 and hop-2 session slots, and let $a_j = \text{id}_{r_j}(v_0)$. The Micro-GT input sequence is

$$\begin{aligned} \mathbf{Z}^{(0)} &= [\mathbf{t}_{v_0} \mid \mathbf{t}_{v_1^{(1)}}, \dots, \mathbf{t}_{v_{K_1}^{(1)}} \mid \mathbf{t}_{v_1^{(2)}}, \dots, \mathbf{t}_{v_{K_2}^{(2)}} \mid \\ &\quad \mathbf{t}_{i_{r_1, a_1}^{(\tau(v_0))}}, \dots, \mathbf{t}_{i_{r_{|\mathcal{R}|}, a_{|\mathcal{R}|}}^{(\tau(v_0))}}] \in \mathbb{R}^{L \times d}, \quad L = 1 + K_1 + K_2 + |\mathcal{R}|. \end{aligned} \quad (7)$$

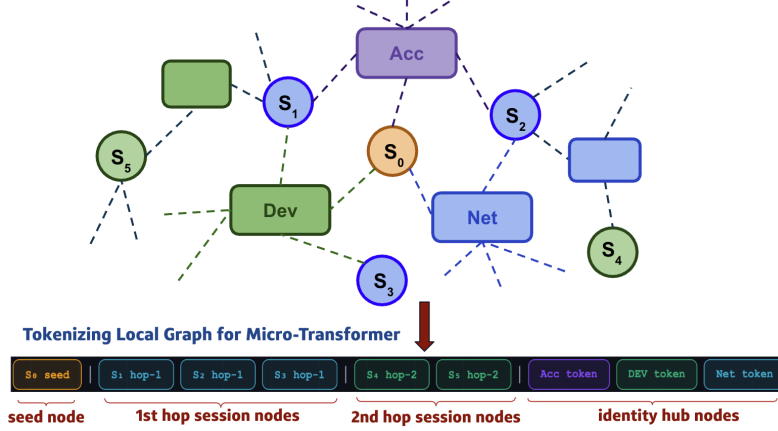


Figure 2: Micro-GT local graph tokenization with the target session, hop-1 and hop-2 predecessors, and target-adjacent time-conditioned identifier states.

Padding positions remain masked throughout Micro-GT.

3.2 Micro-GT: Local Relational Transformer

Let $\mathbf{T} = \mathbf{Z}^{(0)}$ denote the token sequence from Eq. (7). Rather than applying undifferentiated all-pair attention, Micro-GT uses four structured attention masks applied sequentially, each with independent query, key, and value projections. Each masked attention pass follows the pre-norm residual form

$$\mathbf{T} \leftarrow \mathbf{T} + \text{Attn}_m(\text{LN}(\mathbf{T}), M_m), \quad \text{Attn}_m(\mathbf{T}, M_m) = \text{softmax}\left(\frac{\mathbf{Q}_m \mathbf{K}_m^\top}{\sqrt{d_k}} + M_m\right) \mathbf{V}_m, \quad (8)$$

where $\mathbf{Q}_m = \mathbf{T} \mathbf{W}_m^Q$, $\mathbf{K}_m = \mathbf{T} \mathbf{W}_m^K$, and $\mathbf{V}_m = \mathbf{T} \mathbf{W}_m^V$. Each $M_m \in \{0, -\infty\}^{L \times L}$ specifies the permitted interactions for attention pass m . The four masks are illustrated in Figure 3.

(1) Identity-Self mask $\mathbf{M}_1$. The target token and its identifier tokens attend to one another, allowing the model to first establish the target’s relational identity context. Let $\mathcal{I}_0 = \{v_0\} \cup \{i_{r,a_r}^{(v_0)} : r \in \mathcal{R}\}$. Then $\mathbf{M}_1[i, j] = 0$ iff both tokens i and j belong to $\mathcal{I}_0$.

(2) Relation-group mask $\mathbf{M}_2$. For each session token q , let $\rho(q) \subseteq \mathcal{R}$ denote the relation types it shares with the target. Two session tokens may interact when their relation sets overlap, while the target and identifier-context tokens remain accessible to all groups:

$$\mathbf{M}_2[i, j] = 0 \iff \rho(i) \cap \rho(j) \neq \emptyset \text{ or } i \in \mathcal{I}_0 \text{ or } j \in \mathcal{I}_0. \quad (9)$$

This preserves relation-specific interactions while allowing sessions linked through multiple identifiers to participate in multiple groups.

(3) Chain mask $\mathbf{M}_3$. For each hop-1 session, the target, that hop-1 session, and its associated hop-2 predecessors form an attention group. This explicitly models multi-hop relational chains while keeping paths through different hop-1 intermediaries separate.

(4) Full-context mask $\mathbf{M}_4$. All non-padding tokens attend to one another, enabling information exchange across relation groups and propagation chains after the structured passes.

The four passes are applied in order: identity-self, relation-group, chain, full-context, followed by a shared feed-forward sublayer with pre-norm residual. This constitutes one relational transformer block, and HERMES stacks N_L such blocks.

The final local representation $\mathbf{h}_{\text{micro}} \in \mathbb{R}^d$ is obtained through seed-conditioned attention pooling:

$$\mathbf{h}_{\text{micro}} = \mathbf{t}_{v_0}^{(N_L)} + \sum_{i \neq v_0} \alpha_i \mathbf{t}_i^{(N_L)}, \quad \alpha_i = \frac{\exp\left(\mathbf{w}_r^\top [\mathbf{t}_{v_0}^{(N_L)} \parallel \mathbf{t}_i^{(N_L)}]\right)}{\sum_{j \neq v_0} \exp\left(\mathbf{w}_r^\top [\mathbf{t}_{v_0}^{(N_L)} \parallel \mathbf{t}_j^{(N_L)}]\right)}, \quad (10)$$

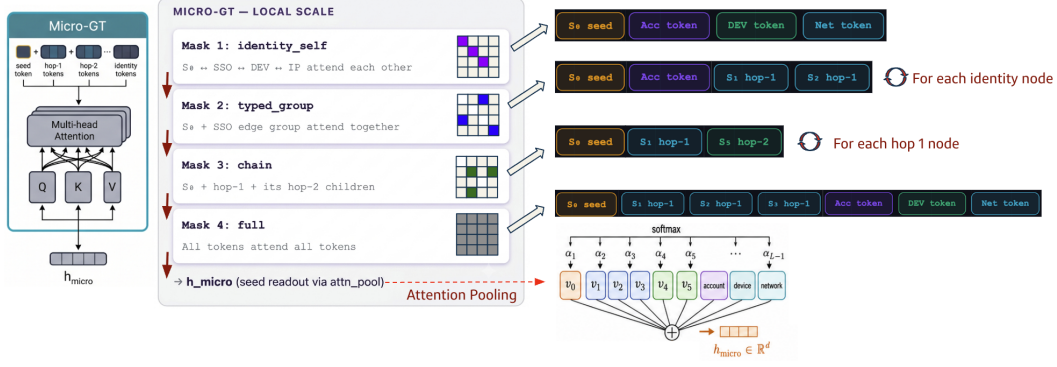


Figure 3: Micro-GT applies four sequential attention patterns over the local token sequence followed by seed-conditioned pooling to obtain $\mathbf{h}_{\text{micro}}$.

where $\mathbf{w}_r \in \mathbb{R}^{2d}$ is learned and padding positions are excluded from the softmax.

3.3 Macro-GT: Ecosystem Context Transformer

Macro-GT models the global ecosystem context $\mathbf{C}_{\tau(v_0)}$ available when target session v_0 is scored. For each calendar date t , we construct typed climate tokens summarizing complementary aspects of the fraud ecosystem: (1) global fraud activity (G), including multi-window fraud-rate trends, session volume, and acceleration; (2) platform-specific behavior (P_1, P_2), capturing per-channel fraud rates, traffic share, and temporal trends; (3) cross-platform shifts (X), reflecting changes in fraud concentration across channels; (4) velocity and anomaly signals (V), capturing sudden changes in fraud rate, volume, and campaign intensity; (5) ecosystem topology (E), summarizing new-entity activity, identifier sharing, and infrastructure concentration; and (6) the current risk regime (R), including label coverage, high-risk infrastructure prevalence, and repeat-fraud activity. All statistics are computed non-anticipatively; in particular, label-dependent quantities use only sessions satisfying $\tau(u) + \ell \leq t$. Complete climate-feature definitions are provided in Appendix A.6.

Each climate feature vector is projected to $\mathbb{R}^d$ and combined with a learned type embedding. Macro-GT additionally receives the target session representation, its time-conditioned identifier states, and adaptive class prototypes representing characteristic fraud and benign patterns. Let $\mathbf{s}_0$ denote the projected target representation, $\mathbf{c}_k$ a climate token, and $\mathbf{i}_r$ the encoded state $i_{r,a_r}^{(\tau(v_0))}$ of the target’s identifier of type r , where $a_r = \text{id}_r(v_0)$. The input sequence is

$$\mathbf{S}^{(0)} = \left[\mathbf{s}_0 \mid \mathbf{c}_G, \mathbf{c}_{P_1}, \mathbf{c}_{P_2}, \mathbf{c}_X, \mathbf{c}_V, \mathbf{c}_E, \mathbf{c}_R \mid \mathbf{i}_1, \dots, \mathbf{i}_{|\mathcal{R}|} \mid \boldsymbol{\mu}_1^+, \dots, \boldsymbol{\mu}_{K_f}^+, \boldsymbol{\mu}_1^-, \dots, \boldsymbol{\mu}_{K_{nf}}^- \right]. \quad (11)$$

A standard pre-norm transformer layer applies full self-attention over $\mathbf{S}^{(0)}$, allowing the target to jointly compare its own state with the current ecosystem context, its identifier history, and representative fraud and benign patterns. The resulting macro representation $\mathbf{h}_{\text{macro}} \in \mathbb{R}^d$ is obtained using the same seed-conditioned attention pooling as in Micro-GT, with $\mathbf{s}_0$ as the query.

3.3.1 Adaptive Class Prototypes

Fraud and benign sessions may exhibit multiple characteristic patterns rather than forming a single cluster. We therefore maintain separate sets of representative prototypes in the Macro-GT embedding space:

$$\mathcal{P}^+ = \{\boldsymbol{\mu}_k^+\}_{k=1}^{K_f}, \quad \mathcal{P}^- = \{\boldsymbol{\mu}_k^-\}_{k=1}^{K_{nf}}, \quad (12)$$

where $\mathcal{P}^+$ and $\mathcal{P}^-$ provide reference patterns for fraud and benign sessions, respectively.

Initialization. The two prototype sets are initialized separately using K-Means++ seeding [Arthur and Vassilvitskii, 2007] on class-stratified training embeddings in batch-normalized representation space.

Adaptive update. After each training batch, a session embedding is assigned only to the nearest prototype of its own class. For class $c \in \{+, -\}$, let $\mathcal{C}_k^c$ denote the batch embeddings assigned to prototype μ_k^c . Assigned prototypes are updated by exponential moving average (EMA):

$$\mu_k^c \leftarrow \alpha \mu_k^c + (1 - \alpha) \frac{1}{|\mathcal{C}_k^c|} \sum_{v \in \mathcal{C}_k^c} \hat{\mathbf{h}}_{\text{macro}}(v), \quad c \in \{+, -\}, \quad (13)$$

where $\hat{\mathbf{h}}_{\text{macro}} = \text{BN}(\mathbf{h}_{\text{macro}})$ and $\alpha \in (0, 1)$ is the EMA momentum. The prototypes are maintained as non-learnable buffers and updated outside the backward pass, decoupling their adaptation from the classification objective. Prototypes that remain unassigned beyond a fixed patience threshold are reinitialized from class-matched batch embeddings to maintain coverage of the representation space.

3.4 Training Objective

The micro and macro embeddings are combined via a learned linear projection:

$$h_{\text{out}} = \mathbf{W}_c [h_{\text{micro}} \parallel h_{\text{macro}}] + \mathbf{b}_c, \quad (14)$$

and a two-layer MLP head with GELU activation and dropout produces the scalar logit. HERMES is trained end-to-end with class-reweighted binary cross-entropy:

$$\mathcal{L} = -\frac{1}{|\mathcal{B}|} \sum_{v \in \mathcal{B}} [w^+ y_v \log \hat{y}_v + (1 - y_v) \log(1 - \hat{y}_v)], \quad w^+ = \frac{|\mathcal{V}_0|}{|\mathcal{V}_1|}, \quad (15)$$

where $\hat{y}_v = \sigma(\text{logit}_v)$, $|\mathcal{V}_0|$ and $|\mathcal{V}_1|$ are the non-fraud and fraud counts in the training set, and $\mathcal{B}$ is a mini-batch sampled via a neighborhood loader. The EMA prototype update is applied after each gradient step, strictly outside the backward pass.

4 Experiments

4.1 Dataset and Evaluation Protocol

We evaluate HERMES on a proprietary dataset of more than 130 million high-risk transaction sessions collected over one year from the digital channels of a leading U.S. financial institution. Each session includes behavioral, platform, and device-related features, together with shared account, device, and network identifiers that define the relational graph. Fraud labels are binary and become observable only after an adjudication delay of ℓ days. Fraud accounts for less than 0.1% of all sessions, resulting in extreme class imbalance. We use a chronological split: the final four months form the test set, the preceding two months form the validation set, and all earlier sessions are used for training. The four test months are also evaluated separately for temporal-stability analysis. All models use the same temporal splits and obey the serve-time and label-maturity constraints defined in Section 2, preventing future-session and immature-label leakage. We report ROC-AUC overall and separately for two digital access channels, denoted P_1 and P_2 for confidentiality. To evaluate deployment-relevant performance, we additionally report two operating-point metrics based on the *flag rate*, i.e., the fraction of sessions flagged as high risk and subjected to friction interventions. Flag rate therefore serves as an operational proxy for customer friction. Specifically, $FR@50R$ denotes the flag rate required to achieve 50% fraud recall (lower is better), while $R@5FR$ denotes fraud recall at a 5% flag rate (higher is better).

4.2 Results

Table 1 compares HERMES with production, tabular, sequential, and graph-based baselines, followed by an incremental ablation of the proposed architecture. All relative changes are reported with respect to the production XGBoost system.

Baseline comparison. *XGBoost (production)* is the gradient-boosted production system that scores sessions independently and serves as the primary reference, achieving 80.26% overall AUC, 78.05% P_1 AUC, and 74.10% P_2 AUC. *MLP* performs nearly identically,

Table 1: Overall performance and component ablation. All values are in %. FR@50R denotes the flag rate at 50% fraud recall ($\downarrow$), and R@5FR denotes fraud recall at a 5% flag rate ($\uparrow$). Parentheses report relative change from the production XGBoost baseline.

Model	AUC	P1 AUC	P2 AUC	FR@50R $\downarrow$	R@5FR $\uparrow$
XGBoost (production)	80.26	78.05	74.10	9.671	39.24
MLP	80.29 (+0.03)	77.99 (−0.08)	75.10 (+1.35)	9.360 (−3.21)	39.78 (+1.39)
FT-Transformer	79.16 (−1.37)	77.02 (−1.31)	66.54 (−10.20)	11.213 (+15.94)	35.42 (−9.74)
Transformer-Seq	81.95 (+2.10)	80.28 (+2.86)	72.97 (−1.52)	9.244 (−4.41)	39.60 (+0.91)
GraphSAGE	82.00 (+2.16)	80.05 (+2.57)	76.20 (+2.84)	8.701 (−10.03)	41.54 (+5.87)
GraphSAGE + LP (ATLAS)	84.02 (+4.69)	82.45 (+5.64)	78.75 (+6.28)	6.325 (−34.59)	46.93 (+19.62)
Micro-GT (full attention)	84.58 (+5.39)	82.83 (+6.13)	80.47 (+8.59)	5.821 (−39.80)	47.75 (+21.70)
Micro-GT (relational masks)	84.78 (+5.64)	83.10 (+6.48)	80.37 (+8.47)	5.801 (−40.01)	47.83 (+21.89)
+ Macro-GT	84.93 (+5.82)	83.24 (+6.65)	81.06 (+9.40)	5.519 (−42.93)	48.44 (+23.45)
Hermes (+centroids)	85.12 (+6.06)	83.46 (+6.94)	81.35 (+9.79)	5.373 (−44.44)	48.91 (+24.66)

indicating that additional model capacity alone provides little benefit without relational information. *FT-Transformer* [Gorishniy et al., 2021] underperforms XGBoost, particularly on P2 (66.54% AUC). *Transformer-Seq* improves overall AUC to 81.95% but remains weak on P2 (72.97%), showing that sequential context alone cannot capture broader cross-session dependencies. *GNN (GraphSAGE)* exploits the shared-identifier graph, increasing overall AUC to 82.00% and P2 AUC to 76.20%, while reducing FR@50R by 10.03% relative to production. *GNN + LP (ATLAS)* [Kerdabadi et al., 2025] further incorporates lag-safe fraud history from connected predecessors, reaching 84.02% overall AUC, 82.45% P1 AUC, and 78.75% P2 AUC. Relative to production, ATLAS reduces FR@50R by 34.59% and improves R@5FR by 19.62%, establishing a strong relational baseline and highlighting the value of adjudicated fraud signals from connected sessions.

Component ablation. Starting from ATLAS, *Micro-GT (full attention)* replaces homogeneous SAGE-style aggregation with a local transformer over the heterogeneous node tokens defined in Eq. (6). This raises overall AUC from 84.02% to 84.58% and P2 AUC from 78.75% to 80.47%, while improving FR@50R from 6.325% to 5.821% and R@5FR from 46.93% to 47.75%. These gains show that incorporating node heterogeneity and direct all-pair interactions captures higher-order dependencies beyond conventional message passing. *Micro-GT (relational masks)* further increases overall AUC to 84.78% and P1 AUC to 83.10%, while maintaining comparable operational performance, supporting structured attention over account, device, network, and multi-hop relational interactions. Adding *Macro-GT (climate tokens)* raises overall AUC to 84.93% and P2 AUC to 81.06%, while reducing FR@50R to 5.519% and increasing R@5FR to 48.44%. The strong P2 improvement indicates that ecosystem-level fraud dynamics, platform shifts, velocity, and infrastructure signals provide complementary information beyond the local neighborhood. Finally, adaptive class prototypes complete HERMES, yielding the best performance across all reported metrics: 85.12% overall AUC, 83.46% P1 AUC, 81.35% P2 AUC, 5.373% FR@50R, and 48.91% R@5FR. Compared with production, HERMES improves overall, P1, and P2 AUC by 4.86, 5.41, and 7.25 points, respectively, while achieving a 44.44% relative reduction in FR@50R and a 24.66% relative improvement in R@5FR. Overall, the results demonstrate complementary gains from expressive local relational modeling, structured relation-aware attention, global ecosystem context, and adaptive class prototypes.

4.3 Temporal Stability

To assess robustness to temporal distribution shift, we evaluate HERMES and key baselines separately across four consecutive held-out test months. Figure 4 shows that HERMES consistently outperforms all baselines throughout the evaluation window, with no degradation relative to the graph-based models in later months. The gains remain stable across periods with changing fraud activity, indicating that the combination of local relational modeling and global ecosystem context improves robustness across evolving fraud regimes. Such temporal stability is particularly important in production settings, where fraud distributions shift continuously, and frequent model recalibration is costly.

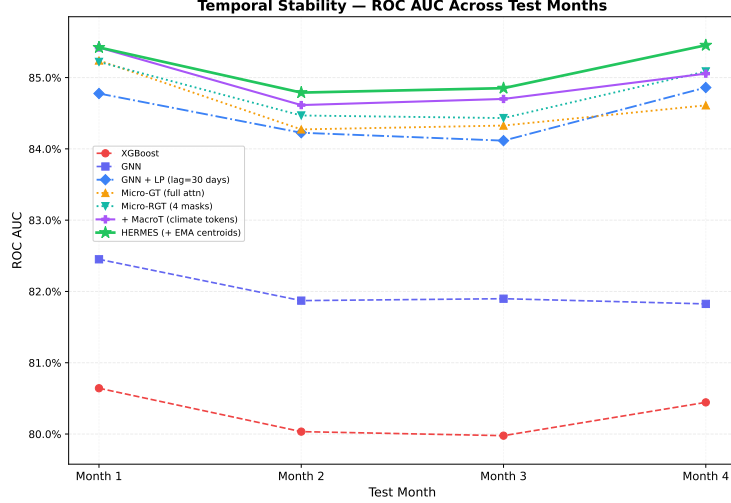


Figure 4: Temporal stability: ROC-AUC of HERMES and key baselines across four consecutive held-out test months. HERMES consistently outperforms all baselines.

5 Related Work

Account takeover detection in operational systems requires balancing fraud prevention against unnecessary intervention on legitimate users. Doerfler et al. [2019] study this tradeoff at Google by evaluating risk-aware login challenges against real-world account hijacking attempts and quantifying the resulting user friction. In an industrial marketplace setting, Kawase et al. [2019] analyze ATO at mobile.de and develop a concept-drift-sensitive detection strategy for evolving attacker behavior.

Relational modeling has also been explored more broadly for financial fraud [Liu et al., 2018, Jiang et al., 2022, Egressy et al., 2024]. TitAnt [Cao et al., 2019] demonstrates large-scale real-time transaction fraud detection using transaction-network information at Ant Financial, while xFraud [Rao et al., 2021] applies heterogeneous graph neural networks to large-scale real-world transaction networks. GTAN [Xiang et al., 2023] constructs temporal transaction graphs and propagates risk through gated temporal attention, and FraudGT [Lin et al., 2024] introduces an edge-aware graph transformer for financial transaction graphs. Beyond fraud-specific architectures, Relational Graph Transformer (RelGT) [Dwivedi et al., 2026] introduces multi-element relational tokenization with local subgraph attention and global centroids for general relational prediction. HERMES builds on this tokenization framework while introducing temporally structured local attention, ecosystem climate tokens, and class-stratified EMA prototypes for non-anticipative ATO detection. Among ATO-specific methods, ATLAS [Kerdabadi et al., 2025] is most closely related, modeling consumer-banking ATO as node classification on a time-respecting session graph that links sessions through shared account, device, and network identifiers using GraphSAGE and serve-time-safe label propagation. Building on this line of relational fraud modeling, HERMES combines structured local reasoning over heterogeneous, time-conditioned session relationships with non-anticipative ecosystem-level context and adaptive fraud and benign reference prototypes.

6 Conclusion

We presented HERMES, a dual-scale graph transformer architecture for ATO fraud detection that combines structured local relational modeling with global ecosystem context. *Micro-GT* captures heterogeneous multi-hop interactions, while *Macro-GT* incorporates non-anticipative climate signals and adaptive class prototypes. Evaluated on more than 130 million high-risk transaction sessions, HERMES achieves 85.12% ROC-AUC, a 44.44% relative reduction in customer friction, and a 24.66% relative improvement in fraud recall over the production system. Ablation and temporal analyses show complementary gains from the local and global components and robust performance across evolving fraud conditions.

References

- David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. In *ACM-SIAM Symposium on Discrete Algorithms*, 2007.
- Shaosheng Cao, XinXing Yang, Cen Chen, Jun Zhou, Xiaolong Li, and Yuan Qi. Titant: online real-time transaction fraud detection in ant financial. *Proc. VLDB Endow.*, 12(12):2082–2093, August 2019. ISSN 2150-8097. doi: 10.14778/3352063.3352126. URL <https://doi.org/10.14778/3352063.3352126>.
- Periwinkle Doerfler, Kurt Thomas, Maija Marincenko, Juri Ranieri, Yu Jiang, Angelika Moscicki, and Damon McCoy. Evaluating login challenges as a defense against account takeover. In *The World Wide Web Conference*, pages 372–382, 2019.
- Vijay Prakash Dwivedi, Sri Jaladi, Yangyi Shen, Federico Lopez, Charilaos Kanatsoulis, Rishi Puri, Matthias Fey, and Jure Leskovec. Relational graph transformer. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust, editors, *International Conference on Learning Representations*, volume 2026, pages 154348–154369, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/fa1ef43326cfb8a808c0649708187f2a-Paper-Conference.pdf.
- Béni Egressy, Luc Von Niederhäusern, Jovan Blanuša, Erik Altman, Roger Wattenhofer, and Kubilay Atasü. Provably powerful graph neural networks for directed multigraphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 11838–11846, 2024.
- Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. In *Advances in Neural Information Processing Systems*, 2021.
- William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, 2017.
- Jiaxin Jiang, Yuan Li, Bingsheng He, Bryan Hooi, Jia Chen, and Johan Kok Zhi Kang. Spade: A real-time fraud detection framework on evolving graphs. *Proceedings of the VLDB Endowment*, 16(3):461–469, 2022.
- Ricardo Kawase, Francesca Diana, Mateusz Czeladka, Markus Schöler, and Manuela Faust. Internet fraud: the case of account takeover in online marketplace. In *Proceedings of the 30th ACM conference on hypertext and social media*, pages 181–190, 2019.
- Mohsen Nayebi Kerdabadi, William Andrew Byron, Xin Sun, and Amirfarrokh Iranitalab. Spatio-temporal directed graph learning for account takeover fraud detection. *arXiv preprint arXiv:2509.20339*, 2025.
- Junhong Lin, Xiaojie Guo, Yada Zhu, Samuel Mitchell, Erik Altman, and Julian Shun. Fraudgt: a simple, effective, and efficient graph transformer for financial fraud detection. In *Proceedings of the 5th ACM International Conference on AI in Finance*, pages 292–300, 2024.
- Ziqi Liu, Chaochao Chen, Xinxing Yang, Jun Zhou, Xiaolong Li, and Le Song. Heterogeneous graph neural networks for malicious account detection. In *ACM International Conference on Information and Knowledge Management*, 2018.
- Susie Xi Rao, Shuai Zhang, Zhichao Han, Zitao Zhang, Wei Min, Zhiyao Chen, Yinan Shan, Yang Zhao, and Ce Zhang. xfraud: explainable fraud transaction detection. *Proc. VLDB Endow.*, 15(3):427–436, November 2021. ISSN 2150-8097. doi: 10.14778/3494124.3494128. URL <https://doi.org/10.14778/3494124.3494128>.
- 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, pages 14557–14565, 2023.

A Appendix

A.1 Token Feature Definitions

Throughout this appendix, $\epsilon > 0$ denotes a small constant for numerical stability, $\text{clip}(x, a, b) = \min(\max(x, a), b)$, $y_u \in \{0, 1\}$ is the fraud label of session u , $\tau(u)$ its timestamp, ℓ the label lag, t_0 the dataset start date, and $\text{id}_r(u)$ the identifier of type $r \in \mathcal{R}$ associated with session u . We use $\text{acc}(u)$, $\text{dev}(u)$, and $\text{net}(u)$ as shorthand for the account, device, and network identifiers, respectively.

A.2 Label-Propagated (LP) Features

For each session v , let

$$\mathcal{L}^-(v) = \bigcup_{r \in \mathcal{R}} \mathcal{L}_r^-(v) \quad (16)$$

denote the set of direct predecessors with adjudicated labels available at score time, where $\mathcal{L}_r^-(v)$ is defined in Section 2. Four aggregate statistics are computed and prepended to $\mathbf{x}_v$:

$$n_\ell(v) = |\mathcal{L}^-(v)| \quad \text{number of labeled predecessors,} \quad (17)$$

$$n_f(v) = \sum_{u \in \mathcal{L}^-(v)} y_u \quad \text{count of fraudulent predecessors,} \quad (18)$$

$$\rho_f(v) = \frac{n_f(v)}{\max(n_\ell(v), 1)} \quad \text{local fraud rate,} \quad (19)$$

$$\mathbb{1}_f(v) = \mathbf{1}[n_f(v) > 0] \quad \text{any-fraud indicator.} \quad (20)$$

A.3 Identifier Token Features

For relation type r and identifier a , each time-conditioned identifier state $i_{r,a}^{(t)}$ carries a 25-dimensional feature vector representing the information available for identifier a at time t . Let

$$\mathcal{A}_{r,a}(t) = \{u \in \mathcal{S} : \text{id}_r(u) = a, \tau(u) < t\} \quad (21)$$

denote all sessions observed through identifier a before t , and let

$$\mathcal{A}_{r,a}^\ell(t) = \{u \in \mathcal{A}_{r,a}(t) : \tau(u) + \ell \leq t\} \quad (22)$$

denote the lag-safe subset. For readability, we suppress the identifier index a from the feature symbols below.

A.4 Fraud rates (lag-safe, 4 features):

$$\rho_r^{(w)}(t) = \frac{\sum_{u \in \mathcal{A}_{r,a}^\ell(t), \tau(u) \geq t-w} y_u}{\max(|\{u \in \mathcal{A}_{r,a}^\ell(t) : \tau(u) \geq t-w\}|, 1)}, \quad w \in \{7, 30, 90, \infty\} \text{ days.} \quad (23)$$

Absolute fraud and clean counts (lag-safe, 2 features):

$$n_f^r(t) = \sum_{u \in \mathcal{A}_{r,a}^\ell(t)} y_u, \quad n_c^r(t) = |\mathcal{A}_{r,a}^\ell(t)| - n_f^r(t). \quad (24)$$

Session volume (no lag, 4 features):

$$c_r^{(w)}(t) = |\{u \in \mathcal{A}_{r,a}(t) : \tau(u) \geq t-w\}|, \quad w \in \{7, 30, 90, \infty\} \text{ days.} \quad (25)$$

Maturity (no lag, 2 features):

$$\text{days_active}_r(t) = t - \min_{u \in \mathcal{A}_{r,a}(t)} \tau(u), \quad (26)$$

$$\text{recency}_r(t) = t - \max_{u \in \mathcal{A}_{r,a}(t)} \tau(u). \quad (27)$$

Co-identifier diversity (no lag, 6 features): Let r' and r'' be the two identifier types other than r . For window $w \in \{7, 30, 90\}$ days, count distinct co-occurring identifiers of type r' observed alongside a :

$$d_{r,r'}^{(w)}(t) = |\{\text{id}_{r'}(u) : u \in \mathcal{A}_{r,a}(t), \tau(u) \geq t - w\}|. \quad (28)$$

Features are computed for both r' and r'' , giving $2 \times 3 = 6$ features.

Fraud-linked co-identifier counts (lag-safe, 4 features): For window $w \in \{30, 90\}$ days and each of the two other identifier types r', r'' :

$$d_{r,r'}^{f,(w)}(t) = |\{\text{id}_{r'}(u) : u \in \mathcal{A}_{r,a}^\ell(t), y_u = 1, \tau(u) \geq t - w\}|. \quad (29)$$

Coverage (3 features):

$$\text{cov}_r^{(w)}(t) = \frac{\min(\text{days_active}_r(t), w)}{w}, \quad w \in \{7, 30, 90\} \text{ days}. \quad (30)$$

A.5 Cross-Identifier Session Features

Each session node v is further augmented with 7 cross-identifier features appended to $\mathbf{x}_v$, capturing pairwise relational risk signals across identifier types. Let

$$\mathcal{P}_{r,r'}(v, t) = \{u \in \mathcal{S} : \text{id}_r(u) = \text{id}_r(v), \text{id}_{r'}(u) = \text{id}_{r'}(v), \tau(u) < t\} \quad (31)$$

denote sessions sharing both identifiers r and r' with v before t , and let

$$\mathcal{P}_{r,r'}^\ell(v, t) = \{u \in \mathcal{P}_{r,r'}(v, t) : \tau(u) + \ell \leq t\} \quad (32)$$

denote the lag-safe subset. W denotes the lookback window for pair-level features. For quantities inherited from the identifier features above, the identifier for relation r is $a_r = \text{id}_r(v)$.

$$\rho_{r,r'}(v, t) = \frac{\sum_{u \in \mathcal{P}_{r,r'}^\ell(v, t)} y_u}{\max(|\mathcal{P}_{r,r'}^\ell(v, t)|, 1)}, \quad (r, r') \in \{(\text{acc}, \text{dev}), (\text{acc}, \text{net}), (\text{dev}, \text{net})\}, \quad (33)$$

$$n_f^{\text{pair}}(v, t) = \sum_{u \in \mathcal{P}_{\text{acc}, \text{dev}}^\ell(v, t)} y_u, \quad (34)$$

$$\rho_{\max}(v, t) = \max_{r \in \mathcal{R}} \rho_r^{(90)}(t), \quad (35)$$

$$\mathbb{K}_{\text{any}}(v, t) = \mathbf{1}[\exists r \in \mathcal{R} : n_f^r(t) > 0], \quad (36)$$

$$\text{cov}_{\text{pair}}(v, t) = \frac{\min(t - t_0, W)}{W}. \quad (37)$$

A.6 Ecosystem Climate Token Features

Climate tokens are precomputed daily at date t . Let $\mathcal{D}(t, w) = \{u \in \mathcal{S} : t - w \leq \tau(u) < t\}$ denote all sessions in a w -day window before t , and let $\mathcal{D}^\ell(t, w) = \{u \in \mathcal{D}(t, w) : \tau(u) + \ell \leq t\}$ denote the lag-safe subset. The dataset start date t_0 is used for coverage normalization. For platform $p \in \{P_1, P_2\}$, let $\mathcal{D}_p(t, w) = \{u \in \mathcal{D}(t, w) : \text{platform}(u) = p\}$ and let $\mathcal{D}_p^\ell(t, w)$ denote the corresponding lag-safe subset. Define the global fraud rate

$$\rho(t, w) = \frac{\sum_{u \in \mathcal{D}^\ell(t, w)} y_u}{\max(|\mathcal{D}^\ell(t, w)|, 1)}$$

and the platform fraud rate

$$\rho_p(t, w) = \frac{\sum_{u \in \mathcal{D}_p^\ell(t, w)} y_u}{\max(|\mathcal{D}_p^\ell(t, w)|, 1)}.$$

Token G — Global fraud climate (12 features):

$$\rho_w^G = \rho(t, w), \quad w \in \{7, 14, 30, 90\} \text{ days}, \quad (38)$$

$$\nu_w^G = \log(1 + |\mathcal{D}(t, w)|), \quad w \in \{7, 30\} \text{ days}, \quad (39)$$

$$\delta_{7/30}^G = \rho(t, 7)/(\rho(t, 30) + \epsilon) - 1, \quad \delta_{30/90}^G = \rho(t, 30)/(\rho(t, 90) + \epsilon) - 1, \quad (40)$$

$$a^G = \delta_{7/30}^G - \delta_{30/90}^G \quad (\text{fraud acceleration}), \quad (41)$$

$$\text{cov}_w^G = \min(t - t_0, w)/w, \quad w \in \{7, 30, 90\}. \quad (42)$$

Tokens P_1, P_2 — Platform-specific climate (10 features each): For platform p :

$$\rho_w^p = \rho_p(t, w), \quad w \in \{7, 30, 90\}, \quad (43)$$

$$\sigma_7^p = |\mathcal{D}_p(t, 7)| / \max(|\mathcal{D}(t, 7)|, 1) \quad (\text{session share}), \quad (44)$$

$$\phi_7^p = \sum_{u \in \mathcal{D}_p^\ell(t, 7)} y_u / \max\left(\sum_{u \in \mathcal{D}^\ell(t, 7)} y_u, 1\right) \quad (\text{fraud share}), \quad (45)$$

$$\delta_{7/30}^p = \rho_p(t, 7)/(\rho_p(t, 30) + \epsilon) - 1, \quad \nu_7^p = \log(1 + |\mathcal{D}_p(t, 7)|), \quad (46)$$

$$\text{cov}_w^p = \min(t - t_0, w)/w, \quad w \in \{7, 30, 90\}. \quad (47)$$

Token X — Cross-platform shift (5 features): Let

$$\phi_w^p = \frac{\sum_{u \in \mathcal{D}_p^\ell(t, w)} y_u}{\max(\sum_{u \in \mathcal{D}^\ell(t, w)} y_u, 1)}$$

denote the share of total fraud attributable to platform p in window w , and let

$$\sigma_w^p = |\mathcal{D}_p(t, w)| / \max(|\mathcal{D}(t, w)|, 1)$$

denote its session share.

$$\xi_w = \rho_{P_1}(t, w)/(\rho_{P_2}(t, w) + \epsilon), \quad w \in \{7, 30\}, \quad (48)$$

$$\gamma_w = \phi_w^{P_1} - \sigma_w^{P_1}, \quad w \in \{7, 30\} \quad (\text{fraud concentration shift}), \quad (49)$$

$$H_7 = -\sigma_7^{P_1} \log(\sigma_7^{P_1} + \epsilon) - (1 - \sigma_7^{P_1}) \log(1 - \sigma_7^{P_1} + \epsilon). \quad (50)$$

Token V — Velocity and anomaly (7 features):

$$z_\rho^{(w)} = \text{clip}\left(\frac{\rho(t, w) - \rho(t, 90)}{\sqrt{\rho(t, 90)(1 - \rho(t, 90)) + \epsilon}}, -5, 5\right), \quad w \in \{7, 30\}, \quad (51)$$

$$z_\nu^{(7)} = \text{clip}\left(\frac{|\mathcal{D}(t, 7)|/7 - |\mathcal{D}(t, 90)|/90}{\sqrt{|\mathcal{D}(t, 90)|/90 + \epsilon}}, -5, 5\right), \quad (52)$$

$$z_f^{(7)} = \text{clip}\left(\frac{f_7/7 - f_{90}/90}{\sqrt{f_{90}/90 + \epsilon}}, -5, 5\right), \quad f_w = \sum_{u \in \mathcal{D}^\ell(t, w)} y_u, \quad (53)$$

$$b^G = \text{clip}\left(\frac{\max_{k=0}^6 \rho(t - k, 1)}{\rho(t, 30) + \epsilon}, 0, 20\right) \quad (\text{fraud burst}), \quad (54)$$

$$\text{cov}_w^V = \min(t - t_0, w)/w, \quad w \in \{7, 30\}. \quad (55)$$

Token E — Ecosystem topology (10 features): Let $\mathcal{X}_r(t, w)$ be the set of unique identifiers of type r observed in $\mathcal{D}(t, w)$, and let $\mathcal{X}_r^{\text{new}}(t, w) \subseteq \mathcal{X}_r(t, w)$ denote those first

observed within $[t - w, t)$.

$$\eta_r^{(w)} = |\mathcal{X}_r^{\text{new}}(t, w)| / \max(|\mathcal{X}_r(t, w)|, 1), \quad r \in \{\text{dev}, \text{acc}\}, \quad w = 7, \quad (56)$$

$$\bar{s}_{\text{dev}}(t) = \text{mean}_{i \in \mathcal{X}_{\text{dev}}(t, 7)} |\{j \in \mathcal{X}_{\text{acc}}(t, 7) : \exists u, \text{dev}(u) = i \wedge \text{acc}(u) = j\}|, \quad (57)$$

$$\kappa_{\text{net}}(t) = \sum_{\text{top-5}} c_i / |\mathcal{D}(t, 7)|, \quad c_i = |\{u \in \mathcal{D}(t, 7) : \text{net}(u) = i\}|, \quad (58)$$

$$\psi(t) = |\mathcal{X}_{\text{acc}}(t, 7)| / \max(|\mathcal{X}_{\text{dev}}(t, 7)|, 1), \quad (59)$$

$$\Delta \eta_{\text{dev}} = \text{clip}\left(\eta_{\text{dev}}^{(7)} / (\eta_{\text{dev}}^{(30)} + \epsilon) - 1, -2, 5\right), \quad (60)$$

$$\Delta \kappa_{\text{net}} = \text{clip}(|\mathcal{X}_{\text{net}}(t, 7)| / (7|\mathcal{X}_{\text{net}}(t, 30)| / 30 + \epsilon) - 1, -2, 5), \quad (61)$$

$$\text{cov}_w^E = \min(t - t_0, w) / w, \quad w \in \{7, 30, 90\}. \quad (62)$$

Token R — Risk regime (7 features): Let

$$\mathcal{F}^\ell(t) = \{u \in \mathcal{S} : \tau(u) + \ell \leq t, y_u = 1\}$$

be all confirmed fraud sessions before t , and define

$$h_r^{(w)}(t) = \frac{|\{u \in \mathcal{D}(t, w) : \text{id}_r(u) \in \{\text{id}_r(f) : f \in \mathcal{F}^\ell(t)\}\}|}{\max(|\mathcal{D}(t, w)|, 1)}$$

as the fraction of sessions in window w that share an identifier of type r with any known-fraud session.

$$\lambda(t) = |\mathcal{D}^\ell(t, 30)| / \max(|\mathcal{D}(t, 30)|, 1) \quad (\text{label coverage}), \quad (63)$$

$$h_r^{(7)}(t) \quad r \in \{\text{net}, \text{dev}\} \quad (\text{high-risk identifier share, 7-day}), \quad (64)$$

$$\chi_{\text{acc}}(t) = h_{\text{acc}}^{(7)}(t) \quad (\text{repeat-fraud account rate}), \quad (65)$$

$$\Delta h_{\text{dev}} = \text{clip}\left(h_{\text{dev}}^{(7)}(t) / (h_{\text{dev}}^{(30)}(t) + \epsilon) - 1, -2, 5\right) \quad (\text{contamination trend}), \quad (66)$$

$$\text{cov}_w^R = \min(t - t_0, w) / w, \quad w \in \{7, 30\}. \quad (67)$$