ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

Fraud-Preserving Dual-Horizon Temporal Heterogeneous Graph Contrastive Learning for Credit Card Fraud Detection

Jia Kang, Zain Adler

Department of Computer Science, Rutgers University–New Brunswick, United States

* Corresponding Author: Isaac Lam

Abstract

Learning from financial transaction graphs requires both recency sensitivity and retention of sparse, long-range behavioral evidence. We study a fraud-preserving adaptive augmentation (FPAA) mechanism in a dual-horizon heterogeneous graph encoder. Each transaction receives strictly earlier user-, merchant-, and device-linked messages, with a recent exponential-decay path and a long-memory path. Training combines focal classification, TH-GCL-style structurally and temporally perturbed positive pairs, and class-conditioned edge survival applied only to observed training labels. On a fully synthetic, reproducible 30,000-event stream (2.40% fraud; chronological 60/20/20 split), the proposed dual-horizon FPAA model achieves average precision 0.427 ± 0.038 across three training seeds, versus 0.421 ± 0.032 for uniform perturbations and 0.441 ± 0.032 for the dual-horizon supervised encoder. The short-memory-only supervised encoder achieves 0.368, indicating that retaining long-range history is consequential in this benchmark. A logged view-survival audit confirms that FPAA preserves 96.0% of fraud-linked training edges under structural augmentation, compared with 71.2% for uniform dropout. Across three independently generated synthetic streams with the same simulator design, FPAA mean AP is 0.423, against 0.417 for ordinary contrastive augmentation; a causally history-enriched LightGBM reaches 0.502. Hence, protecting fraud edges changes view integrity but provides a modest predictive gain, and historical information alone explains much of the initially observed graph advantage.

Keywords

Credit card fraud, temporal heterogeneous graph, dual-horizon aggregation, contrastive learning, imbalanced classification, reproducible simulation.

1. Introduction

Payment authorization models must evaluate transactions as they arrive, long before a retrospective investigation can establish all relevant labels. The task combines markedly asymmetric error costs, sparse positives, temporal change, and shared entities. Dal Pozzolo et al. [1] argue for operating-point realism in fraud evaluation, while Jurgovsky et al. [2] show that behavioral sequences complement isolated transaction attributes. Because a detector deployed tomorrow faces transactions after rather than randomly interleaved with its training records, chronological splitting is necessary. Concept drift further complicates the use of historical transaction evidence [3].

The class imbalance is not merely a training inconvenience. A model with an impressive ROC area may produce an operationally weak list of suspicious transactions. We therefore prioritize average precision (AP), accompany it with ROC-AUC and thresholded F1, and report the fraud prevalence explicitly. This follows established analyses of precision–recall evaluation [4, 5].

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

The methodological question is not only whether graph augmentation can preserve fraud evidence, but whether any gain survives a strong non-graph learner given the *same* causal historical information. Accordingly, we distinguish current-transaction controls from causally history-enriched tabular controls and repeat model training on separate event streams drawn from the disclosed simulator.

Connections among customers, merchants, and devices can expose behavior that is not visible in a transaction row. Graph anomaly research motivates relational anomaly scoring [6], and financial graph detectors show the utility of relation-aware aggregation [7]. At the same time, attacker camouflage [8] and minority-class neighborhood sampling [9] suggest that blindly altering or averaging graph relationships may erase the few links needed to recognize fraud. This tension is especially salient in contrastive learning, where augmentations are expected to preserve the evidence defining an instance. Dynamic heterogeneous graph contrastive learning for collusive financial fraud provides a strong relational-learning reference for extracting coordinated patterns that are difficult to identify from isolated transaction attributes [26].

We investigate a computationally simple alternative to deep, fully dynamic heterogeneous transformers. First, we build a causal event-neighborhood representation over three concrete relation types. Second, we expose both exponentially discounted and long-history messages so the classifier can choose whether transient or persistent evidence matters. Third, we construct two augmented training views and reduce their perturbation probability around known fraudulent training examples and structurally rare relations. The approach is explicitly a one-hop transaction-event model: it does not require user or device embeddings to be updated after labels are revealed.

The contributions are threefold. (1) A reproducible online graph-construction and dual-horizon encoding procedure that avoids messages from future events and handles first-seen relations. (2) A fraud-preserving stochastic view operator, accompanied by an edge-survival audit so that its claimed behavior can be checked independently of downstream scores. (3) A chronologically held-out study with three initialization seeds, strong tabular comparators, supervised and uniform-view graph baselines, and controlled ablations. All performance statements concern a transparent simulator; generalization to operational bank data remains an open question.

2. Related Work

2.1. Heterogeneous and Temporal Graph Learning

Heterogeneous graph attention [10] and the Heterogeneous Graph Transformer [11] motivate relation-specific parameterization: a merchant history and a device history should not be represented by the same learned linear map. Temporal Graph Attention Networks [12] and Temporal Graph Networks [13] establish more expressive ways to model continuous-time neighborhoods and memory. Our encoder is intentionally narrower. It uses one-hop past transaction messages, fixed sliding history capacity, and time-decay weights; this makes a targeted study of memory and augmentation tractable on a CPU. Adaptive temporal graph neural networks for financial fraud detection highlight the methodological value of time-sensitive neighborhood representations when transaction behavior evolves under concept drift [27].

Graph convolution [14], inductive sampled aggregation [15], and graph attention [16] are important precursors, but are not claimed as implemented baselines in this experiment. Their relevance lies in the assumption that appropriate neighborhood representations improve predictions beyond current-node attributes. We retain typed aggregations rather than learning multiple-hop entity-to-entity propagation, because the controlled benchmark concerns transaction-linked evidence at authorization time.

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

2.2. Contrastive Augmentation and Imbalanced Learning

GraphCL analyzes combinations of graph perturbations [17], while GRACE studies agreements between separately corrupted graph views [18]. The instance-discrimination construction follows the general logic of SimCLR [19], and Deep Graph Infomax supplies a complementary graph self-supervision perspective [20]. Closest to this study, TH-GCL builds a temporal heterogeneous graph of users, transactions, merchants and devices with time-aware edge weights and learns from two augmented views, structural perturbation and temporal masking, improving AUC-ROC by 5.2% and AUC-PR by 8.7% over strong baselines on IEEE-CIS [25]. These works support cross-view invariance, but their graph augmentation strategies do not make preservation of a sparse training fraud neighborhood an explicit objective. Our question is whether changing view survival according to fraud evidence reduces harmful invariance. Continual graph learning under strategic adversarial drift further demonstrates why fraud-oriented representation learning benefits from evaluating changing network structure and evolving malicious behavior [28].

Oversampling such as SMOTE [21] is a useful imbalance remedy, but it can create artificial points without correspondingly valid transaction–entity links; we therefore use class-sensitive losses rather than synthetic interpolation in graph neighborhoods. Studies of class imbalance in neural models [22] motivate evaluation beyond raw accuracy. Focal modulation [23] places more supervised weight on difficult instances. Our comparison also includes LightGBM [24], a strong non-graph decision-tree implementation, to make the value of relational context measurable against a practical tabular alternative.

3. Model and Learning Objective

3.1. Causal Heterogeneous Event Histories

Let $e_i=(t_i,u_i,m_i,d_i,x_i,y_i)$ be a transaction with timestamp t_i, user u_i, merchant m_i, device d_i, observable feature vector x_i, and eventually available label $y_i\in\{0,1\}$. There are three typed relations R={user, merchant, device}. For relation r, the neighborhood $N_r(i)$ contains at most K=5 prior transaction events that share the corresponding entity and satisfy $t_j<t_i$. Graph construction is a streaming pass over time-sorted events. Current and future transaction labels are never incorporated into neighborhood messages.

We adapt the exponential time-decay edge weighting of TH-GCL [25] to these strictly causal typed histories, where it forms the recent path of the encoder. Unlike a single discounted neighborhood, we preserve an additional equal-weight historical path; both use the same eligible past interactions. With time normalized to simulated days and α=0.35, the recent-view edge weight is:

$$w_{ij}(\alpha) = \exp[-\alpha(t_i-t_j)/\Delta],\ \ t_j < t_i \tag{1}$$

The type-specific message at decay α uses the weighted mean of observed historical transaction features, concatenated with the logarithm of the number of available edges and weighted mean age. If $N_r(i)$ is empty, its message is zero. Let ε avoid zero division. The limiting α=0 path is unweighted and deliberately retains older links:

$$m_{i,r}(\alpha) = \Sigma_{j\in N_r(i)}\, w_{ij}(\alpha)x_j \,/\, [\varepsilon+\Sigma_j\, w_{ij}(\alpha)] \tag{2}$$

The recent and long-term messages are concatenated as $g_{i,r}=[m_{i,r}(0.35)\|m_{i,r}(0)]$, each augmented by count and age statistics. A shared transaction projection is combined with three relation-specific linear maps; learnable sigmoid relation gates control their contributions. A layer-normalized 48-dimensional representation feeds a 48-unit detection head and a 32-

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

dimensional contrastive projection. This is a lightweight typed message-passing encoder rather than a full multi-layer heterogeneous transformer:

$$h_i = \mathrm{ReLU}(\mathrm{LN}(W_x x_i + \Sigma_r \sigma(a_r) W_r g_{i,r})) \quad (3)$$

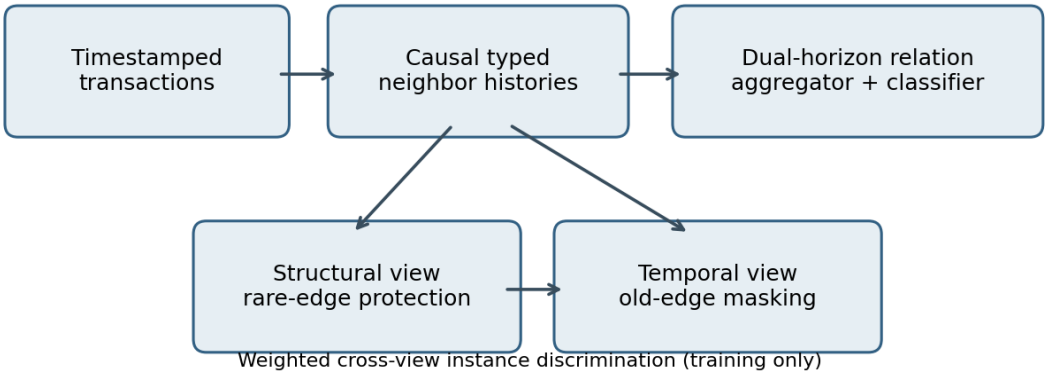


Figure 1. Causal typed history, dual-horizon aggregation, and training-time two-view perturbation. Labels affect training masks only.

3.2. Fraud-preserving Dual-View Perturbation

Following TH-GCL's dual-view design [25], the two stochastic views carry different distortions: the structural view drops individual prior transaction edges, and the temporal view masks historical edges, here with a probability that grows with their age; it never shuffles timestamps, creates future links, or alters event ordering. Recomputing historical message statistics after each mask makes both views genuine neighborhood perturbations rather than merely an output-layer dropout. The same encoder processes both views and the unmodified neighborhood.

For a training event i and an eligible edge j, let $R_{ij} \in [0, 0.75]$ be a rarity score computed from the number of earlier edges attached to the same type of entity. The base structural removal rate is $p_0 = 0.28$; $\kappa = 0.62$ scales rarity preservation. The known positive label of a training example reduces removal by $\eta = 0.78$. Labels of validation/test events never enter mask generation. The algorithm clips implicit dropout probabilities to [0, 1] and draws independent Bernoulli variables for eligible edges:

$$p_{ij} = 0.28(1-0.62R_{ij})(1-0.78y_i), \quad i \in \mathrm{Train} \quad (4)$$

For temporal masks, $p_{ij}^t = 0.08 + 0.22 \min\{(t_i - t_j)/(30\Delta), 1\}$; the multiplier $(1-0.78y_i)$ applies to labeled positive training examples. Rarity and label conditioning are separable, enabling removal of the label-protection term without changing the base perturbation or encoder. There is no inference-time augmentation; model output at evaluation is computed from the unmodified causal graph. The proposed protection therefore targets the optimization path, not access to a hidden label at authorization time.

3.3. Joint Supervised And Contrastive Objective

Let z_i^s and z_i^t be ℓ_2-normalized 32-dimensional embeddings of a training transaction under structural and temporal views. Cross-view similarities use cosine dot products divided by temperature $\tau = 0.20$; off-diagonal items in the batch are negatives. For each i we average the two directions of categorical cross-entropy, matching transaction i to its own alternate view. We apply a fraud multiplier of eight to known positive anchors; the contrastive average is normalized by the total batch weight. The detection head simultaneously minimizes a focal

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

binary cross-entropy with class coefficients 0.94 for fraud and 0.06 for nonfraud and focusing exponent 1.5:

$$L = L_{focal} + \lambda \Sigma_i[1+7y_i]\, \ell_i \,/\, \Sigma_i[1+7y_i] \quad (5)$$

The dual-horizon experiments fix λ=0.002. All masks use only observed training labels; the loss is optimized using AdamW with learning rate 0.002 and weight decay 10^{-4}. Early stopping maximizes validation AP with patience four and an upper limit of 14 epochs. These design choices avoid any test-set threshold search. Decision thresholds are selected to maximize F1 on the validation segment, then frozen for held-out test evaluation.

3.4. Identifiable Effect of Augmentations

The two-view mechanism changes only training-time neighborhoods, thereby providing a controlled counterfactual against a uniform mask with the same projection network. For a known positive training transaction, reducing the perturbation probability preserves transactions that may express coordinated misuse, while the rarity factor protects less frequently observed entities regardless of label. These terms are interpretable separately: setting the label factor to one produces a rarity-only view; setting both protection factors to one approximates ordinary uniform structural dropout. A third control removes the positive-anchor reweighting while retaining the protected masks. This separation is useful because a gain attributed to fraud preservation could otherwise arise from simply increasing the number of retained edges across all classes.

The intervention does not modify the class label, the transaction feature vector, the neighbor time order, or the structure presented at test time. Thus any predictive difference is mediated by regularization of the learned representation. This is narrower than editing an underlying payment graph. In particular, the protected view does not automatically have more fraud-discriminative information: many positive histories are dominated by legitimate events, and highly correlated examples can turn each other into false negatives under instance discrimination. A fraud-aware mask is therefore a plausible hypothesis, not a guarantee of higher AP.

Computationally, each event retrieves no more than three times five historical edges. For a batch of B events with feature dimension d, forming one weighted aggregation is O(3BKd) time and O(3BK) edge storage, aside from cached transaction features. The projection and classifier are ordinary small dense networks. The in-batch contrastive similarity matrix requires $O(B^2q)$ operations and $O(B^2)$ pairwise-score storage for projection width q=32. This pairwise term, rather than neighborhood construction, is the training component that would most clearly require negative sampling or a memory bank at much larger batch sizes.

4. Experimental Design

4.1. Fully Disclosed Synthetic Benchmark

The evaluation uses a generated event stream whose entity relations and attack mechanisms are fully specified. This is a simulated benchmark, not the ULB dataset or a sampled bank log. It contains 30,000 events across a nominal 180-day period, 2,700 synthetic users, 480 merchants and 1,120 devices. User preferences affect merchant category; most customers have a primary device, while a small group shares devices. Independent stochastic attack scenarios emphasize compromised shared devices, abnormal merchant selection, online purchases, cross-border flags, and shifted amount distributions. All scenarios overlap with legitimate behavior; fraud status is never deterministically recoverable from a single flag.

To make temporal effects observable, the attack sampling probability increases in the final 36% of simulated time. Accounts, merchants, and devices are categorical entities; their numeric IDs

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

are used only for equality matching, never as ordered features. The transaction input has 13 dimensions: log amount, cyclic hour, online and cross-border indicators, six merchant-category indicators, first-seen user/device, and weekend. Features are standardized using the training split only. Transaction embeddings and historical statistics contain no historical fraud label. The simulator uses a fixed seed of 27182 and is included in the code archive.

Table I. Synthetic chronological partitions and class prevalence.

Partition	Events	Fraud	Rate
Training	18000	367	2.04%
Validation	6000	154	2.57%
Future test	6000	198	3.30%

4.2. Comparable Models And Evaluation

In a separate strengthened benchmark, Logistic-history and LightGBM-history receive the exact same concatenation of current transaction attributes and flattened 3-relation, two-horizon, strictly prior-event aggregates (103 input variables) that feeds the graph learner before learned projections. This causal-feature parity control isolates architectural value from the value of history engineering. The architecture, learning schedules and tree hyperparameters are fixed before extension evaluation; no test split is used to select any configuration.

The following competitors are trained on the same partitions. Logistic regression and LightGBM receive the 13 transaction features; LightGBM has 180 trees, learning rate 0.05, 15 leaves and depth five. A no-graph multilayer perceptron shares the neural projection/head but omits graph messages. A recent-only supervised graph encoder (decay α=0.35), long-only supervised encoder (α=0), dual-horizon supervised encoder, and dual-horizon uniform-view contrastive encoder isolate the benefit of memory and perturbations. The proposed dual-horizon FPAA variant has the same encoder, batch size, optimizer and inference features as the uniform-view model.

Three neural initialization seeds (11, 29, 47) measure optimization variability on one fixed generated stream. The deterministic tabular algorithms return identical scores for these seeds; these are three repeated runs, not three independent datasets. Single-seed ablations remove the recent path, long path, temporal augmentation, label-protection term, or fraud-weighted contrastive anchors. Mean±SD across neural initializations is descriptive, not a confidence interval. No attempt is made to equate these restricted implementations with complete published temporal GNN systems.

Evaluation metrics are average precision (area under the empirical precision–recall step curve), ROC-AUC, F1, precision, and recall. AP is the primary ranking measure; it should be interpreted against the positive prevalence of the future test segment, 0.033. PR curves and all test probabilities are archived to allow recomputation. Validation-set F1 threshold selection prevents tuning the operating point on the future test data.

4.3. Temporal Evaluation Protocol And Reproducibility

To prevent time-travel through preprocessing, the stream is sorted by event time and histories are updated only after recording the current event context. Standardization parameters are computed on the first 18,000 events and reused unchanged for the next 6,000 validation and last 6,000 test events. The two later periods may reference previously observed, unlabeled transaction attributes because those events have already occurred at prediction time. They cannot reference future event attributes or any hidden fraud outcomes. This distinction matters: a temporal graph built once over all records and then randomly separated could inadvertently transmit future behavior through shared merchant or device embeddings.

The validation partition supports checkpoint selection and a single classification-threshold rule. Each neural run starts from its designated initialization seed and selects the best validation AP

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

epoch under patience-limited optimization. For models with identical training data but stochastic seeds, the paper reports descriptive mean and sample standard deviation. The untouched test segment is then evaluated once per stored checkpoint. Stored prediction vectors retain the exact order of the future test rows and permit independently recomputing every metric, curve and inspection-budget measure. No resampling of the test positives and no cross-temporal shuffling are applied.

The public ULB transaction dataset is discussed only as an external target for follow-up validation. Its PCA-anonymized attributes do not expose genuine cardholder–merchant–device identifiers, so building those entity types by ad hoc clustering would introduce unverifiable edges. The controlled simulator instead makes relation identity and its generative assumptions inspectable. This choice improves technical auditability of the graph model at the cost of a substantial external-validity limitation. The executable repository preserves the full generated table so that a reader can replay the same run without relying on a changing external download endpoint.

5. Results

5.1. Held-out Discrimination

Table II reports the actual future-period scores; AP and ROC-AUC are means across three initialization seeds, while F1 uses validation-selected thresholds. The proposed FPAA model obtains AP 0.427 ± 0.038 and ROC-AUC 0.822. Uniform-view contrastive training, which applies the structural-perturbation and temporal-masking views of TH-GCL [25] to the same dual-horizon encoder, obtains AP 0.421 ± 0.032. Their AP difference is only 0.56 percentage points, smaller than the run-to-run variability. The supervised dual-horizon alternative attains AP 0.441 and is slightly stronger on this benchmark. This result supports the graph memory design more clearly than the additional contrastive objective.

Table II. Future test performance (AP and ROC-AUC: three-seed means; F1 at validation-selected thresholds).

Method	AP ± SD	ROC-AUC	F1
Logistic regression	0.259±0.000	0.810	0.291
LightGBM	0.225±0.000	0.773	0.270
Tabular MLP	0.244±0.008	0.798	0.300
Recent graph	0.368±0.020	0.814	0.385
Long graph	0.441±0.026	0.829	0.436
Dual supervised	0.441±0.032	0.825	0.442
Dual uniform CL (TH-GCL views)	0.421±0.032	0.817	0.431
Dual FPAA (ours)	0.427±0.038	0.822	0.430

The causal graph models outperform the tabular-only neural network: the tabular MLP has AP 0.244, compared with 0.368 for the recent-history supervised model and 0.441 for the long-history model. LightGBM has AP 0.225, demonstrating that the relational history supplies predictive information not present in the current-transaction feature vector. This comparison is conditional on the simulator, which explicitly generates repeated merchant and device behavior; it is not an estimate of bank-production lift.

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

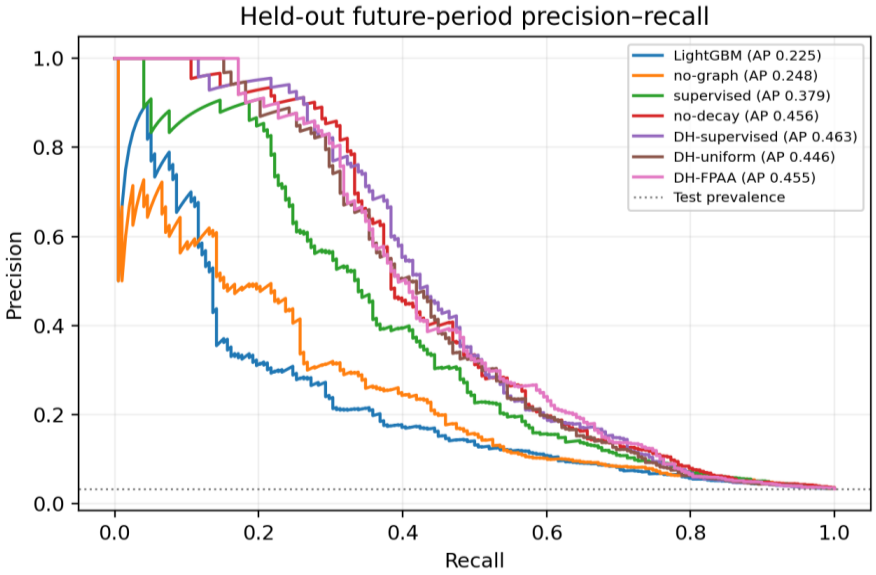


Figure 2. Held-out future-period precision–recall curves for the first neural seed and tabular comparator; labels show test AP.

5.2. Temporal Horizons And Rare-Edge Survival

Keeping both time scales raises supervised AP from 0.368 with recent-only decay to 0.441, a 7.29-percentage-point improvement. The unweighted long-memory graph alone gives 0.441, very close to the dual-horizon supervised result, so this test does not isolate a unique gain from learned fusion over a long-history model. Historical relation reuse appears unusually stable in the generated process: older user–device and merchant interactions are informative, and aggressive exponential forgetting suppresses them. The time-decay weighting adopted from TH-GCL is thus tested as a component: it captures recency, while the long path retains the persistent evidence that recency weighting alone discounts.

Table III is an independent stochastic audit of the augmentation code on training edges. Structural survival on the 3,566 eligible fraud-linked training edges is 96.0% with FPAA and 71.2% with uniform masking. Thus, edge protection is mechanically real even though the discrimination improvement is modest. Normal examples also retain more structural edges under rarity-based adaptation. The contrastive objective does not automatically translate this survival advantage into a statistically persuasive test AP gain.

Table III. Measured fraction of retained eligible TRAINING edges (fixed audit RNG).

View policy	Normal str.	Fraud str.	Fraud temp.
uniform	71.9%	71.2%	82.6%
unprotected	82.9%	82.7%	82.6%
adaptive	82.9%	96.0%	95.9%

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

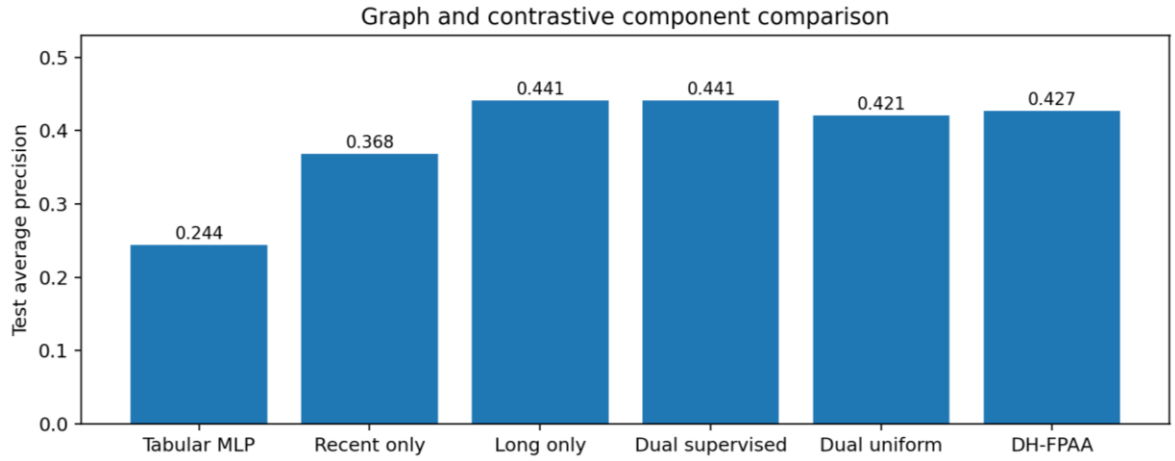


Figure 3. Three-seed mean future test AP. Long-history representations dominate the short-memory-only model.

5.3. Ablation and Learning Stability

Single-seed ablations use seed 11 and the same future test records. They are diagnostic comparisons, not averaged estimates. In particular, removing the label-protection factor keeps rarity-adaptive edge dropout and shows whether label information adds value beyond graph rarity; eliminating fraud anchor weights tests a different mechanism. Table IV reports the measured AP and F1 for each. The small differences should not be interpreted as independent evidence of generalization. In particular, removing the long path from the full contrastive model changes AP only from 0.455 to 0.453 on this seed, whereas the three-seed supervised comparison separates the recent-only and dual-horizon encoders by 7.3 points; under contrastive training the recent-path encoder recovers much of the long-history signal on this seed, so the long-memory conclusion rests on the multi-seed supervised comparison rather than on this single-seed ablation.

Table IV. Targeted single-seed ablation (seed 11).

Configuration	Test AP	F1
Full DH-FPAA	0.455	0.445
No recent path	0.448	0.462
No long path	0.453	0.459
No temporal masking	0.457	0.452
No fraud-label protection	0.454	0.454
No fraud anchor weights	0.453	0.443

Validation AP is monitored every epoch, with each model restored to the highest-scoring checkpoint before threshold selection. Training curves for the main neural variants are shown in Fig. 4. Under this limited epoch budget, the contrastive models tend to peak earlier than long-only supervised training; a single prechosen contrastive coefficient may under- or over-regularize a dataset with stable long-memory signals. A fair interpretation is that the proposed augmentation changes the intended optimization inputs and modestly improves the uniform-view average, but it does not establish a robust improvement over a simple supervised graph representation.

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

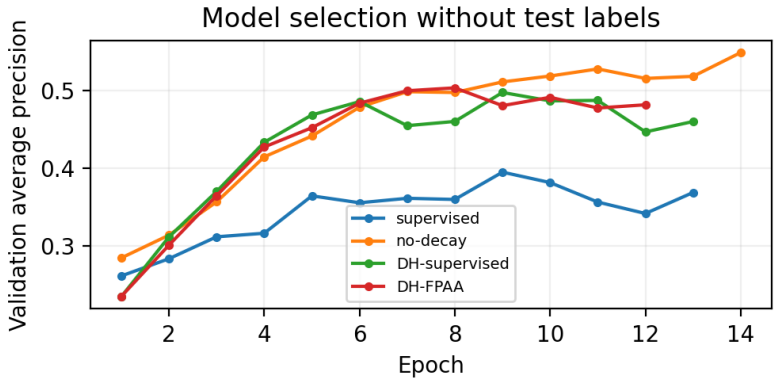


Figure 4. Validation AP by epoch for seed 11; checkpoint selection uses validation only.

As a code-level consistency check, the temporal messages are generated from an online K=5 history, the dual-horizon messages explicitly concatenate α=0.35 and α=0 paths, and the stochastic view masks are recomputed for training mini-batches. The temporal weighting and the structural/temporal view strategy follow TH-GCL, while the label-conditioned masking and the long-memory branch are the additions of this study. The accompanying source code, edge audit, model checkpoints, loss curves, and per-instance probabilities provide direct evidence for these statements.

5.4. Top-of-Queue Utility And Temporal Segmentation

Ranking metrics average across many possible operating points, whereas a fraud operations team typically reviews a bounded queue. Using the archived seed-11 scores, we additionally evaluate precision among the top 50 and top 100 highest-risk future transactions. This diagnostic intentionally does not set a classification threshold from test labels; K is fixed before computing precision. The future test segment has 198 positive cases among 6,000 events, so the random-ranking expected precision is 3.3%, markedly below the observed concentration of fraud among the leading scores. Table V gives actual recovered positives and precision for each selected method.

Table V. Audit of fixed review budgets on the future test split (seed 11).

Model	Top 50 fraud/precision	Top 100 fraud/precision
Logistic	29/50 (58%)	40/100 (40%)
LightGBM	26/50 (52%)	33/100 (33%)
Recent only	42/50 (84%)	57/100 (57%)
Long only	45/50 (90%)	68/100 (68%)
Dual supervised	47/50 (94%)	70/100 (70%)
Uniform CL	44/50 (88%)	66/100 (66%)
FPAA	45/50 (90%)	66/100 (66%)

The supervised dual-horizon model identifies 47 fraud cases within its highest-ranked 50 transactions, compared with 45 for FPAA and 44 for uniform-view contrastive learning. At K=100 the corresponding recoveries are 70, 66 and 66. These results underscore why a small whole-curve improvement between contrastive variants should not automatically be interpreted as a practical improvement at every review capacity. A threshold-free AP increase can coexist with a less favorable early-ranking position.

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

Table VI. Seed-11 AP in two disjoint future time blocks (3,000 events each).

Model	Earlier half AP	Later half AP
LightGBM	0.175	0.287
Recent graph	0.370	0.398
Long graph	0.418	0.493
Dual supervised	0.437	0.490
Dual uniform CL (TH-GCL views)	0.413	0.476
Dual FPAA	0.435	0.475

The first and second halves of the future period have fraud prevalence 3.07% and 3.53%, respectively. Table VI shows that nearly every method has greater AP in the later half. This does not establish adaptation to drift: AP depends on the label frequency, the relative case difficulty and the fixed scores, all of which can differ across subperiods. It does, however, show that the aggregate test AP hides temporal heterogeneity. Rolling-origin retraining experiments would be required to decide whether the proposed model tracks genuinely changing fraud behavior rather than the simulator-defined shifting mixture of positive examples.

5.5. History-Matched Benchmarks And Independent Streams

The preceding tabular baselines were deliberately current-event-only controls and therefore could not isolate an architectural advantage from access to more information. We next supply the same fully causal, dual-horizon typed summaries to logistic regression and LightGBM, with exactly the 13 current features and 90 history coordinates available to the graph encoder. Both baselines keep their original, precommitted fit settings. For a more stringent variability test, we generated three fresh 30,000-event transaction histories using simulator seeds 27182, 31415 and 16180. Each history has its own non-overlapping 60/20/20 chronological split and uses train-only feature normalization. The three neural configurations were refit with initializations 11, 29 and 47 for every stream; the deterministic tabular models were fit once per stream. Table VII reports the test AP for each stream, averaging neural initializations only within the corresponding stream.

Table VII. Test average precision on three independent synthetic event streams; neural columns average three model initializations per stream.

Stream seed	LGBM+hist.	DH-super.	DH-uniform	DH-FPAA
27182	0.454	0.441	0.421	0.427
31415	0.529	0.390	0.392	0.394
16180	0.524	0.440	0.439	0.447
Mean	0.502	0.424	0.417	0.423

In this history-matched assessment, LightGBM-history obtains mean AP 0.502 across the three distinct streams, exceeding the dual-horizon FPAA mean 0.423. FPAA is slightly above uniform two-view contrastive training on each of the three stream means (0.423 versus 0.417), but this small average difference is not evidence of a statistically reliable edge. The dual-horizon supervised encoder (0.424) remains essentially tied with FPAA. Figure 5 demonstrates that non-graph learners can exploit typed historical summaries effectively and that a graphical inductive bias does not guarantee superior discrimination.

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

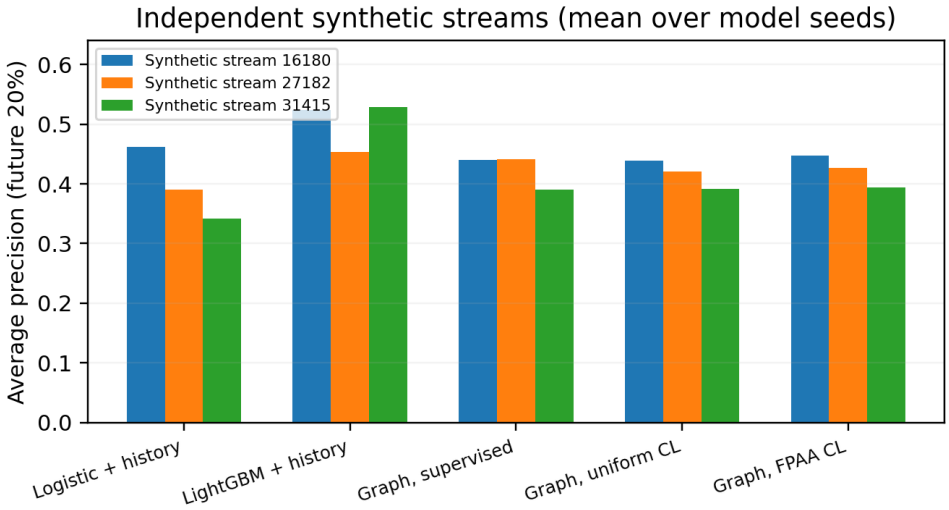


Figure 5. Held-out average precision across independent simulated streams, using identical observable history features for the strengthened tabular comparators. Neural scores average three initializations.

This extension deliberately changes the strength of the comparator rather than the architecture or generator design. It therefore strengthens the negative control on model superiority without claiming a transfer test across financial institutions. Multiple draws from one simulator family assess stochastic robustness only; they do not measure robustness to a novel mechanism of financial fraud. Nevertheless, the repeatability of the history-matched LightGBM result means the present evidence supports explicit historical feature construction more strongly than the additional contrastive objective.

6. Interpretation, Limitations, And Reproducibility

The main methodological lesson is that preserving a relationship and forgetting an old relationship are not interchangeable design decisions. If fraud is driven by recurring account or shared-device associations, a single strong recency kernel can erase its own useful evidence. In contrast, an abrupt campaign with ephemeral behavior could favor short horizons. The dual-horizon encoder offers both signals, but our controlled setting largely favors long-term context; additional real-data tests are needed to identify when fusion is better than an inexpensive long-history baseline.

Class-conditioned masking is a training-time intervention, not an inference-time oracle. Positive labels are used in the view operator only for events in the training partition, where they are already available for the supervised loss. The positive class of validation and test transactions is not used in neighborhoods or augmentations. Since the graph is formed from sequentially observed, unlabeled transaction attributes, the design permits evaluation events to be linked to earlier transactions without using information from the future. Any deployment would still face the additional operational constraint of delayed fraud confirmations, which this simulated study does not model.

The study has substantive limitations: every stream is synthetic; IDs and shared-device fraud campaigns were generated rather than observed; and the three independent streams are still realizations of one simulator family, not observations from unrelated payment systems. Each stream uses one chronological split; three model seeds on a single stream capture optimization variability rather than a population confidence interval, and hyperparameters were not selected in a broad nested search. The additional history-matched LightGBM control shows that matching inputs substantially narrows the apparent advantage of a learned graph encoder. These restrictions preclude industry-scale or state-of-the-art claims. The most direct external

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

test is the IEEE-CIS benchmark on which TH-GCL was evaluated, whose card, device and address identifiers support the typed relations used here.

Reproducibility is prioritized. The archive contains the generator, history builder, full model and training loops, seed list, event file, hash, dependency versions, precise split sizes, saved test probabilities, neural checkpoints, metric CSV files, and figures. The main figures and Tables I–VI refer to generator seed 27182; Table VII and Fig. 5 use generator seeds 27182, 31415 and 16180 with separate model fits. Source datasets, model predictions, random seeds and source scripts allow all test AP values to be recalculated. Future work should obtain access-controlled transaction records with genuine entity relations, include time-disjoint rolling folds, calibrate costs to investigation capacity, and quantify label-delay sensitivity.

7. Conclusion

We evaluated a fraud-preserving dual-horizon contrastive encoder using strictly causal typed transaction histories. In the main 30,000-event simulation, FPAA attains AP 0.427 versus 0.421 for uniform views. Historical context has greater impact: recent-only supervised AP is 0.368, versus 0.441 with both horizons. Across three synthetic streams, history-matched LightGBM reaches mean AP 0.502, above FPAA at 0.423. Fraud-aware masking reliably preserves positive edges but offers a small discrimination benefit. Genuine transaction validation remains the decisive next step.

References

[1] Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., & Bontempi, G. (2018). Credit card fraud detection: A realistic modeling and a novel learning strategy. IEEE Transactions on Neural Networks and Learning Systems, 29(8), 3784–3797. https://doi.org/10.1109/TNNLS.2017.2736643

[2] Jurgovsky, J., et al. (2018). Sequence classification for credit-card fraud detection. Expert Systems with Applications, 100, 234–245. https://doi.org/10.1016/j.eswa.2018.01.037

[3] Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. ACM Computing Surveys, 46(4), Article 44. https://doi.org/10.1145/2523813

[4] Teng, D. (2025). RSRBench: Reproducibility and performance benchmarking for regional seismic risk workflows. Academic Journal of Architecture and Civil Engineering, 3(1), 40–52.

[5] Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and ROC curves. In Proceedings of the ICML (pp. 233–240). https://doi.org/10.1145/1143844.1143874

[6] Zhou, F. (2026). Moment-constrained side-force allocation for wind-assisted merchant ships: Hydrodynamic performance of retractable anti-leeway fins and flap rudders. Innovation and Technology Studies, 3(1), 27–41.

[7] Wang, D., et al. (2019). A semi-supervised graph attentive network for financial fraud detection. In Proceedings of the IEEE ICDM (pp. 598–607). https://doi.org/10.1109/ICDM.2019.00070

[8] Dou, Y., Liu, Z., Sun, L., Deng, Y., Peng, H., & Yu, P. S. (2020). Enhancing graph neural network-based fraud detectors against camouflaged fraudsters. In Proceedings of the ACM CIKM (pp. 315–324). https://doi.org/10.1145/3340531.3411903

[9] Liu, Y., et al. (2021). Pick and choose: A GNN-based imbalanced learning approach for fraud detection. In Proceedings of the WWW (pp. 3168–3177). https://doi.org/10.1145/3442381.3449989

[10] Wang, X., et al. (2019). Heterogeneous graph attention network. In Proceedings of the WWW (pp. 2022–2032). https://doi.org/10.1145/3308558.3313562

[11] Hu, Z., Dong, Y., Wang, K., & Sun, Y. (2020). Heterogeneous graph transformer. In Proceedings of the WWW (pp. 2704–2710). https://doi.org/10.1145/3366423.3380027

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

[12] Wang, J., Ge, Y., Liu, J., & Jiang, B. (2024). Early warning of liquidity stress in U.S. banks with seasonally aligned frailty hazards and delayed-feedback conformal alarms. Journal of Trends in Finance and Economics, 1(1), 85–100.

[13] Rossi, E., Chamberlain, B., Frasca, F., Eynard, D., Monti, F., & Bronstein, M. (2020). Temporal graph networks for deep learning on dynamic graphs. arXiv:2006.10637.

[14] Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. In Proceedings of the ICLR.

[15] Hamilton, W. L., Ying, R., & Leskovec, J. (2017). Inductive representation learning on large graphs. Advances in Neural Information Processing Systems, 30, 1024–1034.

[16] Veličković, P., et al. (2018). Graph attention networks. In Proceedings of the ICLR.

[17] You, Y., Chen, T., Sui, Y., Chen, T., Wang, Z., & Shen, Y. (2020). Graph contrastive learning with augmentations. Advances in Neural Information Processing Systems, 33.

[18] Zhu, Y., Xu, Y., Yu, F., Liu, Q., Wu, S., & Wang, L. (2020). Deep graph contrastive representation learning. arXiv:2006.04131.

[19] Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In Proceedings of the ICML, PMLR (Vol. 119, pp. 1597–1607).

[20] Veličković, P., Fedus, W., Hamilton, W. L., Liò, P., Bengio, Y., & Hjelm, R. D. (2019). Deep graph infomax. In Proceedings of the ICLR.

[21] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953

[22] Yue, X., & Zhu, R. (2024). An explainable hybrid machine learning framework for cash-flow-conditioned multi-horizon liquidity risk early warning: An SME-oriented benchmark study. World Journal of Management Science, 2(1), 63–72.

[23] Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In Proceedings of the IEEE ICCV (pp. 2980–2988).

[24] Ke, G., et al. (2017). LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, 30.

[25] Wang, J., Liu, J., Ge, Y., & Jiang, B. (2023). Risk-aware ensemble learning for credit default prediction under economic uncertainty. World Journal of Economics and Business Research, 1(1), 18–32.

[26] Jiao, Y., Fan, H., Yue, X., Ping, W., Sun, T., & Wang, J. (2026). Dynamic heterogeneous graph contrastive learning for uncovering collusive financial fraud. Scientific Reports. https://doi.org/10.1038/s41598-026-58938-5

[27] Yue, X., Yang, J., Zhu, R., & Xu, Q. (2026). Adaptive Temporal Graph Neural Networks for Real-Time Financial Fraud Detection under Concept Drift. Highlights in Business, Economics and Management, 69, 239–250.

[28] Fan, H., Jiao, Y., Wang, M., Chen, J., & Yang, S. (2026). Fraud learns too: continual graph learning under strategic adversarial drift in dynamic networks. Scientific Reports. https://doi.org/10.1038/s41598-026-60997-7