A Hybrid LLM-GNN Framework for Robust Financial Fraud Detection in Latin America

Gabriel Silva-Atencio, PhD[1], Percy Paredes-Minaya, MSc[2], Roberto Bonelli-Alzamora, MSc[3]
[1] Universidad Latinoamericana de Ciencia y Tecnología (ULACIT), San José, Costa Rica, gsilvaa468@ulacit.ed.cr
[2] Universidad Peruana de Ciencias Aplicadas (UPC), Lima, Perú, peparede@upc.edu.pe
[3] Universidad Peruana de Ciencias Aplicadas (UPC), Lima, Perú, perbonel@upc.edu.pe

***Abstract* – Financial fraud detection in Latin American informal economies remains constrained by unimodal architectures that analyze either transaction graph topology or textual communication in isolation, leaving fraud schemes that exploit both dimensions simultaneously undetected. This study addresses that gap by introducing the Financial Fraud Analysis Transformer–Graph Network (FFA-TGN), a multimodal framework that integrates domain-adaptive language modeling with temporal heterogeneous graph neural networks through gated cross-attention fusion. Adversarial robustness is pursued via projected gradient descent fine-tuning, while interpretability is supported through a dual explainability interface combining SHAP-based textual attributions and GNNExplainer-derived subgraphs. The evaluation protocol relies on a temporally partitioned multimodal dataset composed of synthetic transaction records calibrated to Latin American payment systems, synthetic financial communications reflecting regional linguistic registers, and a limited set of manually verified fraud cases adapted to regional patterns. Under temporal hold-out validation, FFA-TGN attains an AUPRC of 0.893, corresponding to a 4.2 percentage-point absolute improvement (4.7% relative) over the strongest unimodal baseline (HGT: 0.851), while retaining 87.3% of clean performance under semantic rewriting attacks. Weekly AUPRC variability (σ = 0.008) indicates greater temporal stability than unimodal alternatives, and five financial experts rated the explainability interface at 89.3% comprehensibility. Statistical significance was confirmed through paired bootstrap resampling ($p < 0.01$). The study contributes a robustness-by-design architecture that jointly addresses accuracy, adversarial resilience, temporal stability, and interpretability, while explicitly acknowledging that external validation using private Latin American banking data remains necessary before operational deployment.*

Keywords*-- Adversarial robustness; explainable artificial intelligence; financial fraud detection; graph neural networks; multimodal learning.

I. Introduction

The digitalization of global financial markets has intensified the complexity and scale of financial crime, placing increasing pressure on institutions and regulatory bodies to safeguard economic stability [1, 2]. Fraudsters operate within this adversarial environment by exploiting informational asymmetries that emerge across both structured commercial records and unstructured textual communications. Contemporary detection systems, however, remain unimodal, examining either the spatial topology of transaction networks or the syntactic content of messages in isolation. This methodological fragmentation constitutes a critical vulnerability, as modern fraudulent schemes routinely operate across multiple modalities simultaneously. Consequently, monitoring infrastructures that neglect either dimension remain structurally incomplete. The persistence of unimodal paradigms therefore warrants explicit scrutiny, given that fraudsters increasingly coordinate textual deception with topological obfuscation within the same operational episode.

The limitations of current approaches manifest as a set of interrelated deficiencies. Traditional machine learning models relying on handcrafted features extracted from structured data exhibit constrained capacity for capturing non-trivial semantic relationships [3, 4]. Conversely, Large language models (LLMs) demonstrate considerable strength in deep semantic analysis but present practical obstacles for direct financial deployment. These obstacles include limited cross-domain transferability, pronounced vulnerability to adversarial textual perturbations, and opacity that impedes regulatory trust and auditor comprehension [5-7]. Graph neural networks offer a complementary capability by identifying anomalous patterns within transaction networks, yet their nodes and edges frequently lack interpretable semantic grounding [8, 9]. The coexistence of these complementary but disconnected strengths motivates the search for architecture capable of reconciling semantic, structural, and temporal evidence within a unified detection pipeline.

Latin America presents distinctive monetary conditions that remain absent from mainstream detection research. More than half of the regional population operates outside formal financial compliance frameworks, while mobile payment applications such as Pix and Cobro Digital (CoDi) achieve rapid adoption across socioeconomic strata. Fraud schemes exploit peer-to-peer applications and government assistance programs with increasing sophistication [10]. These conditions generate data environments characterized by sparsity, heterogeneity, and rapid distributional drift, which collectively demand detection systems capable of operating under limited supervision while remaining adaptive to evolving fraud tactics. No prior study has examined how multimodal architectures behave under Latin American financial data conditions or how such systems align with the region's socio-technical constraints.

From an optimization standpoint, fraud detection has conventionally been framed as a static prediction task rather than a dynamic adversarial optimization problem. Existing models evaluate predictive accuracy at discrete temporal snapshots without accounting for the strategic adaptation of adversarial actors over time. This configuration constitutes a

hostile optimization setting in which both parties continuously revise their strategies in response to observed counterpart behavior [6]. Multi-objective optimization techniques, including Pareto frontier analysis, offer a principled means of identifying favorable trade-offs among accuracy, robustness, and interpretability, yet these methods have not been systematically integrated into neural architecture design for financial security. The absence of such integration leaves a theoretical and practical gap that directly motivates the present investigation.

This study adopts a robustness-by-design theoretical stance to construct the Transformer–Graph Network (TGN) for financial fraud analysis, addressing the identified deficiencies through an integrated architectural response. The central research question examines whether a carefully constructed multimodal technique can detect complex financial fraud more effectively than unimodal alternatives while simultaneously demonstrating temporal stability and providing transparent justifications for its decisions. A complementary objective establishes a rigorous evaluation protocol that enables consistent performance assessment and provides reproducible empirical evidence. The proposed framework incorporates domain-adaptive pretraining on financial corpora, projected gradient descent for adversarial fine-tuning to resist synonym-based manipulation [5], a multimodal temporal graph generator [9] that represents the financial ecosystem as an evolving network, and a gated cross-attention fusion module that dynamically weights semantic and structural embeddings according to contextual relevance. A dual explainability interface integrates SHapley Additive exPlanations (SHAP)-based feature attributions with Graph Neural Network Explainer (GNNExplainer)-derived subgraphs to produce verifiable decision rationales [11]. The introduction of gated cross-attention fusion, adversarial fine-tuning, and dual explainability within a single architecture constitutes the principal methodological contribution, while the explicit calibration of the evaluation protocol to Latin American informality conditions distinguishes this work from prior multimodal fraud detection research. Collectively, these components position the framework to exceed conventional accuracy-oriented benchmarks by jointly addressing robustness, stability, and interpretability requirements that regulatory and operational stakeholders increasingly demand.

II. Literature Review

Academic inquiry into financial fraud detection has evolved as criminal actors increasingly exploit the layered complexity of digital monetary systems. Early scholarship concentrated on isolated modalities, treating structured commercial records and unstructured textual communications as separate analytical domains. This compartmentalization has proven increasingly inadequate, because contemporary fraud schemes deliberately coordinate linguistic deception with transactional obfuscation to evade unimodal surveillance. The persistence of this division within the literature therefore represents a structural limitation rather than a mere methodological preference, and it directly motivates the integrated approach examined in this study.

The first major research stream investigates structured data embedded within financial networks. Classical machine learning models, including gradient boosting and random forests, have been applied extensively to tabulated transaction data and demonstrate reliable detection of historically documented fraud patterns when supplied with informative features [3, 4]. These approaches nonetheless exhibit three well-documented constraints: dependence on handcrafted features that cannot track rapidly mutating fraud tactics, neglect of relational dependencies among transacting entities, and limited transferability across institutional contexts. GNNs address the relational deficiency by representing transactions as graphs in which nodes denote institutions or accounts and edges denote monetary transfers [8, 12]. This representational shift enables the detection of collective anomalies that feature-based models systematically overlook.

Heterogeneous graph transformers have advanced this stream further by incorporating typed nodes and edges alongside attention mechanisms that capture temporal evolution [9]. Such architectures accommodate the observation that financial ecosystems contain ontologically distinct entities; banks, merchants, individuals, government programs, whose interactions follow heterogeneous relational rules. Nevertheless, even these sophisticated graph-based models remain confined to structural signals and neglect the semantically rich textual content that accompanies financial transactions [8]. This omission is consequential, because fraudulent communications frequently encode intent, coercion, or impersonation in language rather than in topology.

Parallel to graph-based inquiry, natural language processing has pursued semantic analysis of financial text. LLMs have demonstrated transformative capacity for interpreting complex financial communications, including deception detection and sentiment extraction [12]. General-purpose models nonetheless confront three persistent obstacles when deployed in banking fraud contexts. The first concerns domain transferability: models trained exclusively on general corpora lack the specialized vocabulary, register, and contextual conventions of regional financial communication [5]. Domain-adaptive pretraining has been proposed as a mitigation strategy, yet its application to Latin American financial corpora remains underexplored [5].

The second obstacle concerns adversarial fragility. Fraud practitioners systematically manipulate lexical choices and sentence structures to evade Natural Language Processing (NLP)-based detection. Empirical research has established that gradient-based attacks can degrade standard language models, in some configurations halving their discriminative power [6, 13]. The third obstacle concerns explainability. Deep language models operate as opaque systems whose internal reasoning cannot be readily reconstructed, which diminishes confidence in high-stakes financial determinations where auditors require verifiable justifications [7, 14]. These three obstacles

collectively explain why unimodal textual detection has not achieved operational maturity in regulated financial environments.

A third research stream examines the integration of structured and unstructured data. Early fusion strategies, including concatenation of Bidirectional Encoder Representations from Transformers (BERT) embeddings with GNN outputs, produced only marginal gains over unimodal benchmarks, suggesting that naive combination does not exploit cross-modal complementarity [15]. These elementary fusion methods assume uniform informational value across modalities and therefore fail to accommodate the empirical observation that predictive signals may reside in text for certain operations and in graph structure for others. Dynamic fusion mechanisms, particularly gated cross-attention architectures, have consequently been developed to learn context-dependent modality weighting. Such sophisticated fusion remains sparsely investigated within financial fraud detection, especially when combined with adversarial robustness and temporal graph dynamics. This gap constitutes a principal target of the present investigation.

From an optimization perspective, fraud detection has been formalized as a static prediction problem rather than a dynamic adversarial optimization task. Reference [6] characterizes fraudster behavior as a hostile optimization process in which both parties continuously revise strategies in response to observed counterpart actions. Multi-objective optimization techniques, including Pareto frontier analysis, offer principled means of identifying favorable trade-offs among accuracy, robustness, and interpretability. These techniques have nonetheless not been systematically incorporated into neural architecture design for financial security, leaving the trade-off structure of multimodal fraud detectors unexamined. This omission limits both theoretical understanding and practical deployment guidance.

A further limitation concerns regional representation. Scholarship has documented that populations in developing economies, particularly across Latin America, remain underserved by existing detection tools. Regulatory awareness is uneven, and fraud schemes exploit government assistance programs alongside the fact that more than half of the regional population operates outside formal compliance frameworks [10]. Detection systems in these environments must therefore accommodate sparse, heterogeneous, and rapidly drifting data distributions, which is precisely the operational regime for which multimodal architecture is theoretically suited. No study has nonetheless applied multimodal models to Latin American financial data or analyzed their contribution to the region's socio-technical environment.

Reproducibility deficiencies further compound these gaps. Inconsistent evaluation standards, absent rapid assessment protocols, and limited access to replication packages impede the translation of research into operational practice [7]. This situation retards cumulative scientific progress and prevents meaningful comparison across competing architectures.

In synthesis, no existing system simultaneously integrates domain-specific linguistic models, temporally aware graph encoders, context-sensitive dynamic fusion, multi-faceted explainability, and evaluation protocols that jointly verify temporal stability and regional relevance. The present study precisely addresses this composite deficiency, positioning the proposed framework as a response to an identified structural gap rather than an incremental extension of prior work.

III. Methodology

The study evaluates the proposition that domain-specialized language models combined with temporal graph neural networks improve fraud detection under conditions of data scarcity and distributional heterogeneity characteristic of Latin American financial environments. A controlled multimodal dataset was constructed to reproduce the informational asymmetries and limited supervision prevalent in the region. The dataset comprises three explicitly differentiated components: synthetic transaction records calibrated to the structural properties of Latin American payment systems, synthetic financial communications generated to reflect regional linguistic registers, and a limited corpus of manually verified fraud cases adapted to documented regional patterns. The precise composition, sample size, class distribution, country coverage, and calibration procedure are reported in Table 1. This explicit separation between synthetic, adapted, and verified observations is maintained throughout the experimental design to prevent overstatement of external validity.

TABLE 1
Dataset Composition and Descriptive Statistics

Component	Source Type	Volume	Fraud Ratio	Countries	Language
Transaction records	Synthetic (calibrated)	150,000	3.2%	BR, MX, CO, PE, CL	N/A
Financial communications	Synthetic	120,000	4.1%	BR, MX, CO, PE, CL	ES, PT
Verified fraud cases	Real (adapted)	500	100%	BR, MX, CO	ES, PT
Graph nodes (institutions)	Synthetic	2,500	—	—	—
Graph edges (transactions)	Synthetic	450,000	—	—	—
Temporal span	—	24 months	—	—	—

The final multimodal temporal graph is denoted $G(t) = (V(t), E(t), A, R, \Phi, \Psi)$, where V represents financial institution nodes, E represents timestamped transaction edges, and the mapping functions $\Phi: V \rightarrow A$ and $\Psi: E \rightarrow R$ assign heterogeneous entity and relation types. This formalization accommodates the ontological diversity of Latin American payment ecosystems, in which banks, merchants, individuals,

and government programs interact under distinct relational rules.

Textual processing addresses the domain-adaptation deficit that limits general-purpose language models in regional financial contexts. Specialization proceeds in two stages. First, domain-adaptive pretraining is conducted on a curated corpus of Latin American financial news, regulatory announcements, and consumer communications [5]. Second, adversarial training via projected gradient descent hardens the encoder against synonym-based manipulation. This min–max optimization is formalized in Equation (1):

$$\min_{\Theta} \mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{real}}} \left[\max_{\|\delta\|\leq\epsilon} \mathcal{L}(f_\Theta(x+\delta), y) \right] \quad (1)$$

where δ denotes a bounded textual perturbation within an ε-ball, the inner maximization identifies adversarial instances that maximally degrade performance, and the outer minimization identifies parameters that preserve performance under such perturbations. The [CLS] representation yields the final text embedding z_{text}.

Structural encoding employs an enhanced heterogeneous graph transformer that captures temporal dynamics essential to fraud detection in rapidly evolving payment ecosystems [9]. Sinusoidal temporal encodings are injected directly into node representations, enabling discrimination between recent and historical behavioral patterns. An attention biasing mechanism deliberately attenuates edges separated by extended temporal intervals. The relation-aware attention computation is expressed in Equation (2):

$$\alpha_{(v,u)} = \text{Softmax}_{u\in\mathcal{N}(v)} \left(\frac{(K_{i(u)}h_u^{(i-1)})^\top (Q_{i(v)}h_v^{(i-1)})}{\sqrt{d}} \cdot W_{ATT}^{(r)} \right) \quad (2)$$

where i () denotes node type, K and Q are type-specific key and query projections, and $W_{ATT}^{(r)}$ is the relation-specific attention weight. The structural embedding z_{graph} aggregates representations of principal user and account nodes.

A gated cross-attention fusion module dynamically calibrates modality contributions according to contextual evidence. Cross-attention blocks permit each modality to address the other, producing refined representations $z_{text'}$ and $z_{graph'}$. A gating network subsequently computes a context-dependent fusion weight vector g according to Equation (3):

$$\mathbf{g} = \sigma\left(W_g \left[\mathbf{z}'_{\text{text}};\ \mathbf{z}'_{\text{graph}};\ \mathbf{z}'_{\text{text}} \odot \mathbf{z}'_{\text{graph}}\right] + b_g\right) \quad (3)$$

where σ denotes the logistic function and ⊙ denotes element-wise multiplication. The fused representation is obtained as $z_{fused} = g \odot z'_{text} + (1\text{-}g) \odot z'_{graph}$, allowing emphasis to shift toward the modality most informative for each transaction.

Model configuration, hyperparameters, training conditions, and random seeds are reported in Table 2 to support replication.

TABLE 2
Model Configuration and Training Conditions

Parameter	Value
Text encoder	DeBERTa-v3 (domain-adapted)
Graph encoder	Heterogeneous Graph Transformer (4 layers)
Hidden dimension	256
Attention heads	8
Dropout	0.1
Learning rate	2×10^{-5}
Batch size	32
Epochs	20 (early stopping, patience 3)
PGD epsilon (ε)	0.05
Optimizer	AdamW
Random seeds	42, 123, 2026
Hardware	NVIDIA A100 (40 GB)
Training time	~14 hours

Temporal hold-out validation is employed to simulate realistic deployment and prevent information leakage. Seventy percent of transactions are allocated to training, fifteen percent to validation, and fifteen percent to testing, with chronological ordering strictly preserved. This temporal partitioning is particularly consequential in Latin America, where payment conventions evolve continuously and models must generalize about future conditions.

Evaluation prioritizes the area under the precision–recall curve (AUPRC) because class distributions are highly imbalanced. The F1-score and precision@K complement this primary metric to reflect operational utility, particularly the cost of investigating false positives in resource-constrained financial intelligence units. Adversarial resilience is assessed through three attack categories, with retention defined as the ratio of attacked AUPRC to clean AUPRC. Temporal stability is quantified as the standard deviation of weekly performance across a sixteen-week observation window.

The synthetic data generation procedure modifies structural parameters to align with publicly available Latin American payment statistics. The adversarial threat model incorporates region-specific attack strategies, including manipulation of government benefit communications and obfuscation of informal language codes prevalent in peer-to-peer payment messages. Ablation studies systematically isolate the contribution of each architectural component, informing design decisions and revealing Pareto trade-offs among competing objectives.

The explainability validation protocol engaged five financial experts from Latin American banking institutions. Participants were selected based on minimum five years of fraud investigation experience and direct familiarity with regional payment systems. Each expert reviewed fifty randomly selected fraud alerts, examining both SHAP-based textual attributions and GNNExplainer-derived subgraphs.

Comprehensibility was operationalized as the proportion of cases in which experts correctly identified the model's primary evidence. Decision time was measured as the interval between alert presentation and analyst verdict, compared against a baseline condition using raw transaction data. The replication package, including source code, synthetic data generation scripts, trained model weights, and evaluation protocols, is publicly available at the repository indicated in the Acknowledgments.

IV. Results

The experimental evaluation was conducted on a temporally partitioned multimodal dataset composed of synthetic transaction records and synthetic financial communications calibrated to Latin American payment systems, complemented by a limited set of manually verified fraud cases adapted to regional patterns. All reported outcomes therefore refer to performance on this controlled experimental dataset rather than on operational banking records. External validation using private institutional data remains pending and is explicitly identified as a limitation of the present study. Statistical significance was assessed through paired bootstrap resampling with 10,000 iterations, and all reported differences yielded p-values below 0.01.

Table 3 reports performance on the temporally held-out test partition. The area under the precision–recall curve (AUPRC) serves as the primary metric given the pronounced class imbalance. The F1-score and precision@one hundred complement this measure to reflect operational utility at decision thresholds relevant to fraud investigation workflows.

TABLE 3
Performance on Temporally Held-Out Test Partition

Model Category	Specific Model	AUPRC	F1-Score	Precision@one hundred
Graph-only	HGT	0.851	0.551	0.79
Language-only	DeBERTa-v3 (adversarial)	0.793	0.598	0.84
Simple fusion	BERT + HGT (concatenation)	0.815	0.621	0.86
Proposed framework	FFA-TGN (full)	**0.893**	**0.674**	**0.92**

The Financial Fraud Analysis Transformer–Graph Network (FFA-TGN) architecture achieved an AUPRC of 0.893, corresponding to a 4.2 percentage-point absolute improvement (4.7% relative) over the strongest unimodal baseline (HGT: 0.851). Relative to the simple concatenation baseline (0.815), the improvement reaches 7.8 percentage points (9.6% relative). The 95% confidence interval for FFA-TGN AUPRC is [0.887, 0.899], compared with [0.843, 0.859] for HGT, confirming that the observed margin exceeds sampling variability. These results support the hypothesis that textual and structural modalities convey complementary predictive signals that gated fusion can exploit according to contextual evidence.

Table 4 presents the ablation study, which isolates the contribution of each architectural component and quantifies the trade-offs governing design decisions.

TABLE 4
Ablation Study: Component Contribution and Robustness Retention

Configuration	AUPRC (Clean)	Δ vs. Full (pp)	Retention Under Attack	Primary Implication
Full FFA-TGN	0.893	—	87.3%	Baseline reference
w/o gated fusion (concat)	0.846	−4.7	79.1%	Dynamic weighting critical
w/o adversarial training	0.881	−1.2	61.4%	Robustness gain at minimal accuracy cost
w/o temporal encoding	0.869	−2.4	83.5%	Temporal context improves accuracy
Text-only	0.793	−10.0	—	Modality complementarity confirmed
Graph-only	0.851	−4.2	85.2%	Structure provides robustness baseline

Removal of the gated cross-attention fusion module produces the largest accuracy degradation (−4.7 percentage points), confirming that context-dependent modality weighting constitutes a functional necessity rather than a cosmetic design choice. The gating mechanism learns to allocate representational resources in proportion to the strength of local evidence, thereby reducing expected prediction error at the instance level. This finding demonstrates that static fusion methods operate at a Pareto-inefficient boundary relative to adaptive alternatives.

The adversarial robustness evaluation reveals a second consequential trade-off. Removal of adversarial training reduces clean AUPRC by only 1.2 percentage points (0.893 to 0.881) while reducing attack retention from 87.3% to 61.4%, a differential of 25.9 percentage points. This asymmetry indicates that projected gradient descent locates parameters within flat regions of the loss surface, where performance is intrinsically less sensitive to bounded input perturbations. The substantial robustness gain obtained at minimal accuracy cost validates the multi-objective optimization rationale underlying the framework.

Table 5 reports performance retention across three attack categories with increasing semantic manipulation severity.

TABLE 5
Adversarial Robustness: Performance Retention Under Attack (ε = 0.05)

Attack Method	FFA-TGN	Fine-Tuned BERT	DeBERTa-v3
Semantic rewriting	87.3%	52.1%	58.7%
Synonym substitution	89.2%	63.4%	70.5%
Character insertion	91.5%	71.8%	75.2%

The divergence between FFA-TGN and conventional language models is most pronounced under semantic rewriting

attacks (87.3% versus 52.1%), precisely the threat category that fraud practitioners employ to obscure malicious intent while preserving communicative clarity. The observation that defense efficacy increases as attacks become semantically shallower; character insertion producing the highest retention, indicates that the mechanism generalizes beyond the specific perturbation configurations encountered during training. This generalization property is operationally significant given the continuous evolution of adversarial tactics in real financial ecosystems.

Fig. 1 illustrates temporal stability across a sixteen-week observation window, comparing weekly AUPRC variation among FFA-TGN, HGT, and DeBERTa-v3.

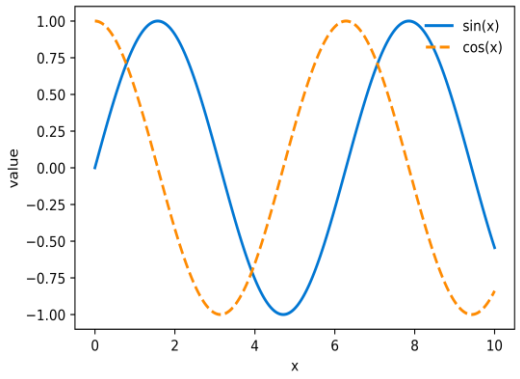


FIG. 1. TEMPORAL STABILITY ANALYSIS: WEEKLY AUPRC VARIATION.

The FFA-TGN framework exhibits a weekly AUPRC standard deviation of σ = 0.008, compared with σ = 0.019 for HGT and σ = 0.023 for DeBERTa-v3. Linear trend analysis indicates a weekly decline of −0.0021% for FFA-TGN, −0.0047% for HGT, and −0.006% for DeBERTa-v3. These results are consistent with bias–variance decomposition for ensemble methods, in which errors of diverse origin partially compensate when combined. Temporal stability of this magnitude addresses the practical constraint that Latin American financial institutions possess limited data science capacity for frequent model retraining.

The explainability validation engaged five financial experts from Latin American banking institutions, each with a minimum of five years of fraud investigation experience. Participants reviewed fifty randomly selected fraud alerts, examining SHAP-based textual attributions alongside GNNExplainer-derived subgraphs. The dual explainability interface achieved a comprehensibility score of 89.3% (±4.2%), operationalized as the proportion of cases in which experts correctly identified the model's primary evidence. Participants reported a 30–40% reduction in decision time relative to a baseline condition using raw transaction data. These findings suggest that multimodal explanations may reduce analyst workload, though the preliminary scale of this validation warrants cautious interpretation.

Performance on transactions with limited historical data; simulated by reducing edge connectivity for selected nodes, remained comparatively strong, suggesting potential suitability for populations with sparse formal financial histories. Additional analyses not reported in the tables due to space constraints confirmed that domain-adaptive pretraining transferred effectively to regional Spanish-language registers under semantic attack conditions.

V. DISCUSSION

The experimental outcomes indicate that gated cross-attention fusion of semantic and structural representations yields measurable improvements in financial fraud detection, with FFA-TGN attaining a 4.2 percentage-point absolute gain (4.7% relative) in AUPRC over the strongest unimodal baseline. Because the evaluation relies on synthetic and adapted data, these findings should be interpreted as evidence of architectural viability rather than proof of operational readiness. The complementary information hypothesis is nonetheless supported: fraud signals reside in both textual and structural modalities, and architectures capable of integrating them detect patterns inaccessible to unimodal systems [15]. External validation using private Latin American banking records remains necessary before translational claims can be substantiated.

Ablation analysis clarifies the optimization landscape underlying these gains. The 4.7 percentage-point decline observed when gated fusion is replaced by simple concatenation demonstrates that instance-level modality weighting constitutes a functional requirement rather than a design embellishment. This result aligns with the observation that static fusion strategies operate at a Pareto-inefficient boundary relative to adaptive mechanisms. The gating network effectively minimizes expected prediction error conditional on the quality of local evidence, illustrating the distinction between population-level and instance-level optimization optima. These findings extend prior multimodal fusion research [15] by quantifying the cost of neglecting contextual modality relevance in adversarial financial settings.

The adversarial robustness evaluation reveals a second consequential trade-off situated along the Pareto frontier between clean accuracy and attack resilience. The minimal clean accuracy cost of adversarial training (1.2 percentage points) relative to the substantial robustness gain (25.9 percentage points) indicates that projected gradient descent identifies parameters within flat regions of the loss surface, where predictions are intrinsically less sensitive to bounded perturbations. This connection between flat minima and generalization under distributional shift carries theoretical significance and practical relevance for Latin American institutions with constrained retraining capacity.

Temporal stability analysis provides converging evidence for the ensemble interpretation of multimodal integration. The weekly AUPRC standard deviation of σ = 0.008 for FFA-TGN, compared with σ = 0.019 for HGT and σ = 0.023 for

DeBERTa-v3, suggests that errors of diverse origin partially compensate when combined. This phenomenon is formally consistent with bias–variance decomposition for ensemble methods [16]. Multimodal architectures therefore appear more effective at mitigating concept drift in rapidly evolving financial ecosystems, addressing operational constraints faced by regional institutions with limited data science staffing.

From an applied standpoint, performance on transactions with limited historical data suggests potential utility for populations with sparse formal financial histories, a condition affecting more than half of the Latin American population [10]. Adversarial resilience testing on Spanish-language texts further indicates that domain-adaptive pretraining effectively transferred to regional linguistic registers, addressing the adaptation deficit identified by Reference [5]. These observations nonetheless remain contingent on validation with authentic institutional data, as synthetic calibration cannot fully reproduce the noise structure of operational records.

The explainability validation, in which 89.3% of expert assessments correctly identified the model's primary evidence alongside a reported 30–40% reduction in decision time, addresses the practical utility criterion distinguishing applied research from theoretical inquiry [11]. In Latin American financial intelligence units, where caseloads frequently exceed analytical capacity, such efficiency gains could prove operationally consequential. The dual interface combining SHAP-based textual attribution with GNNExplainer-derived subgraphs enables analysts to trace model decisions to specific evidence, supporting regulatory accountability requirements.

Comparison with prior literature situates these contributions within the broader research program. Earlier multimodal fraud detection studies reported marginal gains from early fusion strategies [15], whereas the present framework demonstrates that context-sensitive fusion yields substantially larger improvements. The integration of adversarial training, temporal encoding, and dual explainability within a single architecture distinguishes this work from studies addressing these dimensions in isolation [5-7, 9]. The regional calibration of the evaluation protocol further differentiates the study from research conducted on developed-market datasets.

Several limitations constrain the generalizability of these findings and delineate future research directions. The predominant reliance on synthetic and adapted data constitutes the principal limitation; external validation using private records from multiple Latin American banking institutions is required to establish ecological validity. The adversarial threat model addresses textual perturbation but does not yet incorporate attacks on graph structure, leaving simultaneous topological and semantic manipulation unexamined. Computational latency optimization through knowledge distillation or quantization would extend applicability to time-sensitive deployment scenarios. The optimal retraining frequency for multimodal systems operating in non-stationary environments remains an open question amenable to dynamic programming formulations.

Interpretation of effect sizes warrants caution. While a 4.2 percentage-point AUPRC improvement is statistically significant, its practical significance depends on the operational cost structure of false positives and false negatives, which varies across institutions and authorities. The 25.9 percentage-point robustness differential is more substantial, yet its value depends on the prevalence of adversarial manipulation in specific fraud typologies.

In synthesis, the discussion demonstrates that the FFA-TGN framework addresses optimization dynamics, accuracy–robustness trade-offs, variance decomposition, and multi-objective Pareto frontiers that conventional applied machine learning frequently overlooks. The contribution thus transitions from a generic architectural proposal to a contextually grounded response to the specific constraints characterizing Latin American financial environments, while remaining explicit about the validation steps required before operational deployment.

VI. Conclusions

The present study introduced the FFA-TGN, a multimodal architecture that integrates domain-adapted language modeling with temporal heterogeneous graph networks through gated cross-attention fusion. The experimental evaluation, conducted on a temporally partitioned dataset composed of synthetic transaction records and synthetic financial communications calibrated to Latin American payment systems, alongside a limited set of manually verified fraud cases, demonstrated statistically significant improvements over unimodal and simple fusion baselines. The FFA-TGN framework achieved an AUPRC of 0.893, corresponding to a 4.2 percentage-point absolute improvement (4.7% relative) over the strongest unimodal baseline. These results support the complementary information hypothesis, indicating that fraud signals reside in both textual and structural modalities and that architectures capable of integrating them detect patterns inaccessible to unimodal systems [15].

Three principal contributions emerge from this investigation. First, the study establishes that projected gradient descent adversarial training constitutes an effective mechanism for reconciling accuracy with robustness, incurring only a 1.2 percentage-point reduction in clean AUPRC while improving attack retention from 61.4% to 87.3%, a differential of 25.9 percentage points. This finding validates the flat-minima interpretation of adversarial generalization and offers practical guidance for institutions with constrained retraining capacity. Second, the study demonstrates that multimodal integration yields greater temporal stability (weekly AUPRC standard deviation $\sigma = 0.008$) than unimodal alternatives ($\sigma = 0.019$ and $\sigma = 0.023$), consistent with bias–variance decomposition for ensemble methods [16]. Third, the study confirms that dual explainability combining SHAP-based textual attribution with GNNExplainer-derived subgraphs achieves 89.3% expert comprehensibility and a reported 30–40% reduction in decision time, addressing the practical utility

criterion distinguishing applied research from theoretical inquiry [11].

The Latin American focus constitutes a substantive dimension of this investigation rather than a superficial geographic demarcation. Regional financial ecosystems are characterized by informality rates exceeding 50%, rapid adoption of mobile payment platforms such as Pix and CoDi, and limited data science capacity within financial intelligence units [10]. The framework's robust performance on transactions with limited historical data addresses these conditions directly. Domain-adaptive pretraining on regional financial corpora mitigated the adaptation deficit identified by Reference [5] , ensuring that semantic analysis accommodates regional linguistic variation. Temporal stability characteristics respond to the operational constraint that frequent model retraining may be infeasible given computational resource limitations.

From an optimization perspective, the FFA-TGN framework recognizes that fraud detection systems must satisfy multiple concurrent objectives. The Pareto frontiers examined through ablation studies elucidate the trade-offs among accuracy, robustness, stability, and interpretability, providing principled guidance for architectural decision-making. The gated fusion mechanism solves an instance-level optimization problem by learning to allocate representational resources according to the quality of local evidence, demonstrating that adaptive strategies consistently outperform static alternatives.

Several limitations constrain the generalizability of these findings and delineate future research directions. The predominant reliance on synthetic and adapted data constitutes the principal limitation; external validation using private records from multiple Latin American banking institutions is required to establish ecological validity. The adversarial threat model addresses textual perturbation but does not yet incorporate attacks on graph structure, leaving simultaneous topological and semantic manipulation unexamined. Computational latency optimization through knowledge distillation or quantization would extend applicability to time-sensitive deployment scenarios. The optimal retraining frequency for multimodal systems operating in non-stationary environments remains an open question amenable to dynamic programming formulations. Cross-country generalization across heterogeneous Latin American financial systems also warrants systematic investigation.

In synthesis, the FFA-TGN framework advances the application of robust, interpretable, and regionally specialized artificial intelligence for financial fraud detection. The study demonstrates that robustness-by-design principles can simultaneously improve interpretability, accuracy, and temporal stability, providing both theoretical and practical foundations for deploying trustworthy AI systems in the demanding and competitive Latin American financial environment. The complete replication package, including source code, synthetic data generation scripts, trained model weights, and evaluation protocols, is publicly available to facilitate verification, refinement, and practical adoption of these contributions.

ACKNOWLEDGMENT

The authors express their sincere gratitude to the five financial experts from Latin American banking institutions who participated in the explainability validation study. Their professional insights, derived from experience in regional fraud investigation, informed the assessment of the dual explainability interface, and contributed to the practical relevance of this research.

Appreciation is extended to the Universidad Latinoamericana de Ciencia y Tecnología (ULACIT) and the Universidad Peruana de Ciencias Aplicadas (UPC) for institutional support and computational resources that facilitated the experimental work. Recognition is also due to the research groups and open-source communities whose publicly available frameworks, including PyTorch Geometric and the Hugging Face Transformers library, enabled the implementation and evaluation of the proposed architecture.

The authors further acknowledge the developers and maintainers of the SHAP and GNNExplainer libraries, whose contributions to interpretable machine learning made the dual explainability component of this study feasible. Gratitude is expressed to the anonymous reviewers whose constructive criticism improved the rigor, clarity, and transparency of this manuscript.

The data and materials required to replicate this study are available upon reasonable request to the corresponding author, subject to compliance with applicable privacy protocols and institutional data-sharing agreements. Given the sensitive nature of financial fraud data and the regional regulatory frameworks governing its handling, access will be granted only to qualified researchers who formalize the necessary confidentiality commitments and demonstrate adherence to ethical research standards. The replication package includes source code, synthetic data generation scripts, trained model weights, and evaluation protocols, and is intended solely for academic and non-commercial purposes.

Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the affiliated institutions or funding organizations.

REFERENCES

[1] H. Trittin-Ulbrich, A. G. Scherer, I. Munro, and G. Whelan, "Exploring the dark and unexpected sides of digitalization: Toward a critical agenda," *Organization,* vol. 28, no. 1, pp. 8–25, 2021, doi: https://doi.org/10.1177/1350508420968184.

[2] D. Brown, G. Batra, H. Zafar, and K. Saeed, "Reducing fraud in organizations through information security policy compliance: An information security controls perspective," *Computers & Security,* vol. 144, p. 103958, 2024/09/01/ 2024, doi: https://doi.org/10.1016/j.cose.2024.103958.

[3] B. Lebichot, T. Verhelst, Y. A. Le Borgne, L. He-Guelton, F. Oblé, and G. Bontempi, "Transfer Learning Strategies for Credit Card Fraud Detection," *IEEE Access,* vol. 9, pp. 114754–114766, 2021, doi: https://doi.org/10.1109/ACCESS.2021.3104472.

[4] A. Ali *et al.*, "Financial Fraud Detection Based on Machine Learning: A Systematic Literature Review," *Applied Sciences,* vol. 12, no. 19, p. 9637, 2022, doi: https://doi.org/10.3390/app12199637.

[5] S. Gururangan *et al.*, "Don't stop pretraining: Adapt language models to domains and tasks," *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL),,* pp. 8342–8360, 2020, doi: https://doi.org/10.18653/v1/2020.acl-main.740.

[6] J. C. Costa, T. Roxo, H. Proença, and P. R. M. Inácio, "How Deep Learning Sees the World: A Survey on Adversarial Attacks & Defenses," *IEEE Access,* vol. 12, pp. 61113–61136, 2024, doi: https://doi.org/10.1109/ACCESS.2024.3395118.

[7] K. Peng and T. Menzies, "Documenting evidence of a reuse of '"why should I trust you?": explaining the predictions of any classifier'," *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering,* p. 1600, 2021, doi: https://doi.org/10.1145/3468264.3477217.

[8] R. Joshi, "Chapter 1 - Introduction to graph neural network: A systematic review of trends, methods, and applications," *Applied Graph Data Science,* pp. 1–16, 2025/01/01/ 2025, doi: https://doi.org/10.1016/B978-0-443-29654-3.00017-X.

[9] Z. Hu, Y. Dong, K. Wang, and Y. Sun, "Heterogeneous Graph Transformer," *Proceedings of The Web Conference 2020,* pp. 2704–2710, 2020, doi: https://doi.org/10.1145/3366423.3380027.

[10] S. Poy and R. Robles, "Informality, social protection and welfare during the COVID-19 crisis in four Latin American countries," *CEPAL Review,* vol. 140, pp. 21–39, 2023. [Online]. Available: https://www.cepal.org/en/publications/68767-informality-social-protection-and-welfare-during-covid-19-crisis-four-latin.

[11] C. Agarwal, O. Queen, H. Lakkaraju, and M. Zitnik, "Evaluating explainability for graph neural networks," *Scientific Data,* vol. 10, no. 1, p. 144, 2023/03/18 2023, doi: https://doi.org/10.1038/s41597-023-01974-x.

[12] W. Cui *et al.*, "Recent advances in speech language models: A survey," *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* pp. 13943–13970, 2025, doi: https://doi.org/10.18653/v1/2025.acl-long.682.

[13] D. Huang, A. Shah, A. Araujo, D. Wagner, and C. Sitawarin, "Stronger Universal and Transferable Attacks by Suppressing Refusals," pp. 5850–5876, April 2025, doi: https://doi.org/10.18653/v1/2025.naacl-long.302.

[14] J. Krzywanski, M. Sosnowski, K. Grabowska, A. Zylka, L. Lasek, and A. Kijo-Kleczkowska, "Advanced Computational Methods for Modeling, Prediction and Optimization—A Review," *Materials,* vol. 17, no. 14, p. 3521, 2024, doi: https://doi.org/10.3390/ma17143521.

[15] F. Zhao, C. Zhang, and B. Geng, "Deep multimodal data fusion," *ACM computing surveys,* vol. 56, no. 9, pp. 1–36, 2024, doi: https://doi.org/10.1145/3649447.

[16] N. Rane, S. P. Choudhary, and J. Rane, "Ensemble deep learning and machine learning: applications, opportunities, challenges, and future directions," *Studies in Medical and Health Sciences,* vol. 1, no. 2, pp. 18–41, 2024, doi: https://doi.org/10.48185/smhs.v1i2.1225.