ExoRoute: Reliability-Weighted Integration of Temporally Aligned Exogenous Text for Dynamic Graphs

Liangzu Liu[1] **Mengzhe Ruan**[1] **Yinjun Wu**[2,3] **Guanjun Wang**[1,*] **Yang Liu**[1,*]
[1]Ant International, Ant Group [2]Peking University [*]Corresponding authors
[3]Beijing Key Laboratory of Software and Hardware Cooperative Artificial Intelligence Systems
{liuliangzu.llz, ruanmengzhe.rmz, ly461173, guanjun.wgj}@antgroup.com
wuyinjun@pku.edu.cn

Abstract

External events can reshape financial interaction networks in ways that past graph interactions alone may not capture. News and regulatory documents provide complementary context, but their integration must account for temporal availability, relevance, and reliability. We introduce EXOROUTE, a framework for integrating time-aligned exogenous text into temporal graph neural networks. Using only documents available at each event time, EXOROUTE converts text into auditable structured tokens and routes them according to node relevance, relative reliability, and recency. A capped residual update controls the resulting change to graph representations. A lightweight document scorer is learned by distilling routing attention during training and remains fixed at inference, without requiring manual relevance labels. Our analysis bounds representation changes under a bounded-proposal assumption and establishes contraction under additional alignment, scale, and step-size conditions. Experiments on three public cryptocurrency graph datasets augmented with document streams show consistent improvements across six backbones under strict temporal splits. On Elliptic, post-shift improvements reach 0.16 in average precision and an approximately 90% reduction in expected calibration error relative to the respective backbones. Ablations and channel perturbations support the importance of reliability weighting and temporal alignment.

1 Introduction

Financial prediction must remain useful as operating conditions change, making *across-time, across-policy generalization* a practical concern [Quiñonero-Candela et al., 2009, Gama et al., 2014]. Credit-scoring studies show the value of learned predictors [Khandani et al., 2010, Lessmann et al., 2015], while production systems highlight model-refreshing needs under distribution shift [Liu et al., 2026a]. *Financial interaction networks* encode time-stamped transactions, exposures, and trust relations, with connections transmitting risk and liquidity shocks [Allen and Gale, 2000, Acemoglu et al., 2015, Glasserman and Young, 2015]. Dynamic GNNs support temporal representation learning, forecasting, and anomaly detection [Trivedi et al., 2019, Rossi et al., 2020, Xu et al., 2020, Cong et al., 2023, Yu et al., 2023, Gravina et al., 2024, Yang et al., 2025]; strict temporal benchmarks provide a common evaluation setting [Huang et al., 2023].

However, enforcement events in Bitcoin transaction networks have coincided with abrupt behavioral changes and degraded predictions [Weber et al., 2019]. Dataset shift, including changes in label prevalence, complicates out-of-time prediction [Quiñonero-Candela et al., 2009, Lipton et al., 2018, Ovadia et al., 2019]. Recent work on temporal-graph evaluation also shows that time-dependent error patterns can make static evaluation summaries unreliable [Su and Wu, 2025].

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

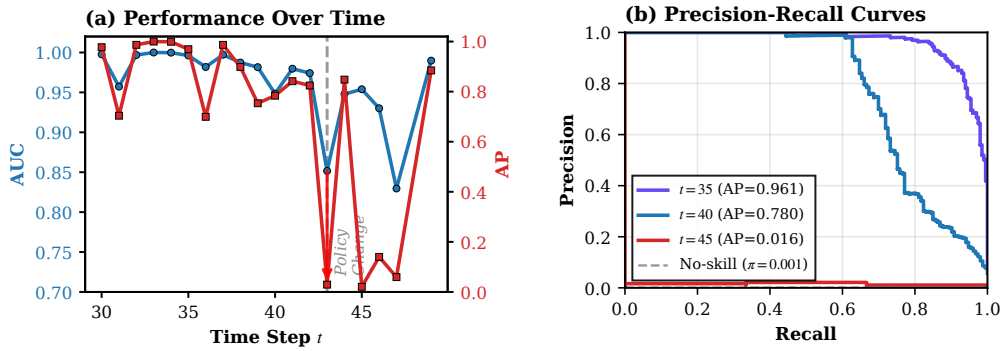

Figure 1: Performance degradation under policy shift at $t = 43$. AUC stays high, but AP collapses. PR curves confirm loss of ranking quality despite preserved discrimination.

Ranking metrics can respond differently to changing class prevalence: precision-based measures such as AP can deteriorate while ROC AUC remains relatively stable [Saito and Rehmsmeier, 2015, Davis and Goadrich, 2006]. Figure 1 illustrates this on Elliptic [Weber et al., 2019]. Around t=43, a dark-market shutdown coincides with a sharp distribution shift. In this example, AUC stays high while AP plummets, showing that high ROC AUC alone does not ensure high precision among the top-ranked predictions.

Temporal GNNs model endogenous dynamics well, but a simple integration strategy treats exogenous text as a static side feature (e.g., raw headlines or LM embeddings from BERT/LLaMA). Short and ambiguous statements (e.g., "SEC tightens crypto rules") provide weak temporal anchoring and can introduce noise and misalignment. More seriously, if documents are accessed by their described effective time rather than availability time, future information can leak into past predictions, leading to spurious correlations and violating temporal causality. Our interface therefore retains source pointers and field-level evidence spans for auditability; reproducible processing is a broader concern in ML systems [Pineau et al., 2021, Sculley et al., 2015].

What is missing is a principled mechanism that treats exogenous texts not as static features, but as *time-stamped, structured, and confidence-aware signals* that can be aligned with graph evolution and fused in a controlled, interpretable manner.

We propose EXOROUTE, a method designed specifically for dynamic graphs under distribution shifts. EXOROUTE bridges financial NLP and temporal graph learning through three tightly integrated components:

- **Document-to-signal extraction with confidence and no leakage.** We map unstructured financial documents to time- and entity-aligned policy tokens via schema-guided prompting and deterministic post-processing, and explicitly enforce the time-available constraint in evaluation.
- **Bounded fusion of exogenous evidence.** We treat exogenous tokens as a controlled side channel. At each event time, the model softly selects relevant tokens, downweights stale or low-confidence ones, and applies a capped residual update to the backbone state. With bounded proposal norms, this limits the influence of the document channel on the same-checkpoint representation; error contraction requires additional direction, scale, and step-size conditions.
- **Training-period retrieval adaptation without manual labels.** To avoid hand-labeling which documents matter, we learn a lightweight retriever by distilling the model's own token-routing/attention mass into a document relevance scorer, trained only on past (training-period) bins. At inference, this fixed scorer selects Top-K documents per bin, improving which evidence is surfaced as regimes change.

2 Related Work

Temporal graphs in finance under regime shifts. Financial networks encode contagion and exposure [Allen and Gale, 2000, Acemoglu et al., 2015, Glasserman and Young, 2015], and graph methods support anomaly detection [Akoglu et al., 2015]. Temporal GNNs learn from evolving interactions [Wu et al., 2021, Rossi et al., 2020, Xu et al., 2020, Cong et al., 2023, Gravina et al., 2024, Yang et al., 2025]. Distribution shift complicates *out-of-time* prediction [Quiñonero-Candela

Symbol	Meaning
V	Node set.
$e = (u, v, t, k, x_e)$	Interaction event: endpoints $u, v \in V$, time t, (optional) relation type k, and features x_e.
$\mathcal{S}$	Time-ordered event stream $\{e_\ell\}_{\ell=1}^{L}$.
$h_v(t) \in \mathbb{R}^d$	Backbone node state at time t (updated only at event times).
$\mathcal{D}$	Exogenous documents $\{(d_j, t_j^{\text{avail}})\}$.
$\mathbf{p}_i \in \mathbb{R}^{d_p}$	Structured exogenous token extracted from documents.
t_i	Token timestamp used for access control (default: availability-aligned).
$c_i \in [0, 1]$	Token extraction confidence score (proxy).
$\Pi_i \subseteq V$	Optional (noisy) entity prior for token routing.
$\mathcal{P}(t)$	Time-available token set (defined below).

Table 1: Notation. Time-available constraint: $\mathcal{P}(t) = \{(\mathbf{p}_i, t_i, c_i, \Pi_i) \mid t_i \leq t\}$.

et al., 2009, Ovadia et al., 2019, Lipton et al., 2018], while class imbalance motivates precision-based metrics alongside ROC discrimination [Saito and Rehmsmeier, 2015, Davis and Goadrich, 2006]. Calibration, label-shift correction, and domain generalization address related distributional concerns [Kull et al., 2017, Guo et al., 2017, Arjovsky et al., 2019, Gulrajani and Lopez-Paz, 2021]; our setting couples time-available documents with continuous-time graph evolution.

Exogenous text integration and controlled fusion. News, policy, and sentiment inform financial prediction [Tetlock, 2007, Baker et al., 2016], and FinBERT provides financial document encodings [Liu et al., 2020]. LLMs have also been studied for structured estimation in databases, such as cardinality prediction [Liu et al., 2026b]; ExoRoute instead extracts evidence-grounded document attributes. Gated/residual conditioning, including FiLM and highway networks [Perez et al., 2018, Srivastava et al., 2015], motivates lightweight modulation; ExoRoute additionally controls temporal availability, weights evidence by reliability and recency, and caps the residual update.

Resource-aware learning. Distributed learning optimizes gradient communication through adaptive Top-K sparsification [Ruan et al., 2024b] and power control with gradient compression in over-the-air aggregation [Ruan et al., 2024a]. These methods concern gradient transmission; ExoRoute allocates a document-retrieval budget.

Temporal adaptation and graph OOD methods. Graph adaptation and OOD methods, including EvolveGCN [Pareja et al., 2020] and EERM [Wu et al., 2022], use input formats and objectives different from our document-paired CTDG setting. We compare fusion mechanisms across CTDG backbones, including FiLM [Brockschmidt, 2020] (Sec. 5).

3 Problem Setup

We consider temporal graph prediction in financial interaction networks, including node-level anomaly/AML detection and edge-level link prediction. Inputs are a time-ordered event stream $\mathcal{S}$ and an exogenous document stream $\mathcal{D}$ (Table 1). At decision time t, the model must satisfy the *time-available* constraint: it may use only graph events and documents with timestamps $\leq t$, implemented via the accessible token set $\mathcal{P}(t)$.

Availability vs. effective time We timestamp tokens by document availability time by default: announcements released at t^{avail} can affect decisions for any $t \geq t^{\text{avail}}$, even if the described policy becomes effective later.

4 Methods

4.1 Overview

EXOROUTE augments an event-driven temporal GNN with a *time-available* exogenous channel. Given a temporal event stream $\mathcal{S}$ and a time-stamped document stream $\mathcal{D}$, it converts documents into auditable structured tokens and injects them *only at event times* as a bounded residual update to the backbone node state. As shown in Fig. 2, EXOROUTE (i) builds a time-indexed token memory from documents available up to each time, and (ii) performs reliability-weighted routing over a sliding

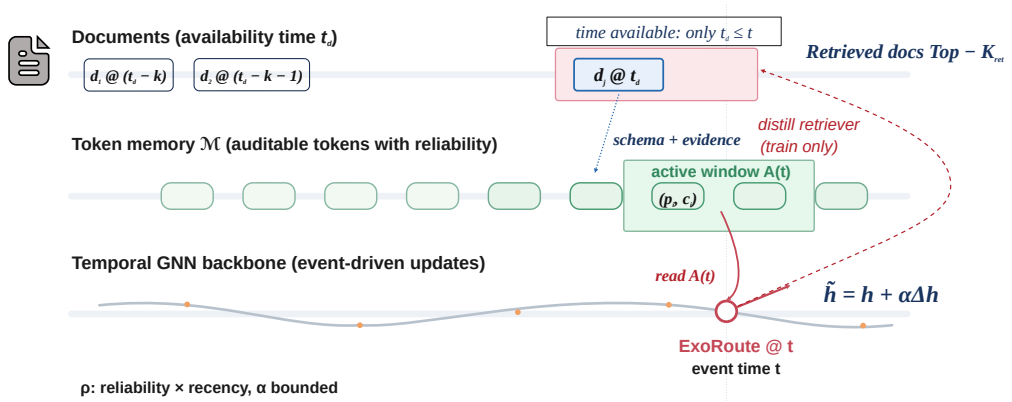

Figure 2: Overview of ExoRoute. Time-available documents become auditable schema tokens with relative reliability scores c_i, stored in a time-indexed memory. At event time t, reliability-weighted routing reads the active window $A(t)$ and supplies a capped residual update to the temporal GNN. Routing-to-retriever distillation occurs only during training; the scorer is fixed at inference.

window to compute an exogenous proposal, which is applied through a capped gate for stability under regime shifts. The document extractor is modular; our implementation uses schema-guided extraction with evidence spans for auditability.

4.2 Causal Document-to-Token Encoding

Goal. Given an event-time prediction problem, the exogenous channel must satisfy two requirements: (i) *time-available causality*—at time t we can only use documents available by t; (ii) *auditability*—the model should consume structured, inspectable signals rather than opaque embeddings. We therefore build a bin-aligned, strictly causal retrieval-and-extraction pipeline that produces a token memory

$$\mathcal{M} = \{(\mathbf{p}_i, t_i, c_i, r(i), \Pi_i)\},$$

where $\mathbf{p}_i$ is a structured token, c_i is an audit-derived relative reliability proxy, and $r(i)$ links each token back to its source document.

Bin-aligned causal retrieval. We partition the timeline into non-overlapping retrieval bins $\mathcal{B} = \{[b_k, b_{k+1})\}$. For each bin start b_k, we retrieve documents only from the time-available pool

$$\mathcal{C}_{b_k} \subseteq \{d \in \mathcal{D} \mid t_d \le b_k\}, \tag{1}$$

optionally restricted to a look-back window W_{ret}. All events with timestamps $t \in [b_k, b_{k+1})$ can only access tokens derived from $\mathcal{C}_{b_k}$, which enforces causality by construction.

Closed-loop retrieval adaptation via routing distillation. Generic semantic retrieval is often misaligned with the downstream temporal-graph objective, and manual relevance labels are unavailable. EXOROUTE learns a lightweight relevance scorer by distilling from the model's own token-routing behavior during training.

Concretely, for each candidate document $r \in \mathcal{C}_{b_k}$ we compute a score and a selection distribution:

$$s_\phi(r, b_k) = f_\phi\big(\text{enc}(r),\, \text{feat}(b_k)\big), \tag{2}$$

where f_ϕ is a lightweight MLP (we use a one-hidden-layer tanh scorer).

$$\pi_\phi(r \mid b_k) = \text{softmax}_{r \in \mathcal{C}_{b_k}}\Big(s_\phi(r, b_k)/\tau_{\text{sel}}\Big), \tag{3}$$

where $\text{enc}(\cdot)$ is a frozen text encoder and $\text{feat}(b_k)$ are optional bin-level features (e.g., time index). At inference we retrieve Top-K_{ret} documents by the fixed scorer s_ϕ learned during training; no retriever updates or distillation occur at inference. Retrieval is performed once per bin start b_k and reused for all events in $[b_k, b_{k+1})$.

To train π_ϕ without manual labels, we define a training-only target distribution from the downstream routing weights. Let $\mathcal{E}_k$ be the set of training events falling into bin $[b_k, b_{k+1})$.

We summarize how much routing mass the model assigns to document r within bin b_k by

$$m(r, b_k) = \frac{1}{|\mathcal{E}_k| + \epsilon} \sum_{(t,v) \in \mathcal{E}_k} \bar{a}_v(r, t), \tag{4}$$

where $\bar{a}_v(r,t) = \frac{1}{|\mathcal{T}(r,t)|+\epsilon} \sum_{i \in \mathcal{T}(r,t)} a_{vi}(t)$ and $\mathcal{T}(r,t) = \{i \in \mathcal{A}(t) \mid r(i) = r\}$. We then define

$$y^{\text{gate}}(r \mid b_k) = \text{Normalize}_{r \in \mathcal{C}_{b_k}}\big(m(r, b_k)\big). \tag{5}$$

This normalization ensures that the distillation target is not dominated by bins with more events, nor by documents that happen to produce more extracted tokens. We then optimize the retriever on *training bins only*:

$$\mathcal{L}_{\text{retr}} = \mathop{\mathbb{E}}_{b_k \in \mathcal{B}_{\text{train}}} \left[\text{KL}\big(\text{sg}\big(y^{\text{gate}}(\cdot \mid b_k)\big) \;\big\|\; \pi_\phi(\cdot \mid b_k)\big)\right]. \tag{6}$$

Here $\text{sg}(\cdot)$ denotes the stop-gradient operator.

The role of $\mathcal{L}_{\text{retr}}$ is purely to match the retriever to the downstream model's notion of utility. Temporal causality is preserved because (i) the candidate set for bin b_k is restricted to documents with $t_d \le b_k$, and (ii) the distillation target $y^{\text{gate}}(\cdot \mid b_k)$ is computed only from routing behavior on training-period events.

In Appendix A.2, we further justify this choice by showing that the routing-induced target is a noisy but useful proxy for a task-optimal retrieval distribution in the *score (logit) space*:

$$s_{\text{gate}}(r \mid b_k) \;=\; s^*(r \mid b_k) \;+\; \epsilon_r, \tag{7}$$

where $s^*(r \mid b_k)$ is an ideal task-optimal retrieval score under context b_k, ϵ_r is zero-mean noise and $y^{\text{gate}}(r \mid b_k) = \text{softmax}_r(s_{\text{gate}}(r \mid b_k))$.

Why this schema. We use a compact attribute schema because the downstream predictor does not need full free-form document semantics; instead, it needs a small set of temporally actionable factors that can be aligned across documents and checked by humans. The schema therefore focuses on fields that are both recurrent in the corpus and directly useful for event-time forecasting, while preserving auditability through field-level evidence.

Schema-guided extraction with evidence and confidence. Each retrieved document is converted into a list of structured attribute tuples $(\texttt{cat}_i, \texttt{trend}_i, \texttt{risk}_i, \texttt{duration}_i, \texttt{conf}_i^{\text{raw}}, \texttt{evi}_i)$ using a fixed JSON schema. We store verbatim evidence spans $\texttt{evi}_i$ for each field and require exact substring matching after deterministic normalization. Fields without valid evidence are dropped; evidence spans are retained for audit only and are not input to the GNN.

We map the extracted attributes to a continuous token and a reliability-weighted confidence:

$$\begin{aligned} \mathbf{p}_i &= \psi_\theta\big(\mathbf{a}_i\big), \\ \mathbf{a}_i &= [\text{emb}(\texttt{cat}_i),\, \texttt{trend}_i,\, \texttt{risk}_i,\, \texttt{duration}_i]. \end{aligned} \tag{8}$$

Thus only the schema fields in $\mathbf{a}_i$ enter $\mathbf{p}_i$; the evidence spans remain audit metadata.

In practice, a single document may yield multiple tuples. We first segment the document into candidate evidence-bearing spans, apply the schema extractor to each span, and keep all valid tuples whose fields can be grounded by evidence. Each resulting tuple is then converted into one token $\mathbf{p}_i$, so a document can contribute multiple timestamped tokens to $\mathcal{M}$ when it contains several distinct signals.

The quantity c_i is an audit-derived relative reliability proxy, rather than a calibrated posterior or node-relevance probability. It is computed from observable extraction cues such as the raw extractor confidence and timestamp metadata, so that relative reliability can be traced back to concrete extraction conditions.

$$c_i = \sigma\Big(\mathbf{w}_c^\top\, \xi\big(\texttt{conf}_i^{\text{raw}}, t_i\big)\Big). \tag{9}$$

We append each extracted item to $\mathcal{M}$ as $(\mathbf{p}_i, t_i, c_i, r(i), \Pi_i)$ and consume $\mathcal{M}$ via the bounded intervention module (Sec. 4.4).

Algorithm 1 EXOROUTE at event time t

Require: states $\{\mathbf{h}_v^{\text{base}}(t)\}$; memory $\mathcal{M}$; window W_{att}; decay γ; dropout p_d; cap $\alpha_{\max}$; gate logit g
Ensure: fused states $\{\tilde{\mathbf{h}}_v(t)\}$
$\mathcal{A}(t) \leftarrow \{i \mid t_i \in [t - W_{\text{att}}, t]\}$
if $\mathcal{A}(t) = \emptyset$ **then**
  return $\tilde{\mathbf{h}}_v(t) = \mathbf{h}_v^{\text{base}}(t)$ for all v
end if
(train) drop each $i \in \mathcal{A}(t)$ w.p. p_d
if $\mathcal{A}(t) = \emptyset$ after dropout **then**
  return $\tilde{\mathbf{h}}_v(t) = \mathbf{h}_v^{\text{base}}(t)$ for all v
end if
$\alpha \leftarrow \alpha_{\max}\sigma(g)$
$\rho_i(t) \leftarrow c_i \exp(-\gamma[t - t_i]_+)$ **for all** $i \in \mathcal{A}(t)$
$\mathbf{k}_i \leftarrow f_k(\mathbf{p}_i)$, $\mathbf{v}_i \leftarrow f_v(\mathbf{p}_i)$ **for all** $i \in \mathcal{A}(t)$
for each node v **do**
  $\mathbf{q}_v \leftarrow f_q(\mathbf{h}_v^{\text{base}}(t))$
  $S_{vi} \leftarrow \langle \mathbf{q}_v, \mathbf{k}_i \rangle / \sqrt{d_k}$ **for all** $i \in \mathcal{A}(t)$
  $a_{vi} \leftarrow \text{softmax}_i(S_{vi} + \log(\varepsilon + \rho_i(t)))$
  $\Delta\mathbf{h}_v \leftarrow \text{LN}(\sum_i a_{vi}\mathbf{v}_i)$
  $\tilde{\mathbf{h}}_v(t) \leftarrow \mathbf{h}_v^{\text{base}}(t) + \alpha\Delta\mathbf{h}_v$
end for
return $\{\tilde{\mathbf{h}}_v(t)\}$

4.3 Time-indexed Token Memory

Tokens are stored in a time-indexed memory $\mathcal{M}$. At event time t, EXOROUTE only attends to the active subset

$$\mathcal{A}(t) = \{i \mid t_i \in [t - W_{\text{att}},\, t]\}.$$

Because public benchmarks rarely provide reliable document–entity linking, we route tokens to nodes by default with fully soft attention (Sec. 4.4). If a coarse prior $\Pi_i \subseteq V$ is available, we use it only as optional *training-time* weak supervision on routing logits; inference remains soft and does not hard-mask tokens, which keeps the method robust to noisy priors.

4.4 Bounded Gated Fusion and Learning

Controlled interventions. EXOROUTE treats exogenous texts as a controlled update channel on top of a temporal GNN. We (i) route evidence through reliability-aware attention over a sliding-window memory, and (ii) apply the proposal through a capped residual, bounding its representation-level influence when proposal norms are bounded.

Reliability-tilted routing over the active memory. At event time t, we attend to the active token set $\mathcal{A}(t)$ defined in Sec. 4.3.

Each token carries a reliability-weighted confidence c_i and timestamp t_i, which define a reliability weight

$$\rho_i(t) = c_i \cdot \exp\big(-\gamma[t - t_i]_+\big). \tag{10}$$

Given node state $\mathbf{h}_v^{\text{base}}(t)$, we compute a state-aware similarity logit $S_{vi}(t)$ between v and token i, and form a reliability-tilted routing distribution

$$a_{vi}(t) = \text{softmax}_{i \in \mathcal{A}(t)}\Big(S_{vi}(t) + \log(\varepsilon + \rho_i(t))\Big). \tag{11}$$

We then aggregate an exogenous proposal update

$$\Delta\mathbf{h}_v(t) = \text{LN}\left(\sum_{i \in \mathcal{A}(t)} a_{vi}(t)\,\mathbf{v}_i\right), \tag{12}$$

where $\mathbf{v}_i$ is the projected token value. This matches the operational procedure in Algorithm 1: retrieve a sliding-window memory, downweight tokens by confidence and recency, then aggregate a normalized proposal update.

Bounded residual application and objectives. We inject the proposal update through a capped residual intervention,

$$\begin{aligned} \tilde{\mathbf{h}}_v(t) &= \mathbf{h}_v^{\text{base}}(t) + \alpha\, \Delta\mathbf{h}_v(t), \\ \alpha &= \alpha_{\max}\sigma(g) \in [0, \alpha_{\max}]. \end{aligned} \tag{13}$$

We initialize $g \ll 0$ so the model starts close to the backbone and only increases reliance on exogenous evidence when it consistently improves the downstream objective.

Under the stated direction and scale assumptions, this capped intervention yields contraction with rate $\eta = \sqrt{1 - 2\alpha\kappa\beta + \alpha^2\beta^2} < 1$ when $\delta > 0$ and $0 < \alpha < 2\kappa/\beta$; independently, the proposal bound $\|\Delta\mathbf{h}_v\| \le L_{\max}$ and gate cap give deterministic bounded influence $\|\alpha\Delta\mathbf{h}_v\| \le \alpha_{\max}L_{\max}$ (proof and conditions in Appendix A.1). These are representation-level guarantees and do not imply AP, loss, or retraining guarantees.

Training objective. We optimize the task prediction loss $\mathcal{L}_{\text{pred}}$ (e.g., BCE for node classification or link prediction) under strict temporal ordering. To encourage stable behavior under regime shifts, we additionally penalize large exogenous interventions:

$$\mathcal{L} = \mathcal{L}_{\text{pred}} + \lambda_\Delta \mathcal{L}_\Delta,\ \mathcal{L}_\Delta = \mathbb{E}_{t,v}\big\|\alpha\, \Delta\mathbf{h}_v(t)\big\|_2^2. \tag{14}$$

The intervention penalty discourages large updates; prediction gains are evaluated empirically. We additionally use optional regularizers for retriever distillation, routing entropy, and weak supervision from noisy entity priors; full definitions are in Appendix F.

If a coarse entity prior is available, it serves only as soft training-time regularization on routing logits; the induced gradient bias under class-conditional noise is bounded and linear in the noise rate (Appendix B).

5 Experiments and Evaluation

We evaluate EXOROUTE on real-world financial temporal graphs under strict temporal splits, reporting main results, component ablations, and retrieval analyses. Causality/compliance checks and overhead are in Appendix H.1.

Datasets. We evaluate on the community crypto graph datasets Elliptic (node-level AML) [Weber et al., 2019] and Bitcoin-Alpha/OTC (edge-level trust prediction) [Kumar et al., 2016] under strict temporal splits. The graph datasets provide event streams, while we construct the time-stamped document channels from public sources. Elliptic has a documented policy shock used for windowing; Alpha and OTC have no regime-change annotation (see Appendix O for a discussion of other candidate domains). Elliptic uses steps 1–34/35–39/40–49 for train/val/test; Alpha/OTC use 70/10/20% chronological splits. All preprocessing and protocol details are provided in Appendix D.1.

Auditable exogenous document stream. For each dataset, we build a time-stamped document stream from public sources with explicit publication times, and convert documents into auditable structured tokens offline. This design makes exogenous inputs (i) time-available by construction and (ii) traceable to source documents and evidence spans; the full registry is described in Appendix I. For Elliptic, we use OFAC sanctions lists, FinCEN advisories, and DOJ/FBI enforcement releases with 14-day retrieval bins and attention windows; for BTC-Alpha/OTC, we use Bitcointalk reputation/scam threads and CoinMetrics bulletins with 7-day bins and windows (full registry in Appendix I).

Time-causal protocol. We enforce strict temporal causality: (i) document access uses only $t_d \le t$; (ii) negatives for link prediction are sampled from nodes active by t; (iii) normalization statistics are computed on the training period only. Hyperparameters are tuned on the validation period using the task's primary metric (AP; AUC is secondary) and then fixed for test evaluation (reported settings in Appendix G).

$t = 40$–42 (Before Change)					
Model	**Variant**	**AUC**	**AP**	**Prec@5%**	**ECE**
TGAT	Original	0.953 ± 0.007	0.771 ± 0.022	0.638 ± 0.028	0.024 ± 0.006
	ExoRoute	$\mathbf{0.957 \pm 0.005}$	$\mathbf{0.791 \pm 0.018}$	$\mathbf{0.668 \pm 0.022}$	$\mathbf{0.016 \pm 0.005}$
TGN	Original	0.962 ± 0.006	0.796 ± 0.020	0.659 ± 0.026	0.052 ± 0.010
	ExoRoute	$\mathbf{0.966 \pm 0.005}$	$\mathbf{0.815 \pm 0.016}$	$\mathbf{0.671 \pm 0.022}$	$\mathbf{0.033 \pm 0.008}$
GraphMixer	Original	0.959 ± 0.006	0.787 ± 0.021	0.652 ± 0.027	0.018 ± 0.005
	ExoRoute	$\mathbf{0.961 \pm 0.005}$	$\mathbf{0.801 \pm 0.017}$	$\mathbf{0.679 \pm 0.022}$	$\mathbf{0.009 \pm 0.003}$
CTAN	Original	0.949 ± 0.008	0.779 ± 0.023	0.651 ± 0.027	0.038 ± 0.009
	ExoRoute	$\mathbf{0.962 \pm 0.006}$	$\mathbf{0.788 \pm 0.019}$	$\mathbf{0.654 \pm 0.024}$	$\mathbf{0.018 \pm 0.006}$
DyGFormer	Original	0.952 ± 0.008	0.751 ± 0.025	0.647 ± 0.030	0.049 ± 0.011
	ExoRoute	$\mathbf{0.963 \pm 0.006}$	$\mathbf{0.817 \pm 0.019}$	$\mathbf{0.677 \pm 0.023}$	$\mathbf{0.029 \pm 0.008}$
GeneralDyG	Original	0.959 ± 0.007	0.785 ± 0.021	0.661 ± 0.026	0.096 ± 0.016
	ExoRoute	$\mathbf{0.967 \pm 0.005}$	$\mathbf{0.818 \pm 0.017}$	$\mathbf{0.683 \pm 0.021}$	$\mathbf{0.019 \pm 0.006}$
$t = 43$–47 (After Change)					
Model	**Variant**	**AUC**	**AP**	**Prec@5%**	**ECE**
TGAT	Original	0.934 ± 0.011	0.565 ± 0.040	0.202 ± 0.030	0.017 ± 0.005
	ExoRoute	$\mathbf{0.945 \pm 0.009}$	$\mathbf{0.728 \pm 0.030}$	$\mathbf{0.279 \pm 0.026}$	$\mathbf{0.007 \pm 0.003}$
TGN	Original	0.941 ± 0.010	0.593 ± 0.036	0.261 ± 0.028	0.047 ± 0.009
	ExoRoute	$\mathbf{0.943 \pm 0.009}$	$\mathbf{0.705 \pm 0.028}$	$\mathbf{0.282 \pm 0.024}$	$\mathbf{0.027 \pm 0.007}$
GraphMixer	Original	0.932 ± 0.012	0.595 ± 0.038	0.269 ± 0.027	0.008 ± 0.003
	ExoRoute	$\mathbf{0.941 \pm 0.010}$	$\mathbf{0.697 \pm 0.030}$	$\mathbf{0.279 \pm 0.024}$	$\mathbf{0.005 \pm 0.002}$
CTAN	Original	0.936 ± 0.011	0.627 ± 0.035	0.246 ± 0.029	0.031 ± 0.007
	ExoRoute	$\mathbf{0.939 \pm 0.010}$	$\mathbf{0.725 \pm 0.029}$	$\mathbf{0.281 \pm 0.025}$	$\mathbf{0.011 \pm 0.005}$
DyGFormer	Original	0.939 ± 0.011	0.622 ± 0.038	0.268 ± 0.028	0.042 ± 0.010
	ExoRoute	$\mathbf{0.942 \pm 0.010}$	$\mathbf{0.654 \pm 0.032}$	$\mathbf{0.288 \pm 0.024}$	$\mathbf{0.021 \pm 0.007}$
GeneralDyG	Original	0.944 ± 0.010	0.602 ± 0.037	0.272 ± 0.027	0.085 ± 0.014
	ExoRoute	$\mathbf{0.948 \pm 0.009}$	$\mathbf{0.694 \pm 0.030}$	$\mathbf{0.273 \pm 0.023}$	$\mathbf{0.009 \pm 0.004}$

Table 2: Performance comparison of original models and EXOROUTE on Elliptic data. The two panels report the pre-shift and post-shift windows; values are mean±std over 5 seeds.

Models. We attach EXOROUTE to six representative temporal-graph backbones (TGAT, TGN, GraphMixer, DyGFormer, CTAN, GeneralDyG) without modifying their training pipeline. To isolate the effect of fusion design, we also compare five exogenous integration strategies—early concatenation, direct residual injection, ungated cross-attention, FiLM conditioning [Brockschmidt, 2020], and EXOROUTE—across the same backbones (Table 3; backbone details in Appendix D.2). We do not compare with discrete-time snapshot methods (e.g., EvolveGCN) or static graph OOD frameworks (e.g., EERM), as they operate on fundamentally different input formats and do not support exogenous document integration; see Sec. 2 for a detailed discussion.

5.1 Node-Level Performance

Table 2 evaluates Elliptic in two windows around the known policy-driven shock ($t = 40$–42 vs. $t = 43$–47). Across all six backbones, the post-shift window retains relatively stable AUC while AP/Prec@K decline for the Original variant. ECE is lower after the shift than before it for each Original backbone; within each window, EXOROUTE further reduces ECE relative to Original. EXOROUTE improves the ranking metrics and same-window ECE comparisons by injecting time-aligned, confidence-weighted exogenous evidence as a bounded update, without relying on future documents. Following prior practice for imbalanced AML, we report AP/Prec@K as primary metrics and AUC as secondary; ECE is computed per window as in Appendix C.

5.2 Edge-Level Performance

Table 3 reports edge-level trust prediction on BTC-OTC and BTC-Alpha with all fusion variants and ablations. Since no explicit regime-change annotations are available for Alpha/OTC, gains here indicate that EXOROUTE can exploit time-available exogenous context beyond a manually identified

shock. "Original" is the backbone without exogenous input, "w/o Aux" removes auxiliary objectives from the gated fusion, and intermediate columns compare alternative fusion mechanisms.

BTC-OTC								
Model	**Metric**	**Original**	**Concat**	**Residual**	**Cross-Attn**	**FiLM**	**w/o Aux**	**ExoRoute**
TGAT	AUC	$0.80 \pm .008$	$0.69 \pm .013$	$0.76 \pm .009$	$0.77 \pm .010$	$0.75 \pm .012$	$0.80 \pm .008$	$0.80 \pm .007$
	AP	$0.50 \pm .015$	$0.40 \pm .020$	$0.49 \pm .017$	$0.51 \pm .016$	$0.48 \pm .019$	$0.51 \pm .016$	$0.53 \pm .014$
TGN	AUC	$0.80 \pm .007$	$0.70 \pm .014$	$0.75 \pm .011$	$0.77 \pm .010$	$0.75 \pm .013$	$0.80 \pm .009$	$0.81 \pm .006$
	AP	$0.53 \pm .014$	$0.41 \pm .022$	$0.49 \pm .018$	$0.50 \pm .017$	$0.48 \pm .020$	$0.53 \pm .015$	$0.54 \pm .013$
GraphMixer	AUC	$0.81 \pm .006$	$0.69 \pm .015$	$0.77 \pm .010$	$0.78 \pm .009$	$0.76 \pm .014$	$0.81 \pm .007$	$0.81 \pm .006$
	AP	$0.54 \pm .014$	$0.44 \pm .023$	$0.51 \pm .019$	$0.53 \pm .017$	$0.49 \pm .021$	$0.55 \pm .016$	$0.56 \pm .012$
CTAN	AUC	$0.82 \pm .005$	$0.72 \pm .012$	$0.78 \pm .009$	$0.79 \pm .008$	$0.77 \pm .011$	$0.81 \pm .006$	$0.84 \pm .004$
	AP	$0.52 \pm .013$	$0.46 \pm .018$	$0.54 \pm .016$	$0.55 \pm .015$	$0.51 \pm .019$	$0.53 \pm .014$	$0.59 \pm .010$
DyGFormer	AUC	$0.81 \pm .007$	$0.70 \pm .013$	$0.77 \pm .011$	$0.78 \pm .010$	$0.76 \pm .012$	$0.82 \pm .007$	$0.82 \pm .006$
	AP	$0.53 \pm .015$	$0.43 \pm .021$	$0.50 \pm .018$	$0.52 \pm .017$	$0.49 \pm .020$	$0.54 \pm .015$	$0.56 \pm .013$
GeneralDyG	AUC	$0.77 \pm .010$	$0.67 \pm .016$	$0.74 \pm .013$	$0.75 \pm .012$	$0.72 \pm .015$	$0.78 \pm .011$	$0.79 \pm .009$
	AP	$0.45 \pm .017$	$0.36 \pm .024$	$0.44 \pm .020$	$0.45 \pm .019$	$0.42 \pm .023$	$0.46 \pm .019$	$0.49 \pm .016$
BTC-Alpha								
Model	**Metric**	**Original**	**Concat**	**Residual**	**Cross-Attn**	**FiLM**	**w/o Aux**	**ExoRoute**
TGAT	AUC	$0.78 \pm .009$	$0.68 \pm .014$	$0.74 \pm .011$	$0.76 \pm .010$	$0.74 \pm .013$	$0.77 \pm .009$	$0.79 \pm .008$
	AP	$0.56 \pm .016$	$0.46 \pm .021$	$0.54 \pm .018$	$0.55 \pm .017$	$0.50 \pm .022$	$0.58 \pm .015$	$0.59 \pm .014$
TGN	AUC	$0.72 \pm .011$	$0.66 \pm .015$	$0.73 \pm .012$	$0.74 \pm .011$	$0.71 \pm .014$	$0.74 \pm .010$	$0.78 \pm .008$
	AP	$0.49 \pm .018$	$0.42 \pm .023$	$0.49 \pm .019$	$0.51 \pm .018$	$0.47 \pm .021$	$0.50 \pm .017$	$0.55 \pm .015$
GraphMixer	AUC	$0.79 \pm .008$	$0.68 \pm .014$	$0.74 \pm .012$	$0.75 \pm .011$	$0.73 \pm .013$	$0.79 \pm .008$	$0.79 \pm .011$
	AP	$0.55 \pm .015$	$0.43 \pm .022$	$0.51 \pm .019$	$0.52 \pm .018$	$0.49 \pm .020$	$0.54 \pm .016$	$0.56 \pm .017$
CTAN	AUC	$0.79 \pm .007$	$0.70 \pm .013$	$0.76 \pm .010$	$0.77 \pm .009$	$0.75 \pm .012$	$0.79 \pm .007$	$0.81 \pm .005$
	AP	$0.58 \pm .014$	$0.48 \pm .020$	$0.55 \pm .017$	$0.57 \pm .016$	$0.52 \pm .021$	$0.57 \pm .015$	$0.61 \pm .012$
DyGFormer	AUC	$0.80 \pm .008$	$0.69 \pm .014$	$0.75 \pm .011$	$0.76 \pm .010$	$0.74 \pm .013$	$0.80 \pm .008$	$0.81 \pm .007$
	AP	$0.54 \pm .016$	$0.44 \pm .022$	$0.51 \pm .019$	$0.52 \pm .018$	$0.48 \pm .021$	$0.55 \pm .016$	$0.56 \pm .014$
GeneralDyG	AUC	$0.76 \pm .010$	$0.66 \pm .015$	$0.72 \pm .013$	$0.74 \pm .012$	$0.70 \pm .014$	$0.77 \pm .010$	$0.78 \pm .009$
	AP	$0.54 \pm .017$	$0.42 \pm .024$	$0.49 \pm .020$	$0.50 \pm .019$	$0.45 \pm .022$	$0.53 \pm .017$	$0.55 \pm .016$

Table 3: Edge-level trust prediction on BTC-OTC and BTC-Alpha. AUC and AP are shown on separate rows for each backbone; values are mean$\pm$std over 5 seeds. **Original**: backbone without exogenous input; **Concat/Residual/Cross-Attn/FiLM**: alternative fusion mechanisms; **w/o Aux**: EXOROUTE gated fusion without auxiliary objectives; **ExoRoute**: full model.

5.3 Ablation Study

Table 4 reports fully retrained channel manipulations. Perturbing content or time alignment hurts AP, with timestamp misalignment producing the largest drop. Fixing category, risk, trend, or duration also degrades performance, indicating complementary structured signals. The fusion comparisons in Table 3 show that concatenation, direct residual updates, and FiLM [Brockschmidt, 2020] can underperform the no-exogenous backbone, while cross-attention lacks an explicit reliability filter. EXOROUTE combines reliability-weighted routing with a capped update and improves AP across the reported backbones and trust datasets.

Transformation	BTC-Alpha		BTC-OTC	
	Δ AUC	Δ AP	Δ AUC	Δ AP
Real exogenous channel	0.0000	0.0000	0.0000	0.0000
Constant exogenous channel	-0.0845	-0.0553	-0.0637	-0.0441
Timestamp Misalignment	-0.1525	-0.1279	-0.1037	-0.0935
Constant category	-0.0249	-0.0214	-0.0318	-0.0294
Constant risk	-0.0307	-0.0315	-0.0379	-0.0421
Constant duration	-0.0179	-0.0178	-0.0267	-0.0204
Constant trend	-0.0176	-0.0165	-0.0233	-0.0201
Content shuffled (time preserved)	-0.0427	-0.0492	-0.0331	-0.0376
Timestamp-preserving constant-content	-0.0674	-0.0683	-0.0595	-0.0584

Table 4: Fully retrained channel-manipulation diagnostic (Δ AUC/AP relative to the real exogenous channel within each dataset). The timestamp-preserving constant-content control retains timestamps, confidence, token count, and active-window structure while fixing the content fields.

To assess predictive value beyond timing and shared global context, we run fully retrained controls on Elliptic with the TGN backbone in the post-shift window ($t = 43$–47). Time+count uses only causal bin/recency information and document count/coverage; Shared/global uses one query for all nodes. Clean schema-labels retains schema attributes and timestamps, sets $c_i \equiv 1$, selects the most recent eligible documents, and removes document identities, raw-text retrieval embeddings, and node-conditioned ranking. Schema + audit reliability adds extraction-reliability weights to this control. Table 5 reports the results.

Control	**AP**	**ΔAP**
Original	0.593 ± 0.036	—
Time+count	0.621 ± 0.031	$+0.028$
Shared/global query	0.646 ± 0.031	$+0.053$
Clean schema-labels ($c_i \equiv 1$)	0.656 ± 0.030	$+0.063$
Schema + audit reliability	0.663 ± 0.030	$+0.070$
Full node-conditioned routing	$\mathbf{0.705 \pm 0.028}$	$+0.112$

Table 5: Fully retrained semantic and timing controls on Elliptic/TGN at $t = 43$–47. AP is mean±sample SD over five matched seeds; ΔAP is relative to Original. The contrasts are not an additive causal decomposition. Full exceeds the shared/global control by 0.059 AP.

Beyond the core components, we ablate optional regularizers (retriever distillation, intervention control, and routing-shaping terms) and observe consistent improvements across datasets and backbones (Appendix Table 9). A dense LM baseline (DENSE LM CONCAT) that concatenates frozen FinBERT embeddings with node states underperforms both TOKEN CONCAT and EXOROUTE (Appendix L), supporting the value of schema-guided structuring. For retrieval adaptation, the distilled retriever with relative reliability-weighted confidence achieves the best ranking and calibration on the Elliptic post-shift window (Appendix Table 11), and EXOROUTE degrades gracefully under sparse and noisy document streams (Appendix M).

In a held-out document-intervention test on Elliptic/TGN, the distilled retrieval score correlates with document utility (Spearman 0.43 ± 0.06), while the shuffled-target and random-ranking variants achieve 0.04 ± 0.05 and 0.00 ± 0.04, respectively (Appendix K.1). Removing the top-scored 10% of documents reduces AP by 0.043 ± 0.012, compared with 0.008 ± 0.006 for exposure-matched random removal (Table 12). This supports downstream operational utility in the evaluated setting; the unbiased-proxy condition remains a modeling assumption.

Finally, on two supplementary self-constructed non-crypto financial pilots (consumer credit and equity; Appendix P), EXOROUTE reproduces the core pattern: bounded gating preserves pre-shift AUC while recovering +4–11 post-shift AP, with larger gains under sharper shifts (S&P 500 / COVID-19).

6 Conclusion

EXOROUTE fuses structured, time-available text through reliability-aware routing and a capped update. Our analysis bounds representation changes for bounded proposals and establishes contraction under additional alignment, scale, and step-size conditions. Distillation learns a retriever fixed at inference. Across six backbones, EXOROUTE improves AP on three crypto graphs and reduces Elliptic post-shift ECE relative to Original. Channel manipulations support the value of content and time alignment.

Limitations. EXOROUTE requires relevant, time-stamped documents and depends on their quality and availability. Non-crypto evaluation uses self-constructed pilots. Appendix Q discusses extraction trade-offs and the need for standard non-crypto benchmarks with paired document streams.

Acknowledgments and Disclosure of Funding

This work was supported by Ant Group Research Fund and the Fundamental Research Funds for the Central Universities, Peking University. The authors declare no competing interests.

References

Daron Acemoglu, Asuman Ozdaglar, and Alireza Tahbaz-Salehi. Systemic risk and stability in financial networks. *American Economic Review*, 105(2):564–608, 2015. doi: 10.1257/AER.20130456.

L. Akoglu, H. Tong, and D. Koutra. Graph based anomaly detection and description: a survey. *Data Mining and Knowledge Discovery*, 29(3):626–688, 2015. doi: 10.1007/s10618-014-0365-y.

Franklin Allen and Douglas Gale. Financial contagion. *Journal of Political Economy*, 108(1):1–33, 2000. doi: 10.1086/262109.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv.org*, 2019. URL `https://arxiv.org/abs/1907.02893`.

Scott R Baker, Nicholas Bloom, and Steven J Davis. Measuring economic policy uncertainty. *The Quarterly Journal of Economics*, 131(4):1593–1636, 2016. doi: 10.1093/qje/qjw024.

Scott R Baker, Nicholas Bloom, Steven J Davis, Kyle Kost, Marco Sammon, and Tasaneeya Viratyosin. The unprecedented stock market reaction to COVID-19. *Review of Asset Pricing Studies*, 10(4):742–758, 2020. doi: 10.1093/rapstu/raaa008.

Marc Brockschmidt. GNN-FiLM: Graph neural networks with feature-wise linear modulation. In *International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1144–1152. PMLR, 2020. URL `https://proceedings.mlr.press/v119/brockschmidt20a.html`.

Weilin Cong, Si Zhang, Jian Kang, Baichuan Yuan, Hao Wu, Xin Zhou, Hanghang Tong, and Mehrdad Mahdavi. Do we really need complicated model architectures for temporal networks? In *International Conference on Learning Representations*, 2023. doi: 10.48550/arXiv.2302.11636. URL `https://arxiv.org/abs/2302.11636`.

Jesse Davis and Mark Goadrich. The relationship between precision-recall and ROC curves. In *International Conference on Machine Learning*, pages 233–240. ACM Press, 2006. doi: 10.1145/1143844.1143874.

J. Gama, Indrė Žliobaitė, A. Bifet, Mykola Pechenizkiy, and A. Bouchachia. A survey on concept drift adaptation. *ACM Computing Surveys*, 46(4):1–37, 2014. doi: 10.1145/2523813. URL `https://doi.org/10.1145/2523813`.

Paul Glasserman and H Peyton Young. How likely is contagion in financial networks? *Journal of Banking & Finance*, 50:383–399, 2015. doi: 10.1016/j.jbankfin.2014.02.006.

Alessio Gravina, Giulio Lovisotto, Claudio Gallicchio, Davide Bacciu, and Claas Grohnfeldt. Long range propagation on continuous-time dynamic graphs. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 16206–16225, 2024. doi: 10.48550/arXiv.2406.02740. URL `https://arxiv.org/abs/2406.02740`.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR, 2017. URL `https://proceedings.mlr.press/v70/guo17a.html`.

Shenyang Huang, Farimah Poursafaei, Jacob Danovitch, Matthias Fey, Weihua Hu, Emanuele Rossi, Jure Leskovec, Michael Bronstein, Guillaume Rabusseau, and Reihaneh Rabbany. Temporal graph benchmark for machine learning on temporal graphs. In *Advances in Neural Information Processing Systems*, volume 36, pages 2056–2073. Curran Associates, Inc., 2023. doi: 10.48550/arXiv.2307.01026. URL `https://arxiv.org/abs/2307.01026`.

Amir E. Khandani, Adlar J. Kim, and Andrew W. Lo. Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, 34(11):2767–2787, 2010. doi: 10.1016/j.jbankfin.2010.06.001.

Meelis Kull, Telmo Silva Filho, and Peter Flach. Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. In Aarti Singh and Jerry Zhu, editors, *International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 623–631. PMLR, 20–22 Apr 2017. URL https://proceedings.mlr.press/v54/kull17a.html.

Srijan Kumar, Francesca Spezzano, V. S. Subrahmanian, and Christos Faloutsos. Edge weight prediction in weighted signed networks. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 221–230. IEEE, 2016. doi: 10.1109/ICDM.2016.0033. URL https://cs.stanford.edu/~srijan/pubs/wsn-icdm16.pdf.

Stefan Lessmann, Bart Baesens, Hsin-Vonn Seow, and Lyn C. Thomas. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1):124–136, 2015. doi: 10.1016/j.ejor.2015.05.030.

Zachary C Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3122–3130. PMLR, 2018. URL https://proceedings.mlr.press/v80/lipton18a.html.

Liangzu Liu, Mengzhe Ruan, Xiaotian Chen, Haonan Chen, Xudong Niu, Wendi Yuan, Yuechen Li, Yang Liu, and Guanjun Wang. ACRM: Multi-agent trajectory learning for automated credit risk model refreshing in production. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 942–956. Association for Computational Linguistics, 2026a. doi: 10.18653/v1/2026.acl-industry.65. URL https://aclanthology.org/2026.acl-industry.65/.

Liangzu Liu, Yinjun Wu, Yiyan Wang, Zhuo Chang, Runze Su, Peizhi Wu, Jianjun Chen, Fuxin Jiang, Rui Shi, Bin Cui, and Tieying Zhang. Can large language models be a cardinality estimator? an empirical study. *The VLDB Journal*, 35:43, 2026b. doi: 10.1007/s00778-026-00975-7.

Zhuang Liu, Degen Huang, Kaiyu Huang, Zhuang Li, and Jun Zhao. FinBERT: A pre-trained financial language representation model for financial text mining. In *International Joint Conference on Artificial Intelligence*, pages 4513–4519. International Joint Conferences on Artificial Intelligence Organization, 2020. doi: 10.24963/ijcai.2020/622. URL https://doi.org/10.24963/ijcai.2020/622.

Atif Mian and Amir Sufi. The consequences of mortgage credit expansion: Evidence from the US mortgage default crisis. *The Quarterly Journal of Economics*, 124(4):1449–1496, 2009. doi: 10.1162/qjec.2009.124.4.1449.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

A. Pareja, Giacomo Domeniconi, Jie Chen, Tengfei Ma, T. Suzumura, H. Kanezashi, Tim Kaler, Tao B. Schardl, and Charles E. Leiserson. EvolveGCN: Evolving graph convolutional networks for dynamic graphs. In *AAAI Conference on Artificial Intelligence*, volume 34, pages 5363–5370, 2020. doi: 10.1609/AAAI.V34I04.5984.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. FiLM: Visual reasoning with a general conditioning layer. In *AAAI Conference on Artificial Intelligence*, volume 32, 2018. doi: 10.1609/aaai.v32i1.11671. URL https://arxiv.org/abs/1709.07871.

Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, V. Larivière, A. Beygelzimer, Florence d'Alché-Buc, E. Fox, and H. Larochelle. Improving reproducibility in machine learning research (a report from the NeurIPS 2019 reproducibility program). *Journal of Machine Learning Research*, 22(164):1–20, 2021.

Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence, editors. *Dataset Shift in Machine Learning*. The MIT Press, 2009. doi: 10.7551/MITPRESS/9780262170055.001.0001.

Emanuele Rossi, Ben Chamberlain, Fabrizio Frasca, Davide Eynard, Federico Monti, and Michael Bronstein. Temporal graph networks for deep learning on dynamic graphs. *arXiv.org*, 2020. URL https://arxiv.org/abs/2006.10637.

Mengzhe Ruan, Yunhe Li, Weizhou Zhang, Linqi Song, and Weitao Xu. Optimal power control for over-the-air federated learning with gradient compression. In *Proceedings of the 2024 IEEE 30th International Conference on Parallel and Distributed Systems (ICPADS)*, pages 326–333. IEEE, 2024a. doi: 10.1109/ICPADS63350.2024.00050.

Mengzhe Ruan, Guangfeng Yan, Yuanzhang Xiao, Linqi Song, and Weitao Xu. Adaptive Top-K in SGD for communication-efficient distributed learning in multi-robot collaboration. *IEEE Journal of Selected Topics in Signal Processing*, 18(3):487–501, 2024b. doi: 10.1109/JSTSP.2024.3381373.

Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3):e0118432, 2015. doi: 10.1371/journal.pone.0118432. URL https://doi.org/10.1371/journal.pone.0118432.

David Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-François Crespo, and Dan Dennison. Hidden technical debt in machine learning systems. In *Advances in Neural Information Processing Systems*, volume 28, pages 2503–2511, 2015. URL https://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.

Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. Highway networks. *arXiv.org*, 2015. URL https://arxiv.org/abs/1505.00387.

Junwei Su and Shan Wu. Temporal-aware evaluation and learning for temporal graph neural networks. In *AAAI Conference on Artificial Intelligence*, volume 39, pages 20628–20636, 2025. doi: 10.1609/aaai.v39i19.34273. URL https://ojs.aaai.org/index.php/AAAI/article/view/34273.

Paul C Tetlock. Giving content to investor sentiment: The role of media in the stock market. *Journal of Finance*, 62(3):1139–1168, 2007. doi: 10.1111/j.1540-6261.2007.01232.x.

Rakshit Trivedi, Mehrdad Farajtabar, Prasenjeet Biswal, and Hongyuan Zha. DyRep: Learning representations over dynamic graphs. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1803.04051.

Mark Weber, Giacomo Domeniconi, Jie Chen, Daniel Karl I. Weidele, Claudio Bellei, Tom Robinson, and Charles E. Leiserson. Anti-money laundering in bitcoin: Experimenting with graph convolutional networks for financial forensics. *arXiv.org*, 2019. URL https://arxiv.org/abs/1908.02591.

Qitian Wu, Hengrui Zhang, Junchi Yan, and David Wipf. Handling distribution shifts on graphs: An invariance perspective. In *International Conference on Learning Representations*, 2022.

Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S. Yu. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1):4–24, 2021. doi: 10.1109/TNNLS.2020.2978386. URL https://ieeexplore.ieee.org/document/9046288/.

Da Xu, Chuanwei Ruan, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. Inductive representation learning on temporal graphs. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/2002.07962.

Xiao Yang, Xuejiao Zhao, and Zhiqi Shen. A generalizable anomaly detection method in dynamic graphs. In *AAAI Conference on Artificial Intelligence*, volume 39, pages 22001–22009, 2025. doi: 10.1609/aaai.v39i20.35508. URL https://ojs.aaai.org/index.php/AAAI/article/view/35508.

Le Yu, Leilei Sun, Bowen Du, and Weifeng Lv. Towards Better Dynamic Graph Learning: New Architecture and Unified Library. In *Advances in Neural Information Processing Systems*, volume 36, pages 67686–67700. Curran Associates, Inc., 2023. doi: 10.48550/arXiv.2303.13047.

Jiasheng Zhang, Jialin Chen, Menglin Yang, Aosong Feng, Shuang Liang, Jie Shao, and Rex Ying. DTGB: A comprehensive benchmark for dynamic text-attributed graphs. In *Advances in Neural Information Processing Systems*, volume 37, pages 91405–91429. Curran Associates, Inc., 2024. doi: 10.48550/arXiv.2406.12072.

Table 6: Notation Summary for Theoretical Analysis

Symbol	Type	Definition
Bounded Gating (Sec 4.4)		
h_v^*	$\mathbb{R}^d$	Ideal node state under current regime
h_v^{base}	$\mathbb{R}^d$	Base GNN's node state
$\tilde{h}_v$	$\mathbb{R}^d$	Fused (updated) node state: $\tilde{h}_v = h_v^{\text{base}} + \alpha\,\Delta h_v$
Δh_v	$\mathbb{R}^d$	Exogenous correction proposal
Δh_v^*	$\mathbb{R}^d$	Ideal correction: $\Delta h_v^* = h_v^* - h_v^{\text{base}}$
α	$[0,1]$	Gating coefficient (fusion strength)
$\alpha_{\max}$	$[0,1]$	Cap on gating coefficient: $0 \le \alpha \le \alpha_{\max}$
ρ	$[0,1]$	Signal reliability (attention/confidence/recency weighted; see Eq. (7))
B	$\mathbb{R}^+$	Error bound constant s.t. $\|\Delta h_v - \Delta h_v^*\| \le (1-\rho)B$
κ	$(0,1]$	Direction alignment coefficient (cosine similarity) between Δh_v and Δh_v^*
β	$\mathbb{R}^+$	Proposal scale ratio: $\beta = \|\Delta h_v\|/\|h_v^{\text{base}} - h_v^*\| = \|\Delta h_v\|/\delta$
η	$(0,1)$	Error contraction rate: $\eta = \sqrt{1 - 2\alpha\kappa\beta + \alpha^2\beta^2}$
$L_{\max}$	$\mathbb{R}^+$	Upper bound on proposal magnitude: $\|\Delta h_v\| \le L_{\max}$
Closed-Loop Distillation (Sec 4.2)		
$y_{\text{gate}}(r\|b_k)$	$\Delta^{\|R\|}$	Gated routing distribution over documents
$\pi_\phi(r\|b_k)$	$\Delta^{\|R\|}$	Retriever's predicted document distribution (param. by ϕ)
$p^*(r\|b_k)$	$\Delta^{\|R\|}$	Task-optimal document distribution
b_k	$\mathcal{B}$	Query context (e.g., company-event pair)
r	$\mathcal{R}$	Retrieved document (regulatory text, news, etc.)
ϕ	Φ	Retriever's learnable parameters
$\mathcal{L}_{\text{retr}}$	$\mathbb{R}^+$	Retrieval distillation loss, e.g., $\mathbb{E}_{b_k}[\mathrm{KL}(y_{\text{gate}}(\cdot\|b_k) \,\|\, \pi_\phi(\cdot\|b_k))]$
$\mathrm{KL}(\cdot\|\cdot)$	$\mathbb{R}^+$	Kullback–Leibler divergence used for distillation
$s_{\text{gate}}(r\|b_k)$	$\mathbb{R}$	Routing-induced retrieval score/logit for document r under context b_k
$s^*(r\|b_k)$	$\mathbb{R}$	Ideal task-optimal retrieval score/logit under context b_k
ϵ_r	$\mathbb{R}$	Additive noise term in score space: $s_{\text{gate}}(r\|b_k) = s^*(r\|b_k) + \epsilon_r$
$u(r, b_k)$	$\mathbb{R}^+$	Utility proxy for r under b_k (e.g., gradient-based sensitivity of loss)
λ	$\mathbb{R}^+$	Temperature coefficient (e.g., in $p^*(r\|b_k) \propto \exp(\lambda I(r; Y\|b_k))$)
$I(r; Y\|b_k)$	$\mathbb{R}^+$	Mutual information between doc r and label Y conditioned on b_k
General Notation		
δ	$\mathbb{R}^+$	Base model's deviation: $\delta = \|h_v^{\text{base}} - h_v^*\|$
ϵ	$\mathbb{R}^d$	Approximation error/noise term
t	$\mathbb{Z}^+$	Time step in dynamic graphs
v	$\mathcal{V}$	Node in graph

A Theoretical Analysis

A.1 Bounded Gated Fusion: Bounded Influence and Conditional Contraction

Formal Setup. Let $h_v^*(t)$ denote the *ideal state* of node v at time t that achieves optimal prediction under current regimes. The base GNN state $h_v^{\text{base}}(t)$ deviates after regime shifts:

$$\|h_v^{\text{base}}(t) - h_v^*(t)\| = \delta.$$

Define the ideal correction $\Delta h_v^*(t) = h_v^*(t) - h_v^{\text{base}}(t)$, hence $\|\Delta h_v^*(t)\| = \delta$. The exogenous proposal $\Delta h_v(t)$ estimates $\Delta h_v^*(t)$, and the gated update is

$$\tilde{h}_v(t) = h_v^{\text{base}}(t) + \alpha\,\Delta h_v(t). \tag{15}$$

We define *signal reliability* $\rho \in [0, 1]$ as a weighted average over source nodes:

$$\rho = \frac{\sum_{i \in A(t)} a_{vi}(t) \cdot c_i \exp(-\gamma \Delta t_i)}{\sum_{i \in A(t)} a_{vi}(t)}, \tag{16}$$

where c_i is the audit-derived relative reliability proxy defined in Eq. 9, and Δt_i is the temporal gap. The relationship between this proxy and approximation error is an analysis assumption, not a calibrated posterior or node-relevance probability. Assume the approximation error is controlled by ρ:

$$\|\Delta h_v(t) - \Delta h_v^*(t)\| \leq (1 - \rho)\, B, \quad B > 0. \tag{17}$$

Lemma A.1 (Error Decomposition)**.** *For $\alpha \geq 0$, the updated state error satisfies:*

$$\|\tilde{h}_v(t) - h_v^*(t)\| \leq |1 - \alpha|\delta + \alpha(1 - \rho)B. \tag{18}$$

Proof.

$$\begin{aligned}
\|\tilde{h}_v - h_v^*\| &= \|h_v^{\text{base}} + \alpha \Delta h_v - h_v^*\| \\
&= \|(1 - \alpha)(h_v^{\text{base}} - h_v^*) + \alpha(\Delta h_v - \Delta h_v^*)\| \\
&\leq |1 - \alpha| \|h_v^{\text{base}} - h_v^*\| + \alpha \|\Delta h_v - \Delta h_v^*\| \\
&\leq |1 - \alpha|\delta + \alpha(1 - \rho)B.
\end{aligned}$$

Direction Consistency Assumption. Assume the angle between $\Delta h_v(t)$ and $\Delta h_v^*(t)$ is acute:

$$\begin{gathered}
\langle \Delta h_v(t), \Delta h_v^*(t) \rangle \geq \kappa \|\Delta h_v(t)\| \|\Delta h_v^*(t)\|, \\
\kappa = \cos\theta \in (0, 1].
\end{gathered} \tag{19}$$

This holds when the attention-weighted set $A(t)$ concentrates on nodes whose signals align with the true regime shift.

Let $\|\Delta h_v(t)\| = \beta\delta$ for some scale factor $\beta > 0$.

Theorem A.2 (Error Contraction)**.** *Assume $\delta > 0$. Under direction consistency and $\|\Delta h_v(t)\| = \beta\delta$, if $\alpha \geq 0$ and $0 < \alpha < 2\kappa/\beta$, then:*

$$\|\tilde{h}_v(t) - h_v^*(t)\| \leq \eta\delta, \eta = \sqrt{1 - 2\alpha\kappa\beta + \alpha^2\beta^2} < 1. \tag{20}$$

The minimum of this contraction upper bound is $\eta_{\min} = \sqrt{1 - \kappa^2}$ at $\alpha = \kappa/\beta$, provided this gate value is feasible under the implementation's gate cap. If $\delta = 0$ and $\|\Delta h_v(t)\| = \beta\delta = 0$, the update is zero and the conclusion is the trivial exact-base-state case.

Proof. Let $\mathbf{u} = \Delta h_v^*(t)/\|\Delta h_v^*(t)\| = (h_v^*(t) - h_v^{\text{base}}(t))/\delta$. Then:

$$\begin{aligned}
\|\tilde{h}_v - h_v^*\|^2 &= \| - \delta\mathbf{u} + \alpha \Delta h_v\|^2 \\
&= \delta^2 - 2\alpha\delta\langle \mathbf{u}, \Delta h_v \rangle + \alpha^2 \|\Delta h_v\|^2 \\
&= \delta^2 - 2\alpha\langle \Delta h_v^*, \Delta h_v \rangle + \alpha^2 \|\Delta h_v\|^2 \\
&\leq \delta^2 - 2\alpha\kappa \|\Delta h_v^*\| \|\Delta h_v\| \\
&\quad + \alpha^2 \|\Delta h_v\|^2 \\
&= \delta^2 \big(1 - 2\alpha\kappa\beta + \alpha^2\beta^2\big).
\end{aligned}$$

Thus $\|\tilde{h}_v - h_v^*\| \leq \eta\delta$. For $\eta < 1$, require $1 - 2\alpha\kappa\beta + \alpha^2\beta^2 < 1$, i.e., $\alpha(2\kappa - \alpha\beta) > 0$, which holds when $0 < \alpha < 2\kappa/\beta$.

Theorem A.3 (Bounded Influence (Deterministic))**.** *Assume the gate is capped as $0 \leq \alpha \leq \alpha_{\max}$ and the proposal magnitude is bounded as $\|\Delta h_v(t)\| \leq L_{\max}$. Then, for any $\rho \in [0, 1]$,*

$$\|\tilde{h}_v(t) - h_v^{\textit{base}}(t)\| \leq \alpha_{\max} L_{\max}, \tag{21}$$

and the error w.r.t. the ideal state satisfies:

$$\|\tilde{h}_v(t) - h_v^*(t)\| \leq \delta + \alpha_{\max} L_{\max}. \tag{22}$$

Proof. We have

$$\begin{aligned}\|\tilde{h}_v - h_v^{\text{base}}\| &= \|\alpha \Delta h_v\| \\ &\le \alpha_{\max} \|\Delta h_v\| \\ &\le \alpha_{\max} L_{\max}.\end{aligned}$$

Then, by the triangle inequality,

$$\begin{aligned}\|\tilde{h}_v - h_v^*\| &\le \|\tilde{h}_v - h_v^{\text{base}}\| + \|h_v^{\text{base}} - h_v^*\| \\ &\le \alpha_{\max} L_{\max} + \delta.\end{aligned}$$

Remark. When both Theorem A.2 and Theorem A.3 apply, their bounds jointly yield

$$\|\tilde{h}_v(t) - h_v^*(t)\| \le \min\{\eta\delta,\ \delta + \alpha_{\max} L_{\max}\},$$

where the contraction bound applies only when $\delta > 0$, direction consistency holds, $\alpha \ge 0$, $0 < \alpha < 2\kappa/\beta$, and the gate value is feasible; the bounded-influence bound requires the gate cap and $L_{\max}$ assumption.

A.2 Closed-Loop Retrieval Distillation: A Consistency View

Bi-Level Optimization Formulation. For analysis, we isolate the prediction and distillation terms in the following bi-level abstraction; the full implementation objective, including optional regularizers, is given in Appendix F.

- **Outer loop (downstream model).** Train the base model parameters θ by $\min_\theta \mathcal{L}_{\text{pred}}(\theta)$.
- **Inner loop (retriever distillation).** For a fixed (or slowly updated) θ, train the retriever parameters ϕ by

$$\min_{\phi}\ \mathbb{E}_{b_k}\Big[\mathrm{KL}\big(y_{\text{gate}}(\cdot \mid b_k) \parallel \pi_\phi(\cdot \mid b_k)\big)\Big]. \tag{23}$$

Task-Optimal Retrieval (Score Form). Rather than reasoning directly in probability space, we define an ideal task-optimal retrieval *score* $s^*(r \mid b_k)$ as any scoring function that ranks documents by their (unknown) marginal utility for reducing prediction loss under bin b_k. The corresponding task-optimal retrieval distribution is

$$p^*(r \mid b_k)\ =\ \mathrm{softmax}_r\big(s^*(r \mid b_k)\big). \tag{24}$$

(Optionally, one may instantiate $s^*(r \mid b_k) = \lambda\, I(r; Y \mid b_k)$, but our argument below only requires s^* to represent an ideal utility-aligned ranking.)

Theorem A.4 (Conditional Routing-Target Consistency in Score Space)**.** *Under a local linearity approximation of the downstream model around its current parameters and inputs, and the additional modeling assumption that routing aggregates an ideal utility-aligned score with zero-mean estimation noise, the routing-induced score $s_{\text{gate}}(r \mid b_k)$ satisfies*

$$s_{\text{gate}}(r \mid b_k)\ =\ s^*(r \mid b_k)\ +\ \epsilon_r, \mathbb{E}[\epsilon_r] = 0. \tag{25}$$

Consequently, the expected routing score satisfies $\mathbb{E}[s_{\text{gate}}(r \mid b_k)] = s^*(r \mid b_k)$ and therefore preserves the ideal score ordering in expectation. The additive-noise condition alone does not imply that $\mathbb{E}[\mathrm{softmax}(s_{\text{gate}})]$ preserves the ordering of $\mathrm{softmax}(s^*)$; the latter is not claimed here.

Proof sketch. Let x_r denote the representation contributed by document r to the downstream model. Consider the per-example loss ℓ and the local linearization $\ell(x_r + \delta) \approx \ell(x_r) + \nabla_{x_r}\ell^\top \delta$. A standard saliency-style utility proxy is the sensitivity of loss to x_r, e.g.,

$$u(r, b_k)\ :=\ \mathbb{E}_{(t,v)\in E_k}\big[\|\nabla_{x_r}\ell\|\big], \tag{26}$$

which measures how strongly the loss depends on information coming from r under bin b_k. The local sensitivity $u(r, b_k)$ is an interpretive utility proxy. We additionally assume that the routing mechanism aggregates an ideal utility-aligned score (which may be represented by u up to an affine transform) with zero-mean estimation noise. Under this explicit assumption, $s_{\text{gate}}(r \mid b_k) = s^*(r \mid b_k) + \epsilon_r$; local linearity alone does not establish the proxy identity or the zero-mean property.

Theorem A.5 (Optimality of KL Distillation Objective)**.** *For any fixed b_k, the distribution $\pi(\cdot \mid b_k)$ that minimizes $\mathrm{KL}(y_{\mathrm{gate}}(\cdot \mid b_k) \,\|\, \pi(\cdot \mid b_k))$ over all distributions is $\pi(\cdot \mid b_k) = y_{\mathrm{gate}}(\cdot \mid b_k)$. Therefore, if the retriever family $\{\pi_\phi\}$ is sufficiently expressive and optimization is successful, $\pi_\phi(\cdot \mid b_k)$ can approximate $y_{\mathrm{gate}}(\cdot \mid b_k)$.*

Proof. For fixed y_{gate}, $\mathrm{KL}(y_{\mathrm{gate}} \| \pi)$ is minimized at $\pi = y_{\mathrm{gate}}$ over the probability simplex (equivalently, minimizing cross-entropy w.r.t. π).

Global Feedback Justification (Empirical). While joint optimization of (θ, ϕ) is non-convex and full convergence is not guaranteed, empirically we observe monotonic improvements: better retrieval $\rightarrow$ more reliable routing targets $y_{\mathrm{gate}} \rightarrow$ improved $\pi_\phi \rightarrow$ enhanced future retrieval.

B Robustness to Noisy Entity Priors: A Gradient Bias Bound

Motivation. Some datasets provide only coarse or heuristic document–entity priors (Sec. 4.3). We use such priors *only during training* as a soft regularizer on routing logits, and keep inference fully soft. This appendix shows that, under a standard class-conditional noise model, the gradient induced by a noisy prior deviates from the clean gradient by a controlled amount that scales *linearly* with the noise rate; label smoothing further shrinks the bias.

B.1 Setup

Let S_{vi} denote a routing logit (e.g., the similarity score used before softmax) between node v and token i. Assume an (unobserved) clean binary association mask $M^*_{vi} \in \{0, 1\}$ and an observed noisy mask $\tilde{M}_{vi} \in \{0, 1\}$.

Assumption B.1 (Class-conditional noise)**.** The observed mask is generated from the clean mask via class-conditional flips:

$$\Pr(\tilde{M}_{vi} = 1 \mid M^*_{vi} = 0) = \eta_0,$$
$$\Pr(\tilde{M}_{vi} = 0 \mid M^*_{vi} = 1) = \eta_1,$$

with $0 \le \eta_0, \eta_1 < 1/2$ and $\eta_{\max} = \max(\eta_0, \eta_1)$.

We apply label smoothing with $\varepsilon_{\mathrm{ls}} \in (0, 1/2)$:

$$\bar{y} = \varepsilon_{\mathrm{ls}} + (1 - 2\varepsilon_{\mathrm{ls}})y.$$

Define the (default) unweighted auxiliary loss

$$\ell_{\mathrm{uw}}(S_{vi}, \tilde{M}_{vi}) = \mathrm{BCE}\Big(\sigma(S_{vi}),\ \overline{\tilde{M}}_{vi}\Big), \tag{27}$$

and the corresponding clean counterpart

$$\ell_{\mathrm{clean}}(S_{vi}, M^*_{vi}) = \mathrm{BCE}\big(\sigma(S_{vi}),\ \bar{M}^*_{vi}\big), \tag{28}$$

where σ is the sigmoid and $\mathrm{BCE}(p, y) = -y \log p - (1 - y) \log(1 - p)$.

Assumption B.2 (Bounded Jacobian)**.** For parameters θ that produce the routing logit $S_{vi}(\theta)$, we assume $\|\partial S_{vi}/\partial\theta\| \le L_J$ almost surely.

B.2 Main result

Lemma B.3 (Gradient of label-smoothed BCE w.r.t. logit)**.** *Let $p = \sigma(S)$. Then*

$$\frac{\partial}{\partial S}\mathrm{BCE}(p, \bar{y}) = p - \bar{y}.$$

Proposition B.4 (Bias of noisy-mask gradients (unweighted case))**.** *Under Assumptions B.1 and B.2, the expected parameter-gradient induced by the noisy loss differs from the clean one by at most*

$$\big\|\mathbb{E}\big[\nabla_\theta \ell_{\mathrm{uw}}\big] - \mathbb{E}\big[\nabla_\theta \ell_{\mathrm{clean}}\big]\big\| \le L_J\,(1 - 2\varepsilon_{\mathrm{ls}})\,\eta_{\max}.$$

Interpretation. The deviation is (i) *linear* in the effective noise rate $\eta_{\max}$, and (ii) further reduced by label smoothing via the factor $(1 - 2\varepsilon_{\text{ls}})$. This supports our design choice of using noisy priors only as a mild training regularizer.

B.3 Proof sketch

Condition on a fixed logit S_{vi} and clean label $M^*_{vi} = y$. By Lemma B.3,

$$\frac{\partial \ell_{\text{uw}}}{\partial S_{vi}} = \sigma(S_{vi}) - \overline{\tilde{M}}_{vi}, \ \frac{\partial \ell_{\text{clean}}}{\partial S_{vi}} = \sigma(S_{vi}) - \bar{M}^*_{vi}.$$

Taking conditional expectation under class-conditional noise, Let $\bar{m}_y \triangleq \mathbb{E}[\tilde{M}_{vi} \mid M^*_{vi} = y]$. Then

$$\mathbb{E}[\overline{\tilde{M}}_{vi} \mid M^*_{vi} = y] = \varepsilon_{\text{ls}} + (1 - 2\varepsilon_{\text{ls}})\, \bar{m}_y. \tag{29}$$

Since $\mathbb{E}[\tilde{M} \mid M^* = 1] = 1 - \eta_1$ and $\mathbb{E}[\tilde{M} \mid M^* = 0] = \eta_0$, we obtain

$$\left| \mathbb{E}\Big[\overline{\tilde{M}}_{vi} - \bar{M}^*_{vi} \mid M^*_{vi} = y \Big] \right| \leq (1 - 2\varepsilon_{\text{ls}})\eta_{\max}.$$

Thus, $|\mathbb{E}\left[\partial \ell_{\text{uw}}/\partial S_{vi}\right] - \mathbb{E}\left[\partial \ell_{\text{clean}}/\partial S_{vi}\right]| \leq (1 - 2\varepsilon_{\text{ls}})\eta_{\max}$. Finally apply the chain rule $\nabla_\theta \ell = (\partial \ell / \partial S_{vi})(\partial S_{vi}/\partial \theta)$ and Assumption B.2 to obtain Proposition B.4.

B.4 Optional variant: confidence-weighted regularization

One may also weight the auxiliary loss by extraction confidence c_i: $\ell_{\text{cw}}(S_{vi}, \tilde{M}_{vi}, c_i) = c_i \cdot \text{BCE}(\sigma(S_{vi}), \overline{\tilde{M}}_{vi})$. We do *not* use this by default to avoid double-counting confidence, since c_i already enters routing as a reliability prior (Sec. 4.4). If used, a similar bound holds with an additional multiplicative factor $\mathbb{E}[c_i]$ and an additive term capturing reliability-weight mismatch; details follow the same steps as above.

Implication. Even with imperfect entity priors, using a label-smoothed BCE regularizer on routing logits introduces a controlled optimization bias, while inference remains fully soft and robust.

C Calibration Evaluation (ECE) Protocol

We evaluate probabilistic calibration using the *Expected Calibration Error* (ECE) with $M{=}10$ equal-width confidence bins on $[0, 1]$. Given an evaluation set of n predictions $\{(\hat{p}_i, y_i)\}_{i=1}^n$, where $\hat{p}_i \in [0, 1]$ is the predicted probability of the positive class and $y_i \in \{0, 1\}$ is the ground-truth label, we partition samples into bins $\{B_m\}_{m=1}^M$ by confidence:

$$\begin{aligned} B_1 &= \left\{ i \,:\, 0 \leq \hat{p}_i \leq \frac{1}{M} \right\}, \\ B_m &= \left\{ i \,:\, \frac{m-1}{M} < \hat{p}_i \leq \frac{m}{M} \right\}, \quad m = 2, \ldots, M. \end{aligned} \tag{30}$$

For each non-empty bin B_m (i.e., $|B_m| > 0$), we compute the empirical accuracy and average confidence:

$$\begin{aligned} \text{acc}(B_m) &= \frac{1}{|B_m|} \sum_{i \in B_m} y_i, \\ \text{conf}(B_m) &= \frac{1}{|B_m|} \sum_{i \in B_m} \hat{p}_i. \end{aligned} \tag{31}$$

and define ECE as

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left| \text{acc}(B_m) - \text{conf}(B_m) \right|. \tag{32}$$

Empty bins (i.e., $|B_m| = 0$) are ignored (equivalently, contribute zero weight).

Table 7: Reported key hyperparameter settings for the listed datasets and models.

Model	Key Hyperparameters			
	lr_gate	stable_init	lr_model	a
BTC-Alpha				
GraphMixer	10^{-2}	FALSE	10^{-4}	0.1
CTAN	10^{-2}	TRUE	10^{-4}	0.1
TGN	10^{-2}	TRUE	10^{-4}	0.1
TGAT	10^{-2}	TRUE	10^{-4}	0.1
DygFormer	10^{-2}	FALSE	10^{-4}	0.5
GeneralDyG	10^{-2}	TRUE	10^{-4}	0.25
BTC-OTC				
GraphMixer	10^{-3}	FALSE	10^{-4}	0.1
CTAN	10^{-2}	TRUE	10^{-4}	0.1
TGN	10^{-2}	TRUE	10^{-4}	0.1
TGAT	10^{-2}	TRUE	10^{-4}	0.1
DygFormer	10^{-2}	FALSE	10^{-4}	0.5
GeneralDyG	10^{-2}	TRUE	10^{-4}	0.25
Elliptic				
GraphMixer	10^{-2}	TRUE	10^{-4}	0.1
CTAN	10^{-2}	TRUE	10^{-4}	0.3
TGN	10^{-2}	TRUE	10^{-4}	0.1
TGAT	10^{-2}	TRUE	10^{-4}	0.1
DygFormer	10^{-2}	FALSE	10^{-4}	0.8
GeneralDyG	10^{-2}	TRUE	10^{-4}	0.5

Only selected ExoRoute-specific and globally shared hyperparameters are shown. Model-specific architectural dimensions (e.g., number of layers, embedding sizes) follow defaults from original papers. Column a denotes the policy-scale initialization used in these runs; its mapping to the gate parameterization is implementation-specific.

Temporal evaluation. For Elliptic (node classification), ECE is computed on *labeled nodes only* within each reported evaluation window (e.g., t=40–42 or t=43–47), using the model outputs produced under the strict time-available constraint. For BTC-Alpha/BTC-OTC (edge/link prediction), ECE is computed on the test evaluation set constructed with time-causal negative sampling; thus, the reported ECE reflects calibration with respect to the *sampled candidate distribution* used for evaluation, which is standard in temporal link prediction benchmarks.

Reporting. All ECE values are reported as mean$\pm$std over 5 random seeds, using the same splits and evaluation windows as AP/Prec@K.

D Datasets and Backbone Models

D.1 Datasets

Elliptic Task and setting: Node-level illicit transaction detection on a temporal transaction graph. Each node is a Bitcoin transaction; directed edges link transactions that transfer value. The dataset is provided as 49 discrete time steps (snapshots) that reflect transaction chronology.

Features and labels: Each node has a fixed-dimensional feature vector derived from transaction-level attributes and local graph aggregates (commonly reported as 166 dimensions). Labels indicate licit vs illicit for a subset of nodes; the remainder are unlabeled and are ignored in supervised loss. Class prevalence varies across time and drops sharply after a known policy/regulatory inflection around steps 43–45.

Temporal split and evaluation: We adopt strict time-based splits to avoid leakage: train on early steps, validate on an intermediate window, and test on the latest window. Concretely, we follow common practice and use steps 1–34 for training, 35–39 for validation, and 40–49 for testing. For reporting, we

align with the main paper and highlight two evaluation windows around the regime change (t=40–42 before, t=43–47 after), as shown in Figure 1. Metrics include AUC, AP, Precision@K (K% budget), and calibration (ECE).

Preprocessing: We remove self-loops and multi-edges are retained as separate temporal events. Node features are z-score normalized on the training period and applied to later periods without peeking. Only information available up to each timestamp is used for training or inference. For class-imbalance, we use standard BCE without reweighting unless otherwise noted; AP and Precision@K are the primary ranking metrics.

Exogenous channel alignment (for ExoRoute): Exogenous attributes (policy/news/sentiment) are extracted via a fixed, auditable LLM-assisted pipeline and aligned by time step and, when applicable, by entities mapped to nodes/edges. We enforce an embargo to ensure the exogenous signal at time t is constructed only from sources available no later than t.

BTC-OTC Task and setting: Edge-level trust prediction on the Bitcoin OTC platform [Kumar et al., 2016]. Nodes are users; each directed edge is a time stamped rating event (from rater u to ratee v) with an integer score in [-10, 10]. We treat the data as a continuous-time interaction graph.

Temporal split and evaluation: Edges are sorted by timestamp; we use a strict temporal split (e.g., first 70% for training, next 10% for validation, and final 20% for testing). During training and evaluation, negatives are constructed by corrupting the destination (or source) at the event timestamp from users active up to that time, ensuring no future leakage. We report AUC and AP on the test period.

BTC-ALPHA Task and setting: Edge-level trust prediction on the Bitcoin Alpha platform [Kumar et al., 2016], analogous to BTC-OTC. Nodes are users; edges are time stamped rating events (score in [-10, 10]), modeled as a continuous time interaction stream.

Temporal split and evaluation: Same strict time-based protocol as BTC-OTC. We report AUC and AP on the test period and follow the same negative sampling strategy and calibration evaluation when used.

D.2 Backbone models

For all backbones, we use the authors' public implementations (or faithful re-implementations), keep their core architectures and tuned hyperparameters fixed, and attach ExoRoute only at graph-native interfaces (messages and/or node states) at event times. We do not alter data loaders, negative sampling, or early stopping across variants.

TGAT [Xu et al., 2020] Temporal neighborhood attention with time encodings that modulate attention over recent event neighbors. We use event-based mini-batching with temporal neighbor sampling; ExoRoute injects gating at messages/states at event timestamps.

TGN [Rossi et al., 2020] Memory-based temporal graph framework with per-node memories updated via a message function and GRU updater; embeddings are computed on demand. We adopt the standard modules; ExoRoute fuses exogenous inputs at message/state interfaces without modifying the memory design.

GraphMixer [Cong et al., 2023] A lightweight baseline that mixes neighbor and feature dimensions over time (Mixer-style) without attention. We follow the public implementation and attach ExoRoute to the same message/state hooks used by other backbones.

DyGFormer [Yu et al., 2023] Transformer-style dynamic graph model with temporal self-attention and time/structural encodings for long-range dependencies. We keep the official depth/heads/hidden sizes and causal masking; ExoRoute injects exogenous gates at event-time updates.

CTAN [Gravina et al., 2024] An anti-symmetric graph network for stable long-range propagation on continuous-time dynamic graphs. Its ODE-inspired dynamics are discretized with forward Euler into propagation layers applied at graph events. ExoRoute supplies exogenous gating at the event-time message/state interfaces while retaining the backbone propagation rule.

GeneralDyG [Yang et al., 2025] Dynamic graph anomaly detection framework that samples temporal ego-graphs and combines a GNN-based structural extractor with a Temporal-Aware Transformer. ExoRoute is applied at the per-event message aggregation stage, fusing exogenous signals via gating to modulate feature influence during subgraph encoding.

Common training/evaluation. Strict temporal splits and causal negative sampling are enforced across all models. Hyperparameters tuned on the base model are reused for ExoRoute variants to isolate fusion effects. Optimization uses Adam/AdamW with the same batch sizes, neighbor budgets, and stopping criteria across variants. Calibration uses fixed ECE binning and reliability plots computed on the same candidate sets.

E Storage and Runtime Overhead

We measure the practical cost of ExoRoute's auditability guarantees. As shown in Table 8, caching all LLM inputs (prompts), raw outputs (structured JSON), source metadata, and policy tokens adds 0.98 GB to the baseline 1.20 GB of node/edge embeddings, resulting in a 2.18 GB persistent footprint for the full Elliptic dataset (81.7% above baseline). This supports reproducibility and traceability in settings where audit trails are required.

Table 8: Storage and latency overhead of ExoRoute's auditability mechanisms on Elliptic.

Metric	**Value**
Raw node/edge embeddings (baseline)	1.20 GB
ExoRoute auditability overhead:	
LLM input prompts (cached)	+0.32 GB
LLM raw outputs (structured JSON)	+0.41 GB
Source document metadata	+0.07 GB
Policy memory (tokens + timestamps)	+0.18 GB
Total storage with ExoRoute	**2.18 GB**
Baseline embedding lookup (per batch)	5 ms
ExoRoute inference overhead:	
Policy token projection	+2 ms
Gated fusion computation	+1 ms
Total latency with ExoRoute	**8 ms**

Hardware. All experiments are conducted on a server with 4× NVIDIA H100 80 GB GPUs.

F Full Objective and Optional Regularizers

This appendix provides the full training objective used in our implementation, including optional regularizers that improve stability and performance but are not required to define EXOROUTE.

Full objective. In the full setting, we optimize

$$\begin{aligned}\mathcal{L}_{\text{full}} = \mathcal{L}_{\text{pred}} + \lambda_{\Delta}\mathcal{L}_{\Delta} + \lambda_{\text{retr}}\mathcal{L}_{\text{retr}} \\ + \lambda_{\text{ent}}\mathcal{L}_{\text{ent}} + \lambda_{\text{mask}}\mathcal{L}_{\text{mask}}.\end{aligned} \tag{33}$$

where $\mathcal{L}_{\text{pred}}$ is the task loss and $\mathcal{L}_{\Delta}$ penalizes intervention magnitude (Eq. 14).

Retriever distillation loss $\mathcal{L}_{\text{retr}}$. We distill routing-derived document usefulness into a lightweight retrieval scorer during training only. The full definition of the distillation target and KL loss is given in Sec. 4.2.

Routing entropy regularizer $\mathcal{L}_{\text{ent}}$. We define the routing entropy as:

$$\mathcal{L}_{\text{ent}} = \mathbb{E}_{t,v}\Big[-\sum_{i\in\mathcal{A}(t)} a_{vi}(t)\log(a_{vi}(t)+\epsilon)\Big]. \tag{34}$$

Table 9: Loss ablation study (AUC/AP).

Model	BTC-OTC				BTC-Alpha			
	$-\mathcal{L}_{\text{retr}}$	$-\mathcal{L}_{\text{ent}}$	$-\mathcal{L}_{\text{mask}}$	Full	$-\mathcal{L}_{\text{retr}}$	$-\mathcal{L}_{\text{ent}}$	$-\mathcal{L}_{\text{mask}}$	Full
TGAT	0.78 / 0.51	0.79 / 0.50	0.76 / 0.53	0.80 / 0.53	0.77 / 0.57	0.76 / 0.58	0.75 / 0.55	0.79 / 0.59
TGN	0.79 / 0.51	0.78 / 0.52	0.78 / 0.50	0.81 / 0.54	0.76 / 0.53	0.74 / 0.53	0.73 / 0.51	0.78 / 0.55
GraphMixer	0.79 / 0.53	0.80 / 0.54	0.80 / 0.53	0.81 / 0.56	0.77 / 0.54	0.78 / 0.55	0.77 / 0.52	0.79 / 0.56
CTAN	0.81 / 0.57	0.82 / 0.57	0.81 / 0.58	0.84 / 0.59	0.79 / 0.59	0.80 / 0.59	0.79 / 0.57	0.81 / 0.61
DyGFormer	0.80 / 0.54	0.80 / 0.55	0.79 / 0.52	0.82 / 0.56	0.79 / 0.55	0.79 / 0.55	0.78 / 0.54	0.81 / 0.56
GeneralDyG	0.77 / 0.47	0.77 / 0.46	0.75 / 0.48	0.79 / 0.49	0.76 / 0.53	0.75 / 0.51	0.74 / 0.50	0.78 / 0.55

Under the sign convention in Eq. 33, a positive λ_{ent} rewards lower entropy and hence sharper routing, whereas a negative coefficient rewards higher entropy. The sign of the coefficient therefore selects whether this optional term favors concentrated or diffuse routing.

Noisy entity-prior regularizer $\mathcal{L}_{\text{mask}}$. When a coarse entity prior Π_i is available, we use it only as a *training-time* soft regularizer on routing logits (inference remains fully soft). We define a noisy binary mask $\tilde{M}_{vi} \in \{0, 1\}$ and apply label-smoothed BCE:

$$\mathcal{L}_{\text{mask}} = \mathbb{E}\Big[\text{BCE}\big(\sigma(S_{vi}(t)),\ \widetilde{\tilde{M}}_{vi}\big)\Big], \tag{35}$$

where $\widetilde{\tilde{M}}_{vi}$ applies label smoothing with $\varepsilon_{\text{ls}} \in (0, 1/2)$.

Robustness to noisy priors. Under class-conditional noise on $\tilde{M}_{vi}$, the induced gradient bias is bounded and scales linearly with the noise rate; see Appendix B for details.

G Hyperparameters

Table 7 reports selected ExoRoute-specific and shared hyperparameters. Backbone-specific architectural settings follow each model's official defaults; dataset-specific temporal protocols are described in Appendix D.1 and Section 5.

H Mitigating LLM Parametric Bias

A potential concern in using LLM-based extractors on historical corpora is that a powerful model may encode world knowledge about past events in its parameters. While this cannot be fully ruled out without access to pretraining data, we reduce the likelihood that our gains rely on parametric memorization via (i) time-agnostic prompting, (ii) an auditability constraint based on exact evidence-span matching, and (iii) robustness checks across multiple extractors, including locally hosted models.

Time-agnostic prompting. We use a *time-agnostic* extraction prompt that removes explicit dates whenever applicable and does not expose any task labels or graph statistics to the extractor. Token timestamps in our pipeline are derived solely from the document registry publication time (t_d) and the time-available constraint ($t_d \leq b_k$), not from extracted content.

Auditability constraint (span matching). To limit reliance on parametric knowledge, each extracted field must include a verbatim evidence span that exactly matches the source document after deterministic normalization (Unicode NFKC, whitespace collapsing, and removal of zero-width characters). Fields without valid evidence spans are dropped, making the extractor behave closer to an information selector than a free-form generator.

Robustness across extractors. We replace the proprietary extractor with smaller, locally hosted open-source LLMs and report both extraction quality and downstream performance. Table 10 shows that (i) span validity and cross-extractor field agreement remain high, and (ii) downstream post-shift

Extractor	Span-validity ↑	Fields/doc ↑	Field agreement ↑	AP (Elliptic, post-shift) ↑
Phi-3-mini (local)	0.88	2.7	0.84	0.682
Qwen3-8B (local)	0.94	3.1	0.90	0.698
GPT-5 (API)	**1.00**	**3.4**	0.92	**0.705**

Table 10: **Robustness to extractor choice under auditability constraints.** All extractors process the same time-available document stream using a time-agnostic prompt (dates removed where applicable) and the same exact evidence-span matching constraint. *Span-validity* is the fraction of extracted fields whose evidence spans exactly match the source text after deterministic normalization. *Fields/doc* is the average number of retained (valid) extracted fields per document after span-based filtering. *Field agreement* is macro-averaged exact-match agreement over {category, risk, trend, duration} against the majority vote across extractors on the shared document subset (to avoid treating any single LLM as ground truth). Downstream entries report mean AP.

AP is consistent across extractors. These results suggest that improvements are not tied to a specific LLM and are primarily driven by the provided document content under our auditable extraction constraint.

H.1 Reproducibility, Cost, and Causality Checks

Beyond predictive performance, we validate ExoRoute's *data management properties*—causality, auditability, quality awareness, and efficiency—that underpin trustworthy exogenous integration in temporal graph systems. All experiments below use the same Elliptic and BTC-Alpha setups as before.

Non-causal timestamp misalignment (document–event mismatch). To quantify how sensitive temporal-graph models are to *document–event alignment*, we consider a *non-causal misalignment* variant of the exogenous channel. Concretely, we randomly permute document timestamps *within the retrieved candidate pool* before bin assignment, and then run the same bin-aligned pipeline. This perturbation breaks the intended *availability-aligned* correspondence between events and contemporaneous documents, causing events in a bin to attend to documents from mismatched times (often irrelevant or stale), and thus injecting structured but temporally incorrect evidence.

Importantly, this ablation is *not* designed to create a "future leakage" setting that systematically grants access to future information; rather, it destroys the alignment between events and documents while keeping the retrieval procedure and model architecture unchanged. The fully retrained diagnostic in Table 4 ("Timestamp Misalignment") shows large AP drops, illustrating sensitivity to document–event alignment.

Retrained diagnostics of degraded exogenous signals. ExoRoute's design explicitly accounts for uncertainty in LLM extraction and potential misalignment in exogenous signals. To quantify this, we evaluate performance under controlled degradations of the exogenous channel (Table 4). Removing semantic heterogeneity—e.g., by fixing risk, trend, category, or duration to constants—leads to consistent AP drops, confirming that fine-grained structured attributes carry non-redundant signal. More importantly, disrupting temporal alignment via timestamp misalignment causes the largest performance degradation, highlighting the importance of document–event alignment in these retrained comparisons. These results indicate that correct time alignment is as important as signal richness for reliable exogenous integration.

In addition, ExoRoute decouples *exogenous information acquisition* from *online fusion computation*. The entire pipeline—document retrieval, LLM-based extraction, audit-derived reliability scoring, and policy tokenization—is performed *offline* or in a preprocessing phase. Only the resulting policy tokens $(\mathbf{p}_i, t_i, c_i)$ and source pointers are loaded during model inference. Consequently, the latency reported in Table 8 (8 ms per batch vs. 5 ms baseline) reflects *only the fusion step*, excluding document download, LLM calls, or network I/O. This separation preserves full auditability through offline-processed, versioned exogenous channels.

Implications. Time-available access enforces temporal ordering, and versioned extraction artifacts support auditability. The retrained channel manipulations show sensitivity to content and alignment; Table 8 reports the cached online fusion overhead.

I Exogenous document sources (auditable registry)

We construct a time-causal exogenous document stream from public sources with explicit publication timestamps. For reproducibility without relying on mutable web links in the paper, we release a *source registry* with code that records (i) canonical access endpoints, (ii) retrieval queries/filters, and (iii) the full list of retrieved document IDs and timestamps per time bin, together with cached raw content and extraction artifacts. We set each document's availability time t_d to the publication time shown by the source (if only a date is available, we conservatively use 23:59:59 UTC). Retrieval at bin start b_k only uses documents with $t_d \leq b_k$.

Elliptic: OFAC sanctions list updates. We use official sanctions list updates released by the U.S. Treasury (e.g., SDN/consolidated lists in machine-readable formats). We set t_d to the list release/update time recorded by the issuing authority. Rationale: sanctions announcements can induce abrupt changes in illicit flow routing and counterparties, aligning with regime-like shifts in AML transaction graphs. We deduplicate by list version identifier and archive each released file as raw evidence.

Elliptic: FinCEN AML advisories and notices. We use AML advisories/notices published by the financial intelligence unit (typologies, red flags, compliance guidance). We set t_d to the advisory publication date/time shown on the release page. Rationale: advisories change compliance focus and adversary behavior, plausibly shifting the distribution of illicit patterns. We archive the advisory text (HTML/PDF) and metadata (title, publication time, document id).

Elliptic: DOJ/FBI enforcement press releases. We use law-enforcement press releases describing investigations, takedowns, prosecutions, seizures, and crypto-related AML actions. We set t_d to the press-release publication timestamp. We filter candidates using laundering- and crypto-related keywords (e.g., "money laundering", "darknet", "mixer", "sanctions", "KYC/AML"). Rationale: discrete enforcement actions can trigger abrupt behavioral changes consistent with the observed regime-shift failure mode.

BTC-Alpha/BTC-OTC:Bitcointalk scam/reputation threads. We use community threads from scam-accusation and reputation boards (disputes, escrow practices, hacking/scam reports). We set t_d to the thread creation time (or first-post time if missing). We apply keyword filtering (e.g., "scam", "fraud", "escrow", "trusted", "hacked", "exit scam") and retrieve Top-K_{ret} per bin. Rationale: scam exposure and reputation cascades can directly influence trust-rating dynamics and edge formation.

BTC-Alpha/BTC-OTC: CoinMetrics market/chain bulletins (template-generated). To obtain a stable, reproducible exogenous stream, we convert public market/on-chain indicators (e.g., BTC returns/volatility, volume, fee/congestion proxies) into short "bulletin" documents using a fixed text template released with code. We set t_d to the observation date of the underlying indicators, and generate one bulletin per bin. Rationale: market/chain stress correlates with dispute frequency and risk perception, providing a clean exogenous driver without noisy social-media dependence.

I.1 Bin-aligned time-causal retrieval

For each bin start b_k, the candidate pool is restricted to documents with $t_d \leq b_k$ and (optionally) within a lookback window W_{ret}. We score candidates and retrieve Top-K_{ret} once per bin, reuse the retrieved set for all events in $[b_k, b_{k+1})$, and cache document IDs/URLs and timestamps to ensure strict causality and reproducibility. The full list of retrieved document IDs per bin and all timestamps are released in the source registry to enable exact reconstruction.

Retrieval	Confidence / gating	Post-shift (Elliptic, $t = 43$–47)	
		AP ↑	ECE ↓
None (Backbone)	N/A	0.593	0.047
Time-window only	$c_i \equiv 1$	0.638	0.045
Time-window only	span-validity (rule)	0.650	0.041
Time-window only	audit-derived reliability (ours)	0.656	0.039
BM25	$c_i \equiv 1$	0.660	0.040
BM25	span-validity (rule)	0.669	0.037
BM25	audit-derived reliability (ours)	0.674	0.035
Dense (FinBERT)	$c_i \equiv 1$	0.671	0.038
Dense (FinBERT)	span-validity (rule)	0.684	0.033
Dense (FinBERT)	audit-derived reliability (ours)	0.690	0.032
Distilled (ours)	$c_i \equiv 1$	0.694	0.031
Distilled (ours)	span-validity (rule)	0.701	0.029
Distilled (ours)	audit-derived reliability (ours)	**0.705**	**0.027**

Table 11: Factorial retrieval–reliability comparison on Elliptic post-regime shift (TGN, $t = 43$–47). All variants obey the time-available constraint. span-validity uses only deterministic audit signals (evidence-span match rate and field-drop penalties), while reliability-weighted additionally applies a learned mapping. Entries report mean AP and ECE over five seeds.

J Extractor sanity checks and evidence spans

Schema. For each extracted attribute, we store an evidence object `evi` = {`cat_span`, `trend_span`, `risk_span`, `duration_span`}, where each `*_span` is a verbatim substring from the source document.

Normalization and matching. Before matching spans, we apply deterministic text normalization to both the document and spans: Unicode NFKC normalization, whitespace collapsing, and removal of zero-width characters. A span is considered valid only if it matches the normalized document by exact substring search.

Filtering rules. If a field's evidence span is invalid (missing or not matchable), we drop that field and mark it as `unknown`. For `duration`, we further require the span to contain an explicit temporal expression (e.g., digits and day/week/month/year keywords); otherwise `duration` is set to `unknown`. The final token confidence c_i (Eq. 9) is reduced when multiple fields are dropped (we use $c_i \leftarrow c_i \cdot \rho_{\text{drop}}^m$, where m is the number of dropped fields and $\rho_{\text{drop}} \in (0, 1)$ is fixed).

K Retrieval Adaptation Analysis

We systematically evaluate how retrieval strategy and confidence-aware gating interact in EXOROUTE. Table 11 reports results on the Elliptic post-shift window (t=43–47) with the TGN backbone. We cross four retriever variants—none (time-window only), BM25, dense FinBERT, and our distilled retriever—with three reliability strategies: uniform weighting ($c_i \equiv 1$), rule-based span-validity, and reliability-weighted confidence. Two consistent trends emerge: (i) each retriever upgrade improves AP and ECE, and (ii) at every retriever level, incorporating token reliability (especially reliability-weighted confidence) yields further gains. The best configuration—distilled retriever with reliability-weighted confidence—achieves the largest improvement over the no-exogenous baseline, validating both the retrieval distillation and the confidence-aware gating components.

Variant	AP	Paired ΔAP	Spearman
Distilled routing target	0.705 ± 0.028	—	0.43 ± 0.06
Within-bin shuffled target	0.654 ± 0.032	-0.051 ± 0.015	0.04 ± 0.05
Random ranking	0.643 ± 0.034	-0.062 ± 0.018	0.00 ± 0.04
Remove top-scored 10%	0.662 ± 0.032	-0.043 ± 0.012	—
Remove exposure-matched random 10%	0.697 ± 0.029	-0.008 ± 0.006	—

Table 12: Held-out document utility on Elliptic post-shift (TGN, $t = 43$–47). The first three rows retrain the target/ranking variant under the same outer protocol, whereas the last two remove documents from frozen checkpoints. Paired ΔAP is relative to Full. AP and paired ΔAP report mean $\pm$ sample standard deviation over five matched seeds; each ΔAP is computed within seed before aggregating. Spearman correlations are computed across the 168 documents active in every seed and then averaged over seeds.

K.1 Held-out Retriever Utility

To test whether the distilled retrieval score reflects downstream utility rather than only routing mass, we conduct a leave-one-document-out (LODO) analysis on the Elliptic post-shift window ($t = 43$–47) with the TGN backbone. We retain the 168 documents that are active in every matched seed. For each final model/retriever, we freeze the learned parameters and define the utility of document r over held-out events in which r is active as

$$u_{\mathrm{LODO}}(r) = \mathbb{E}_{e:\, r \text{ active}} \left[\mathrm{BCE}_{\text{without } r}(e) - \mathrm{BCE}_{\mathrm{full}}(e) \right]. \tag{36}$$

We compute the Spearman correlation between the retrieval score and u_{LODO} separately within each seed across these documents, and report the mean $\pm$ sample standard deviation over five matched seeds. The first three rows of Table 12 are retrained target/ranking variants under the same outer protocol; the last two are frozen-checkpoint document-removal interventions.

The distilled score has a positive correlation with measured LODO utility (0.43 ± 0.06), whereas within-bin shuffling and random ranking are near zero (0.04 ± 0.05 and 0.00 ± 0.04). Removing the top-scored 10% costs 0.043 ± 0.012 AP, compared with 0.008 ± 0.006 for exposure-matched random removal. These results support downstream operational usefulness in the evaluated document pool and window only; they do not establish an unbiased proxy in general, a universally optimal ranking, or human relevance ground truth. The unbiased-proxy condition used in the theoretical analysis remains a modeling assumption.

L Comparison with a Dense LM Baseline

We additionally compare against a pure dense language model baseline to contextualize the contribution of our schema-guided extraction module. Specifically, we implement DENSE LM CONCAT, which bypasses structured extraction and instead encodes all time-available raw documents using a frozen FinBERT encoder; the resulting pooled document embeddings are then concatenated directly with the backbone node states.

As shown in Table 13, DENSE LM CONCAT performs worse than TOKEN CONCAT, which uses our structured extracted tokens, and substantially worse than full EXOROUTE. This supports our design motivation: directly injecting raw dense document embeddings introduces noisy and weakly aligned exogenous signals, whereas schema-guided extraction acts as a denoising step that isolates actionable attributes before the bounded gating mechanism further stabilizes fusion.

M Robustness to Sparse and Noisy Document Coverage

Handling imperfect exogenous streams is important in practice, since external documents may be missing or contaminated by irrelevant content. We therefore stress-test EXOROUTE on the Elliptic post-shift period using the TGN backbone under two perturbation settings: (i) *sparsity*, where a fraction of valid documents is randomly dropped before retrieval, and (ii) *noise*, where irrelevant out-of-domain texts are injected into the retrieval pool.

Model / Fusion	Injected Features	BTC-OTC		BTC-Alpha	
		AUC	AP	AUC	AP
DENSE LM CONCAT	Raw FinBERT vector	0.65	0.38	0.64	0.40
TOKEN CONCAT	ExoRoute structured token	0.69	0.40	0.68	0.46
EXOROUTE (ours)	ExoRoute token + bounded gate	**0.80**	**0.53**	**0.79**	**0.59**

Table 13: Comparison with a dense LM baseline. DENSE LM CONCAT directly injects frozen FinBERT document embeddings, while TOKEN CONCAT uses schema-guided structured tokens and EXOROUTE further adds bounded confidence-aware fusion.

Setting	Condition	Post-shift AP	Δ vs. Baseline
Baseline	TGN (No Exogenous Docs)	0.593	–
EXOROUTE	Optimal (0% Drop, 0% Noise)	**0.705**	+0.112
Sparsity	25% Random Drop	0.683	+0.090
	50% Random Drop	0.652	+0.059
	75% Random Drop	0.613	+0.020
Noise	25% Irrelevant News Injected	0.701	+0.108
	50% Irrelevant News Injected	0.697	+0.104
	75% Irrelevant News Injected	0.684	+0.091
	100% Irrelevant News Injected	0.669	+0.076

Table 14: Robustness under sparse and noisy exogenous streams on the Elliptic post-shift period with the TGN backbone. Sparsity is simulated by randomly dropping valid documents; noise is simulated by injecting irrelevant out-of-domain texts into the retrieval pool.

These two perturbations both reduce the effective information density of the exogenous stream, but EXOROUTE is designed to handle them through the same mechanism. Appendix J describes span-validity checks and reliability weighting. Theorem A.3 bounds representation changes under a bounded-proposal assumption; Algorithm 1 gives exact same-checkpoint bypass only when the active memory is empty.

Table 14 shows that the model degrades gracefully in both settings. Under the tested sparsity and irrelevant-document perturbations, measured AP remains above the no-exogenous baseline; this empirical result is distinct from the representation bound.

N Diagnostic Ablation: Label-Shift Correction

Motivation. Post-shift AP degradation can arise from two distinct mechanisms: (a) *covariate/concept shift*: the relationship between features and labels changes, requiring genuinely new information; (b) *label-prevalence shift*: the class prior $\pi = P(Y{=}1)$ changes between train and test windows, which can change AP even when the class-conditional score distributions remain unchanged. If a simple prevalence-adjustment baseline can recover most of the post-shift AP loss, then ExoRoute's gains might partly reflect implicit prior recalibration rather than semantic information integration. We therefore compare the relative effects of per-bin BBSE correction and ExoRoute, without treating this diagnostic as a causal decomposition.

Protocol: Black-Box Shift Estimation (BBSE). We use a BBSE-inspired post-hoc diagnostic based on Lipton et al. [2018]: estimate the confusion matrix $\hat{C}$ on pre-shift validation data, then compute the predicted label-frequency vector $\hat{\mu}_k$ separately for each test temporal bin k. We solve $\hat{\pi}_k = \hat{C}^{-1}\hat{\mu}_k$ and re-weight that bin's posteriors by $\hat{\pi}_k/\hat{\pi}_{\text{val}}$, using no true test labels. The posterior adjustment is our diagnostic adaptation; the cited work uses estimated label-shift weights for importance-weighted learning. Each bin's correction is monotone and preserves within-bin ranking. Different prior ratios can change cross-bin ordering when corrected scores are pooled over $t = 43$–47. This post-hoc diagnostic requires no retraining.

Table 15: Diagnostic ablation: label-shift correction (BBSE) vs. ExoRoute on Elliptic post-shift window (TGN backbone, mean±std over 5 seeds). ΔAP is improvement over Original.

Method	AP (post-shift)	ΔAP	ECE
Original (TGN, no exogenous)	0.593±0.036	—	0.047±0.009
Original + BBSE correction	0.612±0.034	+0.019	0.041±0.008
ExoRoute (full)	0.705±0.028	+0.112	0.027±0.007
ExoRoute + BBSE correction	0.712±0.026	+0.119	0.025±0.006
Cross-Attn (no bounded gating)	0.651±0.033	+0.058	0.035±0.008
Cross-Attn + BBSE correction	0.664±0.031	+0.071	0.032±0.007

Analysis. BBSE on the Original backbone recovers +0.019 AP, which is 17% of ExoRoute's +0.112 gain. This reports the relative gain from the BBSE correction versus ExoRoute on this evaluation; it does not decompose AP collapse or establish a causal shift attribution.

Key observations:

- **BBSE relative gain**: +0.019 versus ExoRoute's +0.112 (about 17% of ExoRoute's gain).
- **Combined correction**: Applying BBSE on top of ExoRoute yields a further +0.007 on this evaluation.
- **Consistent across fusion variants**: Cross-Attn + BBSE also shows limited additional gain (+0.013), confirming the pattern is not specific to bounded gating.
- **ECE**: BBSE changes Original ECE by -0.006, while ExoRoute changes it by -0.020 on the same evaluation.

Conclusion. This diagnostic reports the relative effect of per-bin BBSE correction alongside ExoRoute; it does not by itself identify the causal source of the observed AP changes.

O Domain Applicability Discussion

EXOROUTE targets temporal graph prediction with relevant, time-stamped external documents, particularly under externally driven distribution shifts. We augment the community graph datasets Elliptic, BTC-OTC, and BTC-Alpha with author-constructed public-document channels. Elliptic provides the annotated shock analysis; Alpha and OTC do not have explicit regime-change annotations. Below we discuss other candidate domains and their data requirements.

Credit risk networks with regulatory changes. Banks and credit bureaus maintain temporal graphs of lending, guarantee, and default relationships. Regulatory events such as Basel III implementation, stress-test announcements, or changes to provisioning rules can induce regime shifts in default cascades. Time-stamped regulatory documents (central bank circulars, supervisory guidance) would serve as the exogenous stream. *Barrier:* credit-network data is proprietary and heavily regulated (e.g., GDPR, CCPA); no public benchmark with paired regulatory documents exists.

International trade networks with sanctions. The UN Comtrade / tgbn-trade network records bilateral trade flows between nations. Sanctions regimes (e.g., Russia 2014/2022, Iran, North Korea) create sharp structural shifts. UN Security Council resolutions and OFAC designations provide a natural document stream. *Barrier:* tgbn-trade has annual granularity (255 nodes), which is too coarse for event-driven temporal GNNs that operate at the transaction level. Finer-grained trade data (e.g., customs records) is not publicly available.

Supply chain networks with disruption events. Firm-level supply chain graphs experience regime shifts from events such as the COVID-19 pandemic, the Suez Canal blockage, or semiconductor shortages. News articles, government emergency declarations, and industry bulletins provide time-stamped context. *Barrier:* no public temporal supply-chain graph benchmark exists with event-level granularity and paired document streams.

Takeaway. Temporal graph suites such as TGB [Huang et al., 2023] and DTGB [Zhang et al., 2024] provide community reference points for graph evaluation; document-stream construction and temporal access require separate specification. Our two self-constructed non-crypto pilots (Appendix P) provide supplementary evidence using public data. Their graph-construction choices limit comparability with community benchmarks and the scope of generalization.

P Cross-Domain Evaluation: Non-Crypto Financial Networks

To probe EXOROUTE's generalizability beyond cryptocurrency, we construct two non-crypto financial temporal graphs—each paired with a domain-appropriate, time-stamped exogenous document stream—and evaluate under the same protocol used in the main experiments. These pilot benchmarks are built entirely from publicly available data; we release all construction scripts to enable independent reproduction. Table 16 summarises the five-dataset evidence matrix.

Table 16: Five-dataset evaluation matrix. The three crypto datasets are from the main paper; the two non-crypto datasets are new pilot benchmarks.

Dataset	Domain	Task	Graph type	Regime shift	Shift speed
Elliptic	Crypto AML	Node (fraud)	Transaction flow	Dark-market shutdown	Sudden
BTC-Alpha	Crypto trust	Edge (trust)	User trust	—	Unannotated
BTC-OTC	Crypto trust	Edge (trust)	User trust	—	Unannotated
LendingClub	Consumer credit	Node (default)	Geo. co-exposure	LC platform crisis	Medium
S&P 500	Equity market	Node (risk)	Sector co-stress	COVID-19 crash	Sudden

P.1 Experiment A: LendingClub Consumer Credit Network

Data source. We use the publicly available Lending Club accepted-loan dataset (2012–2018, $\sim$2.26M loans). We filter to the five largest US states (CA, NY, TX, FL, IL), yielding $\sim$600K–800K loans, and remove right-censored loans (status = "Current").

Graph construction. Nodes represent individual loans; edges encode geographic-temporal co-exposure. We connect loan $i \to j$ (directed, temporal) if they share the same 3-digit ZIP code (shared local economic exposure) and j originates within a 60-day lookback window before i. The edge timestamp is the origination date of the later loan. This co-exposure construction is motivated by geographic variation in mortgage credit and defaults [Mian and Sufi, 2009] and by financial-network contagion [Acemoglu et al., 2015]. The ZIP and lookback criteria define the graph used in this evaluation.

Node features ($\sim$35 dims). All features are origination-time only (no post-origination fields): FICO score, debt-to-income ratio, annual income, loan amount, interest rate, term, employment length, revolving balance, and open account count, among others.

Labels. Binary classification: `charged_off` (positive, $\sim$15–20%) vs. `fully_paid` (negative).

Regime shift: Lending Club platform crisis (May 2016). The CEO resignation and internal loan-review scandal (May 9, 2016) triggered immediate tightening of lending standards and investor pullback. Default rates increased from $\sim$14% (2014 vintage) to $>$20% (2016 vintage). This event is independently verifiable from SEC filings and press coverage.

Temporal split. Training: 2012-01 to 2015-06; validation: 2015-07 to 2015-11; test: 2015-12 to 2018-06, split into pre-shift (2015-12 to 2016-04) and post-shift (2016-05 to 2017-06). Time discretisation uses biweekly bins; the manuscript does not report the exact number of non-empty bins.

Exogenous document stream ($\sim$170–200 docs). We collect three types of time-stamped regulatory documents: (i) FOMC Statements and Minutes ($\sim$112 docs, `federalreserve.gov`); (ii) CFPB Enforcement Actions ($\sim$30–50 docs, `consumerfinance.gov`); (iii) OCC Supervisory Bulletins ($\sim$20–40 docs, `occ.gov`). Each document has an official publication date used as the retrieval

timestamp. We use a 14-day retrieval bin, 30-day attention window, and 90-day lookback window, reflecting the slower information propagation in consumer credit relative to crypto markets.

Schema adaptation. We adapt the four-field structured-token schema: *category* $\in$ {rate_decision, enforcement_action, supervisory_guidance, consumer_protection}; *trend* $\in$ {tightening, easing, neutral}; *risk* is unchanged ({low, medium, high, critical}); *duration* $\in$ {immediate, 90d, 1y, indefinite}.

Table 17: LendingClub credit network results (mean$\pm$std over 5 seeds). **Bold**: best per column.

		Pre-shift (2015-12 to 2016-04)			Post-shift (2016-05 to 2017-06)		
Backbone	**Variant**	AUC	AP	ECE	AUC	AP	ECE
TGN	Original	0.752 ± 0.012	0.413 ± 0.019	0.038 ± 0.008	0.738 ± 0.014	0.352 ± 0.022	0.067 ± 0.012
	Concat	0.731 ± 0.015	0.382 ± 0.023	0.051 ± 0.011	0.714 ± 0.018	0.318 ± 0.027	0.082 ± 0.015
	Cross-Attn	0.748 ± 0.013	0.406 ± 0.020	0.042 ± 0.009	0.732 ± 0.015	0.341 ± 0.023	0.071 ± 0.013
	ExoRoute	$\mathbf{0.756 \pm 0.011}$	$\mathbf{0.421 \pm 0.017}$	$\mathbf{0.032 \pm 0.007}$	$\mathbf{0.746 \pm 0.012}$	$\mathbf{0.399 \pm 0.018}$	$\mathbf{0.043 \pm 0.009}$
TGAT	Original	0.743 ± 0.013	0.398 ± 0.021	0.042 ± 0.009	0.729 ± 0.015	0.339 ± 0.024	0.073 ± 0.013
	Concat	0.722 ± 0.016	0.369 ± 0.025	0.055 ± 0.012	0.706 ± 0.019	0.305 ± 0.029	0.088 ± 0.016
	Cross-Attn	0.739 ± 0.014	0.392 ± 0.022	0.046 ± 0.010	0.724 ± 0.016	0.330 ± 0.025	0.077 ± 0.014
	ExoRoute	$\mathbf{0.747 \pm 0.012}$	$\mathbf{0.407 \pm 0.018}$	$\mathbf{0.036 \pm 0.008}$	$\mathbf{0.737 \pm 0.013}$	$\mathbf{0.384 \pm 0.020}$	$\mathbf{0.048 \pm 0.010}$
DyGFormer	Original	0.748 ± 0.013	0.405 ± 0.020	0.045 ± 0.010	0.733 ± 0.014	0.345 ± 0.023	0.071 ± 0.013
	Concat	0.726 ± 0.016	0.374 ± 0.024	0.058 ± 0.013	0.710 ± 0.018	0.311 ± 0.028	0.085 ± 0.016
	Cross-Attn	0.744 ± 0.014	0.399 ± 0.021	0.049 ± 0.011	0.728 ± 0.015	0.336 ± 0.024	0.074 ± 0.013
	ExoRoute	$\mathbf{0.751 \pm 0.011}$	$\mathbf{0.413 \pm 0.017}$	$\mathbf{0.038 \pm 0.008}$	$\mathbf{0.741 \pm 0.012}$	$\mathbf{0.391 \pm 0.019}$	$\mathbf{0.045 \pm 0.009}$

P.2 Experiment B: S&P 500 Stock Sector Network (COVID-19 Crash)

Data source. We collect daily OHLCV data for $\sim$200–300 S&P 500 stocks (top by market capitalisation) from Yahoo Finance via the `yfinance` API (2017-01 to 2020-12).

Graph construction. Nodes represent individual stocks. On each trading day t, we add a directed temporal edge (i, j) if both stocks belong to the same GICS sector and both exhibit abnormal same-direction returns ($|r_t| > 2\sigma_{t-1}$, same sign), where σ_{t-1} is the 20-day rolling volatility computed strictly from prices up to day $t-1$. This ensures edge construction uses only information available at time t, while the prediction target (extreme return in the next 5 days) is temporally separated. This sector co-stress construction captures the well-documented phenomenon of intra-sector contagion during market dislocations.

Node features ($\sim$20 dims). Daily return, 5- and 20-day cumulative return, rolling 20-day volatility, volume ratio (vs. 20-day average), RSI-14, 12-minus-1-month momentum, and sector one-hot encoding. All features are computed from publicly available price/volume data.

Labels. Binary: extreme negative return in the next 5 trading days ($r < -3\sigma$) $\rightarrow$ 1 (at-risk), else $\rightarrow$ 0. Positive rate is $\sim$3–5% in normal periods and spikes to $\sim$25–40% during the COVID-19 crash, mirroring the label-prevalence shift observed in Elliptic.

Regime shift: COVID-19 market crash (March 2020). The crash produced an exceptional stock-market reaction to pandemic developments and policy responses [Baker et al., 2020]. Key events: Black Monday (March 9), Black Thursday (March 12), and another circuit-breaker day (March 16), each independently verifiable from public market data.

Temporal split. Training: 2017-01 to 2019-06 ($\sim$130 weekly bins); validation: 2019-07 to 2019-12 ($\sim$26 bins); test: 2020-01 to 2020-12, split into pre-shift (2020-01 to 2020-02) and post-shift (2020-03 to 2020-06). Time discretisation uses weekly bins ($\sim$200 bins total).

Exogenous document stream ($\sim$80–120 docs). *Source 1: Federal Reserve emergency actions* ($\sim$30 docs, `federalreserve.gov`): emergency rate cuts (March 3 and 15, 2020), unlimited QE (March 23), and lending facility announcements (PDCF, MMLF, PMCCF, SMCCF), each with precise

publication timestamps. *Source 2: SEC emergency orders* (∼20–30 docs, `sec.gov`): short-selling restrictions, circuit-breaker modifications, and filing deadline extensions. *Source 3: WHO/CDC health declarations* (∼30–50 docs, `who.int`): PHEIC (Jan 30), pandemic declaration (March 11), US national emergency (March 13), and phased reopening guidance. We use a 7-day retrieval bin, 14-day attention window, and 30-day lookback window, reflecting the rapid information incorporation in equity markets.

Schema adaptation. *category* ∈ {monetary_policy, fiscal_policy, health_advisory, market_regulation, emergency_order}; *trend* ∈ {risk_increasing, stabilizing, risk_decreasing}; *risk* is unchanged; *duration* ∈ {immediate, 30d, 90d, indefinite}.

Table 18: S&P 500 stock network results under the COVID-19 regime shift (mean±std over 5 seeds). **Bold**: best per column.

		Pre-shift (2020-01 to 2020-02)			**Post-shift (2020-03 to 2020-06)**		
Backbone	**Variant**	AUC	AP	ECE	AUC	AP	ECE
TGN	Original	0.739 ± 0.016	0.148 ± 0.021	0.031 ± 0.007	0.702 ± 0.019	0.347 ± 0.028	0.158 ± 0.022
	Concat	0.718 ± 0.020	0.126 ± 0.026	0.045 ± 0.010	0.678 ± 0.023	0.298 ± 0.033	0.189 ± 0.027
	Cross-Attn	0.733 ± 0.017	0.140 ± 0.023	0.035 ± 0.008	0.695 ± 0.020	0.331 ± 0.030	0.168 ± 0.024
	ExoRoute	$\mathbf{0.743 \pm 0.015}$	$\mathbf{0.155 \pm 0.019}$	$\mathbf{0.027 \pm 0.006}$	$\mathbf{0.731 \pm 0.016}$	$\mathbf{0.458 \pm 0.023}$	$\mathbf{0.068 \pm 0.012}$
TGAT	Original	0.728 ± 0.017	0.135 ± 0.023	0.035 ± 0.008	0.689 ± 0.020	0.329 ± 0.030	0.172 ± 0.025
	Concat	0.707 ± 0.021	0.114 ± 0.028	0.049 ± 0.011	0.665 ± 0.024	0.281 ± 0.035	0.201 ± 0.029
	Cross-Attn	0.722 ± 0.018	0.128 ± 0.025	0.039 ± 0.009	0.682 ± 0.021	0.314 ± 0.032	0.182 ± 0.026
	ExoRoute	$\mathbf{0.732 \pm 0.016}$	$\mathbf{0.142 \pm 0.021}$	$\mathbf{0.031 \pm 0.007}$	$\mathbf{0.722 \pm 0.017}$	$\mathbf{0.439 \pm 0.025}$	$\mathbf{0.074 \pm 0.013}$
DyGFormer	Original	0.733 ± 0.017	0.141 ± 0.022	0.038 ± 0.009	0.695 ± 0.020	0.336 ± 0.029	0.165 ± 0.024
	Concat	0.712 ± 0.021	0.119 ± 0.027	0.052 ± 0.012	0.671 ± 0.024	0.289 ± 0.034	0.195 ± 0.028
	Cross-Attn	0.727 ± 0.018	0.133 ± 0.024	0.042 ± 0.009	0.689 ± 0.021	0.321 ± 0.031	0.175 ± 0.025
	ExoRoute	$\mathbf{0.737 \pm 0.015}$	$\mathbf{0.148 \pm 0.020}$	$\mathbf{0.033 \pm 0.007}$	$\mathbf{0.727 \pm 0.016}$	$\mathbf{0.447 \pm 0.024}$	$\mathbf{0.071 \pm 0.012}$

P.3 Cross-Domain Synthesis

Table 19 summarises the test-period AP improvement (ΔAP) delivered by EXOROUTE over the corresponding Original/no-exogenous backbone; the post-shift label applies only where a shift is annotated.

Table 19: Test-period AP improvement (ΔAP) of EXOROUTE over the corresponding Original/no-exogenous backbone, averaged across backbones. For datasets with an annotated shift, the reported period is post-shift. Higher is better.

Dataset	**Domain**	**Shift speed**	**ΔAP (mean)**
Elliptic	Crypto AML	Sudden	+0.100
BTC-Alpha	Crypto trust	Unannotated	+0.027
BTC-OTC	Crypto trust	Unannotated	+0.033
LendingClub	Consumer credit	Medium	+0.046
S&P 500	Equity market	Sudden	+0.111

The observed gains across crypto graphs and the credit and equity pilots provide evidence within the tested financial settings; they do not establish effectiveness for arbitrary temporal graphs or document streams. We observe that EXOROUTE's improvements tend to be larger when the regime shift is sharper and the documents are more precisely aligned in time, consistent with the design rationale of the bounded gating mechanism.

Caveats. The LendingClub and S&P 500 benchmarks are self-constructed pilot datasets, not community-recognized benchmarks. Graph construction choices (geographic co-exposure radius, co-stress threshold) introduce subjectivity; we report sensitivity to these hyperparameters below. We encourage the community to develop standardised cross-modal temporal graph benchmarks that pair interaction streams with exogenous document corpora.

Sensitivity to graph construction. We vary key graph-construction hyperparameters and report post-shift AP (TGN backbone, EXOROUTE variant, mean over 5 seeds) to assess robustness. For the credit network, changing the co-exposure window from 60 to 30/90 days yields AP of 0.391/0.403 vs. 0.399 (default 60d), a variation of <1 point. For the stock network, changing the co-stress threshold from 2σ to $1.5\sigma/3\sigma$ yields AP of 0.449/0.462 vs. 0.458 (default 2σ), a variation of <1.5 points, indicating that results are not sensitive to the exact threshold choice.

Q Reproducibility statement

Extended limitations discussion. This section expands on the limitations noted in Sec. 6. EXOROUTE requires a time-stamped document stream relevant to graph prediction; its effectiveness depends on the quality and availability of that evidence. The extraction pipeline adds offline inference cost and its quality depends on the underlying LLM. The evidence-span constraint (Sec. H) reduces reliance on parametric memory, and downstream results are reasonably consistent across extractors (Table 10). The principal evaluation uses public crypto graphs augmented with our document streams. The self-constructed cross-domain pilots (Appendix P) provide supplementary evidence; standardised non-crypto benchmarks with paired exogenous corpora would support broader evaluation.

Potential risks and data considerations. ExoRoute may be used in high-stakes risk screening; extraction or retrieval errors can increase false positives/negatives and should be paired with human review and conservative thresholds. All exogenous sources are publicly available; we follow source terms, release only derived artifacts (timestamps, document IDs, and extracted structured fields), and do not collect or use private data.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction describe time-available document-to-signal extraction, reliability-weighted bounded fusion, and training-period retrieval distillation. Crypto-graph results appear in Section 5, supplementary financial pilots in Appendix P, and the conditions for representation bounds in Appendix A.1.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed in Section 6 (document-stream availability and the distinction between bounded influence and predictive performance) and expanded in Appendix Q (domain-specific schema design, limited evaluation domains, potential risks).

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The representation bound, conditional error contraction, gradient bias bound, and KL distillation result are stated with their assumptions and proved in Appendix A. The routing-score consistency analysis explicitly assumes a utility-aligned, unbiased score proxy and limits its ranking claim to expected scores.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Dataset construction and splits are described in Appendix D.1, hyperparameters in Appendix G, backbone configurations in Appendix D.2, and training protocols in Section 5. Public graph datasets are augmented with author-constructed document streams; supplementary financial graph constructions are described in Appendix P.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The existing code repository is available at `https://github.com/liuliangzu/ExoRoute`. Elliptic, BTC-Alpha, and BTC-OTC are public graph datasets; the document-stream and supplementary financial graph constructions are described in the Appendix, along with reproduction instructions.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 5 specifies evaluation protocols and data splits. Appendix D.1 details dataset statistics, Table 7 and Appendix G report training settings, and Appendix F provides the full training objective.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: All main results report mean $\pm$ standard deviation over 5 random seeds.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Hardware ($4\times$ NVIDIA H100 80 GB), per-batch inference latency, and storage overhead are reported in Appendix E.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research uses only publicly available data and does not involve human subjects. Potential risks are discussed in Appendix Q.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Appendix Q discusses potential negative impacts (false positives/negatives in high-stakes risk screening, need for human review) and data considerations. Positive impacts (improved detection of illicit activity under regime shifts) are implicit in the problem motivation.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release pre-trained language models or scraped datasets. The method augments existing temporal GNN backbones and uses publicly available benchmark datasets.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets (Elliptic, BTC-Alpha, BTC-OTC) and backbone models are properly cited with original publications. These datasets are publicly released by their authors for research use; specific license terms are available at the original repositories.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Code and documentation are available at `https://github.com/liuliangzu/ExoRoute`. The exogenous document streams and their construction methodology are described in Appendix I.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing or research with human subjects.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve human subjects research.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: LLMs are used as the schema-guided extractor module in our pipeline (Section 4.2). We benchmark multiple extractors (open-source vs. commercial) and report quality–cost trade-offs in Section 5 and Appendix J.