WHEN SEMANTICS DEFY CONTEXT: SEMANTIC ANOMALY DETECTION IN TEXT-ATTRIBUTED GRAPHS

Anonymous authors
Paper under double-blind review

ABSTRACT

Graph anomaly detection in text-attributed graphs (TAGs) seeks to identify suspicious entities using textual attributes and relational information. A key challenge is detecting semantic anomalies: nodes whose content is plausible in isolation but incompatible with their relational context. Existing methods that primarily model deviations in node attributes or graph structure can struggle to identify the semantic conflicts underlying such anomalies. Specifically, textual similarity can obscure semantic conflicts, while reliance on a single contextual scale conceals anomalies or misclassifies legitimate semantic differences. Without anomaly labels, constructing informative negative examples is also challenging because strong perturbations create obvious outliers, whereas weak ones preserve contextual compatibility. We propose the Context-Aware Anomaly Reasoning Network (CARNet), an unsupervised framework for assessing semantic compatibility between nodes and their contexts. CARNet compares attribute-based and structure-based community distributions to expose potential conflicts and integrates local and global evidence to infer the semantics expected from the context. Mixing attributes between semantically dissimilar nodes from different communities while preserving the target's relational context provides synthetic mismatches for self-supervised learning. Extensive experiments on five benchmark datasets demonstrate that CARNet achieves state-of-the-art or competitive performance, with further analyses supporting the effectiveness of its key components.

1 INTRODUCTION

Graph anomaly detection (GAD) aims to identify entities whose attributes or relational patterns deviate from expected behavior. In text-attributed graphs (TAGs), this task involves nodes enriched with textual information that conveys their semantic characteristics, such as paper abstracts in citation networks (Pan et al., 2026), user profiles in social networks (Hu et al., 2024), and item descriptions in product networks (Yu et al., 2023). Detecting anomalies in these graphs helps uncover suspicious activities and maintain the integrity and reliability of the underlying systems.

A key challenge in anomaly detection on TAGs is identifying **semantic anomalies**: nodes whose content is plausible in isolation but incompatible with their relational context. Such anomalies underlie out-of-context misinformation and misattributed content on social platforms, where accurate content is repurposed to support a false narrative. Consider a social media graph in which nodes represent posts and edges capture reply or repost relationships. As illustrated in Figure 1(a), a post documenting an evacuation drill in one city fits naturally within a discussion connected to the drill announcement and participants' accounts. In contrast, Figure 1(b) shows the same post circulated as evidence of an ongoing disaster in another city. Although its content remains factually accurate, its use as evidence conflicts with the location and event described by the surrounding posts.

Several recent methods are designed specifically for anomaly detection on TAGs. CMUCL (Xu et al., 2025c) scores each node by the consistency between its text and its structure at multiple scales. CoLL (Xu et al., 2025b) uses multiple LLMs to generate anomaly evidence and fuses this evidence with structural information through a gated GNN. CAMERA (Pan et al., 2026) targets semantic camouflage by adaptively combining graph, semantic, and global deviation cues. These

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

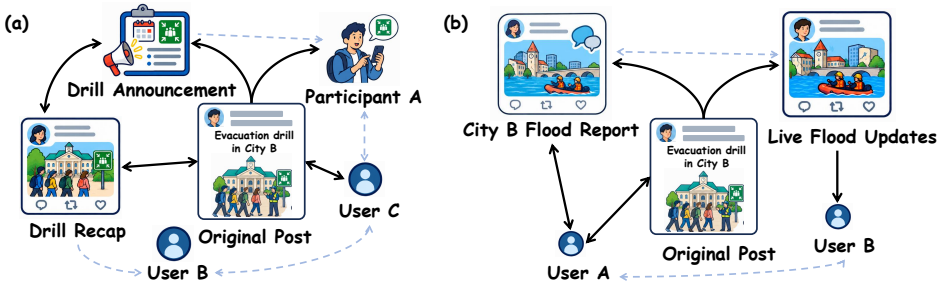

Figure 1: Semantic anomalies in a social media graph. The same post documenting an evacuation drill is (a) compatible with discussions of that drill but (b) anomalous when misattributed as evidence of a disaster in another city.

methods show that text and structure must be considered together. However, all of them score a node by how much it deviates from the learned normal pattern. None of them compares what the node's text says with what its relational context expects. A semantic anomaly is not detected by such scores. Its text is plausible on its own and similar to the text of its neighbors, so it shows little inconsistency and little deviation, even though its meaning conflicts with its context. The lack of anomaly labels makes this problem harder.

Addressing this problem in the absence of anomaly labels requires overcoming three challenges. (1) **Masked Semantic Conflicts.** A node's text expresses its intrinsic semantics, while its connections imply the semantics expected from its relational context. These expectations remain implicit, making direct comparison difficult. Textual similarity between the node and its neighbors can further obscure the semantic conflict. For example, an evacuation drill post shares much vocabulary with disaster reports, so it looks normal when it is linked as evidence of an actual disaster. (2) **Local-Global Mismatch.** Local similarity can make semantic mismatches difficult to detect, while global context alone lacks the relational detail needed to interpret them. Among its direct neighbors, the drill post resembles the flood reports around it, which also mention evacuation, so the local context shows no conflict. Only the wider discussion, which focuses on casualties and property damage, reveals that the post does not belong. The wider context alone, however, cannot decide whether the post is a false report or a legitimate reference to evacuation procedures. Reliable detection therefore requires integrating local relational evidence with global semantic context. (3) **Scarcity of Informative Anomalies.** Without anomaly labels, learning contextual incompatibility relies on constructing informative synthetic anomalies. However, arbitrary perturbations offer unreliable training signals: strong perturbations can create obvious outliers that provide little guidance for detecting semantic conflicts, whereas weak perturbations can leave contextual compatibility intact.

To this end, we propose the **C**ontext-Aware **A**nomaly **R**easoning **Net**work (CARNet), an unsupervised framework that detects semantic anomalies by testing whether a node's context can explain its observed semantics. CARNet addresses the three challenges with three components. Conflict-aware semantic community inference makes the two meanings of a node comparable by projecting both into a shared space of latent communities. Local-global context reasoning combines global community information, which reveals conflicts that local similarity hides, with local relational evidence, which decides whether a revealed conflict is genuine. Cross-context semantic perturbation creates informative synthetic anomalies by mixing a node's attributes with those of the most dissimilar node from a different community while keeping its relational context fixed. Our main contributions are summarized as follows:

- We characterize semantic anomalies in TAGs as conflicts between the semantics a node expresses in its text and those its relational context implies. This characterization shifts the objective of anomaly detection from measuring how far a node deviates from its context to testing whether that context can account for its semantics.
- We propose CARNet, an unsupervised framework for semantic anomaly detection. It compares node semantics with community expectations and integrates local and global context to assess semantic compatibility. Together, these mechanisms reveal semantic conflicts obscured by textual similarity or overlooked at a single contextual scale.
- We construct synthetic anomalies by modifying node semantics under latent community guidance while preserving context. This strategy provides more targeted supervision for contextual incompatibility than random perturbations, enabling CARNet to learn semantic conflicts.

2 Related Work

Due to GAD data annotation demands substantial human effort, many methods have adopted unsupervised learning paradigms. Early methods were mainly based on matrix factorization (Li et al., 2017), clustering (Xu et al., 2007). Recently, with the rapid advancement of GNNs, DOMINANT (Ding et al., 2019), ADA-GAD (He et al., 2024) and GAD-NR (Roy et al., 2024) detect anomalies from errors to reconstruct node attributes, graph structure, or local neighborhoods. One-class methods, including OCGNN (Wang et al., 2021) and TAM (Qiao & Pang, 2023), characterize normality through compact representations or local homophily. CoLA (Liu et al., 2021), AENMONE (Jin et al., 2021) introduce contrastive learning techniques for GAD. SL-GAD (Zheng et al., 2022), GRADATE (Duan et al., 2023), HUGE-GAD (Pan et al., 2025), and HCM (Huang et al., 2022) leverage node subgraph agreement, contextual attribute regression, and structural tasks. Recent studies further explore denoised reconstruction (Pan et al., 2026), diffusion models (Li et al., 2025), and generalist detection (Liu et al., 2024; Lin et al., 2024). Despite their effectiveness, most existing methods do not explicitly assess whether a node's context can explain its semantics.

3 Problem Formulation

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{X}, \mathbf{T})$ be a Text Attributed Graph (TAG), where $\mathcal{V} = \{v_i\}_{i=1}^N$ and $\mathcal{E}$ denote the node and edge sets, respectively, Each node v_i is associated with a d-dimensional attribute vector $\mathbf{x}_i \in \mathbb{R}^d$ and a textual description $\mathbf{t}_i$. Accordingly, $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^\top \in \mathbb{R}^{N \times d}$ denotes the node attribute matrix, while $\mathbf{T} = \{\mathbf{t}_i\}_{i=1}^N$ represents the collection of node-level textual information.

We study unsupervised node semantics anomaly detection, aiming to identify nodes whose intrinsic attributes are inconsistent with their contexts. For each node $v_i \in \mathcal{V}$, let $\mathcal{C}_i = \Phi\big(\mathcal{N}_i, \mathbf{X}_{\mathcal{N}_i}, \mathcal{S}_i\big)$ denote its graph context, where $\mathcal{N}_i$ is the neighborhood of v_i, $\mathbf{X}_{\mathcal{N}_i}$ contains the attributes of its neighboring nodes, $\mathcal{S}_i$ represents higher-order structural information such as community membership, and $\Phi(\cdot)$ is a context-construction function. Unlike conventional node anomalies, which exhibit unusual attributes or connectivity patterns in isolation, semantic anomalies are not necessarily distinguishable from normal nodes based solely on their individual observations. Specifically, a node v_i is considered semantic anomalous when its attribute $\mathbf{x}_i$ is plausible under the attribute distribution but unlikely given its graph context $\mathcal{C}_i$. Formally, $p(\mathbf{x}_i) \gg 0, p(\mathbf{x}_i \mid \mathcal{C}_i) \approx 0$, semantic anomaly detection focuses on identifying violations of the expected attribute–context correspondence.

4 Method

We present the **C**ontext-Aware **A**nomaly **R**easoning **Net**work (CARNet), an unsupervised framework that detects semantic node anomalies by assessing compatibility between node semantics and relational context. As illustrated in Figure 2, CARNet first infers community distributions from attributes and graph structure to expose potential semantic conflicts. It then integrates local relational evidence with global community information to infer expected semantics. Finally, cross-context semantic perturbation mixes each node's attributes with those of the most dissimilar node from a different community, providing negative examples for self-supervised learning.

4.1 Conflict-Aware Semantic Community Inference

Textual similarity can obscure conflicts between a node's intrinsic semantic affiliation and that implied by its relationships. To expose these conflicts, CARNet infers two distributions over shared latent communities: an attribute-based distribution capturing intrinsic semantic affiliation and a structure-based distribution reflecting contextual expectations. Disagreement between these distributions provides an initial indication of semantic conflict.

Given a TAG $\mathcal{G}$, we first propagate node attributes through a lightweight GNN:

$$\mathbf{Z} = f_{\mathrm{GNN}}(\mathbf{X}, \mathcal{E}), \qquad \mathbf{Z} \in \mathbb{R}^{N \times d}. \tag{1}$$

Unlike the original attributes $\mathbf{X}$ which describe nodes independently, $\mathbf{Z}$ incorporates relational evidence from the graph. Consequently, $\mathbf{x}_i$ characterizes the intrinsic semantics expressed by node v_i, while $\mathbf{z}_i$ captures these semantics conditioned on the node's context.

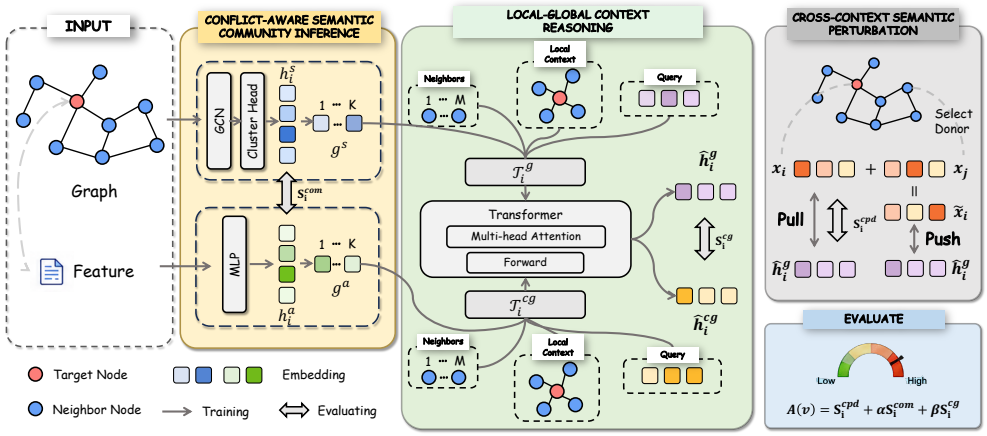

Figure 2: Overall Framework of the Proposed CARNet.

To make these two semantic views directly comparable, we project them into a shared latent identity space consisting of K latent semantic communities:

$$\begin{cases} \mathbf{H}^s = \mathrm{softmax}\left(\dfrac{\mathbf{Z}\mathbf{W}_s^\top + \mathbf{b}_s}{\tau_c}\right) \\ \mathbf{H}^a = \mathrm{softmax}\left(\dfrac{\mathbf{X}\mathbf{W}_a^\top + \mathbf{b}_a}{\tau_c}\right) \end{cases} \qquad \mathbf{H}^s, \mathbf{H}^a \in \mathbb{R}^{N\times K}, \tag{2}$$

where τ_c controls the sharpness of the assignments.

To quantify the disagreement between the two community views, we compute the node-wise KL divergence from the structure-based assignment to the attribute-based assignment:

$$s_i^{\mathrm{com}} = D_{\mathrm{KL}}(\mathbf{h}_i^s \,\|\, \mathbf{h}_i^a)\,. \tag{3}$$

The structure-based assignment further provides a semantic description of the global context surrounding each node. Specifically, we summarize each latent community with a prototype:

$$\mathbf{P}_k = \frac{\sum_{j=1}^N h_{jk}^s \mathbf{z}_j}{\sum_{j=1}^N h_{jk}^s + \epsilon}, \qquad \mathbf{g}_i = \sum_{k=1}^K h_{ik}^s \mathbf{P}_k. \tag{4}$$

Each prototype $\mathbf{P}_k$ captures the shared relational semantics of the k-th latent semantic community, while $\mathbf{g}_i$ aggregates these prototypes according to the structure-based identity of node v_i to form its global semantic context. This soft aggregation allows each node to be associated with multiple semantic communities rather than assigned to a single discrete community. Directly optimizing the module may lead to uninformative uniform assignments or clustering collapse. Therefore, we employ K-means for initialization, with the corresponding procedure detailed in Appendix C.

4.2 Local-Global Context Reasoning

Disagreement between attribute-based and structure-based community distributions indicates a potential semantic conflict, but does not necessarily imply an anomaly. CARNet therefore integrates local relational evidence with global community information to infer the semantics expected from the node's context and assess their compatibility with its observed semantics.

$$\mathbf{t}_{ij}^n = \phi_n(\mathbf{x}_j), \qquad \mathbf{t}_i^l = \phi_l(f_{\mathrm{GAT}}(\mathbf{X}, \mathcal{E}))\,, \qquad \mathbf{t}_i^g = \phi_g(\mathbf{g}_i), \quad v_j \in \mathcal{N}(v_i), \tag{5}$$

where at most M one-hop neighbors are sampled and missing neighbor positions are masked. Together, these tokens capture each node's contextual semantics at multi granularities: $\mathbf{t}^n$ retains fine-grained evidence from individual neighbors, $\mathbf{t}^l$ encodes structural information within the local neighborhood, and $\mathbf{t}^g$ provides a global semantic anchor derived from the prototype mixture.

We organize these contextual signals into a token sequence together with a learnable query token $\mathbf{q}$:

$$\mathcal{T}_i^g = \left[\mathbf{t}_{i1}^n, \ldots, \mathbf{t}_{iM}^n, \mathbf{t}_i^g, \mathbf{t}_i^l, \mathbf{q}\right]. \tag{6}$$

A Transformer-based context reasoner f_θ models interactions among the different contextual signals, and the representation at the query position is decoded as the context-conditioned prediction of the target node:

$$\hat{\mathbf{h}}_i^g = f_\theta(\mathcal{T}_i^g)_{\mathbf{q}}. \tag{7}$$

Since the target attribute $\mathbf{x}_i$ is withheld from the input, the model does not rely on the node's own attributes for prediction. Instead, it uses the contextual evidence encoded in these tokens to assess how well the surrounding context explains the semantics of v_i.

Controlled Context Intervention. CARNet further probes semantic compatibility by testing how contextual predictions change when the global context is aligned with the node's attribute-based identity. We intervene only on the global context while holding immediate and local evidence fixed, allowing the prediction change to isolate the model's sensitivity to global identity disagreement.

Following Eq. equation 2–equation 6, we use the attribute-implied assignments $\mathbf{h}_i^a$ to construct an alternative global context $\mathbf{g}_i^a$ in the shared space $\mathbf{Z}$. Unlike the structure-implied context $\mathbf{g}_i^s$, $\mathbf{g}_i^a$ reflects the node's intrinsic semantic identity. We intervene by replacing only the global context token while holding all other inputs and the context reasoner fixed:

$$\mathbf{t}_i^g = \phi_g(\mathbf{g}_i^s) \quad \longrightarrow \quad \mathbf{t}_i^{cg} = \phi_g(\mathbf{g}_i^a). \tag{8}$$

To train the context reasoner to recover the target semantics from contextual evidence, we minimize the reconstruction error between the predicted representation and the withheld node attribute. Specifically, the reconstruction losses under the original and intervened contexts are defined as

$$\mathcal{L}_{\text{ctx}} = \frac{1}{N}\sum_{i=1}^{N}\|\hat{\mathbf{h}}_i^g - \mathbf{x}_i\|_2^2, \qquad \mathcal{L}_{cg} = \frac{1}{N}\sum_{i=1}^{N}\|\hat{\mathbf{h}}_i^{cg} - \mathbf{x}_i\|_2^2. \tag{9}$$

Here, $\mathcal{L}_{\text{ctx}}$ serves as the primary reconstruction objective, encouraging the context reasoner to recover the target semantics from the observed context. In contrast, $\mathcal{L}_{\text{cg}}$ acts as a counterfactual reconstruction regularizer, preventing the intervened branch from degenerating into arbitrary representations unrelated to the target semantics. The intervention yields an alternative context-conditioned prediction $\hat{\mathbf{h}}_i^{cg}$, whose deviation from the factual prediction is measured by

$$s_i^{cg} = 1 - \cos\left(\hat{\mathbf{h}}_i^g, \hat{\mathbf{h}}_i^{cg}\right). \tag{10}$$

$\mathbf{s}_i^{cg}$ measures the extent to which the disagreement between intrinsic and relationally implied identities affects contextual reasoning.

4.3 Cross-Context Semantic Perturbation

Without anomaly labels, arbitrary attribute perturbations produce obvious outliers or leave contextual compatibility intact. CARNet therefore mixes each target node's attributes with those of its most dissimilar counterpart from a different latent semantic community while keeping the target's relational context fixed. This strategy retains semantic similarity while introducing potential contextual mismatches, providing informative negative examples for learning contextual incompatibility.

At each training step, we sample a subset $\mathcal{V}_{\text{syn}} \subseteq \mathcal{V}$ containing approximately $p_{\text{syn}}N$ nodes. For each selected node v_i, the donor is chosen from a different structure-implied latent community while being as distinct as possible in terms of soft contextual identity:

$$\arg\max_k h_{ik}^s \neq \arg\max_k h_{jk}^s, \qquad j^* = \arg\min_{j\in\mathcal{C}_i} \cos\left(\mathbf{h}_i^s, \mathbf{h}_j^s\right). \tag{11}$$

Here, $\mathcal{C}i$ denotes a randomly sampled candidate set. The hard identity constraint ensures that the donor comes from a different latent community, while the soft identity similarity favors a donor with distinct contextual semantics.

We then construct a synthetic semantic perturbation by interpolating their representations:

$$\widetilde{\mathbf{x}}_i = \lambda \mathbf{x}_i + (1-\lambda)\mathbf{x}_j, \tag{12}$$

The graph position and surrounding context of v_i remain unchanged, while $\widetilde{\mathbf{x}}_i$ introduces semantics from a different latent context. This produces a synthetic anomaly with semantics that conflict with its context, providing an explicit example of contextual incompatibility for the reasoner.

Self-Supervised Context Discrimination. The synthetic perturbations provide supervision for learning which semantics are supported. Specifically, the factual context-conditioned prediction $\hat{\mathbf{h}}_i^g$ should remain close to the original node representation $\mathbf{x}_i$, while being separated from its context-incompatible alternative $\widetilde{\mathbf{x}}_i$. We define the positive and negative similarities as

$$s_i^+ = \frac{\cos\left(\mathbf{x}_i, \hat{\mathbf{h}}_i^g\right)}{\tau}, \qquad s_i^- = \frac{\cos\left(\widetilde{\mathbf{x}}_i, \hat{\mathbf{h}}_i^g\right)}{\tau}. \tag{13}$$

where τ is the contrastive temperature. The corresponding objective is

$$\mathcal{L}_{\text{con}} = -\frac{1}{|\mathcal{V}_{\text{syn}}|} \sum_{i \in \mathcal{V}_{\text{syn}}} \log \frac{\exp(s_i^+)}{\exp(s_i^+) + \exp(s_i^-)}. \tag{14}$$

Minimizing $\mathcal{L}_{\text{con}}$ encourages the reasoner to predict semantics supported by the observed context while suppressing information from incompatible contexts.

The model is thus trained to align its contextual prediction with observed node semantics while deviating from synthetic ones, which teaches it what counts as anomalous. We therefore define the contextual prediction discrepancy as

$$s_i^{\text{cpd}} = 1 - \cos\left(\mathbf{x}_i, \hat{\mathbf{h}}_i^g\right). \tag{15}$$

A high s_i^{cpd} indicates that the node's observed semantics are poorly explained by its context.

4.4 Joint Learning and Anomaly Scoring

CARNet is jointly optimized by the context reconstruction objective $\mathcal{L}_{\text{ctx}}$, the self-supervised discrimination objective $\mathcal{L}_{\text{con}}$, the counterfactual reconstruction regularizer $\mathcal{L}_{\text{cg}}$, and an auxiliary regularization term $\mathcal{L}_{\text{reg}}$ that combines community regularization and intervention sensitivity objectives to prevent community collapse and stabilize node representations. The overall objective is

$$\mathcal{L} = \mathcal{L}_{\text{ctx}} + \mathcal{L}_{\text{con}} + \lambda_{\text{cg}}\mathcal{L}_{\text{cg}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}}, \tag{16}$$

where the detailed formulation of $\mathcal{L}_{\text{reg}}$ is provided in Appendix D.

At inference, CARNet combines the community discrepancy s_i^{com}, contextual prediction discrepancy s_i^{cpd}, and controlled-intervention discrepancy s_i^{cg}. These signals respectively measure cross-view identity conflict, the extent to which the observed semantics cannot be explained by the surrounding context, and the prediction change caused by replacing the global contextual identity. The final anomaly score is defined as

$$\mathcal{A}(v_i) = s_i^{\text{cpd}} + \alpha s_i^{\text{com}} + \beta s_i^{cg}, \tag{17}$$

where α and β weight identity disagreement and intervention sensitivity, respectively.

5 Experiments

5.1 Experimental Setup

Datasets. We evaluate CARNet on five text attributed graphs from citation and social networks: Cora (McCallum et al., 2000), Citeseer (Giles et al., 1998), Pubmed (Sen et al., 2008), Wikics (Mernyei & Cangea, 2020), and Arxiv (Hu et al., 2020). Since these original datasets do not provide semantic anomaly labels, we construct their semantic anomaly detection variants following the anomaly injection protocol of TAG-AD (Xu et al., 2025a). The complete construction procedure and the dataset statistics is provided in Appendix E.1.

Table 1: Overall Anomaly Detection Performance (%). The Best (**Bold**) and Suboptimal (Underlined) Results are Marked, Respectively. OOM: Out-of-Memory Error.

	Cora		Citeseer		Pubmed		Wikics		Arxiv	
Method	**AUC**	**AP**	**AUC**	**AP**	**AUC**	**AP**	**AUC**	**AP**	**AUC**	**AP**
ANOMALOUS	40.04 ± 2.54	6.08 ± 0.38	41.42 ± 1.26	10.14 ± 0.16	48.82 ± 0.97	5.03 ± 0.14	48.60 ± 0.60	8.56 ± 0.11	OOM	
DOMINANT	76.46 ± 0.14	27.36 ± 0.14	65.66 ± 0.42	28.65 ± 0.22	60.88 ± 1.08	8.25 ± 0.15	53.45 ± 0.22	9.76 ± 0.07	OOM	
AnomalyDAE	41.38 ± 2.02	6.65 ± 0.28	50.08 ± 1.11	11.82 ± 0.43	51.00 ± 4.82	5.28 ± 0.68	50.20 ± 2.36	9.03 ± 0.62	OOM	
CoLA	69.93 ± 1.89	16.06 ± 1.34	70.36 ± 1.08	25.11 ± 1.66	69.42 ± 0.77	15.46 ± 1.47	60.53 ± 0.76	15.23 ± 0.83	84.42 ± 0.97	33.54 ± 1.21
CONAD	76.37 ± 1.22	27.21 ± 1.31	65.77 ± 0.27	28.70 ± 0.22	60.87 ± 1.08	8.25 ± 0.15	53.30 ± 0.12	9.75 ± 0.05	OOM	
GCTAM	73.96 ± 1.71	14.09 ± 0.94	72.73 ± 1.75	25.62 ± 1.90	75.56 ± 1.33	12.97 ± 1.24	73.80 ± 1.73	30.84 ± 1.09	78.29 ± 2.51	25.37 ± 1.06
GAD-NR	58.99 ± 2.57	10.50 ± 1.55	65.16 ± 2.61	24.67 ± 1.79	70.24 ± 1.52	11.35 ± 1.49	60.38 ± 2.44	12.42 ± 1.30	69.79 ± 1.04	24.32 ± 2.59
ADA-GAD	72.89 ± 0.36	25.13 ± 0.55	60.27 ± 0.13	25.25 ± 0.02	68.73 ± 0.72	28.05 ± 1.62	69.70 ± 1.99	23.71 ± 1.75	76.52 ± 1.41	29.08 ± 2.11
GRADATE	61.72 ± 3.05	10.19 ± 1.01	57.95 ± 1.89	13.95 ± 0.67	78.60 ± 2.17	40.52 ± 3.02	68.66 ± 1.09	19.64 ± 1.77	OOM	
DAGAD	67.09 ± 2.41	7.03 ± 1.21	69.59 ± 2.06	32.13 ± 1.74	67.72 ± 1.96	24.97 ± 1.72	64.30 ± 2.82	14.85 ± 1.61	OOM	
HUGE-GAD	90.06 ± 1.45	59.77 ± 1.33	78.28 ± 1.57	45.31 ± 1.02	84.12 ± 0.98	39.39 ± 0.77	71.88 ± 2.37	21.75 ± 1.80	88.53 ± 0.45	40.17 ± 1.39
CMUCL	77.16 ± 1.44	31.98 ± 0.86	86.61 ± 1.40	54.54 ± 0.63	81.87 ± 0.74	50.31 ± 0.20	80.51 ± 1.21	38.09 ± 0.32	87.58 ± 2.04	38.35 ± 1.26
CoLL	90.10 ± 1.01	55.79 ± 1.27	81.03 ± 1.04	51.44 ± 0.83	87.65 ± 1.25	51.04 ± 0.51	76.85 ± 1.39	24.61 ± 1.48	90.45 ± 1.45	**48.92 ± 1.02**
CAMERA	**93.63 ± 1.66**	58.72 ± 1.20	85.50 ± 0.97	54.39 ± 0.70	86.47 ± 0.87	39.12 ± 1.39	81.27 ± 2.11	40.90 ± 2.84	91.76 ± 2.13	43.06 ± 2.23
CARNet	92.87 ± 0.50	**64.58 ± 1.78**	**90.09 ± 0.29**	**64.63 ± 1.14**	**89.57 ± 0.24**	**52.88 ± 1.18**	**85.48 ± 0.32**	**46.93 ± 1.40**	**95.50 ± 0.30**	48.01 ± 2.58

Baselines. We compare CARNet with representative graph anomaly detection methods, including, ANOMALOUS (Peng et al., 2018), DOMINANT (Ding et al., 2019), AnomalyDAE (Fan et al., 2020), CoLA (Liu et al., 2021), CONAD (Xu et al., 2022), GRADATE (Duan et al., 2023), GAD-NR (Roy et al., 2024), ADA-GAD (He et al., 2024), GCTAM (Zhang et al., 2025), DAGAD (Liu et al., 2022), HUGE-GAD (Pan et al., 2025), CMUCL (Xu et al., 2025c), CoLL (Xu et al., 2025b), and CAMERA (Pan et al., 2026). Detailed descriptions and implementation settings of the baseline methods are provided in Appendix E.2.

5.2 Overall Detection Performance

We evaluate the overall anomaly detection performance of our proposed method, CARNet, against 14 state-of-the-art baselines across five benchmark datasets using two standard anomaly detection metrics: ROC-AUC (AUC) and average precision (AP). As summarized in Table 1, CARNet achieves leading or competitive performance across the evaluated datasets. On CiteSeer, CARNet attains an AUC of 90.09 and an AP of 64.63, outperforming the second-best results by 4.02% and 18.50%, respectively. On Cora, CARNet achieves a competitive AUC of 92.87, although CAMERA performs slightly better with an AUC of 93.63.

CARNet also performs well on the large-scale Arxiv dataset, where several baselines cannot complete execution due to out-of-memory errors. In contrast, CARNet scales to this dataset, achieving an AUC of 95.50 and an AP of 48.01. Its AUC surpasses CAMERA's by 4.08%, while its AP is close to the best result of 48.92 achieved by CoLL. CARNet outperforms all evaluated baselines on WikiCS, with particularly strong AP performance, indicating its ability to rank anomalous nodes highly. Overall, CARNet achieves leading or competitive performance against a diverse set of baselines across five benchmark datasets, although its advantage varies across datasets and metrics. These results support the overall effectiveness of CARNet in assessing semantic compatibility through community-based modeling and local-global context reasoning. Its performance gains over baselines relying primarily on attribute reconstruction or local consistency suggest that the integrated framework better captures contextual incompatibility in the evaluated settings.

5.3 Detection of Different Anomaly Types

We construct two types of anomalies, structural and semantic, using the same anomaly ratio and report results separately for each type to assess the model's sensitivity to different deviation patterns.Table 2 reports the results for each type across five datasets. CARNet achieves competitive performance in semantic anomaly detection, attaining the highest AUC on four datasets and the highest AP on three. Specifically, it outperforms CAMERA in AP by 0.88%, 1.17%, 0.93%, and 8.24% on CiteSeer, PubMed, WikiCS, and Arxiv, respectively. On Cora, CARNet achieves an AP

Table 2: Detection Performance (%) for Different Anomaly Types . The Best (**Bold**) and Suboptimal (Underlined) Results are Marked, Respectively.

Dataset	Method	Global		Semantic		Structural	
		AUC	AP	AUC	AP	AUC	AP
Cora	HUGE-GAD	90.06	59.77	86.62	47.22	**94.07**	**59.25**
	CoLL	90.10	55.79	88.28	50.21	90.93	53.36
	CAMERA	**93.63**	58.72	**94.20**	**52.86**	92.29	33.36
	CARNet	92.87	**64.58**	92.37	51.45	93.54	55.50
Citeseer	HUGE-GAD	78.28	45.31	76.29	39.06	79.18	41.58
	CoLL	81.03	51.44	79.30	46.14	88.20	49.29
	CAMERA	85.50	54.39	88.31	53.91	75.46	42.38
	CARNet	**90.09**	**64.63**	**89.09**	**61.23**	**91.63**	**51.96**
Pubmed	HUGE-GAD	84.12	39.39	78.39	14.26	91.16	58.03
	CoLL	87.65	51.04	87.89	31.56	92.80	54.62
	CAMERA	86.47	39.12	89.25	40.03	86.44	36.82
	CARNet	**89.57**	**52.88**	**90.30**	**48.30**	**94.45**	**60.39**
Wikics	HUGE-GAD	71.88	21.75	79.16	19.90	74.78	18.20
	CoLL	76.85	24.61	83.25	18.45	78.07	17.60
	CAMERA	81.27	40.90	87.37	41.35	67.61	34.49
	CARNet	**85.48**	**46.93**	**88.18**	**46.96**	**78.17**	**36.26**
Arxiv	HUGE-GAD	88.53	40.17	82.46	35.90	90.78	41.20
	CoLL	90.45	**48.92**	89.86	**47.08**	91.35	**46.60**
	CAMERA	91.76	43.06	90.06	41.79	**93.83**	42.36
	CARNet	**95.50**	48.01	**97.48**	46.56	93.26	45.61

Table 3: Ablation Results (%) on the Experimental Datasets.

Variant	AUC						AP					
	Cora	Citeseer	Pubmed	Wikics	Arxiv	Avg. Drop	Cora	Citeseer	Pubmed	Wikics	Arxiv	Avg. Drop
CARNet	92.87	90.09	89.57	85.48	95.50	-	64.58	64.63	52.88	46.93	48.01	-
w/o CaS	87.76	84.95	85.34	81.26	89.40	-5.45%	40.50	50.35	35.69	32.99	49.76	-22.08%
w/o LC	92.04	89.65	89.33	83.84	94.37	-0.95%	60.24	66.99	52.23	44.77	46.53	-2.11%
w/o GC	91.85	88.59	87.74	84.62	93.49	-1.58%	53.12	58.28	44.49	45.18	45.82	-10.06%
w/o IN	89.06	88.16	88.15	87.00	90.91	-2.17%	49.76	65.69	50.43	42.05	44.41	-8.50%
w/o CCS	73.83	77.12	68.29	66.21	77.93	**-19.92%**	38.22	44.11	39.45	33.75	35.39	**-30.25%**

of 51.45, falling short of CAMERA by 1.41%. Overall, these results demonstrate CARNet's effectiveness in distinguishing semantic anomalies from normal nodes across different datasets. On Cora, CARNet achieves a semantic AP of 51.45, falling short of CAMERA by 1.41%. On Arxiv, it attains the highest semantic AUC of 97.48, while its AP of 46.56 is slightly below CoLL's 47.08. Its effectiveness also extends to structural anomalies, for which it achieves the highest AUC and AP on CiteSeer, PubMed, and WikiCS.

5.4 Ablation Study

We evaluate five ablated variants: **w/o CaS** removes conflict-aware semantic community inference; **w/o LC** removes the local context token; **w/o GC** removes the global context token; **w/o IN** removes controlled context intervention; and **w/o CCS** removes cross-context semantic perturbations. Table 3 presents the performance of the ablated variants. Removing synthetic negatives causes the largest performance decline, with an average AUC reduction of 19.92% across the five datasets. This highlights the importance of informative synthetic anomalies for learning contextual incompatibility. Removing the semantic community inference design also degrades performance, supporting the value of modeling semantic affiliations from both attributes and relational context. Removing either local or global context lowers AUC on all five datasets. Removing global context leads to larger average performance drops, with a particularly pronounced effect on AP. These results support the complementary value of local relational evidence and global community information in assessing semantic compatibility. Removing context intervention generally reduces performance, suggesting that prediction sensitivity to changes in global contextual identity provides additional evidence. However, the improvements in AUC on WikiCS and AP on CiteSeer after its removal indicate that this signal is not equally informative across datasets.

CARNet combines three anomaly signals: s^{cpd}, s^{com}, and s^{cg}. As shown in Figure 3, removing s^{cpd} causes the largest performance drop, indicating that it contributes most to the final result., highlighting the value of assessing whether context can reconstruct a node's observed semantics. Meanwhile,

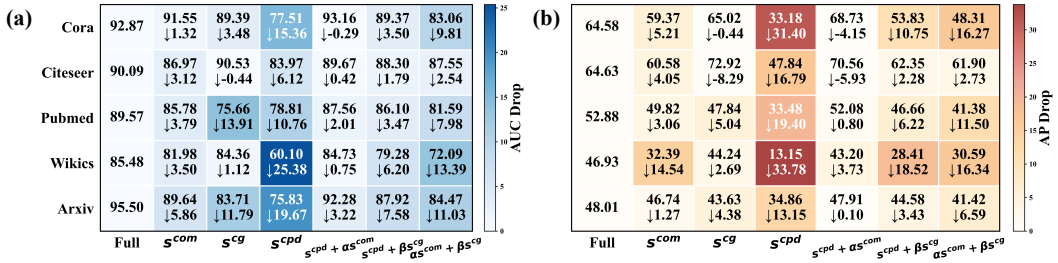

Figure 3: Anomaly Score Composition in CARNet.

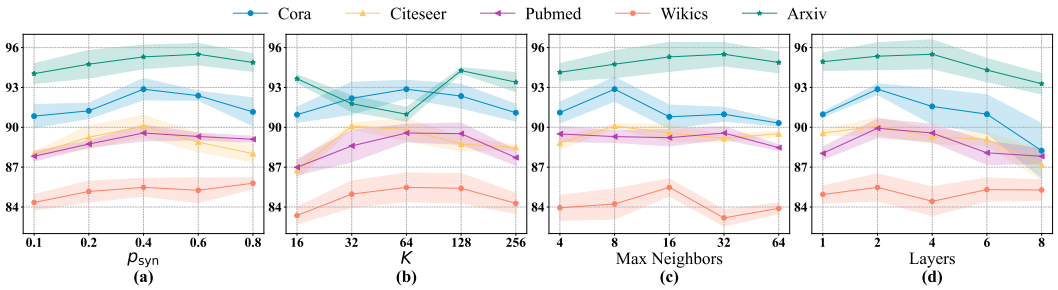

Figure 4: Hyperparameter Sensitivity of CARNet in Terms of AUC on Different Datasets.

s^{com} captures conflicts between attribute-implied and structure-implied identities, and s^{cg} measures sensitivity to global context interventions. Combining s^{cpd} and s^{com} improves detection by jointly assessing contextual reconstruction and identity disagreement, whereas the additional gains from s^{cg} are modest and dataset-dependent.

5.5 Hyperparameter Sensitivity

We further examine sensitivity to the number of communities K, maximum neighbor count M, synthetic ratio p_{syn}, and number of layers L, varying one parameter at a time while keeping the others fixed. The AUC results are presented in Figure 4, and the AP results are provided in Appendix F.2. Moderate values of K and M generally yield better performance. In most cases, $K = 64$ balances coarse community partitions against excessive fragmentation, while small to moderate values of M capture informative neighborhood evidence with limited noise. These trends suggest that contextual reasoning benefits from sufficiently expressive semantic communities and selective neighborhood coverage. For p_{syn}, values around 0.4 achieve the best or near-best results on most datasets, suggesting that synthetic augmentation provides useful supervision but may become less effective when excessive. Finally, shallow architectures with two to four layers generally outperform deeper models, indicating that additional propagation does not necessarily improve semantic anomaly detection. One possible explanation is that excessive propagation blurs the semantic distinctions needed to identify incompatibility between nodes and their contexts.

6 Conclusion

Semantic anomalies in text-attributed graphs can remain hidden when plausible node content conflicts with its context. Detecting such anomalies requires distinguishing genuine contextual incompatibility from legitimate semantic variation. We presented CARNet, an unsupervised framework for addressing this challenge. CARNet compares attribute-based and structure-based community distributions to expose potential semantic conflicts and integrates local and global evidence to assess contextual compatibility. Controlled global context interventions further examine the influence of community disagreement, while community-guided attribute mixing provides informative self-supervision without anomaly labels. Experiments on five benchmark datasets demonstrate competitive detection performance, and ablation studies support the importance of synthetic supervision and the complementary value of local and global contextual evidence.

REFERENCES

Kaize Ding, Jundong Li, Rohit Bhanushali, and Huan Liu. Deep anomaly detection on attributed networks. In *Proceedings of the 2019 SIAM international conference on data mining*, pp. 594–602. SIAM, 2019.

Jingcan Duan, Siwei Wang, Pei Zhang, En Zhu, Jingtao Hu, Hu Jin, Yue Liu, and Zhibin Dong. Graph anomaly detection via multi-scale contrastive learning networks with augmented view. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 7459–7467, 2023.

Haoyi Fan, Fengbin Zhang, and Zuoyong Li. Anomalydae: Dual autoencoder for anomaly detection on attributed networks. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5685–5689. IEEE, 2020.

C. Lee Giles, Kurt D. Bollacker, and Steve Lawrence. Citeseer: an automatic citation indexing system. In *Proceedings of the Third ACM Conference on Digital Libraries*, DL '98, pp. 89–98, New York, NY, USA, 1998. Association for Computing Machinery. ISBN 0897919653. doi: 10.1145/276675.276685. URL https://doi.org/10.1145/276675.276685.

Junwei He, Qianqian Xu, Yangbangyan Jiang, Zitai Wang, and Qingming Huang. Ada-gad: Anomaly-denoised autoencoders for graph anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 8481–8489, 2024.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. *Advances in neural information processing systems*, 33:22118–22133, 2020.

Xinxin Hu, Haotian Chen, Hongchang Chen, Shuxin Liu, Xing Li, Shibo Zhang, Yahui Wang, and Xiangyang Xue. Cost-sensitive gnn-based imbalanced learning for mobile social network fraud detection. *IEEE Transactions on Computational Social Systems*, 11(2):2675–2690, 2024. doi: 10.1109/TCSS.2023.3302651.

Tianjin Huang, Yulong Pei, Vlado Menkovski, and Mykola Pechenizkiy. Hop-count based self-supervised anomaly detection on attributed networks. In *Machine Learning and Knowledge Discovery in Databases*, 2022.

Ming Jin, Yixin Liu, Yu Zheng, Lianhua Chi, Yuan-Fang Li, and Shirui Pan. Anemone: Graph anomaly detection with multi-scale contrastive learning. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, pp. 3122–3126, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384469. doi: 10.1145/3459637.3482057. URL https://doi.org/10.1145/3459637.3482057.

Jinghan Li, Yuan Gao, Jinda Lu, Junfeng Fang, Congcong Wen, Hui Lin, and Xiang Wang. Diffgad: A diffusion-based unsupervised graph anomaly detector. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 37744–37766, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/5ddc53810fcd805697905f40ffe9102d-Paper-Conference.pdf.

Jundong Li, Harsh Dani, Xia Hu, and Huan Liu. Radar: Residual analysis for anomaly detection in attributed networks. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pp. 2152–2158, 2017. doi: 10.24963/ijcai.2017/299. URL https://doi.org/10.24963/ijcai.2017/299.

Yiqing Lin, Jianheng Tang, Chenyi Zi, H. Vicky Zhao, Yuan Yao, and Jia Li. Unigad: unifying multi-level graph anomaly detection. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

Fanzhen Liu, Xiaoxiao Ma, Jia Wu, Jian Yang, Shan Xue, Amin Beheshti, Chuan Zhou, Hao Peng, Quan Z Sheng, and Charu C Aggarwal. Dagad: Data augmentation for graph anomaly detection. In *2022 IEEE international conference on data mining (ICDM)*, pp. 259–268. IEEE, 2022.

Yixin Liu, Zhao Li, Shirui Pan, Chen Gong, Chuan Zhou, and George Karypis. Anomaly detection on attributed networks via contrastive self-supervised learning. *IEEE transactions on neural networks and learning systems*, 33(6):2378–2392, 2021.

Yixin Liu, Shiyuan Li, Yu Zheng, Qingfeng Chen, Chengqi Zhang, and Shirui Pan. Arc: A generalist graph anomaly detector with in-context learning. In *Advances in Neural Information Processing Systems*, 2024.

Andrew Kachites McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. Automating the construction of internet portals with machine learning. *Information Retrieval*, 3(2):127–163, 2000.

Péter Mernyei and Cătălina Cangea. Wiki-cs: A wikipedia-based benchmark for graph neural networks. *arXiv preprint arXiv:2007.02901*, 2020.

Junjun Pan, Yixin Liu, Xin Zheng, Yizhen Zheng, Alan Wee-Chung Liew, Fuyi Li, and Shirui Pan. A label-free heterophily-guided approach for unsupervised graph fraud detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 12443–12451, 2025.

Junjun Pan, Yixin Liu, Yu Zheng, Lianhua Chi, Alan Wee-Chung Liew, and Shirui Pan. Camera: Adapting to semantic camouflage in unsupervised text-attributed graph fraud detection. In *International Joint Conference on Artificial Intelligence*, 2026.

Zhen Peng, Minnan Luo, Jundong Li, Huan Liu, Qinghua Zheng, et al. Anomalous: A joint modeling approach for anomaly detection on attributed networks. In *Ijcai*, volume 18, pp. 3513–3519, 2018.

Hezhe Qiao and Guansong Pang. Truncated affinity maximization: One-class homophily modeling for graph anomaly detection. In *Advances in Neural Information Processing Systems*, 2023.

Amit Roy, Juan Shu, Jia Li, Carl Yang, Olivier Elshocht, Jeroen Smeets, and Pan Li. Gad-nr : Graph anomaly detection via neighborhood reconstruction. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, 2024.

Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Gallagher, and Tina Eliassi-Rad. Collective classification in network data. *AI Mag.*, 29(3):93–106, September 2008. ISSN 0738-4602. doi: 10.1609/aimag.v29i3.2157. URL https://doi.org/10.1609/aimag.v29i3.2157.

Xuhong Wang, Bo Jin, Yao Du, Ping Cui, Yingshui Tan, and Yingshen Yang. One-class graph neural networks for anomaly detection in attributed networks. *Neural Computing and Applications*, 2021.

Haoyan Xu, Ruizhi Qian, Zhengtao Yao, Ziyi Liu, Li Li, Yuqi Li, Yanshu Li, Wenqing Zheng, Daniele Rosa, Daniel Barcklow, et al. Llm-powered text-attributed graph anomaly detection via retrieval-augmented reasoning. *arXiv preprint arXiv:2511.17584*, 2025a.

Xiaowei Xu, Nurcan Yuruk, Zhidan Feng, and Thomas A. J. Schweiger. Scan: a structural clustering algorithm for networks. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '07, pp. 824–833, New York, NY, USA, 2007. Association for Computing Machinery. ISBN 9781595936097. doi: 10.1145/1281192.1281280. URL https://doi.org/10.1145/1281192.1281280.

Yiming Xu, Jiarun Chen, Zhen Peng, Zihan Chen, Qika Lin, Lan Ma, Bin Shi, and Bo Dong. Court of llms: Evidence-augmented generation via multi-llm collaboration for text-attributed graph anomaly detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 2437–2446, 2025b.

Yiming Xu, Xu Hua, Zhen Peng, Bin Shi, Jiarun Chen, Xingbo Fu, Song Wang, and Bo Dong. Text-attributed graph anomaly detection via multi-scale cross-and uni-modal contrastive learning. *arXiv preprint arXiv:2508.00513*, 2025c.

Zhiming Xu, Xiao Huang, Yue Zhao, Yushun Dong, and Jundong Li. Contrastive attributed network anomaly detection with data augmentation. In *Pacific-Asia conference on knowledge discovery and data mining*, pp. 444–457. Springer, 2022.

Jianke Yu, Hanchen Wang, Xiaoyang Wang, Zhao Li, Lu Qin, Wenjie Zhang, Jian Liao, and Ying Zhang. Group-based fraud detection network on e-commerce platforms. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, pp. 5463–5475, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701030. doi: 10.1145/3580305.3599836. URL https://doi.org/10.1145/3580305.3599836.

Xiong Zhang, Hong Peng, Zhenli He, Cheng Xie, Xin Jin, and Hua Jiang. Gctam: Global and contextual truncated affinity combined maximization model for unsupervised graph anomaly detection. In *International Joint Conference on Artificial Intelligence*, 2025.

Yu Zheng, Ming Jin, Yixin Liu, Lianhua Chi, Khoa T. Phan, and Yi-Ping Phoebe Chen. Generative and contrastive self-supervised learning for graph anomaly detection. *IEEE Transactions on Knowledge and Data Engineering*, 2022. doi: 10.1109/TKDE.2021.3134636.

A AI USE STATEMENT

In this work, we used generative AI tools for literature search, code generation, drafting parts of the research paper, and editing the paper to improve readability, brainstorming, sourcing/searching for information, suggest a structure for a research paper,. We have not used generative AI tools for generating synthetic data sets, developing theoretical models. Additionally, we used generative AI tools for suggesting experimental parameters, generating code for creating scientific figures.

We have reviewed all AI-assisted work. Specifically, all AI-generated code was rigorously tested, debugged, and validated by the authors to ensure correctness and reproducibility. The AI-assisted draft and polished text were thoroughly reviewed and revised by all authors to ensure technical accuracy, logical coherence, and adherence to academic standards. We take responsibility for the final content of this work, including text, claims, code, or artifacts produced with the aid of generative AI.

B REPRODUCIBILITY STATEMENT

The experimental setup is summarized in Section 5.1. The Appendix E.1 provides detailed configuration and preprocessing information. The dataset summary are provided in Table 4. We have released the source code at https://anonymous.4open.science/r/CARN-4FB9.

C INITIALIZATION OF THE DUAL-VIEW CLUSTERING HEADS

Directly optimizing the dual-view soft clustering module may result in uninformative uniform assignments or early cluster collapse. We therefore initialize the structure and attribute clustering heads using K-means centroids estimated from $\mathbf{Z}$ and $\mathbf{X}$, respectively.

Given a centroid $\mathbf{c}_k$, the logit associated with community k is initialized as

$$\ell_k(\mathbf{u}) = \mathbf{c}_k^\top \mathbf{u} - \frac{1}{2}\|\mathbf{c}_k\|_2^2. \tag{18}$$

This parameterization follows from

$$-\frac{1}{2}\|\mathbf{u} - \mathbf{c}_k\|_2^2 = \mathbf{c}_k^\top \mathbf{u} - \frac{1}{2}\|\mathbf{c}_k\|_2^2 - \frac{1}{2}\|\mathbf{u}\|_2^2. \tag{19}$$

Because the final term is independent of k, applying a softmax over the resulting logits approximates nearest-centroid assignment. Accordingly, the clustering-head parameters are initialized as

$$\mathbf{W}_k \leftarrow \mathbf{c}_k^\top, \qquad b_k \leftarrow -\frac{1}{2}\|\mathbf{c}_k\|_2^2, \tag{20}$$

and remain trainable during subsequent optimization. We further anneal τ_c during training so that the initially smooth assignments become progressively sharper.

Table 4: Statistics of the Experimental Datasets.

Dataset	Nodes	Edges	Avg. Length	Ano. Count	Ano. Ratio
Cora	2,708	5,879	132.6	208	7.68%
Citeseer	3,186	4,727	150.8	400	12.56%
Pubmed	19,717	45,685	241.2	1,088	5.52%
Wikics	11,701	216,458	414.2	1,089	9.31%
Arxiv	169,343	1,161,174	170.4	6,750	3.98%

D Detailed Learning Objectives

This section provides the detailed definitions of the loss terms used in Eq. equation 16.

D.1 Community-Regularization Objectives

Let $\mathbf{A}$ denote the adjacency matrix, $\mathbf{d}$ the degree vector, and $m = |\mathcal{E}|$. The structure-view assignments are regularized through the soft modularity objective

$$\mathcal{L}_{mod} = -\frac{1}{2m} \operatorname{Tr}\left[(\mathbf{H}^s)^\top \left(\mathbf{A} - \frac{\mathbf{d}\mathbf{d}^\top}{2m} \right) \mathbf{H}^s \right]. \tag{21}$$

In implementation, this objective is evaluated using sparse edge operations without explicitly constructing the dense modularity matrix.

To prevent all nodes from being assigned to a single community, we employ the balance regularizer

$$\mathcal{L}_{col} = \frac{\sqrt{K}}{N} \left\| \sum_{i=1}^{N} \mathbf{h}_i^s \right\|_2 - 1. \tag{22}$$

D.2 Intervention-Sensitivity Objective

The intervention objective encourages stability on observed nodes while requiring synthetic Semantic anomalies to exhibit a CCD larger than the margin m_{cf}:

$$\mathcal{L}_{cf} = \frac{1}{N} \sum_{i=1}^{N} d_i^{cf} + \frac{1}{|\mathcal{V}_{syn}|} \sum_{i \in \mathcal{V}_{syn}} \max\left(0, m_{cf} - \widetilde{s}_i^{cg}\right). \tag{23}$$

Because community assignments are unreliable during early optimization, this objective is gradually activated using the warm-up coefficient $r(t) \in [0, 1]$.

E More Experimental Setup

E.1 Datasets details

We evaluate CARNet on five text-attributed graphs derived from citation and social networks.

- Cora (McCallum et al., 2000) is a citation network of machine learning papers spanning seven research topics, with nodes representing papers and edges indicating citation relationships.
- CiteSeer (Giles et al., 1998) contains computer science publications from six research categories, connected through citation links.
- PubMed (Sen et al., 2008) is a citation network of biomedical publications on diabetes, organized into three research categories.
- Wikics (Mernyei & Cangea, 2020) consists of Wikipedia articles covering ten computer science categories, with edges representing hyperlinks between articles.
- Arxiv (Hu et al., 2020) is a large-scale citation network of arXiv papers spanning 40 computer science subject areas, with textual information derived from paper titles and abstracts.

We construct two types of synthetic anomalies for each dataset, namely semantic anomalies and structural anomalies. The statistics of the resulting datasets are summarized in Table 4 and the detailed construction procedures are described as follows.

Table 5: Overall LLM Anomaly Detection Performance (%).

	Cora		Citeseer		Pubmed		Wikics		Arxiv	
Method	**AUC**	**AP**	**AUC**	**AP**	**AUC**	**AP**	**AUC**	**AP**	**AUC**	**AP**
CARNet$_{Qwen-3.8-Max}$	91.17 ± 0.56	58.26 ± 1.41	89.88 ± 0.20	67.89 ± 0.81	90.71 ± 0.13	56.01 ± 0.25	86.24 ± 0.66	44.36 ± 1.92	93.77 ± 1.02	44.04 ± 1.07
CARNet$_{Deepseek-V4-Pro}$	89.48 ± 0.23	53.56 ± 0.86	86.86 ± 0.39	63.49 ± 0.26	84.45 ± 0.29	42.22 ± 0.75	79.10 ± 0.60	36.54 ± 1.69	90.65 ± 0.30	39.32 ± 1.56
CARNet$_{Gemini-3.8}$	90.78 ± 0.31	59.09 ± 0.77	90.05 ± 0.34	65.76 ± 0.21	88.30 ± 0.54	47.44 ± 0.88	84.30 ± 1.47	40.77 ± 2.97	96.32 ± 0.85	50.29 ± 1.20
CARNet$_{Claude-Opus-5}$	90.34 ± 0.54	57.58 ± 0.26	89.23 ± 0.56	67.15 ± 1.36	91.48 ± 0.32	56.44 ± 1.09	85.16 ± 0.51	42.33 ± 1.46	94.19 ± 0.66	49.42 ± 1.27
CARNet$_{GPT-5.6-Sol}$	92.87 ± 0.50	64.58 ± 1.78	90.09 ± 0.29	66.41 ± 1.14	89.57 ± 0.24	52.88 ± 1.18	85.48 ± 0.32	46.93 ± 1.40	95.50 ± 0.30	48.01 ± 2.58

Semantics anomaly. Since the original datasets lack semantic anomaly annotations, we construct synthetic semantics anomalies by modifying node texts while preserving graph topology, following Xu et al. (2025a). With a fixed random seed, we sample n nodes from the current set of normal nodes. For each selected node v_i, we initialize a candidate category set by excluding its original category. We then sequentially examine neighborhoods at exact hop distances $h = 1, \ldots, k$, progressively eliminating categories that frequently occur at each distance. This procedure terminates when a single candidate remains or all prescribed hop distances have been considered. If multiple candidates remain, we sample the target category with probability proportional to its global frequency. This selection strategy favors categories that are plausible at the dataset level but poorly aligned with the node's locals context.

We employ LLM to rewrite the selected texts using dataset-specific prompts. Each prompt includes the original text, its category, and the target category, and instructs the model to generate content aligned with the target category. The rewritten text replaces the original textual attribute, while the node identity, incident edges, and original category annotation remain unchanged. We subsequently encode all resulting texts with the same text encoder to obtain the updated node-feature matrix. Ground-truth categories are used exclusively for offline benchmark construction and are withheld from the anomaly detectors during training.

Structural anomaly. Since the original datasets do not provide structural anomaly annotations, we construct synthetic structural anomalies by perturbing graph topology while preserving node attributes. Specifically, with a fixed random seed, we sample $n \times m$ nodes exclusively from the current set of normal nodes, thereby preventing overlap with anomalies introduced by other generation procedures. The sampled nodes are randomly partitioned into n disjoint groups, each containing m nodes. For each group, we construct a fully connected subgraph by adding edges between every pair of nodes in both directions, resulting in n artificially dense cliques. Consequently, each generated clique contains $\binom{m}{2}$ undirected connections, and all $n \times m$ participating nodes are labeled as structural anomalies.

Implementation details. We use **GPT-5.6-sol** for text rewriting, with the prompt provided in Figure 5. The resulting texts are encoded using the **all-MiniLM-L6-v2** sentence transformer. Local experiments are conducted on a workstation equipped with an Intel Core i9-13900KF CPU, an NVIDIA GeForce RTX 4090 GPU with 24 GB of memory, and 64 GB of RAM. CARNet is trained for 300 epochs with a learning rate of $\{10^{-3}, 5 \times 10^{-4}\}$ and weight decay of 10^{-5}. We use a two-layer GNN and transformer with a hidden dimension of 64 and 4 attention heads. The number of latent context prototypes is set to $K = \{32, 64, 128\}$, and we sample $\{4, 8, 16, 32\}$ neighbors per node.

Sensitivity to the anomaly-generation LLM. To examine the sensitivity of CARNet to the anomaly-generation LLM, we evaluate it on datasets constructed using five different models. As shown in Table 5, CARNet achieves broadly comparable performance across datasets generated by Qwen, Gemini, Claude, and GPT, particularly in AUC, suggesting limited sensitivity to the generator among these models. In contrast, DeepSeek-generated datasets generally yield lower scores. This discrepancy reflects differences in the quality or semantic clarity of the generated anomalies, further inspection is needed to verify this explanation.

Instruction:

SYSTEM PROMPT: You are a TAG (Text-Attributed Graph) anomaly text generator. Your task is to generate semantics anomalies by partly rewriting node text attributes. Strictly adhere to all provided constraints and requirements.

USER PROMPT:
You are given a normal node from the {Dataset}. The {Dataset} contains Machine Learning publications categorized into {N} topics. Each node's text attribute is the paper title and, optionally, an abstract.
Original node label/topic: ({label name})
Designated target label/topic: ({designated label})
Original text attribute: {raw text}
Task: Rewrite the text attribute so it clearly concerns the designated label/topic: ({designated label}) while remaining a semantics anomaly relative to the original.

Figure 5: Rewriting Prompt for Dataset Construction.

E.2 Baseline Descriptions.

- **ANOMALOUS** (Peng et al., 2018) jointly performs attribute selection and anomaly detection through CUR matrix decomposition and residual analysis. By filtering structurally irrelevant or noisy attributes, it identifies nodes whose selected attribute patterns are inconsistent with the network structure.
- **DOMINANT** (Ding et al., 2019) employs a graph convolutional autoencoder to jointly reconstruct the adjacency matrix and node attributes. Its anomaly score combines structural and attribute reconstruction errors, with nodes that cannot be accurately reconstructed receiving higher abnormality scores.
- **AnomalyDAE** (Fan et al., 2020) uses structure and attribute autoencoders to learn node and attribute embeddings jointly. An attention-based structural encoder and a cross-modal attribute decoder model interactions between topology and attributes, while their reconstruction errors determine node anomalousness.
- **CoLA** (Liu et al., 2021) formulates graph anomaly detection as contrastive self-supervised learning. It contrasts each target node with sampled local subgraphs and estimates abnormality from their agreement. Nodes that are repeatedly inconsistent with their surrounding substructures receive higher anomaly scores.
- **CONAD** (Xu et al., 2022) combines graph data augmentation, contrastive representation learning, and autoencoder-based reconstruction. It constructs diverse anomalous views by perturbing node attributes and graph structure, and learns anomaly-sensitive embeddings by contrasting the original and augmented attributed graphs.
- **GCTAM** (Zhang et al., 2025) extends truncated affinity maximization by jointly considering contextual and global affinity. Its context-aware truncation suppresses affinities associated with suspicious nodes, whereas global truncation strengthens dominant normal relationships, avoiding the rigid node-independent thresholds used in conventional TAM.
- **GAD-NR** (Roy et al., 2024) replaces direct adjacency reconstruction with complete neighborhood reconstruction. From each node representation, it reconstructs the node degree, self-attributes, and the distribution of neighboring attributes, allowing its reconstruction score to capture contextual, structural, and joint anomalies.
- **ADA-GAD** (He et al., 2024) mitigates anomaly overfitting in graph autoencoders through a two-stage denoising framework. It first pretrains autoencoders on multi-level anomaly-reduced graph views and then retrains the decoders on the original graph with anomaly-distribution regularization.
- **GRADATE** (Duan et al., 2023) develops a multi-view, multi-scale contrastive framework over the original and structurally augmented graphs. It jointly performs node–node, node–subgraph, and subgraph–subgraph contrast, thereby capturing complementary anomaly signals at different structural granularities.

Table 6: Overall LLM Anomaly Detection Performance (%). The Best (**Bold**) and Suboptimal (Underlined) Results are Marked, Respectively.

	Cora		Citeseer		Pubmed		Wikics	
Method	**AUC**	**AP**	**AUC**	**AP**	**AUC**	**AP**	**AUC**	**AP**
Qwen3.8-27b	89.53 ± 1.35	45.25 ± 1.40	88.14 ± 1.54	50.73 ± 1.75	76.30 ± 1.22	32.62 ± 0.68	73.64 ± 1.16	31.25 ± 1.65
Qwen3.8-flash	84.81 ± 1.50	41.44 ± 1.81	86.39 ± 1.68	50.94 ± 1.52	74.80 ± 1.50	30.80 ± 2.36	80.44 ± 1.19	38.18 ± 1.04
DeepSeek-V4-Flash	79.07 ± 2.21	23.24 ± 2.66	82.63 ± 1.13	40.30 ± 1.24	76.05 ± 1.88	23.33 ± 2.51	83.46 ± 1.26	42.43 ± 1.47
GLM-5	72.87 ± 1.72	15.18 ± 3.53	81.04 ± 1.25	32.71 ± 2.04	80.55 ± 1.26	31.71 ± 1.54	82.23 ± 1.84	43.59 ± 1.06
GPT-5.6-Luna	81.16 ± 1.38	42.28 ± 1.22	86.31 ± 0.71	52.96 ± 0.63	78.44 ± 1.09	34.31 ± 1.12	**86.18 ± 0.79**	**47.69 ± 1.29**
Claude-sonnet-5	92.54 ± 0.48	63.91 ± 1.06	85.37 ± 0.49	60.83 ± 0.95	81.96 ± 1.47	45.63 ± 1.39	80.15 ± 1.53	45.32 ± 1.94
Gemini-3.5-lite	71.68 ± 2.55	28.60 ± 1.78	77.44 ± 1.84	36.41 ± 2.20	65.94 ± 2.45	19.73 ± 3.38	79.90 ± 1.52	39.68 ± 1.91
Gemini-3.7-flash	81.05 ± 1.63	55.56 ± 1.39	90.04 ± 1.19	63.02 ± 1.31	68.33 ± 1.70	34.91 ± 2.25	81.56 ± 1.30	43.59 ± 1.38
CARNet	**92.87 ± 0.50**	**64.58 ± 1.78**	**90.09 ± 0.29**	**64.63 ± 1.14**	**89.57 ± 0.24**	**52.88 ± 1.18**	85.48 ± 0.32	46.93 ± 1.40

- **DAGAD** (Liu et al., 2022) addresses anomalous sample scarcity and class imbalance in attributed graphs through data augmentation. It combines GNN-based representation learning with generated training samples and an imbalance-tailored learning module to distinguish anomalous nodes from the normal majority.
- **HUGE-GAD** (Pan et al., 2025) performs label-free graph fraud detection by estimating node-level heterophily with the HALO metric. A joint MLP–GNN architecture is trained using ranking and asymmetric alignment objectives to preserve structural evidence while alleviating excessive feature smoothing.
- **CMUCL** (Xu et al., 2025c) jointly trains text and graph encoders for end-to-end anomaly detection on text-attributed graphs. It exploits cross-modal and unimodal multi-scale consistency to learn anomaly-relevant representations and derives node-level anomaly scores through inconsistency mining.
- **CoLL** (Xu et al., 2025b) combines LLM-based semantic reasoning with graph representation learning for text-attributed graph anomaly detection. Multiple LLMs collaboratively generate anomaly-related evidence and interpretable rationales, while a gated GNN adaptively integrates this evidence with textual and high-order structural information.
- **CAMERA** (Pan et al., 2026) targets semantic camouflage in unsupervised text-attributed graph fraud detection. Its ego-decoupled mixture-of-experts captures graph, semantic, and global deviation cues, while a context-informed gating network adaptively combines these cues according to each node and its neighborhood.

F More Experimental Results

F.1 LLMs results on Semantic Anomaly Detection

Given the inherent strengths of LLMs in semantic understanding, we further compare their performance with CARNet on semantic anomaly detection. Table 6 compares CARNet with eight LLMs on four datasets. We exclude Arxiv dataset from this comparison because its large scale would incur prohibitively high LLM inference costs.

CARNet achieves the highest mean AUC and AP on three of the four datasets. Its advantage is most pronounced on PubMed, exceeding the strongest LLM results by 7.61% in AUC and 7.25% in AP. On Cora and CiteSeer, CARNet achieves smaller gains, improving upon the best LLM results by 0.33% and 0.05% in AUC, and 0.67% and 1.61% in AP, respectively. These results suggest that anomalies generated using LLMs are not necessarily readily detectable by other LLMs and support explicitly modeling the compatibility between node semantics and context.

LLM performance also varies across datasets. Claude-sonnet-5 is the strongest LLM on Cora and PubMed, while Gemini-3.7-flash performs best among LLMs on CiteSeer. On WikiCS, GPT-5.6-Luna surpasses CARNet by 0.70% in AUC and 0.76% in AP. Overall, CARNet ranks first or second across all evaluated datasets and metrics, while no single LLM consistently performs best.

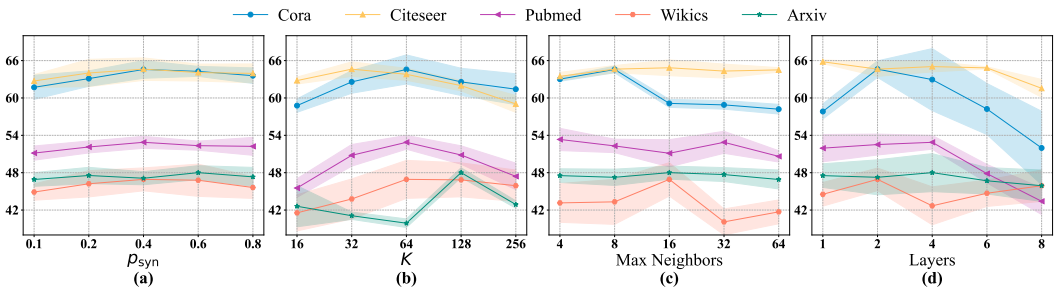

Figure 6: Hyperparameter Sensitivity of CARNet in Terms of AP on Different Datasets.

F.2 Additional Parameter Sensitivity Analysis

The AP results corresponding to the hyperparameter analysis in the main text are presented in Figure 6. We further examine sensitivity to the learning rate, number of training epochs, λ_{con}, and $\lambda_{\text{ctx_cf}}$ across five datasets. Figures 7 and 8 report the AUC and AP results across repeated runs, respectively.

Moderate learning rates of 10^{-3} and 5×10^{-4} provide consistently strong performance. Larger or smaller values generally degrade performance, particularly on Citeseer, Pubmed, and Wikics, indicating the importance of an appropriate optimization step size. Performance improves substantially from 50 to 100–150 epochs, while further training yields dataset-dependent effects. Cora declines after 100 epochs, whereas Citeseer and Pubmed remain relatively stable. Wikics and Arxiv can benefit from longer training, particularly in AP. Increasing λ_{con} generally improves performance, with clearer trends on Cora, Wikics, and Arxiv. Despite some fluctuations on Citeseer and Pubmed, values of 0.8–1.0 provide strong results across datasets. The preferred weight varies across datasets: Cora favors 0.8, PubMed and Wikics favor 0.4, and Arxiv generally benefits from larger values. Citeseer exhibits stable AP but greater variation in AUC. These trends suggest that the intervened context reconstruction objective requires dataset-specific weighting, although several differences between nearby settings are small relative to the reported standard deviations.

F.3 Computational Complexity and Efficiency Analysis

Let N and E denote the numbers of nodes and edges, d the representation dimension, K the number of latent contexts, and M the neighbor sampling budget. With fixed encoder depths, graph encoding costs $\mathcal{O}(Ed + Nd^2)$, while dual-view identity inference and global context construction cost $\mathcal{O}(NKd)$. The factual and intervened reasoning passes cost $\mathcal{O}(N(Td^2 + T^2d))$, where $T = M + 3$. With at most B donor candidates per selected node, synthetic supervision adds $\mathcal{O}(p_{\text{syn}}N(BK + d))$. With fixed encoder depths and donor sampling budgets, the per-epoch time complexity simplifies to

$$\mathcal{O}\big(Ed + Nd(K + Td + T^2) + \mathcal{C}_{\text{reg}}\big)\,,$$

where $T = M + 3$. For fixed model and sampling parameters, this scales linearly with $N + E$, provided that the auxiliary regularization also has linear cost.

We further compare runtime and memory usage on PubMed and Arxiv, as shown in Figure 9. On PubMed, CARNet requires less runtime and memory than CAMERA, although lighter methods incur lower computational costs. On Arxiv, CARNet runs substantially faster than CAMERA, CoLL, and HUGE-GAD while also using less memory. These results indicate that the additional context reasoning and intervention operations remain computationally manageable, supporting the applicability of CARNet to larger graphs.

F.4 Visual Comparison of Anomaly Scores

We visually compare the anomaly scores produced by CARNet and several baseline methods on Cora and WikiCS, as shown in Figure 10 and Figure 11. For each dataset, we project the original node features into 2-dimensions using t-SNE and use the same layout to display the ground-truth anomaly labels and each method's predicted scores. Lighter colors indicate scores closer to 1, corresponding to greater predicted anomalousness.

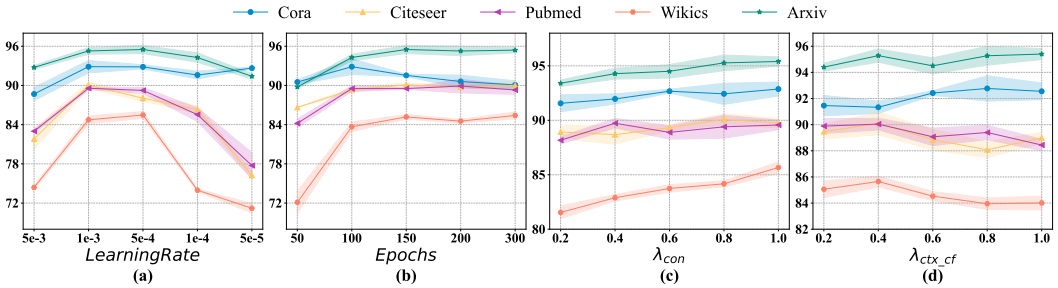

Figure 7: Additional Hyperparameter Sensitivity Analysis of CARNet in Terms of AUC Across Datasets.

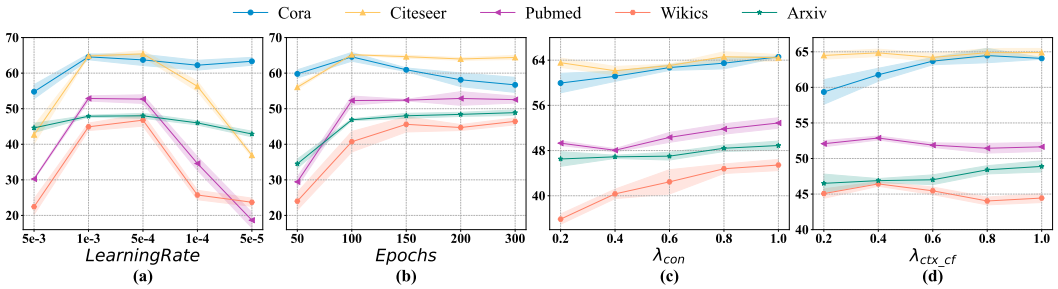

Figure 8: Additional Hyperparameter Sensitivity Analysis of CARNet in Terms of AP Across Datasets.

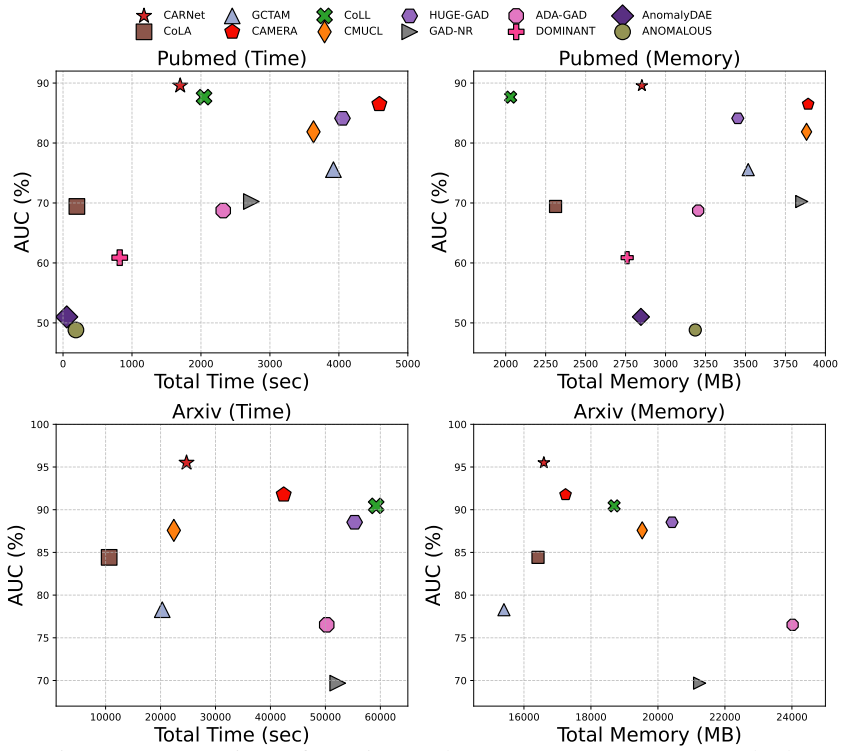

Figure 9: Comparison of Runtime and Memory Usage Across Methods.

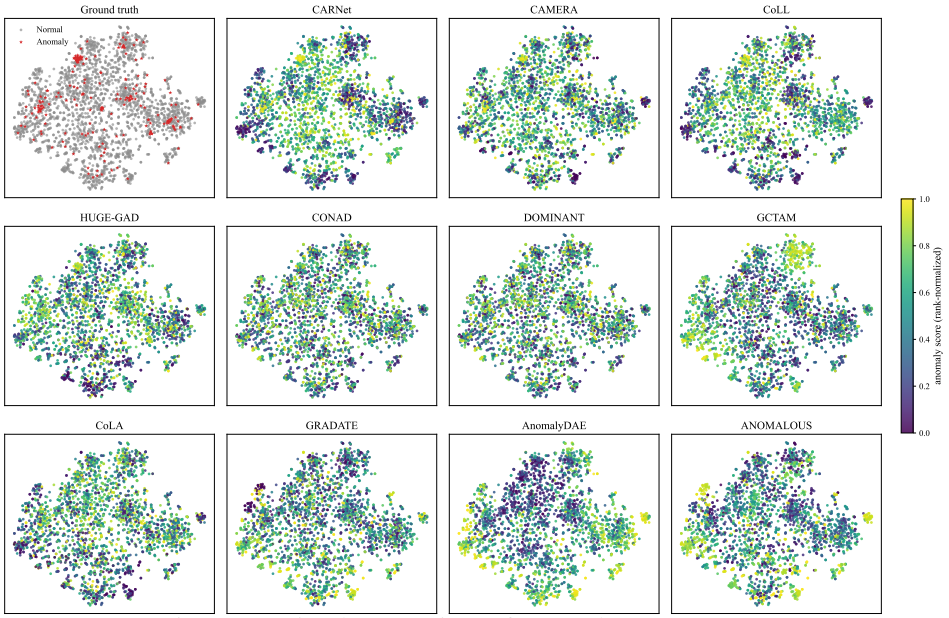

Figure 10: Visual Comparison of Anomaly Scores on Cora.

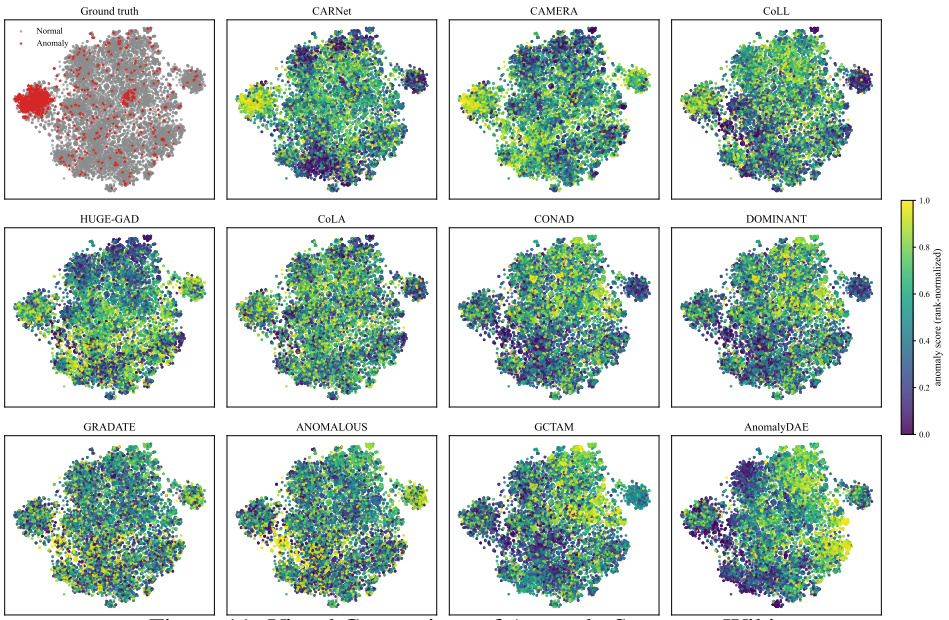

Figure 11: Visual Comparison of Anomaly Scores on Wikics.

On WikiCS, anomalous nodes are concentrated mainly in the left half of the visualization, where CARNet likewise assigns higher scores, indicated by lighter colors. On Cora, anomalies are more dispersed, yet regions with relatively dense anomalies also tend to receive higher scores. This correspondence supports CARNet's effectiveness in identifying anomalies with different spatial distributions in the visualizations. CAMERA and CoLL also show good alignment with the ground-truth anomaly distributions, consistent with their competitive performance.