ANTAGAD: GENERALIST TEXT-ATTRIBUTED GRAPH ANOMALY DETECTION WITH TEXTUAL AND TOPOLOGICAL ANCHORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Representing the rich textual semantics of text-attributed graphs (TAGs) with large language models markedly improves cross-domain generalization, enabling effective anomaly detection across domains as varied as social platforms, financial transaction systems, and multi-agent collaboration environments. Concretely, this generalizability allows detectors to recall a large fraction of anomalies across domains, as reflected in high AUROC. However, this apparent success is undermined by persistently low AUPRC, meaning many flagged nodes are false positives and individual predictions carry unreliable confidence, leading to unacceptable misjudgments in practical deployment. We trace this failure to the widespread reliance on generalizable prototype representations, which enhance cross-domain adaptation at the expense of instance-specific detail and thus blur the distinctions among individual anomalies. We term this the Over-Generalization ($\mathcal{OG}$) problem. To address this, we propose a novel multi-anchor representation scheme to replace prototypes, encoding each node against a set of high-confidence anomaly and normal anchors to improve generalization while preserving the instance-specific characteristics of individual anomalies. Extensive experiments on eight real-world cross-domain datasets show that our method consistently outperforms all competitive baselines, with substantial gains in both AUPRC and AUROC. Beyond these empirical gains, we provide a theoretical characterization of sufficient conditions for controlling $\mathcal{OG}$ together with matched-coverage empirical analyses showing reduced anomaly acceptance on most evaluated settings.

1 INTRODUCTION

Real-world systems often exhibit intricate interactions intertwined with rich semantic information, motivating the emergence of text-attributed graphs (TAGs) Zhang et al. (2025c) as a unified representation paradigm that integrates textual descriptions with relational structures. Such graphs appear widely in security- and risk-sensitive applications, including social platform governance, financial transaction monitoring, cyber threat identification, and multi-agent collaboration networks Liu et al. (2021); Ma et al. (2021); Qiao et al. (2025b), where anomaly nodes often manifest as subtle deviations in either textual semantics or structural behavior Akoglu et al. (2015); Ding et al. (2019). Unlike conventional graph anomaly detection (GAD), for which topological irregularities are the primary focus, TAG anomaly detection requires jointly understanding textual contexts and graph dependencies Yang et al. (2021); Zhao et al. (2022) to identify different anomaly patterns Xu et al. (2025c). Recent advances in LLMs enable richer semantic representations for TAG anomaly detection. By capturing fine-grained textual cues and contextual knowledge, LLM-enhanced representations improve overall anomaly detection and cross-domain transferability Brown et al. (2020); Xu et al. (2025b;a). This capability is considered essential for generalist TAG anomaly detection, with unified detectors being enabled to identify anomalies in unseen domains without dataset-specific training. Therefore, LLM-driven semantic understanding offers a promising direction for scalable and transferable anomaly detection.

Despite progress in cross-domain anomaly ranking, a detector may still accept anomalous nodes as normal at an operating point that retains most normal nodes. We study one possible cause: reference aggregation can produce a normal acceptance region that includes low-density gaps between

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.

distinct normal modes. We refer to this mechanism as over-generalization ($\mathcal{OG}$). Its operational diagnostic is the anomaly acceptance rate at matched normal-node coverage. A high AUROC and a low AUPRC alone do not establish this mechanism, because they summarize different aspects of ranking and AUPRC also depends on anomaly prevalence. Prototype-based methods, including ProMoS Xu et al. (2026), can use multiple references, and IA-GGAD Zhang et al. (2025a) maintains a multi-vector invariant feature pool. Their decision regions therefore need not be a single convex neighborhood. Our analysis identifies conditions under which coarse acceptance regions admit avoidable anomalies; controlled comparisons are needed to attribute an observed $\mathcal{OG}$ difference to reference aggregation. Figure 1 provides qualitative motivation.

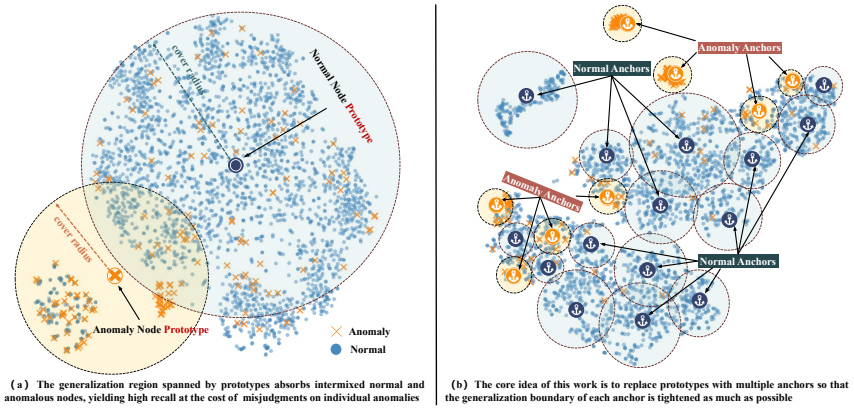


Figure 1: Qualitative t-SNE visualizations on Pubmed: (a) ProMoS Xu et al. (2026); (b) **AnTAGAD**. $\mathcal{OG}$ is evaluated using the actual anomaly scores at matched normal-node coverage, not by measuring regions in this two-dimensional projection.

To alleviate the Over-Generalization problem, we propose **AnTAGAD**, a generalist detector that retains selected textual and topological instances as normal and anomaly references. At inference, a node is scored by the difference between its top-k similarities to the two source-domain anchor banks, together with neighborhood dispersion. This design allows local reference selection without collapsing the entire bank into one mean. Its benefit depends on anchor coverage, class separation, and domain shift. We analyze these conditions for the actual scoring function and evaluate detection performance on eight unseen target graphs.

Contributions. (1) We observed $\mathcal{OG}$ phenomenon that affects TAG anomaly detection precision during cross-domain. (2) We formalize the $\mathcal{OG}$ problem and propose **AnTAGAD** that significantly increases detection effectiveness both in precision (AUPRC) and recall (AUROC).

2 Related Work

2.1 Graph Anomaly Detection

Graph Anomaly Detection (GAD) aims to identify nodes whose attributes or structural behaviors deviate from the dominant graph patterns Tang et al. (2022); Luo et al. (2022); Gao et al. (2023b). Early studies mainly identify anomalies from reconstruction errors or abnormal representations, including methods that reconstruct node attributes and topology Ding et al. (2019); Fan et al. (2020) or preserve irregular patterns during embedding and community modeling Bandyopadhyay et al. (2019; 2020); Luo et al. (2022). Subsequent self-supervised approaches, such as CoLA Liu et al. (2021) and HCM-A Huang et al. (2022), capture local node-context inconsistencies without relying on anomaly labels. Recent advances further exploit the distinctive spectral characteristics of anomalous nodes: BWGNN Tang et al. (2022) characterizes anomaly-induced graph-frequency patterns, and GHRN Gao et al. (2023a) further accounts for anomaly-related heterophily, while TAM Qiao & Pang (2023) and GCTAM Zhang et al. (2025b) suppress suspicious low-affinity propagation. Semi-GGAD Qiao et al. (2024) and CAGAD Xiao et al. (2024) additionally improve robustness by introducing synthetic anomalies. Despite their effectiveness on conventional attributed graphs, many

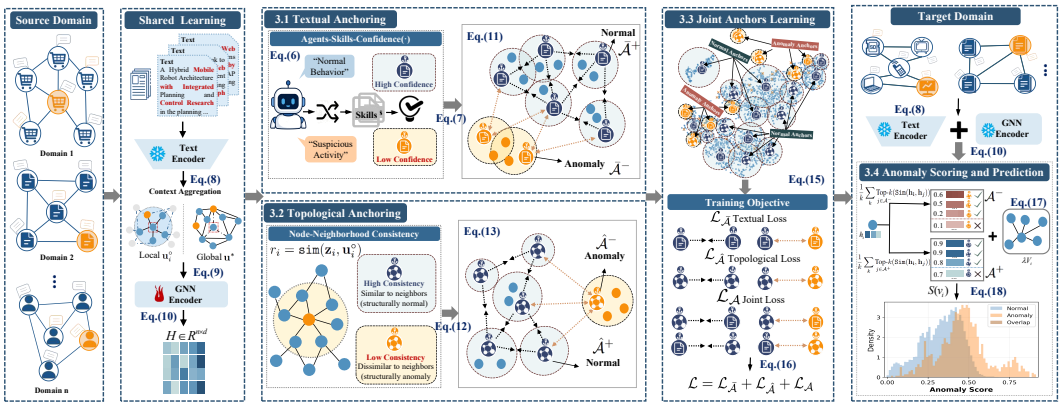


Figure 2: The overall framework of **AnTAGAD**.

of these methods are primarily designed for fixed numerical node attributes and typically follow a graph-specific training paradigm. Consequently, they not only require retraining when transferred to unseen graphs, but also fail to fully capture the fine-grained textual information and text-induced contextual inconsistency inherent in Text-Attributed Graphs.

2.2 Generalist Graph Anomaly Detection

Generalist graph anomaly detection learns transferable scoring rules for unseen graphs. ARC Liu et al. (2024) and UNPrompt Niu et al. (2025) study cross-graph transfer, while AnomalyGFM Qiao et al. (2025a) learns graph-agnostic normal and abnormal prototypes. IA-GGAD Zhang et al. (2025a) combines invariant references with affinity learning, and OWLEYE Zheng et al. (2026) studies transferable anomaly patterns. ProMoS Xu et al. (2026) distills normality priors through multiple learnable prototypes; DR-GGAD Fu et al. (2026) addresses cross-domain anomaly non-discriminativity. These approaches differ in their supervision, reference construction, and adaptation protocols. Our comparison focuses on how reference aggregation and local retrieval affect anomaly acceptance at matched normal coverage. It includes multiple-prototype controls rather than assuming that every prototype method uses a single center.

2.3 Text-Attributed Graph Anomaly Detection

Text-attributed Graph Anomaly Detection (TAGAD) detects anomalous nodes by integrating textual semantics with graph structures. Conventional text embeddings, however, may fail to detect subtle anomaly indicators. This has led to recent research focusing on contrastive learning and large language models. CMUCL Xu et al. (2025c) jointly models textual and structural information through multi-scale contrastive learning. CoLL Xu et al. (2025b) incorporates LLM-generated evidence with GNNs to enhance semantic anomaly understanding, while TAGAD Anonymous (2025) jointly exploits contextual, semantic, and structural features for global and local anomaly detection.

Our study focuses on the effect of reference construction on anomaly acceptance under cross-domain transfer. In particular, we examine when selected instance references preserve local distinctions that may be lost through coarse aggregation, while allowing that sufficiently expressive multi-prototype references can preserve such distinctions as well.

3 Methodology

To mitigate Over-Generalization ($\mathcal{OG}$), **AnTAGAD** uses instance-level textual and topological anchors as normal and anomaly references. The two anchor sources are jointly integrated for representation learning and anomaly scoring. Specifically, we construct textual anchors from semantic reliability, identify topological anchors via node-neighborhood consistency, and merge them into unified anchor banks. During inference, local Top-k anchors and neighborhood dispersion are used for scoring. The overall framework is shown in Figure 2.

Preliminaries. We consider a text-attributed graph (TAG) as

$$\mathcal{G} = (\mathcal{V}, \mathbf{A}, \mathcal{T}),\ \mathcal{T} = \{t_i\}_{i=1}^{N}, \tag{1}$$

where $\mathcal{V} = \{v_i\}_{i=1}^{N}$ is the node set, $\mathbf{A}$ is the adjacency matrix, and $\mathcal{T}$ contains node texts. We learn a detector from source TAGs without using anomaly labels for training and apply it to unseen target TAGs with frozen parameters and source-domain anchors.

3.1 Quantifying Over-Generalization

Let $Y = 0$ denote a normal node and $Y = 1$ an anomaly. Fix a trained detector and its anchor banks, and let S be its finite real-valued anomaly score, with larger values indicating greater abnormality. Write $F_y(t) = \Pr(S \le t \mid Y = y)$ and assume both classes have positive probability. For a prescribed normal coverage $q \in (0, 1)$, define

$$t_q = \inf\{t : F_0(t) \ge q\}, \qquad a_q(S) = \mathbf{1}\{S < t_q\} + \gamma_q \mathbf{1}\{S = t_q\}, \tag{2}$$

where

$$\gamma_q = \begin{cases} \dfrac{q - F_0(t_q^-)}{F_0(t_q) - F_0(t_q^-)}, & F_0(t_q) > F_0(t_q^-), \\ 1, & \text{otherwise.} \end{cases} \tag{3}$$

The same boundary weight is applied to both classes. It gives

$$\mathbb{E}[a_q(S) \mid Y = 0] = q, \qquad \mathcal{OG}(q) = \mathbb{E}[a_q(S) \mid Y = 1]. \tag{4}$$

Thus $\mathcal{OG}$ is the false-negative rate at false-positive rate $1 - q$, corresponding to an operating point on the ROC curve Davis & Goadrich (2006); lower is better. We use this established operating-point error to investigate a specific aggregation mechanism, not as a new statistical error measure. If F_0 has no atom at t_q, then $\mathcal{OG}(q) = F_1(t_q)$.

For anomaly prevalence $\pi = \Pr(Y = 1) \in (0, 1)$, contamination of the normal acceptance region is

$$\mathrm{Cont}(q) = \frac{\pi \mathcal{OG}(q)}{(1 - \pi)q + \pi \mathcal{OG}(q)}. \tag{5}$$

To compare a prototype rule P with an anchor rule A, we report $\Delta_{\mathcal{OG}}(q) = \mathcal{OG}_P(q) - \mathcal{OG}_A(q)$ at the same coverage. Positive values favor A. Attribution to aggregation additionally requires matched encoders, candidate information, and storage budgets. Target labels determine these thresholds only for offline diagnosis; they do not tune the detector or define a label-free deployment threshold.

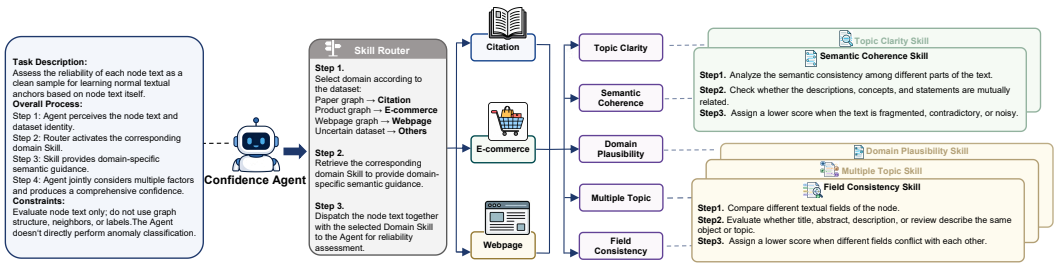


Figure 3: The Agents-Skills-Confidence(·) produces a confidence c_i for each node. Larger scores indicate stronger assessed normality.

3.2 Textual Anchoring

Textual Anchor Selection. Raw text often contains rich cues that provide complementary information to graph topology. Existing text encoders primarily capture general semantics, but may overlook the fine-grained signals of node reliability. To address this issue, we introduce an Agent-Skill-Confidence mechanism illustrated in Figure 3. The Skill Router activates the relevant domain

skill for each node, after which an LLM agent evaluates the text from multiple semantic perspectives. These assessments are aggregated into a semantic reliability score $c_i \in [0, 1]$, which is used to identify reliable confidence for textual anchor construction. Specifically, for node v_i, we obtain

$$c_i = \texttt{Agents-Skills-Confidence}\left(\mathcal{G}_i^t\right), \tag{6}$$

where a larger c_i indicates that the node exhibits stronger agreement with normal semantic patterns, whereas a smaller value suggests potentially anomalous semantic characteristics.

We select textual anchor indices using the source-node normality scores:

$$\bar{\mathcal{A}}^+ = \mathrm{TopK}_K\{c_i : i = 1, \ldots, N\}, \qquad \bar{\mathcal{A}}^- = \mathrm{BottomK}_K\{c_i : i = 1, \ldots, N\}. \tag{7}$$

Here K is the number selected per source graph, branch, and pseudo-class. The operators return node indices, with ties resolved by a fixed ordering. Higher and lower scores are used to assign normal and anomalous pseudo-labels, respectively, rather than establish ground-truth classes. We distinguish the selection budget K from the retrieval size k used in scoring.

Text Encoding. To construct a unified representation space for anchor learning, we first encode the textual attributes of the nodes, before projecting them into a shared latent space.

$$\mathbf{Z} = \mathrm{proj}(\texttt{TextEncoder}(\mathcal{T})) = [\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_N]^\top \in \mathbb{R}^{N \times d}, \tag{8}$$

where $\texttt{TextEncoder}(\cdot)$ extracts semantic representations from node texts, and $\mathrm{proj}(\cdot)$ maps the extracted features into a shared d-dimensional latent space. The resulting representation $\mathbf{z}_i$ serves as the initial semantic representation of node v_i.

Graph Encoding. For nodes with nonempty neighborhoods, define

$$\mathbf{u}_i^\circ = \frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} \mathbf{z}_j, \; \mathbf{u}^* = \frac{1}{N} \sum_{j=1}^{N} \mathbf{z}_j, \; \mathbf{U}^\circ = [\mathbf{u}_1^\circ, \ldots, \mathbf{u}_N^\circ]^\top, \mathbf{U}^* = [\mathbf{u}^*, \ldots, \mathbf{u}^*]^\top. \tag{9}$$

The graph encoder combines semantic features with local and global residuals:

$$\mathbf{H} = \texttt{GNN}(\mathbf{Z} \| (\mathbf{Z} - \mathbf{U}^\circ) \| (\mathbf{Z} - \mathbf{U}^*), \mathbf{A}) = [\mathbf{h}_1, \ldots, \mathbf{h}_N]^\top. \tag{10}$$

These inputs expose deviations to the encoder; they do not by themselves guarantee separation of normal and anomalous nodes. We adopt the same neighborhood definition during training and inference. Given the adjacency matrix $\mathbf{A}$, the neighborhood of node i is defined as $\mathcal{N}(i) = \{j \mid \mathbf{A}_{ij} \neq 0, \; j \neq i\}$. Thus, the central node's own features are excluded when computing the neighborhood mean. For nodes with an empty neighborhood, i.e., $\mathcal{N}(i) = \varnothing$, we define the neighborhood mean as zero. Consequently, their local residual equals their own representation.

Textual Anchor-guided Learning. For a normal-anchor index i, its neighbors provide positive references and textual anomaly anchors provide negatives:

$$\mathcal{L}_{\bar{\mathcal{A}}} = - \sum_{\substack{i \in \bar{\mathcal{A}}^+ \\ |\mathcal{N}(i)| > 0}} \log \frac{\sum_{j \in \mathcal{N}(i)} \exp(\texttt{sim}(\mathbf{h}_i, \mathbf{h}_j)/\tau)}{\sum_{j \in \mathcal{N}(i)} \exp(\texttt{sim}(\mathbf{h}_i, \mathbf{h}_j)/\tau) + \sum_{a \in \bar{\mathcal{A}}^-} \exp(\texttt{sim}(\mathbf{h}_i, \mathbf{h}_a)/\tau)}. \tag{11}$$

Here $\tau > 0$ and similarities are finite. The expression uses a ratio of aggregated positive mass to positive-plus-negative mass. Nodes without positive references are excluded from this sum.

3.3 Topological Anchoring

Topological Anchor Selection. For eligible nodes with defined neighborhood means and similarities, let $r_i = \texttt{sim}(\mathbf{z}_i, \mathbf{u}_i^\circ)$. We select

$$\hat{\mathcal{A}}^+ = \mathrm{TopK}_K\{r_i\}, \qquad \hat{\mathcal{A}}^- = \mathrm{BottomK}_K\{r_i\}. \tag{12}$$

High and low node–neighborhood consistency provide structural normal and anomaly pseudo-labels, respectively. The eligible-node rule and deterministic tie handling are fixed before target evaluation.

Topological Anchor-guided Learning. Using the same aggregation convention as the textual branch,

$$\mathcal{L}_{\hat{\mathcal{A}}} = -\sum_{\substack{i\in\hat{\mathcal{A}}^+\\ |\mathcal{N}(i)|>0}} \log \frac{\sum_{j\in\mathcal{N}(i)} \exp(\mathtt{sim}(\mathbf{h}_i,\mathbf{h}_j)/\tau)}{\sum_{j\in\mathcal{N}(i)} \exp(\mathtt{sim}(\mathbf{h}_i,\mathbf{h}_j)/\tau) + \sum_{a\in\hat{\mathcal{A}}^-} \exp(\mathtt{sim}(\mathbf{h}_i,\mathbf{h}_a)/\tau)}. \tag{13}$$

This objective preserves local consistency around topological normal anchors while separating them from nodes with strong neighborhood inconsistency.

3.4 Joint Textual and Topological Anchoring

Anchor Integration. Since textual and topological anchors characterize node reliability from complementary perspectives, we further integrate them into unified normal and anomaly anchor banks:

$$\mathcal{A}^+ = \{\mathbf{h}_i \mid i \in \{\bar{\mathcal{A}}^+ \cup \hat{\mathcal{A}}^+\}, \mathbf{h}\in\mathbf{H}\},\ \mathcal{A}^- = \{\mathbf{h}_i \mid i \in \{\bar{\mathcal{A}}^- \cup \hat{\mathcal{A}}^-\}, \mathbf{h}\in\mathbf{H}\} \tag{14}$$

In this way, the anchor banks retain representative instances identified from both textual and topological views rather than compressing them into a single global prototype.

Joint Anchor Learning. To further couple the two complementary anchor sources, we apply an additional contrastive objective over the unified anchor banks:

$$\mathcal{L}_{\mathcal{A}} = -\sum_{i\in\mathcal{A}^+} \log \sum_{j\in\mathcal{N}(i)} \exp\left(\frac{\mathtt{sim}(\mathbf{h}_i,\mathbf{h}_j)/\tau}{\sum\limits_{j\in\mathcal{N}(i)} \exp(\mathtt{sim}(\mathbf{h}_i,\mathbf{h}_j)/\tau) + \sum\limits_{k\in\mathcal{A}^-} \exp(\mathtt{sim}(\mathbf{h}_i,\mathbf{h}_k)/\tau)}\right). \tag{15}$$

This joint constraint ensures consistent normal patterns in both views while keeping them separate from the anomaly anchors.

Finally, **AnTAGAD** jointly optimizes the textual, topological, and unified anchor objectives:

$$\mathcal{L} = \mathcal{L}_{\bar{\mathcal{A}}} + \mathcal{L}_{\hat{\mathcal{A}}} + \mathcal{L}_{\mathcal{A}}. \tag{16}$$

The three objectives work together to preserve view-specific information and ensure consistency across views, allowing the model to learn representations that are both transferable and sensitive to individual instances for unseen graphs.

3.5 Anomaly Scoring and Prediction

During inference, an unseen node is evaluated according to its relative similarity to the learned anomaly and normal anchor banks.

For later analysis, we denote the neighborhood-dispersion quantity by

$$V_i = \begin{cases} \dfrac{1}{|\mathcal{N}(v_i)|} \displaystyle\sum_{j\in\mathcal{N}(v_i)} \|\mathbf{z}_j - \mathbf{u}_i^{\circ}\|_2^2, & |\mathcal{N}(v_i)| > 0, \\ 0, & |\mathcal{N}(v_i)| = 0. \end{cases} \tag{17}$$

For an isolated node, we set $V_i = 0$ because no neighborhood-dispersion information is available. In Eq. 18, the neighborhood-dispersion term follows this convention.

Specifically, we define

$$\mathcal{S}(v_i) = \frac{1}{k}\sum_{k} \underset{j\in\mathcal{A}^-}{\text{Top-}k}(\mathtt{Sim}(\mathbf{h_i},\mathbf{h}_j)) - \frac{1}{k}\sum_{k} \underset{j\in\mathcal{A}^+}{\text{Top-}k}(\mathtt{Sim}(\mathbf{h_i},\mathbf{h}_j)) + \lambda V_i \tag{18}$$

The first two terms measure the relative similarity of v_i to anomaly and normal anchors, respectively. Larger scores are assigned to nodes that are closer to the anomaly anchors and farther from the normal ones. The Top-k operator retains only the most informative anchors. The final term measures local neighborhood dispersion. We use V_i to denote this final dispersion term in the following theoretical analysis.

4 EXPERIMENTS

4.1 MAIN RESULT

Table 1 reports AUPRC for 18 baselines and **AnTAGAD** on eight target graphs. **AnTAGAD** ranks first on seven targets with an average rank of 1.25. Its largest improvements over the strongest baseline are 22.14 percentage points on Pubmed and 15.86 points on Texas. The relatively weaker performance on Fitness is reflected in both AUPRC and the $\mathcal{OG}$ results in Table 3, possibly due to stronger overlap between normal and anomalous patterns, which weakens the discriminability of instance-level references and leads to greater anomaly acceptance at high normal-node coverage. These results support improved anomaly ranking on most evaluated targets. They do not identify the cause of the improvement or establish probability calibration. The full AUROC results are reported in Appendix L; the $\mathcal{OG}$ analysis below specifies the additional comparisons needed to test reference aggregation.

Table 1: Anomaly detection in terms of AUPRC (%). The best and second-best results are highlighted in bold and underlined, respectively. Methods marked with * are supervised methods.

Method	Pubmed	Texas	Grocery	Toys	Movies	Cornell	Citeseer	Fitness	Rank
GAD Methods									
BGNN* (2021)	5.85±3.20	12.14±2.10	6.61±0.65	6.54±0.92	6.99±0.91	18.69±6.67	9.01±3.31	5.66±0.53	14.88
BWGNN* (2022)	13.63±5.85	12.93±2.68	9.36±0.82	8.28±0.42	8.17±1.27	15.66±1.35	20.81±3.69	8.59±1.65	11.12
GHRN* (2023)	15.18±10.03	13.45±2.50	9.27±1.26	8.18±0.92	8.11±1.34	16.24±1.59	18.52±1.34	9.04±1.88	10.62
DOMINANT (2019)	9.37±0.93	21.14±1.93	7.60±0.23	8.10±0.23	6.97±0.29	31.16±3.65	9.45±1.32	13.23±1.62	10.50
CoLA (2021)	5.43±0.17	13.88±3.37	6.35±0.21	7.14±0.17	6.04±0.21	10.28±0.77	8.94±1.42	6.33±0.13	16.00
HCM-A (2022)	9.41±0.61	13.99±0.97	12.27±0.67	15.73±0.60	11.14±0.46	18.89±2.69	10.72±0.67	8.86±0.50	9.50
TAM (2023)	11.87±0.22	15.93±0.88	10.01±0.09	12.64±0.09	8.55±0.05	14.29±0.93	3.92±0.00	*OOM*	12.13
Semi-GGAD (2024)	4.20±0.53	14.49±7.91	4.84±0.82	5.16±0.81	4.78±0.70	11.68±4.61	5.78±1.96	4.71±0.81	17.62
CAGAD (2024)	5.59±1.58	21.59±7.46	7.97±1.28	7.37±0.54	8.52±1.38	15.73±2.27	17.03±3.90	7.97±3.75	11.62
GCTAM (2025)	26.77±1.78	24.51±4.05	13.34±0.72	12.35±0.69	10.14±0.39	21.83±3.98	7.69±0.23	*OOM*	8.75
GGAD Methods									
ARC* (2024)	29.70±0.86	31.44±1.77	19.79±0.38	19.86±1.18	17.31±0.81	40.11±3.90	51.36±0.14	**32.29±0.87**	2.13
IA-GGAD* (2025)	10.57±1.58	13.40±1.65	18.82±0.71	17.67±1.10	18.59±0.51	12.49±1.57	35.21±0.33	32.26±0.63	7.38
AnomalyGFM* (2025)	13.88±3.92	16.81±3.21	7.46±0.83	7.25±0.42	6.46±0.26	29.45±3.44	28.61±0.95	8.34±0.81	10.38
UNPrompt* (2025)	9.98±2.40	18.23±4.92	6.07±0.27	5.81±0.23	6.60±0.24	24.83±1.67	12.04±1.11	13.22±3.14	11.88
OWLEYE* (2026)	21.98±0.02	27.93±0.09	12.95±0.40	17.26±0.01	17.72±0.01	35.23±0.11	43.90±0.01	22.38±0.01	3.75
ProMoS (2026)	12.21±0.27	22.86±0.20	12.24±0.34	12.28±0.58	12.59±0.25	29.83±1.02	25.65±0.94	29.61±1.62	6.25
TAGAD Methods									
CMUCL (2025)	5.90±0.32	13.61±3.48	5.49±0.20	5.27±0.04	5.50±0.07	15.03±3.28	14.29±0.66	5.17±0.10	15.62
CoLL (2025)	19.99±1.70	15.01±1.84	11.17±0.74	11.12±1.38	13.18±1.47	22.85±1.97	7.93±0.72	16.98±0.95	8.50
AnTAGAD (Ours)	**51.84±1.29**	**47.30±3.44**	**33.64±1.04**	**32.24±0.29**	**30.37±0.54**	**49.82±0.94**	**53.07±1.31**	30.19±1.20	**1.25**
Δ	↑ 22.14	↑ 15.86	↑ 13.85	↑ 12.38	↑ 11.78	↑ 9.71	↑ 1.71	↓ 2.10	–

Δ denotes the improvement (↑) or degradation (↓) compared with the current best baseline method. *OOM* means out of memory in our computational environment. Rank indicates the average ranking over eight datasets.

4.2 ABLATION STUDY

Table 2 reports component ablations. All variants are independently retrained on the same six source graphs using the same random seeds, with only the corresponding component removed, and are then evaluated in a zero-shot manner on the eight target graphs. Regarding the anchor-bank size, when either **TeA** or **ToA** is removed, the resulting normal and anomalous banks each contain 960 anchors. With both TeA and ToA enabled, the two branch-specific anchor sets are merged and duplicate selections are removed, yielding final bank sizes of 1,900 and 1,854 for normal and anomalous anchors, respectively. The full model consistently achieves the best AUROC across all eight datasets, demonstrating the complementarity of its key components. Moreover, the performance drops observed when removing **Re** or **JL** confirm that both relation modeling and the joint objective contribute to effective anchor integration. These results support

Table 2: AUROC results of **AnTAGAD** and its key components. **Re** denotes V_i in Eq. 18, while **JL** denotes the joint loss. Best per row in bold.

Dataset	w/o TeA	w/o ToA	w/o Re	w/o JL	Ours
Movies	65.93	71.98	66.22	70.48	**74.51**
Pubmed	74.98	77.42	77.77	76.23	**77.81**
Grocery	65.44	71.74	64.51	69.49	**74.21**
Fitness	73.84	78.83	73.45	77.49	**80.90**
Toys	68.73	74.57	68.66	73.46	**76.95**
Texas	53.18	76.78	60.30	67.69	**83.28**
Citeseer	58.52	71.59	52.86	70.56	**73.11**
Cornell	51.39	70.80	54.73	62.32	**81.31**

the usefulness of the combined components, while the corresponding $\mathcal{OG}$**@95** ablation results under the same setting are provided in Appendix F to further assess their contributions to alleviating $\mathcal{OG}$.

4.3 Parameter Sensitivity

The coefficient λ in Eq. 18 controls the contribution of neighborhood dispersion. For a nonempty neighborhood, the dispersion term in Eq. 17 admits the equivalent variance form

$$V_i = \frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} \|\mathbf{z}_j\|_2^2 - \|\mathbf{u}_i^\circ\|_2^2. \tag{19}$$

This identity follows directly from $\mathbf{u}_i^\circ = |\mathcal{N}(i)|^{-1} \sum_{j \in \mathcal{N}(i)} \mathbf{z}_j$ and does not introduce a new definition of V_i. Therefore, $V_i \geq 0$ for nonempty neighborhoods, while $V_i = 0$ for isolated nodes according to Eq. 17.

We vary λ over $\{0, 0.1, \cdots, 1.0\}$ to examine its sensitivity. As shown in Fig. 4, AUROC and AUPRC generally improve with λ at first, but saturate or decrease when λ becomes too large. The role of λ is also consistent with the sufficient margin in Eq. 24: when $v_1 > v_0$, increasing λ enlarges the neighborhood contribution to this sufficient margin, whereas when $v_1 < v_0$, it reduces that contribution. This observation concerns the form of the sufficient condition and does not imply that all conditions of Theorem 2 are satisfied on a particular target graph. An excessively large λ may still overemphasize neighborhood dispersion and impair ranking performance. More detailed parameter sensitivity analyses are provided in Appendix G.

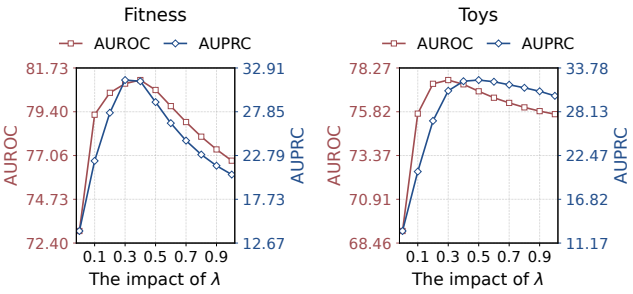


Figure 4: Regularization Weight VS AUROC and AUPRC.

4.4 Over-Generalization Analysis

We evaluate whether localized references reduce anomaly acceptance while retaining the same prescribed fraction of normal nodes. Unlike Theorem 2, which provides a sufficient guarantee under explicit high-probability conditions, the following experiments directly measure $\mathcal{OG}$ from the final anomaly scores. Therefore, empirical improvement in $\mathcal{OG}$ does not require a positive sufficient margin Γ. In the following analysis, $\mathcal{OG}$**@90**, $\mathcal{OG}$**@95**, and $\mathcal{OG}$**@99** denote the anomaly acceptance rates under fixed normal-node coverage levels of 90%, 95%, and 99%, respectively.

Table 3: Comparison of $\mathcal{OG}$ (%) under different coverage levels. Lower values indicate better performance. $\downarrow$ denotes the reduction in $\mathcal{OG}$ achieved by Ours over the second-best method, whereas $\uparrow$ denotes the excess $\mathcal{OG}$ of Ours relative to the best-performing method.

Dataset	$\mathcal{OG}$@90				$\mathcal{OG}$@95				$\mathcal{OG}$@99			
	ProMoS	IA-GGAD	Ours	$\Delta\mathcal{OG}$	ProMoS	IA-GGAD	Ours	$\Delta\mathcal{OG}$	ProMoS	IA-GGAD	Ours	$\Delta\mathcal{OG}$
Citeseer	49.48±0.00	48.84±1.40	**45.31±1.28**	↓ 3.53	51.46±1.00	51.00±1.67	**47.40±0.37**	↓ 3.60	89.79±3.60	70.10±6.15	**54.06±3.38**	↓ 16.04
Pubmed	58.72±1.29	57.89±4.95	**43.80±0.53**	↓ 14.09	95.80±1.21	66.66±2.95	**54.51±1.19**	↓ 12.15	98.94±0.48	80.69±1.58	**57.16±2.27**	↓ 23.53
Grocery	75.41±1.54	65.94±3.24	**55.37±1.79**	↓ 10.57	85.25±0.87	75.62±2.33	**63.85±1.48**	↓ 11.77	94.71±0.55	88.77±1.00	**77.29±0.50**	↓ 11.48
Movies	80.70±0.76	66.76±1.84	**54.02±0.86**	↓ 12.74	87.36±1.14	76.20±1.69	**62.22±0.73**	↓ 13.98	92.92±0.39	88.26±0.58	**77.64±0.72**	↓ 10.62
Toys	71.81±1.96	57.92±0.86	**49.68±0.52**	↓ 8.24	85.40±1.19	70.44±0.93	**61.23±0.57**	↓ 9.21	96.08±0.42	87.06±1.81	**81.21±0.67**	↓ 5.85
Fitness	61.88±1.32	**45.38±4.68**	48.03±0.84	↑ 2.65	69.53±2.35	**54.72±4.26**	60.20±0.67	↑ 5.48	78.23±1.70	**74.10±1.80**	79.87±1.27	↑ 5.77
Cornell	56.24±2.42	**52.38±7.53**	57.14±3.37	↑ 4.76	75.24±3.98	69.52±11.47	**64.76±2.61**	↓ 4.76	99.05±2.13	100.00±0.00	**73.33±2.61**	↓ 25.72
Texas	70.48±6.21	69.52±11.95	**49.52±9.87**	↓ 20.00	86.67±2.13	82.86±10.43	**64.76±2.61**	↓ 18.10	100.00±0.00	99.05±2.13	**80.95±4.76**	↓ 18.10

High-coverage diagnosis. Table 3 shows that **AnTAGAD** achieves the lowest $\mathcal{OG}$ in 20 out of 24 settings, including 6/8 datasets at $\mathcal{OG}$**@90** and 7/8 at both $\mathcal{OG}$**@95** and $\mathcal{OG}$**@99**. The reductions are especially large on Pubmed, Texas, and Cornell, showing that, compared with IA-GGAD and ProMoS, **AnTAGAD** more effectively limits the inclusion of anomalous nodes in the normal region. However, **AnTAGAD** yields higher $\mathcal{OG}$ than IA-GGAD on Fitness at all three coverage levels and on Cornell at $\mathcal{OG}$**@90**, indicating that its advantage is not uniform across datasets. For the small

datasets Texas and Cornell, we additionally report in Appendix D.5 the numbers of anomalous nodes admitted into the normal region by each of the three methods at different normal-coverage levels. Detailed curves under high-coverage settings and paired $\mathcal{OG}$ difference analysis are presented in Appendix D.1 and Appendix D.2.

Controlled reference comparison. In this section, we conduct three controlled experiments to evaluate reference construction, retrieval using the same **AnTAGAD** anchor bank, and different references within the complete scoring model. As shown in Table 4, the construction experiment fixes the encoder, source candidate pool, Top-1 retrieval, and $\lambda = 0$, while comparing Mean Prototype, K-means Centers, Representative Instances, and **AnTAGAD** Anchors over eight target datasets and five random seeds. **AnTAGAD** Anchors yields the lowest anomaly coverage at $\mathcal{OG}$**@90** and $\mathcal{OG}$**@95**, indicating that it admits fewer anomalous nodes into the normal acceptance region. At $\mathcal{OG}$**@99**, its anomaly coverage increases because accepting nearly all normal nodes requires extending the acceptance threshold toward low-density boundary regions, where normal and anomalous samples are more likely to overlap. In this regime, multiple cluster centers may better accommodate heterogeneous normal modes. Detailed results of the other two controlled experiments are presented in Appendix D.3 and Appendix D.4.

Table 4: Comparison of different reference construction strategies under different normal coverage levels.

Method	$\mathcal{OG}$**@90**	$\mathcal{OG}$**@95**	$\mathcal{OG}$**@99**
Mean Prototype	84.12±0.83	94.60±0.67	99.41±0.31
K-means Centers	81.37±1.55	89.14±0.81	94.31±1.39
Representative Instances	82.91±1.59	89.40±1.34	**93.73±1.15**
AnTAGAD Anchors	**78.20±1.25**	**87.63±0.87**	96.41±0.94

Relation to the sufficient guarantee. Theorem 2 provides a sufficient certificate rather than a necessary characterization of low $\mathcal{OG}$. In real target graphs, normal and anomalous scores may overlap, and the sufficient margin Γ need not be positive.

Accordingly, we do not use $\Gamma > 0$ as a prerequisite for empirical improvement. When the sufficient conditions are not satisfied, the theorem is simply inconclusive. The empirical question is instead whether a detector accepts fewer anomalous nodes at the same normal-node coverage, which is measured directly in Table 3.

Uncertainty and interpretation. Training-seed variability on a fixed graph is not a confidence interval for performance on new graphs. A lower $\mathcal{OG}$ at one coverage does not establish dominance at all coverages, and contamination is not independent evidence when coverage and prevalence are fixed. Likewise, empirical overlap between normal and anomalous scores does not contradict Theorem 2, because its guarantee applies only when the stated high-probability conditions are jointly satisfied.

5 Conclusion

We presented **AnTAGAD** for generalist text-attributed graph anomaly detection using textual and topological instance anchors. We formulated $\mathcal{OG}$ as anomaly acceptance at fixed normal-node coverage and derived a sufficient high-probability separation condition for the actual top-k scoring rule. This condition provides a formal certificate when its assumptions hold; it does not require the complete normal and anomalous score distributions to be disjoint on real target graphs. The analysis explains a possible benefit of localized references while showing that additional anchors need not improve $\mathcal{OG}$. The reported results on eight unseen graphs support improved cross-domain detection performance and lower anomaly acceptance at matched normal-node coverage on most evaluated settings. These empirical improvements do not require the sufficient certificate to hold on every target graph.

Limitations. The performance of **AnTAGAD** is sensitive to the number of anchors, and we do not yet have a principled way to determine the optimal anchor budget, which is currently set by trial and error. In addition, anchor-selection quality depends on the LLM agent's capabilities and its configured skills, so poorly specified skills can render the selected anchors ineffective.

REFERENCES

Leman Akoglu, Hanghang Tong, and Danai Koutra. Graph based anomaly detection and description: a survey. *Data mining and knowledge discovery*, 29(3):626–688, 2015.

Anonymous. Towards anomaly detection on text-attributed graphs, 2025. URL https://openreview.net/forum?id=LMKYd9JHgU.

Sambaran Bandyopadhyay, N Lokesh, and M Narasimha Murty. Outlier aware network embedding for attributed networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 12–19, 2019.

Sambaran Bandyopadhyay, Lokesh N, Saley Vishal Vivek, and M Narasimha Murty. Outlier resistant unsupervised deep architectures for attributed network embedding. In *Proceedings of the 13th international conference on web search and data mining*, pp. 25–33, 2020.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pp. 233–240, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595933832. doi: 10.1145/1143844.1143874. URL https://doi.org/10.1145/1143844.1143874.

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.

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.

Changlong Fu, Zhenli He, Xiong Zhang, Cheng Xie, Xin Jin, and Yun Yang. Dr-ggad: Dual residual centering for mitigating anomaly non-discriminativity in generalist graph anomaly detection. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 71971–71997, 2026.

Yuan Gao, Xiang Wang, Xiangnan He, Zhenguang Liu, Huamin Feng, and Yongdong Zhang. Addressing heterophily in graph anomaly detection: A perspective of graph spectrum. In *Proceedings of the ACM Web Conference 2023*, pp. 1528–1538, 2023a.

Yuan Gao, Xiang Wang, Xiangnan He, Zhenguang Liu, Huamin Feng, and Yongdong Zhang. Alleviating structural distribution shift in graph anomaly detection. In *Proceedings of the sixteenth ACM international conference on web search and data mining*, pp. 357–365, 2023b.

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.

Tianjin Huang, Yulong Pei, Vlado Menkovski, and Mykola Pechenizkiy. Hop-count based self-supervised anomaly detection on attributed networks. In *Joint European conference on machine learning and knowledge discovery in databases*, pp. 225–241. Springer, 2022.

Sergei Ivanov and Liudmila Prokhorenkova. Boost then convolve: Gradient boosting meets graph neural networks. *arXiv preprint arXiv:2101.08543*, 2021.

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. *Advances in Neural Information Processing Systems*, 37:50772–50804, 2024.

Xuexiong Luo, Jia Wu, Amin Beheshti, Jian Yang, Xiankun Zhang, Yuan Wang, and Shan Xue. Comga: Community-aware attributed graph anomaly detection. In *Proceedings of the fifteenth ACM international conference on web search and data mining*, pp. 657–665, 2022.

Xiaoxiao Ma, Jia Wu, Shan Xue, Jian Yang, Chuan Zhou, Quan Z Sheng, Hui Xiong, and Leman Akoglu. A comprehensive survey on graph anomaly detection with deep learning. *IEEE transactions on knowledge and data engineering*, 35(12):12012–12038, 2021.

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.

Chaoxi Niu, Hezhe Qiao, Changlu Chen, Ling Chen, and Guansong Pang. Zero-shot generalist graph anomaly detection with unified neighborhood prompts. In James Kwok (ed.), *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pp. 3226–3234. International Joint Conferences on Artificial Intelligence Organization, 8 2025. doi: 10.24963/ijcai.2025/359. URL https://doi.org/10.24963/ijcai.2025/359. Main Track.

Hezhe Qiao and Guansong Pang. Truncated affinity maximization: One-class homophily modeling for graph anomaly detection. *Advances in Neural Information Processing Systems*, 36:49490–49512, 2023.

Hezhe Qiao, Qingsong Wen, Xiaoli Li, Ee-Peng Lim, and Guansong Pang. Generative semi-supervised graph anomaly detection. *Advances in neural information processing systems*, 37: 4660–4688, 2024.

Hezhe Qiao, Chaoxi Niu, Ling Chen, and Guansong Pang. Anomalygfm: Graph foundation model for zero/few-shot anomaly detection. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 2326–2337, 2025a.

Hezhe Qiao, Hanghang Tong, Bo An, Irwin King, Charu Aggarwal, and Guansong Pang. Deep graph anomaly detection: A survey and new perspectives. *IEEE Transactions on Knowledge and Data Engineering*, 37(9):5106–5126, 2025b.

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.

Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Gallagher, and Tina Eliassi-Rad. Collective classification in network data. *AI magazine*, 29(3):93–106, 2008.

Jianheng Tang, Jiajin Li, Ziqi Gao, and Jia Li. Rethinking graph neural networks for anomaly detection. In *International conference on machine learning*, pp. 21076–21089. PMLR, 2022.

Chunjing Xiao, Shikang Pang, Xovee Xu, Xuan Li, Goce Trajcevski, and Fan Zhou. Counterfactual data augmentation with denoising diffusion for graph anomaly detection. *IEEE Transactions on Computational Social Systems*, 11(6):7555–7567, 2024.

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.

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.

Yiming Xu, Zihan Chen, Zhen Peng, Song Wang, Bin Shi, Bo Dong, and Chao Shen. Generalist graph anomaly detection via prototype-based distillation. *arXiv preprint arXiv:2605.26857*, 2026.

Junhan Yang, Zheng Liu, Shitao Xiao, Chaozhuo Li, Defu Lian, Sanjay Agrawal, Amit Singh, Guangzhong Sun, and Xing Xie. Graphformers: Gnn-nested transformers for representation learning on textual graph. *Advances in Neural Information Processing Systems*, 34:28798–28810, 2021.

Xiong Zhang, Zhenli He, Changlong Fu, and Cheng Xie. IA-GGAD: Zero-shot generalist graph anomaly detection via invariant and affinity learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a.

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*, pp. 3642–3650, 2025b.

Zheng Zhang, Yuntong Hu, Bo Pan, Chen Ling, and Liang Zhao. Taga: Text-attributed graph self-supervised learning by synergizing graph and text mutual transformations. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pp. 4263–4272, 2025c.

Jianan Zhao, Meng Qu, Chaozhuo Li, Hao Yan, Qian Liu, Rui Li, Xing Xie, and Jian Tang. Learning on large-scale text-attributed graphs via variational inference. *arXiv preprint arXiv:2210.14709*, 2022.

Lecheng Zheng, Dongqi Fu, Zihao Li, and Jingrui He. Owleye: Zero-shot learner for cross-domain graph data anomaly detection. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 33497–33522, 2026.

APPENDIX

A USE OF LARGE LANGUAGE MODELS

Large language models (LLMs) were used to assist with language editing, including improving the clarity, readability, grammatical correctness, and overall presentation of the manuscript. In our experiment, we used DeepSeek-V4-Flash to assign confidence to source-domain node texts, providing auxiliary semantic guidance for model training. The authors independently developed the research ideas, methodology, experimental setup, interpretation of the results, and the resulting scientific conclusions. All LLM-assisted revisions were carefully reviewed, verified, and further modified where necessary by the authors, who take full responsibility for the final manuscript.

B REPRODUCIBILITY STATEMENT

To facilitate the reproducibility of **AnTAGAD**, we provide the following resources and implementation details:

- **Code Availability:** The source code of **AnTAGAD** is available at `https://anonymous.4open.science/r/AnTAGAD-EDF3`.
- **Experimental Setup:** To ensure that the experiments can be faithfully reproduced, we describe the experimental setup in detail in both the main text and the appendix, including hyperparameter configurations, training procedures, and evaluation protocols.

C CONDITIONAL GUARANTEES FOR $\mathcal{OG}$

Our analysis assumes cosine similarity and nonzero node and anchor vectors. After unit normalization, write them as $\widetilde{\mathbf{h}}$ and $\widetilde{\mathbf{a}}_{y,j}$, where $y = 0$ indexes $\mathcal{A}^+$ and $y = 1$ indexes $\mathcal{A}^-$. For the node being scored, we write $V = V_i$ for notational simplicity. Define

$$D^2_{y,k}(\widetilde{\mathbf{h}}) = \min_{\substack{I \subseteq \{1,\dots,M_y\} \\ |I|=k}} \frac{1}{k} \sum_{j \in I} \|\widetilde{\mathbf{h}} - \widetilde{\mathbf{a}}_{y,j}\|_2^2, \qquad D_{y,k} = \sqrt{D^2_{y,k}}. \tag{20}$$

Lemma 1 (Cosine–distance identity)**.** *Under these assumptions, the actual score satisfies*

$$S = \tfrac{1}{2}(D^2_{0,k} - D^2_{1,k}) + \lambda V. \tag{21}$$

Theorem 2 (Sufficient high-probability separation condition)**.** *Fix the detector, banks, k, and $\lambda \geq 0$. Let $r_0, r_1, d_0, d_1, v_0, v_1$ be finite nonnegative constants, and $\epsilon_0, \epsilon_1 \in [0, 1)$. Suppose the following joint conditions hold on two high-probability class-conditional events:*

$$\begin{array}{lll} D_{0,k} \leq r_0, & D_{1,k} \geq d_0, & V \leq v_0 \\ D_{1,k} \leq r_1, & D_{0,k} \geq d_1, & V \geq v_1 \end{array} \tag{22}$$

Specifically, the first row holds jointly with probability at least $1 - \epsilon_0$ conditional on $Y = 0$, and the second row holds jointly with probability at least $1 - \epsilon_1$ conditional on $Y = 1$. No restrictions are imposed on nodes outside these events; their class-conditional probability masses are at most ϵ_0 and ϵ_1, respectively. Set

$$U_0 = \tfrac{1}{2}(r_0^2 - d_0^2) + \lambda v_0, \quad L_1 = \tfrac{1}{2}(d_1^2 - r_1^2) + \lambda v_1. \tag{23}$$

If

$$\Gamma = L_1 - U_0 = \tfrac{1}{2}(d_1^2 + d_0^2 - r_0^2 - r_1^2) + \lambda(v_1 - v_0) > 0, \tag{24}$$

then each deterministic acceptance region $\{S \leq t\}$ with $t \in [U_0, L_1)$ retains at least $1 - \epsilon_0$ of normal nodes and accepts at most ϵ_1 of anomalies. Consequently,

$$\mathcal{OG}(q) \leq \epsilon_1 \quad \textit{for } 0 < q \leq 1 - \epsilon_0,\ q < 1. \tag{25}$$

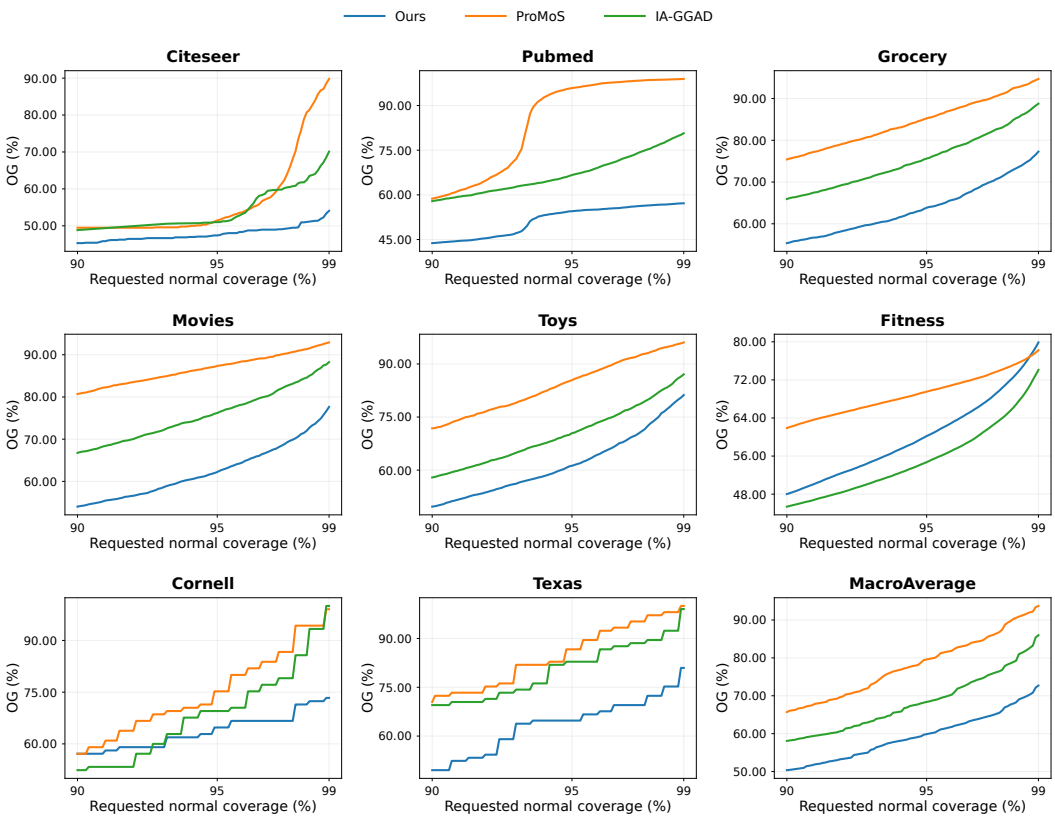


Figure 5: Detailed curves under high-coverage settings.

Proofs and a source-coverage sufficient condition are provided in Appendix E. Theorem 2 gives a sufficient, rather than necessary, condition for controlling anomaly acceptance. In particular, it does not require the complete normal and anomalous score distributions to be disjoint: nodes outside the two high-probability events in Eq. 22 may overlap.

A positive Γ certifies that these high-probability portions of the two classes are sufficiently separated in score to guarantee $\mathcal{OG}(q) \leq \epsilon_1$ for $q \leq 1 - \epsilon_0$. Conversely, $\Gamma \leq 0$ does not imply large $\mathcal{OG}$ or poor detection performance; it only means that this sufficient certificate is inconclusive. Thus, the theorem characterizes a regime in which the actual scoring rule admits a formal guarantee, rather than asserting that this regime must hold on every target graph.

D Preliminary $\mathcal{OG}$ Statistics

D.1 High-Coverage Analysis.

We further compare $\mathcal{OG}$ at normal-coverage levels from 90% to 99%. As shown in Fig. 5, our method achieves lower $\mathcal{OG}$ on most target graphs, indicating that fewer anomalous nodes enter the normal region at the same coverage. The advantage is particularly pronounced on Citeseer and Pubmed, where ProMoS exhibits sharp increases as coverage rises. Our method also maintains the lowest macro-average $\mathcal{OG}$ throughout the evaluated range. However, IA-GGAD performs better on Fitness, and our method's advantage is not uniform across all graphs. The step-like curves on Cornell and Texas are consistent with discrete changes in anomaly acceptance on small graphs.

D.2 Paired $\mathcal{OG}$ Difference Analysis.

We further compare pairwise $\mathcal{OG}$ differences across high normal-coverage levels. As shown in Fig. 6, ProMoS$-$Ours and IA-GGAD$-$Ours remain positive on most target graphs, indicating that

AnTAGAD generally admits fewer anomalous nodes into the normal region under the same coverage. The advantage is especially clear on Pubmed, Grocery, Movies, Toys, and Texas. The step-like curves on Cornell and Texas mainly result from their relatively small graph sizes, where $\mathcal{OG}$ changes discretely as the threshold crosses new node scores. Although the boundary weight γ enables continuous control of normal coverage, the admitted anomaly ratio can still change piecewise.

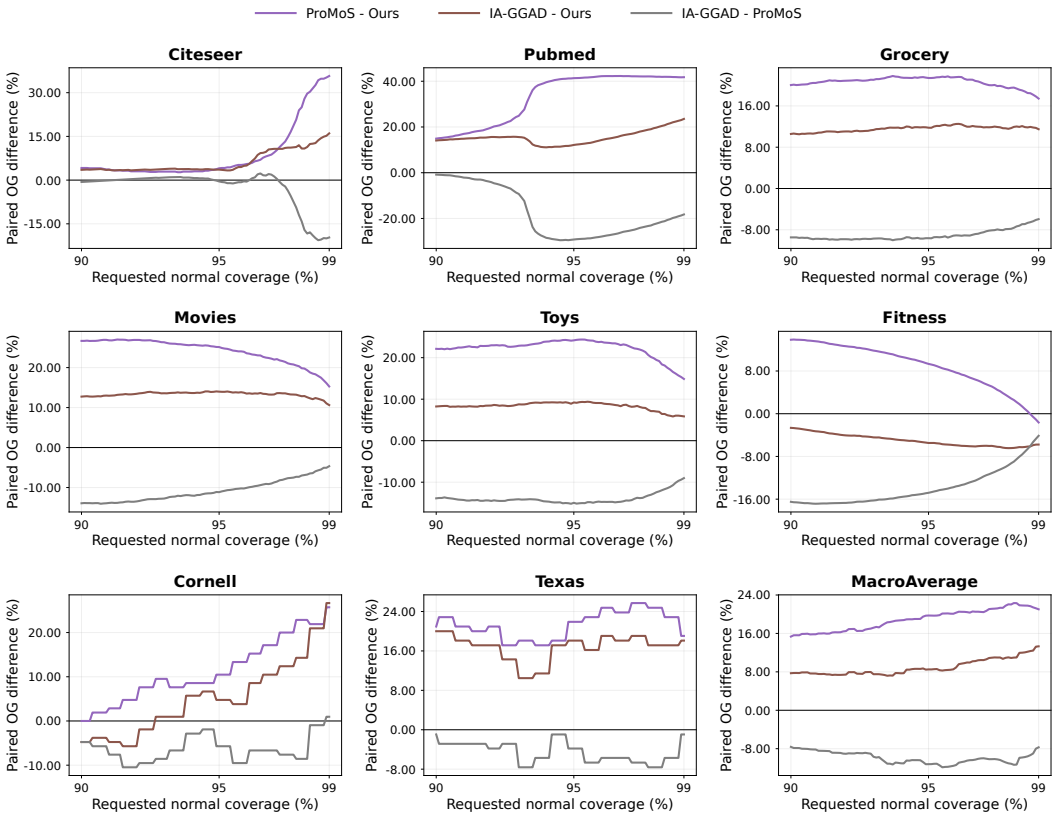


Figure 6: Paired $\mathcal{OG}$ Difference.

D.3 Analysis of Retrieval Volume.

The experiment fixes the frozen encoder and the same **AnTAGAD** anchor bank, comparing only Full Average and Local Top-k retrieval. Neither method incorporates the neighborhood structural score, i.e., $\lambda = 0$, and the average $\mathcal{OG}$ over five random seeds is reported. As shown in Table 5, Local Top-k yields lower anomaly coverage on nearly all datasets and coverage levels, indicating that selecting relevant local anchors for each target node reduces interference from irrelevant references. As normal-node coverage increases to 99%, both methods must expand the normal acceptance region and thus admit more anomalous nodes, reducing Local Top-k's advantage.

D.4 Reference Construction Strategies.

This experiment evaluates different reference construction strategies within the complete model, which combines Local Top-20 reference retrieval with the neighborhood term. With the encoder, candidate pool, and scoring weights fixed, **AnTAGAD** Anchors generally admits fewer anomalous nodes, especially at $\mathcal{OG}$**@99**. On Texas, anomaly coverage increases as the normal acceptance region expands, but **AnTAGAD** Anchors limits this increase more effectively than K-means Centers and Representative Instances.

Table 5: Comparison of $\mathcal{OG}$ (%) under different retrieval volumes.

Dataset	$\mathcal{OG}$@90		$\mathcal{OG}$@95		$\mathcal{OG}$@99	
	Full Average	**Local Top-k**	**Full Average**	**Local Top-k**	**Full Average**	**Local Top-k**
Citeseer	76.67	**72.50**	82.19	**79.69**	100.00	**94.17**
Pubmed	80.66	**47.94**	100.00	**82.99**	100.00	**91.71**
Grocery	89.24	**79.67**	97.93	**89.28**	100.00	**95.31**
Movies	87.52	**78.98**	99.52	**94.64**	100.00	**98.06**
Toys	83.66	**73.40**	93.23	**85.21**	100.00	**95.71**
Fitness	74.60	**70.39**	91.72	**83.98**	98.93	**91.35**
Cornell	90.48	**80.95**	93.33	**85.71**	100.00	**99.05**
Texas	82.86	**72.38**	90.48	**81.90**	**95.24**	96.19

Table 6: Comparison of $\mathcal{OG}$ (%) under different reference construction strategies. Lower values are better, and the best result in each group is highlighted in bold.

Dataset	$\mathcal{OG}$@90			$\mathcal{OG}$@95			$\mathcal{OG}$@99		
	K-means Centers	**Representative Instances**	**AnTAGAD**	**K-means Centers**	**Representative Instances**	**AnTAGAD**	**K-means Centers**	**Representative Instances**	**AnTAGAD**
Citeseer	45.42	**44.79**	45.31	48.02	47.92	**47.40**	67.19	66.67	**54.06**
Pubmed	46.22	46.42	**43.80**	**52.28**	52.45	54.51	62.08	62.75	**57.16**
Grocery	56.82	57.03	**55.37**	65.43	65.37	**63.85**	80.74	80.61	**77.29**
Movies	58.98	58.84	**54.02**	67.50	67.42	**62.22**	83.54	84.18	**77.64**
Toys	53.18	53.26	**49.68**	64.50	64.53	**61.23**	87.53	87.81	**81.21**
Fitness	54.50	54.28	**48.03**	65.95	65.79	**60.20**	84.23	84.59	**79.87**
Cornell	67.37	68.57	**57.14**	72.38	70.48	**64.76**	76.19	77.14	**73.33**
Texas	**46.43**	46.67	49.52	**58.10**	61.90	64.76	85.71	85.71	**80.95**

Table 7: Accepted anomaly counts at different normal-node coverage levels on Cornell and Texas. Values are averaged over five seeds; fractional counts can also result from weighted acceptance of nodes tied at the coverage threshold. Lower is better.

Dataset	$\mathcal{OG}$@90			$\mathcal{OG}$@95			$\mathcal{OG}$@99		
	ProMoS	**IA-GGAD**	**Ours**	**ProMoS**	**IA-GGAD**	**Ours**	**ProMoS**	**IA-GGAD**	**Ours**
Cornell	12.0	**11.0**	12.0	15.8	14.6	**13.6**	20.8	21.0	**15.4**
Texas	14.8	14.6	**10.4**	18.2	17.4	**13.6**	21.0	20.8	**17.0**

D.5 Analysis on Small Graphs.

Since Cornell and Texas contain relatively few nodes, $\mathcal{OG}$ can exhibit noticeable fluctuations across different normal-node coverage levels, as the inclusion of only a few additional anomalous nodes may lead to a relatively large percentage change. To examine this effect further, we report the number of anomalous nodes included in the normal region at different coverage levels. The results suggest that the apparent variation in $\mathcal{OG}$ largely stems from the discrete nature of these small graphs rather than unstable model behavior. In particular, our method consistently includes fewer anomalous nodes on Texas and shows a clearer advantage on Cornell as the normal-node coverage increases, indicating better control of anomaly leakage under high-coverage settings.

E Proofs and Scope of the $\mathcal{OG}$ Analysis

Proof of Lemma 1. For unit vectors,

$$\|\widetilde{\mathbf{h}} - \widetilde{\mathbf{a}}\|_2^2 = 2 - 2\widetilde{\mathbf{h}}^\top \widetilde{\mathbf{a}}.$$

Thus the largest k cosine similarities correspond exactly to the smallest k squared distances, including ties.

Let $M_{k,\mathcal{A}^+}$ and $M_{k,\mathcal{A}^-}$ denote the mean of the k largest cosine similarities to the normal and anomalous anchor banks, respectively. Then

$$M_{k,\mathcal{A}^+} = 1 - \frac{D_{0,k}^2}{2}, \qquad M_{k,\mathcal{A}^-} = 1 - \frac{D_{1,k}^2}{2}.$$

Writing $V = V_i$, substitution into Eq. 18 gives

$$S = M_{k,\mathcal{A}^-} - M_{k,\mathcal{A}^+} + \lambda V = \tfrac{1}{2}(D_{0,k}^2 - D_{1,k}^2) + \lambda V.$$

This proves Eq. 21. Distinct bank entries are counted separately even when their vectors are equal. □

Proof of Theorem 2. On the normal event in Eq. 22,

$$S = \tfrac{1}{2}(D_{0,k}^2 - D_{1,k}^2) + \lambda V \le \tfrac{1}{2}(r_0^2 - d_0^2) + \lambda v_0 = U_0.$$

Since this event has conditional probability at least $1 - \epsilon_0$, at least $1 - \epsilon_0$ of normal nodes have scores no larger than U_0.

Similarly, on the anomalous event,

$$S = \tfrac{1}{2}(D_{0,k}^2 - D_{1,k}^2) + \lambda V \ge \tfrac{1}{2}(d_1^2 - r_1^2) + \lambda v_1 = L_1.$$

Since this event has conditional probability at least $1 - \epsilon_1$, at least $1 - \epsilon_1$ of anomalous nodes have scores no smaller than L_1.

If $\Gamma = L_1 - U_0 > 0$, every threshold $t \in [U_0, L_1)$ accepts every node in the normal event and rejects every node in the anomalous event. Nodes outside these two events are unrestricted and may overlap in score. Hence the normal acceptance probability is at least $1 - \epsilon_0$, whereas the anomaly acceptance probability is at most ϵ_1.

For any $q \le 1 - \epsilon_0$, $F_0(U_0) \ge q$, so the minimal normal quantile threshold satisfies $t_q \le U_0 < L_1$. Therefore no anomaly belonging to the anomalous event can be accepted at t_q, including under the boundary-randomization rule in Eq. 3. Only anomalies outside this event can contribute to acceptance, whose total conditional probability is at most ϵ_1. Hence

$$\mathcal{OG}(q) \le \epsilon_1.$$

No independence among nodes is required. □

E.1 A Geometric Mechanism and Its Limits

Let $C \subset \mathbb{R}^d$ be a nonempty compact normal core and A a finite nonempty normal-anchor set. Suppose finite $\varepsilon, \rho \ge 0$ satisfy

$$\sup_{x \in C} \min_{a \in A} \|x - a\| \le \varepsilon, \qquad \max_{a \in A} \operatorname{dist}(a, C) \le \rho. \tag{26}$$

For $r \ge \varepsilon$, define $R_A = \bigcup_{a \in A} \{x : \|x - a\| \le r\}$ and $C^{[r+\rho]} = \{x : \operatorname{dist}(x, C) \le r + \rho\}$. If a measurable convex region R_P contains C, then

$$C \subseteq R_A \subseteq C^{[r+\rho]}, \quad \operatorname{conv}(C) \subseteq R_P, \quad P_1(R_P) - P_1(R_A) \ge P_1(\operatorname{conv}(C)) - P_1(C^{[r+\rho]}). \tag{27}$$

The first inclusion follows from coverage. For the second, choose an anchor within r of a point in R_A and a core point within ρ of that anchor; the triangle inequality gives distance at most $r + \rho$ to C. Convexity gives the third inclusion, and probability monotonicity gives the bound. If $P_0(C) = 1$ and $P_1(C^{[r+\rho]}) = 0$, both regions retain all normal nodes, whereas R_P accepts at least $P_1(\operatorname{conv}(C))$ of anomalies. Multiple-prototype decision regions need not be convex, and the full score in Eq. 18 need not have a union-of-balls sublevel set.

An exact comparison at matched coverage. Let $a > r > 0$, $0 < b < a - r$, $L > a + r$, $w > 0$, and $\eta \in [0, 1]$. Consider

$$\begin{aligned} X \mid Y = 0 &\sim \tfrac{1}{2}\operatorname{Unif}[-a - r, -a + r] + \tfrac{1}{2}\operatorname{Unif}[a - r, a + r], \\ X \mid Y = 1 &\sim \eta \operatorname{Unif}[-b, b] + (1 - \eta)\operatorname{Unif}[L, L + w]. \end{aligned} \tag{28}$$

The mean-prototype score $s_P(x) = |x|$ has normal quantile $a - r + 2rq$; the two-anchor score $s_A(x) = \min\{|x - a|, |x + a|\}$ has normal quantile rq. Every central anomaly has $s_P \le b < a - r$ and $s_A \ge a - b > r$. Far anomalies exceed both normal thresholds. Consequently, for every $q \in (0, 1)$,

$$\mathcal{OG}_P(q) = \eta, \qquad \mathcal{OG}_A(q) = 0. \tag{29}$$

This is an analytic mechanism example, not a result for learned embeddings on the benchmark graphs.

More clean anchors can be harmful. Let $P_0 = 0.9\,\mathrm{Unif}[-1, 1] + 0.1\delta_{10}$ and $P_1 = \delta_{10.2}$. With nearest-anchor distance and $q = 0.9$, the bank $\{0\}$ gives threshold 1 and $\mathcal{OG}$ 0. Adding the clean normal anchor 10 gives threshold $8/9$, since $0.9(8/9)+0.1 = 0.9$. The anomaly is only 0.2 from that anchor, so $\mathcal{OG}$ becomes 1. Thus clean anchors and larger bank size alone do not imply improvement.

E.2 Source Coverage as a Sufficient Condition

Let C_0, C_1 be nonempty compact subsets of the unit sphere separated by $\Delta = \inf_{x \in C_0, z \in C_1} \|x - z\| > 0$. Assume every class-$y$ anchor lies in C_y, and C_y has a finite closed-ball cover with centers in C_y and radius $\eta_y \ge 0$. Each cover ball contains at least k class-y bank entries. Suppose a target class-y high-probability subset of normalized embeddings lies within distance $\delta_y \ge 0$ of C_y, with $\Delta > \max(\delta_0, \delta_1)$.

For a target normal point in this subset, choose a nearest $x \in C_0$, a cover center within η_0 of x, and k anchors in its cover ball. The triangle inequality bounds each same-class distance by $\delta_0 + 2\eta_0$. Every opposite-class anchor is at distance at least $\Delta - \delta_0$. Exchanging classes yields

$$\begin{aligned} D_{0,k} &\le \delta_0 + 2\eta_0, \quad D_{1,k} \ge \Delta - \delta_0 \quad (Y = 0), \\ D_{1,k} &\le \delta_1 + 2\eta_1, \quad D_{0,k} \ge \Delta - \delta_1 \quad (Y = 1). \end{aligned} \tag{30}$$

If these target-subset and neighborhood conditions jointly hold with the class-conditional probabilities required by Theorem 2, these bounds can be substituted for r_y and d_y. In the symmetric case $\delta_y = \delta$, $\eta_y = \eta$, and $\lambda = 0$, a sufficient condition is

$$\Delta > 2\delta + 2\eta.$$

This deterministic result requires neither independent nodes nor random anchor selection. Selecting the top-K confidence scores alone does not establish its multi-coverage assumption. This construction therefore identifies one sufficient regime in which Theorem 2 becomes informative. Learned embeddings on real target graphs are not required to satisfy this regime, and failure of these sufficient conditions does not imply that localized anchors cannot reduce empirical $\mathcal{OG}$.

E.3 Diagnostic Identities and Boundary Handling

The normal and anomaly accepted masses are $(1 - \pi)q$ and $\pi\mathcal{OG}(q)$, respectively. Dividing the second by their sum proves Eq. 5. At fixed q and π, contamination is strictly increasing in $\mathcal{OG}$, so their simultaneous decrease is not independent evidence. The anomalous-class precision is

$$\mathrm{Precision}(q) = \frac{\pi[1 - \mathcal{OG}(q)]}{\pi[1 - \mathcal{OG}(q)] + (1 - \pi)(1 - q)}. \tag{31}$$

For empirical distributions with the boundary rule in Eq. 3, each normal-score atom of mass w at s contributes $w\{\Pr(S_1 < s) + \Pr(S_1 = s)/2\}$ to the integrated $\mathcal{OG}$ curve. Summing over the atoms gives

$$\int_0^1 \mathcal{OG}(q)\, dq = 1 - \mathrm{AUROC}, \tag{32}$$

where AUROC gives half credit to ties. Thus the full $\mathcal{OG}$ area is not a new measure independent of AUROC. The diagnostic value is its specified operating point and the controlled mechanism comparison. For empirical $q = 1$, use the largest normal score and the same boundary rule; no claim about finite thresholds for unbounded population distributions is needed.

Table 8: Ablation study of the proposed components in terms of $\mathcal{OG}$**@95**. The best result in each column is highlighted in bold.

Components				Datasets							
TeA	**ToA**	**Re**	**JL**	**Movies**	**Pubmed**	**Grocery**	**Fitness**	**Toys**	**Texas**	**Citeseer**	**Cornell**
✓	✓	✓	✓	$\mathbf{62.22 \pm 0.73}$	$\mathbf{54.51 \pm 1.19}$	$\mathbf{63.85 \pm 1.48}$	$\mathbf{60.20 \pm 0.67}$	$\mathbf{61.23 \pm 0.57}$	$\mathbf{64.76 \pm 2.61}$	$\mathbf{47.40 \pm 0.37}$	$\mathbf{64.76 \pm 2.61}$
-	✓	✓	✓	99.70 ± 0.54	99.93 ± 0.15	95.94 ± 1.47	85.12 ± 0.97	89.90 ± 0.84	91.43 ± 2.13	71.98 ± 3.07	95.24 ± 0.00
✓	-	✓	✓	95.14 ± 1.38	83.46 ± 2.96	88.87 ± 0.56	79.18 ± 2.39	82.32 ± 0.61	76.19 ± 5.83	62.60 ± 7.80	78.10 ± 4.26
✓	✓	-	✓	94.64 ± 1.46	82.99 ± 4.49	89.28 ± 0.49	83.98 ± 1.98	85.21 ± 1.05	81.90 ± 3.98	79.69 ± 4.53	85.71 ± 0.00
✓	✓	✓	-	90.08 ± 1.41	75.49 ± 2.25	87.46 ± 0.61	78.36 ± 2.36	81.58 ± 1.34	80.95 ± 3.37	61.67 ± 8.55	85.71 ± 0.00

Implementation of exact empirical coverage. For n_0 normal scores, choose the smallest observed normal threshold whose cumulative count is at least qn_0, and set $\gamma = (qn_0 - n_<)/n_=$, where $n_<$ and $n_=$ count normal scores below and at the threshold. Apply this weight equally to anomalous ties. Do not round a requested rank across an integer by an arbitrary tolerance: anomaly mass may lie inside a normal-score gap. Report the realized coverage and weighted anomaly count.

Scope of the guarantees. The geometric construction explains a possible reference-aggregation failure mode, while Theorem 2 bounds the actual score under explicit sufficient conditions. Neither result proves that the selected pseudo-labels are correct, that training necessarily satisfies the stated distance and dispersion bounds, or that the method dominates every multiple-prototype rule.

In particular, $\Gamma > 0$ is sufficient but not necessary for low $\mathcal{OG}$. When $\Gamma \leq 0$, the theorem provides no certificate but makes no negative claim about the detector. The complete normal and anomalous score distributions also need not be disjoint, since nodes outside the two high-probability events may overlap arbitrarily. Statistical confidence statements must respect graph dependence; repeated seeds on a fixed graph do not create independent target graphs.

F Ablation Study Analysis

Table 8 evaluates the contribution of each component using $\mathcal{OG}$**@95**, where a lower value indicates fewer anomalous nodes being absorbed into the normal region while retaining 95% of normal nodes. The full model achieves the lowest $\mathcal{OG}$**@95** on all eight datasets. Removing **TeA** causes the largest degradation, increasing the average $\mathcal{OG}$ from 59.87% to 91.16%, demonstrating the importance of textual anchors for preserving fine-grained semantic distinctions. Removing **ToA**, **Re**, or **JL** also consistently increases $\mathcal{OG}$, confirming the complementary roles of topology-aware information, neighborhood-dispersion modeling, and joint optimization. Overall, these components work together to effectively alleviate over-generalization.

G Parameter Sensitivity Analysis

The AUROC and AUPRC curves with respect to λ on Citeseer, Pubmed, Grocery, Movies, Texas, and Cornell are shown in Figure 7. Performance is more sensitive to small λ values and gradually stabilizes as λ increases. Increasing λ from zero substantially improves both metrics on most datasets, demonstrating the benefit of appropriately weighting the neighborhood score. However, further increases yield diminishing gains or performance declines. AUROC generally reaches a high level earlier, while AUPRC continues to improve on some datasets, revealing a trade-off between the two metrics. Overall, $\lambda \in [0.4, 0.6]$ generally provides a reasonable balance between AUROC and AUPRC.

Sensitivity of $\mathcal{OG}$**@95** to λ. Figure 8 shows how $\mathcal{OG}$**@95** varies with λ across eight datasets. Lower $\mathcal{OG}$**@95** indicates fewer anomalies incorrectly accepted at 95% normal coverage. As λ increases from zero, $\mathcal{OG}$**@95** decreases substantially across all datasets, suggesting that appropriately weighting the neighborhood score helps reduce anomaly acceptance under high normal coverage. With further increases, some curves stabilize or continue to decline, while others rise again, indicating that an excessively large neighborhood weight does not consistently improve performance. Overall, λ around 0.4 achieves relatively low $\mathcal{OG}$**@95** on most datasets, whereas the benefits of further increases vary across datasets, highlighting the importance of balancing the scoring components.

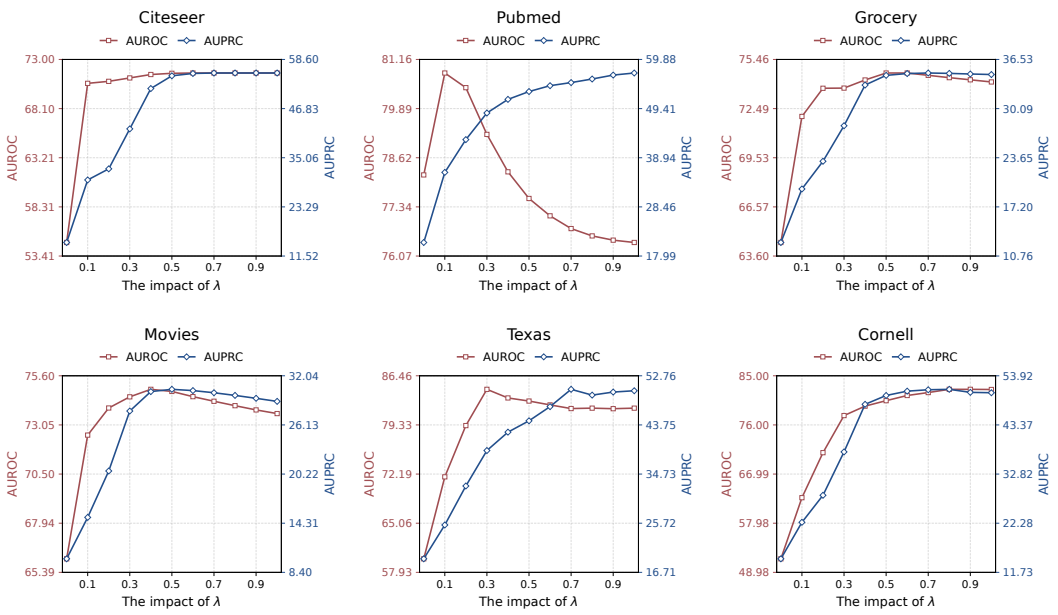


Figure 7: The impact of λ on AUROC and AUPRC across different datasets.

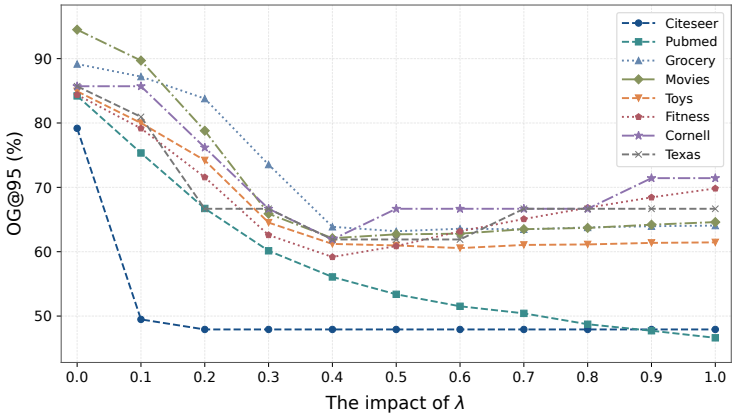


Figure 8: The impact of λ on $\mathcal{OG}$**@95** across eight datasets.

H VISUALIZATION

To further investigate the effectiveness of **AnTAGAD** beyond quantitative performance, we conduct qualitative analyses from both representation and prediction perspectives. Specifically, we visualize the learned node representations to examine the separability between normal and anomalous nodes in the latent space, and analyze the anomaly score distributions to assess the discriminative behavior of the final predictions. These analyses provide complementary insights into how **AnTAGAD** improves fine-grained anomaly discrimination.

H.1 REPRESENTATION VISUALIZATION

Fig.9 presents t-SNE visualizations of the learned node representations produced by UNPrompt, CMUCL, ProMoS, and **AnTAGAD** on Citeseer and Pubmed. Blue circles denote normal nodes, while orange crosses represent anomalous nodes. On Citeseer, the representations learned by UNPrompt, CMUCL, and ProMoS exhibit substantial overlap between normal and anomalous nodes, although UNPrompt separates a small subset of anomalies into compact peripheral clusters. In

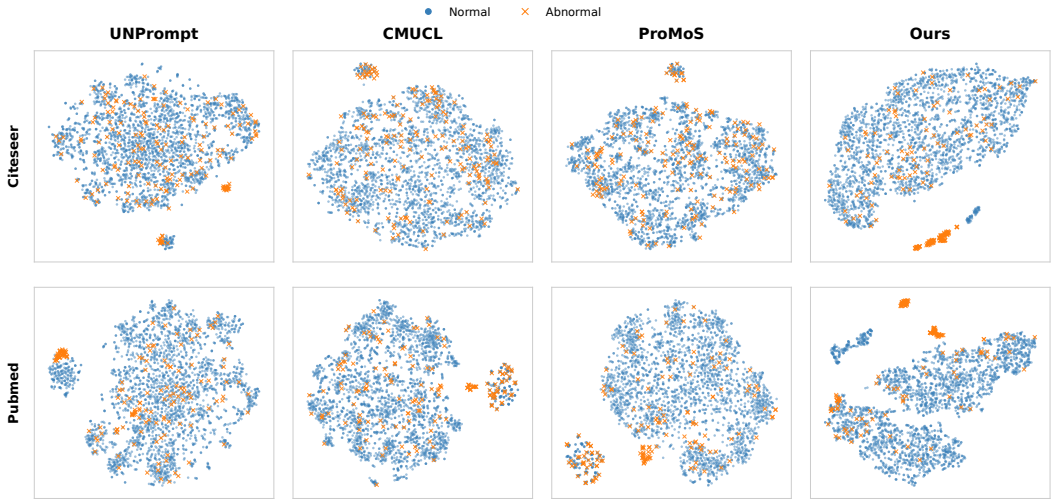


Figure 9: t-SNE visualization of learned node representations on Citeseer (top row) and Pubmed (bottom row) for UNPrompt, CMUCL, ProMoS, and Ours (left to right). Blue circles denote normal nodes, and orange crosses denote abnormal nodes.

contrast, **AnTAGAD** forms a compact anomalous cluster that is clearly separated from the main normal-node distribution, while a smaller number of anomalous nodes remain interspersed with normal nodes. On Pubmed, all methods identify some peripheral anomalous groups, but considerable class overlap remains in the representations produced by UNPrompt, CMUCL, and ProMoS. **AnTAGAD** produces several compact anomalous clusters with clearer separation from the dominant normal regions. These projections offer qualitative views of the representations. Quantitative separation and $\mathcal{OG}$ comparisons must be computed from the original learned embeddings and final scores; the two-dimensional layout does not establish either claim.

H.2 Anomaly Score Distribution Analysis

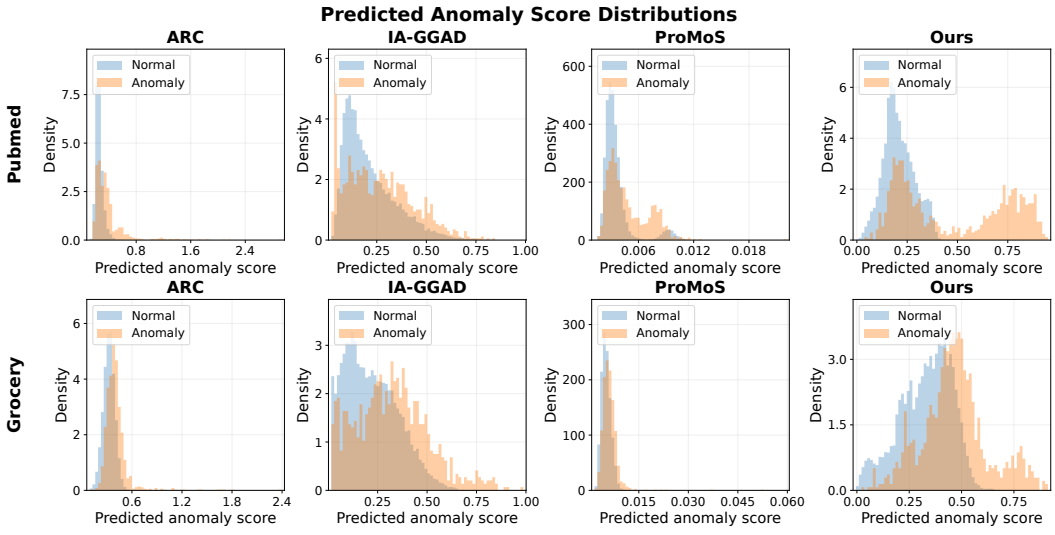


Figure 10: Predicted anomaly score distributions of ARC, IA-GGAD, ProMoS, and Ours on the Pubmed and Grocery datasets.

Figure 10 shows normal and anomalous score distributions on Pubmed and Grocery. The plots suggest differences in class overlap, but raw score ranges are not directly comparable across meth-

ods. An increasing affine rescaling can widen a score distribution without changing its ranking or coverage-matched $\mathcal{OG}$. We therefore use the curves as qualitative evidence and compare the final scores quantitatively at common normal coverage. They do not establish probability calibration.

I Description of Baselines

To ensure a fair and comprehensive evaluation of **AnTAGAD**, we compare it against eighteen representative baseline methods, covering three learning paradigms: GAD, GGAD, and TAGAD approaches. These categories represent common methodological frameworks in graph anomaly detection and include several state-of-the-art (SOTA) models, providing a broad basis for performance comparison under different assumptions and settings.

I.1 GAD Methods

- **BGNN** Ivanov & Prokhorenkova (2021) is an end-to-end framework that combines Gradient Boosted Decision Trees (GBDT) and Graph Neural Networks to model graph-structured data with node features, leading to notable improvements in node prediction performance.
- **BWGNN** Tang et al. (2022) employs spectrally and spatially localized band-pass filters to alleviate the anomaly-induced right-shift phenomenon, thereby improving the model's capability to recognize anomalous nodes.
- **GHRN** Gao et al. (2023a) addresses the heterogeneity issue in graph anomaly detection by accurately removing inter-class edges based on high-frequency graph signals.
- **DOMINANT** Ding et al. (2019) integrates graph convolutional networks with deep autoencoders to simultaneously recover graph topology and node attributes, thereby facilitating effective anomaly detection in attributed networks.
- **CoLA** Liu et al. (2021) derives anomaly scores by constructing contrastive pairs between nodes and their local neighborhood structures, together with GNN-based representation learning and anomaly-aware consistency assessment.
- **HCM-A** Huang et al. (2022) uses hop count prediction as a self-supervised learning objective, incorporates both local and global contextual information, and computes anomaly scores according to prediction errors and uncertainty.
- **TAM** Qiao & Pang (2023) improves anomaly detection and homophily learning by employing local affinity estimation and truncated graph optimization to reduce the adverse effects of heterophilous connections.
- **Semi-GGAD** Qiao et al. (2024) is a semi-supervised anomaly detection approach that constructs pseudo-anomalous nodes with predefined abnormal patterns and uses a limited number of normal nodes to optimize a one-class classifier.
- **CAGAD** Xiao et al. (2024) integrates counterfactual data augmentation with a graph pointer network to discover potential anomalies and form anomalous neighborhoods, thereby learning more discriminative node representations.
- **GCTAM** Zhang et al. (2025b) employs an anomaly truncation strategy that combines contextual information with global affinity, effectively reducing the misclassification caused by fixed-threshold truncation in conventional approaches.

I.2 GGAD Methods

- **ARC** Liu et al. (2024) is the first generalist graph anomaly detection framework designed for few-shot inference, enabling cross-graph transfer without retraining through feature alignment, residual representation learning, and context-aware anomaly scoring.
- **IA-GGAD** Zhang et al. (2025a) proposes an end-to-end zero-shot graph anomaly detection framework that jointly addresses feature-space and structural shifts through invariant representation learning and affinity modeling, enabling direct generalization to unseen graphs.
- **AnomalyGFM** Qiao et al. (2025a) proposes a foundation model for graph anomaly detection that learns prototype-aligned residual representations during pre-training, facilitating cross-graph transfer in both zero-shot and few-shot scenarios.

- **UNPrompt** Niu et al. (2025) develops a neighborhood-prompt-based zero-shot graph anomaly detection approach that enables a unified model to generalize across graphs by exploiting attribute predictability and cross-graph alignment.
- **OWLEYE** Zheng et al. (2026) proposes a zero-shot cross-domain graph anomaly detection framework that combines feature alignment, pattern dictionaries, and truncated reconstruction for anomaly detection on unseen graphs without retraining.
- **ProMoS** Xu et al. (2026) proposes an unsupervised generalist graph anomaly detection framework that transfers normality priors through prototype-guided distillation, allowing direct anomaly detection on novel graphs without further adaptation.

I.3 TAGAD Methods

- **CMUCL** Xu et al. (2025c) develops an end-to-end TAG anomaly detection framework that jointly models text and graph structure through cross- and uni-modal multi-scale contrastive learning with inconsistency-based scoring.
- **CoLL** Xu et al. (2025b) introduces an LLM-enhanced TAG anomaly detection framework that combines multi-LLM evidence generation with gated GNN modeling to capture anomaly-relevant semantics and high-order structure.

J Description of Datasets

In this section, we provide a detailed description of the datasets, with their statistics summarized in Table 9. Specifically, we consider fourteen real-world text-attributed graph datasets spanning three representative domains, including citation, e-commerce, and web networks. To evaluate cross-domain generalization, six datasets are designated as source graphs for training, while the remaining eight are treated as entirely unseen target graphs for zero-shot evaluation. To further provide diverse and realistic anomaly patterns, we retain existing anomaly annotations where available and construct additional contextual, structural, and domain-specific anomalies. This setting enables a comprehensive evaluation of the model's ability to learn transferable anomaly patterns from heterogeneous source domains and generalize to previously unseen TAGs. Detailed descriptions of each dataset are provided below.

- **Citation networks**: **Cora** McCallum et al. (2000), **CiteSeer** Sen et al. (2008), **Pubmed** Sen et al. (2008) and **Arxiv** Hu et al. (2020) are four widely used citation-network benchmarks, where nodes represent scientific publications and edges denote citation links. Cora and CiteSeer contain computer-science papers, Pubmed comprises biomedical publications primarily related to diabetes, and Arxiv consists of computer-science papers collected from arXiv. Each node is associated with textual content comprising the paper title and abstract, which is encoded using a pretrained text encoder to obtain a 768-dimensional vector.
- **E-commerce Networks**: **History**, **Children**, **Products**, **Grocery**, **Movies**, **Toys** and **Fitness** are Amazon product-network benchmarks, in which nodes represent products and edges denote co-purchase relationships. These datasets cover history books, children's books, general merchandise, grocery products, movie- and video-related products, toys and games, and sports and fitness products, respectively. Each node is associated with textual product information: History, Children, Products, Grocery, Movies, and Toys generally provide product titles and descriptions, while Fitness primarily provides product titles. We encode this textual content using a pretrained text encoder to obtain a 768-dimensional feature vector for each node.
- **Web Networks**: **WikiCS**, **Cornell** and **Texas** are three widely used web-page network benchmarks, where nodes represent web pages and edges denote hyperlinks between them. WikiCS is constructed from Wikipedia articles in the field of computer science, and its node text contains article content describing computing concepts, software systems, technologies, organizations, and other related topics. Cornell and Texas are derived from the websites of the computer science departments at Cornell University and the University of Texas at Austin, respectively. Their node text consists of web-page content concerning

faculty and students, courses, research projects, academic programs, and departmental information.

Table 9: Statistics of the 14 TAG datasets.

Dataset	Train	Test	#Nodes	#Edges	Avg. Length	Anomaly
Citation Networks						
Cora	✓	–	2,791	10,990	135.45	164
Citeseer	–	✓	3,262	3,684	153.94	192
Pubmed	–	✓	21,248	92,154	256.08	1,184
Arxiv	✓	–	176,514	1,225,722	179.70	10,162
E-commerce Networks						
History	✓	–	43,503	373,468	228.36	2,494
Children	✓	–	85,137	1,591,763	209.12	4,612
Products	✓	–	46,617	117,064	110.65	2,752
Grocery	–	✓	17,908	147,332	67.36	1,024
Movies	–	✓	17,625	166,690	81.85	1,000
Toys	–	✓	21,283	117,424	74.50	1,240
Fitness	–	✓	187,918	3,113,588	21.87	10,384
Web Networks						
WikiCS	✓	–	14,485	443,558	422.20	704
Cornell	–	✓	197	596	261.27	21
Texas	–	✓	192	608	192.47	21

Anomaly Construction. We construct anomalies at two complementary levels to capture both generic text-structure inconsistencies and domain-specific abnormal behaviors. Following CMUCL Xu et al. (2025c), we first introduce two types of contextual anomalies and two types of structural anomalies. Contextual anomalies are generated by transferring textual content from semantically dissimilar nodes. Specifically, we either insert unrelated content into an eligible field of a target node or replace one of its text spans with content drawn from a semantically mismatched source node. Throughout this process, field boundaries, such as titles, descriptions, and abstracts, are preserved to maintain the original document structure. Structural anomalies are introduced either by organizing selected nodes into dense four-node subgraphs or by connecting a target node to randomly sampled nodes. The number of newly added edges is sampled according to the empirical degree distribution of the corresponding graph. To prevent interference among different perturbation types, these anomalies are assigned to disjoint subsets of initially normal nodes. Beyond these generic perturbations, we further introduce domain-specific anomalies to better reflect representative abnormal behaviors encountered in real-world scenarios. In citation networks, we simulate off-topic or low-quality papers as well as citation-manipulated papers. In e-commerce networks, we construct misleading products associated with unrelated trends and products exhibiting artificially inflated co-purchase relations. In web and encyclopedia networks, we simulate fabricated entries and promotional content. Textual anomalies are introduced through field-preserving insertion or replacement, whereas structural anomalies are generated by incorporating external anomaly-source nodes that either form dense local groups with selected target nodes or connect to them as cross-context neighbors. Existing anomaly annotations are preserved, and the newly constructed anomaly labels are merged with the original annotations to form the final evaluation labels.

K Details of Implementation

K.1 Settings

Datasets. We evaluate the cross-domain generalization capability of **AnTAGAD** on 14 real-world text-attributed graph benchmarks spanning citation, e-commerce, and web graphs as shown in Table 9. Specifically, we use six graphs as source domains for training, denoted as $\mathcal{G}_{train} = \{$Cora, Arxiv, Children, Products, History, WikiCS$\}$, while the remaining eight are treated as entirely unseen target domains for zero-shot evaluation, i.e., $\mathcal{G}_{test} = \{$Citeseer, Pubmed, Grocery, Movies, Toys, Fitness, Cornell, Texas$\}$. We retain ground-truth anomalies where available and construct ad-

ditional anomalies following CMUCL Xu et al. (2025c). Appendix J provides more details on all datasets.

Baselines. We benchmark **AnTAGAD** against 18 representative baselines drawn from three closely related lines of research. The first group comprises conventional GAD approaches, including BGNN Ivanov & Prokhorenkova (2021), BWGNN Tang et al. (2022), GHRN Gao et al. (2023a), DOMINANT Ding et al. (2019), CoLA Liu et al. (2021), HCM-A Huang et al. (2022), TAM Qiao & Pang (2023), Semi-GGAD Qiao et al. (2024), CAGAD Xiao et al. (2024), and GCTAM Zhang et al. (2025b). The second group consists of generalist graph anomaly detection (GGAD) methods, namely ARC Liu et al. (2024), AnomalyGFM Qiao et al. (2025a), UNPrompt Niu et al. (2025), IA-GGAD Zhang et al. (2025a), OWLEYE Zheng et al. (2026), and ProMoS Xu et al. (2026). We also include two TAGAD-specific approaches, CMUCL Xu et al. (2025c) and CoLL Xu et al. (2025b), to compare against methods explicitly designed for text-attributed graphs. Appendix I provides implementation details for all baselines.

Evaluation Protocol. We report AUPRC and AUROC over five training seeds, with means and standard deviations. Models are trained on the six source graphs and evaluated on the eight target graphs without target-specific optimization. Feature adapters, hyperparameters, anchor selection, and stopping rules are fixed using source-domain information before target evaluation. No target labels or target-domain performance are used for model selection.

For $\mathcal{OG}$ diagnosis, we compute the final anomaly score of each method and use Eqs. 2–4 at $q \in \{0.90, 0.95, 0.99\}$, with $q = 0.95$ as the primary operating point. The same boundary-randomization rule is applied to tied scores from both classes. All methods use identical target graphs and paired training seeds. Results are reported separately for each target graph and then macro-averaged with equal graph weights. Target labels are used only for offline evaluation of AUROC, AUPRC, and $\mathcal{OG}$, and are never used to select hyperparameters, source anchors, or stopping epochs.

K.2 Hyper Parameters

The following search ranges are used for source-domain model selection. All hyperparameters are fixed before target-domain evaluation, and no target labels or target-domain performance are used during model selection.

- Hidden layer dimension: {64, 128, 256, 512, 1024}
- Number layer of Graph Encoder: {1, 2, 3, 4}
- Dropout rate: {0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9}
- Learning rate: floats between 10^{-6} and 10^{-3}
- Weight decay floats between 10^{-6} and 10^{-4}
- Anchor selection count: $K = 160$ per source graph, branch, and pseudo-class. After merging the textual and topological selections and removing duplicate indices, the final normal and anomaly anchor banks contain $M_0 = 1900$ and $M_1 = 1854$ anchors, respectively. The retrieval size in the complete model is $k = 20$.

K.3 Evaluation Metrics

AUROC measures anomalous–normal score ranking with half credit for tied pairs. AUPRC summarizes the anomalous-class precision–recall behavior and depends on anomaly prevalence Saito & Rehmsmeier (2015). $\mathcal{OG}$ additionally reports anomaly acceptance at a prescribed normal coverage using Eqs. 2–4. We report per-graph means and standard deviations across five training seeds, along with paired $\mathcal{OG}$ differences. Seed variation describes training randomness on these fixed graphs and is not a confidence interval over new graphs. We compute AUPRC using the non-interpolated average precision implementation in the scikit-learn library, treating anomalies as the positive class and using continuous anomaly scores as inputs. We repeat all methods with five different random seeds.

Table 10: Node anomaly detection performance in terms of AUROC (%). The best and second-best results are highlighted in bold and underlined, respectively. Methods marked with * are supervised methods.

Method	Movies	Texas	Grocery	Fitness	Pubmed	Toys	Citeseer	Cornell	Rank
GAD Methods									
BGNN* (2021)	56.27±4.00	53.05±7.14	53.19±2.30	50.51±2.17	43.56±12.80	51.94±2.43	55.23±5.72	57.74±9.96	14.00
BWGNN* (2022)	55.35±0.97	49.60±7.91	59.20±1.12	49.76±1.53	54.16±4.11	57.82±0.77	61.26±1.82	50.28±3.46	12.50
GHRN* (2023)	52.97±2.31	50.71±6.07	58.65±2.44	49.69±2.49	50.96±4.61	56.98±2.45	59.34±1.56	56.28±2.60	13.00
DOMINANT (2019)	55.87±1.07	67.96±1.19	55.89±1.01	67.66±0.78	56.24±1.96	58.38±0.92	54.18±1.11	66.77±0.87	10.13
CoLA (2021)	52.34±0.81	47.86±6.40	53.22±0.95	54.27±0.44	48.91±1.10	55.90±0.77	57.56±1.91	47.16±4.27	15.00
HCM-A (2022)	64.93±0.36	57.57±1.77	64.68±0.46	65.11±0.79	60.13±1.18	66.55±0.39	60.68±0.75	67.16±1.64	8.00
TAM (2023)	66.77±0.13	59.47±1.18	70.67±0.20	*OOM*	74.90±0.24	74.73±0.13	28.01±0.02	61.07±2.46	8.56
Semi-GGAD (2024)	42.93±6.33	42.62±6.75	42.81±7.73	43.08±7.35	35.87±6.46	44.74±6.77	46.51±14.23	36.75±5.28	18.63
CAGAD (2024)	56.32±3.27	59.69±6.49	55.67±1.61	48.14±6.08	42.11±6.44	56.18±2.60	59.56±4.00	54.41±2.69	12.88
GCTAM (2025)	64.34±0.26	67.30±2.50	65.50±0.56	*OOM*	70.76±0.84	68.64±0.54	49.60±1.11	63.93±2.29	9.56
GGAD Methods									
ARC* (2024)	68.16±0.74	75.66±1.07	70.95±0.36	77.93±0.61	75.22±1.03	74.22±1.01	73.03±0.01	81.25±1.94	2.63
IA-GGAD* (2025)	68.67±0.87	55.27±3.55	66.42±0.41	78.14±0.50	60.11±1.92	69.85±0.55	62.29±1.38	56.84±2.45	6.63
AnomalyGFM* (2025)	51.49±1.68	58.84±2.38	55.93±2.01	59.97±1.49	68.59±2.14	56.21±1.14	56.19±2.18	74.50±1.93	10.75
UNPrompt* (2025)	55.56±1.86	46.80±5.72	54.43±1.93	66.86±2.35	66.99±6.99	53.56±1.63	65.23±1.53	54.94±1.62	11.75
OWLEYE* (2026)	68.24±0.01	77.54±0.06	66.66±0.01	72.82±0.01	71.32±0.03	72.82±0.01	73.02±0.20	74.83±0.06	3.50
ProMoS (2026)	65.77±0.39	74.88±1.03	65.73±0.75	77.24±0.33	70.06±0.58	69.34±0.26	64.28±1.47	74.76±1.20	5.13
TAGAD Methods									
CMUCL (2025)	50.69±0.43	51.86±6.57	49.63±1.56	49.92±0.70	53.13±1.79	48.17±0.29	57.58±1.28	62.41±3.74	14.25
CoLL (2025)	59.80±2.46	45.55±2.49	57.32±2.09	60.02±1.65	56.26±2.33	58.94±0.86	55.79±1.91	51.20±2.83	12.13
Ours	**74.51±0.40**	**83.28±0.88**	**74.21±0.46**	**80.90±0.27**	**77.81±0.28**	**76.95±0.17**	**73.11±1.49**	**81.31±1.03**	**1.00**
Δ	↑ 5.84	↑ 5.74	↑ 3.26	↑ 2.76	↑ 2.59	↑ 2.22	↑ 0.08	↑ 0.06	–

Δ denotes the improvement (↑) or degradation (↓) compared with the current best baseline method. *OOM* means out of memory in our computational environment. Rank indicates the average ranking over eight datasets.

K.4 Experimental Environment

All experiments were run on a Linux server running Ubuntu 20.04. The system used a 13th Gen Intel(R) Core(TM) i7-12700 CPU, 64GB RAM, and an NVIDIA GeForce RTX A6000 GPU with 48GB of memory. We managed the software environment with Anaconda3, using PyCharm as the development interface. Experiments were conducted under Python 3.9.25, CUDA 11.8, and PyTorch 2.2.2.

L Performance Comparison of AUROC

As shown in Table 10, **AnTAGAD** achieves the highest AUROC on all eight unseen target graphs, with an average rank of 1.00, demonstrating consistently strong cross-domain anomaly ranking capability. Compared with the strongest baseline on each dataset, the method is particularly notable on Movies, Texas, Grocery, and Fitness. In contrast, conventional GAD methods show larger performance fluctuations across target graphs, while recent GGAD methods such as ARC, OWLEYE, and IA-GGAD remain relatively dataset-dependent. **AnTAGAD**, however, consistently achieves the best performance across citation, e-commerce, and web domains, indicating more stable cross-domain generalization. Notably, several GGAD baselines exploit supervised anomaly information, whereas **AnTAGAD** still attains the best AUROC on all target graphs. These results support strong ranking performance on the evaluated target graphs. They do not isolate the contribution of reference aggregation, and superiority in AUROC does not imply lower $\mathcal{OG}$ at every coverage. That attribution requires the controlled comparisons specified in the $\mathcal{OG}$ analysis.

M Algorithm and Complexity

M.1 Algorithmic Description

The overall workflow of **AnTAGAD** consists of training and inference phases (see Algorithm 1 and Algorithm 2). During training, the model first evaluates node-level textual reliability and selects high-confidence normal and anomaly anchors. It then encodes node texts into a shared representation space, constructs local and global contextual references, and jointly learns textual and topological anchors through complementary contrastive objectives. The two anchor sources are

further integrated into unified normal and anomaly anchor banks, enabling the model to preserve fine-grained node-specific deviations while learning transferable cross-domain patterns. During inference, **AnTAGAD** directly operates on unseen target graphs with frozen parameters, computing anomaly scores from the relative top-k similarities to normal and anomaly anchors together with local neighborhood dispersion. This design enables efficient zero-shot anomaly detection across different TAGs without target-domain retraining.

Algorithm 1 Training of **AnTAGAD**

Input: Source graphs $\mathcal{G}_{\text{train}}$; training epochs T; per-branch, per-class anchor count K.
Output: Trained model parameters θ; normal and anomaly anchor banks $\mathcal{A}^+, \mathcal{A}^-$.
Initialize model parameters θ.
for each graph $\mathcal{G} \in \mathcal{G}_{\text{train}}$ **do**
    Obtain textual reliability scores $\{c_i\}$ using Agents-Skills-Confidence via Eq. 6.
    Select textual normal and anomaly anchors $\bar{\mathcal{A}}^+, \bar{\mathcal{A}}^-$ via Eq. 7.
end for
for $t = 1, \ldots, T$ **do**
    for each graph $\mathcal{G} \in \mathcal{G}_{\text{train}}$ **do**
        Encode node texts into shared semantic representations $\mathbf{Z}$ via Eq. 8.
        Construct local and global contextual references $\mathbf{U}^{\circ}$, $\mathbf{U}^{*}$ via Eq. 9.
        Compute graph representations $\mathbf{H}$ incorporating local and global deviations via Eq. 10.
        Compute the textual anchor-guided loss $\mathcal{L}_{\bar{\mathcal{A}}}$ via Eq. 11.
        Select topological normal and anomaly anchors $\hat{\mathcal{A}}^+, \hat{\mathcal{A}}^-$ via Eq. 12.
        Compute the topological anchor-guided loss $\mathcal{L}_{\hat{\mathcal{A}}}$ via Eq. 13.
        Combine textual and topological anchors $\mathcal{A}^+, \mathcal{A}^-$ via Eq. 14.
        Compute the joint anchor loss $\mathcal{L}_{\mathcal{A}}$ via Eq. 15.
        Update model parameters by minimizing $\mathcal{L} = \mathcal{L}_{\bar{\mathcal{A}}} + \mathcal{L}_{\hat{\mathcal{A}}} + \mathcal{L}_{\mathcal{A}}$ via Eq. 16.
    end for
end for
Freeze the final encoder parameters θ.
for each source graph $\mathcal{G} \in \mathcal{G}_{\text{train}}$ **do**
    Recompute its representations with the frozen encoder.
    Select its final anchor indices using the specified source-only rules.
end for
Record $M_0 = |\mathcal{A}^+|$ and $M_1 = |\mathcal{A}^-|$.

Algorithm 2 Inference of **AnTAGAD**

Input: Target graphs $\mathcal{G}_{\text{test}}$; frozen model parameters θ; source anchors $\mathcal{A}^+, \mathcal{A}^-$.
Output: Anomaly scores $S(v_i)$ for all target nodes.
for each graph $\mathcal{G} \in \mathcal{G}_{\text{test}}$ **do**
    Compute $\mathbf{Z}$ using Eq. 8.
    Compute neighborhood and global references using Eq. 9.
    Compute $\mathbf{H}$ with the frozen graph encoder using Eq. 10.
    Use cosine similarity with the same normalization convention as the source banks.
    Use the fixed $k \leq \min(M_0, M_1)$ and λ chosen before target evaluation.
    for each node $v_i \in \mathcal{V}$ **do**
        Compute the mean top-k similarity between $\mathbf{h}_i$ and anomaly anchors in $\mathcal{A}^-$.
        Compute the mean top-k similarity between $\mathbf{h}_i$ and normal anchors in $\mathcal{A}^+$.
        Compute the local neighborhood dispersion using the convention in Eq. 17.
        Combine relative anchor similarity and neighborhood dispersion to obtain $S(v_i)$ via Eq. 18.
    end for
end for

M.2 Complexity Analysis

Let n, m denote target nodes and stored edges, d_z the semantic feature dimension, d_h the final graph embedding dimension, G denote the number of source graphs,and $M = M_0 + M_1$ the total number of source-bank entries. Denote text encoding and graph encoding costs by C_{text} and C_{GNN}, respectively. We use BGE-base-en-v1.5 as the text encoder. Let L_t denote the number of Transformer layers and T the maximum token length of node texts. The text encoding cost is $C_{\text{text}} = \mathcal{O}\big(nL_t(T^2 d_z + T d_z^2)\big)$, The graph encoder employs a two-layer graph attention network with four attention heads per layer. Under the current configuration, the feature dimensions of all layers are $\mathcal{O}(d_h)$, yielding a graph encoding cost of $C_{\text{GNN}} = \mathcal{O}\big(n d_h^2 + m d_h\big)$.

Training. For each source graph, neighborhood statistics cost $O((n+m)d_z)$ and sorting-based anchor selection costs $O(n \log n)$. For a branch b with p_b positive pairs and $K_{0,b}, K_{1,b}$ normal and anomaly reference counts, evaluating the stated contrastive objective costs $O((p_b + K_{0,b}K_{1,b})d_h)$. Summing this term over the textual, topological, and joint branches, add encoding costs over all training graphs and epochs. The training complexity is $C_{\text{train}} = \mathcal{O}\Big(T \sum_{g=1}^{G} \big[C_{\text{GNN},g} + n_g \log n_g + m_g \log m_g + d_h \sum_{b \in \mathcal{B}} (p_{g,b} + K_{0,g,b}K_{1,g,b})\big]\Big)$, After training, constructing the final source bank requires one additional gradient-free forward pass with the frozen graph encoder, with cost $C_{\text{bank}} = O\left(\sum_{g=1}^{G} C_{\text{GNN},g}\right)$. The one-time BGE preprocessing cost C_{text} is reported separately from the timed training procedure.

LLM preprocessing cost. We use the DeepSeek-V4-Flash API to score node texts and cache the resulting scores for subsequent training. The billing records attributable to our experiments contain 21,144 requests, consuming 20,221,610 cache-hit input tokens, 60,196,325 cache-miss input tokens, and 44,509,846 output tokens. At the recorded prices of CNY 0.02, 1.00, and 2.00 per million tokens, respectively, the cumulative API cost is

$$C_{\text{API}} = \frac{20,221,610 \times 0.02 + 60,196,325 + 44,509,846 \times 2}{10^6} \approx 149.62 \text{ CNY}.$$

The cached scores are reused across training epochs and repeated runs, avoiding repeated LLM calls.

Inference. Comparing every target embedding with both source banks costs $O(nMd_h)$. A conservative sorting-based bound for top-k retrieval is $O(nM \log M)$; an implemented selection algorithm may reduce this term. Thus a direct implementation costs

$$C_{\text{text}} + C_{\text{GNN}} + O((n+m)d_z + nMd_h + nM \log M)\,. \tag{33}$$

Bank storage is $O(Md_h)$. Materializing all similarities uses $O(nM)$ additional memory, which can be reduced to $O(bM)$ by scoring batches of b target nodes. Target inference uses frozen source parameters and banks without target LLM assessment or optimization.

M.3 Efficiency Analysis

Figure 11 and Table 11 summarize the measured time, memory use, and detection performance of **AnTAGAD** and representative baselines. Across the eight target graphs, **AnTAGAD** achieves average AUROC and AUPRC of 77.8% and 41.1%, respectively. The timing results should be interpreted within their measurement boundaries: the reported training time for **AnTAGAD** excludes one-time BGE text-feature extraction and LLM-based source-node assessment, whose outputs are cached and reused. The key observations are summarized as follows:

- **Measured Training Time and Memory.** The timed training stage of **AnTAGAD** takes 29.43 s and uses 13,724 MiB of memory. Its reported training time is lower than those of ARC (196.09 s), IA-GGAD (322.40 s), AnomalyGFM (515.90 s), and OWLEYE (609.90 s). Because preprocessing and timing boundaries may differ across methods, these figures do not establish an end-to-end training-time advantage.
- **Inference Memory.** **AnTAGAD** uses 2,156 MiB during the measured inference stage, compared with 20,130 MiB for ARC, 24,482 MiB for IA-GGAD, 6,490 MiB for OWLEYE, and 5,063 MiB for CMUCL. These measurements describe the memory used in the evaluated inference setup.

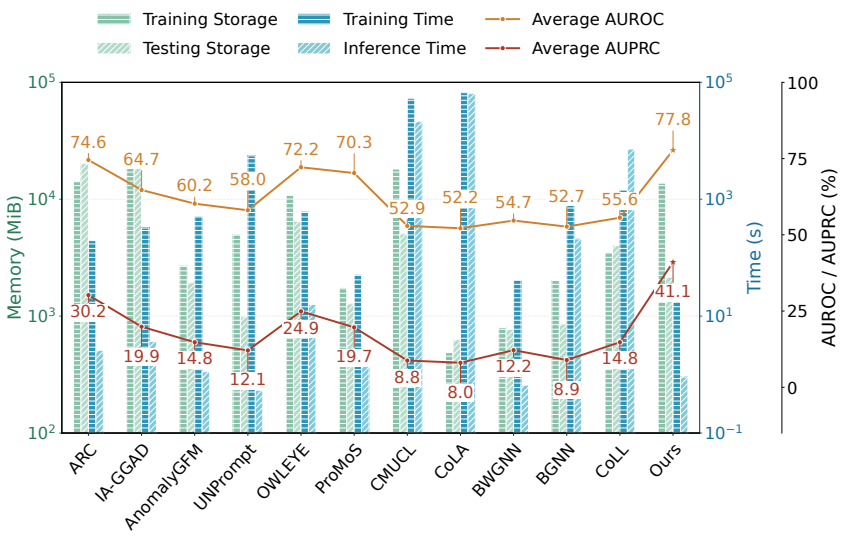


Figure 11: Space and time consumption on Training and inference.

- **Measured Inference Time.** The measured inference stage of **AnTAGAD** takes 0.91 s, compared with 2.61 s for ARC, 3.70 s for IA-GGAD, and 15.90 s for OWLEYE. UNPrompt and BWGNN have lower reported inference times, while **AnTAGAD** achieves higher average AUROC and AUPRC across the evaluated target graphs.
- **Future Optimizations:** Future work will focus on reducing the memory footprint of **AnTAGAD** through memory-efficient training strategies, such as mixed-precision computation, mini-batch/subgraph processing, and more compact anchor storage, while preserving its strong detection performance.

Table 11: Comparison of computational cost and detection performance. Memory is reported in MiB. Average AUROC and AUPRC are computed across eight target datasets. Bold entries indicate the best result among all methods.

Method	Training time (s)	Inference time (s)	Training memory (MiB)	Inference memory (MiB)	Average AUROC (%)	Average AUPRC (%)
ARC	196.09	2.61	14134	20130	74.58	30.23
IA-GGAD	322.40	3.70	18681	24482	64.70	19.88
AnomalyGFM	515.90	1.12	2697	1932	60.22	14.78
UNPrompt	5536.90	**0.54**	4957	985	58.05	12.10
OWLEYE	609.90	15.90	10795	6490	72.16	24.92
ProMoS	51.20	1.50	1760	1276	70.26	19.66
CMUCL	53481.29	21235.95	18143	5063	52.92	8.78
CoLA	66645.10	64815.20	**488**	**628**	52.15	8.05
BWGNN	41.10	0.66	799	770	54.68	12.18
BGNN	1125.10	210.20	2016	856	52.69	8.94
CoLL	1466.00	7240.00	3490	4008	55.61	14.78
AnTAGAD	**29.43**	0.91	13724	2156	**77.8**	**41.1**