Text Meets Topology : Rethinking Out-of-distribution Detection in Text-Rich Networks

Danny Wang, Ruihong Qiu, Guangdong Bai and Zi Huang
The University of Queensland
{danny.wang,r.qiu,g.bai,helen.huang}@uq.edu.au

Abstract

Out-of-distribution (OOD) detection remains challenging in ***text-rich networks***, where textual features intertwine with topological structures. Existing methods primarily address label shifts or rudimentary domain-based splits, overlooking the intricate textual-structural diversity. For example, in social networks, where users represent nodes with textual features (name, bio) while edges indicate friendship status, OOD may stem from the distinct language patterns between bot and normal users. To address this gap, we introduce the **TextTopoOOD** framework for evaluating detection across diverse OOD scenarios: **(1) attribute-level shifts** via text augmentations and embedding perturbations; **(2) structural shifts** through edge rewiring and semantic connections; **(3) thematically-guided label shifts**; and **(4) domain-based divisions**. Furthermore, we propose **TNT-OOD** to model the complex interplay between **T**ext a**N**d **T**opology using: **1) a novel cross-attention module** to fuse local structure into node-level text representations, and **2) a HyperNetwork** to generate node-specific transformation parameters. This aligns topological and semantic features of ID nodes, enhancing ID/OOD distinction across structural and textual shifts. Experiments on **11 datasets across four OOD scenarios** demonstrate the nuanced challenge of **TextTopoOOD** for evaluating OOD detection in text-rich networks. [1]

1 Introduction

Text-rich networks (TrN) have emerged as a powerful paradigm for representing the complex interplay between textual content and relational structures, serving as the lingua franca for modeling intricate real-world systems across diverse domains (Jin et al., 2023; Zou et al., 2023; Tang et al., 2024a,b). Despite the ubiquity of TrNs, machine learning approaches for these hybrid data structures often fail catastrophically when confronted with data that deviates from their training distribution (i.e., out-of-distribution (OOD)) (Gui et al., 2022; Li et al., 2022a; Wang et al., 2024b). To address this challenge, a surge of effective OOD detection techniques have been proposed (Zhao et al., 2020; Guo et al., 2023; Song and Wang, 2022; Chen et al., 2025; Liu et al., 2023b; Du et al., 2023), aiming to identify instances that fall outside the in-distribution (ID) training data.

Nevertheless, the OOD detection problem remains underexplored for TrNs (Lang et al., 2023; Cai et al., 2025; Yang et al., 2021). **Existing methods for OOD detection** in TrN learning primarily focus on **common distribution shifts** like random label-leave-out or temporal splits, largely ignoring the rich textual dimension that characterises real-world networks (Wang et al., 2024c; Xu et al., 2025a). This oversight is critical: in TrNs, semantic shifts in text may precipitate network changes. In product co-purchase networks, for example, shifting product descriptions and customer reviews often precede changes in purchasing patterns and product relationships - shifts that current OOD detection methods fail to capture.

To address this critical gap, we introduce **TextTopoOOD**, a comprehensive evaluation framework for TrNs. Unlike previous benchmarks (Xu et al., 2025b; Wang et al., 2024c), TextTopoOOD explores multiple dimensions of distribution shifts through: **1) Attribute-level modifications** that simulate semantic drift in textual content; **2) Structural alterations** that capture diverse connectivity patterns; **3) Thematically-guided label shifts** that extends beyond random selection; and **4) Domain-based splits** based on dataset's unique properties.

Additionally, prior OOD detection models for structural data typically add post-hoc scoring functions to the classifier (Wu et al., 2023; Yang et al., 2024; Ma et al., 2024). While these methods implic-

[1] Code is available at https://github.com/DannyW618/TNT.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 5494–5523
November 4-9, 2025

itly integrate text and structure through end-to-end training, the shared projection heads and uniform objectives (Hendrycks and Gimpel, 2017; Liu et al., 2020) often fail to capture distribution shifts that manifest differently across modalities and nodes.

In light of this, we propose **TNT-OOD**, a novel framework that models the interplay between **T**ext a**N**d **T**opology. TNT-OOD consists of three key components: **(1) a structure encoder** that learns structure-aware representations from local neighborhoods, **(2) a cross-attention mechanism** that fuses structure-derived and text-derived features to produce contextually grounded representations, and **(3) a HyperNetwork** that generates effective projection parameters to align the fused representations in a contrastive embedding space. This enables TNT-OOD to model heterogeneity at the node level, capturing textual-topological interactions that static encoders or global projection heads come short. Thereby, improving the separability of ID and OOD data when we encounter misalignment from the textual and topological components. We empirically evaluate TNT-OOD on our TextTopoOOD framework with diverse datasets and OOD scenarios, demonstrating consistent improvements over baselines. Our contributions are:

- We introduce **TextTopoOOD**, the first comprehensive framework for evaluating OOD detection in text-rich networks that captures the interplay between textual features and network topology.
- We propose **TNT-OOD**, an effective TrN OOD detector that aligns textual and structural representations using HyperNetwork.
- We validate the **efficacy of TNT-OOD for OOD detection** and the **challenging nature of TextTopoOOD** via 11 TrNs and four OOD scenarios.

2 Preliminaries

A **Text-Rich Network (TrN)** is denoted as $\mathcal{G} = (\mathcal{V}, \mathcal{T}, \mathbf{A})$, where $\mathcal{V} = \{v_1, \ldots, v_n\}$ is the set of n nodes. Each node v_i has a textual content $t_i \in \mathcal{T}$, encoded into a feature vector $\mathbf{x}_i \in \mathbb{R}^d$ through a text encoder (e.g., SBERT) $f_{\text{encoder}} : \mathcal{T} \to \mathbb{R}^d$. This forms the feature matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^T \in \mathbb{R}^{n \times d}$. The network topology is captured by an adjacency matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, where $A_{ij} = 1$ indicates a connection between nodes v_i and v_j. The node labels $\mathbf{Y} = \{y_1, \ldots, y_n\}$ assign each node to one of C classes, with $y_i \in \{1, \ldots, C\}$. This paper studies TrNs with semantic relationships between nodes (e.g., **in e-commerce networks**, nodes represent products with textual descriptions, connected by co-purchase relationships). ***In this paper, we explore two objectives:***

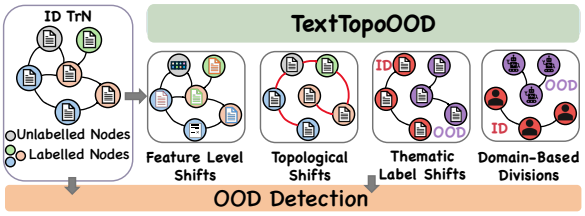


Figure 1: Overview of TextTopoOOD framework.

Objective 1: In-distribution Class Classification. Given the training and test nodes sharing the same distribution, where $P_{train}(\mathcal{G}_{train}) = P_{test}(\mathcal{G}_{test})$, and the conditional distribution $P_{train}(\mathbf{y}|\mathcal{G}_{train}) = P_{test}(\mathbf{y}|\mathcal{G}_{test})$, we aim to design a **classifier** g to accurately predict the label $\mathbf{y} \in \mathbb{R}^n$ for the test nodes, such that: $\mathbf{y} = \text{Softmax}(g(\mathcal{G}_{train}))$.

Objective 2: Node-Level Out-of-distribution Detection. The goal of OOD detection is to detect nodes with a different distribution to the training data at test time. Typically, we consider OOD shifts as $P_{train}(\mathcal{G}_{train}) \neq P_{test}(\mathcal{G}_{test})$, and $P_{train}(\mathbf{y}|\mathcal{G}_{train}) \neq P_{test}(\mathbf{y}|\mathcal{G}_{test})$. The task is to formulate a **detector** F with an **OOD scoring function** S and a given threshold τ_{thresh}, such that:

$$F(\mathcal{G}; g) = \begin{cases} \text{OOD}, & S(\mathcal{G}; g) \geq \tau_{\text{thresh}}, \\ \text{ID}, & S(\mathcal{G}; g) < \tau_{\text{thresh}}. \end{cases} \quad (1)$$

Energy-Based OOD Detection. For a node v_i in a TnR, the energy score is defined as:

$$S(\mathcal{G}_{v_i}; g) = e_i = -\log \sum_{c=0}^{C-1} \exp(\mathbf{z}_{i,c}), \quad (2)$$

where $e_i \in \mathbb{R}$ is the energy score, and $\mathbf{z}_i \in \mathbb{R}^C$ are the logits from the classifier $\mathbf{Z} = g(\mathcal{G})$.

To leverage topology in OOD detection, **energy propagation** is proposed with (Wu et al., 2023):

$$\mathbf{e}^{(k)} = \alpha \mathbf{e}^{(k-1)} + (1 - \alpha)\mathbf{D}^{-1}\mathbf{A}\mathbf{e}^{(k-1)}, \quad (3)$$

where $\mathbf{e}^{(k)}$ represents node energy scores after k propagation steps, α controls energy concentration, and $\mathbf{D}$ is the degree matrix, where $\mathbf{D}_{ii} = \sum_{j=1}^{n} \mathbf{A}_{ij}$ and $\mathbf{D}_{ij} = 0$ for $i \neq j$.

3 TextTopoOOD Framework

This section introduces **TextTopoOOD**, a comprehensive framework that explores diverse OOD sce-

Original Text	Synon. ($\alpha = 0.5, p_{char} = 0.3$)	Anton. ($\alpha = 0.3, p_{char} = 0.3$)
Recent advances in natural language processing have enabled more efficient text encoding for downstream applications.	Curret advanczes in normal language handlin have faacilitated more faster text representation for dowsntream tasks.	Ancient advances in ntaural language aprocessing have disabled more inefficient text encding for upstream appications.
I tried both the new and old reddit design and its like...	I tried either th fres and old rdedit plaan and its similar...	I tried bth the old and young reddit design and its different...

Table 1: Example of text-level feature shift created by TEXTAUGMENT.

narios in text-rich networks. TextTopoOOD analyses OOD scenarios across four dimensions: **(1) attribute-level shifts**, **(2) structural shifts**, **(3) thematically-guided label shifts**, and **(4) domain-based divisions** as shown in Figure 1.

3.1 Attribute-Level Shifts

Attribute-level shifts targets textual or feature components, challenging models to detect semantic changes while preserving network structure. For example, in product networks, descriptions might use new marketing terminology ('eco-friendly' $\rightarrow$ 'sustainable') while co-purchase relationships remain unchanged - requiring models to recognise semantic evolution beyond structural cues. We provide the following attribute-level shift scenarios.

1) Text Augmentation Shift. We generate text augmentation shifts by modifying the nodes' raw textual content. Given a node's original text t_i, we produce a perturbed version $\tilde{t}_i$ using controlled semantic transformations:

$$\tilde{t}_i = \text{TEXTAUGMENT}(t_i, \text{type}, \alpha_{\text{text}}, p_{\text{char}}) \quad (4)$$

where type $\in \{\text{synonym}, \text{antonym}\}$ determines the semantic direction of the shift, and $\alpha_{\text{text}} \in [0, 1]$ controls the noise level or percentage of words modified, p_{char} indicates the probability of character-level edits (insertion, deletion, replacements, swaps). We provide the pseudo-code of the TEXTAUGMENT function in Appendix G.

This approach creates OOD text with preserved syntactic structure but varied semantics. Synonyms challenge detection by maintaining textual meaning while altering word distributions from training data. Antonyms create more pronounced shifts by inverting meanings while preserving grammar, producing contextual inconsistencies. Character-level noise simulates typical human typographical errors. Examples are shown in Table 1 and Appendix C.

2) Feature Mixing Shift. Following prior works, we implement feature mixing to manipulate the encoded embeddings beyond raw texts (Wu et al., 2023). For each encoded node feature vector $\mathbf{x}_i$, we generate a perturbed version:

$$\tilde{\mathbf{x}}_i = (1 - \alpha_{\text{feat}}) \cdot \mathbf{x}_i + \alpha_{\text{feat}} \cdot \mathbf{m}_i, \quad (5)$$

where $\mathbf{m}_i = w \cdot \mathbf{x}_j + (1 - w) \cdot \mathbf{x}_k$ is a convex combination of features from randomly selected nodes j and k, with $w \sim \text{Uniform}(0, 1)$. The parameter α_{feat} controls shift intensity, with higher values creating embeddings that increasingly deviate from the original distribution to become OOD.

3.2 Structural Shifts

Structural shifts alter network connectivity while preserving node attributes, testing models' ability to detect topological changes. In product networks, co-purchase relationships evolve with changing consumer preferences (sustainable items becoming frequently co-purchased), while product descriptions remain unchanged - creating distribution shifts that are topological than semantic.

1) Structure Rewiring. We implement structure shifts using stochastic block models (SBM) (Abbe, 2017) to generate alternative edge distributions that reflect different community structures. Given the original network $\mathcal{G}_{\text{ID}}$ with edge density ρ, we generate an SBM network with node classes (determined by labels $\mathbf{Y}$) as blocks. The probability matrix is defined as:

$$P_{ij} = \begin{cases} p_{ii} = \rho \cdot f_{ii} & \text{if } i = j \\ p_{ij} = \rho \cdot f_{ij} & \text{if } i \neq j \end{cases} \quad (6)$$

where f_{ii} and f_{ij} are scaling factors that control intra-block and inter-block connectivity. The resulting SBM adjacency matrix $\mathbf{A}_{\text{SBM}}$ are then interpolated with the original structure $\mathbf{A}$ following:

$$\mathbf{A}_{\text{OOD}} = (1-\beta) \cdot \text{SAMPLE}(\mathbf{A}) \cup \beta \cdot \text{SAMPLE}(\mathbf{A}_{\text{SBM}}) \quad (7)$$

where $\text{SAMPLE}(\cdot)$ selects edges proportionally to maintain original graph density, and β controls the shift intensity.

This alters community structures while preserving textual content. In citation networks, this resembles when research papers maintain their topics but form new citation patterns with previously unrelated work, controlled via f_{ii} and f_{ij} parameters to reinforce or contradict expected content-connectivity relationships.

2) Semantic Connection Shift. The semantic connection shift reconnects the network based on node feature similarities, creating a correlation between textual semantics and connectivity patterns:

$$\mathbf{A}_{\text{OOD}} = \text{TOPK}(S(\mathbf{X}, \mathbf{X}), k, \text{MODE}) \quad (8)$$

where $S(\mathbf{X}, \mathbf{X})$ is a pairwise cosine similarity matrix between encoded text representations (e.g., $s_{ij} = \text{CosSim}(\mathbf{x}_i, \mathbf{x}_j)$, and k determines the number of edges to select (i.e.,same density as the original network). The selection MODE $\in \{\text{top}, \text{bottom}, \text{threshold percentile}\}$ determines which similarity pairs connect.

3) Text Swap Shift. The text swap shift introduces semantic inconsistencies by exchanging text between nodes, implicitly affecting the structural patterns. This creates a mismatch between node content and network position, challenging models to detect contextual incongruities. We formalise this as a controlled feature permutation operation:

$$\tilde{\mathbf{X}} = \mathbf{P}_{\beta_{\text{swap}}} \cdot \mathbf{X}, \quad (9)$$

where $\mathbf{P}_{\beta_{\text{swap}}}$ is a permutation matrix to swaps features between a proportion $\beta_{\text{swap}} \in [0, 1]$ of nodes.

As shown in Algorithm 3 in Appendix, we offer three variants to **swap texts between**: **1) Intra-class:** Nodes of the same class; **2) Inter-class:** Nodes of different classes; **3) Random:** Any nodes. This simulates real-world scenarios such as hijacked social media accounts posting out-of-character content, citation networks with papers incorrectly categorised, or product listings with mismatched descriptions. The progressive severity of misalignment allows systematic evaluation of model sensitivity to text-structure inconsistencies.

3.3 Label Shift with Thematic Guidance

We extend the traditional label shift approach beyond random selection by incorporating thematic analysis through large language models (LLMs) (i.e., Claude) (Anthropic, 2025). Given the complete set of classes $\mathcal{C}$, we create OOD datasets by withholding specific numbers of classes ($|\mathcal{C}_{\text{OOD}}| \approx 10\% \sim 40\%$ of $|\mathcal{C}|$) according to their thematic relationships:

$$\mathcal{C}_{\text{ID}} = \mathcal{C} \setminus \mathcal{C}_{\text{OOD}}. \quad (10)$$

We implement 3 strategies for OOD class selection:

1) Random selection: Classes are randomly designated as OOD, serving as a baseline approach.

Beyond this, we use LLMs to analyse the class names and descriptions to selects OOD classes by:

2) Thematic similarity: Separating classes that are pairwise similar into ID and OOD, creating scenarios where there exists similar label concepts.

3) Thematic dissimilarity: Grouping classes that most dissimilar to ID classes as OOD, creating scenarios with clearer semantic boundaries between ID and OOD.

For example, closely related topics (i.e., ML & Theory) may share terminology and citation patterns, making them more challenging to distinguish than completely unrelated domains (i.e., ML & Music). The LLM prompt is provided in Appendix C.

3.4 Domain-Based Divisions

Text-rich networks spans diverse domains, with unique properties that could naturally be used to devise OOD data. An example domain-based division using **temporal information** is as follows:

1) Temporal Shift. For datasets with temporal information (e.g., arXiv citation network), we create domain shifts based on publication time:

$$\mathcal{V}_{\text{ID}} = \{v_i \in \mathcal{V} \mid \text{year}(v_i) \in r_{\text{ID}}\} \quad (11)$$

$$\mathcal{V}_{\text{OOD}} = \{v_i \in \mathcal{V} \mid \text{year}(v_i) \in r_{\text{OOD}}\}, \quad (12)$$

where $r_{\text{ID}}, r_{\text{OOD}}$ are disjoint time ranges. This approach creates a series of OOD instances representing increasingly distant future data, simulating the natural evolution of text-rich networks over time.

3.5 Comparison to Related Benchmarks

To highlight TextTopoOOD's novelty, we contrast it with existing benchmarks. OOD-TAG (Wang et al., 2024c) introduced shifts that are derived from common citation network properties like node degree, temporal splits, and word diversity. Recent works like GLIP (Xu et al., 2025b) and GSync-OOD (Xu et al., 2025a) focus solely on random label shifts, lacking comprehensive scenario coverage. In contrast, TextTopoOOD provides a principled evaluation framework that **explicitly targets both textual and topological dimensions**

across diverse network types, including citation, e-commerce, knowledge, and social networks.

4 TNT-OOD Methodology

To address the complex interplay between text and topology, we propose TNT-OOD: a HyperNetwork-enhanced, cross-attentional model for OOD detection in text-rich networks. As shown in Figure 2, the architecture comprises: **1) a GCN-based structure encoder** for structure-aware representation, **2) a Cross-Attention module** that adaptively fuses textual and graph context, and **3) a novel HyperNetwork Projection Head** that generates dedicated weights conditioned on fused representations, enabling heterogeneous alignment.

1) Structure and Textual Encoding. Let $\mathbf{x}_i \in \mathbb{R}^d$ denote the frozen textual embedding of node v_i (e.g., via SBERT (Reimers and Gurevych, 2019)). The structure encoder is a (L)-layer Graph Convolutional Network (GCN) that transforms $\mathbf{x}_i$ via neighborhood aggregation (Kipf and Welling, 2017). The layer-wise update with n nodes is:

$$\mathbf{g}^{(l)} = \sigma\left(\hat{\mathbf{A}}\mathbf{g}^{(l-1)}\mathbf{W}^{(l)}\right), \quad l = 1, \ldots, L, \quad (13)$$

with $\mathbf{g}^{(0)} = \mathbf{X} \in \mathbb{R}^{n\times d}$, and $\hat{\mathbf{A}}$ is the symmetrically normalised adjacency matrix with self-loops, $\mathbf{W}^{(l)}$ are trainable weights for layer l. This produces structural-aware embeddings $\mathbf{g} \in \mathbb{R}^{n\times d_p}$, where d_p is the projection dimension.

2) Cross-Attentional Fusion. To infuse textual embeddings with localised graph-aware context, we introduce a **neighborhood-based cross-attention module**. For each node v_i, we aggregate textual information from its neighbors $\mathcal{N}(i)$ conditioned on the learned structural representations:

$$\mathbf{q} = \mathbf{W}_q\mathbf{g}, \quad \mathbf{k} = \mathbf{W}_k\mathbf{x}, \quad \mathbf{v} = \mathbf{W}_v\mathbf{x}, \quad (14)$$

$$\mathbf{z}_i = \mathbf{x}_i + \sum_{j\in\mathcal{N}(i)} \text{Softmax}_j\left(\frac{\langle \mathbf{q}_i, \mathbf{k}_j\rangle}{\sqrt{d_z}}\right)\cdot \mathbf{v}_j, \quad (15)$$

where $\mathbf{W}_q$, $\mathbf{W}_k$, and $\mathbf{W}_v$ are learnable projection matrices. This yields a fused representation $\mathbf{z}_i$ with structure-aware attention over neighbour content.

3) HyperNetwork-based Projection. Furthermore, to capture the difference between ID and OOD samples, we employ a HyperNetwork to generate *node-specific projections* from the fused text-structure features. The motivation is that using projection weights learned from ID data would expose misalignment when transforming OOD nodes.

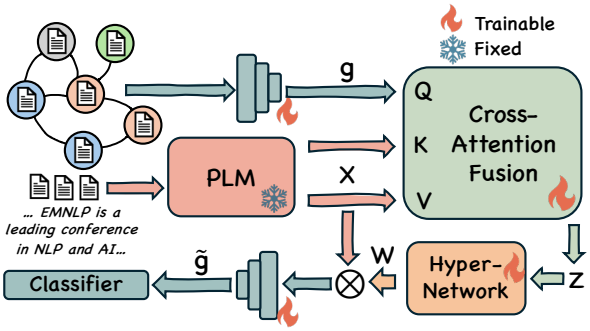


Figure 2: TNT-OOD framework: 1) Structure encoders and Classifier; 2) Text-encoder PLM; 3) Cross-attention fusion module; 4) HyperNetwork projection.

The HyperNetwork is defined as:

$$\mathbf{W}_i = \text{MLP}_{\text{hyper}}(\mathbf{z}_i), \quad (16)$$

where $\mathbf{W}_i \in \mathbb{R}^{d_p\times d}$ is a node-specific weight with projection dimension d_p. The projected text representation for each node $\mathbf{p}_t^i \in \mathbb{R}^{d_p}$ is given by:

$$\mathbf{p}_t^i = \mathbf{W}_i\mathbf{x}_i. \quad (17)$$

To ensure efficiency when working with large scale networks (i.e., over 100,000 nodes), we generate a factorised low-rank representation:

$$(\mathbf{L}_i, \mathbf{R}_i) = \text{LowRankHyper}_{\text{MLP}}(\mathbf{z_i}), \quad (18)$$

where $\mathbf{L}_i \in \mathbb{R}^{d_p\times r}$ and $\mathbf{R}_i \in \mathbb{R}^{r\times d}$ are the left and right factors respectively node v_i, r is the low-rank parameter ($r \ll \min(d_p, d)$). The projected text representation is computed efficiently using:

$$\mathbf{p}_t^i = \mathbf{L}_i \cdot (\mathbf{R}_i \cdot \mathbf{x}_i). \quad (19)$$

This approach reduces memory cost from $O(d_p\times d)$ to $O(d_p \times r + r \times d)$, enable efficient training. The projected representation is further passed through a GCN layer to re-enable structure-aware learning:

$$\tilde{\mathbf{g}} = \text{GCN}_{\text{fuse}}(\mathbf{P}_t, \mathbf{A}) \quad (20)$$

4.1 Contrastive Objective and Classification

To align text and structure-aware embeddings in a shared space, we use a symmetric contrastive loss:

$$\hat{\mathbf{P}}_t = \mathbf{P}_t/\|\mathbf{P}_t\|, \hat{\mathbf{g}} = \tilde{\mathbf{g}}/\|\tilde{\mathbf{g}}\| \quad (21)$$

$$\mathcal{L}_{\text{cont}} = \frac{1}{2}\left(\text{CE}(\hat{\mathbf{P}}_t\hat{\mathbf{g}}^\top/\tau, \mathbf{I}) + \text{CE}(\hat{\mathbf{g}}\hat{\mathbf{P}}_t^\top/\tau, \mathbf{I})\right), \quad (22)$$

where $\hat{\mathbf{P}}_t, \hat{\mathbf{g}} \in \mathbb{R}^{n\times d_p}$, τ is the temperature and $\mathbf{I}$ is the identity target. For ID classification, we include an additional classification layer using $\tilde{\mathbf{g}}$:

$$\mathcal{L}_{\text{cls}} = \text{CE}(\text{GCN}_{\text{cls}}(\tilde{\mathbf{g}}_i), y_i). \quad (23)$$

	Metrics	*Citation Networks* Cora	Citeseer	Arxiv	DBLP	PubMed	*Knowledge & Social Networks* Reddit	WikiCS	*E-commerce Networks* Bookhis	Bookchild	Elephoto	Elecomp
Maha	AUROC (↑)	41.74 ± 9.24	50.63 ± 24.39	67.75 ± 35.39	58.47 ± 24.67	55.32 ± 20.70	42.49 ± 19.65	52.40 ± 24.19	66.65 ± 29.71	68.96 ± 31.61	62.03 ± 26.03	43.97 ± 16.64
	AUPR (↑)	48.75 ± 4.20	55.41 ± 13.37	78.75 ± 28.41	55.74 ± 12.53	54.87 ± 14.89	51.84 ± 10.14	66.68 ± 13.61	89.08 ± 11.22	88.51 ± 12.73	86.97 ± 10.21	87.01 ± 4.18
	FPR95 (↓)	92.15 ± 4.36	87.11 ± 15.16	**56.92 ± 44.11**	75.73 ± 22.72	88.13 ± 8.76	88.77 ± 5.55	82.65 ± 15.50	70.22 ± 33.94	59.06 ± 42.84	76.87 ± 34.46	89.94 ± 13.53
	ID ACC (↑)	86.21 ± 5.48	79.34 ± 7.47	68.18 ± 0.51	77.41 ± 0.21	78.23 ± 0.41	61.85 ± 2.32	82.22 ± 4.18	83.95 ± 0.89	58.52 ± 5.87	82.71 ± 5.81	82.58 ± 7.93
MSP	AUROC (↑)	74.63 ± 5.05	79.33 ± 9.67	70.26 ± 14.51	84.34 ± 5.59	70.49 ± 5.97	47.69 ± 12.05	71.30 ± 4.72	69.65 ± 11.95	64.49 ± 13.06	70.57 ± 12.81	69.05 ± 15.22
	AUPR (↑)	75.58 ± 3.52	81.19 ± 8.66	81.19 ± 21.03	83.68 ± 6.55	66.57 ± 5.09	55.35 ± 6.41	78.41 ± 3.47	90.44 ± 4.63	85.95 ± 4.38	89.69 ± 4.30	94.20 ± 2.79
	FPR95 (↓)	64.27 ± 9.83	61.97 ± 23.93	74.53 ± 23.04	51.13 ± 13.55	70.75 ± 11.72	90.84 ± 3.03	68.15 ± 11.83	68.58 ± 19.97	71.27 ± 25.22	72.90 ± 19.65	70.18 ± 30.06
	ID ACC (↑)	86.21 ± 5.48	79.34 ± 7.47	68.18 ± 0.51	77.41 ± 0.21	78.23 ± 0.41	61.85 ± 2.32	82.22 ± 4.18	83.95 ± 0.89	58.52 ± 5.87	82.71 ± 5.81	82.58 ± 7.93
ODIN	AUROC (↑)	73.71 ± 4.60	78.96 ± 9.13	69.73 ± 14.99	83.44 ± 6.94	70.65 ± 4.12	46.86 ± 16.17	63.28 ± 4.05	68.27 ± 12.01	60.33 ± 13.64	70.14 ± 13.89	69.09 ± 18.31
	AUPR (↑)	74.75 ± 3.49	81.09 ± 7.94	80.96 ± 21.06	82.25 ± 8.63	66.73 ± 3.86	55.04 ± 8.17	71.81 ± 3.29	89.90 ± 4.73	84.10 ± 4.41	89.74 ± 4.72	**94.41 ± 3.45**
	FPR95 (↓)	66.01 ± 8.61	63.32 ± 22.88	75.03 ± 24.10	51.82 ± 14.53	72.73 ± 8.88	89.06 ± 4.73	75.29 ± 7.97	70.21 ± 19.46	74.77 ± 23.13	74.65 ± 20.62	68.95 ± 33.84
	ID ACC (↑)	86.21 ± 5.48	79.34 ± 7.47	68.18 ± 0.51	77.41 ± 0.21	78.23 ± 0.41	61.85 ± 2.32	82.22 ± 4.18	83.95 ± 0.89	58.52 ± 5.87	82.71 ± 5.81	82.58 ± 7.93
NECO	AUROC (↑)	70.77 ± 6.35	68.35 ± 12.47	75.19 ± 14.92	82.89 ± 6.81	66.63 ± 6.48	47.67 ± 10.92	69.09 ± 4.78	72.70 ± 12.18	69.15 ± 20.52	73.65 ± 14.38	67.38 ± 17.91
	AUPR (↑)	73.59 ± 6.67	70.13 ± 11.05	83.40 ± 21.45	82.33 ± 8.48	64.57 ± 5.01	55.38 ± 6.25	77.81 ± 3.77	91.74 ± 4.07	88.38 ± 7.98	91.17 ± 4.80	93.67 ± 3.35
	FPR95 (↓)	73.57 ± 5.56	73.31 ± 19.33	68.32 ± 28.11	52.69 ± 11.63	77.73 ± 9.29	91.08 ± 2.65	74.05 ± 9.64	68.27 ± 20.82	65.35 ± 35.14	70.66 ± 24.65	69.54 ± 32.75
	ID ACC (↑)	86.21 ± 5.48	79.34 ± 7.47	68.18 ± 0.51	77.41 ± 0.21	78.23 ± 0.41	61.85 ± 2.32	82.22 ± 4.18	83.95 ± 0.89	58.52 ± 5.87	82.71 ± 5.81	82.58 ± 7.93
Energy	AUROC (↑)	81.60 ± 3.70	80.11 ± 9.89	**76.96 ± 15.51**	87.31 ± 6.04	71.18 ± 8.84	49.51 ± 6.48	73.57 ± 6.85	**74.03 ± 12.44**	70.25 ± 22.36	**74.18 ± 15.15**	66.93 ± 18.00
	AUPR (↑)	83.74 ± 2.64	81.19 ± 8.64	**84.27 ± 21.89**	**86.73 ± 7.28**	68.03 ± 7.30	55.73 ± 3.36	80.74 ± 5.43	**92.42 ± 3.88**	89.15 ± 9.21	**91.68 ± 5.17**	93.34 ± 3.28
	FPR95 (↓)	59.41 ± 12.56	61.22 ± 27.24	66.28 ± 31.54	46.63 ± 17.56	71.40 ± 13.05	87.85 ± 2.70	66.05 ± 13.87	64.10 ± 22.10	64.01 ± 37.25	70.43 ± 25.61	68.50 ± 21.34
	ID ACC (↑)	86.21 ± 5.48	79.34 ± 7.47	68.18 ± 0.51	77.41 ± 0.21	78.23 ± 0.41	61.85 ± 2.32	82.22 ± 4.18	83.95 ± 0.89	58.52 ± 5.87	82.71 ± 5.81	82.58 ± 7.93
	w/ Prop											
GNNSafe	AUROC (↑)	88.45 ± 4.02	83.29 ± 7.54	37.13 ± 21.81	88.70 ± 4.51	84.33 ± 3.46	57.43 ± 26.51	85.51 ± 10.39	57.33 ± 24.26	67.06 ± 20.56	57.38 ± 14.85	59.95 ± 2.55
	AUPR (↑)	87.77 ± 5.69	79.63 ± 7.77	66.30 ± 13.65	81.71 ± 3.37	80.48 ± 4.54	65.51 ± 14.82	87.49 ± 6.28	84.72 ± 8.22	85.37 ± 6.40	83.08 ± 5.25	89.56 ± 2.69
	FPR95 (↓)	36.06 ± 7.21	43.90 ± 19.75	88.51 ± 8.10	30.46 ± 16.36	39.19 ± 7.12	68.64 ± 15.86	38.73 ± 24.16	65.47 ± 24.12	55.16 ± 34.01	66.26 ± 16.76	64.41 ± 18.20
	ID ACC (↑)	86.21 ± 5.48	79.34 ± 7.47	68.18 ± 0.51	77.41 ± 0.21	78.23 ± 0.41	61.85 ± 2.32	82.22 ± 4.18	83.95 ± 0.89	58.52 ± 5.87	82.71 ± 5.81	82.58 ± 7.93
NODESafe	AUROC (↑)	86.69 ± 6.19	82.09 ± 6.22	38.06 ± 22.48	89.81 ± 3.55	84.66 ± 3.06	54.08 ± 32.33	80.54 ± 24.84	53.31 ± 23.37	57.35 ± 12.69	57.24 ± 9.81	60.87 ± 4.05
	AUPR (↑)	86.11 ± 7.82	77.94 ± 6.27	67.54 ± 12.30	82.59 ± 2.73	80.61 ± 4.21	61.95 ± 20.25	84.48 ± 14.03	83.22 ± 8.18	81.55 ± 4.44	83.24 ± 4.40	90.04 ± 3.47
	FPR95 (↓)	38.54 ± 9.50	45.32 ± 17.70	87.58 ± 8.75	27.04 ± 12.58	38.85 ± 6.10	77.57 ± 28.13	37.75 ± 34.39	73.55 ± 18.97	70.34 ± 21.05	68.73 ± 12.63	60.28 ± 15.51
	ID ACC (↑)	84.94 ± 4.43	79.30 ± 6.62	67.77 ± 1.87	76.82 ± 0.40	78.90 ± 0.14	61.69 ± 0.67	80.92 ± 4.00	83.79 ± 1.26	58.21 ± 5.66	80.09 ± 7.99	80.85 ± 7.96
TNT-OOD	AUROC (↑)	**91.29 ± 4.21**	**86.42 ± 6.53**	47.65 ± 13.80	**89.59 ± 3.50**	**88.82 ± 3.22**	**70.61 ± 37.62**	**89.88 ± 8.19**	72.63 ± 15.71	**79.03± 5.09**	70.09 ± 5.68	**69.67 ± 6.93**
	AUPR (↑)	**89.07± 5.13**	**82.48 ± 6.93**	69.60 ± 13.22	82.12 ± 3.22	**86.67 ± 3.98**	**77.78 ± 24.26**	**89.69 ± 5.10**	89.47 ± 5.75	**89.27 ± 3.82**	88.86 ± 4.96	92.72 ± 4.43
	FPR95 (↓)	**25.71 ± 11.89**	**36.24 ± 18.76**	78.61 ± 10.67	**25.59 ± 9.54**	**33.13 ± 7.47**	**63.28 ± 36.20**	**28.47 ± 23.60**	**42.57 ± 24.62**	**34.98± 17.38**	**45.58 ± 7.76**	**47.76 ± 6.62**
	ID ACC (↑)	86.32 ± 5.34	79.05 ± 7.73	68.62 ± 1.32	77.53 ± 0.16	78.85 ± 0.23	61.78 ± 0.16	82.68 ± 4.20	85.20 ± 1.89	58.93 ± 6.34	87.50 ± 5.69	88.31 ± 5.03

Table 2: Overall performance of OOD detection on TextTopoOOD scenarios. Results are reported as the **aggregated mean and standard deviation of the different OOD scenarios**, over **three runs**. The large variance in OOD detection indicates the different levels of challenging OOD scenarios from TextTopoOOD. The detection results of our TNT-OOD against **with(/without)-score propagation methods** are highlighted by Best and Runner-up (**Best** and Runner-up), respectively. **Full results of each OOD scenario on each datasets is in Appendix I**.

The **final training objective** combines both losses, with a hyperparameter λ:

$$\mathcal{L}_{\text{TNT-OOD}} = \mathcal{L}_{\text{cls}} + \lambda \mathcal{L}_{\text{cont}}. \tag{24}$$

4.2 OOD Scoring Function

At test time, we compute OOD scores as a combination of energy (Eq. 2) and alignment score (Eq. 25):

$$\mathbf{s}_{\text{align}} = \left\langle \hat{\mathbf{P}}_t, \hat{\mathbf{g}} \right\rangle, \tag{25}$$

$$\mathbf{s}_{\text{E-lign}} = \mathbf{e}_{\text{energy}} - T \cdot \mathbf{s}_{\text{align}}, \tag{26}$$

T is the temperature. Notably, energy identifies nodes with low confidence across all classes at the logits level, while alignment detects inconsistencies in text-graph relationships that energy alone might miss - a key motivation of TNT-OOD. Higher score indicates more OOD the node is (i.e., high energy and low alignment). The scores are further refined via a K-layer ***propagation*** **smoothing** as in Eq. 3:

$$\tilde{\mathbf{s}}^{(k)} = \alpha_{\text{score}} \tilde{\mathbf{s}}^{(k-1)} + (1 - \alpha_{\text{score}}) \mathbf{D}^{-1} \mathbf{A} \tilde{\mathbf{s}}^{(k-1)}, \tag{27}$$

where $\tilde{\mathbf{s}}^{(0)} = \mathbf{s}_{\text{E-lign}}$, α_{score} controls concentration, and $\mathbf{D}^{-1}\mathbf{A}$ is the normalised graph Laplacian.

5 Experiments

Datasets. TextTopoOOD and TNT-OOD evaluate OOD scenarios across 11 TrNs with varying scales, structural properties, and domains - including citation, knowledge, social, and e-commerce networks (Chen et al., 2024b). For each dataset, **OOD shifts are generated based on selected TextTopoOOD scenarios at multiple noise levels or modes**. Due to space constraint, the detailed discussion and OOD construction is in Appendix D.

Baselines. We compare TNT-OOD with 7 baselines, including (1) **post-hoc methods**: Mahalanobis (Maha) (Lee et al., 2018), MSP (Hendrycks and Gimpel, 2017), ODIN (Liang et al., 2018), NECO (Liang et al., 2018), and Energy (Liu et al., 2020); (2) **graph-specific OOD detectors** that leverage **propagation schema**: GNNSafe (Wu et al., 2023), and NODESafe (Yang et al., 2024); (3) **LLM** zero shot detection on label shift was conducted with GPT-4o mini (OpenAI, 2024) and Gemini-2.5-flash (Google, 2024).

Metrics. Following prior work, **AUROC** (↑) / **AUPR** (↑) / **FPR95** (↓) was utilised to measure OOD detection, with Accuracy used for ID classification (Wu et al., 2023; Yang et al., 2024). Appendix E provides further details on the metrics.

Implementation. All-MiniLM-L6-v2 (Wang et al., 2020) is used to encode text embeddings. All methods' configurations are determined based on

Cr.Attn.	HyperN.	$\mathcal{L}_{\text{cont.}}$	s_{align}	Cora				Citeseer				Elephoto			
				AUROC(↑)	AUPR(↑)	FPR(↓)	ID Acc(↑)	AUROC(↑)	AUPR(↑)	FPR(↓)	ID Acc(↑)	AUROC(↑)	AUPR(↑)	FPR(↓)	ID Acc(↑)
GNNSafe				89.19	89.59	34.21	82.64	85.20	82.12	40.03	74.06	56.14	82.62	65.14	78.60
✓				88.24	87.34	33.68	82.21	86.25	82.78	36.03	74.60	61.41	86.02	64.17	81.73
✓		✓	✓	**91.65**	**92.46**	27.64	82.45	**88.56**	**85.54**	**33.63**	73.68	60.93	85.79	65.50	81.85
✓	✓			90.82	89.57	27.56	81.72	83.42	80.00	39.24	73.68	66.89	87.44	51.20	83.14
✓	✓	✓		91.35	89.93	**25.49**	82.54	84.74	81.14	36.80	74.11	**67.18**	**87.24**	**49.79**	83.48
TNT-OOD				**91.65**	**90.55**	**24.91**	82.54	**87.96**	**84.65**	**32.28**	74.11	**68.69**	**88.05**	**47.40**	83.48

Table 3: Ablation study. **Cr.Attn.** and **HyperN.** denotes the cross attention and HyperNetwork module, respectively.

the ID classification performance via grid search. For fair comparison, post-hoc baselines used the same classifier configuration, resulting in the **same ID accuracy**. For TNT-OOD, the HyperNetwork is implemented as a two-layer MLP and a rank r of 16. A single GCN layer is used in each component. The projection sizes was set to 128. The number of score propagation K is set to 3 for all relevant methods, with $\alpha_{\text{score}} = 0.5$. Full hyperparameter search and sensitivity analysis are in Appendix F.

5.1 Overall Performance

TNT-OOD consistently outperforms or matches baseline methods across diverse network domains and OOD scenarios. Table 2 shows TNT-OOD's superior performance with significant improvements in key metrics across multiple network types. For citation networks (Cora, Citeseer, Pubmed), TNT-OOD surpasses all baselines, reducing average FPR95 by up to 10%. In knowledge and social networks (Reddit, WikiCS), it delivers substantial gains in AUROC and AUPR (increasing average AUROC from 57.43% to 70.61% on Reddit). For e-commerce networks, TNT-OOD achieves the lowest FPR95 scores across all datasets. While maintaining competitive ID accuracy, TNT-OOD excels in OOD detection, demonstrating its effectiveness in diverse TrNs. However, TextTopoOOD also **reveals challenging scenarios** (e.g., Arxiv) where TNT-OOD and propagation-based methods underperform, validating the framework's efficacy in creating rigorous OOD cases. Full results is in Appendix I. Regarding computational resources, due to the computation of the cross attention and HyperNetwork weight generation module, TNT-OOD has inevitable come at a higher cost in memory usage and computation time as discussed in Appendix H. However, as an initial work on TrN OOD detection, we believe the superior performance of TNT-OOD and the challenging nature of TextTopoOOD paves the way for more efficient and effective methods in the future.

5.2 Extended Analysis on Score Propagation

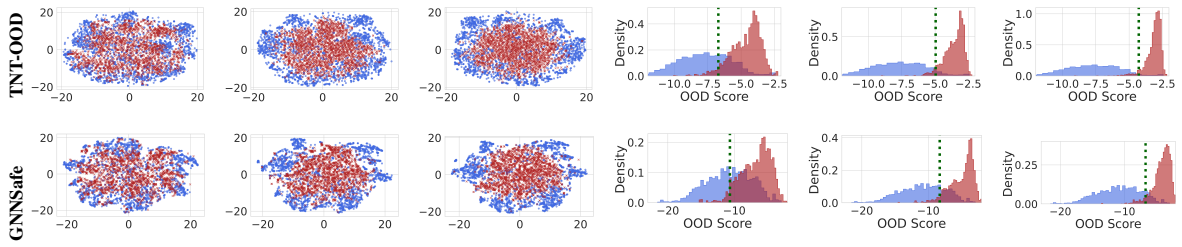
Our framework reveals the nuanced impact of score propagation across different text-rich networks, demonstrating TextTopoOOD's challenging nature. Table 4 shows that in Cora, propagation (prop.) methods substantially enhance performance, with TNT-OOD achieving superior results (25.71% FPR95). Conversely, for Arxiv, prop. degrades performance, with our prop-free variant TNT-wo outperforming all prop. approaches (53.13% FPR95). Bookhis presents a hybrid case where TNT-wo excels in AUROC (88.53%) and FPR95 (38.82%), while TNT-OOD still outperforms prop. baselines. These findings demonstrate our method's adaptability across diverse network characteristics. As propagation is usually unavailable for test-time optimisation, we advocate for future research to enhance OOD detection in TextTopoOOD's challenging scenarios.

	Metrics	w/o Prop.		w/ Prop.	
		Energy	TNT-wo	GNNSafe	TNT-OOD
Cora	AUROC (↑)	81.60	86.24	**88.45**	**91.29**
	AUPR (↑)	83.74	86.57	**87.77**	**89.07**
	FPR95 (↓)	59.41	44.67	**36.06**	**25.71**
Arxiv	AUROC (↑)	**76.96**	**77.99**	37.13	47.65
	AUPR (↑)	**84.27**	**84.11**	66.30	69.60
	FPR95 (↓)	**66.28**	**53.13**	88.51	78.61
Bookhis	AUROC (↑)	74.03	**88.53**	57.33	**75.39**
	AUPR (↑)	**92.42**	**96.89**	84.72	90.06
	FPR95 (↓)	64.10	**38.82**	65.47	**40.13**

Table 4: Detection comparison between **with out (w/o)** and **with (w/)** score propagation. We refer the TNT-OOD method without Prop. as **TNT-wo**.

5.3 Ablation Study

Table 3 presents TNT-OOD's ablation study, with GNNSafe as baseline. The cross-attention mechanism alone improves OOD detection, increasing AUROC from 56.14% to 61.41% on Elephoto. Adding contrastive loss and alignment score delivers strong results on Cora and exceptional performance on Citeseer (88.56% AUROC). While the HyperNetwork shows mixed results across datasets, it greatly improves Elephoto's FPR95 from 64.17%



(a) Emb. $\alpha = 0.5$ (b) $\alpha = 0.7$ (c) $\alpha = 0.9$ (d) Score $\alpha = 0.5$ (e) $\alpha = 0.7$ (f) $\alpha = 0.9$

Figure 3: Embedding representation and OOD score visualisation of ID and OOD data between **TNT-OOD (Top)** and **GNNSafe (Bottom)** at different noise levels on Cora Feature. Dash line indicates FPR95 threshold.

		Cora - Label	Wikics - Label
GPT-4o mini	FPR (↓)	34.89	51.28
TNT-OOD	FPR@ 4o mini TPR (↓)	**25.29**	**30.43**
Gemini-2.5-flash	FPR (↓)	28.58	34.33
TNT-OOD	FPR@ Gemini TPR (↓)	**17.30**	**24.41**

Table 5: Results of LLM as detector on Label shifts.

to 51.20%. When combined with the contrastive objective that aligns text and structure-aware embeddings, it shows consistent improvements. The full TNT-OOD achieves superior performance across metrics and datasets, confirming our components' efficacy for OOD detection in TrNs. Further experiment on the text encoder choice is in Appendix I.

5.4 Comparison with LLM as Detector

We further evaluate LLMs as OOD detectors using only ID categories. Since LLMs cannot produce soft OOD scores, we compare TNT-OOD with FPR calculated at equivalent TPR levels. Table 5 shows TNT-OOD outperforms GPT-4o mini on Cora (reducing FPR from 34.89% to 25.29%) and Gemini-2.5-flash on Wikics (reducing FPR by 10%). The prompt is provided in Appendix Figure 5.

5.5 OOD Score Visualisation

Figure 3 visualises the learned embedding representations and corresponding OOD score distributions across varying noise levels on the **Cora Feature** dataset. The **top row displays results from TNT-OOD**, while the **bottom row shows the GNNSafe** baseline performance. As expected, increasing the noise level (α_{feat} from 0.5 to 0.9) generally reduces the difficulty in distinguishing OOD data, reflected in clearer separations between the ID and OOD samples. The green dashed lines mark the threshold corresponding to an FPR at 95% TPR. Notably, our TNT-OOD framework consistently achieves more pronounced separation between ID and OOD scores, evidenced by a more **distinct gap between distributions** (i.e., the ID data skewed greater to the left of the threshold. In comparison, GNNSafe shows greater overlap in ID and OOD scores across all noise levels, highlighting its relatively weaker discriminative capability.

5.6 Extended Analysis on Coupled Shifts

To further capture the coupled nature of real-world shifts, we conduct an additional experiment where feature mixing is superimposed on top of structure rewiring, as shown in Table 6. A key observation is that these combined shifts amplify representation misalignment, thereby making the discrepancy between ID and OOD nodes more pronounced and ultimately improving detection performance.

Citeseer	**AUROC↑**	**AUPR↑**	**FPR95↓**
StructureRewiring			
mild	71.91	69.69	73.45
strong	94.68	91.74	15.71
FeatureMixing			
$\alpha_{\text{feat}} = 0.5$	84.23	78.89	38.11
$\alpha_{\text{feat}} = 0.7$	91.85	86.63	22.19
Structure+Feature			
mild+$\alpha_{\text{feat}} = 0.5$	94.12	89.50	16.47
mild+$\alpha_{\text{feat}} = 0.7$	96.58	91.74	6.21
strong+$\alpha_{\text{feat}} = 0.5$	96.88	95.01	6.53
strong+$\alpha_{\text{feat}} = 0.7$	96.87	95.02	6.52

Table 6: OOD detection performance (Citeseer) under different OOD shifts.

6 Related Work

Our work explores OOD detection in TrNs across three research trajectories. **1) Post-hoc methods** construct specialised scoring functions using only ID data (Hendrycks and Gimpel, 2017; Sun et al., 2022; Lee et al., 2018; Liang et al., 2018; Ammar et al., 2024; Liu et al., 2020; Lin and Gu, 2023), avoiding model retraining. For **2) graph-structured data**, GNNSafe introduced energy-based scoring with propagation to leverage node-topology interdependencies (Wu et al.,

2023), while NODESafe refined extreme energy scores (Yang et al., 2024) and GRASP enhanced edge augmentation (Ma et al., 2024) - though we exclude comparison with the latter as it uses transductive settings while we focus on inductive tasks. More complex algorithm designs have also been explored, including GOLD and DeGEM (Wang et al., 2025; Chen et al., 2025), which presents a data synthesis strategy to simulate OOD scenarios without requiring actual auxiliary OOD samples. **3) OOD detection in NLP** has made great progress over the year (Lang et al., 2023), employing both established post-hoc methods and newer approaches that leverage LLMs' semantic understanding capabilities (Dai et al., 2023; Xu and Ding, 2025; Liu et al., 2024a). **4) Text-rich Network OOD detection** remains underexplored, with recent LLM-based approaches focusing mainly on label shifts (Xu et al., 2025b; Wang et al., 2024c), neglecting text-structure interactions. We address this gap with TextTopoOOD, which systematically investigates diverse textual and structural shifts, and TNT-OOD, which models text-structure interplay for improved detection. A detailed related work is in Appendix B.

7 Conclusion

We propose **TextTopoOOD**, the first comprehensive framework tailored for evaluating OOD detection in text-rich networks across diverse OOD scenarios; this encompasses feature, structural, label, and domain-based shifts. Moreover, to address the unique challenges posed by the interplay of text and topology, we proposed **TNT-OOD**; a novel detection method that fuses structure-aware context into textual embeddings via a cross-attention mechanism and dynamically adapts to node-level heterogeneity using a HyperNetwork-driven projection module. Extensive experiments on 11 datasets under four distinct OOD scenarios, demonstrates the nuance and challenging nature of TextTopoOOD and the efficacy of TNT-OOD. This work paves the way for more robust and semantically aware models in real-world text rich network applications.

Acknowledgements

This work is supported by Australian Research Council DP230101196, DE250100919 and CE200100025.

Limitations

This study focuses on node-level OOD detection, we hope to expand the research into graph-level OOD detection. Additionally, due to hardware constraints (single NVIDIA RTX A6000 GPU (48GB)), batch-wise and approximation strategies were utilised for similarity calculations and to construct the OOD shifts on larger scale datasets, which may potentially impact the performance. Nevertheless, we have provided extensive experiments across diverse scale and domain datasets, validating the nuance of our TextTopoOOD framework and efficacy of TNT-OOD. Additionally, we have only evaluated a selection of configurations for the proposed TextTopoOOD splits and evaluated on one primary text encoder method, we hope to extend with more experiments as future work. Additionally, in TNT-OOD, the text encoder remain fixed, as text is a key element of TrNs, we leave explorations on fine-tuning the PLMs as future work. Furthermore, with consideration of the textual and structural features, the proposed cross-attention and HyperNetwork architecture inevitably resulted in higher computational cost, both in memory usage and training time as discussed in Section 5. However, we have made an attempted to reduce cost by implementing an efficient low-rank HyperNetwork and using batch calculations. As an initial work that demonstrates great performance and presenting with challenging and nuanced TrN OOD detection scenarios, we hope this can pave the way for future explorations into effective and efficient OOD detection methods on text-rich networks.

References

Momin Abbas, Muneeza Azmat, Raya Horesh, and Mikhail Yurochkin. 2025. Out-of-distribution detection using synthetic data generation. *CoRR*, abs/2502.03323.

Emmanuel Abbe. 2017. Community detection and stochastic block models: Recent developments. *JMLR*.

Mouïn Ben Ammar, Nacim Belkhir, Sebastian Popescu, Antoine Manzanera, and Gianni Franchi. 2024. NECO: neural collapse based out-of-distribution detection. In *ICLR*.

Anthropic. 2025. Claude 3.7 sonnet. Large language model, available at https://www.anthropic.com/claude.

Tianyi Bao, Qitian Wu, Zetian Jiang, Yiting Chen, Jiawei Sun, and Junchi Yan. 2024. Graph out-of-

distribution detection goes neighborhood shaping. In *ICML*.

Tingyi Cai, Yunliang Jiang, Yixin Liu, Ming Li, Changqin Huang, and Shirui Pan. 2025. Out-of-distribution detection on graphs: A survey. *CoRR*.

Vinod Kumar Chauhan, Jiandong Zhou, Ping Lu, Soheila Molaei, and David A. Clifton. 2024. A brief review of hypernetworks in deep learning. *Artif. Intell. Rev.*

Sishuo Chen, Xiaohan Bi, Rundong Gao, and Xu Sun. 2022. Holistic sentence embeddings for better out-of-distribution detection. In *EMNLP Findings*.

Tong Chen, Danny Wang, Xurong Liang, Marten Risius, Gianluca Demartini, and Hongzhi Yin. 2024a. Hate speech detection with generalizable target-aware fairness. In *KDD*.

Yuhan Chen, Yihong Luo, Yifan Song, Pengwen Dai, Jing Tang, and Xiaochun Cao. 2025. Decoupled graph energy-based model for node out-of-distribution detection on heterophilic graphs. *ICLR*.

Zhikai Chen, Haitao Mao, Jingzhe Liu, Yu Song, Bingheng Li, Wei Jin, Bahare Fatemi, Anton Tsitsulin, Bryan Perozzi, Hui Liu, and Jiliang Tang. 2024b. Text-space graph foundation models: Comprehensive benchmarks and new insights. In *NeurIPS*.

Yi Dai, Hao Lang, Kaisheng Zeng, Fei Huang, and Yongbin Li. 2023. Exploring large language models for multi-modal out-of-distribution detection. In *EMNLP Findings*.

Xuefeng Du, Yiyou Sun, Jerry Zhu, and Yixuan Li. 2023. Dream the impossible: Outlier imagination with diffusion models. In *NeurIPS*.

C. Lee Giles, Kurt D. Bollacker, and Steve Lawrence. 1998. Citeseer: An automatic citation indexing system. In *ACM Digital Libraries*.

Google. 2024. Gemini 2.5 flash. Large language model, available at https://gemini.google.com/.

Shurui Gui, Xiner Li, Limei Wang, and Shuiwang Ji. 2022. GOOD: A graph out-of-distribution benchmark. In *NeurIPS*.

Yuxin Guo, Cheng Yang, Yuluo Chen, Jixi Liu, Chuan Shi, and Junping Du. 2023. A data-centric framework to endow graph neural networks with out-of-distribution detection ability. In *KDD*.

David Ha, Andrew M. Dai, and Quoc V. Le. 2017. Hypernetworks. In *ICLR*.

Dan Hendrycks and Kevin Gimpel. 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *ICLR*.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. 2020. Open graph benchmark: Datasets for machine learning on graphs. In *NeurIPS*.

Xuanwen Huang, Kaiqiao Han, Yang Yang, Dezheng Bao, Quanjin Tao, Ziwei Chai, and Qi Zhu. 2024. Can gnn be good adapter for llms? In *WWW*.

Ming Ji, Yizhou Sun, Marina Danilevsky, Jiawei Han, and Jing Gao. 2010. Graph regularized transductive classification on heterogeneous information networks. In *KDD*.

Bowen Jin, Wentao Zhang, Yu Zhang, Yu Meng, Xinyang Zhang, Qi Zhu, and Jiawei Han. 2023. Patton: Language model pretraining on text-rich networks. In *ACL*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *ICLR*.

Hao Lang, Yinhe Zheng, Yixuan Li, Jian Sun, Fei Huang, and Yongbin Li. 2023. A survey on out-of-distribution detection in NLP. *CoRR*.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. 2018. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *NeurIPS*.

Haoyang Li, Xin Wang, Ziwei Zhang, and Wenwu Zhu. 2022a. Out-of-distribution generalization on graphs: A survey. *CoRR*.

Yuhan Li, Peisong Wang, Xiao Zhu, Aochuan Chen, Haiyun Jiang, Deng Cai, Victor Wai Kin Chan, and Jia Li. 2024. Glbench: A comprehensive benchmark for graph with large language models. In *NeurIPS*.

Zenan Li, Qitian Wu, Fan Nie, and Junchi Yan. 2022b. Graphde: A generative framework for debiased learning and out-of-distribution detection on graphs. In *NeurIPS*.

Shiyu Liang, Yixuan Li, and R. Srikant. 2018. Enhancing the reliability of out-of-distribution image detection in neural networks. In *ICLR*.

Haowei Lin and Yuntian Gu. 2023. FLatS: Principled out-of-distribution detection with feature-based likelihood ratio score. In *EMNLP*.

Bo Liu, Liming Zhan, Zexin Lu, Yujie Feng, Lei Xue, and Xiao-Ming Wu. 2024a. How good are llms at out-of-distribution detection? In *COLING*.

Hao Liu, Jiarui Feng, Lecheng Kong, Ningyue Liang, Dacheng Tao, Yixin Chen, and Muhan Zhang. 2024b. One for all: Towards training one graph model for all classification tasks. In *ICLR*.

Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. 2020. Energy-based out-of-distribution detection. In *NeurIPS*.

Yilun Liu, Ruihong Qiu, and Zi Huang. 2023a. Cat: Balanced continual graph learning with graph condensation. In *ICDM*.

Yilun Liu, Ruihong Qiu, Yanran Tang, Hongzhi Yin, and Zi Huang. 2025. PUMA: efficient continual graph learning for node classification with graph condensation. *TKDE*.

Yixin Liu, Kaize Ding, Huan Liu, and Shirui Pan. 2023b. GOOD-D: on unsupervised graph out-of-distribution detection. In *WSDM*.

Longfei Ma, Yiyou Sun, Kaize Ding, Zemin Liu, and Fei Wu. 2024. Revisiting score propagation in graph out-of-distribution detection. In *NeurIPS*.

Jianmo Ni, Jiacheng Li, and Julian J. McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *EMNLP-IJCNLP*.

OpenAI. 2024. Gpt-4o mini. Large language model, available at `https://openai.com/`.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP*.

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

Xu Shen, Yili Wang, Kaixiong Zhou, Shirui Pan, and Xin Wang. 2024. Optimizing OOD detection in molecular graphs: A novel approach with diffusion models. In *KDD*.

Yu Song and Donglin Wang. 2022. Learning on graphs with out-of-distribution nodes. In *KDD*.

Maximilian Stadler, Bertrand Charpentier, Simon Geisler, Daniel Zügner, and Stephan Günnemann. 2021. Graph posterior network: Bayesian predictive uncertainty for node classification. In *NeurIPS*.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. 2022. Out-of-distribution detection with deep nearest neighbors. In *ICML*.

Yanran Tang, Ruihong Qiu, Yilun Liu, Xue Li, and Zi Huang. 2024a. Casegnn: Graph neural networks for legal case retrieval with text-attributed graphs. In *ECIR*.

Yanran Tang, Ruihong Qiu, Hongzhi Yin, Xue Li, and Zi Huang. 2024b. Caselink: Inductive graph learning for legal case retrieval. In *SIGIR*.

Daeho Um, Jongin Lim, Sunoh Kim, Yuneil Yeo, and Yoonho Jung. 2025. Spreading out-of-distribution detection on graphs. In *ICLR*.

Danny Wang, Ruihong Qiu, Guangdong Bai, and Zi Huang. 2025. GOLD: graph out-of-distribution detection via implicit adversarial latent generation. *ICLR*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024a. Multilingual E5 text embeddings: A technical report. *CoRR*.

Luzhi Wang, Dongxiao He, He Zhang, Yixin Liu, Wenjie Wang, Shirui Pan, Di Jin, and Tat-Seng Chua. 2024b. GOODAT: towards test-time graph out-of-distribution detection. In *AAAI*.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *NeurIPS*.

Yiqi Wang, Jiaxin Zhang, Nianhao Xie, Yu Shi, Siwei Wang, Xinwang Liu, En Zhu, and Yusong Tan. 2024c. Towards the effect of large language models on out-of-distribution challenge in text-attributed graphs.

Qitian Wu, Yiting Chen, Chenxiao Yang, and Junchi Yan. 2023. Energy-based out-of-distribution detection for graph neural networks. In *ICLR*.

Haoyan Xu, Zhengtao Yao, Ziyi Wang, Zhan Cheng, Xiyang Hu, Mengyuan Li, and Yue Zhao. 2025a. Graph synthetic out-of-distribution exposure with large language models. *CoRR*.

Haoyan Xu, Zhengtao Yao, Xuzhi Zhang, Ziyi Wang, Langzhou He, Yushun Dong, Philip S. Yu, Mengyuan Li, and Yue Zhao. 2025b. Glip-ood: Zero-shot graph ood detection with foundation model. *CoRR*.

Ruiyao Xu and Kaize Ding. 2025. Large language models for anomaly and out-of-distribution detection: A survey. In *NAACL Findings*.

Hao Yan, Chaozhuo Li, Ruosong Long, Chao Yan, Jianan Zhao, Wenwen Zhuang, Jun Yin, Peiyan Zhang, Weihao Han, Hao Sun, Weiwei Deng, Qi Zhang, Lichao Sun, Xing Xie, and Senzhang Wang. 2023. A comprehensive study on text-attributed graphs: Benchmarking and rethinking. In *NeurIPS*.

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. 2021. Generalized out-of-distribution detection: A survey. *CoRR*.

Shenzhi Yang, Bin Liang, An Liu, Lin Gui, Xingkai Yao, and Xiaofang Zhang. 2024. Bounded and uniform energy-based out-of-distribution detection for graphs. In *ICML*.

Xujiang Zhao, Feng Chen, Shu Hu, and Jin-Hee Cho. 2020. Uncertainty aware semi-supervised learning on graph data. In *NeurIPS*.

Tao Zou, Le Yu, Yifei Huang, Leilei Sun, and Bowen Du. 2023. Pretraining language models with text-attributed heterogeneous graphs. In *EMNLP Findings*.

A Impact Statement

Our work aims to inspire and pave the way for future works on text-rich network OOD detection for real-world applications. This foundational research uses only publicly available datasets and models, with all sources properly acknowledged. We do not identify any direct negative societal impacts requiring specific safeguards or emphasis in this study, though we encourage continuing ethical assessment as the field develops.

B Extended Related Work

Our work explores the challenging problem of OOD detection in text-rich networks, intersecting with several important research trajectories that have evolved independently but are now converging in this complex domain.

Post-hoc OOD Detection Methods. Post-hoc methods represent a fundamental approach where detection relies solely on in-distribution data to construct specialised scoring functions (Hendrycks and Gimpel, 2017; Sun et al., 2022; Lee et al., 2018; Liang et al., 2018; Ammar et al., 2024; Liu et al., 2020; Lin and Gu, 2023). These methods avoid re-training models, providing adaptive detection capabilities across diverse scenarios. Maximum Softmax Probability (MSP)(Hendrycks and Gimpel, 2017) utilises confidence scores from softmax outputs, while ODIN(Liang et al., 2018) enhances this with temperature scaling and input perturbation. Mahalanobis distance-based approaches (Lee et al., 2018) measure deviation from class-conditional Gaussian distributions. These approaches form the methodological foundation upon which specialised network-based detection methods build.

NLP OOD Detection. Recent OOD detection techniques in NLP increasingly leverage PLM/LLMs to improve robustness on novel inputs. Distance-based detectors applied to a model's embedding space have shown strong performance with pre-trained encoders like BERT (Chen et al., 2022). Moreover, research has been conducted on the effectiveness and identifying scenarios where LLM exhibit OOD detection capabilities (Liu et al., 2024a). Another emerging trend is harnessing LLMs to generate synthetic outlier examples for outlier exposure training, which markedly reduces false positives and improves detection without requiring real OOD data (Abbas et al., 2025). Notably, LLM-enhanced strategies has greatly improved the diversity in textual OOD detection, enabling more reliable identification of OOD instances in open-world settings.

Graph-Structured Data OOD Detection. In the domain of graph-structured data, diverse approaches have emerged to leverage network topology alongside node features (Guo et al., 2023; Li et al., 2022b; Stadler et al., 2021; Zhao et al., 2020; Shen et al., 2024; Liu et al., 2023a, 2025; Bao et al., 2024; Um et al., 2025). GNNSafe (Wu et al., 2023) introduced an energy-score based method with a propagation schema to effectively harness the intrinsic interdependencies between node features and network topology. NODESafe (Yang et al., 2024) extended this approach by refining extreme energy scores. GRASP (Ma et al., 2024) proposed an enhanced edge augmentation method to improve propagation effectiveness at test time. More complex algorithm designs have also been explored, including GOLD (Wang et al., 2025), which presents a data synthesis strategy to simulate OOD scenarios without requiring actual auxiliary OOD samples. DeGEM (Chen et al., 2025) provides a novel energy-based modelling architecture featuring multi-hop graph encoders coupled with dedicated energy heads, designed to improve detection over heterophilic network structures. We exclude direct comparison with transductive methods like GRASP as our work focuses on inductive tasks where OOD node features or edge connections are unavailable during, representing a more challenging but realistic scenario.

Text-Rich Network OOD Detection. Text-rich network OOD detection methods remain relatively underexplored, despite their critical importance in real-world applications. Recent attempts have leveraged Large Language Models (LLMs) for generating OOD data or performing zero-shot OOD detection (Xu et al., 2025b,a). However, current research has primarily focused on label shift scenarios, ignoring the intricate interplay between textual semantics and structural properties that characterises real-world distribution shifts (Wang et al., 2024c). In real-world TrN environments, textual content and network structure evolve in tandem, creating complex patterns of distribution shift that cannot be captured by approaches designed for either modality in isolation. To bridge this gap, we propose the TextTopoOOD framework to systematically and thoroughly investigate OOD detection under diverse textual, feature, and structural

Original Text	Synonym ($\alpha = 0.5, p_{text} = 0.3$)	Antonym ($\alpha = 0.3, p_{text} = 0.3$)
Graph neural networks have shown promising results on semi-supervised node classification tasks.	Graph nerve-based netorks have shown promissing results on semi-labeled node categorisation tasks.	Graph neural networks have shown failesd rsults on unsupervised node classificattion tzsks.
Contrastive learning has emerged as a powerful paradigm for self-supervised representation learning.	Contrative learning has enmerged as a strong pardigm for self-labeled feature learning.	Unifying learning has collapsed as a weak heuristic for self-unsupervised representation leearning.
Pre-trained language models are widely used for various natural language understanding tasks.	Pre-traind speech moals are commonly used for various natural textual understanding assignments.	Untraned imzge generaotrs are widely used for unnatural language forgetting tasks.

Table 7: Additional examples of text-level feature shift created by TEXTAUGMENT.

shifts. Complementing this framework, we propose TNT-OOD, a novel OOD detection method that explicitly models the interplay between textual and structural information for improved detection performance on TrNs. Our work thus uniquely contributes to advancing OOD detection capabilities in text-rich network environments, addressing a blind spot in current approaches and establishing both evaluation standards and methodological foundations for this important research direction.

HyperNetwork. Hypernetworks (Ha et al., 2017) are neural networks that generate the weights of another target model, enabling dynamic parameterisation and improved adaptability. They have been widely applied in few-shot learning and domain adaptation, where generalisation to unseen tasks or domains is essential (Chauhan et al., 2024). For instance, it is used in tasks such as hate speech detection to enable detecting unseen domain samples (Chen et al., 2024a).

C Extended Discussion on OOD Shifts in TextTopoOOD.

Label Shift. The LLM prompt for selecting the thematically-guided label-leave-out classes is provided in Prompt 4. We provide three strategies to select OOD classes: 1) Random Selection, 2) Thematically similar labels w.r.t. ID labels, and 3) Thematically dissimilar labels to the ID labels. We used GPT-4o as the LLM model, giving the category names and descriptions (if available) as input.

Text Shift. We provide further examples of text-level shifts involving synonym, antonym replacements as well as character level edits in Table 7.

Domain-Based Sentiment Shift. For datasets with underlying sentiment properties (e.g., review ratings), we can create domain shifts based on the sentiment expressed in the text (i.e., Positive vs. Negative reviews). Node sets can be defined based on their sentiment $s(v_i)$:

$$\mathcal{V}_{\text{ID}} = \{v_i \in \mathcal{V} \mid s(v_i) \in S_{\text{ID}}\} \quad (28)$$

$$\mathcal{V}_{\text{OOD}} = \{v_i \in \mathcal{V} \mid s(v_i) \in S_{\text{OOD}}\}, \quad (29)$$

where S_{ID} is the set of ID sentiments (e.g., {'neutral', 'positive'}) and S_{OOD} is a disjoint set of OOD sentiments (e.g., {'negative'}). The sentiments can be obtained via a sentiment analysis model (i.e., sentiment pre-trained Bert). The textual content associated with these different sentiment groups often exhibits distinct vocabulary, stylistic features, and potentially different graph connectivity patterns if sentiment influences interactions.

D Extended Dataset Description

We have utilised the publicly available text-rich networks provided by TSGFM and GLBench (Chen et al., 2024b; Li et al., 2024), acknowledge and follow the MIT license. We adhere to the provided splits for ID classification, except for label shift and time-shift where we would filter out the associated OOD labels from the ID data. The dataset statistics are presented in Table 8.

Scale: The datasets range from small-scale networks (Cora: ∼2.7K nodes) to large-scale networks

Name	#Nodes	#Edges	#Classes	Domain	Text Characteristics
Citation Networks					
Cora	2,708	10,556	7	CS Citation	Paper title & abstracts
CiteSeer	3,186	8,450	6	CS Citation	Paper title & abstracts
DBLP	14,376	431,326	4	CS Citation	Paper title & abstracts
Arxiv	169,343	1,166,243	40	CS Citation	Paper title & abstracts
PubMed	19,717	88,648	3	Bio Citation	Medical title & abstracts
E-commerce Networks					
History (BookHis)	41,551	358,574	12	E-commerce	Book title & descriptions
Child (BookChild)	76,875	1,554,578	24	E-commerce	Book title & descriptions
Computers (EleComp)	87,229	721,081	10	E-commerce	Product reviews
Photo (ElePhoto)	48,362	500,939	12	E-commerce	Product reviews
Other Networks					
WikiCS	11,701	431,726	10	Knowledge	Encyclopedia articles
Reddit	33,434	302,876	2	Social Forum	Social media posts

Table 8: Statistics of datasets used in TextTopoOOD. The datasets span citation networks, e-commerce networks, and knowledge graphs.

(Products: ∼316K nodes), with edge counts ranging from thousands to millions.

Domains: The datasets span four primary domains, dataset details are provided below:

- **Citation networks** (Cora, CiteSeer, DBLP, Arxiv, PubMed)
- **E-commerce networks** (Bookhis, Bookchild, Elecomp, Elephoto)
- **Knowledge and social networks** (WikiCS, Reddit)

D.1 OOD Test Data Construction.

In this paper, we consider the following configurations of the OOD scenarios of TextTopoOOD as presented in Sec 3. For the OOD shifts, we consider an inductive setting, where OOD node feature/structure (depending on the shift type) and labels are not available during training.

Text Augmentation. For text augmentation (sec 3.1), we utilise NLTK v.3.9.1's WordNet[2] to build a cache of possible **synonyms** and **antonyms** for all unique words in the datasets. Since academic content often lacks suitable replacement terms, we set

$$\alpha_{\text{text}} = 1 \quad \text{and} \quad p_{\text{char}} = 1,$$

in Eq. 4 to introduce sufficient diversity for simulating suitable OOD scenarios.

[2] https://www.nltk.org/

Feature Mixing. For feature mixing (sec 3.1), we implement three different noise levels for each data:

$$\alpha_{\text{feat}} \in \{0.5, 0.7, 0.9\}.$$

This would gradually increase the diversity between the original ID embedding and the transformed OOD embeddings. Example embeddings from a trained classifier are shown in Figure 3.

Structure Rewiring. For structure rewiring shift (sec 3.2), we calculate the density as:

$$\rho = \frac{\text{num ID edges}}{\text{num possible edges}}.$$

This ensures our generated structure does not deviate too far in terms of graph density, mitigating undesired OOD test cases. Following this, we define **three levels of structure shifts** with β, f_{ii}, f_{ij} as follows:

- **Mild**: $\beta = 0.2, f_{ii} = 0.7, f_{ij} = 0.5$
- **Medium**: $\beta = 0.5, f_{ii} = 0.6, f_{ij} = 0.3$
- **Strong**: $\beta = 1.0, f_{ii} = 0.4, f_{ij} = 0.7$

Semantic Connection. For semantic connection shift (sec 3.2), to ensure a challenging and valid OOD shifts, k was selected to be the same number of edges as the original ID graph. Moreover, we define three thresholds to test:

$$\text{threshold} \in \{0.75, 0.85, 0.95\},$$

where we would select edges based on the similarity values at these given thresholds (i.e., $k/2$ edges directly greater & lower than the similarity value at the 0.75% threshold.) This introduces diversity based on a gradual increase in similarity between the connected nodes in the OOD test set.

Text Swap. For text swap shift (sec 3.2), we experiment with all three introduced variants: **1)Intra-class, 2)Inter-class, 3)Random Swap**, all with a swap ratio of

$$\beta_{\text{swap}} = 1.$$

This ensures a maximum distinction against the original ID network.

Label Shift. For label shifts (sec 3.3), we again test the three strategies proposed: **1) Random selection, 2) Thematic similarity, 3) Thematic dissimilarity**. We perform label shift on datasets above four available classes (i.e., excluding datasets like Pubmed, DBLP, and Reddit), this ensures that there is enough ID data for training and our OOD contains diverse labels. The number of OOD classes selected are dependent on the dataset, in general, we select 10% to $\approx$ 40% of the available classes as OOD (i.e., for Cora with 7 classes, we select 3 OOD classes and 4 ID classes). Using the prompt in Figure 4, we can select the thematically similar and dissimilar ID and OOD classes. The detailed OOD classes are shown in Table 9:

Dataset	Random	Similar	Dissimilar
Cora	{0-2}	{1, 3, 4}	{2, 5, 6}
Citeseer	{1, 3, 5}	{2-4}	{0, 1, 5}
WikiCS	{1, 3-5, 8}	{0, 1, 8, 9}	{2-4, 6}
Bookhis	{0, 1, 3, 4, 7}	{1, 2, 6, 10, 11}	{0, 3, 5, 8, 9}
Bookchild	{0, 2, 5, 9-11, 15, 18}	{1, 3, 4, 6, 12, 19, 21, 23}	{0, 2, 5, 7-10, 22}
Elephoto	{1, 3, 5, 7, 9}	{0, 4, 6, 10, 11}	{1-3, 5, 9}
Elecomp	{1-5}	{2-4, 8, 9}	{0, 1, 5-7}

Table 9: OOD Class Configurations by Dataset

Domain Shift. For domain shifts (sec 3.4), we split the Arxiv-dataset based on temporal information. The time ranges are as follow:

- ID Time Range: 1960-2015
- OOD Time Range 1: 2017-2018
- OOD Time Range 2: 2018-2019
- OOD Time Range 3: 2019-2020

Notably, we evaluate three distinct time periods as OOD scenarios. For each temporal split, we include only nodes published within that specific period, with accessible edges limited to citations that existed up to the end date of that period - accurately representing the available information at that point in time.

We provide dataset statistics for each OOD shift type in Table 10, using Cora as an example. Text and feature-level shifts preserve graph structure, maintaining identical node and edge counts as the ID data. Structure-level shifts modify edge counts based on connectivity intensity, though semantic connection shifts maintain the original edge count while reorganising connections. For label shifts, ID statistics vary depending on which classes are designated as OOD, while OOD data maintains the complete edge set to accurately represent the inductive setting where the full graph structure remains accessible during inference with its associated OOD nodes.

Data Split	Nodes	Edges
ID (Main)	2,708	10,556
Feature & Text shifts	2,708	10,556
Structure-level Shifts		
Structural (Medium)	2,708	11,690
Semantic Connect	2,708	10,556
Label Shifts		
Label Shift 1 (ID)	1,412	5,314
Label Shift 1 (OOD)	1,296	10,556
Label Shift 2 (ID)	1,767	6,248
Label Shift 2 (OOD)	941	10,556
Label Shift 3 (ID)	1,121	3,720
Label Shift 3 (OOD)	1,587	10,556

Table 10: ID/OOD Dataset Statistics for Cora

D.2 Dataset description

The detailed descriptions of the dataset are as follows:

CORA This dataset represents a citation network where nodes correspond to academic papers with raw text being the title and abstract, and edges denote citation relationships (Sen et al., 2008). Each paper is classified into one of seven categories.

CITESEER This is another citation network (Giles et al., 1998; Sen et al., 2008), where nodes represent scientific papers classified into one of six classes, and edges denote citation

relationships, text data are also title and abstract. This dataset contains more nodes but fewer edges than Cora.

DBLP This is citation network (Ji et al., 2010), where nodes represent scientific paper classified into one of four categories (area of work), and edges denote co-authorship relationship, text data are paper titles.

ARXIV This large-scale citation network spans research papers from 1960 to 2020 (Hu et al., 2020), with nodes representing papers categorised by subject area, edges indicating citations, and the raw text are titles and abstracts.

PUBMED This is a biomedical paper citation network (Sen et al., 2008), each node is classified into one of three categories. The nodes represent academic papers with title and abstract as text data, while edges are citation relationships.

BOOKHIS & BOOKCHILD These are e-commerce network originated from the Amazon-books dataset: History and Children subsets (Yan et al., 2023). This dataset models an item co-purchasing network, where nodes represent products (books), and edges indicate frequently co-purchased/co-viewed items. Node text capture product descriptions, and labels correspond to product categories (Ni et al., 2019).

ELEPHOTO & ELECOMP These are e-commerce network from the Amazon-electronics dataset. The former is sourced from the Photo subset, while the latter comes from the Computers subset (Yan et al., 2023). This dataset models an item co-purchasing network, where nodes represent products, and edges indicate frequently co-purchased/co-viewed items. Node text capture product reviews (i.e., highest upvoted review of the item), and labels correspond to product categories (Ni et al., 2019).

WIKICS This a knowledge network representing Wikipedia linkage, where nodes represent pages and edges denote reference links between them. Node text include page titles and content, while labels correspond to Wikipedia entry categories (Li et al., 2024; Liu et al., 2024b).

REDDIT This is a social network where nodes represent users and edges indicate reply interactions. Text features include content from users' last three subreddit posts, with labels categorising users as either popular or normal (Li et al., 2024; Huang et al., 2024).

E Metrics

For evaluation of OOD detection methods, we adopt three widely used **manual threshold-independent metrics** (Liu et al., 2020; Wu et al., 2023; Yang et al., 2024; Ma et al., 2024; Ammar et al., 2024): **AUROC, AUPR, and FPR95**. The Area Under the Receiver Operating Characteristic curve (AUROC) provides a global view of the trade-off between true positive rate (TPR) and false positive rate (FPR) across all thresholds, indicating the model's overall discrimination capability. However, AUROC may be less reliable in settings with imbalance scenarios. To address this, we also consider the Area Under the Precision-Recall curve (AUPR), which better reflects model performance in imbalanced scenarios by focusing on both precision and recall. Lastly, the False Positive Rate at 95% True Positive Rate (FPR95) quantifies the proportion of OOD samples falsely detected as ID when the model achieves high TPR, highlighting the robustness of the detection system under stringent sensitivity demands. Collectively, these metrics offer a balanced evaluation of both general and high-sensitivity OOD detection performance.

F Implementation Details

Experiments were conducted using Python 3.9.2 and PyTorch 2.5.1 with Cuda 12.2 on a single NVIDIA RTX A6000 GPU with 48GB of memory. All baseline and TNT-OOD's hyperparameters are searched over the following parameters: num layers $\in \{2, 3, 4\}$, hidden dimension $\in \{64, 128, 256\}$, learning rate $\in \{0.1, \ldots, 0.0001\}$, dropout $\in \{0, \ldots, 0.7\}$. We ran the baseline results to the best of our ability. For text encoding, we use the All-MiniLM-L6-v2; a 22.7M parameter model provided on huggingface under the Apache license 2.0 license[3], this generates a 384 dimensional dense vector for each text sample. All experiments were conducted over three seeds, we report the mean and standard deviation in our paper. The Adam optimiser is used for training (Kingma and Ba, 2015). As described in Section 5, the number of propagation iterations K is set to 3, with α_{score} fixed to 0.5. We further provide the hyperparameter sensitivity analysis for λ in Eq. 24, and τ in Eq. 22

[3] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

in Tables 11 and 12. The experiment were conduct on all shifts except for label shift.

λ	Metric	Cora	Pubmed	Wikics
0	AUROC	90.93 ± 4.29	86.05 ± 3.92	89.80 ± 11.10
	AUPR	89.20 ± 3.85	82.59 ± 5.14	89.94 ± 6.67
	FPR	26.72 ± 11.29	36.38 ± 7.88	26.18 ± 27.85
	ID Acc	82.32 ± 0.20	78.43 ± 0.14	80.15 ± 0.23
0.001	AUROC	91.01 ± 4.30	86.25 ± 3.79	89.92 ± 10.89
	AUPR	89.40 ± 3.82	82.93 ± 4.85	90.01 ± 6.52
	FPR	26.47 ± 11.37	36.12 ± 7.89	26.03 ± 27.65
	ID Acc	82.50 ± 0.18	78.41 ± 0.12	80.19 ± 0.22
0.01	AUROC	90.62 ± 4.08	86.90 ± 3.79	**90.77 ± 8.82**
	AUPR	89.06 ± 3.76	83.99 ± 4.76	**90.59 ± 5.15**
	FPR	27.50 ± 10.75	35.23 ± 7.98	**25.03 ± 24.65**
	ID Acc	82.11 ± 0.97	78.44 ± 0.08	**80.21 ± 0.28**
0.1	AUROC	91.35 ± 4.37	88.24 ± 3.56	92.01 ± 6.91
	AUPR	90.01 ± 3.92	86.18 ± 4.50	91.42 ± 3.94
	FPR	25.40 ± 11.95	34.70 ± 7.73	23.31 ± 23.55
	ID Acc	82.51 ± 0.91	78.49 ± 0.18	80.12 ± 0.21
0.5	AUROC	91.87 ± 4.77	**88.82 ± 3.22**	91.64 ± 9.00
	AUPR	90.81 ± 4.19	**86.67 ± 3.98**	91.30 ± 5.06
	FPR	23.81 ± 13.58	**33.13 ± 7.47**	23.52 ± 29.60
	ID Acc	82.50 ± 0.78	**78.85 ± 0.23**	80.04 ± 0.21
1	AUROC	**91.65 ± 4.61**	88.17 ± 3.17	85.19 ± 20.53
	AUPR	**90.55 ± 4.06**	86.04 ± 3.77	87.52 ± 12.01
	FPR	**24.91 ± 13.11**	35.20 ± 7.40	33.35 ± 34.78
	ID Acc	**82.54 ± 0.25**	78.59 ± 0.28	72.95 ± 0.61

Table 11: Hyperparameter analysis for λ. Bold highlights the parameters selected.

τ	Metric	Bookhis	Elecomp
0.0001	AUROC	34.20 ± 19.87	47.13 ± 19.74
	AUPR	76.50 ± 7.89	87.63 ± 7.68
	FPR	91.91 ± 6.77	77.85 ± 2.47
	ID Acc	72.51 ± 1.42	58.84 ± 0.70
0.001	AUROC	61.55 ± 9.23	64.06 ± 7.06
	AUPR	85.14 ± 4.44	91.28 ± 4.74
	FPR	60.87 ± 10.08	57.93 ± 10.39
	ID Acc	78.86 ± 4.70	78.63 ± 0.20
0.01	AUROC	71.24 ± 11.23	66.37 ± 6.63
	AUPR	88.14 ± 4.49	91.72 ± 4.71
	FPR	49.01 ± 18.80	53.43 ± 10.97
	ID Acc	83.18 ± 0.63	83.14 ± 0.50
0.1	AUROC	**71.75 ± 17.40**	**68.27 ± 6.73**
	AUPR	**88.51 ± 5.87**	**92.11 ± 4.66**
	FPR	**45.57 ± 26.28**	**47.63 ± 7.39**
	ID Acc	**83.86 ± 0.12**	**84.75 ± 0.09**
1	AUROC	70.89 ± 11.44	67.20 ± 6.06
	AUPR	88.19 ± 4.22	91.92 ± 4.43
	FPR	49.71 ± 18.09	51.58 ± 12.03
	ID Acc	83.60 ± 0.30	84.70 ± 0.18

Table 12: Hyperparameter analysis for τ. Bold highlights the optimal parameter.

Algorithm 1 Character-level Edit

function CHAREDIT(w)
  Input: Word w
  Output: Edited word w'
  ops $\leftarrow$ {"ins", "del", "replace", "swap"}
  op $\leftarrow$ Sample(ops)
  pos $\leftarrow$ RandInt$(0, |w| - 1)$
  if op $=$ "ins" **then**
    $w' \leftarrow w[0 : \text{pos}] + \text{RandChar}() + w[\text{pos} :]$
  else if op $=$ "del" and $|w| > 1$ **then**
    $w' \leftarrow w[0 : \text{pos}] + w[\text{pos} + 1 :]$
  else if op $=$ "replace" **then**
    $c \leftarrow \text{RandChar}() \neq w[\text{pos}]$
    $w' \leftarrow w[0 : \text{pos}] + c + w[\text{pos} + 1 :]$
  else if op $=$ "swap" and $|w| > 1$ **then**
    pos $\leftarrow \min(\text{pos}, |w| - 2)$
    $w' \leftarrow w[0 : \text{pos}] + w[\text{pos}+1] + w[\text{pos}]$
    $w' \leftarrow w' + w[\text{pos} + 2 :]$
  else
    $w' \leftarrow w$
  end if
  return w'
end function

Algorithm 2 Text Augmentation

function TEXTAUGMENT(t, type, α, p_{char})
  Input: Text t, type, noise level α, edit prob p_{char}
  Output: Augmented text $\tilde{t}$
  words, tags $\leftarrow$ PosTag(Tokenize(t))
  candidates $\leftarrow \{i : \text{IsEligible}(\text{words}[i])\}$
  $n \leftarrow \lfloor \alpha \cdot |\text{candidates}| \rfloor$
  idxs $\leftarrow$ Sample(candidates, n)
  for $i \in$ idxs **do**
    alts $\leftarrow$ Wordnet(words$[i]$, type)
    if alts $\neq \emptyset$ **then**
      words$[i] \leftarrow$ Sample(alts)
      if Random() $< p_{\text{char}}$ **then**
        words$[i] \leftarrow$ CharEdit(words$[i]$)
      end if
    end if
  end for
  return Detokenize(words)
end function

G Algorithm

Algorithms for the TEXTAUGMENT function in Eq. 4 is provided in Algorithm 2. The associated

Algorithm 3 Text Swap for OOD Generation

function TEXTSWAP($\mathcal{G}(\mathcal{V}, \mathcal{T}, \mathbf{A})$, $\mathbf{Y}$, scope, β)
  Input: Network $\mathcal{G}$, labels $\mathbf{Y}$, swap scope; ratio β
  Output: Network $\mathcal{G}'$ with swapped texts $\tilde{\mathcal{T}}$
  pairs $\leftarrow \{\}$ ▷ Initialize eligible node pairs
  if scope = "intra-class" **then**
    pairs $\leftarrow \{(i,j) \mid i,j \in \mathcal{V}, \mathbf{Y}_i = \mathbf{Y}_j, i \neq j\}$ ▷ Same class
  else if scope = "inter-class" **then**
    pairs $\leftarrow \{(i,j) \mid i,j \in \mathcal{V}, \mathbf{Y}_i \neq \mathbf{Y}_j\}$ ▷ Different classes
  else if scope = "random" **then**
    pairs $\leftarrow \{(i,j) \mid i,j \in \mathcal{V}, i \neq j\}$ ▷ Any nodes
  end if
  $n_{\text{swap}} \leftarrow \lfloor \beta \cdot |\mathcal{V}| \rfloor$ ▷ Number of nodes to swap
  $n_{\text{pairs}} \leftarrow \lceil n_{\text{swap}}/2 \rceil$ ▷ Number of pairs needed
  selected_pairs $\leftarrow$ SampleWithoutReplacement(pairs, n_{pairs})
  $\tilde{\mathcal{T}} \leftarrow \mathcal{T}$ ▷ Initialize with original texts
  for each (i,j) in selected_pairs **do**
    $\tilde{\mathcal{T}}_i, \tilde{\mathcal{T}}_j \leftarrow \mathcal{T}_j, \mathcal{T}_i$ ▷ Swap texts
  end for
  $\mathcal{G}' \leftarrow (\mathcal{V}, \tilde{\mathcal{T}}, \mathbf{A})$ ▷ Update with swapped features
  return $\mathcal{G}'$
end function

character-level edit function is presented in Algorithm 1. The TEXTSWAP function given in Algorithm 3.

H Computational Cost

Table 13 compares the computational costs of TNT-OOD against GNNSafe and NODESafe. As expected, TNT-OOD's cross-attention mechanism and HyperNetwork module increase memory usage and training time compared to standard GCN models with post-hoc detection. However, we significantly reduced these costs by implementing batch-based contrastive objective calculations and a LowRank approach for the HyperNetwork, decreasing memory requirements by up to 80%. Importantly, inference speeds remain comparable across methods, demonstrating TNT-OOD's practical viability for real-world applications – specially

		GNNSafe	NODESafe	TNT-OOD	TNT w/ batch
Cora	Train (s)	2.95	3.50	11.85	8.95
	Infer (s)	0.01	0.02	0.05	0.06
	Mem (MiB)	590	590	1022	852
Citeseer	Train (s)	2.60	2.79	11.32	12.90
	Infer (s)	0.03	0.04	0.07	0.07
	Mem (MiB)	654	654	2029	893
Pubmed	Train (s)	4.50	4.60	36.20	17.58
	Infer (s)	0.15	0.16	0.20	0.22
	Mem (MiB)	852	852	9045	1956
Arxiv	Train (s)	15.00	16.00	OOM	276.48
	Infer (s)	0.76	0.75	OOM	0.82
	Mem (MiB)	3326	3326	OOM	16352
Bookhis	Train (s)	4.50	4.62	522.34	49.82
	Infer (s)	0.18	0.18	0.27	0.25
	Mem (MiB)	1742	1742	12936	5222

Table 13: Computational Performance on a single NVIDIA RTX A6000 GPU (48GB). Train: Convergence time; Infer: Inference time; Mem: Peak memory usage; OOM: Out of memory

		multilingual-E5-Small				All-MiniLM-L6-v2			
Dataset	**Method**	**AUROC**	**AUPR**	**FPR95**	**ID Acc**	**AUROC**	**AUPR**	**FPR95**	**ID Acc**
Cora	Energy	82.60	85.89	60.36	82.50	81.17	83.95	60.42	82.64
	GNNSafe	90.01	91.73	36.66		89.19	89.58	34.21	
	TNT-OOD	**89.52**	**90.31**	**33.61**	**83.27**	**91.65**	**90.55**	**24.91**	82.54
PubMed	Energy	71.88	68.98	74.13	79.07	71.18	68.03	71.40	78.23
	GNNSafe	84.73	81.44	42.74		84.33	80.48	39.19	
	TNT-OOD	**87.29**	**83.64**	**35.76**	**74.93**	**88.82**	**86.67**	**33.13**	78.85
DBLP	Energy	72.82	68.43	63.66	75.91	87.31	86.73	46.63	77.41
	GNNSafe	80.90	74.67	42.66		88.70	81.71	30.46	
	TNT-OOD	**82.94**	**76.85**	**38.29**	**73.07**	**89.40**	**81.91**	**25.78**	77.42

Table 14: Performance Comparison of Text Encoders.

with the superior detection performance demonstrated by TNT-OOD. As an early contribution to this underexplored field, we believe our work establishes a foundation for developing both effective and efficient OOD detection methods for text-rich networks.

I Extended Experiment Results

In this section, we provide the extended results complementing the overall performance table provided in Table 2 in the main paper. Specifically, the OOD detection performance for each OOD scenario at different configurations on each dataset are provided in Tables 15 to 25. The scores reported in the subset tables are averaged across three runs, with variance reflecting performance deviation across the three seeds.

Additionally, to investigate the effectiveness and adptatibility of TNT-OOD and OOD scenarios, we substitute language models from All-MiniLM-L6-v2 to Multilingual-E5-Small (Wang et al., 2024a) with the same embedding size. We tuned the learning rate to ensure comparable ID Acc. As in Table 14, TNT-OOD shows great performance on both PLM text encoders against the best non-propagation and propagation-based baselines.

LLM Prompt for Thematically-Guided OOD Label Selection

You are tasked with selecting [NUM_OOD_CLASSES] classes from the following list to be designated as Out-of-Distribution (OOD), while the remaining classes will be In-Distribution (ID). Your selection must be based on [SELECTION_CRITERIA].
Here are the categories and their descriptions:
[CATEGORY_NAMES_AND_DESCRIPTIONS]
Selection Criteria:

- If [SELECTION_CRITERIA] is "thematic_similarity": Select OOD classes that are pairwise similar to some of the remaining ID classes. This creates a challenging scenario where semantically related classes are separated into ID and OOD.
- If [SELECTION_CRITERIA] is "thematic_dissimilarity": Select OOD classes that are most thematically dissimilar from the remaining ID classes. This creates clearer semantic boundaries between ID and OOD.

Important instructions:

1. You must select EXACTLY [NUM_OOD_CLASSES] classes as OOD.
2. Provide your response as a JSON object with two keys: "id_classes" and "ood_classes".
3. Each key should contain an array of integers representing class indices (0-indexed).
4. Do not include any explanations, reasoning, or additional text.

Example output format:
```
{"id_classes": [0, 1, 3, 5], "ood_classes": [2, 4, 6]}
```

Figure 4: LLM prompt for OOD class selection.

LLM Prompt for LLM as OOD Detector on Label Shifts

You are given a set of known categories indexed from 0 to [NUM_ID_CLASSES]-1, where they are in-distribution (ID) classes. For each test case, you will be given a 'Text'. The given set of known categories are: [CATEGORIES]. Your task is: If the text clearly belongs to one of the ID classes, output the numeric label index (from 0 to [NUM_ID_CLASSES]-1). If the text does not belong to any ID class (i.e., is out-of-distribution), output [NUM_ID_CLASSES].

Important instructions: Output only the predicted label index (an integer from 0 to [NUM_ID_CLASSES]), with no explanation or extra text.

Example output format:
```
3
```

Figure 5: LLM prompt for LLM as Detector methods on label shifts.

OOD Type	Metric	w/o Prop.					w/ Prop.		
		MAHA	MSP	ODIN	NECO	Energy	GNNSafe	NodeSafe	TNT-OOD
Feature (0.50)	AUROC	52.80 ± 1.32	68.77 ± 0.60	66.80 ± 0.51	69.37 ± 3.48	78.16 ± 0.42	82.26 ± 0.77	84.36 ± 0.83	91.22 ± 0.23
	AUPR	54.18 ± 0.92	70.82 ± 0.35	69.12 ± 0.08	73.27 ± 2.89	81.18 ± 0.40	84.37 ± 0.03	86.03 ± 0.70	89.56 ± 0.16
	FPR95	86.22 ± 1.18	73.55 ± 1.71	75.71 ± 1.48	80.13 ± 7.35	65.99 ± 1.41	61.54 ± 8.34	58.16 ± 1.96	32.08 ± 2.55
Feature (0.70)	AUROC	49.58 ± 2.40	79.24 ± 1.02	76.69 ± 1.12	75.22 ± 7.73	89.09 ± 0.21	93.37 ± 0.22	93.92 ± 0.53	97.22 ± 0.19
	AUPR	51.81 ± 1.38	80.61 ± 0.95	78.06 ± 0.94	79.53 ± 6.34	91.11 ± 0.21	94.33 ± 0.24	94.57 ± 0.30	95.89 ± 0.22
	FPR95	85.82 ± 1.40	55.29 ± 2.93	60.01 ± 2.53	69.91 ± 13.27	42.26 ± 3.45	28.08 ± 2.09	27.61 ± 4.07	8.48 ± 1.35
Feature (0.90)	AUROC	47.78 ± 3.58	82.86 ± 1.62	80.01 ± 2.20	82.99 ± 2.85	92.89 ± 0.29	97.20 ± 0.18	97.28 ± 0.22	98.37 ± 0.23
	AUPR	50.73 ± 1.88	83.71 ± 1.76	80.72 ± 2.17	83.88 ± 4.04	94.40 ± 0.28	97.53 ± 0.28	97.15 ± 0.22	97.76 ± 0.36
	FPR95	85.78 ± 1.84	46.58 ± 3.25	51.13 ± 3.34	51.77 ± 3.82	31.43 ± 2.38	11.09 ± 0.90	11.01 ± 1.25	3.42 ± 0.08
Structure (Mild)	AUROC	51.28 ± 1.00	66.22 ± 0.87	68.24 ± 1.04	68.26 ± 0.67	74.01 ± 0.15	81.45 ± 0.63	83.06 ± 0.34	79.08 ± 0.32
	AUPR	54.66 ± 0.63	68.76 ± 0.68	71.07 ± 0.74	72.22 ± 0.95	78.12 ± 0.32	80.77 ± 0.12	82.71 ± 0.57	77.24 ± 0.12
	FPR95	87.75 ± 1.39	79.27 ± 2.70	77.40 ± 2.52	80.48 ± 0.50	75.37 ± 1.38	59.61 ± 7.23	56.87 ± 3.03	62.86 ± 3.64
Structure (Medium)	AUROC	48.05 ± 1.29	67.77 ± 1.14	69.62 ± 1.26	67.40 ± 2.54	75.39 ± 0.55	85.08 ± 0.62	86.61 ± 0.51	83.30 ± 0.28
	AUPR	52.15 ± 0.73	70.31 ± 0.88	72.28 ± 0.98	71.86 ± 2.20	79.50 ± 0.51	85.02 ± 0.45	86.77 ± 0.78	82.06 ± 0.32
	FPR95	90.01 ± 0.64	79.30 ± 2.03	78.16 ± 1.62	83.69 ± 2.67	74.90 ± 1.09	51.05 ± 4.45	49.10 ± 4.08	51.89 ± 1.62
Structure (Strong)	AUROC	36.01 ± 1.88	76.18 ± 1.38	75.56 ± 1.64	70.27 ± 0.58	84.21 ± 0.51	97.03 ± 0.32	97.37 ± 0.25	97.42 ± 0.29
	AUPR	45.50 ± 0.77	78.12 ± 1.32	77.34 ± 1.42	74.70 ± 1.56	87.80 ± 0.37	97.41 ± 0.29	97.53 ± 0.23	96.83 ± 0.53
	FPR95	93.76 ± 0.39	66.51 ± 2.91	67.46 ± 2.80	78.45 ± 4.22	67.84 ± 4.05	11.88 ± 1.90	10.75 ± 1.01	6.50 ± 0.81
Text (Synonym)	AUROC	24.70 ± 7.54	76.01 ± 1.72	73.16 ± 2.07	44.96 ± 34.24	85.44 ± 2.19	91.46 ± 2.19	80.67 ± 11.15	96.57 ± 1.28
	AUPR	41.72 ± 2.31	74.45 ± 1.79	71.21 ± 2.08	46.17 ± 35.07	84.95 ± 3.67	89.57 ± 2.81	79.16 ± 10.06	93.49 ± 1.96
	FPR95	96.65 ± 2.03	52.03 ± 3.96	55.11 ± 3.88	81.82 ± 17.72	44.71 ± 4.40	23.48 ± 4.22	43.04 ± 17.06	7.77 ± 3.74
Text (Antonym)	AUROC	25.31 ± 7.28	74.30 ± 2.37	71.86 ± 2.65	74.46 ± 2.71	82.66 ± 2.55	89.00 ± 2.90	76.77 ± 13.09	95.76 ± 1.63
	AUPR	41.90 ± 2.32	72.87 ± 2.38	69.84 ± 2.35	75.28 ± 2.89	82.55 ± 3.95	86.96 ± 3.48	76.38 ± 10.78	92.44 ± 2.38
	FPR95	96.63 ± 1.99	54.22 ± 4.97	56.22 ± 4.88	65.67 ± 1.87	51.19 ± 3.88	30.30 ± 5.84	48.47 ± 15.03	10.12 ± 4.90
Text Swap (Both)	AUROC	44.08 ± 2.14	81.46 ± 1.48	79.80 ± 1.89	82.86 ± 2.77	90.70 ± 0.47	96.41 ± 0.39	96.64 ± 0.39	97.12 ± 0.20
	AUPR	49.08 ± 1.02	82.34 ± 1.20	80.88 ± 1.38	85.24 ± 1.76	92.75 ± 0.26	96.86 ± 0.45	96.58 ± 0.41	96.20 ± 0.77
	FPR95	90.84 ± 1.13	50.84 ± 4.96	54.14 ± 4.70	54.03 ± 4.03	41.75 ± 3.06	13.81 ± 1.51	14.60 ± 1.31	7.59 ± 1.23
Text Swap (Intra)	AUROC	59.28 ± 1.22	49.48 ± 0.31	49.44 ± 0.34	53.13 ± 1.34	54.24 ± 0.37	56.71 ± 0.26	57.79 ± 0.61	66.71 ± 0.75
	AUPR	59.72 ± 1.06	54.97 ± 0.63	55.35 ± 0.64	59.64 ± 1.57	60.45 ± 0.13	60.93 ± 0.34	63.09 ± 0.47	68.82 ± 0.91
	FPR95	84.46 ± 0.37	91.80 ± 1.06	92.05 ± 0.87	92.99 ± 0.96	92.44 ± 0.64	91.62 ± 2.25	90.28 ± 0.22	87.35 ± 2.27
Text Swap (Inter)	AUROC	43.20 ± 2.38	81.18 ± 1.51	79.52 ± 1.90	80.46 ± 3.80	90.50 ± 0.61	96.20 ± 0.52	96.41 ± 0.52	97.10 ± 0.18
	AUPR	48.72 ± 1.12	82.02 ± 1.44	80.59 ± 1.60	84.19 ± 2.14	92.51 ± 0.33	96.66 ± 0.54	96.35 ± 0.39	96.01 ± 0.74
	FPR95	91.30 ± 0.67	51.50 ± 5.69	55.06 ± 5.66	67.84 ± 9.65	41.15 ± 3.71	15.46 ± 2.73	15.38 ± 3.10	7.50 ± 1.09
Semantic (0.75)	AUROC	41.73 ± 1.74	76.92 ± 1.05	76.44 ± 1.69	77.38 ± 5.99	85.97 ± 0.39	97.52 ± 0.19	97.86 ± 0.19	97.33 ± 0.29
	AUPR	48.11 ± 0.91	79.27 ± 0.95	79.53 ± 1.31	80.57 ± 5.79	89.65 ± 0.34	97.90 ± 0.28	97.89 ± 0.23	96.52 ± 0.51
	FPR95	93.68 ± 1.01	67.15 ± 1.93	68.91 ± 1.73	70.08 ± 4.07	64.60 ± 1.38	9.25 ± 1.25	7.64 ± 0.56	7.04 ± 0.98
Semantic (0.85)	AUROC	43.36 ± 1.70	74.31 ± 0.94	75.03 ± 1.02	70.79 ± 2.09	82.16 ± 0.27	96.07 ± 0.19	96.64 ± 0.25	95.51 ± 0.20
	AUPR	49.14 ± 0.97	77.28 ± 1.08	78.65 ± 0.98	75.49 ± 1.51	86.69 ± 0.45	96.72 ± 0.34	96.95 ± 0.32	94.86 ± 0.27
	FPR95	94.54 ± 0.22	75.58 ± 1.51	76.06 ± 1.04	80.98 ± 1.61	75.97 ± 0.69	18.02 ± 0.90	16.36 ± 0.53	16.89 ± 1.41
Semantic (0.95)	AUROC	44.75 ± 0.92	63.43 ± 0.63	64.60 ± 0.58	63.69 ± 1.44	68.14 ± 0.39	87.90 ± 0.78	89.95 ± 0.43	85.82 ± 0.34
	AUPR	51.12 ± 0.73	67.24 ± 0.62	68.87 ± 0.44	67.86 ± 1.58	73.80 ± 0.56	90.42 ± 0.44	91.99 ± 0.47	87.61 ± 0.20
	FPR95	96.58 ± 0.05	88.12 ± 0.71	87.56 ± 0.85	86.83 ± 1.75	88.70 ± 0.30	61.12 ± 7.45	47.34 ± 3.02	55.16 ± 4.14
	ID Acc	82.64 ± 0.85	82.64 ± 0.85	82.64 ± 0.85	82.64 ± 0.85	82.64 ± 0.85	82.64 ± 0.85	81.80 ± 0.55	82.54 ± 0.25
Label (Random 0 1 2)	AUROC	30.05 ± 8.69	84.77 ± 3.37	83.51 ± 4.05	78.11 ± 0.42	87.17 ± 1.43	90.32 ± 0.31	90.87 ± 0.48	93.25 ± 1.07
	AUPR	42.14 ± 3.38	84.94 ± 4.61	83.57 ± 5.49	81.01 ± 2.52	88.49 ± 1.32	86.34 ± 0.34	86.94 ± 0.51	88.95 ± 0.90
	FPR95	97.67 ± 1.31	66.15 ± 24.00	67.07 ± 23.35	69.09 ± 9.37	49.10 ± 1.80	36.03 ± 1.98	34.31 ± 3.76	22.20 ± 4.39
	ID Acc	91.38 ± 0.50	91.38 ± 0.50	91.38 ± 0.50	91.38 ± 0.50	91.38 ± 0.50	91.38 ± 0.50	86.76 ± 0.32	91.14 ± 0.16
Label (Dissimilar 1 3 4)	AUROC	36.73 ± 1.57	86.82 ± 0.96	86.48 ± 1.31	82.84 ± 0.40	83.65 ± 12.81	81.28 ± 12.76	77.42 ± 10.06	91.63 ± 0.25
	AUPR	55.19 ± 1.34	90.92 ± 1.17	90.93 ± 1.14	89.44 ± 0.56	87.10 ± 12.24	80.63 ± 9.27	78.36 ± 7.68	88.00 ± 0.10
	FPR95	95.61 ± 0.19	44.39 ± 5.60	47.17 ± 7.20	62.08 ± 2.20	47.59 ± 20.45	42.47 ± 26.56	53.29 ± 19.16	20.58 ± 1.40
	ID Acc	91.06 ± 0.87	91.06 ± 0.87	91.06 ± 0.87	91.06 ± 0.87	91.06 ± 0.87	91.06 ± 0.87	89.48 ± 0.49	91.43 ± 0.27
Label (Similar 2 5 6)	AUROC	47.60 ± 4.43	78.44 ± 2.95	77.53 ± 3.13	71.91 ± 9.11	80.42 ± 0.57	82.68 ± 0.62	78.10 ± 6.28	83.65 ± 1.37
	AUPR	39.82 ± 3.63	68.14 ± 6.14	67.04 ± 6.47	63.54 ± 11.55	72.47 ± 1.06	69.22 ± 2.76	63.73 ± 8.60	67.97 ± 1.92
	FPR95	94.83 ± 1.42	61.42 ± 2.40	63.30 ± 1.94	74.61 ± 11.35	66.52 ± 2.70	57.39 ± 2.94	63.47 ± 9.83	46.40 ± 2.75
	ID Acc	87.82 ± 0.61	87.82 ± 0.61	87.82 ± 0.61	87.82 ± 0.61	87.82 ± 0.61	87.82 ± 0.61	87.97 ± 0.97	87.40 ± 0.28

Table 15: Performance comparison of OOD detection methods on **Cora dataset** across different OOD shifts and configurations. MAHA is short for Mahalanobis.

OOD Type	Metric	w/o Prop.					w/ Prop.		
		MAHA	MSP	ODIN	NECO	Energy	GNNSafe	NodeSafe	TNT-OOD
Feature (0.50)	AUROC	74.14 ± 0.09	78.06 ± 0.27	76.59 ± 0.25	69.42 ± 2.61	81.87 ± 0.27	82.43 ± 0.83	79.95 ± 3.56	84.23 ± 6.18
	AUPR	70.51 ± 0.94	78.86 ± 0.27	77.5 ± 0.31	71.34 ± 2.32	82.16 ± 0.09	79.05 ± 0.53	76.21 ± 3.78	78.89 ± 6.01
	FPR95	61.04 ± 0.29	62.95 ± 0.99	66.03 ± 0.77	77.73 ± 4.8	60.56 ± 1.82	52.47 ± 4.11	55.85 ± 5.62	38.11 ± 11.54
Feature (0.70)	AUROC	77.41 ± 0.26	88.38 ± 0.57	87.06 ± 0.51	78.71 ± 5.74	90.31 ± 0.18	91.2 ± 0.74	88.26 ± 3.11	91.85 ± 3.62
	AUPR	72.91 ± 0.85	89.14 ± 0.44	87.9 ± 0.41	80.05 ± 5.96	90.6 ± 0.19	86.61 ± 0.62	82.93 ± 3.69	86.63 ± 3.63
	FPR95	55.2 ± 1.03	42.97 ± 1.58	46.47 ± 1.27	65.55 ± 10.61	40.59 ± 2.34	27.06 ± 4.12	30.24 ± 5.71	22.19 ± 11.22
Feature (0.90)	AUROC	75.06 ± 0.4	90.6 ± 0.7	89.71 ± 0.76	82.99 ± 5.33	91.58 ± 0.18	93.31 ± 0.36	90.77 ± 2.50	94.3 ± 1.57
	AUPR	71 ± 0.78	92.08 ± 0.64	91.3 ± 0.73	84.78 ± 4.96	92.44 ± 0.2	88.49 ± 0.45	85.06 ± 3.14	89.49 ± 1.47
	FPR95	61.12 ± 1.19	41.43 ± 0.84	44.56 ± 1.16	57.25 ± 9.26	40.37 ± 1.87	20.89 ± 2.54	23.82 ± 6.17	16.74 ± 7.04
Structure (Mild)	AUROC	65.61 ± 0.32	67.91 ± 0.7	68.66 ± 0.74	42.78 ± 30.29	70.96 ± 0.61	71.55 ± 0.77	71.27 ± 2.18	71.91 ± 1.09
	AUPR	63.8 ± 0.19	71.92 ± 0.35	73.26 ± 0.37	45.63 ± 32.3	74.03 ± 0.76	69.22 ± 0.48	68.46 ± 2.34	69.69 ± 1.15
	FPR95	80.01 ± 0.93	84.7 ± 1.73	85.04 ± 1.84	91.86 ± 5.77	84.61 ± 1.65	78.27 ± 2.44	80.27 ± 2.38	73.45 ± 0.7
Structure (Medium)	AUROC	62.55 ± 0.22	69.34 ± 0.4	69.92 ± 0.45	65.08 ± 2.15	71.74 ± 0.39	75.86 ± 0.82	75.49 ± 2.76	76.76 ± 1.49
	AUPR	60.91 ± 0.27	73.42 ± 0.2	74.58 ± 0.28	69.36 ± 1.86	75.06 ± 0.77	73.98 ± 0.53	73.02 ± 2.60	75.04 ± 1.32
	FPR95	86.81 ± 0.82	84.66 ± 1.28	84.68 ± 1.24	86.84 ± 1.21	84.93 ± 1.31	70.93 ± 1.93	70.52 ± 4.48	67.47 ± 0.52
Structure (Strong)	AUROC	52.79 ± 0.78	78.28 ± 0.7	77.9 ± 0.76	71.97 ± 1.3	78.74 ± 0.51	92.08 ± 0.43	90.29 ± 2.11	94.68 ± 0.44
	AUPR	53.36 ± 0.5	82.48 ± 0.55	82.34 ± 0.67	75.96 ± 2.48	82.76 ± 0.93	89.15 ± 0.3	87.30 ± 2.11	91.74 ± 0.28
	FPR95	96.28 ± 1.15	83.06 ± 0.09	83.18 ± 0.52	83.65 ± 0.51	83.61 ± 0.67	26.33 ± 3.03	28.11 ± 5.40	15.71 ± 2.66
Text (Synonym)	AUROC	12.35 ± 2.68	95.83 ± 0.42	95.02 ± 0.51	86.59 ± 7.02	96.58 ± 2.15	95.25 ± 1.3	91.77 ± 5.68	96.73 ± 0.24
	AUPR	37.1 ± 0.59	96.1 ± 0.37	95.25 ± 0.43	85.93 ± 8.59	94.97 ± 4.62	90.45 ± 2.24	86.24 ± 7.49	92.36 ± 0.22
	FPR95	99.81 ± 0.16	17.77 ± 2.14	20.42 ± 2.37	39.05 ± 9.35	10.39 ± 3.72	8.61 ± 1.69	15.16 ± 10.14	5.51 ± 1.31
Text (Antonym)	AUROC	10.89 ± 2.52	94.86 ± 0.37	94.07 ± 0.47	88.4 ± 4.44	95.93 ± 2.17	94.76 ± 1.54	91.99 ± 5.16	96.77 ± 0.25
	AUPR	36.78 ± 0.51	95.15 ± 0.44	94.32 ± 0.52	87.04 ± 3.92	94.45 ± 4.57	89.99 ± 2.35	86.73 ± 6.54	92.44 ± 0.22
	FPR95	99.87 ± 0.1	21.2 ± 0.86	23.66 ± 1.08	34.29 ± 13.87	13.8 ± 3.45	11.2 ± 3.84	16.39 ± 10.57	4.94 ± 1.06
Text Swap (Both)	AUROC	62.07 ± 0.43	85.25 ± 0.45	84.7 ± 0.44	47.23 ± 33.4	85.76 ± 0.24	89.81 ± 0.45	87.56 ± 2.12	93.22 ± 0.44
	AUPR	60.49 ± 0.41	87.94 ± 0.4	87.57 ± 0.42	48.56 ± 34.34	88.07 ± 0.19	85.65 ± 0.14	82.55 ± 2.50	89.01 ± 0.34
	FPR95	82.2 ± 0.57	61.69 ± 1.13	62.83 ± 0.91	82.42 ± 12.43	63.02 ± 2.94	38.05 ± 4.42	38.88 ± 4.08	26.86 ± 2.76
Text Swap (Intra)	AUROC	63.25 ± 0.33	56.47 ± 0.72	56.14 ± 0.71	36.21 ± 25.64	59.77 ± 0.44	60.83 ± 0.45	61.85 ± 1.64	75.89 ± 2.1
	AUPR	62.65 ± 0.57	61.17 ± 0.79	61.01 ± 0.81	39.98 ± 28.31	63.65 ± 0.9	62.15 ± 0.77	62.36 ± 1.69	75.63 ± 1.82
	FPR95	82.26 ± 0.38	90.71 ± 0.73	90.93 ± 0.65	94.44 ± 3.94	90.5 ± 0.78	86.15 ± 1.94	85.91 ± 0.93	73.24 ± 4.41
Text Swap (Inter)	AUROC	61.35 ± 0.38	85.19 ± 0.49	84.62 ± 0.46	74.81 ± 4.5	85.63 ± 0.2	89.71 ± 0.38	87.49 ± 2.23	93.12 ± 0.47
	AUPR	60.08 ± 0.46	87.74 ± 0.48	87.38 ± 0.48	78 ± 5.3	87.86 ± 0.08	85.48 ± 0.2	82.41 ± 2.54	88.98 ± 0.31
	FPR95	83.59 ± 0.14	62.08 ± 0.53	63.06 ± 0.57	75.23 ± 3.16	63.86 ± 2.55	37.4 ± 4.47	38.98 ± 4.60	27.4 ± 3.8
Semantic (0.75)	AUROC	66.54 ± 0.17	80.29 ± 0.59	80.67 ± 0.68	72.56 ± 0.66	81.61 ± 0.14	91.04 ± 0.45	89.30 ± 2.49	93.98 ± 0.64
	AUPR	62.9 ± 0.38	81.82 ± 0.65	82.75 ± 0.72	73.88 ± 0.73	83.02 ± 0.91	86.92 ± 0.59	84.74 ± 2.84	90.21 ± 0.66
	FPR95	93.79 ± 0.16	66.8 ± 1.35	67.24 ± 0.94	76.12 ± 1.19	68.89 ± 2.88	24.14 ± 1.37	25.19 ± 4.60	14.42 ± 1.74
Semantic (0.85)	AUROC	65.12 ± 0.12	72.43 ± 0.61	73.15 ± 0.69	66.86 ± 3.04	73.54 ± 0.33	85.84 ± 0.81	85.25 ± 2.73	88.9 ± 1.5
	AUPR	62.46 ± 0.29	73.6 ± 0.6	75.07 ± 0.64	68.82 ± 2.51	75.04 ± 1.19	82.54 ± 1.04	81.54 ± 2.82	85.27 ± 1.47
	FPR95	95.23 ± 0.45	77.19 ± 1.7	77.07 ± 1.51	82.37 ± 1.3	77.99 ± 2.29	37.82 ± 2.25	35.80 ± 4.96	28.49 ± 3.72
Semantic (0.95)	AUROC	60.44 ± 0.18	55.35 ± 0.31	56.08 ± 0.32	51.95 ± 1.09	55.9 ± 0.71	69.31 ± 1.66	71.51 ± 2.53	70.29 ± 1.21
	AUPR	61.53 ± 0.24	58.56 ± 0.36	59.38 ± 0.29	56.66 ± 0.74	59.65 ± 0.99	71.89 ± 1.28	73.17 ± 1.99	71.99 ± 1.01
	FPR95	98.38 ± 0.39	89.31 ± 2.13	89.5 ± 2.15	92.14 ± 1.38	90.02 ± 1.94	71.29 ± 3.88	65.43 ± 3.42	64.42 ± 1.44
	ID Acc	74.06 ± 0.38	74.06 ± 0.38	74.06 ± 0.38	74.06 ± 0.38	74.06 ± 0.38	74.06 ± 0.38	74.62 ± 0.21	74.11 ± 1.11
Label (Random 1 3 5)	AUROC	27.1 ± 1.77	75.59 ± 3.39	74.47 ± 4.06	61.08 ± 5.12	76.12 ± 1.81	76.28 ± 0.28	75.48 ± 4.59	77.67 ± 1.88
	AUPR	37.91 ± 0.53	75.92 ± 2.47	75.04 ± 3.2	64.27 ± 3.3	74.81 ± 4.71	68.94 ± 1.07	67.89 ± 3.06	69.86 ± 1.51
	FPR95	98.07 ± 0.42	77.25 ± 6.93	77.97 ± 6.83	92.03 ± 5.34	76.16 ± 2.66	66.01 ± 5.34	67.10 ± 4.64	60.99 ± 4.37
	ID Acc	88.5 ± 0.5	88.5 ± 0.5	88.5 ± 0.5	88.5 ± 0.5	88.5 ± 0.5	88.5 ± 0.5	88.96 ± 0.27	88.54 ± 0.38
Label (Dissimilar 2 3 4)	AUROC	36.04 ± 1.5	79.91 ± 1.56	80.5 ± 1.5	74.09 ± 1.6	76.32 ± 4.81	71.23 ± 2.79	68.38 ± 0.70	75.16 ± 1.39
	AUPR	51.84 ± 0.53	84.02 ± 1.62	85.36 ± 1.06	78.65 ± 2.17	80.13 ± 5.84	70.84 ± 2.08	69.02 ± 1.37	74.08 ± 1.14
	FPR95	96.31 ± 1.31	61.62 ± 2.13	63.5 ± 2.58	75.62 ± 5.59	63.25 ± 4.11	59.35 ± 4.92	60.42 ± 2.72	57.73 ± 2.12
	ID Acc	77.71 ± 0.31	77.71 ± 0.31	77.71 ± 0.31	77.71 ± 0.31	77.71 ± 0.31	77.71 ± 0.31	76.63 ± 1.28	78.51 ± 0.53
Label (Similar 0 1 5)	AUROC	27 ± 1.37	78.88 ± 4.05	77.56 ± 4.83	72.04 ± 3.49	73.35 ± 6.2	73.72 ± 3.97	79.08 ± 1.24	83.3 ± 0.71
	AUPR	34.15 ± 0.44	75.82 ± 4.27	74.85 ± 5.47	66.86 ± 5.89	67.98 ± 7.74	61.79 ± 3.58	66.83 ± 1.60	70.91 ± 1.04
	FPR95	98.16 ± 0.38	70.6 ± 6.54	71.66 ± 6.2	76.34 ± 7.22	73.37 ± 3.15	64.24 ± 5.03	61.83 ± 0.77	49.42 ± 1.61
	ID Acc	86.53 ± 0.26	86.53 ± 0.26	86.53 ± 0.26	86.53 ± 0.26	86.53 ± 0.26	86.53 ± 0.26	86.35±0.22	87.27 ± 0.26

Table 16: Performance comparison of OOD detection methods on **Citeseer dataset** across different OOD shifts and configurations.

OOD Type	Metric	w/o Prop.					w/ Prop.		
		MAHA	MSP	ODIN	NECO	Energy	GNNSafe	NodeSafe	TNT-OOD
Feature (0.50)	AUROC	89.92 ± 0.23	65.82 ± 0.02	64.10 ± 0.05	69.06 ± 0.16	71.08 ± 0.23	44.97 ± 0.07	44.89 ± 0.09	54.98 ± 1.42
	AUPR	96.73 ± 0.07	87.31 ± 0.03	86.49 ± 0.01	89.16 ± 0.04	90.29 ± 0.07	80.16 ± 0.04	80.09 ± 0.03	84.03 ± 0.50
	FPR95	59.73 ± 1.35	92.93 ± 0.16	94.48 ± 0.14	94.87 ± 0.17	94.94 ± 0.15	95.88 ± 0.13	96.08 ± 0.09	85.40 ± 3.03
Feature (0.70)	AUROC	91.90 ± 0.23	72.14 ± 0.13	70.76 ± 0.08	77.19 ± 0.16	79.59 ± 0.24	46.08 ± 0.08	45.98 ± 0.09	58.29 ± 1.47
	AUPR	97.39 ± 0.07	89.88 ± 0.07	89.29 ± 0.05	92.19 ± 0.04	93.37 ± 0.07	80.14 ± 0.04	80.15 ± 0.02	85.32 ± 0.49
	FPR95	42.58 ± 1.75	83.09 ± 0.34	86.07 ± 0.24	84.60 ± 0.55	84.96 ± 0.51	90.92 ± 0.20	90.55 ± 0.09	73.91 ± 3.65
Feature (0.90)	AUROC	89.19 ± 0.30	69.00 ± 0.26	68.34 ± 0.24	75.97 ± 0.27	78.73 ± 0.31	43.33 ± 0.11	43.73 ± 0.11	58.80 ± 1.29
	AUPR	96.38 ± 0.10	88.36 ± 0.12	88.05 ± 0.11	91.46 ± 0.09	92.89 ± 0.11	77.46 ± 0.04	77.83 ± 0.04	85.25 ± 0.50
	FPR95	51.88 ± 1.16	83.53 ± 0.28	84.52 ± 0.24	77.40 ± 0.71	76.80 ± 0.75	88.50 ± 0.19	87.51 ± 0.08	68.70 ± 4.08
Structure (Mild)	AUROC	99.35 ± 0.01	88.80 ± 0.11	90.02 ± 0.06	93.33 ± 0.06	95.61 ± 0.05	9.88 ± 0.10	10.26 ± 0.12	35.55 ± 0.12
	AUPR	99.83 ± 0.00	96.63 ± 0.04	97.08 ± 0.02	98.06 ± 0.02	98.77 ± 0.02	59.15 ± 0.02	59.24 ± 0.03	65.91 ± 0.12
	FPR95	1.00 ± 0.04	57.92 ± 0.46	55.89 ± 0.39	38.84 ± 0.11	27.21 ± 0.27	98.67 ± 0.04	98.13 ± 0.07	78.11 ± 0.22
Structure (Medium)	AUROC	99.23 ± 0.01	87.84 ± 0.13	89.15 ± 0.07	92.70 ± 0.06	95.14 ± 0.05	9.68 ± 0.10	10.06 ± 0.11	35.16 ± 0.13
	AUPR	99.80 ± 0.00	96.30 ± 0.04	96.79 ± 0.02	97.85 ± 0.02	98.63 ± 0.02	59.10 ± 0.02	59.19 ± 0.03	65.81 ± 0.12
	FPR95	1.42 ± 0.03	60.30 ± 0.36	58.43 ± 0.29	41.83 ± 0.29	30.47 ± 0.22	98.80 ± 0.03	98.28 ± 0.08	79.27 ± 0.23
Structure (Strong)	AUROC	99.97 ± 0.00	98.70 ± 0.09	98.67 ± 0.11	99.66 ± 0.01	99.89 ± 0.02	18.64 ± 0.08	18.35 ± 0.28	40.54 ± 0.08
	AUPR	99.99 ± 0.00	99.50 ± 0.04	99.48 ± 0.05	99.86 ± 0.01	99.92 ± 0.01	61.14 ± 0.02	61.07 ± 0.07	67.15 ± 0.09
	FPR95	0.09 ± 0.01	4.94 ± 0.26	5.06 ± 0.33	1.26 ± 0.03	0.28 ± 0.06	81.88 ± 0.08	82.02 ± 0.30	59.93 ± 0.08
Text (Synonym)	AUROC	11.07 ± 0.52	71.46 ± 0.99	67.27 ± 1.04	82.09 ± 0.94	85.07 ± 0.90	50.22 ± 0.34	50.82 ± 0.24	38.37 ± 1.70
	AUPR	59.83 ± 0.23	89.90 ± 0.44	88.00 ± 0.49	94.49 ± 0.33	95.81 ± 0.28	83.18 ± 0.12	83.51 ± 0.09	73.74 ± 0.67
	FPR95	99.96 ± 0.01	90.26 ± 1.10	93.72 ± 0.73	84.08 ± 1.88	84.89 ± 1.64	89.67 ± 0.65	87.73 ± 0.56	98.34 ± 4.11
Text (Antonym)	AUROC	12.19 ± 0.44	70.40 ± 0.88	66.31 ± 0.91	80.22 ± 0.73	82.93 ± 0.67	49.61 ± 0.25	50.50 ± 0.19	38.85 ± 2.15
	AUPR	60.27 ± 0.20	89.33 ± 0.44	87.41 ± 0.49	93.83 ± 0.27	95.14 ± 0.22	82.97 ± 0.09	83.38 ± 0.07	74.28 ± 0.83
	FPR95	99.95 ± 0.01	91.11 ± 0.63	94.27 ± 0.40	87.02 ± 1.07	88.03 ± 0.85	91.01 ± 0.39	88.60 ± 0.48	98.43 ± 4.95
Text Swap (Both)	AUROC	64.09 ± 0.21	57.57 ± 0.09	57.91 ± 0.10	62.78 ± 0.46	64.24 ± 0.55	34.20 ± 0.18	35.73 ± 0.16	54.40 ± 1.13
	AUPR	86.33 ± 0.12	82.97 ± 0.06	83.15 ± 0.05	86.05 ± 0.20	87.34 ± 0.25	69.39 ± 0.07	70.64 ± 0.09	82.18 ± 0.38
	FPR95	89.78 ± 0.20	91.97 ± 0.08	91.46 ± 0.05	87.76 ± 0.28	87.64 ± 0.31	91.62 ± 0.27	90.73 ± 0.21	74.64 ± 2.88
Text Swap (Intra)	AUROC	66.13 ± 0.11	39.97 ± 0.06	40.10 ± 0.07	40.73 ± 0.16	41.60 ± 0.19	30.84 ± 0.07	32.54 ± 0.08	38.57 ± 1.17
	AUPR	87.71 ± 0.06	73.59 ± 0.05	73.65 ± 0.06	74.48 ± 0.10	75.53 ± 0.11	68.87 ± 0.04	70.15 ± 0.07	74.47 ± 0.41
	FPR95	89.46 ± 0.05	99.23 ± 0.00	99.20 ± 0.01	98.90 ± 0.01	98.86 ± 0.02	99.60 ± 0.01	99.60 ± 0.00	98.92 ± 3.70
Text Swap (Inter)	AUROC	63.74 ± 0.21	57.63 ± 0.07	57.97 ± 0.09	62.82 ± 0.45	64.27 ± 0.54	34.23 ± 0.18	35.74 ± 0.15	54.31 ± 1.12
	AUPR	86.17 ± 0.12	82.99 ± 0.05	83.17 ± 0.04	86.06 ± 0.20	87.35 ± 0.25	69.41 ± 0.08	70.66 ± 0.10	82.13 ± 0.37
	FPR95	90.07 ± 0.22	91.96 ± 0.10	91.45 ± 0.07	87.78 ± 0.24	87.66 ± 0.27	91.62 ± 0.26	90.76 ± 0.18	74.79 ± 2.80
Semantic (0.75)	AUROC	99.79 ± 0.00	96.01 ± 0.12	96.39 ± 0.10	98.61 ± 0.04	99.52 ± 0.06	17.98 ± 0.06	17.84 ± 0.24	40.07 ± 0.08
	AUPR	99.95 ± 0.00	98.58 ± 0.05	98.73 ± 0.04	99.51 ± 0.01	99.81 ± 0.02	60.97 ± 0.02	60.93 ± 0.06	66.98 ± 0.08
	FPR95	0.08 ± 0.00	14.91 ± 0.38	14.18 ± 0.40	5.06 ± 0.13	1.45 ± 0.24	83.04 ± 0.05	83.08 ± 0.25	60.36 ± 0.08
Semantic (0.85)	AUROC	99.36 ± 0.01	90.22 ± 0.09	91.26 ± 0.04	95.08 ± 0.06	97.49 ± 0.14	16.66 ± 0.08	16.59 ± 0.18	39.47 ± 0.09
	AUPR	99.85 ± 0.00	96.63 ± 0.05	97.09 ± 0.04	98.32 ± 0.02	99.15 ± 0.06	60.63 ± 0.02	60.66 ± 0.04	66.77 ± 0.07
	FPR95	0.25 ± 0.03	41.05 ± 0.77	41.06 ± 0.74	19.53 ± 0.20	10.07 ± 0.38	85.22 ± 0.10	85.32 ± 0.20	61.38 ± 0.10
Semantic (0.95)	AUROC	95.50 ± 0.14	53.90 ± 0.11	56.22 ± 0.06	59.18 ± 0.05	62.89 ± 0.02	6.26 ± 0.05	7.64 ± 0.06	23.04 ± 0.07
	AUPR	98.58 ± 0.04	80.23 ± 0.05	82.13 ± 0.03	82.58 ± 0.05	84.65 ± 0.08	58.27 ± 0.01	58.60 ± 0.01	62.52 ± 0.09
	FPR95	23.13 ± 1.07	92.64 ± 0.09	92.31 ± 0.10	88.33 ± 0.10	87.30 ± 0.09	99.08 ± 0.01	99.15 ± 0.06	93.45 ± 0.12
	ID Acc	68.54 ± 0.25	68.54 ± 0.25	68.54 ± 0.25	68.54 ± 0.25	68.54 ± 0.25	68.54 ± 0.25	69.09 ± 0.23	69.55 ± 0.36
Temporal Shift (2018)	AUROC	42.80 ± 0.19	56.56 ± 0.27	56.43 ± 0.29	58.69 ± 0.17	58.73 ± 0.18	64.28 ± 0.08	66.58 ± 0.18	65.56 ± 0.27
	AUPR	34.72 ± 0.08	45.35 ± 0.19	45.18 ± 0.21	46.33 ± 0.10	46.13 ± 0.12	50.63 ± 0.08	55.32 ± 0.42	51.80 ± 0.61
	FPR95	93.47 ± 0.06	87.71 ± 0.47	87.74 ± 0.45	87.20 ± 0.17	87.18 ± 0.17	82.20 ± 0.11	80.00 ± 0.27	79.27 ± 0.43
Temporal Shift (2019)	AUROC	40.93 ± 0.11	57.21 ± 0.31	57.03 ± 0.32	59.28 ± 0.19	59.31 ± 0.22	66.07 ± 0.11	68.56 ± 0.16	66.94 ± 0.22
	AUPR	40.23 ± 0.05	52.74 ± 0.21	52.53 ± 0.22	53.73 ± 0.08	53.53 ± 0.14	58.82 ± 0.09	63.80 ± 0.32	59.56 ± 0.20
	FPR95	94.13 ± 0.05	86.53 ± 0.49	86.63 ± 0.47	86.98 ± 0.24	87.00 ± 0.23	79.58 ± 0.25	77.12 ± 0.30	77.43 ± 0.17
Temporal Shift (2020)	AUROC	42.64 ± 0.18	60.44 ± 0.28	60.44 ± 0.30	64.90 ± 0.19	65.13 ± 0.21	75.51 ± 0.08	78.53 ± 0.26	76.25 ± 0.45
	AUPR	13.75 ± 0.03	21.51 ± 0.15	21.41 ± 0.15	23.09 ± 0.08	23.02 ± 0.10	30.10 ± 0.08	37.08 ± 0.57	30.82 ± 0.33
	FPR95	87.60 ± 0.22	80.81 ± 0.65	80.15 ± 0.59	72.69 ± 0.35	71.87 ± 0.42	55.47 ± 0.25	53.57 ± 0.23	54.28 ± 0.30
	ID Acc	67.82 ± 0.09	67.82 ± 0.09	67.82 ± 0.09	67.82 ± 0.09	67.82 ± 0.09	67.82 ± 0.09	66.44 ± 0.06	67.69 ± 0.09

Table 17: Performance comparison of OOD detection methods on **Arxiv dataset** across different OOD shifts and configurations.

OOD Type	Metric	w/o Prop.					w/ Prop.		
		MAHA	MSP	ODIN	NECO	Energy	GNNSafe	NodeSafe	TNT-OOD
Feature (0.50)	AUROC	72.25 ± 0.54	80.08 ± 2.10	77.91 ± 2.00	79.85 ± 5.36	85.43 ± 8.88	88.54 ± 9.34	90.29 ± 9.64	89.06 ± 1.16
	AUPR	66.23 ± 0.53	78.59 ± 1.73	75.60 ± 1.62	78.64 ± 4.96	85.09 ± 11.85	82.74 ± 9.95	83.66 ± 9.86	82.66 ± 1.11
	FPR95	82.39 ± 1.13	62.10 ± 5.84	65.42 ± 5.71	62.64 ± 8.88	55.12 ± 7.63	42.40 ± 13.10	32.93 ± 16.75	35.09 ± 2.64
Feature (0.70)	AUROC	75.05 ± 0.66	90.10 ± 1.52	87.84 ± 1.57	88.99 ± 0.44	93.31 ± 10.47	93.89 ± 9.47	94.32 ± 9.53	93.66 ± 0.41
	AUPR	68.40 ± 0.70	89.99 ± 1.06	86.83 ± 1.10	89.19 ± 0.67	93.74 ± 15.03	87.15 ± 10.85	87.16 ± 10.63	86.85 ± 0.37
	FPR95	80.37 ± 1.06	40.68 ± 5.57	44.71 ± 5.45	44.65 ± 5.15	35.74 ± 12.70	14.05 ± 20.05	8.64 ± 21.20	13.88 ± 3.24
Feature (0.90)	AUROC	72.19 ± 0.69	93.51 ± 0.99	92.11 ± 1.36	93.28 ± 2.08	94.59 ± 10.28	94.47 ± 8.57	94.61 ± 8.60	94.35 ± 0.16
	AUPR	65.90 ± 0.74	94.40 ± 0.86	92.44 ± 1.45	94.51 ± 1.45	95.64 ± 15.51	87.42 ± 10.48	87.25 ± 10.32	87.34 ± 0.14
	FPR95	85.54 ± 0.88	37.51 ± 4.30	40.23 ± 4.47	41.66 ± 9.15	36.11 ± 11.74	8.16 ± 18.76	6.67 ± 19.31	9.22 ± 0.91
Structure (Mild)	AUROC	75.70 ± 0.67	86.91 ± 1.25	87.59 ± 0.94	84.88 ± 1.23	92.13 ± 10.53	92.69 ± 10.32	91.83 ± 10.69	91.10 ± 0.71
	AUPR	63.82 ± 0.58	87.42 ± 1.94	88.30 ± 1.66	85.84 ± 1.36	91.75 ± 15.27	83.85 ± 12.11	83.07 ± 12.28	81.94 ± 0.58
	FPR95	43.09 ± 0.63	52.70 ± 2.01	52.39 ± 2.72	60.97 ± 1.73	37.97 ± 8.76	18.38 ± 13.92	23.37 ± 14.27	23.48 ± 5.30
Structure (Medium)	AUROC	73.97 ± 0.71	87.04 ± 1.17	87.78 ± 0.81	87.94 ± 1.51	92.25 ± 10.44	92.88 ± 10.10	91.96 ± 10.39	91.27 ± 0.76
	AUPR	62.17 ± 0.62	87.39 ± 1.81	88.40 ± 1.43	88.43 ± 1.18	91.87 ± 15.11	84.01 ± 11.94	83.14 ± 12.04	82.08 ± 0.62
	FPR95	45.36 ± 0.61	51.34 ± 2.46	51.36 ± 3.08	50.88 ± 2.25	36.94 ± 9.34	17.37 ± 14.28	22.09 ± 14.28	22.48 ± 5.67
Structure (Strong)	AUROC	70.24 ± 1.39	99.41 ± 0.63	98.73 ± 1.62	99.41 ± 2.44	99.78 ± 9.88	95.76 ± 8.07	95.67 ± 7.96	95.71 ± 0.02
	AUPR	57.29 ± 1.01	99.13 ± 0.77	98.05 ± 2.58	99.27 ± 2.67	99.48 ± 16.32	86.81 ± 10.86	86.63 ± 10.75	86.73 ± 0.03
	FPR95	39.60 ± 1.29	2.11 ± 2.49	4.12 ± 3.69	2.54 ± 8.90	0.55 ± 14.89	4.25 ± 8.83	4.34 ± 9.08	4.32 ± 0.05
Text (Synonym)	AUROC	15.47 ± 0.40	77.26 ± 10.27	73.59 ± 11.38	81.37 ± 16.06	80.31 ± 15.02	84.53 ± 16.50	88.63 ± 17.89	91.34 ± 1.09
	AUPR	34.36 ± 0.09	73.05 ± 11.94	67.73 ± 12.93	75.93 ± 16.57	74.79 ± 14.57	78.75 ± 14.79	82.81 ± 15.57	84.72 ± 1.08
	FPR95	96.71 ± 0.10	62.46 ± 13.46	64.27 ± 13.31	54.29 ± 14.32	56.17 ± 9.09	48.18 ± 13.71	38.24 ± 19.04	23.91 ± 5.58
Text (Antonym)	AUROC	15.24 ± 0.14	77.60 ± 10.10	74.18 ± 11.18	65.73 ± 10.81	80.49 ± 13.38	84.11 ± 14.54	87.90 ± 15.51	90.39 ± 1.21
	AUPR	34.30 ± 0.02	75.03 ± 11.46	69.73 ± 12.61	61.93 ± 11.29	76.48 ± 13.88	79.01 ± 13.80	82.64 ± 14.34	83.90 ± 1.19
	FPR95	96.89 ± 0.03	66.87 ± 14.05	68.32 ± 13.48	62.54 ± 17.05	61.29 ± 7.45	54.90 ± 11.42	44.40 ± 17.65	28.96 ± 5.90
Text Swap (Both)	AUROC	60.79 ± 0.72	90.52 ± 1.00	89.31 ± 1.38	90.01 ± 1.98	91.39 ± 9.69	92.91 ± 8.49	93.18 ± 8.52	92.92 ± 0.09
	AUPR	56.69 ± 0.66	91.96 ± 1.04	90.16 ± 1.69	91.83 ± 2.18	93.44 ± 14.60	85.83 ± 9.97	85.85 ± 9.85	85.79 ± 0.06
	FPR95	93.28 ± 0.27	53.80 ± 2.99	55.41 ± 2.62	57.73 ± 7.66	58.94 ± 5.39	22.74 ± 17.46	16.73 ± 18.37	19.56 ± 1.22
Text Swap (Intra)	AUROC	60.65 ± 0.39	59.85 ± 1.84	58.93 ± 1.62	58.49 ± 1.34	63.83 ± 3.22	66.04 ± 3.97	67.85 ± 4.35	69.52 ± 0.15
	AUPR	57.31 ± 0.36	61.14 ± 1.33	59.83 ± 1.18	60.55 ± 0.99	66.19 ± 4.93	64.77 ± 4.06	65.81 ± 4.10	66.67 ± 0.48
	FPR95	93.71 ± 0.17	88.63 ± 1.45	89.08 ± 1.22	90.89 ± 2.09	88.07 ± 0.63	84.86 ± 1.33	84.93 ± 2.69	78.92 ± 2.92
Text Swap (Inter)	AUROC	59.77 ± 0.76	89.98 ± 1.06	88.77 ± 1.31	88.26 ± 1.13	90.90 ± 9.81	92.72 ± 8.77	93.05 ± 8.78	92.59 ± 0.10
	AUPR	55.89 ± 0.69	91.44 ± 1.06	89.64 ± 1.50	89.90 ± 1.16	92.94 ± 14.47	85.66 ± 10.07	85.74 ± 9.93	85.51 ± 0.03
	FPR95	93.45 ± 0.14	55.20 ± 2.24	56.72 ± 3.03	58.50 ± 6.18	61.46 ± 5.18	24.18 ± 17.52	18.21 ± 17.79	21.53 ± 1.51
Semantic (0.75)	AUROC	76.48 ± 0.80	96.78 ± 0.48	97.24 ± 1.15	97.06 ± 4.53	98.40 ± 9.67	95.74 ± 7.02	95.55 ± 6.65	95.45 ± 0.14
	AUPR	63.82 ± 0.81	96.14 ± 0.38	96.53 ± 1.39	96.84 ± 4.25	98.25 ± 14.48	86.85 ± 9.69	86.50 ± 9.19	86.43 ± 0.23
	FPR95	44.63 ± 0.42	13.46 ± 3.76	10.44 ± 4.48	14.30 ± 9.87	8.03 ± 22.65	4.32 ± 10.10	4.57 ± 10.69	5.00 ± 0.39
Semantic (0.85)	AUROC	72.58 ± 0.50	89.83 ± 0.61	91.51 ± 0.91	86.13 ± 2.97	93.77 ± 9.12	95.28 ± 7.23	95.00 ± 7.02	93.73 ± 0.76
	AUPR	61.44 ± 0.55	88.14 ± 0.44	90.30 ± 0.79	85.92 ± 2.57	93.66 ± 12.12	86.45 ± 9.38	85.88 ± 8.98	84.23 ± 1.05
	FPR95	59.63 ± 0.85	37.08 ± 3.64	32.35 ± 2.90	52.81 ± 12.37	28.03 ± 18.54	5.85 ± 15.14	6.24 ± 16.01	9.98 ± 2.15
Semantic (0.95)	AUROC	61.28 ± 0.40	68.84 ± 0.68	72.19 ± 0.11	68.46 ± 3.43	72.67 ± 4.39	76.68 ± 3.13	79.03 ± 3.51	68.98 ± 2.67
	AUPR	54.13 ± 0.25	67.40 ± 1.50	71.44 ± 0.80	67.22 ± 2.37	72.00 ± 4.14	67.47 ± 2.43	69.99 ± 2.61	59.51 ± 1.71
	FPR95	84.52 ± 0.70	78.30 ± 1.66	76.19 ± 2.42	77.54 ± 6.74	76.29 ± 3.57	55.73 ± 8.23	52.99 ± 7.94	63.99 ± 4.30
	ID Acc	77.41 ± 0.21	77.41 ± 0.21	77.41 ± 0.21	77.41 ± 0.21	77.41 ± 0.21	77.41 ± 0.21	76.82 ± 0.40	77.42 ± 0.12

Table 18: Performance comparison of OOD detection methods on **DBLP dataset** across different OOD shifts and configurations.

OOD Type	Metric	w/o Prop.					w/ Prop.		
		MAHA	MSP	ODIN	NECO	Energy	GNNSafe	NodeSafe	TNT-OOD
Feature (0.50)	AUROC	79.57 ± 0.54	71.24 ± 0.51	68.81 ± 0.65	62.69 ± 5.54	72.43 ± 1.48	77.47 ± 2.08	77.96 ± 2.58	84.04 ± 1.80
	AUPR	77.08 ± 0.82	67.25 ± 0.46	65.34 ± 0.51	61.67 ± 4.63	70.16 ± 1.15	76.68 ± 1.77	76.52 ± 1.88	82.75 ± 0.53
	FPR95	81.08 ± 2.29	71.15 ± 1.39	77.18 ± 1.17	83.88 ± 3.24	77.82 ± 2.31	59.22 ± 2.47	58.39 ± 2.30	49.64 ± 6.66
Feature (0.70)	AUROC	81.92 ± 0.59	78.50 ± 0.64	76.31 ± 0.80	71.16 ± 1.22	78.63 ± 1.93	85.90 ± 2.32	86.38 ± 2.98	93.26 ± 1.17
	AUPR	78.81 ± 1.03	73.33 ± 0.69	71.91 ± 0.71	68.33 ± 0.21	74.27 ± 2.09	83.04 ± 2.66	82.97 ± 3.17	92.31 ± 0.98
	FPR95	72.57 ± 3.70	54.01 ± 1.20	61.24 ± 1.24	69.90 ± 3.53	61.41 ± 3.71	36.53 ± 3.54	35.28 ± 4.04	22.64 ± 4.33
Feature (0.90)	AUROC	78.78 ± 0.51	82.47 ± 0.57	82.40 ± 0.58	76.14 ± 4.49	81.27 ± 2.08	90.65 ± 2.22	90.92 ± 2.89	97.10 ± 0.59
	AUPR	74.58 ± 1.07	76.75 ± 0.63	77.44 ± 0.44	71.36 ± 4.36	74.90 ± 2.42	85.10 ± 3.42	84.95 ± 4.34	96.37 ± 0.76
	FPR95	71.50 ± 3.03	45.08 ± 1.50	46.75 ± 1.47	67.84 ± 8.56	49.23 ± 4.33	22.82 ± 3.36	21.82 ± 4.57	10.33 ± 1.96
Structure (Mild)	AUROC	59.89 ± 0.15	61.99 ± 0.40	64.39 ± 0.26	61.11 ± 2.31	62.91 ± 0.84	74.60 ± 1.42	75.03 ± 1.71	76.94 ± 1.50
	AUPR	55.91 ± 0.13	59.73 ± 0.27	61.39 ± 0.24	61.09 ± 2.12	61.86 ± 0.60	68.83 ± 1.29	69.23 ± 1.57	71.67 ± 0.81
	FPR95	91.22 ± 0.12	86.66 ± 0.30	85.53 ± 0.20	90.69 ± 0.31	88.53 ± 0.68	63.29 ± 1.75	64.04 ± 2.07	64.83 ± 3.60
Structure (Medium)	AUROC	57.44 ± 0.25	63.96 ± 0.45	66.26 ± 0.35	63.72 ± 1.96	64.41 ± 1.13	80.01 ± 1.54	80.50 ± 1.93	83.02 ± 1.52
	AUPR	53.14 ± 0.22	61.31 ± 0.41	62.79 ± 0.31	62.45 ± 1.10	63.25 ± 0.90	75.44 ± 1.60	75.89 ± 2.00	79.46 ± 0.98
	FPR95	91.12 ± 0.10	84.97 ± 0.31	83.80 ± 0.24	86.36 ± 1.76	86.90 ± 0.70	53.12 ± 1.73	53.23 ± 2.84	52.36 ± 5.43
Structure (Strong)	AUROC	45.24 ± 0.50	68.86 ± 0.58	69.37 ± 0.50	61.46 ± 6.19	67.47 ± 1.84	90.95 ± 1.57	91.41 ± 2.14	95.51 ± 0.92
	AUPR	43.91 ± 0.26	64.94 ± 0.56	64.60 ± 0.32	60.35 ± 3.30	64.68 ± 1.94	88.56 ± 2.03	88.90 ± 2.66	95.16 ± 0.99
	FPR95	91.81 ± 0.24	79.87 ± 0.77	80.19 ± 1.15	85.05 ± 5.28	81.65 ± 0.75	26.59 ± 2.72	24.89 ± 3.72	18.21 ± 3.83
Text (Synonym)	AUROC	22.35 ± 1.03	76.59 ± 0.53	74.36 ± 0.68	78.24 ± 3.56	82.78 ± 1.02	90.22 ± 1.47	89.58 ± 1.58	93.69 ± 1.23
	AUPR	35.51 ± 0.28	72.21 ± 0.50	70.31 ± 0.58	74.78 ± 3.52	78.44 ± 1.13	87.91 ± 1.56	87.24 ± 1.82	92.23 ± 2.55
	FPR95	98.50 ± 0.08	60.87 ± 1.11	67.57 ± 1.14	61.75 ± 6.22	53.17 ± 2.22	27.87 ± 3.00	31.31 ± 2.79	23.83 ± 2.62
Text (Antonym)	AUROC	23.47 ± 0.81	74.67 ± 0.57	72.46 ± 0.64	71.37 ± 6.49	81.08 ± 0.97	88.52 ± 1.49	87.85 ± 1.46	92.01 ± 1.39
	AUPR	35.81 ± 0.23	70.20 ± 0.53	68.45 ± 0.50	67.94 ± 6.03	76.35 ± 0.96	85.52 ± 1.54	84.87 ± 1.61	90.23 ± 2.87
	FPR95	98.53 ± 0.02	65.83 ± 1.01	71.98 ± 1.14	69.72 ± 6.98	57.43 ± 1.74	30.68 ± 2.93	34.85 ± 3.48	29.00 ± 1.51
Text Swap (Both)	AUROC	62.17 ± 0.19	78.60 ± 0.49	80.37 ± 0.45	76.16 ± 2.72	78.29 ± 2.05	89.69 ± 2.42	90.07 ± 3.13	96.04 ± 0.94
	AUPR	58.13 ± 0.31	73.32 ± 0.47	75.51 ± 0.37	71.32 ± 2.82	72.77 ± 2.18	83.46 ± 3.56	83.59 ± 4.57	95.29 ± 1.23
	FPR95	83.58 ± 0.99	55.92 ± 1.17	54.38 ± 1.08	62.86 ± 2.82	57.58 ± 2.75	26.31 ± 3.98	24.90 ± 4.89	15.37 ± 3.38
Text Swap (Intra)	AUROC	63.70 ± 0.07	55.93 ± 0.25	55.61 ± 0.37	55.99 ± 2.01	59.11 ± 0.87	64.39 ± 1.74	65.31 ± 2.54	71.46 ± 3.08
	AUPR	61.20 ± 0.06	53.19 ± 0.17	53.06 ± 0.17	54.60 ± 1.23	56.37 ± 0.70	58.57 ± 1.26	59.29 ± 1.92	65.81 ± 2.08
	FPR95	83.90 ± 0.62	87.30 ± 0.32	89.04 ± 0.16	90.64 ± 2.45	87.77 ± 0.75	70.61 ± 2.18	68.86 ± 2.59	62.48 ± 3.07
Text Swap (Inter)	AUROC	59.94 ± 0.19	75.86 ± 0.79	76.64 ± 0.80	71.44 ± 4.38	76.54 ± 1.99	87.20 ± 2.51	87.74 ± 3.32	93.59 ± 1.41
	AUPR	56.60 ± 0.26	70.81 ± 0.76	72.11 ± 0.69	67.77 ± 3.92	71.98 ± 2.05	82.28 ± 3.37	82.42 ± 4.22	92.89 ± 1.41
	FPR95	87.23 ± 0.89	61.04 ± 1.42	62.25 ± 1.34	70.67 ± 7.37	64.33 ± 3.07	34.31 ± 4.48	32.45 ± 5.94	25.55 ± 5.38
Semantic (0.75)	AUROC	60.68 ± 0.84	71.29 ± 0.50	73.57 ± 0.43	67.97 ± 1.80	67.80 ± 1.85	90.61 ± 1.65	91.27 ± 2.41	94.56 ± 0.86
	AUPR	53.45 ± 0.63	67.01 ± 0.48	68.93 ± 0.40	65.76 ± 2.39	65.62 ± 1.63	87.09 ± 2.07	87.62 ± 2.92	93.04 ± 0.88
	FPR95	84.59 ± 0.31	72.76 ± 0.57	70.95 ± 0.73	79.96 ± 1.86	76.88 ± 1.37	24.71 ± 2.78	22.10 ± 4.85	17.65 ± 3.29
Semantic (0.85)	AUROC	58.58 ± 0.80	66.01 ± 0.30	68.36 ± 0.18	57.22 ± 3.99	62.19 ± 1.42	87.26 ± 1.62	88.13 ± 2.52	90.23 ± 1.16
	AUPR	52.76 ± 0.59	62.59 ± 0.32	64.24 ± 0.19	59.17 ± 2.26	61.16 ± 1.01	83.66 ± 1.71	84.44 ± 2.66	87.36 ± 0.89
	FPR95	92.51 ± 0.16	81.22 ± 0.30	79.93 ± 0.17	89.43 ± 2.09	82.94 ± 0.75	32.35 ± 2.88	29.97 ± 4.07	28.06 ± 4.27
Semantic (0.95)	AUROC	53.14 ± 0.51	55.77 ± 0.16	57.43 ± 0.14	49.97 ± 1.32	50.82 ± 0.66	78.13 ± 1.27	79.01 ± 2.09	78.02 ± 1.54
	AUPR	50.50 ± 0.42	54.77 ± 0.11	55.51 ± 0.12	50.56 ± 1.10	51.17 ± 0.31	74.33 ± 0.93	75.21 ± 1.77	74.32 ± 0.84
	FPR95	95.28 ± 0.15	91.16 ± 0.15	90.32 ± 0.23	91.39 ± 2.01	90.04 ± 0.34	50.20 ± 1.93	47.51 ± 2.15	50.63 ± 4.71
	ID Acc	78.23 ± 0.41	78.23 ± 0.41	78.23 ± 0.41	78.23 ± 0.41	78.23 ± 0.41	78.23 ± 0.41	78.90 ± 0.14	78.85 ± 0.23

Table 19: Performance comparison of OOD detection methods on **PubMed dataset** across different OOD shifts and configurations.

OOD Type	Metric	w/o Prop.					w/ Prop.		
		MAHA	MSP	ODIN	NECO	Energy	GNNSafe	NodeSafe	TNT-OOD
Feature (0.50)	AUROC	52.44 ± 4.42	48.52 ± 10.26	53.00 ± 10.29	47.70 ± 10.73	51.00 ± 11.35	65.45 ± 25.84	49.36 ± 1.63	76.44 ± 2.30
	AUPR	54.83 ± 3.38	55.88 ± 6.55	58.93 ± 7.57	55.46 ± 6.60	55.46 ± 7.33	69.94 ± 20.29	53.84 ± 1.42	78.38 ± 2.56
	FPR95	86.72 ± 1.48	90.05 ± 5.98	88.67 ± 4.67	90.20 ± 6.17	85.25 ± 8.34	63.98 ± 42.59	95.10 ± 0.12	84.41 ± 0.87
Feature (0.70)	AUROC	51.60 ± 6.24	48.96 ± 11.44	55.36 ± 13.72	48.02 ± 11.93	51.24 ± 12.31	66.01 ± 26.33	50.23 ± 1.99	84.86 ± 2.42
	AUPR	53.99 ± 4.66	56.31 ± 7.29	61.00 ± 10.14	55.74 ± 7.30	55.49 ± 7.81	70.32 ± 20.39	54.26 ± 1.71	88.05 ± 3.18
	FPR95	85.71 ± 2.12	88.84 ± 7.74	85.36 ± 8.90	88.96 ± 7.94	84.03 ± 9.05	63.79 ± 42.50	94.89 ± 0.13	77.23 ± 1.52
Feature (0.90)	AUROC	53.02 ± 5.94	49.77 ± 10.96	53.78 ± 11.14	48.89 ± 11.46	52.04 ± 11.66	66.00 ± 25.53	52.12 ± 2.02	87.38 ± 1.74
	AUPR	55.13 ± 4.64	56.79 ± 7.01	59.58 ± 8.09	56.34 ± 7.09	56.09 ± 7.60	70.02 ± 19.99	54.95 ± 1.89	91.13 ± 2.69
	FPR95	85.56 ± 1.88	89.75 ± 6.30	87.75 ± 5.86	89.92 ± 6.46	84.96 ± 8.40	63.97 ± 42.42	94.24 ± 0.16	75.26 ± 1.90
Structure (Mild)	AUROC	63.86 ± 5.78	50.13 ± 7.18	46.11 ± 3.44	49.32 ± 8.17	52.10 ± 9.80	70.56 ± 21.00	58.79 ± 1.89	84.76 ± 1.14
	AUPR	64.27 ± 5.19	55.57 ± 4.75	52.11 ± 4.24	55.15 ± 5.31	56.06 ± 6.97	69.98 ± 18.39	57.08 ± 1.63	80.57 ± 1.22
	FPR95	77.72 ± 5.66	88.98 ± 4.56	89.28 ± 2.84	89.55 ± 5.09	84.98 ± 6.79	59.32 ± 38.47	85.09 ± 0.10	42.95 ± 2.54
Structure (Medium)	AUROC	64.20 ± 5.95	49.68 ± 6.88	45.70 ± 3.68	48.86 ± 7.89	51.96 ± 9.72	72.09 ± 19.95	60.44 ± 2.05	85.17 ± 0.95
	AUPR	64.68 ± 5.42	55.34 ± 4.44	51.78 ± 4.45	54.95 ± 5.04	55.88 ± 6.94	71.20 ± 17.82	58.29 ± 1.85	80.97 ± 0.98
	FPR95	77.72 ± 5.66	88.72 ± 4.56	89.00 ± 2.85	89.33 ± 5.09	84.40 ± 7.01	59.10 ± 38.50	84.76 ± 0.21	41.58 ± 2.84
Structure (Strong)	AUROC	57.87 ± 2.98	47.91 ± 2.37	43.19 ± 4.78	47.14 ± 3.16	49.74 ± 6.01	90.58 ± 2.93	86.80 ± 2.47	89.22 ± 2.00
	AUPR	61.10 ± 3.35	53.67 ± 1.85	50.47 ± 5.10	53.15 ± 2.61	53.99 ± 5.00	91.77 ± 4.40	82.05 ± 3.66	88.81 ± 1.86
	FPR95	86.91 ± 1.52	88.02 ± 2.82	87.98 ± 1.20	88.91 ± 2.47	86.15 ± 2.34	40.52 ± 3.75	29.89 ± 2.93	40.80 ± 7.68
Text (Synonym)	AUROC	27.13 ± 20.56	35.95 ± 6.90	23.20 ± 17.88	38.71 ± 9.34	47.63 ± 16.53	84.52 ± 8.13	87.80 ± 5.51	3.77 ± 9.45
	AUPR	44.49 ± 10.07	48.11 ± 6.10	43.68 ± 9.52	49.32 ± 6.38	55.81 ± 11.22	83.67 ± 8.84	89.54 ± 4.62	35.74 ± 10.76
	FPR95	92.91 ± 5.40	94.49 ± 2.33	94.81 ± 4.87	93.84 ± 1.38	88.65 ± 8.13	52.32 ± 32.48	57.16 ± 23.50	98.70 ± 28.31
Text (Antonym)	AUROC	22.97 ± 23.19	20.06 ± 17.07	18.17 ± 20.72	22.07 ± 15.94	28.51 ± 20.71	50.76 ± 32.16	97.79 ± 0.95	4.72 ± 8.13
	AUPR	43.65 ± 10.58	42.72 ± 9.15	42.70 ± 10.04	43.16 ± 8.87	45.54 ± 10.59	59.70 ± 23.52	98.12 ± 1.08	35.89 ± 10.85
	FPR95	93.45 ± 5.75	96.33 ± 3.06	95.61 ± 5.44	96.22 ± 2.98	91.26 ± 9.97	66.65 ± 38.43	8.74 ± 6.72	98.55 ± 9.01
Text Swap (Both)	AUROC	55.06 ± 2.58	51.67 ± 6.46	51.21 ± 4.81	51.14 ± 6.67	53.67 ± 6.07	66.17 ± 21.35	55.56 ± 1.57	85.66 ± 0.51
	AUPR	57.93 ± 2.99	57.43 ± 3.68	56.74 ± 2.73	57.14 ± 3.76	57.40 ± 4.33	69.40 ± 17.47	57.50 ± 1.70	90.32 ± 0.50
	FPR95	88.75 ± 2.48	92.10 ± 2.22	91.22 ± 2.90	92.34 ± 2.51	89.34 ± 5.12	67.78 ± 37.68	93.13 ± 0.08	83.29 ± 1.54
Text Swap (Intra)	AUROC	56.15 ± 1.16	50.61 ± 5.65	49.84 ± 2.95	50.11 ± 5.93	52.72 ± 5.89	64.70 ± 21.60	50.89 ± 1.24	84.22 ± 2.24
	AUPR	59.03 ± 1.31	56.70 ± 3.28	55.57 ± 1.45	56.46 ± 3.39	56.84 ± 4.27	68.18 ± 17.78	54.41 ± 1.15	89.16 ± 2.71
	FPR95	89.05 ± 2.75	92.62 ± 2.03	91.76 ± 2.94	92.78 ± 2.26	90.47 ± 4.44	70.17 ± 34.69	94.53 ± 0.21	85.18 ± 1.35
Text Swap (Inter)	AUROC	54.27 ± 1.59	51.13 ± 4.93	50.89 ± 3.97	50.72 ± 5.09	52.71 ± 4.65	64.25 ± 18.64	55.15 ± 1.07	79.30 ± 0.91
	AUPR	57.84 ± 1.72	56.63 ± 2.91	56.44 ± 2.38	56.41 ± 2.94	56.91 ± 3.19	67.45 ± 15.00	58.12 ± 1.20	84.21 ± 1.02
	FPR95	90.55 ± 2.20	92.47 ± 1.82	91.59 ± 2.90	92.62 ± 2.02	91.52 ± 2.84	80.99 ± 19.19	93.38 ± 0.03	87.44 ± 0.92
Semantic (0.75)	AUROC	17.65 ± 6.87	59.94 ± 21.07	65.89 ± 16.75	59.36 ± 21.46	51.84 ± 12.36	5.17 ± 4.01	2.14 ± 0.03	99.72 ± 11.52
	AUPR	38.80 ± 1.75	61.95 ± 16.39	65.08 ± 14.28	62.13 ± 16.54	57.68 ± 13.50	37.24 ± 0.89	37.18 ± 0.02	99.81 ± 11.60
	FPR95	92.42 ± 0.64	83.39 ± 10.56	75.17 ± 16.34	84.71 ± 10.29	84.58 ± 9.34	96.77 ± 3.46	100.00 ± 0.00	0.62 ± 1.51
Semantic (0.85)	AUROC	17.97 ± 7.07	60.65 ± 19.42	65.48 ± 15.63	60.22 ± 19.72	53.13 ± 10.97	7.72 ± 6.17	3.58 ± 0.02	99.52 ± 9.46
	AUPR	38.86 ± 1.81	62.73 ± 15.56	65.25 ± 13.33	63.16 ± 15.63	58.57 ± 12.54	38.44 ± 1.44	38.53 ± 0.01	99.67 ± 9.46
	FPR95	93.56 ± 0.10	85.07 ± 10.15	76.68 ± 17.50	85.87 ± 9.41	86.03 ± 8.53	95.28 ± 5.32	99.99 ± 0.02	1.07 ± 1.55
Semantic (0.95)	AUROC	18.16 ± 7.09	62.40 ± 15.78	60.41 ± 15.23	62.43 ± 15.98	56.23 ± 7.88	19.76 ± 19.03	7.80 ± 0.15	95.87 ± 7.18
	AUPR	38.88 ± 1.90	64.98 ± 14.45	63.15 ± 12.52	65.90 ± 14.46	63.58 ± 8.80	43.58 ± 5.76	41.50 ± 0.02	97.11 ± 7.03
	FPR95	97.27 ± 1.07	96.35 ± 1.10	95.84 ± 1.32	95.98 ± 1.63	96.16 ± 1.39	89.53 ± 12.70	99.72 ± 0.10	25.05 ± 1.68
	ID Acc	61.85 ± 2.32	61.85 ± 2.32	61.85 ± 2.32	61.85 ± 2.32	61.85 ± 2.32	61.85 ± 2.32	61.69 ± 0.67	61.78 ± 0.16

Table 20: Performance comparison of OOD detection methods on **Reddit dataset** across different OOD shifts and configurations.

OOD Type	Metric	w/o Prop.					w/ Prop.		
		MAHA	MSP	ODIN	NECO	Energy	GNNSafe	NodeSafe	TNT-OOD
Feature (0.50)	AUROC	64.60 ± 0.74	65.46 ± 0.62	62.03 ± 0.71	62.19 ± 2.48	67.71 ± 0.60	82.58 ± 0.98	84.92 ± 1.33	91.69 ± 0.15
	AUPR	76.28 ± 0.86	76.84 ± 0.27	74.01 ± 0.39	73.96 ± 2.45	79.24 ± 0.49	87.07 ± 0.57	89.24 ± 0.65	92.05 ± 0.05
	FPR95	89.68 ± 0.36	80.34 ± 0.90	83.11 ± 0.99	84.82 ± 0.95	80.65 ± 0.61	62.59 ± 3.49	71.80 ± 1.02	29.10 ± 2.99
Feature (0.70)	AUROC	68.79 ± 1.73	76.27 ± 0.43	69.96 ± 0.70	72.49 ± 3.01	77.99 ± 0.77	92.55 ± 0.46	91.78 ± 1.46	96.05 ± 0.15
	AUPR	78.97 ± 1.90	84.00 ± 0.17	79.00 ± 0.15	81.36 ± 3.11	86.40 ± 0.66	92.65 ± 0.18	92.25 ± 0.76	94.13 ± 0.05
	FPR95	86.63 ± 0.47	62.89 ± 2.56	73.23 ± 1.96	79.68 ± 3.64	67.42 ± 1.59	22.58 ± 1.83	34.49 ± 11.90	6.01 ± 0.11
Feature (0.90)	AUROC	73.26 ± 3.10	81.95 ± 2.84	69.56 ± 3.11	79.39 ± 6.64	85.75 ± 1.07	96.10 ± 0.34	95.42 ± 1.30	96.97 ± 0.01
	AUPR	81.70 ± 3.06	87.22 ± 2.32	76.39 ± 2.94	85.90 ± 5.86	91.13 ± 0.62	94.16 ± 0.13	93.61 ± 0.69	94.38 ± 0.04
	FPR95	80.27 ± 3.48	49.11 ± 6.25	66.30 ± 2.44	60.48 ± 14.32	49.21 ± 4.76	6.62 ± 1.88	11.03 ± 8.42	3.30 ± 0.08
Structure (Mild)	AUROC	69.09 ± 1.29	74.48 ± 1.15	68.52 ± 2.24	72.10 ± 3.74	77.10 ± 0.37	93.25 ± 0.52	96.95 ± 0.06	93.86 ± 0.43
	AUPR	75.76 ± 1.54	82.56 ± 1.72	73.37 ± 2.42	82.12 ± 4.03	86.63 ± 0.70	91.95 ± 0.26	93.68 ± 0.05	92.08 ± 0.26
	FPR95	76.77 ± 0.98	79.43 ± 1.21	72.14 ± 1.05	81.78 ± 1.90	82.10 ± 0.69	29.82 ± 4.22	3.27 ± 0.18	24.68 ± 1.28
Structure (Medium)	AUROC	68.23 ± 1.34	74.31 ± 1.07	69.02 ± 2.34	74.68 ± 2.49	76.78 ± 0.37	93.02 ± 0.50	96.93 ± 0.06	93.68 ± 0.42
	AUPR	74.92 ± 1.51	82.57 ± 1.60	74.16 ± 2.67	84.04 ± 2.47	86.38 ± 0.67	91.84 ± 0.25	93.67 ± 0.06	91.98 ± 0.26
	FPR95	77.03 ± 0.86	79.61 ± 0.98	72.62 ± 0.87	80.35 ± 2.55	82.07 ± 0.67	30.60 ± 3.77	3.27 ± 0.18	25.30 ± 1.22
Structure (Strong)	AUROC	69.53 ± 5.79	74.93 ± 6.32	51.67 ± 4.10	72.40 ± 9.28	84.86 ± 2.79	90.58 ± 0.04	97.02 ± 0.01	97.06 ± 0.01
	AUPR	70.82 ± 3.70	73.98 ± 5.10	60.61 ± 1.77	76.05 ± 7.35	81.66 ± 2.58	91.77 ± 0.07	93.72 ± 0.02	93.78 ± 0.02
	FPR95	43.54 ± 6.37	32.54 ± 6.49	55.21 ± 4.25	45.68 ± 9.12	21.41 ± 3.28	40.52 ± 0.04	2.98 ± 0.01	2.94 ± 0.00
Text (Synonym)	AUROC	12.22 ± 0.70	63.40 ± 3.65	59.75 ± 4.33	66.70 ± 3.12	64.09 ± 5.37	73.28 ± 12.31	32.43 ± 10.52	80.81 ± 2.58
	AUPR	47.46 ± 0.17	73.81 ± 3.51	70.60 ± 3.80	76.05 ± 3.38	74.39 ± 4.47	80.28 ± 7.93	57.91 ± 5.55	84.91 ± 0.96
	FPR95	98.91 ± 0.02	83.77 ± 3.56	84.71 ± 2.48	78.50 ± 4.49	80.86 ± 3.67	64.27 ± 15.03	95.42 ± 1.13	56.47 ± 17.31
Text (Antonym)	AUROC	11.41 ± 1.15	60.54 ± 4.63	56.72 ± 4.28	54.73 ± 13.69	58.23 ± 4.88	61.57 ± 12.90	30.23 ± 11.92	69.78 ± 6.34
	AUPR	47.27 ± 0.30	72.25 ± 3.81	69.10 ± 3.45	69.48 ± 8.66	70.25 ± 3.78	73.85 ± 8.24	56.43 ± 6.13	78.18 ± 3.39
	FPR95	99.09 ± 0.23	87.34 ± 1.73	88.36 ± 2.35	90.64 ± 7.11	86.28 ± 2.66	83.68 ± 5.06	95.75 ± 1.04	78.13 ± 11.78
Text Swap (Both)	AUROC	70.53 ± 2.99	82.65 ± 2.56	68.93 ± 1.88	79.33 ± 5.07	86.66 ± 0.74	96.60 ± 0.17	96.20 ± 0.50	96.95 ± 0.05
	AUPR	79.47 ± 2.73	87.41 ± 2.21	75.52 ± 1.89	86.25 ± 4.73	91.59 ± 0.47	94.05 ± 0.06	93.72 ± 0.29	94.03 ± 0.05
	FPR95	85.45 ± 1.27	46.51 ± 4.37	65.59 ± 2.13	62.43 ± 6.94	45.77 ± 2.21	4.49 ± 0.59	5.85 ± 2.43	3.18 ± 0.14
Text Swap (Intra)	AUROC	65.91 ± 0.42	52.31 ± 0.12	51.81 ± 0.31	53.16 ± 1.19	54.26 ± 0.48	62.32 ± 0.89	63.42 ± 1.55	81.99 ± 0.92
	AUPR	77.81 ± 0.58	68.10 ± 0.13	67.74 ± 0.27	69.07 ± 0.71	70.34 ± 0.36	73.67 ± 0.54	76.37 ± 1.31	86.04 ± 0.41
	FPR95	91.47 ± 0.20	91.40 ± 0.96	97.29 ± 3.84	93.65 ± 0.50	90.70 ± 0.34	87.42 ± 0.46	93.51 ± 0.19	63.55 ± 2.17
Text Swap (Inter)	AUROC	69.89 ± 2.94	82.66 ± 2.04	69.37 ± 2.02	79.60 ± 1.44	86.14 ± 0.67	96.72 ± 0.06	96.23 ± 0.51	96.95 ± 0.06
	AUPR	79.09 ± 2.68	87.69 ± 1.67	76.24 ± 2.10	84.96 ± 0.88	91.39 ± 0.48	94.12 ± 0.06	93.76 ± 0.29	94.07 ± 0.08
	FPR95	85.86 ± 1.49	49.43 ± 5.24	65.93 ± 2.54	61.75 ± 9.09	47.98 ± 3.45	4.03 ± 0.34	5.67 ± 2.49	3.19 ± 0.14
Semantic (0.75)	AUROC	69.82 ± 4.03	75.40 ± 4.45	54.75 ± 2.62	81.06 ± 0.74	84.30 ± 2.01	97.12 ± 0.03	97.03 ± 0.01	97.05 ± 0.00
	AUPR	72.19 ± 2.70	77.49 ± 3.83	62.80 ± 1.49	84.34 ± 1.18	84.32 ± 1.88	93.93 ± 0.06	93.73 ± 0.02	93.78 ± 0.00
	FPR95	47.64 ± 4.67	37.23 ± 5.43	58.58 ± 2.58	40.84 ± 5.24	26.21 ± 2.71	2.90 ± 0.04	2.98 ± 0.01	2.96 ± 0.00
Semantic (0.85)	AUROC	65.89 ± 2.28	74.71 ± 2.58	60.23 ± 3.14	73.51 ± 6.21	81.64 ± 1.01	97.09 ± 0.02	97.01 ± 0.02	97.04 ± 0.01
	AUPR	70.10 ± 1.39	78.59 ± 1.99	67.70 ± 2.19	80.89 ± 5.48	83.84 ± 1.17	93.96 ± 0.05	93.76 ± 0.02	93.91 ± 0.02
	FPR95	56.20 ± 2.95	44.55 ± 3.49	60.60 ± 2.51	62.23 ± 3.54	35.38 ± 0.64	2.94 ± 0.00	2.98 ± 0.01	2.99 ± 0.04
Semantic (0.95)	AUROC	52.13 ± 0.57	66.92 ± 0.37	63.27 ± 0.55	63.16 ± 0.87	66.86 ± 0.41	95.60 ± 0.30	96.35 ± 0.21	96.40 ± 0.17
	AUPR	63.57 ± 0.23	75.52 ± 0.46	73.13 ± 0.33	73.79 ± 0.79	75.26 ± 0.27	93.57 ± 0.13	93.72 ± 0.08	93.95 ± 0.10
	FPR95	81.60 ± 0.35	79.47 ± 1.17	80.52 ± 1.95	73.63 ± 1.88	72.78 ± 0.76	7.48 ± 1.11	3.38 ± 0.07	6.41 ± 1.48
	ID Acc	79.26 ± 0.28	79.26 ± 0.28	79.26 ± 0.28	79.26 ± 0.28	79.26 ± 0.28	79.26 ± 0.28	78.09 ± 0.23	80.22 ± 0.28
Label (Random 1 3 4 5 8)	AUROC	27.94 ± 1.34	64.14 ± 2.30	59.48 ± 2.49	63.91 ± 3.70	64.33 ± 0.89	74.13 ± 1.57	74.75 ± 3.53	84.31 ± 2.05
	AUPR	58.82 ± 0.49	80.85 ± 1.34	78.07 ± 1.38	80.52 ± 1.77	80.66 ± 0.43	87.20 ± 0.54	85.90 ± 1.76	91.38 ± 0.81
	FPR95	97.75 ± 0.43	100.00 ± 0.00	100.00 ± 0.00	88.33 ± 4.82	93.08 ± 2.07	91.73 ± 5.62	79.32 ± 5.22	60.59 ± 7.78
	ID Acc	87.41 ± 0.10	87.41 ± 0.10	87.41 ± 0.10	87.41 ± 0.10	87.41 ± 0.10	87.41 ± 0.10	86.03 ± 0.32	88.42 ± 0.18
Label (Disimilar 0 1 8 9)	AUROC	41.24 ± 2.47	73.01 ± 1.53	70.10 ± 1.38	68.38 ± 3.74	75.17 ± 1.41	87.00 ± 1.85	85.18 ± 0.94	86.42 ± 0.93
	AUPR	34.35 ± 1.54	59.49 ± 2.56	55.81 ± 2.45	54.57 ± 6.24	61.38 ± 2.74	75.83 ± 2.67	72.31 ± 1.72	74.48 ± 1.64
	FPR95	96.94 ± 0.15	73.53 ± 2.59	77.06 ± 1.04	80.23 ± 4.09	69.81 ± 1.34	43.46 ± 2.86	44.73 ± 0.78	42.66 ± 0.77
	ID Acc	82.88 ± 0.10	82.88 ± 0.10	82.88 ± 0.10	82.88 ± 0.10	82.88 ± 0.10	82.88 ± 0.10	80.72 ± 0.29	88.42 ± 0.18
Label (Similar 2 3 4 6)	AUROC	30.98 ± 5.53	78.32 ± 3.29	75.68 ± 4.19	66.15 ± 3.29	71.23 ± 2.33	82.22 ± 3.71	86.45 ± 2.33	85.52 ± 1.20
	AUPR	64.38 ± 2.24	89.89 ± 1.62	88.55 ± 1.92	84.40 ± 1.38	86.20 ± 1.11	87.82 ± 1.60	89.65 ± 1.26	89.76 ± 0.53
	FPR95	93.79 ± 2.82	64.01 ± 7.29	67.48 ± 7.62	83.38 ± 3.19	73.66 ± 3.09	38.08 ± 9.20	27.54 ± 3.39	33.63 ± 7.82
	ID Acc	85.22 ± 0.41	85.22 ± 0.41	85.22 ± 0.41	85.22 ± 0.41	85.22 ± 0.41	85.22 ± 0.41	84.51 ± 0.66	86.05 ± 0.98

Table 21: Performance comparison of OOD detection methods on **WikiCS dataset** across different OOD shifts and configurations.

OOD Type	Metric	w/o Prop.					w/ Prop.		
		MAHA	MSP	ODIN	NECO	Energy	GNNSafe	NodeSafe	TNT-OOD
Feature (0.50)	AUROC	84.88 ± 0.68	60.38 ± 0.71	58.73 ± 0.68	62.06 ± 0.92	62.66 ± 0.96	61.82 ± 0.76	60.37 ± 0.70	76.98 ± 1.44
	AUPR	95.89 ± 0.19	87.46 ± 0.26	86.58 ± 0.27	88.24 ± 0.33	88.61 ± 0.35	87.12 ± 0.23	86.45 ± 0.25	91.71 ± 0.39
	FPR95	48.00 ± 1.66	88.73 ± 1.06	89.88 ± 1.02	89.75 ± 1.05	89.79 ± 1.14	70.92 ± 1.21	82.61 ± 0.97	46.20 ± 3.31
Feature (0.70)	AUROC	87.53 ± 0.69	60.04 ± 1.03	57.95 ± 0.98	63.08 ± 1.07	64.44 ± 1.08	64.08 ± 0.84	61.57 ± 0.92	80.46 ± 1.41
	AUPR	96.81 ± 0.20	86.35 ± 0.50	85.27 ± 0.50	88.08 ± 0.46	89.11 ± 0.44	87.38 ± 0.26	86.47 ± 0.31	92.60 ± 0.41
	FPR95	46.57 ± 1.65	86.51 ± 0.75	88.49 ± 1.06	88.33 ± 0.98	88.31 ± 1.02	65.84 ± 1.40	77.09 ± 1.48	36.06 ± 3.93
Feature (0.90)	AUROC	83.60 ± 0.64	58.65 ± 0.92	56.75 ± 0.87	62.38 ± 0.92	63.94 ± 0.92	64.23 ± 0.76	61.43 ± 0.90	81.12 ± 1.31
	AUPR	95.83 ± 0.20	85.64 ± 0.56	84.67 ± 0.53	87.71 ± 0.48	88.93 ± 0.43	86.99 ± 0.25	86.12 ± 0.27	92.60 ± 0.50
	FPR95	60.06 ± 0.85	87.05 ± 0.63	88.96 ± 0.64	88.59 ± 0.63	88.52 ± 0.63	62.77 ± 0.92	73.88 ± 1.86	32.31 ± 3.71
Structure (Mild)	AUROC	95.93 ± 0.32	68.85 ± 1.17	68.49 ± 1.06	75.23 ± 1.03	77.76 ± 1.03	67.37 ± 0.69	62.96 ± 0.63	78.89 ± 0.67
	AUPR	98.83 ± 0.07	89.19 ± 0.52	89.11 ± 0.47	91.73 ± 0.45	93.05 ± 0.42	83.30 ± 0.24	81.39 ± 0.22	87.12 ± 0.21
	FPR95	14.45 ± 1.73	66.85 ± 1.13	67.87 ± 1.12	63.74 ± 1.82	63.18 ± 1.96	54.15 ± 1.37	55.89 ± 0.52	29.77 ± 2.01
Structure (Medium)	AUROC	95.61 ± 0.30	68.51 ± 1.17	68.04 ± 1.08	74.92 ± 1.03	77.45 ± 1.04	67.86 ± 0.67	63.35 ± 0.62	79.17 ± 0.67
	AUPR	98.69 ± 0.06	89.04 ± 0.53	88.92 ± 0.50	91.58 ± 0.45	92.88 ± 0.43	83.41 ± 0.24	81.47 ± 0.21	87.20 ± 0.21
	FPR95	14.82 ± 1.66	66.09 ± 1.15	67.21 ± 1.09	62.57 ± 1.67	62.11 ± 1.87	52.09 ± 1.59	54.11 ± 0.70	28.28 ± 2.18
Structure (Strong)	AUROC	96.21 ± 0.19	64.05 ± 1.75	61.88 ± 1.64	72.59 ± 1.44	76.14 ± 1.44	73.60 ± 0.55	66.92 ± 0.94	82.00 ± 0.37
	AUPR	98.16 ± 0.07	83.75 ± 0.92	82.62 ± 0.79	88.04 ± 0.78	90.42 ± 0.73	84.72 ± 0.23	82.06 ± 0.33	88.00 ± 0.16
	FPR95	8.32 ± 0.62	55.83 ± 1.99	58.09 ± 1.87	49.65 ± 2.11	47.96 ± 2.28	30.72 ± 0.70	37.93 ± 1.19	18.57 ± 0.52
Text (Synonym)	AUROC	25.38 ± 0.33	86.41 ± 0.83	85.24 ± 0.81	86.56 ± 0.51	86.10 ± 0.42	80.86 ± 0.46	69.67 ± 2.72	83.54 ± 0.25
	AUPR	71.58 ± 0.14	96.40 ± 0.25	95.99 ± 0.28	96.54 ± 0.15	96.31 ± 0.12	92.56 ± 0.11	88.73 ± 1.02	93.53 ± 0.09
	FPR95	99.49 ± 0.58	47.40 ± 2.06	49.61 ± 1.80	54.04 ± 1.83	55.73 ± 1.85	34.93 ± 2.24	59.05 ± 4.37	26.06 ± 1.79
Text (Antonym)	AUROC	22.63 ± 0.92	90.48 ± 0.56	89.31 ± 0.55	90.46 ± 0.58	89.89 ± 0.56	82.53 ± 0.43	73.56 ± 2.41	84.45 ± 0.14
	AUPR	70.56 ± 0.54	97.59 ± 0.16	97.23 ± 0.16	97.66 ± 0.18	97.36 ± 0.17	93.16 ± 0.13	90.00 ± 0.88	93.81 ± 0.06
	FPR95	99.63 ± 0.66	37.18 ± 1.87	39.57 ± 1.93	42.42 ± 2.43	44.54 ± 2.41	28.25 ± 2.15	47.93 ± 4.88	21.00 ± 0.70
Text Swap (Both)	AUROC	66.58 ± 0.56	57.61 ± 0.64	56.32 ± 0.61	60.29 ± 0.68	61.34 ± 0.68	63.00 ± 0.62	61.47 ± 0.36	77.86 ± 1.15
	AUPR	90.56 ± 0.20	85.90 ± 0.35	85.28 ± 0.32	87.40 ± 0.36	88.28 ± 0.35	86.08 ± 0.18	85.64 ± 0.12	91.30 ± 0.33
	FPR95	90.09 ± 0.10	89.78 ± 0.67	90.49 ± 0.63	89.92 ± 0.35	89.86 ± 0.35	70.55 ± 0.82	73.58 ± 0.41	42.24 ± 1.55
Text Swap (Intra)	AUROC	66.91 ± 0.32	52.04 ± 0.19	52.04 ± 0.23	53.39 ± 0.19	53.66 ± 0.19	54.42 ± 0.02	54.11 ± 0.15	68.76 ± 1.42
	AUPR	90.74 ± 0.12	84.75 ± 0.06	84.75 ± 0.08	85.22 ± 0.09	85.28 ± 0.11	84.45 ± 0.03	84.33 ± 0.03	88.87 ± 0.39
	FPR95	90.45 ± 0.02	93.60 ± 0.20	93.65 ± 0.28	93.32 ± 0.19	93.33 ± 0.18	88.75 ± 0.36	95.97 ± 0.09	64.46 ± 2.98
Text Swap (Inter)	AUROC	63.25 ± 0.57	58.14 ± 0.47	56.81 ± 0.40	60.52 ± 0.58	61.48 ± 0.59	63.04 ± 0.53	62.66 ± 0.42	75.06 ± 0.98
	AUPR	89.26 ± 0.23	86.11 ± 0.28	85.47 ± 0.25	87.50 ± 0.31	88.34 ± 0.32	86.17 ± 0.17	85.99 ± 0.13	90.64 ± 0.28
	FPR95	90.55 ± 0.20	89.47 ± 0.47	90.25 ± 0.39	89.82 ± 0.33	89.76 ± 0.33	71.72 ± 0.33	71.01 ± 0.25	52.36 ± 1.15
Semantic (0.75)	AUROC	95.17 ± 0.26	79.67 ± 0.68	77.74 ± 0.66	87.76 ± 0.24	91.71 ± 0.12	3.54 ± 0.02	3.15 ± 0.01	32.39 ± 1.88
	AUPR	99.12 ± 0.09	91.90 ± 0.34	91.04 ± 0.34	95.40 ± 0.15	97.29 ± 0.05	65.70 ± 0.01	65.32 ± 0.02	74.10 ± 0.41
	FPR95	25.38 ± 1.04	36.67 ± 0.63	40.04 ± 0.26	27.18 ± 0.21	23.38 ± 0.24	99.99 ± 0.00	100.00 ± 0.00	85.90 ± 1.48
Semantic (0.85)	AUROC	92.01 ± 0.24	79.52 ± 0.52	77.96 ± 0.52	86.59 ± 0.18	89.93 ± 0.13	7.52 ± 0.16	5.73 ± 0.01	42.43 ± 1.09
	AUPR	98.53 ± 0.09	92.98 ± 0.25	92.32 ± 0.25	95.63 ± 0.11	97.03 ± 0.05	67.78 ± 0.04	66.81 ± 0.04	78.78 ± 0.29
	FPR95	85.89 ± 0.39	47.93 ± 0.25	50.45 ± 0.19	41.58 ± 1.07	39.59 ± 1.12	99.98 ± 0.01	100.00 ± 0.00	85.88 ± 0.78
Semantic (0.95)	AUROC	85.17 ± 0.30	74.21 ± 0.34	74.09 ± 0.32	80.31 ± 0.16	82.83 ± 0.18	21.73 ± 0.64	11.51 ± 0.01	49.07 ± 0.50
	AUPR	97.20 ± 0.10	92.29 ± 0.15	92.36 ± 0.14	94.31 ± 0.08	95.22 ± 0.06	74.27 ± 0.15	70.54 ± 0.05	83.76 ± 0.22
	FPR95	94.48 ± 0.35	66.32 ± 0.46	66.97 ± 0.54	61.98 ± 0.59	61.55 ± 0.55	99.92 ± 0.01	100.00 ± 0.00	90.96 ± 0.52
	ID Acc	83.42 ± 0.21	83.42 ± 0.21	83.42 ± 0.21	83.42 ± 0.21	83.42 ± 0.21	83.42 ± 0.21	82.90 ± 0.06	83.86 ± 0.12
Label (Random 0 1 3 4 7)	AUROC	42.14 ± 0.90	55.80 ± 1.13	53.31 ± 0.51	57.32 ± 7.18	59.71 ± 0.93	38.15 ± 0.48	37.99 ± 2.02	64.84 ± 0.11
	AUPR	98.65 ± 0.03	99.21 ± 0.05	99.16 ± 0.03	99.12 ± 0.27	99.34 ± 0.03	98.61 ± 0.02	98.57 ± 0.04	99.27 ± 0.08
	FPR95	99.30 ± 0.12	94.94 ± 0.44	96.16 ± 0.33	86.30 ± 9.13	93.54 ± 0.54	93.72 ± 0.77	99.21 ± 0.43	42.15 ± 0.12
	ID Acc	80.30 ± 0.20	80.30 ± 0.20	80.30 ± 0.20	80.30 ± 0.20	80.30 ± 0.20	80.30 ± 0.20	82.29 ± 0.86	84.33 ± 0.62
Label (Disimilar 1 2 6 10 11)	AUROC	35.19 ± 1.16	68.91 ± 2.40	65.57 ± 2.74	69.16 ± 2.26	68.56 ± 2.13	62.58 ± 1.92	59.77 ± 1.55	80.69 ± 0.11
	AUPR	81.43 ± 0.35	93.16 ± 0.64	91.83 ± 0.94	92.96 ± 0.56	92.66 ± 0.55	89.39 ± 0.49	88.91 ± 0.32	94.39 ± 0.03
	FPR95	98.65 ± 0.25	77.15 ± 2.24	79.15 ± 2.07	79.06 ± 1.51	79.18 ± 1.60	70.66 ± 1.96	80.25 ± 3.20	22.60 ± 0.18
	ID Acc	79.40 ± 0.38	79.40 ± 0.38	79.40 ± 0.38	79.40 ± 0.38	79.40 ± 0.38	79.40 ± 0.38	78.86 ± 0.31	81.05 ± 0.16
Label (Similar 0 3 5 8 9)	AUROC	37.43 ± 1.98	82.01 ± 0.65	81.37 ± 0.75	77.44 ± 1.38	76.87 ± 1.52	73.88 ± 0.75	71.74 ± 0.25	85.63 ± 0.02
	AUPR	60.46 ± 1.19	89.22 ± 0.39	89.07 ± 0.48	87.09 ± 0.76	86.70 ± 0.92	80.99 ± 0.46	79.86 ± 0.12	89.13 ± 0.03
	FPR95	98.35 ± 0.15	60.58 ± 1.69	62.33 ± 1.54	72.33 ± 3.40	72.54 ± 3.54	51.87 ± 2.49	61.97 ± 0.87	17.98 ± 0.03
	ID Acc	94.02 ± 0.12	94.02 ± 0.12	94.02 ± 0.12	94.02 ± 0.12	94.02 ± 0.12	94.02 ± 0.12	92.88 ± 0.11	94.21 ± 0.24

Table 22: Performance comparison of OOD detection methods on **BookHis dataset** across different OOD shifts and configurations.

OOD Type	Metric	w/o Prop.					w/ Prop.		
		MAHA	MSP	ODIN	NECO	Energy	GNNSafe	NodeSafe	TNT-OOD
Feature (0.50)	AUROC	78.01 ± 0.44	64.64 ± 0.89	61.26 ± 1.08	70.09 ± 0.97	70.76 ± 0.96	74.84 ± 0.89	61.87 ± 3.68	81.16 ± 2.66
	AUPR	93.24 ± 0.14	88.58 ± 0.43	86.91 ± 0.46	90.64 ± 0.40	91.12 ± 0.37	91.08 ± 0.28	86.73 ± 1.16	92.77 ± 0.83
	FPR95	56.93 ± 0.95	81.84 ± 0.87	84.26 ± 1.35	78.09 ± 1.43	78.07 ± 1.54	61.11 ± 1.46	78.84 ± 2.97	35.42 ± 5.42
Feature (0.70)	AUROC	82.07 ± 0.38	69.01 ± 1.33	64.08 ± 1.52	74.85 ± 1.13	75.58 ± 1.07	80.22 ± 0.79	65.52 ± 5.14	83.88 ± 3.14
	AUPR	94.76 ± 0.13	89.78 ± 0.69	87.37 ± 0.67	92.29 ± 0.50	92.92 ± 0.40	92.27 ± 0.24	87.53 ± 1.54	93.52 ± 0.93
	FPR95	52.06 ± 0.38	74.97 ± 0.80	79.11 ± 2.07	70.11 ± 2.04	69.93 ± 2.26	39.11 ± 2.21	70.38 ± 4.87	23.91 ± 8.06
Feature (0.90)	AUROC	80.52 ± 0.15	69.05 ± 1.60	63.65 ± 1.70	74.25 ± 1.07	74.94 ± 0.91	80.41 ± 0.66	67.93 ± 5.05	84.15 ± 3.56
	AUPR	94.65 ± 0.07	89.45 ± 0.88	86.77 ± 0.77	92.20 ± 0.54	93.00 ± 0.36	91.92 ± 0.20	87.88 ± 1.48	93.49 ± 0.50
	FPR95	62.45 ± 0.35	75.10 ± 0.41	79.80 ± 2.00	72.62 ± 1.65	72.93 ± 1.67	37.20 ± 2.08	63.30 ± 5.99	22.20 ± 9.78
Structure (Mild)	AUROC	97.93 ± 0.14	78.17 ± 1.96	76.51 ± 2.10	90.88 ± 1.12	94.14 ± 0.81	82.57 ± 0.39	62.76 ± 6.35	82.21 ± 0.10
	AUPR	99.54 ± 0.03	90.18 ± 0.99	88.77 ± 0.92	96.20 ± 0.57	98.39 ± 0.29	88.48 ± 0.13	81.42 ± 2.14	88.23 ± 0.06
	FPR95	11.48 ± 0.88	48.07 ± 2.62	48.72 ± 4.48	27.53 ± 2.22	23.32 ± 2.10	20.41 ± 1.27	57.14 ± 7.63	19.90 ± 0.24
Structure (Medium)	AUROC	97.89 ± 0.14	78.14 ± 1.93	76.59 ± 2.12	90.82 ± 1.13	94.06 ± 0.81	82.52 ± 0.41	63.12 ± 6.09	82.16 ± 0.10
	AUPR	99.52 ± 0.03	90.37 ± 0.97	88.96 ± 0.93	96.26 ± 0.57	98.36 ± 0.29	88.47 ± 0.14	81.56 ± 2.05	88.21 ± 0.06
	FPR95	11.38 ± 0.88	48.33 ± 2.66	48.90 ± 4.51	27.64 ± 2.32	23.36 ± 2.14	20.60 ± 1.33	56.79 ± 7.46	20.03 ± 0.24
Structure (Strong)	AUROC	99.61 ± 0.07	78.22 ± 5.12	70.07 ± 3.11	91.35 ± 2.23	94.88 ± 1.34	83.91 ± 0.08	60.36 ± 9.41	83.39 ± 0.04
	AUPR	99.82 ± 0.04	87.29 ± 2.39	83.83 ± 1.29	94.41 ± 1.32	97.36 ± 0.76	88.85 ± 0.04	79.60 ± 3.18	88.59 ± 0.04
	FPR95	0.92 ± 0.16	27.90 ± 5.67	37.98 ± 3.43	12.92 ± 2.96	9.46 ± 2.09	16.16 ± 0.10	41.21 ± 9.85	16.66 ± 0.04
Text (Synonym)	AUROC	15.38 ± 3.11	43.87 ± 8.69	37.65 ± 9.60	38.43 ± 2.58	37.40 ± 1.50	32.44 ± 1.59	32.30 ± 7.72	72.39 ± 5.91
	AUPR	68.16 ± 1.00	79.80 ± 3.78	77.17 ± 4.16	77.54 ± 1.41	76.72 ± 0.59	77.05 ± 0.52	77.21 ± 2.41	90.32 ± 1.80
	FPR95	99.95 ± 0.04	94.80 ± 3.43	96.63 ± 2.41	98.17 ± 0.24	98.27 ± 0.21	97.87 ± 0.58	97.55 ± 1.70	65.33 ± 5.66
Text (Antonym)	AUROC	14.57 ± 2.77	41.46 ± 6.79	34.72 ± 7.17	35.49 ± 1.98	34.61 ± 1.17	29.68 ± 0.83	33.05 ± 10.44	73.93 ± 6.10
	AUPR	67.91 ± 0.92	78.55 ± 2.65	75.81 ± 3.02	75.93 ± 0.98	75.31 ± 0.47	76.14 ± 0.30	77.37 ± 3.15	90.70 ± 1.80
	FPR95	99.97 ± 0.03	95.63 ± 2.39	97.89 ± 1.12	98.53 ± 0.13	98.58 ± 0.10	98.58 ± 0.34	96.31 ± 4.40	59.52 ± 7.06
Text Swap (Both)	AUROC	69.23 ± 0.18	62.84 ± 1.05	58.68 ± 0.96	66.15 ± 0.65	66.64 ± 0.48	73.37 ± 0.40	69.21 ± 2.04	81.55 ± 3.65
	AUPR	91.60 ± 0.08	87.62 ± 0.56	85.53 ± 0.49	89.88 ± 0.38	90.63 ± 0.23	89.04 ± 0.11	87.77 ± 0.54	92.36 ± 1.08
	FPR95	88.12 ± 0.37	85.33 ± 0.16	87.93 ± 0.48	85.59 ± 0.17	85.61 ± 0.19	58.74 ± 0.60	67.75 ± 2.80	31.77 ± 9.11
Text Swap (Intra)	AUROC	67.88 ± 0.20	56.39 ± 0.56	54.62 ± 0.70	60.51 ± 0.47	61.34 ± 0.40	65.91 ± 0.40	61.70 ± 1.46	77.18 ± 3.04
	AUPR	91.05 ± 0.11	85.75 ± 0.31	84.79 ± 0.33	87.83 ± 0.28	88.55 ± 0.21	87.31 ± 0.12	86.05 ± 0.36	91.21 ± 0.96
	FPR95	88.60 ± 0.48	91.12 ± 0.23	91.80 ± 0.53	89.99 ± 0.09	89.94 ± 0.09	79.90 ± 0.61	82.13 ± 1.76	54.25 ± 4.97
Text Swap (Inter)	AUROC	68.74 ± 0.18	62.62 ± 1.07	58.58 ± 1.01	65.94 ± 0.65	66.40 ± 0.47	73.15 ± 0.41	69.05 ± 2.03	81.23 ± 3.40
	AUPR	91.37 ± 0.08	87.59 ± 0.56	85.56 ± 0.48	89.80 ± 0.37	90.50 ± 0.22	89.01 ± 0.10	87.75 ± 0.54	92.28 ± 1.03
	FPR95	88.17 ± 0.41	85.83 ± 0.18	88.21 ± 0.46	85.60 ± 0.08	85.61 ± 0.09	59.65 ± 0.80	69.04 ± 2.65	33.67 ± 8.23
Semantic (0.75)	AUROC	99.25 ± 0.06	77.42 ± 4.62	70.80 ± 2.87	91.08 ± 1.95	94.74 ± 1.13	83.87 ± 0.10	59.81 ± 9.32	83.35 ± 0.04
	AUPR	99.78 ± 0.04	88.04 ± 2.05	85.70 ± 1.17	95.79 ± 0.95	98.25 ± 0.40	88.84 ± 0.05	79.47 ± 3.15	88.74 ± 0.04
	FPR95	1.10 ± 0.15	31.22 ± 5.57	40.83 ± 3.67	15.01 ± 2.94	11.32 ± 2.10	16.23 ± 0.12	42.38 ± 9.86	16.73 ± 0.04
Semantic (0.85)	AUROC	98.64 ± 0.07	76.92 ± 3.66	71.59 ± 2.77	90.72 ± 1.55	94.39 ± 0.89	83.82 ± 0.12	61.40 ± 8.40	83.27 ± 0.05
	AUPR	99.70 ± 0.02	88.91 ± 1.55	87.20 ± 1.09	96.28 ± 0.66	98.39 ± 0.28	88.82 ± 0.06	80.10 ± 2.87	88.98 ± 0.03
	FPR95	1.97 ± 0.33	35.88 ± 4.91	44.63 ± 3.87	18.19 ± 2.86	14.03 ± 1.89	16.30 ± 0.12	41.70 ± 9.14	16.88 ± 0.07
Semantic (0.95)	AUROC	96.17 ± 0.15	76.16 ± 2.34	73.04 ± 2.71	89.47 ± 1.11	92.97 ± 0.69	83.67 ± 0.19	65.28 ± 6.26	83.07 ± 0.09
	AUPR	99.25 ± 0.04	90.38 ± 0.95	89.47 ± 1.00	96.56 ± 0.42	98.25 ± 0.21	88.85 ± 0.07	81.99 ± 2.15	89.82 ± 0.05
	FPR95	16.25 ± 1.62	46.17 ± 3.19	52.82 ± 4.91	25.92 ± 2.87	20.89 ± 2.28	16.57 ± 0.27	40.15 ± 7.44	17.50 ± 0.12
	ID Acc	54.37 ± 0.28	54.37 ± 0.28	54.37 ± 0.28	54.37 ± 0.28	54.37 ± 0.28	54.37 ± 0.28	54.20 ± 0.30	54.45 ± 0.09
Label (Random 0 2 5 9 10 11 15 18)	AUROC	59.48 ± 0.26	63.99 ± 0.93	62.93 ± 1.01	64.15 ± 0.32	63.92 ± 0.21	57.26 ± 0.12	57.43 ± 0.19	79.26 ± 0.04
	AUPR	92.58 ± 0.11	93.53 ± 0.21	93.32 ± 0.23	93.75 ± 0.09	93.74 ± 0.06	90.20 ± 0.08	89.98 ± 0.06	95.02 ± 0.01
	FPR95	93.14 ± 0.42	86.91 ± 1.18	88.41 ± 1.15	88.98 ± 0.54	89.04 ± 0.54	81.31 ± 1.05	86.24 ± 1.05	24.46 ± 0.05
	ID Acc	66.51 ± 0.17	66.51 ± 0.17	66.51 ± 0.17	66.51 ± 0.17	66.51 ± 0.17	66.51 ± 0.17	64.89 ± 0.21	67.89 ± 0.13
Label (Disimilar 1 3 4 6 12 19 21 23)	AUROC	46.79 ± 0.14	58.30 ± 0.41	56.36 ± 0.41	53.64 ± 0.43	52.25 ± 0.45	51.65 ± 0.64	52.66 ± 0.92	55.56 ± 0.34
	AUPR	54.07 ± 0.20	61.95 ± 0.15	60.76 ± 0.18	58.73 ± 0.36	57.34 ± 0.41	56.34 ± 0.33	56.79 ± 0.65	58.06 ± 0.25
	FPR95	97.00 ± 0.19	90.24 ± 0.41	91.33 ± 0.56	93.44 ± 0.29	93.65 ± 0.37	91.07 ± 0.33	92.10 ± 1.68	86.92 ± 0.21
	ID Acc	59.98 ± 0.18	59.98 ± 0.18	59.98 ± 0.18	59.98 ± 0.18	59.98 ± 0.18	59.98 ± 0.18	60.12 ± 0.35	59.85 ± 0.71
Label (Similar 0 2 5 7 8 9 10 22)	AUROC	54.12 ± 0.31	60.88 ± 0.47	58.65 ± 0.53	59.99 ± 0.55	59.46 ± 0.55	56.78 ± 0.63	56.20 ± 0.22	81.55 ± 0.54
	AUPR	88.21 ± 0.12	90.11 ± 0.15	89.31 ± 0.21	89.94 ± 0.16	89.80 ± 0.16	87.32 ± 0.18	86.82 ± 0.03	94.05 ± 0.88
	FPR95	93.63 ± 0.26	88.36 ± 0.49	89.31 ± 0.64	89.70 ± 0.65	89.69 ± 0.59	83.79 ± 0.34	86.11 ± 0.62	22.04 ± 0.72
	ID Acc	61.53 ± 0.49	61.53 ± 0.49	61.53 ± 0.49	61.53 ± 0.49	61.53 ± 0.49	61.53 ± 0.49	61.62 ± 0.28	62.50 ± 0.50

Table 23: Performance comparison of OOD detection methods on **BookChild dataset** across different OOD shifts and configurations.

OOD Type	Metric	w/o Prop.					w/ Prop.		
		MAHA	MSP	ODIN	NECO	Energy	GNNSafe	NodeSafe	TNT-OOD
Feature (0.50)	AUROC	69.18 ± 0.51	65.44 ± 0.18	64.30 ± 0.32	68.04 ± 0.54	68.34 ± 0.55	66.29 ± 0.34	61.19 ± 0.21	73.13 ± 0.62
	AUPR	89.82 ± 0.10	88.81 ± 0.07	88.58 ± 0.15	90.05 ± 0.24	90.42 ± 0.26	88.33 ± 0.20	87.32 ± 0.20	91.97 ± 0.25
	FPR95	72.72 ± 2.09	84.24 ± 0.25	86.55 ± 0.08	86.13 ± 0.59	86.28 ± 0.54	63.75 ± 0.82	80.31 ± 0.09	47.91 ± 1.97
Feature (0.70)	AUROC	70.67 ± 0.52	65.73 ± 0.33	64.44 ± 0.28	71.28 ± 0.41	72.00 ± 0.44	68.51 ± 0.23	65.23 ± 0.28	75.18 ± 0.53
	AUPR	90.59 ± 0.13	88.87 ± 0.11	88.50 ± 0.10	91.16 ± 0.16	91.90 ± 0.18	88.73 ± 0.16	88.04 ± 0.17	92.72 ± 0.20
	FPR95	70.07 ± 1.83	83.79 ± 0.66	87.02 ± 0.57	80.90 ± 0.84	81.13 ± 0.70	54.22 ± 0.44	63.58 ± 0.63	41.30 ± 1.53
Feature (0.90)	AUROC	67.48 ± 0.31	61.37 ± 0.28	60.86 ± 0.18	67.74 ± 0.33	68.52 ± 0.33	67.04 ± 0.18	65.09 ± 0.35	74.99 ± 0.53
	AUPR	89.86 ± 0.11	87.40 ± 0.10	87.26 ± 0.07	90.09 ± 0.12	90.99 ± 0.14	87.57 ± 0.11	86.98 ± 0.13	92.54 ± 0.50
	FPR95	75.66 ± 0.89	89.08 ± 0.58	90.39 ± 0.36	83.36 ± 0.75	83.47 ± 0.65	55.00 ± 0.45	57.42 ± 0.98	40.55 ± 1.56
Structure (Mild)	AUROC	97.07 ± 0.17	82.29 ± 0.17	85.40 ± 0.31	90.82 ± 0.28	93.02 ± 0.24	60.25 ± 0.17	56.51 ± 0.18	61.89 ± 0.32
	AUPR	99.32 ± 0.06	92.84 ± 0.08	94.75 ± 0.14	96.60 ± 0.08	98.02 ± 0.09	79.49 ± 0.05	78.38 ± 0.05	79.94 ± 0.11
	FPR95	13.08 ± 0.09	44.14 ± 0.24	42.52 ± 0.69	32.42 ± 0.79	31.06 ± 0.72	43.59 ± 0.28	50.50 ± 0.29	39.97 ± 0.53
Structure (Medium)	AUROC	97.00 ± 0.18	82.19 ± 0.20	85.25 ± 0.30	90.79 ± 0.27	93.01 ± 0.24	60.27 ± 0.16	56.46 ± 0.18	61.88 ± 0.32
	AUPR	99.29 ± 0.07	92.86 ± 0.12	94.75 ± 0.14	96.59 ± 0.08	98.00 ± 0.10	79.50 ± 0.05	78.37 ± 0.06	79.94 ± 0.10
	FPR95	12.99 ± 0.14	44.05 ± 0.42	42.53 ± 0.71	31.87 ± 0.71	30.54 ± 0.78	43.54 ± 0.23	50.50 ± 0.32	39.99 ± 0.52
Structure (Strong)	AUROC	99.17 ± 0.25	73.82 ± 0.51	75.99 ± 0.56	90.70 ± 0.61	95.18 ± 0.37	63.27 ± 0.14	61.92 ± 0.13	63.74 ± 0.26
	AUPR	99.70 ± 0.12	86.02 ± 0.20	87.45 ± 0.27	94.47 ± 0.28	97.64 ± 0.09	80.32 ± 0.05	79.85 ± 0.05	80.47 ± 0.09
	FPR95	2.19 ± 0.34	36.60 ± 0.76	36.29 ± 0.76	15.46 ± 0.88	10.25 ± 0.91	36.99 ± 0.16	38.39 ± 0.11	36.34 ± 0.30
Text (Synonym)	AUROC	32.42 ± 3.23	68.48 ± 1.01	64.07 ± 1.83	68.69 ± 1.50	68.10 ± 1.43	66.68 ± 0.71	64.54 ± 0.44	72.78 ± 1.64
	AUPR	74.62 ± 1.16	89.93 ± 0.42	88.33 ± 0.75	90.20 ± 0.55	90.12 ± 0.52	87.41 ± 0.31	87.41 ± 0.11	91.81 ± 0.76
	FPR95	99.05 ± 0.39	81.11 ± 1.28	87.15 ± 1.48	84.73 ± 1.40	85.15 ± 1.12	59.15 ± 2.17	64.27 ± 1.44	49.12 ± 5.08
Text (Antonym)	AUROC	30.76 ± 3.74	63.45 ± 1.28	58.54 ± 2.14	65.06 ± 2.07	64.79 ± 1.95	64.69 ± 1.06	60.61 ± 0.47	73.59 ± 1.86
	AUPR	74.02 ± 1.41	87.72 ± 0.45	85.72 ± 0.84	88.28 ± 0.84	88.39 ± 0.78	86.45 ± 0.45	85.94 ± 0.12	92.14 ± 0.85
	FPR95	99.11 ± 0.37	85.96 ± 1.30	91.09 ± 1.28	87.61 ± 1.33	87.86 ± 1.11	64.73 ± 2.70	74.15 ± 1.49	46.96 ± 5.82
Text Swap (Both)	AUROC	51.05 ± 0.10	55.42 ± 0.44	55.48 ± 0.40	57.85 ± 0.34	58.02 ± 0.33	61.62 ± 0.22	64.32 ± 0.45	71.63 ± 0.55
	AUPR	84.72 ± 0.20	85.61 ± 0.18	85.77 ± 0.20	86.99 ± 0.18	87.53 ± 0.20	84.59 ± 0.02	85.53 ± 0.13	90.79 ± 0.21
	FPR95	93.63 ± 0.14	92.46 ± 0.50	92.73 ± 0.39	89.21 ± 0.29	89.31 ± 0.35	66.13 ± 0.95	56.72 ± 1.27	45.84 ± 2.15
Text Swap (Intra)	AUROC	50.96 ± 0.11	47.91 ± 0.12	48.43 ± 0.13	48.27 ± 0.07	48.39 ± 0.32	53.32 ± 0.03	53.54 ± 0.19	60.83 ± 0.66
	AUPR	84.51 ± 0.02	82.21 ± 0.12	82.48 ± 0.12	82.47 ± 0.10	82.70 ± 0.09	82.90 ± 0.04	83.44 ± 0.13	87.05 ± 0.22
	FPR95	93.54 ± 0.06	94.95 ± 0.10	94.78 ± 0.04	94.81 ± 0.15	94.82 ± 0.15	88.70 ± 0.18	87.99 ± 0.07	79.29 ± 1.46
Text Swap (Inter)	AUROC	50.40 ± 0.05	55.19 ± 0.36	55.21 ± 0.31	57.46 ± 0.32	57.61 ± 0.32	61.32 ± 0.22	63.93 ± 0.46	71.10 ± 0.49
	AUPR	84.30 ± 0.20	85.43 ± 0.15	85.57 ± 0.17	86.71 ± 0.19	87.22 ± 0.20	84.55 ± 0.03	85.46 ± 0.14	90.64 ± 0.19
	FPR95	93.87 ± 0.14	92.27 ± 0.52	92.56 ± 0.49	89.22 ± 0.30	89.34 ± 0.39	66.96 ± 0.98	59.26 ± 1.32	47.97 ± 1.88
Semantic (0.75)	AUROC	92.84 ± 0.32	94.90 ± 0.15	95.64 ± 0.06	95.79 ± 0.07	96.01 ± 0.16	9.90 ± 0.21	35.55 ± 2.77	62.91 ± 1.50
	AUPR	98.56 ± 0.16	98.43 ± 0.09	99.07 ± 0.03	98.96 ± 0.02	99.11 ± 0.10	67.15 ± 0.02	73.80 ± 0.61	83.20 ± 0.35
	FPR95	79.02 ± 0.63	32.11 ± 0.29	37.71 ± 0.10	34.08 ± 0.99	35.29 ± 1.22	97.33 ± 0.27	74.45 ± 2.76	43.74 ± 3.20
Semantic (0.85)	AUROC	86.67 ± 0.45	91.71 ± 0.16	92.14 ± 0.09	92.70 ± 0.14	92.91 ± 0.23	23.29 ± 0.89	37.62 ± 3.09	64.32 ± 1.03
	AUPR	97.26 ± 0.23	97.88 ± 0.06	98.40 ± 0.02	98.37 ± 0.02	98.50 ± 0.11	72.62 ± 0.25	76.08 ± 0.67	85.69 ± 0.35
	FPR95	90.04 ± 0.53	60.36 ± 1.03	63.70 ± 0.90	59.67 ± 1.47	60.14 ± 1.30	92.82 ± 0.77	74.59 ± 4.25	48.19 ± 3.29
Semantic (0.95)	AUROC	72.28 ± 0.51	82.42 ± 0.12	82.59 ± 0.19	83.89 ± 0.24	84.08 ± 0.32	49.93 ± 2.01	39.62 ± 2.61	69.19 ± 0.89
	AUPR	93.61 ± 0.31	95.99 ± 0.03	96.29 ± 0.04	96.51 ± 0.06	96.60 ± 0.12	82.72 ± 0.40	80.25 ± 0.49	89.82 ± 0.64
	FPR95	92.86 ± 0.32	78.90 ± 0.11	78.68 ± 0.63	76.85 ± 0.74	76.93 ± 0.85	82.29 ± 0.70	98.81 ± 3.15	55.79 ± 2.73
	ID Acc	78.60 ± 0.08	78.60 ± 0.08	78.60 ± 0.08	78.60 ± 0.08	78.60 ± 0.08	78.60 ± 0.08	74.44 ± 0.28	83.48 ± 0.28
Label (Random 1 3 5 7 9)	AUROC	41.23 ± 1.33	77.32 ± 0.42	77.51 ± 0.64	75.44 ± 0.90	75.09 ± 1.05	65.47 ± 0.67	62.18 ± 0.22	76.71 ± 0.38
	AUPR	86.67 ± 0.47	95.86 ± 0.05	96.07 ± 0.12	95.59 ± 0.18	95.48 ± 0.26	92.36 ± 0.17	91.79 ± 0.13	95.99 ± 0.12
	FPR95	98.64 ± 0.26	68.14 ± 0.89	69.13 ± 1.03	73.62 ± 1.55	73.69 ± 1.56	65.71 ± 1.85	73.94 ± 0.14	35.84 ± 0.35
	ID Acc	87.58 ± 0.08	87.58 ± 0.08	87.58 ± 0.08	87.58 ± 0.08	87.58 ± 0.08	87.58 ± 0.08	86.31 ± 0.14	93.52 ± 0.28
Label (Disimilar 0 4 6 10 11)	AUROC	36.79 ± 0.83	59.17 ± 1.54	57.76 ± 1.85	57.75 ± 1.48	57.32 ± 1.44	58.69 ± 0.83	55.75 ± 0.26	77.75 ± 0.16
	AUPR	56.62 ± 0.44	73.29 ± 0.96	72.65 ± 1.14	72.46 ± 1.02	72.05 ± 0.98	70.44 ± 0.65	69.81 ± 0.22	86.35 ± 0.18
	FPR95	99.09 ± 0.22	95.67 ± 0.48	96.42 ± 0.33	95.99 ± 0.26	95.99 ± 0.26	89.45 ± 1.02	94.65 ± 0.50	36.42 ± 0.40
	ID Acc	86.27 ± 0.19	86.27 ± 0.19	86.27 ± 0.19	86.27 ± 0.19	86.27 ± 0.19	86.27 ± 0.19	85.74 ± 0.14	89.15 ± 0.04
Label (Similar 1 2 3 5 9)	AUROC	38.99 ± 0.07	77.46 ± 0.16	77.69 ± 0.22	76.61 ± 0.20	76.37 ± 0.25	66.56 ± 0.09	63.68 ± 0.22	76.83 ± 0.56
	AUPR	87.65 ± 0.09	96.45 ± 0.01	96.62 ± 0.03	96.36 ± 0.06	96.31 ± 0.10	93.38 ± 0.09	93.21 ± 0.07	96.52 ± 0.10
	FPR95	99.08 ± 0.07	64.75 ± 0.82	65.24 ± 0.76	69.69 ± 0.66	69.90 ± 0.75	60.45 ± 0.78	68.34 ± 0.92	37.18 ± 0.45
	ID Acc	86.60 ± 0.05	86.60 ± 0.05	86.60 ± 0.05	86.60 ± 0.05	86.60 ± 0.05	86.60 ± 0.05	85.16 ± 0.07	91.14 ± 0.15

Table 24: Performance comparison of OOD detection methods on **ElePhoto dataset** across different OOD shifts and configurations.

OOD Type	Metric	w/o Prop.					w/ Prop.		
		MAHA	MSP	ODIN	NECO	Energy	GNNSafe	NodeSafe	TNT-OOD
Feature (0.50)	AUROC	43.68 ± 1.27	60.82 ± 0.52	58.66 ± 0.49	58.57 ± 0.39	57.93 ± 0.34	61.49 ± 0.25	61.83 ± 0.30	73.64 ± 0.13
	AUPR	86.70 ± 0.40	92.38 ± 0.23	92.04 ± 0.23	91.33 ± 0.16	90.80 ± 0.09	90.69 ± 0.01	92.10 ± 0.03	95.78 ± 0.04
	FPR95	97.54 ± 0.16	89.51 ± 0.32	90.87 ± 0.29	91.80 ± 0.37	91.85 ± 0.36	75.82 ± 0.40	76.73 ± 0.93	53.69 ± 0.57
Feature (0.70)	AUROC	42.86 ± 1.82	68.76 ± 0.73	67.76 ± 0.62	64.96 ± 0.79	63.93 ± 0.76	66.17 ± 0.44	66.78 ± 0.25	76.45 ± 0.08
	AUPR	86.17 ± 0.59	94.21 ± 0.28	94.31 ± 0.21	93.04 ± 0.29	92.36 ± 0.19	91.27 ± 0.05	92.72 ± 0.03	96.27 ± 0.04
	FPR95	97.45 ± 0.30	76.36 ± 0.84	77.50 ± 0.82	80.87 ± 1.35	81.15 ± 1.33	56.57 ± 1.37	51.13 ± 0.67	43.30 ± 0.15
Feature (0.90)	AUROC	45.06 ± 1.38	66.22 ± 1.02	66.52 ± 1.23	63.50 ± 0.80	62.67 ± 0.74	65.87 ± 0.43	66.21 ± 0.13	76.76 ± 0.07
	AUPR	87.08 ± 0.48	93.81 ± 0.33	94.15 ± 0.31	92.93 ± 0.30	92.33 ± 0.17	91.05 ± 0.03	91.85 ± 0.08	96.27 ± 0.50
	FPR95	96.80 ± 0.12	79.10 ± 1.42	78.47 ± 2.08	80.62 ± 1.34	80.78 ± 1.26	54.03 ± 1.56	47.56 ± 0.35	40.65 ± 0.05
Structure (Mild)	AUROC	79.03 ± 1.50	86.54 ± 0.69	91.91 ± 0.51	89.13 ± 0.46	89.22 ± 0.29	57.33 ± 0.33	54.93 ± 0.46	59.37 ± 0.09
	AUPR	95.63 ± 0.41	97.83 ± 0.20	98.90 ± 0.10	98.08 ± 0.16	97.68 ± 0.11	85.76 ± 0.08	85.14 ± 0.10	86.09 ± 0.02
	FPR95	61.76 ± 1.55	46.41 ± 0.77	36.47 ± 0.86	40.45 ± 0.83	39.60 ± 0.88	50.14 ± 0.23	51.68 ± 0.38	44.27 ± 0.07
Structure (Medium)	AUROC	79.31 ± 1.41	85.91 ± 0.67	91.45 ± 0.50	88.66 ± 0.45	88.78 ± 0.29	57.08 ± 0.31	54.63 ± 0.49	59.24 ± 0.10
	AUPR	95.72 ± 0.41	97.70 ± 0.20	98.83 ± 0.10	97.97 ± 0.16	97.57 ± 0.11	85.72 ± 0.08	85.10 ± 0.10	86.07 ± 0.03
	FPR95	60.79 ± 1.28	47.28 ± 0.89	37.49 ± 0.86	41.31 ± 0.80	40.63 ± 0.72	50.62 ± 0.17	52.50 ± 0.54	44.63 ± 0.10
Structure (Strong)	AUROC	37.79 ± 8.42	90.91 ± 4.39	93.87 ± 2.62	92.55 ± 1.58	92.00 ± 0.99	61.38 ± 0.35	58.76 ± 0.56	62.19 ± 0.01
	AUPR	82.55 ± 2.04	97.07 ± 1.25	98.13 ± 0.80	97.26 ± 0.45	96.86 ± 0.24	86.42 ± 0.08	85.78 ± 0.13	86.62 ± 0.00
	FPR95	80.03 ± 5.64	14.26 ± 6.89	10.84 ± 4.41	11.57 ± 2.68	12.27 ± 1.86	38.80 ± 0.37	41.36 ± 0.61	37.97 ± 0.04
Text (Synonym)	AUROC	18.09 ± 2.49	61.65 ± 4.77	54.99 ± 5.89	53.99 ± 7.79	53.09 ± 7.90	58.76 ± 5.38	66.28 ± 1.40	73.60 ± 0.92
	AUPR	79.35 ± 0.76	92.09 ± 1.32	90.91 ± 1.58	90.40 ± 2.17	90.05 ± 2.19	90.44 ± 1.28	92.70 ± 0.57	95.68 ± 0.23
	FPR95	99.77 ± 0.07	87.54 ± 4.80	91.21 ± 3.68	92.36 ± 3.21	92.52 ± 3.12	78.04 ± 9.67	58.21 ± 3.75	53.48 ± 2.98
Text (Antonym)	AUROC	17.15 ± 2.97	59.09 ± 5.16	52.38 ± 6.42	51.63 ± 8.17	50.84 ± 8.26	57.01 ± 5.69	65.62 ± 1.05	72.87 ± 0.85
	AUPR	79.04 ± 0.91	91.32 ± 1.46	90.11 ± 1.78	89.60 ± 2.34	89.32 ± 2.35	90.10 ± 1.39	92.51 ± 0.43	95.51 ± 0.22
	FPR95	99.79 ± 0.06	89.88 ± 4.32	93.07 ± 2.98	93.60 ± 2.69	93.69 ± 2.62	81.42 ± 8.58	60.83 ± 3.63	55.79 ± 3.01
Text Swap (Both)	AUROC	42.73 ± 0.37	60.18 ± 0.68	61.99 ± 0.73	57.66 ± 0.69	57.19 ± 0.65	63.15 ± 0.49	65.79 ± 0.10	74.07 ± 0.15
	AUPR	87.68 ± 0.28	92.65 ± 0.22	93.33 ± 0.19	92.08 ± 0.19	91.82 ± 0.12	91.19 ± 0.06	91.60 ± 0.02	95.48 ± 0.07
	FPR95	98.24 ± 0.14	87.23 ± 0.67	86.09 ± 0.75	88.47 ± 0.33	88.59 ± 0.33	61.29 ± 1.69	49.70 ± 0.11	42.92 ± 0.20
Text Swap (Intra)	AUROC	47.45 ± 0.23	45.51 ± 0.07	46.48 ± 0.06	45.29 ± 0.14	45.35 ± 0.14	52.71 ± 0.08	54.68 ± 0.18	61.97 ± 0.24
	AUPR	89.65 ± 0.09	88.62 ± 0.05	88.93 ± 0.06	88.57 ± 0.03	88.59 ± 0.02	89.89 ± 0.01	90.29 ± 0.03	93.10 ± 0.09
	FPR95	95.94 ± 0.08	96.38 ± 0.09	96.63 ± 0.17	96.46 ± 0.09	96.46 ± 0.09	90.05 ± 0.46	89.80 ± 0.20	81.30 ± 0.59
Text Swap (Inter)	AUROC	42.85 ± 0.34	59.91 ± 0.62	61.81 ± 0.70	57.43 ± 0.61	56.97 ± 0.57	63.05 ± 0.44	65.70 ± 0.12	73.84 ± 0.13
	AUPR	87.72 ± 0.27	92.57 ± 0.21	93.28 ± 0.19	92.01 ± 0.17	91.76 ± 0.10	91.15 ± 0.05	91.58 ± 0.03	95.44 ± 0.06
	FPR95	98.28 ± 0.15	87.27 ± 0.63	86.19 ± 0.69	88.57 ± 0.33	88.70 ± 0.28	61.42 ± 1.55	50.27 ± 0.20	43.73 ± 0.34
Semantic (0.75)	AUROC	41.79 ± 7.87	91.86 ± 3.84	94.86 ± 2.25	93.11 ± 1.39	92.43 ± 0.91	61.48 ± 0.36	58.87 ± 0.55	62.36 ± 0.05
	AUPR	84.81 ± 2.03	98.00 ± 0.84	98.90 ± 0.49	98.05 ± 0.30	97.51 ± 0.15	86.47 ± 0.09	85.82 ± 0.13	86.94 ± 0.03
	FPR95	84.29 ± 4.57	14.56 ± 7.20	10.48 ± 4.54	12.20 ± 2.86	13.11 ± 2.03	38.81 ± 0.40	41.37 ± 0.62	37.95 ± 0.03
Semantic (0.85)	AUROC	54.18 ± 6.31	92.58 ± 1.68	94.43 ± 1.39	93.34 ± 0.83	92.60 ± 0.84	61.66 ± 0.36	58.81 ± 0.50	62.43 ± 0.04
	AUPR	88.90 ± 1.47	98.47 ± 0.31	99.02 ± 0.24	98.37 ± 0.25	97.77 ± 0.12	86.61 ± 0.07	85.92 ± 0.12	87.47 ± 0.02
	FPR95	80.68 ± 5.13	17.16 ± 3.81	13.73 ± 3.39	13.42 ± 2.25	13.99 ± 2.29	38.86 ± 0.40	41.86 ± 0.59	38.28 ± 0.05
Semantic (0.95)	AUROC	74.52 ± 1.66	81.42 ± 1.00	86.95 ± 1.22	83.73 ± 0.74	83.64 ± 0.58	60.25 ± 0.26	56.91 ± 0.49	62.06 ± 0.08
	AUPR	94.75 ± 0.47	96.52 ± 0.20	98.02 ± 0.22	96.72 ± 0.09	96.39 ± 0.06	86.71 ± 0.09	86.09 ± 0.08	89.26 ± 0.04
	FPR95	71.19 ± 4.36	52.31 ± 0.36	45.82 ± 0.94	46.41 ± 1.11	46.06 ± 1.20	44.22 ± 0.35	47.98 ± 0.44	41.90 ± 0.23
	ID Acc	76.97 ± 0.15	76.97 ± 0.15	76.97 ± 0.15	76.97 ± 0.15	76.97 ± 0.15	76.97 ± 0.15	75.22 ± 0.07	84.75 ± 0.09
Label (Random 1 2 3 4 5)	AUROC	36.27 ± 2.43	49.96 ± 1.18	47.76 ± 1.18	51.10 ± 0.00	51.11 ± 1.00	55.89 ± 0.63	54.75 ± 0.40	79.36 ± 0.41
	AUPR	92.00 ± 0.56	95.19 ± 0.13	94.86 ± 0.12	95.38 ± 0.00	95.39 ± 0.11	95.40 ± 0.10	95.32 ± 0.13	98.39 ± 0.04
	FPR95	98.82 ± 0.27	100.00 ± 0.00	100.00 ± 0.00	97.44 ± 0.00	97.44 ± 0.19	93.16 ± 1.27	94.61 ± 0.54	38.06 ± 0.28
	ID Acc	92.07 ± 0.20	92.07 ± 0.20	92.07 ± 0.20	92.07 ± 0.20	92.07 ± 0.20	92.07 ± 0.20	90.46 ± 0.30	93.60 ± 0.56
Label (Disimilar 2 3 4 8 9)	AUROC	37.52 ± 1.54	64.15 ± 0.60	62.58 ± 0.73	62.18 ± 1.19	61.93 ± 1.25	59.31 ± 0.76	57.27 ± 0.42	69.07 ± 1.79
	AUPR	82.04 ± 0.32	91.99 ± 0.24	91.92 ± 0.18	91.59 ± 0.40	91.54 ± 0.41	89.50 ± 0.27	89.52 ± 0.22	91.34 ± 0.48
	FPR95	98.33 ± 0.74	89.32 ± 0.42	94.15 ± 4.14	89.43 ± 1.10	89.47 ± 1.09	78.75 ± 0.39	81.76 ± 0.82	73.09 ± 2.02
	ID Acc	84.35 ± 0.33	84.35 ± 0.33	84.35 ± 0.33	84.35 ± 0.33	84.35 ± 0.33	84.35 ± 0.33	83.238 ± 0.28	90.89 ± 0.25
Label (Similar 0 1 5 6 7)	AUROC	33.54 ± 1.04	57.13 ± 0.08	55.45 ± 0.35	53.26 ± 0.46	53.13 ± 0.43	58.63 ± 0.30	61.90 ± 0.29	81.56 ± 0.11
	AUPR	87.25 ± 0.20	93.44 ± 0.02	93.15 ± 0.05	92.67 ± 0.08	92.63 ± 0.05	93.48 ± 0.01	94.13 ± 0.08	97.71 ± 0.02
	FPR95	99.43 ± 0.11	100.00 ± 0.00	100.00 ± 0.00	93.71 ± 0.17	93.71 ± 0.17	87.57 ± 0.17	88.11 ± 0.68	34.01 ± 0.03
	ID Acc	88.14 ± 0.32	88.14 ± 0.32	88.14 ± 0.32	88.14 ± 0.32	88.14 ± 0.32	88.14 ± 0.32	85.7 ± 0.10	91.11 ± 0.43

Table 25: Performance comparison of OOD detection methods on **EleComp dataset** across different OOD shifts and configurations.