PLACEGUARD: RECORD-PLACEMENT VERIFICATION FOR MULTIMODAL GRAPH OOD DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Open-world catalog admission must determine whether a submitted text-image record belongs at a claimed graph identity before assigning a known class. Existing multimodal OOD detectors evaluate record familiarity, and graph detectors evaluate a node with its topology, yet neither view directly verifies the record-position pairing. We study record-placement OOD, where a record and graph position are individually in-distribution but their assignment is unsupported. An in-distribution record can therefore be out of place. A position-controlled study reassigns intact records while keeping recipients, relations, and target-excluded contexts constant. One-to-one reassignment makes every deterministic record-only score exactly nondiscriminative and leaves position-only scores invariant, whereas record-position compatibility separates cross-class and within-class reassignment. PlaceGuard turns that observation into three necessary checks for within-record validity, record-position fit, and context familiarity. It multiplies the first two supports so that both must be high and uses a one-way context check that increases unknownness only beyond the clean validation tail. Across six public graphs and three class splits, PlaceGuard reaches conventional semantic OOD AUROC 0.913 and macro weakest-locus AUROC 0.780, compared with 0.504 for the strongest method under the shared protocol. The results show that assignment verification complements conventional semantic detection when a familiar record is submitted to an unsupported relational position.

1 INTRODUCTION

Product catalogs and social platforms rarely store entities as isolated feature vectors. A product has a title and image as well as co-view or co-purchase history. An image-caption post may be linked to earlier posts from the same author. Multimodal attributed graphs (MAGs) represent the former as node attributes and the latter as inter-entity relations (Ektefaie et al., 2023; Peng et al., 2024; Zhu et al., 2025; Yan et al., 2025; Wan et al., 2026). In an open-world workflow, admission must determine whether a newly supplied record supports a known class at the identifier and relations it claims, or should be routed to review before it enters the relational store.

A concrete catalog workflow has two stages. First, an upstream identity service maps a submitted title and image to an existing catalog identifier using seller metadata, stock identifiers, or deduplication. Second, a verifier receives that candidate record together with the claimed identifier and the relations already stored for it. Identity retrieval, link creation, and catalog repair remain upstream operations. The verifier returns a known category and admits the submitted record at the claimed position, or routes the assignment to review before the catalog is updated.

This decision contains two objects with different roles. The *multimodal record* is the target's text and image. The *graph position* is the target identifier together with its stored incident relations, which summarize how that identifier has interacted with other entities. For example, catalog integration or identifier reuse may attach a perfectly plausible title-image pair to the co-purchase history of another product. Neither the record nor the position must be unusual in isolation. Their assignment can still be unsupported. The graph is not a third modality in this task. It is the record's claimed relational position, so the verifier must evaluate whether these familiar objects correspond.

Multimodal OOD detection measures semantic familiarity, modality-wise confidence, and within-record agreement (Dong et al., 2024; Li et al., 2025). Graph OOD detection uses node features and

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

topology to refine an OOD score (Wu et al., 2023; Yang et al., 2024; Bao et al., 2024; Ma et al., 2024). Contextual graph anomaly methods likewise compare attributes with graph structure (Ding et al., 2019; Xu et al., 2022). These families identify anomalous records, neighborhoods, and complete node states that differ from the training distribution. The unresolved decision is whether a record and graph position can each be in-distribution while their observed assignment is OOD at the claimed relational identity and should therefore be rejected before admission.

We answer this question with a position-controlled comparison. For the same known recipient nodes, we hold the identifier, degree, target-excluded context, and context posterior constant, and replace only the attached text, image, or intact text-image record. A one-to-one intact-record reassignment preserves the complete record multiset. Consequently, every deterministic record-only score has the same clean and reassigned score multiset and obtains AUROC 0.5. Position-only scores are also exactly invariant because the recipient positions never change. Joint record-position compatibility detects cross-class and within-class reassignment with AUROC $0.866/0.758$ on Movies and $0.972/0.885$ on EleFashion. Thus, the missing signal resides in the dependence between the record and its claimed position, which remains visible after both marginal objects are held constant across every comparison condition in the controlled study.

This observation calls for a different decision structure. Admission should require the record to be internally valid, to fit its claimed position, and to arrive in a familiar context. PlaceGuard implements these requirements as necessary checks. It multiplies within-record validity with record-position fit to form factorized assignment support. Since the product never exceeds either factor, a strong check cannot hide a failed one. A separate context path is neutral for ordinary clean contexts and can only increase rejection in the unfamiliar tail. Training uses known nodes with structured text, image, cross class, and same class record replacements. Clean known validation nodes define the score ranks and final threshold. The resulting verifier learns from known-class records and calibrates admission on clean known validation nodes, preserving the information available in an actual known-only deployment workflow for catalog updates and social-content admission decisions.

Our Contributions. (1) **In-depth Investigation.** We formalize record-placement OOD and establish a controlled case in which record-only and position-only scores are exactly uninformative, while their dependence identifies the unsupported assignment. (2) **Novel Method.** We introduce PlaceGuard, which factorizes assignment support into two necessary factors and gives context a one-way path that can increase rejection but cannot rescue a failed assignment. (3) **State-of-the-art Performance.** Under the shared fixed-position protocol, our verifier leads the comparison on 5/5 shift-type AUROC columns and achieves macro weakest locus AUROC 0.780 over six public graphs, compared with 0.504 for the strongest evaluated method under the same protocol.

2 Preliminaries and Problem Formulation

2.1 Multimodal attributed graph

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a graph with $N = |\mathcal{V}|$ nodes. Each node $i \in \mathcal{V}$ carries a text feature $\mathbf{x}_i^t \in \mathbb{R}^{d_t}$ and an image feature $\mathbf{x}_i^v \in \mathbb{R}^{d_v}$. An edge $(i, j) \in \mathcal{E}$ records an observed inter-entity relationship. We distinguish the current multimodal record attached to identifier i from its *graph position*, meaning identifier i together with its stored incident relations. The record contains entity-internal content. The graph position supplies cross-entity context. In the product graphs used in this work, edges arise from co-purchase, `also_buy`, or `also_view` relations. In the Reddit graphs, nodes are image-caption posts and the released construction connects posts through same-author `also_posted` groups. Connected nodes may share a class or express similarity, complementarity, shared authorship, or heterophily (Zuo et al., 2026; Guo et al., 2025; Zheng et al., 2026b).

The known label set is $\mathcal{Y}_{\mathrm{K}} = \{1, \ldots, C\}$. Training and validation contain only nodes from $\mathcal{Y}_{\mathrm{K}}$. At test time, a record may belong to a known class or an unseen class in $\mathcal{Y}_{\mathrm{U}}$, where $\mathcal{Y}_{\mathrm{K}} \cap \mathcal{Y}_{\mathrm{U}} = \varnothing$. The system outputs a class in $\mathcal{Y}_{\mathrm{K}}$ or $\perp$. Here, $\perp$ rejects an observed node state because its record is semantically unknown or its record-position assignment is unsupported. Fitting, checkpoint selection, and calibration use only known-class data.

The claimed identifier is observed at inference. Its existing relations may summarize prior co-view, co-purchase, or same-author activity. The new text and image are the only arriving record. This order

matches moderation and catalog update systems in which identity resolution precedes admission. It also separates our task from entity retrieval and link prediction.

2.2 SEMANTIC OOD AND RECORD-PLACEMENT OOD

We evaluate two reasons for returning $\perp$. A *semantic unknown* has a multimodal record outside the semantic support of the known classes, matching conventional open-set node classification and graph OOD detection (Wu et al., 2020; 2023; Gui et al., 2022). *Record-placement OOD* instead keeps an intact familiar text-image record but attaches it to a graph position unsupported by known record-position pairings. Its record may therefore have a known semantic class while the observed assignment is OOD. The rejection target is this assignment, which distinguishes a familiar record from a supported record at a particular position.

For a target node i, let $\mathbf{r}_i = (\mathbf{x}_i^t, \mathbf{x}_i^v)$ denote its complete record and let $\mathbf{c}_i$ denote context computed from connected support nodes without copying $\mathbf{r}_i$ into that context. A graph-free detector has the form $s_i = f(\mathbf{r}_i)$, whereas a context-aware detector may use $s_i = g(\mathbf{r}_i, \mathbf{c}_i)$. This information boundary makes $\mathbf{c}_i$ a summary of connected entities that excludes the target record itself.

2.3 OPEN-WORLD DECISION AND EVALUATION

Let $q_i(c)$ denote the support assigned to known class c and let u_i be a scalar unknownness score. A decision rule must return exactly one element of $\mathcal{Y}_{\mathrm{K}} \cup \{\perp\}$. It predicts the best-supported known class when u_i is below a threshold and returns $\perp$ otherwise. The threshold is estimated from clean known validation nodes. Consequently, neither semantic unknowns nor record-placement shifts can determine the operating point. We evaluate discrimination by comparing clean known test nodes separately against semantic unknowns and against each controlled replacement condition. AUROC and average precision (AP) measure threshold-independent ranking. For FPR95, unknown nodes are positives and scores are oriented so that larger values indicate greater unknownness. On the complete empirical ROC, FPR95 is the clean-known false-positive rate at the first point whose unknown TPR reaches 95%, without interpolation. Unknown-detection TPR is always paired with the realized clean acceptance. Known-class accuracy is measured before rejection, separating prediction quality from the effect of the admission threshold.

The observable graph is also restricted at a target decision. In the intended inductive admission setting, the recipient's links to an already admitted support store are available, so support-node membership is known even though labels for the arriving record are not. Context may use those support nodes connected to the target, but it excludes the target record and all evaluation-evaluation paths. This yields an inductive information boundary. Each decision compares a target record with relationship patterns learned from known support nodes, and each evaluated node is verified independently against that support. The setting differs from ordinary semantic graph OOD because a shifted record can remain inside known marginal support, and from attributed-graph anomaly detection because the output must retain a known-class prediction under known-only calibration.

2.4 INTACT-RECORD REASSIGNMENT AND IDENTIFIABILITY

Let π be a one-to-one permutation of selected test nodes, with $\pi(i) \neq i$. Intact-record reassignment replaces $\mathbf{r}_i$ by $\mathbf{r}_{\pi(i)}$ while leaving identifier i, its incident relations, and its target-excluded context constant. Donors are selected from a different known class in the primary stress test. Thus, a donor record remains semantically valid in isolation, but the record-position assignment need not be supported. The permutation preserves the complete record multiset exactly.

Proposition 1 (Graph-free score invariance)**.** *For any deterministic graph-free score f, the multiset $\{f(\mathbf{r}_i) : i \in \mathcal{S}\}$ equals $\{f(\mathbf{r}_{\pi(i)}) : i \in \mathcal{S}\}$ for every finite selected set $\mathcal{S}$ and every permutation π of $\mathcal{S}$. Consequently, no threshold or rank statistic based only on $f(\mathbf{r})$ can distinguish the original records from their one-to-one reassignment when the two populations are compared as multisets.*

With equal-sized populations and standard tie handling, this equality gives AUROC exactly 0.5 even when f varies across nodes. The result follows from permutation invariance of a multiset. Appendix C provides the formal argument. The proposition applies to deterministic record-only

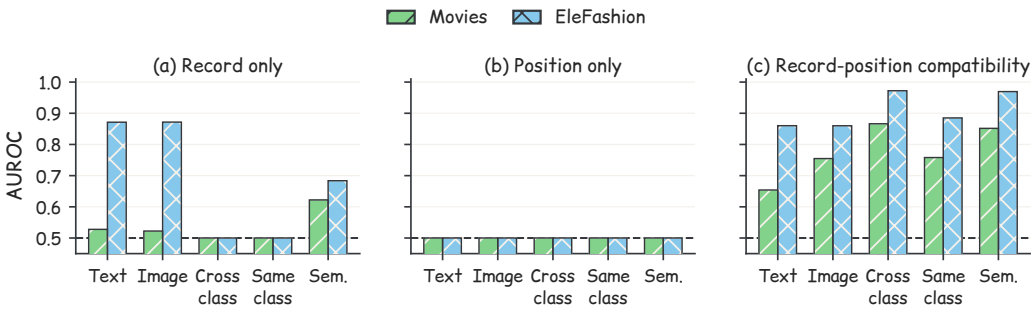

Figure 1: **Assignment-level OOD.** Holding recipients constant makes both marginal scores blind to intact-record reassignment, while record-position compatibility remains informative.

scores under one-to-one reassignment. It isolates a setting in which cross-entity relationships contain information that is unavailable to any record-only multimodal detector.

3 Empirical Study of Assignment OOD

Existing OOD detectors usually search for an unusual record, position, or complete node state. We instead hold the recipient position constant and test whether the OOD signal can disappear from both marginals while remaining in their assignment. This study establishes the information available to the verifier before we introduce PlaceGuard.

Position-controlled comparison. We use Movies and EleFashion because their released product relations are stored separately from the current text-image record. For each graph, we select 512 known recipient nodes with at least one connected known support node. Every condition uses the same recipient identifiers, labels, degrees, target-excluded contexts, and context posteriors. We then replace only (1) text, (2) image, (3) an intact record from another known class, (4) an intact record from the same known class, or (5) an unseen-class record. The first four donor sets contain only known nodes. Unseen records are used only in terminal evaluation. We compare a record-only confidence score, two position-only scores, modality-specific record-context cosine scores, and their multimodal average. All thresholds and rank transforms use clean known validation nodes.

Neither marginal identifies intact-record placement. Fig. 1(a) and (b) show the exact information boundary. Record-only scoring obtains AUROC 0.500 for cross-class and within-class placement because a one-to-one reassignment preserves its score multiset. Position-only degree and context uncertainty obtain AUROC 0.500 for all five conditions because the positions are unchanged. Both results follow exactly from withholding the record-position correspondence from the score.

The dependence remains identifiable. When the complete record is compared with its unchanged target-excluded context, Fig. 1(c) shows cross-class/within-class placement AUROC $0.866/0.758$ on Movies and $0.972/0.885$ on EleFashion. The same score obtains semantic-record AUROC 0.852 and 0.969, respectively. Thus, two familiar marginals do not imply a familiar assignment. Graph information becomes useful only through its correspondence with the attached record.

The signal follows the observed pairing. A separate six-graph control keeps the record and context collections and preserves coarse recipient degree groups, but breaks which record is paired with which context. As shown in Fig. 2, the observed pairing separates intact-record reassignment on every graph, while the degree-stratified null remains near chance and also lowers clean acceptance. The assignment signal therefore follows record-position correspondence beyond coarse degree structure alone.

Each modality sees only its own mismatch. The text-context score is exactly nondiscriminative for image replacement, while the image-context score is exactly nondiscriminative for text replacement. A detector that treats one modality as a proxy for the complete record therefore leaves a predictable blind spot. The complete decision must evaluate both modalities against the same position and preserve within-record validity.

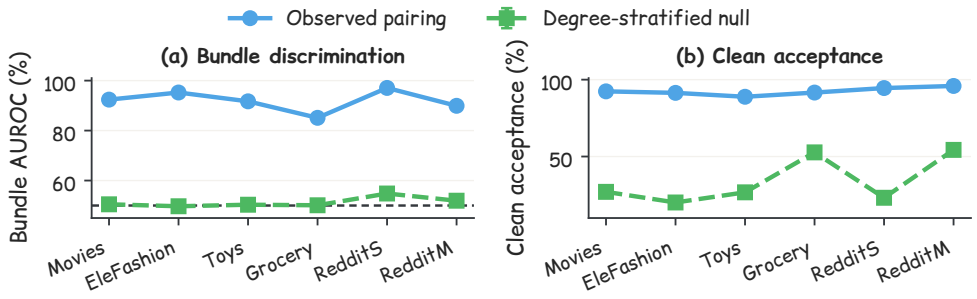

Figure 2: **Pairing specificity.** Observed pairing separates reassignment across six graphs, while a degree-stratified null remains near chance.

Taken together, these observations require one decision that checks complete record validity, record position fit, and context familiarity. Every accepted record also receives the prediction of the frozen known class classifier. This decision rule motivates PlaceGuard.

4 PlaceGuard Record-Placement Verification

4.1 Necessary checks without score compensation

PlaceGuard follows one principle. Admission requires every necessary check to pass. The first two checks ask whether the text and image form a valid record and whether that record fits the claimed position. Their product defines factorized assignment support. The third check asks whether the target excluded context remains familiar under the known classes. It is neutral over the ordinary clean region and gains influence only in the clean validation tail. This ordered information flow prevents one strong score from cancelling a failed check.

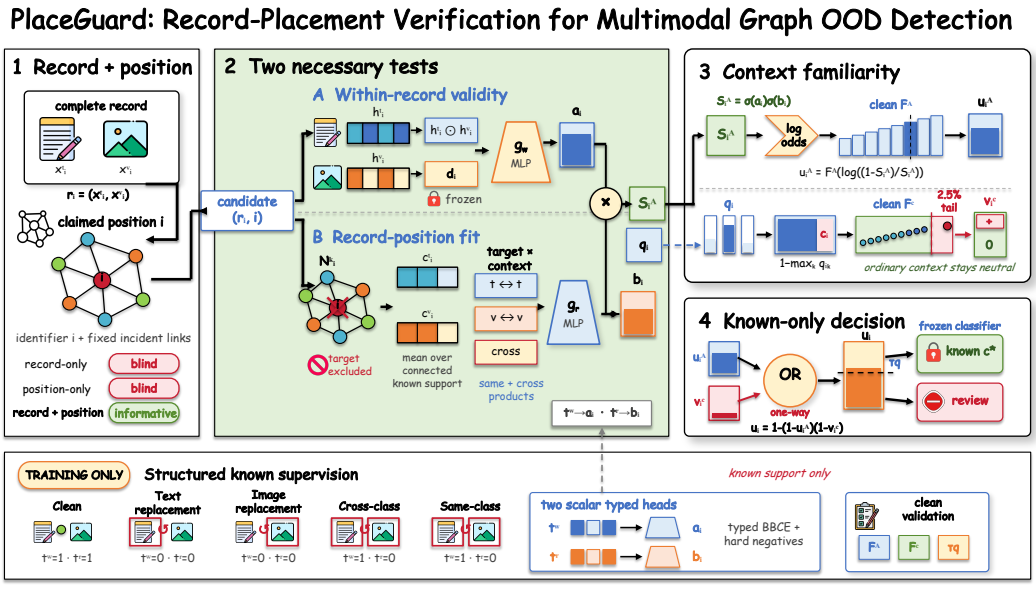

Figure 3: **PlaceGuard overview.** The verifier checks the record, its claimed position, and context familiarity.

4.2 FACTORIZED ASSIGNMENT SUPPORT

We first project each target modality into a shared hidden space. A separate projection summarizes the matching modality over connected known support nodes. For $m \in \{t, v\}$, the target and context representations are defined together as

$$\mathbf{h}_i^m = \phi_m(\mathbf{x}_i^m), \qquad \mathbf{c}_i^m = \frac{1}{|\mathcal{N}_i^{\mathrm{K}}|} \sum_{j \in \mathcal{N}_i^{\mathrm{K}}} \phi_m^c(\mathbf{x}_j^m), \qquad m \in \{t, v\}. \tag{1}$$

The neighborhood $\mathcal{N}_i^{\mathrm{K}}$ contains connected known support nodes and excludes the target record and every evaluation to evaluation path. PlaceGuard next forms one scalar for within record validity and one scalar for record position fit. The two parallel heads are

$$\begin{aligned} a_i &= w_a^\top g_w([\mathbf{h}_i^t, \mathbf{h}_i^v, \mathbf{h}_i^t \odot \mathbf{h}_i^v, \mathbf{d}_i]), \\ b_i &= w_b^\top g_r([\mathbf{h}_i^t, \mathbf{h}_i^v, \mathbf{h}_i^t \odot \mathbf{c}_i^t, \mathbf{h}_i^v \odot \mathbf{c}_i^v, \tfrac{1}{2}(\mathbf{h}_i^t \odot \mathbf{c}_i^v + \mathbf{h}_i^v \odot \mathbf{c}_i^t)]). \end{aligned} \tag{2}$$

Here, $\mathbf{d}_i$ is the discrepancy between frozen text and image predictions, and g_w and g_r are multilayer perceptrons. Context enters only b_i through its interaction with the target record. The frozen class posterior is reserved for the accepted class prediction.

Each head expresses a necessary check, so strong support from one head must not compensate for weak support from the other. PlaceGuard therefore defines factorized assignment support as

$$S_i^A = \sigma(a_i)\sigma(b_i) \le \min\{\sigma(a_i), \sigma(b_i)\}. \tag{3}$$

The product is high only when both checks pass. The inequality makes the noncompensatory role of the two factors explicit.

4.3 LEARNING FACTOR ROLES FROM KNOWN RECORDS

Training uses clean known nodes and four structured negative types drawn only from known support. They are text replacement, image replacement, cross class record replacement, and same class record replacement. An intact donor record remains valid as a record but does not fit the recipient position. A one modality replacement violates both factors. Accordingly, t_i^w is one for clean records and intact record replacements, while t_i^r is one only for clean records.

Let $\mathcal{S}_t$ and $\mathcal{S}_v$ contain the one modality replacements and let $\mathcal{S}_-$ contain every structured negative. Substituting the typed and hard terms into the complete objective gives

$$\begin{aligned} \mathcal{L} = \mathcal{L}_{\mathrm{member}}(S^A, t^r) + \frac{\lambda_f}{4}\Big[&\mathrm{BBCE}(a, t^w) + \mathrm{BBCE}(b, t^r) \\ &+ \mathbb{E}_{i \in \mathcal{S}_t \cup \mathcal{S}_v} \mathrm{softplus}(a_i) + \mathbb{E}_{i \in \mathcal{S}_-} \mathrm{softplus}(b_i)\Big]. \end{aligned} \tag{4}$$

Here, balanced binary cross entropy gives equal weight to positive and negative examples. $\mathcal{L}_{\mathrm{member}}$ gives equal weight to clean and structured records. The hard term suppresses a surviving record logit for one modality changes and a surviving position logit for every structured change. The known class classifier remains frozen throughout training. Structured validation selects the checkpoint by worst-type detection, AUROC, and retained known-class accuracy. Validation negatives are disjoint from training negatives. Fitting, calibration, and checkpoint selection use known-class records only.

4.4 ONE-WAY CONTEXT CHECK AND FINAL DECISION

Let $\widehat{F}_A$ be the empirical clean validation distribution of assignment log odds. Let q_{ik} be the known class posterior computed from target-excluded context and let $\widehat{F}_C$ be the clean validation distribution of its uncertainty. The two calibrated violation scores are

$$u_i^A = \widehat{F}_A\left(\log \frac{1 - S_i^A}{S_i^A}\right), \qquad v_i^C = \left[\frac{\widehat{F}_C(1 - \max_k q_{ik}) - 0.975}{0.025}\right]_0^1. \tag{5}$$

The clipping operator limits a value to $[0, 1]$. Ordinary clean contexts have $v_i^C = 0$ and leave assignment unknownness unchanged. Beyond that region, PlaceGuard combines the assignment and context violations with the one-way union

$$u_i = 1 - (1 - u_i^A)(1 - v_i^C) \ge u_i^A. \tag{6}$$

The inequality guarantees that a context result can increase rejection but cannot make a poor assignment acceptable. PlaceGuard returns rejection when $u_i > \tau_q$ and otherwise returns $\arg\max_k \pi_{ik}$ from the frozen classifier. The threshold τ_q is estimated from clean known validation scores under the same known-only information policy used for training.

4.5 Computational complexity

Let h be the hidden width, $d_i = |\mathcal{N}_i^{\mathrm{K}}|$, C be the number of known classes, and n_A and n_C be the two calibration set sizes. With cached support projections and sorted calibration scores, verification of one target has time and working memory

$$T_i = \mathcal{O}\big(d_i(h + C) + h^2 + \log n_A + \log n_C\big), \qquad M_i = \mathcal{O}(h + C). \tag{7}$$

The neighborhood term is linear in the incident known support edges and the quadratic term comes from the two small scalar heads. One training epoch costs $\mathcal{O}(M_{\mathrm{tr}}(h + C) + N_{\mathrm{tr}}h^2)$ for N_{tr} structured records and M_{tr} support incidences. Independent target verification uses cached support representations and therefore avoids full graph propagation for each arriving record.

5 Experiments

5.1 Experimental setup

Datasets. We evaluate Movies, EleFashion, Toys, Grocery, RedditS, and RedditM from public MAG benchmarks (Yan et al., 2025; Zhu et al., 2025; Wan et al., 2026). The four product graphs contain text, images, and pooled product or co-purchase relations. RedditS and RedditM connect image-caption posts through the released same-author relation. Fitting and calibration expose only known classes, while disjoint classes serve as semantic unknowns.

Protocols. We first report conventional semantic node OOD. Clean known test nodes and held-out-class nodes are scored at their released positions. We use exactly the semantic records and clean recipients from the fixed-recipient evaluation. Each protocol therefore contains the same paired clean and semantic records. In the conventional view, every record remains at its released position. In the fixed-recipient view, the semantic record is assigned to the same known recipient. We also test text, image, cross-class placement, and within-class placement changes while keeping the claimed identifier and its target-excluded context constant.

Both protocol views use deterministic class-holdout seeds 42, 43, and 44 on all six graphs. Each seed separates known support, structured validation, clean calibration, clean test, and unseen-class populations. Every retained recipient has a same-class non-self donor and a known support neighbor. Context uses support-to-target edges, excludes the target record, and removes evaluation-to-evaluation paths. Training donors come only from known support. Repeated-run percentages are mean $\pm$ sample standard deviation over the three splits. All three splits were selected before this batch and all three outcomes are retained without changing the method or selection rule.

Comparisons. Both main tables contain the same eight methods in the same order. They include GNNSafe (Wu et al., 2023), GRASP (Ma et al., 2024), NODESAFE (Yang et al., 2024), TopoOOD (Bao et al., 2024), TNT-OOD (Wang et al., 2025a), DPU (Li et al., 2025), TIDE (Wang et al., 2026), and PlaceGuard. TIDE is the directly compatible ICML 2026 node OOD method for this node-level decision object and information boundary.

All main-table methods use known-class training data without external outlier exposure. PlaceGuard constructs its mismatched record-position pairs from known training records, with held-out classes and evaluation records reserved for terminal testing. All baseline implementations are adapted to the shared protocol and use the same splits, frozen features, target-excluded support context, three-setting selection budget, metrics, recipients, donors, and terminal evaluation. Role-matched controls include maximum softmax confidence (Hendrycks & Gimpel, 2017), prototype and component compatibility, transparent record context scores, and factorized assignment support that forms the learned core of PlaceGuard.

Implementation and validation. All learned rows receive the same frozen text/image features and target-excluded support context. Three predeclared configurations are selected by structured

known-only validation. AdamW trains for at most 120 epochs with early stopping. Saved artifacts include the checkpoint, calibration, per-node scores, predictions, donor assignments, timing, and environment. Appendix E gives the full boundary, objective, and grid. The encoder and support-fitted PCA remain unchanged across seeds, isolating placement from representation pretraining.

Metrics. Unknown examples are positive and larger scores indicate greater unknownness. We report area under the receiver operating characteristic curve (AUROC), average precision (AP), and the clean false positive rate at 95% unknown recall (FPR95). Appendix B reports per graph boundaries, complete locus metrics, operating points, and known class accuracy. Table 1 reports per-graph semantic OOD AUROC. The fixed-recipient comparison reports all five shift AUROCs and the weakest-locus AUROC, AP, and FPR95. In either table, 50.0% AUROC is chance for clean examples versus one shifted condition. A weakest-locus value near chance identifies at least one shift that is indistinguishable from clean records, even when the same method performs normally on released-position semantic OOD.

5.2 Main comparison

Table 1: **Conventional semantic OOD.** AUROC percentages over 3 splits.

Method	Movies	EleF.	Toys	Grocery	RedditS	RedditM	Avg.
GNNSafe (ICLR'23)	62.9 ± 7.0	89.7 ± 2.6	84.0 ± 5.4	79.0 ± 3.6	97.7 ± 1.6	90.9 ± 1.0	84.0 ± 1.4
GRASP (NeurIPS'24)	76.3 ± 3.3	90.8 ± 4.7	91.5 ± 3.7	84.4 ± 4.5	97.3 ± 1.3	97.0 ± 0.6	89.9 ± 1.9
NODESAFE (ICML'24)	71.4 ± 0.7	89.5 ± 7.3	87.9 ± 5.3	81.9 ± 4.8	97.1 ± 1.3	92.4 ± 0.9	86.7 ± 1.4
TopoOOD (ICML'24)	73.5 ± 5.0	91.8 ± 3.2	91.7 ± 5.0	80.8 ± 6.0	96.0 ± 2.5	97.6 ± 0.7	88.7 ± 2.3
TNT-OOD (EMNLP'25)	76.7 ± 2.3	92.2 ± 3.6	94.1 ± 2.2	86.6 ± 1.6	97.2 ± 0.6	97.2 ± 0.6	90.9 ± 1.2
DPU (CVPR'25)	55.4 ± 3.5	69.0 ± 10.7	73.0 ± 7.1	74.2 ± 3.0	89.0 ± 3.4	84.0 ± 0.6	74.1 ± 0.6
TIDE (ICML'26)	77.0 ± 2.5	92.7 ± 3.9	94.4 ± 1.6	**90.2** ± 1.6	97.4 ± 0.4	97.8 ± 0.5	90.5 ± 1.1
PlaceGuard (ours)	**77.2** ± 0.7	**92.9** ± 4.4	**94.6** ± 1.3	87.2 ± 2.0	**98.1** ± 1.2	**98.0** ± 0.4	**91.3** ± 0.9

Released positions provide the conventional semantic OOD reference. Table 1 compares clean known nodes with held-out-class nodes at their own released positions. It uses the same methods, semantic records, clean records, and class splits as the fixed-recipient table. All eight methods attain conventional macro AUROC between 74.1 and 92.5, establishing their ordinary semantic OOD behavior before placement changes. TopoOOD reaches 88.7 after its released ID-positive Dirichlet score is converted to the common OOD-positive convention specified by the official evaluation code.

Table 2: **Fixed-recipient OOD.** Macro percentages with mean and sample standard deviation over 3 splits.

Method	Record OOD			Placement OOD		Weakest locus		
	Text	Image	Semantic	Cross class	Same class	AUROC	AP	FPR95
GNNSafe (ICLR'23)	55.1 ± 0.5	53.4 ± 0.5	59.6 ± 0.8	48.0 ± 0.1	49.4 ± 0.0	48.0 ± 0.1	46.7 ± 0.1	94.4 ± 0.2
GRASP (NeurIPS'24)	50.2 ± 0.0	50.2 ± 0.0	50.5 ± 0.1	50.0 ± 0.0	50.0 ± 0.0	49.9 ± 0.0	50.0 ± 0.0	95.9 ± 0.2
NODESAFE (ICML'24)	50.3 ± 0.1	50.8 ± 0.1	50.8 ± 0.1	50.9 ± 0.1	50.2 ± 0.0	50.2 ± 0.0	50.3 ± 0.0	94.9 ± 0.0
TopoOOD (ICML'24)	48.9 ± 0.0	47.5 ± 0.1	46.9 ± 0.1	46.6 ± 0.2	49.1 ± 0.1	46.6 ± 0.2	47.0 ± 0.0	96.7 ± 0.1
TNT-OOD (EMNLP'25)	50.7 ± 0.0	51.1 ± 0.0	51.5 ± 0.1	51.5 ± 0.0	50.4 ± 0.0	50.3 ± 0.0	50.4 ± 0.1	94.9 ± 0.1
DPU (CVPR'25)	69.9 ± 0.7	69.9 ± 0.6	74.1 ± 0.6	50.0 ± 0.0	50.0 ± 0.0	50.0 ± 0.0	50.0 ± 0.0	95.1 ± 0.0
TIDE (ICML'26)	51.1 ± 0.1	51.7 ± 0.1	52.4 ± 0.1	52.5 ± 0.1	50.4 ± 0.0	50.4 ± 0.0	50.5 ± 0.1	94.9 ± 0.1
PlaceGuard (ours)	**91.2** ± 0.6	**94.9** ± 0.4	**90.8** ± 0.7	**92.6** ± 0.2	**78.0** ± 0.8	**78.0** ± 0.8	**76.7** ± 0.9	**73.1** ± 1.1

Changing the verified object exposes an assignment blind spot. Table 2 places familiar text, image, and semantic record shifts next to cross-class and same-class placement shifts under one fixed-recipient protocol. PlaceGuard leads 5/5 shift-type AUROC columns and reaches macro weakest-locus AUROC 0.780 versus 0.504 for the strongest compared method. The weakest-locus value is the minimum after scoring all five shifts. PlaceGuard's 0.780 therefore means that every shift remains well above chance, whereas each compared method has at least one near-chance condition. PlaceGuard also reaches 91.3 on the released-position reference. DPU obtains 69.9, 69.9, and 74.1 AUROC on text, image, and semantic record shifts, followed by 50.0 and 50.0 on the two placement shifts. Proposition 1 explains the exact chance placement result for record-only scoring. Graph and text-topology methods recompute their scores and still approach chance on at least one shift under

this protocol. Together, the two main tables separate semantic OOD recognition at a released position from verification of a familiar record at a claimed known position.

Placement remains detectable after removing class conflict. Across six graphs, PlaceGuard reaches macro cross-class and within-class placement AUROC 0.926 and 0.780. Within-class reassignment removes the class-label conflict and therefore requires finer assignment information. Semantic-record AUROC is 0.908. The remaining variation across RedditM splits is retained in the reported standard deviation.

Performance varies across graphs and operating regimes. Worst-locus AP is 0.767 and FPR95 is 0.731. Movies has the lowest conventional AUROC at 0.772, and RedditM has the lowest fixed-recipient AUROC at 0.676. The per-graph and operating-point results in Appendix B show where thresholded admission is most likely to require review.

5.3 Core verification choices

Table 3: **Verification structure.** Macro AUROC with mean and sample standard deviation over 3 splits.

Verification rule	Conv.	Text	Image	Cross	Same	Semantic	Worst
Class-specific support product	$87.0_{\pm 0.9}$	$91.0_{\pm 0.8}$	$94.5_{\pm 1.3}$	$89.1_{\pm 4.1}$	$75.7_{\pm 2.9}$	$90.5_{\pm 1.7}$	$75.7_{\pm 2.9}$
Factorized assignment support	$87.0_{\pm 0.5}$	$\mathbf{92.4}_{\pm 0.4}$	$\mathbf{95.3}_{\pm 0.2}$	$91.8_{\pm 0.4}$	$77.1_{\pm 0.8}$	$89.9_{\pm 0.8}$	$77.1_{\pm 0.8}$
PlaceGuard (ours)	$\mathbf{91.3}_{\pm 0.9}$	$91.2_{\pm 0.6}$	$94.9_{\pm 0.4}$	$\mathbf{92.6}_{\pm 0.2}$	$\mathbf{78.0}_{\pm 0.8}$	$\mathbf{90.8}_{\pm 0.7}$	$\mathbf{78.0}_{\pm 0.8}$

Table 3 isolates the verification structure. Factorized assignment support reaches 0.771 and forms the central record-placement mechanism. Adding the one-way context check raises the weakest-locus result to 0.780 and the conventional result from 0.870 to 0.913. This gain gives context a complementary role for semantic unknowns at released positions. The context path is neutral across the ordinary clean region and activates only in the clean calibration tail, so it adds rejection without compensating for a failed assignment factor. This ablation supports distinct decision roles for record validity, position fit, and context familiarity. The same frozen classifier supplies the accepted class prediction for every verification rule.

The factorized row demonstrates that assignment support alone is already a strong fixed-recipient verifier. Open-world admission also includes semantic unknowns at their released positions. The one-way composition adds this complementary signal while preserving assignment unknownness as a lower bound on the final score. PlaceGuard therefore uses the factorized verifier as its core and extends it with a noncompensatory path for conventional semantic coverage.

6 Conclusion

Open-world admission must verify whether a multimodal record is familiar and whether it belongs at the graph position it claims. Our controlled reassignment study shows that record-only and position-only scores can both be exactly nondiscriminative even when record-position compatibility remains informative. PlaceGuard turns this observation into separate checks for within-record validity, record-position fit, and context familiarity. Factorized assignment support requires both record checks to pass, and the one-way context path adds rejection without rescuing a failed assignment. Across six graphs and three class splits, PlaceGuard reaches conventional semantic OOD AUROC 0.913, cross-class placement AUROC 0.926, same-class placement AUROC 0.780, and the strongest macro weakest-locus result under the shared protocol. These results identify record-position assignment as a distinct open-world decision and show how semantic OOD detection can extend to records submitted at persistent relational identities. The resulting verifier returns a known class for supported assignments and routes unsupported assignments for review before they enter the relational store. Controlled reassignment establishes assignment-level observability, and the current reliability boundary supports a review-based admission workflow.

AI USE STATEMENT

Generative AI tools assisted literature organization, code development, experiment orchestration, and language editing. The authors checked citations, code, protocols, metrics, and claims against saved artifacts and take responsibility for the paper.

REPRODUCIBILITY STATEMENT

The problem formulation, method, and appendix specify the information boundary, equations, calibration, training, data construction, baselines, and hyperparameters. Released artifacts will include splits, donor assignments, hashes, scores, calibration, timing, configurations, and metric reconstruction. Tables are generated from them. Anonymous code will accompany the submission.

REFERENCES

Reza Averly and Wei-Lun Chao. Unified out-of-distribution detection: A model-specific perspective. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1453–1463, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Averly_Unified_Out-Of-Distribution_Detection_A_Model-Specific_Perspective_ICCV_2023_paper.html.

Tianyi Bao, Qitian Wu, Zetian Jiang, Yiting Chen, Jiawei Sun, and Junchi Yan. Graph out-of-distribution detection goes neighborhood shaping. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 2923–2943. PMLR, 2024. URL https://proceedings.mlr.press/v235/bao24c.html.

Tingyi Cai, Yunliang Jiang, Yixin Liu, Ming Li, Changqin Huang, and Shirui Pan. Out-of-distribution detection on graphs: A survey. *arXiv preprint arXiv:2502.08105*, 2025.

Ming Chen, Zhewei Wei, Zengfeng Huang, Bolin Ding, and Yaliang Li. Simple and deep graph convolutional networks. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1725–1735. PMLR, 2020.

Kaize Ding, Jundong Li, Rohit Bhanushali, and Huan Liu. Deep anomaly detection on attributed networks. In *Proceedings of the 2019 SIAM International Conference on Data Mining*, pp. 594–602. SIAM, 2019.

Hao Dong, Yue Zhao, Eleni Chatzi, and Olga Fink. MultiOOD: Scaling out-of-distribution detection for multiple modalities. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/e969483cbae5a24df038fb5847f5b8b0-Abstract-Conference.html.

Yasha Ektefaie, George Dasoulas, Ayush Noori, Maha R. Farhat, and Marinka Zitnik. Multimodal learning with graphs. *Nat. Mac. Intell.*, 5(4):340–350, 2023.

Dongzhe Fan, Yi Fang, Jiajin Liu, Djellel Difallah, and Qiaoyu Tan. Mlaga: Multimodal large language and graph assistant. *CoRR*, abs/2506.02568, 2025.

Haoyi Fan, Fengbin Zhang, and Zuoyong Li. Anomalydae: Dual autoencoder for anomaly detection on attributed networks. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5685–5689, 2020. doi: 10.1109/ICASSP40776.2020.9053387.

Yi Fang, Bowen Jin, Jiacheng Shen, Sirui Ding, Qiaoyu Tan, and Jiawei Han. GRAPHGPT-O: synergistic multimodal comprehension and generation on graphs. In *CVPR*, pp. 19467–19476. Computer Vision Foundation / IEEE, 2025.

Vojtech Franc, Jakub Paplham, and Daniel Prusa. SCOD: From heuristics to theory. In *European Conference on Computer Vision*, 2024. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/11303_ECCV_2024_paper.php.

Shurui Gui, Xiner Li, Limei Wang, and Shuiwang Ji. Good: A graph out-of-distribution benchmark. *Advances in Neural Information Processing Systems*, 35:2059–2073, 2022.

Zhaochen Guo, Zhixiang Shen, Xuanting Xie, Liangjian Wen, and Zhao Kang. Disentangling homophily and heterophily in multimodal graph clustering. *CoRR*, abs/2507.15253, 2025.

Zhiqiang Guo, Jianjun Li, Guohui Li, Chaoyang Wang, Si Shi, and Bin Ruan. Lgmrec: Local and global graph learning for multimodal recommendation. In *AAAI*, pp. 8454–8462. AAAI Press, 2024.

William L. Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *NIPS*, pp. 1024–1034, 2017.

Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *SIGIR*, pp. 639–648. ACM, 2020.

Yufei He, Yuan Sui, Xiaoxin He, Yue Liu, Yifei Sun, and Bryan Hooi. Unigraph2: Learning a unified embedding space to bind multimodal graphs. In *WWW*, 2025. doi: 10.1145/3696410.3714818. URL https://arxiv.org/abs/2502.00806.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Hkg4TI9xl.

Zhenyu Hou, Yufei He, Yukuo Cen, Xiao Liu, Yuxiao Dong, Evgeny Kharlamov, and Jie Tang. Graphmae2: A decoding-enhanced masked self-supervised graph learner. In *WWW*, pp. 737–746. ACM, 2023.

Jun Hu, Bryan Hooi, Bingsheng He, and Yinwei Wei. Modality-independent graph neural networks with global transformers for multimodal recommendation. In *AAAI*, pp. 11790–11798. AAAI Press, 2025.

Jun Hu, Yufei He, Yuan Li, Bryan Hooi, and Bingsheng He. NTSFormer: A self-teaching graph transformer for multimodal isolated cold-start node classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 14856–14864, 2026. doi: 10.1609/aaai.v40i17.38506.

Ming Jin, Yixin Liu, Yu Zheng, Lianhua Chi, Yuan-Fang Li, and Shirui Pan. Anemone: Graph anomaly detection with multi-scale contrastive learning. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pp. 3122–3126, 2021.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=SJU4ayYgl.

Guohao Li, Matthias Müller, Bernard Ghanem, and Vladlen Koltun. Training graph neural networks with 1000 layers. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6437–6449. PMLR, 2021.

Jundong Li, Harsh Dani, Xia Hu, and Huan Liu. Radar: Residual analysis for anomaly detection in attributed networks. In *IJCAI*, pp. 2152–2158, 2017.

Shawn Li, Huixian Gong, Hao Dong, Tiankai Yang, Zhengzhong Tu, and Yue Zhao. DPU: Dynamic prototype updating for multimodal out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10193–10202, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Li_DPU_Dynamic_Prototype_Updating_for_Multimodal_Out-of-Distribution_Detection_CVPR_2025_paper.html.

Xunkai Li, Zhengyu Wu, Zekai Chen, Henan Sun, Daohan Su, Guang Zeng, Hongchao Qin, Rong-Hua Li, and Guoren Wang. Lion: A clifford neural paradigm for multimodal-attributed graph learning, 2026a. URL https://arxiv.org/abs/2601.21453.

Yuan Li, Jun Hu, Jiaxin Jiang, Bryan Hooi, and Bingsheng He. Robust multimodal recommendation via graph retrieval-enhanced modality completion. In *SIGIR*, 2026b. doi: 10.1145/3805712.3809607. URL https://arxiv.org/abs/2605.00670.

Jiajin Liu, Dongzhe Fan, Jiacheng Shen, Chuanhao Ji, Daochen Zha, and Qiaoyu Tan. Graph-mllm: Harnessing multimodal large language models for multimodal graph learning. *CoRR*, abs/2506.10282, 2025.

Kay Liu, Yingtong Dou, Yue Zhao, Xueying Ding, Xiyang Hu, Ruitong Zhang, Kaize Ding, Canyu Chen, Hao Peng, Kai Shu, George H. Chen, Zhihao Jia, and Philip S. Yu. Pygod: A python library for graph outlier detection. *arXiv preprint arXiv:2204.12095*, 2022a.

Kay Liu, Yingtong Dou, Yue Zhao, Xueying Ding, Xiyang Hu, Ruitong Zhang, Kaize Ding, Canyu Chen, Hao Peng, Kai Shu, Lichao Sun, Jundong Li, George H. Chen, Zhihao Jia, and Philip S. Yu. Benchmarking node outlier detection on graphs. *arXiv preprint arXiv:2206.10071*, 2022b.

Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21464–21475, 2020.

Wenrui Liu, Hong Chang, Ruibing Hou, Shiguang Shan, and Xilin Chen. Component-based out-of-distribution detection. *CoRR*, abs/2604.21546, 2026. URL https://arxiv.org/abs/2604.21546.

Longfei Ma, Yiyou Sun, Kaize Ding, Zemin Liu, and Fei Wu. Revisiting score propagation in graph out-of-distribution detection. *Advances in Neural Information Processing Systems*, 37:4341–4373, 2024.

Harikrishna Narasimhan, Aditya Krishna Menon, Wittawat Jitkrittum, and Sanjiv Kumar. Plugin estimators for selective classification with out-of-distribution detection. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/d78d68cae595fabadd187b583ee8708e-Abstract-Conference.html.

Ciyuan Peng, Jiayuan He, and Feng Xia. Learning on multimodal graphs: A survey. *CoRR*, abs/2402.05322, 2024.

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

Maximilian Stadler, Bertrand Charpentier, Simon Geisler, Daniel Zügner, and Stephan Günnemann. Graph posterior network: Bayesian predictive uncertainty for node classification. *Advances in Neural Information Processing Systems*, 34:18033–18048, 2021.

Daohan Su, Xiong Yongfu, Yi Liu, Hao Liu, Xunkai Li, Yinlin Zhu, Hongchao Qin, Rong-Hua Li, and Guoren Wang. Campa: Efficient and aligned multimodal graph learning via decoupled propagation and aggregation, 2026. URL https://arxiv.org/abs/2605.11468.

Zhulin Tao, Yinwei Wei, Xiang Wang, Xiangnan He, Xianglin Huang, and Tat-Seng Chua. MGAT: multimodal graph attention network for recommendation. *Inf. Process. Manag.*, 57(5):102277, 2020.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.

Petar Veličković, William Fedus, William L. Hamilton, Pietro Liò, Yoshua Bengio, and R. Devon Hjelm. Deep graph infomax. In *International Conference on Learning Representations*. OpenReview.net, 2019. URL https://openreview.net/forum?id=rklz9iAcKQ.

Chenxi Wan, Xunkai Li, Yilong Zuo, Haokun Deng, Sihan Li, Bowen Fan, Hongchao Qin, Rong-Hua Li, and Guoren Wang. OpenMAG: A comprehensive benchmark for multimodal-attributed graph. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/64650.

Danny Wang, Ruihong Qiu, Guangdong Bai, and Zi Huang. Text meets topology: Rethinking out-of-distribution detection in text-rich networks. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025a.

Danny Wang, Ruihong Qiu, Guangdong Bai, and Zi Huang. GOLD: Graph out-of-distribution detection via implicit adversarial latent generation. In *International Conference on Learning Representations*. OpenReview.net, 2025b. URL https://openreview.net/forum?id=y5einmJ0Yx.

Danny Wang, Ruihong Qiu, and Zi Huang. What information matters? graph out-of-distribution detection via tri-component information decomposition. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Yang Wang, Xinye Wang, Chengxin He, Xiaocong Chen, Zhaohang Luo, Lei Duan, and Jie Zuo. Community-guided contrastive learning with anomaly-aware reconstruction for anomaly detection on attributed networks. In *Database Systems for Advanced Applications - 29th International Conference*, pp. 199–209, 2024.

Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. MMGCN: multi-modal graph convolution network for personalized recommendation of micro-video. In *ACM Multimedia*, pp. 1437–1445. ACM, 2019.

Man Wu, Shirui Pan, and Xingquan Zhu. Openwgl: Open-world graph learning. In *2020 IEEE international conference on data mining (icdm)*, pp. 681–690. IEEE, 2020.

Qitian Wu, Yiting Chen, Chenxiao Yang, and Junchi Yan. Energy-based out-of-distribution detection for graph neural networks. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=zoz7Ze4STUL.

Zhengyu Wu, Xu Wang, Hongchao Qin, Xunkai Li, Guang Zeng, Rong-Hua Li, and Guoren Wang. Multimodal graph negative learning, 2026a. URL https://arxiv.org/abs/2606.12863.

Zhengyu Wu, Xu Wang, Hongchao Qin, Xunkai Li, Guang Zeng, Rong-Hua Li, and Guoren Wang. Smgfm: Spectral multimodal graph pretraining for multimodal-attributed graphs, 2026b. URL https://arxiv.org/abs/2606.12867.

Guoxuan Xia and Christos-Savvas Bouganis. Augmenting the softmax with additional confidence scores for improved selective classification with out-of-distribution data. *International Journal of Computer Vision*, 132:3714–3752, 2024. doi: 10.1007/s11263-024-02029-3.

Haoyan Xu, Zhengtao Yao, Yushun Dong, Ziyi Wang, Ryan A. Rossi, Mengyuan Li, and Yue Zhao. Few-shot graph out-of-distribution detection with LLMs. In *Machine Learning and Knowledge Discovery in Databases. Research Track*, volume 16016 of *Lecture Notes in Computer Science*, pp. 307–324. Springer Nature Switzerland, 2025. doi: 10.1007/978-3-032-06078-5_18. URL https://doi.org/10.1007/978-3-032-06078-5_18.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *ICLR*. OpenReview.net, 2019.

Zhiming Xu, Xiao Huang, Yue Zhao, Yushun Dong, and Jundong Li. Contrastive attributed network anomaly detection with data augmentation. In *Pacific-Asian Conference on Knowledge Discovery and Data Mining (PAKDD)*, 2022.

Hao Yan, Chaozhuo Li, Zhigang Yu, Jun Yin, Ruochen Liu, Peiyan Zhang, Weihao Han, Mingzheng Li, Zhengxin Zeng, Hao Sun, Weiwei Deng, Feng Sun, Qi Zhang, and Senzhang Wang. When graph meets multimodal: Benchmarking on multimodal attributed graphs learning. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2025. doi: 10.1145/3711896.3737404.

Yujun Yan, Milad Hashemi, Kevin Swersky, Yaoqing Yang, and Danai Koutra. Two sides of the same coin: Heterophily and oversmoothing in graph convolutional neural networks. In *2022 IEEE International Conference on Data Mining*, pp. 1287–1292. IEEE, 2022. doi: 10.1109/ICDM54844.2022.00169. URL https://doi.org/10.1109/ICDM54844.2022.00169.

Jingkang Yang, Kaiyang Zhou, and Ziwei Liu. Full-spectrum out-of-distribution detection. *International Journal of Computer Vision*, 131(10):2607–2622, 2023. doi: 10.1007/s11263-023-01811-z.

Shenzhi Yang, Bin Liang, An Liu, Lin Gui, Xingkai Yao, and Xiaofang Zhang. Bounded and uniform energy-based out-of-distribution detection for graphs. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 56216–56234. PMLR, 2024. URL https://proceedings.mlr.press/v235/yang24n.html.

Rex Ying, Ruining He, Kaifeng Chen, Pong Eksombatchai, William L. Hamilton, and Jure Leskovec. Graph convolutional neural networks for web-scale recommender systems. In *KDD*, pp. 974–983. ACM, 2018.

Qin Zhang, Qincai Li, Xiaojun Chen, Peng Zhang, Shirui Pan, Philippe Fournier-Viger, and Joshua Zhexue Huang. A dynamic variational framework for open-world node classification in structured sequences. In *2022 IEEE International Conference on Data Mining (ICDM)*, pp. 703–712. IEEE, 2022.

Qingyang Zhang, Qiuxuan Feng, Joey Tianyi Zhou, Yatao Bian, Qinghua Hu, and Changqing Zhang. The best of both worlds: On the dilemma of out-of-distribution detection. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/8143b8c73073a9a23b9c18e400066471-Abstract-Conference.html.

Sirui Zhang, Xu Wang, and Zhengyu Wu. Context-aware modality-topology co-alignment for multimodal attributed graphs, 2026. URL https://arxiv.org/abs/2606.14172.

Haoran Zheng, Renchi Yang, Hongtao Wang, and Jianliang Xu. Cross-contrastive clustering for multimodal attributed graphs with dual graph filtering. In *KDD '26: Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, volume 1, pp. 2020–2030. Association for Computing Machinery, 2026a. doi: 10.1145/3770854.3780236. URL https://doi.org/10.1145/3770854.3780236.

Xin Zheng, Yi Wang, Yixin Liu, Ming Li, Miao Zhang, Di Jin, Philip S. Yu, and Shirui Pan. Graph neural networks for graphs with heterophily: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 38(7):4385–4404, 2026b. doi: 10.1109/TKDE.2026.3680353. URL https://doi.org/10.1109/TKDE.2026.3680353.

Jing Zhu, Yuhang Zhou, Shengyi Qian, Zhongmou He, Tong Zhao, Neil Shah, and Danai Koutra. Mosaic of modalities: A comprehensive benchmark for multimodal graph learning. In *CVPR*, pp. 14215–14224. Computer Vision Foundation / IEEE, 2025.

Yinlin Zhu, Xunkai Li, Di Wu, Wang Luo, Miao Hu, and Guocong Quan. Tmte: Effective multimodal graph learning with task-aware modality and topology co-evolution, 2026. URL https://arxiv.org/abs/2603.27723.

Yilong Zuo, Xunkai Li, Zhihan Zhang, Ronghua Li, and Guoren Wang. Rolemag: Learning neighbor roles in multimodal graphs, 2026. URL https://arxiv.org/abs/2604.12271.

A Related Works

Multimodal graph learning. MAG learning combines node modalities with cross-entity structure, and public benchmarks cover product, social, art, and media graphs (Ektefaie et al., 2023; Peng et al., 2024; Yan et al., 2025; Zhu et al., 2025; Wan et al., 2026). Existing methods study recommendation, modality-specific propagation, topology interaction, completion, and foundation models (Wei et al., 2019; Tao et al., 2020; Guo et al., 2024; Hu et al., 2025; Zheng et al., 2026a; Guo et al., 2025; He et al., 2025; Zhu et al., 2026; Zhang et al., 2026; Su et al., 2026; Zuo et al., 2026; Li et al., 2026a;b; Hu et al., 2026; Liu et al., 2025; Fang et al., 2025; Fan et al., 2025; Wu et al., 2026a;b). These lines learn or complete node representations. Our task instead verifies whether an intact record belongs to its claimed relationship context before returning a known class.

Out-of-distribution and open-world detection. Entity-level OOD scores include softmax confidence, energy, full-spectrum and selective variants, and multimodal discrepancy or prototype models (Hendrycks & Gimpel, 2017; Liu et al., 2020; Yang et al., 2023; Averly & Chao, 2023; Zhang et al., 2024; Narasimhan et al., 2024; Franc et al., 2024; Xia & Bouganis, 2024; Dong et al., 2024; Li et al., 2025). Graph OOD methods further use uncertainty, topology, propagation, information decomposition, text-topology interaction, generated outliers, or transfer (Gui et al., 2022; Stadler et al., 2021; Wu et al., 2020; Zhang et al., 2022; Wu et al., 2023; Yang et al., 2024; Bao et al., 2024; Wang et al., 2026; 2025a; Ma et al., 2024; Wang et al., 2025b; Xu et al., 2025; Cai et al., 2025). Our formulation changes the verified object. The record marginal remains unchanged, and the missing signal is whether that record belongs to the observed graph position.

Graph anomaly and compositional detection. Attributed-graph anomaly detectors reconstruct or contrast attributes and structure, model local agreement, or learn neighborhood statistics (Ding et al., 2019; Fan et al., 2020; Xu et al., 2022; Jin et al., 2021; Li et al., 2017; Roy et al., 2024; Wang et al., 2024; Liu et al., 2022a;b). CoOD scores local component shifts and cross-component consistency (Liu et al., 2026), making composition the closest conceptual boundary. Our task instead requires an intact text-image record and its target-specific cross-entity relations to support the same known-class decision. The implementation uses standard convolution, attention, inductive aggregation, expressive message passing, self-supervision, heterophilous learning, and recommendation-graph primitives (Kipf & Welling, 2017; Veličković et al., 2018; Hamilton et al., 2017; Xu et al., 2019; Chen et al., 2020; Li et al., 2021; Veličković et al., 2019; Hou et al., 2023; Yan et al., 2022; Zheng et al., 2026b; Ying et al., 2018; He et al., 2020). The open-world information boundary and assignment-verification decision define the contribution built from these established operators.

B More Experiments

B.1 Known-class prediction before rejection

Table 4 reports accuracy before the reject rule is applied. The frozen classifier supplies the PlaceGuard prediction. The rejection score therefore changes only acceptance and leaves the predicted known class unchanged.

Table 4: **Known-class accuracy before rejection.** Values are mean ± sample standard deviation.

Dataset	Frozen classifier	PlaceGuard before rejection
Movies	**51.6** ± 2.4	**51.6** ± 2.4
EleFashion	**79.4** ± 4.0	**79.4** ± 4.0
Toys	**68.6** ± 2.5	**68.6** ± 2.5
Grocery	**65.8** ± 2.2	**65.8** ± 2.2
RedditS	**91.6** ± 1.0	**91.6** ± 1.0
RedditM	**78.2** ± 3.0	**78.2** ± 3.0
Macro	**72.5** ± 0.3	**72.5** ± 0.3

B.2 VARIATION ACROSS CLASS-HOLDOUT SEEDS

Table 5 reports per graph variation across three predeclared class holdout seeds. Movies is the weakest conventional graph. RedditM is the weakest graph under the fixed-recipient protocol. The per-graph view makes both performance boundaries explicit.

Table 5: **Per-graph boundaries.** Values are mean ± sample standard deviation over 3 splits.

Dataset	Conventional AUROC	Fixed weakest AUROC	Known accuracy
Movies	77.2 ± 0.7	82.8 ± 1.3	51.6 ± 2.4
EleFashion	92.9 ± 4.4	88.2 ± 0.5	79.4 ± 4.0
Toys	94.6 ± 1.3	79.2 ± 1.6	68.6 ± 2.5
Grocery	87.2 ± 2.0	74.5 ± 1.7	65.8 ± 2.2
RedditS	98.1 ± 1.2	75.4 ± 1.5	91.6 ± 1.0
RedditM	98.0 ± 0.4	67.6 ± 0.8	78.2 ± 3.0

The fixed-recipient protocol removes degree zero shortcuts because every evaluated recipient has at least one known support neighbor. The remaining variation among positive-degree recipients shows that useful assignment support depends on informative relationships as well as their presence.

B.3 FAILURE LOCATION PROFILES ACROSS GRAPHS

Table 6 gives the exact macro metrics for each failure location. Fig. 4 then shows how the same metrics vary across the six graphs. AUROC and AP follow similar ordering, while FPR95 exposes a harder high recall regime. Same class placement is the recurring limit, especially on RedditM. Image replacement and cross class placement remain strong across the six graphs.

Table 6: **Metrics across failure locations.** Results are percentages over three splits.

Failure location	AUROC	AP	FPR95
Text replacement	91.2 ± 0.6	90.9 ± 0.8	37.2 ± 1.4
Image replacement	94.9 ± 0.4	94.6 ± 0.5	23.3 ± 1.5
Semantic record	90.8 ± 0.7	88.8 ± 0.8	36.0 ± 3.8
Cross class placement	92.6 ± 0.2	91.0 ± 0.2	28.5 ± 0.6
Same class placement	78.0 ± 0.8	76.7 ± 0.9	73.1 ± 1.1

B.4 KNOWN-ONLY OPERATING POINTS

Table 7 converts the continuous score into two admission rules estimated only from clean known validation scores. The $q = .90$ rule gives higher weakest shift recall, while the $q = .95$ rule accepts more clean records. RedditM remains the hardest graph at both operating points. Threshold selection must therefore reflect the relative cost of review and missed rejection rather than follow AUROC alone.

Table 7: **Known-only operating points.** Clean acceptance and weakest shift recall are percentages over three splits.

Dataset	$q = .90$		$q = .95$	
	Clean accept.	Weakest recall	Clean accept.	Weakest recall
Movies	90.9 ± 1.8	55.5 ± 2.9	95.0 ± 1.4	42.3 ± 6.9
EleFashion	88.7 ± 3.8	71.2 ± 8.4	94.7 ± 3.0	55.3 ± 15.2
Toys	86.6 ± 1.2	51.0 ± 6.9	95.7 ± 3.7	22.2 ± 20.7
Grocery	90.6 ± 2.4	33.4 ± 2.4	94.9 ± 1.4	18.5 ± 1.7
RedditS	91.1 ± 0.6	37.8 ± 6.5	95.4 ± 0.4	25.9 ± 6.2
RedditM	88.4 ± 4.6	29.8 ± 6.9	95.0 ± 4.9	14.0 ± 12.1

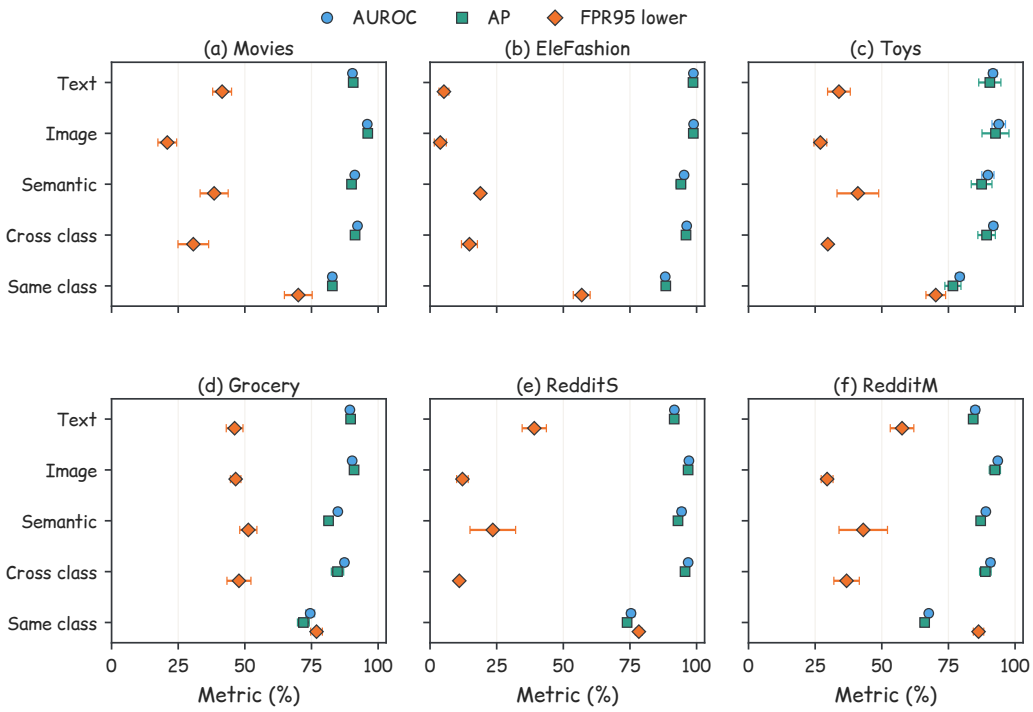

Figure 4: **Failure location profiles.** Per graph results over three class splits.

C Proof of Graph-Free Score Invariance

Proof of Proposition 1. Because $\pi : \mathcal{S} \to \mathcal{S}$ is a bijection, every index $j \in \mathcal{S}$ occurs exactly once as $j = \pi(i)$. Therefore,

$$\{f(\mathbf{r}_{\pi(i)}) : i \in \mathcal{S}\} = \{f(\mathbf{r}_j) : j \in \mathcal{S}\}$$

as multisets. Every statistic that depends only on the unordered collection of scores, including a fixed-threshold acceptance rate and any rank-based two-sample statistic, is consequently unchanged. The argument applies to a deterministic record-only score and a one-to-one reassignment. Detectors that observe graph context and shifts that alter the record multiset fall outside this proposition. □

D Inference Procedure

PlaceGuard uses Algorithm 1 for every target record.

Algorithm 1 PlaceGuard verification

Input Target record $\mathbf{r}_i$, claimed position i, connected known support $\mathcal{N}_i^{\mathrm{K}}$, frozen class predictions, empirical distributions $\widehat{F}_A$ and $\widehat{F}_C$, admission threshold τ_q
Output Known class prediction or rejection $\perp$

1. Project the target modalities and aggregate each modality over $\mathcal{N}_i^{\mathrm{K}}$ after excluding the target and every evaluation path.
2. Compute a_i and b_i using Eq. 2, then obtain S_i^A using Eq. 3.
3. Map the assignment log odds through $\widehat{F}_A$ to obtain u_i^A. Compute context uncertainty c_i from the target excluded context posterior.
4. Activate v_i^C only beyond the clean 97.5 percent context tail and compute $u_i = 1 - (1 - u_i^A)(1 - v_i^C)$.
5. Return $\perp$ when $u_i > \tau_q$. Otherwise return the frozen classifier prediction.

The same order is used for every dataset, class split, and replacement type. The released run directory stores the selected checkpoint, clean calibration scores, target scores, class predictions, donor assignments, timing, and software environment. These files allow the tables to be rebuilt without using labels or outcomes from unknown records during selection.

E Experimental Details

E.1 Information policy and splits

For each class seed, known support is divided deterministically into structured training and validation populations. Clean calibration, clean test, and semantic-unknown populations are disjoint. Training donors come from known support, and evaluation donors are reserved for terminal evaluation. Context excludes the target and evaluation-to-evaluation edges, and thresholds use only clean known validation nodes.

E.2 Replacement construction

Text and image replacements draw one modality from a known support donor while keeping the other target modality and recipient position constant. Cross-class placement moves a complete record from a different known class. The harder within-class placement condition moves a complete record from the recipient class. Donor assignments are one-to-one wherever the construction requires multiset preservation. Training, validation, and terminal evaluation use disjoint donor pools. Semantic unknown records are reserved for terminal evaluation after checkpoint and threshold selection.

E.3 Feature processing and optimization

Text and image embeddings are projected separately to 64 dimensions using support-fitted principal component analysis (PCA) and then ℓ_2 normalized. The frozen system provides text, image, fused, and graph logits. Support heads are two-layer MLPs with layer normalization, GELU, and dropout. AdamW runs for at most 120 epochs with patience 20. Configurations contain every optimizer, loss, seed, and checkpoint choice.

E.4 Shared search and calibration settings

Table 8 reports the complete three-setting search shared by all datasets. The classifier is frozen, so the legacy distillation field in the shared configuration file is not optimized. The same search grid is retained for every dataset after results are observed.

Table 8: **Shared configuration grid.** All datasets use the same three candidates.

Setting	Hidden	Dropout	Learning rate	Weight decay	Factor weight
1	64	0.1	10^{-3}	10^{-4}	1
2	128	0.1	10^{-3}	10^{-4}	1
3	128	0.2	5×10^{-4}	5×10^{-4}	2

Table 9: **Selection and calibration rules.** Every stage excludes terminal unknown results.

Stage	Fixed rule
Fitting	Known support records and structured known mismatches only
Configuration selection	Known-support structured validation using weakest negative recall, then AUROC, then clean known-class accuracy
Clean calibration	Clean known validation records only
Method development	Class seeds 53, 54
Terminal cohort	Class seeds 42, 43, 44 with all outcomes retained
Context tail	Clean context rank above 0.975
Admission thresholds	Clean validation quantiles at 0.90 and 0.95
Unknown access	No unknown record or terminal result enters fitting, selection, calibration, or thresholding

Table 10: **Terminal cohort outcomes.** AUROC percentages retain every split.

Class seed	Conventional	Weakest location
42	91.9	77.2
43	91.8	77.9
44	90.2	78.8

Table 11: **Audited artifact chain.** Identifiers are SHA-256 prefixes.

Stage	Saved record	Identifier
Development choice	Frozen candidate	`140b19c90b44`
Terminal protocol	Run contract	`66b441278aea`
Terminal outcomes	Run summary	`b230df320942`

Table 9 records the data boundary for every model decision. The score formula and the 0.975 context tail were chosen on development seeds 53 and 54 before the terminal cohort on seeds 42, 43, and 44. All three terminal outcomes are retained, and none changes fitting, selection, calibration, or thresholding. Table 10 shows each retained terminal outcome before aggregation. Table 11 links the development choice, terminal protocol, and terminal summary through saved identifiers.

E.5 Training objective details

Let $\mathcal{S}_+$ and $\mathcal{S}_-$ denote clean and structured training samples. For a probability s_i and a binary target t_i, define the balanced binary loss $\mathcal{B}$ below. The factor loss applies the same definition after the sigmoid transform.

$$\begin{aligned} \mathcal{B}(s,t) &= -\tfrac{1}{2}\mathbb{E}_{t_i=1}\log s_i - \tfrac{1}{2}\mathbb{E}_{t_i=0}\log(1-s_i), \\ \mathrm{BBCE}(z,t) &= \mathcal{B}(\sigma(z),t). \end{aligned} \tag{8}$$

Thus $\mathcal{L}_{\mathrm{member}}(S^A, t^r) = \mathcal{B}(S^A, t^r)$. The definition balances clean and structured records as well as positive and negative factor targets. No class loss is optimized because the accepted class prediction is frozen.

E.6 Environment and metric reconstruction

The terminal runs use Python 3.10.18, PyTorch 2.8.0 with CUDA 12.8, NumPy 1.26.4, and scikit-learn 1.7.2 on Linux. Each dataset directory records elapsed time, peak GPU memory, peak process memory, the selected configuration, and the complete validation trace.

Metric reconstruction starts from continuous per record scores. Unknown records are positive for AUROC, AP, and FPR95. FPR95 is read from the complete empirical ROC at the first point reaching 95% unknown recall. Operating thresholds use empirical quantiles of clean known validation scores.

E.7 Comparison controls

All rows use the same frozen features, class split, recipients, and terminal conditions. Maximum softmax probability reads fused record logits. Prototype distance compares each modality with known class support prototypes. Preassigned role compatibility follows CoOD's component consistency principle. Multimodal cosine averages text context and image context incompatibility. These controls describe the corresponding information paths.

E.8 Compared OOD methods

The broad comparison preserves the decision operations of seven OOD method families while replacing their original dataset wrapper with the fixed-recipient MAG protocol. NODESAFE (Yang et al., 2024) and TopoOOD (Bao et al., 2024) receive the complete normalized text-image record as the node feature and the known support graph as topology. TIDE (Wang et al., 2026) retains its joint,

structure, feature, bottleneck, reconstruction, and information-decomposition components. TNT-OOD (Wang et al., 2025a) retains its text-topology fusion, energy-alignment score, and propagation operator, with the complete multimodal record occupying its content input. DPU (Li et al., 2025) retains its dynamic prototype training and energy score. GNNSafe (Wu et al., 2023) propagates classifier energy and GRASP (Ma et al., 2024) propagates maximum-softmax confidence from frozen clean anchors. Every query uses an isolated symmetric support-plus-query graph with messages restricted to known support and the current query. Training follows the known-class information policy in Section E.7.

All graph adapters use one-way support-to-query messages. Checkpoints are selected on the same known validation partition with one frozen source recipe, and clean calibration opens only after selection. Seeds, recipients, donors, five shifts, metrics, and terminal evaluation are identical. The rows are comparable under the shared task.

F Dataset and Relationship Details

Movies, Toys, and Grocery combine text and images with released `also_buy`/`also_view` neighborhoods. EleFashion uses product co-purchase relations. RedditS and RedditM are image-caption posts labeled by subreddit, with neighborhoods built from author-grouped `also_posted` lists that encode same-author relationships. Reassignment moves the complete text-image record while keeping node identifiers and released relations unchanged. It changes record-relation correspondence while leaving both marginal collections unchanged.

Table 12 records graph scale, feature dimensions, and predefined support, calibration, and test populations from recorded run summaries.

Table 12: **Dataset scale.** Counts and feature dimensions come from recorded run summaries.

Dataset	Nodes	Edges	Text dim.	Image dim.	Support	Calibration/Test
Movies	16,672	109,195	768	768	2,048	256/512
EleFashion	97,766	199,602	1024	2048	2,048	256/512
Toys	20,695	63,443	768	768	2,048	256/512
Grocery	17,074	85,670	768	768	2,048	256/512
RedditS	15,894	283,080	768	768	3,072	256/512
RedditM	99,638	583,594	768	768	3,072	256/512

G Limitations and Broader Impact

Intact-record reassignment is designed as a controlled identifiability test and estimates neither a natural anomaly distribution nor anomaly prevalence. Deployment assumes an inductive admission setting with incident links to an admitted support store. New records arrive without labels, and records evaluated together do not exchange messages. Sparse or uninformative support can remove the placement signal and reduce the decision to record-only information. The frozen encoder and support-fitted PCA leave representation drift untested. Relational context may also encode popularity, community, or demographic bias. Deployment should audit relations, protect privacy, and treat relational incompatibility separately from semantic novelty.