BEYOND CONFIDENCE THRESHOLDING: MITIGATING CACHE CORRUPTION IN TEST-TIME OOD DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Out-of-distribution (OOD) detection aims to identify test samples deviating from in-distribution (ID) data. Recent test-time adaptation methods tackle this challenge by dynamically updating visual caches to retrieve negative labels that delineate the ID decision boundary. However, these methods implicitly assume balanced test streams and suffer catastrophic performance degradation under realistic, ID-dominated streaming deployments in challenging Near-OOD scenarios. In this work, we reveal that this failure stems from an inherent objective misalignment between margin-driven negative label retrieval and confidence-based cache updates. By projecting high-dimensional representations onto a scalar confidence metric, existing cache updates discard local neighborhood geometry, mistakenly purging informative boundary negatives in favor of coarse, uninformative concepts. To resolve this dilemma, we propose a graph-driven training-free framework that replaces heuristic scalar thresholding with a mutual k-NN bipartite graph constructed between visual embeddings and textual concepts, introducing a topological connectivity gate to govern cache admission via geometric proximity. By preserving the cross-modal manifold structure, our approach reliably mines informative negative labels that are acutely sensitive to the fine-grained boundary between ID and Near-OOD data. Extensive evaluations on challenging Near-OOD benchmarks demonstrate that our framework effectively prevents boundary collapse, outperforming state-of-the-art baselines by up to 10% AUROC under severe streaming imbalances while remaining highly competitive in balanced settings.

1 INTRODUCTION

Deploying artificial intelligence systems into open-world environments inevitably exposes them to out-of-distribution (OOD) inputs outside predefined categories. Existing vision models often produce overconfident predictions for these anomalous samples (Scheirer et al., 2012; Nguyen et al., 2015), posing severe risks in safety-critical domains such as autonomous driving (Filos et al., 2020) and medical diagnosis (Fu et al., 2023). Consequently, equipping models with the capability to reliably distinguish in-distribution (ID) data from unknown OOD instances remains an indispensable requirement for AI safety.

Early visual OOD detection methods predominantly operated on unimodal features extracted from closed-set classifiers, leveraging output confidence (Hendrycks & Gimpel, 2017; Basart et al., 2022), feature distances (Lee et al., 2018; Sun et al., 2022), or energy scores (Liu et al., 2020). With the advent of vision-language models (VLMs) such as CLIP (Radford et al., 2021), exploiting extensive textual semantic knowledge has emerged as a promising paradigm for zero-shot OOD detection (Esmaeilpour et al., 2022; Ming et al., 2022; Wang et al., 2023). A prominent line of work, exemplified by NegLabel (Jiang et al., 2024), achieves superior performance by mining vast corpora of negative textual concepts distant from known ID labels and scoring test images according to their relative alignment with ID versus negative text anchors. Building upon this, recent test-time adaptation methods (Zhang & Zhang, 2024; Zhang et al., 2026; Zhu et al., 2026b) introduce dynamic visual caches. By monitoring online test streams (*i.e.*, sequences of test images arriving at the cache), these approaches retrieve activated negative labels that capture empirical OOD characteristics while suppressing irrelevant concepts, achieving substantial gains on standard benchmarks. Despite these advances, existing frameworks implicitly presume balanced arrivals between ID and OOD samples. In realistic deployments, however, test streams are overwhelmingly ID-dominated, as practical sys-

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.

tems operate within designated domains where inputs predominantly mirror the training distribution and anomalies emerge only intermittently. Under such streaming imbalance, existing methods suffer catastrophic performance degradation, particularly in challenging Near-OOD settings where visual concepts between ID and OOD classes share fine-grained semantics.

In this work, we reveal that this failure mode stems from a fundamental objective misalignment between negative label retrieval and cache admission rules. On one hand, negative label retrieval seeks to maximize the functional separation margin between ID and OOD representations by mining boundary-sensitive hard negatives. On the other hand, existing cache updates govern sample admission via confidence thresholding, pursuing nominal classification purity during test-time adaptation. Projecting high-dimensional representations onto a scalar confidence metric discards local neighborhood geometry, mistakenly purging informative boundary negatives in favor of uninformative coarse concepts. This in turn reinforces the confidence-based partitioning bias and locks cache updates into a self-reinforcing failure loop. To resolve this dilemma, we propose a graph-driven training-free framework that unifies negative label selection and cache evolution through bidirectional topological consensus. Specifically, we construct a bipartite graph between visual embeddings and text concepts via mutual k-nearest-neighbor (k-NN) matching. Building on this graph representation, we filter out uninformative coarse negative labels that offer marginal utility for Near-OOD detection, and introduce a sub-bipartite topological connectivity gate to govern cache admission through geometric proximity. By prioritizing cross-modal geometric alignment over heuristic prediction confidence, our approach ensures that memory admission preserves the margin expansion objective inherent to hard negative selection. The main contributions of this paper are summarized as follows:

- We identify a critical failure mode of test-time Near-OOD adaptation under streaming imbalance, providing a mechanistic analysis that reveals an inherent objective misalignment between margin-driven negative retrieval and confidence-based cache updates.
- We propose a structure-driven training-free framework that unifies negative label selection and cache evolution via mutual k-NN bipartite graph consensus, using a sub-bipartite topological connectivity gate to govern cache admission through geometric proximity rather than heuristic scalar confidence.
- Extensive evaluations across standard Near-OOD benchmarks show that our framework prevents boundary collapse, significantly outperforming state-of-the-art baselines under severe streaming imbalances while remaining competitive in balanced settings.

2 Related Works

VLM-Based OOD Detection. With the development of pre-trained vision-language models (VLMs) (Jia et al., 2021; Radford et al., 2021), utilizing rich textual semantics to identify visual distribution shifts has gained significant momentum (Esmaeilpour et al., 2022; Li et al., 2024; Le et al., 2026). A common paradigm involves fine-tuning model branches through prompt learning or auxiliary objectives, as demonstrated in LoCoOp (Miyai et al., 2023), CLIPN (Wang et al., 2023), and LAPT (Zhang et al., 2024b). However, training-based approaches typically introduce additional computational overhead and, in some cases, require access to auxiliary OOD data, which can limit their practicality in real-world deployments.

Consequently, training-free approaches have attracted growing interest (Zhang et al., 2024a; Ansari et al., 2025; Peng et al., 2026). Early attempts such as MCM (Ming et al., 2022) detect anomalies by analyzing the softmax similarity distribution over ID text prompts. To establish a clearer decision boundary, NegLabel (Jiang et al., 2024) explores thousands of candidate negative concepts mined from external lexical corpora based on cosine distances from ID classes. Subsequent studies incorporate text-image consistency constraints or conjugated semantic pools to improve negative label quality (Cao et al., 2024; Chen et al., 2024). Nonetheless, since these negative label spaces are selected statically prior to inference, they remain agnostic to the actual deployment environment and inevitably include uninformative or irrelevant concepts that degrade OOD discriminability.

Test-Time Adaptation for OOD Detection. Test-time adaptation enables models to dynamically adjust to evolving data streams during inference (Liang et al., 2025). While test-time optimization strategies refine model parameters online, their iterative gradient updates introduce substantial latency that impedes real-time execution (Gao et al., 2023; Fan et al., 2024). To maintain efficiency,

optimization-free methods rely on dynamic memory mechanisms (Zhang & Zhang, 2024; Yang et al., 2025). For example, AdaNeg (Zhang & Zhang, 2024) couples negative labels with adaptive image proxies. TANL (Zhang et al., 2026) introduces dynamic FIFO queues to mine activated negative labels on the fly by monitoring online test streams. More recently, DDE (Zhu et al., 2026b) attempt to accommodate incoming test diversity by expanding external lexical prototypes and negative dictionaries during test time.

Despite their success, existing test-time adaptation methods implicitly assume balanced arrivals between ID and OOD samples. While Far-OOD instances reside well outside the ID semantic space and remain largely unaffected by distributional shifts, Near-OOD instances share fine-grained visual characteristics with ID data, making them acutely vulnerable under ID-dominated streaming streams. In such imbalanced settings, scalar confidence partitioning populates positive caches with boundary ID instances, inflating the penalty against boundary concepts and purging informative hard negatives. Unlike prior frameworks relying on confidence heuristics, our approach unifies negative label selection and cache evolution through bidirectional bipartite graph consensus, consistently preserving boundary-sensitive negative labels to ensure robust separation margins.

3 METHODOLOGY

In this section, we present our training-free test-time adaptation framework, focusing on the challenging Near-OOD detection setting where semantic shifts are subtle and visual boundaries remain highly ambiguous. Under realistic ID-dominated streams, existing methods suffer from severe performance degradation caused by an inherent objective misalignment between scalar confidence partitioning and margin-driven negative label selection. We resolve this failure via two key components: (1) revealing that dual-tower vision-language contrastive alignment shares structural homologies with bipartite attention aggregation, which justifies our mutual k-NN bipartite graph as a structural reconstruction of the pre-trained semantic topology; and (2) introducing two new graph-based negative label selection strategies alongside an aligned cache update rule governed by a sub-bipartite topological connectivity gate.

3.1 PRELIMINARIES

Let $\mathcal{X}$ denote the image space and $\mathcal{Y}^+ = \{y_c\}_{c=1}^C$ represent the set of ID class labels. In zero-shot OOD detection with VLMs, an image $\mathbf{x} \in \mathcal{X}$ is mapped to a normalized visual feature $\mathbf{v} = f_{\text{img}}(\mathbf{x}) \in \mathbb{R}^D$, while textual class names are embedded as normalized text vectors $\mathbf{t}_c = f_{\text{txt}}(\rho(y_c)) \in \mathbb{R}^D$ via a prompt template $\rho(\cdot)$. An OOD detection framework employs a scoring function $S(\mathbf{x})$ as confidence to distinguish ID samples from unknown OOD instances using a decision threshold γ (Lee et al., 2018; Liu et al., 2020):

$$\text{Predcition} = \begin{cases} \text{ID}, & \text{if } S(\mathbf{x}) \geq \gamma; \\ \text{OOD}, & \text{otherwise.} \end{cases} \tag{1}$$

More recently, NegLabel (Jiang et al., 2024) introduces an external text corpus $\mathcal{Y}^{\text{cor}} = \{\hat{y}_j\}_{j=1}^{N_{\text{cor}}}$ beyond the ID labels $\mathcal{Y}^+$, and mines a subset of M negative textual labels $\mathcal{Y}^- = \{\tilde{y}_m\}_{m=1}^M \subset \mathcal{Y}^{\text{cor}}$ that are distant from the ID class names in the embedding space. It evaluates test instances using an expanded Softmax probability score $S_{\text{nl}}(\mathbf{v})$, which computes the sum of probabilities assigned to ID classes normalized over both ID and mined negative labels:

$$S_{\text{nl}}(\mathbf{v}) = \sum_{i=1}^{C} \frac{\exp(\tau \mathbf{v}^\top \mathbf{t}_i)}{\sum_{j=1}^{C} \exp(\tau \mathbf{v}^\top \mathbf{t}_j) + \sum_{m=1}^{M} \exp(\tau \mathbf{v}^\top \tilde{\mathbf{t}}_m)}, \tag{2}$$

where $\mathbf{t}_j$ and $\tilde{\mathbf{t}}_m$ denote the text embeddings of the j-th ID class and m-th mined negative label, respectively, and τ is a temperature scaling parameter. ID samples typically yield higher cosine similarities with ID classes and obtain larger $S_{\text{nl}}(\mathbf{v})$ values. Conversely, OOD samples align more closely with negative labels, expanding the denominator sum and suppressing the score toward zero.

3.2 CONTRASTIVE LEARNING AS BIPARTITE GRAPH ATTENTION NETWORKS

To theoretically ground our graph-guided framework, we reveal the intrinsic equivalence between dual-tower contrastive optimization and message passing in Graph Attention Networks (GATs). For an image embedding $\mathbf{v}_i$ and candidate text embeddings $\mathbf{t}_j \in \mathcal{B}_{\text{txt}}$, the InfoNCE loss $\mathcal{L}_i = -\log p_{i,y_i}$ yields the visual gradient:

$$\nabla_{\mathbf{v}_i}\mathcal{L}_i = \tau\left(\sum_{\mathbf{t}_j\in\mathcal{B}_{\text{txt}}} p_{ij}\mathbf{t}_j - \mathbf{t}_{y_i}\right), \quad \text{where} \quad p_{ij} = \frac{\exp(\tau\mathbf{v}_i^\top\mathbf{t}_j)}{\sum_{\mathbf{t}_k\in\mathcal{B}_{\text{txt}}}\exp(\tau\mathbf{v}_i^\top\mathbf{t}_k)}. \tag{3}$$

A single gradient update with step size η updates the visual representation as:

$$\mathbf{v}_i^{t+1} = \mathbf{v}_i^t - \eta\nabla_{\mathbf{v}_i}\mathcal{L}_i = \mathbf{v}_i^t + \eta\tau\mathbf{t}_{y_i} - \eta\tau\sum_{\mathbf{t}_j\in\mathcal{B}_{\text{txt}}} p_{ij}\mathbf{t}_j. \tag{4}$$

This formulation is structurally equivalent to directed message aggregation over a cross-modal bipartite graph. Here, visual and textual embeddings act as bipartite nodes, normalized softmax affinities p_{ij} function as dynamic attention weights, and visual representations are updated through self-retention alongside message aggregation from connected textual neighbors.

Motivation for Topological Graph Reconstruction. Standard contrastive pre-training computes dense bipartite interactions across the entire candidate pool $\mathcal{B}_{\text{txt}}$. For Near-OOD detection, however, arbitrary negative pairings offer negligible boundary utility; instead, the most informative negative labels are those sharing fine-grained semantic affinity with visual instances. Unlike prior methods that retrieve negative labels indiscriminately from the full corpus $\mathcal{Y}^{\text{cor}}$, our approach reconstructs the intrinsic pre-training topology via mutual k-NN connectivity. This guarantees that selected negative labels retain bidirectional cross-modal correspondences with visual samples, establishing an effective semantic scaffold to separate hard boundary anomalies.

3.3 DUAL-SCALE BIPARTITE GRAPH CONSTRUCTION

Let $\mathcal{X}_{\text{pos}} = \{\mathbf{v}_i\}_{i=1}^{L_{\text{pos}}} \subset \mathbb{R}^D$ and $\mathcal{X}_{\text{neg}} = \{\mathbf{v}_j\}_{j=1}^{L_{\text{neg}}} \subset \mathbb{R}^D$ denote the positive and negative visual caches of capacities L_{pos} and L_{neg}, which store confident ID and OOD samples at test time, respectively. The unified visual cache matrix is formed by concatenating both memory pools:

$$\mathbf{V}_{\text{cache}} = [\mathcal{X}_{\text{pos}}; \mathcal{X}_{\text{neg}}] \in \mathbb{R}^{L_{\text{all}}\times D}, \quad \text{where } L_{\text{all}} = L_{\text{pos}} + L_{\text{neg}}. \tag{5}$$

Similarly, we combine the ID text features $\mathbf{T}_{\text{id}} = [\mathbf{t}_1, \ldots, \mathbf{t}_C] \in \mathbb{R}^{D\times C}$ with the negative corpus text features $\mathbf{T}_{\text{neg}} = [\tilde{\mathbf{t}}_1, \ldots, \tilde{\mathbf{t}}_{N_{\text{cor}}}] \in \mathbb{R}^{D\times N_{\text{cor}}}$ into a comprehensive textual matrix:

$$\mathbf{T}_{\text{all}} = [\mathbf{T}_{\text{id}}, \mathbf{T}_{\text{neg}}] \in \mathbb{R}^{D\times(C+N_{\text{cor}})}. \tag{6}$$

We compute the cross-modal affinity matrix $\mathbf{W} \in \mathbb{R}^{L_{\text{all}}\times(C+N_{\text{cor}})}$ representing the pairwise similarities between all cached visual nodes and candidate textual nodes:

$$\mathbf{W} = \mathbf{V}_{\text{cache}}\mathbf{T}_{\text{all}}. \tag{7}$$

To reconstruct the intrinsic topology and identify salient cross-modal connections while ensuring global connectivity, we construct a dual-scale bipartite adjacency mask $\mathbf{A} \in \{0,1\}^{L_{\text{all}}\times(C+N_{\text{cor}})}$ composed of a mutual k-NN graph and a skeleton graph:

1. **Mutual k-NN Topology ($\mathbf{A}_{\text{mutual}}$).** To accurately capture high-density local manifolds, we enforce a bidirectional k-nearest neighbor consensus:

$$(\mathbf{A}_{\text{mutual}})_{ij} = \mathbb{I}\left(j \in \text{TopK}_{j'}(\mathbf{W}_{ij'}, k)\right) \cdot \mathbb{I}\left(i \in \text{TopK}_{i'}(\mathbf{W}_{i'j}, k)\right), \tag{8}$$

where $\text{TopK}_{j'}(\mathbf{W}_{ij'}, k)$ selects the top-k text candidate indices for visual node i, $\text{TopK}_{i'}(\mathbf{W}_{i'j}, k)$ selects the top-k visual candidate indices for text node j, and $\mathbb{I}(\cdot)$ is the indicator function.

2. **Top-1 Skeleton Topology ($\mathbf{A}_{\text{skeleton}}$).** To prevent nodes from becoming isolated under strict mutual filtering, we retain Top-1 matches to ensure global graph connectivity:

$$(\mathbf{A}_{\text{skeleton}})_{ij} = \mathbb{I}\left(j = \arg\max_{j'} \mathbf{W}_{ij'}\right) \vee \mathbb{I}\left(i = \arg\max_{i'} \mathbf{W}_{i'j}\right). \tag{9}$$

The final bipartite adjacency mask $\mathbf{A}$ is formed via the union of both topologies:

$$\mathbf{A} = \mathbf{A}_{\text{mutual}} \vee \mathbf{A}_{\text{skeleton}}. \tag{10}$$

3.4 Graph-Based Negative Label Selection and Confidence Scoring

Using the bipartite adjacency mask $\mathbf{A}$ and the global similarity matrix $\mathbf{W}$, we select the most informative negative textual labels by evaluating the topological disparity between the positive and negative visual caches. To account for both global semantics and sample-specific visual manifold variations, we compute three complementary disparity scores:

1. **Unconstrained Disparity ($\mathbf{R}_1$).** Following the baseline activation formulation in TANL (Zhang et al., 2026), this metric measures the raw aggregate cross-modal similarity difference of negative texts toward the entire negative and positive memory pools:

$$\mathbf{R}_1 = \sum_{j=1}^{L_{\text{neg}}} (\mathbf{W}_{\text{neg}})_j - \sum_{i=1}^{L_{\text{pos}}} (\mathbf{W}_{\text{pos}})_i \in \mathbb{R}^{1 \times N_{\text{cor}}}, \tag{11}$$

where $\mathbf{W}_{\text{pos}} = \mathbf{W}_{1:L_{\text{pos}}, C+1:C+N_{\text{cor}}}$ and $\mathbf{W}_{\text{neg}} = \mathbf{W}_{L_{\text{pos}}+1:L_{\text{all}}, C+1:C+N_{\text{cor}}}$ denote the similarity sub-matrices of the negative corpus $\mathbf{T}_{\text{neg}}$ with respect to the positive cache $\mathcal{X}_{\text{pos}}$ and negative cache $\mathcal{X}_{\text{neg}}$, respectively. Although $\mathbf{R}_1$ captures overall activation toward OOD samples, unweighted summation admits uninformative negatives. To address this limitation, we introduce two graph-based disparity formulations.

2. **Topological Global Disparity ($\mathbf{R}_2$).** Evaluates the disparity strictly along cross-modal edges validated by the dual-scale topological mask $\mathbf{A}$:

$$\mathbf{R}_2 = \sum_{j=1}^{L_{\text{neg}}} (\mathbf{W}_{\text{neg}} \odot \mathbf{A}_{\text{neg}})_j - \sum_{i=1}^{L_{\text{pos}}} (\mathbf{W}_{\text{pos}} \odot \mathbf{A}_{\text{pos}})_i \in \mathbb{R}^{1 \times N_{\text{cor}}}, \tag{12}$$

where $\mathbf{A}_{\text{pos}} = \mathbf{A}_{1:L_{\text{pos}}, C+1:C+N_{\text{cor}}}$ and $\mathbf{A}_{\text{neg}} = \mathbf{A}_{L_{\text{pos}}+1:L_{\text{all}}, C+1:C+N_{\text{cor}}}$ denote the corresponding topological adjacency submatrices, and $\odot$ is the Hadamard product. By pruning spurious background connections, $\mathbf{R}_2$ isolates informative boundary negatives. Nevertheless, pooling across the entire cache assigns a uniform negative label subset to all test samples, overlooking instance specific visual variations.

3. **Topological Instance Disparity ($\mathbf{R}_3$).** Modulates topological connections by the visual affinity between the test batch $\mathbf{V}_{\text{test}} \in \mathbb{R}^{B \times D}$ and the cached visual representations, conditioning negative label retrieval on the local manifold of each query:

$$\mathbf{R}_3 = (\mathbf{V}_{\text{test}} \mathcal{X}_{\text{neg}}^{\top})(\mathbf{W}_{\text{neg}} \odot \mathbf{A}_{\text{neg}}) - (\mathbf{V}_{\text{test}} \mathcal{X}_{\text{pos}}^{\top})(\mathbf{W}_{\text{pos}} \odot \mathbf{A}_{\text{pos}}) \in \mathbb{R}^{B \times N_{\text{cor}}}. \tag{13}$$

Multi-Scale Negative Label Retrieval. For each disparity score $\mathbf{R}_r$ ($r \in \{1, 2, 3\}$), we extract the indices of the top-β highest-scoring negative text features:

$$\mathcal{I}_r = \text{TopK}(\mathbf{R}_r, \beta) \subset \{1, \ldots, N_{\text{cor}}\}, \tag{14}$$

and retrieve the corresponding selected negative text embeddings $\mathbf{T}_{\text{neg}}^r = \mathbf{T}_{\text{neg}}[:, \mathcal{I}_r] \in \mathbb{R}^{D \times \beta}$.

Composite Confidence Scoring. For an incoming test image $\mathbf{v}$, we evaluate its prediction against the selected negative subset $\mathbf{T}_{\text{neg}}^r = [\tilde{\mathbf{t}}_1^r, \ldots, \tilde{\mathbf{t}}_\beta^r] \in \mathbb{R}^{D \times \beta}$ under each scale $r \in \{1, 2, 3\}$ using the scoring formulation defined in Eq. 2:

$$S_{\text{nl}}^r(\mathbf{v}) = \sum_{i=1}^{C} \frac{\exp(\tau \mathbf{v}^{\top} \mathbf{t}_i)}{\sum_{j=1}^{C} \exp(\tau \mathbf{v}^{\top} \mathbf{t}_j) + \sum_{m=1}^{\beta} \exp(\tau \mathbf{v}^{\top} \tilde{\mathbf{t}}_m^r)}. \tag{15}$$

The overall confidence score $S_{\text{final}}(\mathbf{v})$ aggregates predictions across all three topological scales:

$$S_{\text{final}}(\mathbf{v}) = \sum_{r=1}^{3} S_{\text{nl}}^{r}(\mathbf{v}). \tag{16}$$

Concurrently, the unique union of activated negative indices $\mathcal{I}_{\text{active}} = \bigcup_{r=1}^{3} \mathcal{I}_r$ is collected to structure the active sub-graph for subsequent cache evolution.

3.5 Cache Update and Initialization

Cache Update via Topological Connectivity. Existing test-time methods rely on scalar confidence thresholding to partition samples into positive and negative visual caches. Under ID-dominated streams, this introduces a fundamental objective misalignment: scalar projection discards local neighborhood geometry, purging informative boundary negatives in favor of coarse concepts and locking cache evolution into a self-reinforcing failure loop (detailed in Appx. A.1). To realign cache admission with boundary separation, we replace scalar thresholding with a decision gate based on local topological connectivity over the active text subspace. Given active negative indices $\mathcal{I}_{\text{active}} \subset \{1, \ldots, N_{\text{cor}}\}$, we form the compact text matrix:

$$\mathbf{T}_{\text{active}} = [\mathbf{T}_{\text{id}}, \mathbf{T}_{\text{neg}}[:, \mathcal{I}_{\text{active}}]] \in \mathbb{R}^{D \times (C + N_{\text{active}})}. \tag{17}$$

For a streaming test batch $\mathbf{V}_{\text{test}} \in \mathbb{R}^{B \times D}$, we compute similarity $\mathbf{W}_{\text{sub}} = \mathbf{V}_{\text{test}} \mathbf{T}_{\text{active}}$ and derive its dual scale adjacency $\mathbf{A}_{\text{sub}} \in \{0, 1\}^{B \times (C + N_{\text{active}})}$ as detailed in Sec. 3.3. For each sample $\mathbf{v}_i$, we compute mean topological connectivity toward ID and active negative label subsets:

$$\bar{E}_{\text{id}}(\mathbf{v}_i) = \frac{1}{C} \sum_{c=1}^{C} (\mathbf{A}_{\text{sub}} \odot \mathbf{W}_{\text{sub}})_{i,c}, \quad \bar{E}_{\text{neg}}(\mathbf{v}_i) = \frac{1}{N_{\text{active}}} \sum_{j=1}^{N_{\text{active}}} (\mathbf{A}_{\text{sub}} \odot \mathbf{W}_{\text{sub}})_{i,C+j}. \tag{18}$$

The visual memory queues are then dynamically updated via the following consensus rule:

- If $\bar{E}_{\text{id}}(\mathbf{v}_i) > \bar{E}_{\text{neg}}(\mathbf{v}_i)$, sample $\mathbf{v}_i$ is identified as an ID instance and appended to $\mathcal{X}_{\text{pos}}$.
- If $\bar{E}_{\text{neg}}(\mathbf{v}_i) > \bar{E}_{\text{id}}(\mathbf{v}_i)$, sample $\mathbf{v}_i$ is identified as a potential outlier and appended to $\mathcal{X}_{\text{neg}}$.

Both caches follow a FIFO replacement policy to maintain maximum capacities L_{pos} and L_{neg}.

Cache Initialization. At the beginning of testing (cold-start phase), we initialize both visual caches using structured semantic prototypes derived directly from the text embeddings:

$$\mathcal{X}_{\text{pos}} = \text{KMeans}(\mathbf{T}_{\text{id}}^{\top}, K = L_{\text{pos}}), \quad \mathcal{X}_{\text{neg}} = \text{KMeans}(\mathbf{T}_{\text{neg}}^{\top}, K = L_{\text{neg}}), \tag{19}$$

where $\text{KMeans}(\mathbf{X}, K)$ yields K cluster centroids from feature matrix $\mathbf{X}$. If available class embeddings are fewer than the cache capacity, text features are used directly.

4 Experiments

4.1 Setup

Datasets. We evaluate our framework using ImageNet-1k (Deng et al., 2009) as the ID dataset within the OpenOOD (Yang et al., 2022; Zhang et al., 2023) benchmark. For Near-OOD evaluation, we employ SSB-hard (Vaze et al., 2022) and NINCO (Bitterwolf et al., 2023), which exhibit subtle, fine-grained semantic shifts from ID classes. For Far-OOD evaluation, we use iNaturalist (Van Horn et al., 2018), Textures (Cimpoi et al., 2014), and OpenImage-O (Wang et al., 2022). In addition, we evaluate under the full-spectrum (FS) setting (Yang et al., 2023), which introduces covariate-shifted ID streams, specifically ImageNet-C (Hendrycks & Dietterich, 2019), ImageNet-R (Hendrycks et al., 2021), and ImageNet-V2 (Recht et al., 2019), to assess detector robustness against non-semantic variations, including synthetic corruptions and natural distribution shifts.

Implementation Details. Our primary visual backbone is the ViT-B/16 image encoder pre-trained via CLIP (Radford et al., 2021), with additional architectures evaluated in Appx. A.4. We fix the

Table 1: Near-OOD detection performance on SSB-hard and NINCO (mean ± std over 5 runs). **Bold** and underline denote the best and second-best results, respectively. FS denotes the full-spectrum evaluation setting incorporating covariate-shifted ID instances.

Methods	OOD datasets									
	SSB-hard		NINCO		SSB-hard-FS		NINCO-FS		Average	
	AUROC↑	FPR@95↓	AUROC↑	FPR@95↓	AUROC↑	FPR@95↓	AUROC↑	FPR@95↓	AUROC↑	FPR@95↓
MCM	$65.07_{\pm 0.61}$	$85.71_{\pm 0.67}$	$75.22_{\pm 0.28}$	$68.53_{\pm 0.06}$	$52.74_{\pm 0.17}$	$89.50_{\pm 0.46}$	$63.23_{\pm 0.08}$	$79.80_{\pm 0.33}$	66.07	79.39
NegLabel	$71.88_{\pm 0.59}$	$76.94_{\pm 0.15}$	$76.87_{\pm 0.37}$	$\underline{61.28}_{\pm 0.19}$	$67.09_{\pm 0.28}$	$82.91_{\pm 0.27}$	$72.53_{\pm 0.09}$	$\underline{68.80}_{\pm 0.35}$	72.09	72.48
AdaNeg	$70.87_{\pm 0.21}$	$77.68_{\pm 0.44}$	$74.36_{\pm 0.17}$	$62.63_{\pm 0.59}$	$63.95_{\pm 0.16}$	$82.55_{\pm 0.60}$	$67.54_{\pm 0.28}$	$69.59_{\pm 0.90}$	69.18	73.11
ANTS	$68.93_{\pm 0.29}$	$82.42_{\pm 0.74}$	$76.40_{\pm 0.35}$	$69.68_{\pm 1.46}$	$55.99_{\pm 0.27}$	$89.66_{\pm 0.37}$	$65.43_{\pm 0.56}$	$79.74_{\pm 1.22}$	66.69	80.38
DDE	$74.12_{\pm 1.86}$	$72.39_{\pm 2.00}$	$77.09_{\pm 1.78}$	$65.79_{\pm 2.87}$	$66.60_{\pm 0.51}$	$80.79_{\pm 0.64}$	$70.93_{\pm 0.38}$	$72.89_{\pm 1.41}$	72.19	72.97
TANL	$\underline{75.69}_{\pm 0.97}$	$\underline{71.24}_{\pm 1.51}$	$\underline{79.24}_{\pm 0.75}$	$62.76_{\pm 1.77}$	$\underline{67.91}_{\pm 0.06}$	$\underline{79.76}_{\pm 0.09}$	$\underline{73.40}_{\pm 0.45}$	$71.00_{\pm 0.50}$	$\underline{74.06}$	$\underline{71.19}$
Ours	$\mathbf{79.64}_{\pm 0.37}$	$\mathbf{65.07}_{\pm 0.69}$	$\mathbf{83.66}_{\pm 0.18}$	$\mathbf{59.58}_{\pm 0.95}$	$\mathbf{72.56}_{\pm 0.19}$	$\mathbf{74.77}_{\pm 0.24}$	$\mathbf{78.65}_{\pm 1.03}$	$\mathbf{68.11}_{\pm 1.88}$	**78.63**	**66.88**

number of retrieved negative labels to $\beta = 1000$, temperature to $\tau = 1$ in Eq. 2, cache capacities to $L_{\text{pos}} = L_{\text{neg}} = 300$ in Eq. 5, and the mutual neighborhood size to $k = 70$ in Eq. 8. Following standard evaluation protocols (Huang & Li, 2021; Ming et al., 2022; Jiang et al., 2024), we report FPR95 and AUROC as the primary performance metrics.

4.2 Experiment Results

Our framework is specifically designed to tackle Near-OOD detection under streaming imbalance. Consequently, we focus primarily on Near-OOD evaluation in this section, while full evaluation results on Far-OOD benchmarks are deferred to Appx. A.2. To simulate realistic streaming deployments characterized by rare anomaly arrivals, we construct imbalanced test streams by subsampling Near-OOD datasets to 2%, 5%, 10%, 20%, and 50% of their original size. We benchmark our method against two static zero-shot baselines (MCM (Ming et al., 2022) and NegLabel (Jiang et al., 2024)), whose instance-independent scoring functions remain invariant to stream ordering and class ratios, alongside four representative test-time adaptation methods: ANTS (Zhu et al., 2026a), AdaNeg (Zhang & Zhang, 2024), DDE (Zhu et al., 2026b), and TANL (Zhang et al., 2026).

As summarized in Tab. 1, we report the mean and standard deviation of FPR95 and AUROC averaged across the subsampling ratios over 5 independent runs. Our framework consistently achieves top-ranked performance across all evaluation settings, attaining an average AUROC of **78.63%** and reducing FPR95 to **66.88%**. Under severe streaming imbalance, existing adaptation methods such as AdaNeg and ANTS degrade drastically, falling well behind the static NegLabel baseline. Under the more demanding FS evaluation, DDE is similarly surpassed by NegLabel, underscoring the acute vulnerability of confidence-driven test-time adaptation when subjected to imbalanced streams.

We further illustrate the dynamic AUROC trajectories across varying OOD data proportions (from 2% to 100%) in Fig. 1. Existing test-time adaptation approaches experience severe performance degradation in the low-OOD regime (2% to 10%). In contrast, our framework maintains an exceptionally flat and resilient trajectory, outperforming TANL by approximately **10%** AUROC at the 2% extreme while matching its competitiveness as streams become balanced. Crucially, when we ablate our framework by substituting our topological gate with TANL's cache update rule (*Ours (TANL-Cache)*), which governs cache admission via scalar confidence S_{nl}, performance drops dramatically under low-OOD proportions. This observation verifies our primary assumption that the objective misalignment between confidence-based cache partitioning and margin-driven negative selection is the key problem causing Near-OOD detection failure under imbalanced test environments.

4.3 Ablation Studies

To investigate the individual contributions of our framework's components, we conduct comprehensive ablation studies on the negative ranking disparity formulations and the graph neighborhood size k. Tab. 2 reports the empirical evaluation across standard and full-spectrum (FS) settings for both Near-OOD and Far-OOD benchmarks.

Effect of disparity scoring strategies. As shown in the first row of Tab. 2, relying exclusively on the Unconstrained Disparity ($\mathbf{R}_1$) yields suboptimal results (78.01% AUROC on Near-OOD), demon-

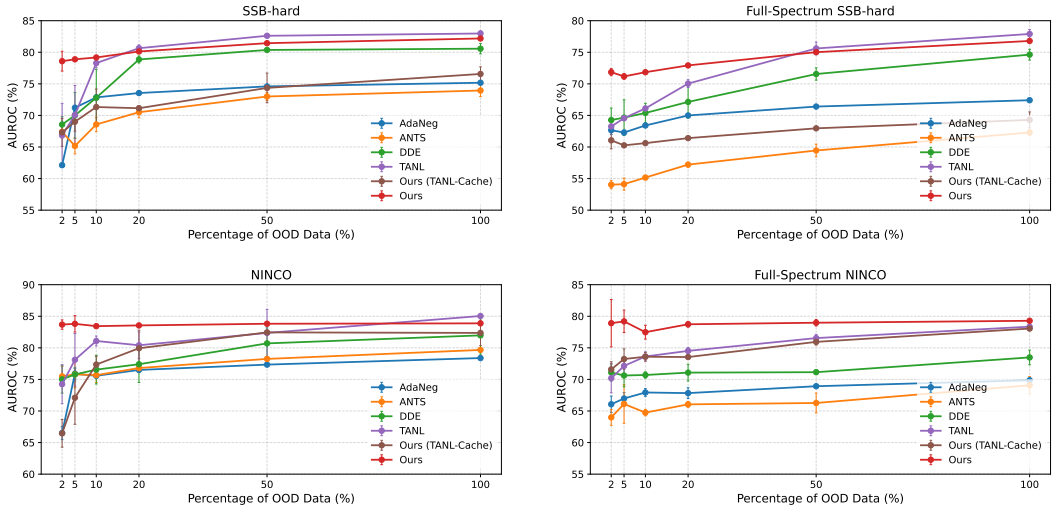


Figure 1: AUROC performance across varying Near-OOD data proportions (from 2% to 100%). While existing test-time adaptation baselines collapse sharply when Near-OOD instances arrive sparsely ($\leq 10\%$), our framework exhibits robust invariance to extreme streaming imbalances and remains highly competitive in balanced settings.

Table 2: Ablation results across disparity scoring formulations and graph neighborhood sizes k on Near-OOD and Far-OOD benchmarks (mean $\pm$ std across 5 runs). ✓ and ✗ indicate the presence or absence of each module.

Components				OOD Types							
Disparity Scores			k	Near-OOD		Far-OOD		Near-OOD-FS		Far-OOD-FS	
$\mathbf{R}_1$	$\mathbf{R}_2$	$\mathbf{R}_3$		AUROC↑	FPR@95↓	AUROC↑	FPR@95↓	AUROC↑	FPR@95↓	AUROC↑	FPR@95↓
✓	✗	✗	70	$78.01_{\pm 0.09}$	$67.28_{\pm 0.68}$	$88.42_{\pm 0.85}$	$38.63_{\pm 1.59}$	$73.09_{\pm 0.60}$	$74.95_{\pm 0.51}$	$86.74_{\pm 1.53}$	$45.29_{\pm 1.33}$
✓	✓	✗	70	$81.09_{\pm 0.29}$	$63.38_{\pm 1.11}$	$92.50_{\pm 0.12}$	$30.96_{\pm 1.52}$	$75.59_{\pm 0.55}$	$\mathbf{71.19}_{\pm 0.59}$	$90.74_{\pm 0.25}$	$36.95_{\pm 0.45}$
✓	✗	✓	70	$81.10_{\pm 0.30}$	$62.63_{\pm 0.59}$	$92.11_{\pm 0.22}$	$31.80_{\pm 0.49}$	$75.37_{\pm 0.09}$	$72.07_{\pm 0.41}$	$90.56_{\pm 0.18}$	$37.44_{\pm 0.17}$
✗	✓	✓	70	$81.35_{\pm 0.21}$	$\mathbf{62.01}_{\pm 0.89}$	$92.49_{\pm 0.09}$	$31.32_{\pm 0.34}$	$75.07_{\pm 0.72}$	$71.40_{\pm 1.44}$	$89.83_{\pm 0.25}$	$40.14_{\pm 1.36}$
✓	✓	✓	30	$80.66_{\pm 0.16}$	$64.20_{\pm 0.64}$	$92.53_{\pm 0.03}$	$\mathbf{30.21}_{\pm 0.67}$	$74.84_{\pm 0.35}$	$71.82_{\pm 0.23}$	$90.51_{\pm 0.14}$	$37.84_{\pm 0.39}$
✓	✓	✓	150	$81.10_{\pm 0.10}$	$63.22_{\pm 0.51}$	$92.50_{\pm 0.20}$	$31.52_{\pm 0.79}$	$75.60_{\pm 0.29}$	$70.26_{\pm 0.45}$	$90.56_{\pm 0.24}$	$37.29_{\pm 0.43}$
✓	✓	✓	300	$80.09_{\pm 0.36}$	$65.49_{\pm 0.65}$	$91.62_{\pm 0.15}$	$35.04_{\pm 0.84}$	$75.32_{\pm 0.11}$	$71.14_{\pm 0.84}$	$90.22_{\pm 0.12}$	$38.27_{\pm 0.77}$
✓	✓	✓	70	$\mathbf{81.49}_{\pm 0.36}$	$62.36_{\pm 0.79}$	$\mathbf{92.54}_{\pm 0.16}$	$30.74_{\pm 0.63}$	$\mathbf{75.61}_{\pm 0.47}$	$71.44_{\pm 0.96}$	$\mathbf{90.58}_{\pm 0.19}$	$\mathbf{37.26}_{\pm 0.23}$

strating that unweighted topological connectivity alone cannot reliably select effective boundary negatives. Introducing either the Topological Global/Instance Disparity ($\mathbf{R}_{2/3}$) improves Near-OOD AUROC by over **3.0**% and Far-OOD AUROC by over **4.0**%. Fusing all three formulations ($\mathbf{R}_1 + \mathbf{R}_2 + \mathbf{R}_3$) achieves the strongest overall discriminability, reaching 81.49% AUROC on Near-OOD and 75.61% on Near-OOD-FS.

Sensitivity to neighborhood size k. We also examine sensitivity to the graph neighborhood parameter $k \in \{30, 70, 150, 300\}$ in Tab. 2. While performance remains generally stable, setting $k = 30$ overly constrains graph connectivity and discards informative boundary candidates, lowering Near-OOD AUROC to 80.66%. Conversely, an excessively dense graph ($k = 300$) admits spurious cross-modal edges, degrading discriminability across all benchmarks (e.g., FPR@95 rises to 65.49% on Near-OOD and 35.04% on Far-OOD). An intermediate value of $k = 70$ reliably balances local boundary precision with topological completeness.

4.4 Semantic Alignment of Selected Negative Labels

To uncover the geometric mechanism governing the efficacy of our framework under imbalanced Near-OOD scenario (for Far-OOD, analysis are detailed in Appx. A.3), we analyze the semantic alignment of mined negative labels against ground truth instances across varying streaming ratios.

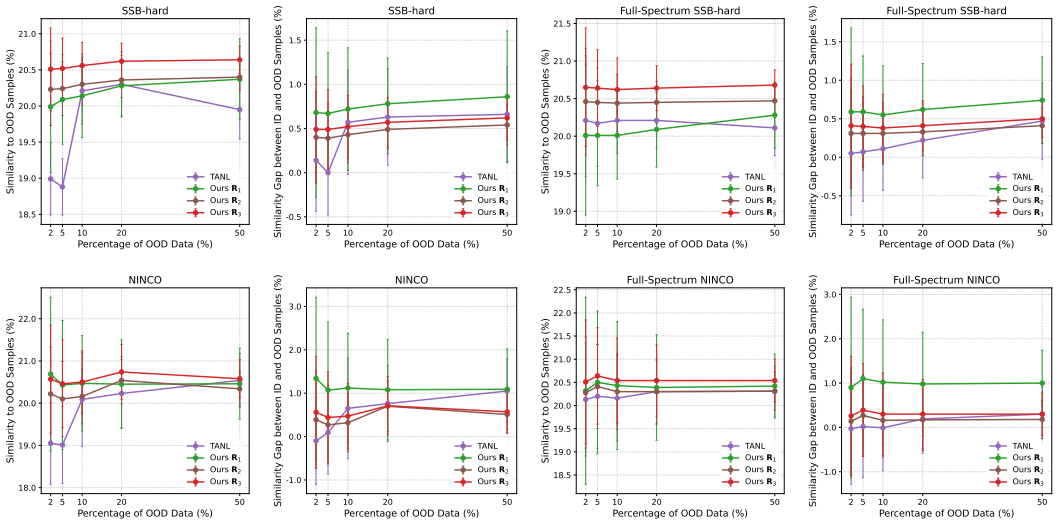


Figure 2: Semantic alignment of selected negative labels across varying OOD data proportions from 2% to 50% on Near-OOD benchmarks and their full-spectrum variants. Columns 1 and 3 display the average cosine similarity between selected negative labels and ground truth OOD images. Columns 2 and 4 show the similarity margin gap between OOD and ID representations. While TANL undergoes severe semantic degradation under extreme streaming imbalance, our graph-driven selections consistently preserve high OOD affinity and sustain robust separation margins.

Specifically, we evaluate two metrics over the test stream: first, the average cosine similarity between selected negative labels and ground truth OOD images, and second, the margin separation gap obtained by subtracting the similarity to ID images from the similarity to OOD images.

As visualized in Fig. 2, our findings reveal two critical phenomena. First, across both SSB hard and NINCO, as well as their full-spectrum counterparts, our visual modulated dynamic strategy (Ours $\mathbf{R}_3$) consistently achieves the highest cosine similarity to true OOD queries. It outperforms TANL as well as its negative label selection under our cache update (Ours $\mathbf{R}_1$) across all subsampling ratios, validating that TANL's Unconstrained Disparity fails to select suitable negative labels for Near-OOD samples and that cache update misalignment further worsens this issue. Second, examining the similarity gap, TANL margin drops close to zero or turns negative at low OOD proportions, such as reaching negative gaps on SSB hard and full-spectrum benchmarks at 2% OOD arrival. A negative or vanishing gap implies that the mined negative labels align as strongly with peripheral ID samples as with true anomalies, providing experimental proof for our mechanism analysis in Appx. A.1 that TANL tends to select easy negative labels located far away from both ID and OOD test samples under an imbalanced Near OOD scenario. In contrast to TANL, our graph driven strategies ($\mathbf{R}_1$, $\mathbf{R}_2$, $\mathbf{R}_3$) maintain strictly positive and wide separation margins across all imbalance regimes, with the advantage becoming especially prominent under low OOD proportions.

5 Conclusion

In this work, we identified a critical failure mode in test time Near-OOD detection under streaming imbalance, demonstrating that heuristic confidence thresholding causes severe objective misalignment by discarding local manifold geometry. To address this dilemma, we established the mathematical equivalence between dual-tower contrastive alignment and message passing over bipartite graphs, proposing a training-free framework that reconstructs this semantic topology via mutual k nearest neighbor matching. By introducing topological disparity metrics alongside a sub bipartite topological connectivity gate, our method successfully aligns cache admission with boundary separation. Comprehensive evaluations across extensive benchmarks confirm that our approach effectively prevents boundary collapse, consistently delivering substantial gains under severe streaming imbalances while preserving competitive performance in balanced settings.

AI USE STATEMENT

During the preparation of this manuscript, generative AI tools were used solely for language editing and polishing to improve readability. Generative AI was not used for "generate synthetic data sets", "help develop theoretical models or conceptual frameworks", "formulate mathematical claims", "provide critical ingredients for proving mathematical claims", "assist in the writing of proofs", "propose or refine hypotheses", "design or provide feedback on research methodology or experiments", "implement methods", "assist with translation", "clean and reformat dataset", "support qualitative and thematic data analysis", or "interpret results". We have reviewed and edited all generated content and take full responsibility for the final manuscript, including its integrity, accuracy, and claims.

REFERENCES

Amirhossein Ansari, Ke Wang, and Pulei Xiong. Negrefine: Refining negative label-based zero-shot ood detection. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 573–582. IEEE, 2025.

Steven Basart, Mazeika Mantas, Mostajabi Mohammadreza, Steinhardt Jacob, and Song Dawn. Scaling out-of-distribution detection for real-world settings. In *International Conference on Machine Learning*, 2022.

Julian Bitterwolf, Maximilian Mueller, and Matthias Hein. In or out? fixing imagenet out-of-distribution detection evaluation. *arXiv preprint arXiv:2306.00826*, 2023.

Chentao Cao, Zhun Zhong, Zhanke Zhou, Yang Liu, Tongliang Liu, and Bo Han. Envisioning outlier exposure by large language models for out-of-distribution detection. *arXiv preprint arXiv:2406.00806*, 2024.

Mengyuan Chen, Junyu Gao, and Changsheng Xu. Conjugated semantic pool improves ood detection with pre-trained vision-language models. *Advances in Neural Information Processing Systems*, 37:82560–82593, 2024.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3606–3613, 2014.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, 29(6):141–142, 2012.

Sepideh Esmaeilpour, Bing Liu, Eric Robertson, and Lei Shu. Zero-shot out-of-distribution detection based on the pre-trained model clip. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 6568–6576, 2022.

Ke Fan, Tong Liu, Xingyu Qiu, Yikai Wang, Lian Huai, Zeyu Shangguan, Shuang Gou, Fengjian Liu, Yuqian Fu, Yanwei Fu, et al. Test-time linear out-of-distribution detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23752–23761. IEEE, 2024.

Angelos Filos, Panagiotis Tigkas, Rowan McAllister, Nicholas Rhinehart, Sergey Levine, and Yarin Gal. Can autonomous vehicles identify, recover from, and adapt to distribution shifts? In *International Conference on Machine Learning*, pp. 3145–3153. PMLR, 2020.

Wei Fu, Yufei Chen, Wei Liu, Xiaodong Yue, and Chao Ma. Evidence reconciled neural network for out-of-distribution detection in medical images. In *International conference on medical image computing and computer-assisted intervention*, pp. 305–315. Springer, 2023.

Zhitong Gao, Shipeng Yan, and Xuming He. Atta: Anomaly-aware test-time adaptation for out-of-distribution detection in segmentation. *Advances in Neural Information Processing Systems*, 36: 45150–45171, 2023.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8320–8329. IEEE, 2021.

Rui Huang and Yixuan Li. Mos: Towards scaling out-of-distribution detection for large semantic space. In *2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 8706–8715. IEEE, 2021.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

Xue Jiang, Feng Liu, Zhen Fang, Hong Chen, Tongliang Liu, Feng Zheng, and Bo Han. Negative label guided ood detection with pretrained vision-language models. In *International Conference on Learning Representations*, volume 2024, pp. 14913–14941, 2024.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Fayi Le, Wenwu He, Chentao Cao, Dong Liang, and Zhuo-Xu Cui. Dualcnst: enhancing zero-shot out-of-distribution detection via text-image consistency in vision-language models. *Advances in Neural Information Processing Systems*, 38:154226–154265, 2026.

Yann Le, Xuan Yang, et al. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

Jinglun Li, Xinyu Zhou, Kaixun Jiang, Lingyi Hong, Pinxue Guo, Zhaoyu Chen, Weifeng Ge, and Wenqiang Zhang. Tagood: A novel approach to out-of-distribution detection via vision-language representations and class center learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 1225–1234, 2024.

Jian Liang, Ran He, and Tieniu Tan. A comprehensive survey on test-time adaptation under distribution shifts. *International Journal of Computer Vision*, 133(1):31–64, 2025.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020.

Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. Delving into out-of-distribution detection with vision-language representations. *Advances in neural information processing systems*, 35:35087–35102, 2022.

Atsuyuki Miyai, Qing Yu, Go Irie, and Kiyoharu Aizawa. Locoop: Few-shot out-of-distribution detection via prompt learning. *Advances in Neural Information Processing Systems*, 36:76298–76310, 2023.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, pp. 4. Granada, 2011.

Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 427–436, 2015.

Bo Peng, Jie Lu, Guangquan Zhang, and Zhen Fang. An information-theoretical framework for understanding out-of-distribution detection with pretrained vision-language models. *Advances in Neural Information Processing Systems*, 38:1930–1957, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pp. 5389–5400. PMLR, 2019.

Walter J Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E Boult. Toward open set recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(7): 1757–1772, 2012.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International conference on machine learning*, pp. 20827–20840. PMLR, 2022.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 8769–8778. IEEE, 2018.

Maria De La Iglesia Vayá, Jose Manuel Saborit, Joaquim Angel Montell, Antonio Pertusa, Aurelia Bustos, Miguel Cazorla, Joaquin Galant, Xavier Barber, Domingo Orozco-Beltrán, Francisco García-García, et al. Bimcv covid-19+: a large annotated dataset of rx and ct images from covid-19 patients. *arXiv preprint arXiv:2006.01174*, 2020.

Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need. In *International conference on learning representations*, 2022.

Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4911–4920. IEEE, 2022.

Hualiang Wang, Yi Li, Huifeng Yao, and Xiaomeng Li. Clipn for zero-shot ood detection: Teaching clip to say no. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1802–1812. IEEE, 2023.

Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, et al. Openood: Benchmarking generalized out-of-distribution detection. *Advances in Neural Information Processing Systems*, 35:32598–32611, 2022.

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

Yifeng Yang, Lin Zhu, Zewen Sun, Hengyu Liu, Qinying Gu, and Nanyang Ye. Oodd: Test-time out-of-distribution detection with dynamic dictionary. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30630–30639. IEEE, 2025.

Boxuan Zhang, Jianing Zhu, Zengmao Wang, Tongliang Liu, Bo Du, and Bo Han. What if the input is expanded in ood detection? *Advances in Neural Information Processing Systems*, 37: 21289–21329, 2024a.

Jingyang Zhang, Jingkang Yang, Pengyun Wang, Haoqi Wang, Yueqian Lin, Haoran Zhang, Yiyou Sun, Xuefeng Du, Yixuan Li, Ziwei Liu, et al. Openood v1. 5: Enhanced benchmark for out-of-distribution detection. *arXiv preprint arXiv:2306.09301*, 2023.

Yabin Zhang and Lei Zhang. Adaneg: Adaptive negative proxy guided ood detection with vision-language models. *Advances in Neural Information Processing Systems*, 37:38744–38768, 2024.

Yabin Zhang, Wenjie Zhu, Chenhang He, and Lei Zhang. Lapt: Label-driven automated prompt tuning for ood detection with vision-language models. In *European conference on computer vision*, pp. 271–288. Springer, 2024b.

Yabin Zhang, Maya Varma, Yunhe Gao, Jean-Benoit Delbrouck, Jiaming Liu, Chong Wang, and Curtis Langlotz. Activation matters: Test-time activated negative labels for ood detection with vision-language models. *arXiv preprint arXiv:2603.25250*, 2026.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.

Wenjie Zhu, Yabin Zhang, Xin Jin, Wenjun Zeng, and Lei Zhang. Ants: Adaptive negative textual space shaping for ood detection via test-time mllm understanding and reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20–30, 2026a.

Wenjie Zhu, Yabin Zhang, Liang Xu, Xin Jin, Wenjun Zeng, and Lei Zhang. Dual distribution estimation for zero-shot noisy test-time adaptation with vlms. *arXiv preprint arXiv:2606.25758*, 2026b.

A APPENDIX

A.1 MECHANISTIC ANALYSIS

In this section, we provide the formal mechanistic analysis previewed in the main text, detailing the root cause of cache degradation under streaming imbalance and analytically verifying our structural remedy. We focus our analysis on the Near-OOD setting, as Far-OOD instances reside well outside the ID semantic manifold and remain relatively stable to such distributional shifts. Specifically, we examine: (1) why cache construction should prioritize maximizing the geometric separation gap rather than pursuing confidence-based label purity, and how variance thresholding triggers a degenerate feedback loop under streaming imbalance ($N_{\text{id}} \gg N_{\text{ood}}$) while remaining stable in balanced regimes ($N_{\text{id}} \approx N_{\text{ood}}$); and (2) how our mutual graph topology and sub-bipartite topological connectivity gate mitigate this failure by partitioning visual caches based on high-dimensional geometric structure rather than prediction confidence.

Geometric Decomposition. Let $\mathcal{M}_{\text{id}} \subset \mathbb{S}^{D-1}$ denote the compact support manifold of ID visual embeddings with boundary $\partial\mathcal{M}_{\text{id}}$, and let $\Omega_\epsilon = \{\mathbf{v} \in \mathbb{S}^{D-1} \mid \text{dist}(\mathbf{v}, \partial\mathcal{M}_{\text{id}}) < \epsilon\}$ define its boundary neighborhood. In the Near-OOD evaluation regime, all OOD test instances originate from semantically related exterior distributions, their visual features reside in the exterior shell $\mathbf{v}_{\text{near}} \in \Omega_\epsilon \setminus \mathcal{M}_{\text{id}}$, differing primarily in their relative distance to the boundary $\partial\mathcal{M}_{\text{id}}$. We partition the candidate negative corpus $\mathbf{T}_{\text{neg}}$ into boundary-sensitive hard negatives $\mathcal{T}_{\text{hard}}$ (concepts sharing partial semantic attributes with classes along $\partial\mathcal{M}_{\text{id}}$) and distant coarse negatives $\mathcal{T}_{\text{coarse}} = \mathbf{T}_{\text{neg}} \setminus \mathcal{T}_{\text{hard}}$ ($|\mathcal{T}_{\text{coarse}}| \gg |\mathcal{T}_{\text{hard}}|$).

A.1.1 THE OBJECTIVE OF CACHE EVOLUTION: GAP MAXIMIZATION VS. CONFIDENCE PARTITIONING

In test-time OOD detection, visual caches $\mathcal{X}_{\text{pos}}$ and $\mathcal{X}_{\text{neg}}$ do not merely serve as static ground-truth storage. Instead, they act as opposing operators parameterizing the directional disparity score $\mathbf{R}_1(\tilde{\mathbf{t}})$:

$$\mathbf{R}_1(\tilde{\mathbf{t}}) = \sum_{\mathbf{v} \in \mathcal{X}_{\text{neg}}} \mathbf{v}^\top \tilde{\mathbf{t}} - \sum_{\mathbf{v} \in \mathcal{X}_{\text{pos}}} \mathbf{v}^\top \tilde{\mathbf{t}}. \tag{20}$$

However, recent methods update $\mathcal{X}_{\text{pos}}$ and $\mathcal{X}_{\text{neg}}$ to pursue nominal label purity within the caches rather than maximizing the functional separation margin between ID and Near-OOD representations. While this distinction is inconsequential in balanced test environments, it becomes decisive under streaming imbalance.

Unimodal Collapse under Extreme Streaming Imbalance ($N_{\text{id}} \gg N_{\text{ood}}$). Prior methods (e.g., TANL (Zhang et al., 2026)) govern cache admission by minimizing within-cluster variance over the confidence scores of historical test samples $\mathcal{V}_{\text{hist}}$:

$$\gamma^* = \arg\min_{\gamma} \left(\text{var}(\mathcal{O}^+) + \text{var}(\mathcal{O}^-) \right), \tag{21}$$

where $\mathcal{O}^+ = \{S_{\text{nl}}(\mathbf{v}) \mid \mathbf{v} \in \mathcal{V}_{\text{hist}},\ S_{\text{nl}}(\mathbf{v}) \geq \gamma\}$ and $\mathcal{O}^- = \{S_{\text{nl}}(\mathbf{v}) \mid \mathbf{v} \in \mathcal{V}_{\text{hist}},\ S_{\text{nl}}(\mathbf{v}) < \gamma\}$. Once γ^* is determined, samples in the current test batch with $S_{\text{nl}}(\mathbf{v}) \geq \gamma^*$ and $S_{\text{nl}}(\mathbf{v}) < \gamma^*$ are routed to $\mathcal{X}_{\text{pos}}$ and $\mathcal{X}_{\text{neg}}$, respectively. When ID arrivals heavily dominate the stream, the historical score distribution degenerates into a single continuous unimodal density over ID instances. This confidence thresholding routes peripheral ID instances into the positive cache. Consequently, their visual similarity to boundary concepts inflates the subtractive penalty against hard negatives in Eq. 11. Deprived of sufficient Near-OOD arrivals in $\mathcal{X}_{\text{neg}}$ to provide counterbalancing positive alignment, informative hard negatives are systematically displaced from selection:

- Peripheral ID instances located along the boundary ($\mathbf{v}_{\text{edge}} \in \Omega_\epsilon \cap \mathcal{M}_{\text{id}}$) retain relatively high confidence and are systematically placed into $\mathcal{X}_{\text{pos}}$. Since $\mathbf{v}_{\text{edge}}$ shares visual attributes with hard negative concepts $\tilde{\mathbf{t}}_{\text{hard}}$, the empirical mean similarity over the positive cache remains elevated:

$$\mathbf{R}_{1,\text{pos}}(\tilde{\mathbf{t}}_{\text{hard}}) = \frac{1}{L_{\text{pos}}} \sum_{\mathbf{v} \in \mathcal{X}_{\text{pos}}} \mathbf{v}^\top \tilde{\mathbf{t}}_{\text{hard}} \gg 0. \tag{22}$$

- Meanwhile, since OOD arrivals are extremely sparse ($N_{\text{ood}} \ll N_{\text{id}}$), the negative cache $\mathcal{X}_{\text{neg}}$ rarely receives instances from the immediate boundary exterior. Instead, scalar confidence filtering populates $\mathcal{X}_{\text{neg}}$ predominantly with low-confidence ID instances or the subset of Near-OOD instances that reside farther from the boundary $\partial\mathcal{M}_{\text{id}}$. For these peripheral and distant instances, semantic overlap with the fine-grained boundary negative labels $\tilde{\mathbf{t}}_{\text{hard}}$ is minimal:

$$\mathbf{R}_{1,\text{neg}}(\tilde{\mathbf{t}}_{\text{hard}}) = \frac{1}{L_{\text{neg}}} \sum_{\mathbf{v} \in \mathcal{X}_{\text{neg}}} \mathbf{v}^\top \tilde{\mathbf{t}}_{\text{hard}} \ll \mathbf{R}_{1,\text{pos}}(\tilde{\mathbf{t}}_{\text{hard}}). \tag{23}$$

For an uninformative coarse negative $\tilde{\mathbf{t}}_{\text{coarse}} \in \mathcal{T}_{\text{coarse}}$, its semantic distance from ID classes ensures that $\mathbf{R}_{1,\text{pos}}(\tilde{\mathbf{t}}_{\text{coarse}}) \approx 0$. Furthermore, because the instances populating $\mathcal{X}_{\text{neg}}$ reside far from the boundary $\partial\mathcal{M}_{\text{id}}$, their alignment with coarse concepts $\mathbf{R}_{1,\text{neg}}(\tilde{\mathbf{t}}_{\text{coarse}})$ remains elevated compared to their alignment with fine-grained hard negative concepts $\mathbf{R}_{1,\text{neg}}(\tilde{\mathbf{t}}_{\text{hard}})$. Consequently, the resulting disparity satisfies:

$$\mathbf{R}_{1,\text{neg}}(\tilde{\mathbf{t}}_{\text{hard}}) - \mathbf{R}_{1,\text{pos}}(\tilde{\mathbf{t}}_{\text{hard}}) < \mathbf{R}_{1,\text{neg}}(\tilde{\mathbf{t}}_{\text{coarse}}) - \mathbf{R}_{1,\text{pos}}(\tilde{\mathbf{t}}_{\text{coarse}}). \tag{24}$$

Assuming normalized cache capacities $L_{\text{pos}} = L_{\text{neg}} = L$, this inequality maps directly to an inversion of ranking scores:

$$\mathbf{R}_1(\tilde{\mathbf{t}}_{\text{hard}}) < \mathbf{R}_1(\tilde{\mathbf{t}}_{\text{coarse}}). \tag{25}$$

Consequently, coarse negative labels consistently outrank hard negatives, systematically displacing boundary-informative concepts from the top-β retrieved subset ($\mathcal{Y}^- \subset \mathcal{T}_{\text{coarse}}$).

Margin Collapse and the Feedback Cycle. When coarse negatives dominate $\mathcal{Y}^-$, cross-modal differentiation vanishes, causing the negative denominator sum in Eq. 2 to yield indistinguishable values for boundary queries:

$$\sum_{m=1}^{\beta} \exp(\tau \mathbf{v}_{\text{near}}^\top \tilde{\mathbf{t}}_m^{\text{coarse}}) \approx \sum_{m=1}^{\beta} \exp(\tau \mathbf{v}_{\text{edge}}^\top \tilde{\mathbf{t}}_m^{\text{coarse}}) \implies S_{\text{nl}}(\mathbf{v}_{\text{near}}) \approx S_{\text{nl}}(\mathbf{v}_{\text{edge}}). \tag{26}$$

Since the score difference between true Near-OOD instances ($\mathbf{v}_{\text{near}}$) and boundary ID instances ($\mathbf{v}_{\text{edge}}$) contracts toward zero:

$$\Delta S = |S_{\text{nl}}(\mathbf{v}_{\text{edge}}) - S_{\text{nl}}(\mathbf{v}_{\text{near}})| \to 0, \tag{27}$$

the separation margin collapses, leading to the misclassification of Near-OOD instances as ID. In subsequent iterations, confidence partitioning admits these false positives into $\mathcal{X}_{\text{pos}}$, reinforcing $\mathbf{R}_{1,\text{pos}}(\tilde{\mathbf{t}}_{\text{hard}})$ and locking the system into a persistent cycle that continually selects coarse negatives.

Transition to Balanced Streams ($N_{\text{id}} \approx N_{\text{ood}}$). As the stream balances, the historical score density resolves back into a distinct bimodal distribution. Minimizing variance places γ^* into the low-density valley between the ID and Near-OOD modes, allowing sufficient purer ID and Near-OOD instances to enter $\mathcal{X}_{\text{pos}}$ and $\mathcal{X}_{\text{neg}}$, respectively. This structural clarification shifts all three terms simultaneously:

1. $\mathbf{R}_{1,\text{neg}}(\tilde{\mathbf{t}}_{\text{hard}})$ increases, as authentic Near-OOD boundary samples replenish $\mathcal{X}_{\text{neg}}$ and provide strong cross-modal alignment.
2. $\mathbf{R}_{1,\text{pos}}(\tilde{\mathbf{t}}_{\text{hard}})$ decreases, as $\mathcal{X}_{\text{pos}}$ collects purer ID instances deeper in the interior of $\mathcal{M}_{\text{id}}$ rather than ambiguous boundary samples.
3. $\mathbf{R}_{1,\text{neg}}(\tilde{\mathbf{t}}_{\text{coarse}})$ decreases, because replacing distant outliers with boundary-focused Near-OOD instances reduces spurious alignments with coarse concepts.

Consequently, the previous disparity inequality is reversed:

$$\mathbf{R}_{1,\text{neg}}(\tilde{\mathbf{t}}_{\text{hard}}) - \mathbf{R}_{1,\text{pos}}(\tilde{\mathbf{t}}_{\text{hard}}) \geq \mathbf{R}_{1,\text{neg}}(\tilde{\mathbf{t}}_{\text{coarse}}) - \mathbf{R}_{1,\text{pos}}(\tilde{\mathbf{t}}_{\text{coarse}}). \tag{28}$$

Even if hard negatives do not outrank every single coarse candidate across the corpus, this positive shift forms a localized signal peak that preserves their inclusion in the top-β set, allowing baseline methods to expand the denominator in Eq. 2 under balanced conditions.

A.1.2 Structure-Guided Cache Allocation to Enlarge the Separation Gap

Rather than relying on 1D scalar confidence or attempting strict ground-truth classification, our framework governs cache updates directly through geometric proximity to active semantic subspaces:

$$\bar{E}_{\text{id}}(\mathbf{v}) = \frac{1}{C}\sum_{c=1}^{C}(\mathbf{A}_{\text{sub}} \odot \mathbf{W}_{\text{sub}})_c, \quad \bar{E}_{\text{neg}}(\mathbf{v}) = \frac{1}{N_{\text{active}}}\sum_{j=1}^{N_{\text{active}}}(\mathbf{A}_{\text{sub}} \odot \mathbf{W}_{\text{sub}})_{C+j}. \tag{29}$$

Under this criterion, an incoming sample is admitted into $\mathcal{X}_{\text{neg}}$ if and only if its local topological connectivity toward the active negative text pool exceeds that toward ID classes ($\bar{E}_{\text{neg}}(\mathbf{v}) > \bar{E}_{\text{id}}(\mathbf{v})$).

Geometry-Driven Subspace Partitioning. This connectivity comparison acts as a structural distance filter:

1. **Concentrating Boundary Support in $\mathcal{X}_{\text{neg}}$.** Samples routed to $\mathcal{X}_{\text{neg}}$, whether they are true Near-OOD arrivals or boundary ID instances with outward feature orientations, are explicitly selected because they reside geometrically closer to the hard negative subspace $\mathbf{T}_{\text{neg}}^{\text{active}}$ than to the ID center. In the absence of abundant OOD data, these boundary-proximal instances provide the necessary localized visual support inside $\mathcal{X}_{\text{neg}}$, directly driving up $\mathbf{R}_{1,\text{neg}}(\tilde{\mathbf{t}}_{\text{hard}})$.
2. **Constraining $\mathcal{X}_{\text{pos}}$ to the Invariant ID Subspace.** Conversely, admission into $\mathcal{X}_{\text{pos}}$ requires unambiguous geometric dominance toward ID prototypes ($\bar{E}_{\text{id}}(\mathbf{v}) > \bar{E}_{\text{neg}}(\mathbf{v})$). Instances satisfying this condition lie deeper in the interior of the ID manifold, systematically reducing the empirical affinity between the positive cache and boundary negative labels ($\mathbf{R}_{1,\text{pos}}(\tilde{\mathbf{t}}_{\text{hard}})$ decreases).
3. **Inherent Suppression of Coarse Negatives.** Since coarse negatives $\mathcal{T}_{\text{coarse}}$ remain semantically distant from both the ID manifold and the boundary region, their geometric distances to samples in both $\mathcal{X}_{\text{pos}}$ and $\mathcal{X}_{\text{neg}}$ remain consistently large, preventing them from benefiting from this boundary-focused update.

Restoration of the Selection Margin and Denominator Stability. By explicitly aligning cache membership with cross-modal geometric distance rather than prediction confidence, our update drives $\mathbf{R}_{1,\text{neg}}(\tilde{\mathbf{t}}_{\text{hard}})$ up while suppressing $\mathbf{R}_{1,\text{pos}}(\tilde{\mathbf{t}}_{\text{hard}})$. This preserves the relative disparity margin even under severe streaming imbalance:

$$\mathbf{R}_{1,\text{neg}}(\tilde{\mathbf{t}}_{\text{hard}}) - \mathbf{R}_{1,\text{pos}}(\tilde{\mathbf{t}}_{\text{hard}}) > \mathbf{R}_{1,\text{neg}}(\tilde{\mathbf{t}}_{\text{coarse}}) - \mathbf{R}_{1,\text{pos}}(\tilde{\mathbf{t}}_{\text{coarse}}). \tag{30}$$

Furthermore, our proposed graph-based disparity scores $\mathbf{R}_2$ and $\mathbf{R}_3$ naturally filter out coarse negative labels that lack dense mutual connections with cached visual features. As a result, hard negative labels maintain competitive rankings and are reliably retained in the active candidate subset ($\tilde{\mathbf{t}}_{\text{hard}} \in \mathcal{Y}^-$). Crucially, retaining hard negative labels in $\mathcal{Y}^-$ prevents the negative denominator sum in Eq. 2 from collapsing into uninformative background noise, ensuring that the negative regularizer functions stably whether the incoming stream is balanced or heavily skewed.

A.2 Far-OOD Results on the ImageNet-1k

We report the empirical evaluation on Far-OOD benchmarks across varying streaming proportions in Fig. 3. Compared to the Near-OOD regime, performance across all methods remains consistently elevated on Far-OOD data ($> 90\%$ AUROC), as Far-OOD instances reside semantically distant from the ID manifold and are less prone to distributional confusion. However, our framework exhibits a moderate performance trade-off relative to baselines, which stems from two primary factors:

1. **Filtering of Coarse Negatives.** Far-OOD detection fundamentally benefits from coarse, globally dispersed negative labels. By construction, our mutual k-NN bipartite graph preferentially preserves localized boundary edges with high semantic proximity, naturally discarding semantically distant negative labels that would otherwise provide broad, uniform suppression against distant Far-OOD queries.
2. **Cache Disparity Under Boundary Candidates.** Since our graph-guided negative label selection concentrates exclusively on hard boundary negatives, ID instances exhibit

non-negligible topological affinity toward these selected candidates compared to distant concepts. Consequently, sub-bipartite topological connectivity gating operates under a narrower margin, whereas simple confidence thresholding easily isolates Far-OOD samples due to their vanishing predicted likelihoods.

Notably, when our topological cache update is replaced with TANL's confidence thresholding (*Ours (TANL-Cache)*), Far-OOD performance improves across all proportions, verifying that confidence partitioning remains effective primarily when semantic gaps are sufficiently large. Nevertheless, since Far-OOD anomalies are distinctly separated and easily identifiable, the marginal drop of approximately 1%–2% AUROC in the Far-OOD regime is a justifiable trade-off for the substantial gains ($+10\%$ AUROC) achieved under the more critical and practically Near-OOD streaming imbalance.

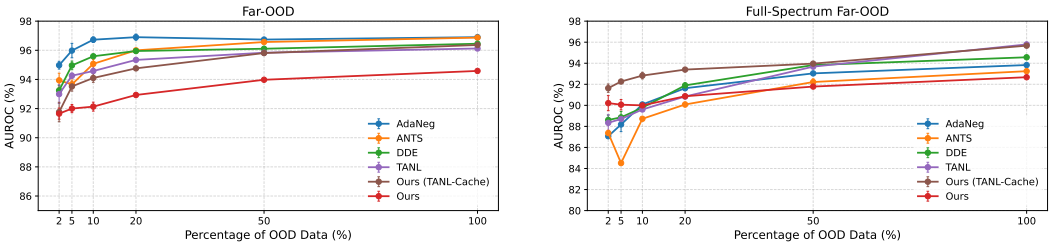


Figure 3: AUROC performance across varying Far-OOD data proportions (from 2% to 100%). While prioritizing localized boundary negatives incurs a slight trade-off on distant Far-OOD instances, all methods retain high discriminability ($> 90\%$ AUROC), demonstrating that our design appropriately optimizes for the more challenging Near-OOD failure mode.

A.3 Semantic Alignment of Selected Negative Labels

To thoroughly investigate our framework behavior across distributional regimes, we evaluate the semantic alignment and margin separation of selected negative labels on Far-OOD benchmarks: iNaturalist, Textures, and OpenImage-O, alongside their full-spectrum variants. Fig. 4 illustrates the average cosine similarity to true OOD images as well as the similarity gap between ID and OOD samples across streaming proportions.

Analysis of Far-OOD Alignment and Inherent Limitations. As shown in Fig. 4, our refined strategies ($\mathbf{R}_2$ and $\mathbf{R}_3$) display lower absolute cosine similarity to true Far-OOD samples compared to TANL when OOD arrival rates increase, which consequently yields smaller margin gaps on datasets such as Textures and OpenImage-O. This empirical observation highlights a clear design trade-off in our selection mechanism. By design, our bipartite graph preferentially captures hard boundary negatives that share localized mutual connections with positive classes. Since Far-OOD instances are situated far away from the ID manifold, our model prioritizes negative labels near the decision boundary rather than broad, globally dispersed concepts. This proximity to ID boundaries narrows the relative gap against true Far-OOD samples.

Mitigation via Unconstrained Disparity and Future Outlook. Notably, the Unconstrained Disparity ($\mathbf{R}_1$) mitigates this limitation, maintaining higher similarity and wider separation margins across multiple benchmarks, such as iNaturalist and full-spectrum iNaturalist. This indicates that static graph degree connectivity provides broader label coverage that is better suited for coarse distribution shifts. Consequently, it is promising to upweight the confidence score derived from the unweighted formulation S_{nl}^1 under Far-OOD conditions, or to develop an automated meta weighting scheme that dynamically calibrates disparity weights according to estimated distribution shifts. We leave such adaptive weighting schemes as an impactful direction for future research.

In summary, as discussed in Appx. A.2, Far-OOD anomalies remain inherently distinct and readily separable, rendering the minor performance variance of approximately 1% to 2% AUROC an entirely acceptable trade-off for the substantial gains ($+10\%$ AUROC) achieved under the more critical and practically hazardous Near-OOD streaming imbalance.

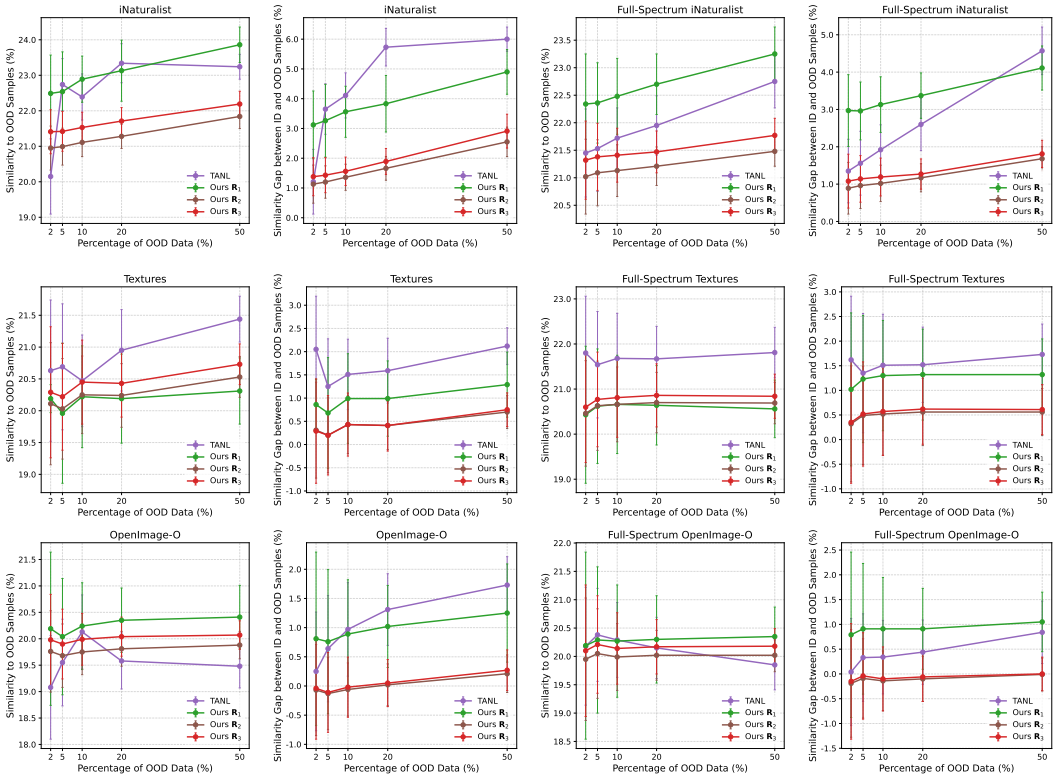


Figure 4: Semantic alignment of selected negative labels across varying OOD data proportions from 2% to 50% on Far-OOD benchmarks and their full-spectrum variants. Columns 1 and 3 report average cosine similarity between selected negative labels and ground truth OOD images. Columns 2 and 4 display the similarity margin gap between OOD and ID representations. Prioritizing localized boundary negative labels narrows the margin gap on distant anomalies, though the unweighted formulation $\mathbf{R}_1$ substantially alleviates this effect across benchmark distributions.

Table 3: Near- and Far-OOD detection performance on ImageNet-1k (mean ± std over 5 runs) un der different backbones. FS denotes the full-spectrum evaluation setting incorporating covariate-shifted ID instances.

Methods	OOD Types							
	Near-OOD		Far-OOD		Near-OOD-FS		Far-OOD-FS	
	AUROC↑	FPR@95↓	AUROC↑	FPR@95↓	AUROC↑	FPR@95↓	AUROC↑	FPR@95↓
ResNet50								
TANL	$71.75_{\pm 1.25}$	$73.18_{\pm 0.98}$	$\mathbf{92.54}_{\pm 0.15}$	$\mathbf{28.13}_{\pm 0.52}$	$62.09_{\pm 0.74}$	$82.70_{\pm 0.23}$	$\mathbf{91.61}_{\pm 0.34}$	$\mathbf{33.11}_{\pm 1.38}$
Ours	$\mathbf{77.89}_{\pm 0.27}$	$\mathbf{67.62}_{\pm 0.53}$	$90.45_{\pm 0.05}$	$34.91_{\pm 0.13}$	$\mathbf{72.16}_{\pm 0.07}$	$\mathbf{75.55}_{\pm 0.30}$	$89.33_{\pm 0.25}$	$39.00_{\pm 0.77}$
VIT-B/32								
TANL	$74.85_{\pm 1.54}$	$70.57_{\pm 1.54}$	$\mathbf{92.92}_{\pm 0.32}$	$\mathbf{27.89}_{\pm 0.27}$	$69.81_{\pm 0.25}$	$77.74_{\pm 0.18}$	$\mathbf{91.26}_{\pm 0.15}$	$\mathbf{33.78}_{\pm 0.76}$
Ours	$\mathbf{80.30}_{\pm 0.17}$	$\mathbf{65.21}_{\pm 0.41}$	$90.62_{\pm 0.45}$	$35.88_{\pm 1.08}$	$\mathbf{75.33}_{\pm 0.25}$	$\mathbf{73.20}_{\pm 0.46}$	$89.45_{\pm 0.21}$	$38.46_{\pm 0.20}$
VIT-B/16								
TANL	$77.47_{\pm 0.76}$	$66.99_{\pm 1.50}$	$\mathbf{94.6}_{\pm 0.22}$	$\mathbf{22.14}_{\pm 1.34}$	$70.66_{\pm 0.25}$	$75.38_{\pm 0.09}$	$\mathbf{92.93}_{\pm 0.15}$	$\mathbf{29.52}_{\pm 0.15}$
Ours	$\mathbf{81.49}_{\pm 0.36}$	$\mathbf{62.36}_{\pm 0.79}$	$92.54_{\pm 0.16}$	$30.74_{\pm 0.63}$	$\mathbf{75.61}_{\pm 0.47}$	$\mathbf{71.44}_{\pm 0.96}$	$90.58_{\pm 0.19}$	$37.26_{\pm 0.23}$
VIT-L/14								
TANL	$80.97_{\pm 0.63}$	$64.31_{\pm 0.59}$	$\mathbf{94.25}_{\pm 0.08}$	$\mathbf{22.79}_{\pm 0.55}$	$75.22_{\pm 0.15}$	$75.04_{\pm 0.14}$	$\mathbf{92.28}_{\pm 0.22}$	$\mathbf{29.95}_{\pm 1.28}$
Ours	$\mathbf{83.19}_{\pm 0.20}$	$\mathbf{58.55}_{\pm 0.83}$	$92.95_{\pm 0.08}$	$29.67_{\pm 0.18}$	$\mathbf{78.93}_{\pm 0.28}$	$\mathbf{65.46}_{\pm 0.79}$	$90.92_{\pm 0.07}$	$35.01_{\pm 0.55}$

A.4 Results with Different VLM Backbones

To assess architectural generalizability, we evaluate our framework alongside the competitive baseline TANL across four distinct pre-trained vision encoders: ResNet-50, ViT-B/32, ViT-B/16, and ViT-L/14. As summarized in Tab. 3, the empirical findings are consistent with our primary conclusions across all architectural variants.

Scaling visual backbone capacity systematically elevates overall detection performance for both approaches, moving from ResNet-50 to ViT-L/14. In particular, Near-OOD AUROC advances from 77.89% to 83.19% for our method and from 71.75% to 80.97% for TANL. Nevertheless, the relative performance characteristics between the two frameworks remain pronounced across all scales. Our framework consistently establishes a decisive lead across every Near-OOD and Near-OOD full-spectrum benchmark, reducing Near-OOD FPR@95 by up to 5.56% on ResNet-50, 5.36% on ViT-B/32, 4.63% on ViT-B/16, and 5.76% on ViT-L/14. Under the covariate-shifted Near-OOD-FS setting, our method widens the margin even further, outperforming TANL by up to 10.07% AUROC on ResNet-50.

Conversely, TANL retains a moderate advantage on Far-OOD benchmarks, achieving higher AUROC values across backbones (for instance, 94.25% versus 92.95% on ViT-L/14). This confirms our analysis in Appx A.2 and Appx A.3: while scaling model capacity enhances representation richness, the fundamental trade-off governed by negative label mining remains invariant. TANL's reliance on broader semantic suppression benefits distant outliers, whereas our graph-guided negative selection focuses on tight boundary mining to resolve challenging, fine-grained Near-OOD ambiguities.

A.5 Results on CIFAR-100 and Medical Dataset

To assess framework generalizability beyond ImageNet scale distributions, we evaluate detection performance on CIFAR-100 Krizhevsky et al. (2009) alongside the clinical chest radiograph COVID-19 dataset Vayá et al. (2020) within the OpenOOD benchmark suite. For the CIFAR-100 setting, CIFAR-10 and TIN Le et al. (2015) serve as Near-OOD datasets exhibiting fine grained semantic overlap, while MNIST Deng (2012), SVHN Netzer et al. (2011), Textures Cimpoi et al. (2014), and Places365 Zhou et al. (2017) constitute Far-OOD benchmarks. For the clinical benchmark, BIMCV Vayá et al. (2020) is designated as the in distribution dataset for discriminating COVID-19 pathology from normal chest radiographs. Near-OOD shifts comprise CT-SCAN images (featuring cross modality lung scans) and X-Ray-Bone radiographs (representing cross anatomical shifts), whereas MNIST, CIFAR-10, Textures, and TIN function as Far-OOD benchmarks.

Table 4: Near- and Far-OOD detection performance on CIFAR-100 (mean ± std over 5 runs). **Bold** denotes the best result and FS denotes the full-spectrum evaluation setting incorporating covariate-shifted ID instances.

Methods	OOD Types							
	Near-OOD		Far-OOD		Near-OOD-FS		Far-OOD-FS	
	AUROC↑	FPR@95↓	AUROC↑	FPR@95↓	AUROC↑	FPR@95↓	AUROC↑	FPR@95↓
NegLabel	$72.46_{\pm 0.07}$	$76.85_{\pm 0.74}$	$79.58_{\pm 0.13}$	$61.70_{\pm 0.49}$	$68.91_{\pm 0.02}$	$84.83_{\pm 0.03}$	$71.08_{\pm 0.01}$	$81.41_{\pm 0.02}$
TANL	$73.61_{\pm 0.45}$	$74.91_{\pm 1.57}$	$82.77_{\pm 0.20}$	$50.86_{\pm 0.30}$	$71.06_{\pm 0.01}$	$84.48_{\pm 0.04}$	$74.58_{\pm 0.02}$	$80.79_{\pm 0.05}$
Ours	$\mathbf{74.18}_{\pm 0.10}$	$\mathbf{74.14}_{\pm 1.27}$	$\mathbf{93.87}_{\pm 0.01}$	$\mathbf{28.66}_{\pm 0.79}$	$\mathbf{73.10}_{\pm 0.04}$	$\mathbf{80.44}_{\pm 0.06}$	$\mathbf{78.45}_{\pm 0.03}$	$\mathbf{75.54}_{\pm 0.10}$

Table 5: Near- and Far-OOD detection performance on COVID-19 (mean ± std over 5 runs). **Bold** denotes the best result and FS denotes the full-spectrum evaluation setting incorporating covariate-shifted ID instances.

Methods	OOD Types							
	Near-OOD		Far-OOD		Near-OOD-FS		Far-OOD-FS	
	AUROC↑	FPR@95↓	AUROC↑	FPR@95↓	AUROC↑	FPR@95↓	AUROC↑	FPR@95↓
NegLabel	$99.20_{\pm 0.15}$	$4.41_{\pm 0.49}$	$\mathbf{100.00}_{\pm 0.00}$	$\mathbf{0.00}_{\pm 0.00}$	$98.78_{\pm 0.21}$	$5.39_{\pm 1.11}$	$\mathbf{100.00}_{\pm 0.00}$	$\mathbf{0.00}_{\pm 0.00}$
TANL	$97.84_{\pm 0.37}$	$9.85_{\pm 1.81}$	$\mathbf{100.00}_{\pm 0.00}$	$\mathbf{0.00}_{\pm 0.00}$	$97.94_{\pm 0.87}$	$9.97_{\pm 1.76}$	$\mathbf{100.00}_{\pm 0.00}$	$\mathbf{0.00}_{\pm 0.00}$
Ours	$\mathbf{99.68}_{\pm 0.38}$	$\mathbf{4.27}_{\pm 0.67}$	$\mathbf{100.00}_{\pm 0.00}$	$\mathbf{0.00}_{\pm 0.00}$	$\mathbf{99.75}_{\pm 0.08}$	$\mathbf{4.69}_{\pm 0.33}$	$\mathbf{100.00}_{\pm 0.00}$	$\mathbf{0.00}_{\pm 0.00}$

As reported in Table 4, our method consistently outperforms NegLabel and TANL across every standard and full-spectrum regime on CIFAR-100. On Near-OOD evaluation, our method achieves 74.18% AUROC and reduces FPR@95 to 74.14%, advancing to 73.10% AUROC under the covariate-shifted Near-OOD full-spectrum setting. Notably, on Far-OOD detection, our method substantially surpasses TANL, improving AUROC from 82.77% to 93.87% and decreasing FPR@95 from 50.86% to 28.66%. This finding reconciles cleanly with the trade-off discussed in Appx. **??**. On large-scale ImageNet corpora, pruning coarse negative labels selectively trades off distant Far-OOD margins to resolve fine-grained Near-OOD ambiguities. Conversely, on low-resolution benchmarks such as CIFAR-100, degraded visual fidelity severely inflates prediction ambiguity under streaming imbalance. Under these conditions, baseline scalar confidence thresholding collapses, admitting noisy samples and corrupting cache updates. In contrast, our bipartite topological gate robustly filters boundary noise, preserving semantic fidelity and allowing our framework to outperform baselines across both Near-OOD and Far-OOD regimes.

As detailed in Tab. 5, the distinctive clinical feature space allows all methods to achieve saturated discriminability on Far-OOD data, attaining 100.00% AUROC and 0.00% FPR@95. On the challenging fine grained Near-OOD benchmarks, our framework attains the top ranked performance, yielding 99.68% AUROC and reducing FPR@95 to 4.27% on standard Near-OOD, while retaining 99.75% AUROC and 4.69% FPR@95 under the Near-OOD-FS protocol. In contrast, TANL experiences an increase in FPR@95 (9.85% and 9.97% across Near OOD settings), corroborating that our bipartite topological formulation provides superior boundary calibration across diverse imaging domains.