Semantic-Aware Hierarchical Negative Label Selection for Zero-Shot Out-of-Distribution Detection

Fei Ding[1,2], Feng Zhang[3,4], Wei Chen[3,4*], Tengjiao Wang[3,4], Yan Zhang[1]
[1]School of Intelligence Science and Technology, Peking University, Beijing, China
[2]Institute for Artificial Intelligence, Peking University, Beijing, China
[3]School of Computer Science, Peking University, Beijing, China
[4]Institute of Computational Social Science, Peking University (Qingdao), Qingdao, China
{dingfei,zhangfeng}@stu.pku.edu.cn, {pekingchenwei,tjwang,zhyzhy001}@pku.edu.cn

***Abstract*—Zero-shot out-of-distribution (OOD) detection aims to determine whether an input image belongs to any in-distribution (ID) category relying solely on textual labels. Existing negative label-based methods typically construct negative label sets by selecting labels that are semantically distant from ID labels within a large semantic pool to represent potential OOD categories. While effective, this strategy inherently favors easy negatives and overlooks hard negatives that lie close to the decision boundary. In addition, directly aggregating all negative labels may lead to erroneous predictions when multiple semantically similar negative labels collectively dominate the scoring function. To address these issues, we propose HiNeg, a novel framework for zero-shot OOD detection that jointly models negative label diversity and cluster-level relative dominance. HiNeg introduces a hierarchical negative label selection strategy HiSelect that captures both semantically distant easy negatives and boundary-adjacent hard negatives, providing complementary coverage of potential OOD semantics. Furthermore, a cluster-relative aggregated score is designed to evaluate negative evidence relative to the strongest ID response, suppressing redundant contributions from correlated negative labels within the same cluster. Extensive experiments on standard zero-shot OOD detection benchmarks demonstrate the effectiveness of the proposed approach.**

***Index Terms*—out-of-distribution detection, negative label, zero-shot detection**

I. Introduction

Out-of-distribution (OOD) detection, which aims to identify samples that do not belong to any known in-distribution (ID) class, is crucial for ensuring the reliability of AI systems in open-world environments. While traditional supervised methods are limited by their reliance on large labeled datasets and computationally intensive training, pretrained vision-language models (VLMs) like CLIP [1] provide a scalable alternative. By aligning images and text in a shared semantic space, VLMs enable zero-shot OOD detection relying solely on textual labels, eliminating the need for downstream image supervision. Within this framework, the construction and selection of negative labels representing potential OOD categories is critical, as they help delineate the decision boundaries of ID classes in the textual space, significantly enhancing the model's accuracy in identifying OOD samples.

Existing methods, including NegLabel [2], CSP [3], and NegRefine [4], mine a semantic pool derived from a large-scale corpus for labels with low semantic similarity to ID classes, thereby approximating potential OOD categories and improving the discriminative capability of VLMs. While effective, this NegMining strategy heavily favors labels distant from ID labels in the semantic space. This creates a bias towards easy negatives and overlooks challenging hard negatives. As illustrated in Fig. 1(a), easy negative labels that are semantically distant from ID labels (e.g., *red-brown area* relative to the ID label *gong*) can effectively identify certain OOD samples. However, such easy negatives are insufficient for challenging OOD images whose semantics partially overla with ID concepts. Fig. 1(b) demonstrates that hard negative labels located closer to the ID decision boundary (e.g., *latchstring* relative to *electrical switch*) are critical for detecting these difficult cases. The absence of boundary-adjacent labels in existing NegMining-based approaches leads to missed detections of challenging OOD samples. Overcoming this inherent NegMining bias requires a structured mechanism capturing both easy and hard negatives to enable more comprehensive modeling of OOD semantics.

Nevertheless, directly aggregating the selected set of hierarchical negatives can still cause erroneous OOD predictions for ID samples. As illustrated in Fig. 1(c), correlated negative labels within the same semantic cluster can collectively dominate the negative score. In this example, despite the ID label *knot* provides strong evidence for an ID image, multiple related negative labels (e.g., *hammer throw*, *stopper knot*, and *untying*) jointly overwhelm the scoring function, resulting in an incorrect OOD prediction. This highlights the flaw of naive negative aggregation, where semantically clustered negative labels exert redundant influence.

To address these challenges, we propose HiNeg, a novel framework for zero-shot OOD detection. First, HiNeg employs HiSelect, a hierarchical negative label selection strategy that jointly captures easy and hard negatives via two-level

*Corresponding author.

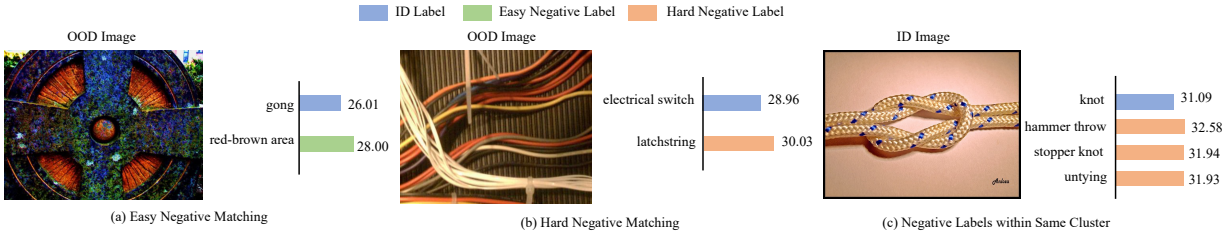

Fig. 1. ID and OOD images from the ImageNet-1K benchmark, along with their highest-scoring ID and negative labels based on CLIP similarity ($\tau = 0.01$).

clustering for complementary semantic coverage. Second, it introduces a cluster-relative aggregated score that calibrates negative evidence at the cluster level by comparing it against the strongest corresponding ID response. By emphasizing relative dominance near the decision boundary and mitigating redundant contributions from closely related negative labels, HiNeg achieves robust OOD detection without additional training. Our main contributions are summarized as follows:

- We develop a hierarchical negative label selection strategy HiSelect that organizes candidate labels through two-level clustering, enabling the simultaneous inclusion of semantically distant and boundary-adjacent negatives for zero-shot OOD detection.
- We propose a cluster-relative aggregation score that evaluates negative evidence in relation to the most confident ID label within each cluster, mitigating the impact of correlated negative labels in the final OOD decision.
- Extensive experiments on zero-shot OOD detection benchmark demonstrate that HiNeg consistently outperforms existing zero-shot baselines.

II. Related Work

Traditional OOD detection methods are designed for single-modal image classifiers, utilizing either training-time regularization or post-hoc scoring techniques [5]–[7]. The development of vision-language models, such as CLIP [1], has encouraged the use of multi-modal features for OOD detection [8], [9]. Based on these models, zero-shot OOD detection recognizes OOD samples by computing the semantic alignment between visual inputs and ID category descriptions, removing the need for downstream training images [10], [11]. To explicitly represent OOD semantics, negative label-based methods emerge recently. NegLabel [2] first introduces the retrieval of negative labels from large-scale external corpora, and CSP [3] broadens the negative label pool using superclass labels modified by adjectives. NegRefine [4] improves this selection process by applying large language models to filter out proper nouns and subcategories of ID classes from negative candidates.

III. Problem Definition

Given an ID label set $Y_{\text{in}} = \{y_1, \ldots, y_C\}$, OOD detection aims to construct a score function $S(x)$ that estimates the likelihood of an input image x belonging to the ID label set. The image x is identified as ID when its score exceeds a predefined threshold γ, and is considered OOD otherwise. For zero-shot OOD detection, we adopt a pretrained vision-language model CLIP, which comprises an image encoder $f_{\text{img}}(\cdot)$ and a text encoder $f_{\text{text}}(\cdot)$ that map images and texts into a shared semantic embedding space. Given an input image x and a label $y_i \in Y_{\text{in}}$, the corresponding image and text embeddings are obtained as $h = f_{\text{img}}(x)$ and $l_i = f_{\text{text}}(\text{prompt}(y_i))$, respectively, where a fixed textual template (e.g., "a photo of a y_i") is used to construct the prompt.

Existing negative label-based methods enhance the ability to identify potential OOD images by selecting a set of negative labels $Y_{\text{neg}} = \{\tilde{y}_1, \ldots, \tilde{y}_M\}$ from the semantic pool derived from large-scale lexical corpus like WordNet [12], which are used together with ID labels to define the detection score function. A basis approach is NegLabel [2], which defines the OOD detection score as

$$S_{\text{NegLabel}}(x) = \frac{\sum_{i=1}^{C} e^{\text{sim}(x,y_i)/\tau}}{\sum_{i=1}^{C} e^{\text{sim}(x,y_i)/\tau} + \sum_{j=1}^{M} e^{\text{sim}(x,\tilde{y}_j)/\tau}}, \quad (1)$$

where $\text{sim}(x, y)$ denotes the cosine similarity between the ℓ_2-normalized image embedding h and label embedding l, and the temperature parameter is set to $\tau = 0.01$. Subsequent methods like CSP [3] and NegRefine [4] share this scoring formulation, differing primarily in semantic pool construction and negative label filtering, respectively. In this work, we focus on improving negative label-based zero-shot OOD detection by addressing the limitations of existing negative label selection strategy and further provide an aggregation schemes.

IV. Methodology

To enhance negative label-based zero-shot OOD detection, we introduce two key components. First, a hierarchical selection strategy employs two-level clustering to jointly capture easy and hard negatives for complementary semantic coverage. Second, a cluster-relative aggregated score calibrates negative evidence against the strongest ID response within each cluster, which reduces the impact of redundant negatives and improves robustness near the ID–OOD decision boundary. Together, these components significantly enhance both negative label quality and scoring reliability.

A. Hierarchical Negative Label Selection

To construct an informative and diverse set of negative labels, we adopt a coarse-to-fine two-level clustering strategy

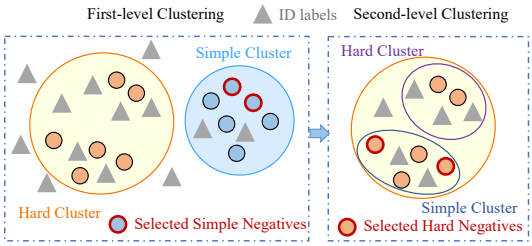

Fig. 2. Framework of hierarchical negative label selection strategy HiSelect.

HiSelect for hierarchical negative label selection. Following CSP [3], we use the same semantic pool as candidate negative labels. Unlike prior methods [2]–[4] that select negative labels with low semantic similarity to ID labels, which tends to favor easy negatives, our approach aims to jointly capture both easy and hard negatives through hierarchical refinement. As illustrated in Fig. 2, we first perform a coarse-grained clustering over ID labels and candidate negative labels to distinguish semantically easy clusters from hard clusters based on the proportion of ID labels within each cluster. Easy negatives are selected from easy clusters, while labels from hard clusters are further refined through a second-level clustering to identify hard negatives that lie closer to the ID decision boundary.

First-level Clustering. We first apply the K-means algorithm [13], [14] to cluster the ℓ_2-normalized embeddings of ID labels together with those of candidate negative labels from the semantic pool into K coarse clusters. By minimizing within-cluster variance in the embedding space, this coarse partitioning captures the global semantic structure of the label space and enables efficient processing of large-scale label sets.

Separation of Easy and Hard Clusters. To distinguish easy and hard clusters, we compute for each cluster q the proportion of ID labels within the cluster as

$$r_q = \frac{n_{\text{pos}}}{n_{\text{pos}} + n_{\text{neg}}}, \tag{2}$$

where n_{pos} and n_{neg} denote the numbers of ID labels and candidate negative labels in cluster q, respectively. We then compare r_q with the overall proportion of ID labels in the entire label set, denoted as

$$r_{\text{global}} = \frac{C}{C + N_{\text{neg}}}, \tag{3}$$

where C and N_{neg} represent the total numbers of ID labels and candidate negative labels, respectively. Clusters satisfying $r_q \leq r_{\text{global}}$ are regarded as easy clusters, while clusters with $r_q > r_{\text{global}}$ are treated as hard clusters.

Easy Negatives Selection. Within each easy cluster obtained from the first-level clustering, we select easy negatives by ranking candidate negative labels according to their semantic similarity to ID labels. Specifically, for each candidate negative label $\tilde{y}_i$, we compute

$$s_i = \max_{y \in Y_{\text{in}}} \text{sim}(\tilde{y}_i, y), \tag{4}$$

where $\text{sim}(\cdot, \cdot)$ denotes the cosine similarity between the corresponding normalized label embeddings. We then select the lowest $p_1\%$ of candidates with the smallest similarity values within each cluster as easy negatives:

$$Y_{\text{easy}} = \{\tilde{y}_i \mid s_i \leq \text{Percentile}_p(\{s_i\})\}. \tag{5}$$

This selection strategy ensures that the chosen easy negatives are semantically distant from ID labels, while sampling from multiple clusters contributes to a diverse set of easy negatives.

Second-level Clustering. While easy negatives provide broad semantic coverage, clusters with higher ID label ratios often contain negative labels that are semantically close to ID classes. Such hard clusters require further refinement to uncover informative hard negatives that are more challenging for OOD detection. For each hard cluster identified at the first level, we perform a fine-grained clustering using the Leiden algorithm [15], a graph-based community detection method that is well suited for discovering local semantic structures.

Specifically, within each hard cluster, we construct a k-nearest neighbor (kNN) graph over the ℓ_2-normalized label embeddings, where each node represents a label and is connected to its k most semantically similar neighbors, with edge weights defined by cosine similarity. On this neighbor graph, the Leiden algorithm partitions nodes by maximizing a modularity objective

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \gamma \frac{k_i k_j}{2m} \right] \delta(c_i, c_j), \tag{6}$$

where A_{ij} denotes the edge weight between nodes i and j, k_i is the degree of node i, m is the total edge weight in the graph, γ is a resolution parameter controlling cluster granularity, c_i denotes the cluster assignment, and $\delta(\cdot, \cdot)$ is the Kronecker delta function.

Through iterative local optimization and multilevel aggregation, Leiden produces well-connected and locally optimal communities. Compared to K-means, this graph-based refinement is particularly effective for handling hard clusters with diverse sizes and complex internal semantics. As a result, the second-level clustering reveals coherent subclusters that facilitate the identification of hard negatives lying closer to the ID decision boundary.

Hard Negatives Selection. For the subclusters produced by the second-level clustering under a first-level cluster q, we apply the same criterion as in the first level to distinguish easy and hard subclusters based on the proportion of ID labels. Subclusters satisfying $r_{q,t} \leq r_{\text{global}}$ are regarded as easy subclusters. Hard negatives are sampled exclusively from these easy subclusters to avoid selecting negative labels that are overly close to ID semantics, which could otherwise lead to increased false negatives for ID samples.

To obtain a discriminative yet diverse set of hard negatives, we employ a Determinantal Point Process (DPP) [16] sampler to select $p_2\%$ candidate negatives from each easy subcluster. We first define a diversity matrix using a linear kernel over label embeddings:

$$D_{ij} = (\tilde{l}_i)^\top \tilde{l}_j, \tag{7}$$

where $\tilde{l}_i$ denotes the ℓ_2-normalized label embedding of a candidate negative label $\tilde{y}_i$ in the subcluster. The diversity matrix is then combined with an identity term to control the strength of diversity:

$$S = \mu D + (1 - \mu) I, \tag{8}$$

where $\mu \in [0, 1]$ balances diversity against quality.

To further encourage separation of selected negative labels from ID semantics, we assign each candidate a quality score based on its maximum similarity to ID labels within the corresponding first-level cluster q:

$$s_i = \max_{y \in Y_{\text{in}}^{(q)}} \text{sim}(\tilde{y}_i, y), \qquad u_i = 1 - s_i, \tag{9}$$

where $\text{sim}(\cdot, \cdot)$ denotes cosine similarity between label embeddings. Following standard quality–diversity DPP construction, we form the DPP kernel as

$$Z = Q^{\frac{1}{2}} S Q^{\frac{1}{2}}, \qquad Q = \text{diag}(\hat{u}_1, \ldots, \hat{u}_n), \tag{10}$$

where $\hat{u}_i$ denotes the min-max normalized quality score of candidate $\tilde{y}_i$, and n is the number of candidates in the subcluster. We then sample $p_2\%$ hard negatives Y_{hard} according to the DPP defined by Z, which favors subsets that are both diverse and well separated from ID semantics.

The final negative label set is obtained by taking the union of easy and hard negatives:

$$Y_{\text{neg}} = Y_{\text{easy}} \cup Y_{\text{hard}}. \tag{11}$$

This combined set integrates both semantically distant and boundary-adjacent negatives, offering broad semantic coverage while capturing complementary OOD characteristics.

B. Cluster-Relative Aggregated Score

Although the negative labels selected by our hierarchical strategy provide broad and informative OOD coverage, directly aggregating all negative labels can still lead to false OOD predictions for ID samples. This issue often arises when multiple semantically similar negative labels within the same cluster jointly amplify negative evidence, even in the presence of strong ID responses near the decision boundary.

To address this problem, we propose the cluster-relative aggregated score, which calibrates negative evidence by comparing it with ID evidence at the level of first-level semantic clusters. Given an input image x, let $Y_{\text{in}}^{(q)}$ denote the ID labels assigned to the q-th first-level cluster, and let $Y_{\text{neg}}^{(q)}$ denote the selected negative labels associated with the same cluster. We define the strongest ID evidence within cluster q as

$$M_q(x) = \max_{y \in Y_{in}^{(q)}} e^{\text{sim}(x,y)/\tau}, \tag{12}$$

and the strongest negative evidence within the cluster as

$$N_q(x) = \max_{\tilde{y} \in Y_{\text{neg}}^{(q)}} e^{\text{sim}(x,\tilde{y})/\tau}. \tag{13}$$

The cluster-relative aggregated score is then defined as

$$S_{\text{CRA}}(x) = \max_q \frac{M_q(x)}{M_q(x) + N_q(x)}, \tag{14}$$

which emphasizes the relative dominance of ID evidence over negative evidence within each semantic cluster. By summarizing negative responses at the cluster level, this formulation suppresses redundant contributions from closely related negative labels and prevents correlated negatives from dominating the final OOD decision.

Finally, we integrate the proposed score into the overall OOD scoring function as

$$S(x) = S_{\text{NegLabel}}(x) + \alpha\, S_{\text{CRA}}(x), \tag{15}$$

where $S_{\text{NegLabel}}(x)$ is the original negative-label-based score and α balances the contribution of the proposed cluster-relative refinement.

V. Experiments

A. Experimental Setup

Datasets and Baselines. Following [4], we adopt ImageNet-1K [19] as ID dataset, and utilize iNaturalist [20], OpenImage-O [7], Clean [21], and NINCO [21] as OOD datasets. These OOD datasets are selected due to their minimal in-distribution contamination, offering a cleaner and more reliable evaluation protocol than those used in earlier studies. Baselines consist of recent zero-shot OOD detection methods that operate without any training data.

Implementation Details. We adopt CLIP [1] with a ViT-B/16 backbone [22] as the pre-trained vision–language model. The candidate semantic pool is constructed following CSP [3]. In the hierarchical negative label selection module, we employ a two-level clustering strategy, where K-means clustering is performed with $K = 30$ at the first level, and Leiden clustering is applied at the second level with the number of neighbors $k = 5$ and the resolution parameter $\lambda = 0.25$. For hard negative selection, we use a Determinantal Point Process sampler with the diversity–quality trade-off parameter $\mu = 0.5$. The sample ratio of easy and hard negatives is set to $p_1 = 5\%$ and $p_2 = 40\%$, respectively. In the final scoring function, the balancing weight α is set to 0.1.

Evaluation Metrics. We evaluate OOD detection performance with two metrics: the false positive rate of OOD samples at 95% true positive rate of ID samples (FPR95) and the area under the receiver operating characteristic curve (AUROC).

B. Main Results

OOD Detection on ImageNet-1K benchmark. As summarized in Table I, HiNeg outperforms existing zero-shot OOD baselines, achieving average improvements of 1.15% in FPR95 and 0.24% in AUROC over the strongest baseline NegRefine. HiNeg shows consistent gains across three of the four OOD datasets, with slightly lower performance on iNaturalist, likely because the already high performance of the CSP semantic pool reduces the need for further fine-grained label selection and aggregation.

Hard OOD Detection. We evaluate HiNeg on hard OOD detection benchmarks utilizing ImageNet-1K subsets and the Spurious OOD benchmark [23]. Following [4], [10], we consider ImageNet-10, 20, and 99 as ID datasets and the rest

TABLE I
OOD DETECTION PERFORMANCE ON IMAGENET-1K AS THE ID DATASET. BOLD INDICATES THE BEST PERFORMANCE.

Methods	iNaturalist		OpenImage-O		Clean		NINCO		Average	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
ZOC [17]	90.98	79.53	70.04	80.61	81.07	72.88	78.67	67.96	80.19	75.24
MCM [10]	32.20	94.59	41.03	92.00	57.79	83.24	79.33	74.34	52.59	86.04
GL-MCM [11]	17.42	96.44	34.52	92.91	51.49	84.78	74.40	76.03	44.46	87.54
CLIPN [18]	19.11	96.20	30.36	92.22	41.56	87.31	66.51	78.72	39.39	88.61
NegLabel [2]	1.91	99.49	28.62	93.74	41.22	86.79	68.70	77.30	35.11	89.33
CSP [3]	1.54	**99.60**	28.94	94.09	38.75	88.32	68.65	77.88	34.47	89.97
NegRefine [4]	**1.47**	99.57	23.85	95.00	33.18	90.65	61.96	81.92	30.12	91.79
HiNeg (ours)	1.80	99.53	**22.11**	**95.47**	**32.96**	**90.71**	**59.02**	**82.40**	**28.97**	**92.03**

TABLE II
OOD DETECTION PERFORMANCE ON HARD OOD DETECTION TASKS.

ID datasets	ImageNet-10		ImageNet-10		ImageNet-20		ImageNet-99		WaterBirds	
OOD datatasets	ImageNet-20		ImageNet-99		ImageNet-10		ImageNet-10		Placesbg	
Methods	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
NegLabel	5.00	98.80	1.89	99.45	11.60	98.04	52.89	89.53	29.16	87.99
CSP	**3.30**	**99.02**	**1.78**	**99.50**	3.40	98.79	44.35	91.21	12.07	92.88
NegRefine	3.89	98.95	2.24	99.21	2.28	99.19	40.99	91.68	11.96	93.13
HiNeg (ours)	4.03	98.92	1.96	99.37	**2.06**	**99.24**	**39.27**	**91.72**	**11.48**	**93.17**

TABLE III
ABLATION STUDY RESULTS. EASY AND HARD DENOTE VARIANTS THAT USE ONLY THE EASY NEGATIVES OR ONLY THE HARD NEGATIVES SELECTED BY HISELECT, RESPECTIVELY. THE FIRST ROW REPORTS THE PERFORMANCE OF CSP WITH THE NEGMINING STRATEGY AS THE BASELINE.

HiSelect		S_{CRA}	iNaturalist		OpenImage-O		Clean		NINCO		Average	
Easy	Hard		FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
×	×	×	1.54	99.60	28.94	94.09	38.75	88.32	68.65	77.88	34.47	89.97
✓	×	×	7.45	98.40	24.42	94.77	37.20	88.48	67.01	78.03	34.02	89.92
×	✓	×	**1.28**	**99.64**	26.76	94.57	40.48	87.75	60.02	81.98	32.14	90.99
✓	✓	×	1.63	99.57	23.07	95.27	34.29	90.27	59.24	81.90	29.56	91.75
✓	✓	✓	1.80	99.53	**22.11**	**95.47**	**32.96**	**90.71**	**59.02**	**82.40**	**28.97**	**92.03**

as OOD, resulting in semantically challenging OOD settings. We also evaluate spurious OOD detection with WaterBirds as ID and Placesbg as OOD. As shown in Table II, HiNeg performs slightly worse on ImageNet-10 as ID, likely due to the limited number of ID labels and the relatively clear ID–OOD separation in this setting. In contrast, HiNeg achieves the best performance on more challenging ImageNet subset configurations and the Spurious benchmark, demonstrating its superiority in hard OOD detection over methods restricted to semantic distance for negative label selection.

C. Ablation Study and Discussions

Ablation Study. As shown in Table III, removing the cluster-relative aggregated score S_{CRA} degrades performance, with FPR95 increasing by 0.59% and AUROC decreasing by 0.28%. This indicates that S_{CRA} prevents closely related negative labels from dominating OOD decisions. Furthermore, replacing the hierarchical negative label selection strategy with NegMining startegy of CSP, which selects only the most distant negative labels from ID labels, results in a significant performance drop, increasing FPR95 by 4.91% and decreasing AUROC by 1.78%. This highlights the necessity of our hierarchical approach to capture diverse negative semantics. Finally, removing either the selected easy or hard negatives reduces performance, confirming they are complementary and essential for comprehensive coverage of potential OOD categories, thereby enhancing detection robustness.

Backbone Analysis. To verify its generalization capability, we evaluate HiNeg with two distinct vision backbones, ResNet50 and ViT-B/32. As Table IV shows, HiNeg consistently achieves the best average performance on both architectures. Specifically, with ResNet50, it surpasses strong baselines by reaching an average FPR95 of 33.83% and an AUROC of 90.33%. A similar trend with ViT-B/32 indicates strong generalization of HiNeg across VLM architectures. Furthermore, the identical hyperparameters utilized across all backbones highlights the robustness of our method.

Hyperparameter Sensitivity Analysis. As shown in Fig. 3, we analyze key hyperparameters including easy and hard negative sampling ratios p_1 and p_2, and the balancing weight

TABLE IV
COMPARISON WITH DIFFERENT BACKBONES ON IMAGENET-1K AS THE ID DATASET.

Backbone	Methods	iNaturalist		OpenImage-O		Clean		NINCO		Average	
		FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
ResNet50	NegLabel	2.88	99.24	33.35	92.00	45.52	84.45	73.53	74.20	38.82	87.47
	CSP	**1.95**	**99.46**	33.91	91.83	42.06	86.08	72.86	75.53	37.70	88.22
	NegRefine	2.30	99.38	31.05	92.57	38.97	88.26	65.62	80.73	34.48	90.23
	HiNeg (ours)	2.26	99.39	**29.97**	**92.61**	**38.29**	**88.28**	**64.81**	**81.03**	**33.83**	**90.33**
ViT-B/32	NegLabel	3.73	99.11	31.26	92.87	44.16	85.20	71.45	75.30	37.65	88.12
	CSP	2.37	**99.46**	31.01	93.37	40.15	87.42	70.62	76.96	36.04	89.30
	NegRefine	**2.20**	99.45	26.11	94.22	35.47	89.90	60.70	82.38	31.12	91.49
	HiNeg (ours)	2.48	99.42	**24.93**	**94.36**	**35.08**	**90.02**	**59.37**	**82.42**	**30.47**	**91.56**

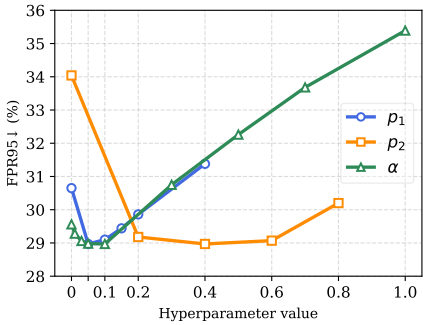

Fig. 3. Sensitivity analysis of p_1, p_2, and α on average performance.

α. HiNeg exhibits robust performance when p_1 ranges from 0.05 to 0.2 and p_2 from 0.2 to 0.6. Excessively small ratios yield insufficient semantic coverage, while overly large ones introduce negatives semantically close to ID labels and risk ID misclassification. Finally, optimal performance occurs at $\alpha = 0.1$, where the cluster-relative score effectively suppresses the joint amplification of semantically similar negatives to ensure a robust and well-balanced OOD decision.

VI. CONCLUSION

In this work, we introduce HiNeg, a hierarchical framework for zero-shot OOD detection that improves negative label-based modeling. By jointly selecting semantically distant and boundary-adjacent negative labels and calibrating negative evidence at the cluster level, HiNeg enables more effective discrimination between ID and OOD samples. Extensive experiments on standard OOD detection benchmarks demonstrate the robustness and effectiveness of the proposed approach, highlighting the importance of structured negative label selection and aggregation for zero-shot OOD detection.

ACKNOWLEDGMENT

This research is supported by the Natural Science Foundation of Qingdao(No.25-3-1-4-zyyd-jch).

REFERENCES

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *ICML*, 2021.

[2] X. Jiang, F. Liu, Z. Fang, H. Chen, T. Liu, F. Zheng, and B. Han, "Negative label guided OOD detection with pretrained vision-language models," in *ICLR*, 2024.

[3] M. Chen, J. Gao, and C. Xu, "Conjugated semantic pool improves OOD detection with pre-trained vision-language models," in *NeurIPS*, 2024.

[4] A. Ansari, K. Wang, and P. Xiong, "Negrefine: Refining negative label-based zero-shot ood detection," in *ICCV*, 2025.

[5] S. Vaze, K. Han, A. Vedaldi, and A. Zisserman, "Open-set recognition: A good closed-set classifier is all you need?" in *ICLR*, 2022.

[6] Y. Sun, Y. Ming, X. Zhu, and Y. Li, "Out-of-distribution detection with deep nearest neighbors," in *ICML*, 2022.

[7] H. Wang, Z. Li, L. Feng, and W. Zhang, "Vim: Out-of-distribution with virtual-logit matching," in *CVPR*, 2022.

[8] S. Park, J. Mok, D. Jung, S. Lee, and S. Yoon, "On the powerfulness of textual outlier exposure for visual ood detection," in *NeurIPS*, 2023.

[9] L. Tao, X. Du, J. Zhu, and Y. Li, "Non-parametric outlier synthesis," in *ICLR*, 2023.

[10] Y. Ming, Z. Cai, J. Gu, Y. Sun, W. Li, and Y. Li, "Delving into out-of-distribution detection with vision-language representations," in *NeurIPS*, 2022.

[11] A. Miyai, Q. Yu, G. Irie, and K. Aizawa, "Zero-shot in-distribution detection in multi-object settings using vision-language foundation models," *CoRR*, 2023.

[12] G. A. Miller, "Wordnet: A lexical database for english," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[13] S. P. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. 28, no. 2, pp. 129–136, 1982.

[14] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1. University of California Press, 1967, pp. 281–297.

[15] V. A. Traag, L. Waltman, and N. J. van Eck, "From louvain to leiden: guaranteeing well-connected communities," *CoRR*, 2018.

[16] A. Kulesza and B. Taskar, "Determinantal point processes for machine learning," *Found. Trends Mach. Learn.*, vol. 5, pp. 123–286, 2012.

[17] S. Esmaeilpour, B. Liu, E. Robertson, and L. Shu, "Zero-shot out-of-distribution detection based on the pre-trained model CLIP," in *AAAI*, 2022, pp. 6568–6576.

[18] H. Wang, Y. Li, H. Yao, and X. Li, "CLIPN for zero-shot OOD detection: Teaching CLIP to say no," in *ICCV*, 2023.

[19] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009.

[20] G. V. Horn, O. M. Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, and S. J. Belongie, "The inaturalist species classification and detection dataset," in *CVPR*, 2018, pp. 8769–8778.

[21] J. Bitterwolf, M. Müller, and M. Hein, "In or out? fixing imagenet out-of-distribution detection evaluation," in *ICML*, 2023.

[22] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.

[23] Y. Ming, H. Yin, and Y. Li, "On the impact of spurious correlation for out-of-distribution detection," in *AAAI*, 2022.