Appendix Table of Contents

A. Pseudo-code for Modality Inversion

Here, we provide the pseudo-code for the modality inversion process in Algorithm 2. This procedure aims to transform a high-confidence OOD image into an extra negative text embedding by optimizing a set of pseudo-tokens.

Algorithm 2 Modality Inversion From Image To Text [42]

Require: High-confidence OOD image $\boldsymbol{x}$, number of pseudo-tokens T, number of optimization steps S;

Initialize $\boldsymbol{v} = \{v_1, v_2, \cdots, v_T\}$;
Extract image embedding $\boldsymbol{h} = \mathcal{I}(\boldsymbol{x})$;
for $s = 1$ to S **do**
    Form $\bar{\boldsymbol{v}} = [\mathcal{E}(''a\ photo\ of''), \boldsymbol{v}]$;
    Extract text embedding $\boldsymbol{e}_v^- = \mathcal{T}(\bar{\boldsymbol{v}})$;
    Calculate the cosine loss $\mathcal{L} = 1 - cos(\boldsymbol{e}_v^-, \boldsymbol{h})$;
    Update $\boldsymbol{v}$ to minimize $\mathcal{L}$;
end for
return Negative text embedding $\boldsymbol{e}_v^- = \mathcal{T}(\bar{\boldsymbol{v}})$.

B. Additional Results

B.1. Full results on the OpenOOD Benchmark

Table 4. Full results of our method with ID dataset of ImageNet-1K on the OpenOOD benchmark.

Near / Far OOD	Datasets	FPR95 ↓	AUROC ↑
Near-OOD	SSB-hard [55]	69.96	80.24
	NINCO [3]	60.90	84.16
	Mean	**65.43**	**82.20**
Far-OOD	iNaturalist [24]	0.40	99.71
	Textures [7]	18.17	96.74
	OpenImage-O [56]	32.30	93.69
	Mean	**16.96**	**96.71**

B.2. More Results on the OpenOOD Benchmark

Table 5. OOD detection results with ID dataset of CIFAR100 on the OpenOOD benchmark using CLIP ViT-B/16 architecture. Full results are available in Table 6.

Methods	FPR95 ↓		AUROC ↑	
	Near-OOD	Far-OOD	Near-OOD	Far-OOD
Methods requiring training on ID or extra data				
GEN [38]	–	–	81.31	79.68
VOS [13] + EBO [37]	–	–	80.93	81.32
SCALE [65]	–	–	80.99	81.42
OE [21] + MSP [20]	–	–	88.30	81.41
Zero-shot methods (no training on ID or extra data)				
MCM [41]	75.20	59.32	71.00	76.00
NegLabel [29]	71.44	40.92	70.58	89.68
AdaNeg [74]	**59.07**	29.35	84.60	95.25
InterNeg	62.54	**20.02**	**85.45**	**96.39**

Table 6. Full results of our method with ID dataset of CIFAR100 on the OpenOOD benchmark.

Near / Far OOD	Datasets	FPR95 ↓	AUROC ↑
Near-OOD	CIFAR100 [30]	60.10	84.19
	TIN [31]	64.99	86.71
	Mean	**62.54**	**85.45**
Far-OOD	MNIST [9]	0.01	99.97
	SVHN [44]	3.23	99.41
	Texture [7]	21.16	96.84
	Places365 [77]	55.68	89.36
	Mean	**20.02**	**96.39**

Table 7. OOD detection results with ID dataset of CIFAR10 on the OpenOOD benchmark using CLIP ViT-B/16 architecture. Full results are available in Table 8.

Methods	FPR95 ↓		AUROC ↑	
	Near-OOD	Far-OOD	Near-OOD	Far-OOD
Methods requiring training on ID or extra data				
PixMix [23] + KNN [52]	–	–	93.10	95.94
OE [21] + MSP [20]	–	–	94.82	96.00
PixMix [23] + RotPred [22]	–	–	94.86	98.18
Zero-shot methods (no training on ID or extra data)				
MCM [41]	30.86	17.99	91.92	95.54
NegLabel [29]	28.75	6.60	94.58	98.39
AdaNeg [74]	**20.40**	2.79	94.78	99.26
InterNeg	23.93	**2.59**	**95.13**	**99.29**

Table 8. Full results of our method with ID dataset of CIFAR10 on the OpenOOD benchmark.

Near / Far OOD	Datasets	FPR95 ↓	AUROC ↑
Near-OOD	CIFAR100 [30]	37.26	91.81
	TIN [31]	10.60	98.45
	Mean	**23.93**	**95.13**
Far-OOD	MNIST [9]	0.00	100.00
	SVHN [44]	0.06	99.97
	Texture [7]	0.41	99.57
	Places365 [77]	9.89	97.61
	Mean	**2.59**	**99.29**

Following AdaNeg, we also evaluate our method on small-scale datasets from the OpenOOD benchmark. Specifically, we adopt CIFAR-10/100 [30] as an ID dataset while utilizing CIFAR-100/10 along with TinyImageNet (TIN) [31] as Near-OOD datasets. For Far-OOD datasets, we consider MNIST [9], SVHN [44], Texture [7], and Places365 [77]. For hyperparameters, we maintain consistency with the experimental setup described in Section C, modifying only M (set to 70,000) to align with AdaNeg and ensure fair comparison. All other parameters remain unchanged. In the Near-OOD setting, our method achieves the best results on the AUROC metric.

As illustrated in Table 5 and Table 7, our method still achieves state-of-the-art performance in the Far-OOD setting, even outperforming those training methods.

B.3. Impact of ID/OOD Dataset Ordering.

To evaluate the robustness of our method to ID/OOD dataset ordering, we generate different dataset permutations by varying random seeds. As shown in Table 9, our method maintains stable performance with standard deviations of merely 0.05

Table 9. Average results of AUROC (↑) with five different seeds on both Four-OOD and Near-OOD benchmarks.

Benchmark	Seed					Mean	Std
	0	1	2	3	4		
Four-OOD	97.43	97.38	97.38	97.45	97.51	97.43	0.05
Near-OOD	82.20	81.56	81.50	82.29	82.59	82.03	0.46

(Four-OOD) and 0.46 (Near-OOD), demonstrating strong insensitivity to ordering effects.

B.4. Ablation of Different CLIP Architectures

As shown in Table 10, we evaluate multiple CLIP backbone architectures in the Four-OOD setting using ImageNet-1K as the ID dataset. The experimental results demonstrate that our method consistently outperforms all baseline approaches by a substantial margin across different backbone architectures. This robust performance advantage highlights both the effectiveness and robustness of our proposed method.

Table 10. OOD detection results with ID dataset of ImageNet-1k and traditional Four-OOD datasets using different CLIP backbone architectures.

Backbones	Methods	OOD Datasets								Average	
		iNaturalist		SUN		Places		Textures			
		AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
ResNet50	NegLabel	99.24	2.88	94.54	26.51	89.72	42.60	88.40	50.80	92.97	30.70
	CSP	99.46	1.95	95.73	19.05	90.39	38.58	92.41	32.66	94.50	23.06
	AdaNeg	**99.58**	1.18	97.37	10.56	**93.84**	43.19	94.18	35.00	96.24	22.48
	InterNeg	99.56	**1.16**	**98.35**	**7.99**	93.71	**37.82**	**96.05**	**21.92**	**96.92**	**17.22**
ResNet101	NegLabel	99.27	3.11	94.96	24.55	89.42	44.82	87.22	52.78	92.72	31.32
	CSP	99.47	2.04	95.71	19.50	90.27	39.57	90.59	38.67	94.01	24.95
	AdaNeg	**99.69**	**0.78**	97.65	10.61	**94.00**	40.38	93.59	39.44	96.23	22.80
	InterNeg	99.64	0.87	**98.54**	**7.02**	93.70	**38.32**	**95.94**	**24.74**	**96.96**	**17.74**
ViT-B/32	NegLabel	99.11	3.73	95.27	22.48	91.72	34.94	88.57	50.51	93.67	27.92
	CSP	99.46	2.37	96.49	15.01	92.42	**31.47**	93.64	25.09	95.50	18.49
	AdaNeg	99.67	0.87	97.74	9.62	**93.98**	36.45	94.58	33.26	96.49	20.05
	InterNeg	**99.68**	**0.70**	**98.74**	**5.79**	93.65	38.05	**96.02**	**23.55**	**97.02**	**17.02**
ViT-B/16	NegLabel	99.49	1.91	95.49	20.53	91.64	35.59	90.22	43.56	94.21	25.40
	CSP	99.61	1.54	96.69	13.82	92.85	29.69	93.78	25.78	95.73	17.71
	AdaNeg	99.71	0.59	97.44	9.50	94.55	34.34	94.93	31.27	96.66	18.93
	InterNeg	**99.79**	**0.40**	**98.68**	**6.78**	**95.01**	**27.11**	**96.26**	**21.85**	**97.43**	**14.04**
ViT-L/14	NegLabel	99.53	1.77	95.63	22.33	93.01	32.22	89.71	42.92	94.47	24.81
	CSP	99.72	1.21	96.73	14.88	93.58	28.41	92.71	28.16	95.69	18.17
	AdaNeg	99.82	0.26	97.97	7.94	**95.12**	28.67	94.24	38.28	96.79	18.79
	InterNeg	**99.88**	**0.21**	**98.75**	**5.88**	95.03	**26.82**	**96.07**	**22.89**	**97.43**	**13.95**

B.5. Cross-Domain Analysis

In cross-domain settings, there is usually no ID training set provided for constructing ID image proxies. To evaluate our method in a cross-domain scenario, we conduct two additional experiments as follows:

- **ImageNet ID Image Proxies:** Since ImageNet-V2 is a variant of ImageNet, we use the image proxies computed from the original ImageNet training set as a substitute.
- **ImageNet-V2 ID Image Proxies:** Alternatively, we randomly select a small subset of ID images (e.g., 4 images per class) from ImageNet-V2 itself to compute the ID image proxies, and use the rest for testing.

Specifically, we set ImageNet-V2 as the ID dataset and Four-OOD (iNaturalist, SUN, Places, Textures) as the OOD datasets. Table 11 shows the average results across the Four-OOD datasets. These results demonstrate that our method remains effective and robust in cross-domain scenarios.

Table 11. Cross-domain OOD detection performance on ImageNet-V2 with Four-OOD datasets.

ID Image Proxies Source	Method	AUROC ↑	FPR95 ↓
ImageNet	NegLabel	93.08	29.77
	AdaNeg	96.24	19.21
	InterNeg	**96.84**	**18.40**
ImageNet-V2	NegLabel	93.90	27.64
	AdaNeg	96.19	19.97
	InterNeg	**96.96**	**18.18**

B.6. ID Misclassification with Different OOD Datasets

Since both AdaNeg and our method dynamically adjust the OOD score according to the test images during inference, it is important to evaluate additional OOD datasets to better understand the phenomenon of ID misclassification. In Figure 2, we use SUN as the OOD dataset. Here, we extend the evaluation to other OOD datasets included in the Four-OOD benchmark. As shown in Figure 5, our method consistently outperforms AdaNeg across various OOD datasets, demonstrating the robustness and effectiveness of InterNeg.

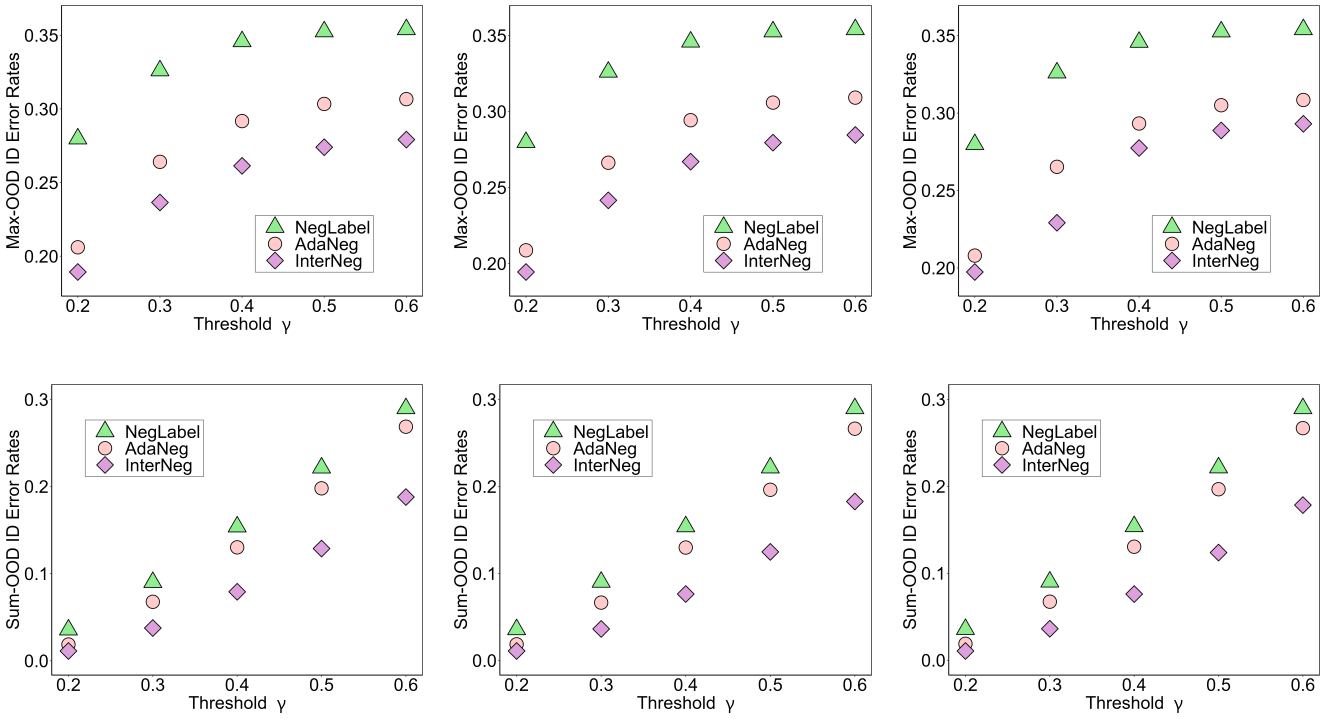

Figure 5. Max-OOD and Sum-OOD ID error rates on different OOD datasets. *Left:* iNaturalist. *Middle:* Places. *Right:* Textures.

B.7. Inference Cost

Method	**Mean**	**iNaturalist**	**Places**	**Textures**	**SUN**
AdaNeg	0.0058	0.0068	0.0046	0.0093	0.0025
InterNeg	0.0067	0.0074	0.0055	0.0113	0.0029

Table 12. Comparison of inference time (seconds per image).

We evaluate the inference efficiency of our approach on the traditional Four-OOD benchmark, using ImageNet-1K as the ID dataset (see Table 12). When measured on an NVIDIA RTX 3090 GPU, our method incurs only a **minor** computational overhead compared to AdaNeg.

B.8. Sensitivity to Corpus Choice

To investigate the impact of the underlying corpus, we substitute WordNet with the Common-20K and Part-of-Speech corpora. As shown in Table 13, our method consistently outperforms the strongest baseline, AdaNeg, reducing the average FPR95 by **3.62% and 4.76%**, respectively. This demonstrates that our performance improvements are robust and agnostic to the choice of the source corpus.

Source	Method	Average	
		AUROC ↑	FPR95 ↓
Common-20K	NegLabel	90.50	43.02
	CSP	92.06	36.56
	AdaNeg	93.12	32.39
	InterNeg	**94.58**	**28.77**
Part-of-Speech	NegLabel	92.71	32.12
	CSP	94.21	24.42
	AdaNeg	95.07	23.17
	InterNeg	**95.98**	**18.41**

Table 13. Evaluation with different corpus sources on the traditional Four-OOD benchmark, using ImageNet-1K as the ID dataset.

C. Experimental Setup

Datasets. Following previous work [6, 16, 29, 74], we evaluate our method on the large-scale ImageNet-1K Four-OOD detection benchmark [27]. This widely-used benchmark utilizes the ImageNet-1K [8] dataset as ID data and iNaturalist [24], SUN [63], Places [77], Textures [7] as four OOD datasets, where the labels of the four OOD datasets that overlap with ImageNet-1K have been manually removed. Furthermore, we also conduct experiments on the OpenOOD benchmark [66, 73] following [74], This benchmark also uses ImageNet-1K as the ID dataset, while categorizing OOD data into two distinct groups based on the empirical performance of OOD detectors: Near-OOD (*e.g.*, SSB-hard [55], NINCO [3]) and Far-OOD (*e.g.*, iNaturalist [24], Textures [7], OpenImage-O [56]). Also, each OOD dataset has no classes that overlap with the ID dataset.

Implementation Details. In this paper, we implement various CLIP [47] backbone architectures, including ResNet50, ResNet101, ViT-B/32, ViT-B/16 and ViT-L/14. Unless otherwise specified, we adopt the CLIP ViT-B/16 model as the pre-trained VLM, which consists of a visual encoder based on ViT-B/16 Vision Transformer [11] and a text encoder built on Text Transformer [54]. For hyperparameters, we set the number N of ID images per class as 16, the number M of negative texts as 2000, the maximum size K of extra negative text embeddings as 2000, temperature $\tau = 1.0$ and threshold $\beta = 0.35$ in all experiments. Following [29, 47], we employ the text prompt template of `"The nice [class]"`. All experiments are conducted using NVIDIA RTX 3090 GPUs.

Evaluation Metrics. Following common practice [6, 16, 29, 74], we adopt the following metrics to evaluate the OOD detection performance: (1) AUROC, the area under the receiver operating characteristic curve; (2) FPR95, the false positive rate of the OOD data when the true positive rate of ID data is 95%.

D. Further Discussion

D.1. Comparison with ArGue and SimLabel.

It is crucial to distinguish our approach from recent methods such as SimLabel [78] and ArGue [53]. First, we differentiate our focus on **metric consistency**—aligning the OOD detection metric with CLIP's inter-modal training objective—from

SimLabel's **semantic consistency**, which focuses on aligning ID labels with semantic synonyms. Second, our method and ArGue **differ fundamentally** in their core paradigms. ArGue employs prompt tuning with negative attributes to suppress spurious visual features, aiming for robust ID classification. In contrast, InterNeg leverages inter-modal guided negative texts explicitly tailored for OOD detection.

D.2. Clarification on Zero-Shot OOD Detection and ID Data Dependency

Definition of Zero-Shot OOD Detection. In the context of Out-of-Distribution (OOD) detection, we align our setting with the established consensus in prior literature. For instance, NegLabel defines the zero-shot capability as predicting the correct label "*without prior training on that specific class*." Similarly, EOE characterizes it as operating "*without re-training on any unseen ID data*." Therefore, the term "zero-shot" typically describes the training pipeline, especially for the CLIP module. Given this constraint, it is a common practice for the baselines in this field to introduce external knowledge or additional modules. For example, ZOC /CLIPN trains a captioner/encoder to generate candidate unknown classes/negative semantics within images, respectively. To mitigate the training burden, recent advances such as NegLabel and AdaNeg turn to proxies to enhance the representations of potential OOD classes. Our method **follows this paradigm** by fixing the fundamental inconsistency hidden in the proxy generation, **without training on either ID or external data**.

Dependency on ID Data. Furthermore, assuming access to a small set of ID data is highly realistic for real-world OOD detection deployments (e.g., autonomous driving and medical diagnosis), where ID classes are inherently known and well-defined. Crucially, in our framework, we utilize ID samples **solely for proxy calculation, not model training**. As validated in Figure 4, our approach achieves state-of-the-art performance with a minimal ID data dependency of merely 4 images per class, underscoring its practical applicability.