Mitigating the Modality Gap: Few-Shot Out-of-Distribution Detection with Multi-modal Prototypes and Image Bias Estimation

Yimu Wang, Evelien Riddell, Adrian Chow, Sean Sedwards, Krzysztof Czarnecki
University of Waterloo

{yimu.wang,evelien.riddell,ahtchow,sean.sedwards,krzysztof.czarnecki}@uwaterloo.ca

Abstract

Existing vision-language model (VLM)-based methods for out-of-distribution (OOD) detection typically rely on similarity scores between input images and in-distribution (ID) text prototypes. However, the modality gap between image and text often results in high false positive rates, as OOD samples can exhibit high similarity to ID text prototypes. To mitigate the impact of this modality gap, we propose incorporating ID image prototypes along with ID text prototypes. We present theoretical and empirical evidence indicating that this approach enhances VLM-based OOD detection performance without any additional training. To further reduce the gap between image and text, we introduce a novel few-shot tuning framework, SUPREME, comprising biased prompt generation (BPG) and image-text consistency (ITC) modules. BPG enhances image-text fusion and improves generalization (prevents overfitting on the training data) by conditioning ID text prototypes on the Gaussian-based estimated image domain bias; ITC reduces the modality gap by minimizing intra- and inter-modal distances. Moreover, inspired by our theoretical and empirical findings, we introduce a novel OOD score S_{GMP}, leveraging uni- and cross-modal similarities. Finally, extensive experiments demonstrate that SUPREME consistently outperforms existing VLM-based OOD detection methods.

1. Introduction

Detecting out-of-distribution (OOD) samples [11, 18, 21, 38, 41, 54, 58] is essential for the real-world deployment of machine learning models [16, 51, 63], as novel samples may emerge and should be flagged for careful consideration. Recently, inspired by the power of vision-language foundation models (VLMs) [14, 20, 51], novel approaches to OOD detection using VLMs [1–3, 10, 11, 24, 35, 36, 40, 44, 45, 50, 59, 69, 70] have gained significant attention. Early VLM-based OOD detection works [3, 24, 44, 59] mainly focus on using CLIP [51] in a zero-shot setting, where only
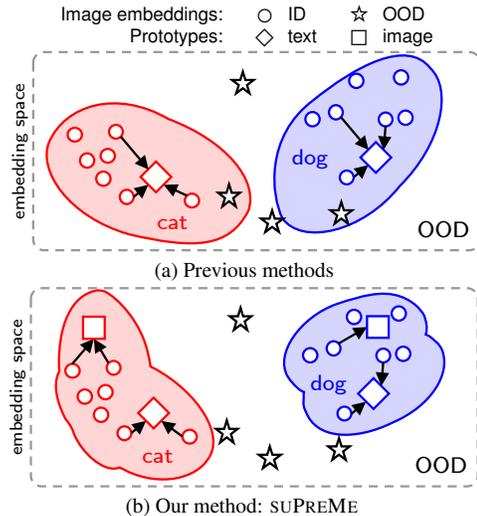

Figure 1. (a) Previous VLM-based OOD detection [36, 44, 45] uses only ID text prototypes ($\diamond$) to identify OOD samples. (b) SUPREME employs ID image prototypes ($\square$) to complement ID text prototypes, reducing the impact of the image-text modality gap [39] and sharpening the boundary between ID and OOD.

the in-distribution (ID) class names are utilized. For example, Maximum Concept Matching (MCM) [44] measures the similarity between input images and text embeddings of ID classes, also referred to as ID *text prototypes*, in the joint vision-language representation space of CLIP. MCM then uses this similarity score to differentiate OOD and ID samples, on the basis that ID samples should have higher similarity scores.

Following MCM, other zero-shot methods [3, 24, 59] aim at generating OOD (negative) text prototypes by querying large language models (LLMs) [26, 48] or WordNet [43]. These methods obtain OOD scores by evaluating the similarity difference between ID and OOD text prototypes. To further improve performance, various few-shot tuning OOD detection methods [2, 36, 40, 47, 70] have been proposed. These methods mainly target learning OOD (negative) text prototypes. For example, NegPrompt [36] learns OOD text prototypes by minimizing the similarity between OOD text
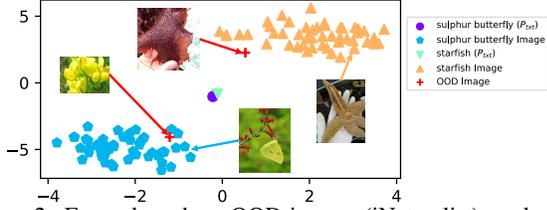
Figure 2. Examples where OOD images (iNaturalist) are located closer to ID text prototypes compared to ID images (ImageNet-1k) generated by CLIP (ViT-B/16). We also visualize the embeddings generated by our SUPREME in Fig. 5, which has smaller gaps between image and text embeddings.

prototypes and ID data with contrastive learning.

While these approaches have shown promise using CLIP, they use only text prototypes and thus are limited by the modality gap [39] between image and text modalities. Liang *et al.* [39] demonstrate that CLIP's image and text embeddings are located in two clearly separate regions of the embedding space. As a result, OOD images may also exhibit high similarity to ID text prototypes (Fig. 2, more examples are in Fig. 6), since high similarity can stem from either true semantic similarity or mere spatial closeness to the ID text prototypes, causing a risk of increased false positives. This observation raises a critical question about the impact of the modality gap on VLM-based OOD detection, leading to our first research question:

> *RQ1: What is the impact of the modality gap on VLM-based OOD detection methods?*

In response to RQ1, we hypothesize that the modality gap negatively impacts performance and that incorporating image and text (multi-modal) prototypes, as opposed to only text prototypes, could alleviate this effect. We illustrate this in Fig. 1. To explore this, we compute class-specific ID image prototypes by averaging embeddings extracted from the ID data within each class. We extend the MCM score S_{MCM} [44] to incorporate the image prototypes alongside the text prototypes. Compared to S_{MCM}, our empirical analysis shows that, without any training, using the extension of S_{MCM} with multi-modal prototypes improves average FPR95 and AUROC from 32.4 and 94.5 to 24.2 and 95.8, respectively, on Imagenet-100 [6] (ID) across four OOD datasets (Tab. 1). For theoretical support, in Theorem 1, we demonstrate that incorporating image prototypes increases the score separation between ID and OOD data, leading to improved performance.

Given these findings, *i.e.*, that the modality gap significantly affects VLM-based OOD detection, we conclude that it is necessary to explore approaches mitigating the gap to improve reliability. This leads us to our second question:

> *RQ2: How can we mitigate the impact of modality gap to improve VLM-based OOD detection?*

To address RQ2, we propose a novel few-**s**hot tuning, m**U**lti-modal **PR**ototyp**E**-based **ME**thod for OOD detection

with CLIP, termed SUPREME. SUPREME comprises biased prompt generation (BPG) and image-text consistency (ITC) modules. BPG introduces a Gaussian-based image domain bias to enhance image-text fusion and generalization on ID data unseen during fine-tuning by preventing overfitting on the limited training ID data. ITC reduces the modality gap by minimizing inter- and intra-modal distances as CLIP remains frozen. SUPREME also makes use of a new OOD score, which we call the **G**eneralized **M**ulti-modal **P**rototype OOD score, denoted S_{GMP}.

BPG. In contrast to previous OOD detection methods that only utilize learnable contexts for text prototypes [2, 36, 45], BPG conditions text prototypes on three components: learnable contexts, the image embedding, and the ID image domain bias. The image embedding facilitates image-text fusion, enhancing cross-modal alignment. The third component, the Gaussian-based ID image domain bias, captures the distribution of ID image embeddings to improve generalization on unseen ID data and avoid overfitting.

ITC. As the CLIP remains frozen during training, to minimize the modality gap, we introduce intra- and inter- modal losses with two mappings. First, the image embedding I is mapped to the text domain as $I' = f_{img\text{-}txt}(I)$ with the image-to-text mapping $f_{img\text{-}txt}$. To ensure it aligns with ID text prototypes, we employ the inter-modal loss ℓ_{inter}. Additionally, to avoid information loss, the intra-modal loss ℓ_{intra} is applied between the original embedding I and the reconstructed image embedding $\hat{I} = f_{txt\text{-}img}(I')$ using the text-to-image mapping $f_{txt\text{-}img}$.

S_{GMP}. Building on our findings, we introduce a new OOD score, S_{GMP}, which integrates uni- and cross-modal similarities. While previous methods rely solely on the similarity between ID text prototypes and input image embedding, S_{GMP} incorporates the similarity between multi-modal input embeddings—the image embedding I and the mapped image embedding $f_{img\text{-}txt}(I)$—and ID multi-modal (image and text) prototypes, respectively.

Our contributions are summarized as follows:

- (RQ1) We empirically and theoretically demonstrate that multi-modal (image and text) prototypes reduce the negative impact of the modality gap, resulting in performance improvements without additional training.
- (RQ2) To further mitigate the modality gap, we propose a novel few-shot tuning framework, SUPREME, comprising biased prompt generation (BPG) and image-text consistency (ITC) modules.
- (RQ2) Building on our empirical and theoretical results, we design a new OOD score, S_{GMP}, exploiting ID multi-modal prototypes and multi-modal input embeddings to enhance performance and robustness.
- (RQ2) Extensive experiments across multiple benchmarks [6, 68] demonstrate that SUPREME outperforms existing OOD detection methods.

2. Related Work

Out-of-distribution (OOD) detection. OOD detection [18, 22, 41, 63, 68] aims to discriminate ID samples and OOD samples. Traditional methods include post-hoc [30, 49, 53, 54], confidence enhancement [8, 17, 46, 60, 62], and outlier exposure methods [9, 11, 19, 21, 23, 28, 42, 72]. Recently, the field has shifted towards exploiting large-scale pre-trained VLMs [14, 20, 51], *i.e.*, CLIP [51], to enhance OOD detection. MCM [44] employs CLIP and explores the effects of softmax and temperature scaling in a zero-shot manner. Subsequent works can be roughly categorized into zero-shot and few-shot tuning methods. Zero-shot methods [3, 24, 44, 59] primarily focus on identifying negative classes by querying LLMs [26, 48] or WordNet [43]. In parallel, few-shot tuning methods [2, 36, 40, 47, 70] also mainly target finding negative text prompts (OOD text prototypes). For example, CATEX [40], LSN [47], and Neg-Prompt [36] learn negative prompts by minimizing the similarity between negative prompts and ID training data. Following those works, SynOOD [35] generates OOD images using image generative models and LLMs to finetune the image and text encoders. On the other side, AdaNeg [69] introduces a memory bank during testing to selectively cache discriminative features from test images, representing the targeted OOD distribution. The memory bank further helps in distinguishing new OOD images.

In contrast to previous VLM-based methods that use only text prototypes, SUPREME enhances OOD detection by employing multi-modal prototypes and encouraging cross-modal alignment to reduce the modality gap.

Prompt tuning. Prompt tuning originates in Natural Language Processing [31, 37, 52] as a method for automating template/prompt creation in models such as BERT [7] and GPT [48]. For example, AutoPrompt [52] is a gradient-based approach for identifying "optimal" prompts, replacing manually designed prompts. Recently, prompt tuning has been applied in computer vision models [27, 73, 74]. Notably, CoOp [74] and CoCoOp [73], as representative methods in visual prompt tuning (VPT), employ learnable prompts optimized by minimizing classification loss to improve CLIP's performance.

Different from previous VPT works [2, 40, 45, 47, 73, 74] that generate text prompts with text and individual image data, SUPREME generates text prompts/prototypes with a novel Gaussian-based image domain bias for estimating the distribution of image data, which enables enhanced image-text fusion with improved generalization ability.

3. Preliminaries

3.1. Problem Setup

Let and $\mathcal{Y} = \{y_1, \dots, y_C\}$ represent the set of ID class labels, where C is the number of ID classes.

OOD detection. In real-world applications, AI models are trained on ID data and may misclassify OOD data into ID classes with high confidence [25]. To tackle this problem, OOD detection [64] is proposed to identify OOD samples using a score function $S(\cdot)$ [32, 33]. A sample is classified as OOD if $S(\mathbf{x}) \leq \gamma$, where γ is a predefined threshold.

Few-shot tuning. In contrast to existing works that either utilize the entire ID training dataset [30, 34, 58] or avoid tuning entirely [3, 24, 44, 59], we study the scenario where the model is fine-tuned using only a subset of the ID training data (16 images per class) without access to OOD data or other additional data.

3.2. Revisiting CLIP, MCM, and Prompt Tuning

CLIP [51] is a foundational VLM pre-trained on a web-scale image-text dataset using self-supervised contrastive learning [4]. Specifically, for classification, CLIP first incorporates class labels $\mathcal{Y} = \{y_c\}_{c=1}^C$, *e.g.*, "cat", into fixed pre-designed, rather than learned, text prompts, *e.g.*, "a photo of a *label*". These prompts combined with class labels are processed by CLIP's text encoder $f_{text}(\cdot)$ to generate text embeddings $\{P_{txt,c}\}_{c=1}^C$, where $P_{txt,c} = f_{text}$("a photo of a $\{y_c\}$"). Given an image $\mathbf{x}$, the image embedding I is obtained by the image encoder $f_{image}(\cdot)$ as $I = f_{image}(\mathbf{x})$. The index of the predicted class label is given by $\arg\max_{c \in \{1..C\}} \cos(I, P_{txt,c})$, where $\cos(\cdot, \cdot)$ denotes the cosine similarity.

CLIP for OOD Detection (MCM) [44]. Beyond its strong classification capabilities, MCM demonstrates that pre-trained CLIP models also exhibit robust zero-shot OOD detection capabilities. Specifically, MCM defines the maximum classification score as the OOD score S_{MCM},

$$S_{MCM}(I, \{P_{txt,c}\}_{c=1}^C) = \\ \max_{c \in \{1..C\}} \frac{\exp(\cos(I, P_{txt,c})/\tau)}{\sum_{j=1}^C \exp(\cos(I, P_{txt,j})/\tau)}, \quad (1)$$

where τ is the temperature and $\exp(\cdot)$ is the exponential function.

Visual prompt tuning [15, 65, 73, 74]. To improve CLIP's performance when target training data is accessible, CoOp [74] replaces manually designed prompt templates with learnable soft prompts (context tokens) $CoOp_c = [V_1, \dots, V_L, y_c]$ for class index c. L is the length of learnable prompts and each V_i represents a learnable vector. The class-wise text embedding is then generated as $P_{txt,c}^{CoOp} = f_{text}(CoOp_c)$ for class index c. With CLIP's text and image encoder parameters frozen, the learnable prompts are optimized using the cross-entropy loss:

$$\ell_{ID}(I) = -\log \frac{\exp(\cos(I, P_{txt,c}^{CoOp})/\tau)}{\sum_{j=1}^C \exp(\cos(I, P_{txt,j}^{CoOp})/\tau)}, \quad (2)$$
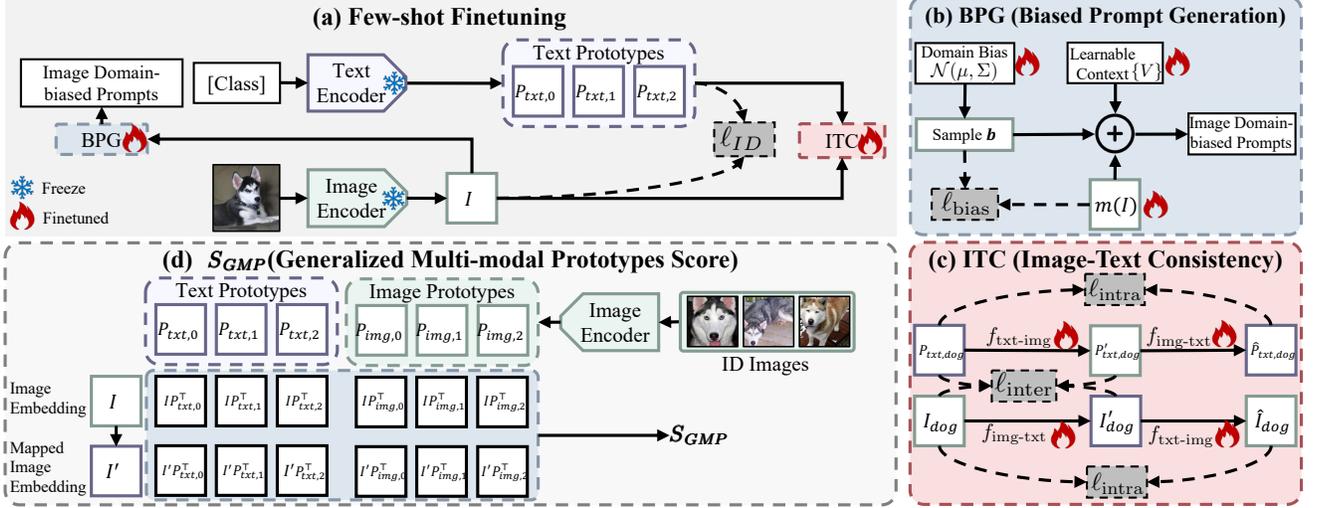
Figure 3. Overview of SUPREME. The two novel modules, *i.e.*, BPG (biased prompt generation) and ITC (image-text consistency), are designed to minimize the modality gap. (a) During the few-shot fine-tuning stage, (b) BPG generates image domain-biased prompts, conditioned on the estimated image domain bias and the mapped image embedding $m(I)$ for better image-text fusion. (c) ITC minimizes the modality gap directly by the intra- and inter-modal losses (ℓ_{intra} and ℓ_{inter}) with the text-to-image mapping $f_{txt\text{-}img}$ and the image-to-text mapping $f_{img\text{-}txt}$. (d) During inference, image prototypes are obtained by averaging each class's base ID image embeddings (Eq. (3)). The proposed S_{GMP} (Eq. (10)) is calculated based on the maximum similarity between the multi-modal embeddings (the image embedding I and mapped image embedding I') and ID multi-modal prototypes (*i.e.*, text prototypes $\{P_{txt,c}\}_{c=1}^C$ and image prototypes $\{P_{img,c}\}_{c=1}^C$). S_{MCM} refers to the MCM score [44].

where c is the index of the ground truth label for the input image. Following CoOp, to avoid potential overfitting on the training data, CoCoOp [73] conditions prompts on the image embedding I as $CoCoOp_c = [V_1 + m(I), \ldots, V_L + m(I), y_c]$, where $m(\cdot)$ is a two-layer multi-layer perception (MLP, Linear-ReLU-Linear).

4. Methodology

In this section, we first answer RQ1 by presenting our hypothesis and findings in Sec. 4.1, where we demonstrate that the modality gap negatively impacts performance and that incorporating image prototypes alleviates this effect. Then, to mitigate the modality gap (RQ2), we introduce SUPREME in Sec. 4.2 with two modules and a novel OOD score, S_{GMP}. We provide an overview of SUPREME in Fig. 3.

4.1. RQ1: Multi-modal Prototypes

CLIP has been applied in different areas [13, 56]. However, a recent study [39] demonstrates that representations of images and texts are clearly separated, creating what is known as the modality gap. It remains unclear, however, whether this gap positively or negatively impacts OOD detection.

We hypothesize that the modality gap negatively impacts OOD detection performance and that it can lead to increased false positives, as OOD images may exhibit high similarity to ID text prototypes due to either semantic similarity or small spatial distance.

To test our hypothesis, we demonstrate that incorporat-

ing multi-modal (image and text) prototypes can mitigate the negative impact of the modality gap and reduce false positives. We obtain ID text prototypes $\{P_{txt,c}\}_{c=1}^C$ as in MCM [44]. For ID image prototypes $\{P_{img,c}\}_{c=1}^C$, we collect ID base images $\{\mathbf{x}_i\}_{i=1}^{N_{base}}$ with corresponding labels $\{y_i\}_{i=1}^{N_{base}}$, where N_{base} is the number of base images. ID image prototypes are then calculated by averaging the image embeddings for each class label y_c:

$$P_{img,c} = \frac{\sum_{i=1}^{N_{base}} \mathbb{1}(y_i = y_c) f_{image}(\mathbf{x}_i)}{\sum_{j=1}^{N_{base}} \mathbb{1}(y_j = y_c)}, \quad (3)$$

where $\mathbb{1}(\cdot)$ is the indicator function and $f_{image}(\cdot)$ is CLIP's image encoder. With these image prototypes, we extend S_{MCM} (Eq. (1)) [44] to a new OOD score, which we term the Multi-Modal Prototypes score, S_{MMP}, defined as

$$S_{MMP} = \frac{S_{MCM}(I, \{P_{img,c}\}_{c=1}^C) + S_{MCM}(I, \{P_{txt,c}\}_{c=1}^C)}{2}. \quad (4)$$

This approach balances contributions from both modalities, potentially mitigating the impact of the modality gap.

Empirical evidence. We begin by empirically validating the effectiveness of image prototypes and S_{MMP}. We use ImageNet-100 [6] as the ID dataset, with four OOD datasets: iNaturalist [57], SUN [61], Places [71], and Texture [5]. With all training images of ImageNet-100 used for obtaining image prototypes, we observe in Tab. 1 that, on average, S_{MMP} outperforms S_{MCM} in FPR95 and AUROC, demonstrating the effectiveness of image prototypes and S_{MMP}.

Methods	FPR95 (%) ↓	AUROC (%) ↑
ImageNet-100 → 4 Datasets (Average)		
S_{MCM} [44]	32.58	94.48
S_{MMP} (Ours, Eq. (4))	**24.18**	**95.79**

Table 1. Comparison of OOD detection results on ImageNet-100 (ID) → 4 OOD datasets (OOD) with CLIP-B/16. FPR95 represents the false positive rate of OOD images when the true positive rate of ID images is at 95%. AUROC is the area under the receiver operating characteristic curve. Best in **bold**.

Theoretical observation. To further understand why image prototypes improve performance, we present a theoretical analysis in Theorem 1 of the Appendix. The theorem shows that S_{MMP} increases the expected score separation between ID and OOD samples compared to using text prototypes alone (as in S_{MCM}) under general assumptions on OOD and ID distributions used by previous works [12, 39].

While incorporating image prototypes improves performance , inspired by recent research [36, 45], we further fine-tune CLIP in a few-shot tuning manner to mitigate the impact of the modality gap (RQ2).

4.2. RQ2: SUPREME

To alleviate the negative impact of the image-text modality gap [39] in CLIP (RQ2), we propose SUPREME, which is summarized in Fig. 3. SUPREME incorporates two modules, *i.e.*, the biased prompt generation module (BPG) and the image-text consistency module (ITC), along with a new OOD score. BPG enhances image-text fusion using image domain-biased prompts, while ITC minimizes both intra- and inter-modal distances. SUPREME's novel OOD score, S_{GMP}, exploits uni- and cross-modal similarities to improve robustness and performance.

Biased Prompt Generation (BPG). Drawing inspiration from CoCoOp [74], we employ a set of learnable context tokens $\{V_i \in \mathcal{R}^{1 \times N_{\text{lm}}}\}_{i=1}^L$ and the mapped image embedding $m(I) \in \mathcal{R}^{1 \times N_{\text{lm}}}$ to generate the text prototypes, where L is the length of learnable contexts, $m(\cdot)$ is a two-layer MLP, and N_{lm} is the dimension of text encoder's input. Similar to previous methods [45, 73], learnable context tokens replace the manually designed prompt templates, *e.g.*, "a photo of a *class*", for generating text prototypes.

To improve generalization on unseen ID data and avoid overfitting on few-shot training data, we introduce a third component, a Gaussian-based estimated image domain bias $\mathcal{N}(\mu, \Sigma)$. This bias counteracts the modality gap by pre-biasing text prototypes toward ID images and augments the training data to better approximate the true distribution of mapped ID image embeddings. Using a Gaussian distribution is a reasonable a priori choice and is mathematically tractable. Our ablation study (Tab. 6) confirms that this choice is valid and is successful in reducing FPs. After that,

we generate image domain-biased prompts (IDBP) conditioned on these three components.

Estimating image domain bias. In each training iteration, we sample $\mathbf{b} \in \mathcal{R}^{1 \times N_{\text{lm}}}$ from $\mathcal{N}(\mu, \Sigma)$ as

$$\mathbf{b} = \mu + \sigma n, n \sim \mathcal{N}(0, \mathcal{I}), \tag{5}$$

where $\Sigma = \sigma\sigma^\top$ (based on the Cholesky decomposition) and $\mathcal{N}(0, \mathcal{I})$ represents the standard Gaussian distribution, given identity matrix $\mathcal{I}$. To align the domain bias with the distribution of training ID image embeddings, we employ the following loss:

$$\ell_{bias}(I) = \|\mu - m(I)\|_1 + \|\mathbf{b} - m(I)\|_1, \tag{6}$$

where $\|\cdot\|_1$ is the ℓ_1 distance. In the first term of the loss (Eq. (6)), μ is updated. In the second term, as $\mathbf{b}$ is sampled using μ and Σ, μ and Σ will be updated during the few-shot training procedure. The first term, $\|\mu - m(I)\|_1$, pushes the bias distribution's mean μ toward the centroid of training image embeddings, while the second term, $\|\mathbf{b} - m(I)\|_1$, ensures sampled bias vectors $\mathbf{b}$ are within the training image embedding space.

Generating image domain-biased prompts (IDBP). We condition IDBPs on three components: the learnable contexts $\{V_i \in \mathcal{R}^{1 \times N_{\text{lm}}}\}_{i=1}^L$, mapped image embedding $m(I) \in \mathcal{R}^{1 \times N_{\text{lm}}}$, and a sample $\mathbf{b} \in \mathcal{R}^{1 \times N_{\text{lm}}}$ from $\mathcal{N}(\mu, \Sigma)$ as

$$IDBP_c = [V_1 + m(I) + \mathbf{b}, \dots, V_L + m(I) + \mathbf{b}, y_c]. \tag{7}$$

The final ID text prototype $P_{txt,c}$ for the c-th class is obtained as $P_{txt,c} = f_{text}(IDBP_c)$.

We thus establish a foundation for cross-modal fusion by conditioning prompts on learned contexts, image embeddings, and the estimated image domain bias.

Image-Text Consistency (ITC). While BPG reduces the gap by image-text fusion, image-text consistency (ITC) directly reduces the modality gap by aligning image and text embeddings through inter- and intra-modal distances at a fine-grained level. It consists of two mappings, *i.e.*, the image-to-text mapping $f_{img\text{-}txt}(\cdot)$ and text-to-image mapping $f_{txt\text{-}img}(\cdot)$, which are used for mapping embeddings into the other modality.

The first component of ITC is the *inter-modal loss*. It is designed to ensure that mapped text/image embeddings $(\{f_{txt\text{-}img}(P_{txt,c})\}_{c=1}^C$ and $f_{img\text{-}txt}(I))$ remain close to their original counterparts, thus enabling effective alignment. It is designed to minimize the cross-entropy loss:

$$\ell_{inter}(I) = -\log \frac{\exp(\cos(I, f_{txt\text{-}img}(P_{txt,c}))/\tau)}{\sum_{j=1}^C \exp(\cos(I, f_{txt\text{-}img}(P_{txt,j}))/\tau)}$$
$$-\log \frac{\exp(\cos(f_{img\text{-}txt}(I), P_{txt,c})/\tau)}{\sum_{j=1}^C \exp(\cos(f_{img\text{-}txt}(I), P_{txt,j})/\tau)}, \tag{8}$$

Methods	Venue	Shots	iNaturalist		SUN		Places		Texture		Average	
			FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑
MOS† [21] (BiT)	CVPR'21	-	9.28	98.15	40.63	92.01	49.54	89.06	60.43	81.23	39.97	90.11
Fort *et al.*† [11] (ViT-B)	NeurIPS'21	-	15.07	96.64	54.12	86.37	57.99	85.24	53.32	84.77	45.12	88.25
Energy† [41] (ViT-B)	NeurIPS'20	-	21.59	95.99	34.28	93.15	36.64	91.82	51.18	88.09	35.92	92.26
MSP† [18] (ViT-B)	ICLR'17	-	40.89	88.63	65.81	81.24	67.90	80.14	64.96	78.16	59.89	82.04
ODIN‡ [38]	ICLR'18	-	30.22	94.65	54.04	87.17	55.06	85.54	51.67	87.85	47.75	88.80
ViM‡ [58]	CVPR'22	-	32.19	93.16	54.01	87.19	60.67	83.75	53.94	87.18	50.20	87.82
KNN‡ [54]	ICML'22	-	29.17	94.52	35.62	92.67	39.61	91.02	64.35	85.67	42.19	90.97
VLM-based OOD Detection (CLIP of ViT-B/16)												
MCM [44]	NeurIPS'22	0	30.91	94.61	37.59	92.57	44.69	89.77	57.77	86.11	42.74	90.77
CLIPN [59]	CVPR'23	0	23.94	95.27	26.17	93.93	33.45	92.28	40.83	90.93	31.10	93.10
EOE [3]	ICML'24	0	12.29	97.52	20.40	95.73	30.16	92.95	57.53	85.64	30.09	92.96
CoOp* [74]	IJCV	16	29.47	94.89	31.34	93.36	40.28	90.07	54.25	87.58	38.83	91.47
CoCoOp* [73]	CVPR'22	16	30.74	94.73	31.18	93.15	38.75	90.63	53.84	87.92	38.63	91.61
NPOS [55]	ICLR'23	-	16.58	96.19	43.77	90.44	45.27	89.44	46.12	88.80	37.93	91.22
LoCoOp [45]	NeurIPS'23	16	16.05	96.86	23.44	95.07	32.87	91.98	42.28	90.19	28.66	93.52
CATEX [40]	NeurIPS'23	-	10.18	97.88	33.87	92.83	41.43	90.48	33.17	92.73	29.66	93.48
LSN [47]	ICLR'24	-	21.56	95.83	26.32	94.35	34.48	91.25	38.54	90.42	30.22	92.96
ID-Like [2]	CVPR'24	4	8.98	98.19	42.03	91.64	44.00	90.57	**25.27**	94.32	30.07	93.68
NegPrompt [36]	CVPR'24	16	6.32	**98.73**	22.89	95.55	27.60	93.34	35.21	91.60	23.01	94.81
SCT [66]	NeurIPS'24	16	13.94	95.86	20.55	95.33	29.86	92.24	41.51	89.06	26 47	93.37
Local-Prompt [67]	ICLR'25	16	8.63	98.07	23.23	95.12	31.74	92.42	34.50	92.29	24.52	94.48
SUPREME		16	**5.91**	**98.73**	**15.85**	**96.08**	26.18	93.19	34.88	91.93	20.70	95.38
SUPREME†		16	8.27	98.29	19.40	95.84	26.69	**93.56**	26.77	**94.45**	**20.28**	**95.54**

Table 2. OOD detection results with ID data of ImageNet-1k and four OOD datasets using CLIP (ViT-B/16). Best in **bold**. Second best is underlined. "*" is cited from LSN [47]. "†" represents results from MCM [44]. "‡" represents results from NPOS [55]. SUPREME† employs 16 images per class for training and the whole training split for calculating ID image prototypes.

where c is the index of the ground truth label class for the input image and τ is the temperature.

Next, to prevent information loss, we introduce *intra-modal loss*. Specifically, we reconstruct the original embedding by mapping the mapped embedding back to its initial modality. Then, we minimize the ℓ_1 distance between the reconstructed and original embeddings as follows:

$$\ell_{intra}(I) = \| I - f_{txt\text{-}img}(f_{img\text{-}txt}(I)) \|_1$$
$$+ \sum_{j=1}^{C} \| P_{txt,j} - f_{img\text{-}txt}(f_{txt\text{-}img}(P_{txt,j})) \|_1 . \quad (9)$$

The inter- and intra-modal losses serve to reduce the modality gap between image and text, and to preserve all information during mapping.

Training, Inference, and S_{GMP}. Our overall training objectives are a linear combination of the four losses: $\mathcal{L} = \ell_{ID} + \alpha(\ell_{intra} + \ell_{inter}) + \beta \ell_{bias}$, where α and β are trade-off parameters. During inference, we use the mean μ of the image domain bias as the sample **b** to generate image domain-biased prompts.

Generalized multi-modal prototypes OOD score S_{GMP}. Building on our empirical and theoretical findings in Sec. 4.1, with the image-to-text mapping $f_{img\text{-}txt}$, we introduce the generalized multi-modal prototypes OOD score S_{GMP}. S_{GMP} uses the average of the maximum similarities between the multi-modal input embedding (I and $I' = f_{txt\text{-}img}(I)$) and multi-modal prototypes:

$$S_{GMP}(I, I', \{P_{txt,c}\}_{c=1}^{C}, \{P_{img,c}\}_{c=1}^{C}) =$$

$$\left(S_{MCM}(I, \{P_{txt,c}\}_{c=1}^{C}) + S_{MCM}(I, \{P_{img,c}\}_{c=1}^{C}) \right. \quad (10)$$
$$\left. + S_{MCM}(I', \{P_{txt,c}\}_{c=1}^{C}) + S_{MCM}(I', \{P_{img,c}\}_{c=1}^{C}) \right)/4 .$$

Remark. S_{GMP} extends S_{MCM} by incorporating both uni- and cross-modal similarities. In different scenarios, the importance of intra- and inter-modal similarity is different. However, we **average the contributions** from the image and text modalities as a starting point to ensure the generalization ability of S_{GMP} and reduce the gap. This approach enhances the separation between ID and OOD samples, improving the OOD detection performance of VLM-based methods, as shown in Fig. 4.

5. Experiments

5.1. Experimental Details

Datasets and benchmarks. Our experiments primarily use the ImageNet-1k [6] and ImageNet-100 [6] dataset as ID data. Aligning with standards from prior works [44, 59], we evaluate on four diverse OOD datasets: iNaturalist [57], SUN [61], Places [71], and Texture [5]. OpenOOD benchmark [68] is also used for evaluation on ImageNet-1k. We test on ImageNet-10 [6] and ImageNet-20 [6] for hard OOD. Details are in the Appendix.

Baselines methods. We compare SUPREME with several baseline methods [2, 3, 11, 18, 21, 24, 36, 38, 40, 41, 44, 45, 47, 54, 55, 58, 59, 66, 67, 73, 74].

Implementation details. For each class, 16 images are used for few-shot tuning and the calculation of ID image proto-

| Methods | ImageNet-100 → OOD Datasets | | | | | | | | Average | |
| | iNaturalist | | SUN | | Places | | Texture | | | |
	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑
MSP	23.55	95.92	37.02	92.45	40.76	91.23	24.40	94.90	31.43	93.63
ViM	20.11	96.22	38.56	93.12	44.01	87.33	33.12	93.24	33.95	92.48
MCM	18.13	96.77	36.45	94.54	34.52	94.36	41.22	92.25	32.58	94.48
CoOp	9.30	97.95	11.64	97.61	17.45	96.53	15.94	96.90	13.58	97.25
CoCoOp	11.76	97.84	14.28	97.13	15.16	96.73	18.27	96.54	14.86	97.06
SUPREME	**2.31**	**99.28**	8.26	**98.38**	12.20	97.12	11.10	97.53	8.47	**98.08**
SUPREME†	2.54	99.21	**8.08**	98.28	**12.17**	97.15	9.98	97.69	8.19	**98.08**

Table 3. OOD detection results with ID data of ImageNet-100 and four OOD datasets using CLIP (VIT-B/16). Best in **bold**. Second best is underlined.

types. We use the image and text encoders of VIT-B/16 pre-trained by CLIP [51] for all the experiments. To avoid overfitting, instead of using deep MLPs, we employ linear transformations without bias as the mappings ($f_{img\text{-}txt}$ and $f_{txt\text{-}img}$). The parameters of CLIP are frozen. We use SGD to optimize other parameters, *e.g.*, estimated image domain bias and learnable context, with a momentum of 0.9. Training epochs, learning rate, batch size, and learnable context length are set to $50, 0.002, 32$, and 16, respectively. We set $\alpha = 0.005$ and $\beta = 0.1$. Experiments are conducted on a single Nvidia V100 with 32 GB memory. All the results for SUPREME are the average of three trials.

Metrics. We employ two widely accepted metrics: FPR95, the false positive rate of OOD images when the true positive rate of ID images is at 95%, and AUROC, the area under the receiver operating characteristic curve.

5.2. Main Results

ImageNet-1k and ImageNet-100 as the ID dataset. In Tab. 2, we present a comprehensive analysis of ImageNet-1k (ID) across four OOD datasets. SUPREME demonstrates superior performance, achieving the lowest FPR95 and highest AUROC scores across most datasets. Notably, SUPREME achieves an average FPR95 of 20.70 and an average AUROC of 95.38. With all the training data for calculating ID prototypes, SUPREME† is further boosted with an average FPR95 of 20.28 and an average AUROC of 95.54. In Tab. 3, we observe consistent trends using ImageNet-100 as the ID dataset. SUPREME achieves the best average results, with FPR95 of 8.47 and AUROC of 98.08, showing substantial improvements over other methods.

ImageNet-1k on OpenOOD. The results on the OpenOOD benchmark are shown in Tab. 4. Compared to previous methods, SUPREME achieves the best performance in both near and far OOD scenarios. Specifically, SUPREME achieves an FPR95 of 71.44 and AUROC of 75.95 in the near OOD scenario, and an FPR95 of 20.92 and AUROC of 95.45 in the far OOD scenario.

ImageNet-10 ↔ ImageNet-20. To evaluate SUPREME under challenging conditions where ID and OOD datasets

| Methods | NearOOD | | FarOOD | |
	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑
MCM	85.37	58.97	69.87	77.11
LoCoOp	87.95	64.97	59.88	83.23
SUPREME	**71.44**	**75.95**	**20.92**	**95.45**

Table 4. OOD Detection results with ID data of ImageNet-1k on the OpenOOD benchmark. Best in **bold**.

| Methods | ImageNet-10 → ImageNet-20 | | ImageNet-20 → ImageNet-10 | |
	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑
Energy	10.30	97.94	16.40	97.37
CLIPN	7.80	98.07	13.67	97.47
MCM	5.00	98.71	17.40	98.87
LoCoOp*	5.60	98.47	5.40	98.92
SUPREME	**4.10**	**98.96**	**3.60**	**99.05**

Table 5. Comparison on ImageNet-10 ↔ ImageNet-20 with CLIP-B/16. Best in **bold**. "*" represents our reproduction.

share similar semantics, we follow previous work [24, 44] and conduct experiments on ImageNet-10 and ImageNet-20 as shown in Tab. 5. SUPREME achieves the best performance, with the lowest FPR95 and highest AUROC scores in both scenarios, demonstrating its robustness.

5.3. Ablation Study

We present an ablation study using CLIP (ViT-B/16) on ImageNet-1k → four OOD datasets. The computational complexity analysis (Tab. 14), compatibility with post-hoc methods (Tab. 12), impacts of shot count (Fig. 8), hyperparameter sensitivity (Figs. 7, 9 and 10), and the use of out-of-distribution prototypes (Tab. 12) are in the Appendix.

Effectiveness of BPG, ITC, and S_{GMP}. In Tab. 6, we evaluate the effectiveness of BPG and ITC, as well as the proposed score S_{GMP}, for enhancing OOD detection. We see that excluding BPG (-BPG) raises FPR95 to 28.12 and reduces AUROC to 92.77, highlighting BPG's importance in reducing false positives. Similarly, removing ITC (-ITC) significantly degrades performance, increasing FPR95 to 28.74 and lowering AUROC to 93.27. We also observe that the Gaussian bias term in ITC is important, as evidenced by the drop in

Methods	FPR95 ↓	AUROC ↑
SUPREME	**20.70**	**95.38**
- BPG	28.12	92.77
- ITC	28.74	93.27
- Bias in ITC	24.61	94.90
- S_{GMP}	26.02	94.79
w. MLP	25.24	93.06

Table 6. Effectiveness of different modules and the proposed OOD score S_{GMP}.

Methods	Top-1 Acc. (%)
MCM	67.0
CLIPN	68.5
ID-like	68.3
SUPREME	**71.8**

Table 7. Comparison in ID accuracy on ImageNet-1K val set.

$\left\|\overline{\Delta}_{gap}\right\|$	ImageNet (ID) ↓	iNaturalist ↓	Places ↓	Texture ↓	SUN ↓
MCM	0.8679	0.9839	0.9101	0.9373	0.9252
SUPREME	**0.6135**	**0.7159**	**0.8961**	**0.9273**	**0.9002**

Table 8. Quantifying the modality gap with different methods. We use CLIP (ViT-B/16) as the base model.

performance (fourth row). Furthermore, when the model uses S_{MCM} instead of S_{GMP}, the performance also declines, with FPR95 at 26.02 and AUROC at 94.79. Our full model, incorporating ITC, BPG, and S_{GMP}, achieves the best performance. The detailed table is in Tab. 11 in the Appendix.

Using MLP for image-text consistency. $f_{img\text{-}txt}$ and $f_{txt\text{-}img}$ in ITC can sufficiently address the modality gap, we replace them with a two-layer MLP (Linear-ReLU-Linear) and present the experimental results in Tab. 6 (w. MLP). Results indicate that using MLP negatively impacts performance, which might be because, with limited training data, sophisticated modules such as MLP could overfit.

SUPREME improves the ID performance. In Tab. 7, we compare the Top-1 accuracy of different methods on the ImageNet-1K validation set. Compared to other OOD detection methods, *e.g.*, CLIPN (68.5%) and ID-like (68.3%), SUPREME shows a considerable gain with the highest accuracy, demonstrating its effectiveness in maintaining high ID accuracy while also excelling in OOD detection.

Effectiveness of S_{GMP}. We visualize the empirical cumulative distribution of S_{MCM} and our S_{GMP} in Fig. 4 with the ID data of ImageNet-1k and the OOD data of iNaturalist using the same model (SUPREME). S_{GMP} clearly improves the gap between ID and OOD data in the empirical CDF figure. Furthermore, the Kolmogorov–Smirnov test statistic is enlarged from 0.7270 (S_{MCM}) to 0.8555 (S_{GMP}), indicating that using S_{GMP} yields more separable scores than S_{MCM}.

Reduced modality gap (t-SNE visualization and quantitative assessment of modality gap). We choose two arbitrary classes from ImageNet-1k (Val) and use t-SNE to visualize embeddings in Fig. 5. It is clear that ITC reduces the modality gap, as the distances between mapped image embeddings I' and text prototypes P_{txt} are smaller than that between I and P_{txt}. Similarly, we observe that image embeddings I are located close to image prototypes P_{img}. We also quantitatively assess the modality gap following Liang et al. [39] as shown in Tabs. 8 and 15, which shows that SUPREME achieves smaller modality gaps compared with
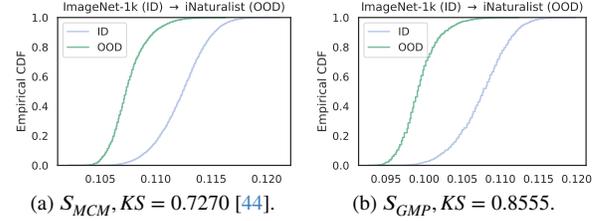

Figure 4. Comparison of (a) S_{MCM} [44] and (b) our S_{GMP} on ImageNet-1k (ID) to iNaturalist (OOD). The scores are multiplied by 100 for better illustration. KS is the Kolmogorov–Smirnov statistic. Higher KS values indicate a greater difference between the distributions. Best viewed in color.
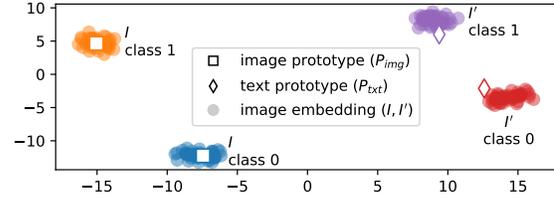

Figure 5. The t-SNE visualization of embeddings from two arbitrary classes in ImageNet-1k (Val) using SUPREME.

previous methods.

6. Conclusion

We have introduced SUPREME, a novel VLM-based few-shot tuning OOD detection method. SUPREME incorporates two key modules: BPG, which enhances image-text fusion, and ITC, which minimizes the inter- and intra-modal distances. We also proposed the generalized multi-modal prototype OOD score, S_{GMP}, using multi-modal prototypes and embeddings to improve robustness and reduce false positives. Experiments show that SUPREME consistently outperforms existing VLM-based OOD detection methods.

Limitations. In some scenarios, the ID data may not be available. Exploring methods such as image generation or retrieval from a web-scale dataset could help alleviate the dependency on the ID data. Moreover, exploring more advanced distribution models rather than the Gaussian distribution we use may further improve the performance.

Acknowledgement

References

[1] Nikolas Adaloglou, Felix Michels, Tim Kaiser, and Markus Kollmann. Adapting Contrastive Language-Image Pretrained (CLIP) Models for Out-of-Distribution Detection. *Transactions on Machine Learning Research*, 2023. 1

[2] Yichen Bai, Zongbo Han, Bing Cao, Xiaoheng Jiang, Qinghua Hu, and Changqing Zhang. ID-like Prompt Learning for Few-Shot Out-of-Distribution Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17480–17489, 2024. 1, 2, 3, 6

[3] 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. In *ICML*, 2024. 1, 3, 6

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, pages 1597–1607. PMLR, 2020. 3

[5] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and and A. Vedaldi. Describing Textures in the Wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014. 4, 6

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 2, 4, 6

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*, 2019. arXiv: 1810.04805. 3

[8] Terrance DeVries and Graham W. Taylor. Learning Confidence for Out-of-Distribution Detection in Neural Networks, 2018. arXiv:1802.04865 [stat]. 3

[9] Xuefeng Du, Yiyou Sun, Jerry Zhu, and Yixuan Li. Dream the Impossible: Outlier Imagination with Diffusion Models. In *Advances in Neural Information Processing Systems*, pages 60878–60901, 2023. 3

[10] 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*, pages 6568–6576, 2022. arXiv:2109.02748 [cs]. 1

[11] Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. Exploring the Limits of Out-of-Distribution Detection. In *Advances in Neural Information Processing Systems*, pages 7068–7081. Curran Associates, Inc., 2021. 1, 3, 6

[12] Hao Fu, Naman Patel, Prashanth Krishnamurthy, and Farshad Khorrami. CLIPScope: Enhancing Zero-Shot OOD Detection with Bayesian Scoring. pages 5346–5355, 2025. 5, 2, 7

[13] Zijian Gao, Jingyu Liu, Weiqi Sun, Sheng Chen, Dedan Chang, and Lili Zhao. CLIP2TV: Align, Match and Distill for Video-Text Retrieval, 2022. arXiv:2111.05610 [cs]. 4

[14] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. ImageBind: One Embedding Space To Bind Them All, 2023. arXiv:2305.05665 [cs]. 1, 3

[15] Tao Guo, Song Guo, and Junxiao Wang. pFedPrompt: Learning Personalized Prompt for Vision-Language Models in Federated Learning. In *Proceedings of the ACM Web Conference 2023*, pages 1364–1374, New York, NY, USA, 2023. Association for Computing Machinery. 3

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, Las Vegas, NV, 2016. IEEE Computer Society. 1, 3

[17] Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. Why ReLU Networks Yield High-Confidence Predictions Far Away From the Training Data and How to Mitigate the Problem. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 41–50, Long Beach, CA, USA, 2019. IEEE. 3

[18] Dan Hendrycks and Kevin Gimpel. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *International Conference on Learning Representations*, 2017. 1, 3, 6

[19] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep Anomaly Detection with Outlier Exposure. In *International Conference on Learning Representations*, 2018. 3

[20] Georg Hess, Adam Tonderski, Christoffer Petersson, Kalle Åström, and Lennart Svensson. LidarCLIP or: How I learned to talk to point clouds. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision (WACV)*, pages 7438–7447, 2024. 1, 3

[21] 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)*, pages 8706–8715, Nashville, TN, USA, 2021. IEEE. 1, 3, 6, 2

[22] Rui Huang, Andrew Geng, and Yixuan Li. On the Importance of Gradients for Detecting Distributional Shifts in the Wild. In *Advances in Neural Information Processing Systems*, pages 677–689. Curran Associates, Inc., 2021. 3

[23] Wenyu Jiang, Hao Cheng, MingCai Chen, Chongjun Wang, and Hongxin Wei. DOS: Diverse Outlier Sampling for Out-of-Distribution Detection. In *International Conference on Learning Representations*, 2024. 3

[24] 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 *ICLR*, 2024. 1, 3, 6, 7

[25] K J Joseph, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Towards Open World Object Detection. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5826–5836, Nashville, TN, USA, 2021. IEEE. 3

[26] Omar Khattab and Matei Zaharia. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT, 2020. arXiv:2004.12832 [cs]. 1, 3

[27] Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating Prompts: Foundational Model

Adaptation without Forgetting. In *International Conference on Computer Vision*, pages 15190–15200, 2023. 3

[28] Tung Kieu, Bin Yang, and Christian S. Jensen. Outlier Detection for Multidimensional Time Series Using Deep Neural Networks. In *IEEE International Conference on Mobile Data Management*, pages 125–134, Aalborg, Denmark, 2018. IEEE. 3

[29] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, pages 1097–1105, 2012. 6

[30] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2018. 3

[31] Brian Lester, Rami Al-Rfou, and Noah Constant. The Power of Scale for Parameter-Efficient Prompt Tuning, 2021. arXiv:2104.08691 [cs]. 3

[32] Jingyao Li, Pengguang Chen, Zexin He, Shaozuo Yu, Shu Liu, and Jiaya Jia. Rethinking Out-of-Distribution (OOD) Detection: Masked Image Modeling Is All You Need. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 11578–11589, 2023. 3

[33] Jingyao Li, Pengguang Chen, Shaozuo Yu, Shu Liu, and Jiaya Jia. MOODv2: Masked Image Modeling for Out-of-Distribution Detection, 2024. arXiv:2401.02611 [cs]. 3

[34] 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 *ACM MM*. arXiv, 2024. arXiv:2408.15566 [cs]. 3

[35] Jinglun Li, Kaixun Jiang, Zhaoyu Chen, Bo Lin, Yao Tang, Weifeng Ge, and Wenqiang Zhang. Synthesizing near-boundary OOD samples for out-of-distribution detection. In *Int. Conf. on Computer Vision (ICCV)*, 2025. 1, 3

[36] Tianqi Li, Guansong Pang, Xiao Bai, Wenjun Miao, and Jin Zheng. Learning Transferable Negative Prompts for Out-of-Distribution Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17584–17594, 2024. 1, 2, 3, 5, 6

[37] Xiang Lisa Li and Percy Liang. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online, 2021. Association for Computational Linguistics. 3

[38] Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks, 2020. arXiv:1706.02690 [cs]. 1, 6

[39] Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *Advances in neural information processing systems*, 2022. 1, 2, 4, 5, 8, 7

[40] Kai Liu, Zhihang Fu, Chao Chen, Sheng Jin, Ze Chen, Mingyuan Tao, Rongxin Jiang, and Jieping Ye. Category-Extensible Out-of-Distribution Detection via Hierarchical Context Descriptions. In *Advances in Neural Information Processing Systems*, pages 33241–33261, 2023. 1, 3, 6

[41] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based Out-of-distribution Detection. In *Advances in Neural Information Processing Systems*, pages 21464–21475. Curran Associates, Inc., 2020. 1, 3, 6

[42] Yuyuan Liu, Choubo Ding, Yu Tian, Guansong Pang, Vasileios Belagiannis, Ian Reid, and Gustavo Carneiro. Residual Pattern Learning for Pixel-Wise Out-of-Distribution Detection in Semantic Segmentation. In *International Conference on Computer Vision*, pages 1151–1161, 2023. 3

[43] George A. Miller. WordNet: A lexical database for english. *Communications of The Acm*, 38(11):39–41, 1995. tex.bibsource: dblp computer science bibliography, https://dblp.org tex.timestamp: Fri, 24 Mar 2023 16:31:07 +0100. 1, 3

[44] Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. Delving into out-of-distribution detection with vision-language representations. In *Advances in neural information processing systems*, 2022. 1, 2, 3, 4, 5, 6, 7, 8

[45] Atsuyuki Miyai, Qing Yu, Go Irie, and Kiyoharu Aizawa. LoCoOp: Few-Shot Out-of-Distribution Detection via Prompt Learning. In *Advances in Neural Information Processing Systems*, pages 76298–76310, 2023. 1, 2, 3, 5, 6, 4

[46] Harikrishna Narasimhan, Aditya Krishna Menon, Wittawat Jitkrittum, and Sanjiv Kumar. Plugin estimators for selective classification with out-of-distribution detection. In *International Conference on Learning Representations*, 2024. 3

[47] Jun Nie, Yonggang Zhang, Zhen Fang, Tongliang Liu, Bo Han, and Xinmei Tian. Out-of-Distribution Detection with Negative Prompts. In *ICLR*, 2024. 1, 3, 6

[48] OpenAI. GPT-4 Technical Report, 2023. arXiv:2303.08774 [cs]. 1, 3

[49] Jaewoo Park, Yoon Gyo Jung, and Andrew Beng Jin Teoh. Nearest Neighbor Guidance for Out-of-Distribution Detection. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1686–1695, Paris, France, 2023. IEEE. 3

[50] Sangha Park, Jisoo Mok, Dahuin Jung, Saehyung Lee, and Sungroh Yoon. On the Powerfulness of Textual Outlier Exposure for Visual OoD Detection. In *Advances in Neural Information Processing Systems*, pages 51675–51687, 2023. 1

[51] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. ISSN: 2640-3498. 1, 3, 7

[52] Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts, 2020. arXiv:2010.15980 [cs]. 3

[53] Yiyou Sun, Chuan Guo, and Yixuan Li. ReAct: Out-of-distribution Detection With Rectified Activations. In *Advances in Neural Information Processing Systems*, pages 144–157. Curran Associates, Inc., 2021. 3

[54] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-Distribution Detection with Deep Nearest Neighbors. In *Proceedings of the 39th International Conference on Machine Learning*, pages 20827–20840. PMLR, 2022. ISSN: 2640-3498. 1, 3, 6

[55] Leitian Tao, Xuefeng Du, Jerry Zhu, and Yixuan Li. Non-parametric Outlier Synthesis. In *ICLR*, 2023. 6

[56] Michael Tschannen, Basil Mustafa, and Neil Houlsby. CLIPPO: Image-and-Language Understanding from Pixels Only, 2023. arXiv:2212.08045 [cs]. 4

[57] 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*, pages 8769–8778, Salt Lake City, UT, 2018. IEEE. 4, 6

[58] Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. ViM: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 4921–4930, 2022. 1, 3, 6

[59] 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)*, pages 1802–1812, Paris, France, 2023. IEEE. 1, 3, 6

[60] Hongxin Wei, Renchunzi Xie, Hao Cheng, Lei Feng, Bo An, and Yixuan Li. Mitigating Neural Network Overconfidence with Logit Normalization, 2022. Version Number: 2. 3

[61] Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3485–3492, San Francisco, CA, USA, 2010. IEEE. 4, 6

[62] Mingyu Xu, Zheng Lian, Bin Liu, and Jianhua Tao. VRA: Variational Rectified Activation for Out-of-distribution Detection. In *Advances in Neural Information Processing Systems*, 2023. 3

[63] Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, Xuefeng Du, Kaiyang Zhou, Wayne Zhang, Dan Hendrycks, Yixuan Li, and Ziwei Liu. OpenOOD: Benchmarking Generalized Out-of-Distribution Detection, 2022. Version Number: 1. 1, 3

[64] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized Out-of-Distribution Detection: A Survey, 2024. arXiv:2110.11334 [cs]. 3

[65] Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-Language Prompt Tuning with Knowledge-Guided Context Optimization. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6757–6767, Vancouver, BC, Canada, 2023. IEEE. 3

[66] Geng Yu, Jianing Zhu, Jiangchao Yao, and Bo Han. Self-Calibrated Tuning of Vision-Language Models for Out-of-Distribution Detection. 2024. 6

[67] Fanhu Zeng, Zhen Cheng, Fei Zhu, Hongxin Wei, and Xu-Yao Zhang. Local-Prompt: Extensible Local Prompts for Few-Shot Out-of-Distribution Detection. 2025. 6

[68] Jingyang Zhang, Jingkang Yang, Pengyun Wang, Haoqi Wang, Yueqian Lin, Haoran Zhang, Yiyou Sun, Xuefeng Du, Kaiyang Zhou, Wayne Zhang, Yixuan Li, Ziwei Liu, Yiran Chen, and Hai Li. OpenOOD v1.5: Enhanced Benchmark for Out-of-Distribution Detection, 2023. arXiv:2306.09301 [cs]. 2, 3, 6

[69] Yabin Zhang and Lei Zhang. AdaNeg: Adaptive negative proxy guided OOD detection with vision-language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 1, 3

[70] Yabin Zhang, Wenjie Zhu, Chenhang He, and Lei Zhang. LAPT: Label-driven Automated Prompt Tuning for OOD Detection with Vision-Language Models. In *ECCV*. arXiv, 2024. arXiv:2407.08966 [cs]. 1, 3

[71] 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, 2018. 4, 6

[72] Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. Learning Placeholders for Open-Set Recognition. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4399–4408, Nashville, TN, USA, 2021. IEEE. 3

[73] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2022. 3, 4, 5, 6

[74] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to Prompt for Vision-Language Models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 3, 5, 6, 4

Mitigating the Modality Gap: Few-Shot Out-of-Distribution Detection with Multi-modal Prototypes and Image Bias Estimation

Supplementary Material

In this Appendix, we present the details and complete results corresponding to Tab. 1 in Sec. 8.1, full theoretical motivation in Sec. 8.2, additional ablation studies in Sec. 9.2, benchmark dataset specifications in Sec. 9.3, and the proof of Theorem 1 in Sec. 10.

7. OOD Examples

In Figs. 2 and 6, we visualize the embedding space using t-SNE projection across four different OOD datasets (iNaturalist, Texture, Places, and SUN) generated with CLIP (ViT-B/16). A pattern emerges: OOD images (red crosses) are consistently positioned closer to ID text prototypes than their corresponding ID images across all four scenarios. For instance, in Fig. 6(a), OOD images (red cross) are positioned closer to the "sulphur butterfly" text prototype (purple circle) than the corresponding ID images (blue pentagon). This counterintuitive positioning is systematically observed across diverse semantic categories—from biological entities (sulphur butterfly, starfish) to textures (honeycomb, cherimoya) and environmental scenes (jay/cougar, rapeseed/hay).

This phenomenon reveals a fundamental limitation in current embedding spaces: *positional proximity does not guarantee semantic similarity*. The t-SNE visualization demonstrates that geometric distance in the learned representation space can be misleading, as semantically unrelated OOD samples may appear closer to class prototypes than genuine in-distribution examples. This misalignment between spatial proximity and semantic similarity undermines the reliability of distance-based OOD detection methods. It motivates our design of multimodal prototypes that better capture the true semantic boundaries of in-distribution data.

8. RQ1

8.1. More Empirical Evidence for RQ1

To empirically justify the effectiveness of multi-modal (image and text) prototypes, we present experiments using S_{MCM} and S_{MMP} on four OOD datasets (*i.e.*, iNaturalist, SUN, Places, and Texture) as shown in Tab. 9. We use CLIP-B/16 as the base model for S_{MCM} and S_{MMP}. All the training images from ImageNet-100 are used for generating image prototypes. We also include S_{MCM} with image prototypes as a comparison baseline.

Our method, S_{MMP}, consistently demonstrates superior performance in terms of FPR95 across all four tasks, achieving the lowest values, with notable improvements on iNat-

Methods	FPR95 (%) ↓	AUROC (%) ↑
ImageNet-100 → 4 datasets (Average)		
Text prototypes	32.58	94.48
Image prototypes	31.09	94.74
Text + Image prototypes	24.18	95.79
ImageNet-100 → iNaturalist		
Text prototypes	18.13	96.77
Image prototypes	26.06	93.09
Text + Image prototypes	**14.76**	**97.36**
ImageNet-100 → SUN		
Text prototypes	36.45	**94.54**
Image prototypes	39.25	93.17
Text + Image prototypes	**30.28**	92.51
ImageNet-100 → Places		
Text prototypes	34.52	**94.36**
Image prototypes	40.59	96.43
Text + Image prototypes	**34.04**	93.92
ImageNet-100 → Texture		
Text prototypes	41.22	92.25
Image prototypes	18.46	96.27
Text + Image prototypes	**17.66**	**96.66**

Table 9. Comparison of OOD detection results on four OOD tasks. We use CLIP-B/16 for S_{MCM} and S_{MMP} (ours, Eq. (4)). FPR95 represents the false positive rate of OOD images when the true positive rate of ID images is at 95% while AUROC is the area under the receiver operating characteristic curve. Best in **bold**.

uralist (14.76 vs. 18.13) and Texture (17.66 vs. 41.22), highlighting its robustness in reducing false positives. For AUROC, S_{MMP} achieves the highest score on iNaturalist (97.36) and Texture (96.66). However, in the SUN and Places datasets, while S_{MMP} maintains competitive performance, it slightly underperforms S_{MCM} in AUROC (92.51 vs. 94.54 on SUN and 93.92 vs. 94.36 on Places).

Similarly, we observe that S_{MMP} outperforms S_{MCM} with image prototypes on most of the metrics, while S_{MCM} with image prototypes has slightly better performance compared to S_{MCM} (with text prototypes). This shows that the modality gap negatively impacts the performance.

Overall, the results underscore the effectiveness of S_{MMP} and multi-modal prototypes in reducing false positives and the performance of VLM-based OOD detection.

8.2. Theoretical Evidence for RQ1

To further understand why image prototypes improve performance, we present a theoretical analysis in Theorem 1.
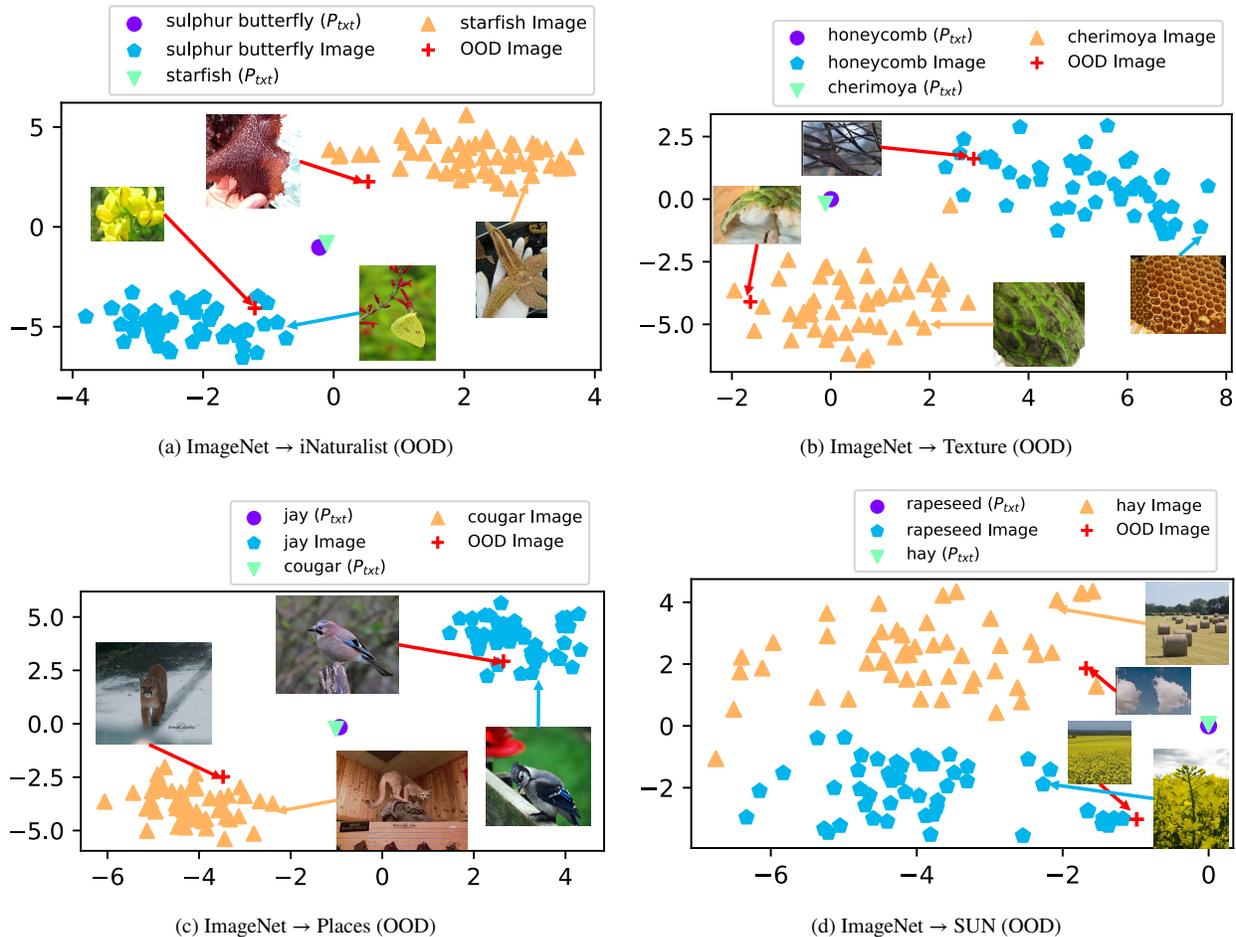
Figure 6. Examples where OOD images from four OOD datasets are located closer to ID text prototypes compared to ID images (ImageNet-1k) generated with CLIP (ViT-B/16).

Theorem 1 (Multi-modal Prototypes Increase Score Separation between ID and OOD Data). *Assuming that the OOD data is not drawn from any ID distribution, we have,*

$$\mathbb{E}\left[S_{MMP}(I_{ID}) - S_{MMP}(I_{OOD})\right] \geq \mathbb{E}\left[S_{MCM}(I_{ID}) - S_{MCM}(I_{OOD})\right],$$
(11)

where I_{ID} and I_{OOD} are the image embeddings of ID and OOD samples. We omit multi-modal prototypes for clarity.

This greater score separation enhances the separation between ID and OOD data, improving detection performance, under general assumptions used by previous works [12, 39]. And it further improves performance. The proof is provided in Sec. 10.

9. Experiments

9.1. Results on OpenOOD (CSID)

We evaluate the performance of SUPREME on the OpenOOD benchmark (CSID) [21] using the ImageNet-

Methods	FPR95 (%) ↓	AUROC (%) ↑
LoCoOp	85.60	60.73
SUPREME	**64.97**	**75.32**

Table 10. OOD Detection results with ID data of ImageNet-1k on the OpenOOD benchmark (CSID). Best in **Bold**.

1k dataset as the ID data in Tab. 10. The results show that SUPREME outperforms LoCoOp, achieving the lowest FPR95 and highest AUROC, demonstrating its effectiveness in OOD detection across different benchmarks.

9.2. Ablations

To understand the effectiveness of SUPREME, we present the results of several ablation studies using CLIP (ViT-B/16) on ImageNet-1k → four OOD datasets (*i.e.*, iNaturalist, SUN, Places, and Texture). Our experiments examine the effectiveness of the proposed components (*i.e.*, ITC, BPG,

ITC	BPG	S_{GMP}	iNaturalist		SUN		Places		Texture		Average	
			FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑
			30.91	94.61	37.59	92.57	44.69	89.77	57.77	86.11	42.74	90.77
✓	✓		8.56	98.48	22.36	95.38	33.60	92.55	39.57	**92.75**	26.02	94.79
✓		✓	20.86	95.06	22.04	94.92	30.22	91.64	39.38	89.46	28.12	92.77
	✓	✓	16.10	96.50	23.81	94.53	32.63	91.94	42.41	90.11	28.74	93.27
✓	✓(w/o Bias)	✓	21.55	97.79	16.37	**98.00**	26.25	92.24	34.25	91.56	24.61	94.90
✓	✓	✓	**5.91**	**98.73**	15.85	96.08	**26.18**	**93.19**	**34.88**	91.93	**20.70**	**95.38**
	w. ResNet-50		21.93	96.22	37.73	91.00	43.27	87.41	28.14	93.51	32.77	92.04
	w. ResNet-101		6.13	98.58	20.61	95.03	33.16	90.94	41.31	90.70	25.30	93.81

Table 11. Effectiveness of different modules, the proposed OOD score S_{GMP}, and different backbones. When S_{GMP} is not combined with BPG, it degenerates to S_{MCM}.

ReAct	S_{EOE}	iNaturalist		SUN		Places		Texture		Average	
		FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑
		5.91	**98.73**	15.85	96.08	26.18	93.19	34.88	91.93	20.70	95.38
✓		8.03	97.93	17.39	95.51	24.66	93.32	**33.21**	**92.13**	20.82	94.72
	✓	**4.85**	98.71	14.90	96.13	23.82	93.68	34.96	91.22	19.63	94.93
✓	✓	7.27	98.32	**11.79**	**97.23**	**20.28**	**94.36**	38.87	90.97	**19.55**	**95.22**

Table 12. Performance of SUPREME combined with post-hoc methods, i.e., ReAct [53] and S_{EOE} [3].

Methods	ResNet-50		ResNet-101	
	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑
NegPrompt	57.19	85.70	50.35	88.04
LoCoOp	34.45	91.95	-	-
SUPREME	**30.55**	**92.90**	**25.30**	**93.81**

Table 13. Performance comparison using different base models.

and S_{GMP}), the impact of combining SUPREME with post-hoc methods, training data efficiency (number of shots), the statistical evaluation of S_{GMP}, and sensitivity to hyperparameters.

Evaluating the effectiveness of the proposed modules (BPG and ITC) and S_{GMP} (full results). To evaluate the effectiveness of the proposed modules and S_{GMP}, we present detailed results of a combinatorial ablation study in Tab. 11. Performance is evaluated across four datasets (i.e., iNaturalist, SUN, Places, and Texture).

Combining SUPREME with post-hoc methods. To understand how SUPREME combines with existing representative post-hoc methods, i.e., ReAct [53] and S_{EOE} [3], we present a systematic study in Tab. 12. We directly apply ReAct and S_{EOE} on our best model as post-hoc methods, which means we do not use them during training. For ReAct, we set the threshold at 0.95, while for S_{EOE}, we follow the standard parameters. We see that using ReAct alone makes the performance worse with an average FPR95 of 20.82 and AUROC of 94.72. On the other hand, S_{EOE} boosts average

FPR95 from 20.70 to 19.63. When combining ReAct and S_{EOE}, the FPR95 is boosted to 19.55 on average. However, using ReAct and/or S_{EOE} decreases the average AUROC compared to vanilla SUPREME, showing that these post-hoc methods might only help reduce false positives. The experiments show that SUPREME is compatible with these post-hoc methods without performance drops.

Effectiveness of SUPREME with different backbones. The results in Tab. 13 show that SUPREME consistently outperforms other methods across different backbones, including ResNet-50 [16] and ResNet-101 [16]. Specifically, SUPREME achieves 30.55 for FPR95 and 92.90 for AUROC with ResNet-50, and 25.30 for FPR95 and 93.81 for AUROC with ResNet-101, demonstrating its robustness and effectiveness across various model architectures.

Computational cost of the training and inference. Table 14 compares parameter efficiency and inference scalability between the baseline MCM method and our proposed SUPREME, using CLIP-B/16 as the base model. While both methods exhibit comparable total parameter sizes, SUPREME introduces 567,842 trainable parameters to achieve its enhanced functionality, representing a negligible increase of 0.46% in total parameters. SUPREME obtains image prototypes with a complexity of $O(N_{base})$, while maintaining inference complexity at $O(1)$ given one test input. Also, the calculation of image prototypes can be conducted before inference, and image prototypes can be shared among all the test samples, which avoids introducing overhead during inference compared with MCM. It demonstrates

Methods	Parameters	Trainable Parameters	Image Prototypes	Inference
MCM	124,323,841	0	-	$O(1)$
SUPREME	124,891,683 (↑567,842)	567,842	$O(N_{base})$	$O(1)$

Table 14. Comparison of (trainable) parameters and inference efficiency between MCM and SUPREME with the base model as CLIP-B/16.
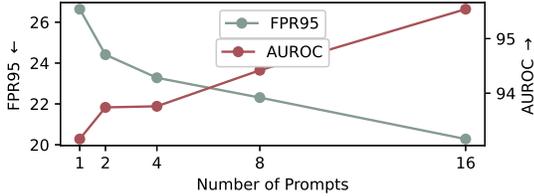

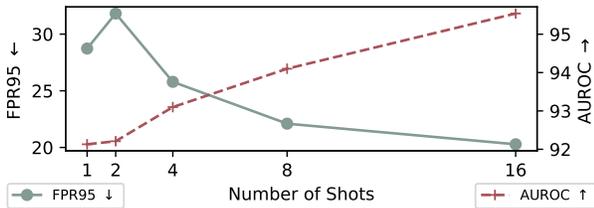
Figure 7. Performance comparison on the length of prompts.

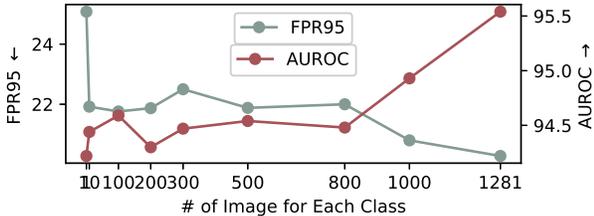Figure 8. The impact of different sizes of fine-tuning data.

Figure 9. Performance comparison on the number of based images used for obtaining image prototypes. "1281" corresponds to using all the training images as the base images.

Figure 10. Hyperparameter sensitivity (α and β). We report the average FPR95 and AUROC.

the scalability of SUPREME without computational overhead during inference, effectively addressing challenges of learnable flexibility and inference efficiency.

Impact of the length of prompts. In the main experiments, we follow previous works [45, 73, 74] and set the length of prompts to $L = 16$. Here, we would like to validate that 16 is the best choice for SUPREME. We conducted experiments with SUPREME † and results shown in Fig. 7 indicate that among the choices from 1 to 16, 16 obtains the lowest FPR95 and highest AUROC. Increasing the length of prompts improves the OOD detection performance.

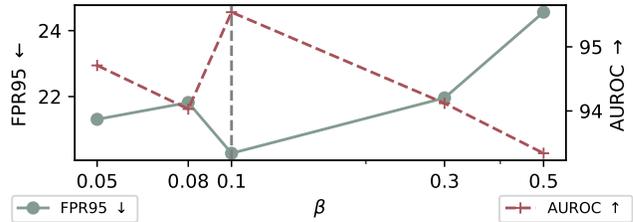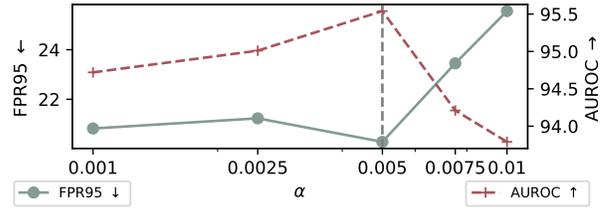**How much data do we need for few-shot tuning?** We eval-

uate the effect of the number of shots for few-shot tuning on the model's performance in Fig. 8 with SUPREME †. As the number of shots increases from 1 to 16, we observe a consistent improvement in both metrics. Specifically, FPR95 decreases from 28.73 at 1-shot to 20.28 at 16-shot, indicating that more shots lead to better separability between ID and OOD data. Similarly, AUROC improves from 92.13 at 1-shot to a peak of 95.54 at 16-shot, demonstrating enhanced overall detection performance. These results highlight the robustness of SUPREME in leveraging additional labeled examples, with diminishing returns on FPR95 as data size increases. While higher shot numbers provide the best performance, the model performs reasonably well even with a limited number of shots, showcasing its flexibility in low-data scenarios.

Impact of base image size for constructing ID image prototypes. In S_{GMP}, we use a set of base images to construct image prototypes $\{P_{i,c}\}_{c=1}^{C}$. To understand how the size of base images affects performance, we conduct experiments varying its size by randomly selecting a subset of the training images. The results in Fig. 9 show that by increasing the number of images for each class, the performance of SUPREME is improved, as image prototypes are closer to the real distribution. FPR95 and AUROC reach the best when using all the images. However, as shown in Tab. 2, the performance is already competitive when using the same set of
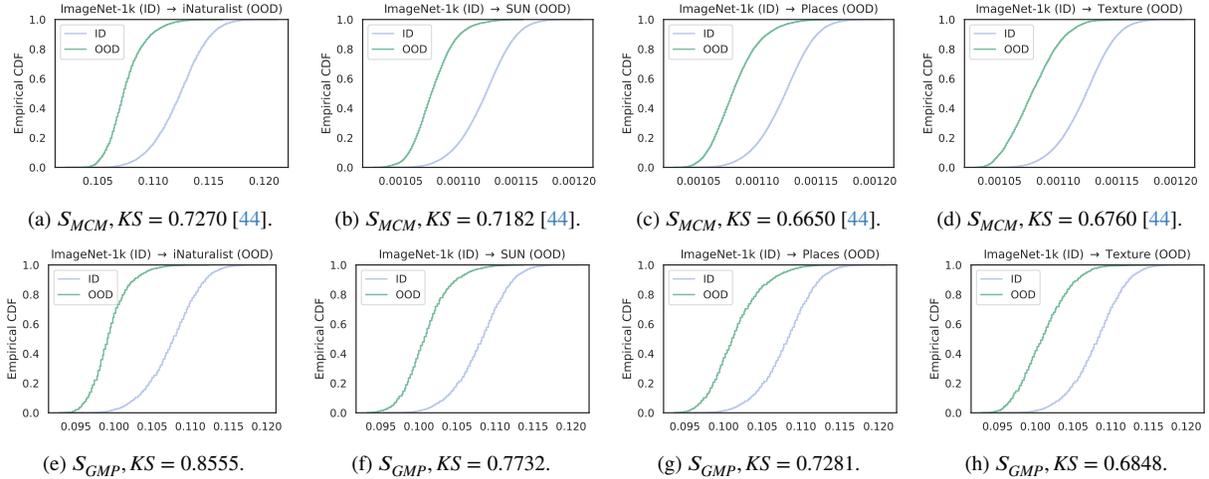
| (a) $S_{MCM}, KS = 0.7270$ [44]. | (b) $S_{MCM}, KS = 0.7182$ [44]. | (c) $S_{MCM}, KS = 0.6650$ [44]. | (d) $S_{MCM}, KS = 0.6760$ [44]. |
| (e) $S_{GMP}, KS = 0.8555$. | (f) $S_{GMP}, KS = 0.7732$. | (g) $S_{GMP}, KS = 0.7281$. | (h) $S_{GMP}, KS = 0.6848$. |

Figure 11. The comparison between S_{MCM} [44] and our proposed S_{GMP} score (Eq. (10)) on ImageNet-1k (ID) to four OOD datasets, *i.e.*, iNaturalist, SUN, Places, and Texture. Best viewed in color.

$\left\|\vec{\Delta}_{\text{gap}}\right\|$	Modality 1	Modality 2	ImageNet (ID) ↓	iNaturalist ↓	Places ↓	Texture ↓	SUN ↓
MCM	P_{txt}	I	0.8679	0.9839	0.9101	0.9373	0.9252
SUPREME	P_{txt}	I	0.6135	0.7159	0.8961	0.9273	0.9002
	P_{txt}	$f_{\text{img-txt}}(I)$	0.1938	**0.2238**	**0.2376**	**0.2559**	**0.2454**
	P_{img}	I	**0.1348**	0.4531	0.3474	0.3333	0.3447

Table 15. Quantifying the modality gap with different methods. We use CLIP (ViT-B/16) as the base model.

few-shot data for fine-tuning, indicating that the model is robust to the size of base images.

Hyperparameter (α and β) sensitivity. We analyze the effect of the trade-off hyperparameters α and β (Sec. 4.2) on the model's performance, as shown in Fig. 10 with SUPREME †. For α, we observe a trade-off between FPR95 and AUROC as its value increases. Specifically, as α grows from 0.001 to 0.01, FPR95 decreases initially from 20.82 to 20.28 at $\alpha = 0.005$, indicating improved separability, but then rises to 25.56 at $\alpha = 0.01$. Meanwhile, AUROC peaks at 95.54 for $\alpha = 0.005$ before declining, indicating that excessively large values of α may overemphasize certain features, leading to diminished overall performance. Similarly, β demonstrates a comparable trend. The FPR95 achieves its lowest value of 20.28 at $\beta = 0.1$, while AUROC peaks at 95.54. The results of both hyperparameters suggest that SUPREME is not sensitive to the parameters and that moderate values ($\alpha = 0.005, \beta = 0.1$) balance the trade-off between false positive rate and overall discriminative ability.

Full statistical evaluation of S_{GMP}. We visualize the empirical cumulative distribution (empirical CDF) and the Kolmogorov–Smirnov (KS) test statistics for S_{MCM} and our proposed S_{GMP} in Fig. 11. The evaluation uses our best model trained on the ImageNet-1k (ID) dataset, with

iNaturalist, SUN, Places, and Texture serving as OOD datasets. The empirical CDF clearly shows that S_{GMP} improves the gap between ID and OOD data more effectively. Furthermore, the KS test statistics improve ceonsiderably from $0.7270, 0.7182, 0.6650$, and 0.6760 (S_{MCM}) to $0.8724, 0.7798, 0.7219$, and 0.6874 (S_{GMP}) across the four OOD datasets. These results indicate that S_{GMP} achieves better separation between ID and OOD scores compared to S_{MCM}.

Quantifying modality gap. To quantify the modality gap, following [39], we formulate the modality gap as the ℓ_2 distance between the center of image embeddings and (class) text embeddings, *i.e.*, $\left\|\vec{\Delta}_{\text{gap}}\right\| = \left\|\frac{1}{C}\sum_{c=1}^{C} P_{\text{txt},c} - \frac{1}{N_{\text{image}}}\sum_{i=1}^{N_{\text{image}}} I_i\right\|_2$, where $P_{\text{txt},c}$ is the embedding of c-th class, N_{image} is the number of images, and I_i is the image embedding of the i-th image. We use the test set of each dataset for calculating the modality gap. We also calculate the modality gap between the center of mapped image embeddings $f_{\text{txt-img}}(I)$ and image ID prototypes P_{img} as shown in Tab. 15. The results demonstrate that SUPREME significantly reduces the modality gap across all datasets. Specifically, compared to the baseline MCM, SUPREME reduces the original modality gap by approximately 25-30%,

which shows that the design of SUPREME could effectively reduce the gap. Moreover, when comparing text prototypes with our mapped image embeddings $f_{t2i}(I)$, we achieve a dramatic reduction of over 75% in the modality gap (row 3), indicating that our mapping function effectively bridges the gap between modalities. Additionally, the comparison between image ID prototypes and original images (row 4) shows substantial improvement.

9.3. Details of Benchmarking Datasets

ImageNet-100, ImageNet-10, and ImageNet-20. Following MCM [44], we choose 100/10/20 classes from ImageNet-1k [29] to form ImageNet-100, ImageNet-10, and ImageNet-20. The chosen classes for each dataset are as follows:

- **ImageNet-100**: n03877845, n03000684, n03110669, n03710721, n02825657, n02113186, n01817953, n04239074, n02002556, n04356056, n03187595, n03355925, n03125729, n02058221, n01580077, n03016953, n02843684, n04371430, n01944390, n03887697, n04037443, n02493793, n01518878, n03840681, n04179913, n01871265, n03866082, n03180011, n01910747, n03388549, n03908714, n01855032, n02134084, n03400231, n04483307, n03721384, n02033041, n01775062, n02808304, n13052670, n01601694, n04136333, n03272562, n03895866, n03995372, n06785654, n02111889, n03447721, n03666591, n04376876, n03929855, n02128757, n02326432, n07614500, n01695060, n02484975, n02105412, n04090263, n03127925, n04550184, n04606251, n02488702, n03404251, n03633091, n02091635, n03457902, n02233338, n02483362, n04461696, n02871525, n01689811, n01498041, n02107312, n01632458, n03394916, n04147183, n04418357, n03218198, n01917289, n02102318, n02088364, n09835506, n02095570, n03982430, n04041544, n04562935, n03933933, n01843065, n02128925, n02480495, n03425413, n03935335, n02971356, n02124075, n07714571, n03133878, n02097130, n02113799, n09399592, n03594945.
- **ImageNet-10**: n04552348, n04285008, n01530575, n02123597, n02422699, n02107574, n01641577, n03417042, n02389026, n03095699.
- **ImageNet-20**: n04147183, n02951358, n02782093, n04389033, n03773504, n02917067, n02317335, n01632458, n01630670, n01631663, n02391049, n01693334, n01697457, n02120079, n02114367, n02132136, n03785016, n04310018, n04266014, n04252077.

Other datasets. Similarly, we use subsets from iNaturalist [57], SUN [61], Places [71], and Texture [5] as OOD datasets, which are created by Huang and Li [21].

- **iNaturalist** contains images from the natural world with images from 5089 classes, belonging to 13 super-categories, such as Plantae (Plant), Insecta (Insect), Aves (Bird), Mammalia (Mammal). The subset containing 110 plant classes not present in ImageNet-1k is chosen as the OOD test set. The classes are as follows: *Coprosma lucida, Cucurbita foetidissima, Mitella diphylla, Selaginella bigelovii, Toxicodendron vernix, Rumex obtusifolius, Ceratophyllum demersum, Streptopus amplexifolius, Portulaca oleracea, Cynodon dactylon, Agave lechuguilla, Pennantia corymbosa, Sapindus saponaria, Prunus serotina, Chondracanthus exasperatus, Sambucus racemosa, Polypodium vulgare, Rhus integrifolia, Woodwardia areolata, Epifagus virginiana, Rubus idaeus, Croton setiger, Mammillaria dioica, Opuntia littoralis, Cercis canadensis, Psidium guajava, Asclepias exaltata, Linaria purpurea, Ferocactus wislizeni, Briza minor, Arbutus menziesii, Corylus americana, Pleopeltis polypodioides, Myoporum laetum, Persea americana, Avena fatua, Blechnum discolor, Physocarpus capitatus, Ungnadia speciosa, Cercocarpus betuloides, Arisaema dracontium, Juniperus californica, Euphorbia prostrata, Leptopteris hymenophylloides, Arum italicum, Raphanus sativus, Myrsine australis, Lupinus stiversii, Pinus echinata, Geum macrophyllum, Ripogonum scandens, Echinocereus triglochidiatus, Cupressus macrocarpa, Ulmus crassifolia, Phormium tenax, Aptenia cordifolia, Osmunda claytoniana, Datura wrightii, Solanum rostratum, Viola adunca, Toxicodendron diversilobum, Viola sororia, Uropappus lindleyi, Veronica chamaedrys, Adenocaulon bicolor, Clintonia uniflora, Cirsium scariosum, Arum maculatum, Taraxacum officinale officinale, Orthilia secunda, Eryngium yuccifolium, Diodia virginiana, Cuscuta gronovii, Sisyrinchium montanum, Lotus corniculatus, Lamium purpureum, Ranunculus repens, Hirschfeldia incana, Phlox divaricata laphamii, Lilium martagon, Clarkia purpurea, Hibiscus moscheutos, Polanisia dodecandra, Fallugia paradoxa, Oenothera rosea, Proboscidea louisianica, Packera glabella, Impatiens parviflora, Glaucium flavum, Cirsium andersonii, Heliopsis helianthoides, Hesperis matronalis, Callirhoe pedata, Crocosmia × crocosmiiflora, Calochortus albus, Nuttallanthus canadensis, Argemone albiflora, Eriogonum fasciculatum, Pyrrhopappus pauciflorus, Zantedeschia aethiopica, Melilotus officinalis, Peritoma arborea, Sisyrinchium bellum, Lobelia siphilitica, Sorghastrum nutans, Typha domingensis, Rubus laciniatus, Dichelostemma congestum, Chimaphila maculata, Echinocactus texensis.*
- **SUN** contains 899 classes that cover indoor, urban, and natural places. We use the subset that contains 50 natural objects that do not overlap with ImageNet-1k. The classes we use are *badlands, bamboo forest, bayou, botanical*

garden, canal (natural), canal (urban), catacomb, cavern (indoor), corn field, creek, crevasse, desert (sand), desert (vegetation), field (cultivated), field (wild), fishpond, forest (broadleaf), forest (needleleaf), forest path, forest road, hayfield, ice floe, ice shelf, iceberg, islet, marsh, ocean, orchard, pond, rainforest, rice paddy, river, rock arch, sky, snowfield, swamp, tree farm, trench, vineyard, waterfall (block), waterfall (fan), waterfall (plunge), wave, wheat field, herb garden, putting green, ski slope, topiary garden, vegetable garden, formal garden.

- **Places** contains photos labeled with scene semantic categories from three macro-classes: Indoor, Nature, and Urban. We use a subset sampled from 50 categories that are not present in ImageNet-1k. The classes we use are *badlands, bamboo forest, canal (natural), canal (urban), corn field, creek, crevasse, desert (sand), desert (vegetation), desert road, field (cultivated), field (wild), field road, forest (broadleaf), forest path, forest road, formal garden, glacier, grotto, hayfield, ice floe, ice shelf, iceberg, igloo, islet, japanese garden, lagoon, lawn, marsh, ocean, orchard, pond, rainforest, rice paddy, river, rock arch, ski slope, sky, snowfield, swamp, swimming hole, topiary garden, tree farm, trench, tundra, underwater (ocean deep), vegetable garden, waterfall, wave, wheat field.*

- **Texture** contains images of textures and abstracted patterns. As no categories overlap with ImageNet1k, we use the entire dataset.

10. Proof of Theorem 1

We first introduce three necessary assumptions before proceeding with the proof of our theorem.

Assumption 1 (ID image embeddings follow the distribution centered at ID image prototypes). *Image embeddings $\mathbf{x}_c$ are drawn from the symmetric distribution $\mathcal{D}_c$ which is centered at the image prototypes $P_{i,c}$ in the embedding space.*

Assumption 2 (The image-text modality gap exists). *The image embeddings are closer to the ground truth class's image prototypes than the text prototypes due to the modality gap.*

Remark. *These two assumptions are the basic assumptions about the multi-modal prototypes and the modality gap. Based on the findings from a recent study [39], these two hold true for most real-world scenarios.*

Assumption 3 (Models are well trained). *Cosine similarity between embeddings from the same class distribution $\mathcal{D}_i$ is higher on average than between embeddings from different distributions. And the image embeddings from the c-th class are equally dissimilar to other classes.*

Remark. *Here we only consider the optimal representation extractors without any other assumptions on the models.*

Assumption 4 (OOD embeddings are not sampled from any ID distribution [12]). *OOD image embeddings $\mathbf{x}_{OOD}$ are drawn from a distribution $\mathcal{D}_{OOD}$ different from any $\mathcal{D}_c^*, \forall c \in \{1..C\}$ and $\forall * \in \{text, image\}$. In addition, the OOD image embeddings are equally similar to all ID text/image classes.*

Remark. *This assumption is a general basic assumption used by previous studies [12] on the OOD data, which provides us with an overall characteristic of the OOD data.*

Under the above assumptions, we first would like to prove the following inequality holds:

$$\Delta_{image} = \mathbb{E}\left[\max_{i\in\{1..C\}} \sigma_i^{image}(\mathbf{x}_{ID}) - \max_{i\in\{1..C\}} \sigma_i^{image}(\mathbf{x}_{OOD})\right]$$

$$\geq \Delta_{text} = \mathbb{E}\left[\max_{i\in\{1..C\}} \sigma_i^{text}(\mathbf{x}_{ID}) - \max_{i\in\{1..C\}} \sigma_i^{text}(\mathbf{x}_{OOD})\right]$$

$$\text{s.t. } \sigma_i^* = \frac{\exp(s_{*,i})}{\sum_{k=1}^N \exp(s_k)} \text{ and } s_{*,i} = cos(\mathbf{x}, P_{*,i})$$

$$\forall * \in \{image, text\}, \tag{12}$$

where $\mathbf{x}_{ID}$ and $\mathbf{x}_{OOD}$ are samples from any ID distributions and OOD samples.

First, for ID samples, with Assumption 3, we have

$$\mathbb{E}\left[\max_{c\in\{1..C\}} \sigma_c^*(\mathbf{x}_{ID})\right] = \frac{\exp\left(\mu_{intra}^*\right)}{\exp\left(\mu_{intra}^*\right) + (C-1)\exp\left(\mu_{inter}^*\right)}$$

$$\forall * \in \{text, image\}, \tag{13}$$

where $\mu_{intra}^* = \max_{c\in\{1..C\}} \mathbb{E}[cos(\mathbf{x}_{ID}, P_{*,c})]$, $\mu_{inter}^* = \mathbb{E}[cos(\mathbf{x}_{ID}, I_{c_j})], \forall c_j \in \{1..C\} \setminus \{c_{ID}\}$, and C_{ID} is the ground-truth class for the input $\mathbf{x}_{ID}$.

With assumption 4, for OOD samples, we have

$$\mathbb{E}\left[\max_{c\in\{1..C\}} \sigma_c^{image}(\mathbf{x}_{OOD})\right] = \mathbb{E}\left[\max_{c\in\{1..C\}} \sigma_c^{text}(\mathbf{x}_{OOD})\right] = \frac{1}{C}. \tag{14}$$

Then we have

$$\Delta_{image} = \mathbb{E}\left[\max_{c\in\{1..C\}} \sigma_c^{image}(\mathbf{x}_i)\right] - \frac{1}{C}. \tag{15}$$

Similarly, we have

$$\Delta_{text} = \mathbb{E}\left[\max_{c\in\{1..C\}} \sigma_c^{text}(\mathbf{x}_i)\right] - \frac{1}{C}. \tag{16}$$

Since $\mu_{intra}^{image} > \mu_{intra}^{text}$ (Assumption 2), we have

$$\Delta_{image} > \Delta_{text}. \tag{17}$$

By simple algebra, we have

$$\mathbb{E}\left[\frac{\max_{i\in\{1..C\}}\sigma_i^{image}(\mathbf{x}_{ID}) + \max_{i\in\{1..C\}}\sigma_i^{text}(\mathbf{x}_{ID})}{2}\right.$$
$$\left. -\frac{\max_{i\in\{1..C\}}\sigma_i^{image}(\mathbf{x}_{OOD}) + \max_{i\in\{1..C\}}\sigma_i^{text}(\mathbf{x}_{OOD})}{2}\right]$$
$$\geq\mathbb{E}\left[\max_{i\in\{1..C\}}\sigma_i^{text}(\mathbf{x}_{ID}) - \max_{i\in\{1..C\}}\sigma_i^{text}(\mathbf{x}_{OOD})\right],$$
$$\mathbb{E}\left[S_{MMP}(\mathbf{x}_{ID}) - S_{MMP}(\mathbf{x}_{OOD})\right]$$
$$\geq\mathbb{E}\left[S_{MCM}(\mathbf{x}_{ID}) - S_{MCM}(\mathbf{x}_{OOD})\right] .$$

$$(18)$$

This completes the proof.