Enhanced OoD Detection through Cross-Modal Alignment of Multi-Modal Representations

Supplementary Material

A. Detailed Experimental Settings

A.1. Datasets

MOS Benchmark Datasets. Our experiments utilize the MOS benchmark dataset [12], which has been used in prior studies [14, 31, 34, 46]. The MOS benchmark includes ImageNet-1k [3] as the in-distribution (ID) dataset with a validation set with 50,000 images. The out-of-distribution (OoD) dataset consists of four datasets: SUN [48], Places [50], Textures [2], and iNaturalist [42], with no overlap with ImageNet-1k. The test OoD datasets include 10,000 images each from iNaturalist, SUN, Places, and 5,640 images from Textures.

OpenOOD v1.5 Benchmark Datasets. OpenOOD v1.5 [49] includes six benchmarks: four standard OoDD benchmarks and two full-spectrum OoDD benchmarks. For our experiments, we utilize the standard OoDD benchmark with ImageNet-1k, which consists of 45,000 images for testing and 5,000 images for validation. This benchmark dataset includes two scenarios: Near-OoD and Far-OoD scenarios. The Near-OoD scenario includes SSB-hard [43] (49,000 images across 980 categories) and NINCO [1] (5,879 images), while the Far-OoD scenario contains iNaturalist (10,000 images), Textures (5,640 images), and OpenImage-O [45] (1,763 images).

A.2. Implementation details

Our proposed method, Cross-Modal Alignment (CMA), was implemented using Python 3.9.18 and PyTorch 1.12.0+cu116. All experiments were conducted on 8 NVIDIA A6000 GPUs, each with 48GB of memory, running on Ubuntu 20.04.6 LTS.

We conduct experiments across *zero-shot* (ZS), *prompt learning* (PL), *single-modal fine-tuning* (SMFT), and *multi-modal fine-tuning* (MMFT), following the default configurations of each baseline for fair comparison. For ZS, we use the post-hoc methods MCM [31] and NegLabel [14], which leverage text information without additional training. These OoD scoring methods are also employed for PL and MMFT experiments.

As in previous works [14, 31], we use the prompts "a photo of a [label]" for MCM in the ZS setting and "The nice [label]" for NegLabel with temperature scaling set to 1 and 0.01, respectively. NegLabel [14] highlights that word choice in prompt engineering significantly impacts OoD detection performance: negative prompts degrade performance, positive prompts improve it, and neutral prompts require careful tuning. Among these, the "The nice [label]" template provides optimal results in NegLabel. However, in MMFT methods such as FLYP [8], template choice does not lead to significant accuracy variations. To explore this further, we conduct an ablation study using three different prompts, finding that the prompt choice does not significantly affect OoDD performance in MMFT. Detailed results are provided in Appendix B.3.

Unlike MCM, NegLabel requires additional hyperparameters due to its use of negative texts. Specifically, negative labels are mined from large word corpora like WordNet [6]. We follow the NegMining from Jiang et al. [14], which extracts $M = 10,000$ negative labels with a percentile $\eta = 0.05$ from WordNet. The extracted texts are then used for OoD scoring, as described in Eq. 7. Additionally, the results in Tables 1 and 2 do not incorporate the grouping strategy. Detailed results on the grouping strategy are provided in Appendix B.2.

For CoOp [52] and LoCoOp [34], we also adopt the settings proposed in the original papers. However, PL employs few-shot fine-tuning with a 16-shot setting, making direct comparisons with SMFT and MMFT potentially unfair. To address this, we not only follow the original settings but also evaluate with increasing shot counts, up to the full-shot setting where all ID data are utilized, as detailed in Appendix B.1. In the 16-shot setting, we report the averaged results from three repeated experiments with seeds 0, 1, and 2 using the original codebase [34, 52]. For other shot settings, we do not perform repeated experiments.

In SMFT, we use LP, FFT, and LP-FT [21], each of which exclusively utilizes the visual encoder without relying on textual information. Following Goyal et al. [8], we perform a hyperparameter sweep with learning rates {1e-4, 1e-5, 1e-6} and weight decay values {0.0, 0.1}. Models are trained for 10 epochs, selecting the best based on in-distribution (ID) accuracy, as OoD data is not directly available for validation in real-world OoDD settings. Through this procedure, the learning rates are set to 1e-4 for LP and 1e-5 for FFT and LP-FT, with a weight decay of 0.0.

For MMFT, we compare FLYP [8] and m^2-mix [35] with our proposed approach, using the same hyperparameter sweep as SMFT. Our method explores learning rates {1e-4, 1e-5, 1e-6}, weight decay values {0.0, 0.1}, and alignment strengths (i.e., λ) {1e-1, 1e-2, 1e-3}, with a batch size of 512. Early stopping is based on the accuracy of the ID validation set. In m^2-mix, the mixup weight is con-

trolled using the λ value instead of alignment strength.

Fig. 3 presents a snippet of code to illustrate our proposed method. In summary, the CLIP image and text encoders extract corresponding embeddings, which are projected into the same dimension and undergo L_2 normalization, projecting them onto a shared hyperspherical space. For CLIP and FLYP, contrastive learning optimizes cosine similarity by increasing it for matching image-text pairs while reducing it for non-matching pairs. Our method extends this approach by calculating `CMA_text` and `CMA_img` to derive the `total_loss`. When the alignment strength parameter λ is set to 0, the training process is equivalent to FLYP.

```
# extract image and text embeddings
img_emb, text_emb, scale = model(images, texts)

# joint hypersphirical embeddings
img_emb /= img_emb.norm(dim=-1, keepdim=True)
text_emb /= text_emb.norm(dim=-1, keepdim=True)

# scaled cosine similarity
logits = (scale) * (img_emb @ text_emb.T)

# clip symmetric loss function
gt = torch.arange(bs)
img_loss = Cross_Entropy_Loss(logits, gt, axis=0)
text_loss = Cross_Entropy_loss(logits, gt, axis=1)

# cross-modal-alignment regularization
CMA_img = -torch.logsumexp(logits.per_img,dim=1)
CMA_text = -torch.logsumexp(logits.per_text,dim=1)

# total CMA loss
total_loss = (img_loss + args.lam * CMA_img.mean())/2
           + (text_loss + args.lam * CMA_text.mean())/2
```

Figure 3. Pytorch-like pseudo-code of CMA

B. Additional Ablations

B.1. PL results with various-shot settings

In Tables 1 and 2 of the main paper, we present PL results based on the default 16-shot settings from CoOp [52] and LoCoOp [34]. To ensure a fair comparison, we extend these implementations to consider additional shot settings, including full-shot, and compare them with SMFT and MMFT, which employ full-shot configurations. Specifically, we conduct experiments using 256, 512, 1024, and full-shot settings, as shown in Tables 5, 6, and 7. All settings use a batch size of 512, consistent with the SMFT and MMFT configurations.

Our observations reveal that increasing the number of shots generally improves both OoDD performance and ID accuracy, suggesting that prompt learning benefits from more training data. However, full-shot settings do not always yield better results; in some benchmarks, performance at full-shot is even worse than ZS. Additionally, the observed improvements are not sufficient to outperform other baselines.

For CoOp$_{\text{NegLabel}}$, the highest ID accuracy of 74.07% is achieved in the 1024-shot setting, as shown in Table 7. In contrast, the best OoDD performance is observed in the 256-shot setting on the MOS benchmark and the 512-shot setting on the OpenOOD v1.5 benchmark. These results indicate that while increasing the number of shots from 16 to full-shot provides incremental gains, determining an optimal setting remains difficult. Nonetheless, our approach consistently outperforms CoOp and LoCoOp across all shot settings.

B.2. The effect of grouping strategy

The NegMining algorithm expands textual information by selecting words maximally distant from ID texts, thereby reducing the risk of high similarity between ID images and negative labels, as described in Algorithm 1. However, increasing the number of negative labels raises the variance in OoD scores, which can lead to more false positives. To address this, NegLabel [14] has proposed a grouping strategy that divides the negative labels into multiple groups to balance the benefits of additional information with the risk of false positives.

We report the performance of the grouping strategy proposed by NegLabel at $n = 100$ in Table 8. Applying the grouping strategy improves OoDD performance as shown in the table. To highlight the inherent capabilities of CMA, we do not apply additional performance-enhancing techniques, such as the grouping strategy, in our main experiments. Nevertheless, our method achieves state-of-the-art performance without the grouping strategy and shows further improvements when it is applied.

Algorithm 1 NegMining (proposed in NegLabel [14])

Input: Candidate labels $\mathcal{Y}^c$, ID labels $\mathcal{Y}$, Text encoder f^{text}
Output: Negative labels $\mathcal{Y}^-$

// Calculate text embeddings
for $y_i \in \mathcal{Y}$ **do**
  $e_i = f^{\text{text}}(\text{prompt}(y_i))$
end for
for $\tilde{y}_i \in \mathcal{Y}^c$ **do**
  $\tilde{e}_i = f^{\text{text}}(\text{prompt}(\tilde{y}_i))$
  // Measure candidate-ID label distance.
  $d_i = \text{percentile}_\eta(\{-\cos(\tilde{e}_i, e_k)\}_{k=1}^K)$
end for
// Choose M negative labels from top-k distances.
$\mathcal{Y}^- = \text{topk}([d_1, d_2, \ldots, d_C], \mathcal{Y}^c, M)$

B.3. The impact of prompts on OoDD performance

To evaluate the impact of prompts on performance, we conduct an ablation study using three prompts: "A photo of a [label]", "The nice [label]", and no prompt, as shown in Table 9. These prompts are derived from the ablation study of NegLabel [14]. Models are trained and evaluated with the same prompts. Our results indicate that altering the prompt does not lead to significant changes in performance. Specifically, in MCM, the performance difference across prompts does not exceed 1% in terms of average AUROC and FPR95. While positive prompts demonstrate slightly better OoDD performance, the differences are not significant enough to affect its performance superiority. These results show that the choice of prompt during MMFT has a negligible impact on OoDD performance.

B.4. The impact of λ values on OoDD performance

We select the λ value based on ID accuracy, as actual OoD data is not available for evaluation. To determine the optimal value of λ, we compare different λ values {1e-1, 1e-2, 1e-3, 5e-4}, and 1e-3 yields the highest ID accuracy, which is also aligned with the best OoDD performance, as shown in Table 10. Notably, an increase in alignment strength does not consistently improve ID accuracy or OoDD performance, highlighting the need for careful tuning. In our experiments, CMA demonstrates strong OoDD performance when the λ value is optimized for ID accuracy, even without access to OoD data.

Table 5. OoDD performance across different shot settings (16-, 256-, 512-, 1024-, and full-shot) for CoOp and LoCoOp on the MOS benchmark

	iNaturalist		SUN		Places		Textures		Average	
Methods	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
Zero-Shot										
MCM	32.28	94.40	39.33	92.28	44.94	89.83	57.98	85.99	43.63	90.63
NegLabel	**2.30**	**99.37**	**23.23**	**95.14**	39.85	90.98	46.49	89.64	27.97	93.78
16-shot										
CoOp$_{\text{MCM}}$	26.37	94.49	35.23	92.59	43.29	89.67	41.47	90.62	36.59	91.84
CoOp$_{\text{NegLabel}}$	4.95	98.90	25.76	94.59	30.07	93.33	44.35	89.59	26.28	94.10
LoCoOp$_{\text{MCM}}$	23.08	95.46	33.39	93.25	40.74	90.52	40.75	91.14	34.49	92.59
LoCoOp$_{\text{NegLabel}}$	3.19	99.25	46.63	90.58	55.44	87.65	46.03	89.85	37.83	91.83
256-shot										
CoOp$_{\text{MCM}}$	28.26	94.14	34.69	92.83	42.05	90.15	41.67	90.53	36.67	91.91
CoOp$_{\text{NegLabel}}$	4.28	99.00	29.34	94.41	**29.07**	94.23	34.25	93.01	**24.23**	**95.16**
LoCoOp$_{\text{MCM}}$	18.80	96.12	34.46	92.92	42.04	90.32	39.77	91.55	33.77	92.73
LoCoOp$_{\text{NegLabel}}$	4.37	99.09	49.39	90.53	64.01	85.82	51.45	88.95	42.31	91.10
512-shot										
CoOp$_{\text{MCM}}$	24.78	94.80	33.63	92.89	40.61	90.46	39.45	91.17	34.62	92.33
CoOp$_{\text{NegLabel}}$	3.59	99.14	34.54	93.30	30.80	93.79	**31.01**	**93.64**	24.98	94.97
LoCoOp$_{\text{MCM}}$	22.00	95.50	30.06	93.80	36.27	91.37	40.89	91.38	32.30	93.02
LoCoOp$_{\text{NegLabel}}$	4.85	98.93	40.15	92.29	58.99	86.76	60.78	85.40	41.19	90.84
1024-shot										
CoOp$_{\text{MCM}}$	22.83	95.15	33.60	92.80	40.96	90.46	39.40	91.33	34.20	92.44
CoOp$_{\text{NegLabel}}$	4.54	98.93	33.76	93.65	30.19	**94.26**	31.73	93.51	25.05	95.09
LoCoOp$_{\text{MCM}}$	22.10	95.27	32.58	93.58	38.50	91.16	39.52	91.52	33.18	92.88
LoCoOp$_{\text{NegLabel}}$	3.80	99.10	41.17	91.97	56.59	87.95	56.93	87.43	39.62	91.61
full-shot										
CoOp$_{\text{MCM}}$	23.88	94.98	35.74	92.49	41.72	90.13	38.93	91.14	29.10	92.19
CoOp$_{\text{NegLabel}}$	5.14	98.87	32.80	93.72	32.23	93.80	32.81	93.16	25.75	94.89
LoCoOp$_{\text{MCM}}$	20.25	96.01	32.72	93.25	38.82	90.87	39.96	91.39	32.94	92.88
LoCoOp$_{\text{NegLabel}}$	4.48	99.02	43.44	91.03	66.05	83.16	53.51	88.34	41.87	90.39

Table 6. OoDD performance across different shot settings (16-, 256-, 512-, 1024-, and full-shot) for CoOp and LoCoOp on the OpenOOD v1.5 benchmark

	SSB-hard		NINCO		Near-OoD		iNaturalist		Textures		Openimage-O		Far-OoD	
Methods	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
Zero-Shot														
MCM	89.45	64.11	82.70	69.82	86.08	66.97	61.94	87.62	54.26	87.71	53.80	88.60	56.67	87.98
NegLabel	81.87	71.32	69.82	77.09	75.85	74.21	**2.32**	**99.36**	44.98	90.56	31.10	93.10	26.14	94.34
16-shot														
CoOp$_{\text{MCM}}$	86.10	67.72	77.24	74.59	81.67	71.16	26.31	94.50	38.38	91.92	37.64	92.08	34.11	92.83
CoOp$_{\text{NegLabel}}$	68.20	78.72	57.71	84.20	62.96	81.46	4.96	98.90	42.74	90.28	23.18	95.25	23.63	94.81
LoCoOp$_{\text{MCM}}$	87.38	66.23	77.04	73.46	82.21	69.84	22.98	95.48	37.89	92.27	37.44	92.09	32.77	93.28
LoCoOp$_{\text{NegLabel}}$	76.03	74.42	66.51	80.96	71.27	77.69	3.70	99.16	43.87	90.59	29.29	93.65	25.62	94.47
256-shot														
CoOp$_{\text{MCM}}$	86.87	68.10	76.04	74.52	81.45	71.31	28.23	94.16	38.11	91.93	37.27	92.22	34.54	92.77
CoOp$_{\text{NegLabel}}$	**63.67**	81.66	52.80	86.29	58.23	83.98	4.29	98.99	32.51	93.65	20.76	95.96	19.19	96.20
LoCoOp$_{\text{MCM}}$	86.63	66.46	76.07	73.77	81.35	70.11	18.78	96.14	36.55	92.70	34.53	92.76	29.95	93.86
LoCoOp$_{\text{NegLabel}}$	68.88	81.54	70.86	79.89	69.87	80.72	4.40	99.08	50.31	89.83	29.87	93.97	28.19	94.29
512-shot														
CoOp$_{\text{MCM}}$	85.89	68.85	76.06	74.85	80.98	71.85	24.75	94.81	35.96	92.49	35.28	92.57	32.00	93.29
CoOp$_{\text{NegLabel}}$	68.48	79.54	**50.01**	**87.31**	59.25	83.43	3.60	99.14	**29.28**	**94.22**	**18.90**	96.34	**17.26**	**96.57**
LoCoOp$_{\text{MCM}}$	86.40	66.30	75.30	73.94	80.85	70.12	21.93	95.51	37.89	92.55	34.44	92.77	31.42	93.61
LoCoOp$_{\text{NegLabel}}$	67.23	81.48	70.40	78.99	68.82	80.23	4.86	98.92	59.40	86.31	30.01	93.91	31.42	93.05
1024-shot														
CoOp$_{\text{MCM}}$	86.04	68.76	75.94	75.58	80.99	72.17	22.78	95.17	36.11	92.56	34.69	92.71	31.19	93.48
CoOp$_{\text{NegLabel}}$	68.92	79.90	50.16	87.23	59.54	83.57	4.56	98.93	30.07	94.10	19.83	96.20	18.15	96.41
LoCoOp$_{\text{MCM}}$	86.37	66.53	75.78	74.48	81.07	70.51	21.94	95.28	36.55	92.55	35.59	92.46	31.36	93.43
LoCoOp$_{\text{NegLabel}}$	65.79	82.29	71.85	77.62	68.82	79.95	3.84	99.09	55.83	88.11	29.45	93.84	29.71	93.68
full-shot														
CoOp$_{\text{MCM}}$	86.13	68.92	75.96	75.17	81.04	72.04	23.84	94.99	35.45	92.41	34.70	92.77	31.33	93.39
CoOp$_{\text{NegLabel}}$	63.73	**82.39**	52.27	86.90	**58.00**	**84.64**	5.16	98.87	31.10	93.73	19.02	**96.42**	18.43	96.34
LoCoOp$_{\text{MCM}}$	85.89	67.35	74.84	75.03	80.36	71.19	20.17	96.02	37.09	92.52	33.62	92.95	30.29	93.83
LoCoOp$_{\text{NegLabel}}$	69.13	80.92	72.60	77.84	70.86	79.38	4.50	99.01	51.92	89.25	30.11	93.68	28.84	93.98

Table 7. ID accuracy across different shot settings (16-, 256-, 512-, 1024-, and full-shot) for CoOp and LoCoOp on ImageNet-1k

Methods	Acc.
16-shot	
CoOp	71.95
LoCoOp	71.72
256-shot	
CoOp	72.96
LoCoOp	72.76
512-shot	
CoOp	73.61
LoCoOp	73.12
1024-shot	
CoOp	**74.07**
LoCoOp	73.28
full-shot	
CoOp	73.97
LoCoOp	73.44

Table 8. The effect of the grouping strategy ($n = 100$) on the MOS benchmark. The symbol * represents the result with the grouping strategy.

Methods	iNaturalist		SUN		Places		Textures		Average	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
Zero-Shot (ZS)										
MCM	32.28	94.40	39.33	92.28	44.94	89.83	57.98	85.99	43.63	90.63
NegLabel	2.30	99.37	23.23	95.14	39.85	90.98	46.49	89.64	27.97	93.78
NegLabel*	1.55	99.58	17.96	95.82	33.53	91.97	44.34	89.86	24.35	94.31
Multi-modal Fine-tuning (MMFT)										
FLYP$_{\text{MCM}}$	24.86	94.35	39.81	90.58	47.92	87.16	41.19	89.34	38.44	90.36
FLYP$_{\text{NegLabel}}$	3.16	99.31	23.48	94.82	37.23	90.86	41.70	89.27	26.39	93.57
FLYP$_{\text{NegLabel}^\star}$	2.41	99.45	20.38	95.40	32.64	91.66	38.49	89.83	23.48	94.08
m^2-mix$_{\text{MCM}}$	22.41	95.61	39.18	91.85	47.07	88.72	43.44	90.13	38.02	91.58
m^2-mix$_{\text{NegLabel}}$	2.39	99.43	23.03	94.86	35.55	91.21	36.65	90.68	24.40	94.05
m^2-mix$_{\text{NegLabel}^\star}$	1.85	99.53	20.13	95.41	31.91	91.96	34.22	91.17	22.03	94.52
CMA$_{\text{MCM}}$ (Ours)	22.95	95.65	40.01	91.78	48.83	88.41	44.93	89.87	39.18	91.43
CMA$_{\text{NegLabel}}$ (Ours)	1.65	99.62	16.84	96.36	27.65	93.11	33.58	91.64	19.93	95.13
CMA$_{\text{NegLabel}^\star}$ (Ours)	**1.38**	**99.66**	**16.11**	**96.55**	**26.52**	**93.48**	**33.09**	**91.90**	**19.27**	**95.40**

Table 9. Comparison with different prompt settings (e.g., positive, neutral, and no prompts) for FLYP and CMA on the MOS benchmark

Methods	iNaturalist		SUN		Places		Textures		Average	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
"$< class >$"										
FLYP$_{\text{MCM}}$	25.42	94.06	38.93	90.84	47.29	87.30	40.60	89.45	38.06	90.41
FLYP$_{\text{NegLabel}}$	3.64	99.23	22.61	95.22	35.70	91.49	43.44	88.55	26.35	93.62
CMA$_{\text{MCM}}$ (Ours)	21.85	95.73	39.53	91.85	48.18	88.56	45.62	89.78	38.80	91.48
CMA$_{\text{NegLabel}}$ (Ours)	2.15	99.55	18.57	96.15	28.98	93.05	34.01	**91.86**	20.93	**95.15**
"a photo of a $< class >$"										
FLYP$_{\text{MCM}}$	26.13	94.04	39.04	90.64	47.63	87.02	41.12	89.90	38.48	90.40
FLYP$_{\text{NegLabel}}$	4.51	99.06	29.23	93.99	42.58	89.46	43.83	88.03	30.04	92.63
CMA$_{\text{MCM}}$ (Ours)	22.07	95.80	38.82	91.91	47.70	88.62	44.08	89.94	38.17	91.57
CMA$_{\text{NegLabel}}$ (Ours)	1.92	99.55	20.72	96.03	32.28	92.50	35.27	91.07	22.55	94.79
"The nice $< class >$"										
FLYP$_{\text{MCM}}$	24.86	94.35	39.81	90.58	47.92	87.16	41.19	89.34	38.44	90.36
FLYP$_{\text{NegLabel}}$	3.16	99.31	23.48	94.82	37.23	90.86	41.70	89.27	26.39	93.57
CMA$_{\text{MCM}}$ (Ours)	22.95	95.65	40.01	91.78	48.83	88.41	44.93	89.87	39.18	91.43
CMA$_{\text{NegLabel}}$ (Ours)	**1.65**	**99.62**	**16.84**	**96.36**	**27.65**	**93.11**	**33.58**	91.64	**19.93**	95.13

Table 10. Comparison of different λ values for CMA on the MOS benchmark

Methods	iNaturalist		SUN		Places		Textures		Average		Acc
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	
$\lambda = 0.1$											
CMA$_{\text{MCM}}$ (Ours)	26.20	94.96	53.54	86.56	57.73	83.45	49.73	88.00	46.80	88.24	81.12
CMA$_{\text{NegLabel}}$ (Ours)	5.61	98.91	32.57	93.75	40.97	90.44	47.30	89.68	31.61	93.19	
$\lambda = 0.01$											
CMA$_{\text{MCM}}$ (Ours)	22.70	95.86	43.95	90.64	52.63	87.33	45.80	90.09	41.27	90.98	81.96
CMA$_{\text{NegLabel}}$ (Ours)	2.89	99.41	20.24	95.81	31.11	<u>92.78</u>	39.15	91.30	23.35	94.83	
$\lambda = 0.001$											
CMA$_{\text{MCM}}$ (Ours)	22.95	95.65	40.01	91.78	48.83	88.41	44.93	89.87	39.18	91.43	**82.64**
CMA$_{\text{NegLabel}}$ (Ours)	**1.65**	**99.62**	**16.84**	**96.36**	**27.65**	**93.11**	<u>33.58</u>	<u>91.64</u>	**19.93**	**95.13**	
$\lambda = 0.0005$											
CMA$_{\text{MCM}}$ (Ours)	22.20	95.73	38.72	91.96	47.59	88.17	42.13	90.16	37.66	91.50	<u>82.56</u>
CMA$_{\text{NegLabel}}$ (Ours)	<u>1.85</u>	<u>99.61</u>	<u>17.70</u>	<u>96.08</u>	<u>29.50</u>	92.50	**31.54**	**92.12**	<u>20.15</u>	<u>95.08</u>	

B.5. Additional Near-OoD experiments

To thoroughly evaluate performance in Near-OoD scenarios, we conduct experiments on challenging ImageNet-1k splits from [36]. Specifically, we adopt the P_1, P_2, and P_3 protocols in [36]. Each of these includes known, negative, and unknown classes. For example, in P_1, the known classes consist of 116 fine-grained dog breeds from ImageNet. The unknown classes include 166 non-animal categories that are semantically distant from the known classes. Additionally, 67 four-legged animal classes are designated as negative classes, which are semantically closer to the known classes but remain distinct. The negative classes are originally intended to aid the model in distinguishing known classes from unknown classes during training.

For our Near-OoD experiments, we treat negative classes, along with unknown classes, as OoD since a simple zero-shot (ZS) method using NegLabel yields near-perfect performance on P_1 and P_2, making it difficult to evaluate the benefits of MMFT approaches. As shown in Table 12, ZS NegLabel achieves AUROC scores of 99.96% and 99.42% for P_1 and P_2, respectively. These results indicate that the unknown classes can be effectively distinguished using pre-trained textual information. Since the model already separates the unknown set too well, it becomes challenging to evaluate the contribution of textual information in MMFT. To address this, we construct more challenging splits P_1', P_2', and P_3' by designating additional negative datasets as OoD (i.e., negative classes + unknown classes), as described in Table 11. We perform a hyperparameter search based on FPR95 using the validation sets of known and negative classes. Note that negative classes are used only as validation/test datasets, and are not included in training.

As shown in Table 13, our method achieves the highest AUROC scores, maintaining robust OoDD performance even under challenging conditions, while also achieving the highest accuracy among all compared methods. However, we observe that although AUROC remains higher than that of $\text{FFT}_{\text{Energy}}$ (i.e., SMFT), which does not utilize textual information, the average FPR95 is comparable. To gain a deeper understanding, we analyze each protocol in sequence.

Starting with P_1', we observe that $\text{FFT}_{\text{Energy}}$ underperforms in both FPR95 and AUROC compared to methods that utilize textual information through NegLabel. This can be attributed to the fact that in P_1', the semantic distance between unknown/negative classes and known classes is sufficiently large, allowing textual information such as negative concept labels to effectively distinguish them. This observation aligns with prior findings on the effectiveness of textual information in Far-OoD scenario. Next, in P_2', we observe that all NegLabel-based methods, except for our $\text{CMA}_{\text{NegLabel}}$, underperform compared to $\text{FFT}_{\text{Energy}}$. This indicates that CMA effectively reduces the modality gap, thereby improving the utilization of textual information. Similarly, in P_3', while our method performs worse than SMFT in terms of FPR95, it achieves a higher AUROC score.

These findings indicate that $\text{FFT}_{\text{Energy}}$, which relies solely on visual features, can effectively distinguish between subclasses within a broader category (e.g., various types of "Hunting Dog" in P_2') solely based on visual cues. In contrast, existing NegLabel-based approaches struggle to separate ID and OoD classes when they belong to the same or semantically related categories, likely due to the modality gap. Our method addresses this challenge by mitigating the modality gap, thereby improving detection performance in Near-OoD scenarios.

	ID (Known)	**OOD (Negative+Unknown)**
P_1'	All dog classes 29055 / 5800	Other 4-legged animal classes Non-animal classes 17420 / 11650 (3350+8300)
P_2'	Half of hunting dog classes 7224 / 1500	Half of hunting dog classes Other 4-legged animal classes 7949 / 4300 (1550+2750)
P_3'	Mix of common classes 38633 / 7550	Mix of common classes 24549 / 13050 (4850+8200)

Table 11. More challenging ImageNet-1k splits. The numbers represent the number of validation/test samples.

C. Additional Visualization

To achieve more intuitive visualization, we use PCA with 1,000 image prototypes and class prototypes from ImageNet-1k, along with 1,000 random OoD images (from the MOS benchmark datasets) and negative texts. As shown in Fig. 4, methods such as ZS and FLYP exhibit a clear modality gap between ID image and ID text embeddings (orange and blue points). This gap is also observed in OoD image and text embeddings, as illustrated in Figs. 5 and 6.

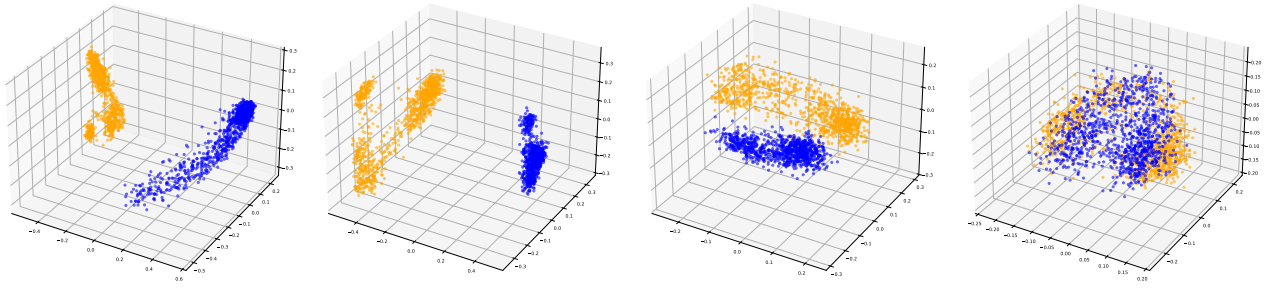
Our findings indicate that eliminating this modality gap among ID embeddings is essential for fully leveraging textual information, such as negative concept texts, as discussed in Section 5. In CMA, which addresses this modality gap, the orange and blue points are clustered closer together, as are the red and green points.

Table 12. ZS OoDD Performance on splits P_1, P_2, and P_3 (ID = Known, OoD = Unknown)

Methods	P_1		P_2		P_3		Average	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
Zero-Shot (ZS)								
MCM	14.27	96.96	56.07	88.89	35.11	89.64	35.15	91.83
NegLabel	**0.23**	**99.96**	**3.35**	**99.42**	**29.41**	**90.10**	**11.00**	**96.49**

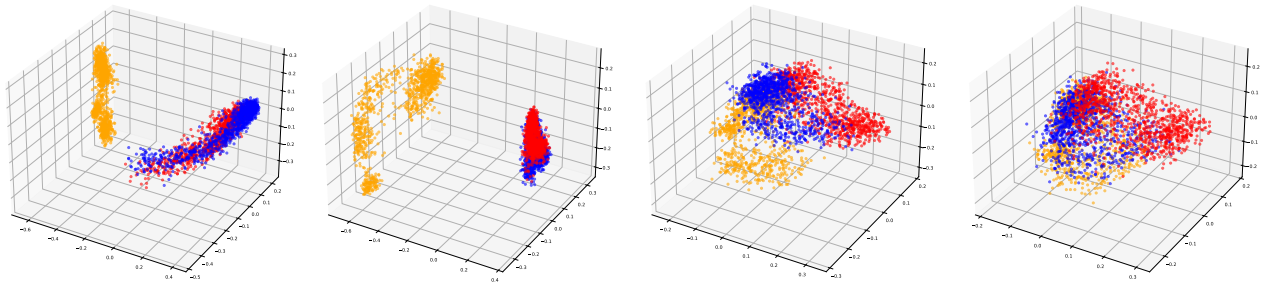
Table 13. Comparison of OoDD performance on our splits P_1', P_2', and P_3' (ID = Known, OoD = Negative + Unknown)

Methods	P_1'		P_2'		P_3'		Average		Acc
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	
Zero-Shot (ZS)									
MCM	29.18	93.07	64.30	85.34	52.41	83.32	48.63	87.24	73.43
NegLabel	1.17	99.74	24.09	93.34	48.50	82.90	24.59	91.99	
Prompt-Learning									
CoOp$_{\text{MCM}}$	24.83	93.91	63.03	84.75	52.78	83.64	46.88	87.43	75.10
CoOp$_{\text{NegLabel}}$	1.20	99.73	26.75	91.65	48.82	83.72	25.59	91.70	
LoCoOp$_{\text{MCM}}$	22.47	94.87	62.95	84.70	52.25	83.20	45.89	87.59	74.95
LoCoOp$_{\text{NegLabel}}$	**0.90**	**99.78**	25.25	91.85	60.13	79.47	28.76	90.37	
Single-modal Fine-tuning (SMFT)									
FFT$_{\text{MSP}}$	19.18	95.45	61.44	87.04	56.46	93.78	45.69	89.98	
FFT$_{\text{ODIN}}$	3.24	99.30	25.88	93.78	**38.00**	89.55	22.37	94.21	84.74
FFT$_{\text{Energy}}$	2.60	99.41	18.19	94.52	39.92	88.40	**20.24**	94.11	
Multi-modal Fine-tuning (MMFT)									
FLYP$_{\text{MCM}}$	9.14	98.16	42.67	89.94	42.05	87.11	31.29	91.74	85.30
FLYP$_{\text{NegLabel}}$	2.33	99.42	19.35	94.45	41.76	86.79	21.15	93.56	
CMA$_{\text{MCM}}$	9.21	98.16	43.02	89.95	41.26	89.43	31.16	92.51	**85.55**
CMA$_{\text{NegLabel}}$	2.29	99.42	**18.07**	**94.76**	40.97	**89.71**	20.44	**94.63**	



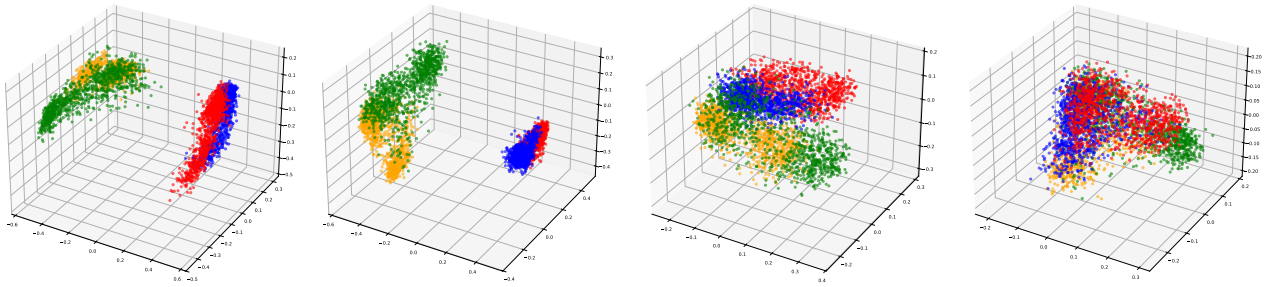
(a) ZS (b) FLYP (c) m^2-mix (d) CMA

Figure 4. Visualization of image and text embeddings using PCA on ImageNet-1k. Orange and blue points represent ID image and ID text embeddings, respectively.



(a) Zero-Shot (b) FLYP (c) m^2-mix (d) CMA

Figure 5. Visualization of image and text embeddings using PCA on ImageNet-1k and negative texts. Orange and blue points represent ID image and ID text embeddings, respectively, while red points denote negative text embeddings.



(a) Zero-Shot (b) FLYP (c) m^2-mix (d) CMA

Figure 6. Visualization of image and text embeddings using PCA on ImageNet-1k, MOS benchmark datasets, and negative texts. Orange and blue points represent ID image and ID text embeddings, respectively, while green and red points denote OoD image and negative text embeddings.