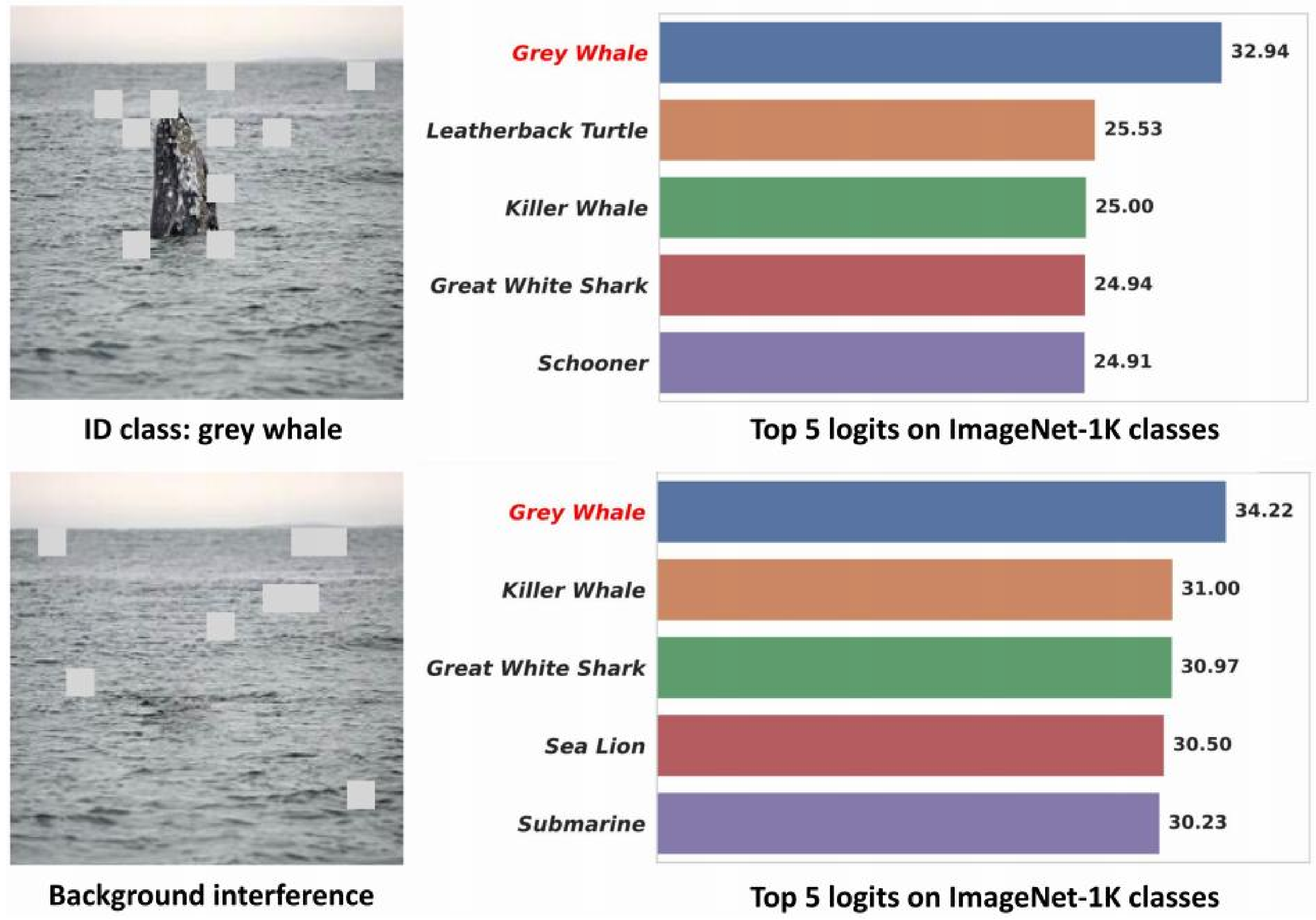
Overcoming Shortcut Problem in VLM for Robust Out-of-Distribution Detection

Zhuo Xu[1], Xiang Xiang[1,2,*] ,Yifan Liang[1]

[1] National Key Lab of Multi-Spectral Information Intelligent Processing Technology
School of Artificial Intelligence and Automation,Huazhong University of Science and Technology, China
[2] Peng Cheng National Laboratory, Shenzhen, China

CVPR Nashville JUNE 11-15, 2025

Motivation



ID class: grey whale

Top 5 logits on ImageNet-1K classes

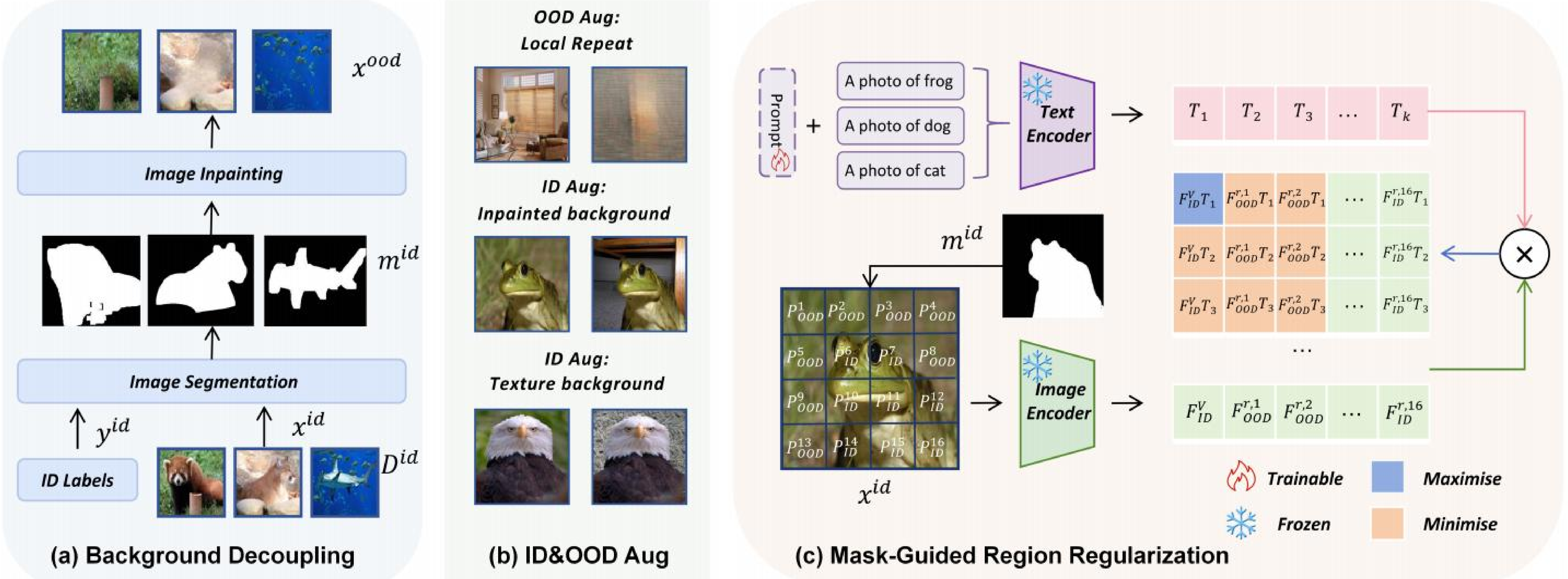
Background interference

Top 5 logits on ImageNet-1K classes

➢ VLMs (e.g CLIP) suffers from serious shortcut problem, may output rediculous higher logits on background interference.

➢ Existing methods neglect this critical problem.

Methods


x^{ood}
Image Inpainting
m^{id}
Image Segmentation
y^{id}
x^{id}
ID Labels
D^{id}
OOD Aug:
Local Repeat
ID Aug:
Inpainted background
ID Aug:
Texture background
Prompt
+
A photo of frog
A photo of dog
A photo of cat
Text Encoder
T_1 T_2 T_3 … T_k
$F^{V}_{ID}T_1$ $F^{r,1}_{OOD}T_1$ $F^{r,2}_{OOD}T_1$ … $F^{r,16}_{ID}T_1$
$F^{V}_{ID}T_2$ $F^{r,1}_{OOD}T_2$ $F^{r,2}_{OOD}T_2$ … $F^{r,16}_{ID}T_2$
$F^{V}_{ID}T_3$ $F^{r,1}_{OOD}T_3$ $F^{r,2}_{OOD}T_3$ … $F^{r,16}_{ID}T_3$
…
×
m^{id}
P^1_{OOD} P^2_{OOD} P^3_{OOD} P^4_{OOD}
P^5_{OOD} P^6_{ID} P^7_{ID} P^8_{OOD}
P^9_{OOD} P^{10}_{ID} P^{11}_{ID} P^{12}_{ID}
P^{13}_{OOD} P^{14}_{ID} P^{15}_{ID} P^{16}_{ID}
x^{id}
Image Encoder
F^{V}_{ID} $F^{r,1}_{OOD}$ $F^{r,2}_{OOD}$ … $F^{r,16}_{ID}$
Trainable
Frozen
Maximise
Minimise

(a) Background Decoupling

(b) ID&OOD Aug

(c) Mask-Guided Region Regularization

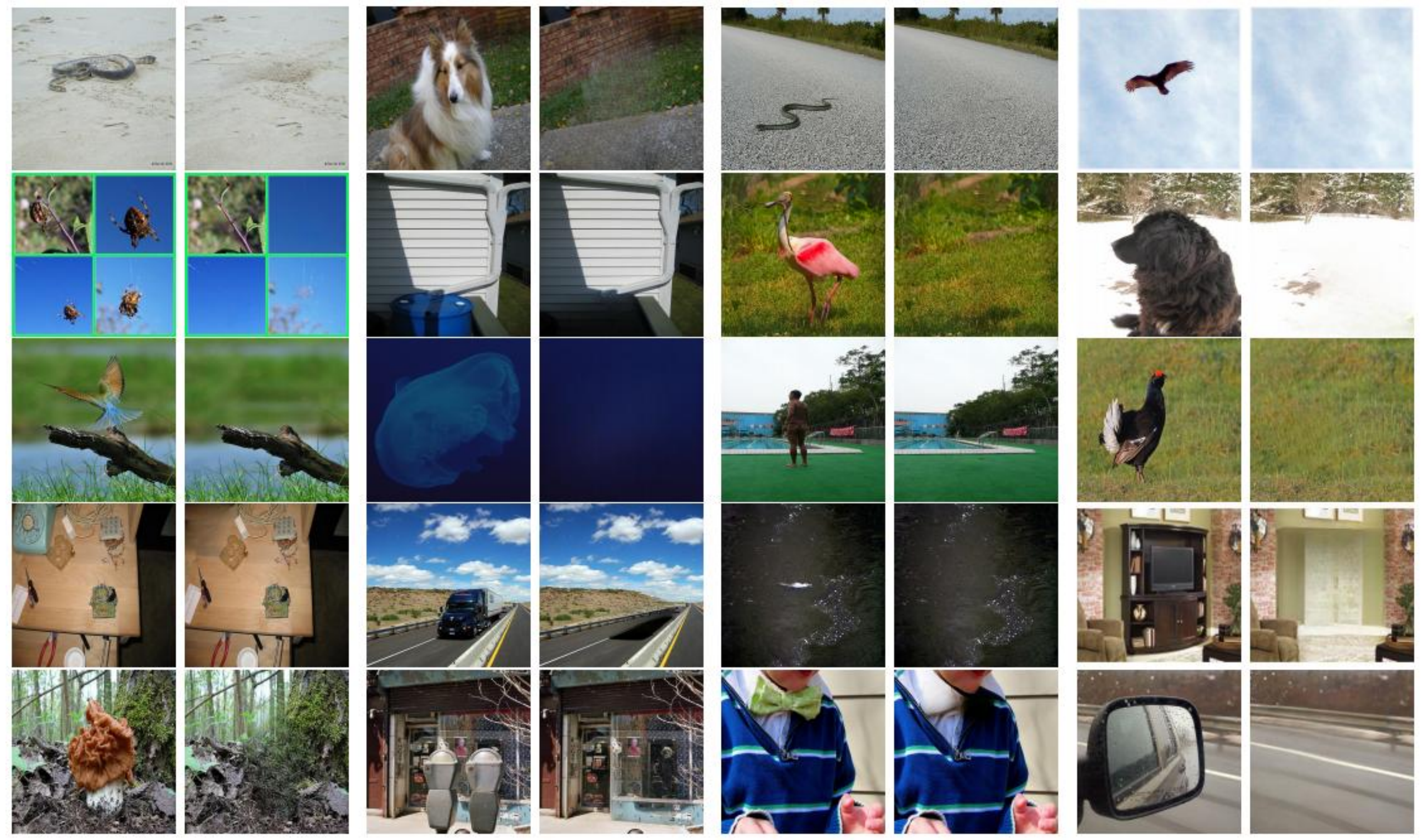
➢ Decoupling images into foreground and background, removing the foreground with ID information to generate background-only images. Inpainting the removed ID region with background information to generate natural background samples.

➢ Repeat the local regions and replace the background of ID samples with diverse background for data augmentation.

➢ Mask-guided region regularization to constrain ID-irrelevant areas and reduce the model's response to these regions.

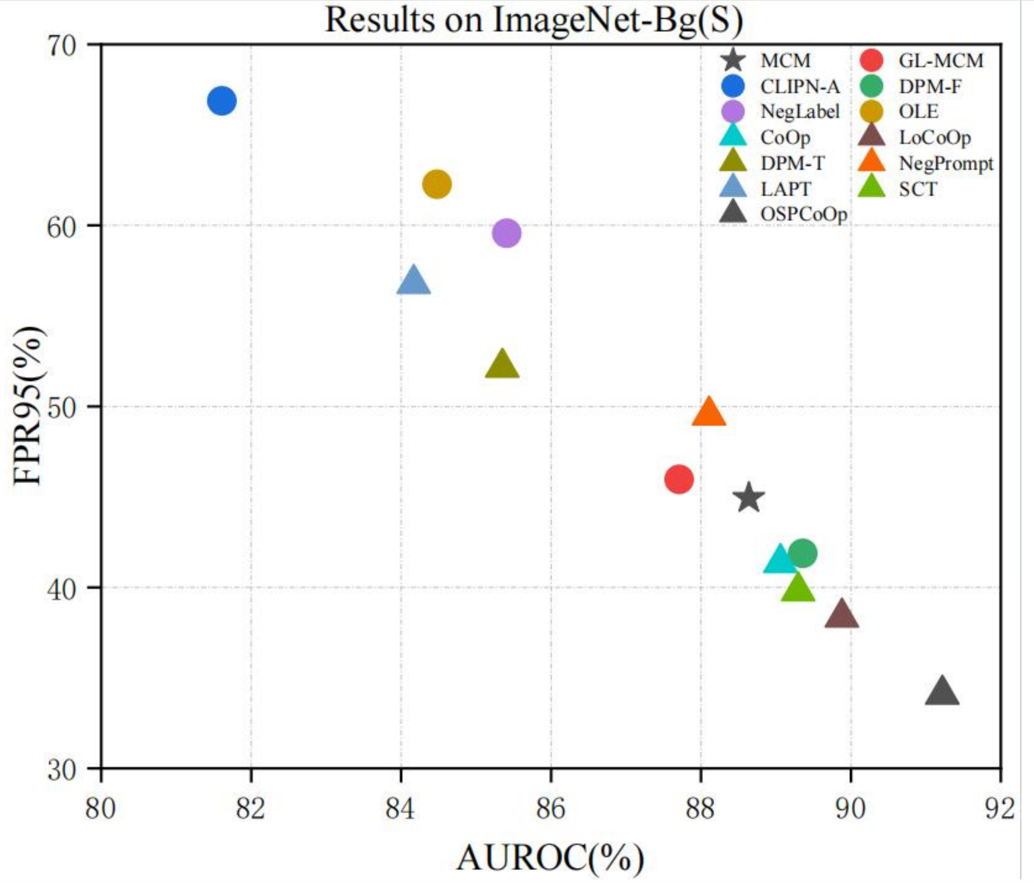
Background Datasets

- Removing corresponding ID-relevant regions from samples in the ImageNet validation set to evaluate the robustness of the model against background inteference.
- ImageNet-Bg, with 48,285 background images.
- ImageNet-Bg(S), sampling from ImageNet-Bg, with 24,863 cleaner background images.



Quantitative Results

Method	iNaturalist		SUN		Places		Texture		Avg	
	AUROC↑	FPR95 ↓	AUROC↑	FPR95 ↓	AUROC↑	FPR95 ↓	AUROC↑	FPR95 ↓	AUROC↑	FPR95 ↓
Training-free methods										
ZOC [10]	86.09	87.30	81.20	81.51	83.39	73.06	76.46	98.90	81.79	85.19
MCM [33]	94.61	30.94	92.56	37.67	89.76	44.76	86.10	57.91	90.76	42.82
GL-MCM [39]	96.71	15.18	93.09	30.42	89.90	38.85	83.63	57.93	90.83	35.47
CLIPN-A [49]	95.27	23.94	93.93	26.17	90.93	40.83	92.28	33.45	93.10	31.10
DPM-F [57] †	96.84	15.26	91.78	42.58	89.60	45.99	85.74	57.55	90.99	40.35
Outlier-label exposure methods										
NegLabel [21]	99.48	1.99	95.43	21.05	91.95	34.95	90.90	44.79	94.25	25.69
LAPT [55]	99.63	1.16	96.01	19.12	92.01	33.01	91.06	40.32	94.68	23.40
EOE [4]	97.52	12.29	95.73	20.40	92.94	30.16	85.64	57.53	92.96	30.09
OLE [8]	98.33	7.61	94.87	22.44	92.45	31.73	92.40	34.70	94.51	24.12
Requires few-shot training (or w/ fine-tuning)										
CoOp [60]	96.62	14.60	92.65	28.48	89.98	36.49	88.03	43.13	91.82	30.67
LoCoOp [35]	96.86	16.05	95.07	23.44	91.98	32.87	90.19	42.28	93.52	28.66
SCT [51]	95.86	13.94	95.33	20.55	92.24	29.86	89.06	41.51	93.37	26.47
DPM-T [57] †	97.04	14.47	93.19	33.06	89.78	39.46	87.49	49.73	91.88	34.18
ID-like [2]	**98.19**	**8.98**	91.64	42.03	91.15	41.74	**94.38**	**26.77**	93.84	29.88
NegPrompt [27] †	90.69	45.97	92.18	39.43	91.65	37.49	90.01	44.84	91.13	41.93
OSPCoOp (Ours)	97.13	15.25	**96.74**	**18.26**	**94.01**	**25.74**	91.13	41.26	**94.75**	**25.13**

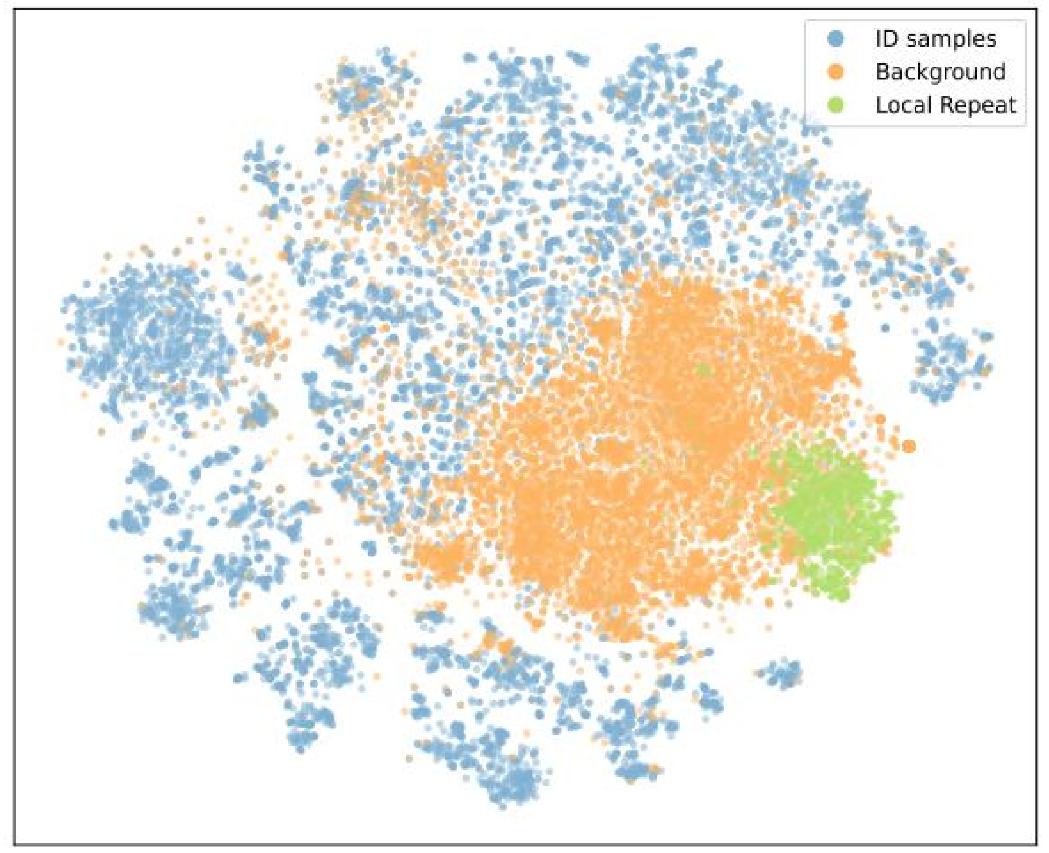


Analysis

➢ The pseudo-OOD data partly show a clustering trend, with a small portion distributed around the ID samples, which are easy to take shortcuts.

➢ Constrain both regional and global logits for backgrounds can mitigate shortcuts.

λ^{r}_{ood}	λ^{g}_{ood}	AUROC↑	FPR95↓
×	×	92.05	32.78
✓	×	93.94	28.45
×	✓	94.14	25.35
✓	✓	**94.75**	**25.13**



Conclusions

➢ We observe that VLMs (e.g CLIP) often relies on background information, which can lead to failures in OOD detection,especially for background interference.

➢ we present ImageNet-Bg, a novel OOD evaluation benchmark designed to facilitate a comprehensive assessment of model robustness against background interference.

➢ Decoupling background information to generate pseudo-OOD supervision, constraining the model's responses to ID-irrelevant regions, can effectively mitigate the shortcut problem.

*** Correspondence to xex@hust.edu.cn**