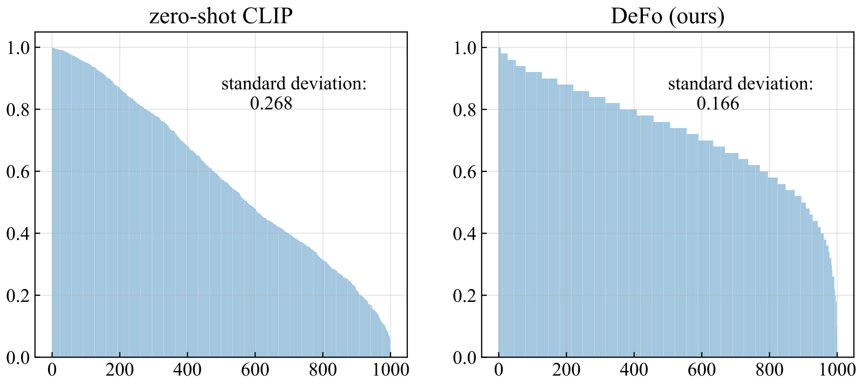


Figure 4: Class-wise accuracy distribution on ImageNet.

Table 6: Image retrieval with off-the-shelf features (mAP, %)

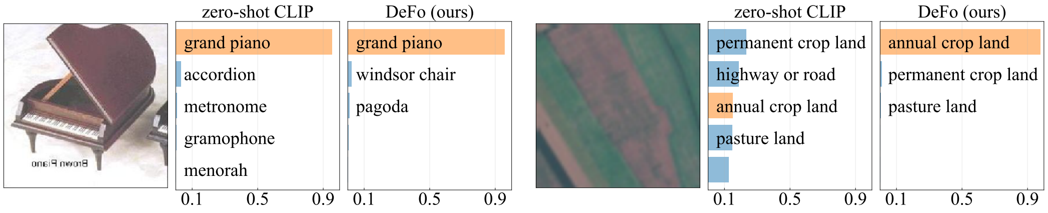
Method	$\mathcal{R}$Ox (M)	$\mathcal{R}$Ox (H)	$\mathcal{R}$Par (M)	$\mathcal{R}$Par (H)
CLIP	36.1	11.8	56.4	27.9
DeFo (ours)	36.3	11.8	56.6	28.0

APPENDIX

Quantitative analysis of the expressive sensitivity challenge. Here we give further evidence of the expressive sensitivity challenge for CLIP and how DeFo addresses it. As shown in Figure 4, CLIP yields unstable performance over the 1000 classes in ImageNet, with 0.268 standard deviation of class-wise accuracy. In contrast, the deviation is decreased to 0.166 when DeFo is utilized. While the predicted probabilities of CLIP are directly decided by the textual targets (i.e., the class names, see Equation 1), this high-deviation result indicates that the weak predictive performance of the zero-shot CLIP is caused by the intrinsic limitations of textual targets, rather than the limitation of the image encoder's capacity in visual representation.

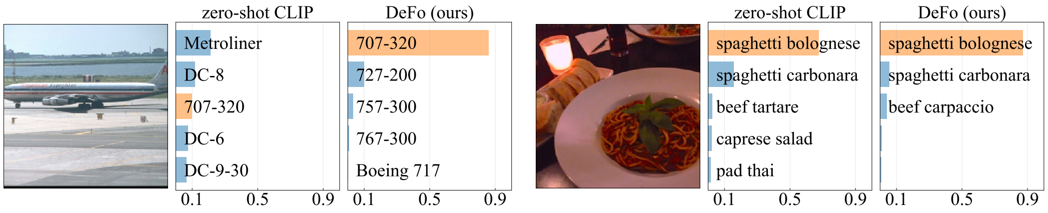
Examples of CLIP's and DeFo's predictions. We also give some examples of the predicted probabilities of DeFo and compare them with the zero-shot CLIP. As shown in Figure 5, the examples are randomly selected from the test splits of the eight datasets and we report their top-5 predictions. We observe that aside from CLIP's higher mis-classification rate over the eight datasets, its failed cases are often rare or very abstract objects. For example, CLIP fails to recognize the satellite image (Figure 5b) and the type of aircraft (Figure 5c). Also, despite CLIP correctly classifies the type of cars (Figure 5g), the gap between the predicted probabilities of the top two classes is unclear. These results show the challenge of conceptual sensitivity to CLIP-like inference protocols, while our DeFo in contrast performs robustly in recognizing a variety of visual objects.

Image retrieval performance. To further demonstrate DeFo's potential in domain generalization, we explore the performance in retrieval of off-the-shelf features in the "revisited Oxford and Paris" dataset. Following DINO (Caron et al., 2021), we report the mAP scores for the Medium (M) and Hard (H) splits as below (with ResNet-50 image encoder). We find DeFo to slightly outperform the original CLIP model. Despite better performance, DeFo hence also generalizes equally well.



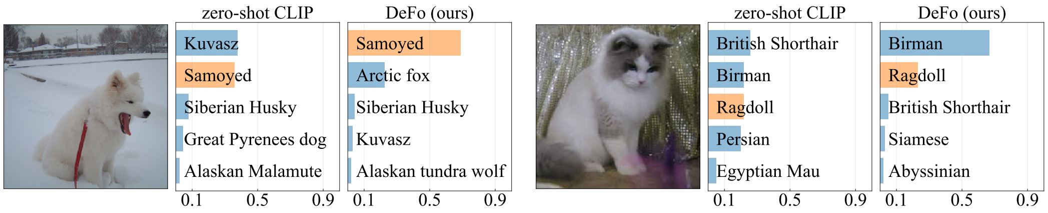
(a) Caltech-101

(b) EuroSAT



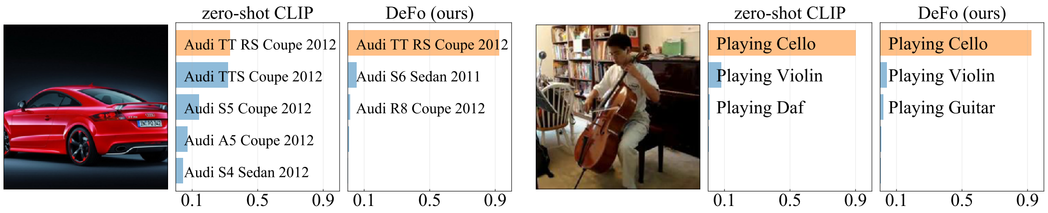
(c) FGVCAircraft

(d) Food101



(e) ImageNet

(f) OxfordPets



(g) StanfordCars

(h) UCF101

Figure 5: Examples of predicted probabilities (top-5) of CLIP and our DeFo. The ground truth of each image is in orange.