

Supplementary Material

BLIP t-SNE

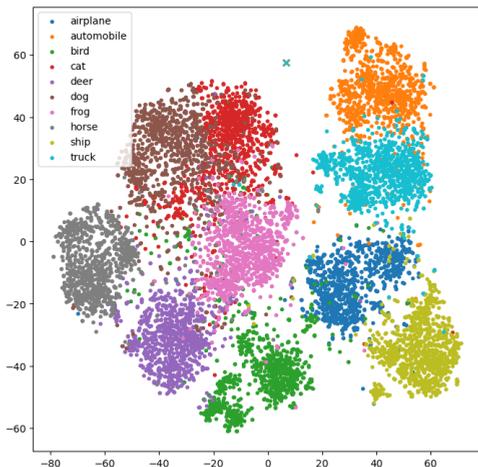


Figure 1: Plot of t-SNE projections of BLIP embeddings of CIFAR-10 labels (crosses) and images (dots), coloured according to class. We see a similar pattern to CLIP; the class label embeddings are highly clustered together far away from all of the image embeddings, which are more dispersed in the latent space.

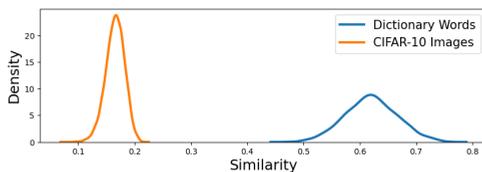


Figure 2: Density of similarities between dictionary text embeddings and class label text embeddings (blue) versus the similarities between class label text embeddings and image embeddings from the multi-modal BLIP model. We see a similar pattern to CLIP; text embeddings of different types are vastly more similar to each other than they are to image embeddings despite contrastive training.

Text Prompt Formulation

	# Normal Classes		
Text Prompt	1	6	9
"This is a photo of a {}"	99.14	97.27	96.02
A photo of a {}"	99.12	97.19	96.00
"{}"	99.07	96.93	96.62

Table 1: Performance of BLISS when using different formulations of text prompts, where {} is replaced by a word. We see performance is robust across all prompt formulations.

Few-shot Performance

<i>N</i> -shot	1 class	6 classes	9 classes
1	98.91±0.68	93.42±3.19	91.49±4.97
5	98.93±0.60	94.95±1.79	93.91±3.02
10	99.06±0.51	96.14±1.38	95.23±2.28
50	99.14±0.44	97.09±0.86	96.01±1.87
100	99.14±0.43	97.19±0.81	95.91±2.11
1000	99.14±0.43	97.27±0.76	96.02±1.93

Table 2: Mean and standard deviation in AUROC score of BLISS under different *N*-shot settings, where *N* is the number of samples per normal class, averaged over five runs. The decline in performance with fewer training samples is surprisingly small. In the 1 normal class setting, using just 1 training samples results in a 0.2% lower score than using all samples. The drop is more noticeable for more normal classes; 4% and 4.5% for 6 and 9 normal classes respectively. Nevertheless, BLISS with 1 example still outperforms ZOC as well as almost all full-shot baselines. Close to optimal performance is reached with at least 50 samples. The training samples are only used to compute the mean and standard deviation statistics for score calibration, so the number of training samples is relatively less important for BLISS than it is for other methods which need training samples for model optimisation or nearest neighbours.

Dictionary Sources

<i>C</i>	CIFAR-10	CIFAR-10	CIFAR-10	CIFAR-100	TinyImageNet
	1 class	6 classes	9 classes	20 classes	20 classes
Places365	98.5	95.5	93.6	85.6	84.5
OpenImages	99.1	96.8	96.0	89.4	90.5

Table 3: BLISS with Places365 and OpenImages dictionaries for each dataset (*C10(1)* indicates CIFAR-10 with 1 normal class). Places365 contains 365 labels describing room types while OpenImages contains more than 20,000 words of a broad range of objects. Performance drops with the Places365 dictionary, which can be explained by not only the much more narrow scope of the words (all describing room types), but also by its smaller size (365 words). Even so, it still outperforms most baselines. Performance with the much larger and broader OpenImages dictionary is very similar to that of the ImageNet dictionary.

Class splits

Normal/Anomaly Class	Biased-CLIP	BLISS
Horse/Deer	91.30	96.34
Deer/Horse	97.24	97.77
Automobile/Truck	90.33	97.48
Truck/Automobile	86.60	90.74

Table 4: For challenging cases, we devised an experimental setup with two classes (one normal and one anomaly class) in CIFAR-10 which are semantically similar to each other. We chose the horse/deer classes and automobile/truck classes. Table 4 show the performance of Naive-CLIP versus BLISS in this setup. We see that BLISS clearly outperforms Naive-CLIP in all of these challenging cases.

TopK

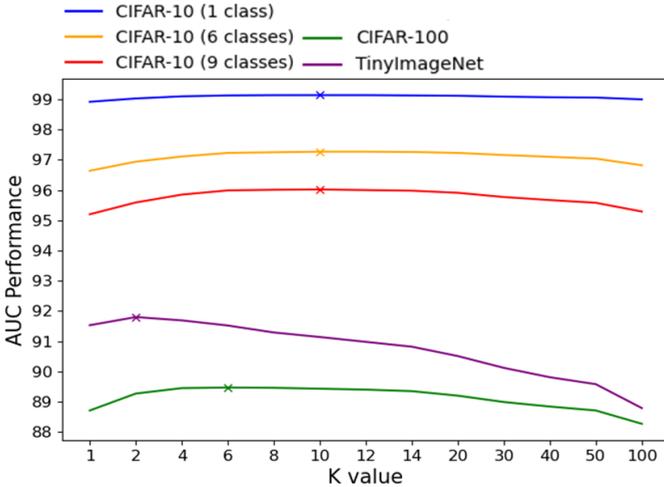


Figure 3: Mean AUROC Performance of BLISS with different settings of K in the topK words in the External Text Score. The maximum score is marked by a cross. We see that performance is robust to a wide range of K values overall.

FPR95

	Internal class score only		BLISS
w/o (5) with (5)			
1 class	9.27 \pm 4.0	9.27 \pm 4.0	3.87 \pm 2.2
6 classes	34.88 \pm 9.7	30.82 \pm 8.2	11.89 \pm 3.4
9 classes	39.25 \pm 15.2	37.82 \pm 14.9	15.20 \pm 9.0

Table 5: Mean and standard deviation in FPR95 scores. Lower scores are better.