	Class Name	GPT Desc	Llama Desc	Class + Llama	Llama-2	Class + Llama-2
CIFAR-10	87.1	87.8	88.4	89.1	88.8	89.5
CIFAR-100	59.0	63.0	59.4	63.2	61.2	63.8
Mean	73.0	75.4	73.9	76.2	75.0	76.6

Table 1: Accuracy (%) for a ViT-B/32 model with different text generation strategies for training the text-only classifier. *Class + Llama* refers to prepending class names to descriptions from Llama.

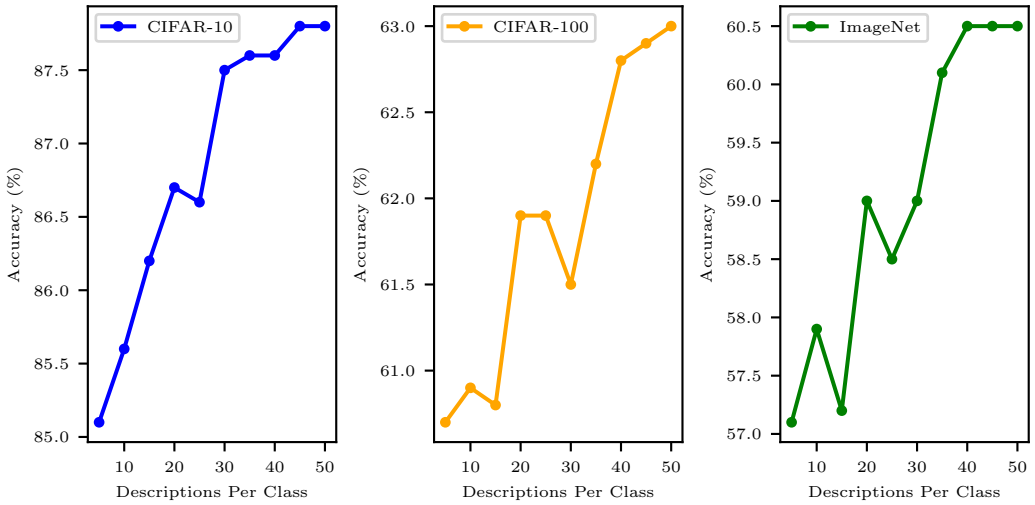
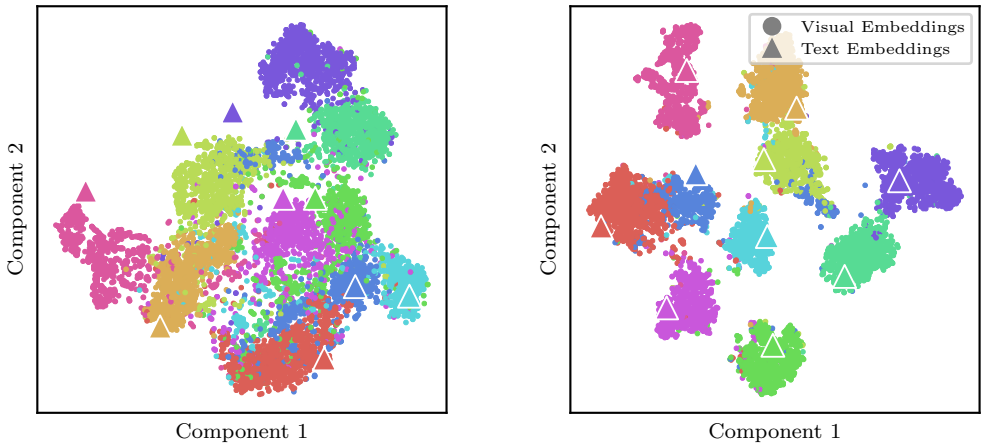


Figure 1: Effect of diversity of descriptions on the Top-1 Accuracy (%). We train the text classifier by randomly choosing (with an increment of 5) a certain number of descriptions per class for each evaluation step. In our paper, we use a maximum of 50 descriptions per class.



(a) Base CLIP (ViT-B/32). (b) LaFTer Adapted CLIP (ViT-B/32).

Figure 2: TSNE projections for visual *(circles)* and text *(triangles)* embeddings for 10 classes of the EuroSAT dataset from (a) Base *(un-adapted)* CLIP ViT-B/32 model and (b) after adaptation with LaFTer. For the Base CLIP model, the text embeddings are the output feature vector from the CLIP text encoder and for LaFTer we use the weights of the adapted text-classifier. Visual embeddings are always the output features from the vision encoder.