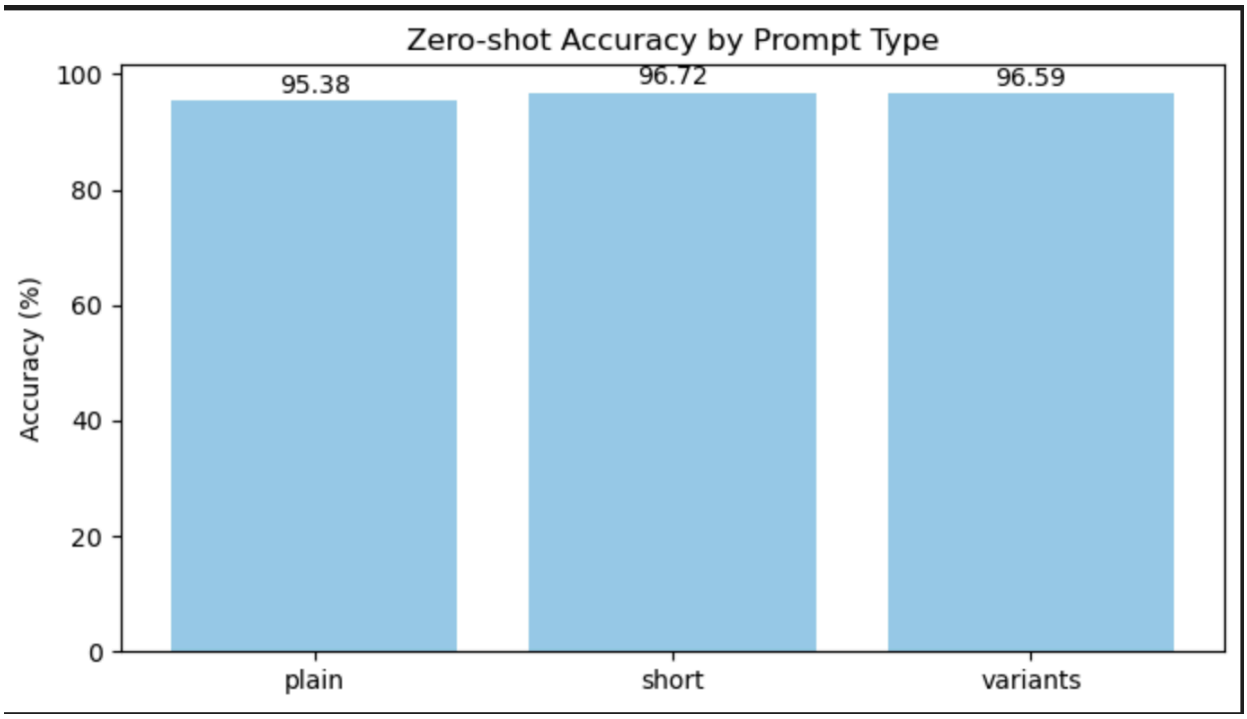
Task5:

1. Zero-Shot Classification on STL-10

Zero-shot accuracy using different prompting techniques was calculated. The results are shown in Fig (5.1)



(Fig 5.1)

Accuracies follow the trend: Short > Variants > Plain. Since CLIP was trained on large-scale web-scraped captions, which often appear in short or variant style, performance is better for these than plain raw labels.

2. Exploring the Modality Gap

CLIP is a multimodal model with separate encoders for text (Transformer) and images (Vision encoder). Due to differences in representation spaces and inherent noise in web data (e.g., incomplete or misleading captions), a modality gap arises. Ideally, embeddings from text and images should live in the same semantic space

so that “a photo of a dog” and a dog image are close together. In practice, they occupy different regions. This problem is addressed using dimensionality reduction (t-SNE, UMAP) and alignment techniques (Procrustes with and without L2 normalization).

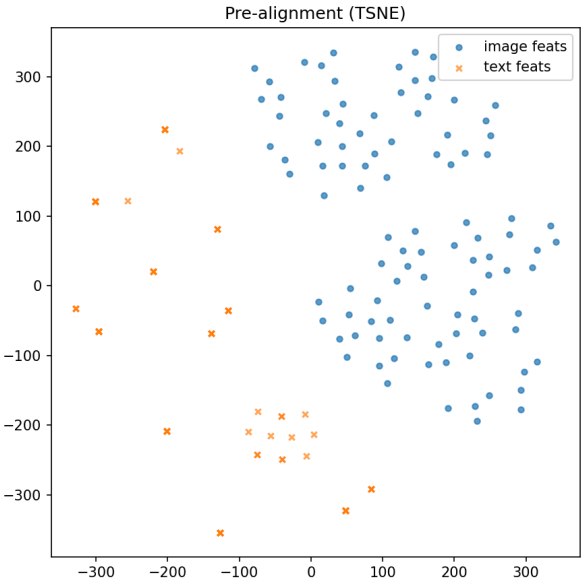


Fig (5.2)

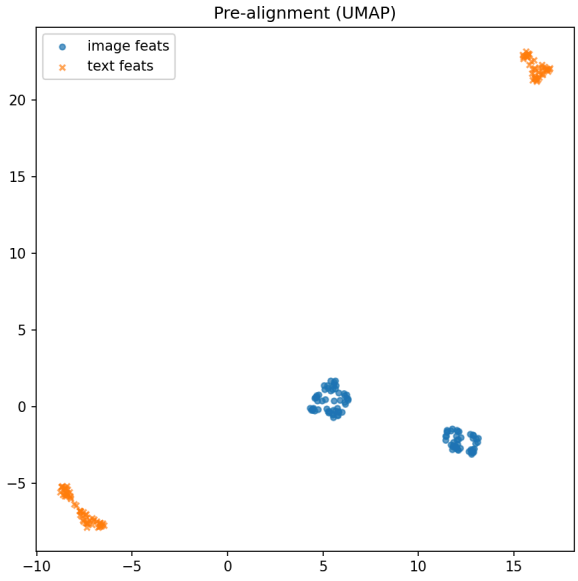


Fig (5.3)

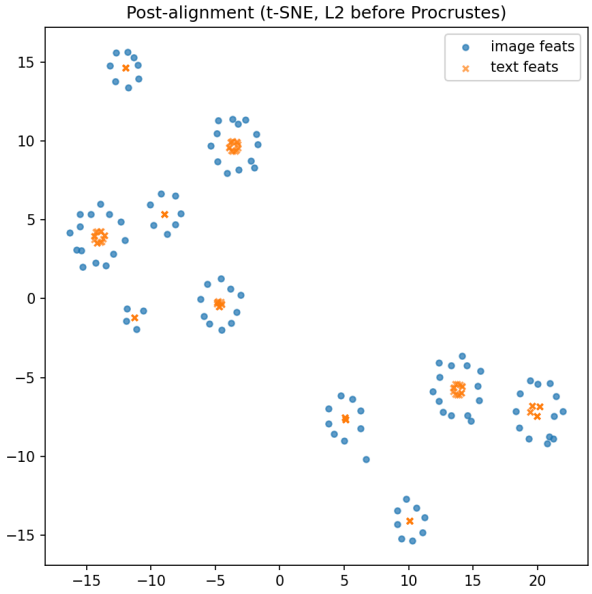


Fig (5.4)

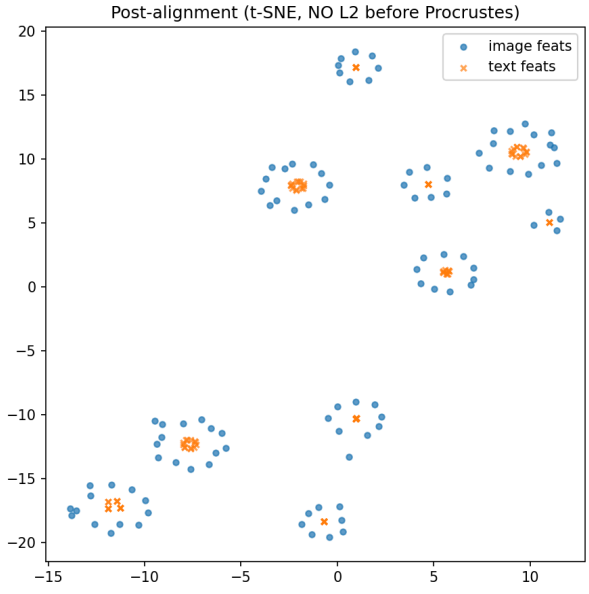


Fig (5.5)

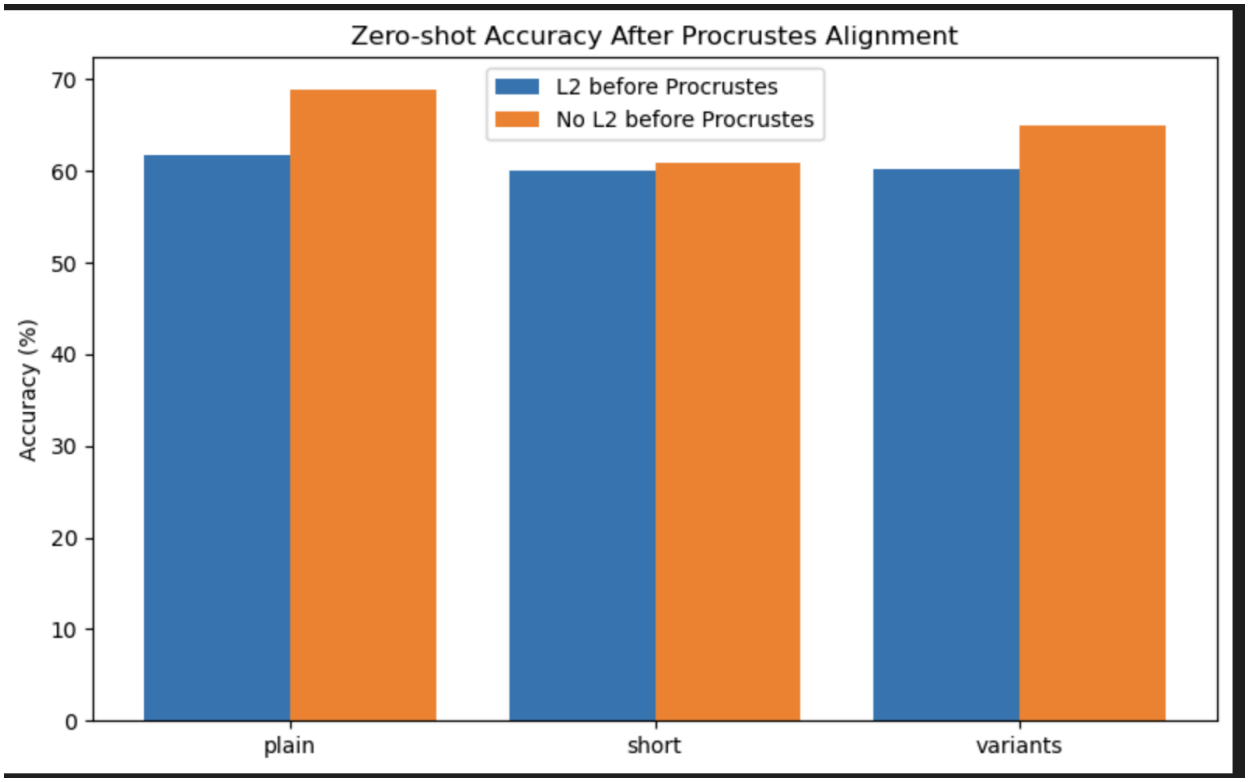
Findings

Pre-alignment: Images and texts are clearly separable forming distinct clusters in Pre-TSNE (Fig 5.2) and Pre-Umap alignment (Fig 5.3)
Post-alignment (Procrustes, with and without L2 normalization): Clusters overlapped significantly, indicating reduced modality gap.

Cosine Similarity

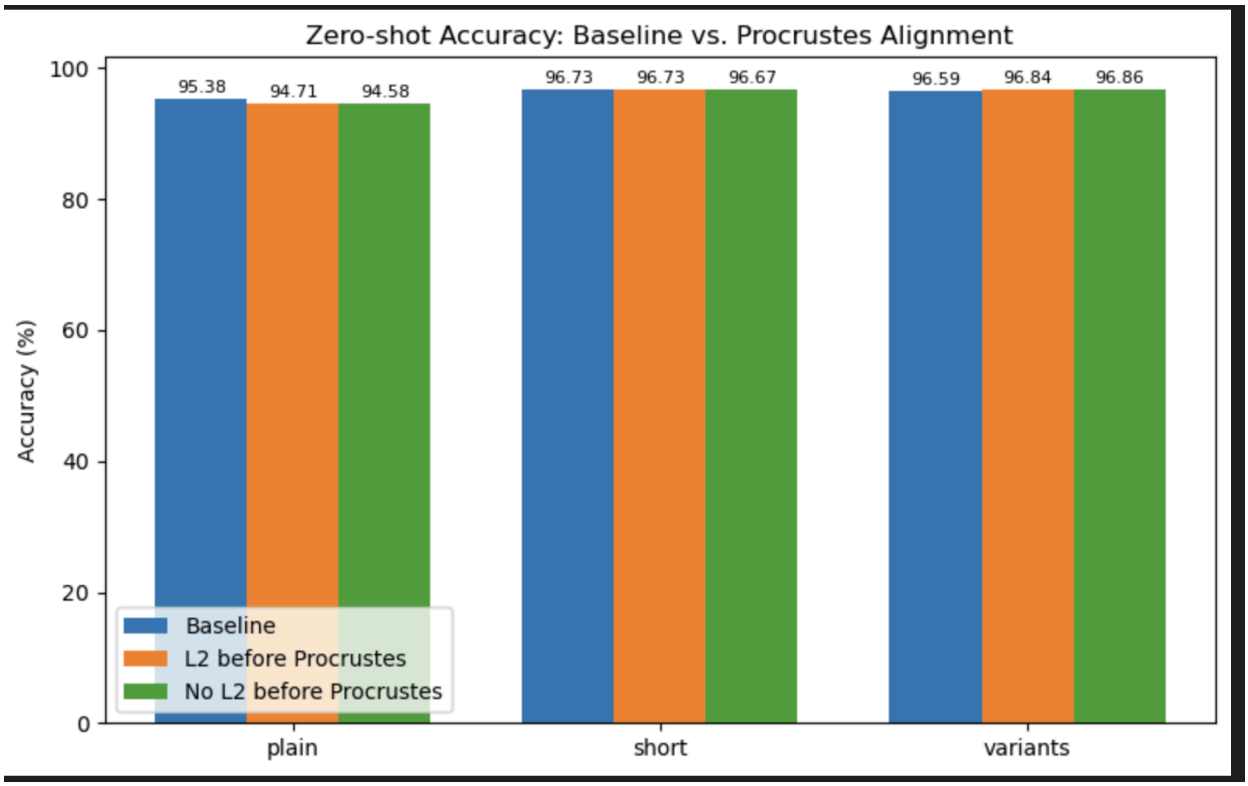
Mean cosine similarity was observed which came out to be:
Mean-Cosine-Pre: 0.2704313099384308
Mean-Cosine-Post-L2: 0.8707431554794312
Mean-Cosine-Post-NoL2: 0.8707432746887207

Alignment increases similarity between paired embeddings and improves cluster overlap, but reduces accuracy when only images are aligned (Fig. 5.6). When both images and prompts are aligned, accuracy remains high (Fig. 5.7).



(Fig 5.6)

If alignment is applied consistently to both images and prompts then accuracy stays high as shown in Fig (5.7)



(Fig 5.7)

Conclusion

CLIP performs well even with modality gap, due to preserving semantic structures during pretraining. Alignment further reduces modality gap but needs consistent application across both modalities to avoid hurting zero-shot accuracy.