AI 623 - Deep Vision Language Models
Zero Assignment

Maria Rafique [*] [1]

Abstract

This document presents the experimentation and results for the Zero-Week assignment, which consists of four tasks designed to explore foundational concepts in modern deep vision learning. Each task investigates a different aspect of representation learning and model behavior through controlled experiments. I have described the experimental setup, report quantitative and qualitative results, and discuss key observations derived from the outcomes of each task.

1. Inner Workings of ResNet-152

1.1. Baseline Setup

In this task resNet-152 model head is unfreezed only, freezed backbone and trained head only on new dataset which is CIFAR-10. CIFAR-10 images resized to 224x224 to match the IMAGENET images sizes on which resNet-152 is pre-trained. When freezed almost everything but only the last layer FC layer, The accuracies came out decent quickly which suggested that the backbone features were general and reusable. This transfer learning on new dataset helped and suggested we can leverage already learnt features/edges. As resNet-152 has oo many parameters almost 60M. CIFAR-10 has only 50k training images, so from-scratch training will overfit unless you use heavy regularization + augmentation + careful training recipes and training ResNet-152 from scratch is expensive required many epochs and lots of GPU time. Pretrained features already encode useful edges/textures/shapes, we can leverage that.

Epoch	Train Acc.	Val. Acc.
1	0.7848	0.8303
2	0.8346	0.8470
3	0.8467	0.8447
4	0.8507	0.8512

Table 1. Training and Validation Accuracy per Epoch on CIFAR-10 with resNet-152

1.2. Residual Connections in Practice

I unfroze Layer 4 and the Fully Connected (FC) head, then fine-tuned the model under two conditions: with skip connections enabled and with them disabled. When skip connections were disabled, accuracy dropped drastically. Conversely, with skip connections enabled, the loss curve remained smooth and gradients stayed stable throughout training, allowing the network to converge quickly.

	Skips ON (A)		Skips OFF (B)	
Epoch	Train Acc	Val Acc	Train Acc	Val Acc
1	0.8381	0.9023	0.1633	0.2454
2	0.9163	0.9160	0.2808	0.3045
3	0.9437	0.9154	0.3341	0.3471
4	0.9635	0.9167	0.3633	0.3753
5	0.9741	**0.9232**	—	—

Table 2. Comparison of Fine-tuning Layer 4 and Head with Skip Connections ON vs. OFF ($lr = 10^{-4}$).

Skip connections prevented gradient vanishing and smoothen the loss landscape, helped in significantly faster convergence and higher accuracy compared to plain architectures. By facilitating direct gradient flow through the $+1$ term, skip connection mitigate the degradation problem inherent in very deep networks.

1.3. Feature Hierarchies and Representations

The t-SNE visualizations of layer-1, layer-2, layer-3 revealed a clear progression in feature abstraction. In the early 2 layers, representations exhibited low class separability, with data points clustered primarily by low-level visual similarities like color and texture. As the data reaches Layer 4, I observed a significant increase in semantic separability, where points of the same class (color) form tight, distinct clusters. This confirms that the network successfully transforms raw pixel data into high-level and linearly separable categorical representations.

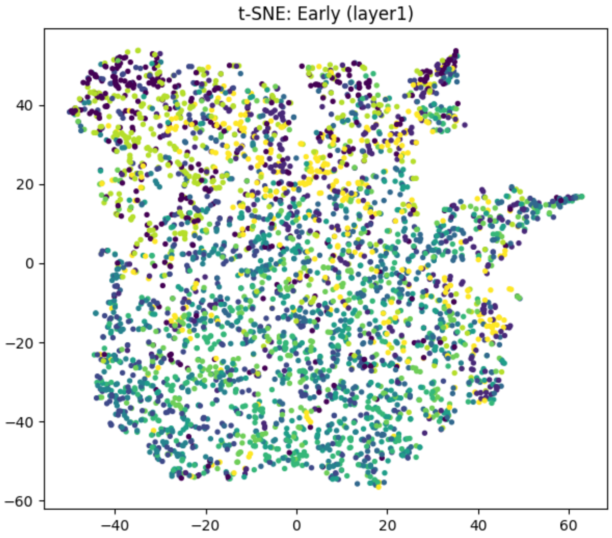

Figure 1. ResNet-152 Layer1 representation

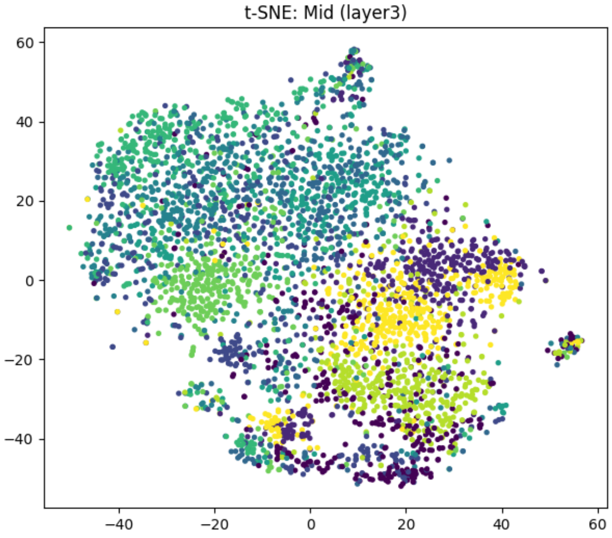

Figure 2. ResNet-152 Layer3 representation

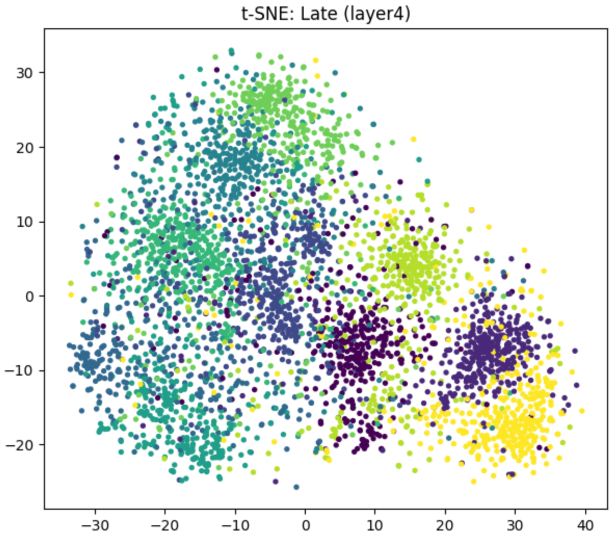

Figure 3. ResNet-152 Layer4 representation.

2. Understanding Vision Transformers (ViT)

2.1. Top-1 prediction

In this task I have loaded the ViT model from pytorch and took a following sample image.

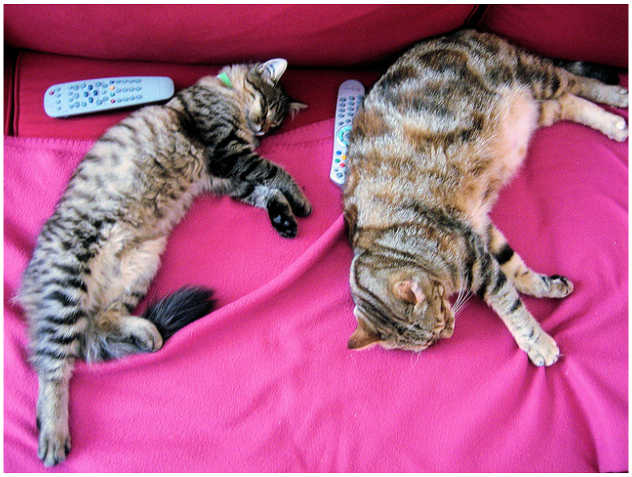

Figure 4. Top-1 Predicted class name for this figure is **tabby**

2.2. Visualizing Patch Attention

Visualize attention map over image.

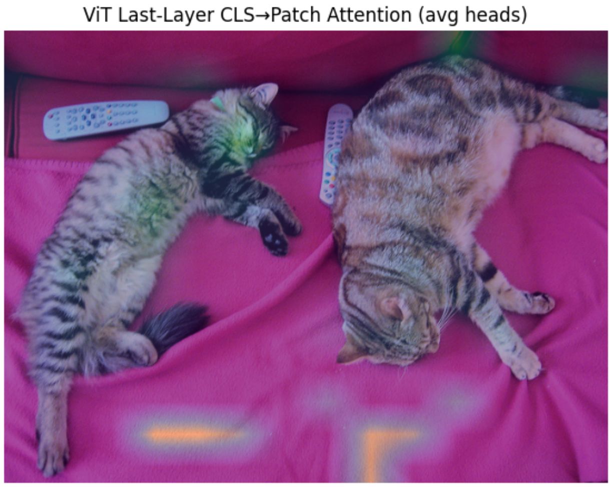

Figure 5. Aggregated Attention heads over image

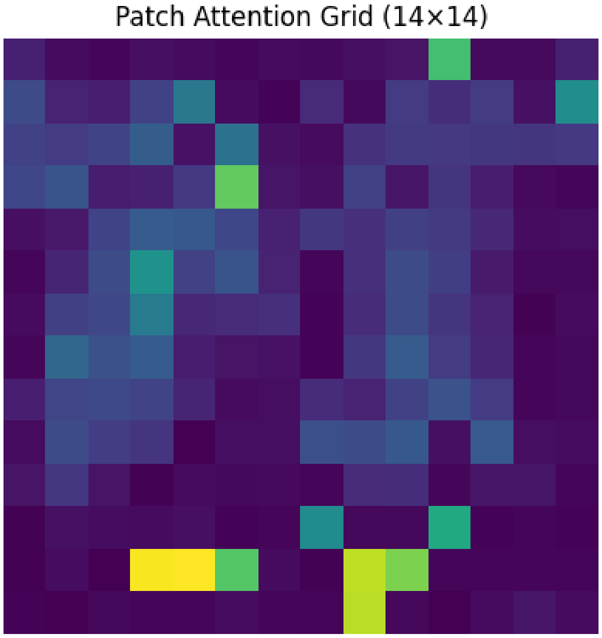

Figure 6. Attention Maps heat map

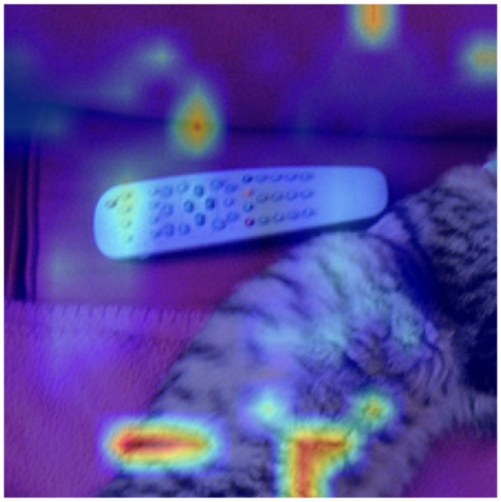

Figure 7. Visualizing Model Interpretability via Patch Attention

2.3. Analyze the Attention Map

The attention-map overlay shows how the ViT's classification token attends to different image regions in the final transformer layer. In our example, which contains a remote control placed on a couch, the attention is partially concentrated near the remote, indicating that the model uses features of the object for prediction. However, several strong attention regions also appear on the surrounding couch and blanket areas. This suggests that the model does not rely exclusively on the object itself but also incorporates contextual background information when making its decision.

Unlike convolutional neural networks (CNNs), whose Grad-CAM visualizations typically produce compact, object-shaped heatmaps due to their strong local inductive bias, the ViT's attention is more globally distributed. Since self-attention allows every patch to interact with every other patch, the model aggregates information from the entire image rather than focusing strictly on localized regions. As a result, the attention map appears more diffuse and spread out compared to CNN-based explanations.

A key advantage of transformers is that attention weights are explicitly available as part of the forward pass, making interpretability straightforward without requiring gradient-based techniques. Additionally, multi-head attention enables analysis of multiple perspectives, where different heads may specialize in focusing on object parts, textures, or global context.

One noticeable behavior is that some high-attention regions correspond to background areas rather than the object itself. This may indicate that the model is leveraging contextual cues or dataset biases. Overall, the visualization shows that

ViTs use both object-specific and contextual information for classification, reflecting their global reasoning capability

3. Training Variational Autoencoders

3.1. Training

Trained provided VAE architecture on MNIST dataset.Trained for 20 epochs, validation and training losses is decreasing and curves are smooth. Signify no overfitting. Refer to Figure 4

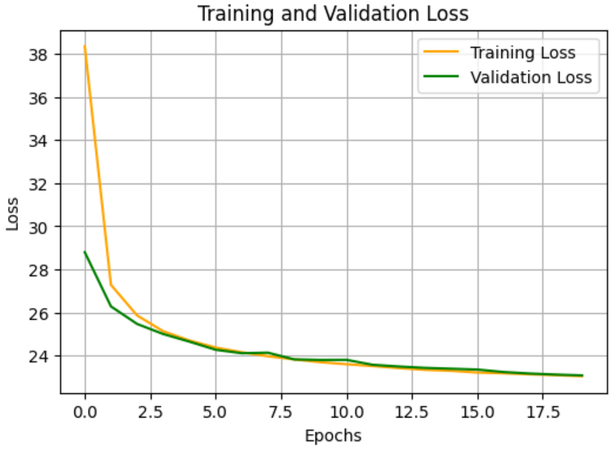

Figure 8. Training and validation loss curved for VAE on MNIST data set.

3.2. Visualize Reconstructions and Generations

Reconstructed images preserved the overall structure and class characteristics of the input images. But finer details appear a bit blurred which is expected due to the stochastic latent representation and the use of MSE error reconstruction loss. Refer to figure 5

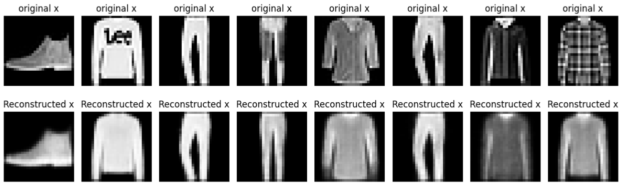

Figure 9. Original and reconstructed images comparison

Image sampling from the prior $z \sim \mathcal{N}(0, I)$. Refer to figure 6. Sampled generated from gaussian distribution. Images resemble clothing-like shapes Some samples ambiguous or unrealistic Shows model learned distribution, not memorization Samples generated from the standard normal prior resemble plausible FashionMNIST items, capturing the general structure of the dataset. However, some samples lack sharpness or clear class identity, reflecting the generative uncertainty of the model.

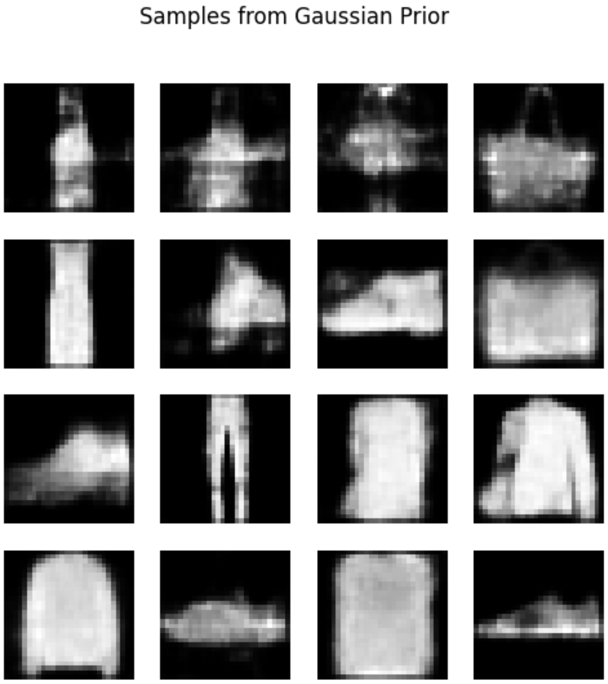

Figure 10. Sampling from Gaussian prior dist

When sampling from a Laplacian prior, the generated images exhibit increased artifacts and reduced structural coherence. This behavior is expected since the VAE was trained assuming a Gaussian latent prior, and sampling from a mismatched distribution leads to latent values outside the learned manifold.

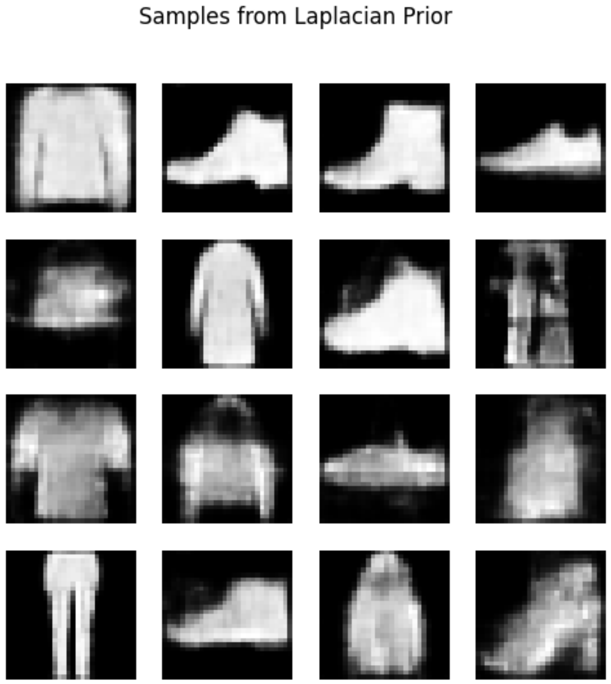

Figure 11. Sampling from Laplacian prior dist

3.3. Posterior Collapse Investigation

Objective function VAE Lose e.g. ELBO consists of a reconstruction term and a KL divergence term. In my implementation the reconstruction term is the negative log likelihood of the decoder output, which is implemented as mean squared error under a Gaussian likelihood assumption and minimizing MSE plus KL divergence is equivalent to minimizing the negative ELBO.

KL divergence is not decreasing toward zero but instead increasing initially and then stabilizes at a non-zero value. Posterior collapse was investigated by analyzing the KL divergence and the encoder outputs. The KL term does not collapse toward zero during training; instead, it increases in early epochs and stabilizes at a non-zero value for both training and validation sets. Additionally, the latent means exhibit non-zero variance across inputs, indicating that the encoder produces input-dependent representations. Therefore, the posterior is not collapsing to the prior, and the latent variables are being meaningfully utilized. See Figure 08.

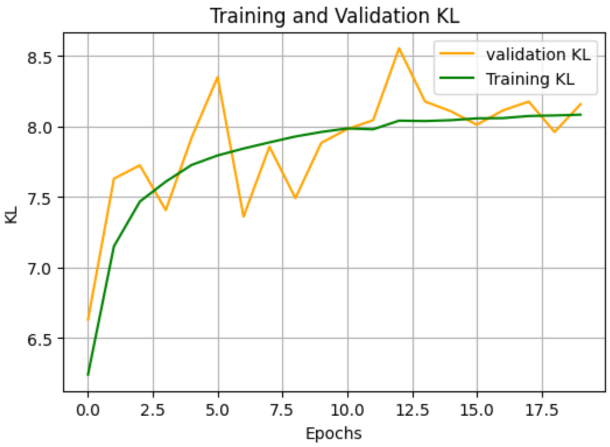

Figure 12. Sampling from Laplacian prior dist

If the decoder is too powerful, it may ignore the latent variable. Posterior collapse typically occurs when the decoder is significantly more expressive than the encoder, allowing it to model the data distribution without relying on the latent variables. In such cases, the encoder is encouraged to match the prior, causing the posterior to become independent of the input. Additionally, a strong KL regularization term early in training can force the posterior toward the prior before the encoder has learned meaningful representations, leading to collapse. This effect is especially pronounced when the KL term dominates the reconstruction loss during early optimization.

3.4. Mitigating Posterior Collapse

3.5. Transfer Learning and Generalization

4. Modality Gap

Clip is sensitive to formulation of text prompts, In this task I have studied the modality gap between text embeddings and visual embeddings by implementing different text prompt strategies and analyzed how different strategies effect zero-short classification.

4.1. Zero-Shot Classification on STL-10

Here is a brief summary of different Prompt strategies results on zero-shot classification:

4.2. Exploring the Modality Gap

Figure 1: Shows a 2D projection (UMAP/t-SNE) of CLIP image embeddings (blue) and text embeddings (orange) extracted from the STL-10 dataset. This separation indicates the presence of a modality gap. CLIP does not collapse image and text representations into an identical distribu-

Prompts	True Label	Predicted
horse	deer	horse
ship	airplane	ship
bird	airplane	bird

Table 3. Example zero-shot predictions using the *plain label* prompting strategy. While overall accuracy is **96.26%**, most errors occur between visually similar categories (e.g., deer–horse, airplane–ship), illustrating typical failure modes of CLIP.

Prompt	True Label	Predicted
a photo of a horse	deer	horse
a photo of a ship	airplane	ship
a photo of a car	truck	car

Table 4. Example zero-shot predictions using the *template-based* prompting strategy (“a photo of a {label}”). This simple natural-language formulation improves text–image alignment and achieves the highest accuracy of **97.36%**. Remaining errors mainly occur between visually similar classes (e.g., deer–horse, airplane–ship, truck–car).

tion, even though they are mapped into a shared embedding space.

Normalization constrains both image and text embeddings to lie on a shared hypersphere, which reduces scale differences between modalities and makes cosine similarity meaningful. Normalization reduces but does not eliminate the modality gap, aligning the modalities geometrically while preserving modality-specific structure.CLIP learns a shared semantic geometry where cross-modal alignment is sufficient for comparison, even if the modalities remain partially distinct. This explains why CLIP achieves high zero-shot accuracy despite a persistent modality gap. Hence These results show that CLIP aligns image and text representations at a semantic level rather than fully merging their distributions, enabling effective zero-shot inference through relative similarity.

Prompt	True Label	Predicted
a photo of a horse standing on grass	deer	horse
a photo of a ship or boat on the water	airplane	ship
a photo of a bird with wings and feathers	cat	bird

Table 5. Example predictions using *descriptive prompts*. Although more detailed text was expected to improve alignment, performance decreases to **94.00%**. Overly specific descriptions introduce contextual bias (e.g., “on grass”, “on the water”, “with wings”), which can shift embeddings away from the true class and increase confusion between semantically related categories.

Table 6. Summary of Model Accuracy

Strategy	Accuracy
Plain	96.26%
Template	97.36%
Descriptive	94.00%

Table 7. On the STL-10 test set, CLIP achieved high zero-shot accuracy across all prompting strategies. Using plain labels yielded **96.26%**, while a simple natural-language template (“a photo of a label”) improved performance to **97.36%**. However, the descriptive prompt variant reduced accuracy to **94.00%** indicating that adding extra attributes can degrade alignment when the description does not match the dataset’s visual distribution.

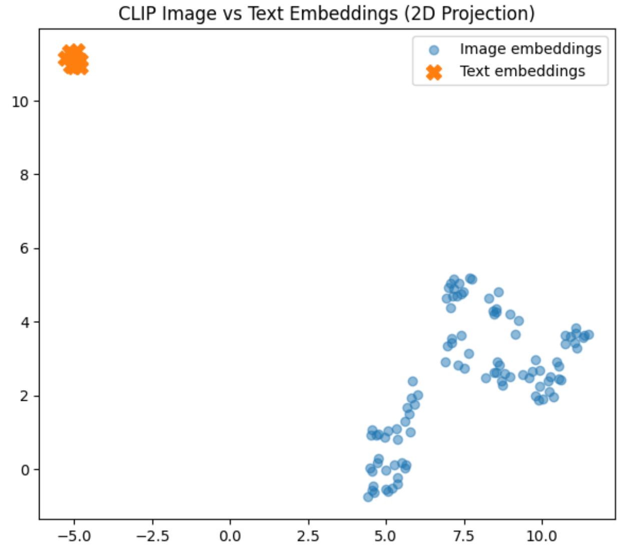

Figure 13. Visualizing the modality gap in CLIP using STL-10.

4.3. Bridging the Modality Gap

After applying the orthogonal Procrustes transform, the image embeddings shift toward the text embedding region in the 2D projection, reducing the visible separation between modalities. Since the learned transform is constrained to be orthogonal, it preserves angles and (approximately) vector norms, but changes the orientation of the image subspace to better align with the text subspace under a least-squares objective. See the plots below in **Figure 2:** and **Figure 2:**

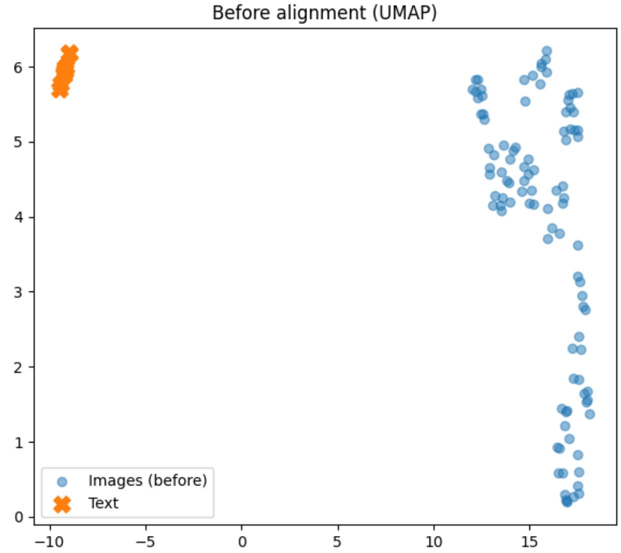

Figure 14. Visualizing the modality gap in CLIP using STL-10.

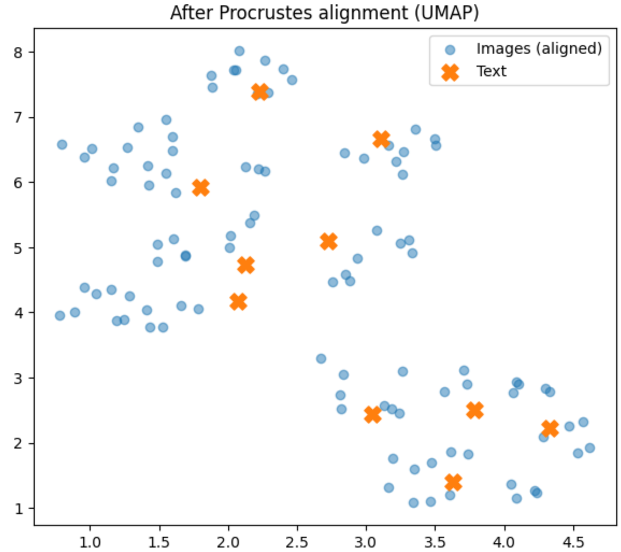

Figure 15. Visualizing the modality gap in CLIP using STL-10.

Shape of learned Rotation matrix (512, 512)

Baseline accuracy (no alignment): 97.36Aligned accuracy (Procrustes): 97.79

Table 8. Accuracy comparison before and after alignment

Strategy	**Accuracy**
Baseline accuracy (no alignment)	96.36%
Aligned accuracy (Procrustes):	97.79%

Table 9. On the STL-10 test set, CLIP achieved high zero-shot accuracy across all prompting strategies. Using plain labels yielded **96.26%**, while a simple natural-language template (“a photo of a label”) improved performance to **97.36%**. However, the descriptive prompt variant reduced accuracy to **94.00%** indicating that adding extra attributes can degrade alignment when the description does not match the dataset’s visual distribution.

References

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.