Knowledge Distillation using Whitened-CLIP

Muhammad Faraz Ahmad [1] **Amna Iftikhar** [1] **Areeb Khalid** [1]

Code Repository

The notebooks are available at: `GitHub`.

Abstract

While Vision-Language Models like CLIP have shown excellent zero-shot performance on a variety of computer vision tasks, the huge compute and memory needs of state-of-the-art CLIP models limit their practical deployment on edge devices like IoT sensors. Knowledge Distillation (KD) is a widely used model compression technique that solves this by distilling the essence of a larger teacher model into smaller student models. Although standard KD works reasonably well, we identify that standard KD strategies are hindered by the double-ellipsoid geometry of the teacher CLIP, which forces the student models to learn geometric biases rather than semantic content. In this work, we propose Whitened-CLIP KD, a geometry-aware distillation framework that uses ZCA whitening to transform the teacher's embedding space into an isotropic hypersphere. We hypothesize and validate that this transformation of the teacher's latent space enables significantly faster convergence and superior semantic alignment. Our experiments on MS-COCO (training) and CIFAR-10 (zero-shot evaluation) demonstrate that whitening significantly helps low-capacity students: our whitened MobileNet-V3 student outperforms the raw baseline by over 8% in accuracy, while the ResNet-18 student achieves near-convergence performance in a single epoch. Geometric analysis using t-SNE plots reveals that our method mitigates the modality gap and prevents feature collapse.

1. Introduction

Vision-Language Models (VLMs) like CLIP have revolutionized the field of computer vision (Radford et al., 2021), enabling various tasks such as object classification, detection, segmentation, and captioning. By aligning image and text embeddings in a shared embedding space, CLIP enables excellent zero-shot inference across various downstream tasks. However, state-of-the-art CLIP models have huge parameter counts (304.0M for `ViT-L/14` image encoder and 86.2M for `ViT-B/16` image encoder), making them impractical for deployment on edge devices like mobile phones or IoT sensors. To address this, Knowledge Distillation (KD) is used to distill the essence of a larger teacher model to a smaller student model. Standard knowledge distillation strategies try to match the image and text embeddings of the teacher and the student by either minimizing the MSE or the KL-divergence between the embeddings, or employing gradient matching or contrastive strategies. Although these strategies have shown decent results in KD, they suffer from a fundamental inefficiency rooted in the geometry of the teacher's latent space.

Recent studies have revealed that the latent space of pre-trained CLIP is not a uniform hypersphere, as ideally intended. Work by (Levi & Gilboa, 2025) shows that CLIP instead has a double-ellipsoid geometry, where the image and the text embeddings reside on two separate, tilted ellipsoids centered away from the origin. While they show that this geometry actually helps the original CLIP mitigate false negatives, we argue that this geometry makes it harder for low-capacity student models to learn from teacher CLIP models. The student models, in trying to mimic the raw embeddings of the teacher, waste capacity by learning the ellipsoid bias rather than learning the pure semantic content. This results in student models that achieve low training loss but fail to generalize in zero-shot tasks.

In this work, we propose using a geometry-aware knowledge distillation framework inspired by the whitening transform proposed by (Betser et al., 2025). Instead of distilling raw teacher embeddings, we apply a whitening transform on the embeddings and use them for distilling knowledge. This forces the teacher's latent space into an isotropic hypersphere, removing the ellipsoid bias and forcing the student model to learn semantic content. To the best of our knowledge, we are the first to propose the use of the whitening transform for knowledge distillation.

Our contributions are given as follows:

- We propose a novel KD framework, W-CLIP KD, that uses ZCA whitening to transform CLIP's double ellipsoid geometry to an isotropic hypersphere, enabling efficient distillation to low-capacity models.
- We test our method by doing KD training on MS-COCO and evaluation on CIFAR-10, demonstrating

the performance improvements of our method over the standard KD process.

- We perform detailed geometric analysis using t-SNE plots and SVD for feature collapse analysis, identifying how our method solves the failures of standard KD.

2. Related Works

Our work builds upon foundational research in VLMs like CLIP, knowledge distillation using CLIP, geometric analysis of CLIP's latent space, and latent alignment using whitening.

2.1. Vision-Language Models (VLMs)

Vision-language models like CLIP (Radford et al., 2021) have become the standard approach for learning joint image-text representations. CLIP performs excellently on zero-shot inference tasks because of its strong generalization ability. However, the huge size and memory requirements of state-of-the-art CLIP models limit their practical use. To solve this, various model compression methods are used.

2.2. Knowledge Distillation using CLIP

Knowledge distillation is one of the most widely used techniques for model compression. Originally proposed by (Hinton et al., 2015), knowledge distillation transfers knowledge by training the student model to match the soft logits of a teacher model. For CLIP, the state-of-the-art approach is CLIP-KD (Yang et al., 2024), in which many distillation strategies are discussed, and simple feature distillation (MSE between the embeddings) is found to outperform other relation-based and contrastive-based distillation strategies. Other works like TinyCLIP (Wu et al., 2023) and MobileCLIP (Vasu et al., 2024) focus on architectural efficiency, using weight inheritance and hybrid architectures, respectively, to create efficient student models. However, all these distillation strategies use raw teacher embeddings, thereby inheriting the geometric flaws of the teacher's latent space.

2.3. CLIP's Latent Geometry

Different recent works have analyzed the geometry of CLIP's latent space and found interesting results. (Liang et al., 2022) discovered the modality gap in the latent space: text and image embeddings lie on different parts in the space, separated by a constant gap. (Fahim et al., 2024) further studied this gap and found that it emerges due to the two-encoder contrastive loss and therefore termed it as contrastive gap. (Liang et al., 2022) also discovered the thin-shell phenomenon where most of the embedding mass is concentrated within a small range from the mean. Recently, (Levi & Gilboa, 2025) characterized CLIP's geometry as two tilted ellipsoids (one for each modality) centered away from the origin. These works highlight the flaws in CLIP's latent space that may hinder learning for low-capacity student models.

2.4. Whitening Transform

Recently, (Betser et al., 2025) proposed Whitened CLIP primarily for likelihood estimation of images and captions. By whitening the raw embeddings of the teacher, they forced the embeddings into a standard normal distribution, effectively transforming the latent into uniform hyperspheres (one for each modality). While (Betser et al., 2025) focused on using W-CLIP for log-likelihood estimations, we extend this concept to knowledge distillation. We integrate the whitening process as a pre-distillation step and hypothesize that training with whitened geometry will lead to more efficient learning by the student compared to training with tilted ellipsoids.

3. Methodology

This section contains the detailed methodology of our work. The computation of the whitening matrix for the training dataset is done slightly differently compared to how it's done in the whitening paper by (Betser et al., 2025). The technical details about that are given in section 3.1. The details of our experimental setup are given in section 3.2.

3.1. Theoretical Basis:

As previously explained in section 2.3, the embeddings of standard CLIP lie in narrow cones where features are highly correlated. To transform the embedding space into a uniform hypersphere where features are decorrelated, we learn a whitening transform W directly from the training data and apply it to the raw embeddings generated by the teacher, before distilling them into the student. However, we don't follow the exact setup for computing W as done by (Betser et al., 2025). The paper uses PCA (Principal Component Analysis) for computing W for likelihood estimation, but we use ZCA (Zero-phase Component Analysis) for computing W for knowledge distillation. Here, using ZCA for W helps our cause of knowledge distillation, which was not possible with using PCA. (Betser et al., 2025) explicitly note in the appendix that using PCA results in the loss of original geometry, as PCA aligns the data along the axes of maximum variance (principal components), effectively rotating the latent space.

For knowledge distillation, this rotation is detrimental. If we consider a simple MSE objective $L_{\text{MSE}} = ||S - T||^2$, the student S simply attempts to mimic the features of teacher T. If we use PCA-based whitening, the semantic meaning of feature dimensions gets scrambled; the dimension that originally encoded ”texture” may encode a completely different feature after rotation. This will force the student model to learn an additional rotation matrix in addition to the semantic features, wasting the model's capacity.

In contrast, ZCA-based whitening transforms the data to be isotropic ($\Sigma = I$) but still rotates the latent space back to the original dimensions, preserving as much semantic information as possible. Explanation of the calculation of W is given in the Appendix 7.1.

3.2. Experimental Setup:

3.2.1. Datasets:

We primarily used the MS-COCO dataset for training and the CIFAR-10 dataset for evaluation.

- **MS-COCO 2017:** We train all the student models on the MS-COCO 2017 dataset's training split downloaded from Kaggle under `awsaf49/coco-2017-dataset`. It contains approximately 118,000 image-caption pairs. CLIP-KD paper by (Yang et al., 2024) used CC(3M+12M) for training, but that dataset was huge, so we had to use MS-COCO. MS-COCO is still diverse enough for our purposes.
- **CIFAR-10**: To assess the student models, we perform zero-shot classification on the CIFAR-10 dataset. The student model predicts the class of an unseen image by comparing its embedding to the text embeddings of prompts like "a photo of a [class]" (e.g., "a photo of a car", "a photo of a frog"). This checks the generalization capabilities of the student models as they never saw CIFAR-10 during training, and a good performance on CIFAR-10 implies good generalization capabilities.

3.2.2. Model Architectures:

Teacher Model: We use the `ViT-B/16` variant of CLIP, pre-trained on the LAION-2B dataset (OpenCLIP) as our teacher model. We chose this model because it's very commonly used as the teacher in literature, and its high zero-shot capabilities with a huge parameter count of ≈ 86.2M make it a good candidate for being a teacher model.

Student Models: We used several student models to test the effectiveness of our method, including `MobileNetV3`, `MobileViT-S`, and `ResNet-18`. The details of their architecture are given in the Appendix 7.2.

3.2.3. Embedding Storage:

One important step that we did before training was pre-computing the embeddings for the entire dataset. MS-COCO is a huge dataset, and computing the embeddings while training takes a lot of time. To avoid that, we compute the embeddings of the teacher over all 118k images and captions and store them as static NumPy arrays as `coco_teacher_img.npy` and `coco_teacher_txt.npy`. Then, we compute the mean μ and the whitening matrix W from these embeddings before the training begins.

3.2.4. Hyperparameters & Training details:

For all the student models, we used the AdamW optimizer with a learning rate of 1e-4. We used a batch size of 128 for most training setups (only used 256 for one). 128 was the maximum batch size we could use, given the limited compute and memory requirements. The paper used a 1024 batch size with 8 dedicated A800 GPUs, which we did not have access to. Our training was done mostly on a single A100 GPU or T4 GPU.

We trained the models for 5 epochs, sufficient to show reasonable trends. We used the following weights for different components of the loss function (as used in CLIP-KD): $\lambda_{\text{InfoNCE}} = 1.0$, $\lambda_{ICL} = 1.0$, $\lambda_{FD} = 2000.0$. Our loss function for the whitened student was:

$$L = L_{\text{InfoNCE}} + 2000 L_{\text{FD}} + L_{\text{ICL}}$$

We also used linear projection heads to project the embeddings of all encoders to 512 dimensions, allowing us to compute the components of the loss.

4. Results

This section contains detailed results for the student models we tested. Our results demonstrate decent gains in accuracy and convergence speed with whitening over standard knowledge distillation. Detailed discussions of each model are provided in the corresponding discussion sections (5.1, 5.2, 5.3).

4.1. ResNet-18 Results

For ResNet-18, we found that whitening significantly affected the convergence speed for the student model and consistently got a better accuracy on CIFAR-10 compared to the unwhitened student. We can see this from the log in 4.1.1.

4.1.1. Training Log:

Table 1 shows student ResNet-18 models' accuracy over 5 epochs.

4.1.2. Loss plots:

Fig. 1 represents the loss plots for each individual component of the loss function.

Epoch	Standard-KD Acc.	Whitened-KD Acc.
1	48.97%	57.15%
2	48.29%	55.91%
3	46.70%	56.85%
4	54.01%	57.27%
5	56.71%	56.96%

Table 1. ResNet-18 Training Log

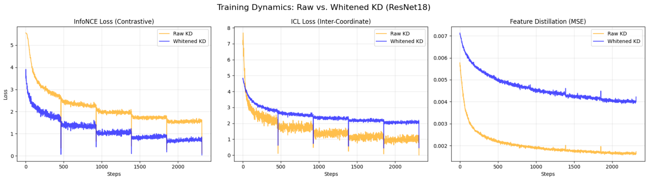

Figure 1. Loss Curves (Epochs 1-5)

4.1.3. t-SNE Plots:

Fig. 2 (a) shows the t-SNE plots for the teacher and different students, each student with a different configuration of loss. Fig. 2 (b) shows the correlation matrices while Fig. 2 (c) shows the feature collapse analysis, indicating how much each model is using each of the 512 dimensions. The ablation of losses and accuracies is given in the appendix 7.3.1 in Fig. 5.

4.2. MobileNet-V3 Results

MobileNet-V3 got the biggest impact from whitening and we saw the highest accuracy gap, of over 8% in the last epoch. The training log in Table 2 shows this gap.

Epoch	Standard-KD Acc.	Whitened-KD Acc.
1	36.93%	55.85%
2	42.26%	54.65%
3	45.26%	55.53%
4	44.40%	54.08%
5	47.19%	55.46%

Table 2. MobileNet-V3 Training Log

4.2.1. t-SNE Plots:

The t-SNE plots for MobileNet students, correlation matrices, and feature collapse plots are given in Fig. 3. The detailed discussion of these results is in Section 5.2.

4.3. MobileViT-S Results

For MobileViT-S, we observe different training dynamics than previously observed. We can see from Table 3 that the whitened MobileViT-S student initially struggles against the raw student, lagging behind in accuracy by some percent. From epoch 3, it takes over and maintains a better accuracy.

Epoch	Standard-KD Acc.	Whitened-KD Acc.
1	48.56%	44.46%
2	55.92%	53.67%
3	57.31%	59.13%
4	59.63%	61.32%
5	61.60%	61.70%

Table 3. MobileViT-S Training Log

4.3.1. t-SNE Plots:

The t-SNE plots for the teacher and MobileViT-S students for all loss configurations are given in Fig. 4.

5. Discussion

5.1. ResNet-18 Discussion

As we can see from the training log in Table 1, the whitened student consistently outperformed the student trained with raw teacher embeddings. The interesting observation here is the convergence speed of whitened student vs. raw student: the whitened student got to $\approx$ 57% accuracy in the first epoch, while the raw student needed 4 more epochs to reach that accuracy. This shows that whitening accelerates the training process, and a whitened student may only require a fraction of epochs to reach a decent accuracy compared to the raw student.

Loss Plots: We see many interesting observations from the loss curves in Fig. 1. We observe that the InfoNCE loss for the whitened student is consistently lower than that of the raw student, as expected. The ICL loss for both students remains close, although the raw student gets a lower loss here. But, for the Feature Distillation (FD) loss, we consistently see that the whitened student has way higher loss compared to that of the raw student. This is expected, and it is because of the new uniform hypersphere geometry that comes out of the whitening process. This geometry has larger average Euclidean distances (as it's spread out); however, this actually helps the student model to learn superior semantic alignment, resulting in better generalization capabilities. This is evident from lower InfoNCE loss and higher accuracies on CIFAR-10 over all epochs.

t-SNE Plots: From Fig. 2 (a), we can see the t-SNE plots for the teacher and all the students with different configurations of losses. As expected, we see a huge modality gap in the teacher model and the raw student model (of 0.7816 and 0.8838, respectively). For the whitened student, this gap reduces to only 0.1229, and we start to see good clusters. We see one counterintuitive thing in the t-SNEs of ICL (Raw and White), which should show the modality

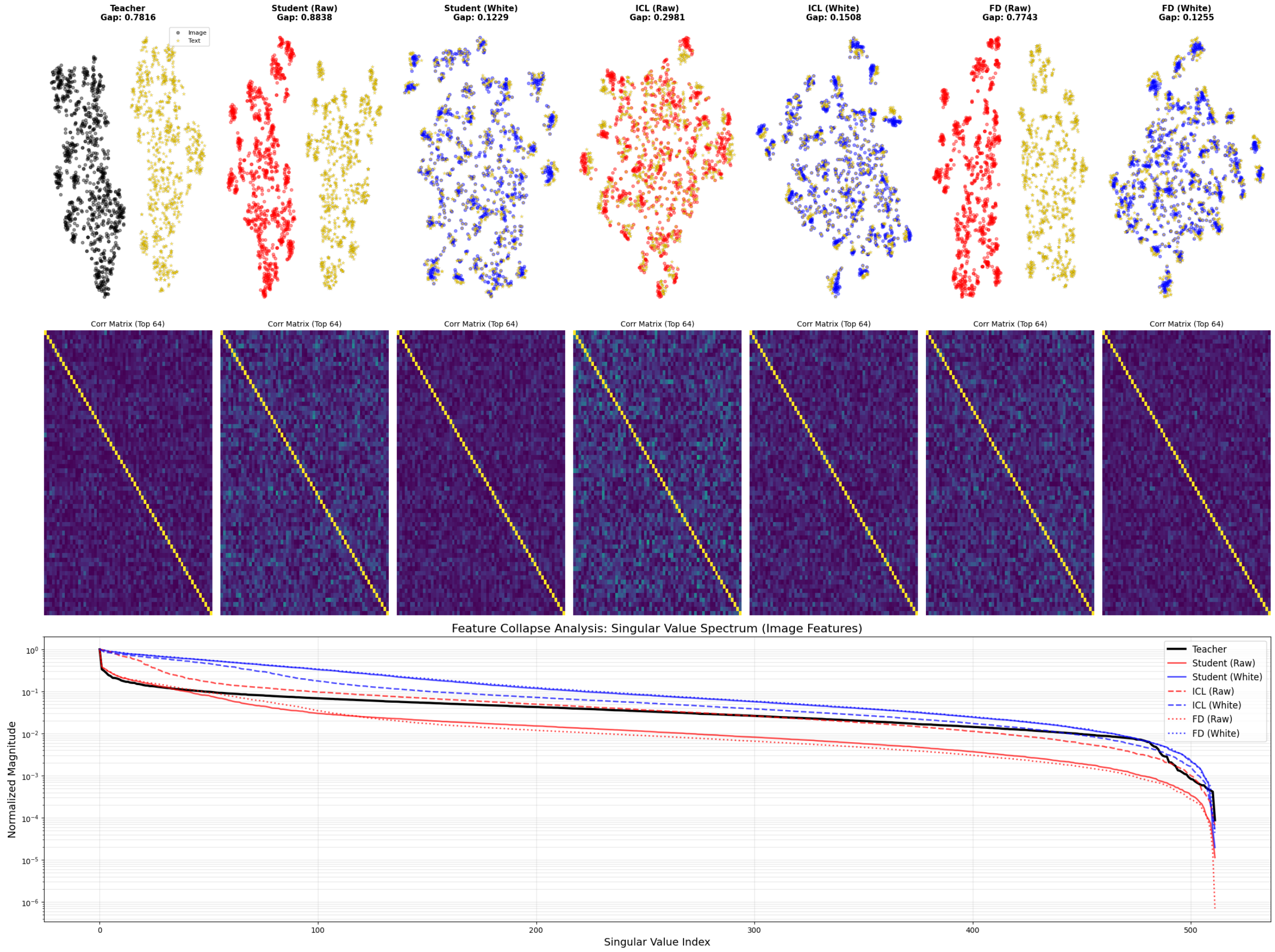

Figure 2. ResNet-18's (a) t-SNEs (b) Correlation Matrices (c) Feature Collapse Analysis

gap, but they don't. The effect can be explained by the fact that ResNet-18 has a lot more parameters ($\approx 11.4M$) than MobileNet-V3, because of which it is able to brute-force the ICL loss. It finds a way to pull the modalities together to maximize similarity scores, even without explicit geometric guidance. On the other hand, we see expected t-SNEs in raw FD (with a modality gap), and whitened FD (with the modality gap removed).

Correlation Matrices: The correlation matrices of the top 64 features also show the results we expect: all the correlation matrices of whitened students are much darker, implying close-to-zero covariance with other features. On the other hand, the raw students have lighter entries in the matrices, implying correlation between features.

Feature Collapse Analysis: Fig. 2 (c) shows the feature collapse analysis plots, which tell how much information is stored in each of the 512 dimensions. We can see that the whitened student and the FD-only whitened student preserve the most information across all dimensions. The raw students show a sharp decay, indicating that they store most of their information in only a few dimensions (e.g., the top 50).

A slower decay indicates a higher effective rank, which means that the model is utilizing a larger portion of the available dimensions, which is the case we see in the whitened students.

5.2. MobileNet-V3 Discussion

For MobileNet-V3, we observe the highest impact of whitening among the three tested models. From the training log in Table 2, we can see that the whitened student is up in accuracy by a huge margin for all epochs, even better by around $\approx 18.92\%$ in the first epoch. We again confirm here that the convergence speed is much better for whitened models compared to raw models; whitening accelerates the training process and only requires a few epochs to converge. The loss trends for MobileNet-V3 also match the ones for ResNet-18.

t-SNE Plots: Fig. 3 (a) shows the t-SNE plots for the teacher

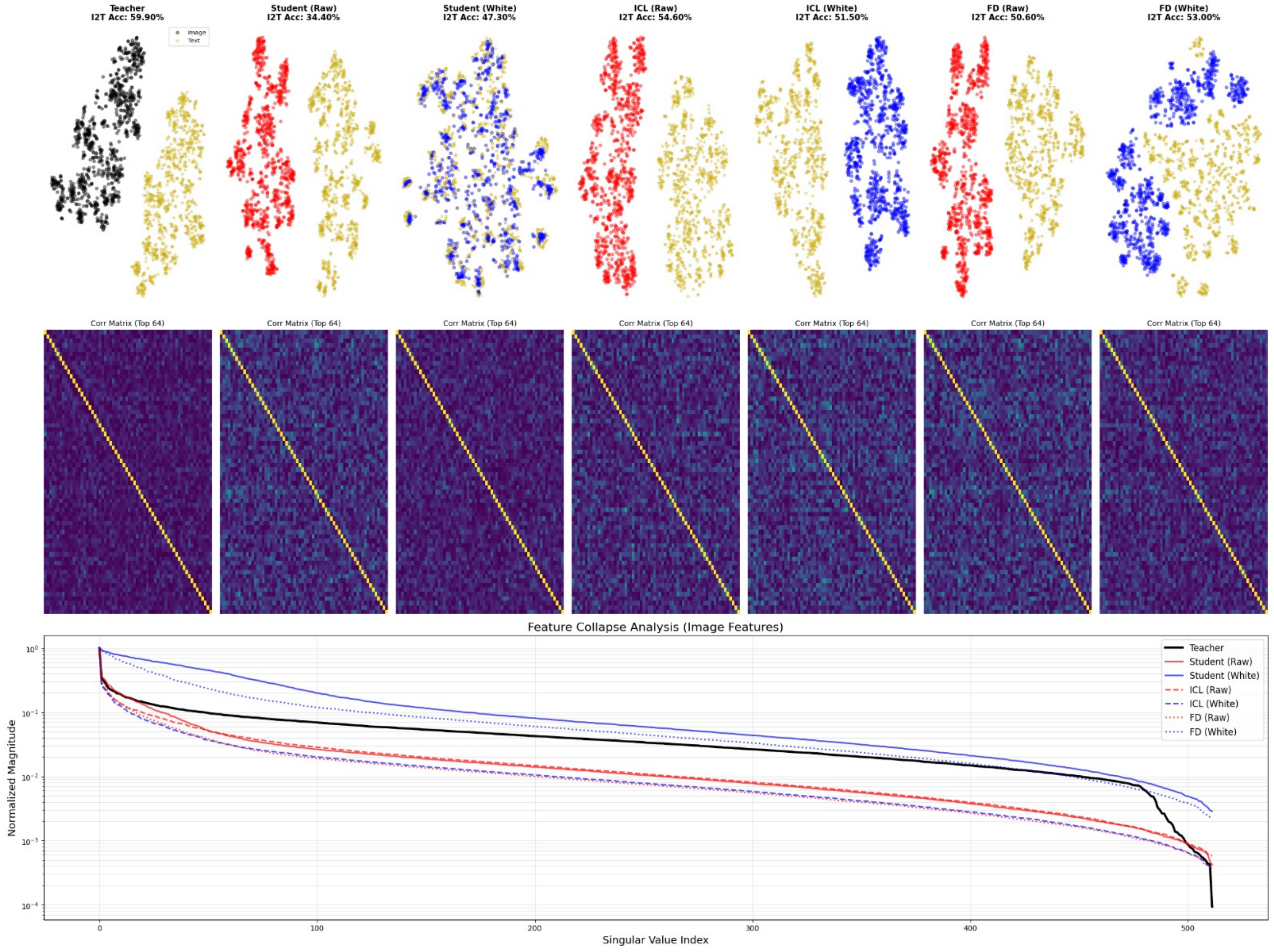

Figure 3. MobileNet-V3's (a) t-SNEs (b) Correlation Matrices (c) Feature Collapse Analysis

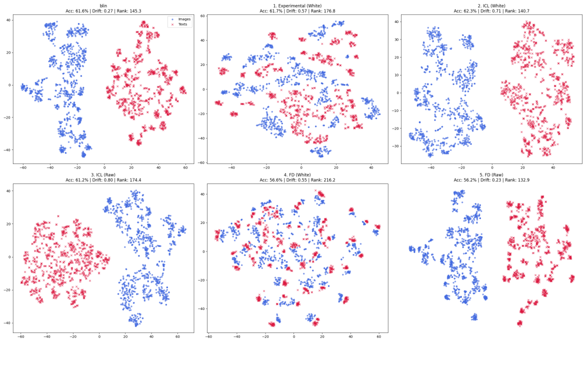

Figure 4. MobileViT-S t-SNE plots

and all the students. The t-SNE plots for the teacher, the raw student, and the whitened student (essentially the first 3) are very similar to the ones we saw for ResNet in 2 (a). The main difference here is that both raw and whitened ICL students now show a modality gap, which vanished for ResNet due to its greater capacity. MobileNet is too small (2M parameters) to get clusters without geometric guidance by FD; therefore, it defaults to keeping images and text separate. FD Raw also shows the gap, and it is only after whitening FD that we see the modality gap starting to reduce. And because of the whitened FD component of the total loss, the total whitened student shows good clusters where semantics dominate.

Correlation Matrices & Feature Collapse: Just as for ResNet-18, the correlation matrices for the whitened students are darker than the ones for raw students, indicating very low feature correlation. One different observation is for the correlation matrices of ICL (White) and FD (White), which contain relatively lighter entries compared to their ResNet counterparts. This is due to the models not getting trained until convergence because of low compute access. As for feature collapse analysis from Fig. 3 (c), the trends are very similar to ResNet's trends.

5.3. MobileViT-S Discussion

We can see from Table 3 that for MobileViT-S, the whitened student model starts with a lower accuracy than the raw student model. The whitened student lags behind the raw student for two epochs, and from epoch 3, it begins leading the raw student and maintains its lead for epochs 4 and 5, albeit by a small margin. This lagging behavior is unexpected, but we hypothesize that it is due to the hybrid architecture of Mobile-ViT (CNN+Tranformer), which makes it difficult to model the whitened geometry initially. The whitened student here requires more training steps to model this geometry, but once it's learned, the whitened student overtakes the raw student in accuracy.

t-SNE Plots: The t-SNE plots in Fig. 4 show expected results, very similar to the t-SNE plots in MobileNet-V3. As expected, the raw-embeddings' models have a huge modality gap, and the whitened-embeddings' models have a lower modality gap (except for whitened ICL, as whitening majorly affects FD only). We see a lot of clusters forming for FD-Whitened for similar images and texts, confirming that whitened geometry helps InfoNCE a lot to form the clusters more easily compared to without whitening.

6. Conclusion

In this work, we presented a novel KD framework based on correcting the geometry of teacher CLIP models by applying ZCA whitening to the raw image and text embeddings. We tested the working of our framework by comparing the accuracy scores of students with and without whitening, and found that all three models got improvements in accuracy scores with whitening. Possible future directions of our work include testing our method on larger datasets, such as CC(3M+12M), with a larger batch size; testing the method on students with different architectures; and tuning the hyperparameters and weights of the components in the loss function or adding/removing components in the loss function.

References

Betser, R., Levi, M. Y., and Gilboa, G. Whitened clip as a likelihood surrogate of images and captions, 2025. URL https://arxiv.org/abs/2505.06934.

Fahim, A., Murphy, A., and Fyshe, A. It's not a modality gap: Characterizing and addressing the contrastive gap, 2024. URL https://arxiv.org/abs/2405.18570.

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network, 2015. URL https://arxiv.org/abs/1503.02531.

Levi, M. Y. and Gilboa, G. The double-ellipsoid geometry of clip, 2025. URL https://arxiv.org/abs/2411.14517.

Liang, W., Zhang, Y., Kwon, Y., Yeung, S., and Zou, J. Mind the gap: Understanding the modality gap in multimodal contrastive representation learning, 2022. URL https://arxiv.org/abs/2203.02053.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Vasu, P. K. A., Pouransari, H., Faghri, F., Vemulapalli, R., and Tuzel, O. Mobileclip: Fast image-text models through multi-modal reinforced training, 2024. URL https://arxiv.org/abs/2311.17049.

Wu, K., Peng, H., Zhou, Z., Xiao, B., Liu, M., Yuan, L., Xuan, H., Valenzuela, M., Xi, Chen, Wang, X., Chao, H., and Hu, H. Tinyclip: Clip distillation via affinity mimicking and weight inheritance, 2023. URL https://arxiv.org/abs/2309.12314.

Yang, C., An, Z., Huang, L., Bi, J., Yu, X., Yang, H., Diao, B., and Xu, Y. Clip-kd: An empirical study of clip model distillation, 2024. URL https://arxiv.org/abs/2307.12732.

7. Appendix

7.1. W calculation:

Given the centered teacher embeddings $T \in \mathbb{R}^{N \times D}$ computed over the training set, where N is the size of the training set and D is the embedding dimension, we compute the covariance matrix $\Sigma = \frac{1}{N-1} T^T T$. Decomposing $\Sigma = U \Lambda U^T$, the ZCA whitening matrix W_{ZCA} is computed as:

$$W_{ZCA} = U(\Lambda + \epsilon I)^{-1/2} U^T$$

Λ represents the matrix for eigenvalues, while U matrix contains the corresponding eigenvectors. The term U^T rotates the data into the eigenbasis for scaling, and the leading U rotates it back to the original orientation. The whitened target T_{white} is then:

$$T_{white} = T W_{ZCA}$$

7.2. Student Architectures:

Image Encoder	Text Encoder
MobileNet-V3 (2.0M)	L: 12, d_h: 384, h=6 (21.3M)
MobileViT-S (5.3M)	L: 12, d_h: 384, h=6 (21.3M)
ResNet-18 (11.4M)	L: 12, d_h: 384, h=6 (21.3M)

L represents the number of layers in the transformer, d_h represents the width/each head's dimension, h represents the number of heads.

7.3. Additional Results

7.3.1. ResNet-18 Results

From Fig. 5, we can see that FD with whitening gives the highest accuracies across all epochs.

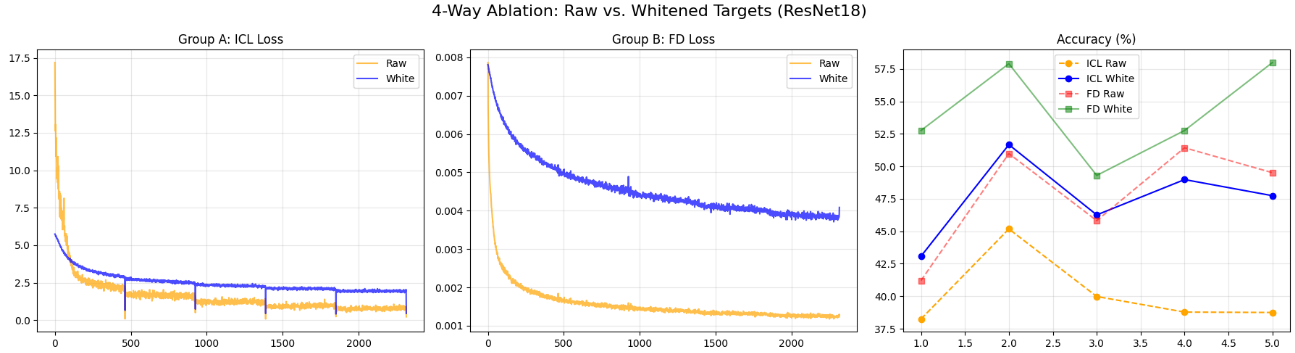

Figure 5. Ablation of Loss and Accuracy comparison