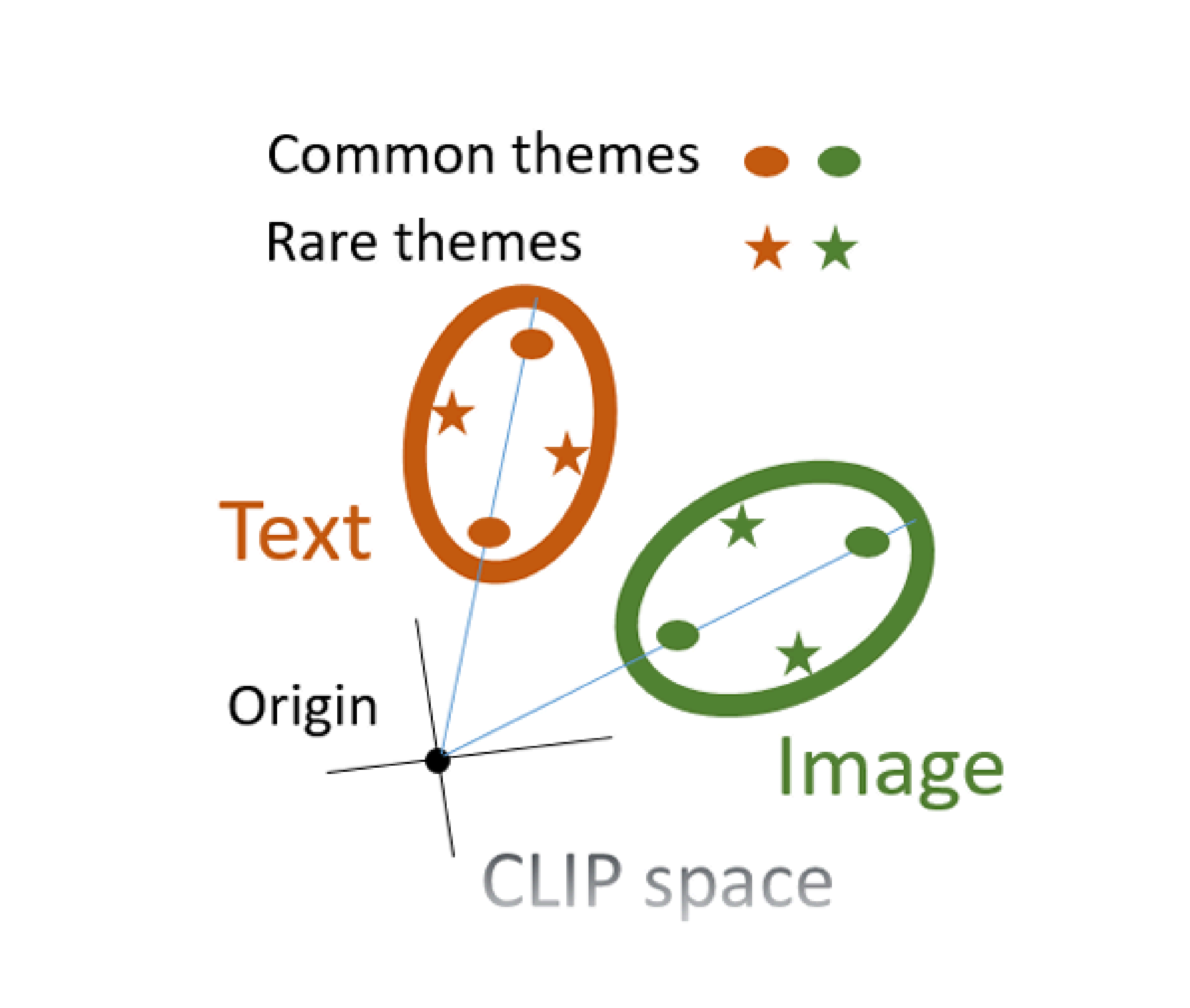
KNOWLEDGE DISTILLATION USING WHITENED-CLIP

Group 18
Muhammad Faraz Ahmad
Areeb Khalid Kidwai
Amna Iftikhar

CLIP-KD LIMITATION: LATENT GEOMETRY

- Recent studies prove that pre-trained CLIP's embeddings suffer from the modality gap and the narrow cone effect
- "Double-Ellipsoid" paper by Levi & Gilboa (2025) shows that both text and image modalities lie on separate tilted ellipsoids
- **Failure Mode:** Standard CLIP-KD forces students to mimic this complex, distorted geometry
- **Claim:** Low-capacity students (e.g., MobileNet) waste capacity learning this geometric bias rather than semantic content, resulting in poor zero-shot generalization

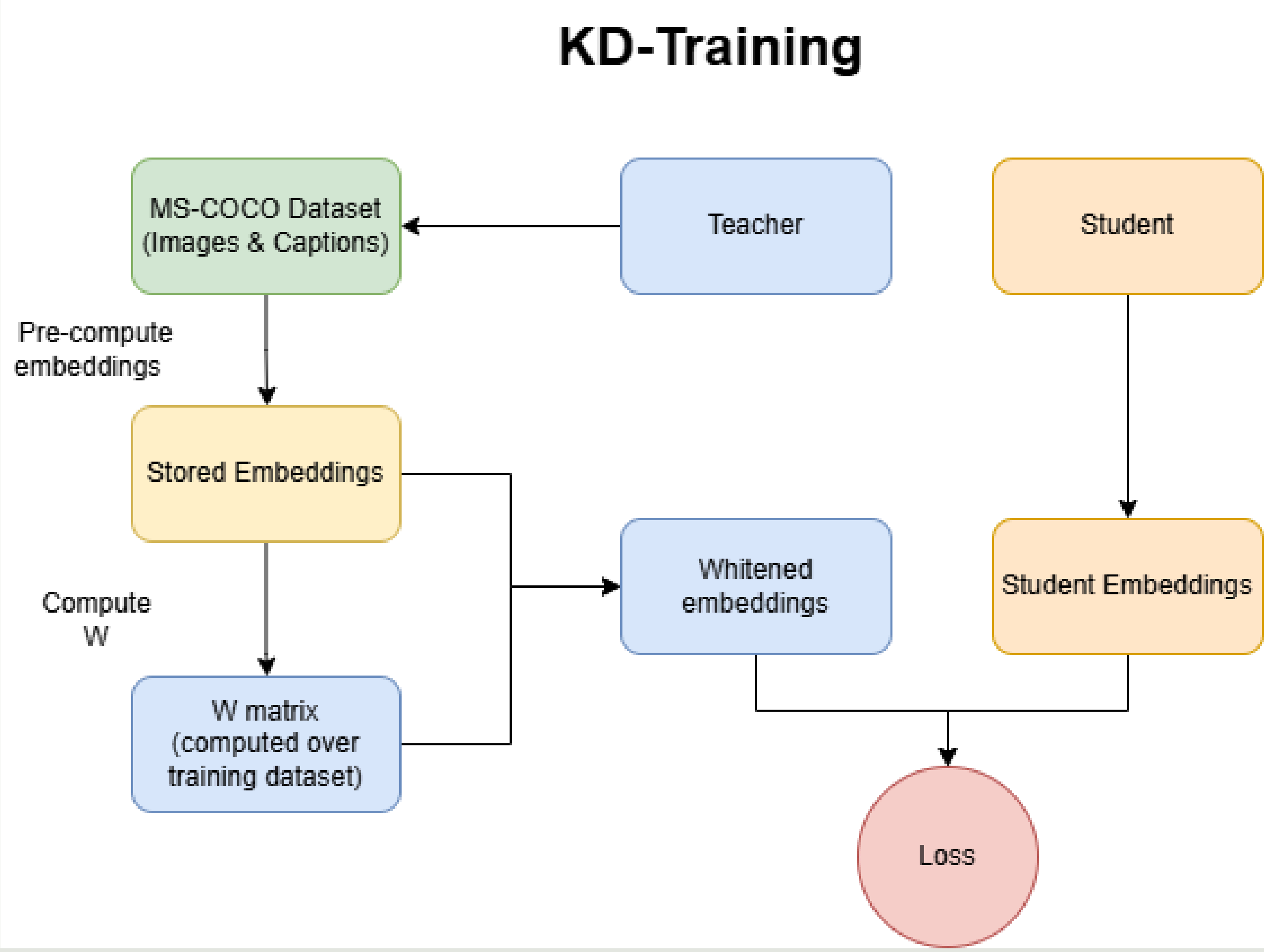


SOLUTION: WHITENED-CLIP KD

- **Proposed Solution:** Using ZCA whitening to transform CLIP's latent space into an isotropic hypersphere (Σ=I)

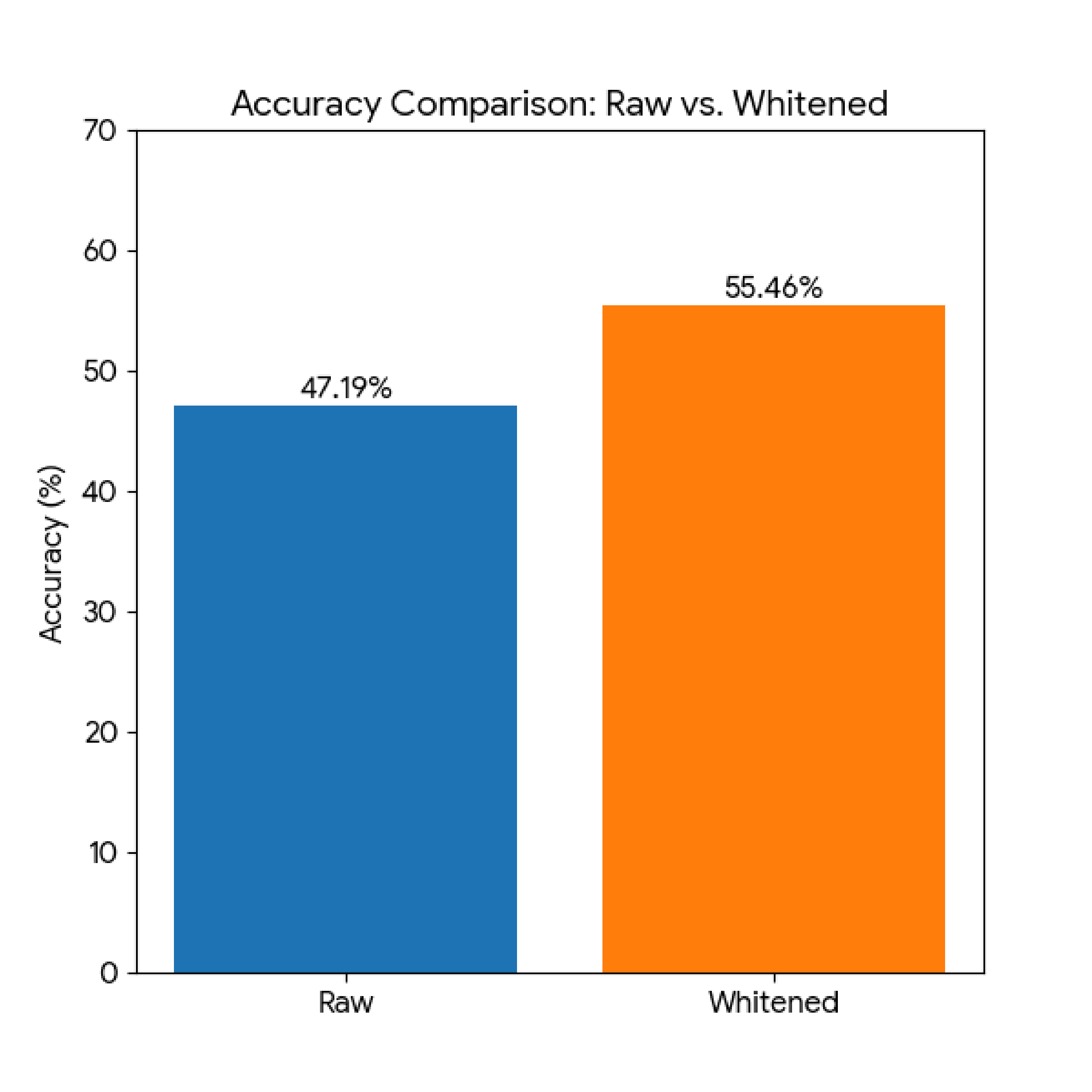
$$T_{white} = (T - \mu)W_{ZCA}$$

- **Novelty:** We use ZCA whitening instead of PCA-based whitening used by the W-CLIP paper because ZCA rotates the data back into its original axes, preserving semantic alignment
- **Setup:**
    - **Teacher:** CLIP *ViT-B/16* (~86.2M, pre-computed embeddings)
    - **Students:** MobileNetV3 (~2M), ResNet-18 (~11.4M), MobileViT (~5.3M)
    - Training on MS-COCO, Zero-shot evaluation on CIFAR-10



RESULTS

- In our experiments, the whitened student showed significant improvements in:
  - Convergence speed
  - Accuracy Gains
- **MobileNet-V3:**
  - **Epoch 1 Accuracy:** 36.93% (Raw) vs. 55.85% (Whitened) (a ~19% increase)
  - **Epoch 5 Accuracy:** 47.19% (Raw) vs. 55.46% (Whitened) (still an 8% difference)
- **ResNet-18:**
  - Whitened student achieves 57.15% in one epoch, which takes 5 epochs for the raw student
  - Whitened student consistently maintains an accuracy lead over the raw student
- For MobileNet-V3, whitening acted as a necessity (huge performance increase), while for ResNet-18, it acted as a convergence accelerator



GEOMETRIC ANALYSIS

- Closing the modality Gap:
  Whitening forces the teacher embeddings to spread out,
  allowing Feature distillation and contrastive clustering to pair them closer, reducing the modality gap.
- Feature Independence:
  As per the sparse correlation matrices, whitening promotes feature Independence, and allows for easier modal
- Mitigating Feature Collapse:
  As per the eigenvalue decomposition, we can observe that individual feature variance remains consistent throughout all dimensions, preventing collapse

CONCLUSION & FUTURE WORK

- **Tradeoff:** Our experiments suggest that there is a tradeoff between model capacity and geometry
    - Low-capacity models like MobileNet-V3 benefit a lot from whitening: whitening not only accelerates convergence, but it also significantly improves performance
    - Higher-capacity models like ResNet-18 get faster convergence from whitening, but after a few epochs, both models converge to around the same accuracy
- **Practical use case:** Whitened-KD can help very small models achieve very decent performance, which can then be deployed in IoT sensors
- **Future Work:** Using larger datasets (CC12M), evaluating the method on more student models specifically hybrid-architecture students, parameter tuning

APPENDIX SLIDES

W CALCULATION

- Given the teacher's embeddings $T \in \mathbb{R}^{N \times D}$, where N is the size of the training dataset, and D is the embedding dimension, we compute the covariance matrix using:

$$\Sigma = \frac{1}{N-1}(T-\mu)^T(T-\mu)$$

- We then decompose Σ using Singular Value Decomposition into:

$$\Sigma = U \Lambda U^T$$

- We can then compute the ZCA-based whitening matrix W using:

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

- Finally, we can compute whitening embeddings using:

$$T_{white} = (T-\mu) W_{ZCA}$$

STUDENT ARCHITECTURES

Image Encoders	Text Encoders
1. MobileNet-V3 (2.0M)	L: 12, dh: 384, h=6 (21.3M)
2. ResNet-18 (11.4M)	L: 12, dh: 384, h=6 (21.3M)
3. MobileViT-S (5.3M)	L: 12, dh: 384, h=6 (21.3M)

L: no. of layers, dh: dimension of each head, h: no. of heads

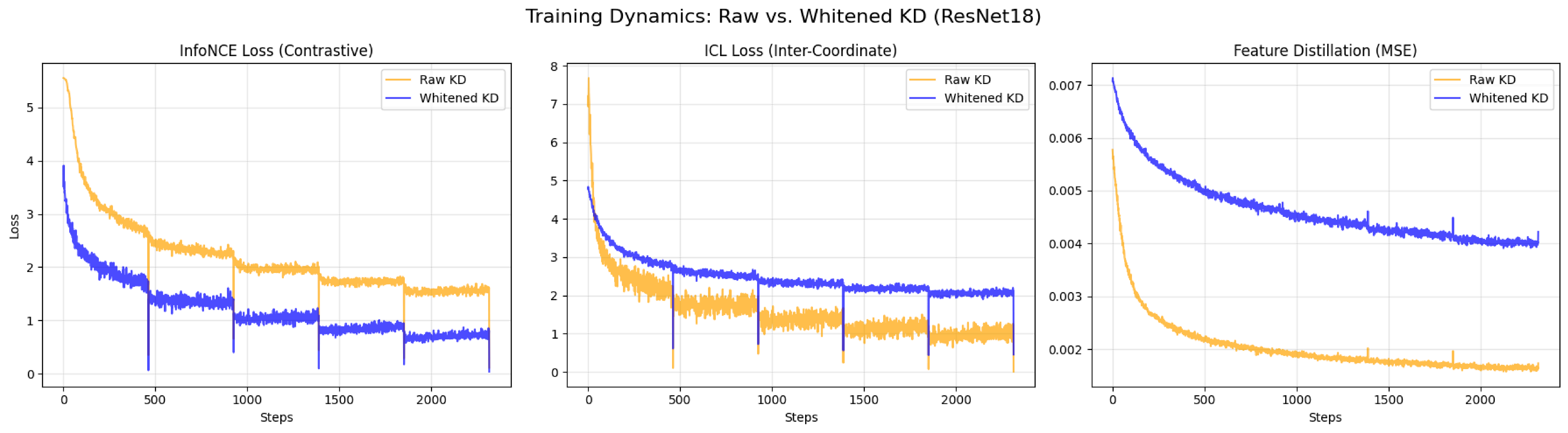
TRAINING CONFIGURATIONS

- **Batch size:** We used a batch size of 128, which was the maximum we could use given compute limitations
- **Learning rate:** We used an AdamW optimizer with a learning rate of 1e-4
- **Loss function:**

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

- λ_InfoNCE=1.0, λ_FD = 2000, λ_ICL=1.0 (standard from CLIP-KD)

RESNET-18 LOSS PLOTS

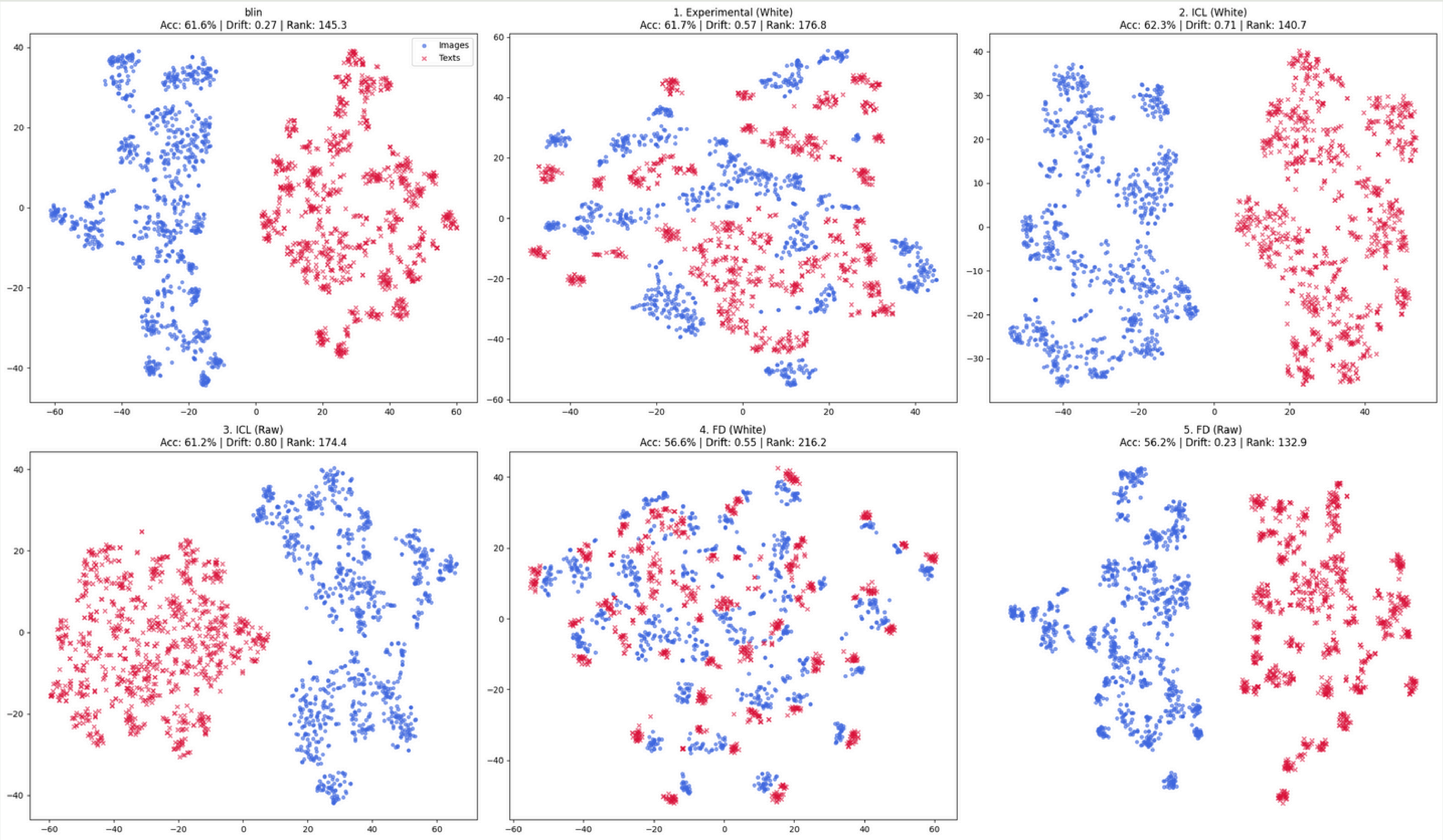


MOBILEVIT-S RESULTS

- Different training dynamics for MobileViT:

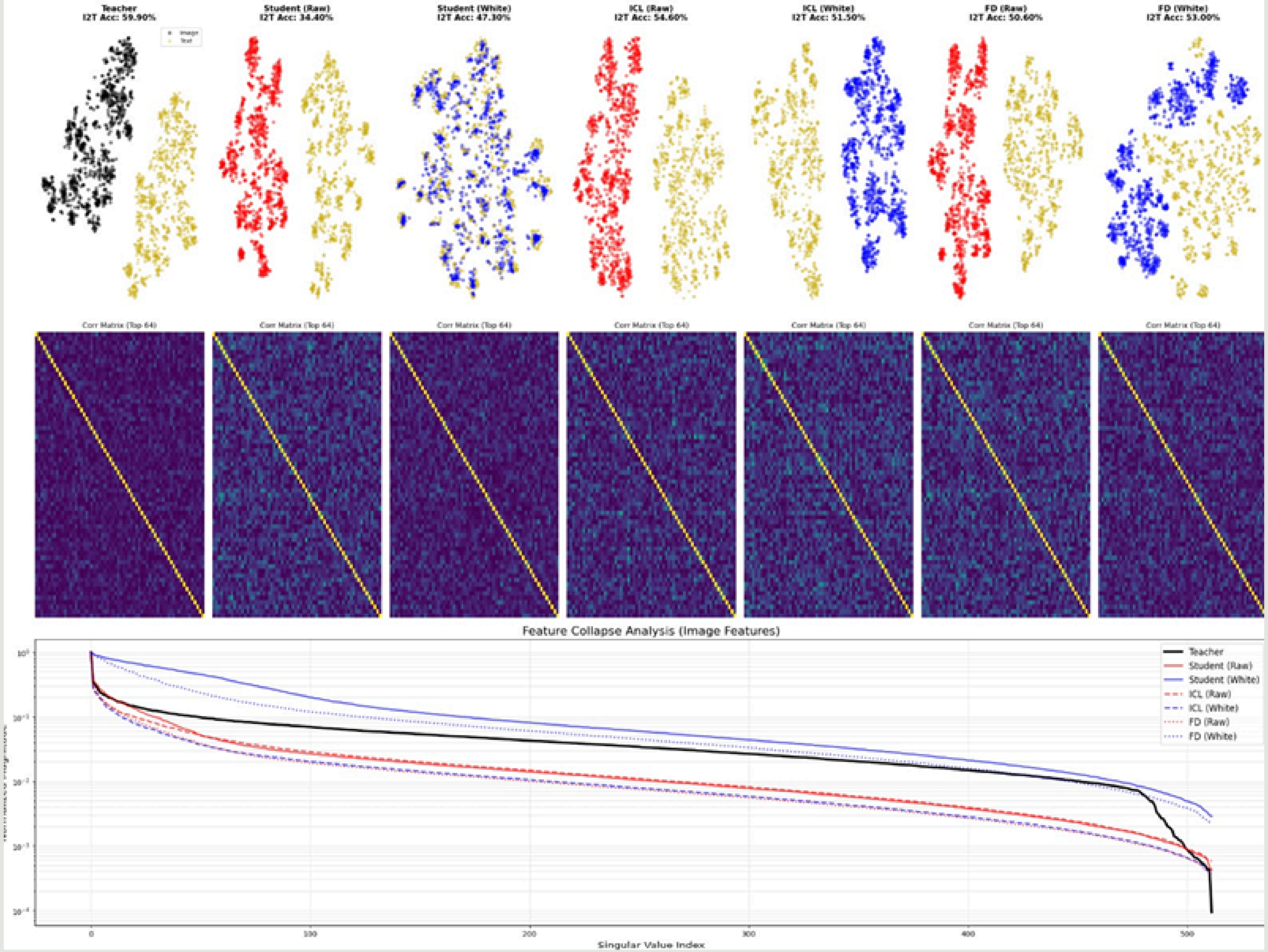
Epoch	Standard KD	Whitened
1	48.56%	44.46%
2	55.92%	53.67%
3	57.31%	59.13%
4	59.63%	61.32%
5	61.60%	61.70%

(a) Training Log



(b) t-SNE plots for MobileViT-S

MobileNetV3 Ablation



REFERENCES

Name: CLIP-KD: An Empirical Study of CLIP Model Distillation
Link: https://arxiv.org/abs/2307.12732

Name: Whitened CLIP as a Likelihood Surrogate of Images and Captions
Link: https://arxiv.org/abs/2505.06934

Name: The Double-Ellipsoid Geometry of CLIP
Link: https://arxiv.org/abs/2411.14517

Thank You

For your attention