The Double-Ellipsoid Geometry of CLIP(ICML 25')

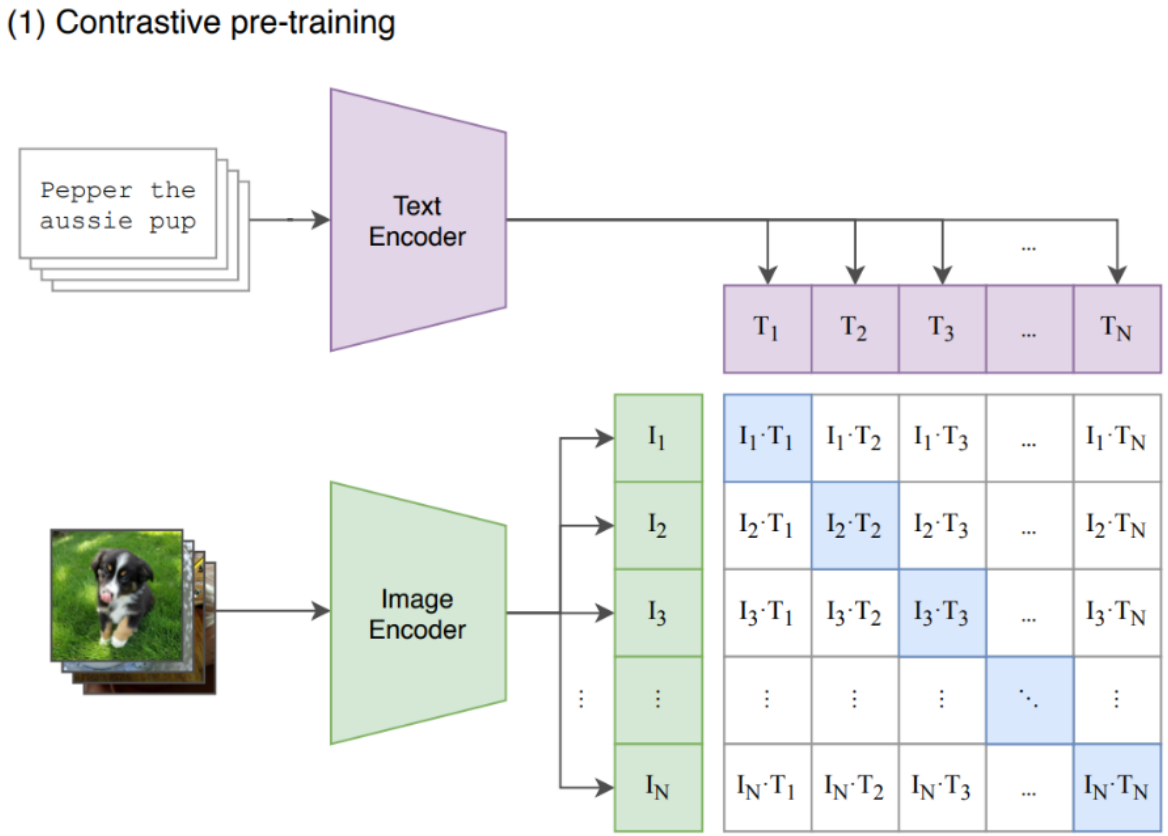
MEIR YOSSEF LEVI

SUPERVISOR: GUY GILBOA

TECHNION
Israel Institute of Technology

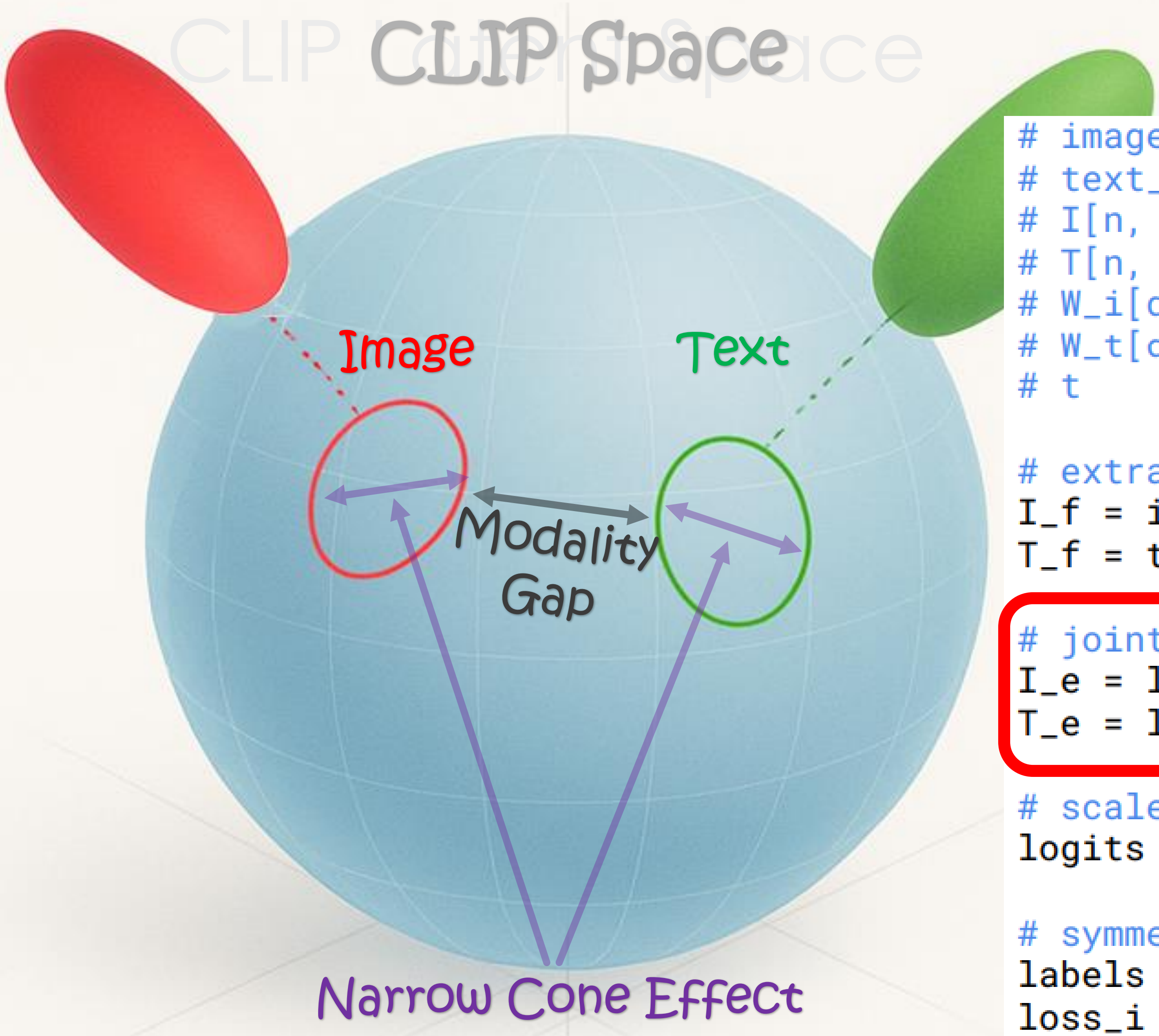
HTTPS://GITHUB.COM/YOSSILEVII100/DOUBLE-ELLIPSOID-CLIP

Contrastive Language-Image Pre-Training (CLIP)



Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *ICML*, 2021.

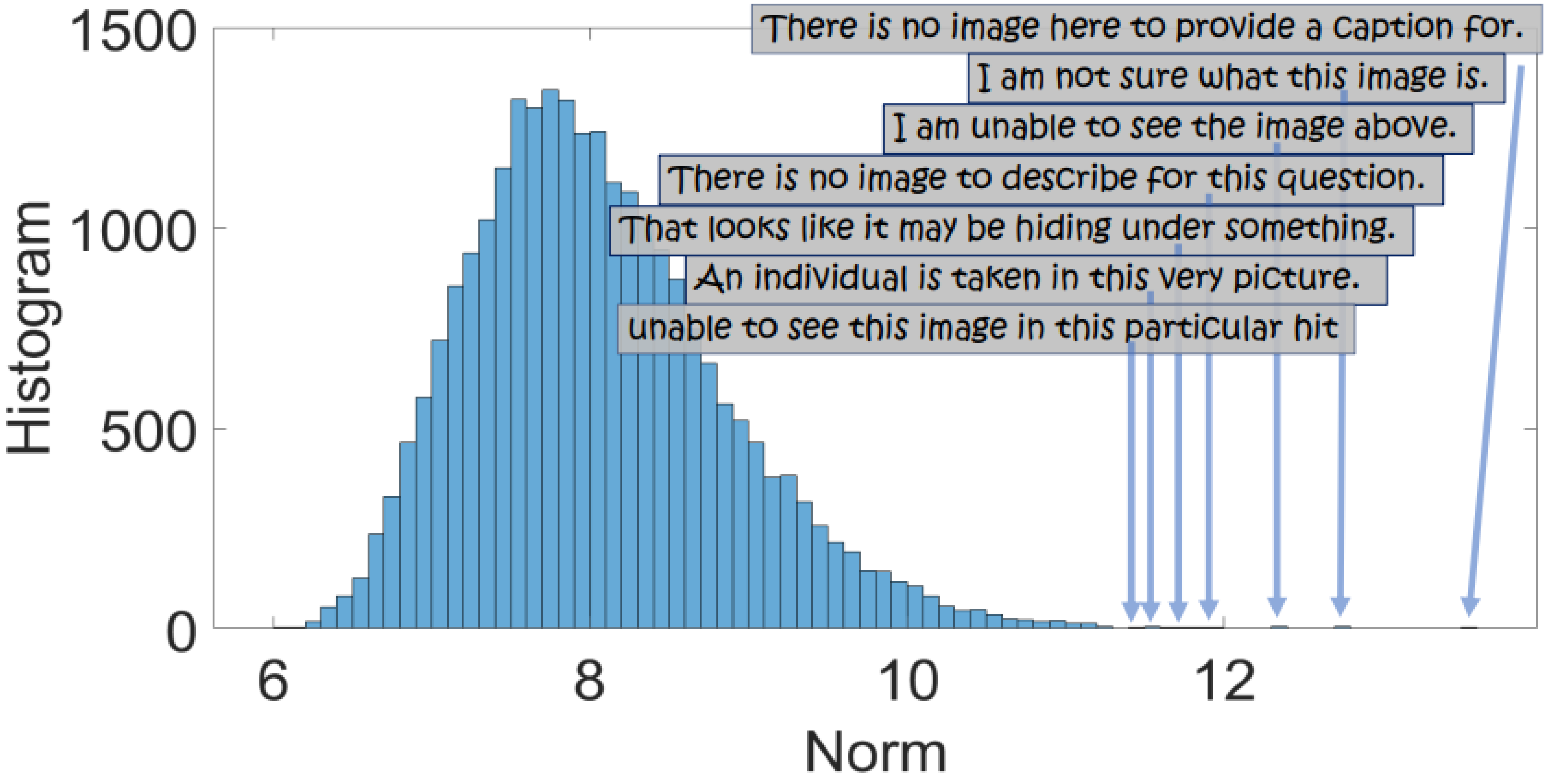
- Widely used
- Latent space geometry is poorly understood




```
# image_encoder - ResNet or Vision Transformer
# text_encoder  - CBOW or Text Transformer
# I[n, h, w, c] - minibatch of aligned images
# T[n, l]       - minibatch of aligned texts
# W_i[d_i, d_e] - learned proj of image to embed
# W_t[d_t, d_e] - learned proj of text to embed
# t             - learned temperature parameter

# extract feature representations of each modality
I_f = image_encoder(I) #[n, d_i]
T_f = text_encoder(T)  #[n, d_t]

# joint multimodal embedding [n, d_e]
I_e = l2_normalize(np.dot(I_f, W_i), axis=1)
T_e = l2_normalize(np.dot(T_f, W_t), axis=1)

# scaled pairwise cosine similarities [n, n]
logits = np.dot(I_e, T_e.T) * np.exp(t)

# symmetric loss function
labels = np.arange(n)
loss_i = cross_entropy_loss(logits, labels, axis=0)
loss_t = cross_entropy_loss(logits, labels, axis=1)
loss   = (loss_i + loss_t)/2
```

Why analyze before normalization?

- Analyze the earliest point possible
- Projection is an information reducing operation
- Norm is actually matters!



Geometric Properties

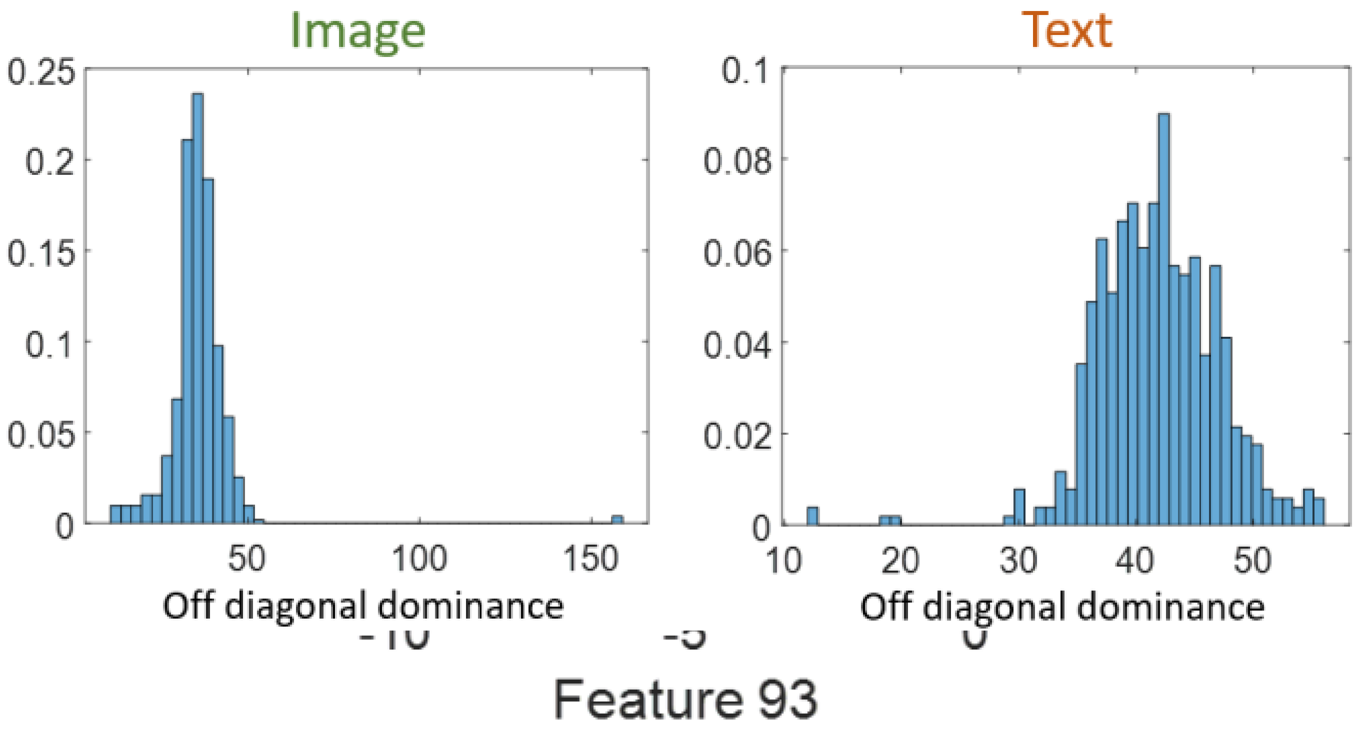
Property 1: Image and text reside on separate subspaces, $\mathcal{X}_i \cap \mathcal{X}_t \approx \emptyset$.

Property 2: The mass of each modality is concentrated within a thin shell, with zero mass near the mean of the distribution.

Property 3: The embedding of both text and image is of an ellipsoid shell.

Property 4: The ellipsoids of both modalities are tilted.

Property 5: The ellipsoids are not centered near the origin.

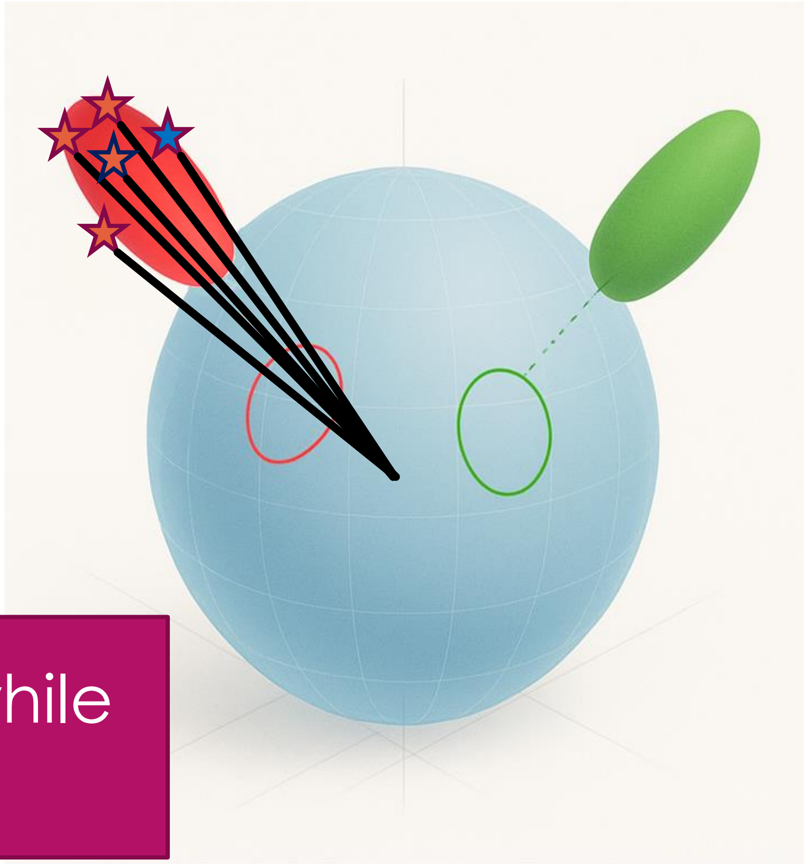


Additional Geometric observation

- Another key observation on CLIP latent space is: ***Conformity***
- Estimate how common a sample is within a given group by:

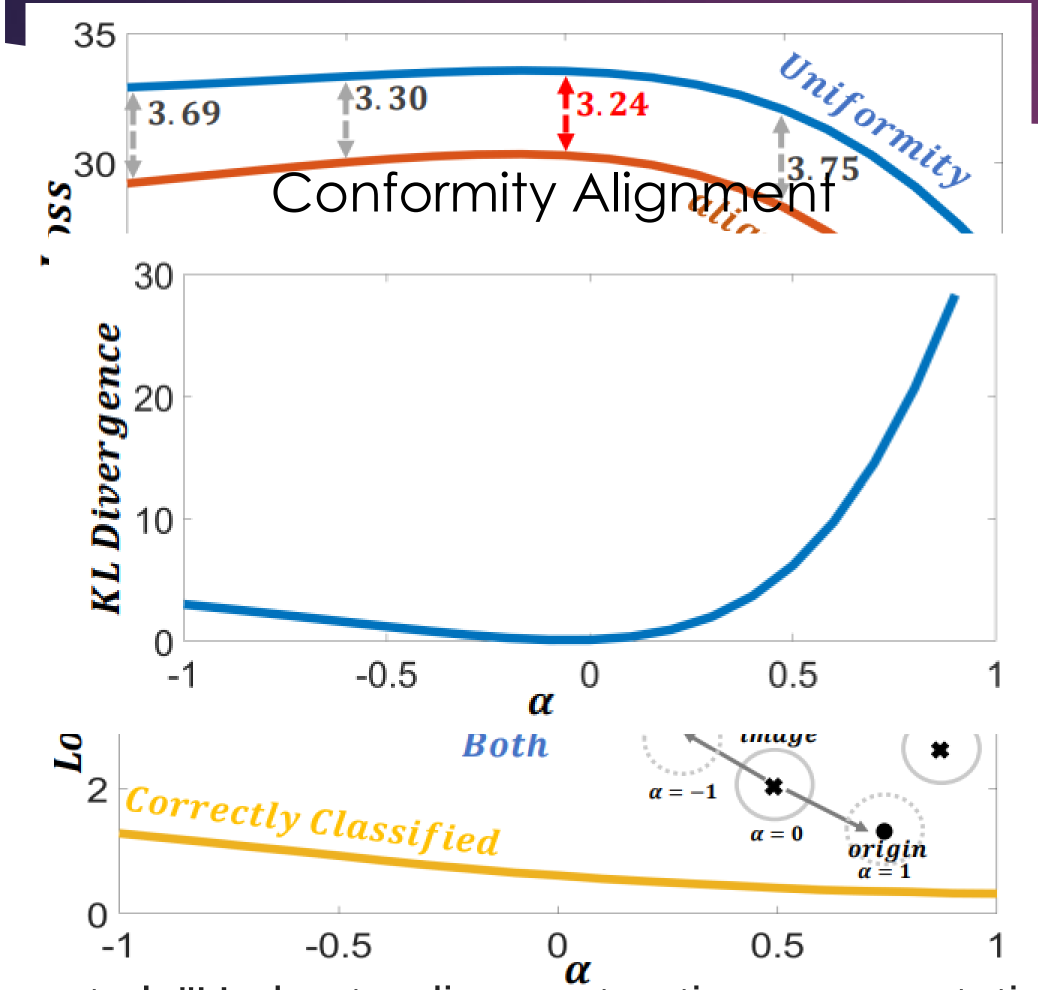
$$C(v^j) = \mathop{\mathbb{E}}_{\substack{v^k \in S \\ j \neq k}} [\cos(v^j, v^k)]$$

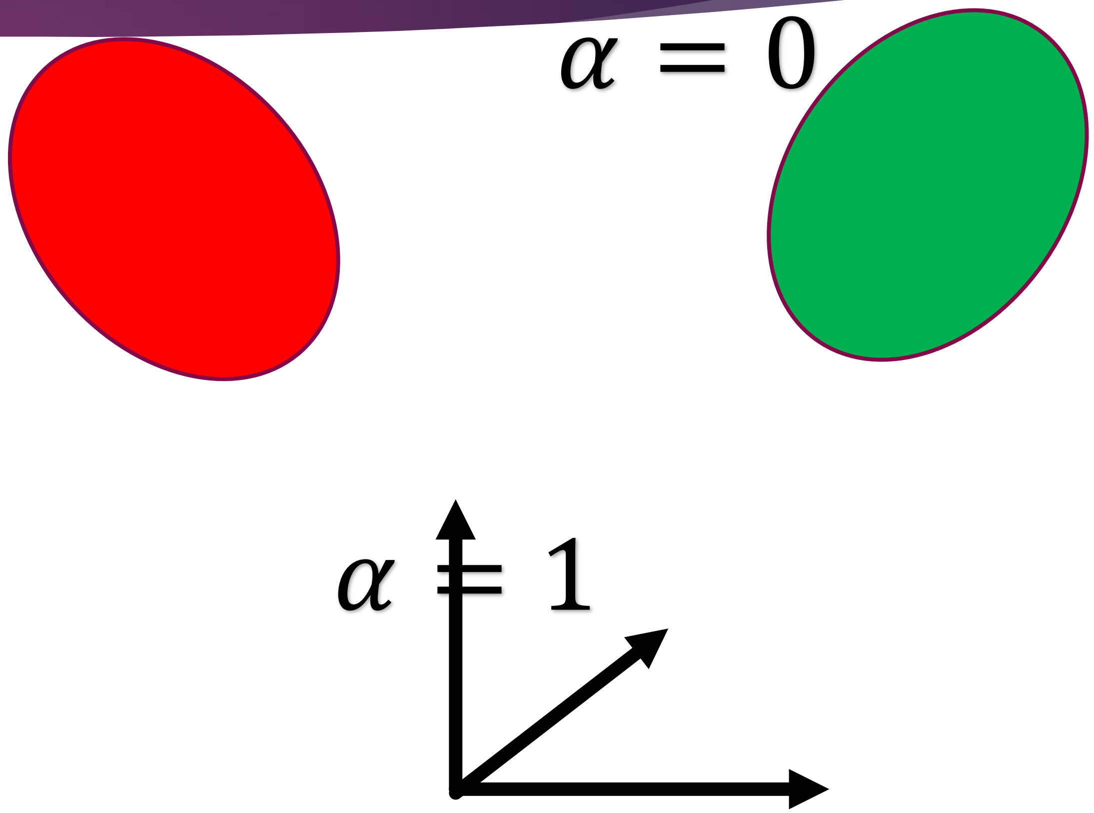
- We prove that this property is proportional to:



common concepts lie near the modality mean, while rare ones are pushed farther away.

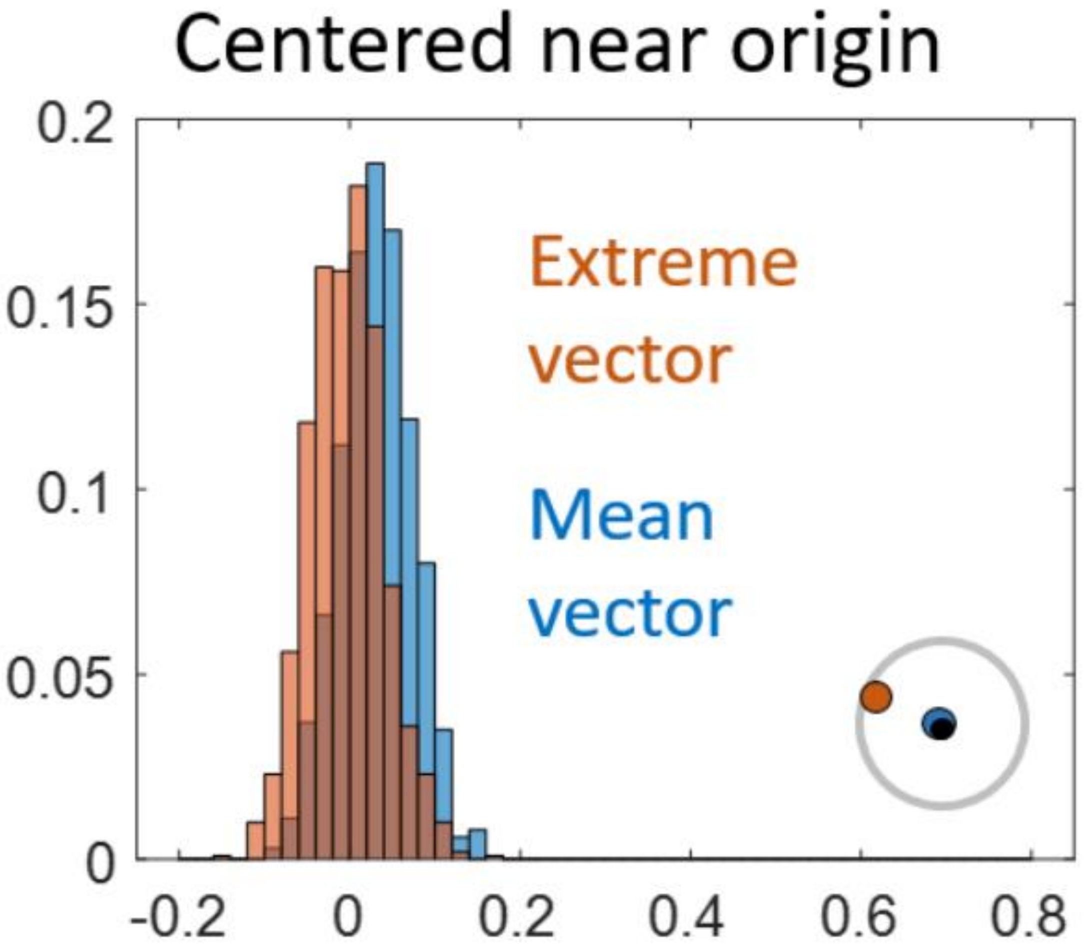
Is non-origin-centered is beneficial?

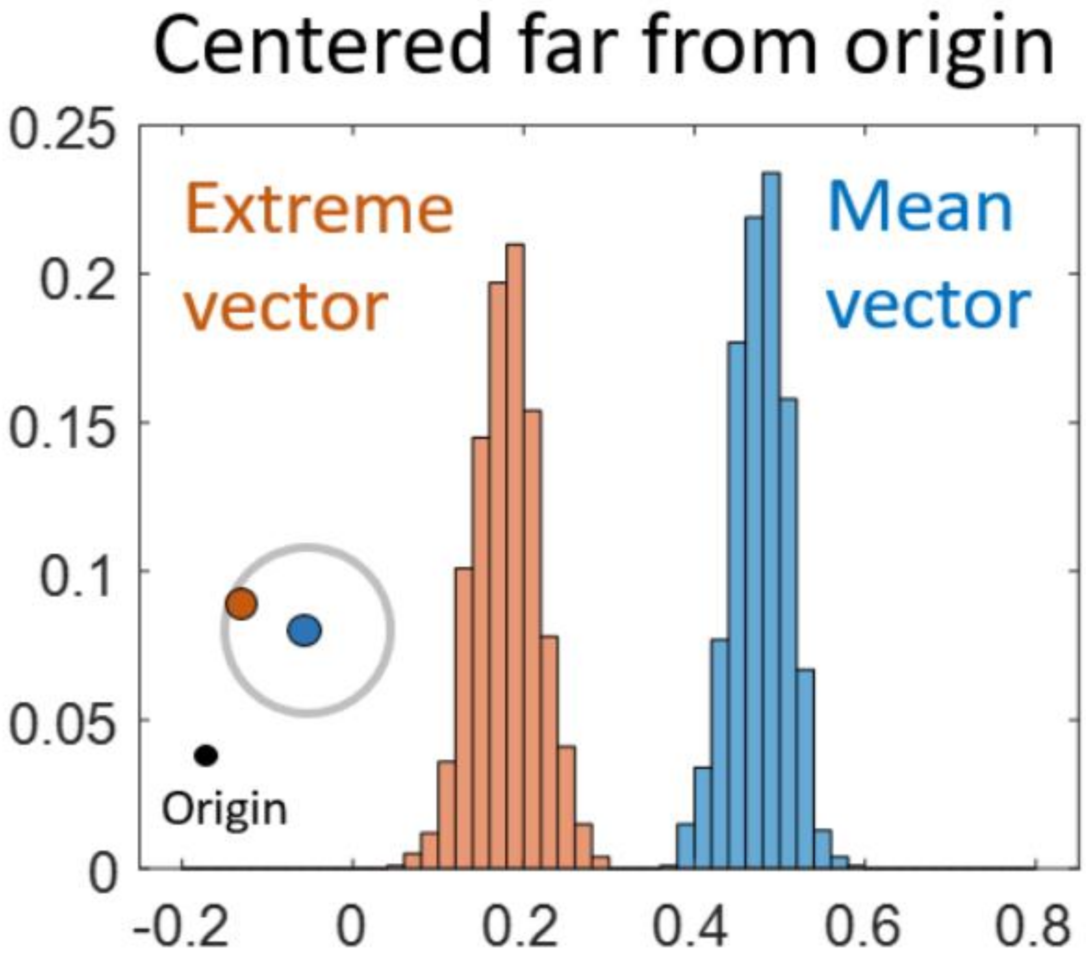




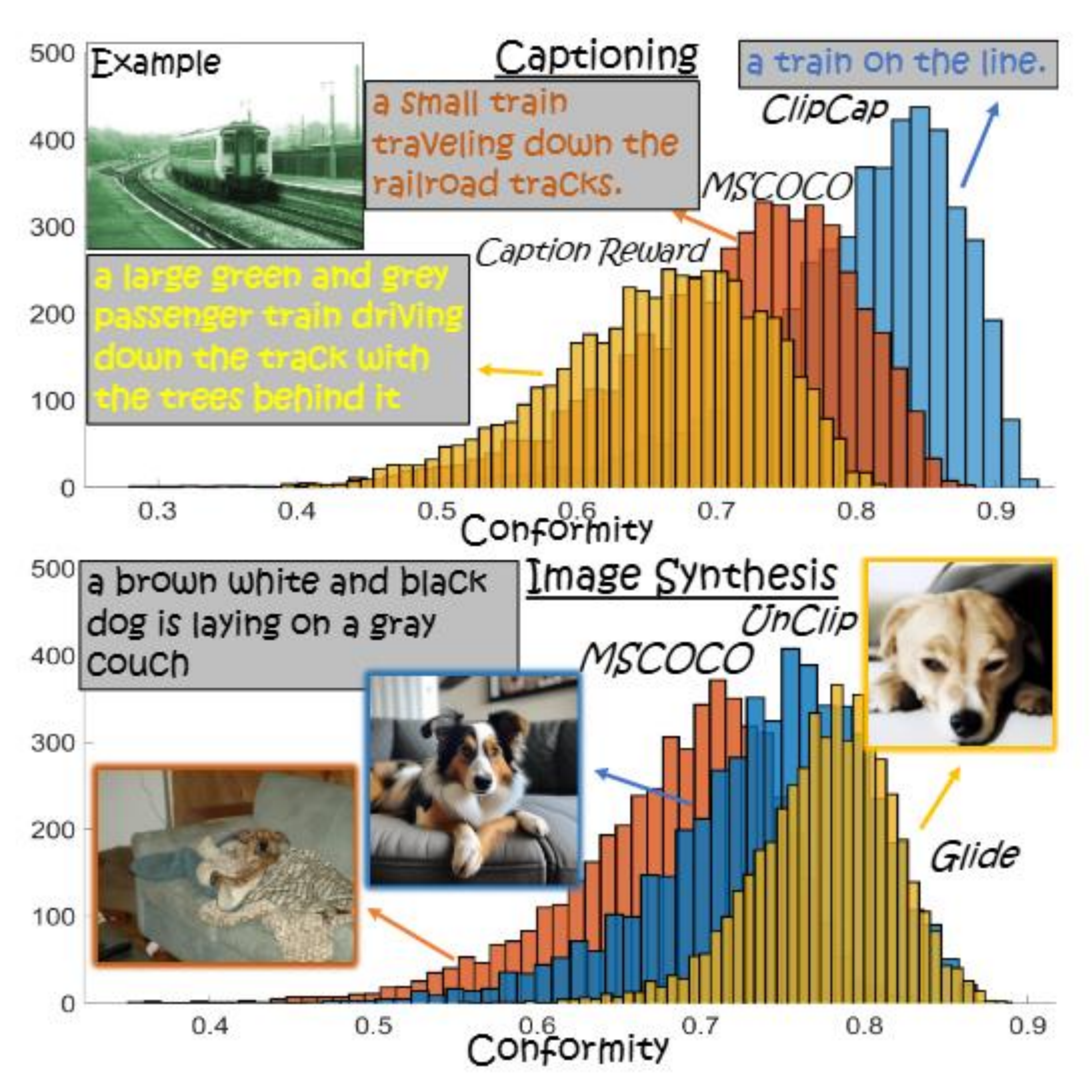
Wang et al. "Understanding contrastive representation learning through alignment and uniformity on the hypersphere." *ICML* 2020.

Why non-origin-centered is beneficial?





Why is it good?



Thank You!