Connect, Collapse, Corrupt: Learning Cross-Modal Tasks with Uni-Modal Data

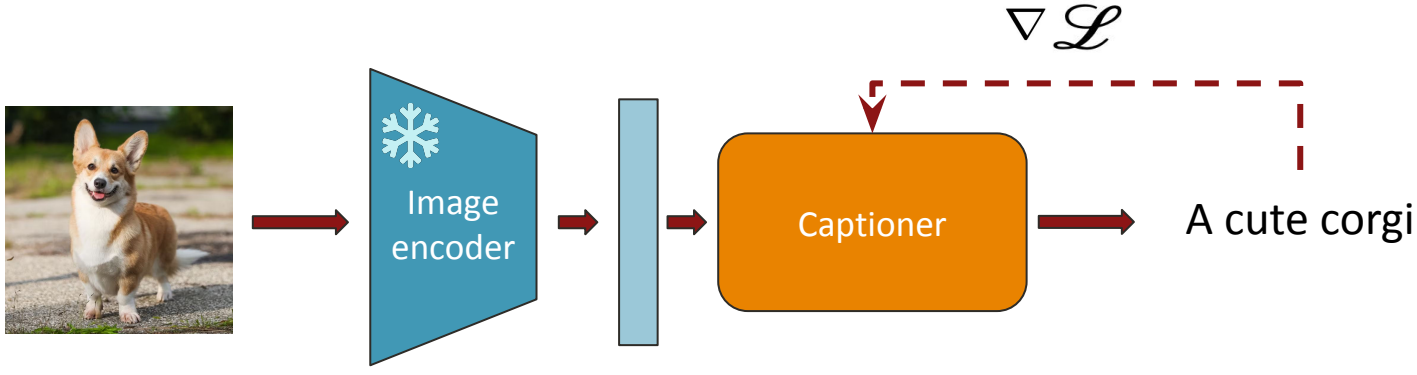
Yuhui Zhang*, Elaine Sui*, Serena Yeung-Levy

ICLR 2024

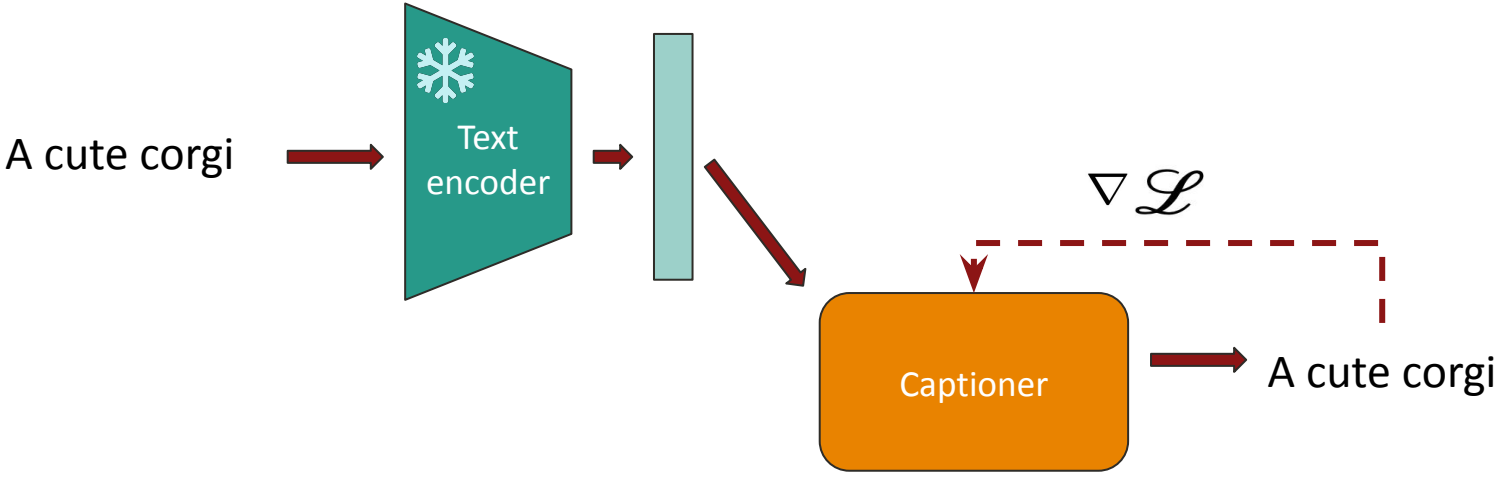
* equal contribution

Stanford University

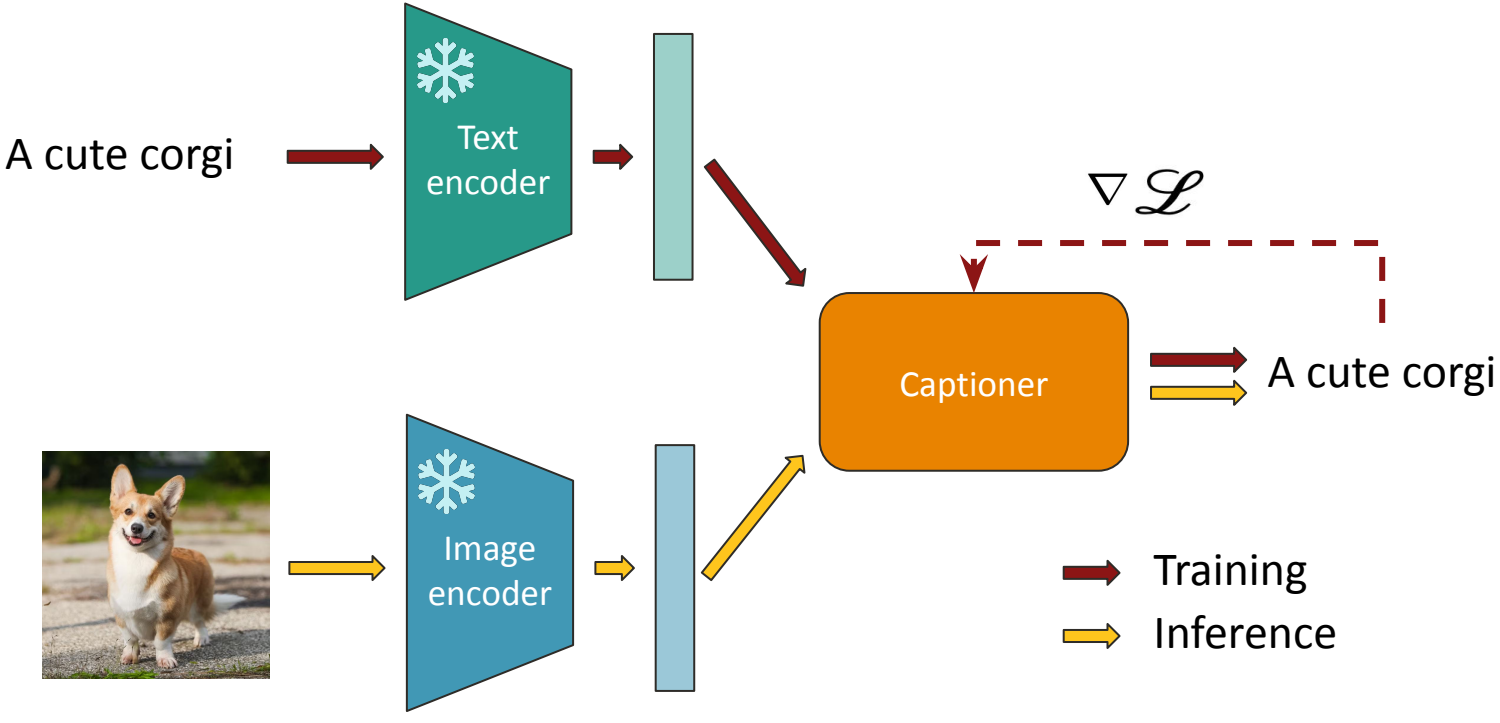
Traditional Image Captioning Training



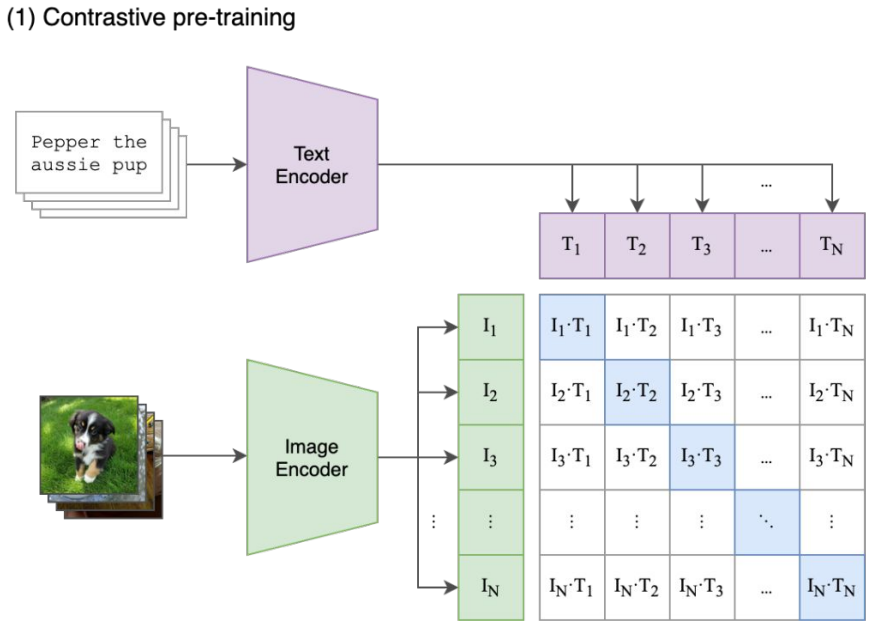
Text-Only Image Captioning Training

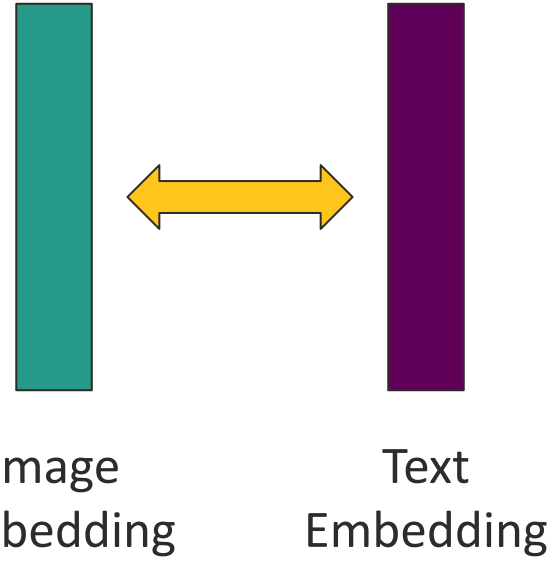


Text-Only Image Captioning Training



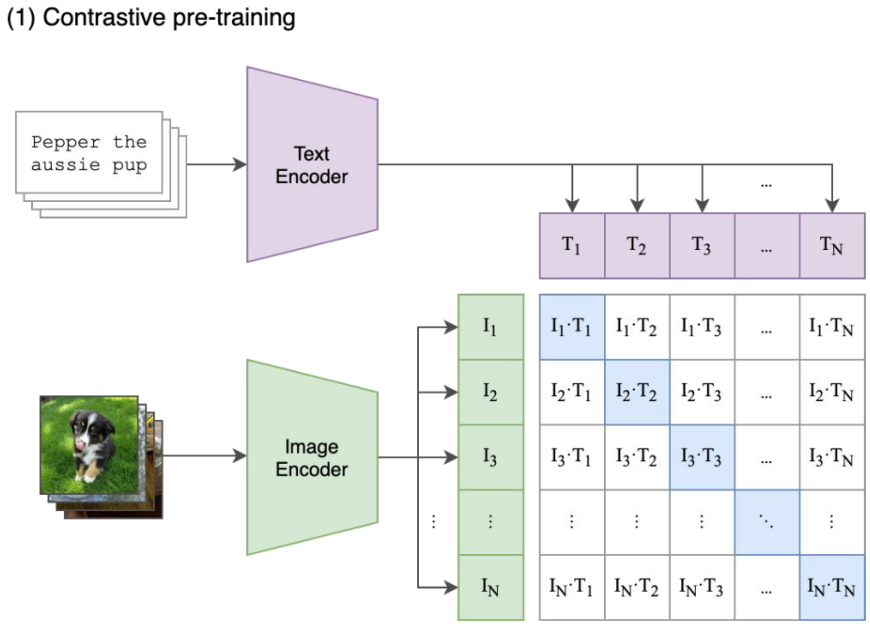
Large-Scale Multi-Modal Contrastive Learning

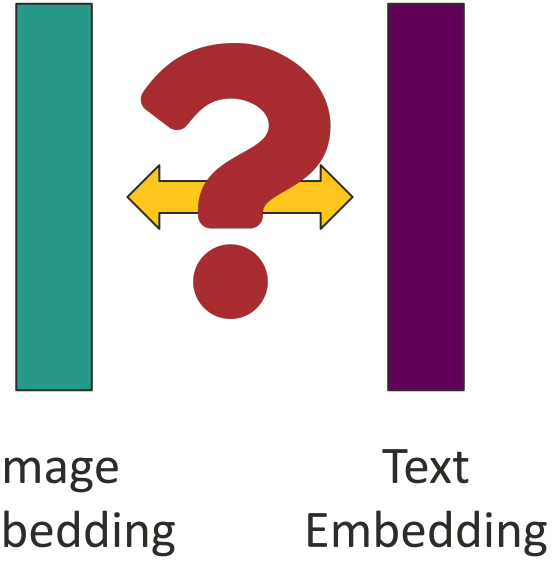




(Radford et al., 2021)

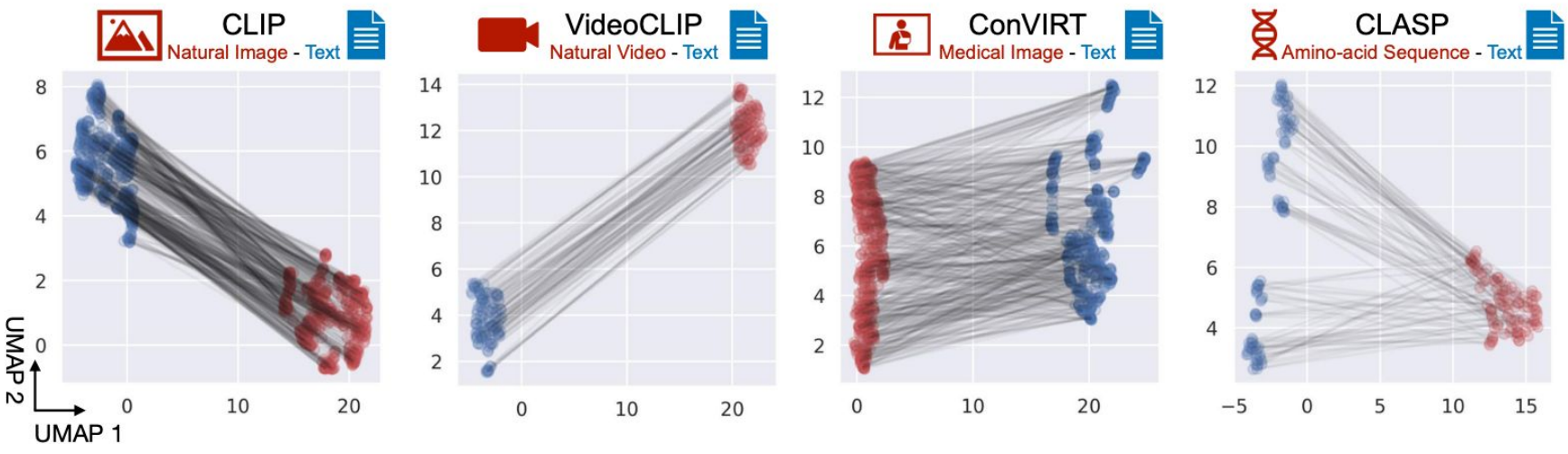
Large-Scale Multi-Modal Contrastive Learning





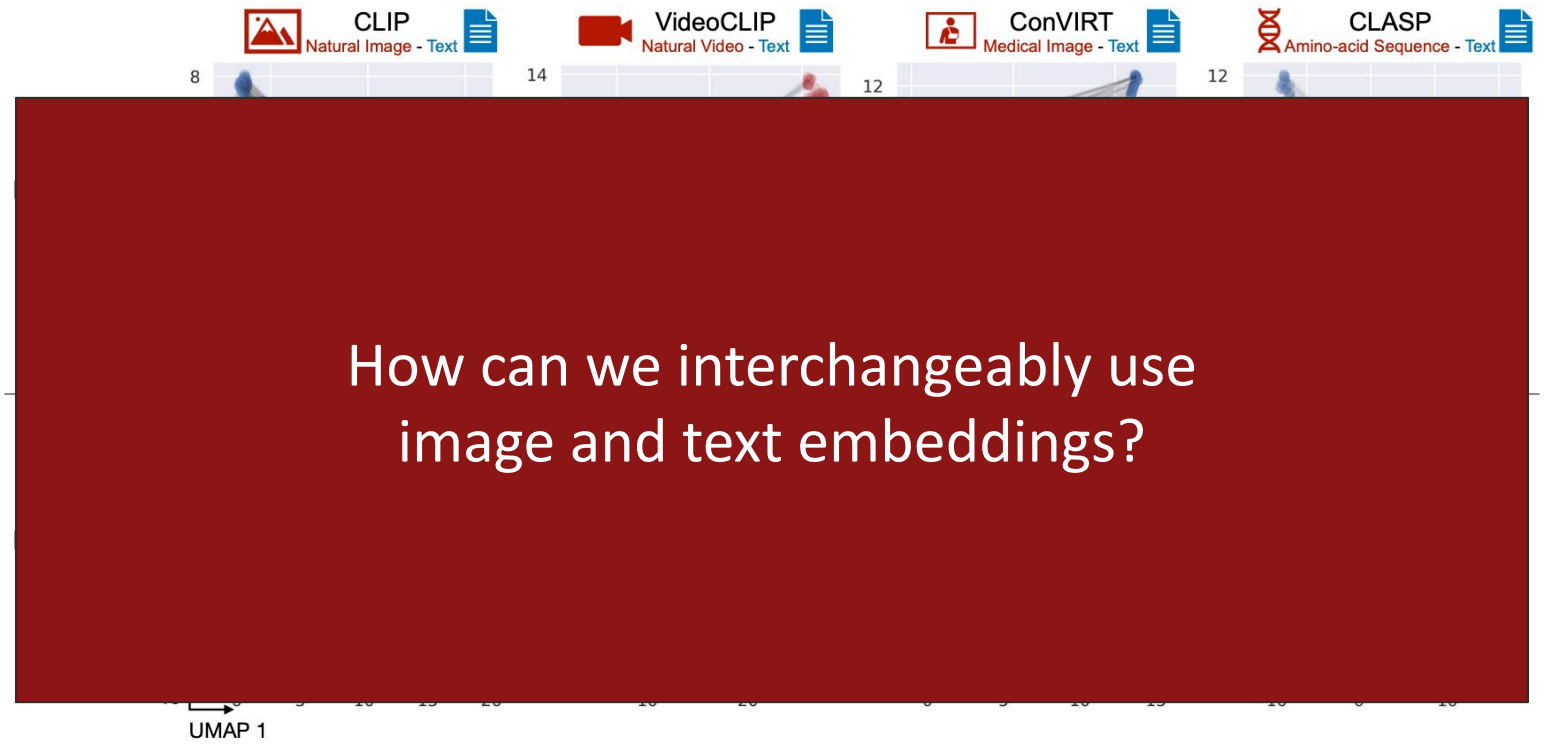
(Radford et al., 2021)

Modality Gap



(Liang et al., 2022)

Modality Gap



(Liang et al., 2022)

Representation Space Geometry

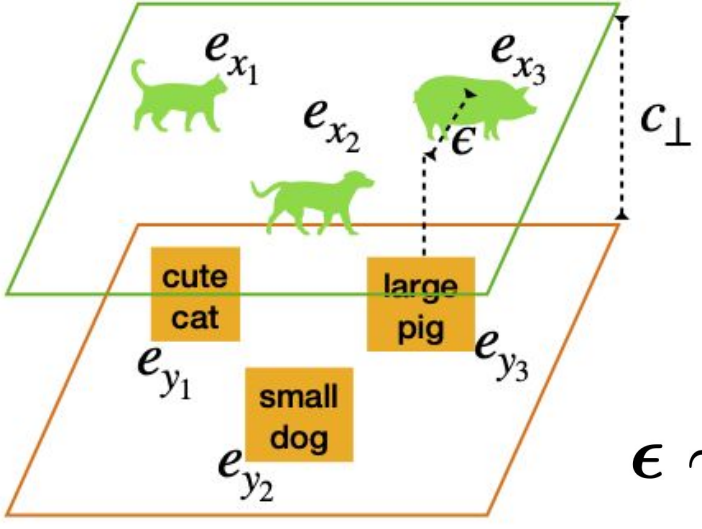
$$e_x - e_y = c_\perp + \epsilon$$

Empirically Verified

Modality 1 Embedding

$e_x - e_y = c_\perp + \epsilon$

Modality 2 Embedding



$c_\perp$ Modality Gap

$\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})$ Alignment Noise

Theoretical Explanation of Geometry

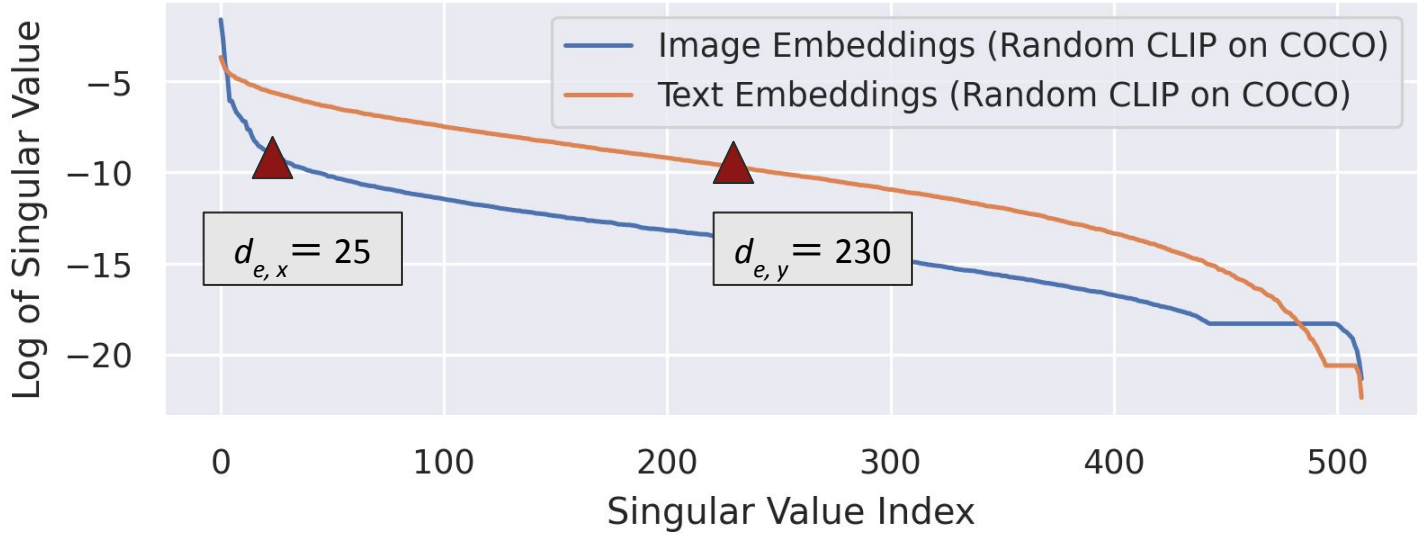
$\boldsymbol{c}_{\perp}$ Modality Gap

- Dimensional Collapse of Model Initialization
- Gradients of Contrastive Optimization

$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})$ Alignment Noise ← Stable Region Controlled by Temperature

Modality Gap Emerges at Initialization

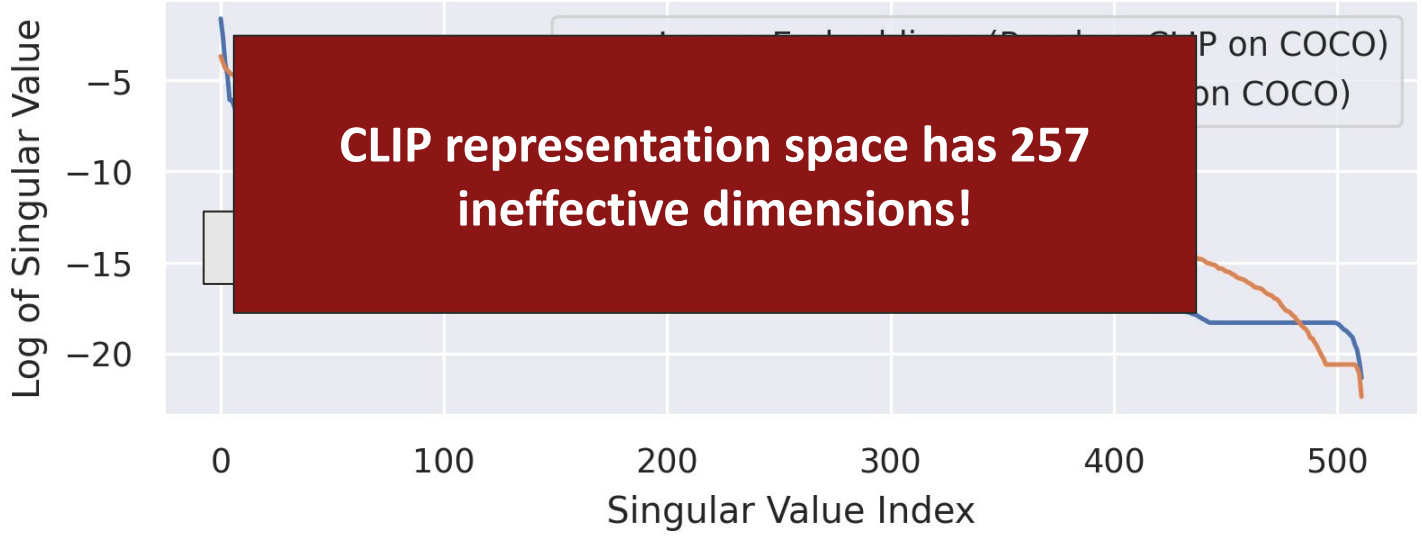
Dimensional Collapse of the Representation Space



Effective dimensions are small given a randomly initialized 512-d CLIP embeddings.

Modality Gap Emerges at Initialization

Dimensional Collapse of the Representation Space

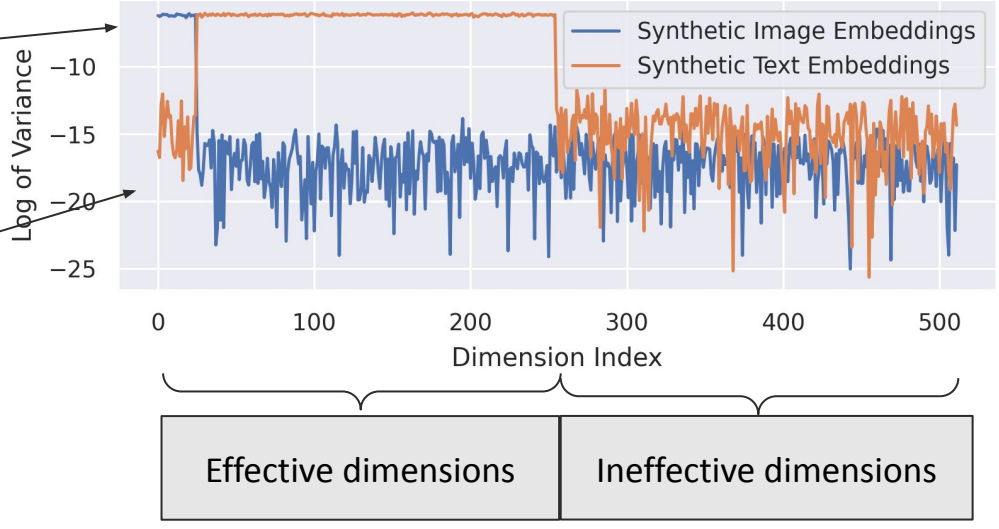


Effective dimensions are small given a randomly initialized 512-d CLIP embeddings.

Modality Gap is Preserved after Optimization

Dimensions sampled from standard Gaussian

Dimensions initialized to random constant

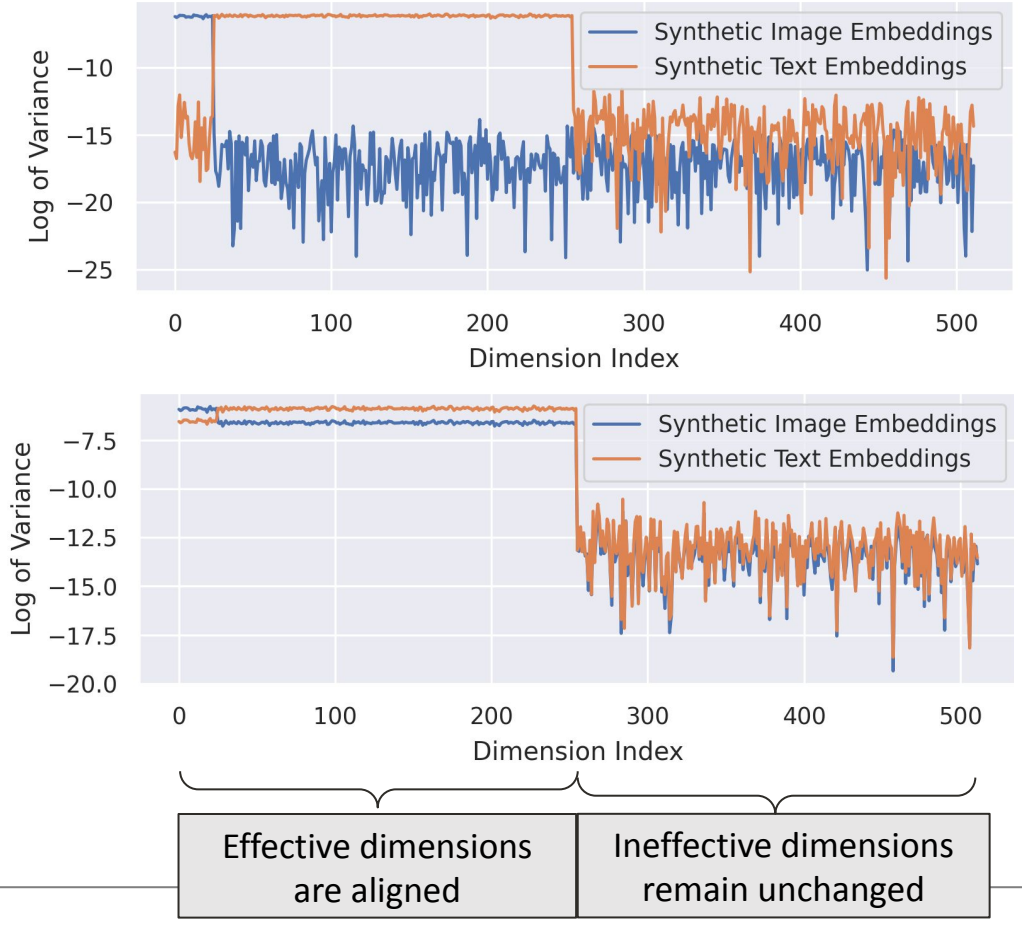


Before optimization

Modality Gap is Preserved after Optimization

Before optimization

After optimization



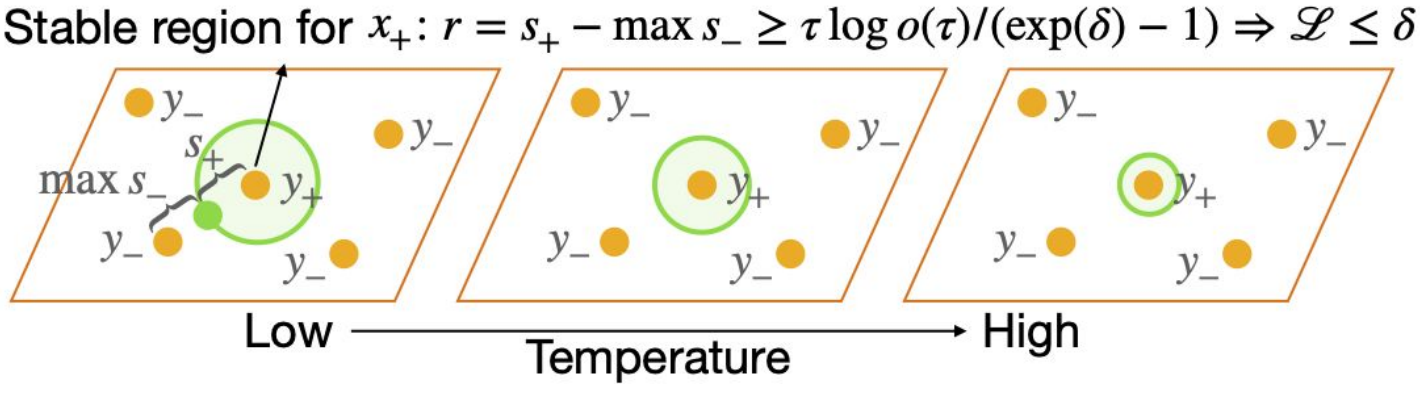
See gradient analysis in paper.

Alignment Noise

Given the contrastive loss $\mathcal{L}_i = -\log \frac{\exp(\boldsymbol{e}_{x_i} \cdot \boldsymbol{e}_{y_i}/\tau)}{\sum_{j=1}^{n} \exp(\boldsymbol{e}_{x_i} \cdot \boldsymbol{e}_{y_j}/\tau)}$,

the stable region r is a function of temperature τ.

Within the stable region, the optimization halts.



C^3: Connect, Collapse, Corrupt

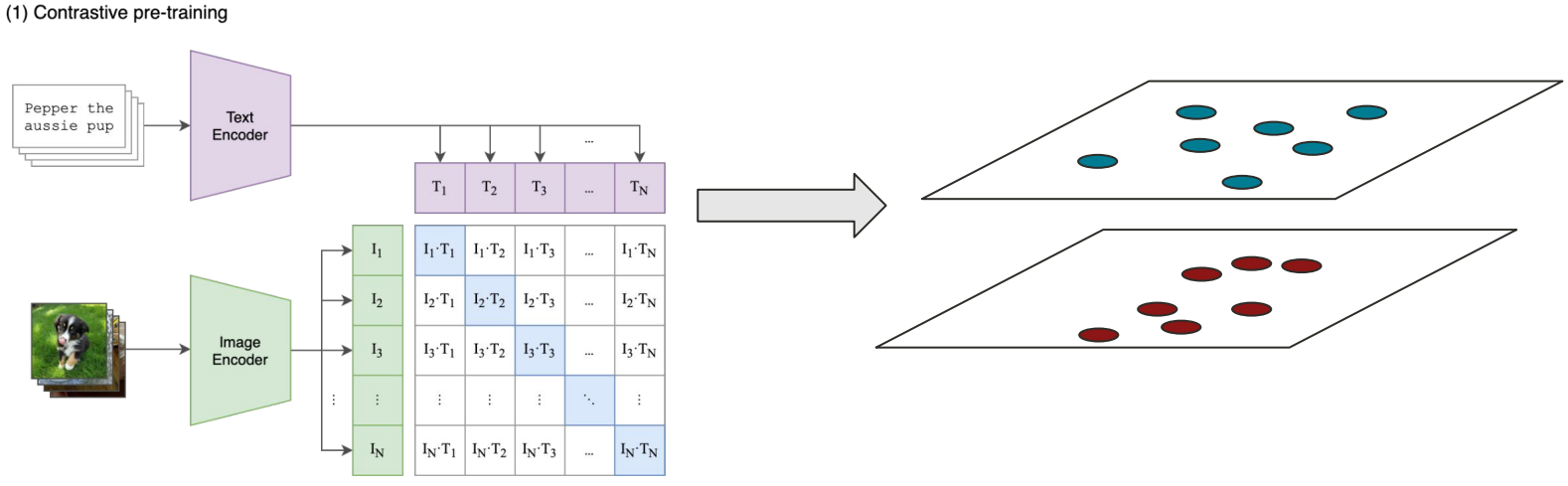
Connect: Relate concepts with multi-modal contrastive learning.

Collapse: Close the modality gap by subtracting the embedding mean from each modality.

Corrupt: Add noise to the embeddings as regularization during training.

1. Connect

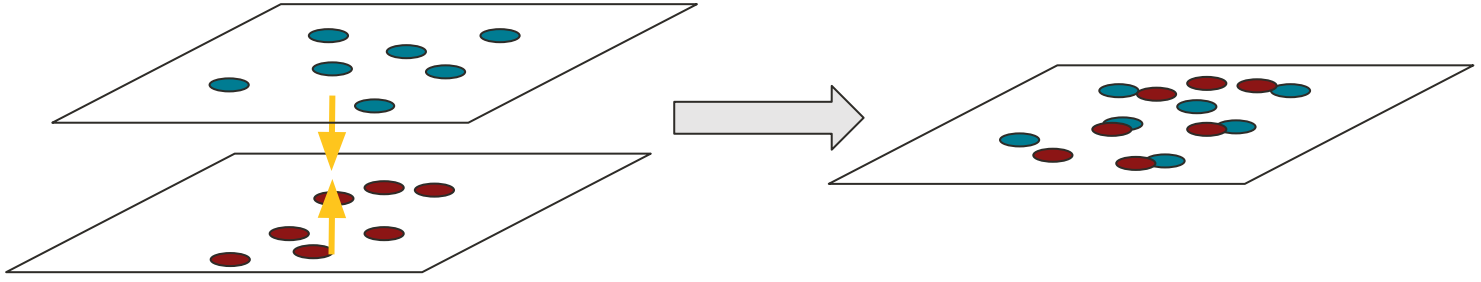
Relate concepts with multi-modal contrastive learning.



(Radford et al., 2021; Liang et al., 2022)

2. Collapse

Close the modality gap by subtracting the embedding mean from each modality.



3. Corrupt

Add noise to the embeddings as regularization during training.

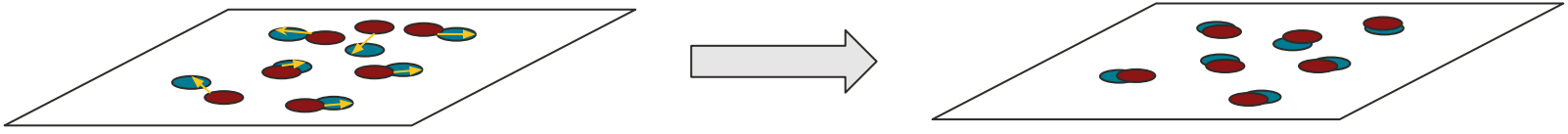
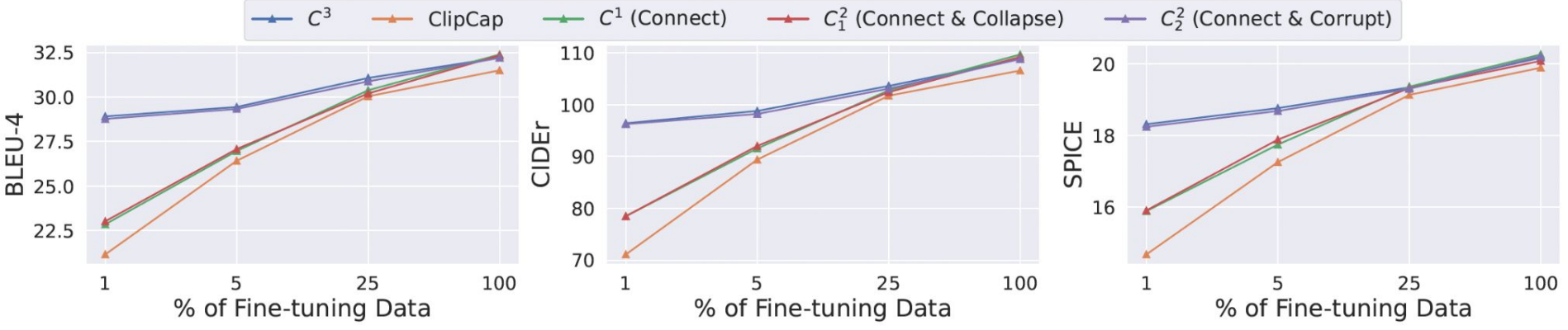


Image Captioning on MSCOCO

Method	Conn.	Coll.	Corr.	BLEU-1↑	BLEU-4↑	METEOR↑	ROUGE-L↑	CIDEr↑	SPICE↑
Baselines									
ZeroCap (2022)	✗	✗	✗	49.8	7.0	15.4	31.8	34.5	-
MAGIC (2022)	✗	✗	✗	56.8	12.9	17.4	39.9	49.3	11.3
ESPER (2022)	✗	✗	✗	-	21.9	21.9	-	78.2	-
CLIPRe (2023)	✓	✗	✗	-	4.6	13.3	-	25.6	9.2
DeCap (2023)	✓	✗	✗	-	8.9	17.5	-	50.6	13.1
WS-ClipCap (2023)	✓	✗	✗	50.3	9.6	15.2	37.5	33.7	8.6
WS-ClipCap-Multi (2023)	✓	✗	✓	65.5	22.1	22.2	48.0	74.6	14.9
CapDec (2022)	✓	✗	✓	69.2	26.4	**25.1**	51.8	91.8	-
Ours									
C^1	✓	✗	✗	28.1	2.4	12.2	25.4	13.0	6.8
C_1^2	✓	✓	✗	44.4	6.1	15.5	33.6	25.2	9.2
C_2^2	✓	✗	✓	69.0	25.5	24.3	50.8	87.6	17.6
C^3	✓	✓	✓	**71.0**$\pm$0.1	**27.7**$\pm$0.1	25.0$\pm$0.0	**52.0**$\pm$0.0	**93.3**$\pm$0.3	**18.3**$\pm$0.1

Zero-shot image captioning results.
Train: caption-to-caption. Test: image-to-caption.

Image Captioning in the Low-Data Regime



Semi-supervised image captioning results.
Train: caption-to-caption. Fine-tune: image-to-caption. Test: image-to-caption.

Image Captioning Qualitative Results

C^1: *Dog sniffing the back of the black and white dog in the cobblestone.*

C_1^2: *Dog behind a man's ear in the street.*

C_2^2: *A black dog looking at a dog on the sidewalk.*

C^3: *A dog looking at a black cat on the street.*

Ground truth: A black cat is scared by a large dog.

Connect

Connect + Collapse

Connect + Corrupt

Connect + Collapse + Corrupt

C^1: *A red and white bus is en route to the International Space Station.*

C_1^2: *A PERSON IS DRIVING THE SEA WAVE*

C_2^2: *A group of people in a boat with a large white boat in the background.*

C^3: *A group of people on a boat in the water.*

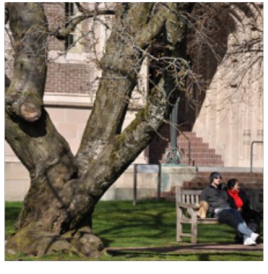
Ground truth: A boat in the water filled with people

Connect

Connect + Collapse

Connect + Corrupt

Connect + Collapse + Corrupt



C^1: *A person and their luggage, resting in the courtyard of the University of Southern California, observing the Eiffel Tower.*

C_1^2: *A person on a bench in the midst of the street to take a walk with their library.*

C_2^2: *A man sitting on a bench with a book.*

C^3: *A man and woman sitting on a park bench.*

Ground truth: A man and woman sitting on a park bench under a huge tree

Connect

Connect + Collapse

Connect + Corrupt

Connect + Collapse + Corrupt

Text-to-Image Generation on MSCOCO

Language-only

Method	Conn.	Coll.	Corr.	FID↓	IS↑
Baselines					
DALL-E (2021)	✗	✗	✗	27.5	17.9
CogView (2021)	✗	✗	✗	27.1	18.2
LAFITE$_G$ (2022c)	✓	✗	✓	20.9	24.9
Ours					
C^1	✓	✗	✗	29.8	22.4
C_1^2	✓	✓	✗	21.7	24.4
C_2^2	✓	✗	✓	19.8	25.5
C^3	✓	✓	✓	**19.6**	**26.0**

Zero-shot text-to-image generation results.
Train: image-to-image. Test: text-to-image.

Text-to-Image Generation Qualitative Results

A small town street with old buildings

Ground Truth

$LAFITE_G$

C^1

C_1^2

C_2^2

C^3

A street scene with a double-decker bus on the road.

Ground Truth

$LAFITE_G$

C^1

Connect

C_1^2

Connect + Collapse

C_2^2

Connect + Corrupt

C^3

Connect + Collapse + Corrupt

Thank You