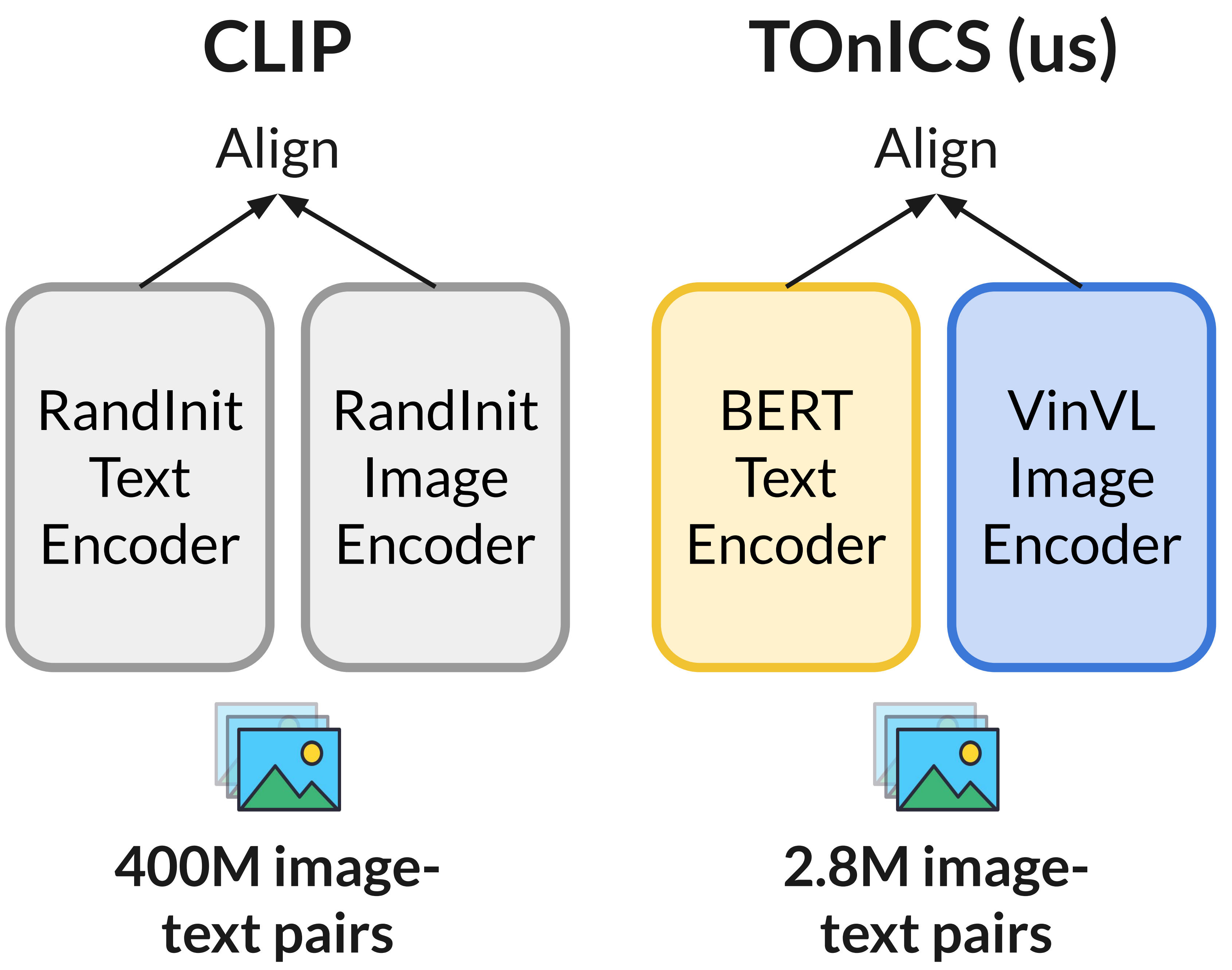
Curriculum Learning for Data-Efficient Vision-Language Alignment

Tejas Srinivasan, Xiang Ren, Jesse Thomason

Introduction

Goal: Efficiently align language and vision representations to each other



CLIP

Align

RandInit Text Encoder

RandInit Image Encoder

400M image-text pairs

TOnICS (us)

Align

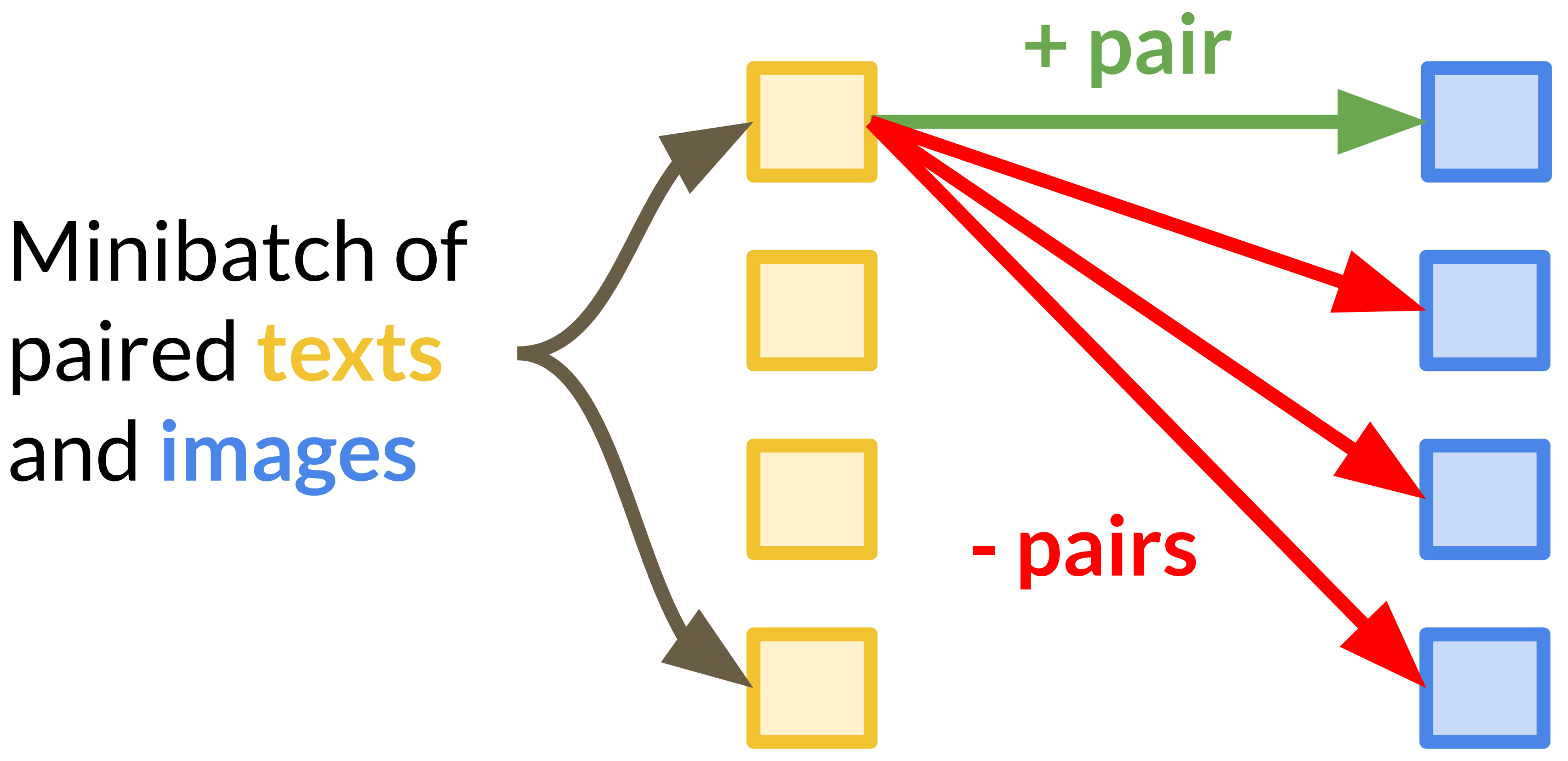
BERT Text Encoder

VinVL Image Encoder

2.8M image-text pairs

Contrastive Image-Text Alignment

Training Objective: Cross-entropy loss of correctly matching text to corresponding image



Minibatch of paired texts and images

+ pair

- pairs

Random Minibatches: Easy contrastive task

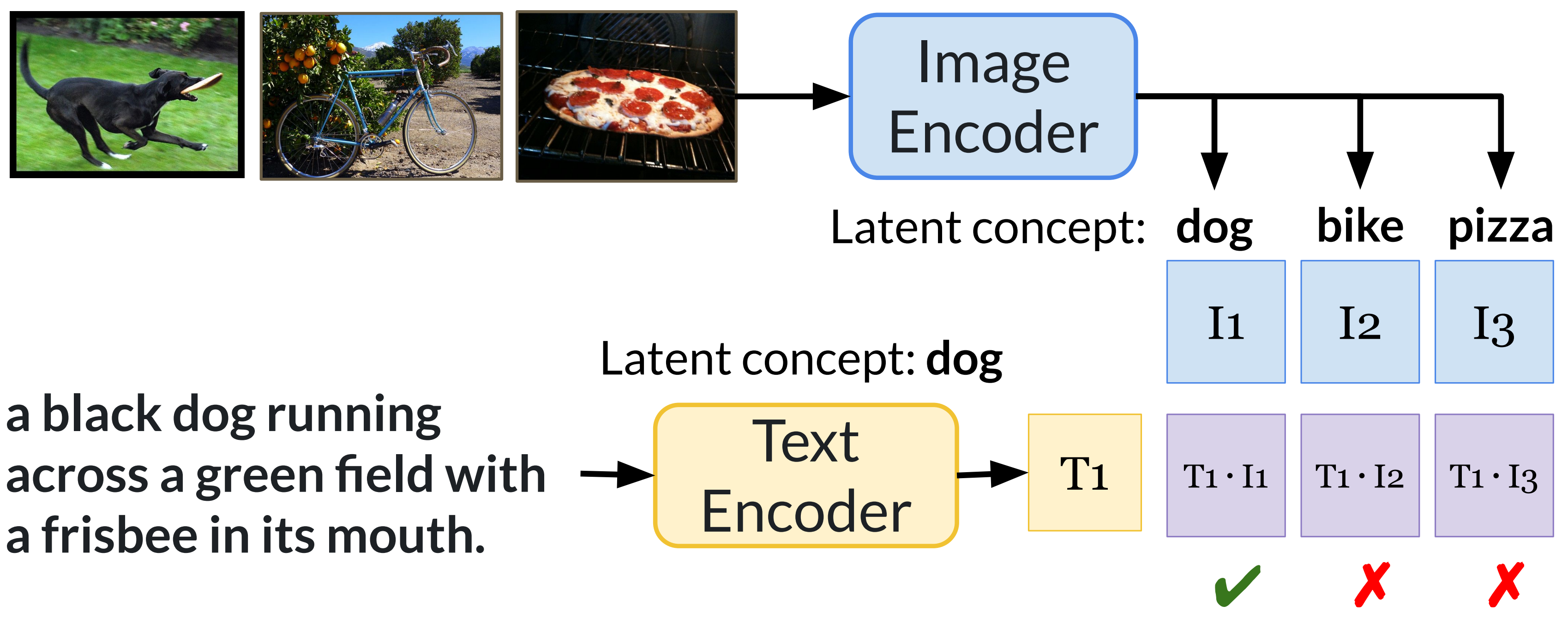


Image Encoder

Latent concept: **dog** **bike** **pizza**

I1 I2 I3

Latent concept: **dog**

a black dog running across a green field with a frisbee in its mouth.

Text Encoder

T1

T1 · I1 T1 · I2 T1 · I3

✔ ✗ ✗

Images contain same object: Harder task

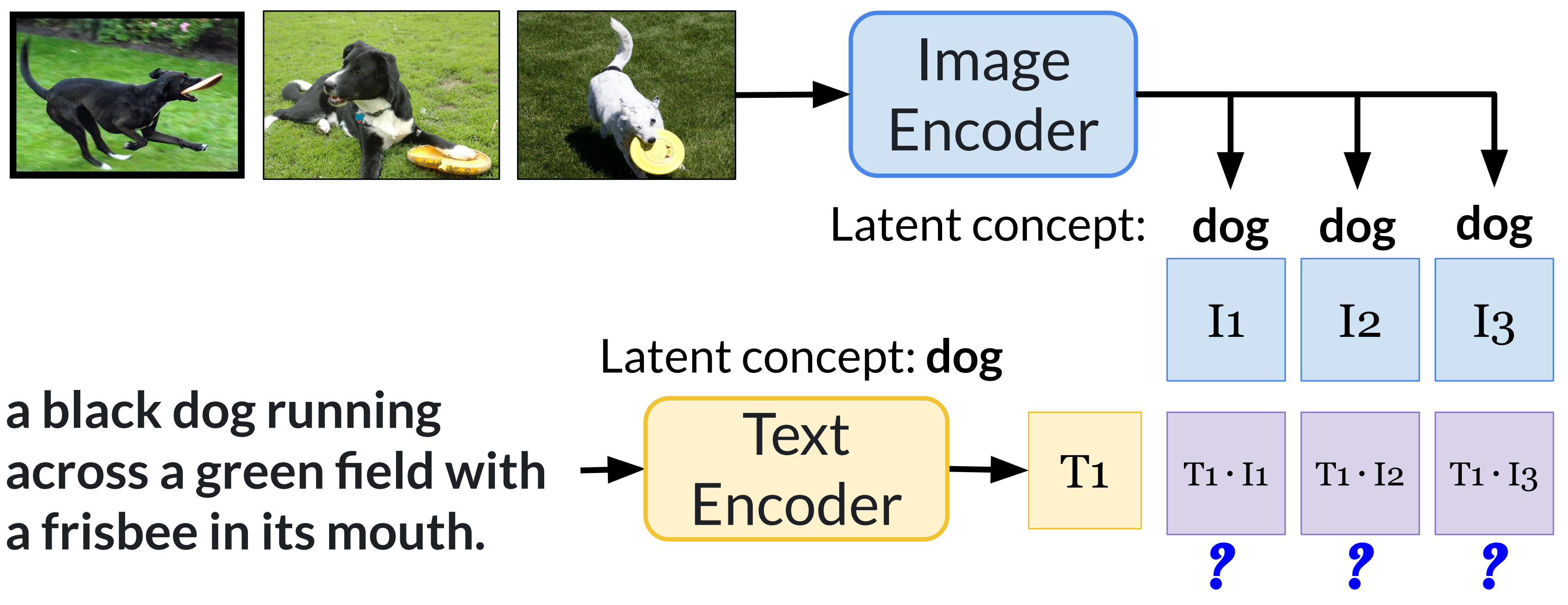


Image Encoder

Latent concept: **dog** **dog** **dog**

I1 I2 I3

Latent concept: **dog**

a black dog running across a green field with a frisbee in its mouth.

Text Encoder

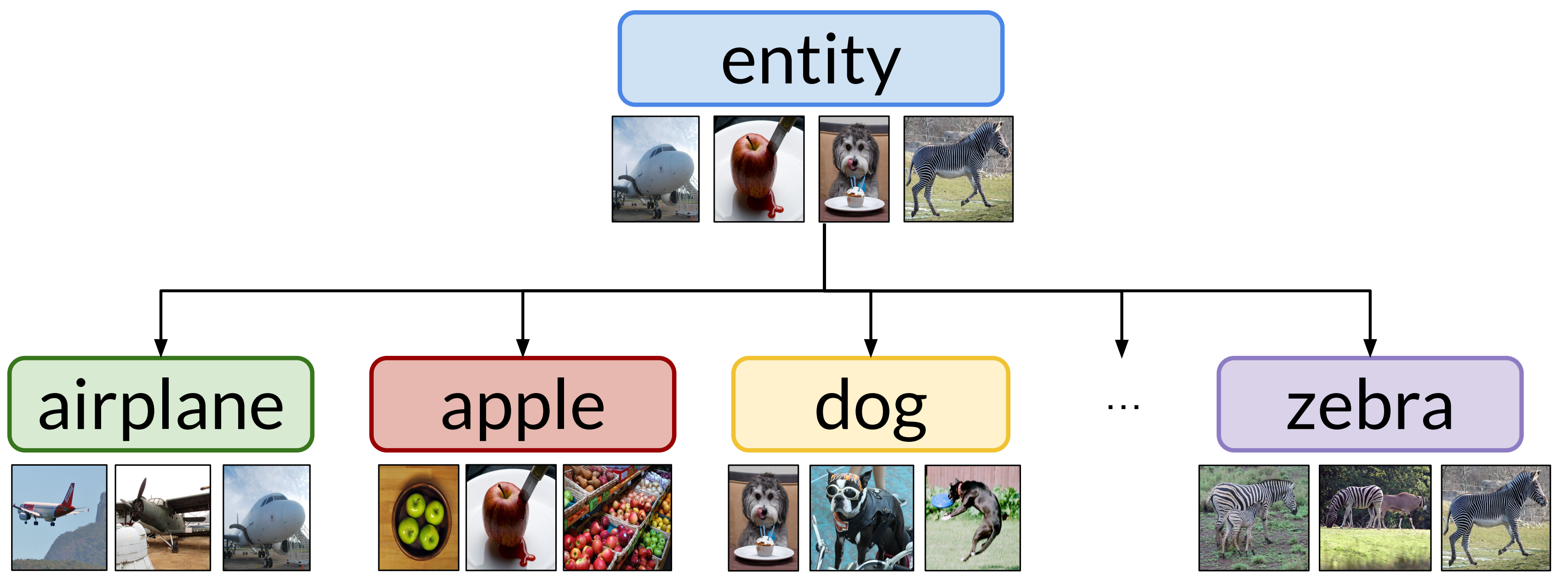
T1

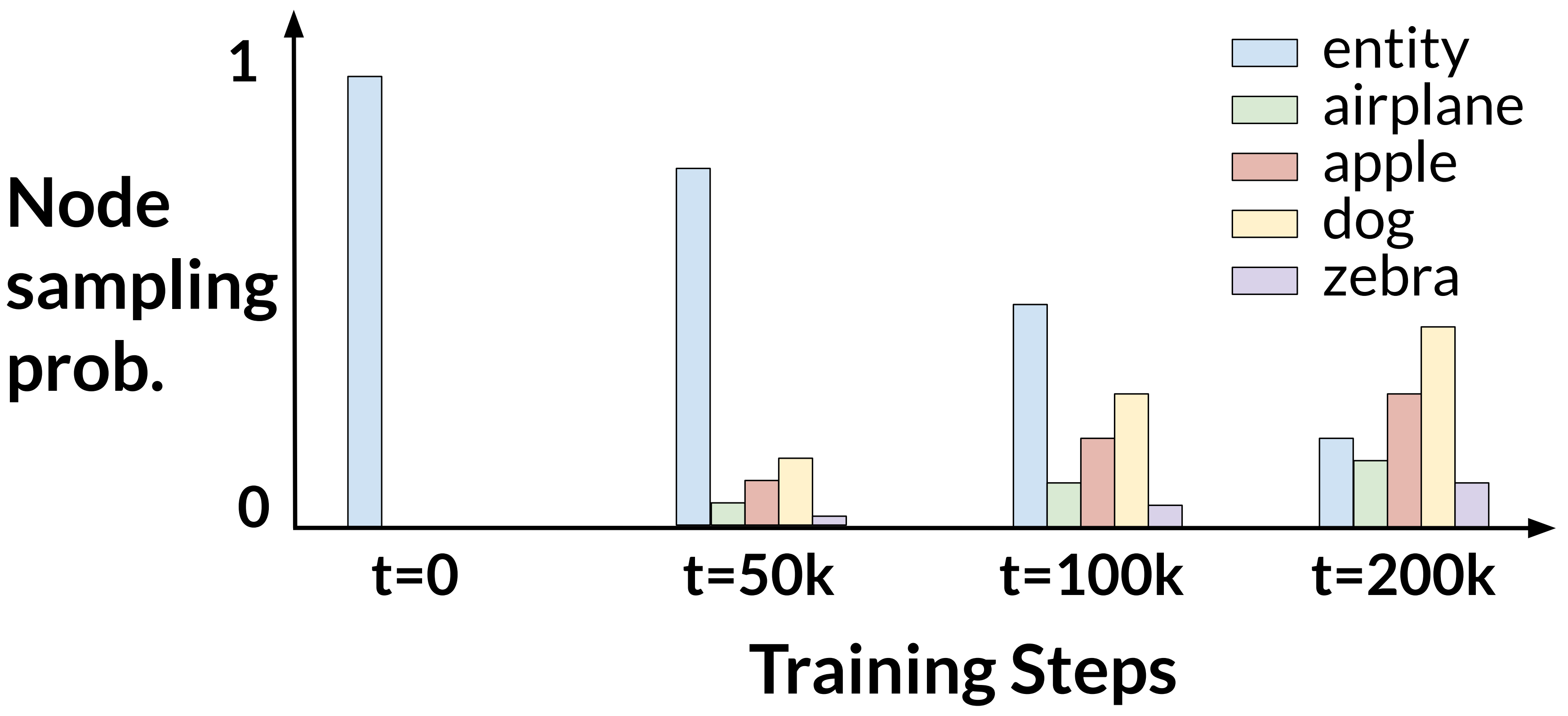
T1 · I1 T1 · I2 T1 · I3

? ? ?

TOnICS: Training with Ontology-Informed Contrastive Sampling

Curriculum learning algorithm for minibatch sampling

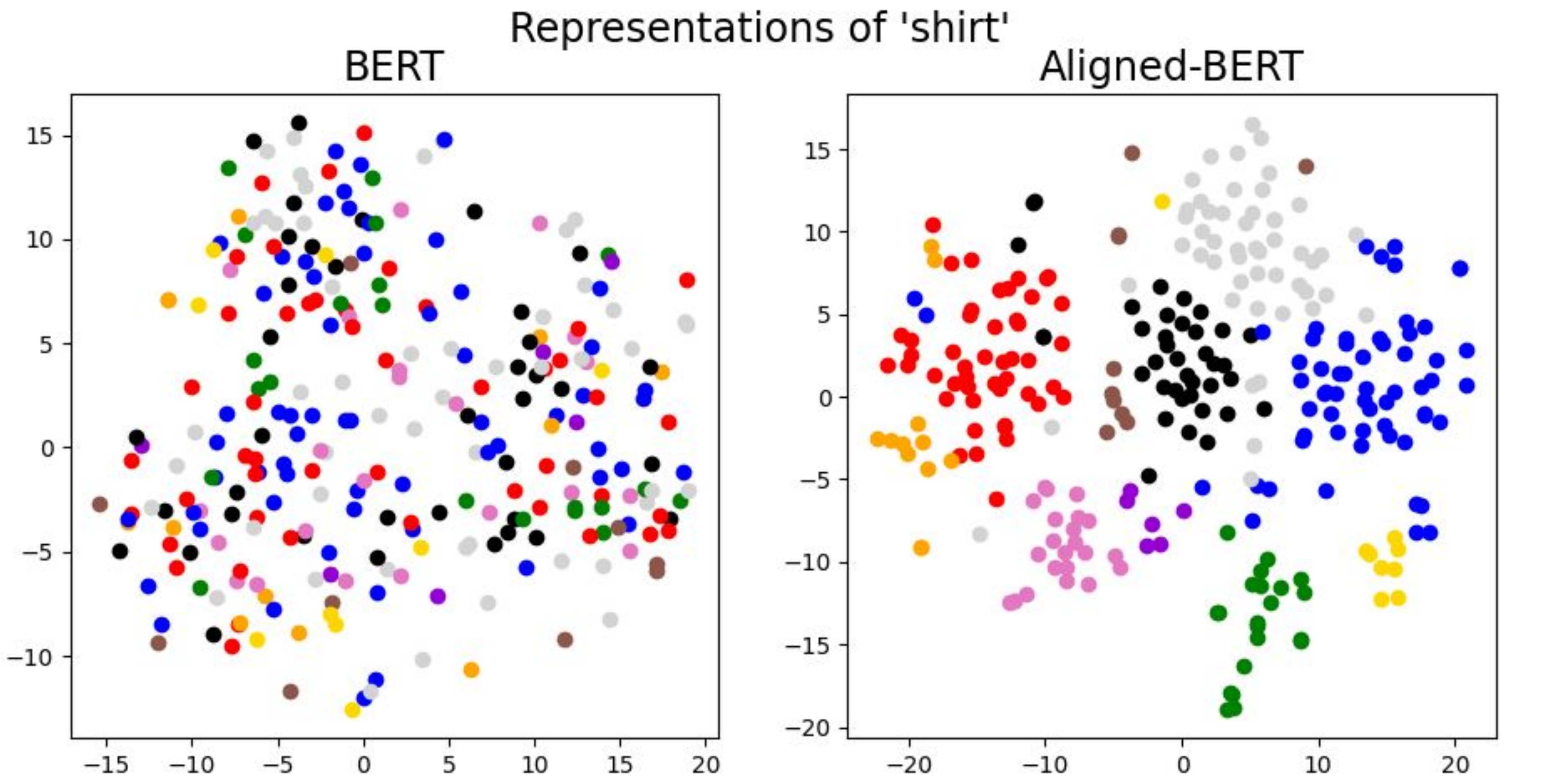




Results

- **Data:** MS-COCO + Google Conceptual Captions
- **Training:** 500K steps, 6 days on single GPU
- **Evaluation:** Zero-Shot Flickr30K Retrieval

			Image Retrieval		Text Retrieval	
Model	Minibatch Sampling Method	2-way Loss	R@1	R@5	R@1	R@5
CLIP-ViT-B/32	Random	-	58.6	83.4	**79.2**	**95.0**
BERT-VinVL Aligner (ours)	Random	✗	58.2	84.2	22.2	47.9
	TOnICS	✗	**60.3**	85.1	24.4	49.0
	Random	✔	58.9	84.6	76.1	93.3
	TOnICS	✔	59.7	**85.2**	76.6	94.1



USC Viterbi
School of Engineering