

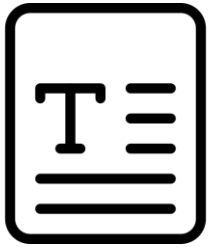
Multimodality

Vision-Language models (CLIP)

Yong-Sheng Chen
Dept. Computer Science, NYCU

What is modality

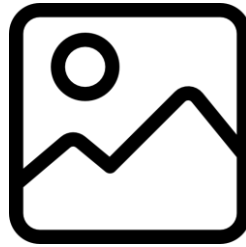
- Modality is the way in which information is expressed or perceived



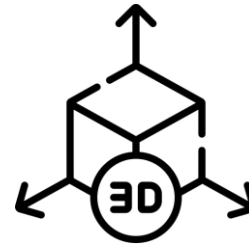
Text



Signal



Image



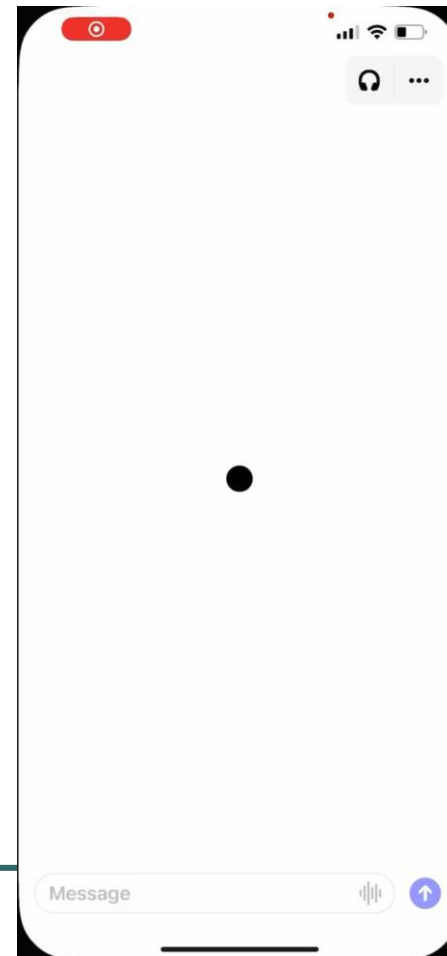
3D models



Sensor data

What is multimodal

- Multimodal involves the study of **heterogeneous** and **interconnected** data
- Examples
 - Image Captioning
 - Visual question answering
 - ...



Why multimodal

- **Enhanced Information Completeness:** Different modalities capture complementary aspects of the same scenario
- **Improved Performance:** Integrating multiple data sources often increases accuracy, robustness, and interpretability
- **Real-world Alignment:** Real environments involve diverse sensory inputs, making multimodal systems more practical

Multimodal is hot right now

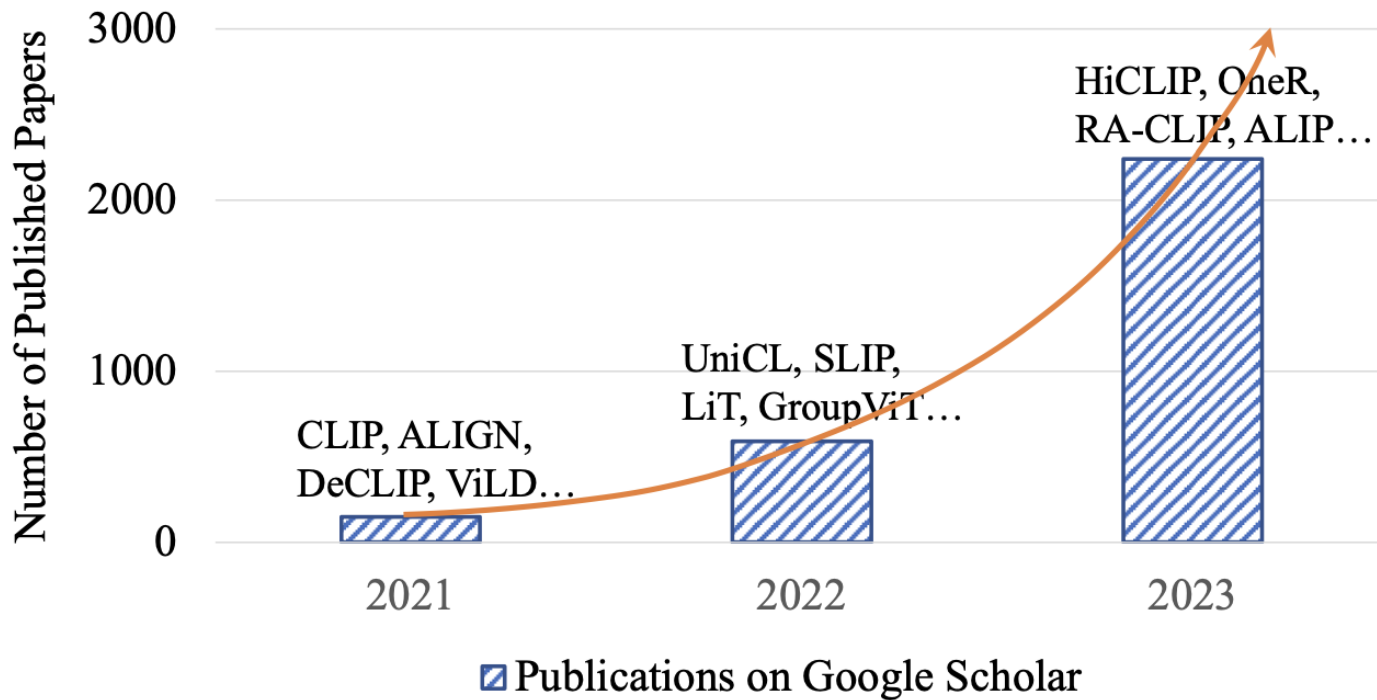
Multimodality stands at the forefront of the new wave of foundation model breakthroughs

Language Is Not All You Need: Aligning Perception with Language Models

Shaohan Huang*, Li Dong*, Wenhui Wang*, Yaru Hao*, Saksham Singhal*, Shuming Ma*
Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, Qiang Liu, Kriti Aggarwal
Zewen Chi, Johan Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, Furu Wei[†]
Microsoft

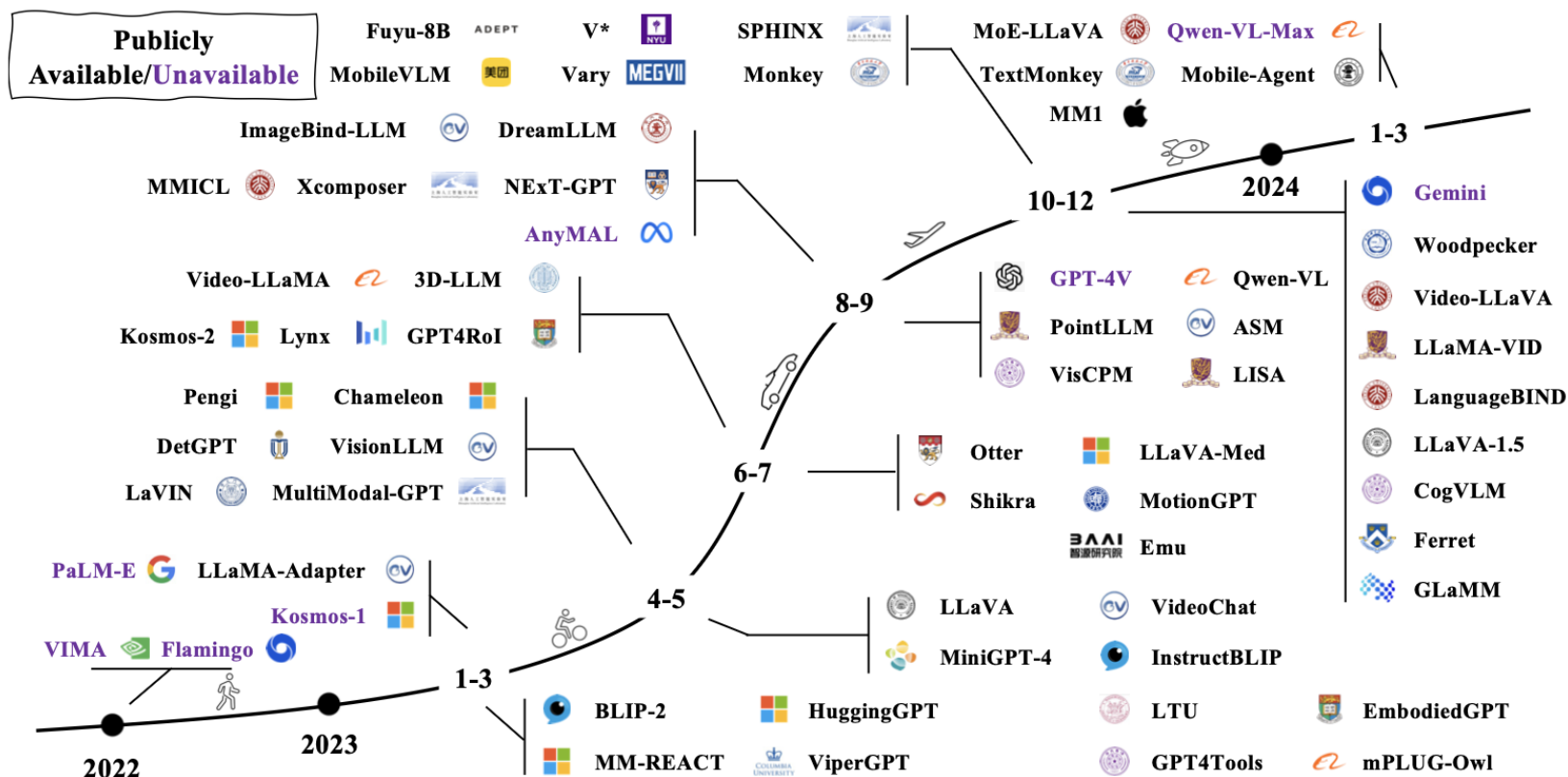
<https://github.com/microsoft/unilm>

Multimodal is hot right now



The number of publications on visual recognition Vision Language Models (VLMs) (from Google Scholar). The publications have grown exponentially since the pioneer study CLIP in 2021.

Multimodal is hot right now



A timeline of representative Multimodal Large Language Models (MLLMs).

CLIP: Contrastive Language-Image Pre-Training

Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PMLR, 2021.

Language vs. vision

● Language

- Self-supervised learning
- Zero-shot transfer
- Large dataset (from web)



● Vision

- Supervised learning
- Large dataset?



Language vs. vision

● Language

- Self-supervised learning
- Zero-shot transfer
- Large dataset (from web)



● Vision

- Supervised learning
- Large dataset?



CLIP: Alignment between language concept and vision concept

- Availability of large dataset
- Enable zero-shot transfer

Sufficiently Large Dataset

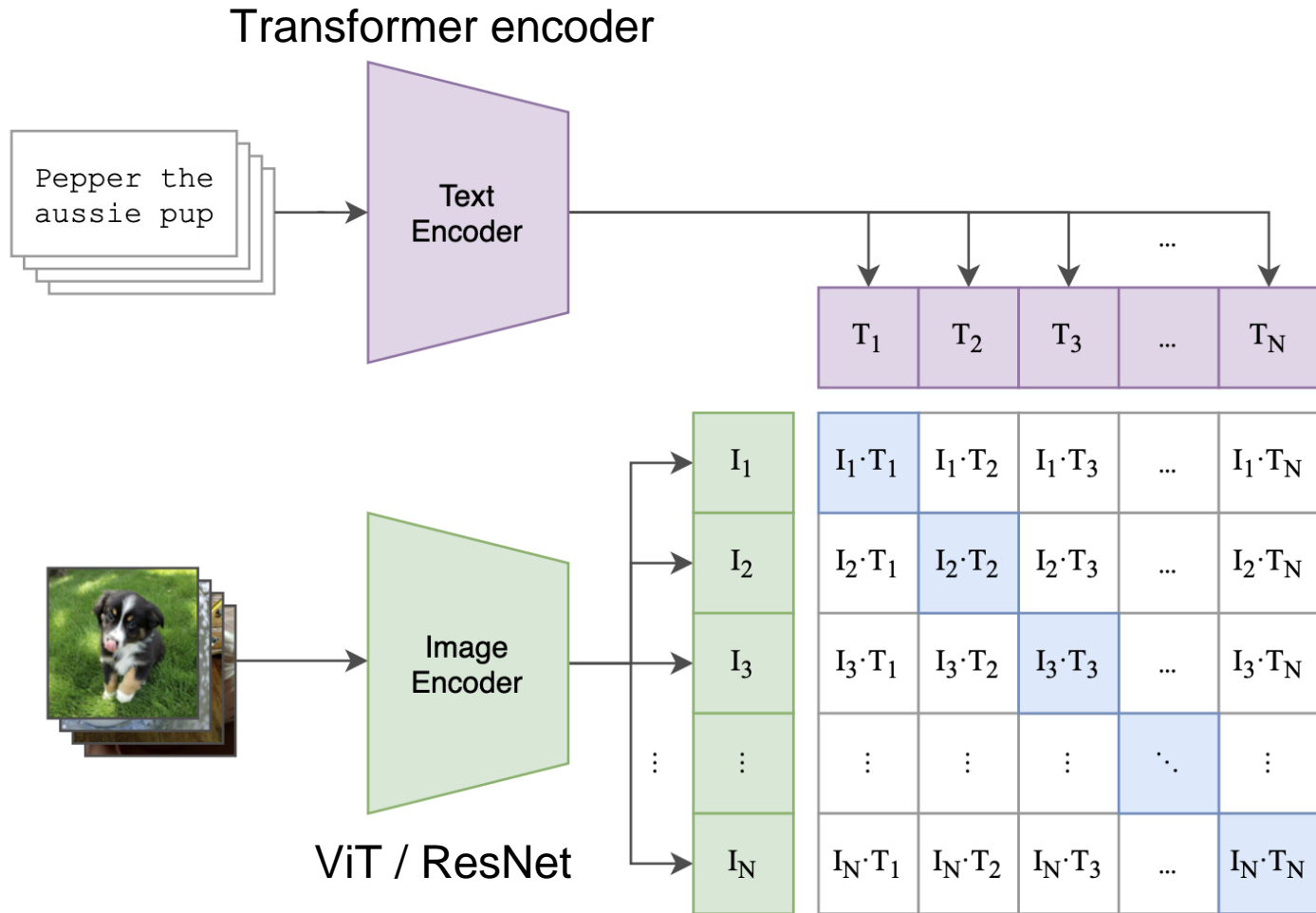
- **WebImageText (WIT) dataset**

- Consists of 400M (image, text) pairs sourced from the Internet
- Filtered using 500,000 high-frequency words from web
- Capped at 20,000 pairs per keyword to maintain diversity

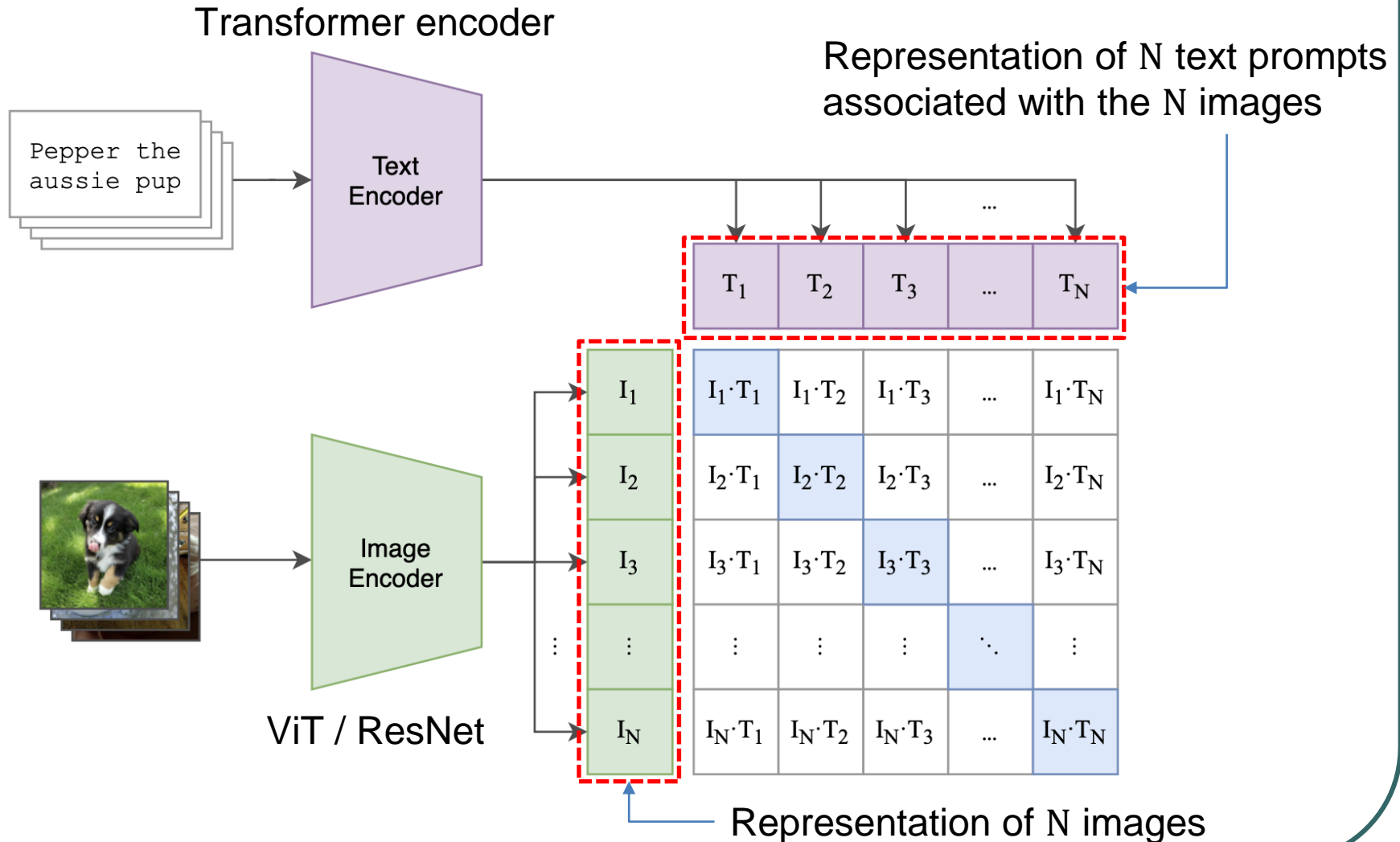
Dataset	Size	Annotation
WIT (CLIP)	400M pairs	✗
YFCC100M	100M images	✗
ImageNet	1.4M images	✓
COCO	330K images	✓
JFT-300M	300M images	✓

The WIT dataset has a similar total word count as the WebText dataset used to train GPT-2

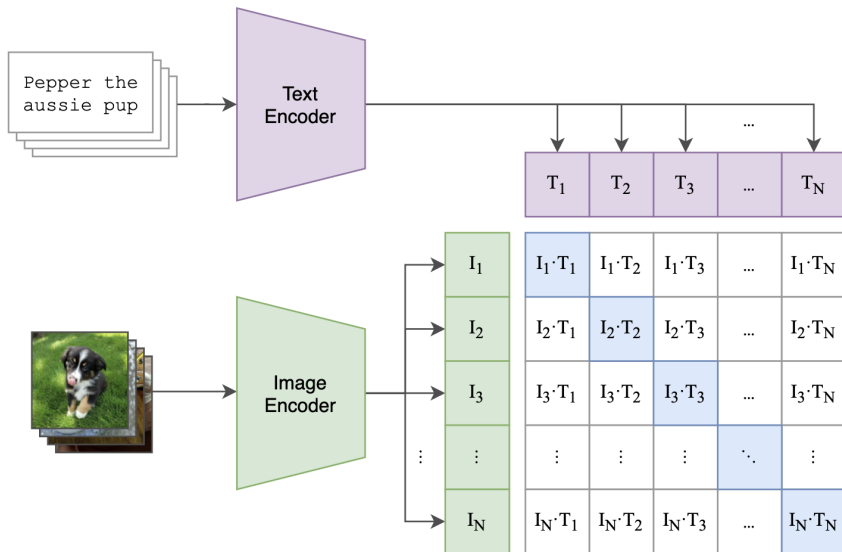
Contrastive Pre-Training



Contrastive Pre-Training



Contrastive Pre-Training



```
# 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
```

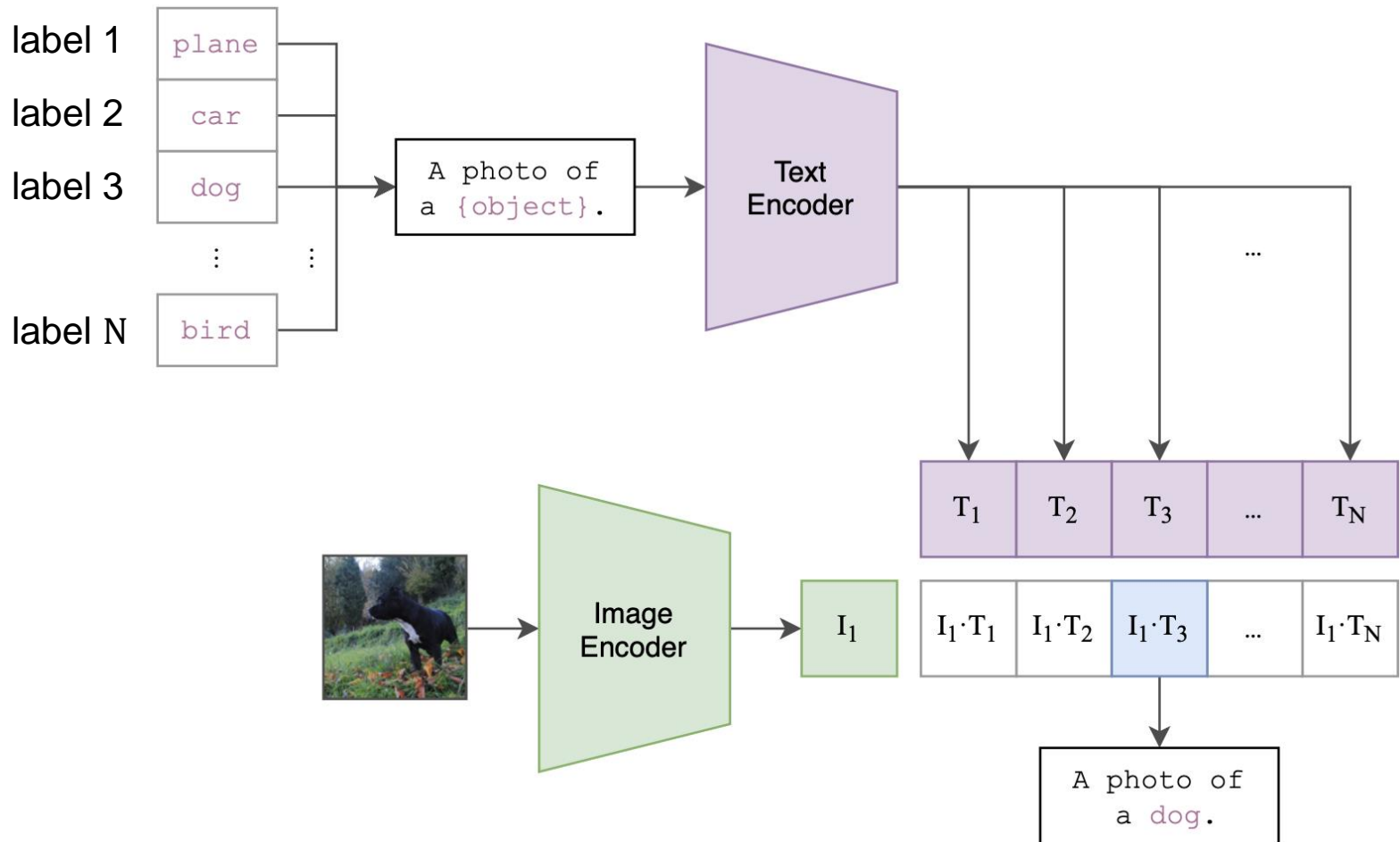
```
# extract feature representations of each modality
I_f = image_encoder(I) #[n, d_i]
T_f = text_encoder(T)  #[n, d_t]
```

```
# 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)
```

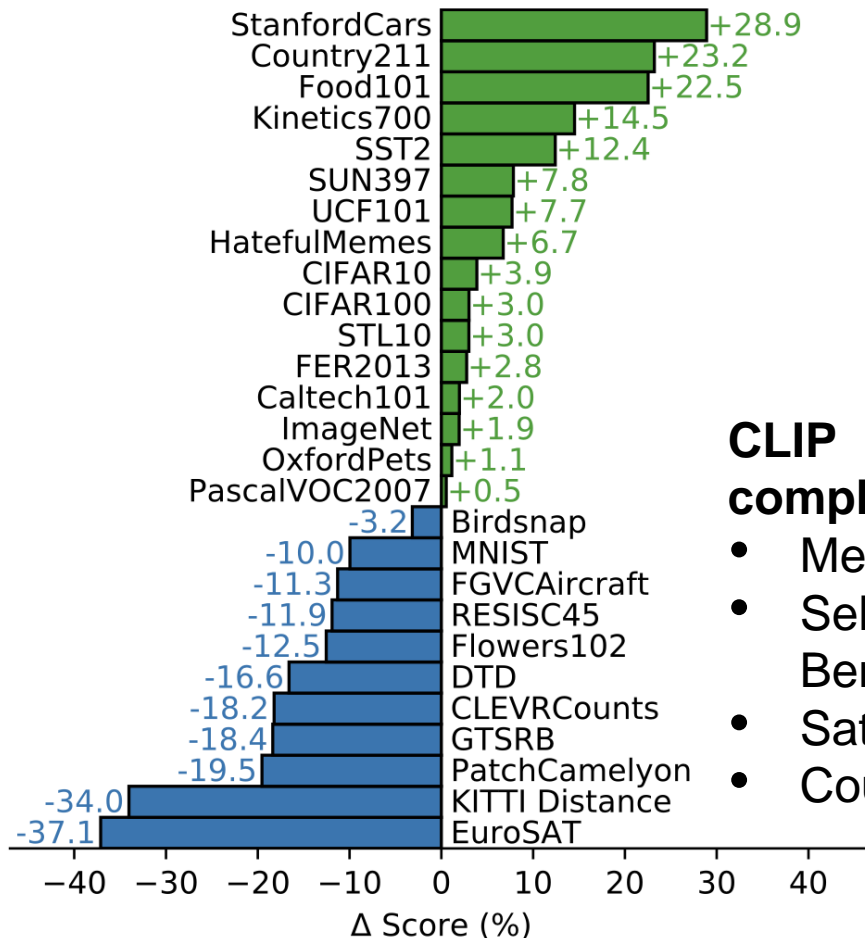
```
# symmetric loss function
labels = np.arange(n)
loss_i = cross_entropy_loss(logits, labels, axis=0)
loss_t = cross_entropy_loss(logits, labels, axis=1)
loss   = (loss_i + loss_t)/2
```

Zero-Shot Prediction



Zero-Shot Prediction Performance

Zero-shot CLIP vs. Linear Probe on ResNet50

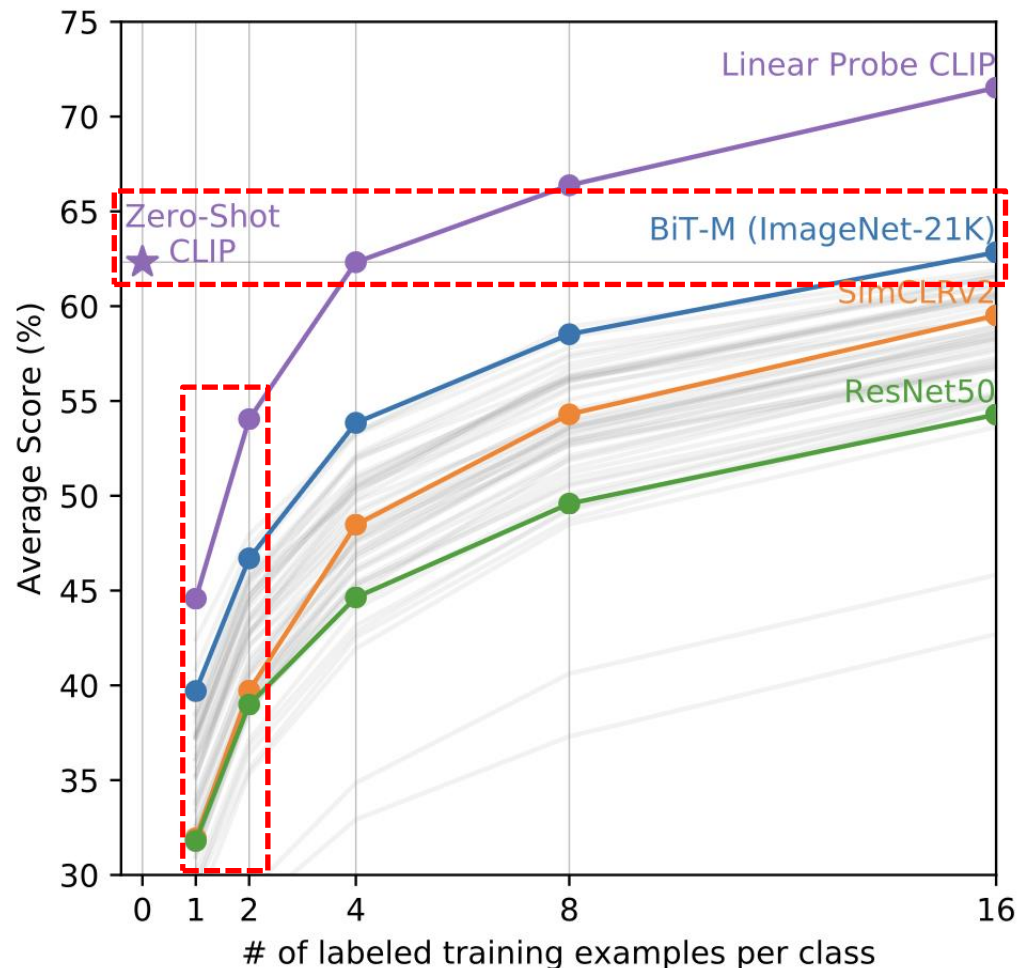


CLIP is not good at handling specialized, complex, or abstract task dataset

- Medical: Patch Camelyon
- Self-driving: German Traffic Sign Recognition Benchmark (GTSRB) & KITTI distance
- Satellite: EuroSAT & RESISC45
- Counting task: CLEVRCounts





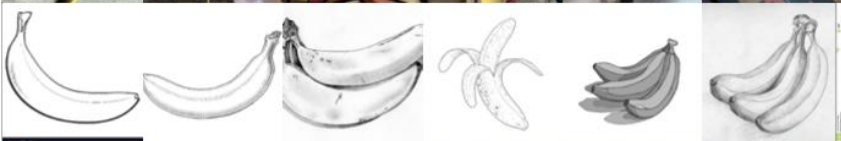

Zero-Shot Prediction Performance

Zero-shot CLIP vs. Few-shot Linear Probes



Zero-Shot Prediction Performance

Zero-shot CLIP vs. ImageNet model on natural distribution shifts

	Dataset Examples	ImageNet ResNet101	Zero-Shot CLIP	Δ Score
ImageNet		76.2	76.2	0%
ImageNetV2		64.3	70.1	+5.8%
ImageNet-R		37.7	88.9	+51.2%
ObjectNet		32.6	72.3	+39.7%
ImageNet Sketch		25.2	60.2	+35.0%
ImageNet-A		2.7	77.1	+74.4%