IMEN574/IDSC574 Programming for Data Science

19. **Multimodal Models I**

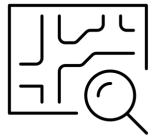
Spring 2026

Instructor: **Sangwoo Mo**

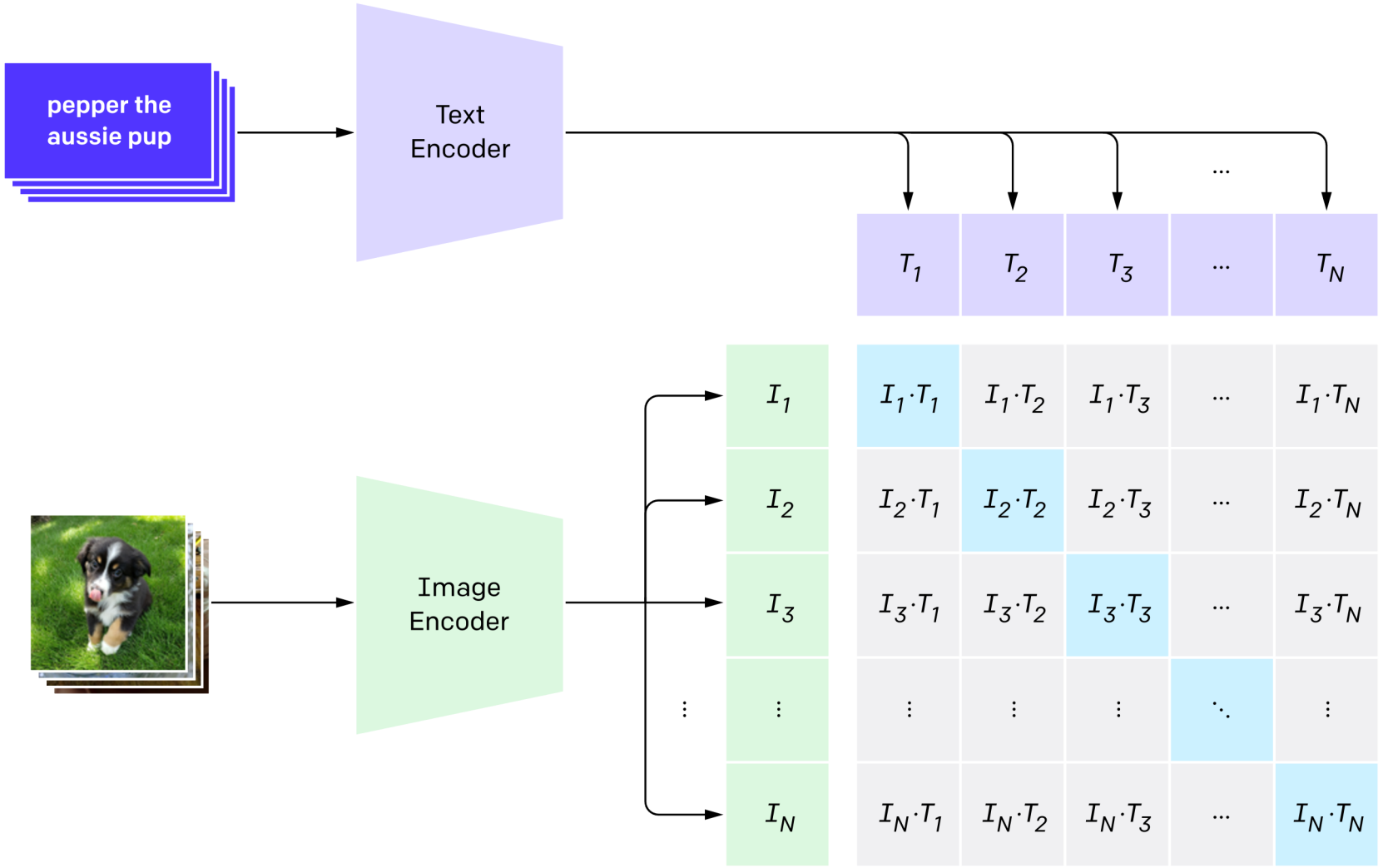
Abstract Perception – Semantic Cue

	Abstract Perception “what” (high-level, conceptual)	**Dense Perception** “where” (low-level, spatial)
Visual	Self-Supervised Learning (e.g., DINO)	1. Spatialization of Abstract Perception 2. Dense Supervision (e.g., SAM)
Semantic		



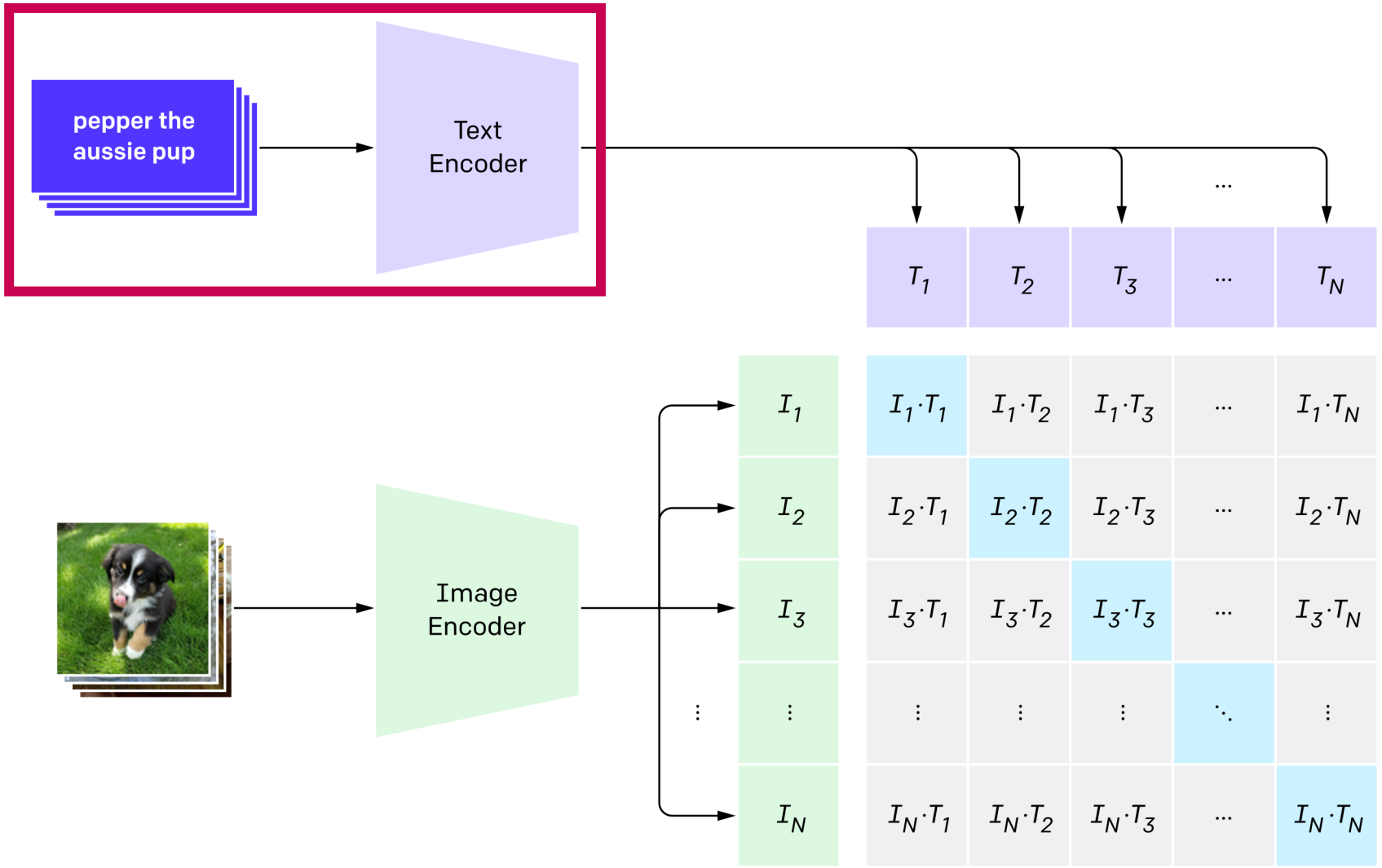


Semantic Cue: Vision-Language Model (e.g., CLIP)



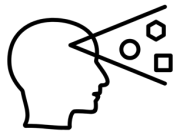
Radford et al. Learning Transferable Visual Models From Natural Language Supervision. ICML 2021.

Before Diving Into CLIP...

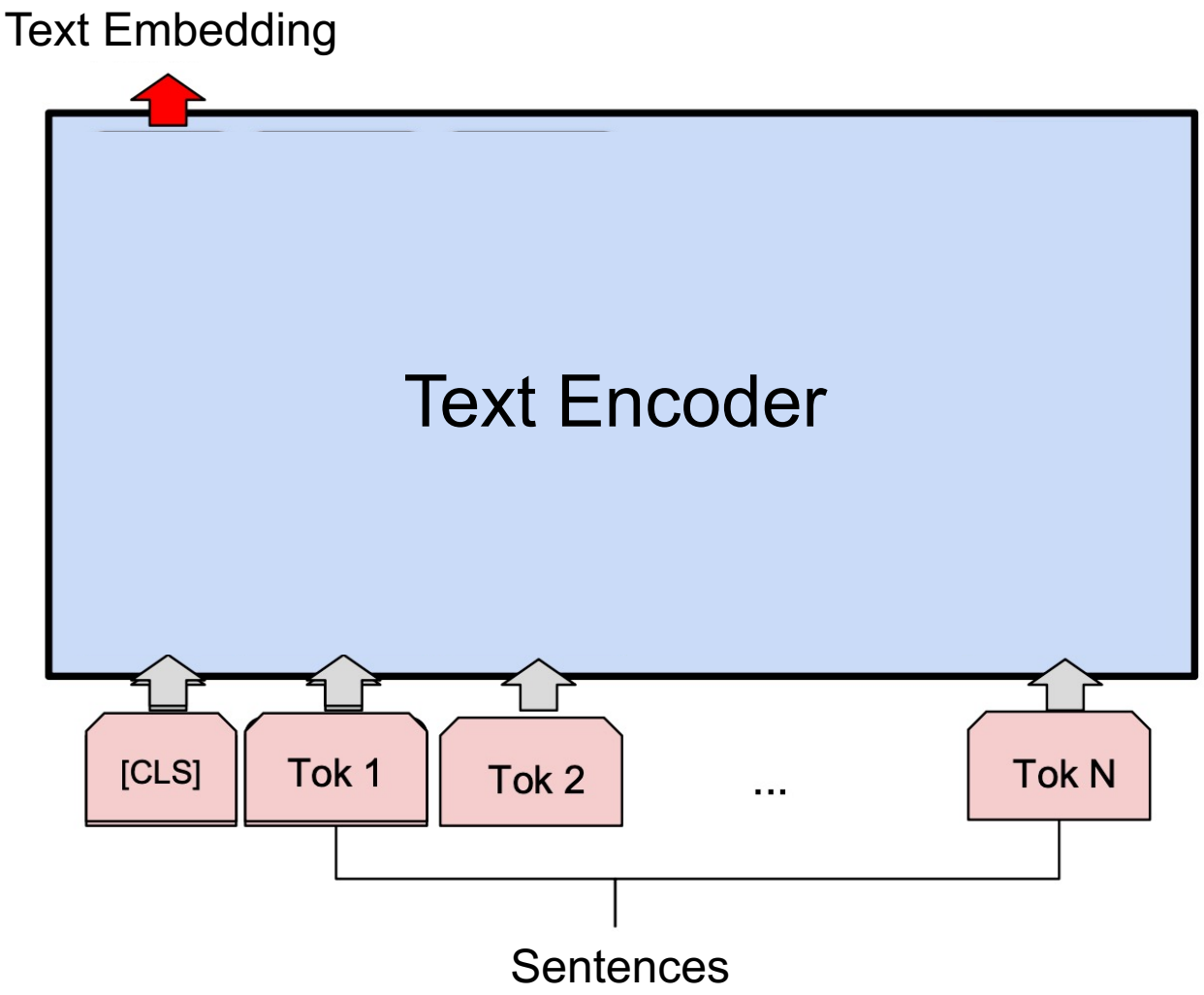


Radford et al. Learning Transferable Visual Models From Natural Language Supervision. ICML 2021.

Language Understanding

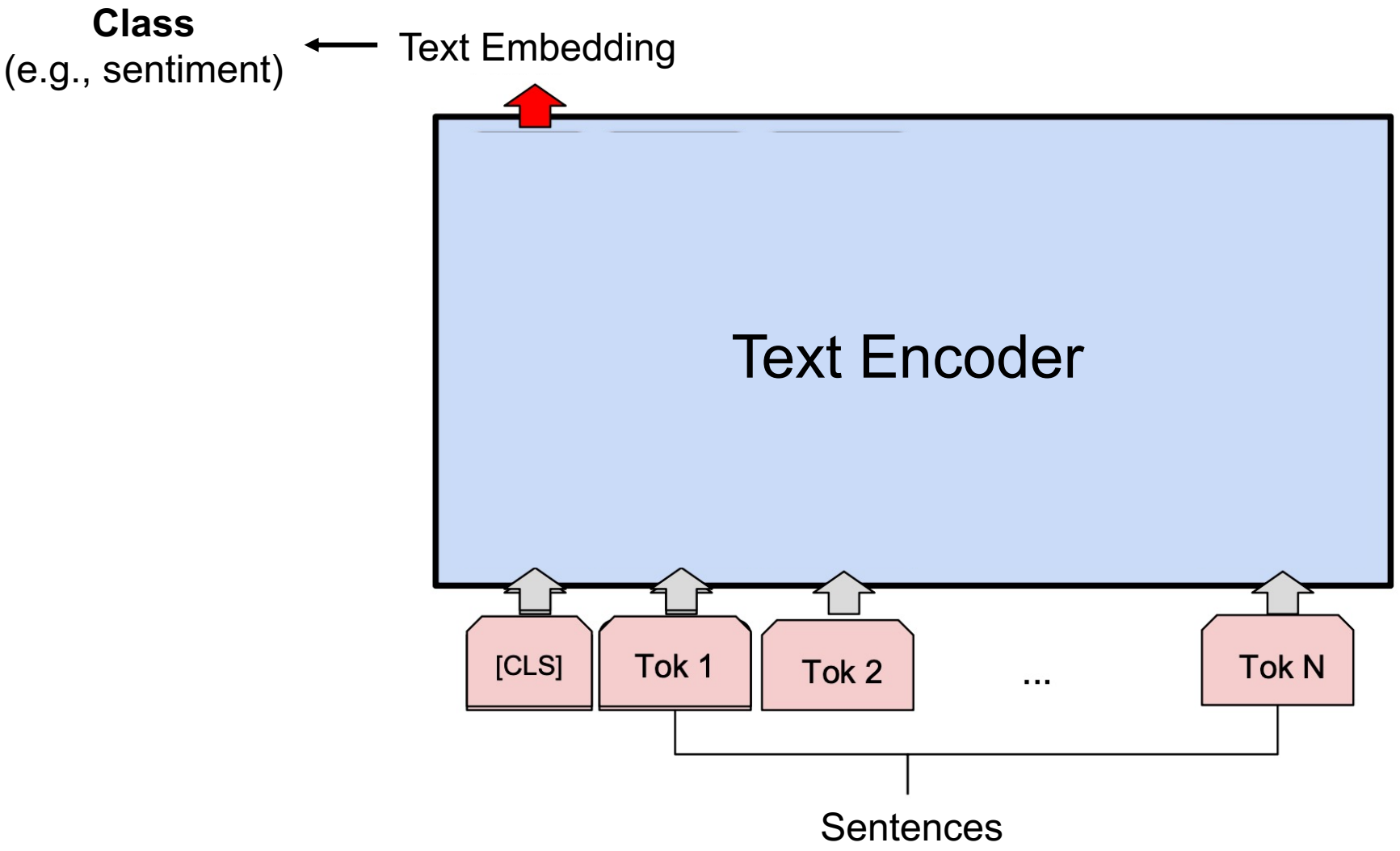
	Perception 	Generation 	Action 
Vision	DINO, CLIP, SAM, etc.		
Language	BERT		

Language Understanding



Delvin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL 2019.

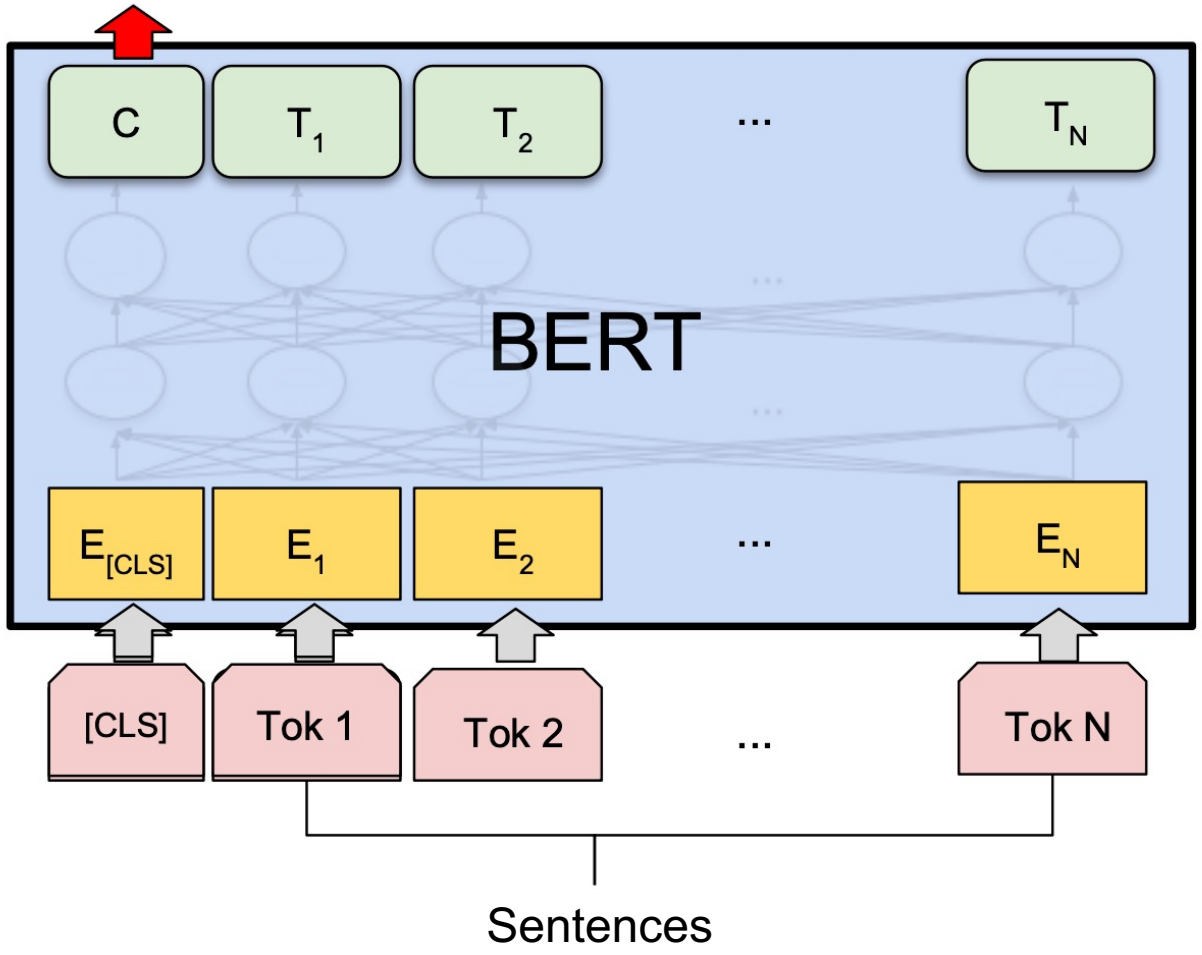
Language Understanding



Delvin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL 2019.

Language Understanding

Self-Supervised Learning (e.g., BERT)



Delvin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL 2019.

BERT

BERT

Bidirectional Encoder Representations from Transformers

Like *Bidirectional LSTMs*, let's look in **both** directions



BERT

BERT

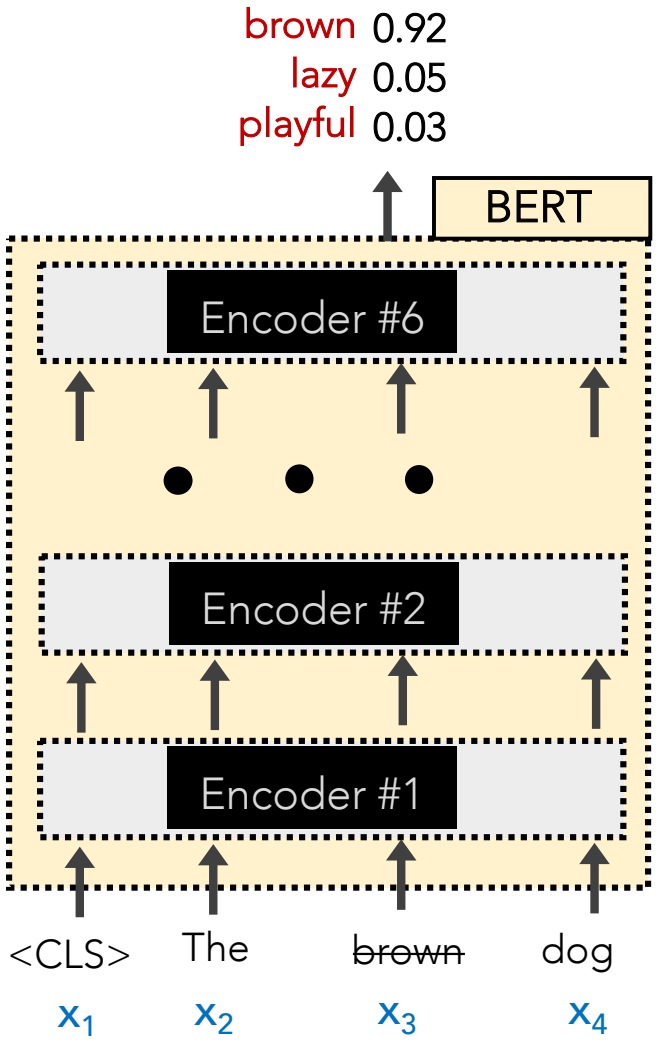
Bidirectional Encoder Representations from Transformers

Let's only use Transformer *Encoders*, no Decoders



BERT

BERT

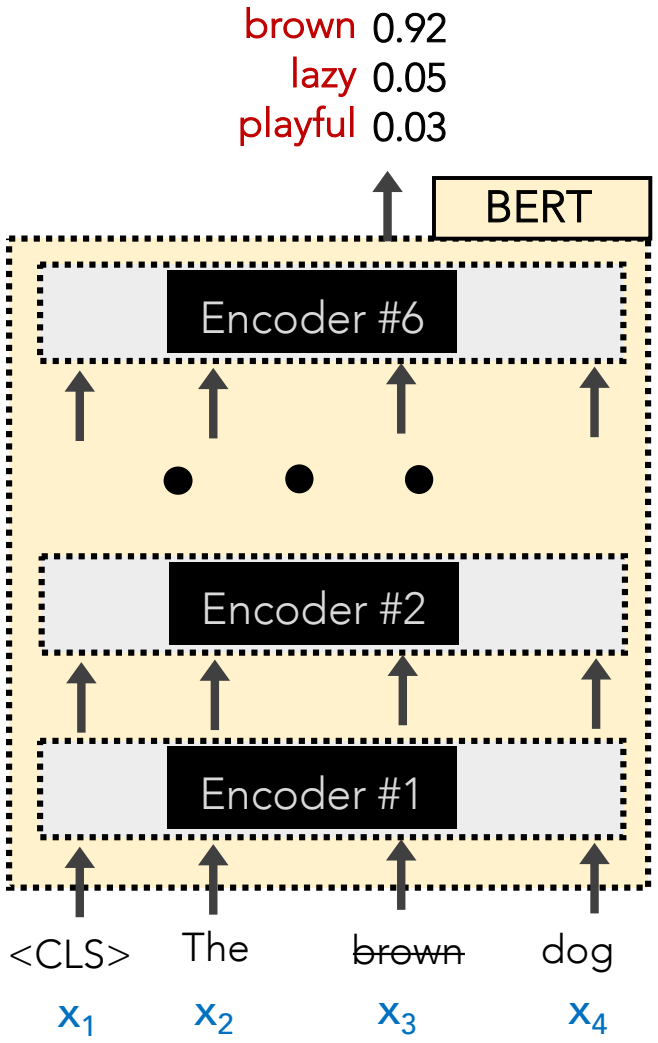


BERT:

- Model: several Transformer Encoders. Input sentence or sentence pairs, [CLS] token, subword embeddings
- Objective: MLM and next-sentence prediction
- Data: BooksCorpus and Wikipedia

BERT

BERT



BERT has 2 training objectives:

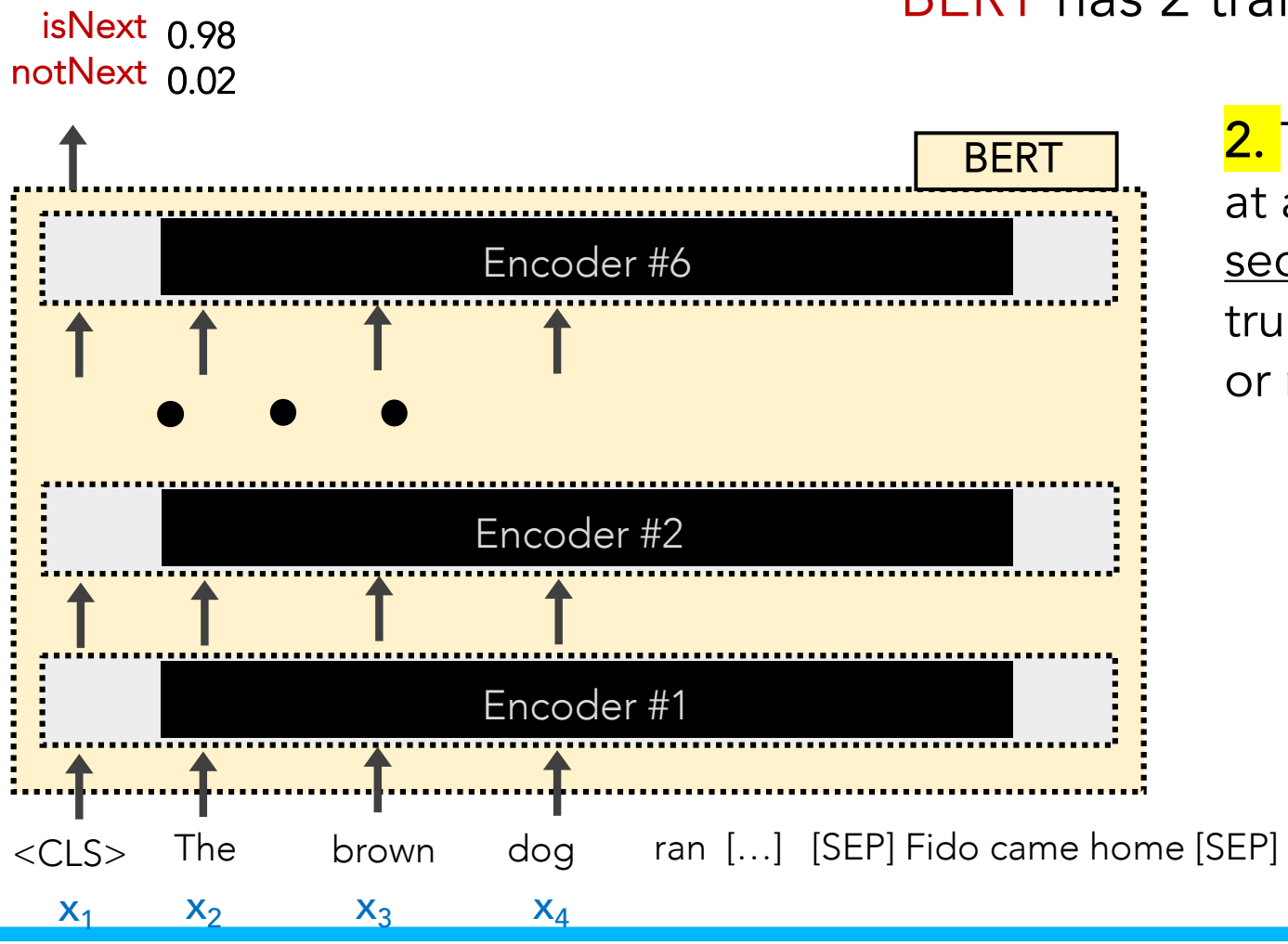
1. Predict the **Masked word** (a la CBOW)

15% of all input words are randomly masked.

- 80% become [MASK]
- 10% become revert back
- 10% become are deliberately corrupted as wrong words

BERT

BERT



BERT has 2 training objectives:

2. Two sentences are fed in at a time. Predict the if the second sentence of input truly follows the first one or not.

BERT

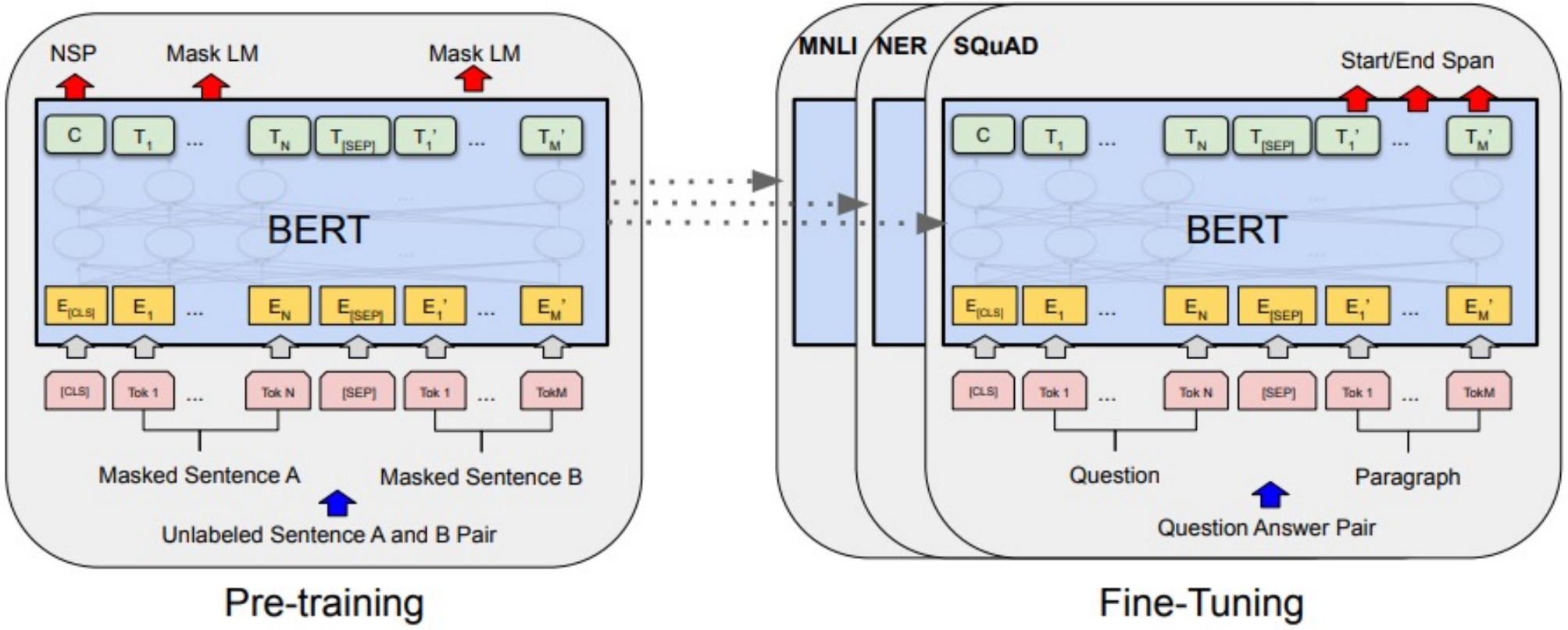
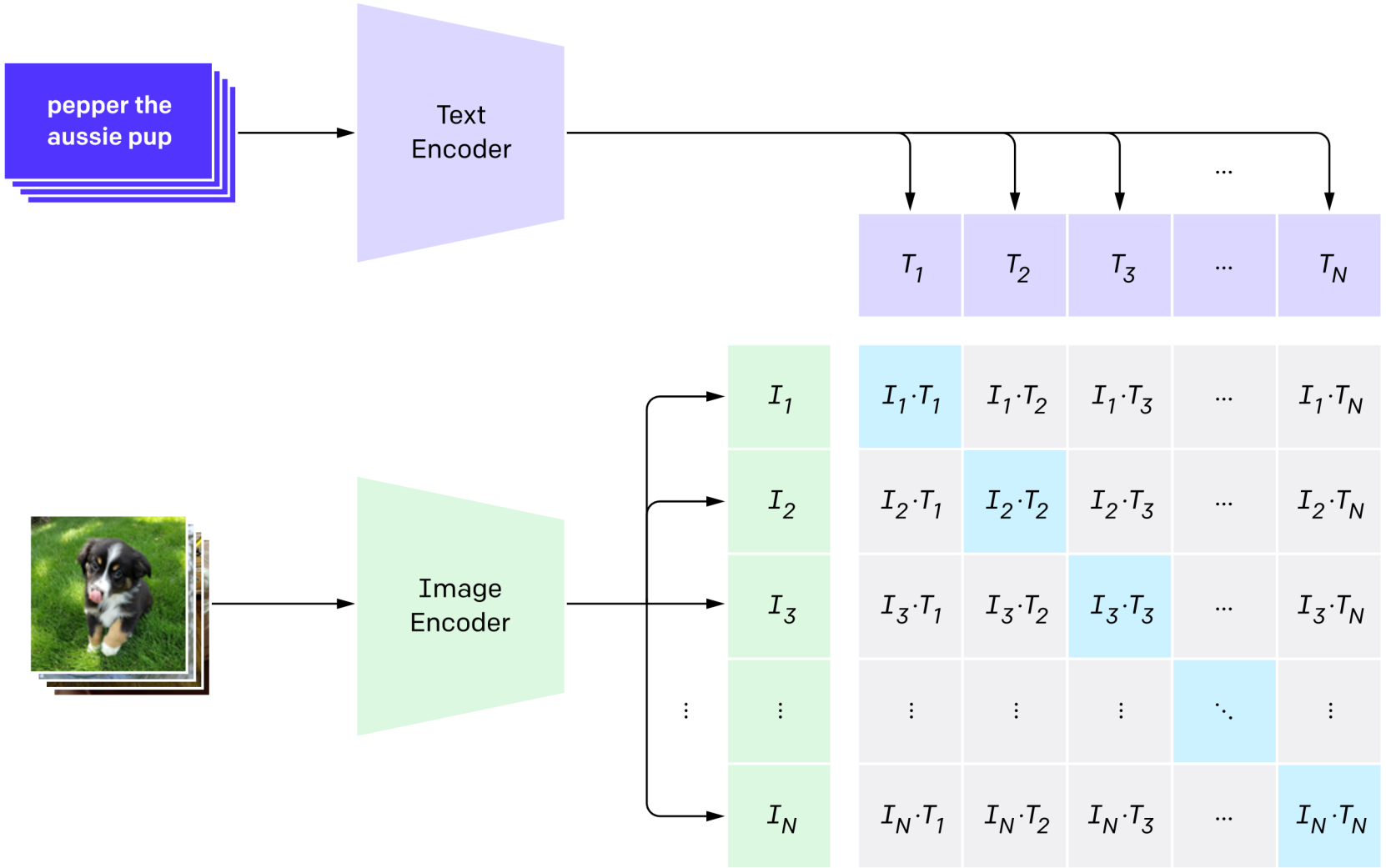


Figure 1: Overall pre-training and fine-tuning procedures for BERT. Apart from output layers, the same architectures are used in both pre-training and fine-tuning. The same pre-trained model parameters are used to initialize models for different down-stream tasks. During fine-tuning, all parameters are fine-tuned. `[CLS]` is a special symbol added in front of every input example, and `[SEP]` is a special separator token (e.g. separating questions/answers).

https://aclanthology.org/N19-1423.pdf

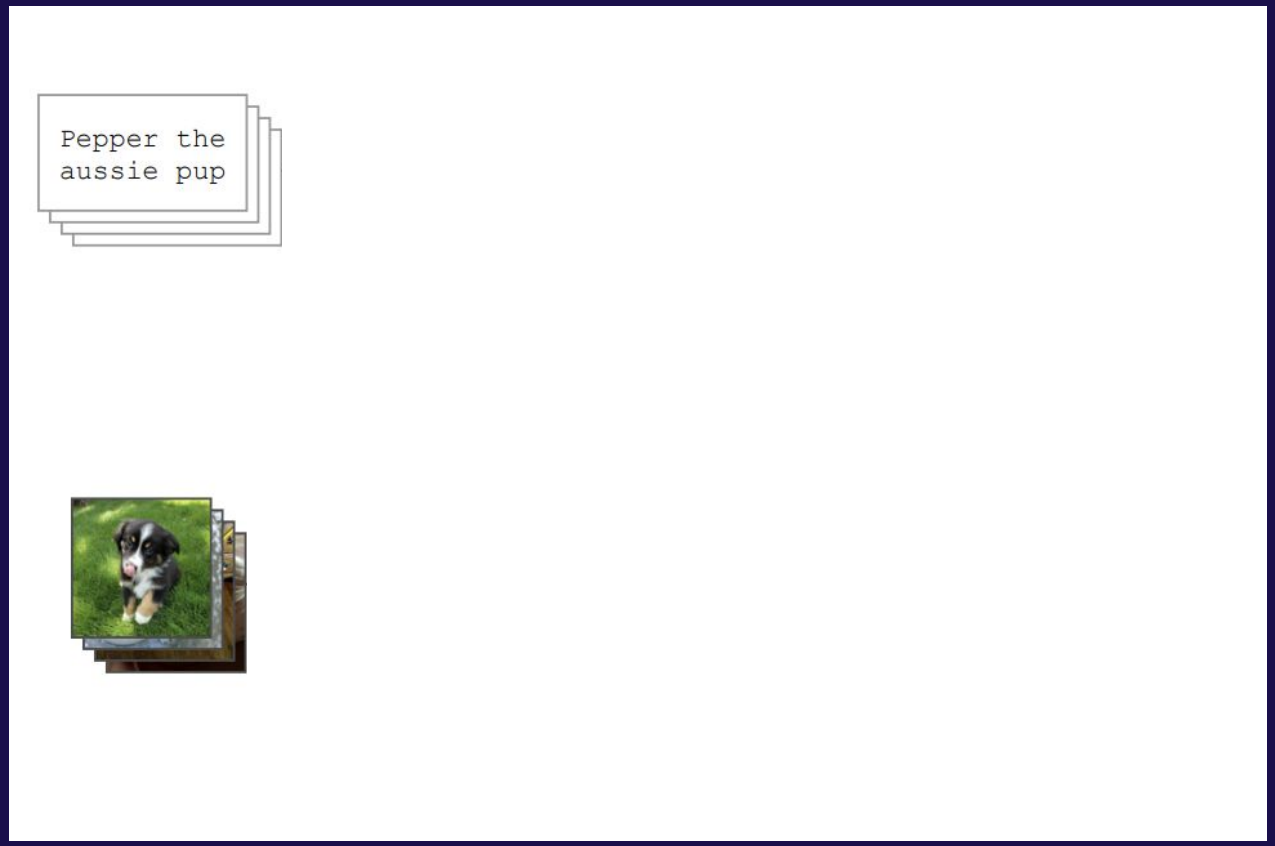
CLIP



Radford et al. Learning Transferable Visual Models From Natural Language Supervision. ICML 2021.

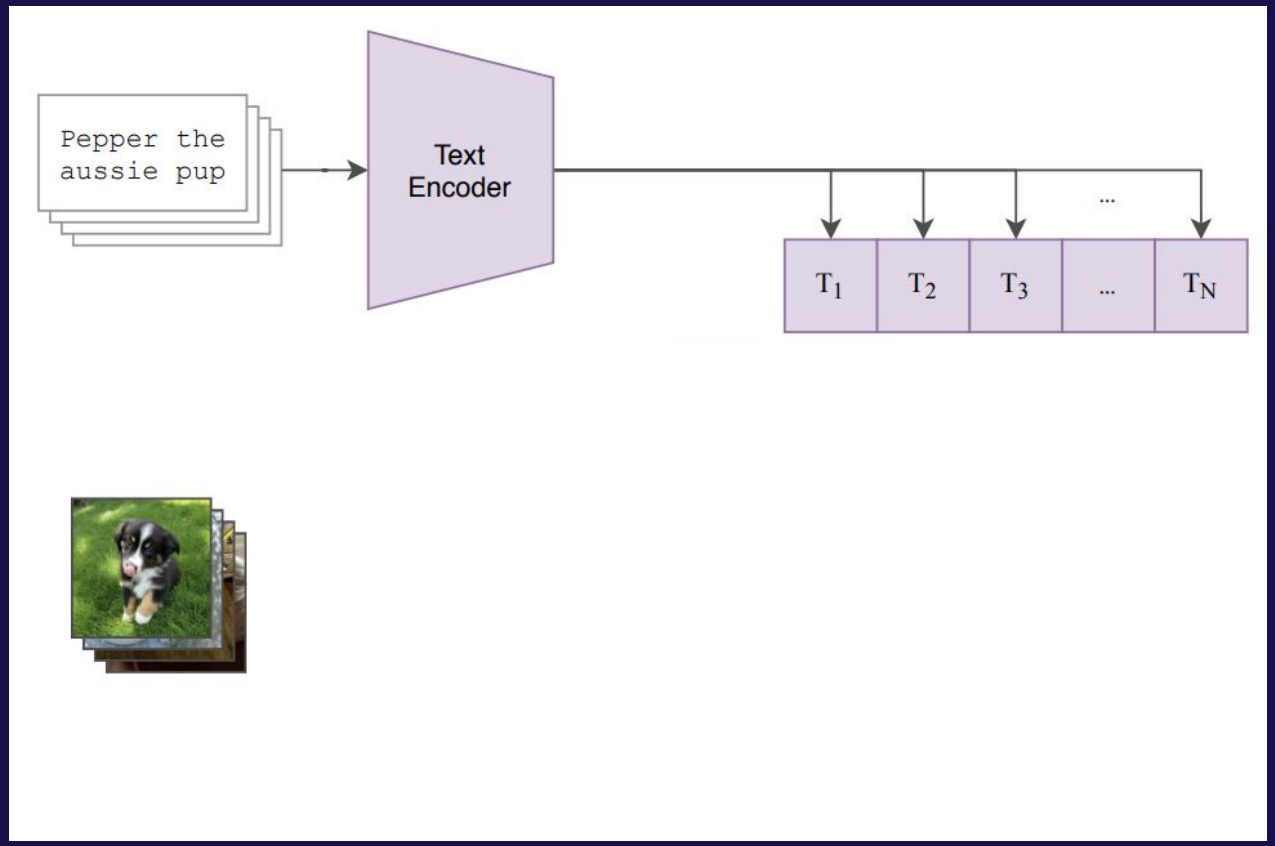
CLIP

CLIP: Contrastive Language-Image Pre-training



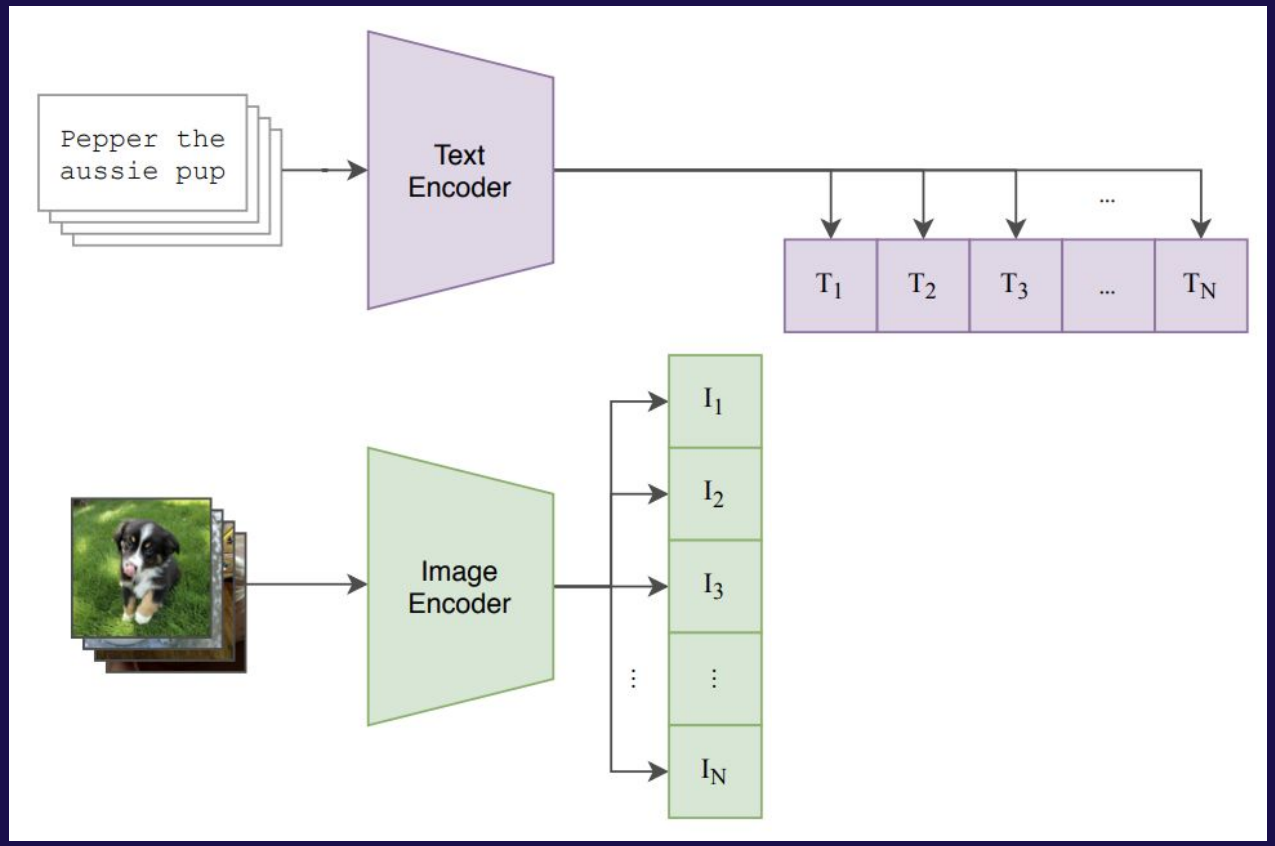
CLIP

CLIP: Contrastive Language-Image Pre-training



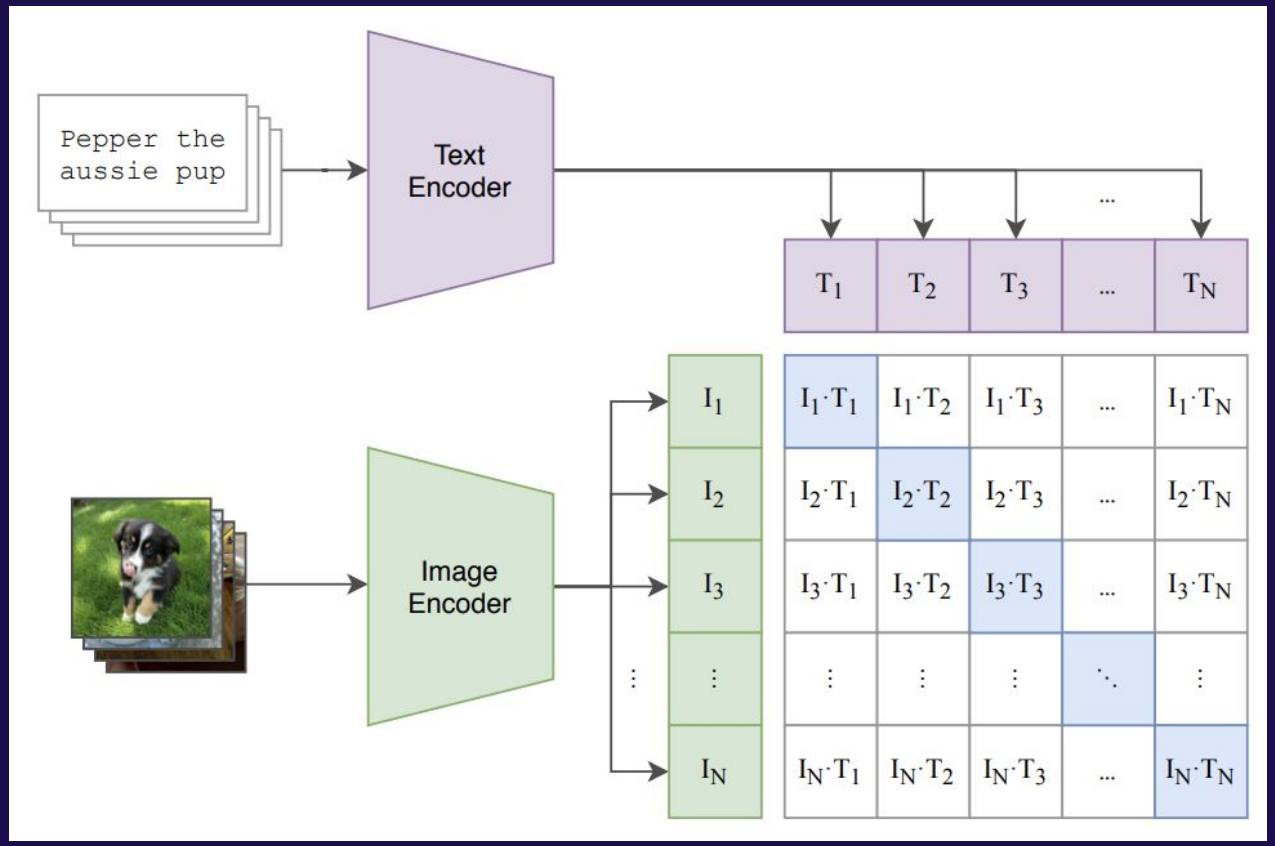
CLIP

CLIP: Contrastive Language-Image Pre-training



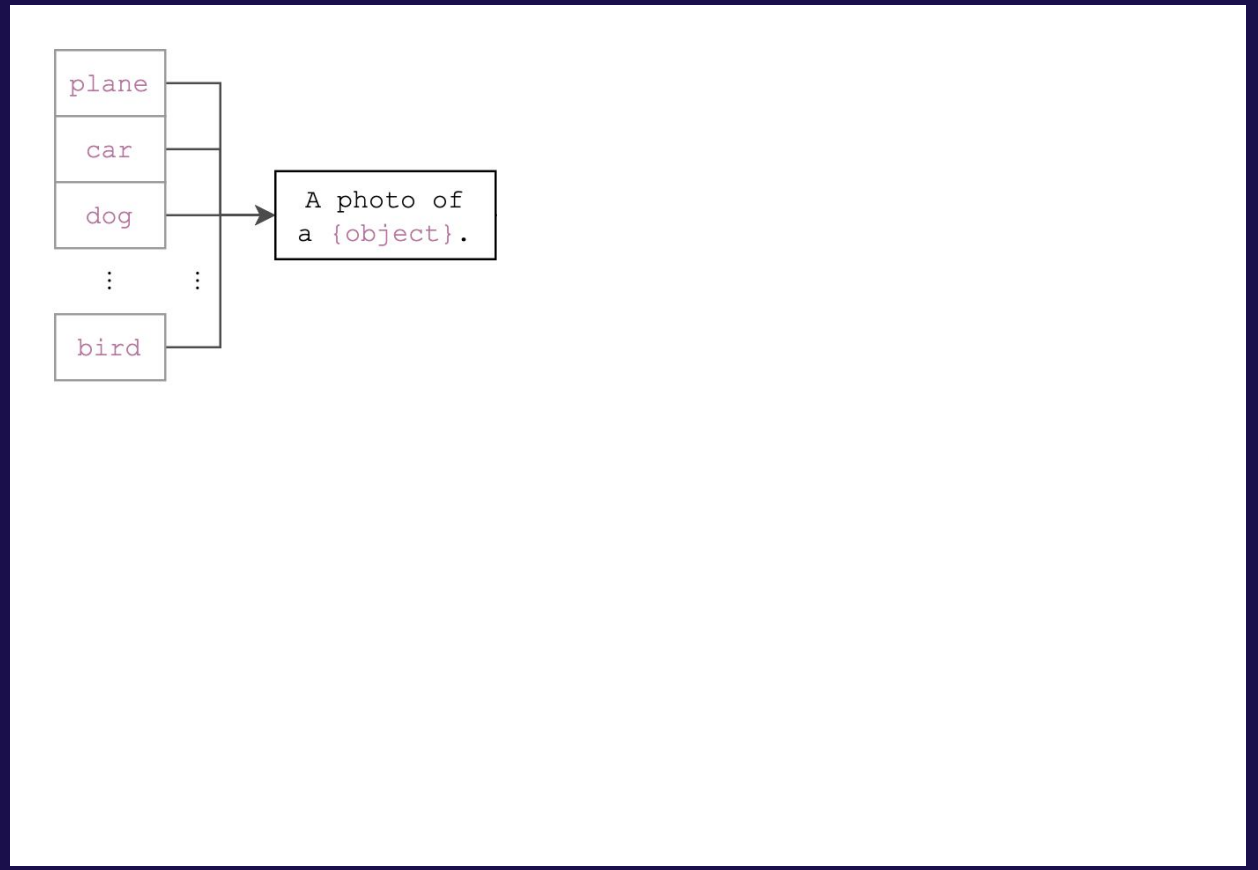
CLIP

CLIP: Contrastive Language-Image Pre-training



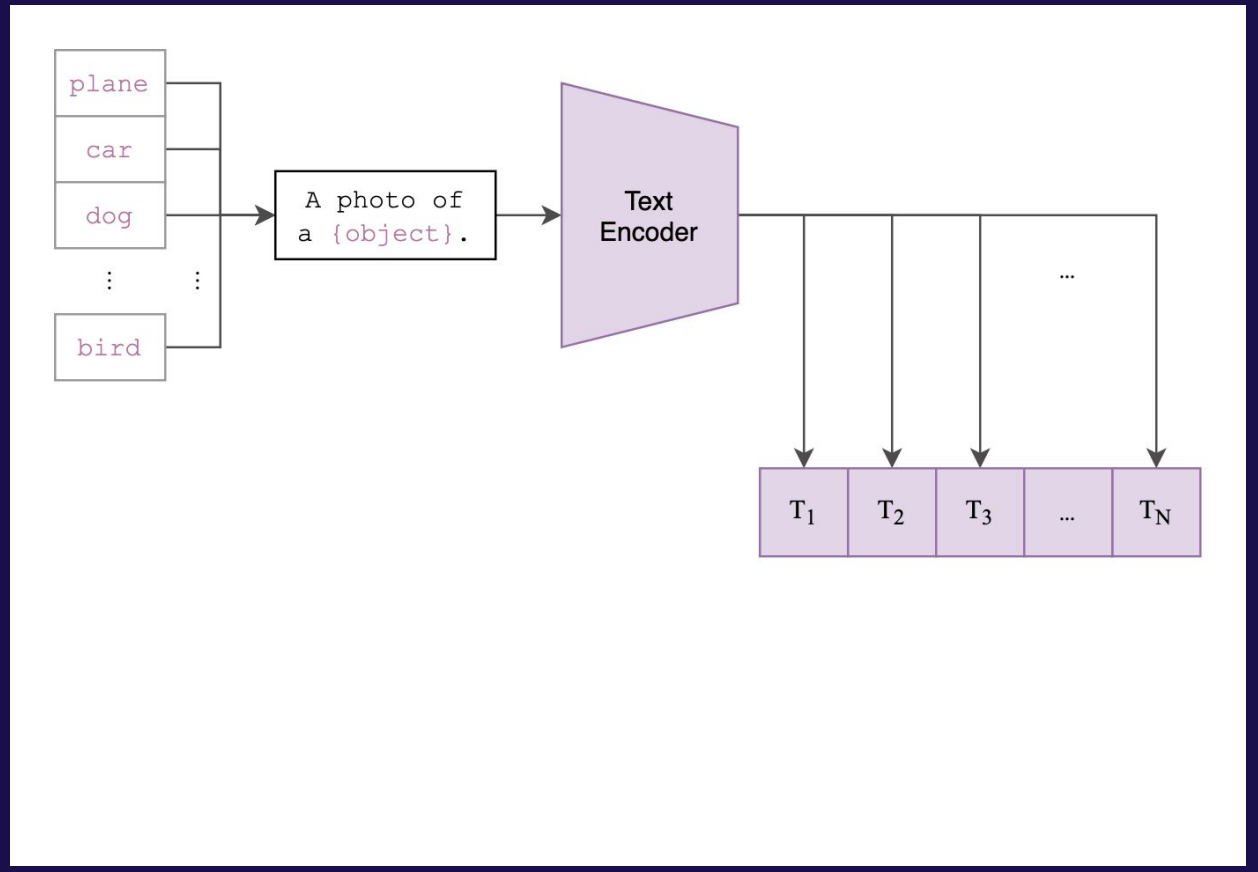
CLIP

Zero-shot image classification



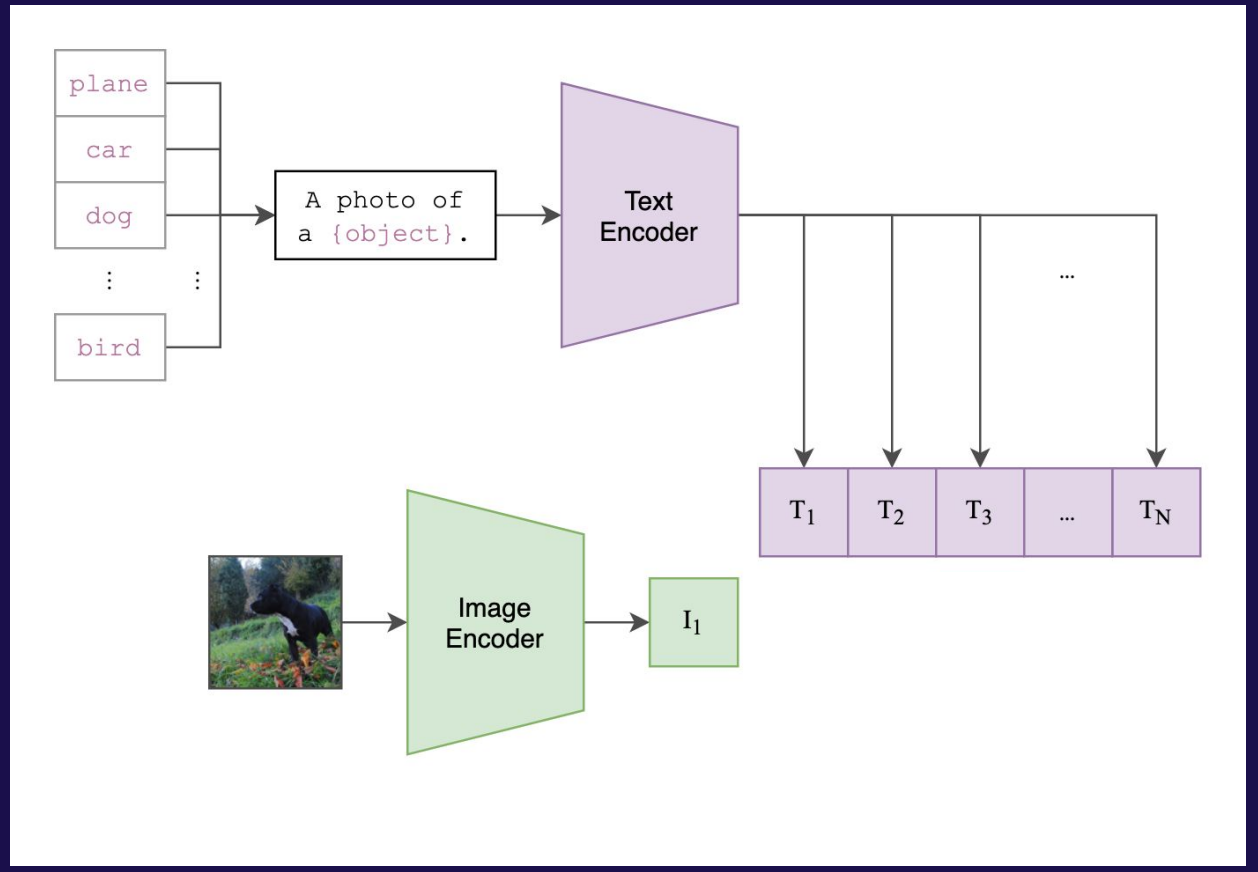
CLIP

Zero-shot image classification



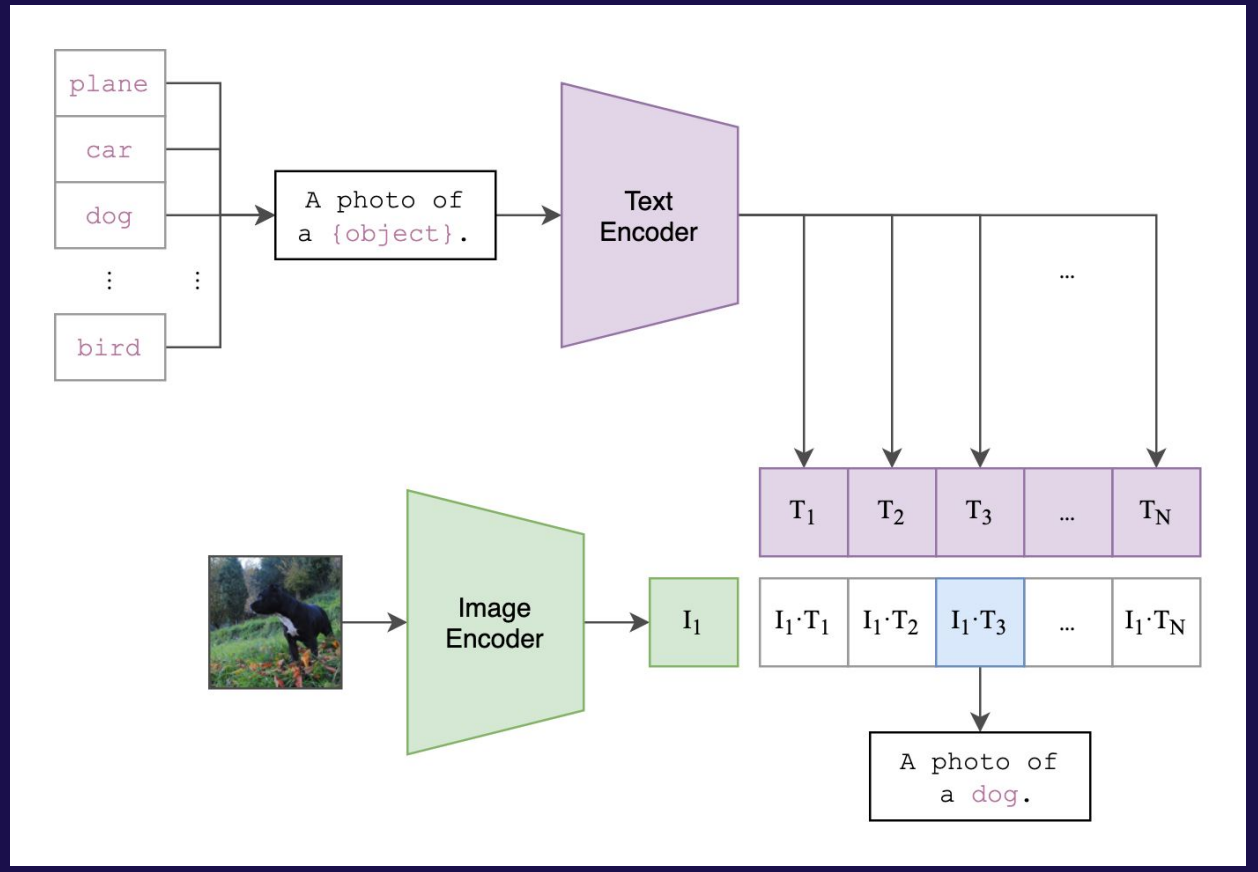
CLIP

Zero-shot image classification



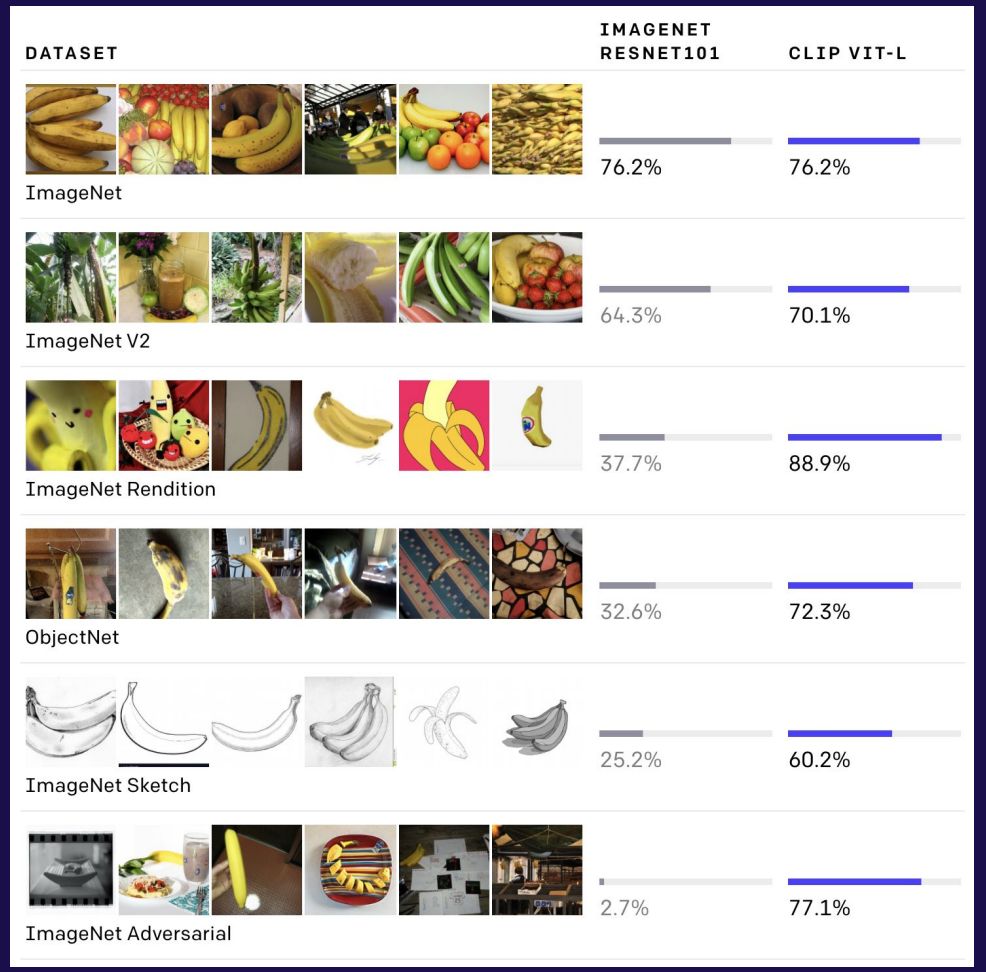
CLIP

Zero-shot image classification



CLIP

Zero-shot CLIP is much more robust

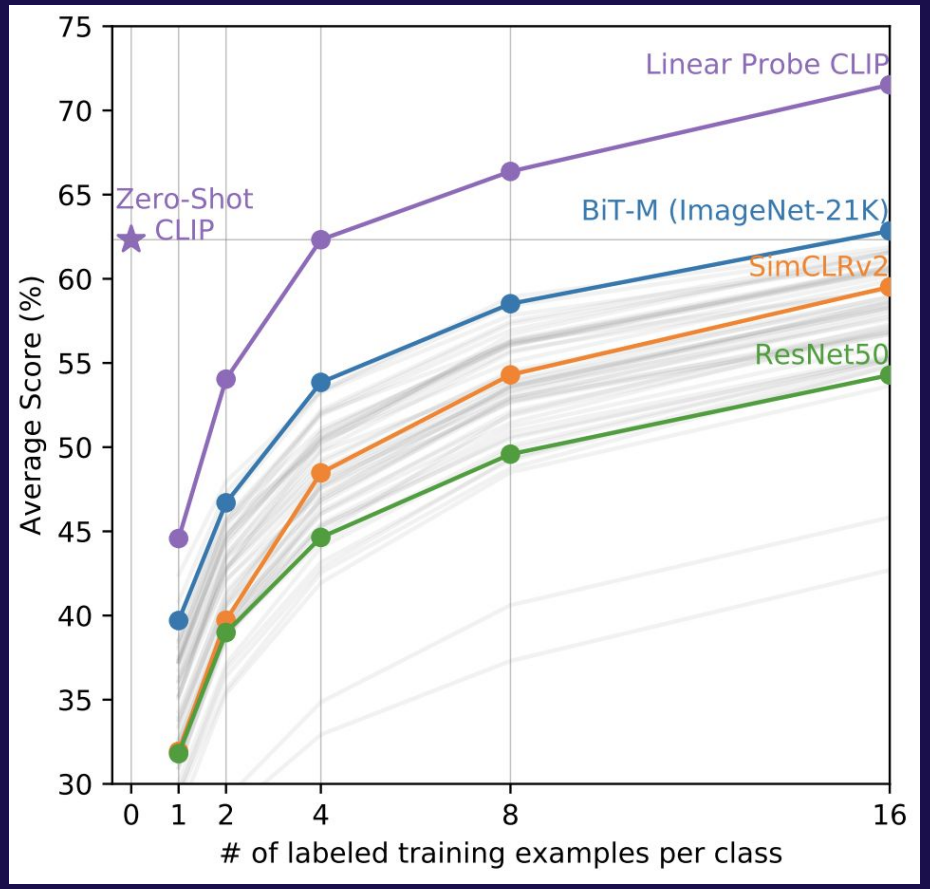


CLIP

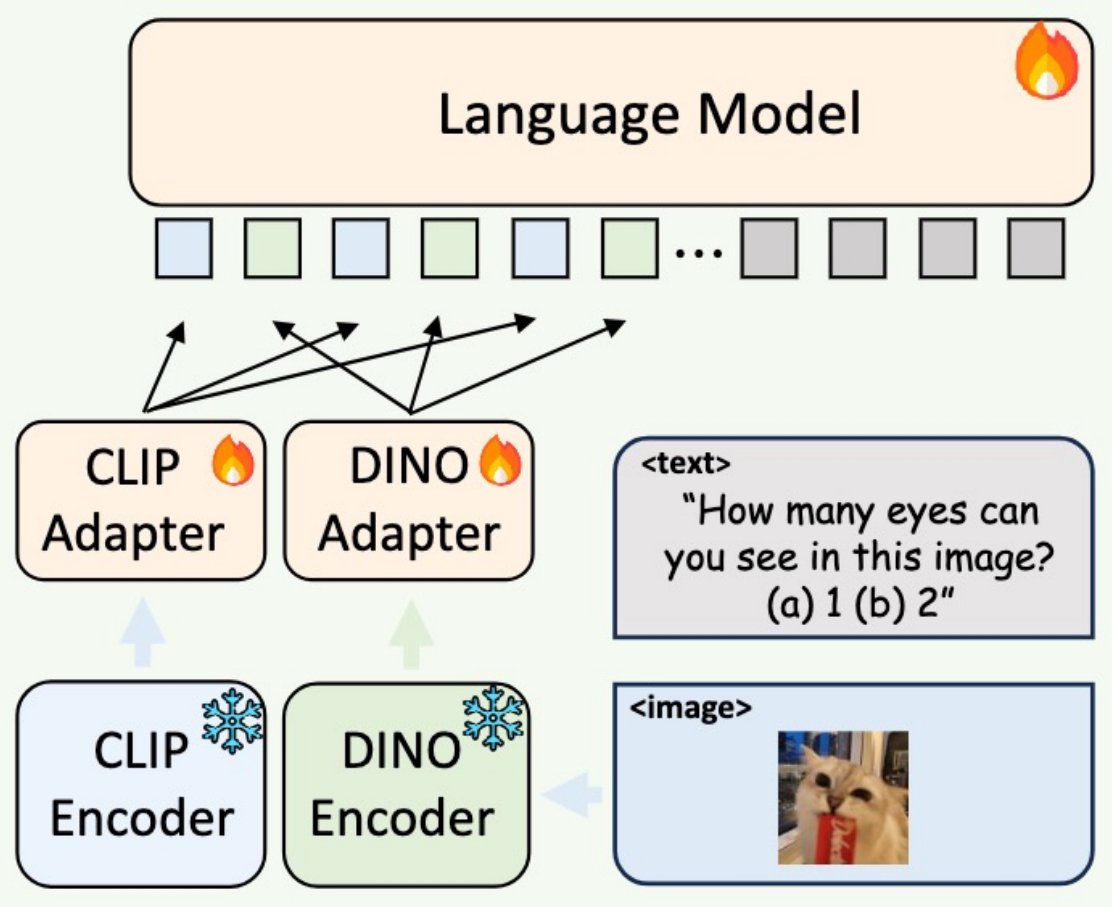
Zero-shot CLIP vs Few-shot linear probes

Zero-shot CLIP is as good as

- 4-shot linear-probe CLIP
- 16-shot BiT-M



Complementarity of Visual and Semantic Cues

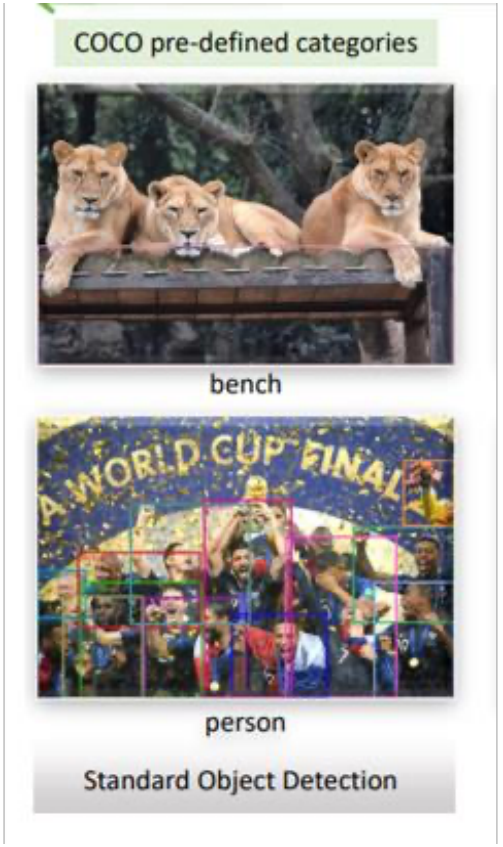


Tong et al. Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs. CVPR 2024.

Grounding + CLIP

Closed-set Detection

Detect objects of pre-defined classes



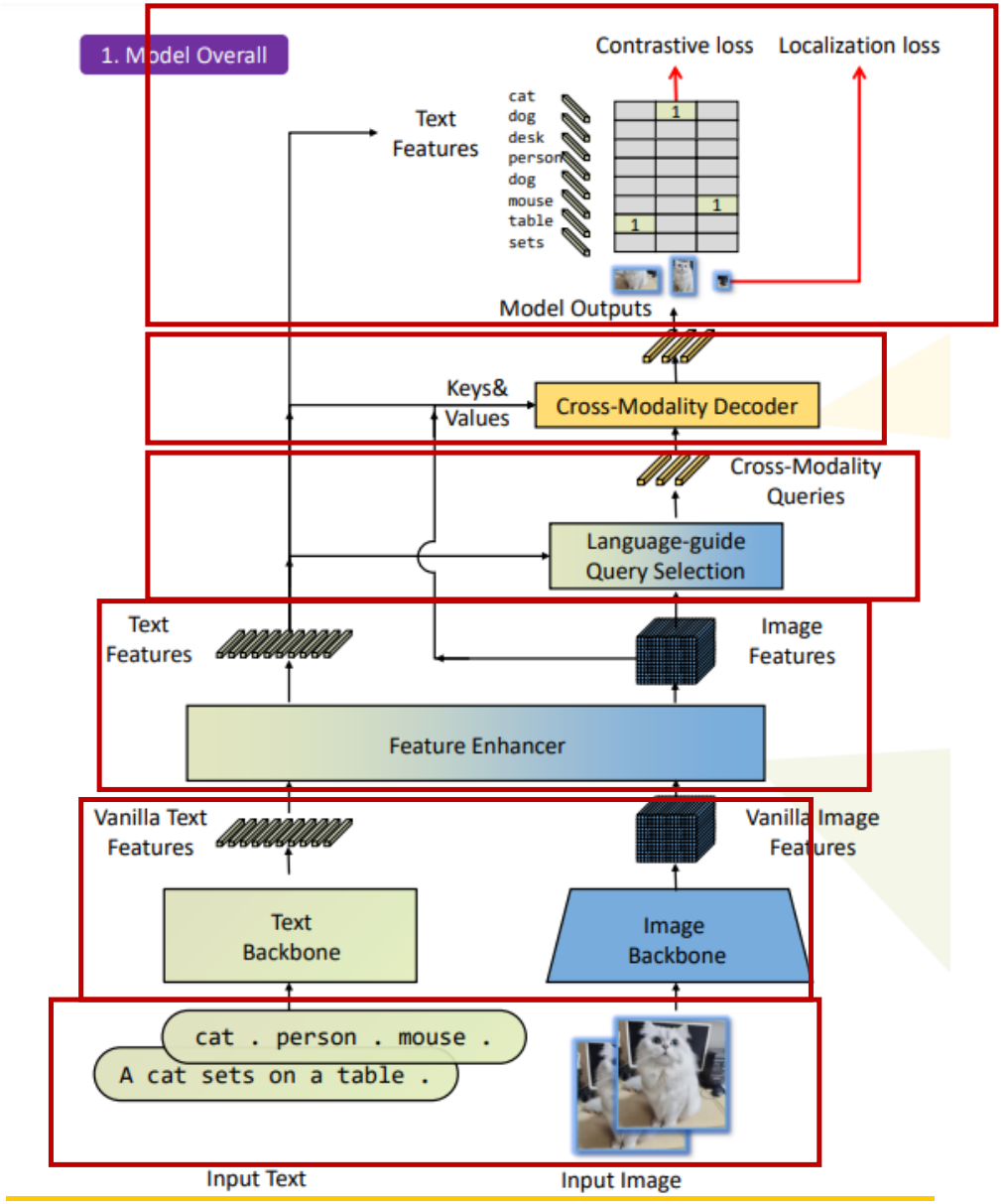
Open-set Detection

Detect objects that were not present in training data



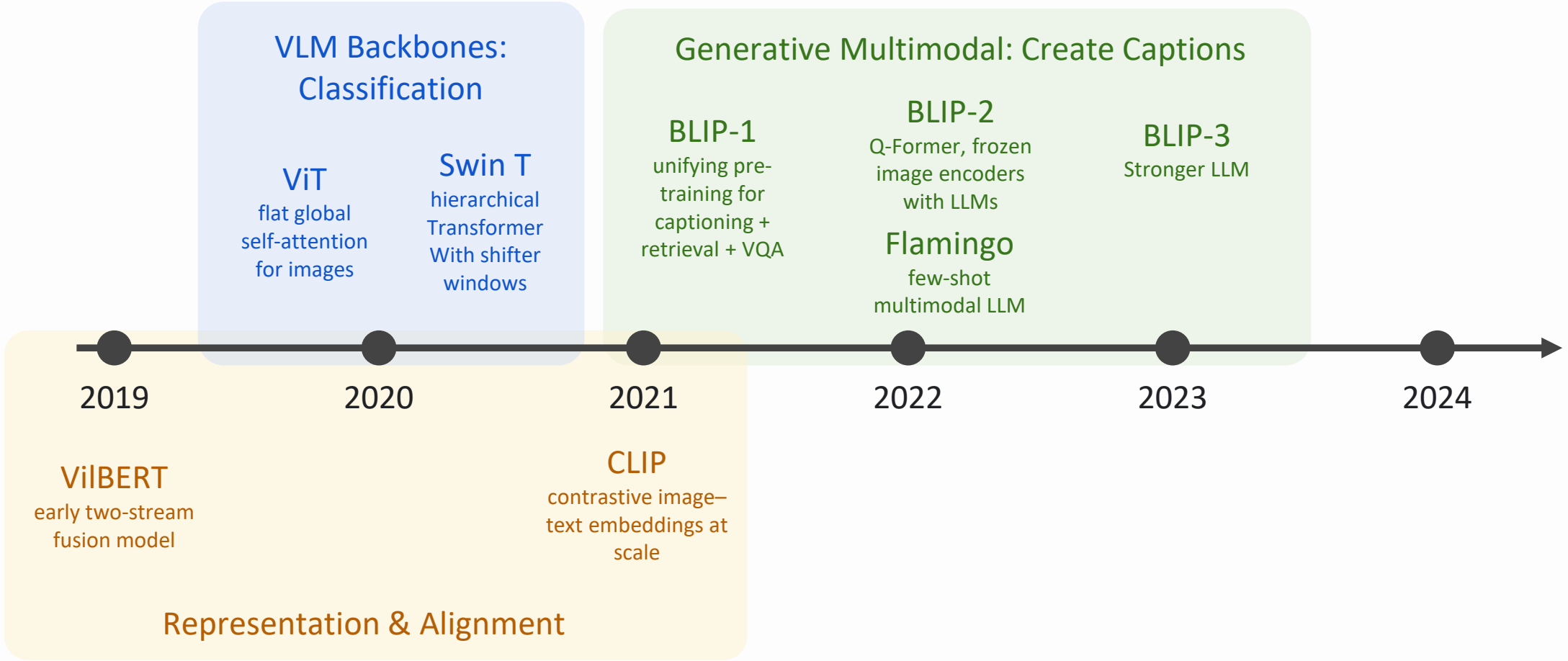
Grounding + CLIP

- Dual-encoder-single-decoder architecture
- Key Components :
  - Image Backbone
  - Text Backbone
  - Feature Enhancer
  - Language-Guided Query Selection
  - Cross-Modality Decoder



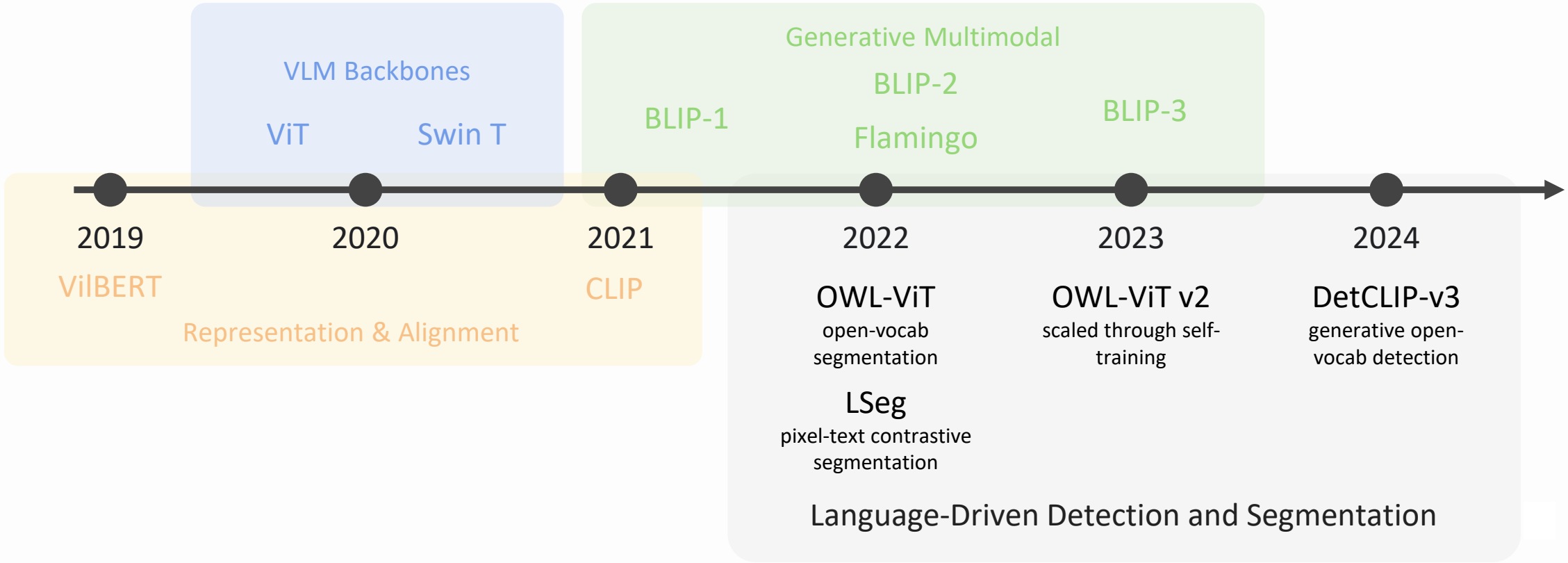
Grounding + CLIP

VLM Timeline: What have we seen so far?



Grounding + CLIP

VLM Timeline: What will we cover today?



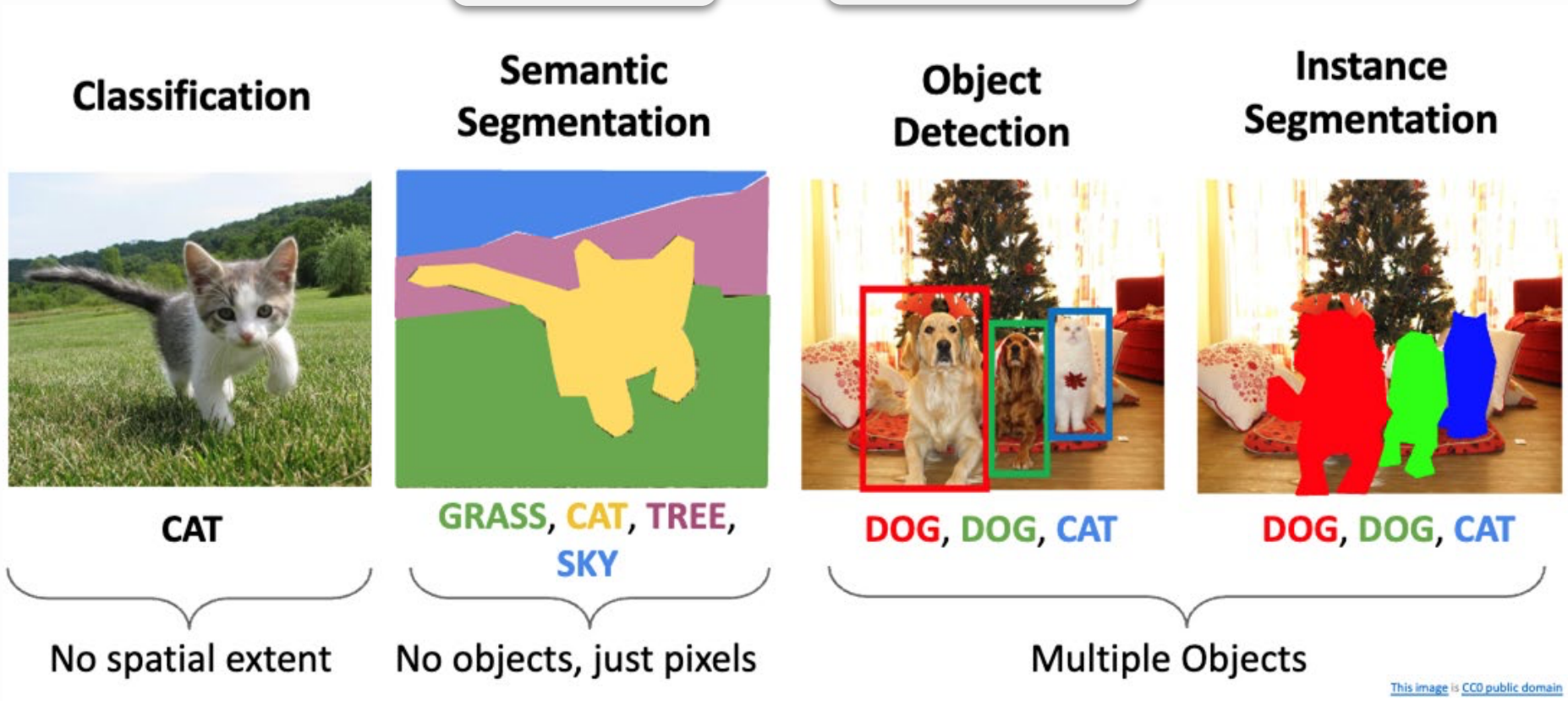
Grounding + CLIP

Computer Vision Tasks

OWL-ViT

LSeg

DetCLIP-v3



Source: UMich CV Lecture Slides

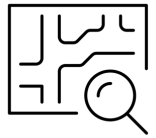
Grounding + CLIP

Visualization (OVD)



Slide inspired by Lewei Yao

Progress So Far – Part III

	Abstract Perception “what” (high-level, conceptual) 	**Dense Perception** “where” (low-level, spatial) 
Visual	Self-Supervised Learning (e.g., DINO)	1. Spatialization of Abstract Perception 2. Dense Supervision (e.g., SAM)
Semantic	Vision-Language Model (e.g., CLIP)	

Progress So Far – Part III

	Perception	Generation	Action
Vision	DINO, CLIP, SAM, etc.		
Language	BERT		

