CLIP and CLIP-based Models

Learning Transferable Visual Models From Natural Language Supervision

Alec Radford [*1] Jong Wook Kim [*1] Chris Hallacy [1] Aditya Ramesh [1] Gabriel Goh [1] Sandhini Agarwal [1]
Girish Sastry [1] Amanda Askell [1] Pamela Mishkin [1] Jack Clark [1] Gretchen Krueger [1] Ilya Sutskever [1]

OPENAI/ICML2021/Cite:4634

Zequn Zeng 2023.4.26

AIGC-Text2Image

CLIP： Bridge between Image and Text

vibrant portrait painting of Salvador Dalí with a robotic half face

a shiba inu wearing a beret and black turtleneck

a close up of a handpalm with leaves growing from it

an espresso machine that makes coffee from human souls, artstation

panda mad scientist mixing sparkling chemicals, artstation

a corgi's head depicted as an explosion of a nebula

From DALLE-2

OUTLINE

- Contrastive Language-Image Pre-training (CLIP)
- Visual Concepts in CLIP
- CLIP-based models for downstream tasks
- Limitations of current Multi-Modal Pretraining

Contrastive Language-Image Pre-training(CLIP)

- Natural language Supervision
- 400M Image text pairs
- Zero-shot Capability

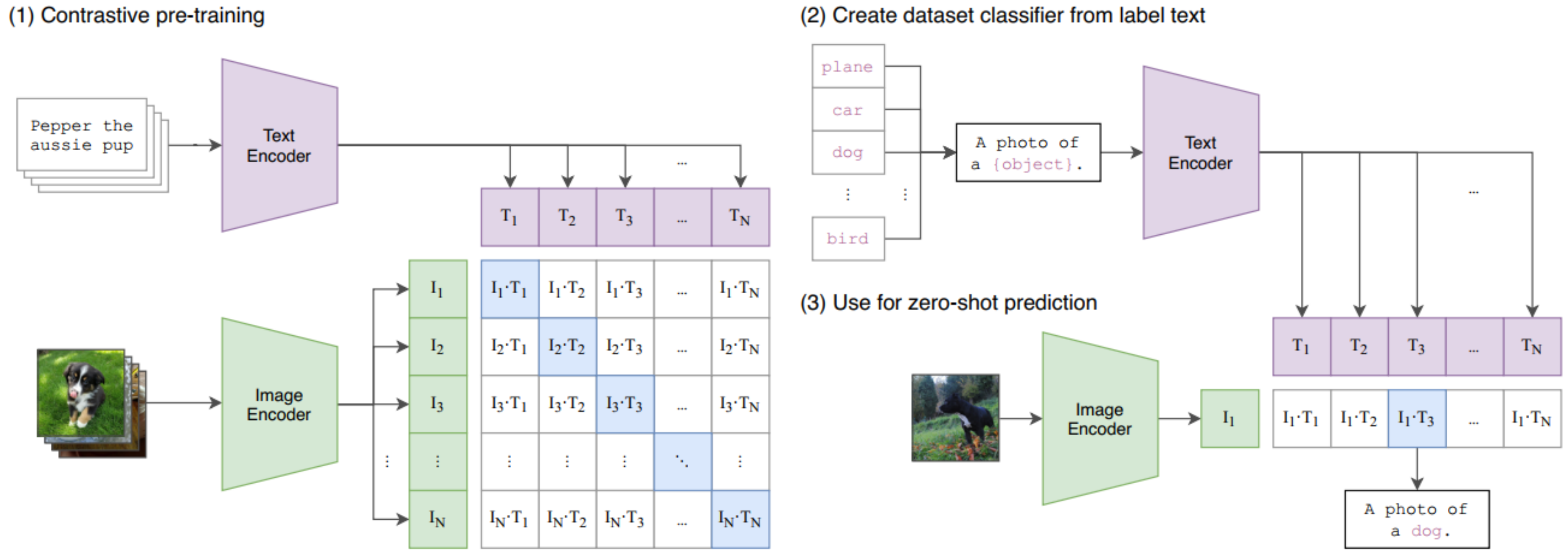
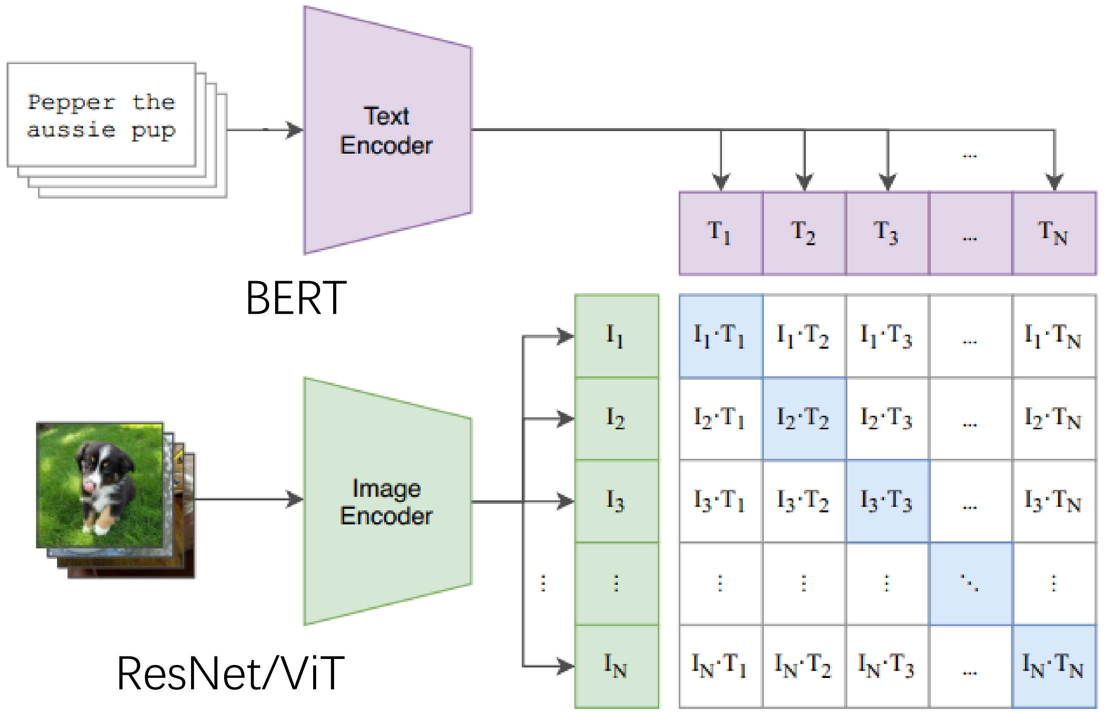


Figure 1. Summary of our approach. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the learned text encoder synthesizes a zero-shot linear classifier by embedding the names or descriptions of the target dataset's classes.

Contrastive Language-Image Pre-training(CLIP)

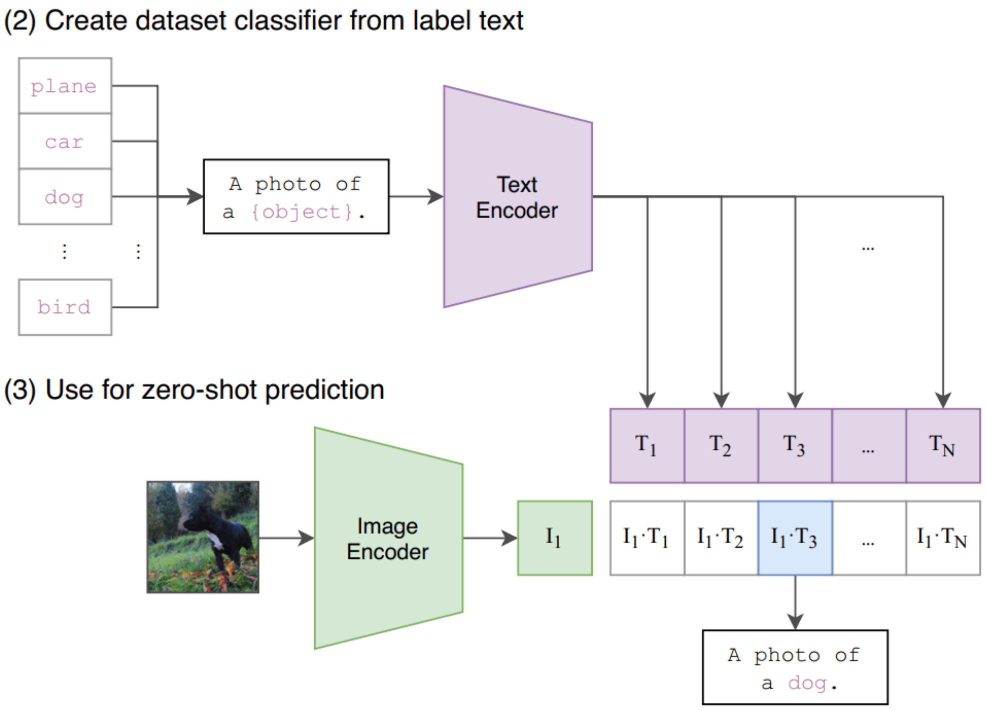
- Natural language Supervision vs Class label
  - Fine-grained label: Entity, attribute, relation



- 400M noisy Image-Text pairs
  - attempt to cover as broad a set of visual concepts as possible

Contrastive Language-Image Pre-training(CLIP)

Zero-shot Capability

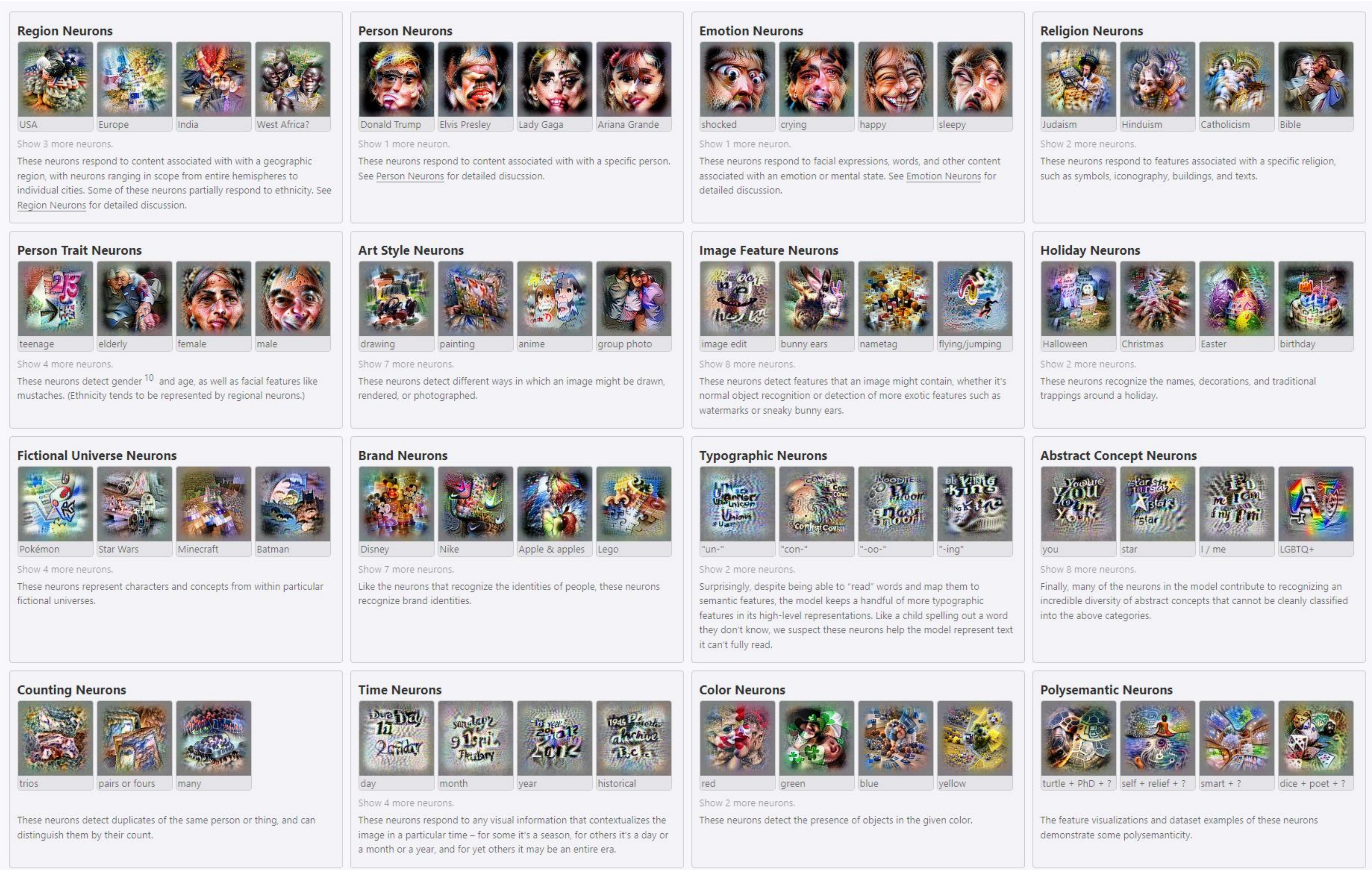


☐ Multimodal Neurons in Artificial Neural Networks (Visual Concepts in CLIP)

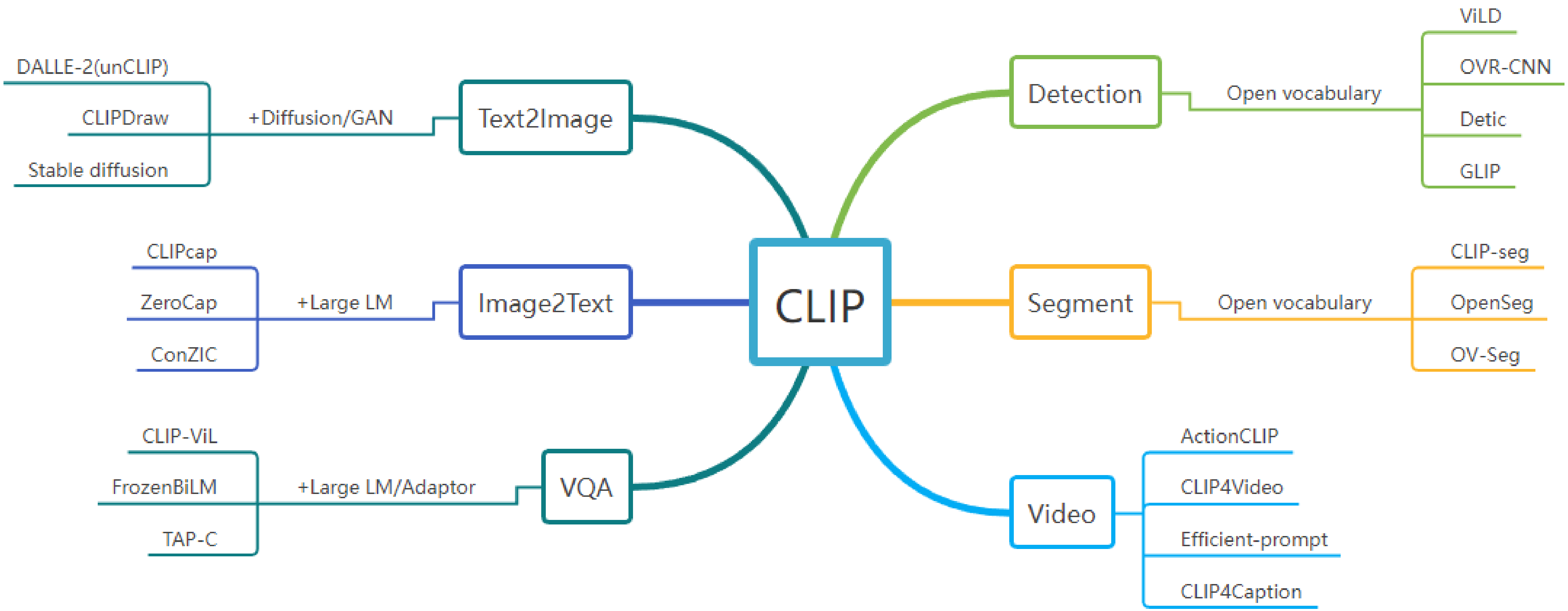
- Visual Concept with Multiple Visual Styles
- Neurons in CLIP Resnet layer

Biological Neuron		CLIP Neuron		Previous Artificial Neuron		
Probed via depth electrodes		Neuron 244 from penultimate layer in CLIP RN50_4x		Neuron 483, generic person detector from Inception v1		
Halle Berry		Spiderman		human face		
	Responds to photos of Halle Berry and Halle Berry in costume ✓	 ↻ view more	Responds to photos of Spiderman in costume and spiders ✓		Responds to faces of people ✓	Photorealistic images
	Responds to skeches of Halle Berry ✓	 ↻ view more	Responds to comics or drawings of Spiderman and spider-themed icons ✓		Does not respond significantly to drawings of faces ✗	Conceptual drawings
	Responds to the text "Halle Berry" ✓	 ↻ view more	Responds to the text "spider" and others ✓		Does not respond significantly to text ✗	Images of text

☐ Multimodal Neurons in Artificial Neural Networks (Visual Concepts in CLIP)



□CLIP-based models for downstream tasks



Cite:4634

Open-Vocabulary Detection: ViLD

- Base categories in Training data
- Novel categories not in Training data

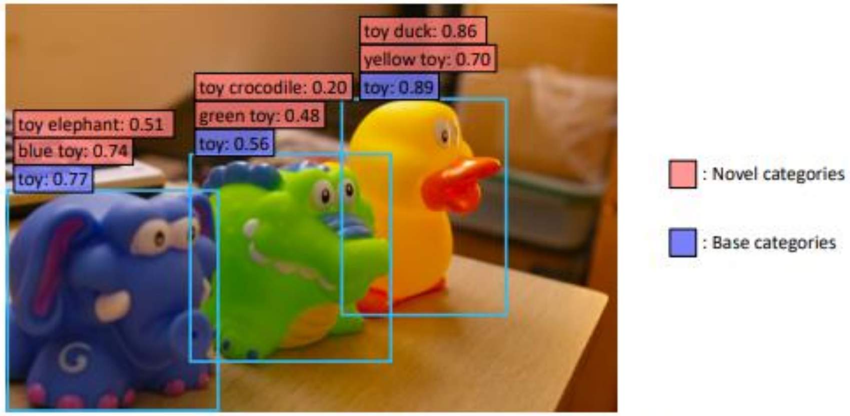
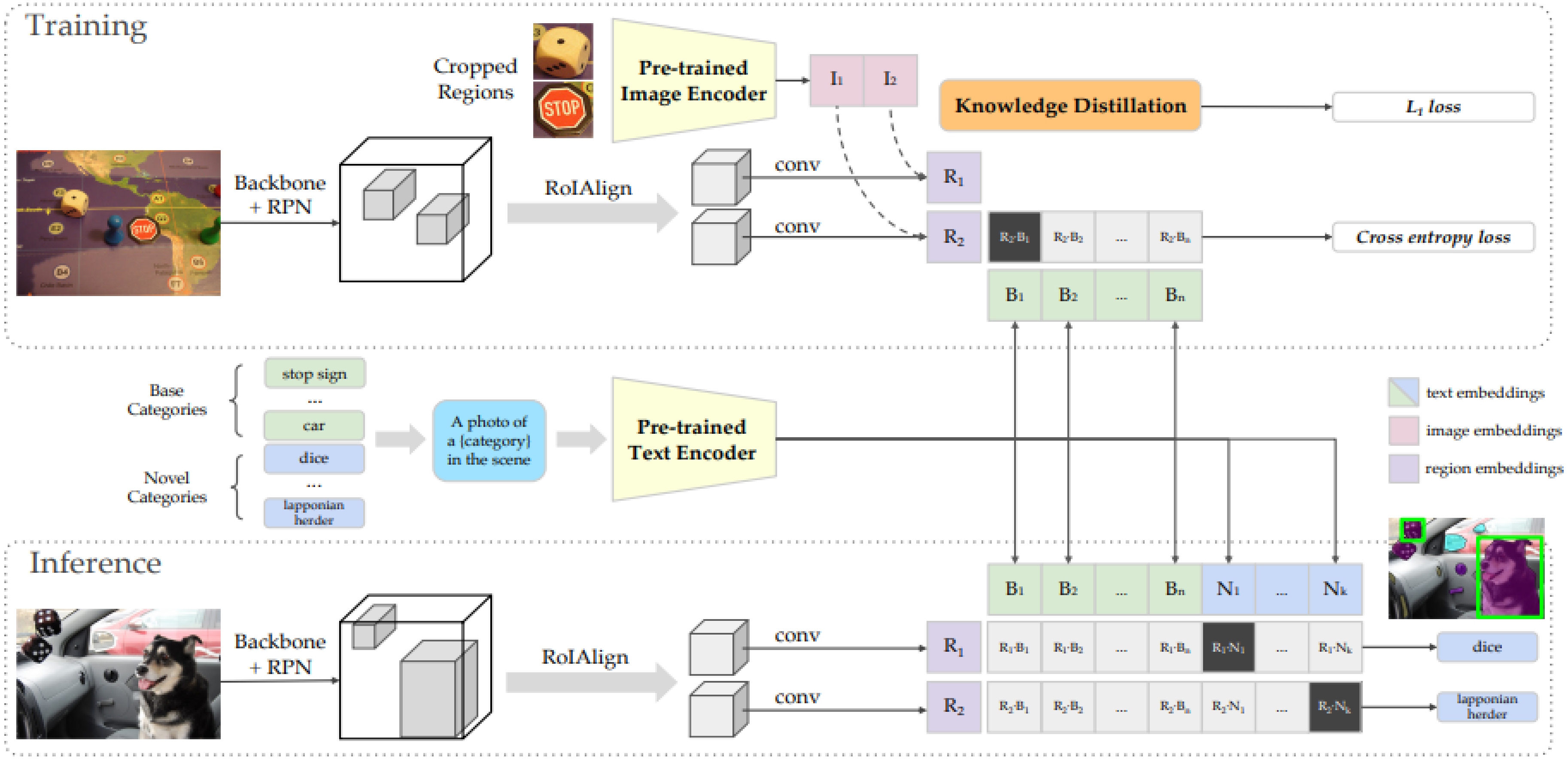


Figure 1: **An example of our open-vocabulary detector with arbitrary texts.** After training on base categories (purple), we can detect novel categories (pink) that are not present in the training data.

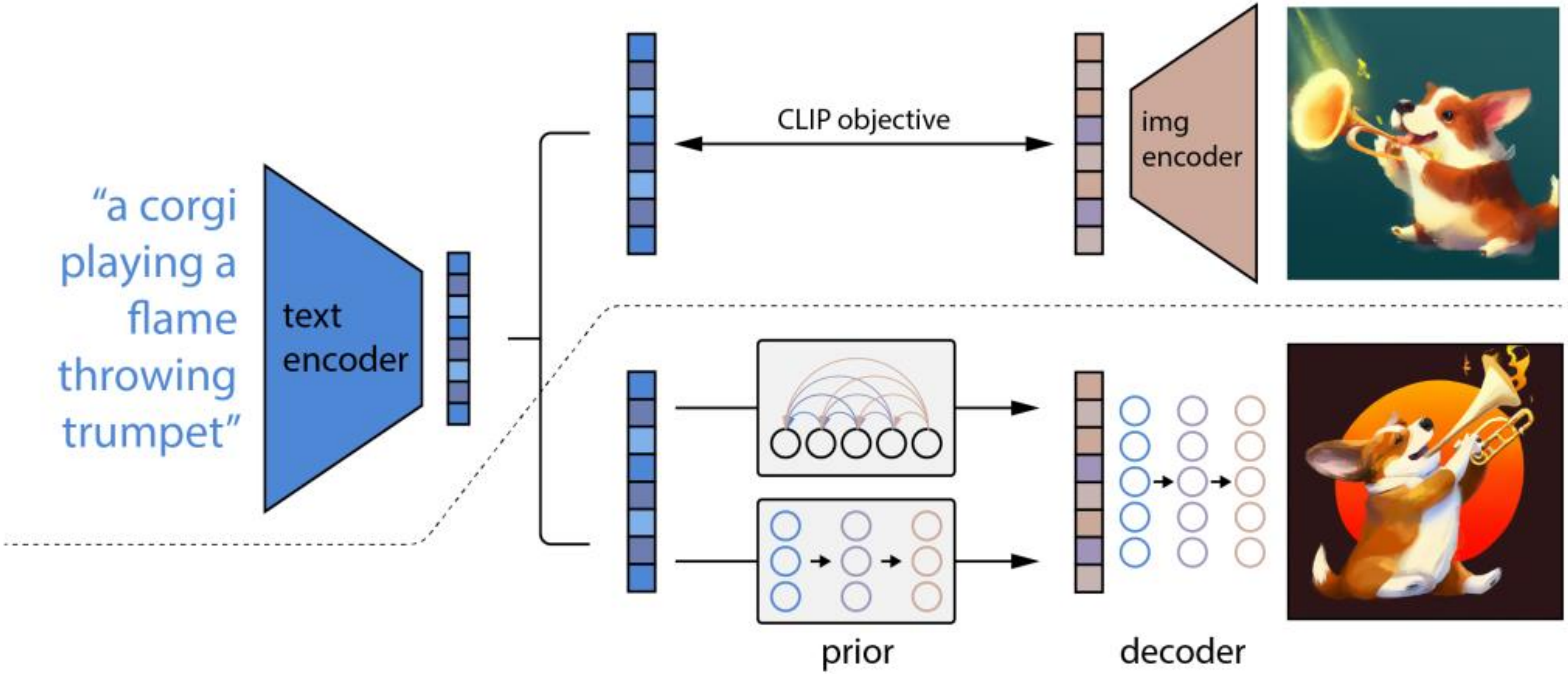
Open-Vocabulary Detection: ViLD

Framework



☐Text2Image: DALLE-2(unCLIP)

- Framework



□Text2Image: DALLE-2(unCLIP)

an espresso machine that makes coffee from human souls, artstation

panda mad scientist mixing sparkling chemicals, artstation

a corgi's head depicted as an explosion of a nebula

a dolphin in an astronaut suit on saturn, artstation

a propaganda poster depicting a cat dressed as french emperor napoleon holding a piece of cheese

a teddy bear on a skateboard in times square

Text2Image: DALLE-2(unCLIP)

- binding wrong attributes to objects (Reconstructions)

- wrong spelling

Figure 16: Samples from unCLIP for the prompt, “A sign that says deep learning.”

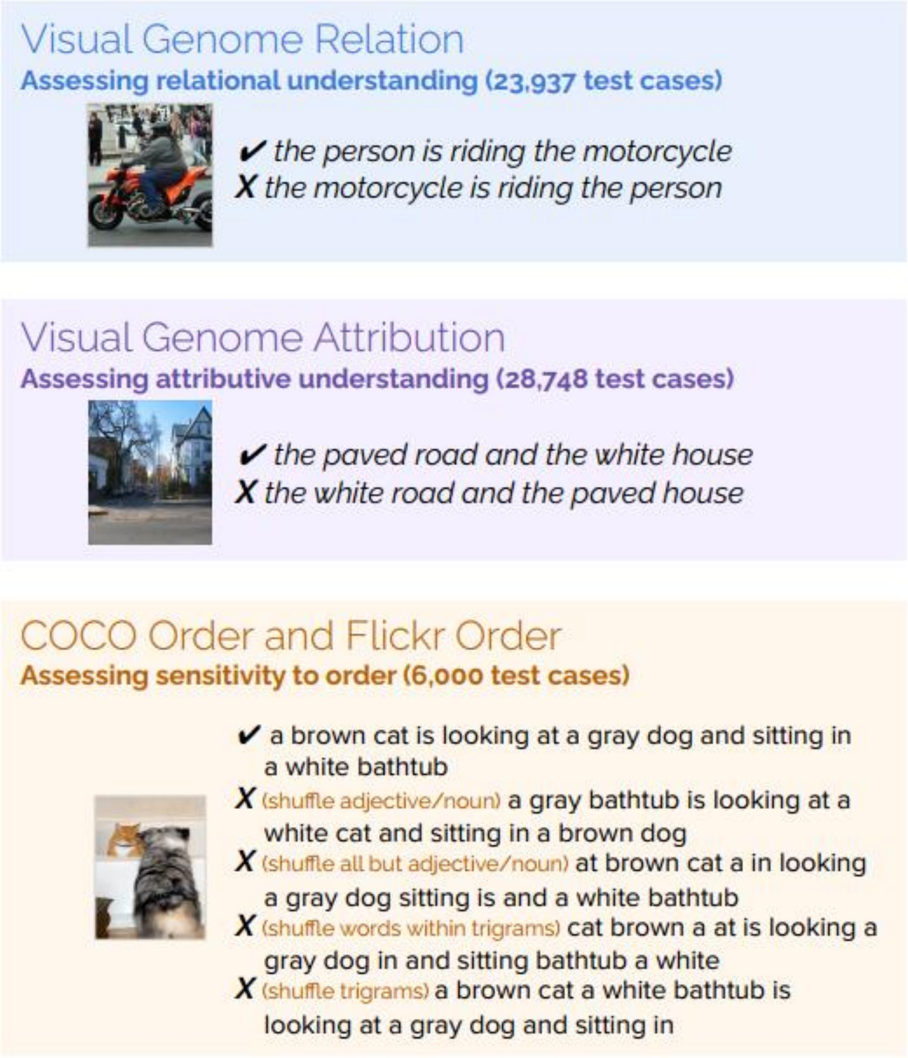
□Limitations of current Multi-Modal Pretraining(ICLR2023 oral)

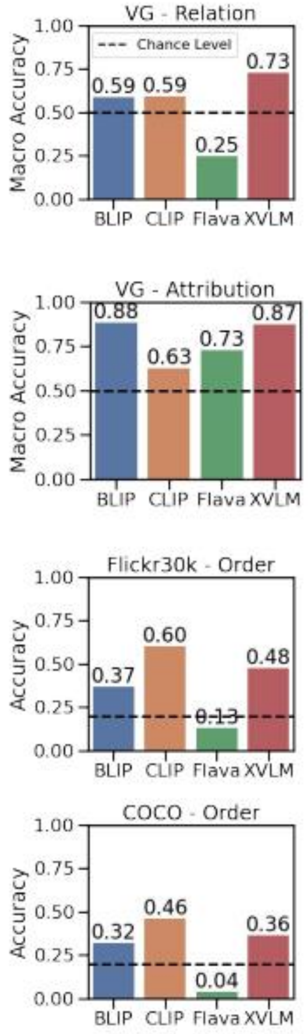
WHEN AND WHY VISION-LANGUAGE MODELS BEHAVE LIKE BAGS-OF-WORDS, AND WHAT TO DO ABOUT IT?

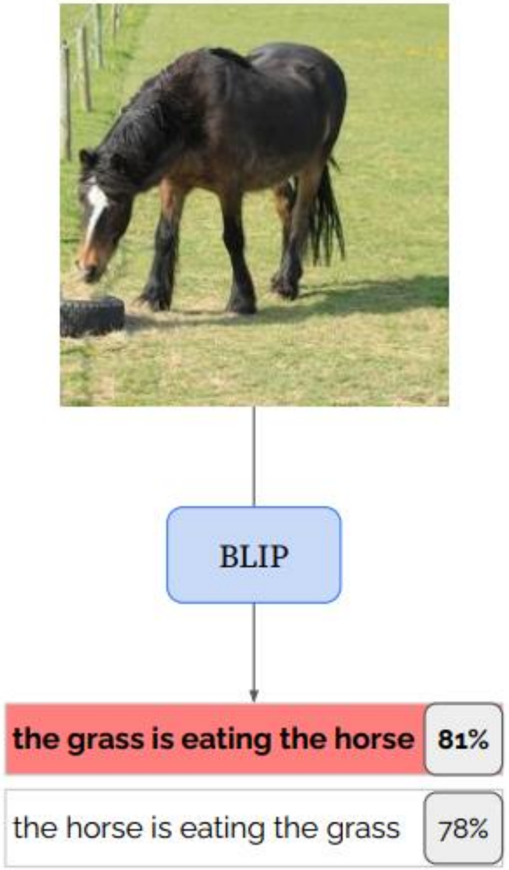
Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, James Zou
Stanford University
Stanford, CA 94305
{merty, fede, pkalluri, jurafsky, jamesz}@stanford.edu

- **Lack of Compositional Understanding and Order information**
  - **Attributes: understanding of objects' properties**
  - **Relation: relational understanding**
  - **Order: order sensitivity**

□Limitations of current Multi-Modal Pretraining(ICLR2023 oral)

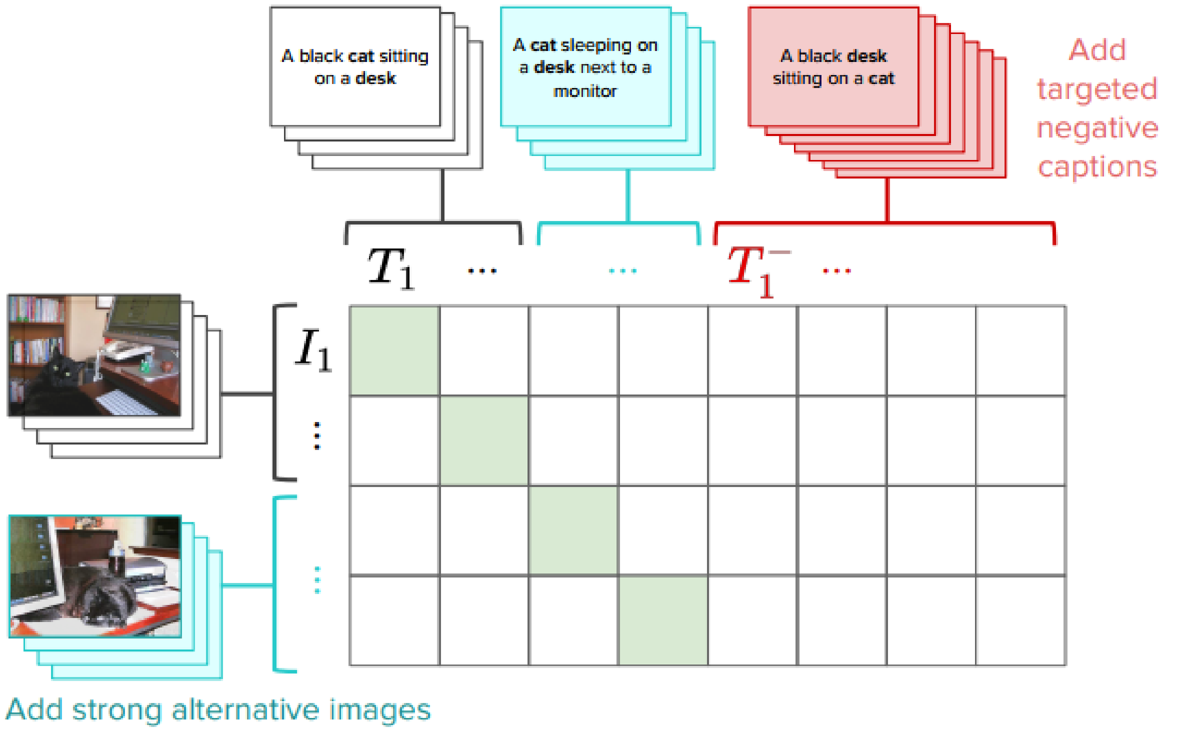


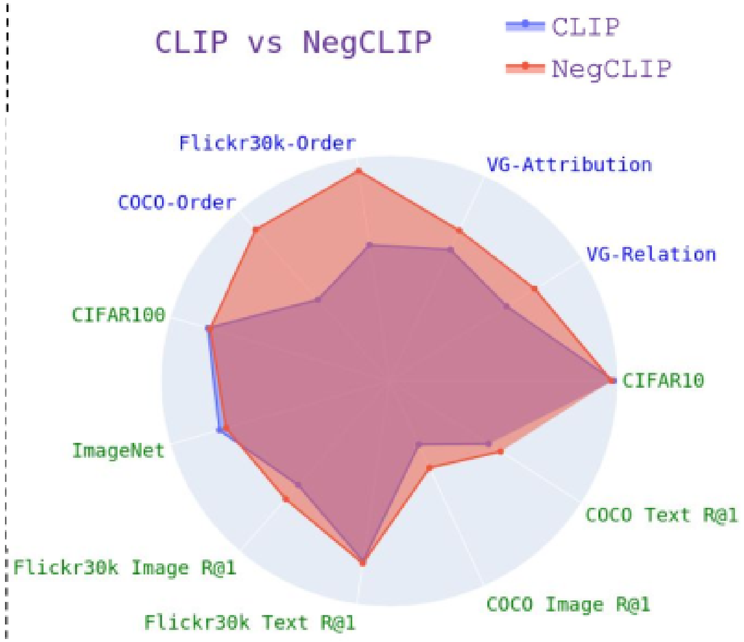


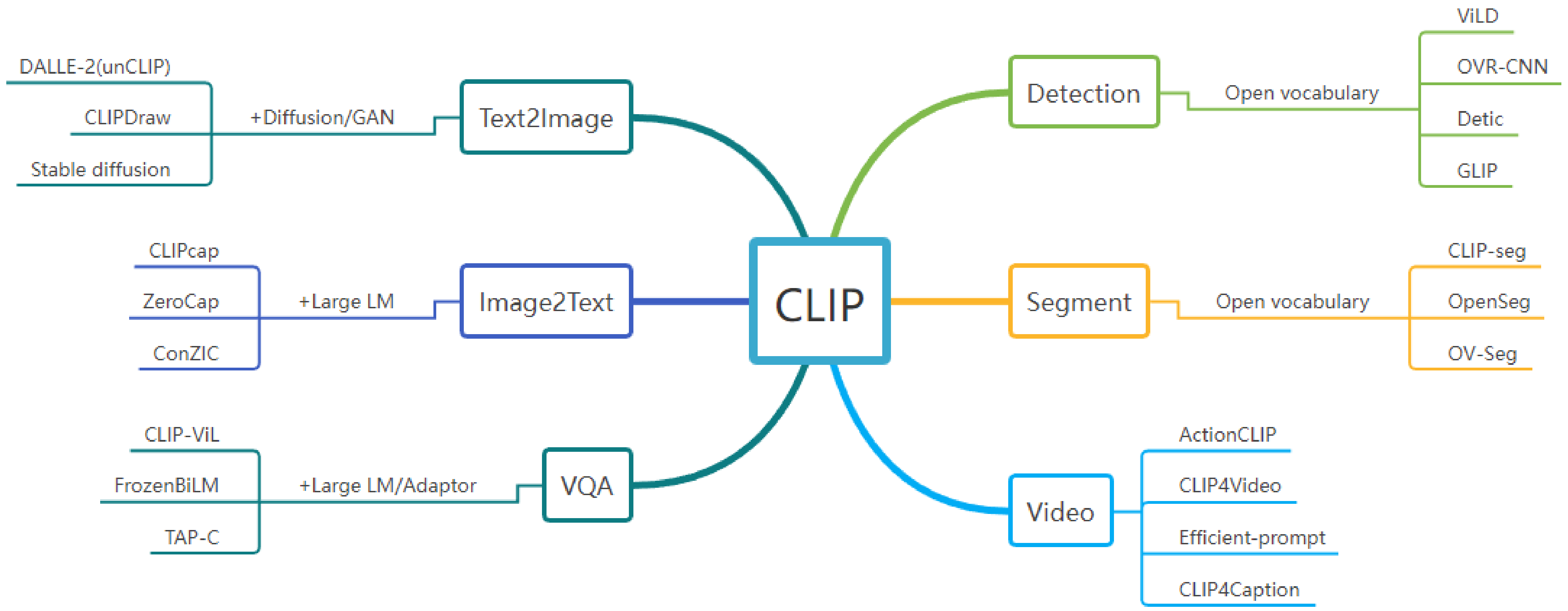


☐Limitation of current Multi-Modal Pretraining(ICLR2023 oral)

- Cause: shortcut of CL pretraining
  - Scheme: Compositional hard negatives







CLIP
Text2Image
+Diffusion/GAN
DALLE-2(unCLIP)
CLIPDraw
Stable diffusion
Image2Text
+Large LM
CLIPcap
ZeroCap
ConZIC
VQA
+Large LM/Adaptor
CLIP-ViL
FrozenBiLM
TAP-C
Detection
Open vocabulary
ViLD
OVR-CNN
Detic
GLIP
Segment
Open vocabulary
CLIP-seg
OpenSeg
OV-Seg
Video
ActionCLIP
CLIP4Video
Efficient-prompt
CLIP4Caption

Q&A

Thanks for your listening!