

🔗 CLIP and CLIP-based Models

Learning Transferable Visual Models From Natural Language Supervision

Alec Radford^{*1} Jong Wook Kim^{*1} Chris Hallacy¹ Aditya Ramesh¹ Gabriel Goh¹ Sandhini Agarwal¹
Girish Sastry¹ Amanda Askell¹ Pamela Mishkin¹ Jack Clark¹ Gretchen Krueger¹ Ilya Sutskever¹

OPENAI/ICML2021/Cite:4634

🧠 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

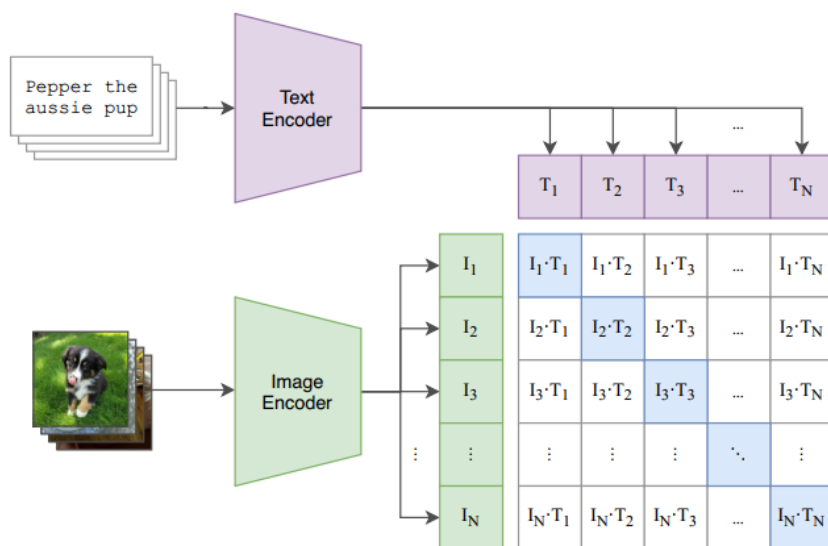
OUTLINE

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

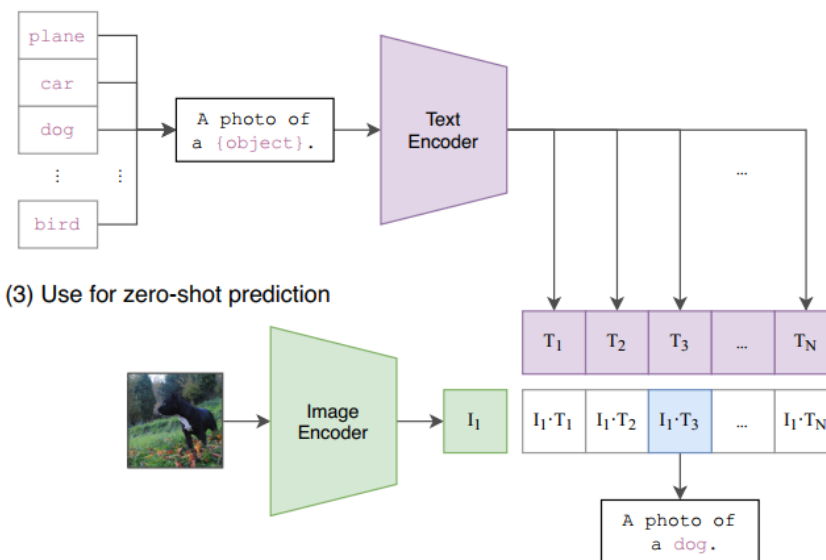
□ Contrastive Language-Image Pre-training (CLIP)

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

(1) Contrastive pre-training



(2) Create dataset classifier from label text



(3) Use for zero-shot prediction

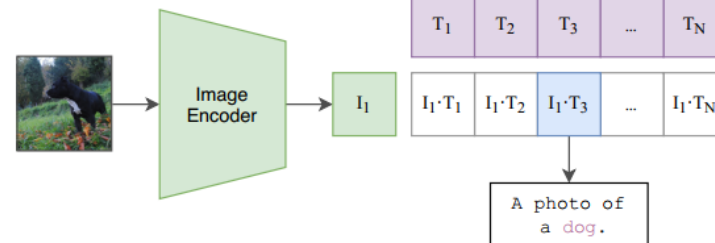
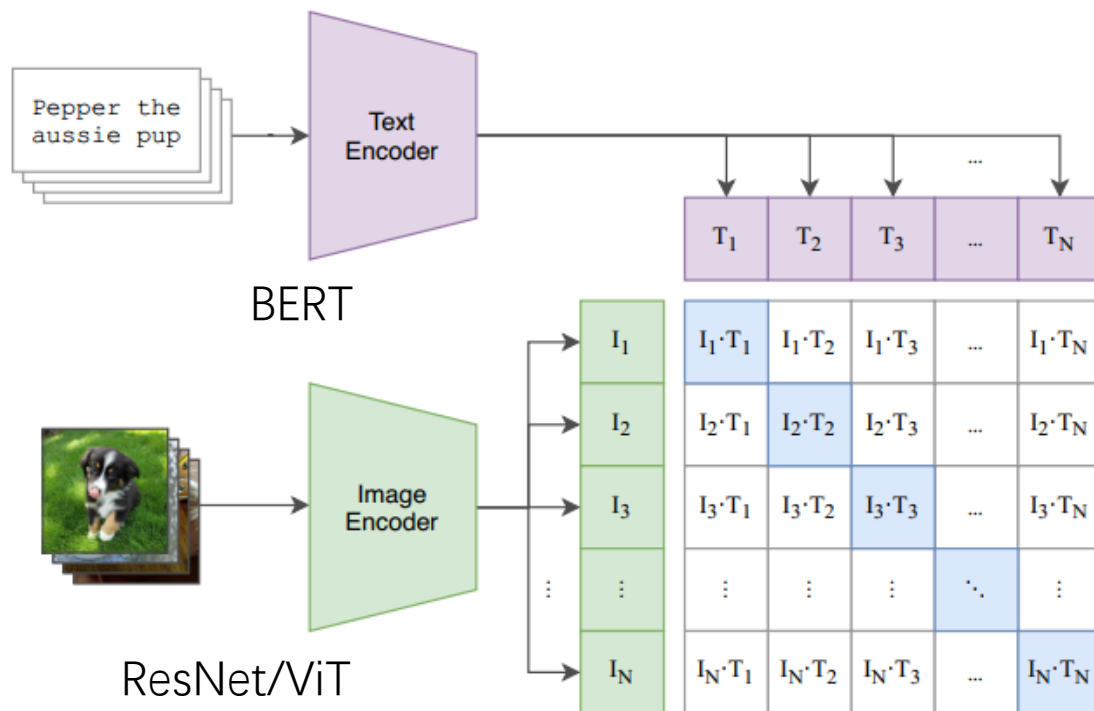


Figure 1. Summary of our approach. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the 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



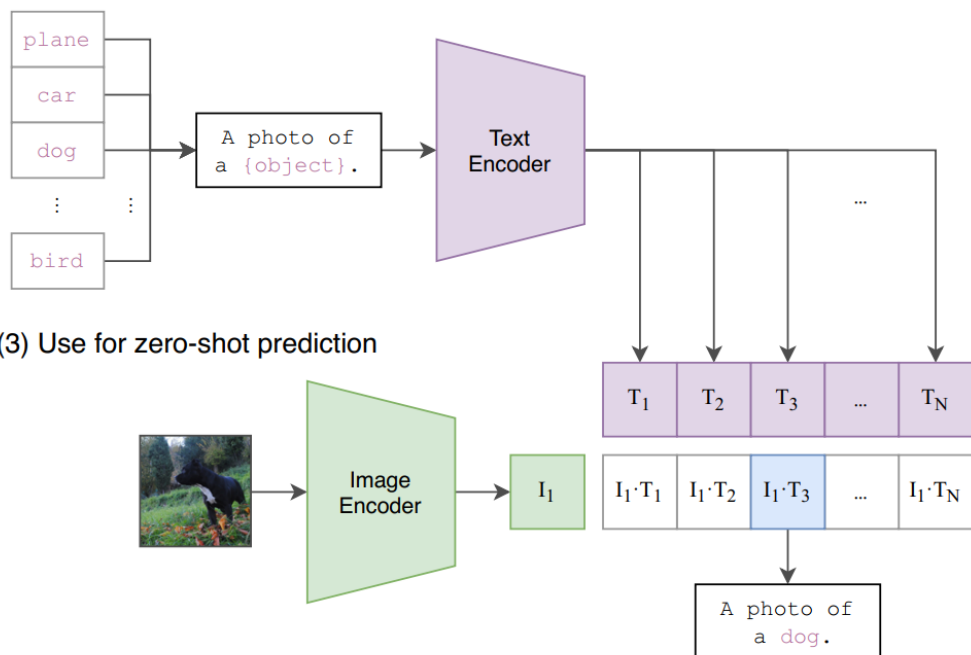
★ ★ 400M noisy Image-Text pairs

• attempt to cover as broad a set of visual concepts as possible

Contrastive Language-Image Pre-training (CLIP)

★ Zero-shot Capability

(2) Create dataset classifier from label text












(3) Use for zero-shot prediction

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%	

□ 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 ✓	 Responds to photos of Spiderman in costume and spiders ✓ view more	 Responds to faces of people ✓	Photorealistic images
 Responds to sketches of Halle Berry ✓	 Responds to comics or drawings of Spiderman and spider-themed icons ✓ view more	 Does not respond significantly to drawings of faces ✗	Conceptual drawings
 Responds to the text "Halle Berry" ✓	 Responds to the text "spider" and others ✓ view more	 Does not respond significantly to text ✗	Images of text

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

Region Neurons



USA Europe India West Africa?

Show 3 more neurons.

These neurons respond to content associated with a geographic region, with neurons ranging in scope from entire hemispheres to individual cities. Some of these neurons partially respond to ethnicity. See [Region Neurons](#) for detailed discussion.

Person Neurons



Donald Trump Elvis Presley Lady Gaga Ariana Grande

Show 1 more neuron.

These neurons respond to content associated with a specific person. See [Person Neurons](#) for detailed discussion.

Emotion Neurons



shocked crying happy sleepy

Show 1 more neuron.

These neurons respond to facial expressions, words, and other content associated with an emotion or mental state. See [Emotion Neurons](#) for detailed discussion.

Religion Neurons



Judaism Hinduism Catholicism Bible

Show 2 more neurons.

These neurons respond to features associated with a specific religion, such as symbols, iconography, buildings, and texts.

Person Trait Neurons



teenage elderly female male

Show 4 more neurons.

These neurons detect gender¹⁰ and age, as well as facial features like mustaches. (Ethnicity tends to be represented by regional neurons.)

Art Style Neurons



drawing painting anime group photo

Show 7 more neurons.

These neurons detect different ways in which an image might be drawn, rendered, or photographed.

Image Feature Neurons



image edit bunny ears nametag flying/jumping

Show 8 more neurons.

These neurons detect features that an image might contain, whether it's normal object recognition or detection of more exotic features such as watermarks or sneaky bunny ears.

Holiday Neurons



Halloween Christmas Easter birthday

Show 2 more neurons.

These neurons recognize the names, decorations, and traditional trappings around a holiday.

Fictional Universe Neurons



Pokémon Star Wars Minecraft Batman

Show 4 more neurons.

These neurons represent characters and concepts from within particular fictional universes.

Brand Neurons



Disney Nike Apple & apples Lego

Show 7 more neurons.

Like the neurons that recognize the identities of people, these neurons recognize brand identities.

Typographic Neurons



"un-" "con-" "-oo-" "-ing"

Show 2 more neurons.

Surprisingly, despite being able to "read" words and map them to semantic features, the model keeps a handful of more typographic features in its high-level representations. Like a child spelling out a word they don't know, we suspect these neurons help the model represent text it can't fully read.

Abstract Concept Neurons



you star I / me LGBTQ+

Show 8 more neurons.

Finally, many of the neurons in the model contribute to recognizing an incredible diversity of abstract concepts that cannot be cleanly classified into the above categories.

Counting Neurons



trios pairs or fours many

Show 4 more neurons.

These neurons detect duplicates of the same person or thing, and can distinguish them by their count.

Time Neurons



day month year historical

Show 4 more neurons.

These neurons respond to any visual information that contextualizes the image in a particular time – for some it's a season, for others it's a day or a month or a year, and for yet others it may be an entire era.

Color Neurons



red green blue yellow

Show 2 more neurons.

These neurons detect the presence of objects in the given color.

Polysemantic Neurons

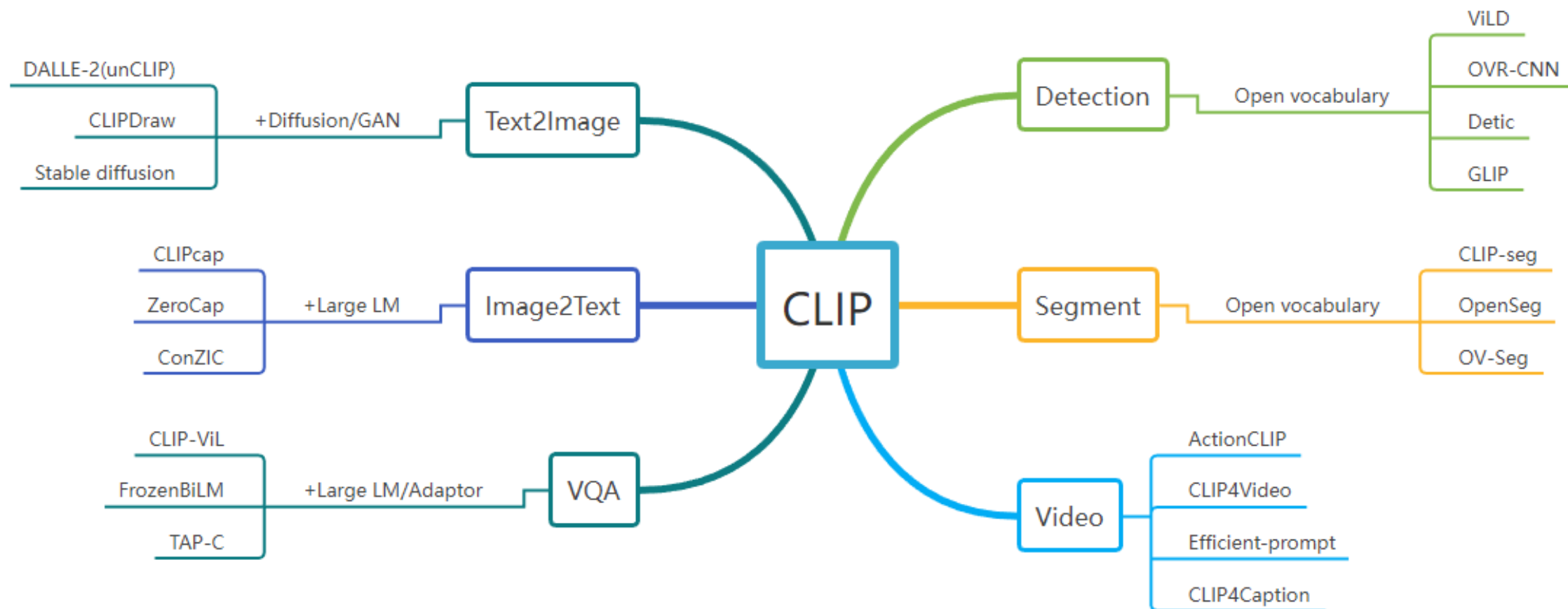


turtle + PhD + ? self + relief + ? smart + ? dice + poet + ?

Show 2 more neurons.

The feature visualizations and dataset examples of these neurons demonstrate some polysemanticity.

CLIP-based models for downstream tasks



Cite:4634

□ Open-Vocabulary Detection: ViLD

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

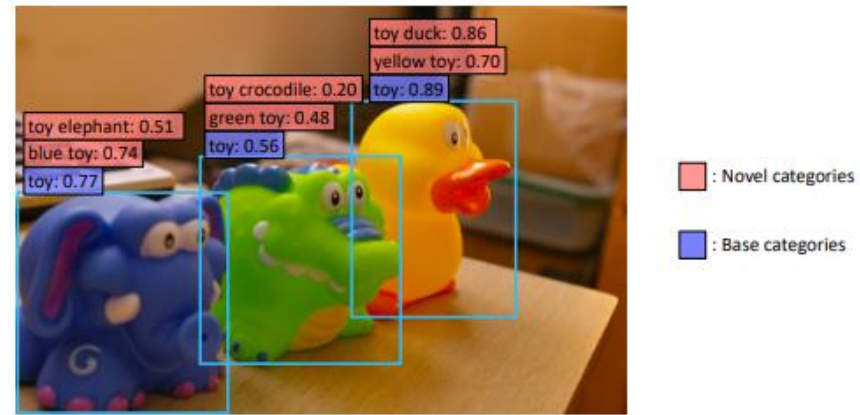
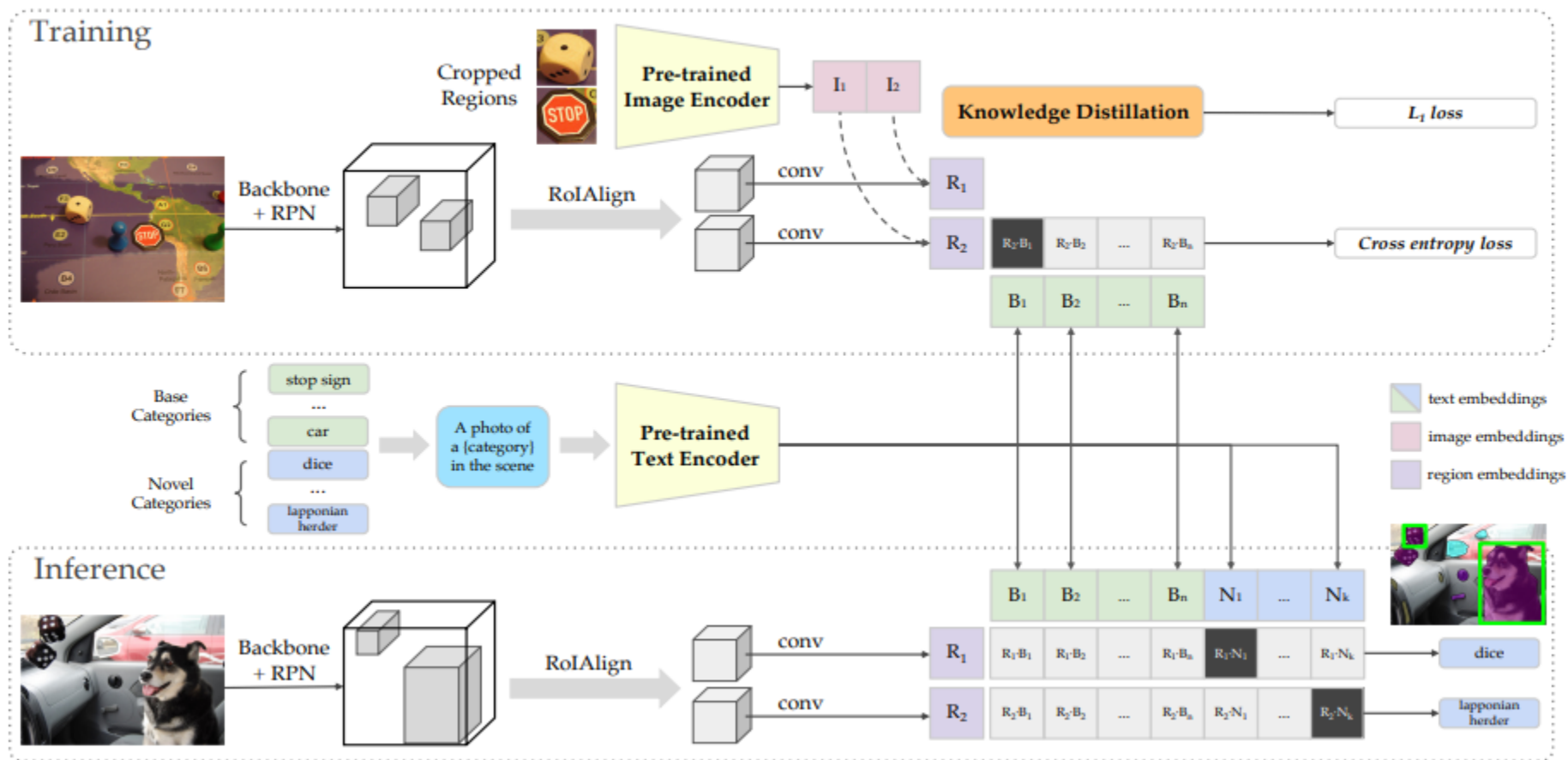


Figure 1: **An example of our open-vocabulary detector with arbitrary texts.** After training on base categories (purple), we can detect novel categories (pink) that are not present in the training data.

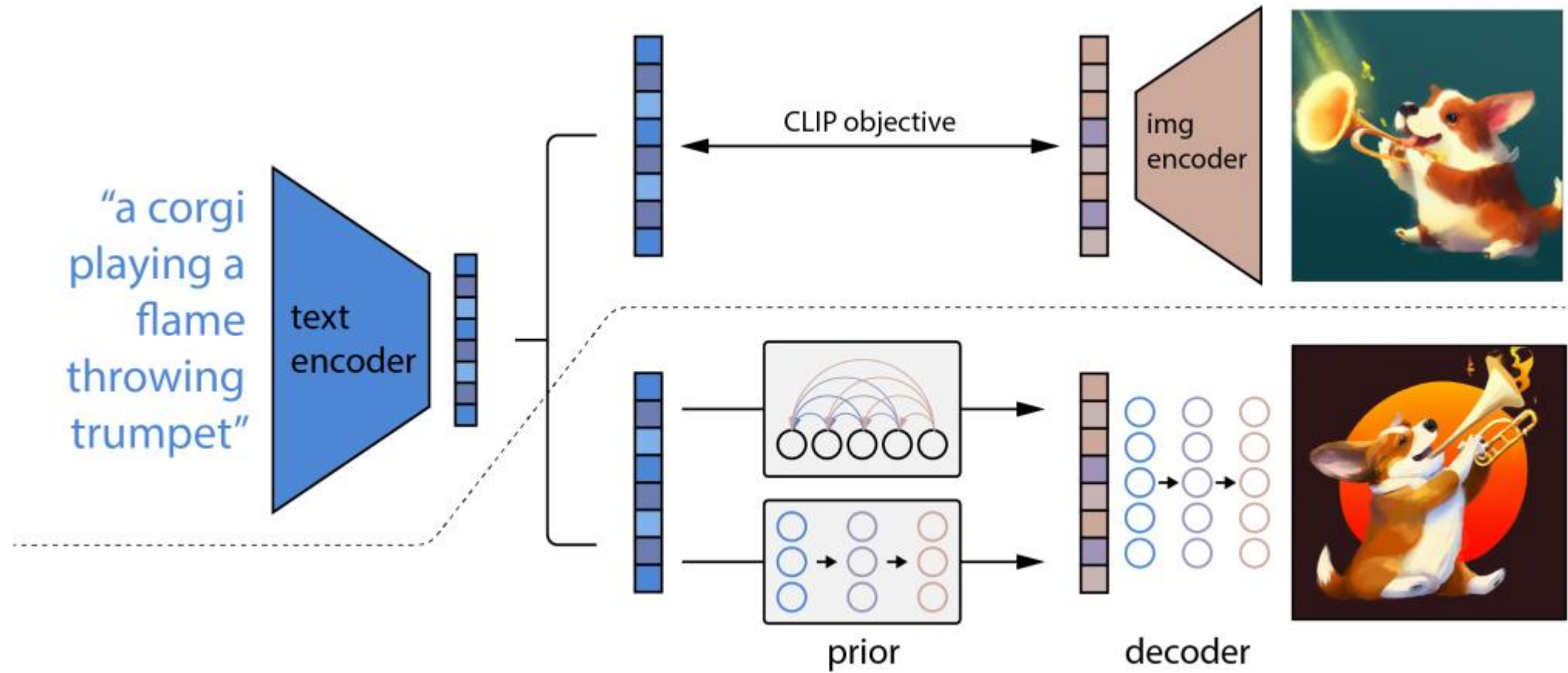
□ Open-Vocabulary Detection: ViLD

☆☆ Framework



□Text2Image: DALLE-2(unCLIP)

✧✧ Framework



□Text2Image: DALLE-2(unCLIP)



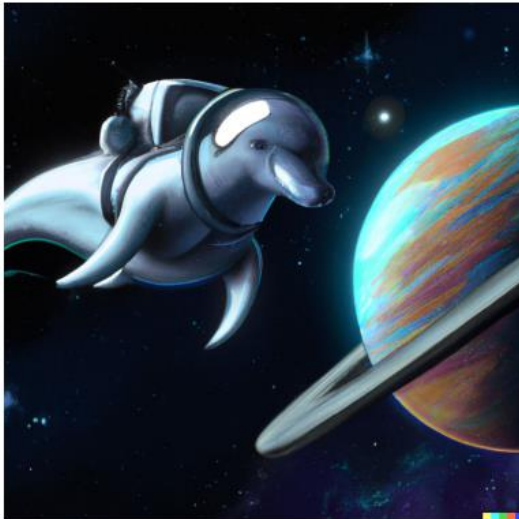
an espresso machine that makes coffee from human souls, artstation



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: DALLÉ-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 BE-
HAVE 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

- ① Attributes: understanding of objects' properties
- ① Relation: relational understanding
- ① Order: order sensitivity

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

Visual Genome Relation

Assessing relational understanding (23,937 test cases)



- ✓ the person is riding the motorcycle
- ✗ the motorcycle is riding the person

Visual Genome Attribution

Assessing attributive understanding (28,748 test cases)



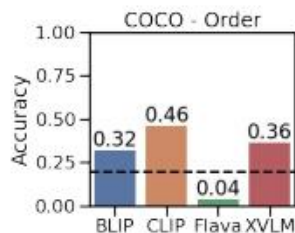
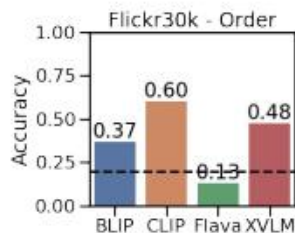
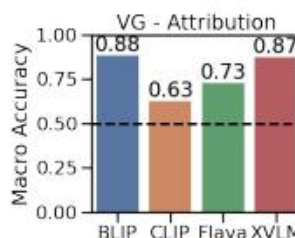
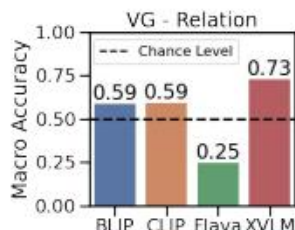
- ✓ the paved road and the white house
- ✗ the white road and the paved house

COCO Order and Flickr Order

Assessing sensitivity to order (6,000 test cases)



- ✓ a brown cat is looking at a gray dog and sitting in a white bathtub
- ✗ (shuffle adjective/noun) a gray bathtub is looking at a white cat and sitting in a brown dog
- ✗ (shuffle all but adjective/noun) at brown cat a in looking a gray dog sitting is and a white bathtub
- ✗ (shuffle words within trigrams) cat brown a at is looking a gray dog in and sitting bathtub a white
- ✗ (shuffle trigrams) a brown cat a white bathtub is looking at a gray dog and sitting in



BLIP

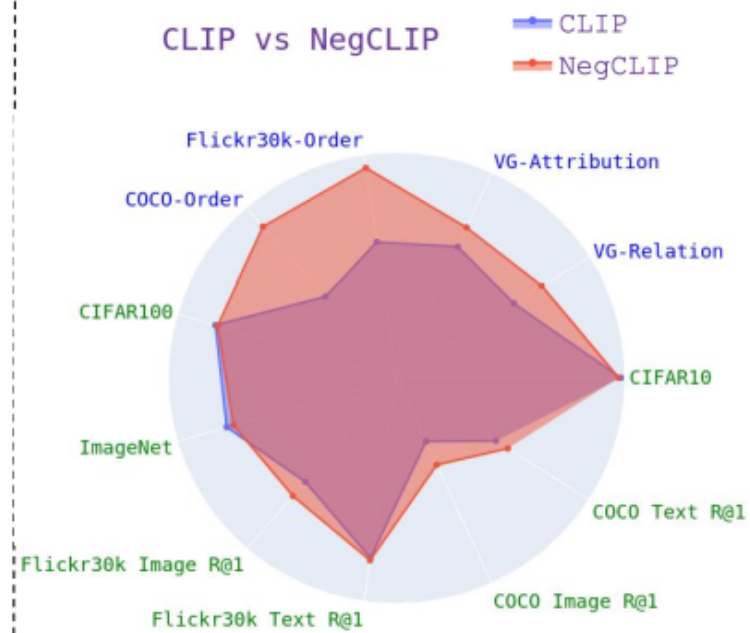
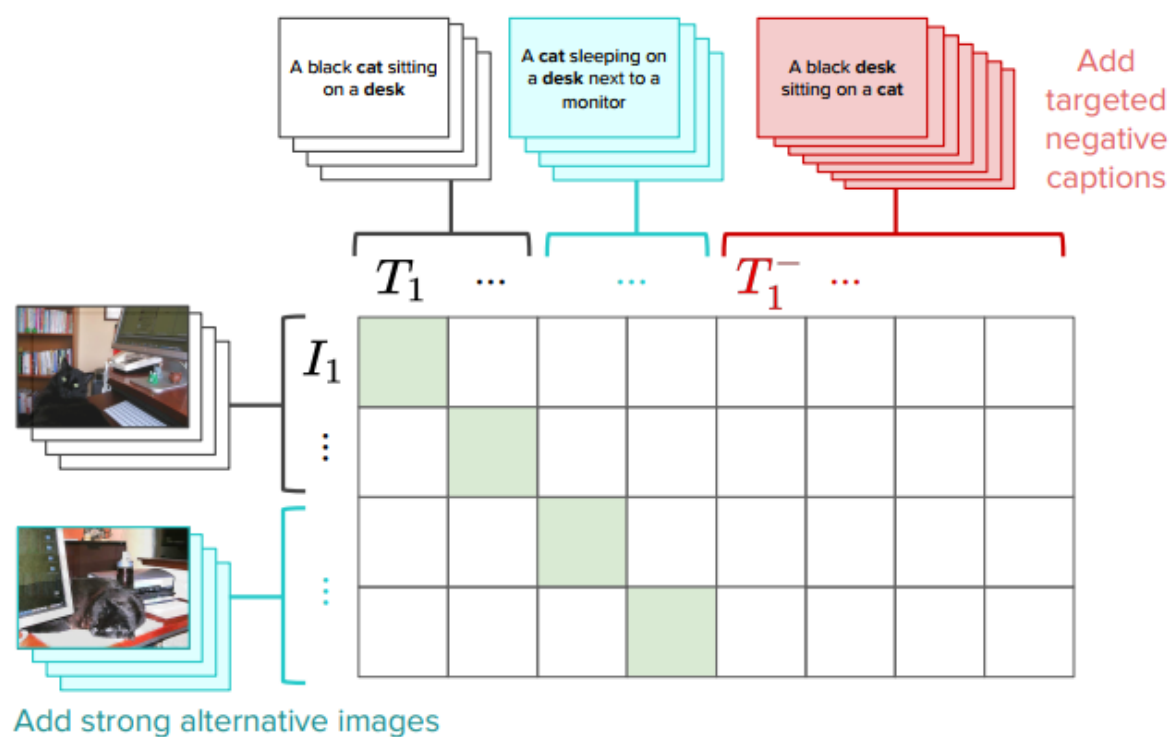
the grass is eating the horse 81%

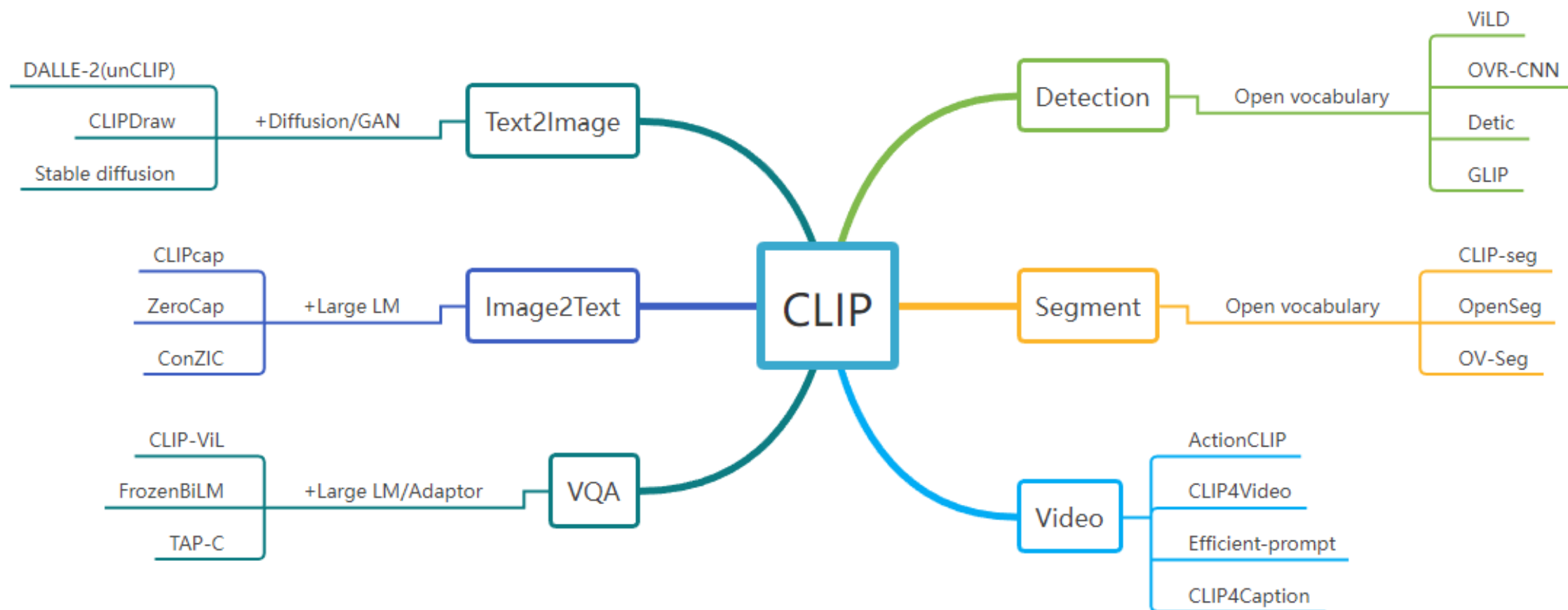
the horse is eating the grass 78%

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

☆☆ Cause: shortcut of CL pretraining

☆☆ Scheme: Compositional hard negatives





Q&A

Thanks for your listening!