Foundations of Vision-Language Models: Concepts and Roadmap

Kaiyang Zhou

香港浸會大學
HONG KONG BAPTIST UNIVERSITY

DEPARTMENT OF
COMPUTER SCIENCE
計算機科學系

Outline

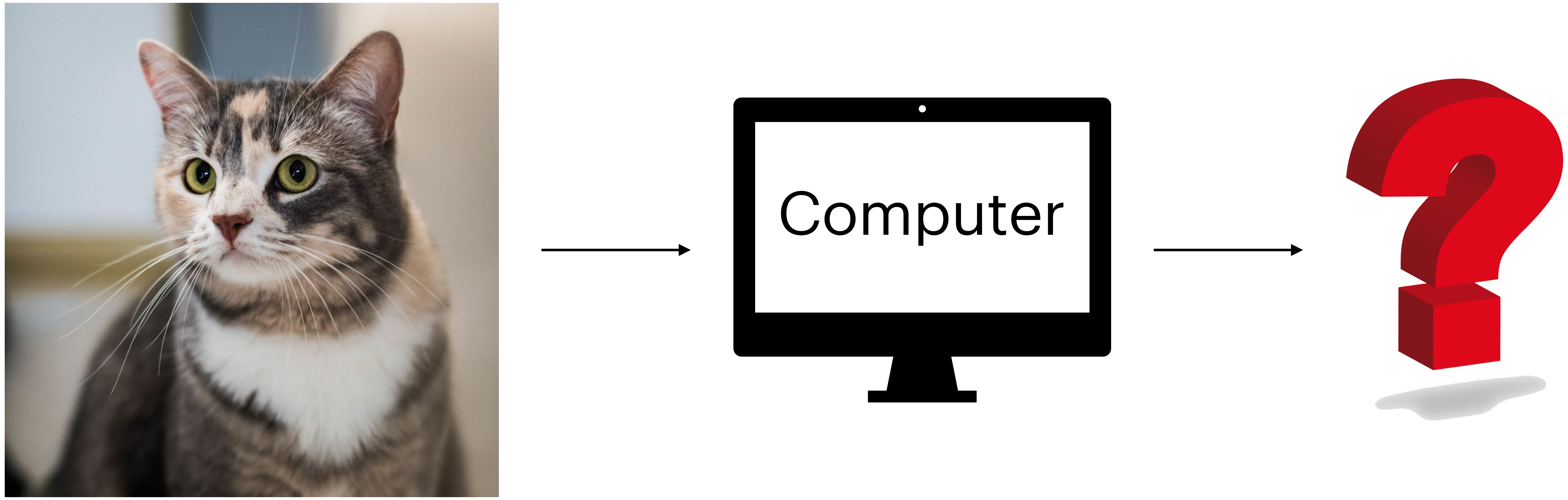
- History
- Pre-training
- Prompting
- Applications

Outline

- History
- Pre-training
- Prompting
- Applications

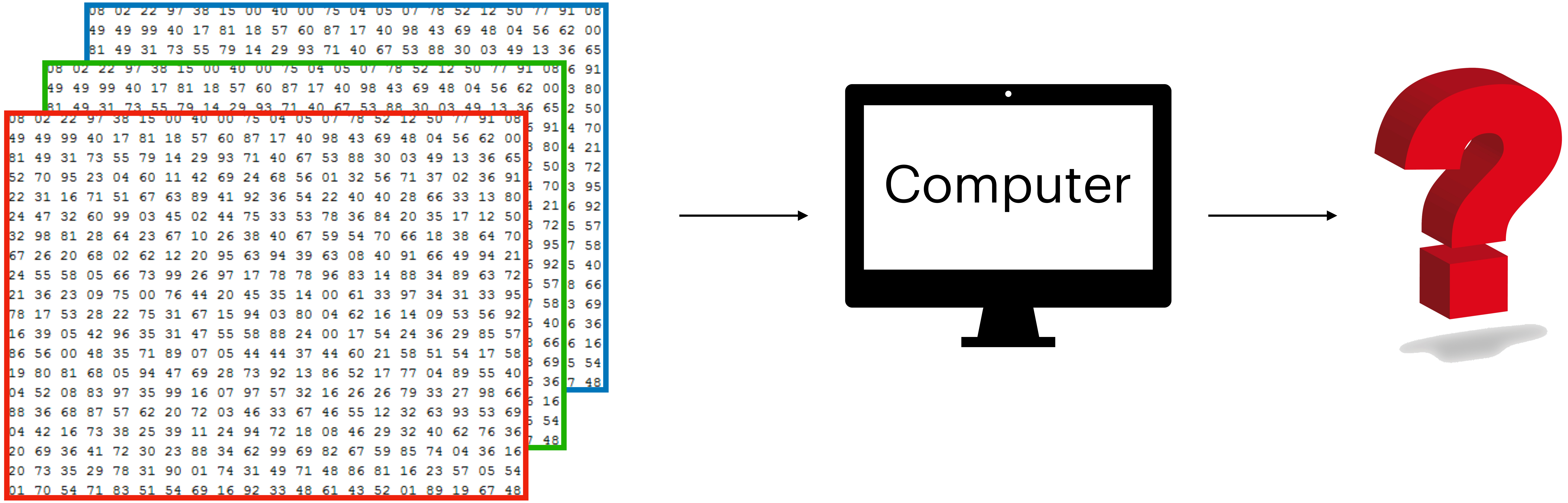
Vision Models

Teach computers to see

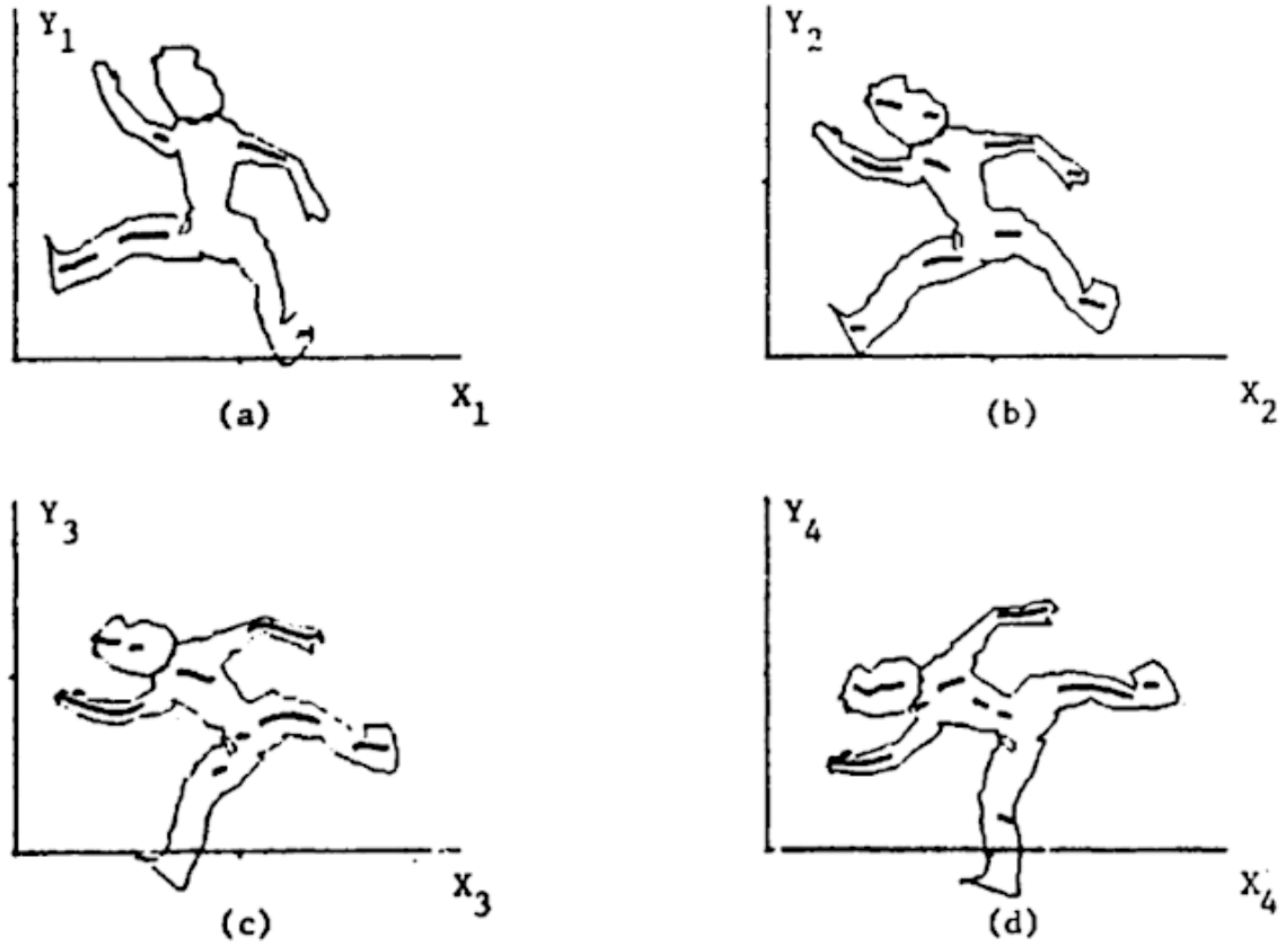


Teach computers to see

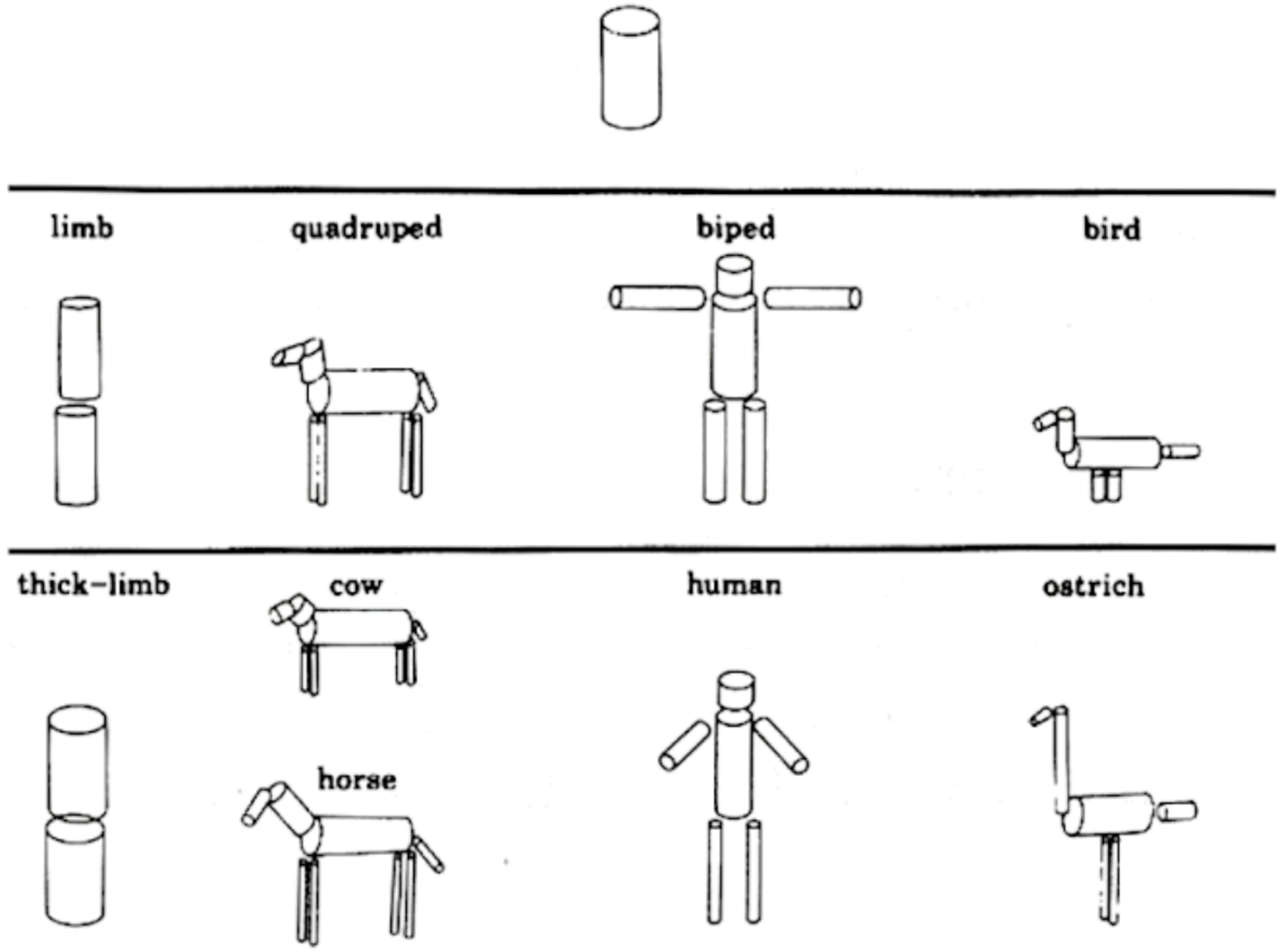
The key question is how to build discriminative visual representations



Visual representations



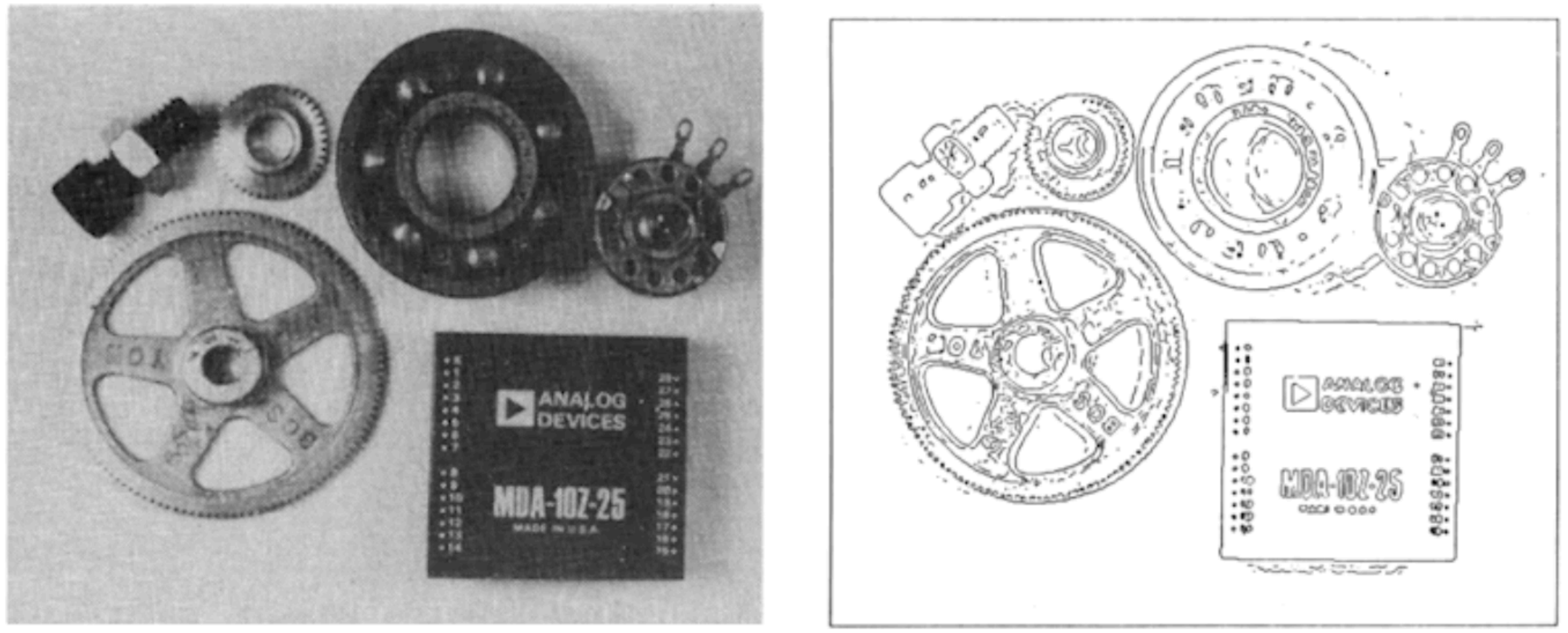
Curves
Nevatia & Binford '77



Cylinders
Brooks & Binford '79

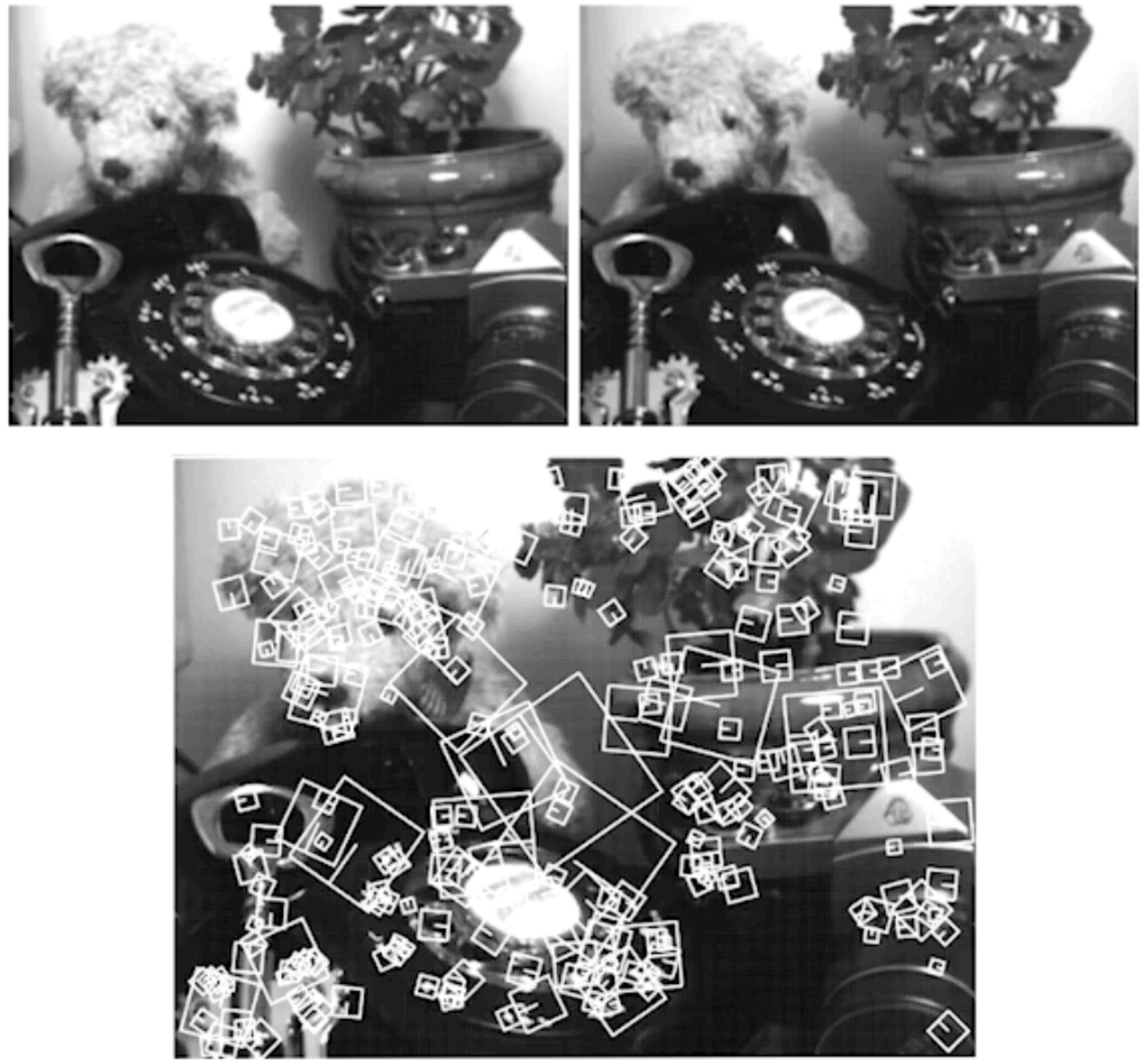
Image credit: Kaiming He.

Visual representations



(a) (b)

Edges
Canny '86

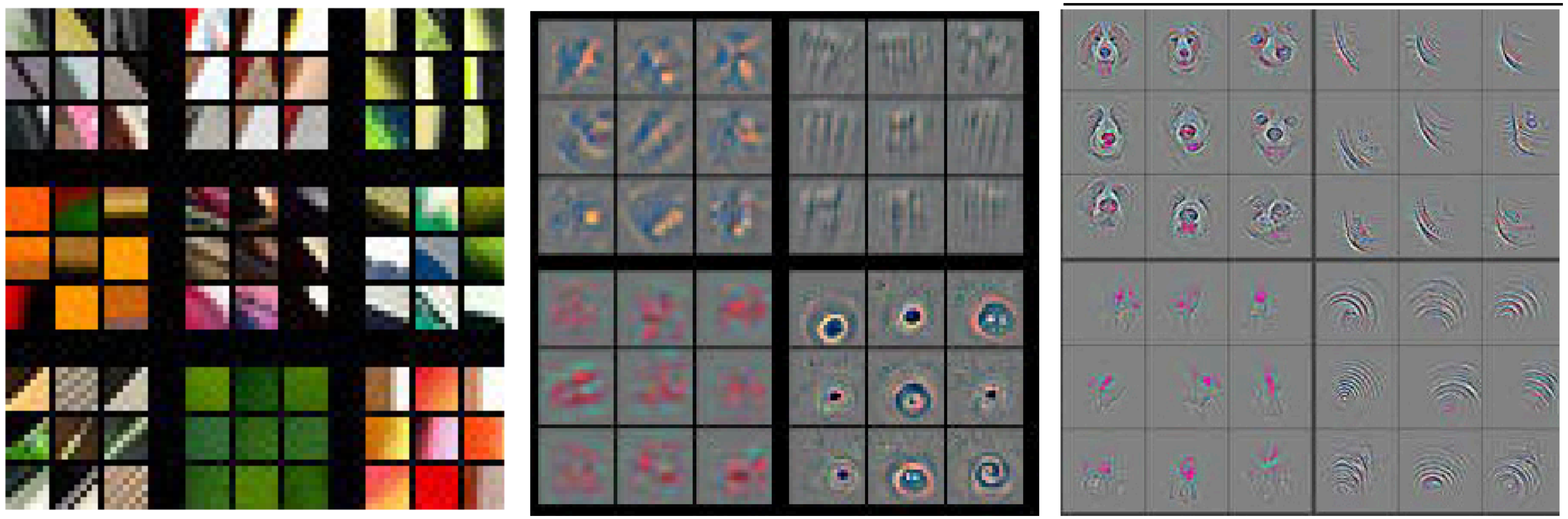


Local Features (SIFT)
Lowe '99

Image credit: Kaiming He.

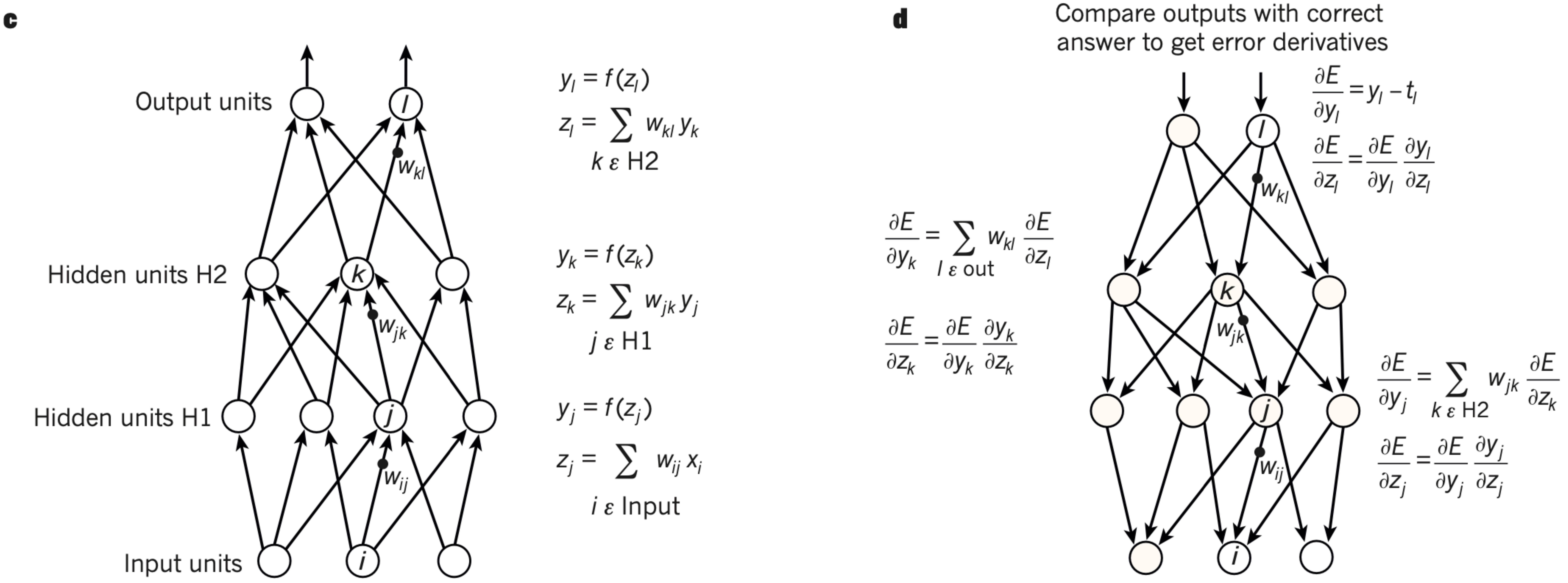
Visual representations

low-level features → mid-level features → high-level features →



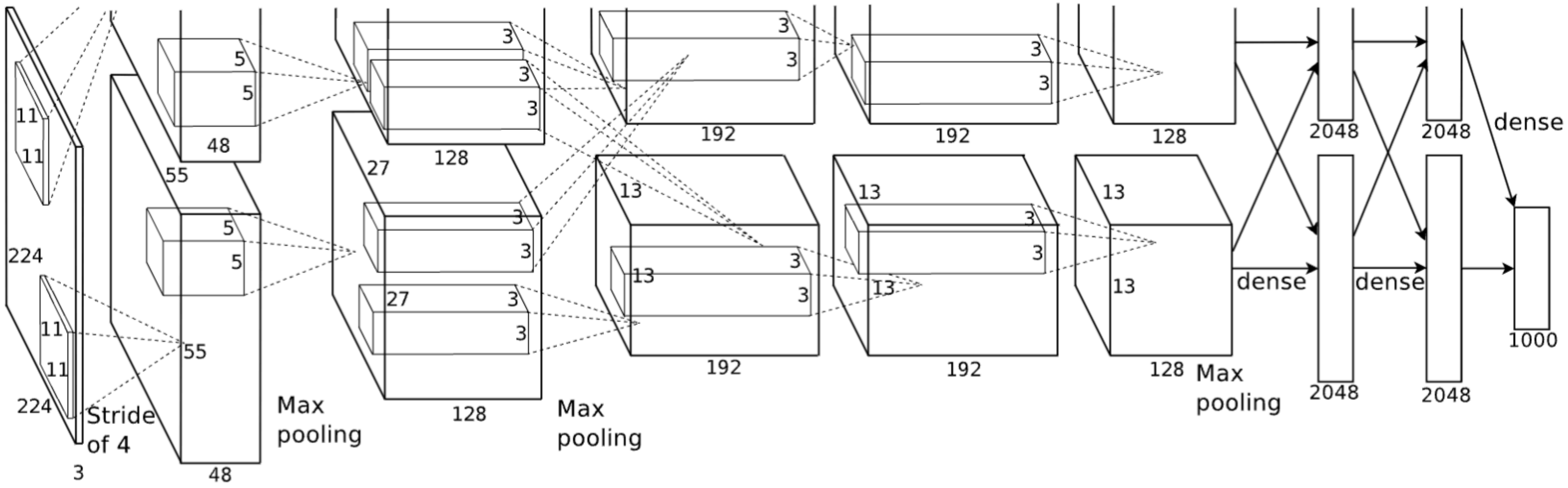
Zeiler and Fergus. Visualizing and Understanding Convolutional Networks. ECCV'14.

Deep neural network



Lecun et al. Deep learning. Nature’15.

Convolutional neural network

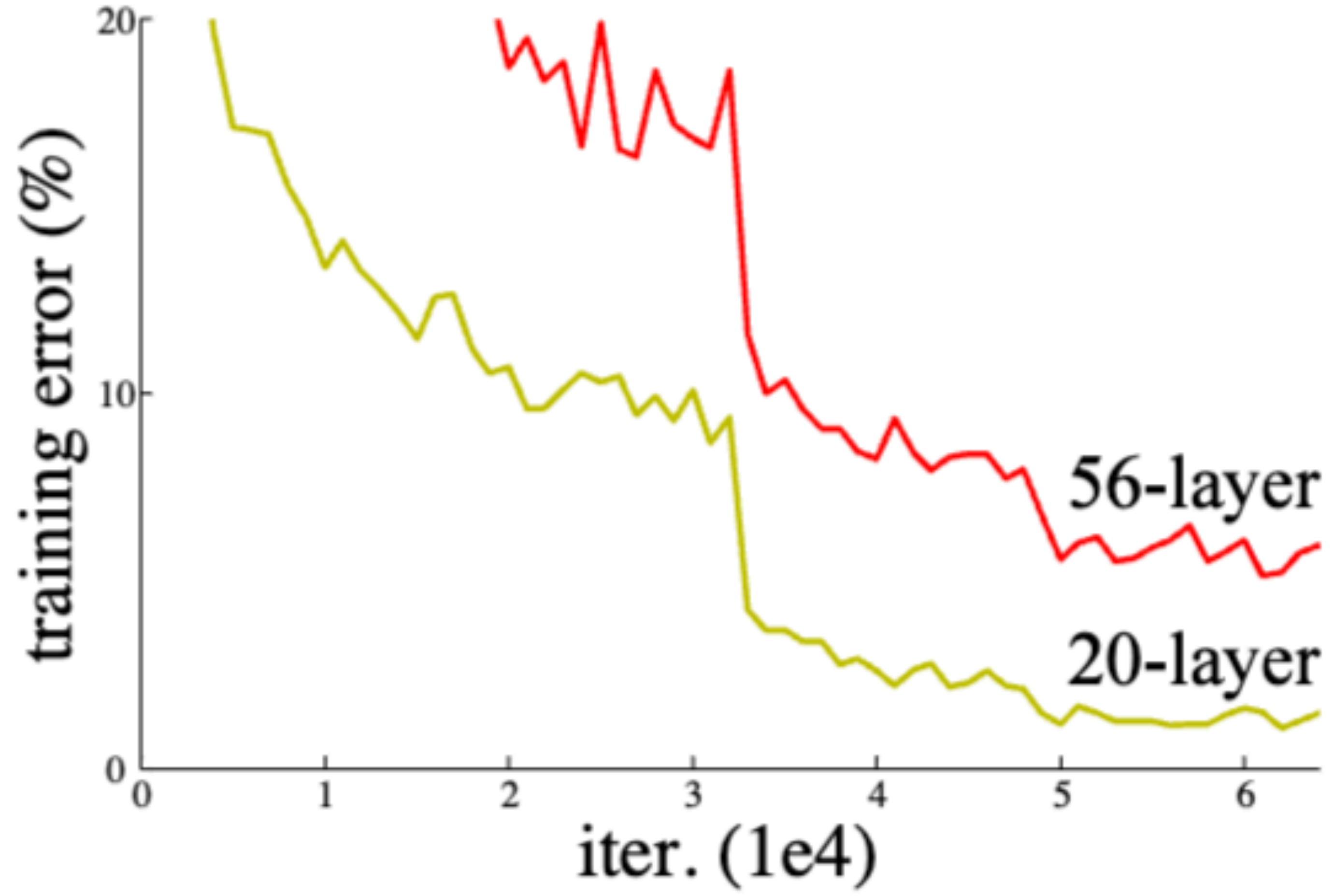


- ReLU non-linearity
- Feature normalization
- Dropout
- Data augmentation
- Multi-GPU training

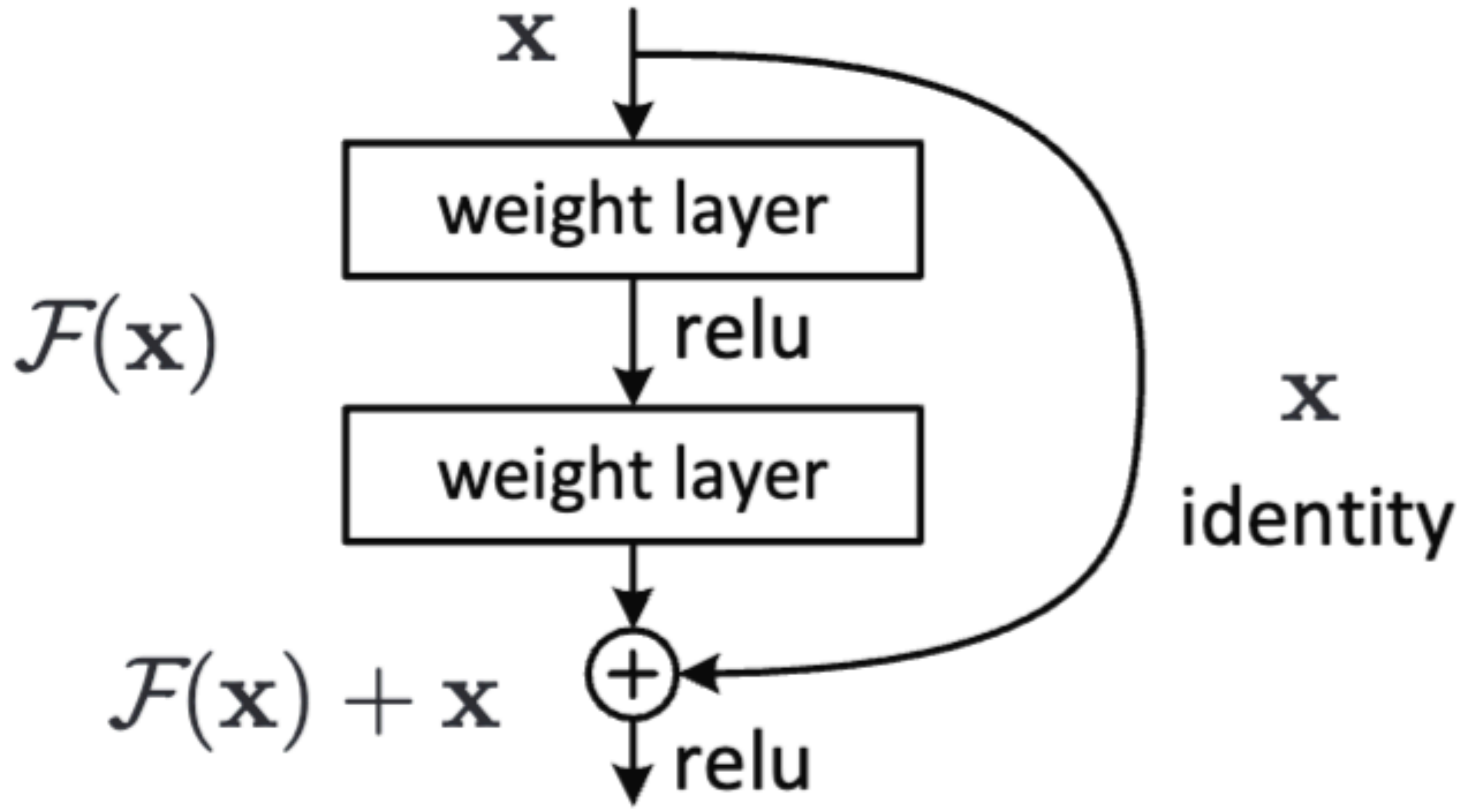
Krizhevsky et al. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems 25 (2012).

Deep residual network

Problem: Deeper networks are difficult to train

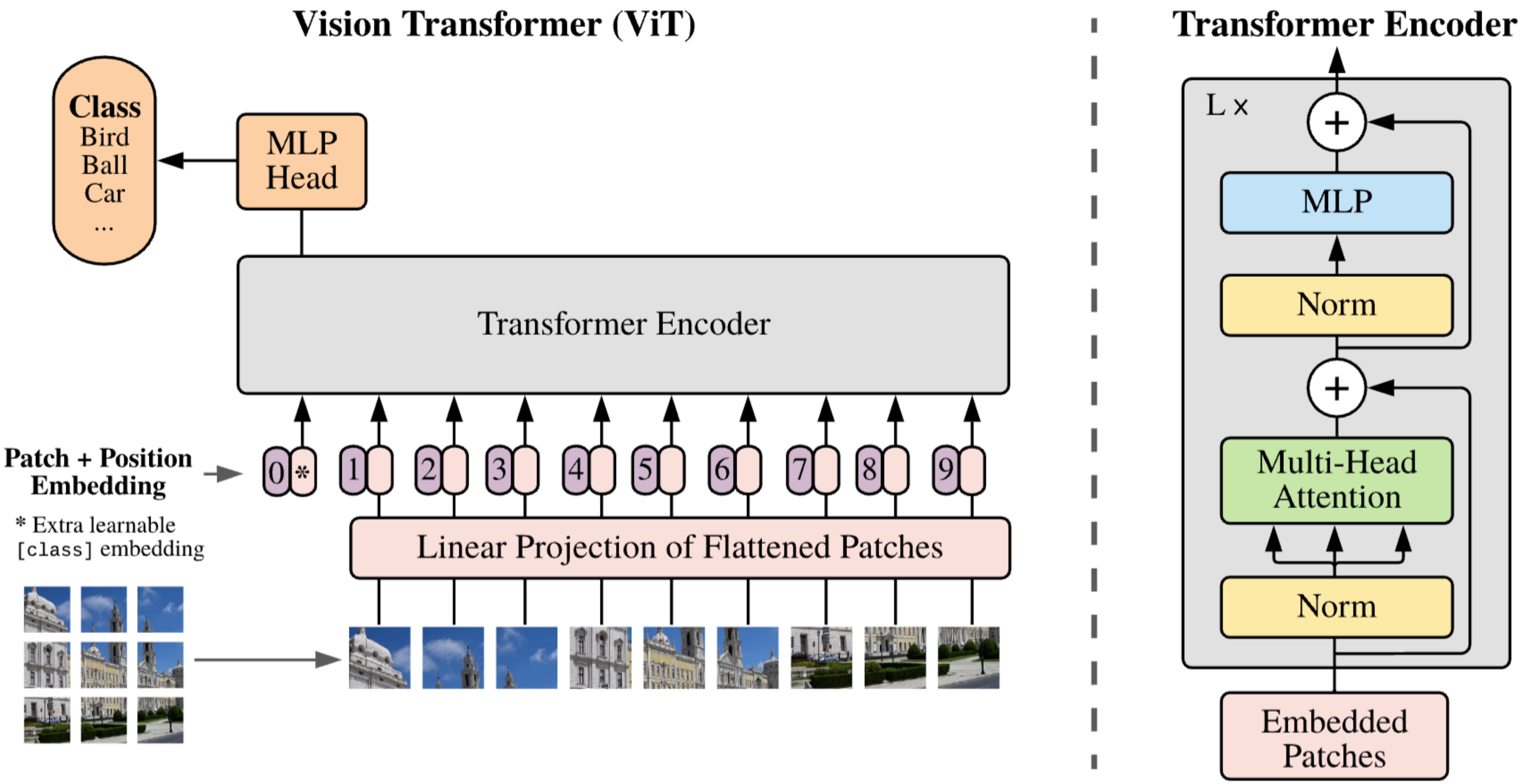


Solution: Residual connection



He et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

Vision Transformer



Dosovitskiy et al. “An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).

https://github.com/facebookresearch/detectron2

Language Models

Word representations (embeddings)

Sentiment Analysis

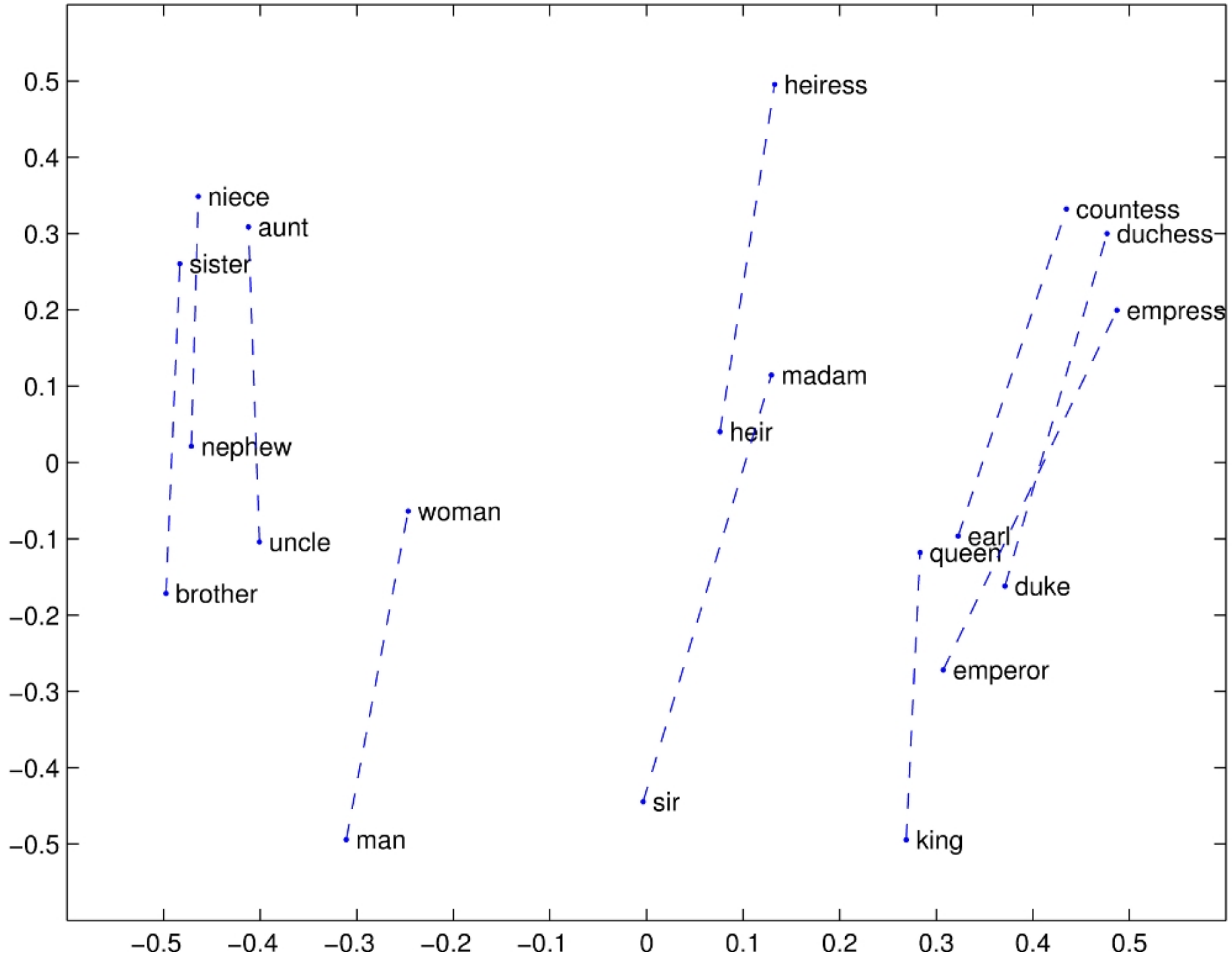
Machine Translation

Text Summarization

Email Filtering

Chatbot

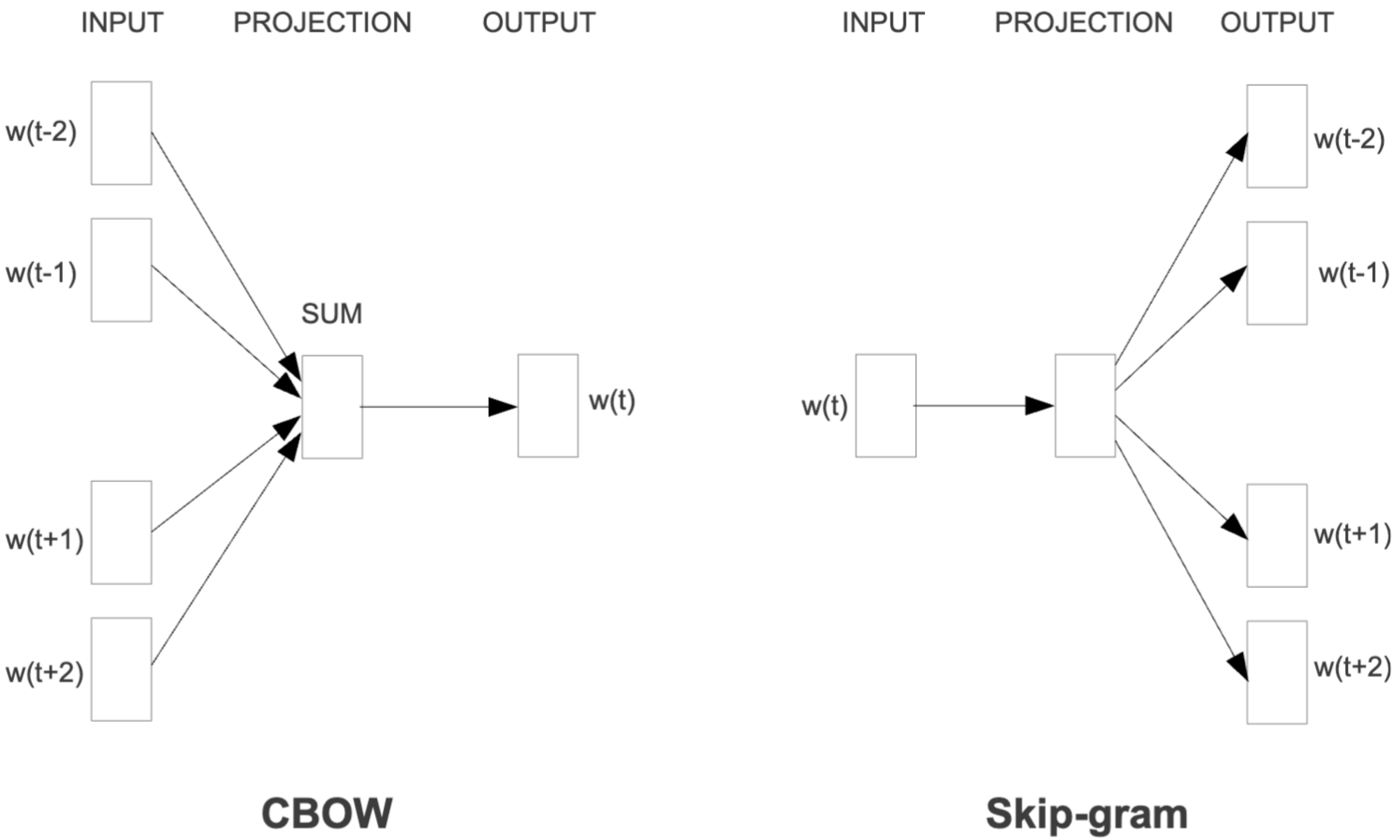
etc.



Pennington et al. "Glove: Global vectors for word representation." EMNLP. 2014.

Word2vec

Predicts the current word based on the context



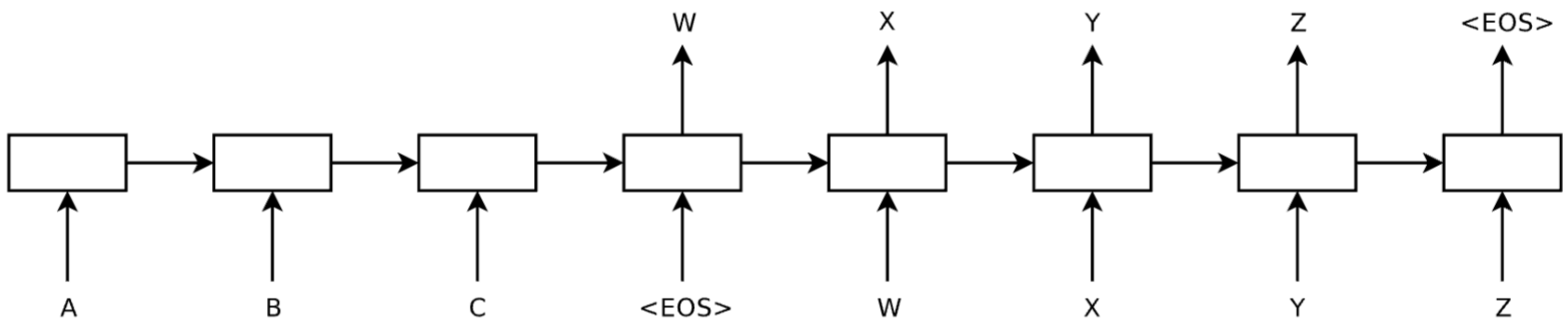
Predicts context words given the current word

Mikolov et al. "Efficient estimation of word representations in vector space." *arXiv preprint arXiv:1301.3781* 3781 (2013).

Window Size	Text	Skip-grams
2	[The **wide** road shimmered] in the hot sun.	wide, the wide, road wide, shimmered
	The [wide road **shimmered** in the] hot sun.	shimmered, wide shimmered, road shimmered, in shimmered, the
	The wide road shimmered in [the hot **sun**].	sun, the sun, hot
3	[The **wide** road shimmered in] the hot sun.	wide, the wide, road wide, shimmered wide, in
	[The wide road **shimmered** in the hot] sun.	shimmered, the shimmered, wide shimmered, road shimmered, in shimmered, the shimmered, hot
	The wide road shimmered [in the hot **sun**].	sun, in sun, the sun, hot

https://www.tensorflow.org/text/tutorials/word2vec

Seq2seq



Encoder-Decoder

- Autoregressive
- “Large” NN
- “Large” datasets

Sutskever et al. "Sequence to Sequence Learning with Neural Networks." *arXiv preprint arXiv:1409.3215* (2014).

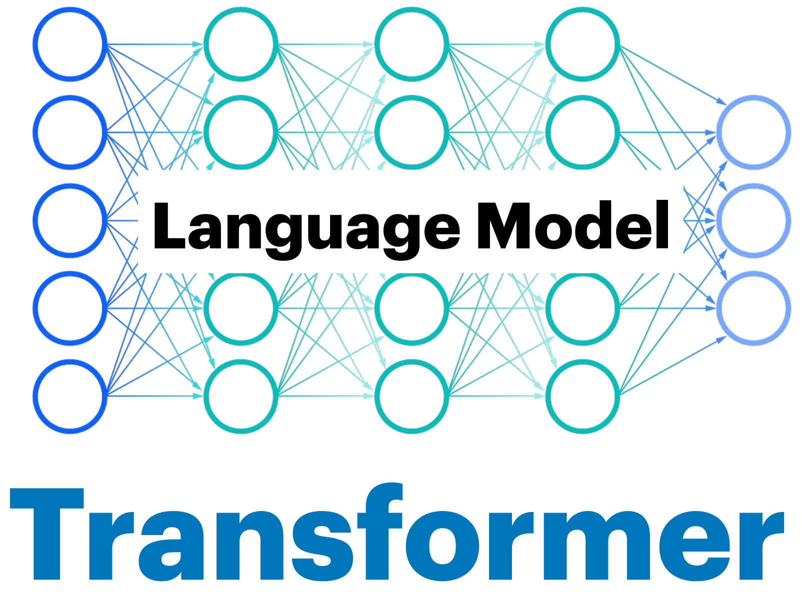
GPT

Autoregressive training (scaled up)

Improving Language Understanding by Generative Pre-Training

Alec Radford OpenAI alec@openai.com
Karthik Narasimhan OpenAI karthikn@openai.com
Tim Salimans OpenAI tim@openai.com
Ilya Sutskever OpenAI ilyasu@openai.com

"AI is the new" →



→

Word	Probability
a	0.000001
ah	0.000002
…	…
elect	0.000022
electricity	0.03
…	…
zip	0.000034

$$\max_{\theta} \sum_{i} \log P(\underbrace{u_i}_{\text{Next token}} \mid \underbrace{u_{i-k}, \ldots, u_{i-1}}_{\text{Previous tokens}}; \underbrace{\theta}_{\text{Model parameters}})$$

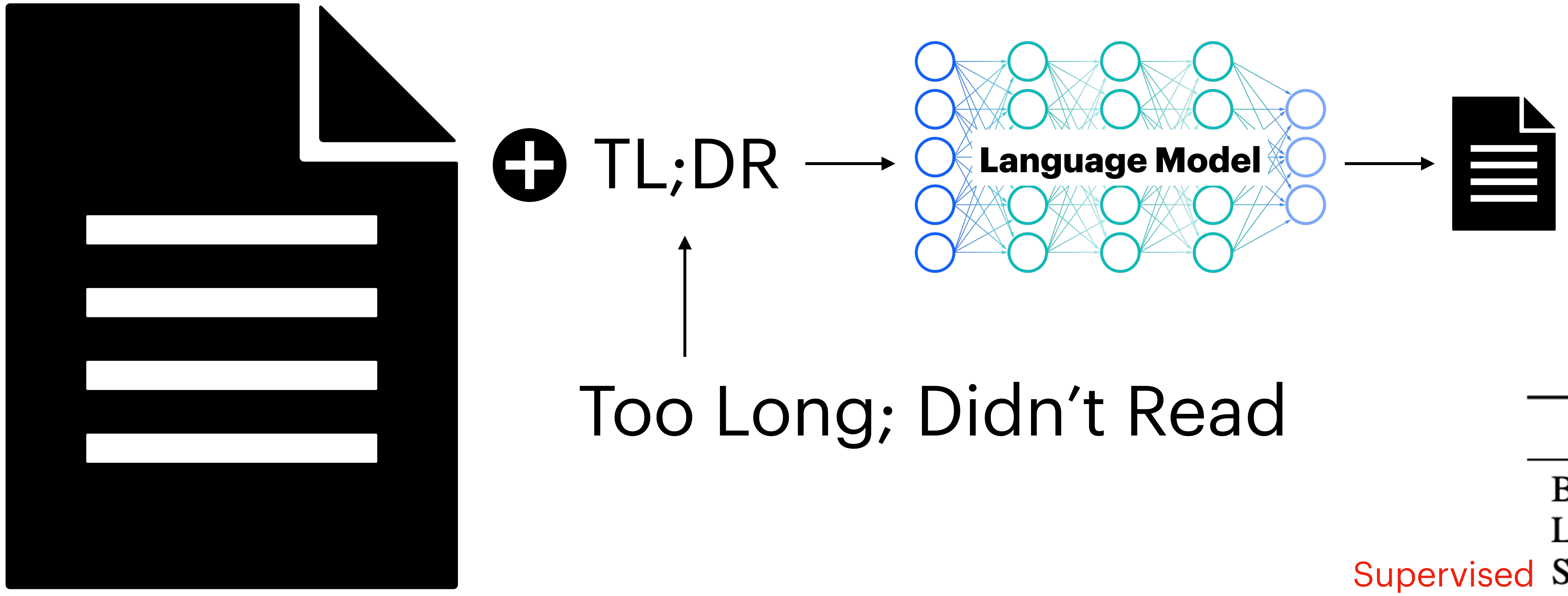
Radford et al. "Improving language understanding by generative pre-training." (2018).

GPT

Language Models are Unsupervised Multitask Learners

Alec Radford [*1] Jeffrey Wu [*1] Rewon Child [1] David Luan [1] Dario Amodei [**1] Ilya Sutskever [**1]

Zero-shot (autocomplete)



		R-1	R-2	R-L	R-AVG
	Bottom-Up Sum	**41.22**	**18.68**	**38.34**	**32.75**
	Lede-3	40.38	17.66	36.62	31.55
Supervised	Seq2Seq + Attn	31.33	11.81	28.83	23.99
	GPT-2 `TL;DR:`	29.34	8.27	26.58	21.40
Random baseline	Random-3	28.78	8.63	25.52	20.98
	GPT-2 no hint	21.58	4.03	19.47	15.03

Table 4. Summarization performance as measured by ROUGE F1 metrics on the CNN and Daily Mail dataset. Bottom-Up Sum is the SOTA model from (Gehrmann et al., 2018)

Radford et al. Language Models are Unsupervised Multitask Learners. 2019.

GPT

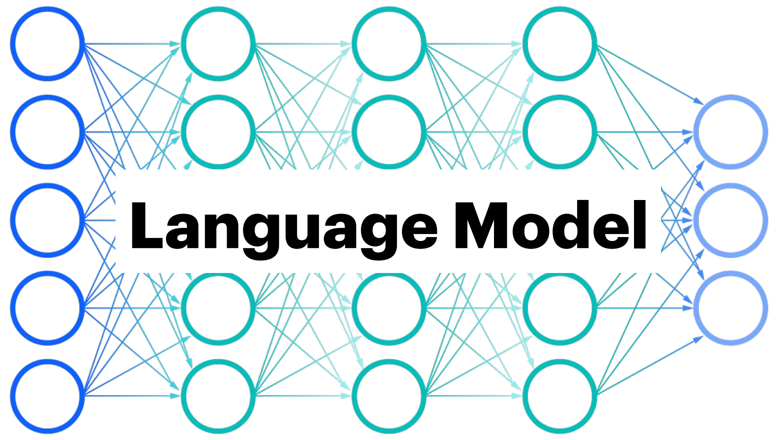
Language Models are Few-Shot Learners

Tom B. Brown* Benjamin Mann* Nick Ryder* Melanie Subbiah*
Jared Kaplan† Prafulla Dhariwal Arvind Neelakantan Pranav Shyam Girish Sastry
Amanda Askell Sandhini Agarwal Ariel Herbert-Voss Gretchen Krueger Tom Henighan
Rewon Child Aditya Ramesh Daniel M. Ziegler Jeffrey Wu Clemens Winter
Christopher Hesse Mark Chen Eric Sigler Mateusz Litwin Scott Gray
Benjamin Chess Jack Clark Christopher Berner
Sam McCandlish Alec Radford Ilya Sutskever Dario Amodei

OpenAI

Few-shot in-context learning

"Translate English into French:
Hello => Bonjour
Thank you => Merci
Goodbye => Au revoir
Excuse me => __"



"À plus tard"

Setting	En→Fr	Fr→En	En→De	De→En	En→Ro	Ro→En
SOTA (Supervised)	**45.6**[a]	35.0 [b]	**41.2**[c]	40.2[d]	**38.5**[e]	**39.9**[e]
XLM [LC19]	33.4	33.3	26.4	34.3	33.3	31.8
MASS [STQ+19]	37.5	34.9	28.3	35.2	35.2	33.1
mBART [LGG+20]	-	-	29.8	34.0	35.0	30.5
GPT-3 Zero-Shot	25.2	21.2	24.6	27.2	14.1	19.9
GPT-3 One-Shot	28.3	33.7	26.2	30.4	20.6	38.6
GPT-3 Few-Shot	32.6	39.2	29.7	40.6	21.0	39.5

Good at X -> En

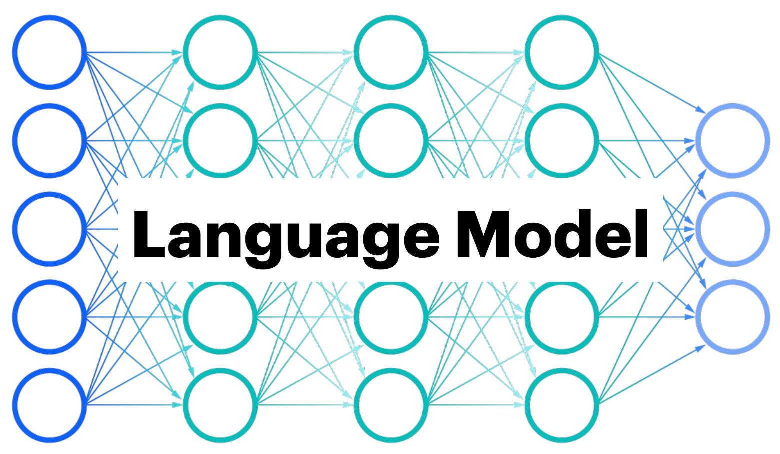
Brown et al. Language Models are Few-Shot Learners. NeurIPS'20.

Chain-of-thought prompting

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

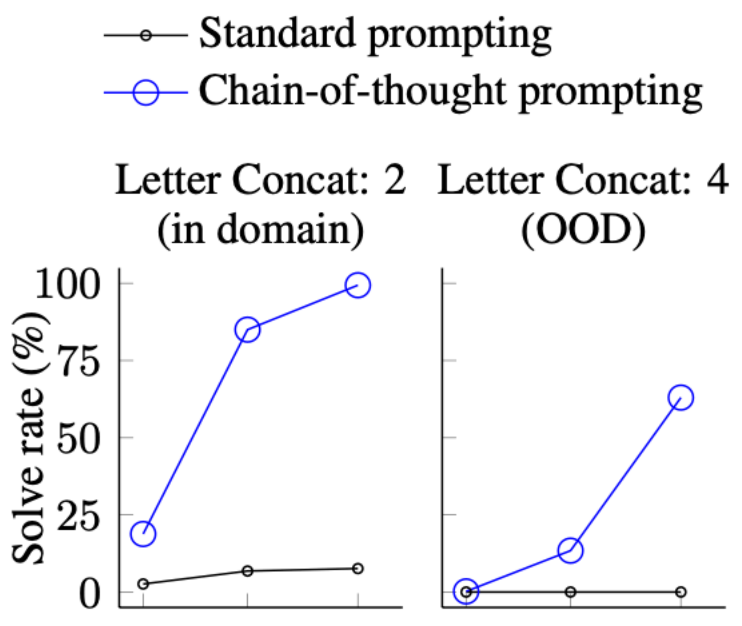
A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

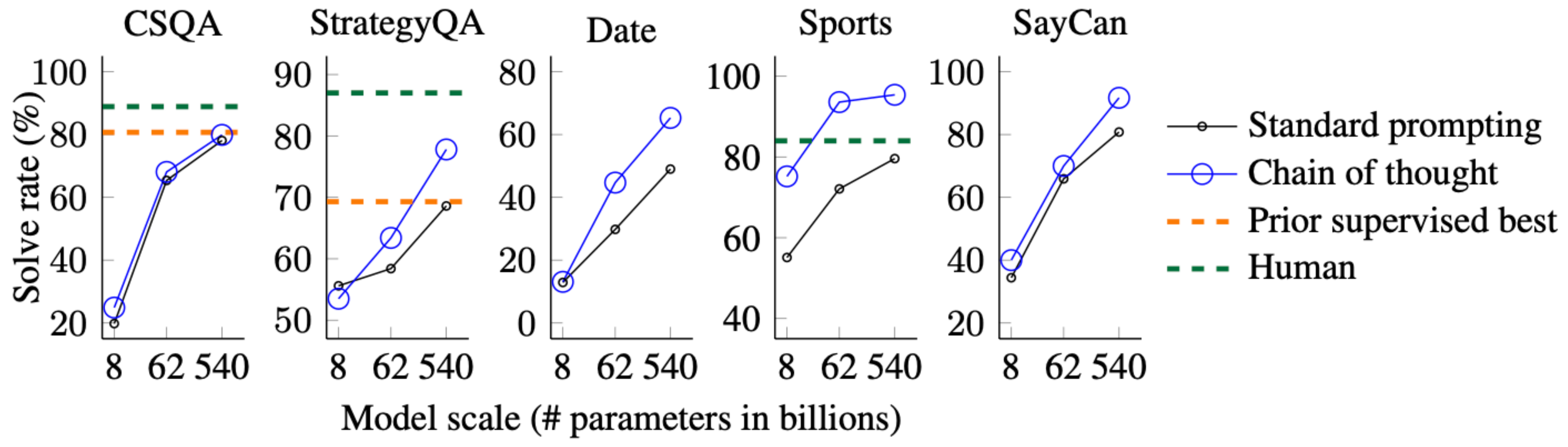


A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9. The answer is 9.

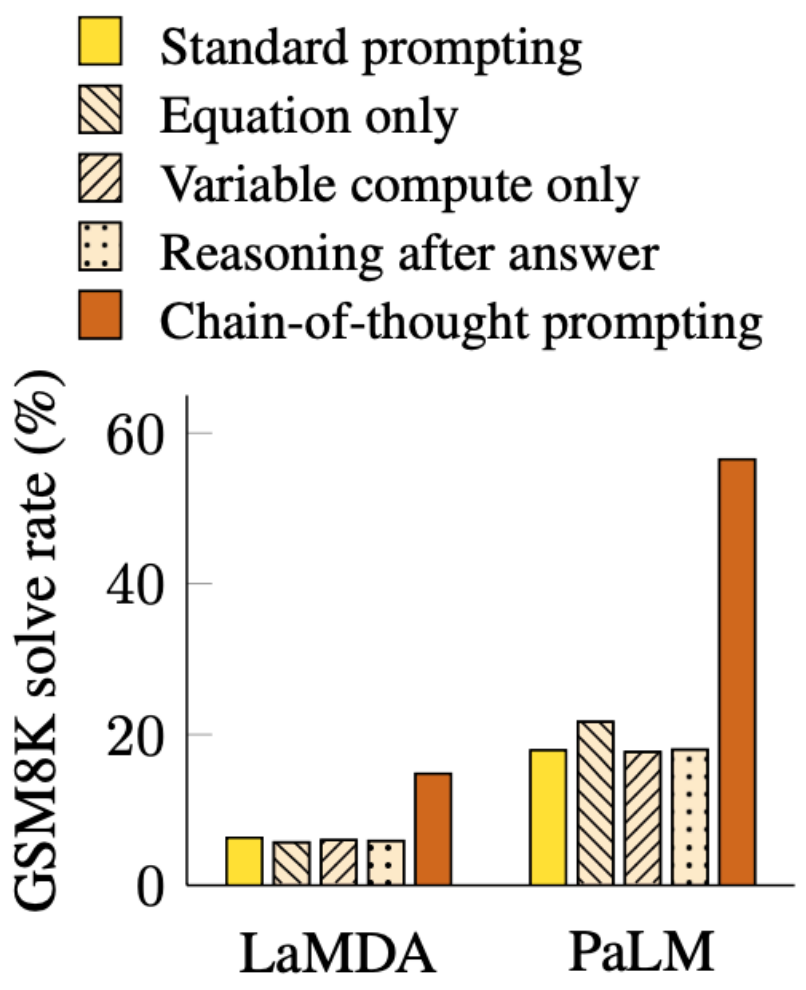
Symbolic reasoning



Commonsense reasoning



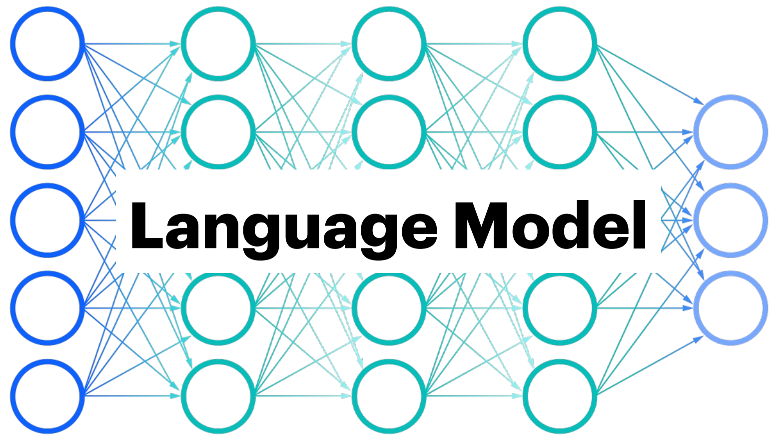
Math



Wei et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. NeurIPS'22.

Zero-shot chain-of-thought prompting

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A: Let's think step by step.



There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls.

		MultiArith	GSM8K
Zero-Shot		**17.7**	**10.4**
Few-Shot (2 samples)		33.7	15.6
Few-Shot (8 samples)		33.8	15.6
Zero-Shot-CoT	Significantly beats zero-shot ←	**78.7**	**40.7**
Few-Shot-CoT (2 samples)		84.8	41.3
Few-Shot-CoT (4 samples : First) (*1)		89.2	-
Few-Shot-CoT (4 samples : Second) (*1)		90.5	-
Few-Shot-CoT (8 samples)	Manual CoT is still better ←	93.0	48.7

Kojima et al. Large Language Models are Zero-Shot Reasoners. NeurIPS'22.

Chain-of-Thought Prompting Elicits Reasoning in Large Language Models

Large Language Models are Zero-Shot Reasoners

t.koj

Mach
Google

Tree of Thoughts: Deliberate Problem Solving with Large Language Models

Shuny
Princeton U

Thom
Prince

LEAST-TO-MOST PROMPTING ENABLES COMPLEX REASONING IN LARGE LANGUAGE MODELS

Denny Zhou[†*] **Na**
Dale Schuurmans[†]
[†]Google Research, B

SELF-CONSISTENCY IMPROVES CHAIN OF THOUGHT REASONING IN LANGUAGE MODELS

Xuezhi Wang[†‡] **Jason Wei**[†] **Dale Schuurmans**[†] **Quoc Le**[†] **Ed H. Chi**[†]
Sharan Narang[†] **Aakanksha Chowdhery**[†] **Denny Zhou**[†§]
[†]Google Research, Brain Team
[‡]xuezhiw@google.com, [§]dennyzhou@google.com

Forbes

FORBES > INNOVATION > ENTERPRISE TECH

The Hot New Job That Pays Six Figures: AI Prompt Engineering

Bloomberg

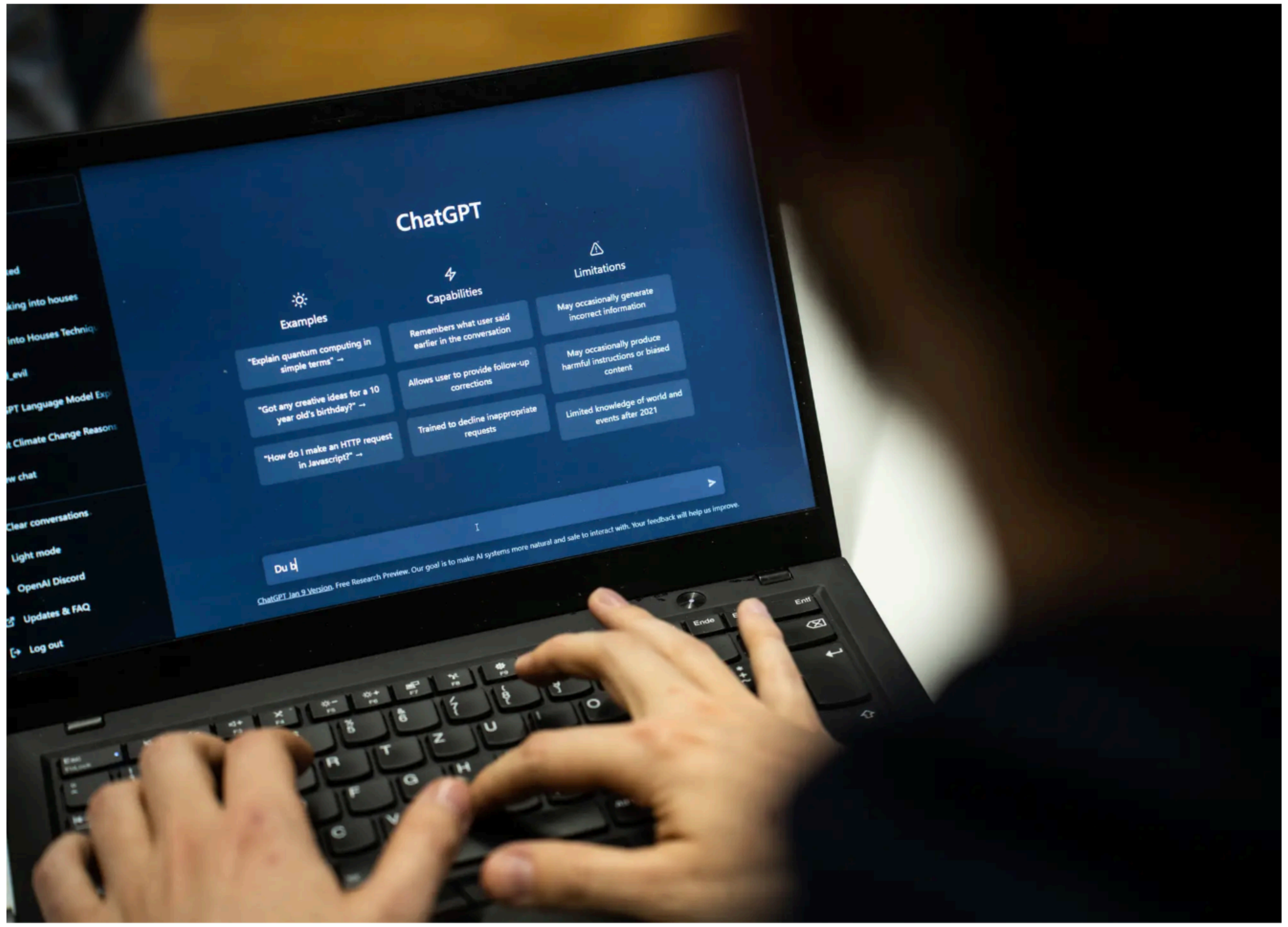
AI's Hottest Job: Prompt Engineer

So-called AI whisperers can earn six-figure salaries, no programming experience necessary. Conrad Quilty-Harper met one of these prompt engineers to find out how to coax the best out of a large-language model. (Source: Bloomberg)

July 5th, 2023, 11:27 PM GMT+0800

AI 'prompt engineer' jobs can pay up to $375,000 a year and don't always require a background in tech

Britney Nguyen May 1, 2023, 11:34 PM GMT+8 Share Save



The rise of generative AI tools like ChatGPT is creating a hot market for "prompt engineers" who test and improve chatbot answers.

Getty Images

BUSINESS INSIDER

Convergence of Vision and Language Models

Traditional vision models struggle to generalize

A classifier trained to recognise horse images would not be able to recognise zebra, though the latter is just like horse but with black-and-white stripes

Why traditional vision models struggle to generalize?

Each row in $\boldsymbol{W}$ can be viewed as a class prototype

$$\longrightarrow \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} \longrightarrow \text{Softmax}(\underbrace{\boldsymbol{W}}_{3\times d}\underbrace{\boldsymbol{x}}_{d\times 1}) \longrightarrow \begin{bmatrix} 0.20 \\ 0.75 \\ 0.05 \end{bmatrix} \begin{array}{l} p(\text{“dog”}|\boldsymbol{x}) \\ p(\text{“cat”}|\boldsymbol{x}) \\ p(\text{“bird”}|\boldsymbol{x}) \end{array}$$

$\boldsymbol{x} \in \mathbb{R}^d$

$$p(y=0|\boldsymbol{x}) = \frac{\exp(z_0)}{\exp(z_0)+\exp(z_1)+\exp(z_2)}$$

$$p(y=1|\boldsymbol{x}) = \frac{\exp(z_1)}{\exp(z_0)+\exp(z_1)+\exp(z_2)}$$

$$p(y=2|\boldsymbol{x}) = \frac{\exp(z_2)}{\exp(z_0)+\exp(z_1)+\exp(z_2)}$$

Why traditional models struggle to generalize?

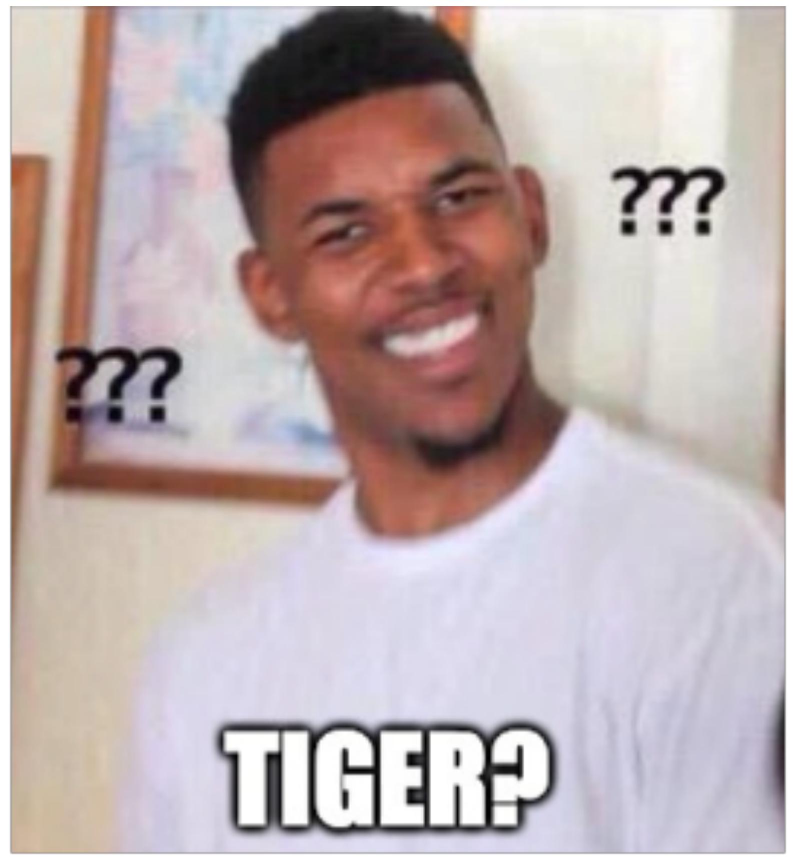
$$\longrightarrow \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} \longrightarrow \text{Softmax}(\underbrace{\boldsymbol{W}}_{\boxed{3\times d}}\ \underbrace{\boldsymbol{x}}_{d\times 1}) \longrightarrow \begin{array}{cl} \begin{bmatrix} 0.20 \\ 0.75 \\ 0.05 \end{bmatrix} & \begin{array}{l} p(\text{“dog”}|\boldsymbol{x}) \\ p(\text{“cat”}|\boldsymbol{x}) \\ p(\text{“bird”}|\boldsymbol{x}) \end{array} \end{array}$$

$\boldsymbol{x} \in \mathbb{R}^d$

Each row in ***W*** can be viewed as a class prototype

There is no class prototype for *tiger*

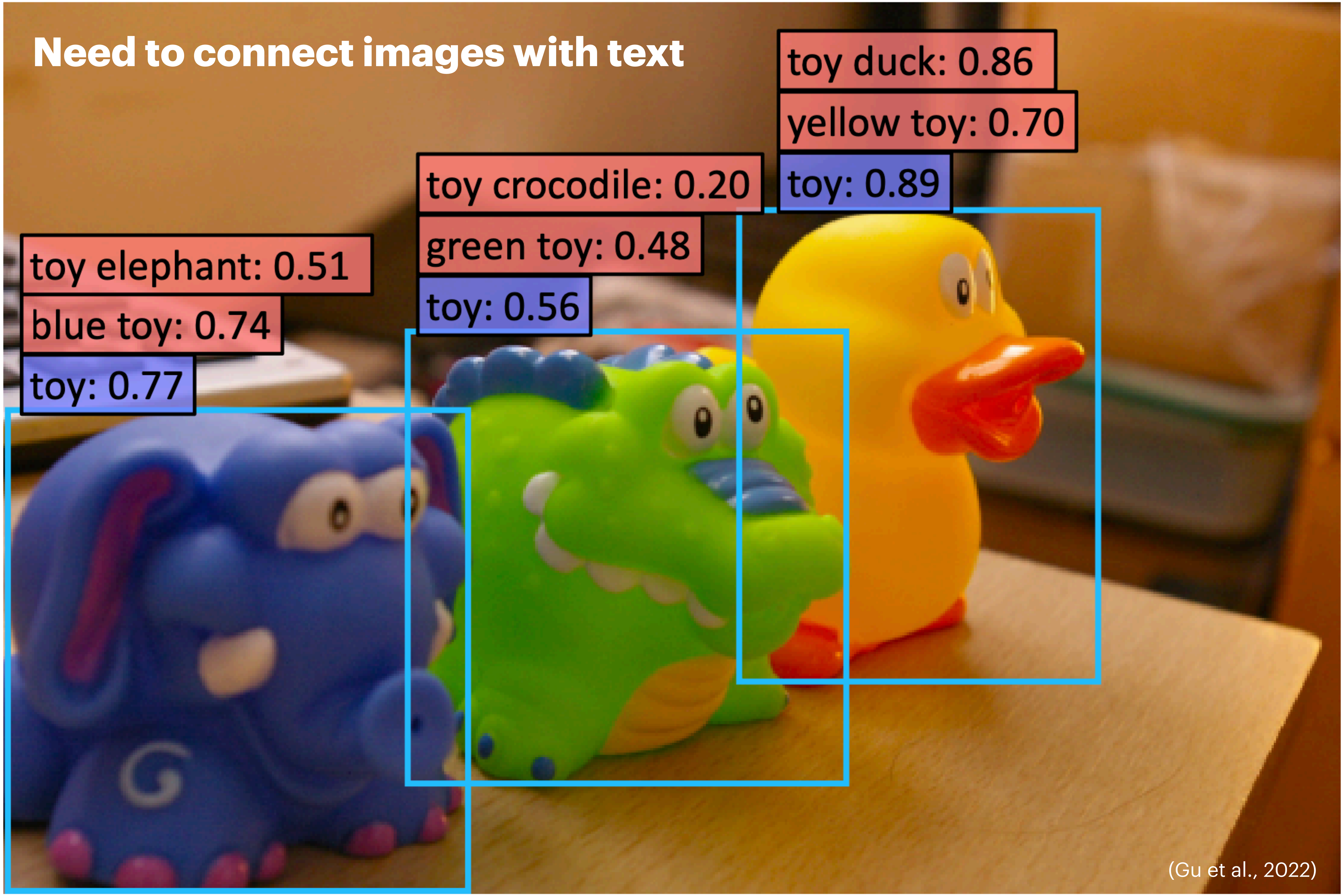
Need to re-train the model!



Traditional models cannot handle open-vocabulary tasks

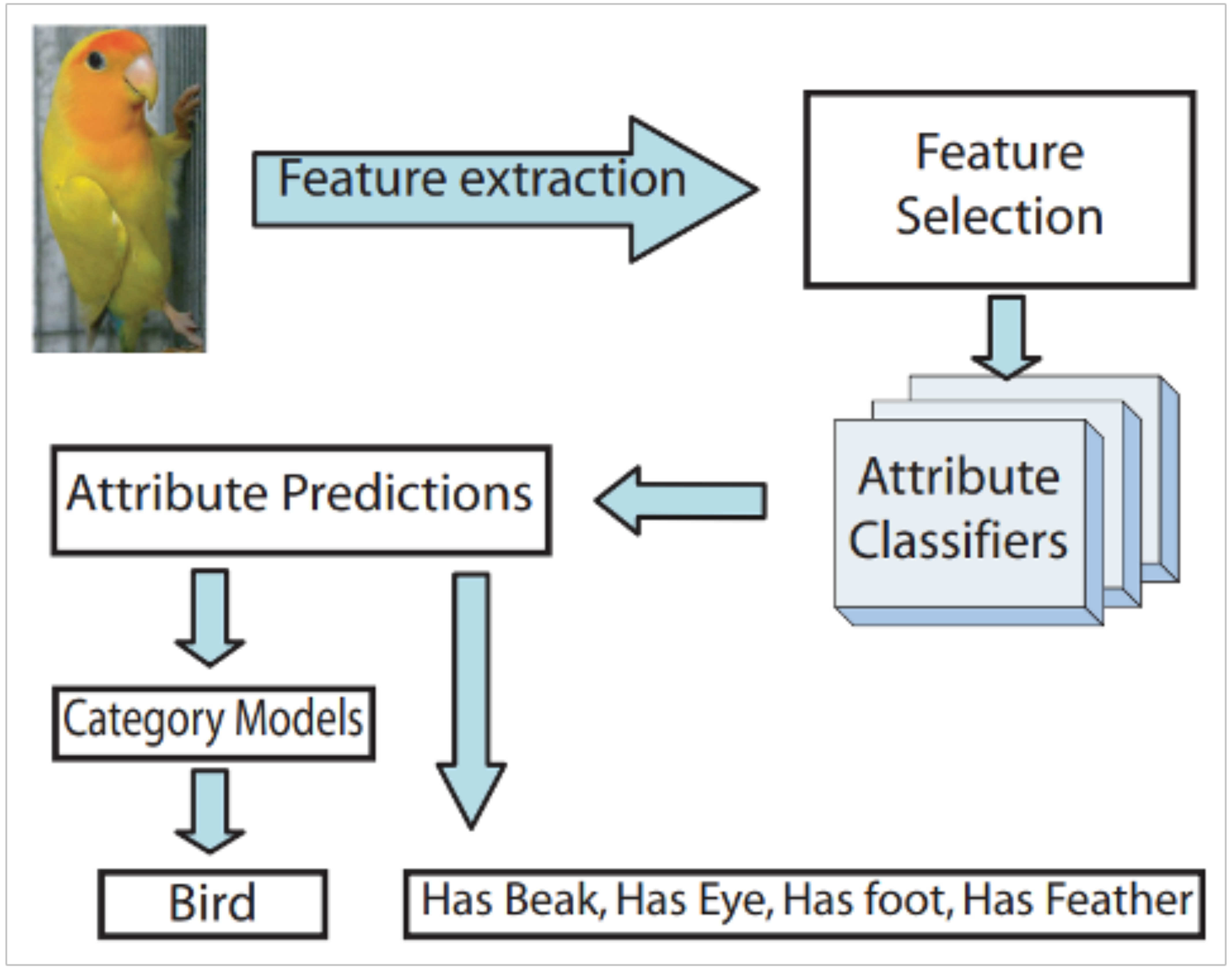
Image generated by OpenAI's DALL-E

Need to connect images with text



(Gu et al., 2022)

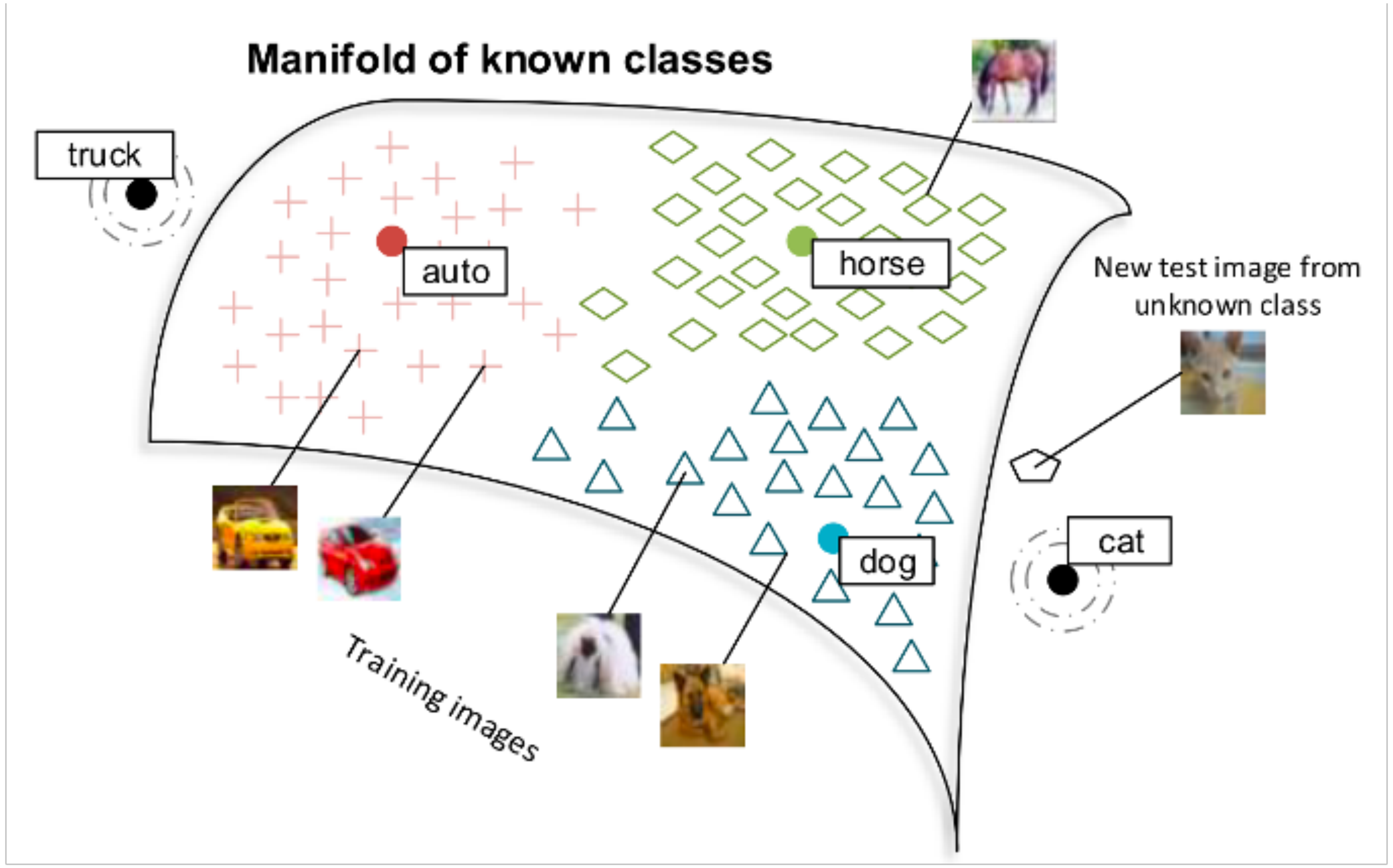
Early methods



Associate classes with auxiliary information like **attributes**, which encode distinguishing properties of objects

Farhadi et al. Describing Objects by their Attributes. CVPR'09.

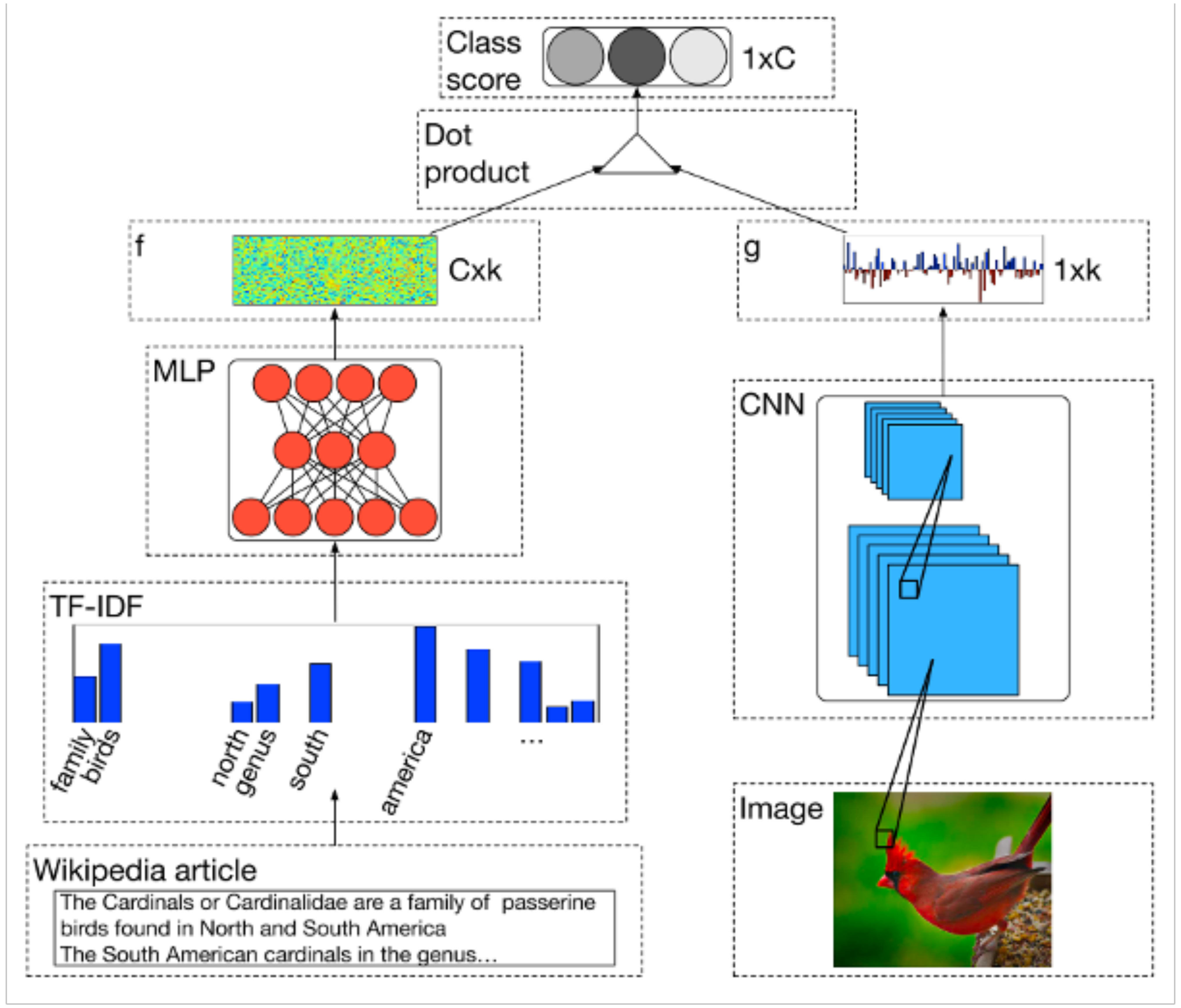
Early methods



Associate images with semantic **word vectors** (i.e., word2vec)

Socher et al. Zero-Shot Learning Through Cross-Modal Transfer. NeurIPS'13.

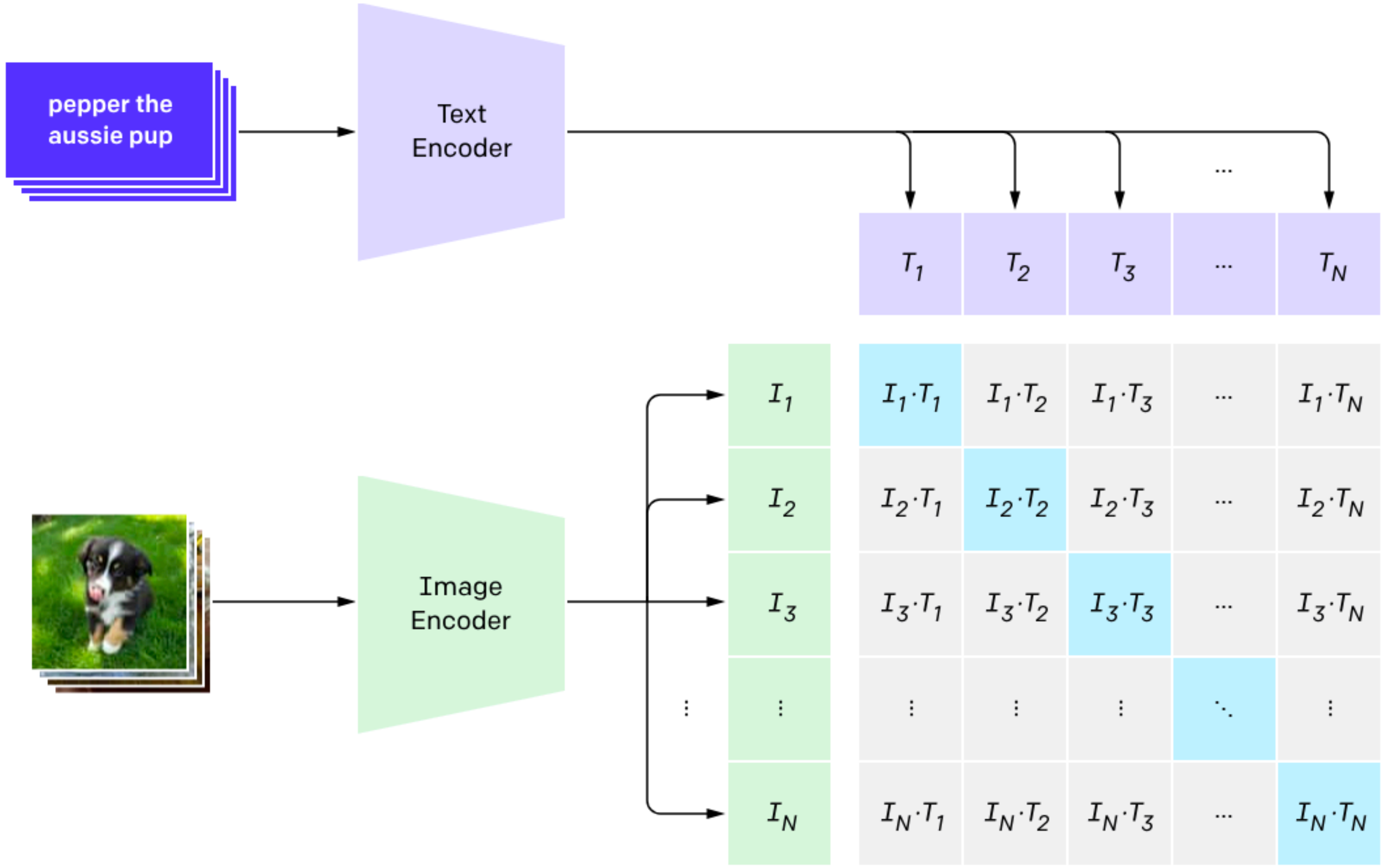
Early methods



Learn a **joint embedding space** for images and text

Ba et al. Predicting Deep Zero-Shot Convolutional Neural Networks using Textual Descriptions. ICCV'15.

Today's methods



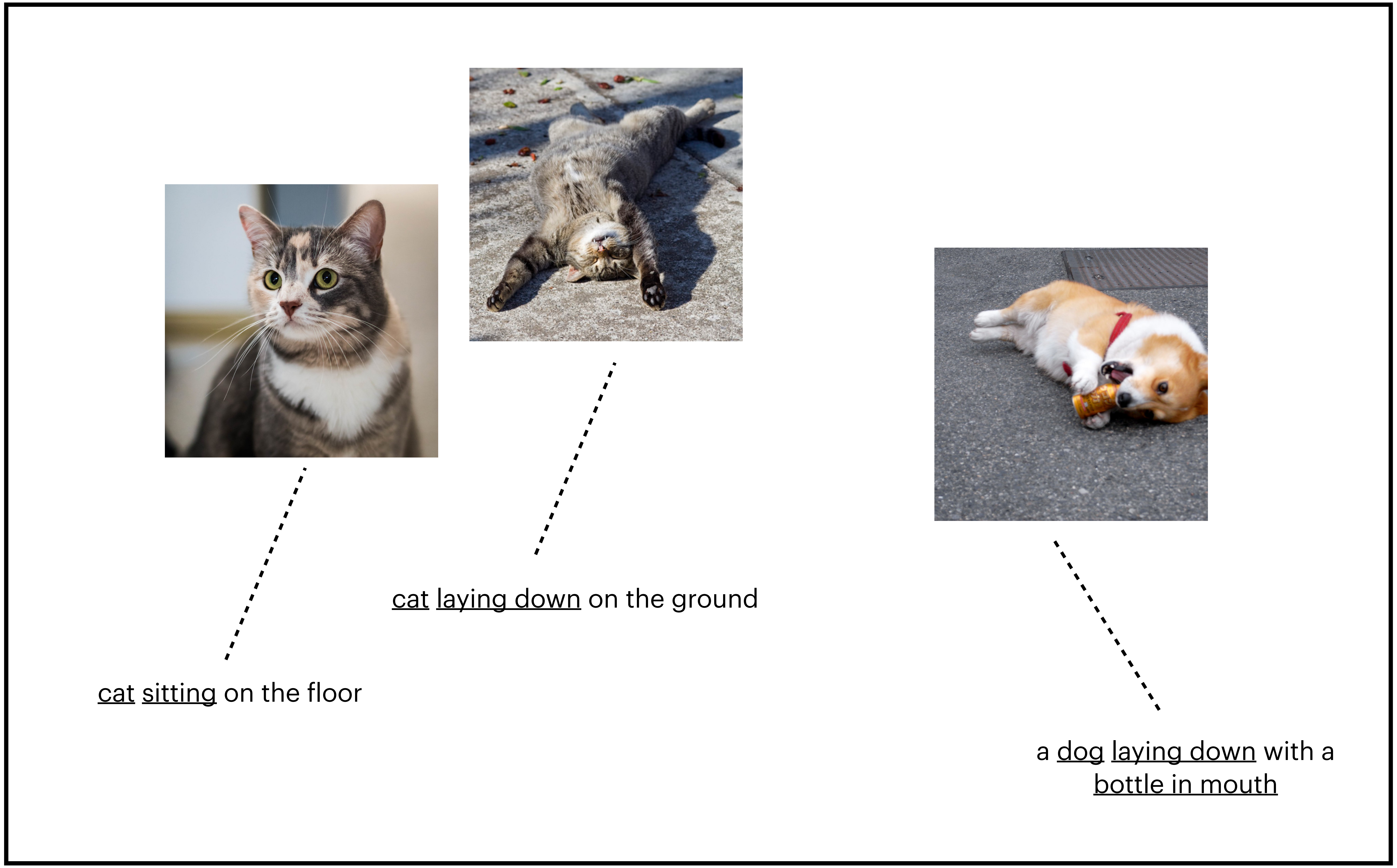
Learn a **joint embedding space** for images and text, **using large-capacity models and web-scale datasets**

Radford et al. Learning Transferable Visual Models From Natural Language Supervision. ICML'21.

Outline

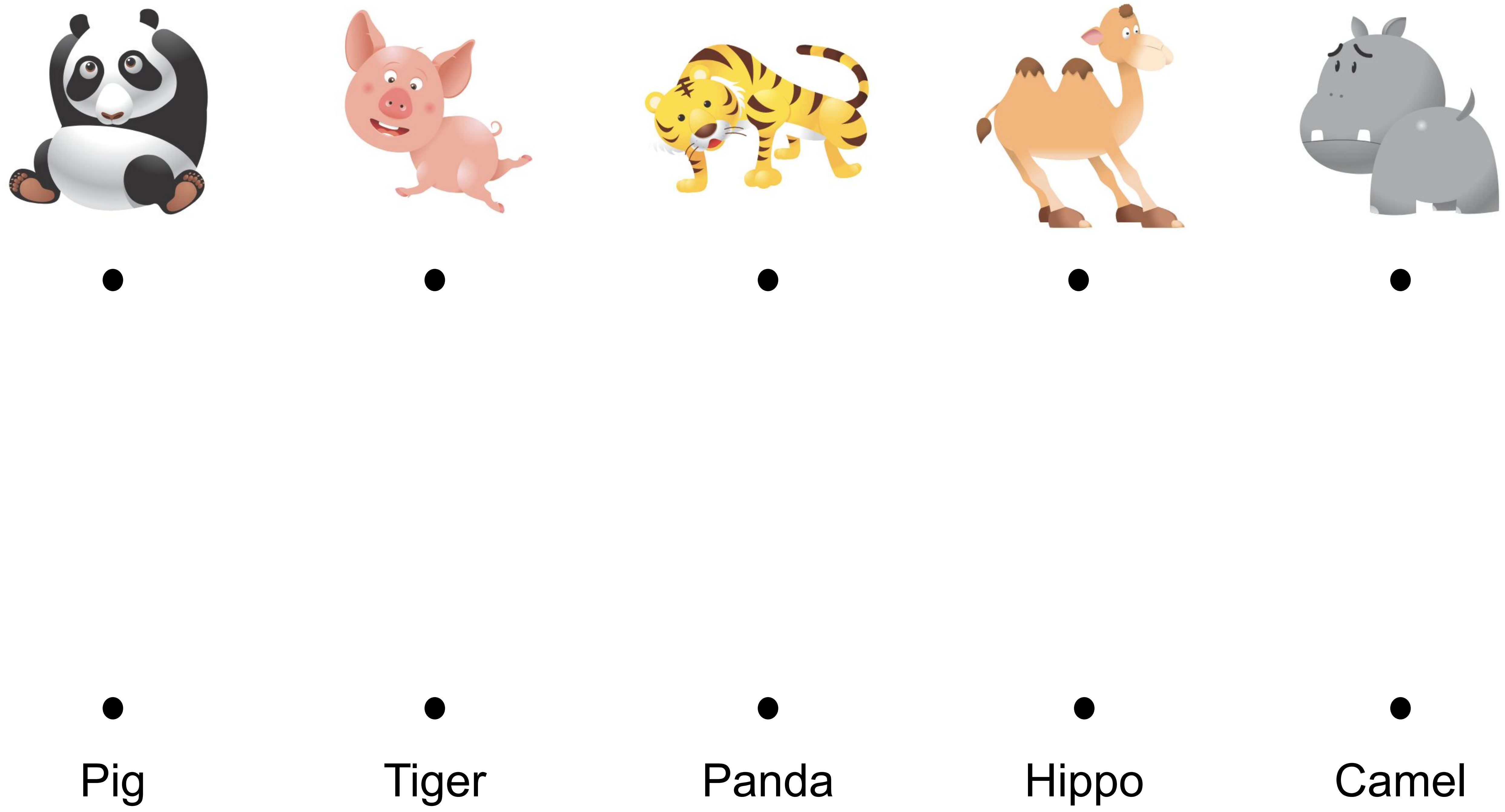
- History
- Pre-training
- Prompting
- Applications

Key idea: joint embedding space learning



Contrastive learning

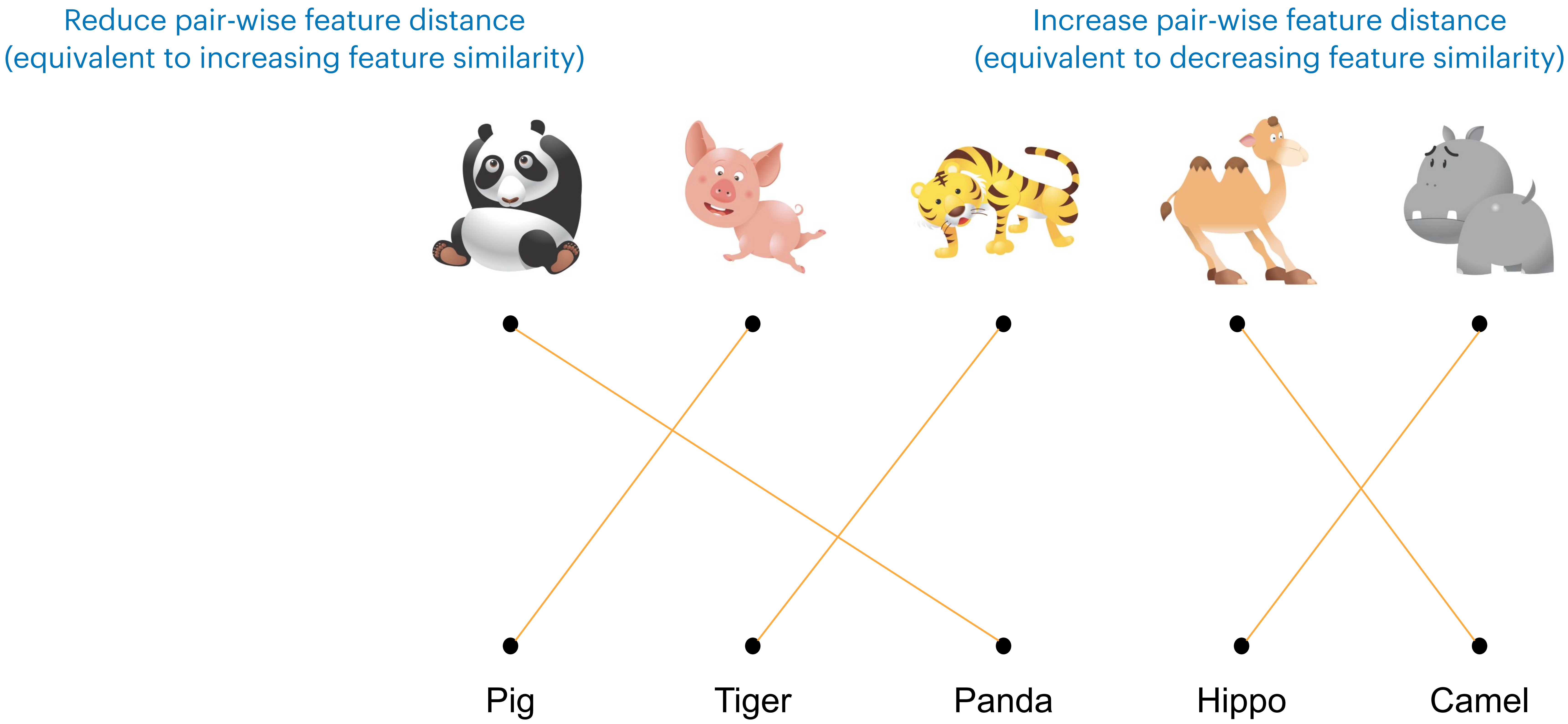
The goal is to associate each image with the correct label



https://icml.cc/media/icml-2021/Slides/9193.pdf

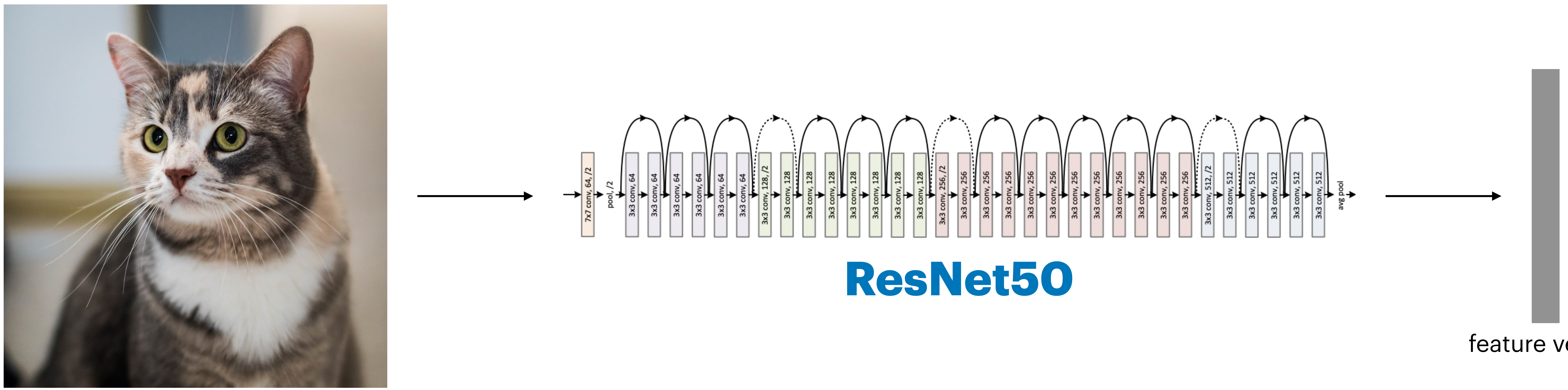
Contrastive learning

Pull together matched pairs while push away unmatched pairs

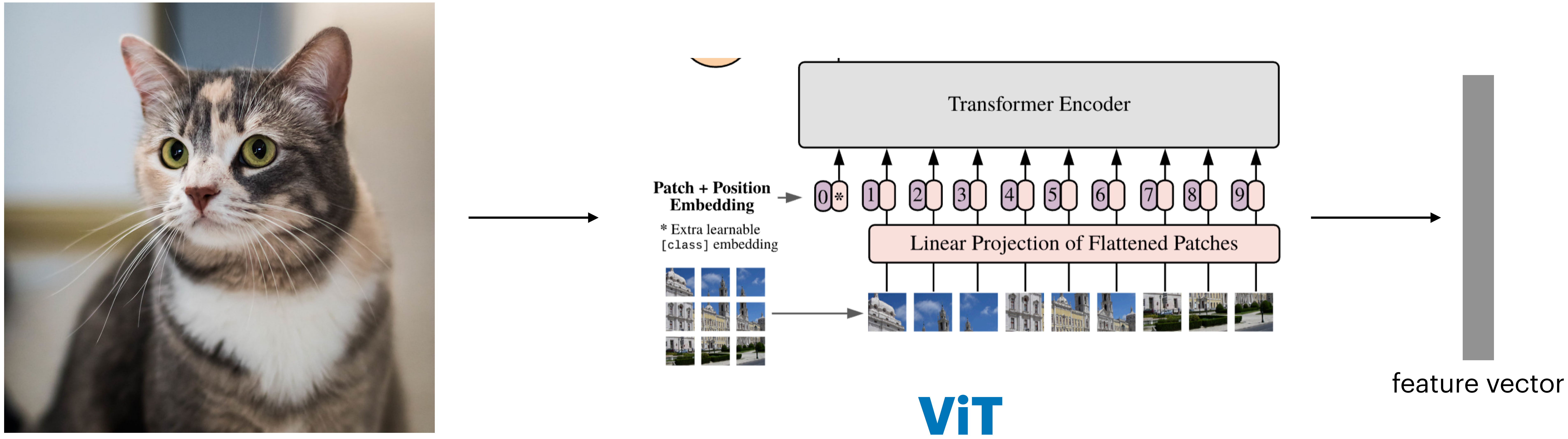


https://icml.cc/media/icml-2021/Slides/9193.pdf

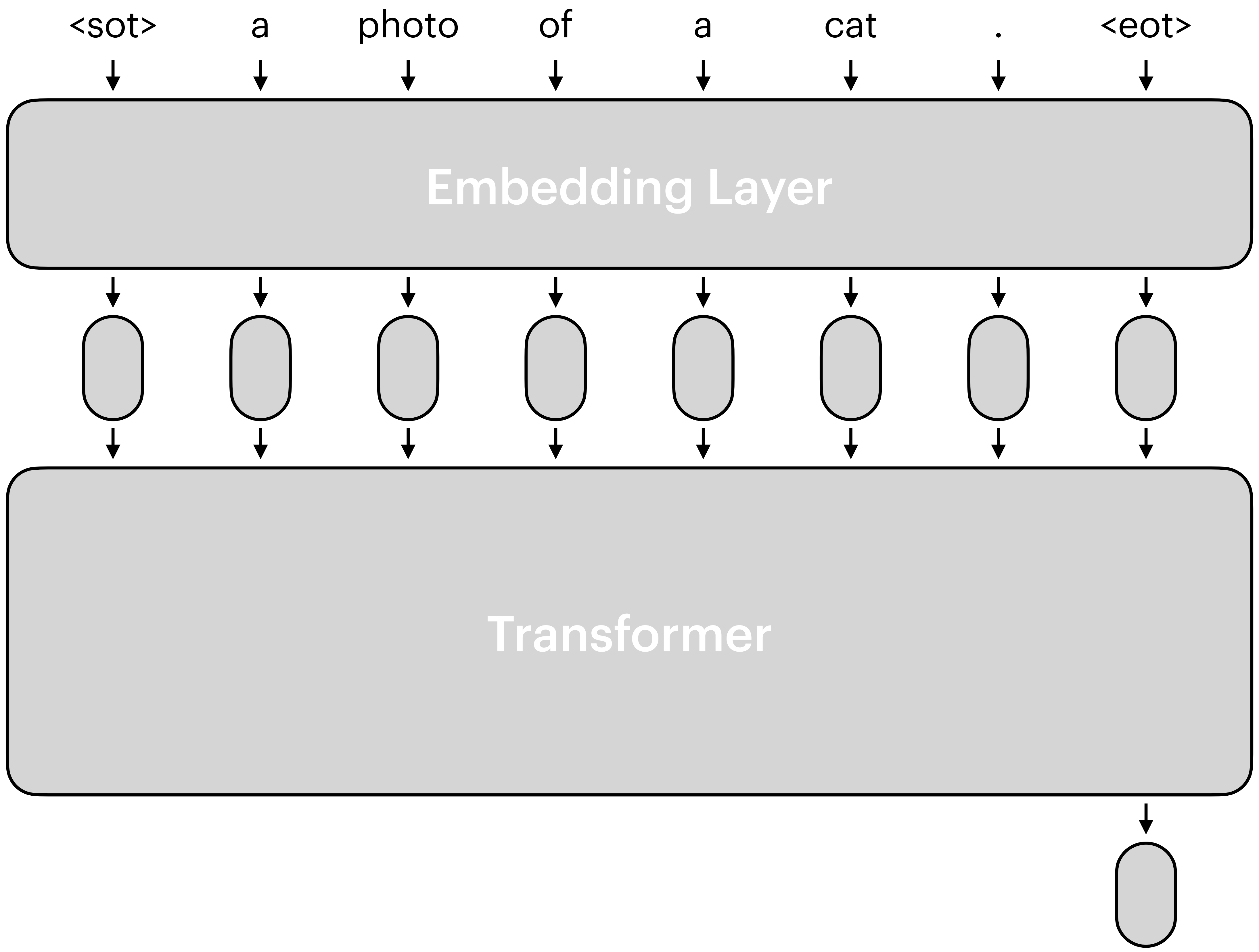
Architecture: image encoder

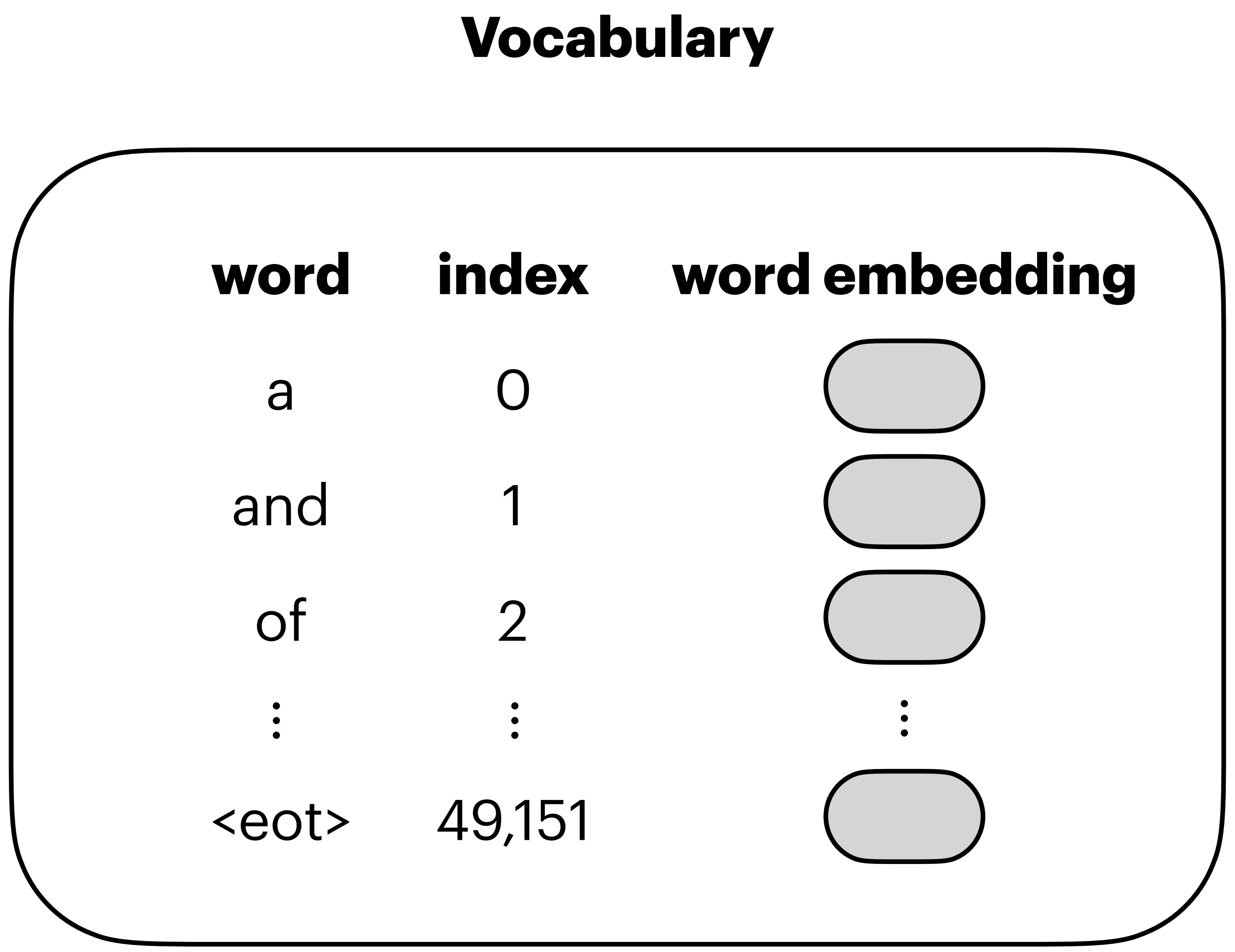


Architecture: image encoder

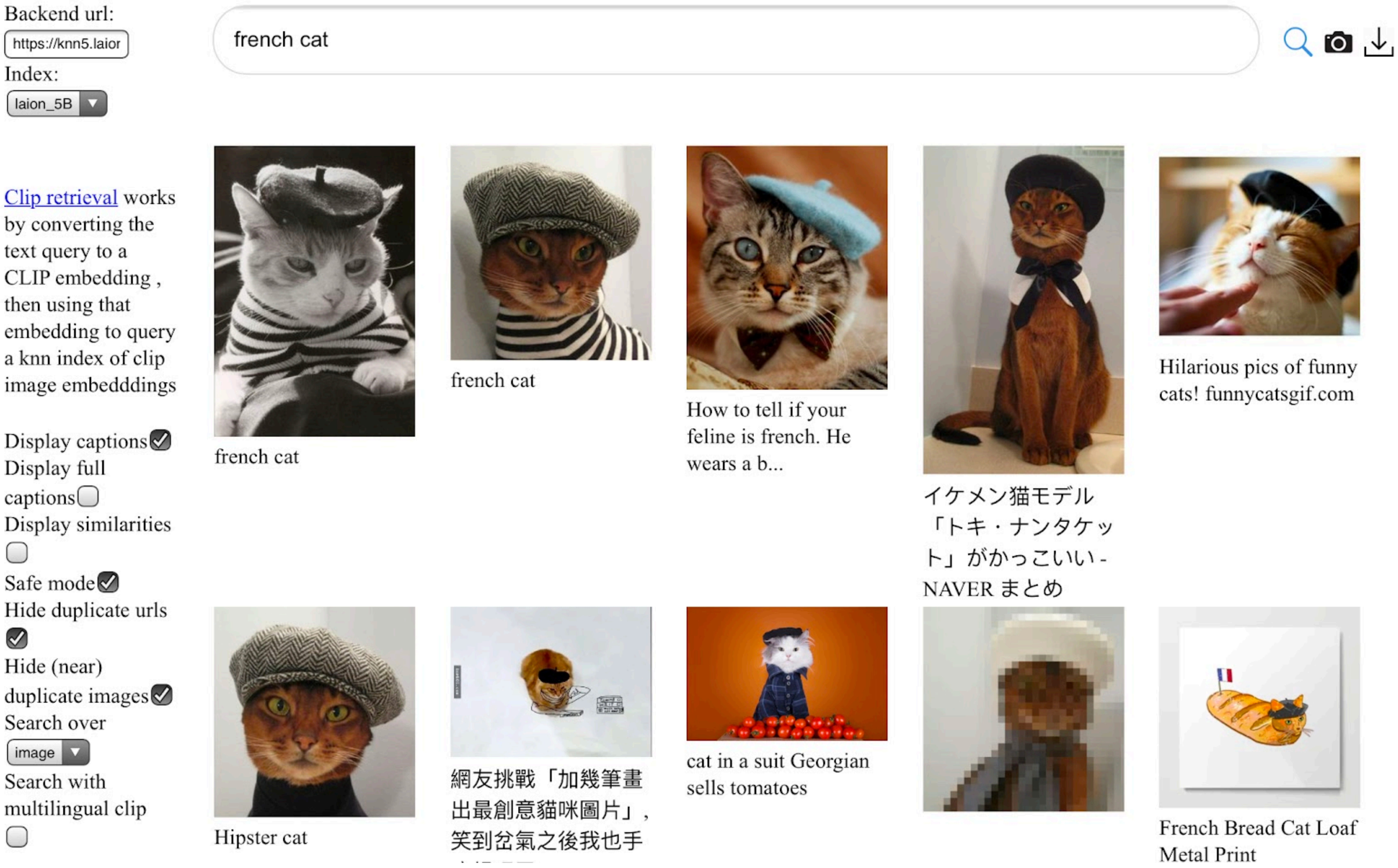


Architecture: text encoder



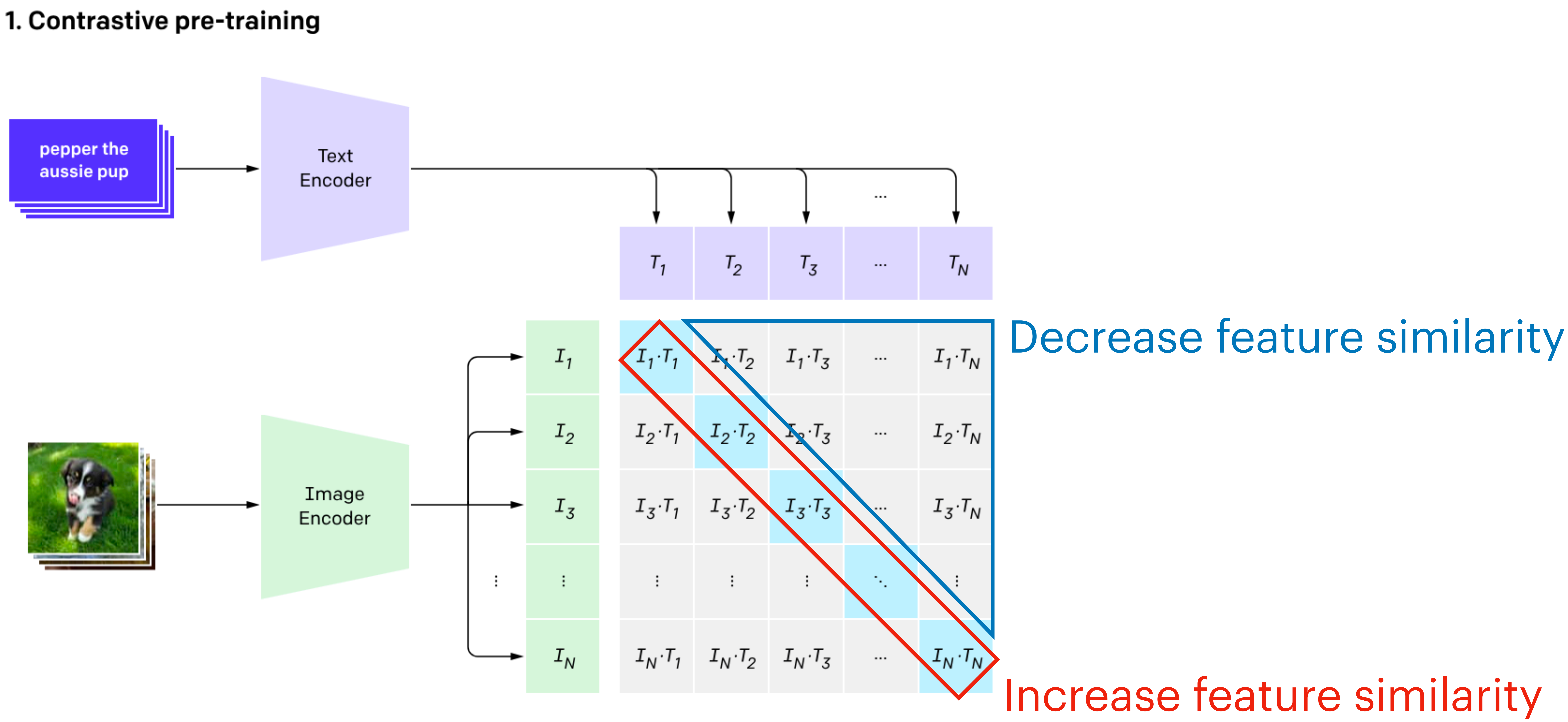


Data: LAION-5B



Schuhmann et al. LAION-5B: An open large-scale dataset for training next generation image-text models. NeurIPS'22.

Contrastive Language-Image Pre-training (CLIP)

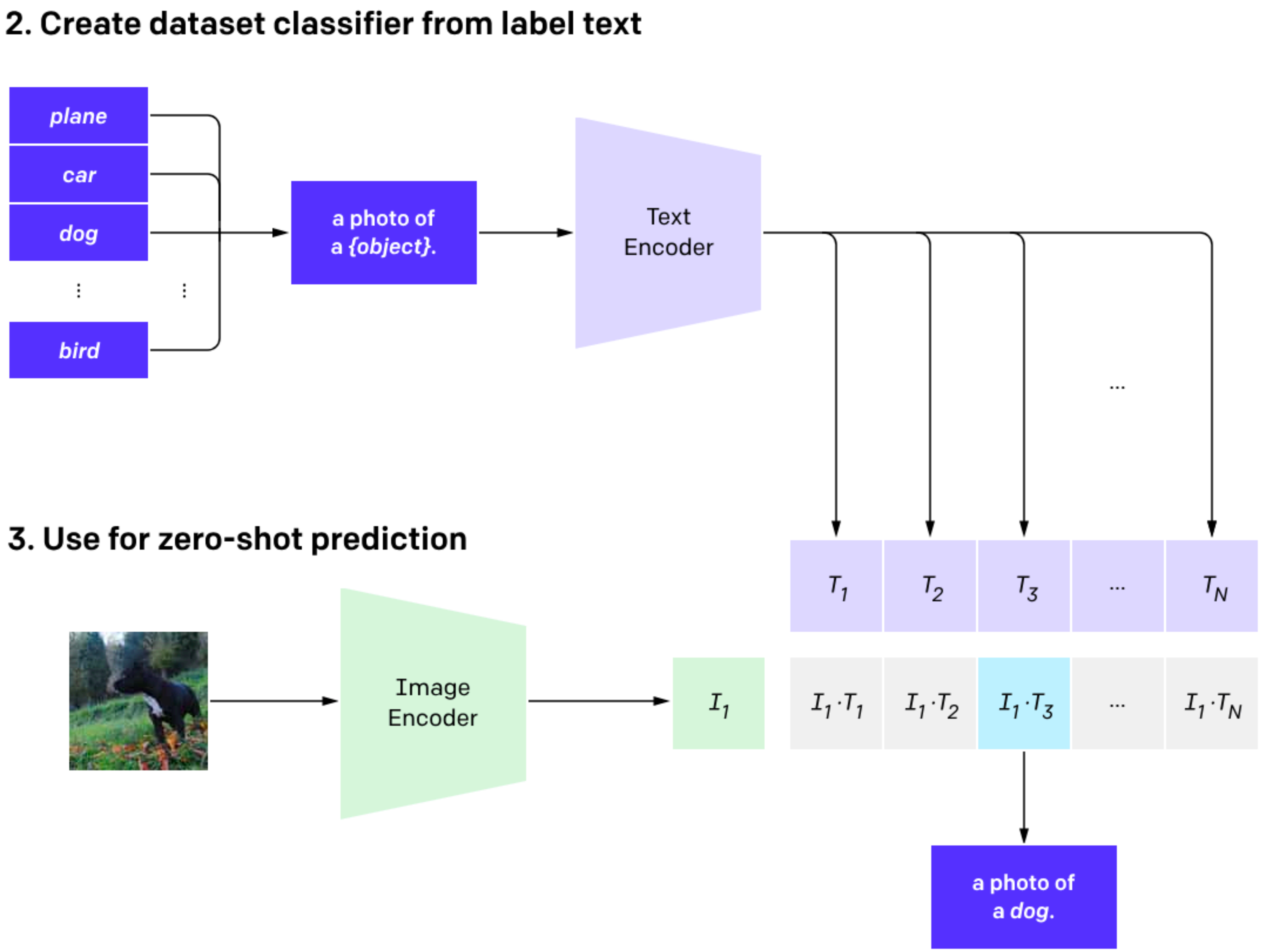


Radford et al. Learning Transferable Visual Models From Natural Language Supervision. ICML'21.

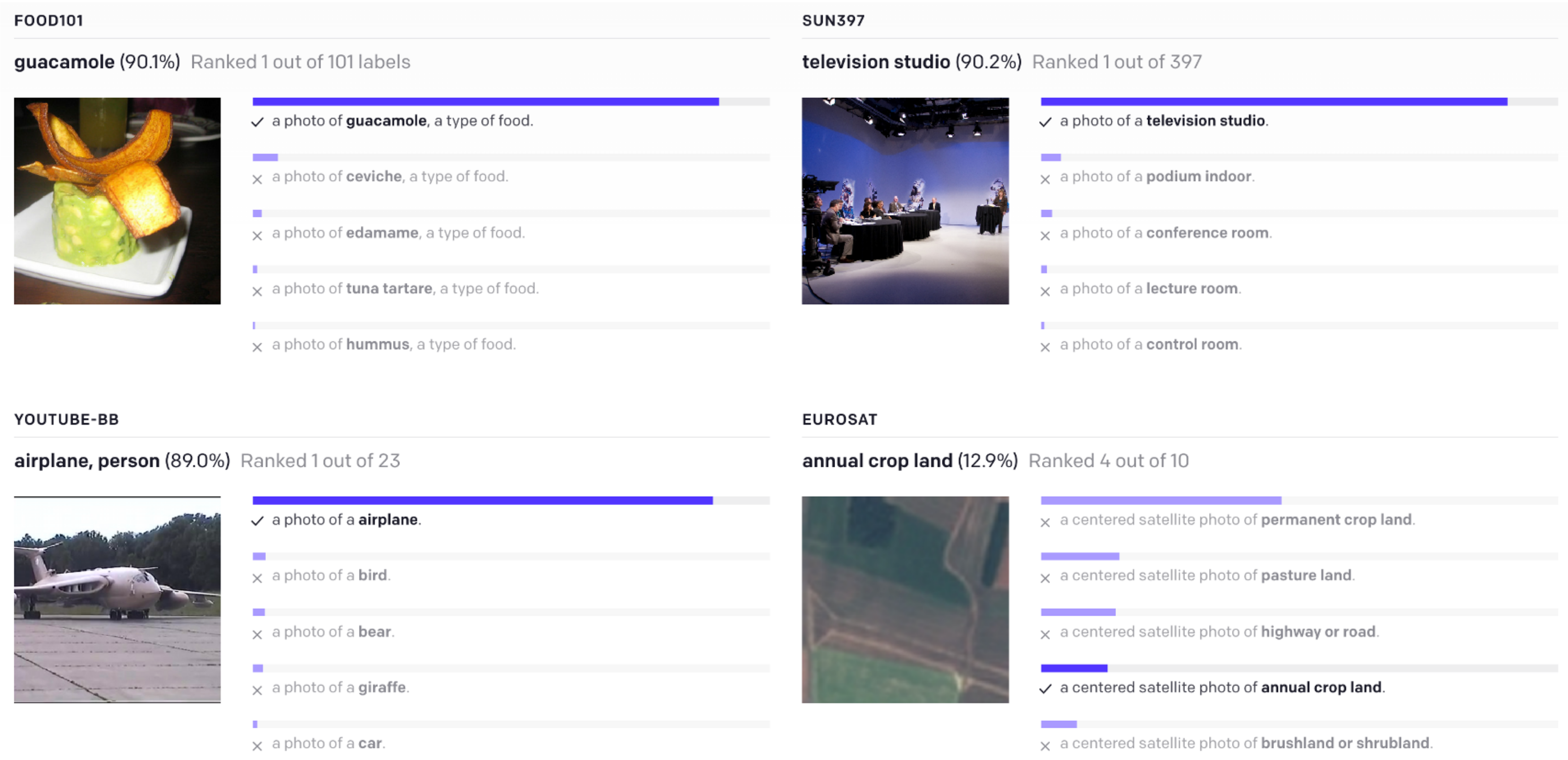
Outline

- History
- Pre-training
- Prompting
- Applications

Zero-shot prompting



Radford et al. Learning Transferable Visual Models From Natural Language Supervision. ICML'21.



Radford et al. Learning Transferable Visual Models From Natural Language Supervision. ICML’21.

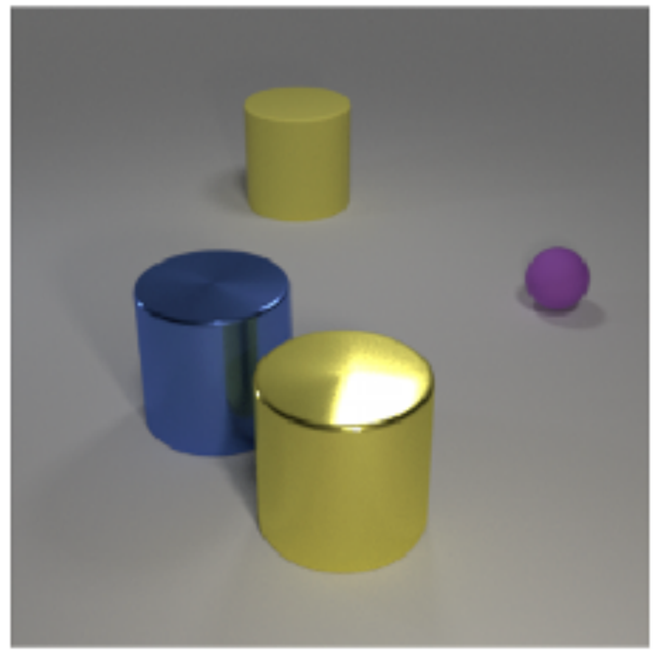
CIFAR-10

bird (40.9%) Ranked 1 out of 10 labels

- ✓ a photo of a **bird**.
- × a photo of a **cat**.
- × a photo of a **deer**.
- × a photo of a **frog**.
- × a photo of a **dog**.

CLEVR COUNT

4 (17.1%) Ranked 2 out of 8



- × a photo of **3** objects.
- ✓ a photo of **4** objects.
- × a photo of **5** objects.
- × a photo of **6** objects.
- × a photo of **10** objects.

FACIAL EMOTION RECOGNITION 2013 (FER2013)

angry (8.2%) Ranked 5 out of 7

- × a photo of a **happy** looking face.
- × a photo of a **neutral** looking face.
- × a photo of a **surprised** looking face.
- × a photo of a **fearful** looking face.
- ✓ a photo of a **angry** looking face.

UCF101

Volleyball Spiking (99.3%) Ranked 1 out of 101

- ✓ a photo of a person **volleyball spiking**.
- × a photo of a person **jump rope**.
- × a photo of a person **long jump**.
- × a photo of a person **soccer penalty**.
- × a photo of a person **table tennis shot**.

Radford et al. Learning Transferable Visual Models From Natural Language Supervision. ICML'21.

STANFORD CARS

2012 Honda Accord Coupe (63.3%) Ranked 1 out of 196

✓ a photo of a **2012 honda accord coupe**.

× a photo of a **2012 honda accord sedan**.

× a photo of a **2012 acura tl sedan**.

× a photo of a **2012 acura tsx sedan**.

× a photo of a **2008 acura tl type-s**.

SUN

kennel indoor (98.6%) Ranked 1 out of 723

✓ a photo of a **kennel indoor**.

× a photo of a **kennel outdoor**.

× a photo of a **jail cell**.

× a photo of a **jail indoor**.

× a photo of a **veterinarians office**.

KINETICS-700

country line dancing (99.0%) Ranked 1 out of 700

✓ a photo of **country line dancing**.

× a photo of **square dancing**.

× a photo of **swing dancing**.

× a photo of **dancing charleston**.

× a photo of **salsa dancing**.

FLOWERS-102

great masterwort (74.3%) Ranked 1 out of 102

✓ a photo of a **great masterwort**, a type of flower.

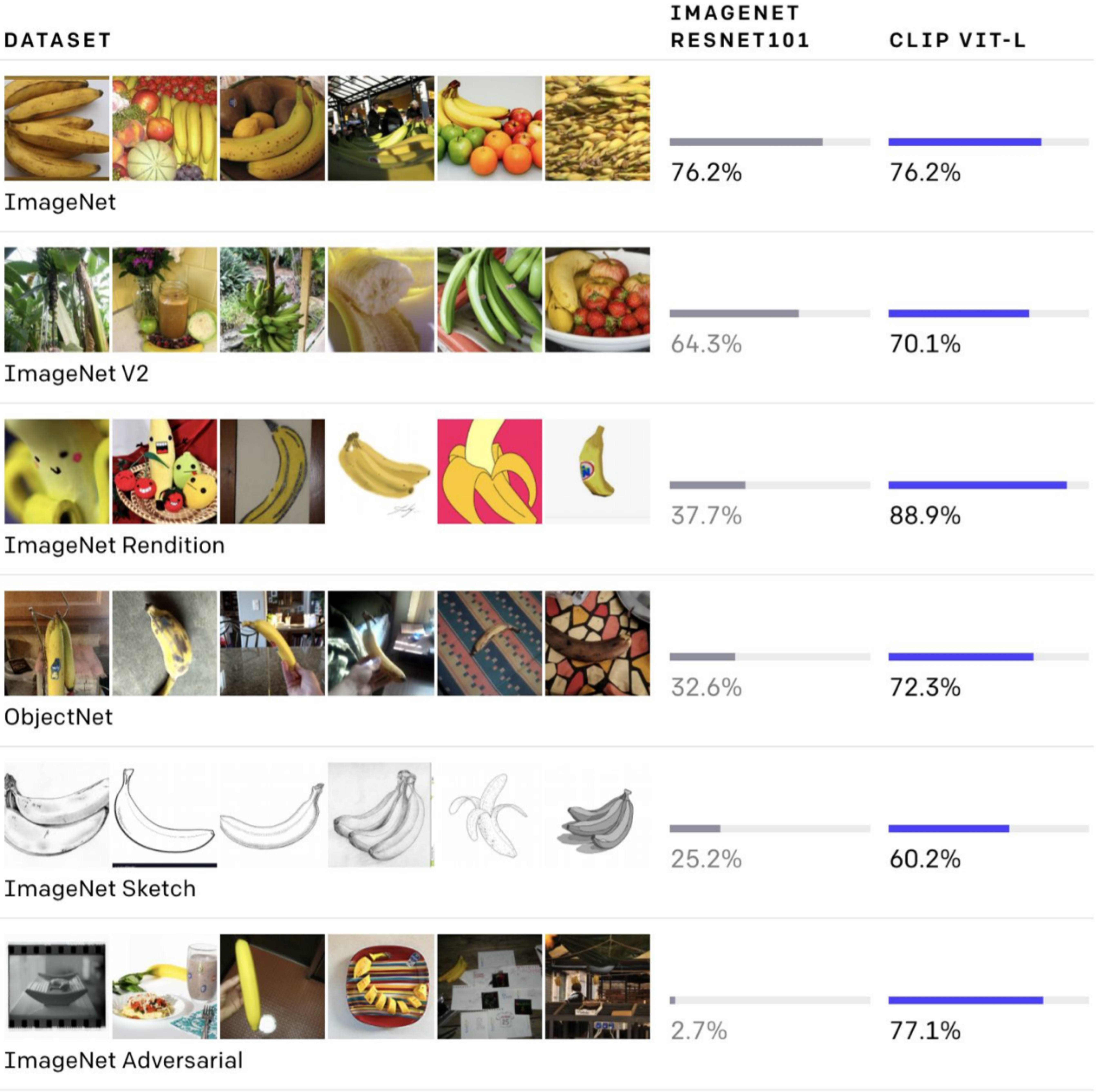
× a photo of a **bishop of llandaff**, a type of flower.

× a photo of a **pincushion flower**, a type of flower.

× a photo of a **globe flower**, a type of flower.

× a photo of a **prince of wales feathers**, a type of flower.

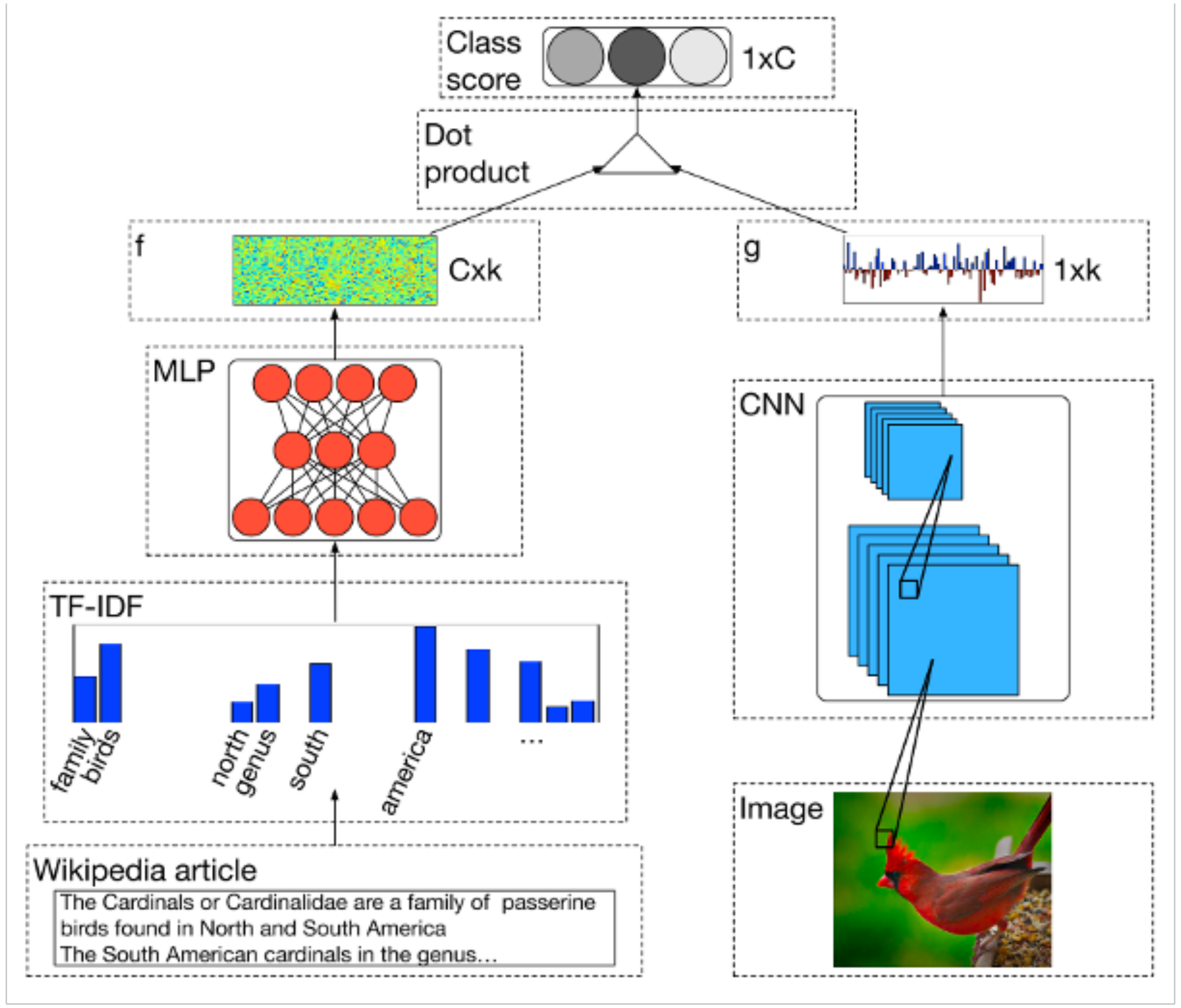
Radford et al. Learning Transferable Visual Models From Natural Language Supervision. ICML'21.



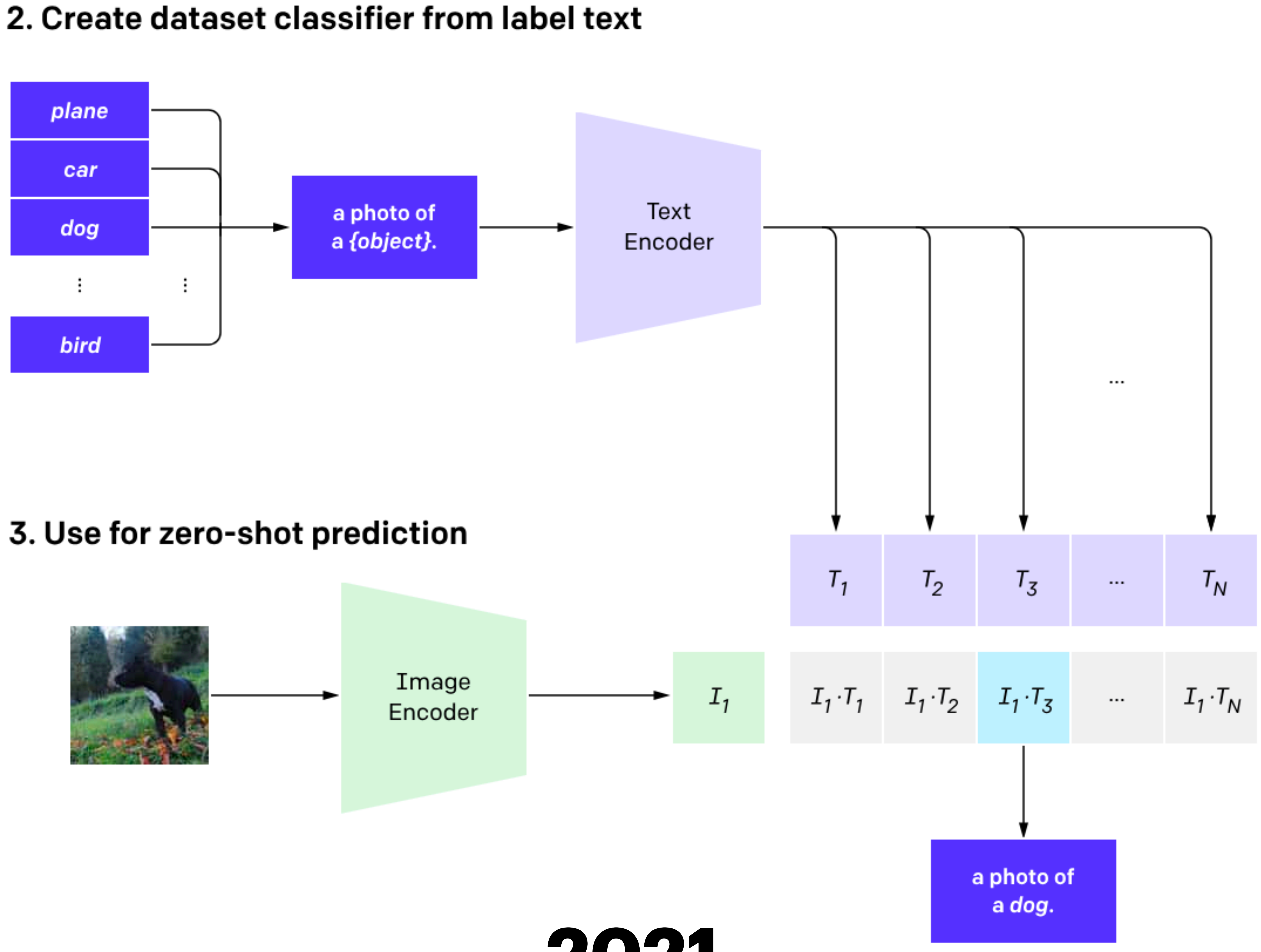
Radford et al. Learning Transferable Visual Models From Natural Language Supervision. ICML'21.

Why CLIP works?

- Scaling (model & data)
- Transformer
- Contrastive learning



2015



2021

Prompt engineering

Text prompts
A photo of a {dog}
A photo of a {cat}
A photo of a {bird}
...
A photo of a {tiger}

a bad photo of a {}.
a photo of many {}.
a sculpture of a {}.
a photo of the hard to see {}.
a low resolution photo of the {}.
a rendering of a {}.
graffiti of a {}.
a bad photo of the {}.
a cropped photo of the {}.
a tattoo of a {}.
the embroidered {}.
a photo of a hard to see {}.
a bright photo of a {}.
a photo of a clean {}.
a photo of a dirty {}.
a dark photo of the {}.
a drawing of a {}.
a photo of my {}.
the plastic {}.
a photo of the cool {}.
a close-up photo of a {}.
a black and white photo of the {}.
a painting of the {}.
a painting of a {}.

a pixelated photo of the {}.
a sculpture of the {}.
a bright photo of the {}.
a cropped photo of a {}.
a plastic {}.
a photo of the dirty {}.
a jpeg corrupted photo of a {}.
a blurry photo of the {}.
a photo of the {}.
a good photo of the {}.
a rendering of the {}.
a {} in a video game.
a photo of one {}.
a doodle of a {}.
a close-up photo of the {}.
a photo of a {}.
the origami {}.
the {} in a video game.
a sketch of a {}.
a doodle of the {}.
a origami {}.
a low resolution photo of a {}.
the toy {}.
a rendition of the {}.

a photo of the clean {}.
a photo of a large {}.
a rendition of a {}.
a photo of a nice {}.
a photo of a weird {}.
a blurry photo of a {}.
a cartoon {}.
art of a {}.
a sketch of the {}.
a embroidered {}.
a pixelated photo of a {}.
itap of the {}.
a jpeg corrupted photo of the {}.
a good photo of a {}.
a plushie {}.
a photo of the nice {}.
a photo of the small {}.
a photo of the weird {}.
the cartoon {}.
art of the {}.
a drawing of the {}.
a photo of the large {}.
a black and white photo of a {}.
the plushie {}.

https://github.com/openai/CLIP/blob/main/notebooks/Prompt_Engineering_for_ImageNet.ipynb

Prompt engineering is hard

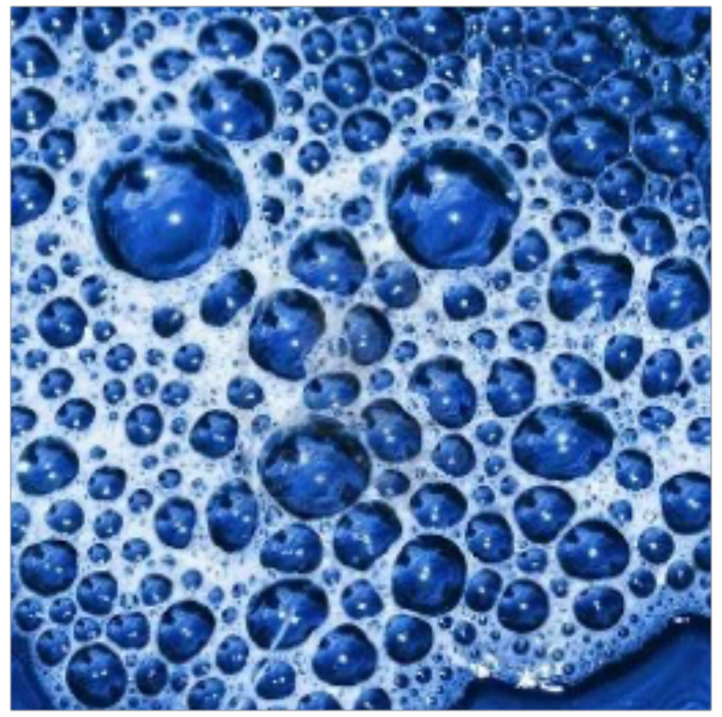
Caltech101

Prompt	Accuracy
a [CLASS].	82.68
a photo of [CLASS].	80.81
a photo of a [CLASS].	86.29
$[V]_1$ $[V]_2$ … $[V]_M$ [CLASS].	**91.83**

Flowers102

Prompt	Accuracy
a photo of a [CLASS].	60.86
a flower photo of a [CLASS].	65.81
a photo of a [CLASS], a type of flower.	66.14
$[V]_1$ $[V]_2$ … $[V]_M$ [CLASS].	**94.51**

Describable Textures (DTD)



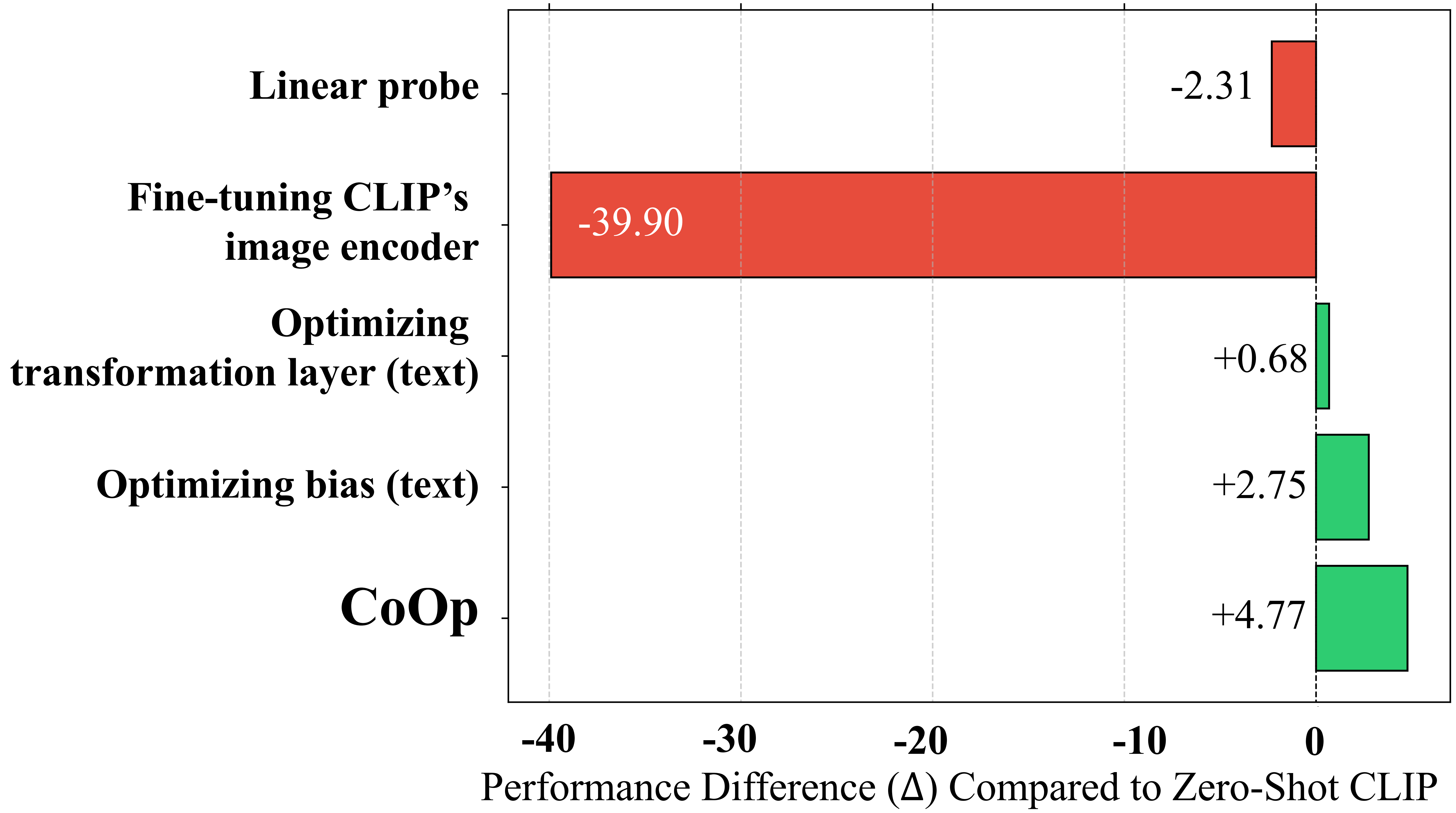
Prompt	Accuracy
a photo of a [CLASS].	39.83
a photo of a [CLASS] texture.	40.25
[CLASS] texture.	42.32
$[V]_1$ $[V]_2$ … $[V]_M$ [CLASS].	**63.58**

EuroSAT

Prompt	Accuracy
a photo of a [CLASS].	24.17
a satellite photo of [CLASS].	37.46
a centered satellite photo of [CLASS].	37.56
$[V]_1$ $[V]_2$ … $[V]_M$ [CLASS].	**83.53**

Zhou et al. "Learning to prompt for vision-language models." *International Journal of Computer Vision* 130.9 (2022): 2337-2348.

Fine-tuning is also hard



Overfitting! The model is too big

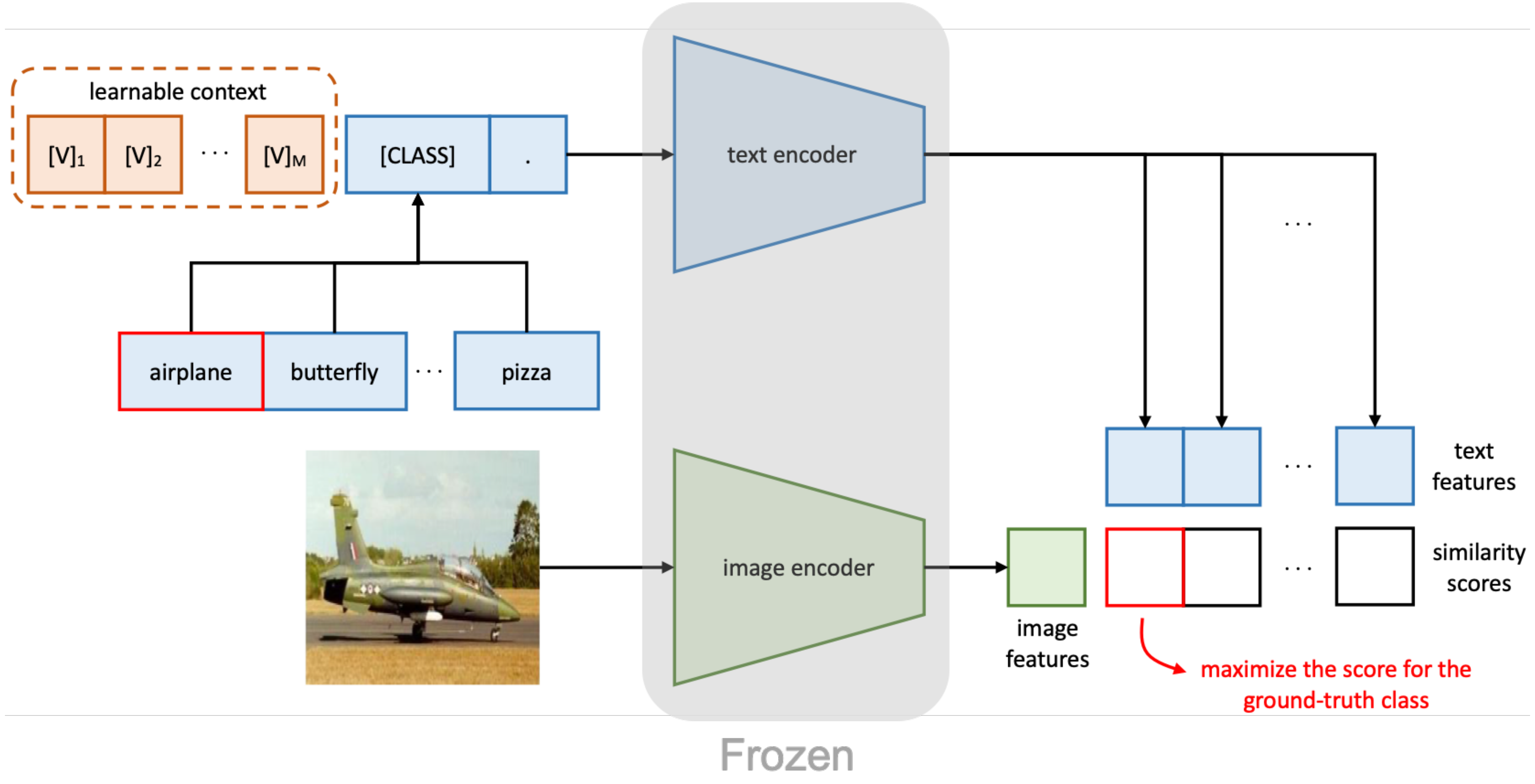
Partial fine-tuning seems working (but which parts?)

Prompt learning works the best!

Zhou et al. "Learning to prompt for vision-language models." *International Journal of Computer Vision* 130.9 (2022): 2337-2348.

Context Optimization (CoOp)

/ku:p/



By forwarding a prompt $\boldsymbol{t}$ to the text encoder $g(\cdot)$, we can obtain a classification weight vector representing a visual concept (still from the [EOS] token position). The prediction probability is computed as

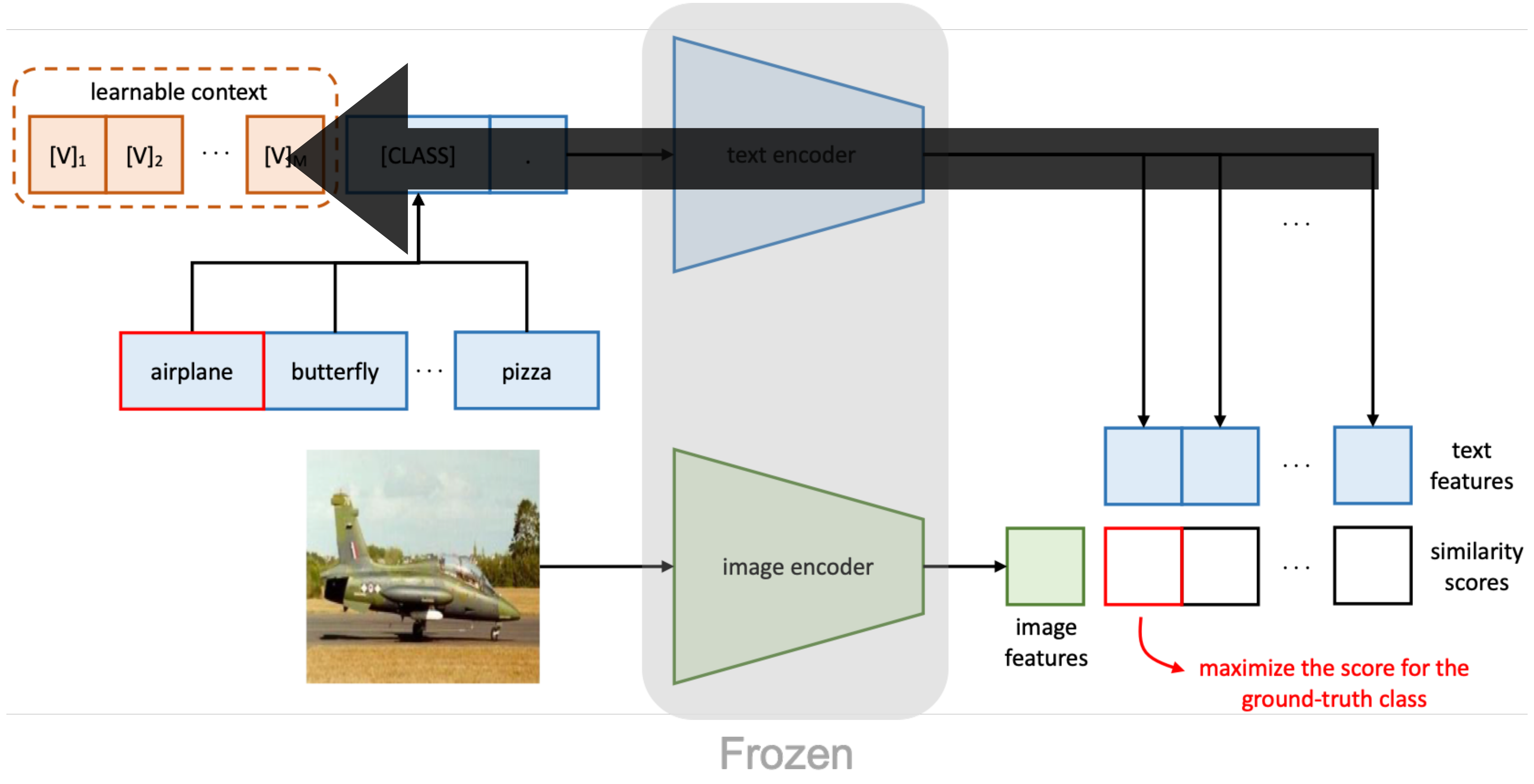
$$p(y=i|\boldsymbol{x}) = \frac{\exp(\cos(g(\boldsymbol{t}_i), \boldsymbol{f})/\tau)}{\sum_{j=1}^{K} \exp(\cos(g(\boldsymbol{t}_j), \boldsymbol{f})/\tau)}, \tag{3}$$

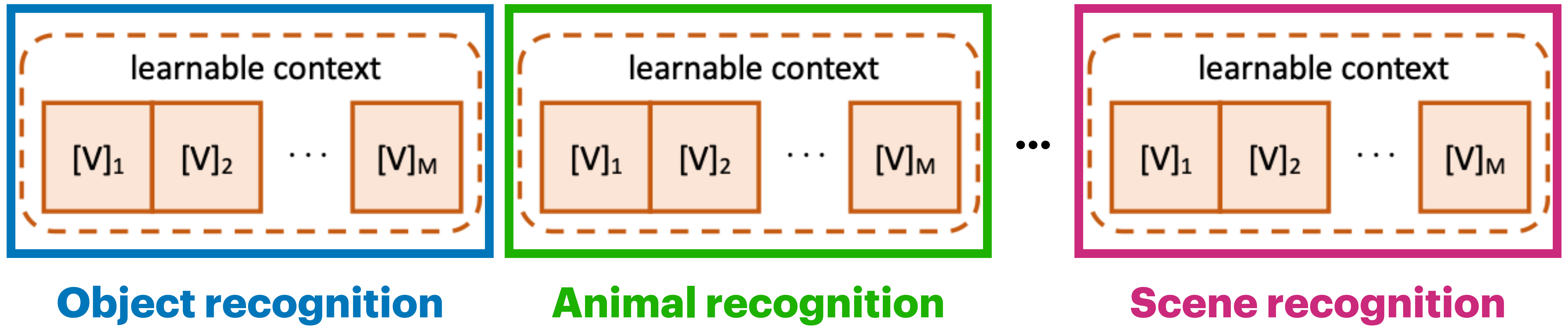
where the class token within each prompt $\boldsymbol{t}_i$ is replaced by the corresponding word embedding vector(s) of the i-th class name.

Zhou et al. "Learning to prompt for vision-language models." *International Journal of Computer Vision* 130.9 (2022): 2337-2348.

Why do prompt learning?

- Enjoys rich gradient information
- Mitigates overfitting (few parameters)
- Reduces storage cost (one per task or user)

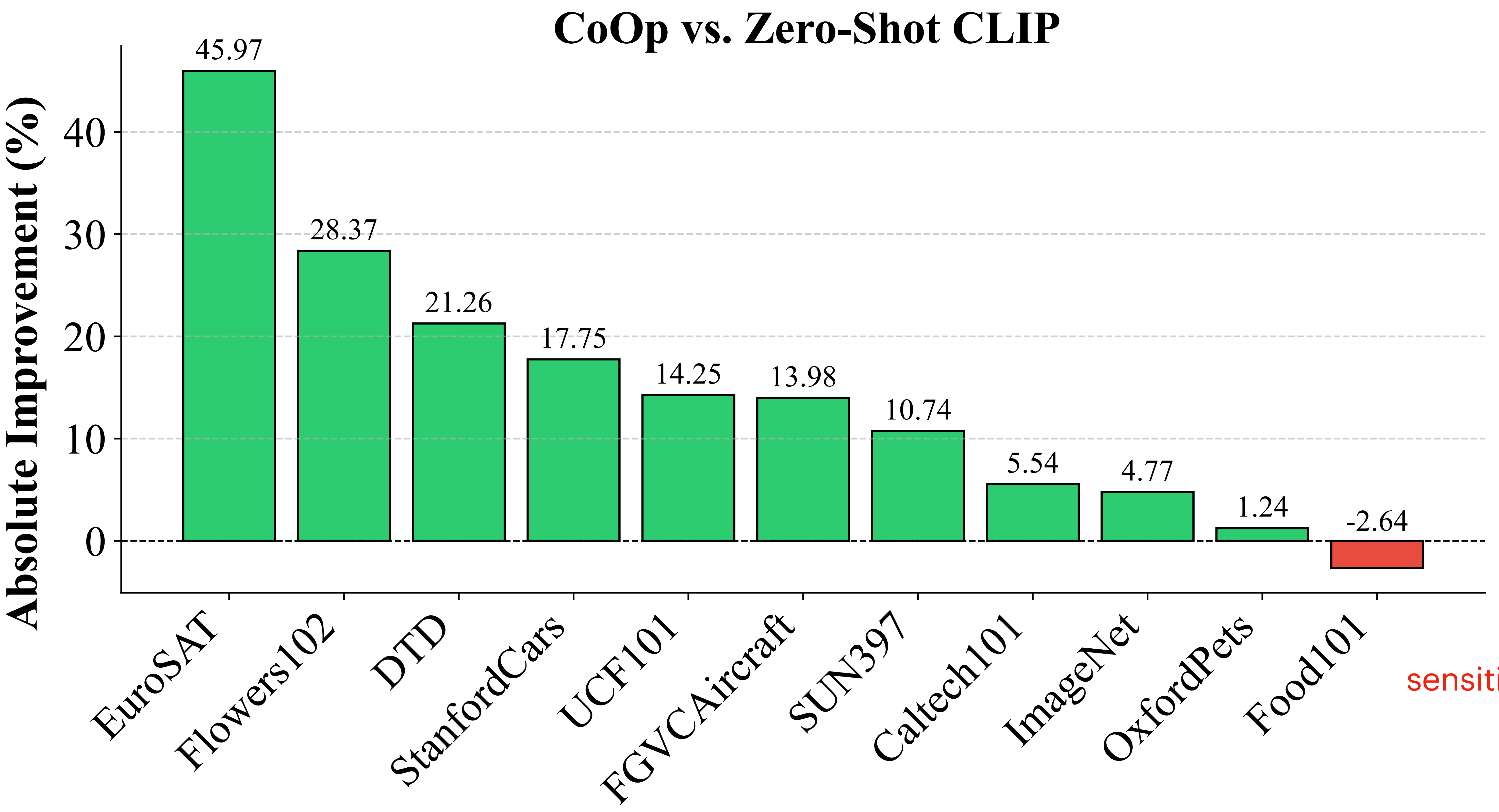




Zhou et al. "Learning to prompt for vision-language models." *International Journal of Computer Vision* 130.9 (2022): 2337-2348.

Few-shot learning

- Works on diverse tasks (objects, animals, scenes, actions, etc.)
- Significantly beats hand-crafted prompts (also needs labels for tuning)



sensitive to label noise in Food101

Zhou et al. "Learning to prompt for vision-language models." *International Journal of Computer Vision* 130.9 (2022): 2337-2348.

Domain generalization

- Train on one dataset but test on a different one with domain shifts
- Still beats hand-crafted prompts despite being a learning-based approach

	CLIP	Ours
ImageNet (source)	58.18	**63.33**
V2 (target)	51.34	**55.40**
Sketch (target)	33.32	**34.67**
Adversarial (target)	21.65	**23.06**
Rendition (target)	56.00	**56.60**

Zhou et al. "Learning to prompt for vision-language models." *International Journal of Computer Vision* 130.9 (2022): 2337-2348.

Interpretable? Not really

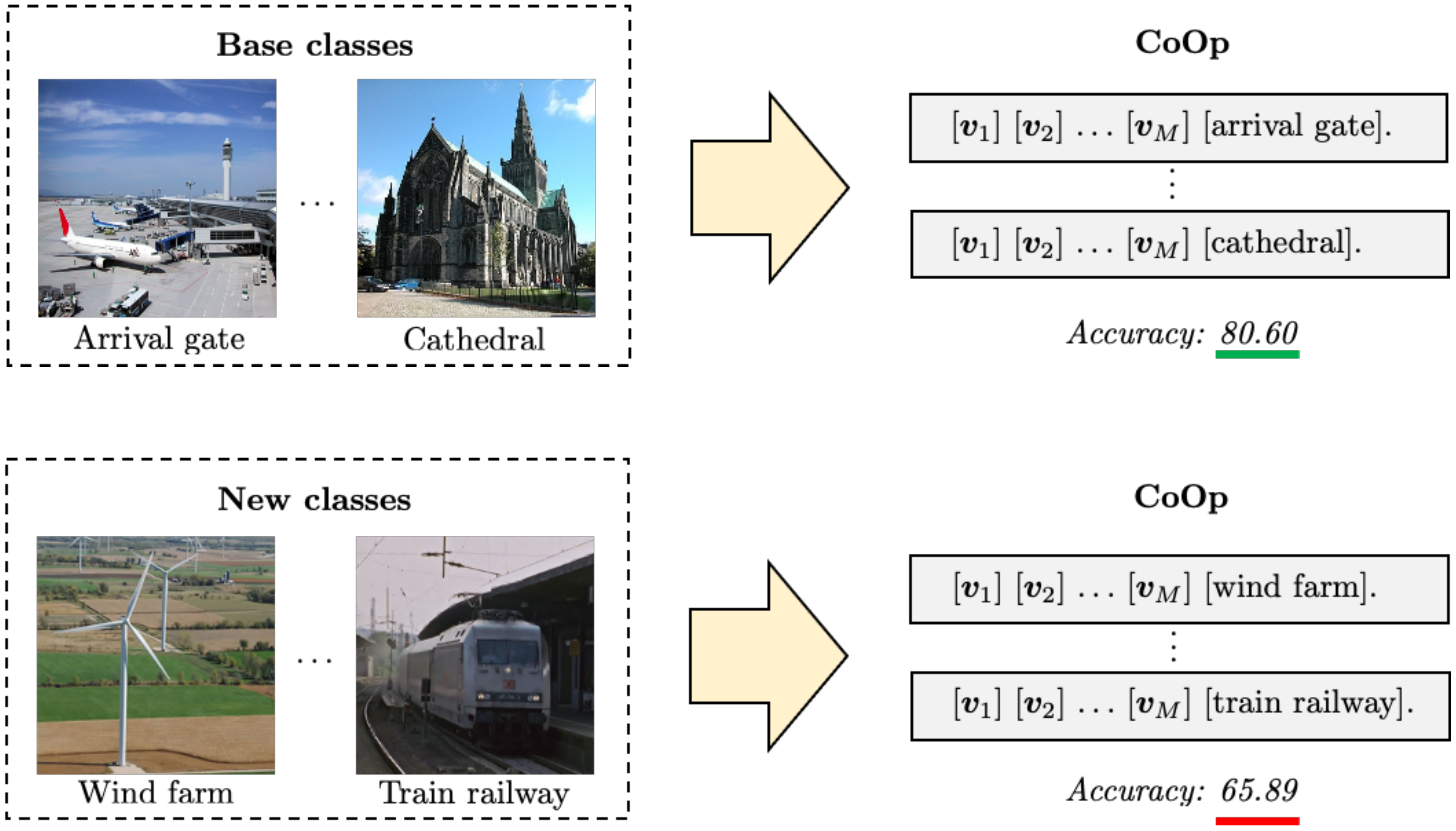
Finding 1: few are somewhat relevant

Finding 2: the whole prompt does not make much sense

#	ImageNet	Food101	OxfordPets	DTD	UCF101
1	Potd (1.7136)	Lc (0.6752)	Tosc (2.5952)	Boxed (0.9433)	Meteorologist (1.5377)
2	That (1.4015)	Enjoyed (0.5305)	Judge (1.2635)	Seed (1.0498)	Exe (0.9807)
3	Filmed (1.2275)	Beh (0.5390)	Fluffy (1.6099)	Anna (0.8127)	Parents (1.0654)
4	Fruit (1.4864)	Matches (0.5646)	Cart (1.3958)	Mountain (0.9509)	Masterful (0.9528)
5	,... (1.5863)	Nytimes (0.6993)	Harlan (2.2948)	Eldest (0.7111)	Fe (1.3574)
6	°(1.7502)	Prou (0.5905)	Paw (1.3055)	Pretty (0.8762)	Thof (1.2841)
7	Excluded (1.2355)	Lower (0.5390)	Incase (1.2215)	Faces (0.7872)	Where (0.9705)
8	Cold (1.4654)	N/A	Bie (1.5454)	Honey (1.8414)	Kristen (1.1921)
9	Stery (1.6085)	Minute (0.5672)	Snuggle (1.1578)	Series (1.6680)	Imam (1.1297)
10	Warri (1.3055)	~ (0.5529)	Along (1.8298)	Coca (1.5571)	Near (0.8942)
11	Marvelcomics (1.5638)	Well (0.5659)	Enjoyment (2.3495)	Moon (1.2775)	Tummy (1.4303)
12	.: (1.7387)	Ends (0.6113)	Jt (1.3726)	Ih (1.0382)	Hel (0.7644)
13	N/A	Mis (0.5826)	Improving (1.3198)	Won (0.9314)	Boop (1.0491)
14	Lation (1.5015)	Somethin (0.6041)	Srsly (1.6759)	Replied (1.1429)	N/A
15	Muh (1.4985)	Seminar (0.5274)	Asteroid (1.3395)	Sent (1.3173)	Facial (1.4452)
16	.# (1.9340)	N/A	N/A	Piedmont (1.5198)	During (1.1755)

Zhou et al. "Learning to prompt for vision-language models." *International Journal of Computer Vision* 130.9 (2022): 2337-2348.

Generalize beyond the training labels?



… only works for a subset of classes (overfitting)

Zhou et al. "Conditional prompt learning for vision-language models." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

ImageNet

↓8.86%

Caltech101

↓8.19%

Flowers102

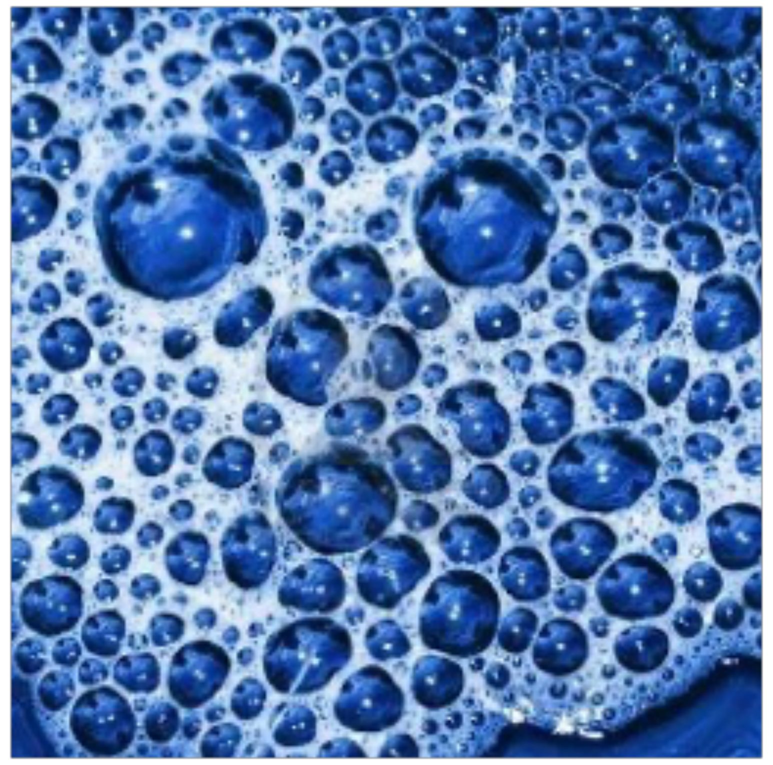
↓37.93%

StanfordCars

↓17.72%

FGVCAircraft

↓18.14%



DTD

↓38.26%

EuroSAT

↓37.45%

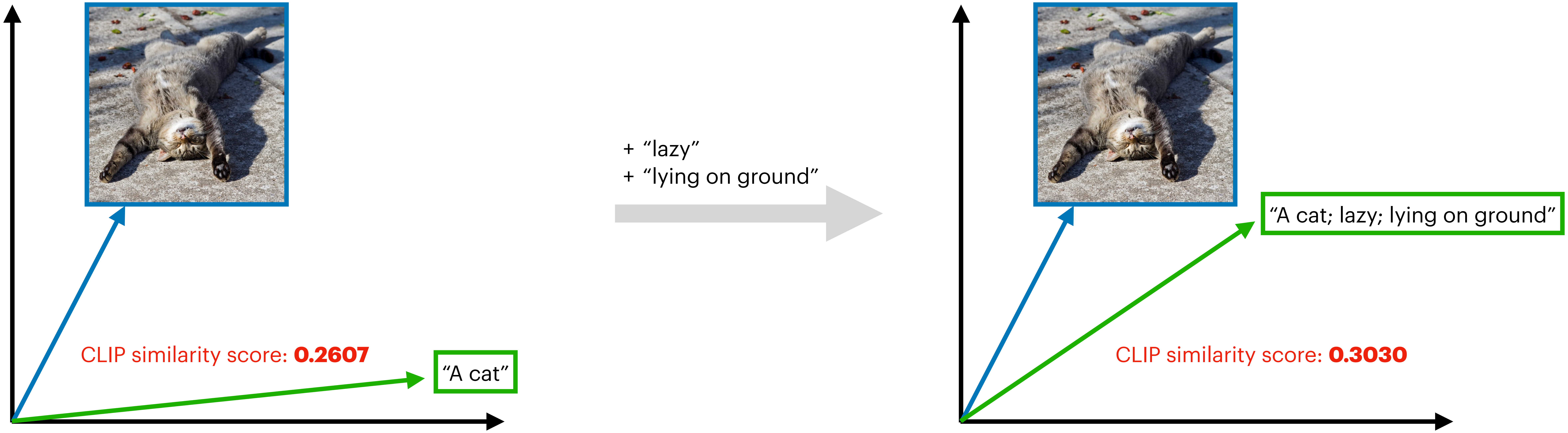
UCF101

↓28.64%

Zhou et al. "Conditional prompt learning for vision-language models." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

What is a good prompt?

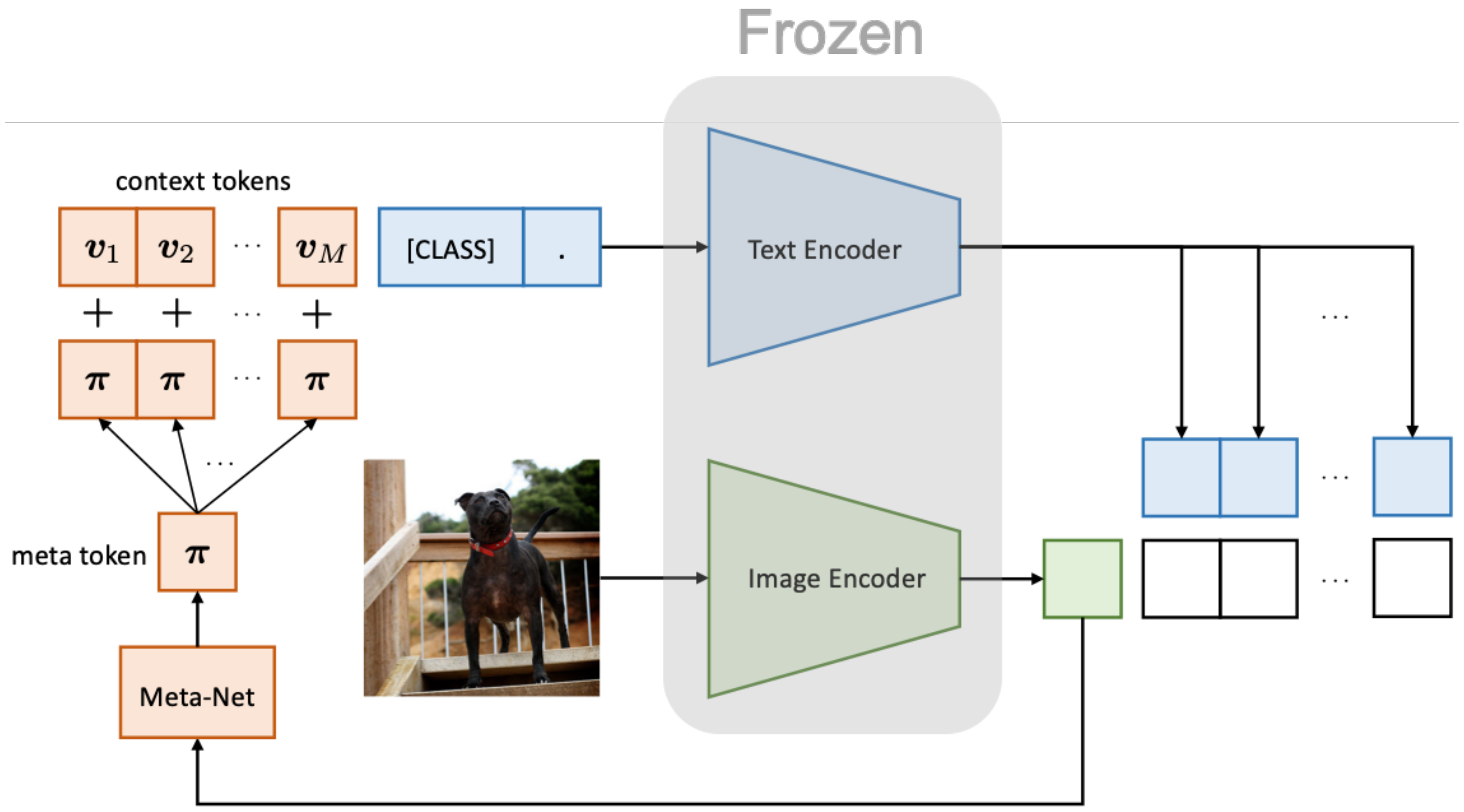
- Contains instance-specific information
- Pushes text features closer to image features



Zhou et al. "Conditional prompt learning for vision-language models." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

Conditional Context Optimization (CoCoOp)

/kəʊˌkuːp/



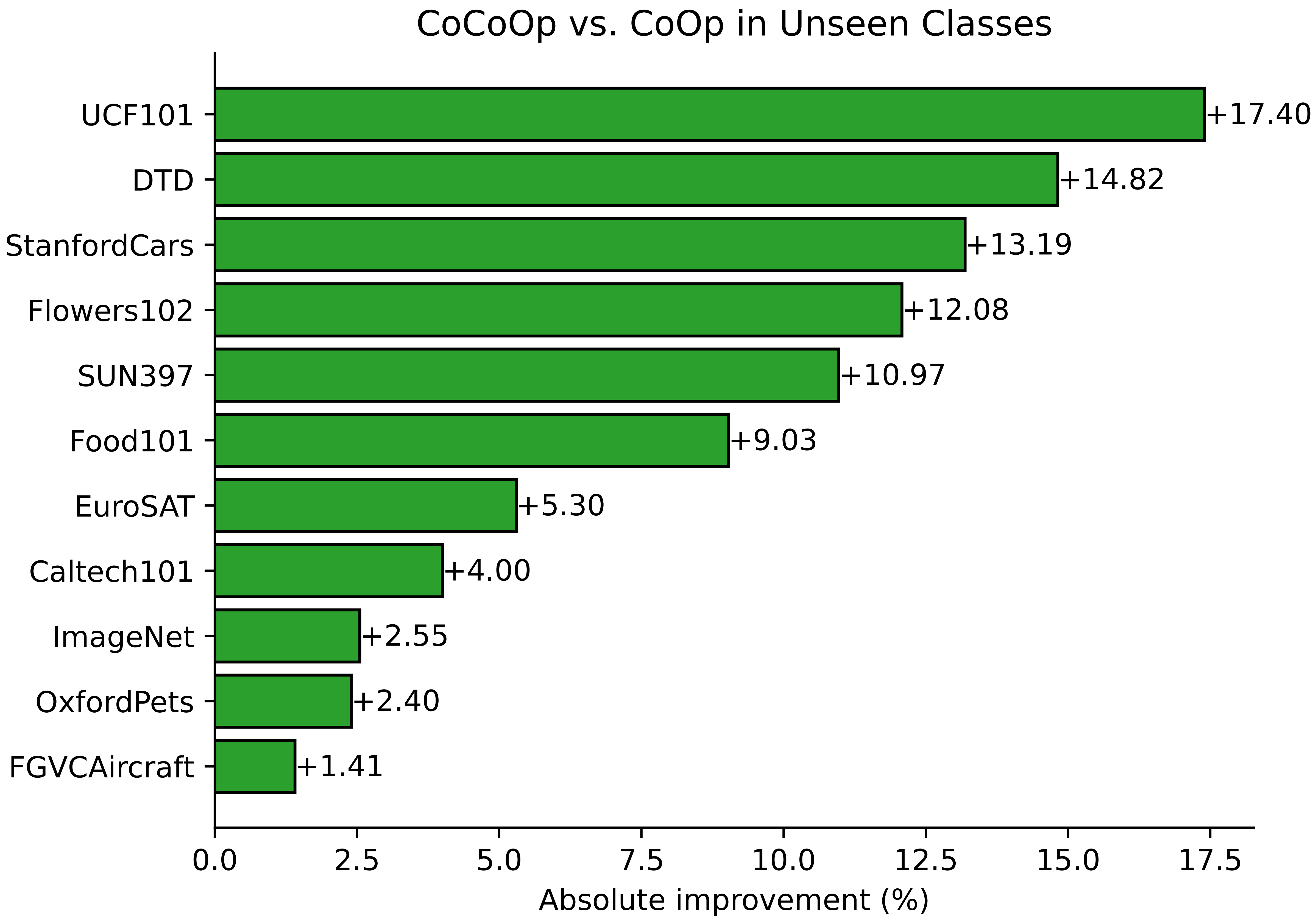
Let $h_{\boldsymbol{\theta}}(\cdot)$ denote the Meta-Net parameterized by $\boldsymbol{\theta}$, each context token is now obtained by $\boldsymbol{v}_m(\boldsymbol{x}) = \boldsymbol{v}_m + \boldsymbol{\pi}$ where $\boldsymbol{\pi} = h_{\boldsymbol{\theta}}(\boldsymbol{x})$ and $m \in \{1, 2, ..., M\}$. The prompt for the i-th class is thus conditioned on the input, i.e., $\boldsymbol{t}_i(\boldsymbol{x}) = \{\boldsymbol{v}_1(\boldsymbol{x}), \boldsymbol{v}_2(\boldsymbol{x}), \ldots, \boldsymbol{v}_M(\boldsymbol{x}), \boldsymbol{c}_i\}$. The prediction probability is computed as

$$p(y|\boldsymbol{x}) = \frac{\exp(\text{sim}(\boldsymbol{x}, g(\boldsymbol{t}_y(\boldsymbol{x})))/\tau)}{\sum_{i=1}^{K} \exp(\text{sim}(\boldsymbol{x}, g(\boldsymbol{t}_i(\boldsymbol{x}))/\tau)}. \quad (3)$$

Zhou et al. "Conditional prompt learning for vision-language models." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

Key messages

1. Conditional prompt learning is more generalizable



Zhou et al. "Conditional prompt learning for vision-language models." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

Key messages

1. Conditional prompt learning is more generalizable
2. Conditional prompt learning is more transferable

Table 2. **Comparison of prompt learning methods in the cross-dataset transfer setting**. Prompts applied to the 10 target datasets are learned from ImageNet (16 images per class). Clearly, CoCoOp demonstrates better transferability than CoOp. Δ denotes CoCoOp's gain over CoOp.

	Source	Target										
	ImageNet	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	FGVCAircraft	SUN397	DTD	EuroSAT	UCF101	*Average*
CoOp [62]	**71.51**	93.70	89.14	64.51	68.71	85.30	18.47	64.15	41.92	**46.39**	66.55	63.88
CoCoOp	71.02	**94.43**	**90.14**	**65.32**	**71.88**	**86.06**	**22.94**	**67.36**	**45.73**	45.37	**68.21**	**65.74**
Δ	**-0.49**	**+0.73**	**+1.00**	**+0.81**	**+3.17**	**+0.76**	**+4.47**	**+3.21**	**+3.81**	**-1.02**	**+1.66**	**+1.86**

Zhou et al. "Conditional prompt learning for vision-language models." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

Key messages

1. Conditional prompt learning is more generalizable
2. Conditional prompt learning is more transferable
3. Conditional prompt learning is more robust

Table 3. **Comparison of manual and learning-based prompts in domain generalization**. CoOp and CoCoOp use as training data 16 images from each of the 1,000 classes on ImageNet. In general, CoCoOp is more domain-generalizable than CoOp.

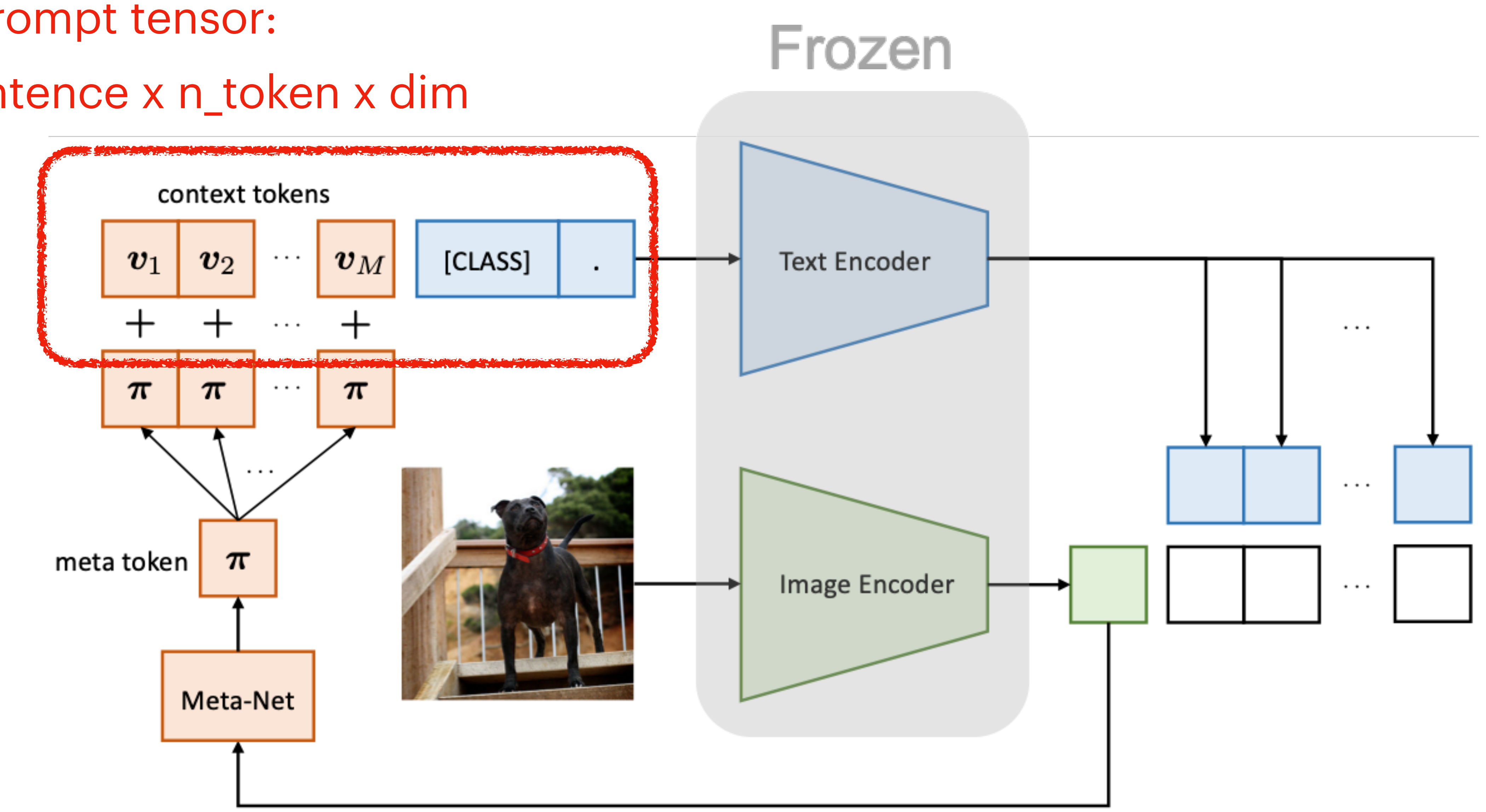
		Source	Target			
	Learnable?	ImageNet	ImageNetV2	ImageNet-Sketch	ImageNet-A	ImageNet-R
CLIP [40]		66.73	60.83	46.15	47.77	73.96
CoOp [62]	✓	**71.51**	**64.20**	47.99	49.71	75.21
CoCoOp	✓	71.02	64.07	**48.75**	**50.63**	**76.18**

Zhou et al. "Conditional prompt learning for vision-language models." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

Key messages

1. Conditional prompt learning is more generalizable
2. Conditional prompt learning is more transferable
3. Conditional prompt learning is more robust
4. Conditional prompt learning is very slow to train (batch_size=1)

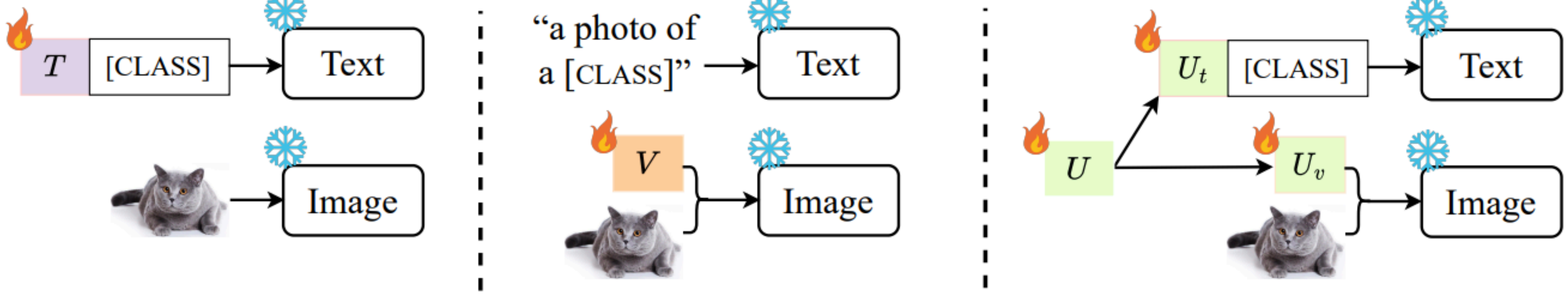
3D prompt tensor:
n_sentence x n_token x dim



Zhou et al. "Conditional prompt learning for vision-language models." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

Multimodal prompt learning

- Idea: simultaneously adjust text and image features
- Same performance but much faster training

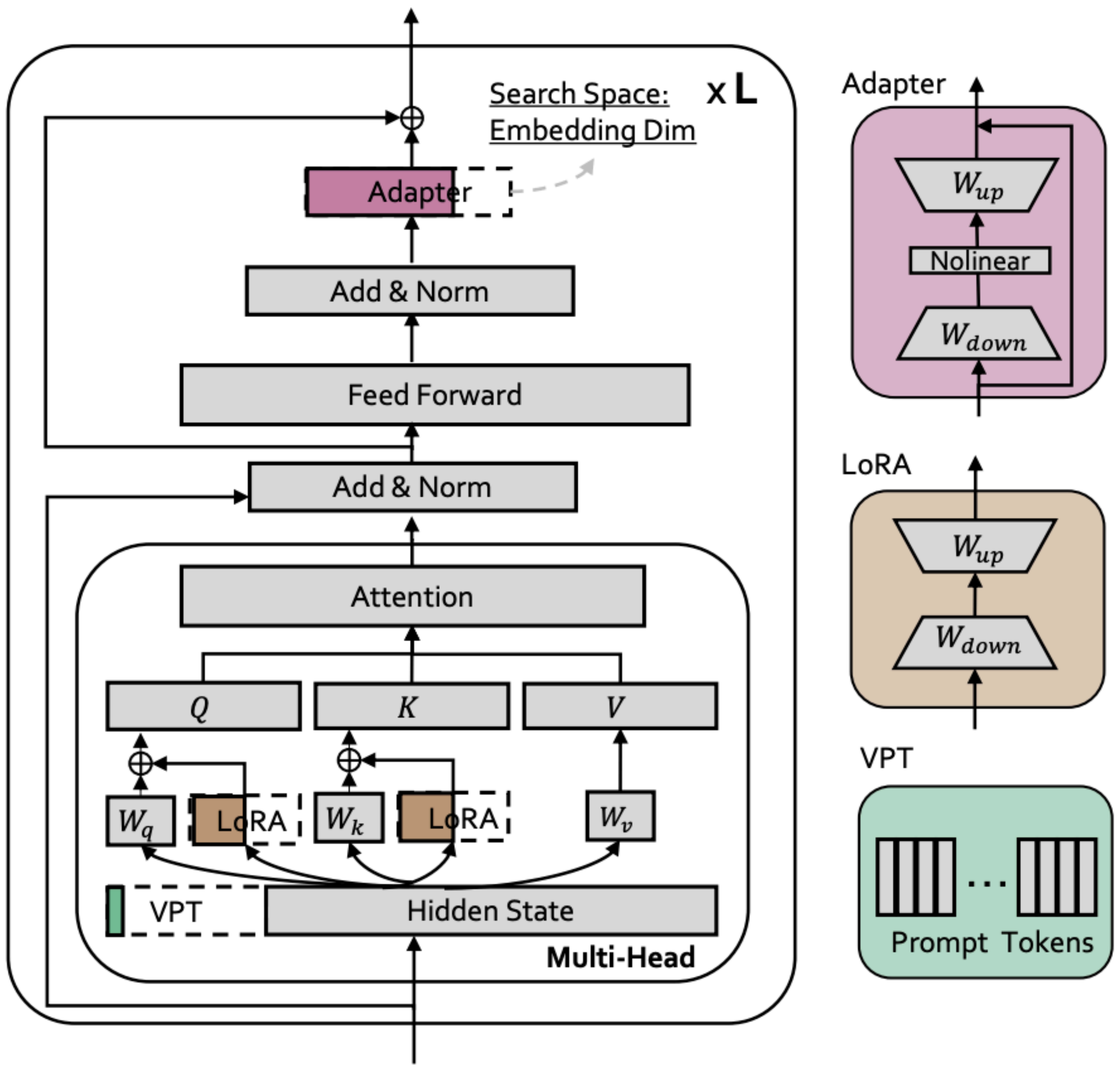


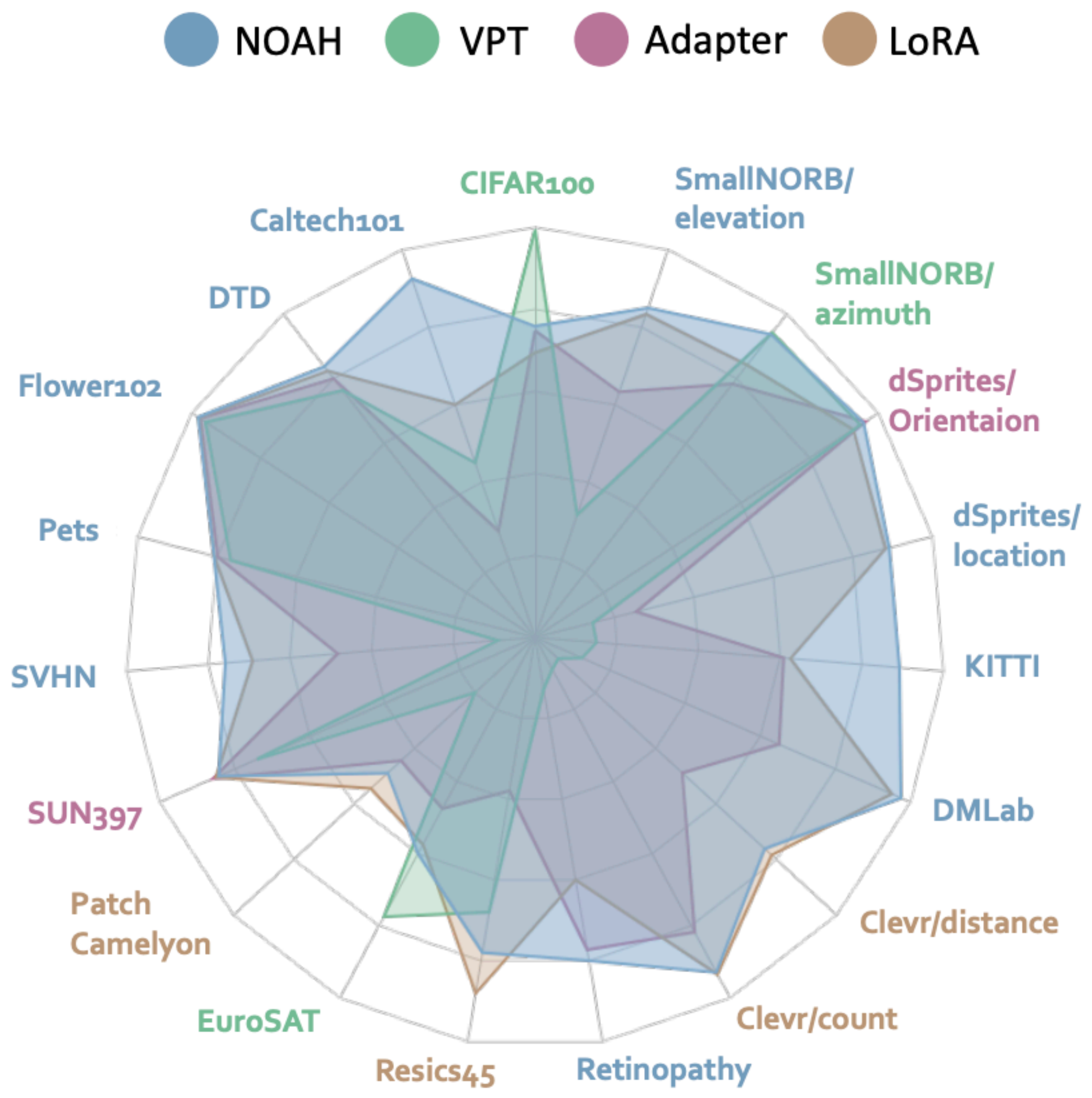
(a) Text Prompt - CoOp (b) Visual Prompt - VPT (c) Unified Prompt - Ours

#	Method	Source	Target				Average	*OOD* Average
		ImageNet	-V2	-S	-A	-R		
1	CoOp	71.51	64.20	47.99	49.71	75.21	61.72	59.28
2	CoCoOp	71.02	64.07	**48.75**	50.63	76.18	62.13	59.91
3	VPT-shallow	68.98	62.10	47.68	47.19	76.10	60.38	58.27
4	VPT-deep	70.57	63.67	47.66	43.85	74.42	60.04	57.40
5	UPT	**72.63**	**64.35**	48.66	**50.66**	**76.24**	**62.51**	**59.98**

Zang et al. "Unified vision and language prompt learning." *arXiv preprint arXiv:2210.07225* (2022).

Have more compute? Do prompt search





Zhang et al. "Neural prompt search." *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).

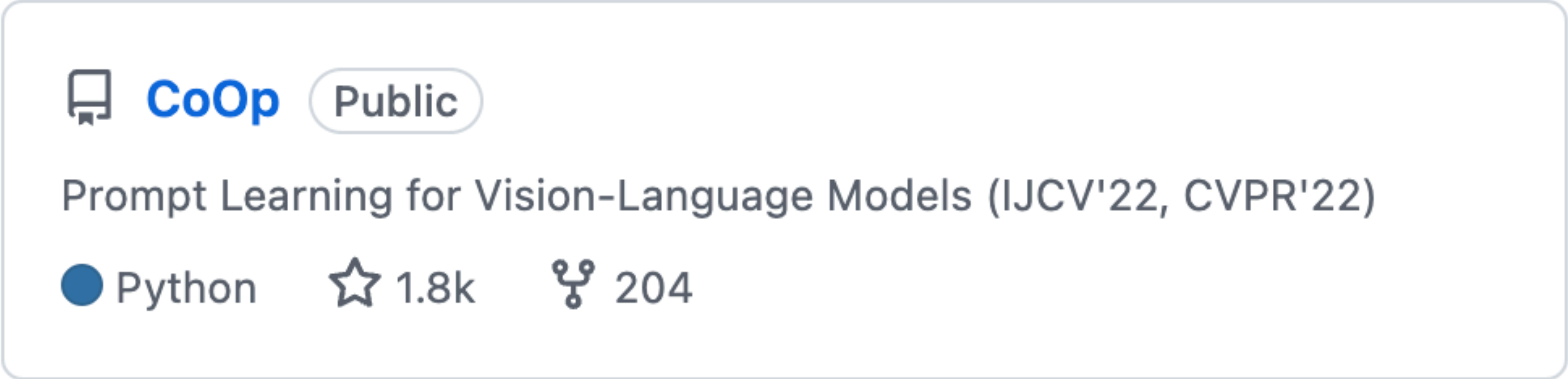
Take home messages

- VLMs largely reshaped the landscape of visual recognition
- Deploying VLMs in the real world is a non-trivial problem
- Prompt learning is a data-efficient adaptation method
- Conditional prompt learning works better but is too slow
- Multimodal prompt learning strikes a good balance between performance and speed
- Do NAS to search for the best adaptation modules if more compute is available

Relevant prompting papers

- Learning to Prompt for Vision-Language Models
- Conditional Prompt Learning for Vision-Language Models
- Unified Vision and Language Prompt Learning
- Neural Prompt Search

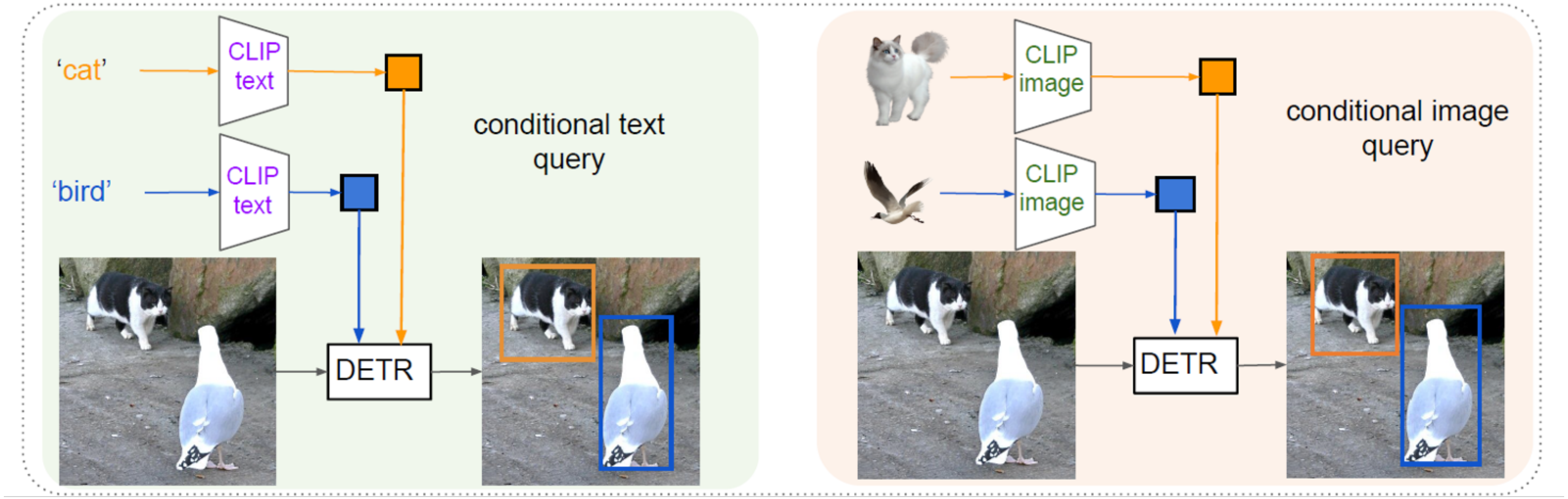
Open-source code: https://github.com/KaiyangZhou/CoOp

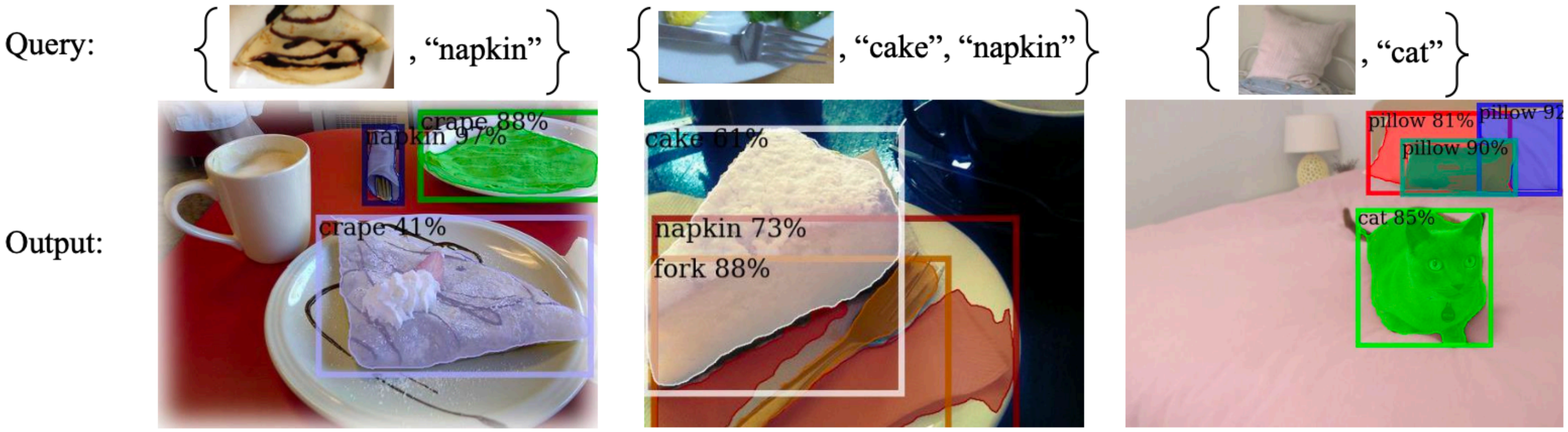


Outline

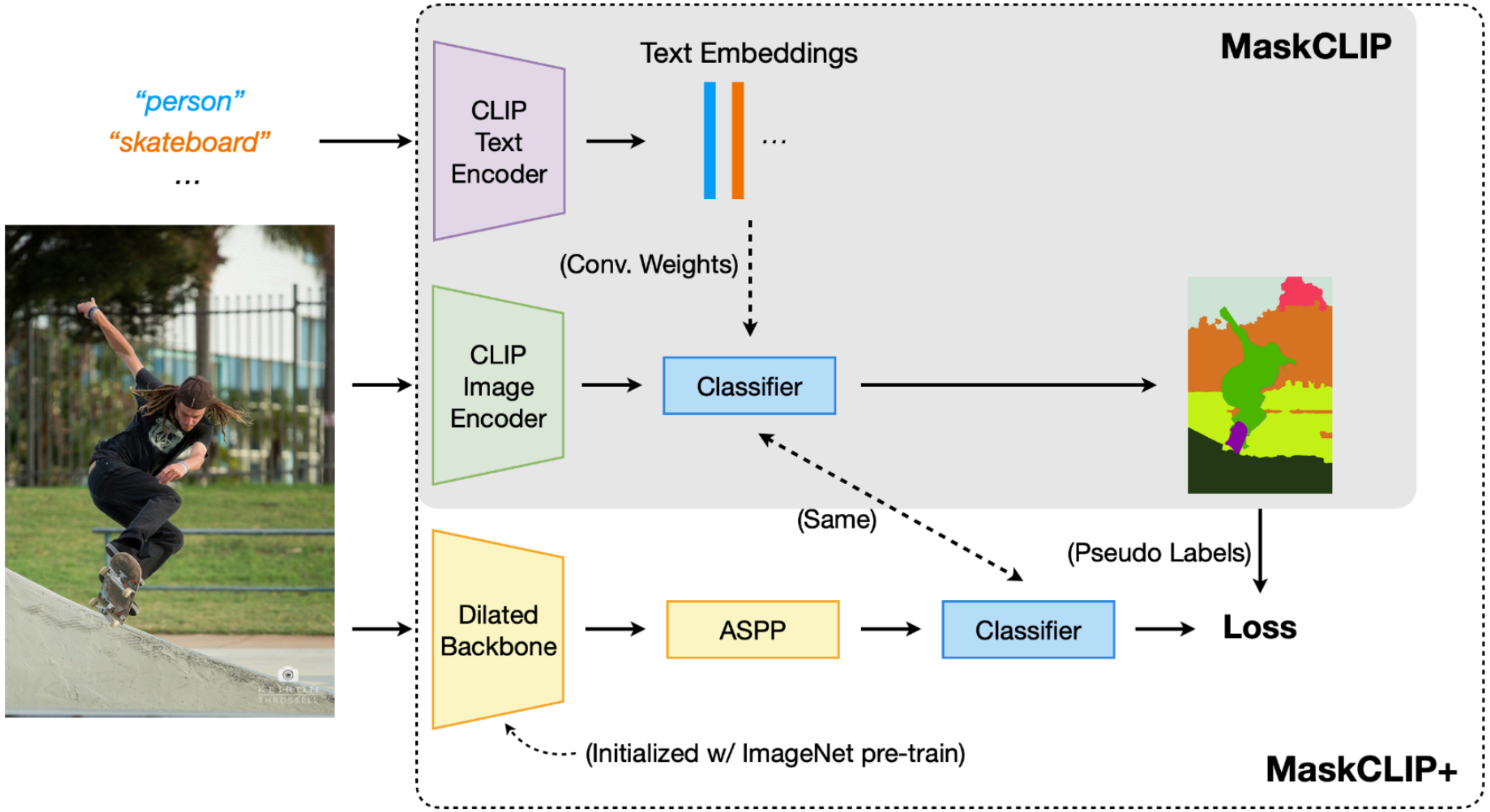
- History
- Pre-training
- Prompting
- Applications

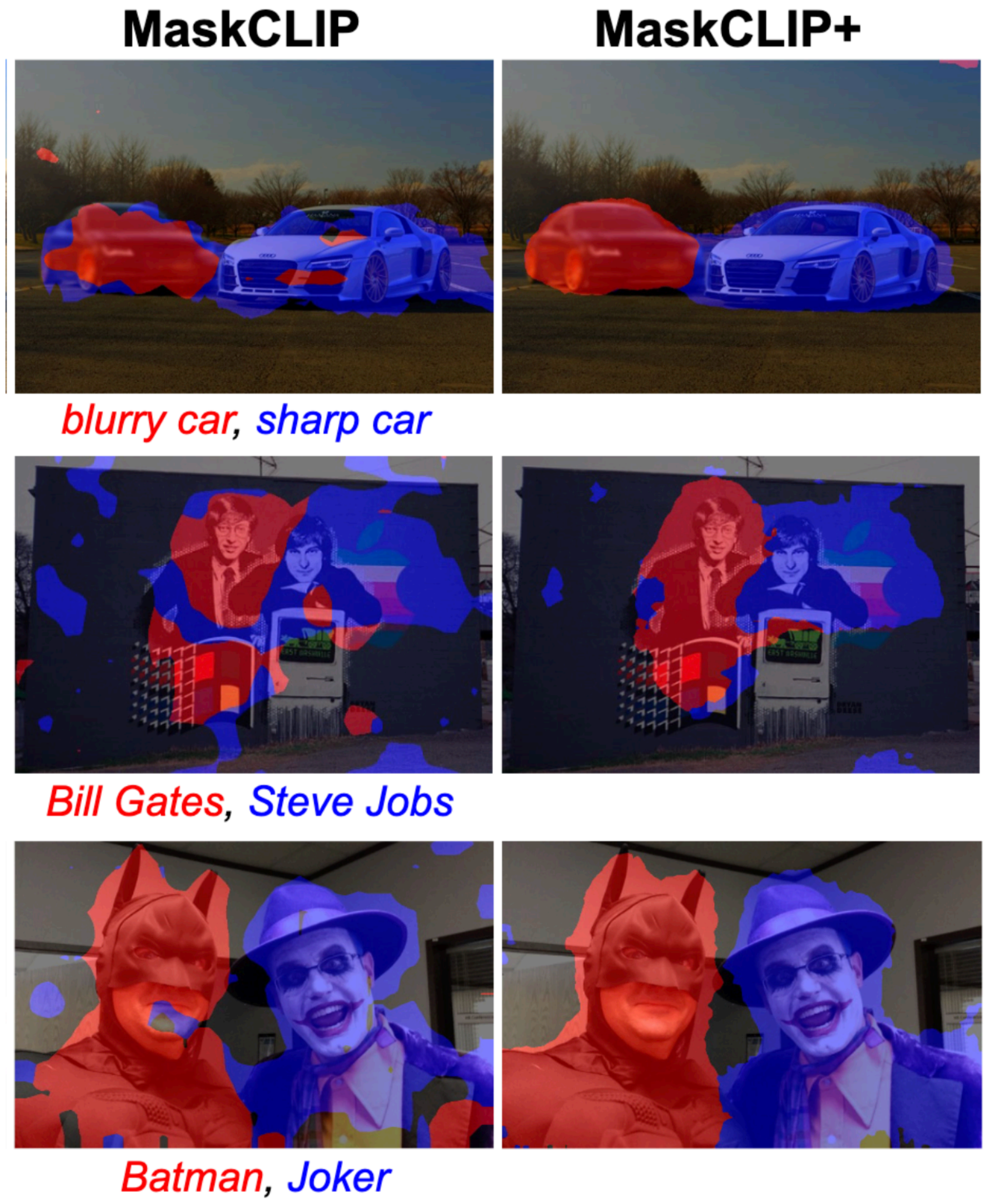
Open-Vocabulary Perception





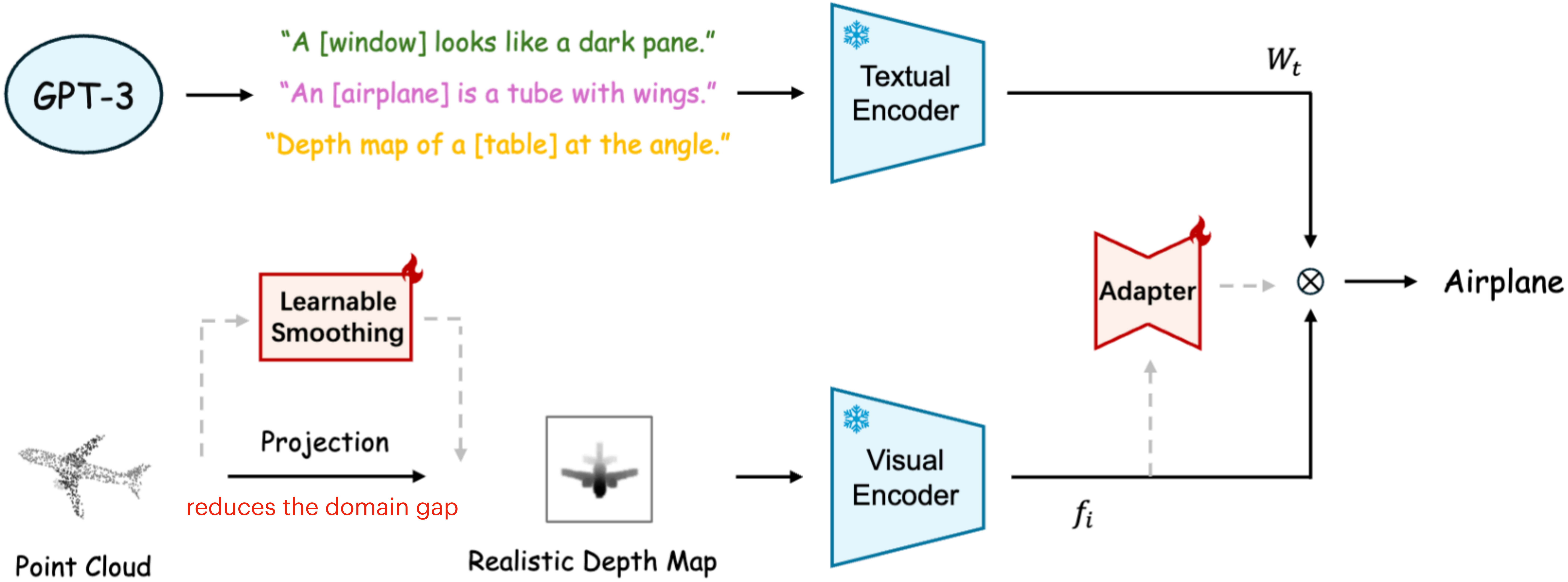
Zang et al. "Open-vocabulary detr with conditional matching." *European Conference on Computer Vision*. 2022.



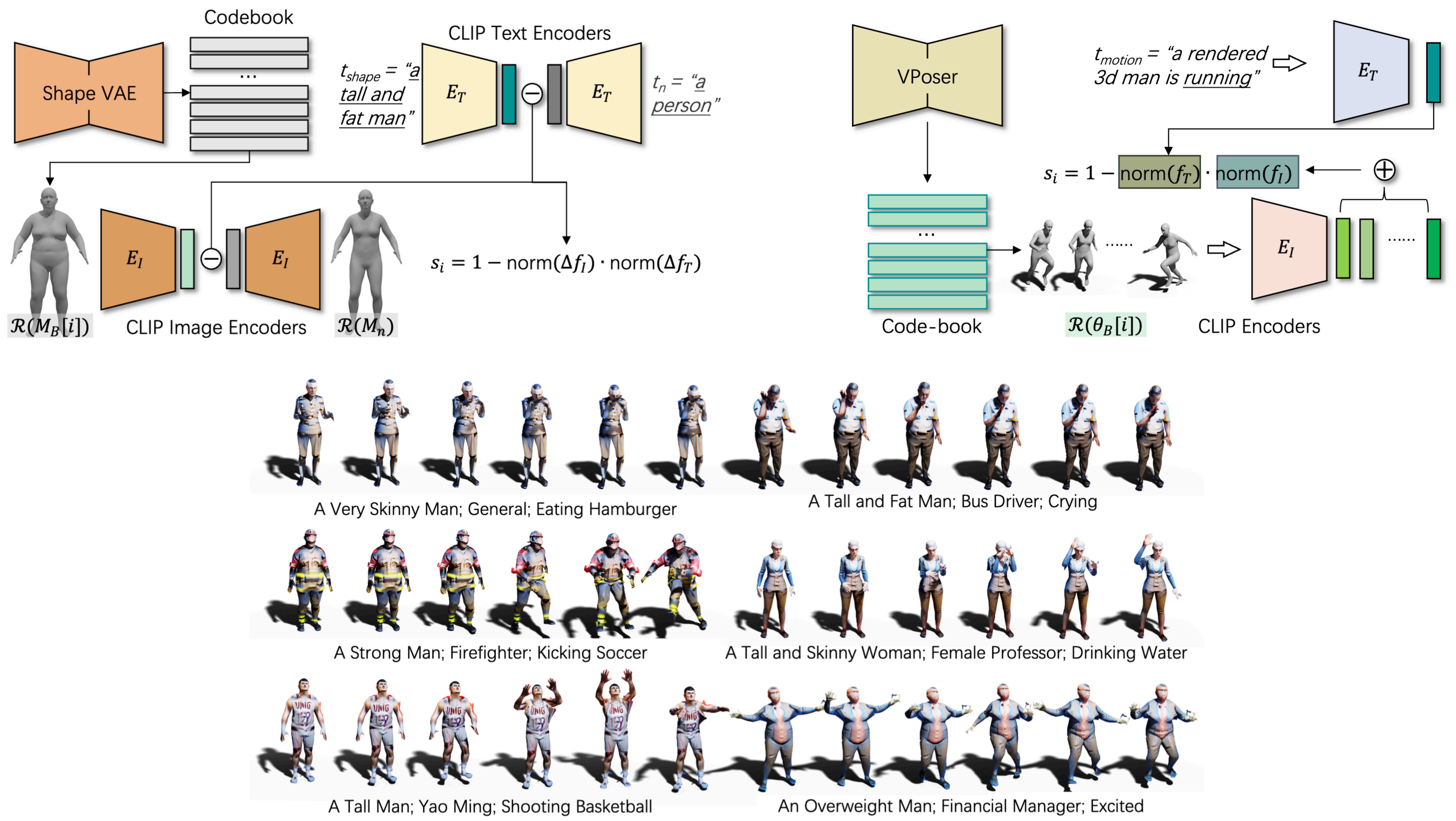


Zhou et al. "Extract free dense labels from clip." *European Conference on Computer Vision*. 2022.

3D Understanding and Generation

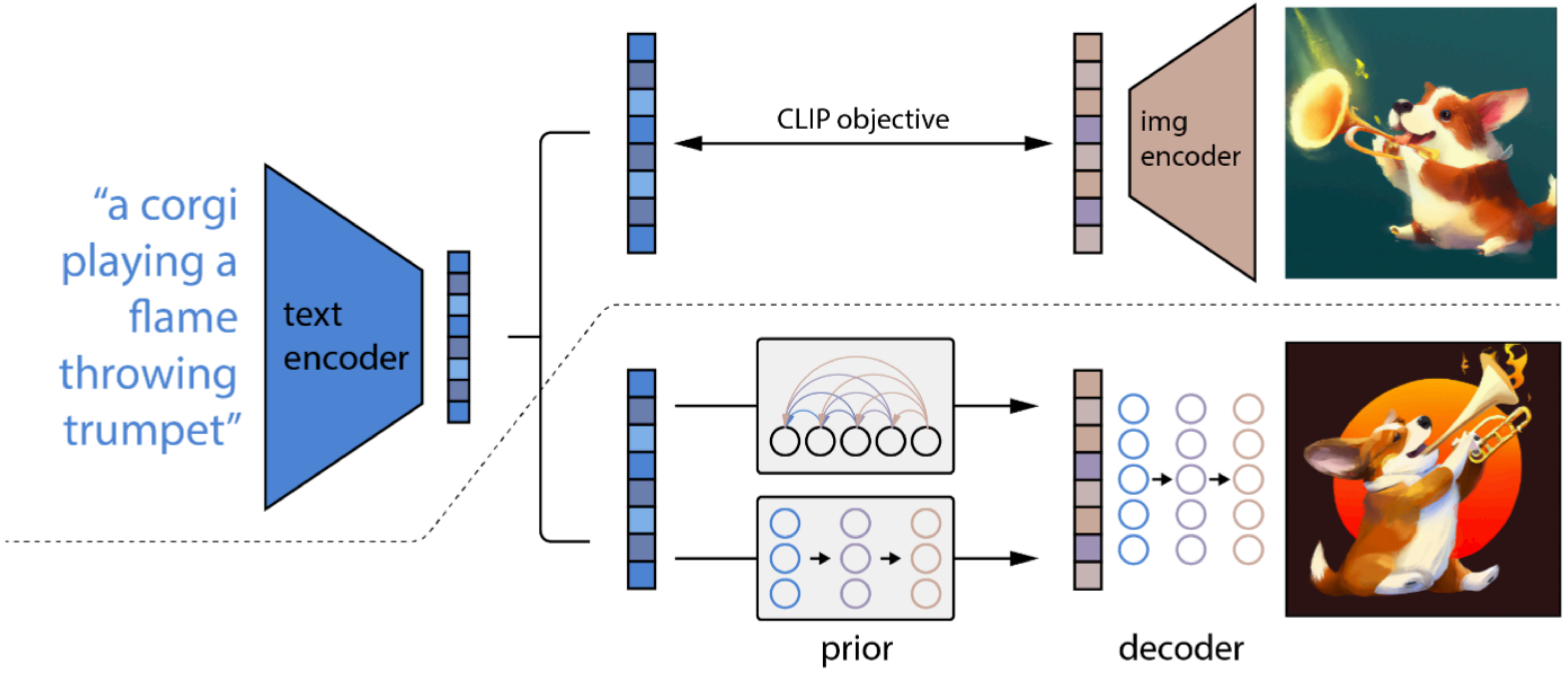


Zhu et al. "Pointclip v2: Prompting clip and gpt for powerful 3d open-world learning." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023.



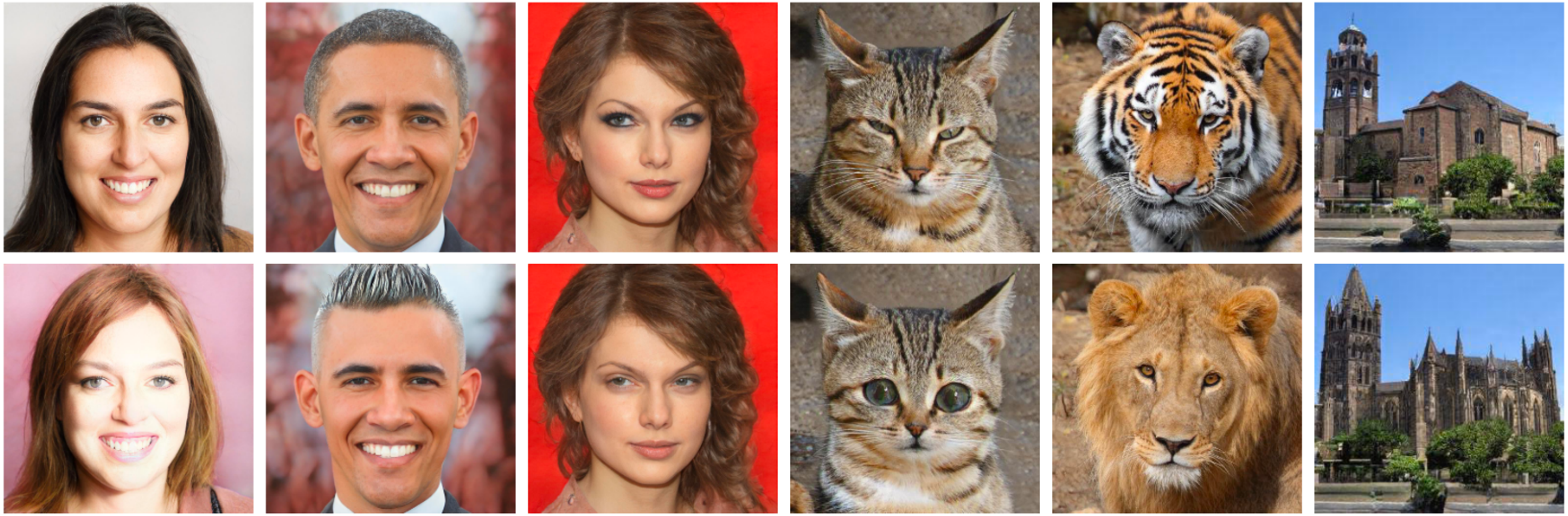
Hong et al. "AvatarCLIP: zero-shot text-driven generation and animation of 3D avatars." *ACM Transactions on Graphics (TOG)* 41.4 (2022): 1-19.

Generative Models and Creativity



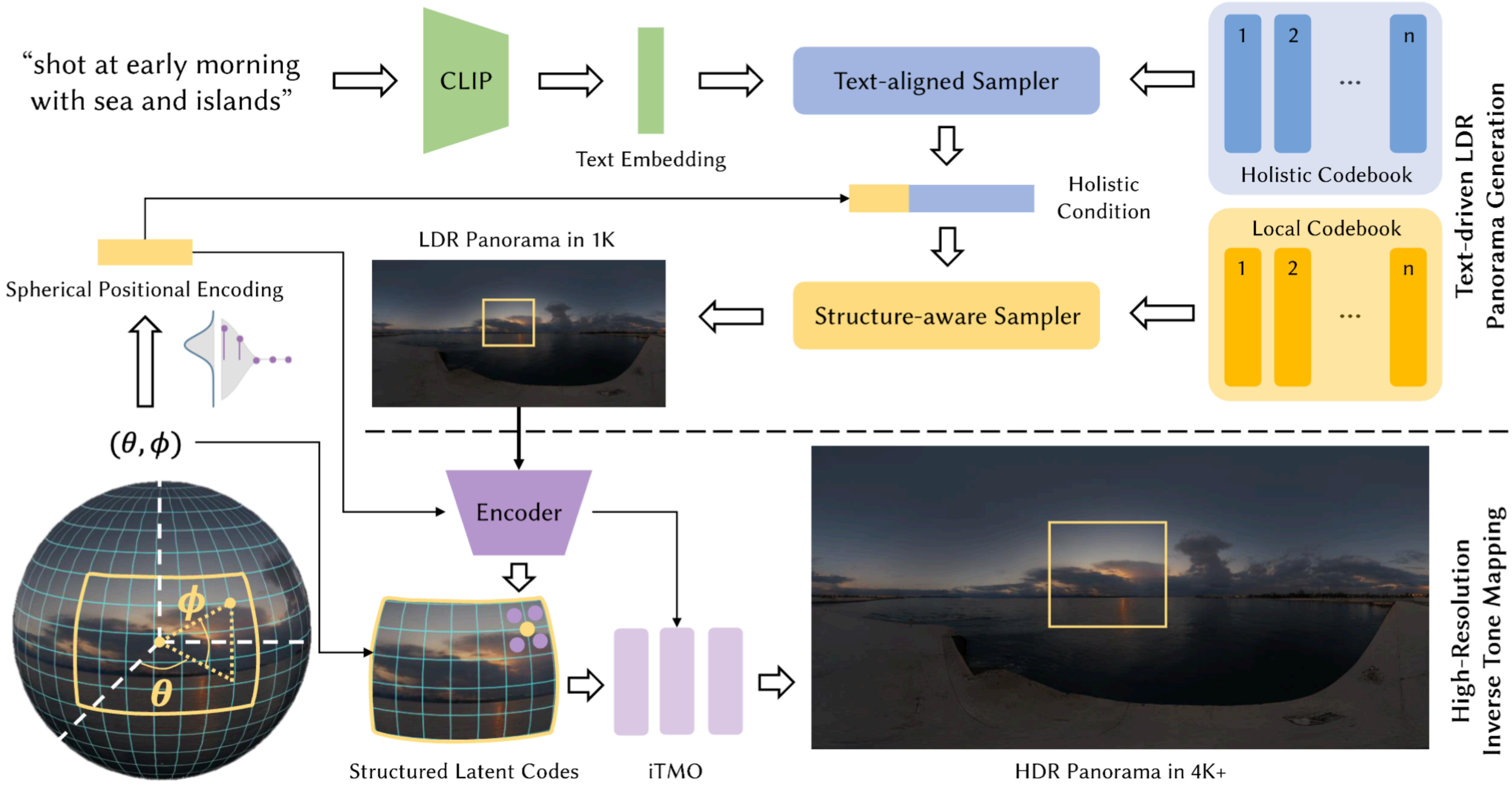
panda mad scientist mixing sparkling chemicals, artstation

Ramesh et al. Hierarchical Text-Conditional Image Generation with CLIP Latents.



“Emma Stone” “Mohawk hairstyle” “Without makeup” “Cute cat” “Lion” “Gothic church”

Patashnik et al. StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery. ICCV'21.



Chen et al. "Text2light: Zero-shot text-driven hdr panorama generation." *ACM Transactions on Graphics (TOG)* 41.6 (2022): 1-16.

Recap

- **History**:
  - evolution of vision and language models, convergence to VLMs
- **Pre-training**
  - contrastive learning, dual encoders, image-text pairs
- **Prompting**
  - prompt engineering, prompt learning
- **Applications**
  - open-vocabulary perception, 3D, GenAI