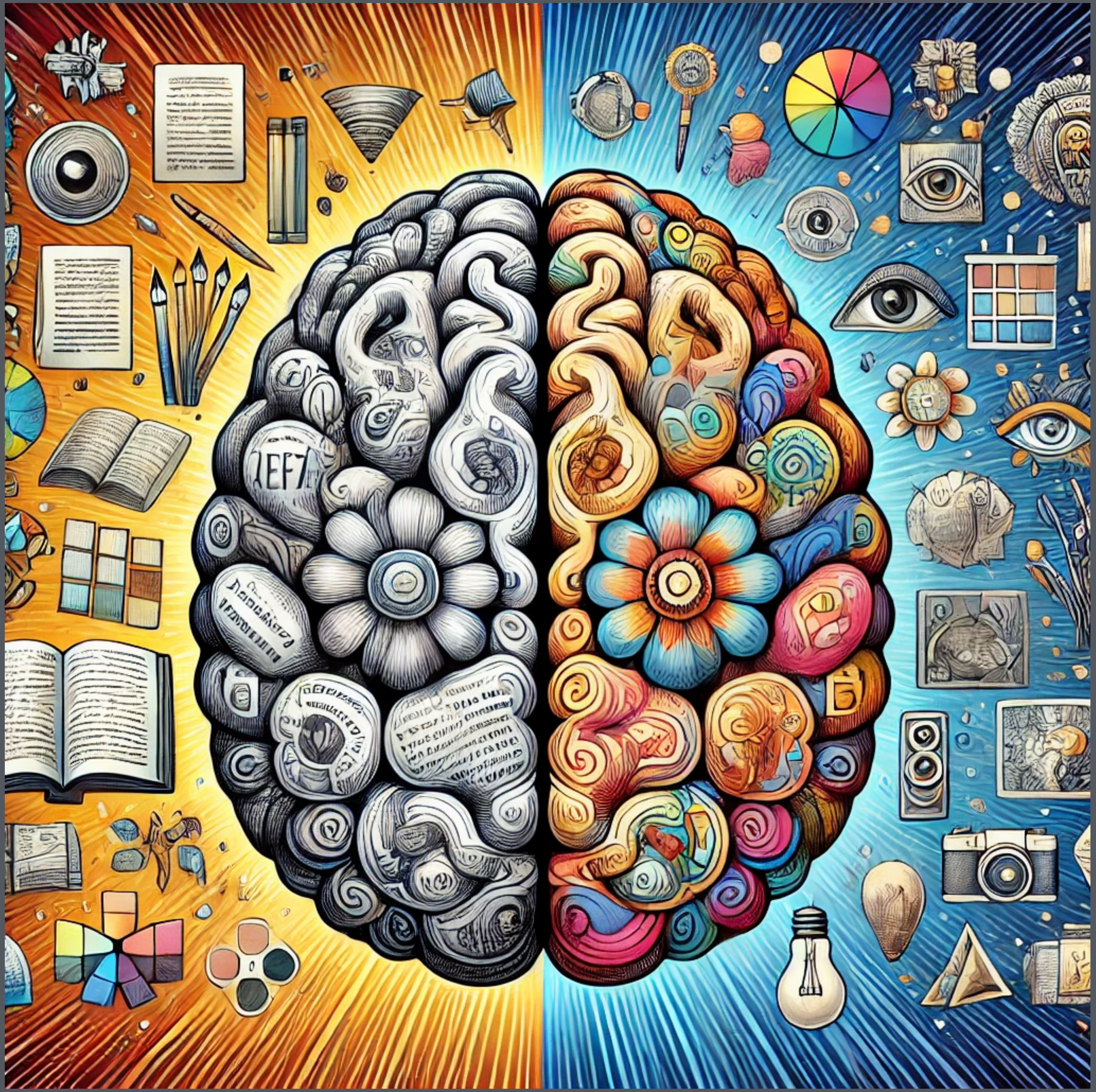
Self-supervised Learning and Vision-Language



NO COFFEE
NO PROBLEM

@ IAISS
TWITTER: @Y_M_ASANO

YUKI M. ASANO

Hi, I'm Yuki

- Currently: Full Prof at the University of Technology Nuremberg (UTN)
  - Self-supervised Learning
  - Multimodal Learning
  - Large Model Adaptation
  - Large Language Models
- Happy to collaborate on works in these topics
- I love running/hiking

focus of today

- More info: https://yukimasano.github.io/ yuki.asano@utn.de

Fundamental AI Lab YUKI ASANO UTN

Self-Supervised Representation Learning

- Novel SSL algorithms
- Better visual Foundation Models
- Synthetic data and generative models

Vision-language Learning

- Vision-Language Models
- Data-efficient training
- Fundamental understanding
- Bias, Privacy and Fairness

Fundamental AI Lab

Video and Temporal Learning

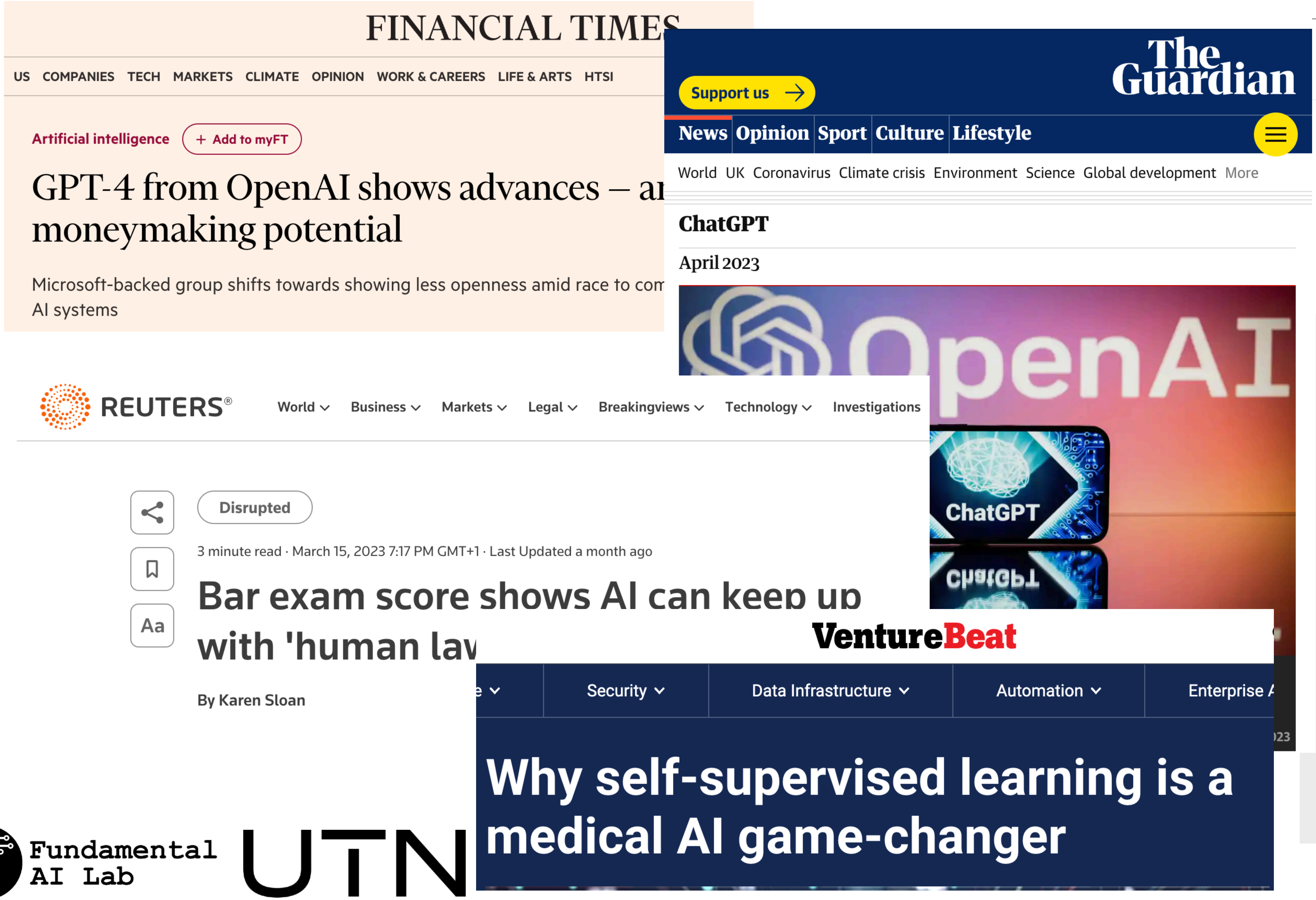
- Learning image models from video signals
- Cross-modal and multimodal learning frameworks
- Better video architectures and tasks

Large Language Models

- Instruction-Tuning
- Reasoning, planning
- Parameter-Efficient Finetuning
- Bias, Fairness

Nuremburg image via Unsplash

Self-supervised and vision-language learning is everywhere now



FINANCIAL TIMES

US COMPANIES TECH MARKETS CLIMATE OPINION WORK & CAREERS LIFE & ARTS HTSI

Artificial intelligence + Add to myFT

GPT-4 from OpenAI shows advances – a
moneymaking potential

Microsoft-backed group shifts towards showing less openness amid race to com
AI systems

The Guardian

Support us →

News Opinion Sport Culture Lifestyle

World UK Coronavirus Climate crisis Environment Science Global development More

ChatGPT

April 2023

REUTERS®

World ⌄ Business ⌄ Markets ⌄ Legal ⌄ Breakingviews ⌄ Technology ⌄ Investigations

Disrupted

3 minute read · March 15, 2023 7:17 PM GMT+1 · Last Updated a month ago

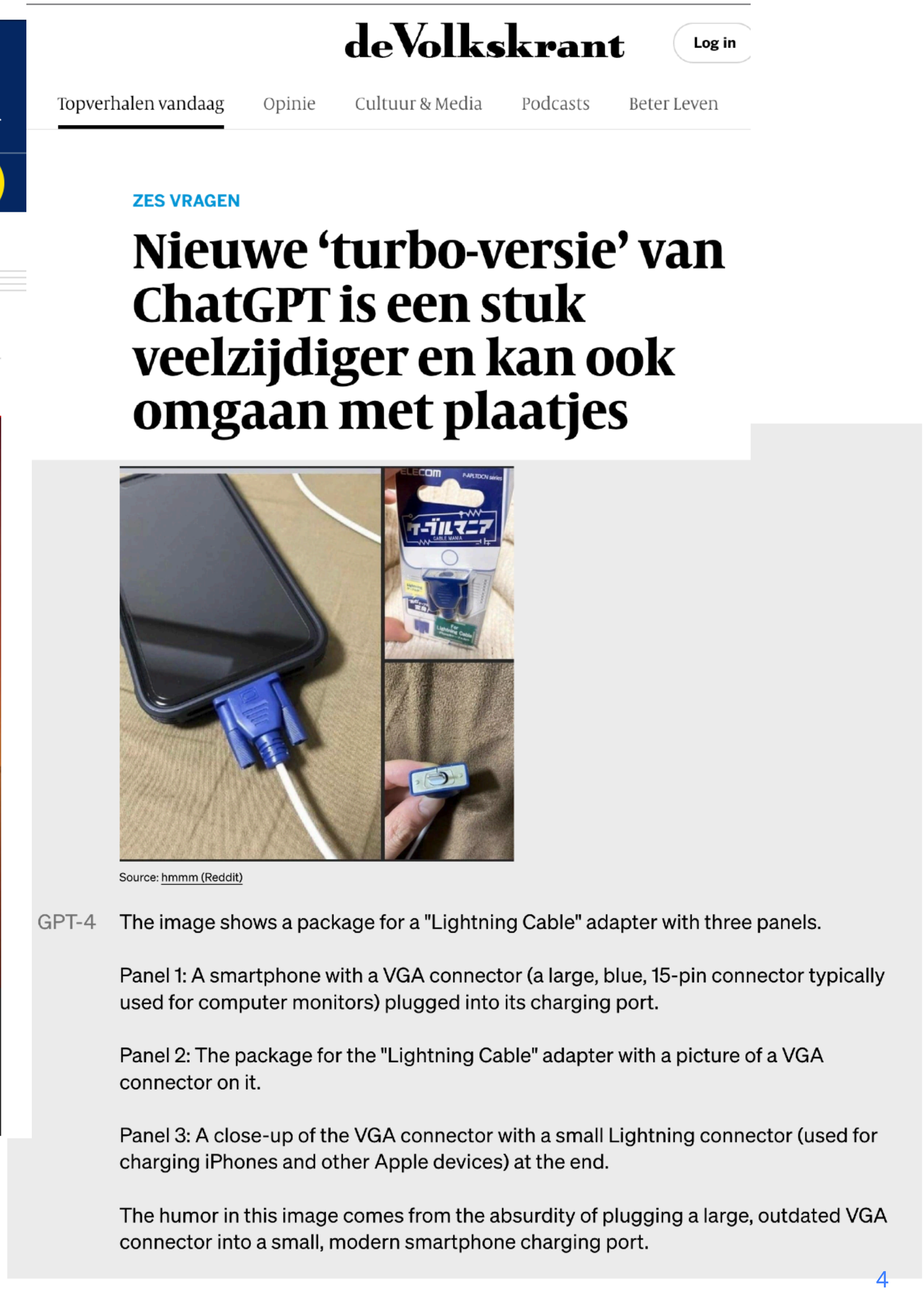
Bar exam score shows AI can keep up
with 'human lav

By Karen Sloan

VentureBeat

e ⌄ Security ⌄ Data Infrastructure ⌄ Automation ⌄ Enterprise A

Why self-supervised learning is a medical AI game-changer



deVolkskrant Log in

Topverhalen vandaag Opinie Cultuur & Media Podcasts Beter Leven

ZES VRAGEN

Nieuwe 'turbo-versie' van ChatGPT is een stuk veelzijdiger en kan ook omgaan met plaatjes

Source: hmmm (Reddit)

GPT-4 The image shows a package for a "Lightning Cable" adapter with three panels.

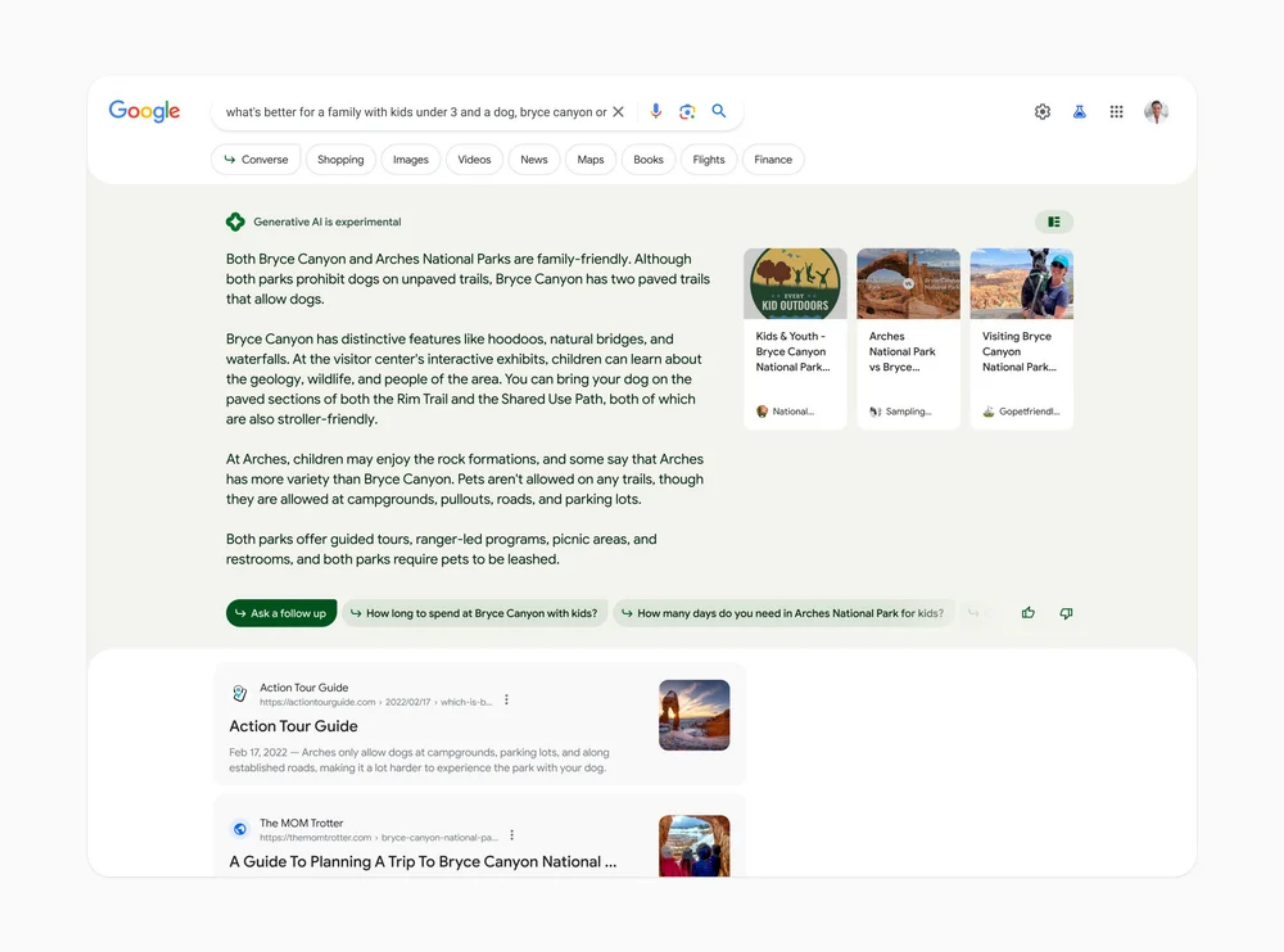
Panel 1: A smartphone with a VGA connector (a large, blue, 15-pin connector typically used for computer monitors) plugged into its charging port.

Panel 2: The package for the "Lightning Cable" adapter with a picture of a VGA connector on it.

Panel 3: A close-up of the VGA connector with a small Lightning connector (used for charging iPhones and other Apple devices) at the end.

The humor in this image comes from the absurdity of plugging a large, outdated VGA connector into a small, modern smartphone charging port.

How your grandma might even know vision-language deep learning:


Google
Converse
Shopping
Images
Videos
News
Maps
Books
Flights
Finance
KID OUTDOORS
Ask a follow up
How long to spend at Bryce Canyon with kids?
How many days do you need in Arches National Park for kids?

what's better for a family with kids under 3 and a dog, bryce canyon or

Generative AI is experimental

Both Bryce Canyon and Arches National Parks are family-friendly. Although both parks prohibit dogs on unpaved trails, Bryce Canyon has two paved trails that allow dogs.

Bryce Canyon has distinctive features like hoodoos, natural bridges, and waterfalls. At the visitor center's interactive exhibits, children can learn about the geology, wildlife, and people of the area. You can bring your dog on the paved sections of both the Rim Trail and the Shared Use Path, both of which are also stroller-friendly.

At Arches, children may enjoy the rock formations, and some say that Arches has more variety than Bryce Canyon. Pets aren't allowed on any trails, though they are allowed at campgrounds, pullouts, roads, and parking lots.

Both parks offer guided tours, ranger-led programs, picnic areas, and restrooms, and both parks require pets to be leashed.

Kids & Youth - Bryce Canyon National Park...
National...

Arches National Park vs Bryce...
Sampling...

Visiting Bryce Canyon National Park...
Gopetfriendl...

Action Tour Guide
https://actiontourguide.com › 2022/02/17 › which-is-b...

Action Tour Guide

Feb 17, 2022 — Arches only allow dogs at campgrounds, parking lots, and along established roads, making it a lot harder to experience the park with your dog.

The MOM Trotter
https://themomtrotter.com › bryce-canyon-national-pa...

A Guide To Planning A Trip To Bryce Canyon National ...

Google's Gemini; OpenAI's Sora

Philosophy

There's a lot going on.

We will not cover everything.

But we will cover the core foundational works and principles and recent works that represent the diversity of research in this field.

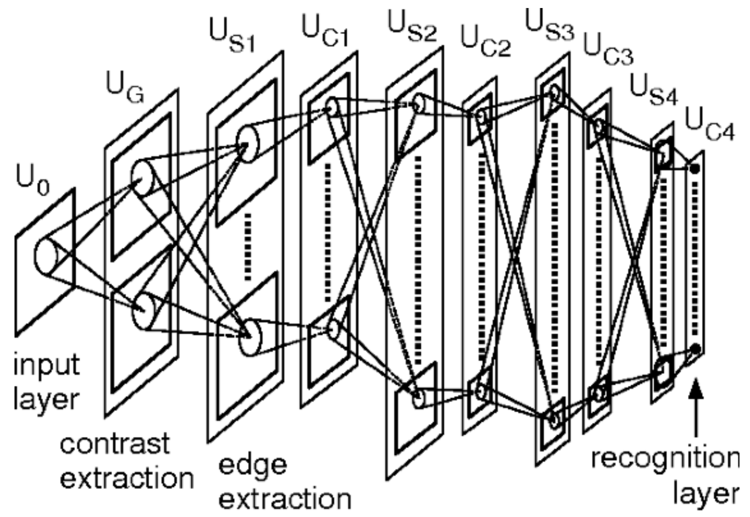
I've achieved my goal if after this lecture you think:
vision-language learning is exciting and impactful *and the lecture + tutorial gave me ideas on* ***how to get started*** *working in this field*

Representation Learning

The field of AI has made rapid progress, the crucial fuel is data

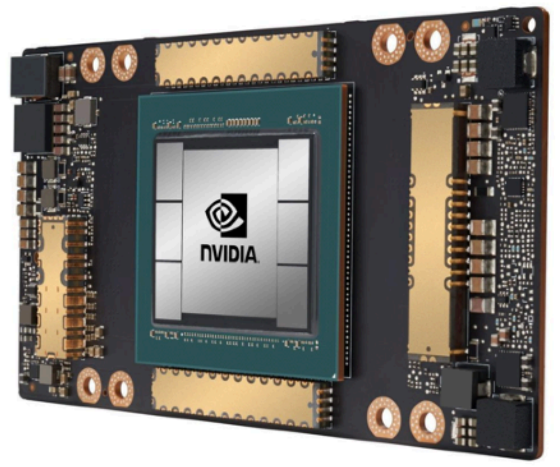
Algorithms

Deep neural networks



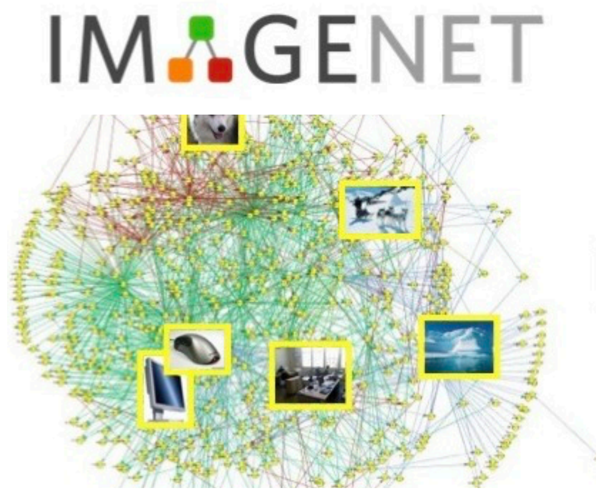
Hardware

GPUs



Data

Large scale datasets

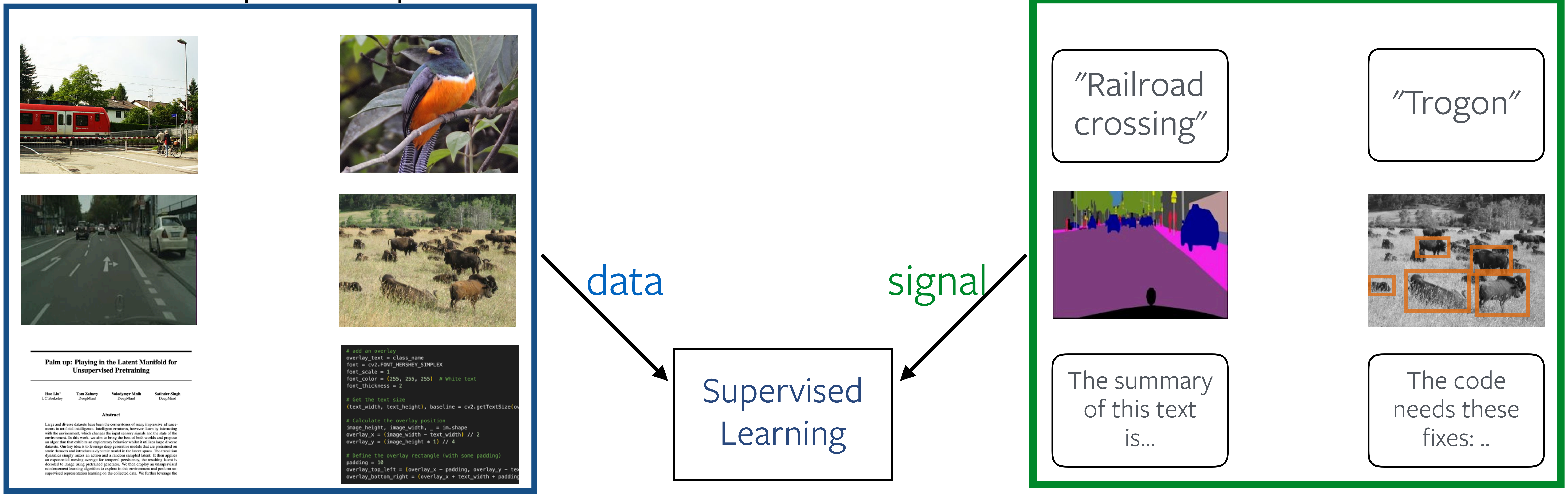


Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Fukushima, K., Biol. Cybernetics 1980
Object Recognition with Gradient-Based Learning. LeCun et al. Shape, Contour and Grouping in Computer Vision 1999
ImageNet: A Large-Scale Hierarchical Image Database. Deng, et al. CVPR, 2009.
ImageNet Classification with Deep Convolutional Neural Networks., Krizhevsky et al., NeurIPS 2012

However, manual data annotations for supervised learning is limiting.

Data is cheap & ubiquitous

But annotations are expensive and often require experts



Palm up: Playing in the Latent Manifold for Unsupervised Pretraining

Hao Liu* UC Berkeley
Tom Zahavy DeepMind
Volodymyr Mnih DeepMind
Satinder Singh DeepMind

Abstract

Large and diverse datasets have been the cornerstones of many impressive advancements in artificial intelligence. Intelligent creatures, however, learn by interacting with the environment, which changes the input sensory signals and the state of the environment. In this work, we aim to bring the best of both worlds and propose an algorithm that exhibits an exploratory behavior whilst it utilizes large diverse datasets. Our key idea is to leverage deep generative models that are pretrained on static datasets and introduce a dynamic model in the latent space. The transition dynamics simply mixes an action and a random sampled latent. It then applies an exponential moving average for temporal persistency, the resulting latent is decoded to image using pretrained generator. We then employ an unsupervised reinforcement learning algorithm to explore in this environment and perform unsupervised representation learning on the collected data. We further leverage the

```
# add an overlay
overlay_text = class_name
font = cv2.FONT_HERSHEY_SIMPLEX
font_scale = 1
font_color = (255, 255, 255)  # White text
font_thickness = 2

# Get the text size
(text_width, text_height), baseline = cv2.getTextSize(ov

# Calculate the overlay position
image_height, image_width, _ = im.shape
overlay_x = (image_width - text_width) // 2
overlay_y = (image_height * 1) // 4

# Define the overlay rectangle (with some padding)
padding = 10
overlay_top_left = (overlay_x - padding, overlay_y - tex
overlay_bottom_right = (overlay_x + text_width + padding
```

data

signal

Supervised Learning

”Railroad crossing”

”Trogon”

The summary of this text is…

The code needs these fixes: ..

ImageNet: A Large-Scale Hierarchical Image Database. Dong et al. CVPR 2009
The Cityscapes Dataset for Semantic Urban Scene Understanding. Cordts et al. CVPR 2016
Scene parsing through ADE20K dataset. Zhou et al. CVPR 2017.

Self-supervised learning solves the problem of annotations.

Palm up: Playing in the Latent Manifold for Unsupervised Pretraining

Hao Liu* UC Berkeley
Tom Zahavy DeepMind
Volodymyr Mnih DeepMind
Satinder Singh DeepMind

Abstract

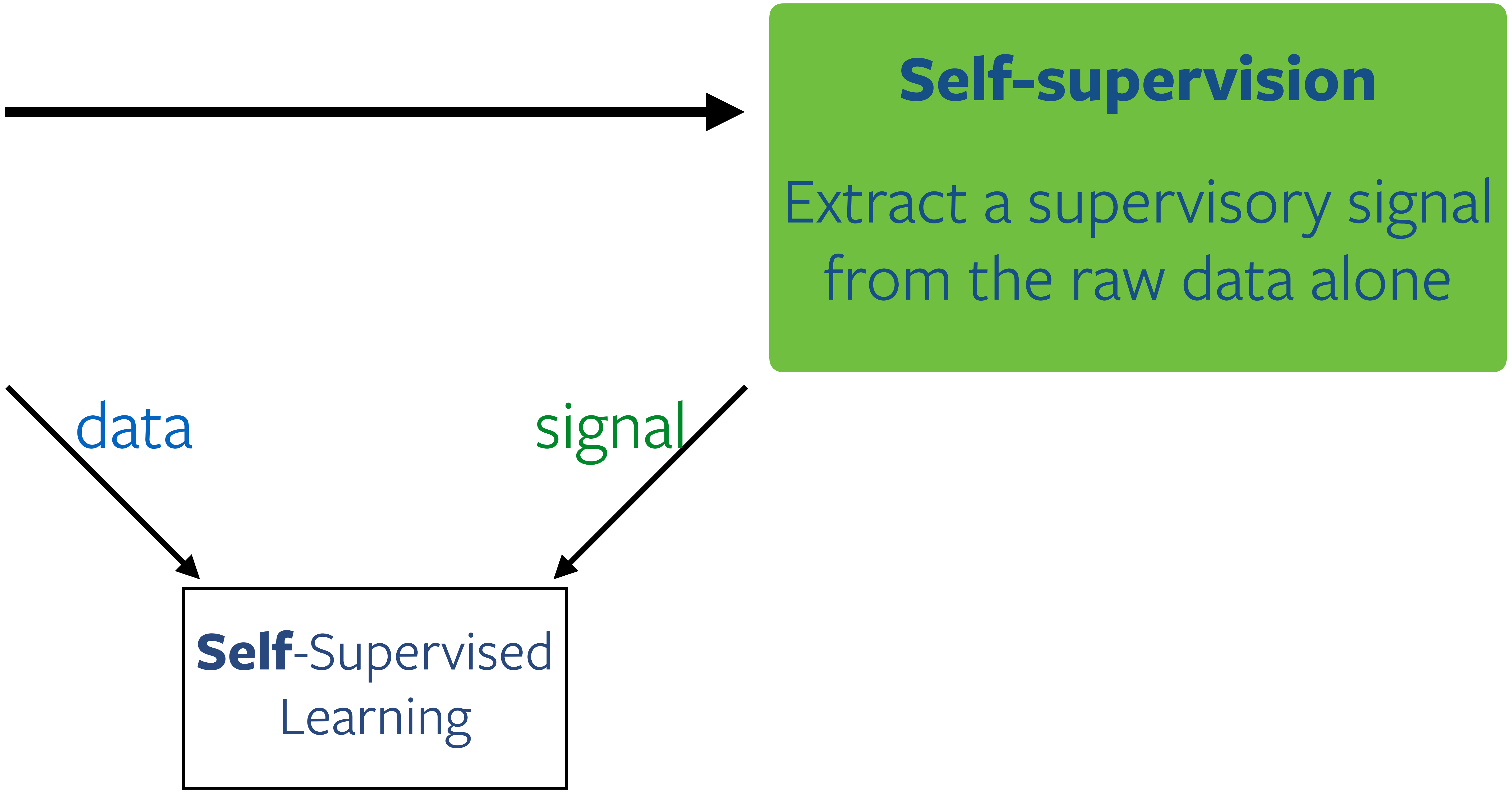
Large and diverse datasets have been the cornerstones of many impressive advancements in artificial intelligence. Intelligent creatures, however, learn by interacting with the environment, which changes the input sensory signals and the state of the environment. In this work, we aim to bring the best of both worlds and propose an algorithm that exhibits an exploratory behavior whilst it utilizes large diverse datasets. Our key idea is to leverage deep generative models that are pretrained on static datasets and introduce a dynamic model in the latent space. The transition dynamics simply mixes an action and a random sampled latent. It then applies an exponential moving average for temporal persistency, the resulting latent is decoded to image using pretrained generator. We then employ an unsupervised reinforcement learning algorithm to explore in this environment and perform unsupervised representation learning on the collected data. We further leverage the

```
# add an overlay
overlay_text = class_name
font = cv2.FONT_HERSHEY_SIMPLEX
font_scale = 1
font_color = (255, 255, 255)  # White text
font_thickness = 2

# Get the text size
(text_width, text_height), baseline = cv2.getTextSize(ov

# Calculate the overlay position
image_height, image_width, _ = im.shape
overlay_x = (image_width - text_width) // 2
overlay_y = (image_height * 1) // 4

# Define the overlay rectangle (with some padding)
padding = 10
overlay_top_left = (overlay_x - padding, overlay_y - tex
overlay_bottom_right = (overlay_x + text_width + padding
```



Self-supervision

Extract a supervisory signal from the raw data alone

data

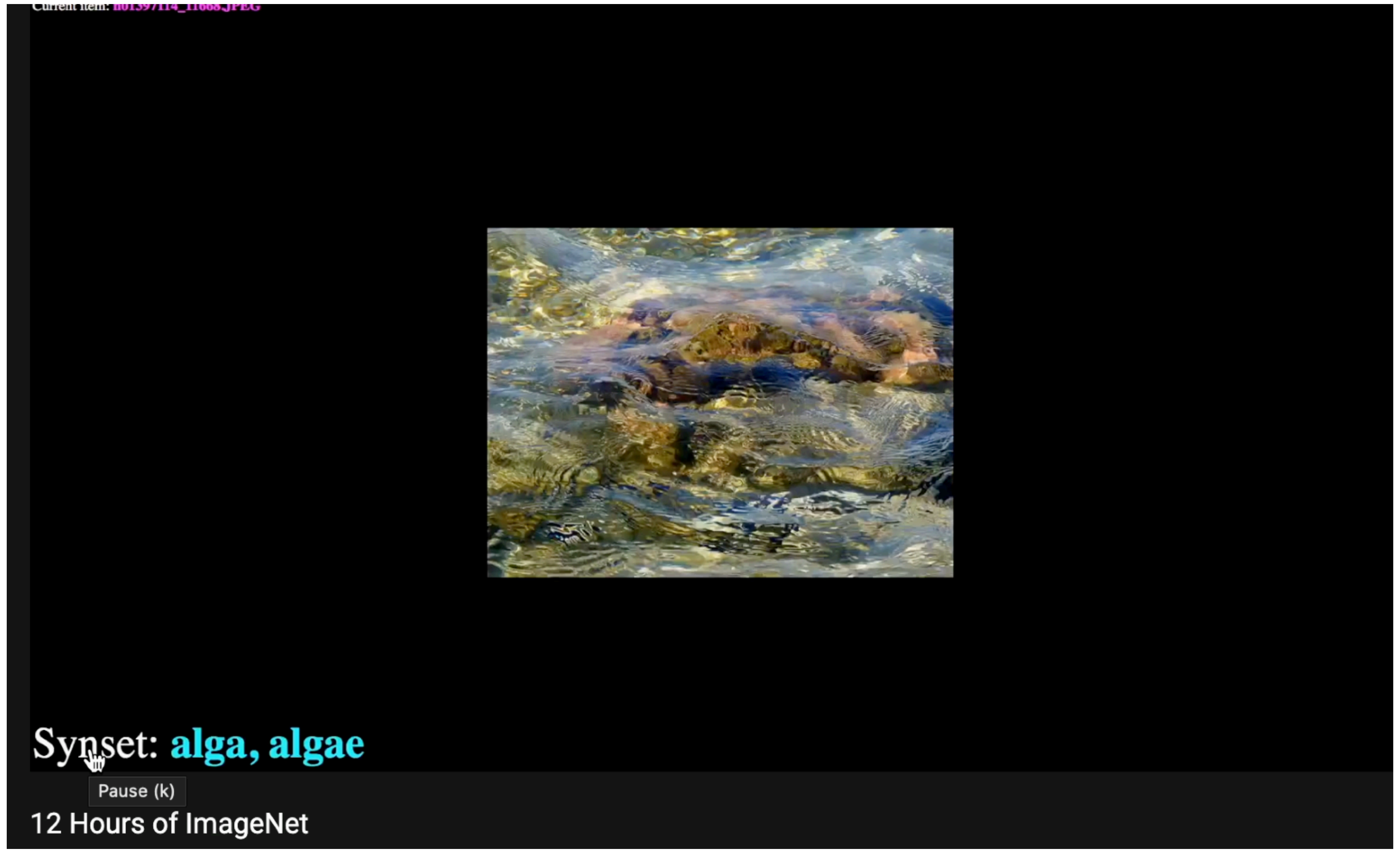
signal

Self-Supervised Learning

self-supervised learning: why?

Reason 1: Scalability

(above) x 50 =1.2M images

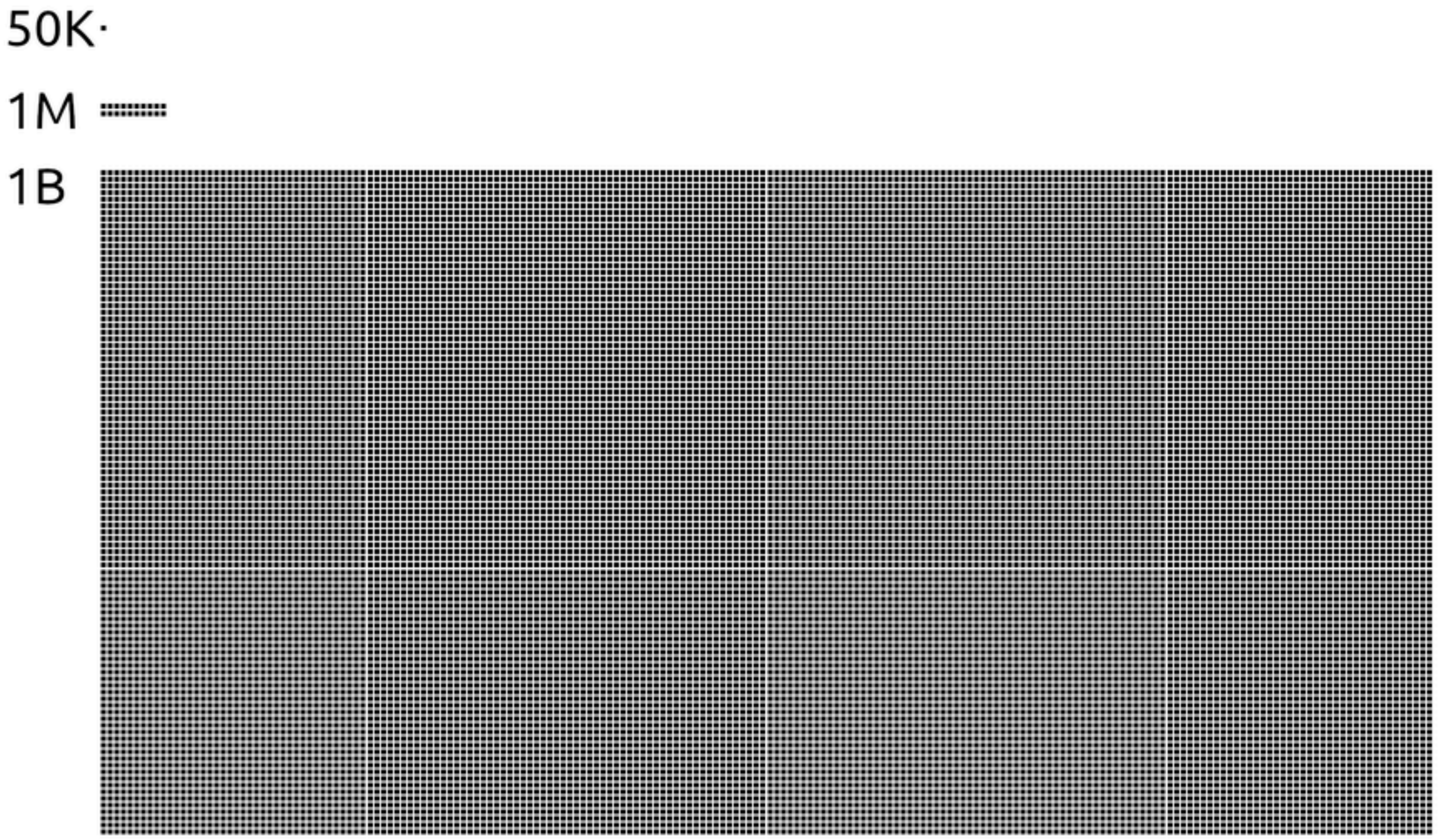


90ms * 1.2M = 30h

ImageNet: A Large-Scale Hierarchical Image Database. Deng, et al. CVPR, 2009.
12 hours of ImageNet: https://www.youtube.com/watch?v=PC60JL-lMzA

Reason 1: Scalability

Instagram: >50B images

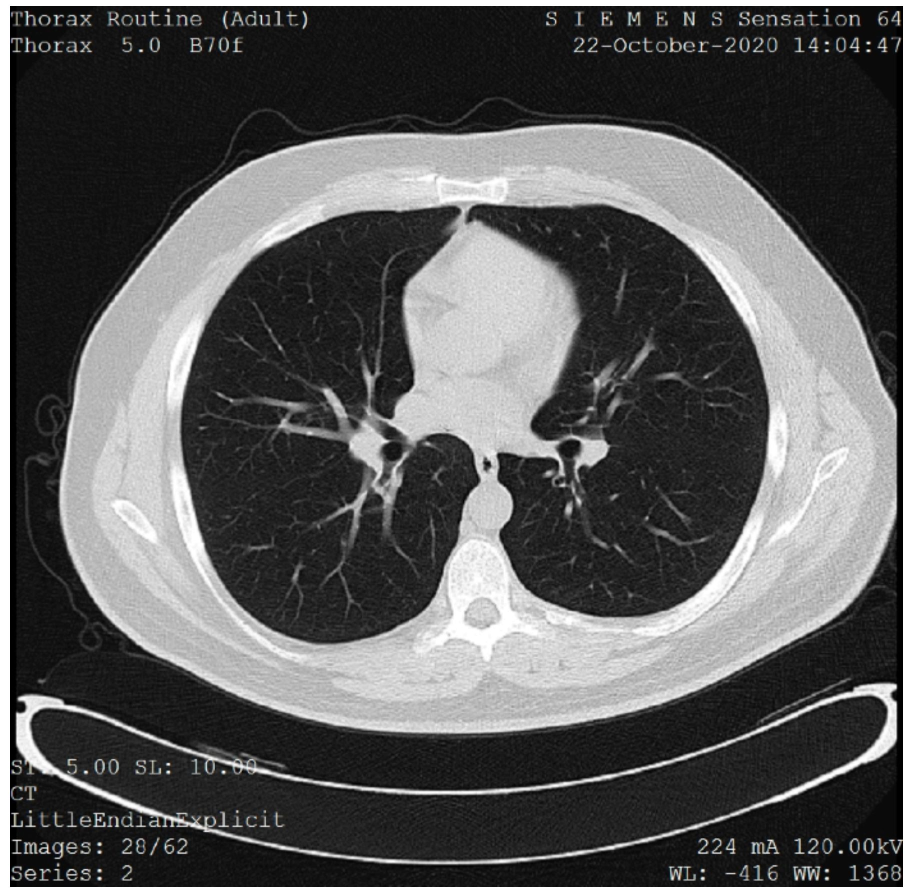


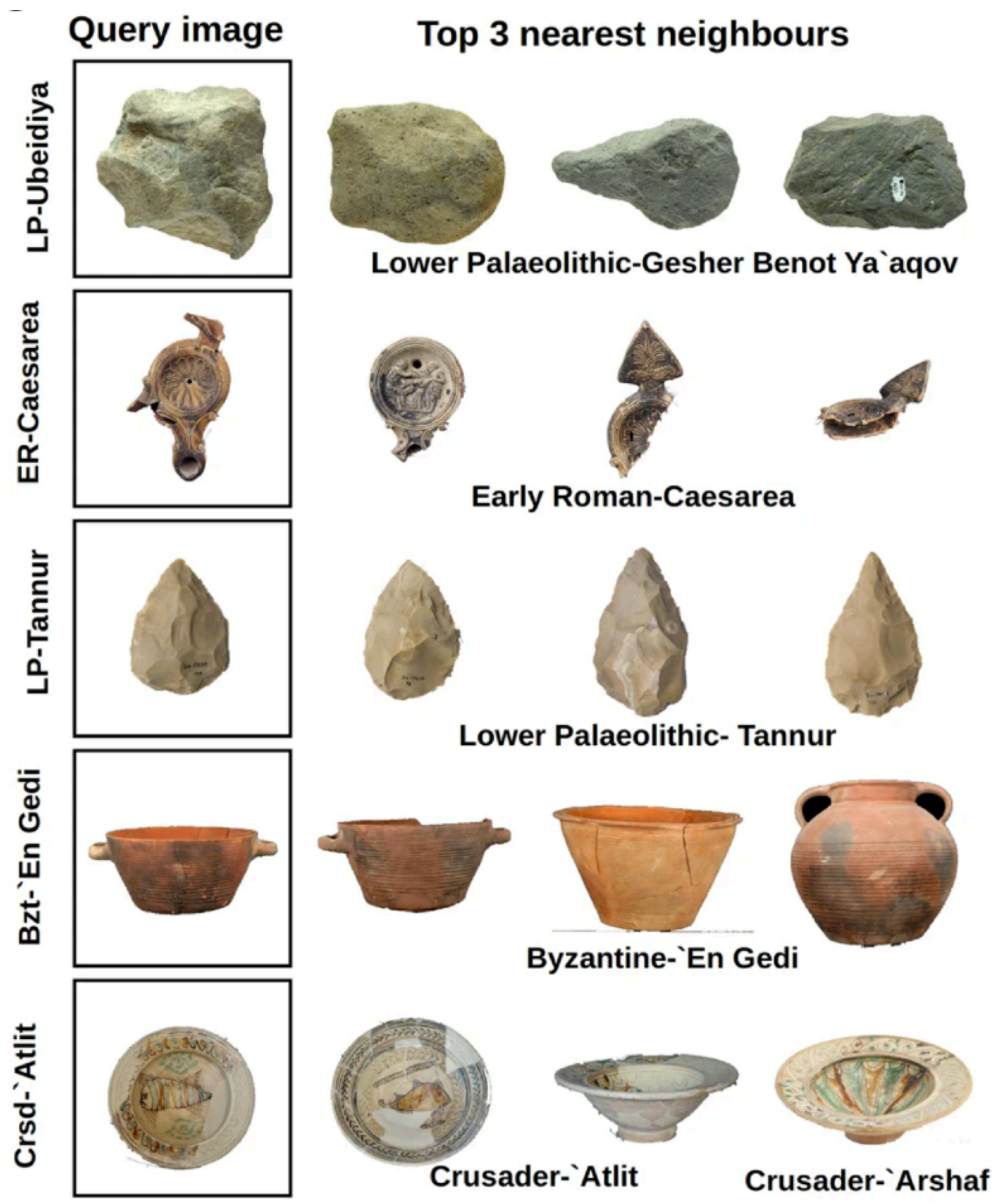
Annotation is expensive, yet datasets keep getting bigger.

Reason 2: Constantly changing domains

Unclear when & what to relabel. Again, large costs just to "keep up".

Reason 2: Accessibility & generalisability



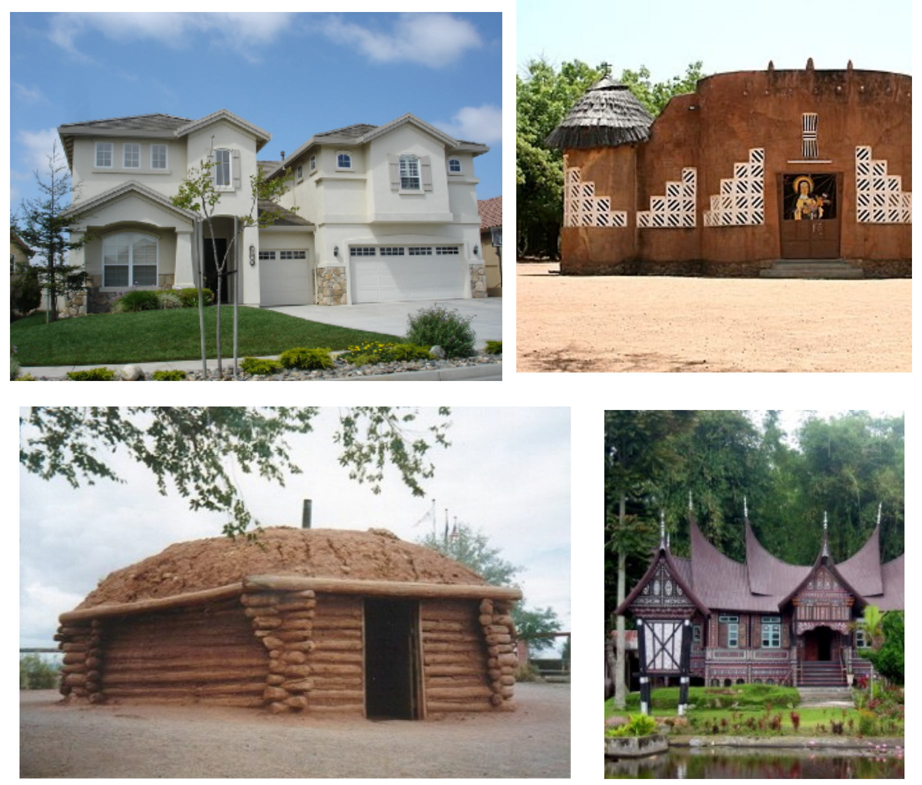


Pretrained models are very useful for a variety of tasks.

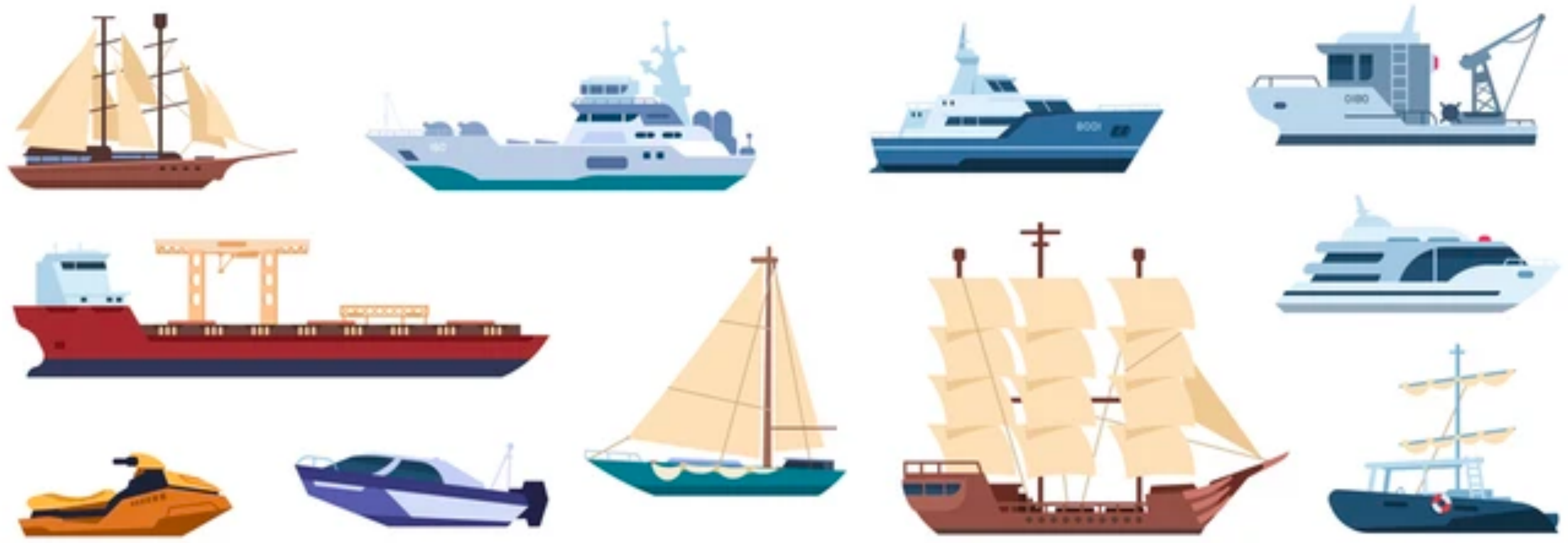
Fundamental AI Lab

https://www.kaggle.com/c/herbarium-2019-fgvc6, https://en.wikipedia.org/wiki/Medical_imaging#/media/File:CT_Scan_General_Illustration.jpg
Schaefer et al. Deep convolutional neural networks as strong gravitational lens detectors. Astronomy & Astrophysics.
Resler et al. A deep-learning model for predictive archaeology and archaeological community detection. Nature Humanities & Social Sciences Communications.

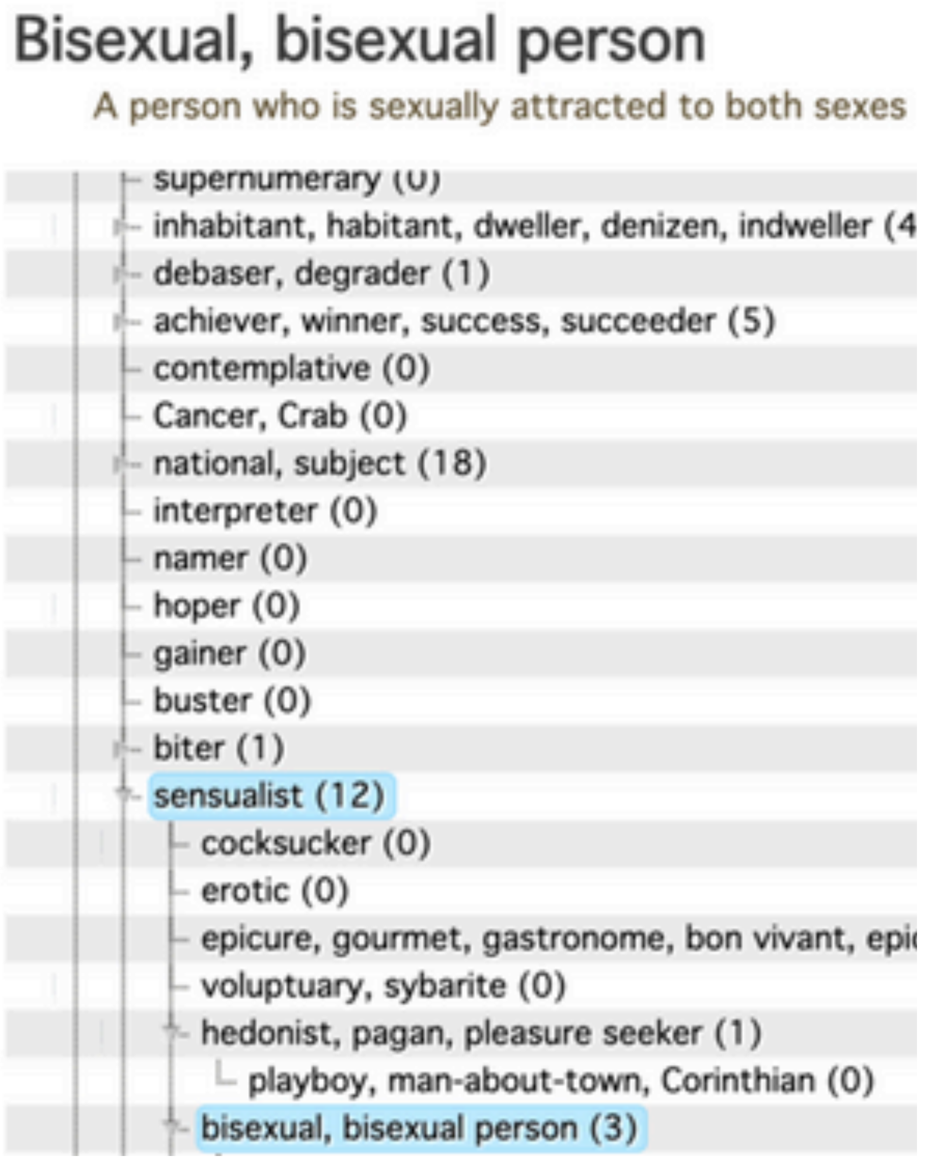
Reason 3: Ambiguity of labels



"A house"?



"A boat"?



Bisexual, bisexual person

A person who is sexually attracted to both sexes

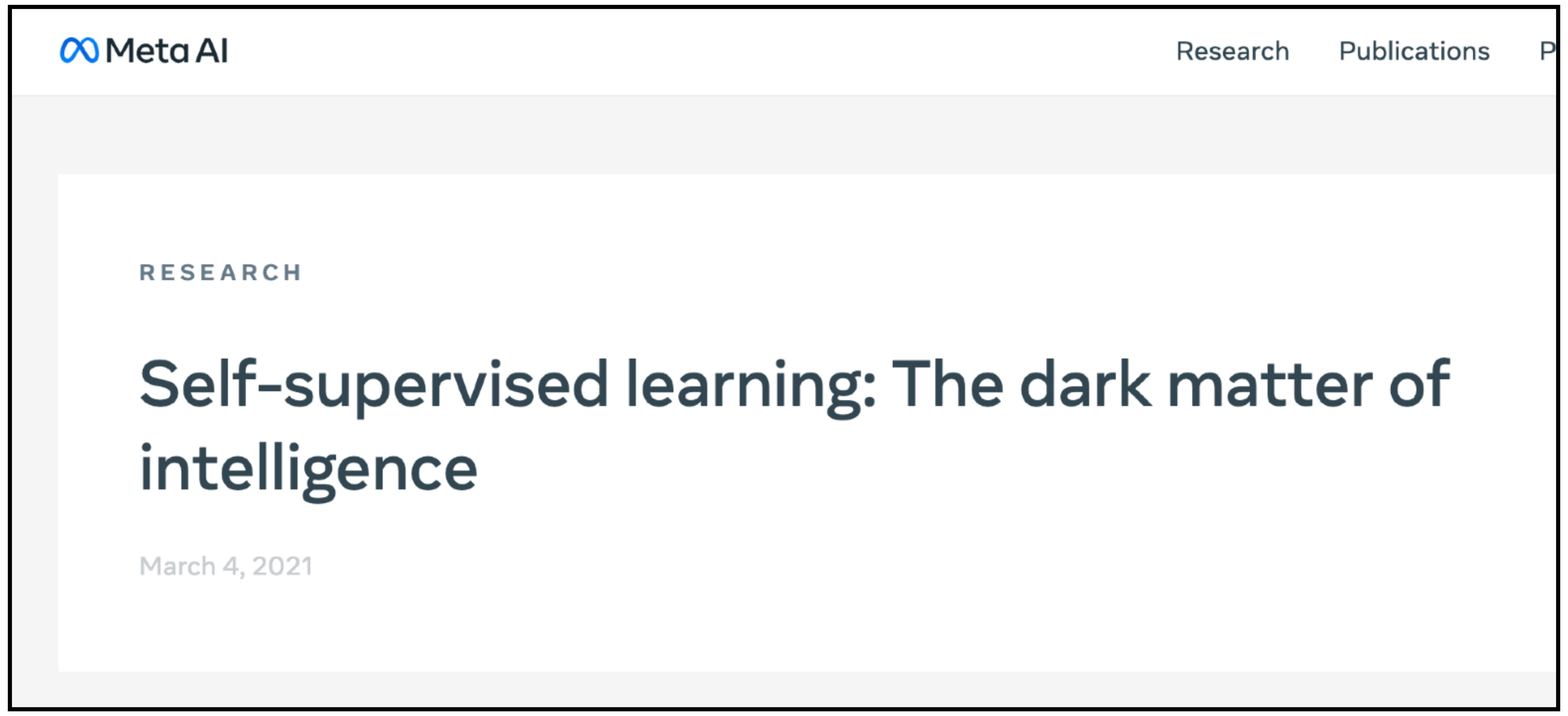
- supernumerary (0)
- inhabitant, habitant, dweller, denizen, indweller (4
- debaser, degrader (1)
- achiever, winner, success, succeeder (5)
- contemplative (0)
- Cancer, Crab (0)
- national, subject (18)
- interpreter (0)
- namer (0)
- hoper (0)
- gainer (0)
- buster (0)
- biter (1)
- sensualist (12)
  - cocksucker (0)
  - erotic (0)
  - epicure, gourmet, gastronome, bon vivant, epi
  - voluptuary, sybarite (0)
  - hedonist, pagan, pleasure seeker (1)
    - playboy, man-about-town, Corinthian (0)
  - bisexual, bisexual person (3)

Nonsensical visual labels

Labels are ambiguous at best, discriminating and bias-propagating at worst.
Do we really wish to provide our models with these priors?

https://en.wikipedia.org/wiki/List_of_house_styles
https://www.shutterstock.com/image-illustration/flat-ships-sailing-yachts-marine-sailboats-1903407259
https://excavating.ai/ Crawford & Paglen

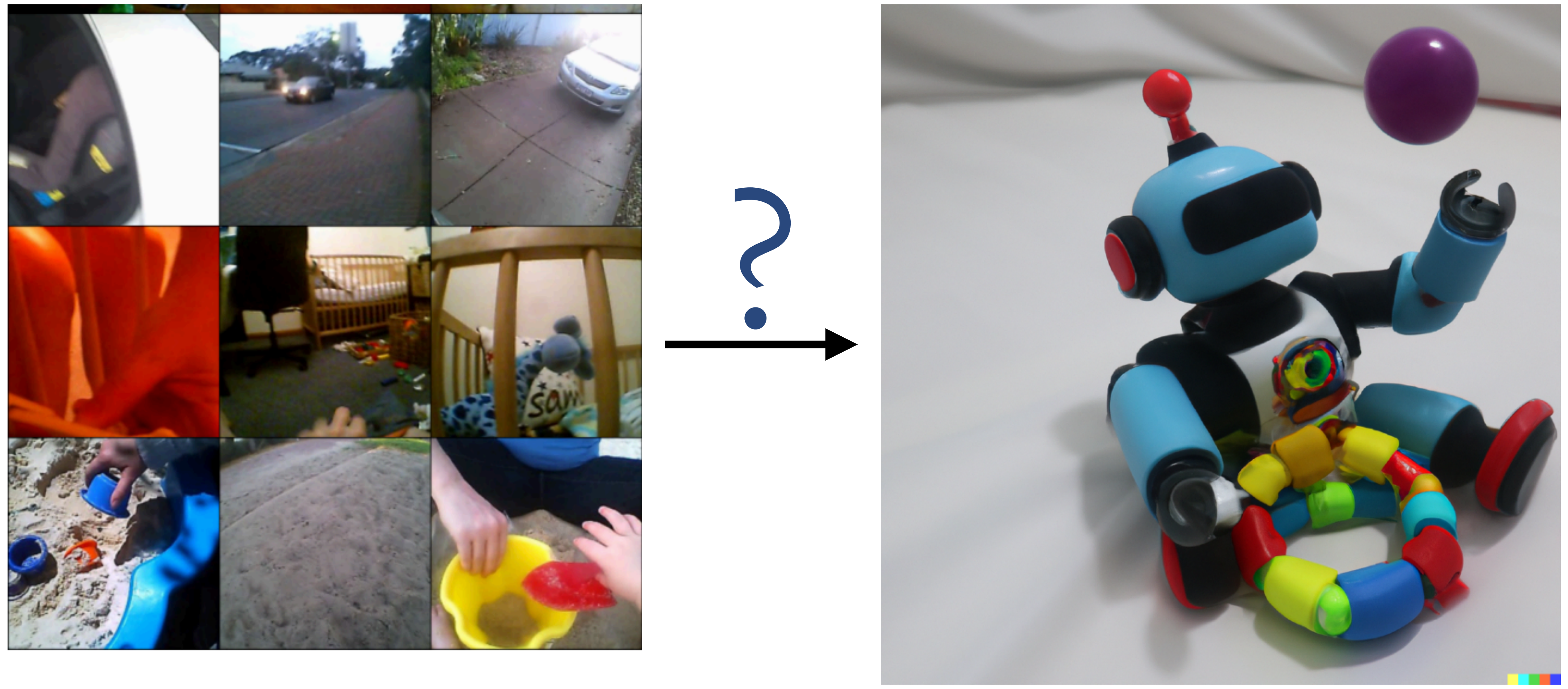
Reason 4: Investigating the fundamentals of visual understanding



RESEARCH

Self-supervised learning: The dark matter of intelligence

March 4, 2021



As babies, we learn how the world works largely by observation. We form generalized predictive models about objects in the world by learning concepts such as object permanence and gravity. Later in life, we observe the world, act on it, observe again, and build hypotheses to explain how our actions change our environment by trial and error.

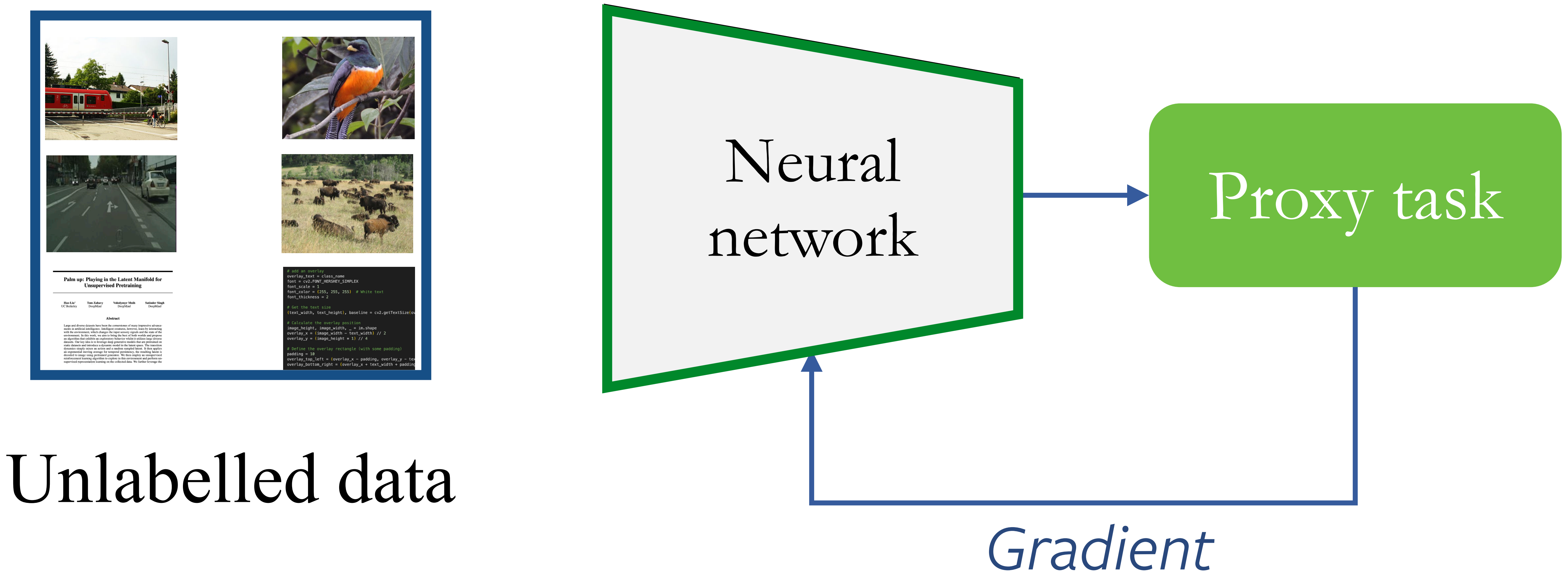
What, if there are, are the limits of learning without labels?

https://ai.facebook.com/blog/self-supervised-learning-the-dark-matter-of-intelligence/
Orhan et al. Self-supervised learning through the eyes of a child. NeurIPS 2020

Self-supervised Learning

General procedure of self-supervised learning.

Phase 1: Pretraining


Unlabelled data

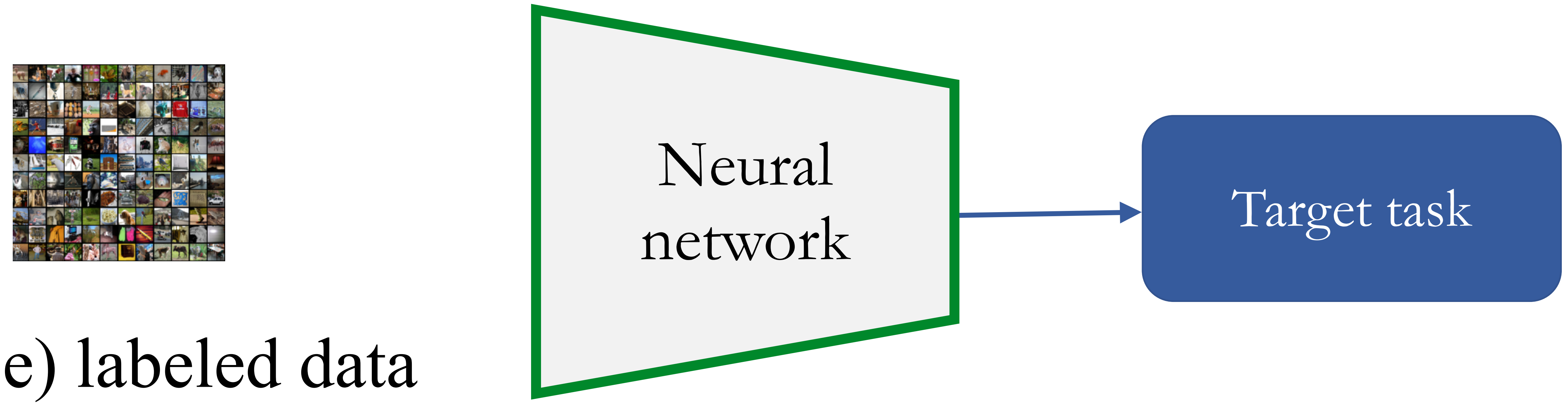
Neural network

Proxy task

Gradient

Neural Network can be Language Model or Vision Model...

Phase 2: Downstream tasks


(Sparse) labeled data

Neural network

Target task

Advantage of having phase 1:

- Better performance in phase 2
- Less labels required in phase 2

Downstream semi-supervised tasks: Self-supervised Learning helps

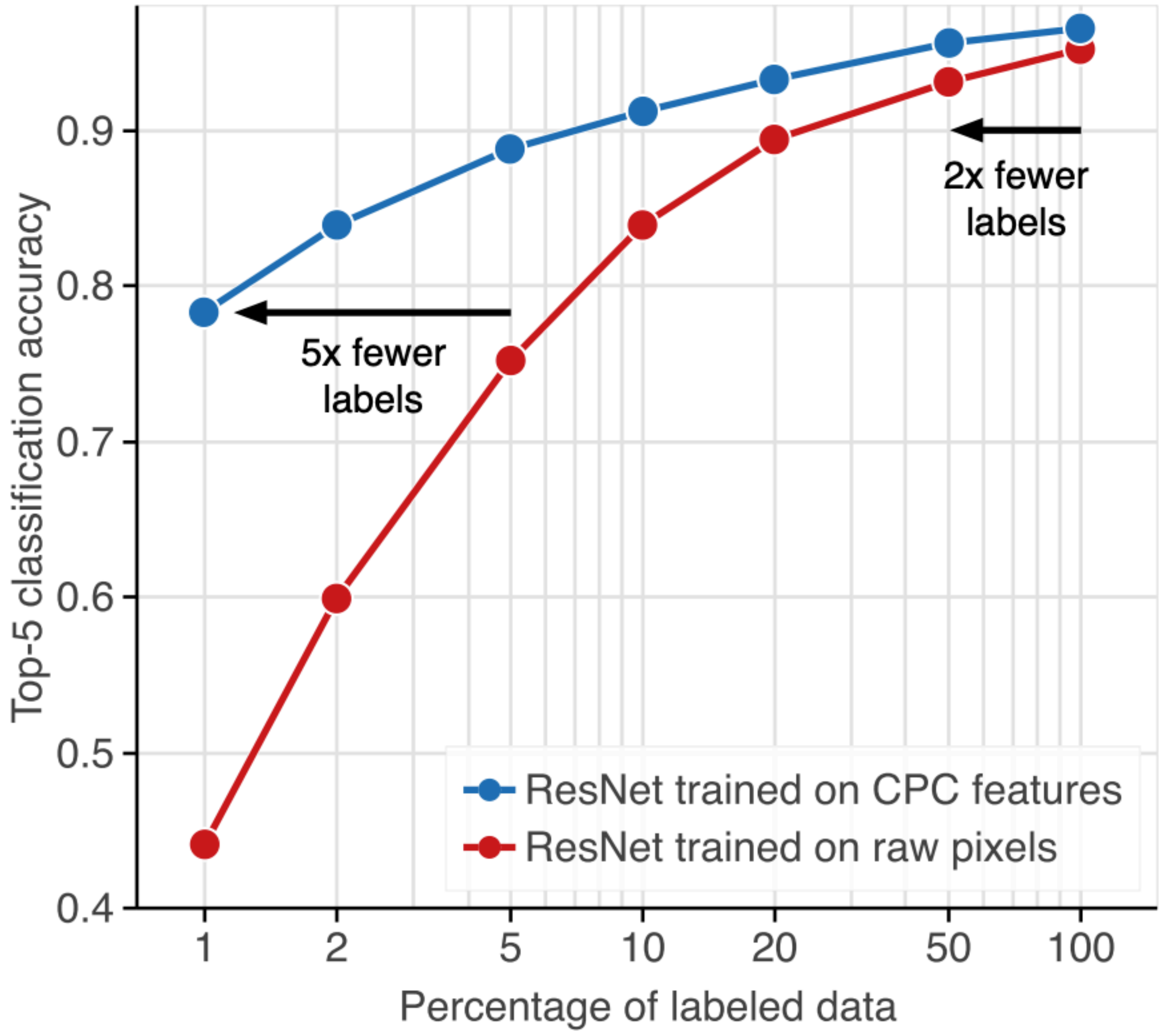


Figure 1. Data-efficient image recognition with Contrastive Predictive Coding. With decreasing amounts of labeled data, supervised networks trained on pixels fail to generalize (red). When trained on unsupervised representations learned with CPC, these networks retain a much higher accuracy in this low-data regime (blue). Equivalently, the accuracy of supervised networks can be matched with significantly fewer labels (horizontal arrows).

Once pretrained, self-supervised networks good for quick transfer learning even with few labels

Achieves much better performance for low number of annotated data

This is the case if you were to found a startup and tackle a new problem (annotation=expensive)

Fundamental
AI Lab

Data-Efficient Image Recognition with Contrastive Predictive Coding. Henaff et al.

Question 1: How does one learn without labels?

Question 2: How for images? How for text?

Question 3: How can we combine them?

Question 4: Why is everyone so hyped about this?

Question 5: I want GPT4V now!?

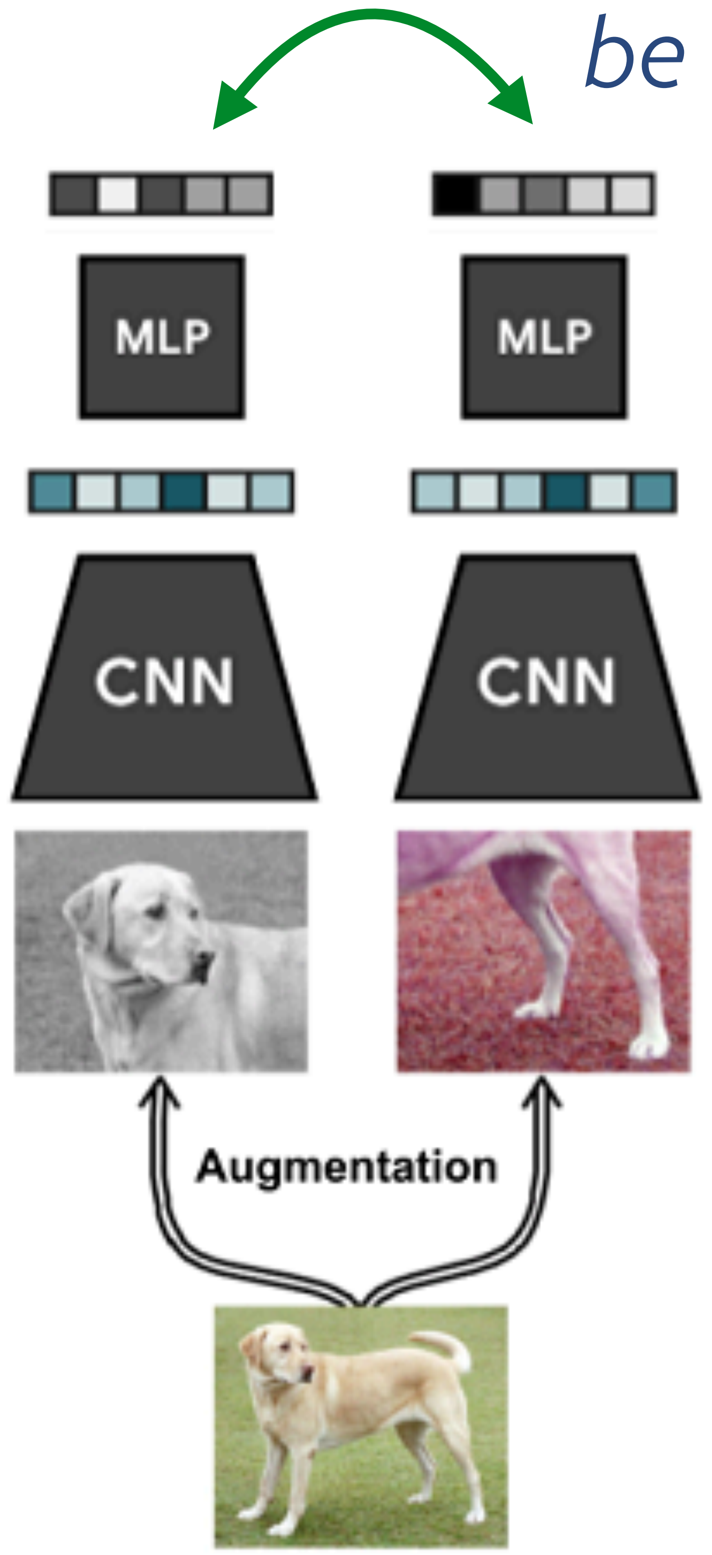
Fundamental AI Lab

UTN

One way to train: Noise-contrastive self-supervised learning

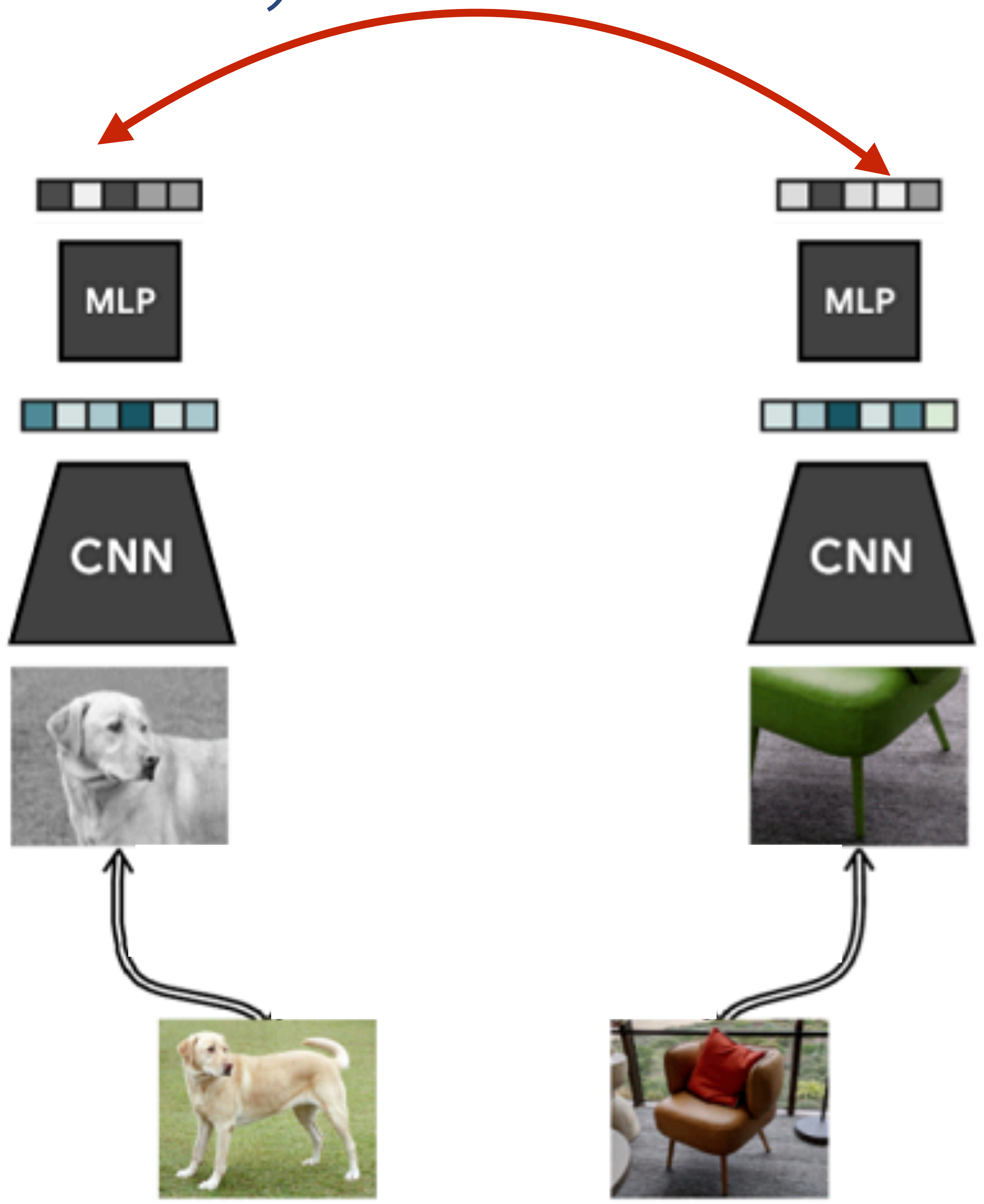
Core idea:

1) these should be the same



Augmentations should not change the embedding of an image

2) these should be different

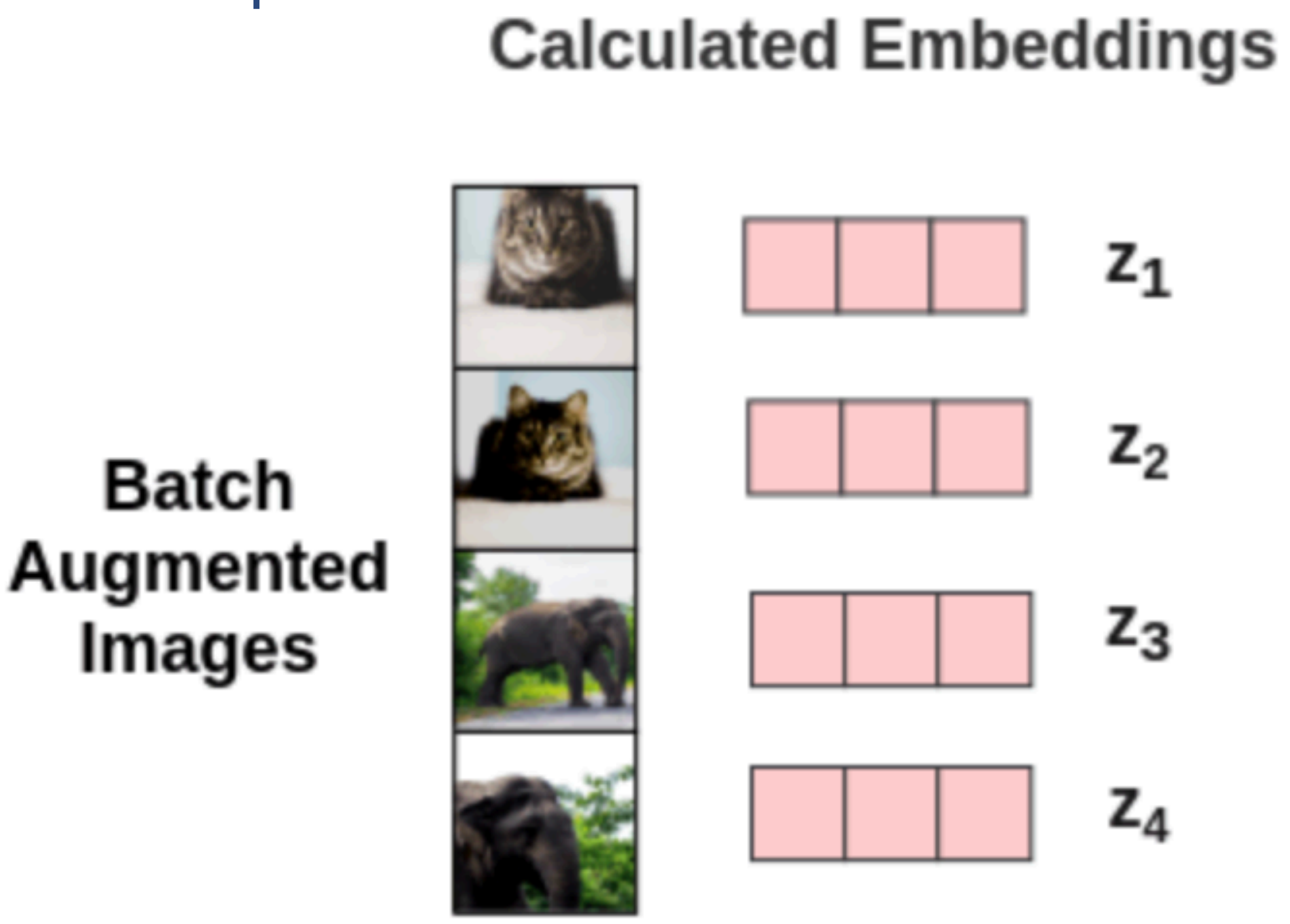


Different images should have different embeddings

Dosovitskiy et al. Discriminative Unsupervised Feature Learning with Exemplar Convolutional Neural Networks. 2015
Wu et al. *Unsupervised Feature Learning via Non-Parametric Instance-level Discrimination.* CVPR 2018
Chen et al. *A Simple Framework for Contrastive Learning of Visual Representations.* ICML 2020

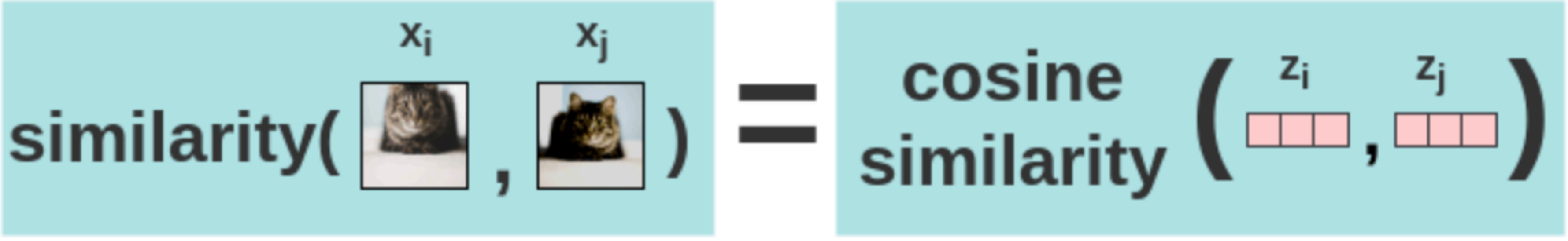
How SimCLR works in detail

Step 1



Step 2

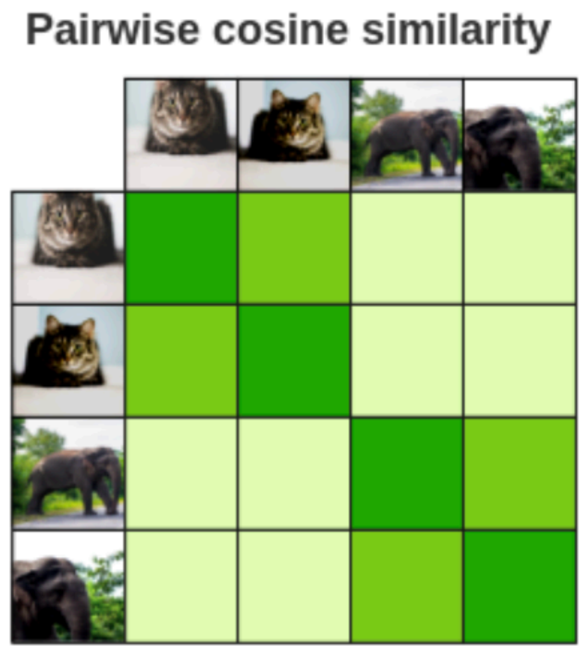
Similarity Calculation of Augmented Images



$$s_{i,j} = \frac{z_i^T z_j}{(\tau ||z_i|| ||z_j||)}$$

- τ is the adjustable temperature parameter. It can scale the inputs and widen the range `[-1, 1]` of cosine similarity
- ||z_i|| is the norm of the vector.

Step 3



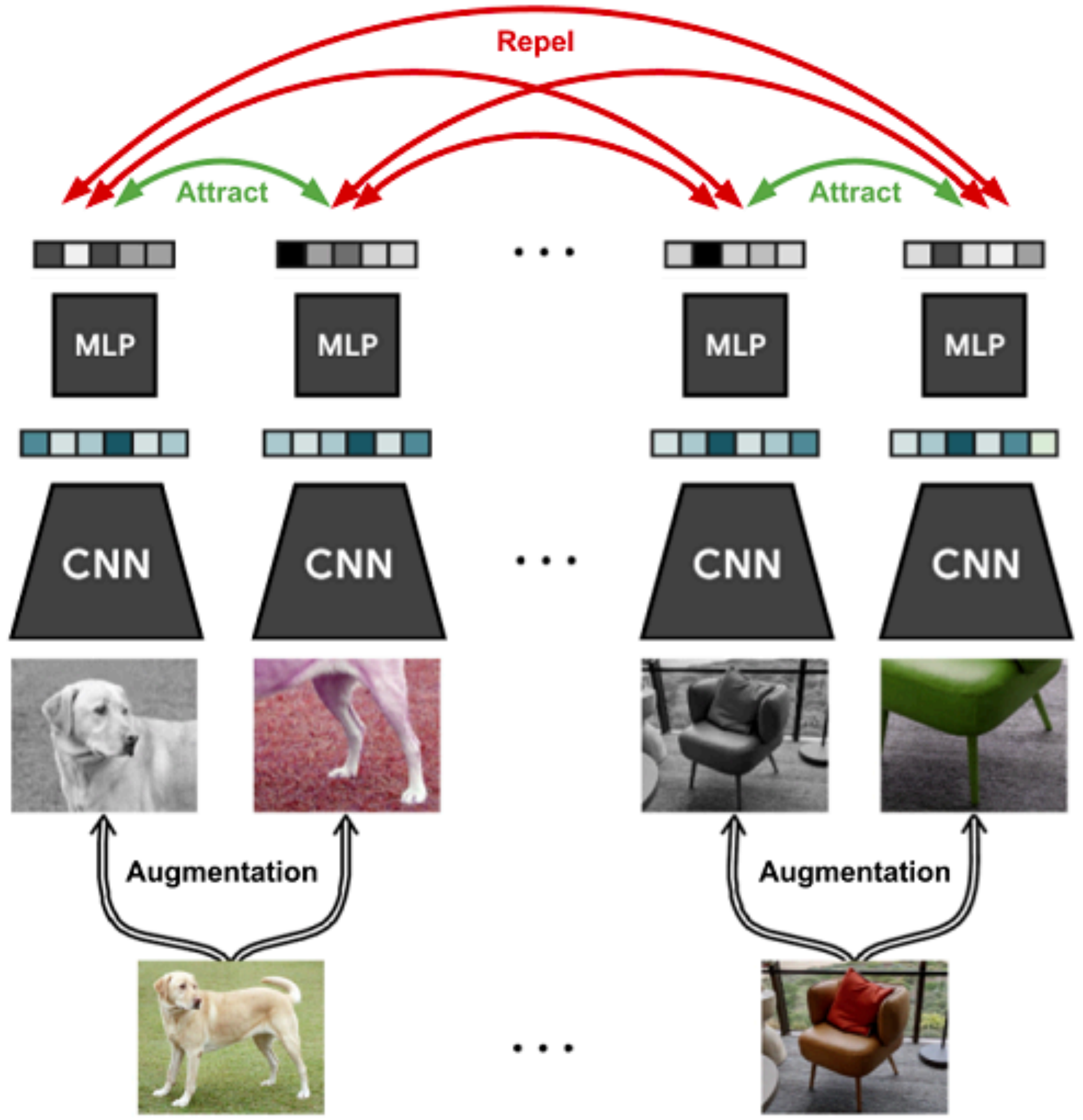
Loss: *relatively* increase similarity for pairs, decrease rest

What happens if you only try to increase the diagonal?

Fundamental AI Lab UTN

https://amitness.com/2020/03/illustrated-simclr/

Putting it into a loss function



SimCLR

The contrastive loss for positive pairs i,j:

$$\ell_{i,j} = -\log \frac{\exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \;_{[k\neq i]} \exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)},$$

with z_i, z_k embeddings for images *i* and *k*,

τ a temperature, sim() *is the dot-product*

"non-parametric" softmax

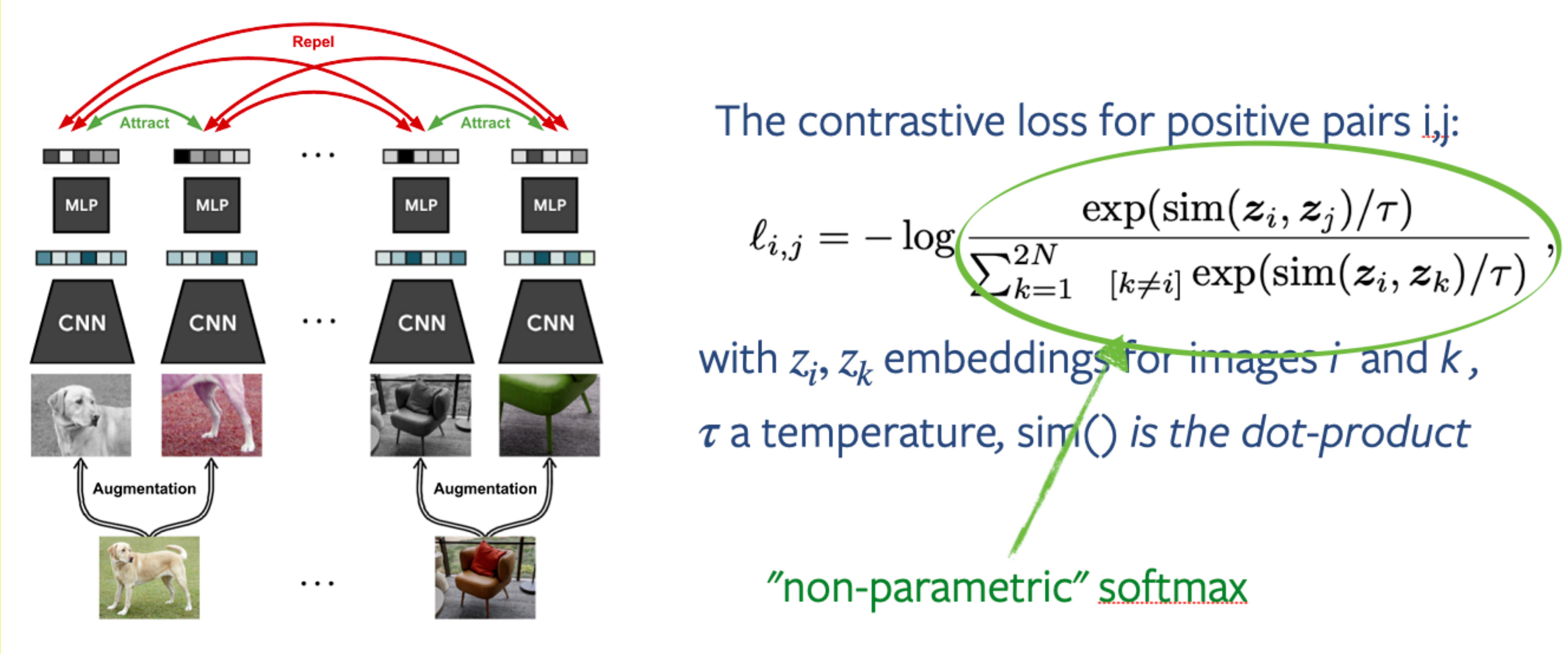
Enforces image-uniqueness and

enforces augmentation-invariance

Fundamental AI Lab UTN

Wu et al. *Unsupervised Feature Learning via Non-Parametric Instance-level Discrimination.* CVPR 2018
Chen et al. *A Simple Framework for Contrastive Learning of Visual Representations.* ICML 2020

Turn to your neighbor and answer + discuss these questions! (2min)



The contrastive loss for positive pairs i,j:

$$\ell_{i,j} = -\log \frac{\exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} {}_{[k\neq i]} \exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)},$$

with z_i, z_k embeddings for images i and k,

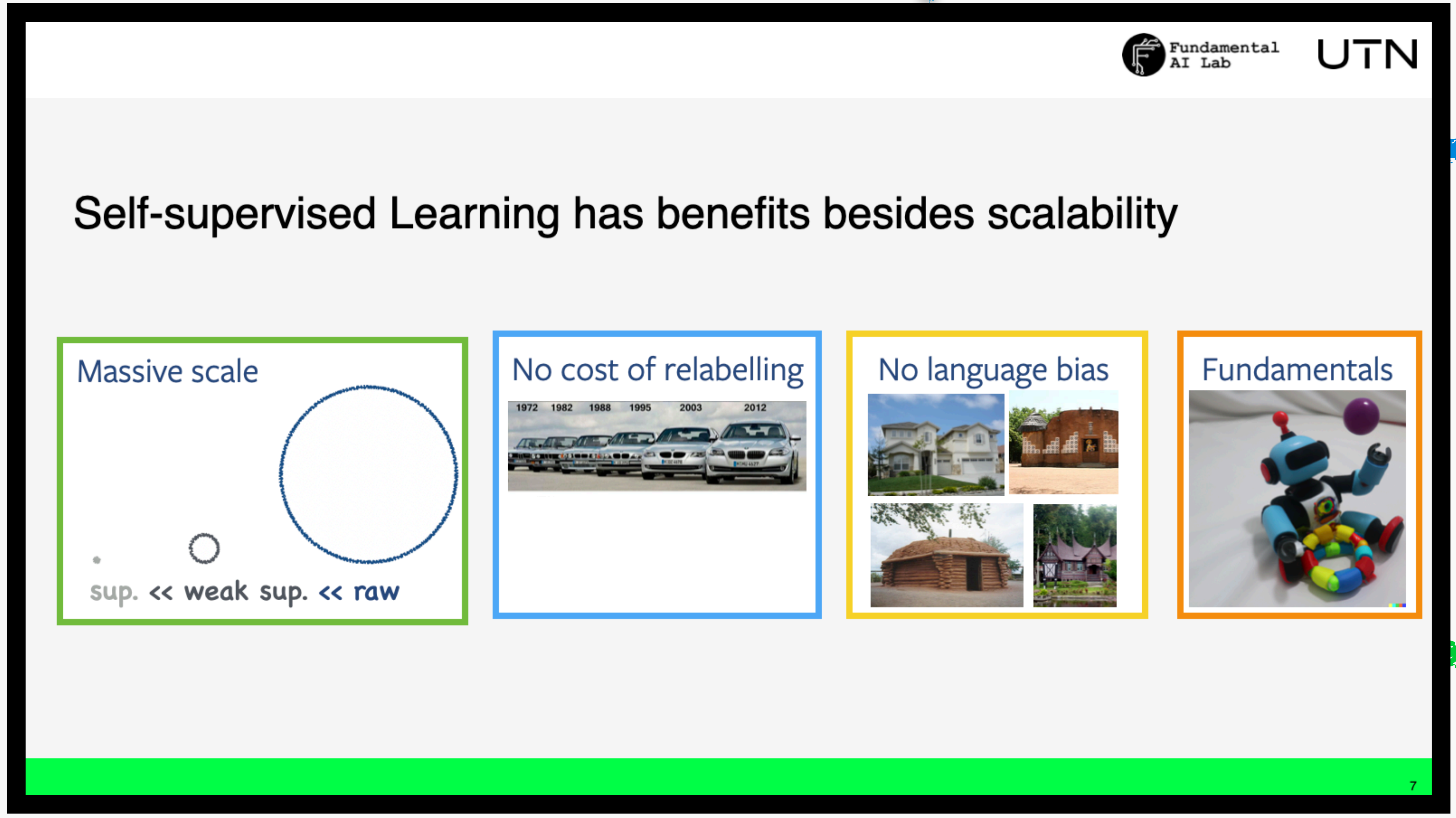
τ a temperature, sim() *is the dot-product*

"non-parametric" softmax

Q1: Neighbor right to neighbor left: explain what the loss is doing, exactly. Loss here denotes "i,j". How many of these losses do we have in total for a N different images?

Q2 Neighbor left to neighbor right: explain why it's called non-parametric softmax. How is it different from, .e.g a softmax at the end of a ImageNet-1k classification network?

Self-supervised vision Foundation Models are hard to obtain


Fundamental AI Lab UTN

Self-supervised Learning has benefits besides scalability

Massive scale

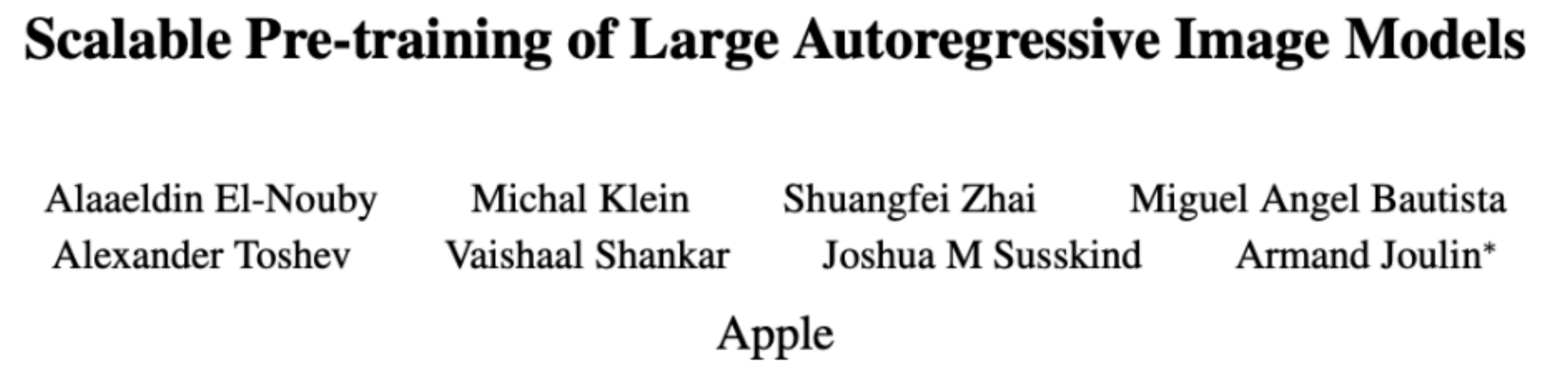
sup. << weak sup. << raw

No cost of relabelling

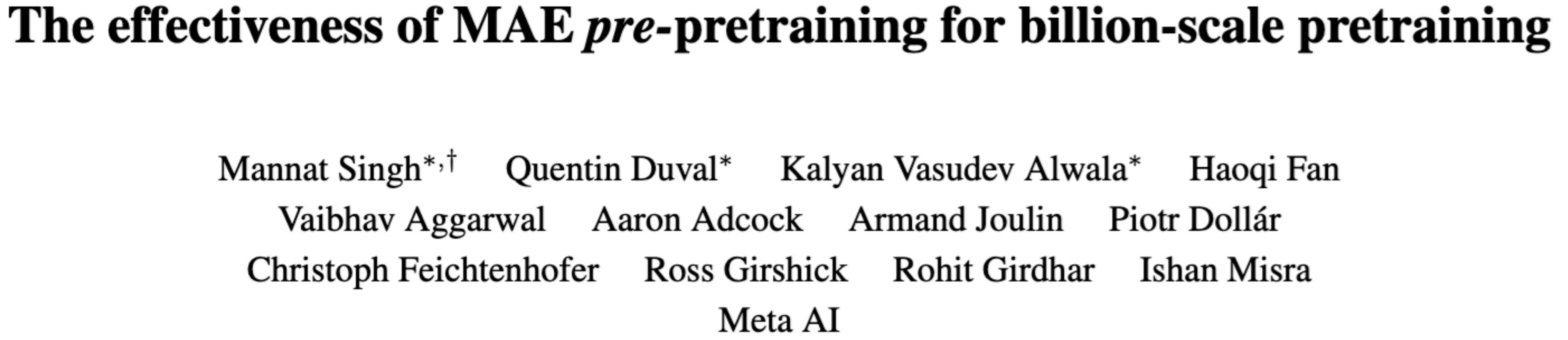
No language bias

Fundamentals

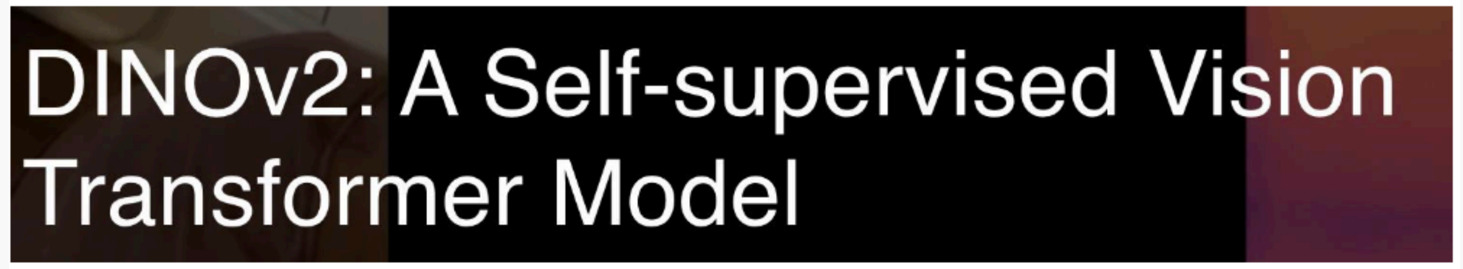
7

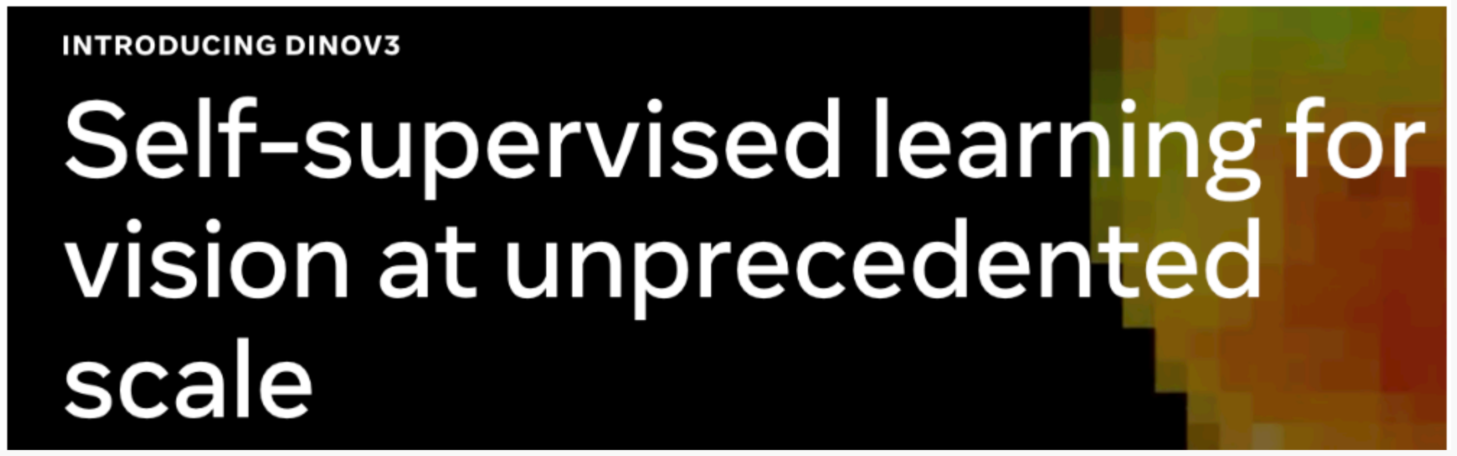

Scalable Pre-training of Large Autoregressive Image Models

Alaaeldin El-Nouby Michal Klein Shuangfei Zhai Miguel Angel Bautista
Alexander Toshev Vaishaal Shankar Joshua M Susskind Armand Joulin*
Apple

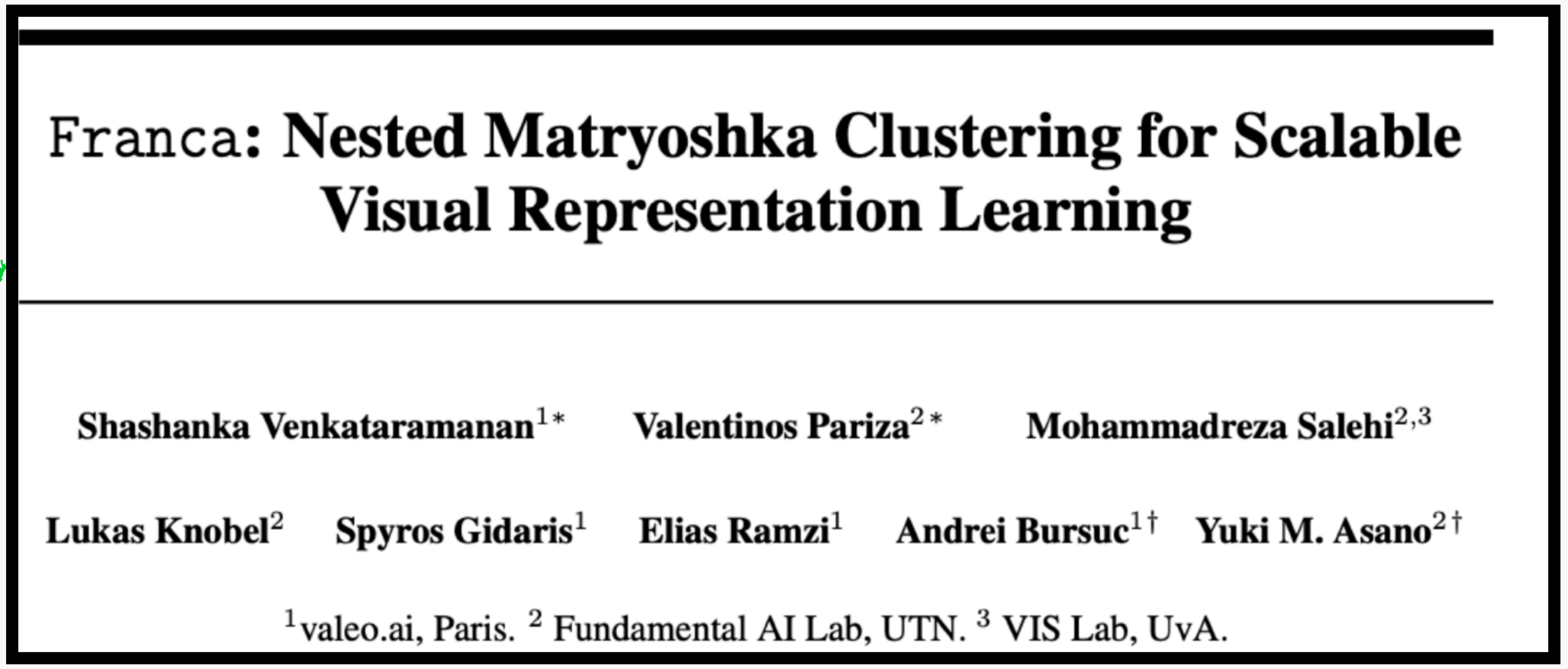

The effectiveness of MAE *pre*-pretraining for billion-scale pretraining

Mannat Singh*,† Quentin Duval* Kalyan Vasudev Alwala* Haoqi Fan
Vaibhav Aggarwal Aaron Adcock Armand Joulin Piotr Dollár
Christoph Feichtenhofer Ross Girshick Rohit Girdhar Ishan Misra
Meta AI


DINOv2: A Self-supervised Vision Transformer Model

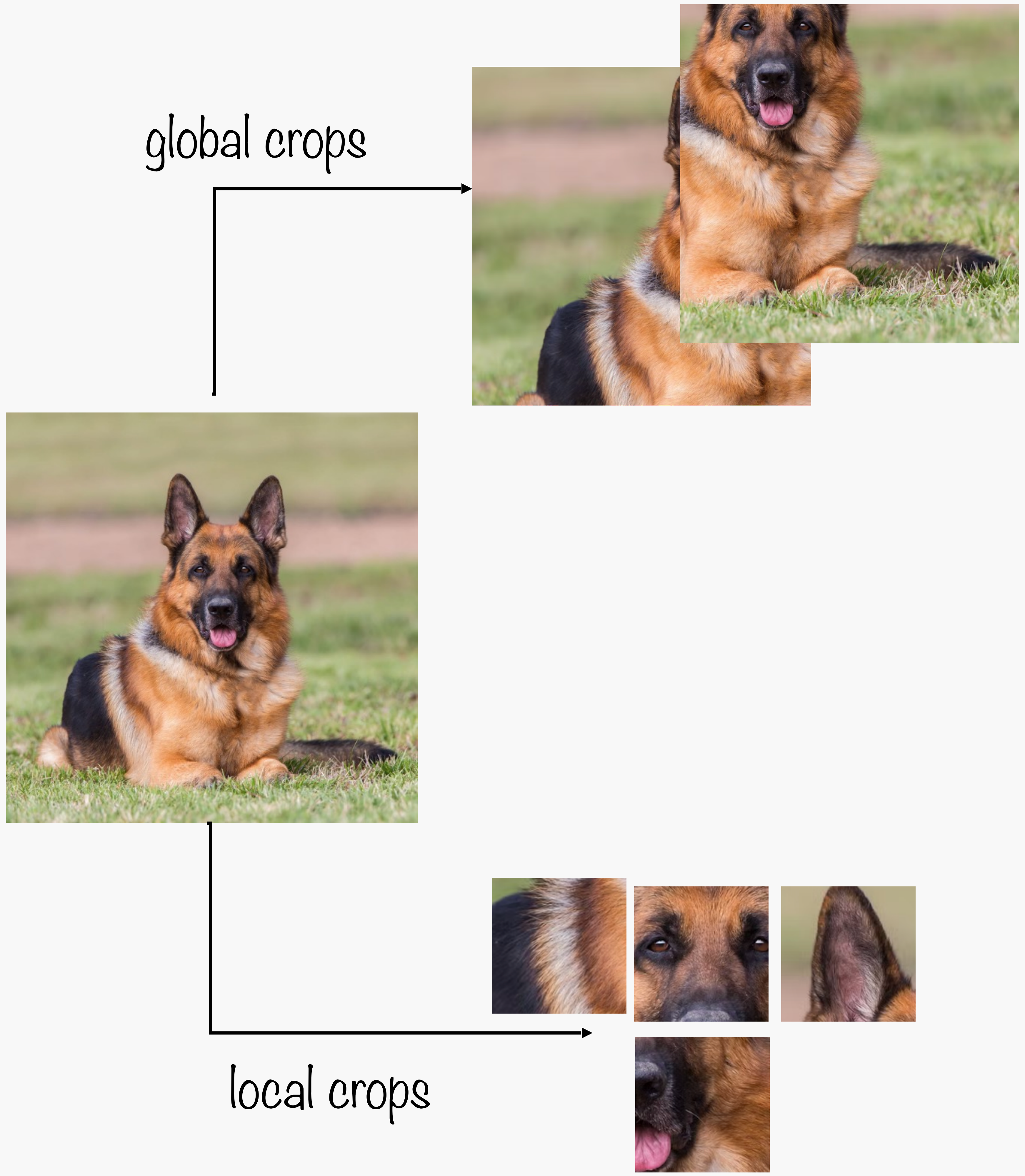

INTRODUCING DINOV3

Self-supervised learning for vision at unprecedented scale


`Franca`: Nested Matryoshka Clustering for Scalable Visual Representation Learning

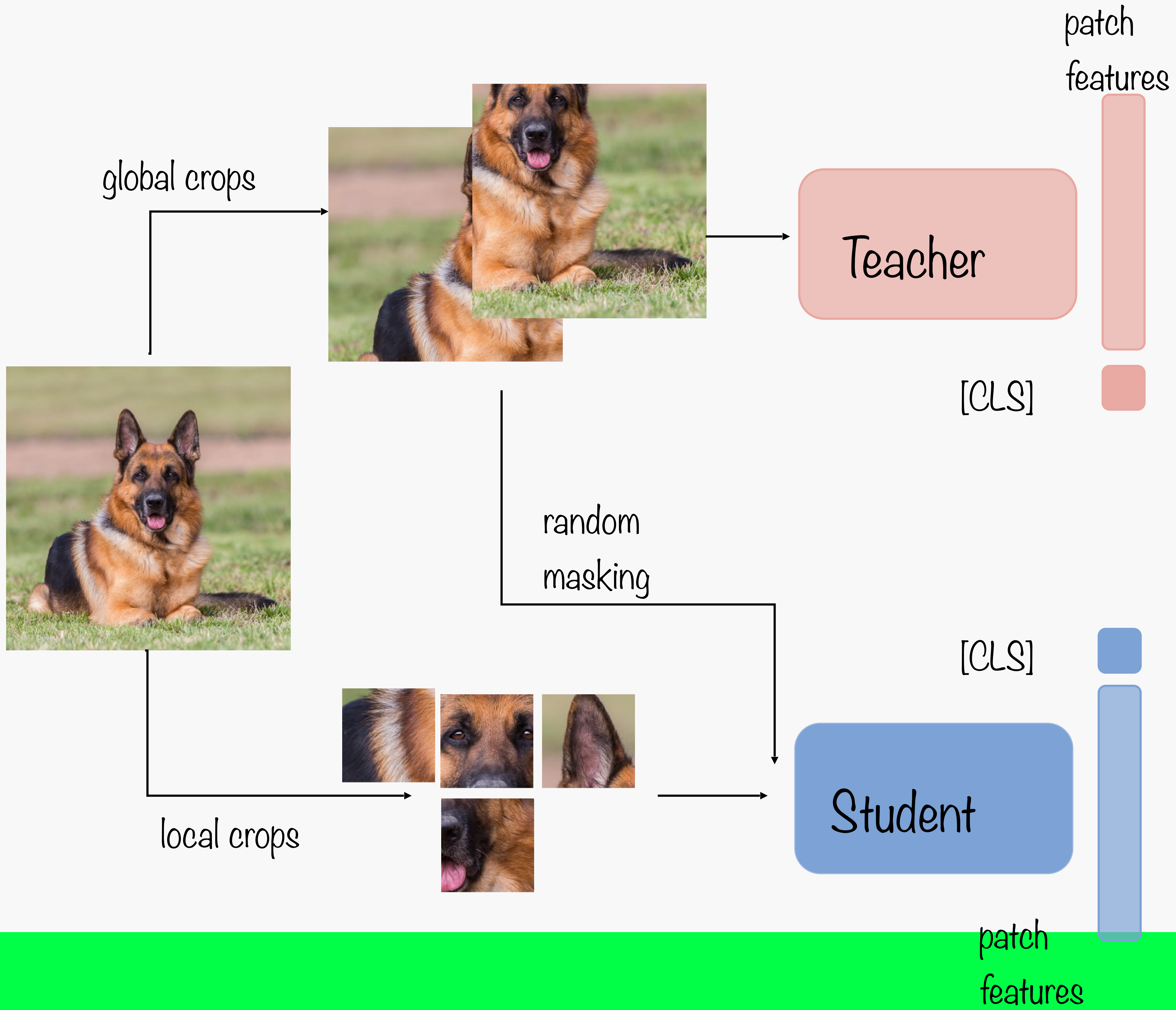
Shashanka Venkataramanan[1]* Valentinos Pariza[2]* Mohammadreza Salehi[2,3]
Lukas Knobel[2] Spyros Gidaris[1] Elias Ramzi[1] Andrei Bursuc[1]† Yuki M. Asano[2]†
[1]valeo.ai, Paris. [2] Fundamental AI Lab, UTN. [3] VIS Lab, UvA.

Understanding DINOv2/v3

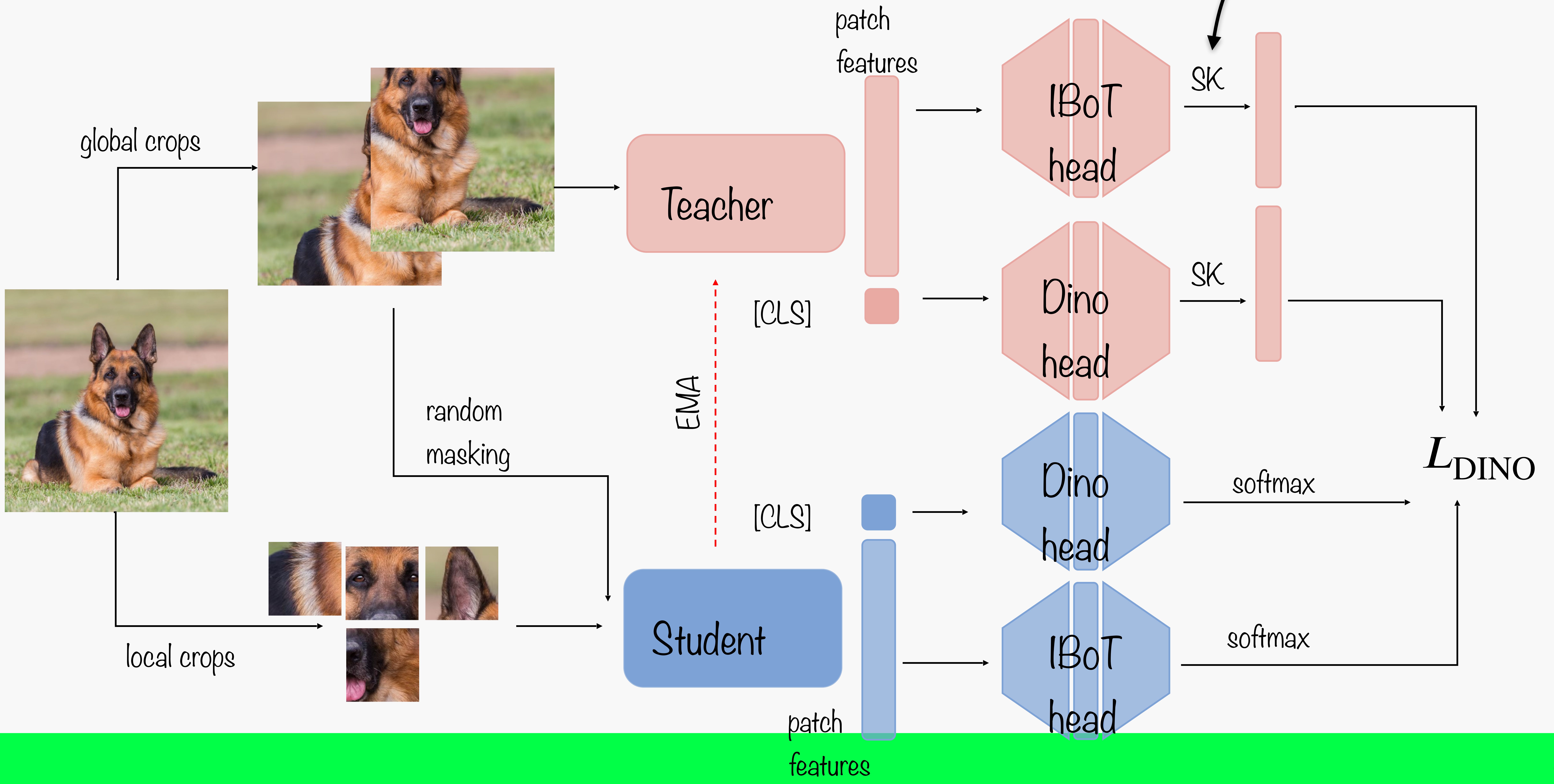


Understanding DINOv2/v3

UTN

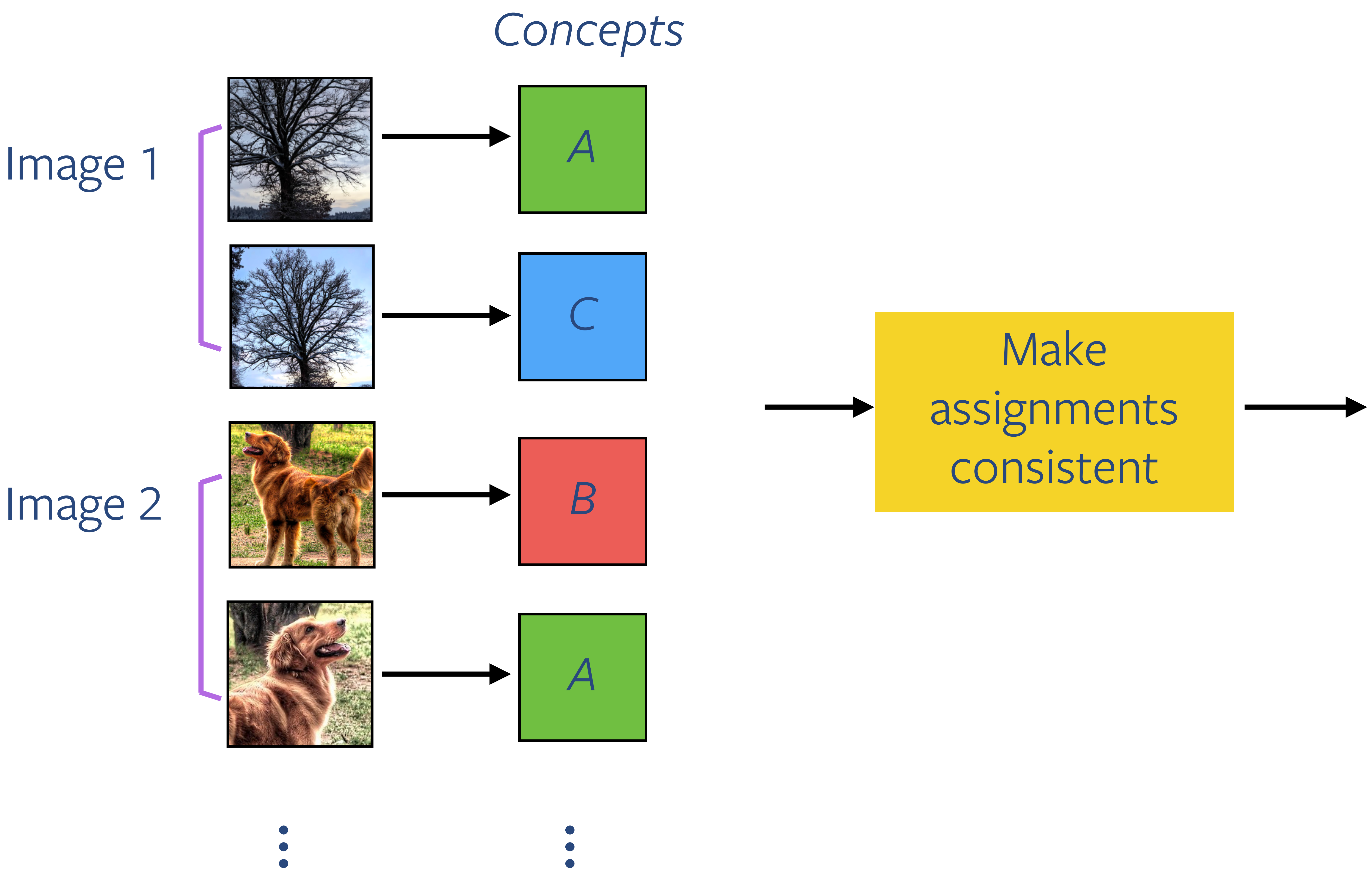


Understanding DINOv2/v3



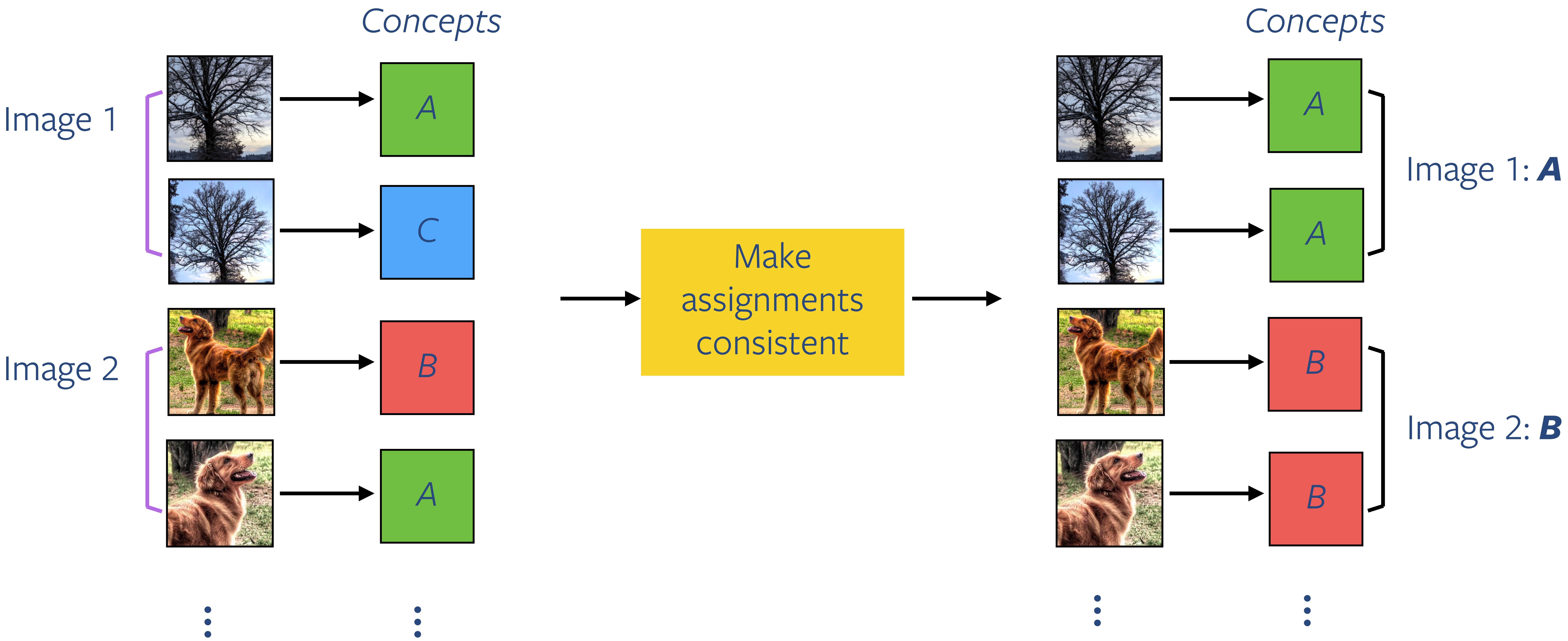
How the ”SK” step in DINOv2/DINOv3 works

Our work applies the idea of augmentation invariance to assign concepts.



Self-labelling via simultaneous clustering and representation learning. Asano et al. ICLR 2020

Our work applies the idea of transformation invariance to assign concepts.



Self-labelling via simultaneous clustering and representation learning. Asano et al. ICLR 2020

How can we optimize the labels and make assignments consistent?

If we had ground-truth labels

$$\min_{\Phi} L(y, \Phi),$$

where $L(y, \Phi) = \frac{1}{N} \sum_{i=1}^{N} \log p(y_i \mid \mathbf{x}_i, \Phi)$

- L is the loss (cost) function
- Φ is the deep neural network model
- y are the labels

Our novel contribution *without* ground-truth

Solution sketch:

1. Represent via an assignment table q and optimize:

$$L(q, \Phi) = \frac{1}{N} \sum_{i=1}^{N} \sum_{y} q(y \mid \mathbf{x}_i) \log p(y \mid \mathbf{x}_i, \Phi)$$

But: The trivial solution for q is to set all labels to be the same

2. Use pseudolabels an equal number of times: 🟥🟥🟩🟩🟦🟦 🟥🟥🟩🟩🟦🟦

3. Pose as approximate optimal transport:

$$\min_{q, \Phi} L(q, \Phi) \quad \text{s.t.} \quad \sum_{i=1}^{N} q(y \mid \mathbf{x}_i) = \frac{N}{K},$$

Fundamental AI Lab

Self-labelling via simultaneous clustering and representation learning. Asano et al. ICLR 2020
Sinkhorn distances: Lightspeed computation of optimal transport. Cuturi. NeurIPS 2013

SK optimisation

$$\min_{P\in U} F(P) = \min_{P\in U}\Big[\langle Q, -\log P\rangle - \lambda h(P)\Big]$$

$$0 = \frac{\mathrm{d}}{\mathrm{d}p_{ij}} F$$

$$= \frac{\mathrm{d}}{\mathrm{d}p_{ij}}\Big[\sum_{ij} Q_{ij}P_{ij} + \lambda P_{ij}\log(P_{ij}) + \sum_i \alpha_i(\sum_{ij} P_{ij} - 1) + \sum_j \beta_j(\sum_{ij} P_{ij} - 1)\Big]$$

$$= Q_{ij} + \lambda\log(P_{ij}) + \lambda + \alpha_i + \beta_j$$

Hence

$$P_{ij} = \exp\big(-\lambda^{-1}\alpha_i - \lambda^{-1}Q_{ij} - 1 - \lambda^{-1}\beta_j\big)$$

$$= u_i e^{-\lambda^{-1}Q_{ij}} v_j \quad = u_i e^{\lambda^{-1}log(q)} v_j$$

SK optimisation of assignments Q

$$\min_{Q\in U} L = \min_{Q\in U}\left[\langle Q, \underbrace{-\log P}_{C\geq 0,\quad \text{costs}}\rangle - \frac{1}{\lambda}h(Q)\right]$$

using $H(Q) = H(r) + H(c) - D_{KL}(Q\|rc^{\mathrm{T}}) = \log(NK) - D_{KL}(Q\|rc^{\mathrm{T}})$

$$\min_{Q\in U} L = \min_{Q\in U}\left[\langle Q, C\rangle + \frac{1}{\lambda}D_{KL}(Q\|rc^{\mathrm{T}})\right] + \text{const}\,.$$

Find minimum:

$$0 = \frac{\mathrm{d}}{\mathrm{d}q_{ij}}F \quad = \frac{\mathrm{d}}{\mathrm{d}q_{ij}}\Big[\sum_{ij} Q_{ij}C_{ij} + \frac{1}{\lambda}Q_{ij}\log(Q_{ij}) + \sum_{i}\alpha_i(\sum_{ij}Q_{ij} - 1) + \sum_{j}\beta_j(\sum_{ij}Q_{ij} - 1)\Big]$$

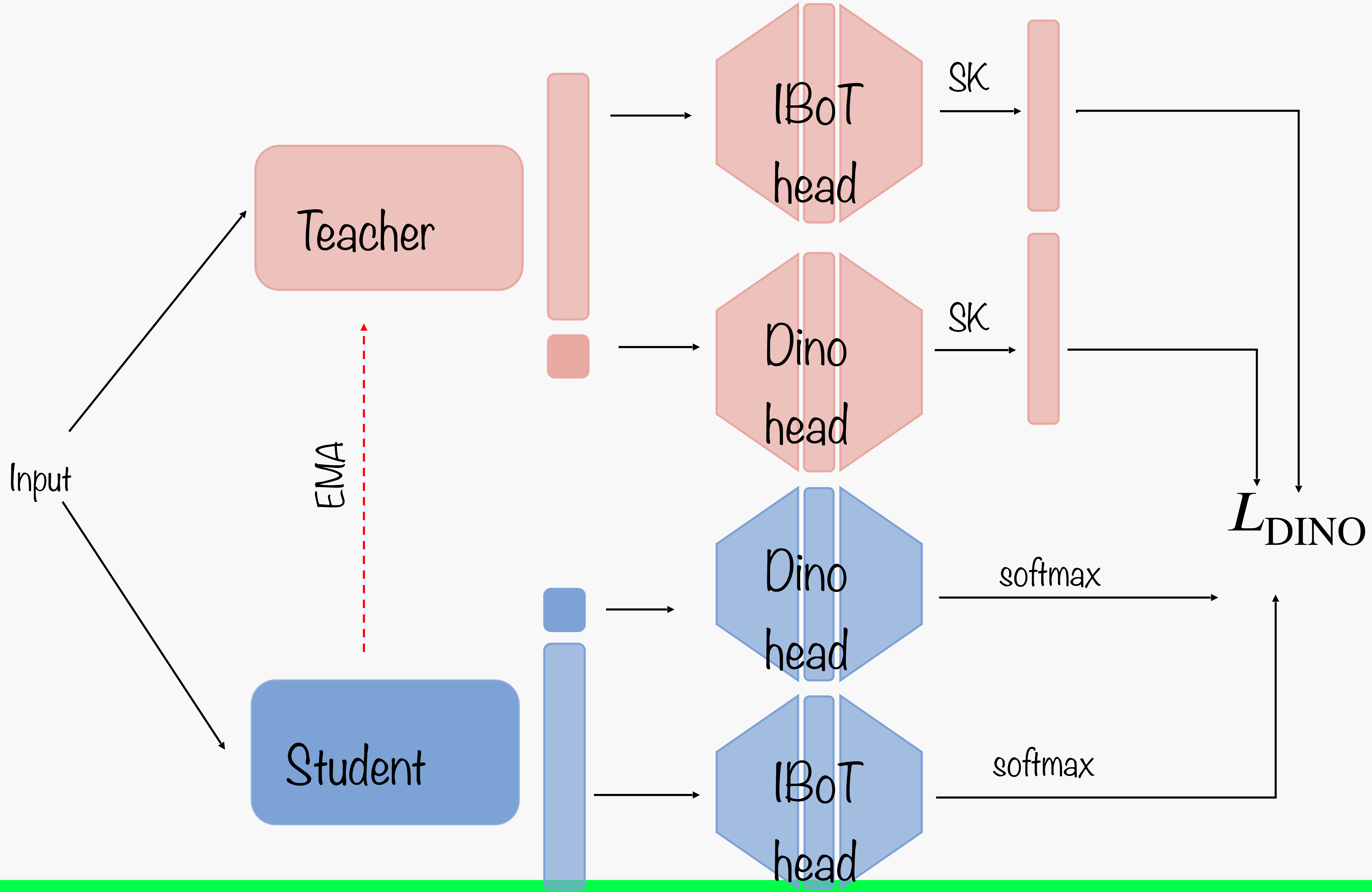
$$= C_{ij} + \frac{1}{\lambda}\log(Q_{ij}) + \lambda + \alpha_i + \beta_j$$

Hence:

$$Q_{ij} = \exp\big(-\lambda\alpha_i - \lambda C_{ij} - 1 - \lambda\beta_j\big)$$

$$= u_i e^{-\lambda C_{ij}} v_j = u_i e^{\lambda\log(p)} v_j = u_i p^{\lambda} v_j$$

Challenges in training DINOv2



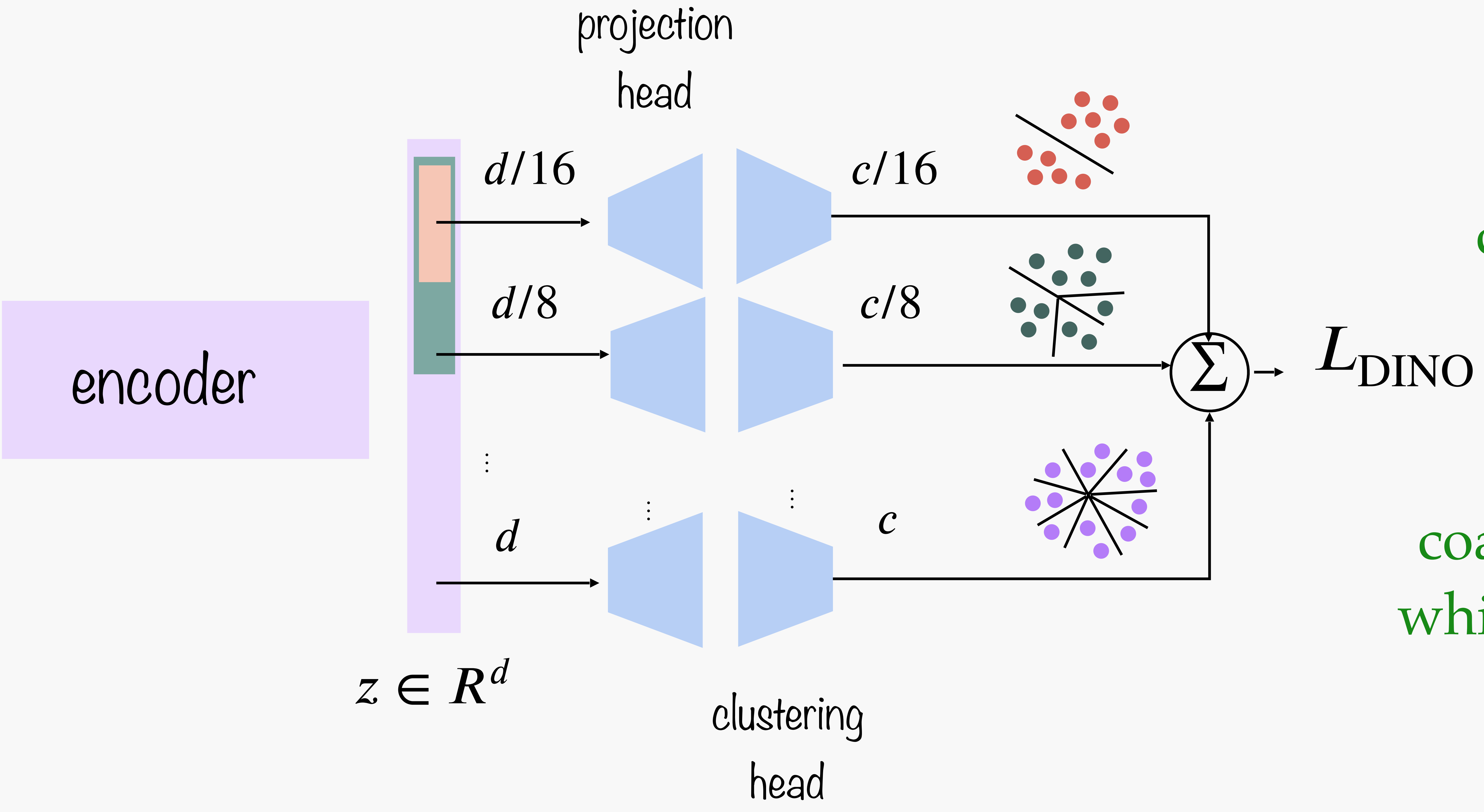
Maps features to very large codebook (131K in DINOv2, 262K in DINOv3)

Model extremely sensitive to hyper-params

Features have a high positional bias

Proprietary & "CVPR" curated data

Our Nested Matryoshka Clustering

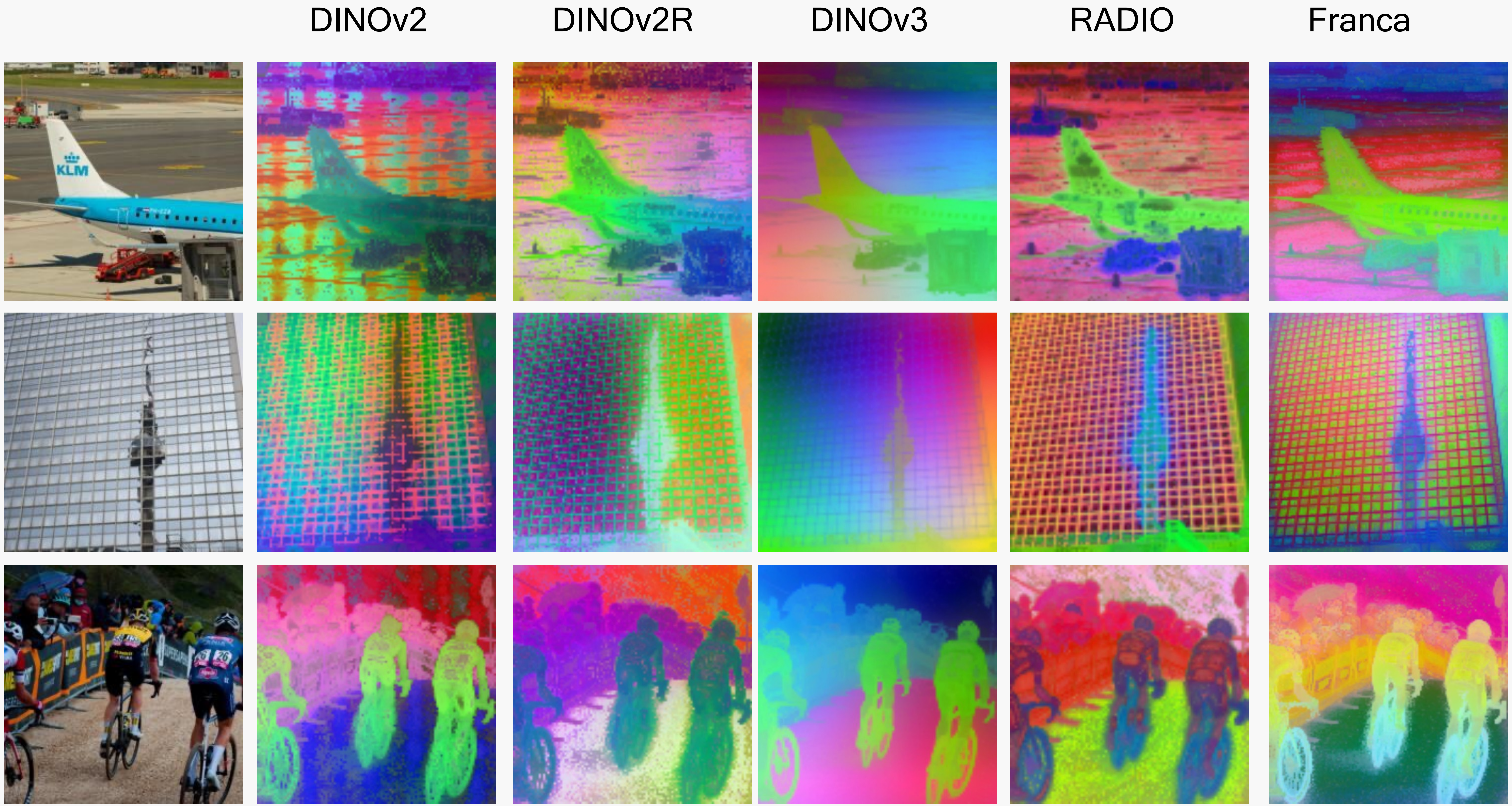


dedicated projection head and clustering objective to each subspace

coarse heads capture global semantics, while fine heads focus on local structure

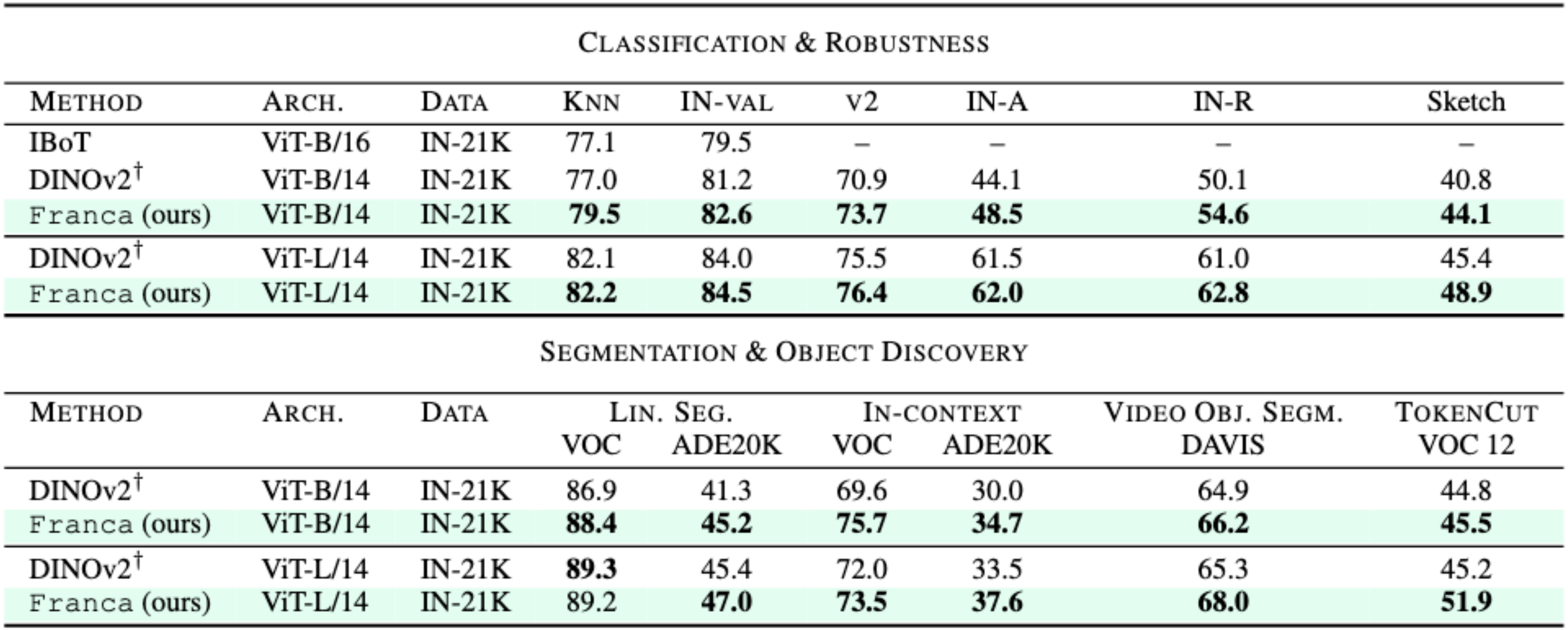
Franca: Nested Matryoshka Clustering for Scalable Visual Representation Learning. arXiv 2025. Venkataramanan, Pariza, Salehi, Knobel, Gidaris, Ramzi, Bursuc, Asano

Semantic Coherence Emerges in PCA Visualizations



Images were selected randomly with np.random.randint(seed=42)

Franca vs DINOv2 on *equal pretraining data*



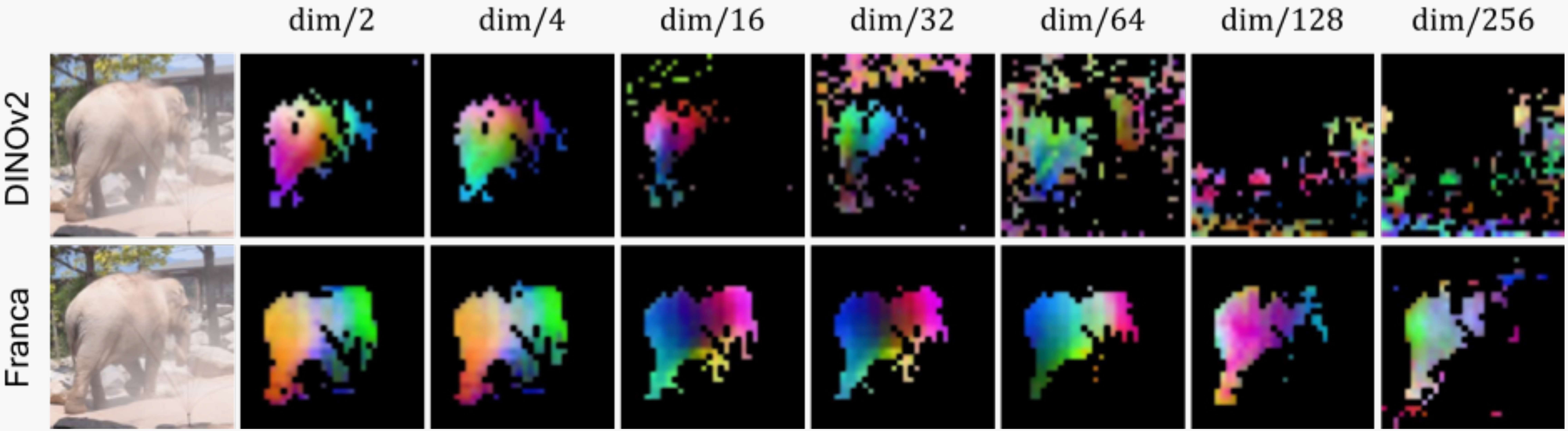
CLASSIFICATION & ROBUSTNESS

METHOD	ARCH.	DATA	KNN	IN-VAL	V2	IN-A	IN-R	Sketch
IBoT	ViT-B/16	IN-21K	77.1	79.5	–	–	–	–
DINOv2†	ViT-B/14	IN-21K	77.0	81.2	70.9	44.1	50.1	40.8
`Franca` (ours)	ViT-B/14	IN-21K	**79.5**	**82.6**	**73.7**	**48.5**	**54.6**	**44.1**
DINOv2†	ViT-L/14	IN-21K	82.1	84.0	75.5	61.5	61.0	45.4
`Franca` (ours)	ViT-L/14	IN-21K	**82.2**	**84.5**	**76.4**	**62.0**	**62.8**	**48.9**

SEGMENTATION & OBJECT DISCOVERY

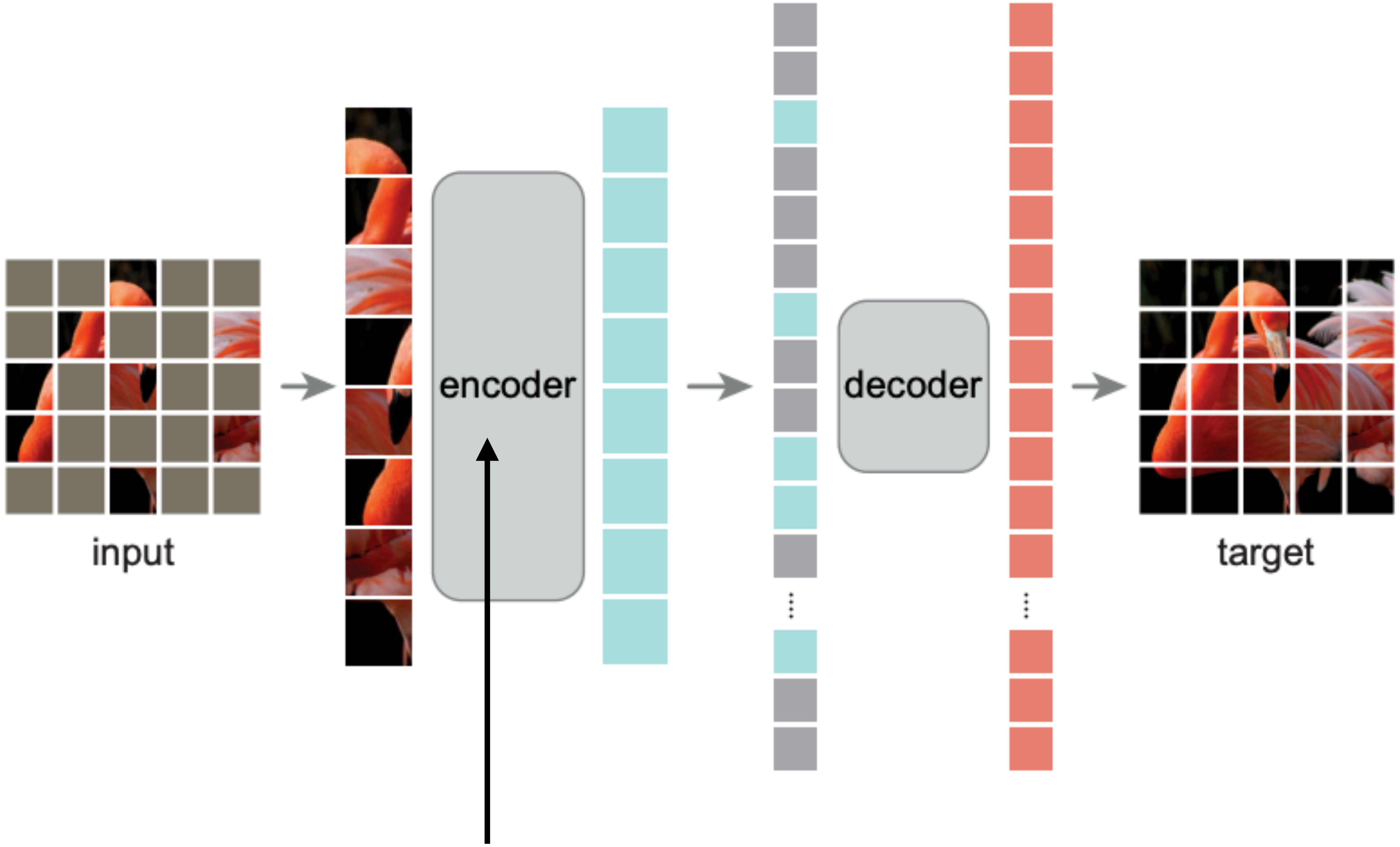
METHOD	ARCH.	DATA	LIN. SEG.		IN-CONTEXT		VIDEO OBJ. SEGM.	TOKENCUT
			VOC	ADE20K	VOC	ADE20K	DAVIS	VOC 12
DINOv2†	ViT-B/14	IN-21K	86.9	41.3	69.6	30.0	64.9	44.8
`Franca` (ours)	ViT-B/14	IN-21K	**88.4**	**45.2**	**75.7**	**34.7**	**66.2**	**45.5**
DINOv2†	ViT-L/14	IN-21K	**89.3**	45.4	72.0	33.5	65.3	45.2
`Franca` (ours)	ViT-L/14	IN-21K	89.2	**47.0**	**73.5**	**37.6**	**68.0**	**51.9**

Nested Matryoshka Maintains coherent part structures



Franca retains coherent part-level structure well beyond its trained dimensions

Another self-supervised task for images: Masked Image Modelling



Vision Transformer

The task:

- Mask out parts of the image
- Let the model predict the missing part

The motivation comes from Language Modelling, where we predict masked-out words in a sentence.

He et al. *Masked Autoencoders Are Scalable Vision Learners*. CVPR'21
Xie et al. *SimMIM: A Simple Framework for Masked Image Modeling. ArXiv*
Dosovitskiy et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. ICLR'21
https://www.sbert.net/examples/unsupervised_learning/MLM

Language Modelling in a nutshell

OH MY GOD, THE
VOLCANO IS ERU_
ERUDITE?

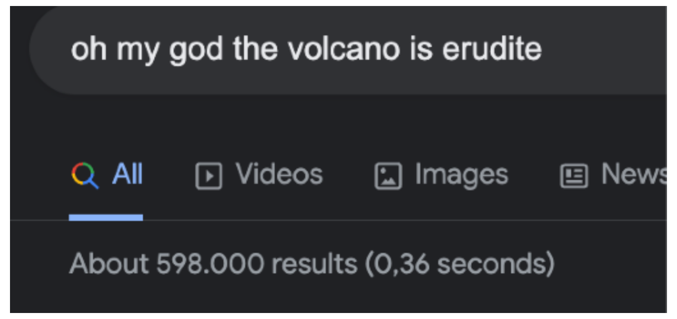
https://thegradient.pub/understanding-evaluation-metrics-for-language-models/

Language Modelling via next-word prediction: the most common way.

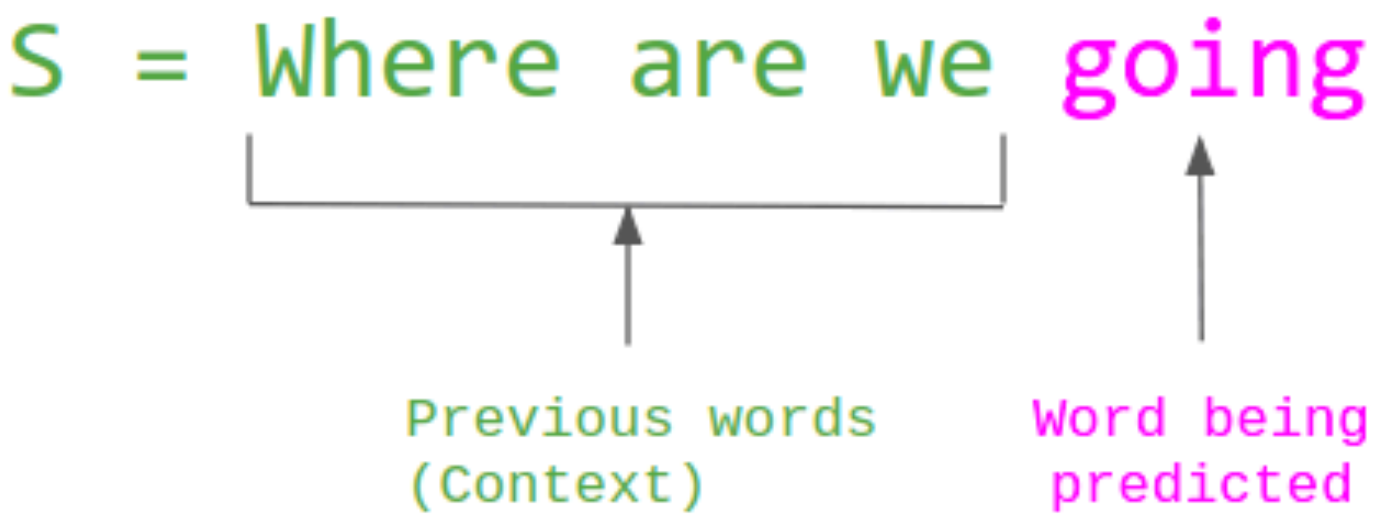
Why ”erudite” is not a good guess

Factor the probability of a datapoint (w_1,..., w_n):

$$P_{(w_1,w_2,...,w_n)} = p(w_1)p(w_2|w_1)p(w_3|w_1,w_2)...p(w_n|w_1,w_2,..,w_{n-1})$$
$$= \prod_{i=1}^{n} p(w_i|w_1,...,w_{i-1})$$

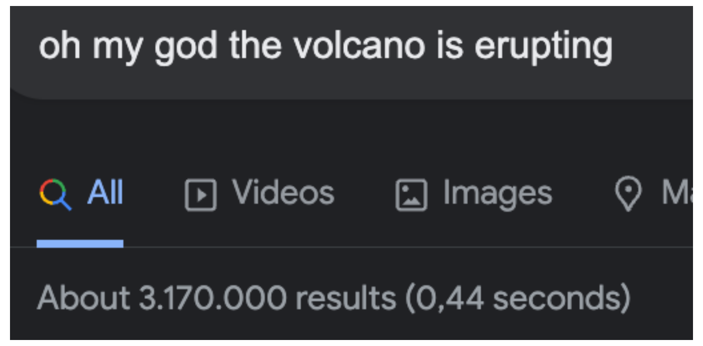

oh my god the volcano is erudite

About 598.000 results (0,36 seconds)


S = Where are we going

Previous words (Context)

Word being predicted


oh my god the volcano is erupting

About 3.170.000 results (0,44 seconds)

P(S) = P(Where) x P(are | Where) x P(we | Where are) x P(going | Where are we)

https://thegradient.pub/understanding-evaluation-metrics-for-language-models/
Radford et al. Improving Language Understanding by Generative Pre-Training . 2018

Who here knows how GPT-2/3 works?

raise your hand!

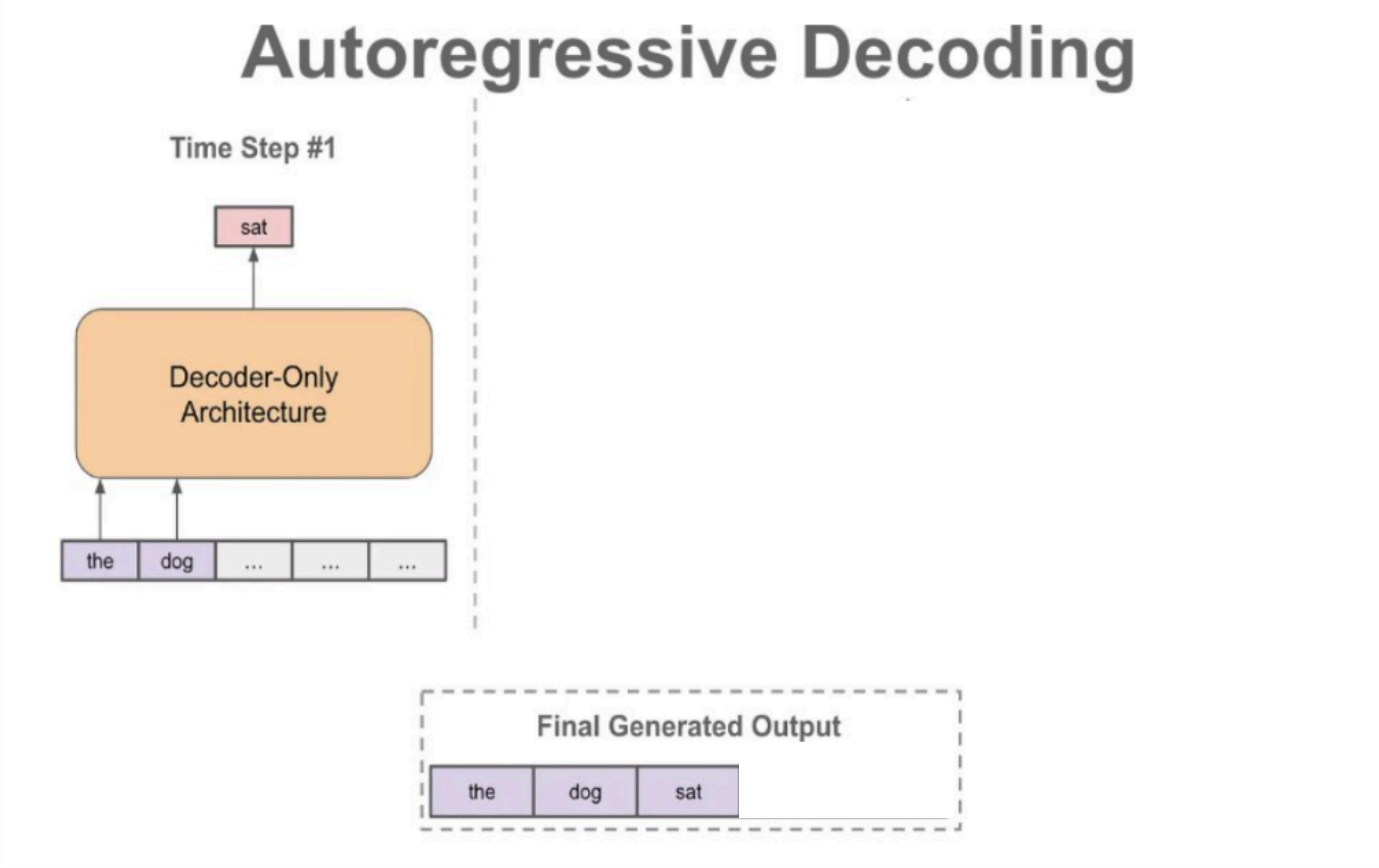
Who here knows what a tokenizer is?

raise your hand!

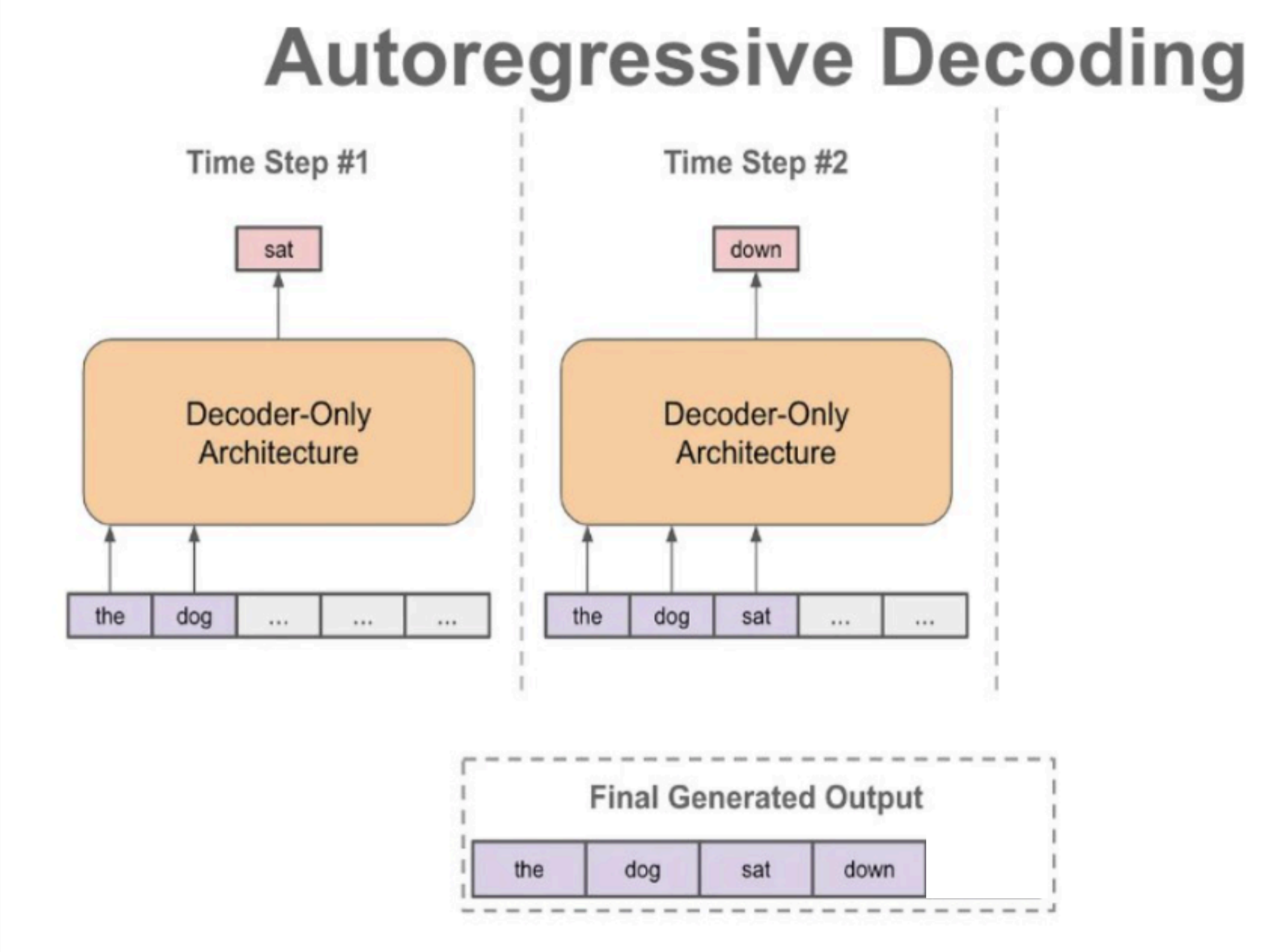
Who here knows why training via next-word prediction is so good?

raise your hand!

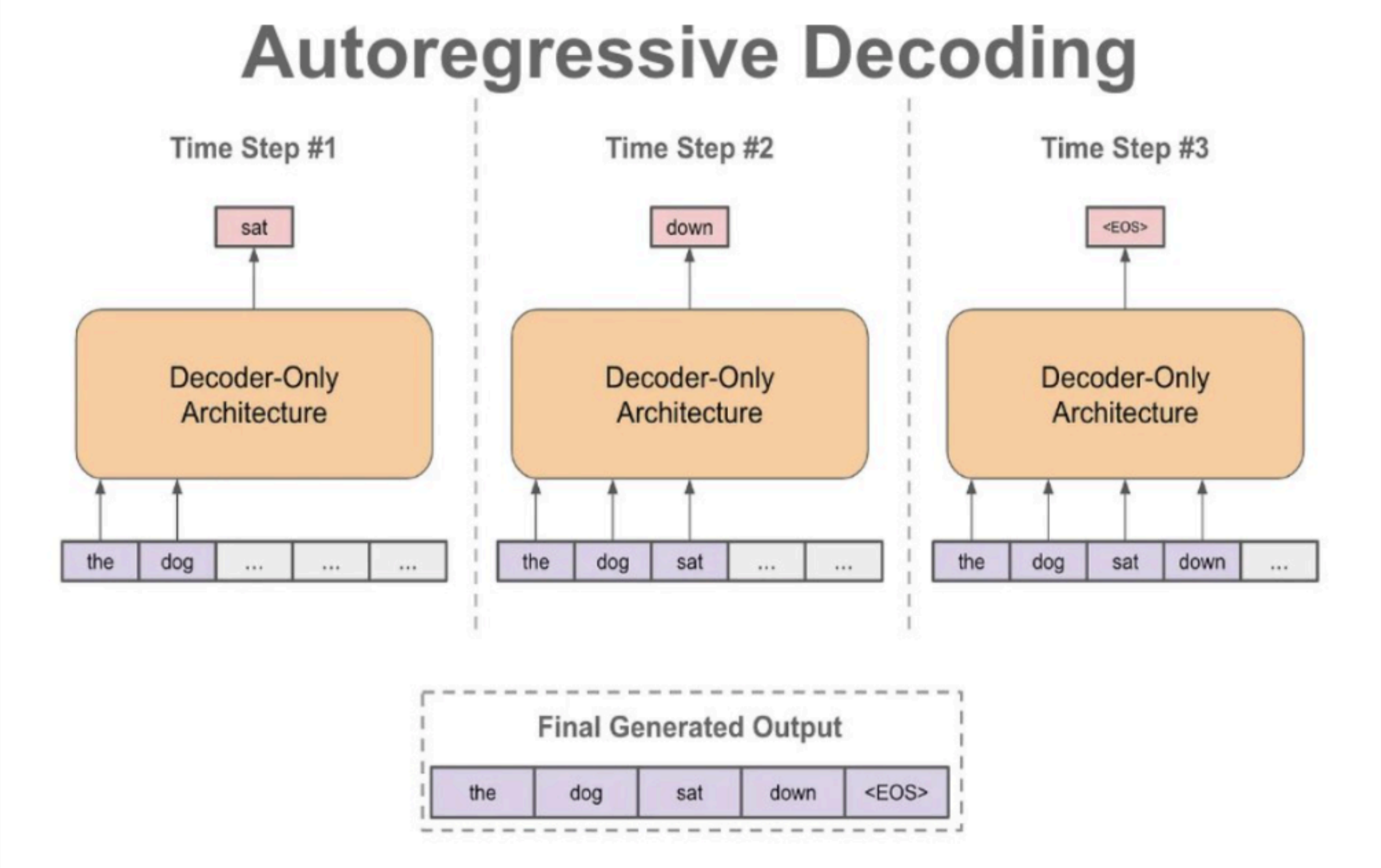
LLM Inference: how is the word after that one generated?



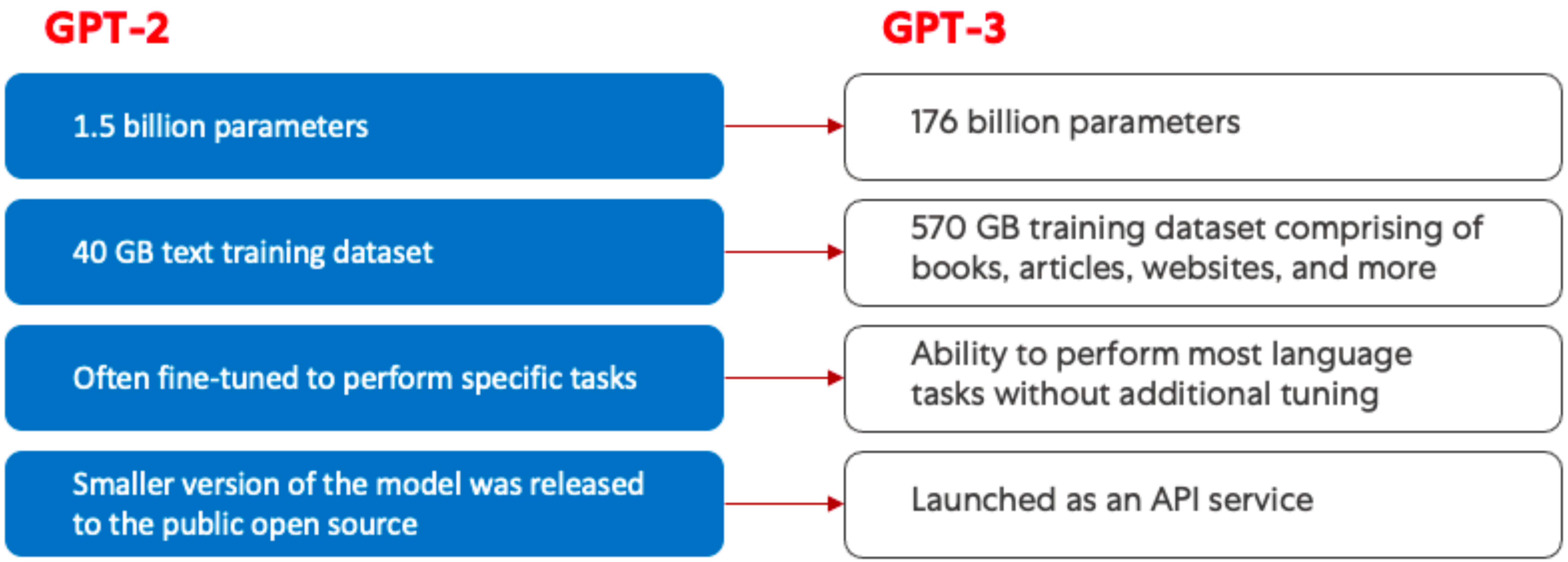
LLM Inference: how is the word after that one generated?



LLM Inference: how is the word after that one generated?



GPT-1,2,3: same loss. different training data and model sizes



GPT-1	GPT-2	GPT-3
117 million parameters	1.5 billion parameters	176 billion parameters
1.2 GB sized training dataset	40 GB text training dataset	570 GB training dataset comprising of books, articles, websites, and more
Supervised finetuning afterwards	Often fine-tuned to perform specific tasks	Ability to perform most language tasks without additional tuning
No release	Smaller version of the model was released to the public open source	Launched as an API service

https://towardsdatascience.com/how-chatgpt-works-the-models-behind-the-bot-1ce5fca96286

Post-Training: Instruction Tuning and Alignment

1 Pre-Training

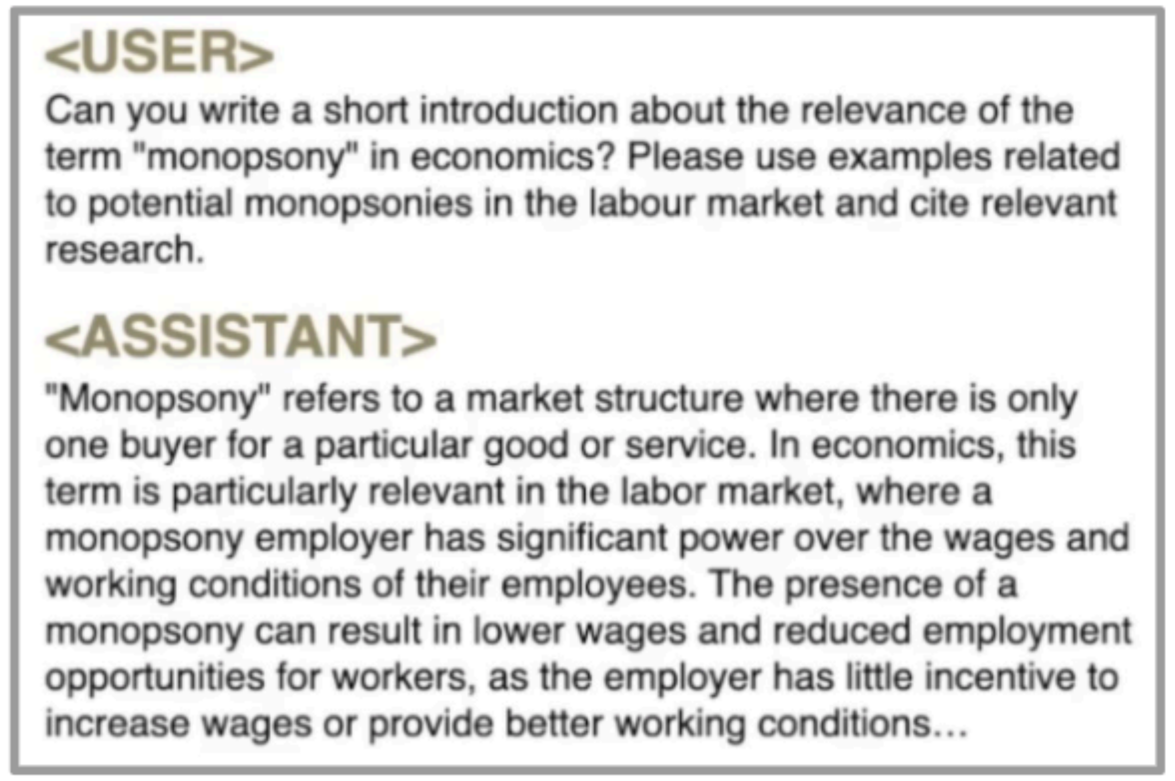
Massive amounts of data from Internet, books, etc.

Problem: A model that can babble on about anything, but not aligned with what we want (e.g. Question-Answering)

SSL

2 Instruction Fine-tuning

Teach model to respond to instructions.



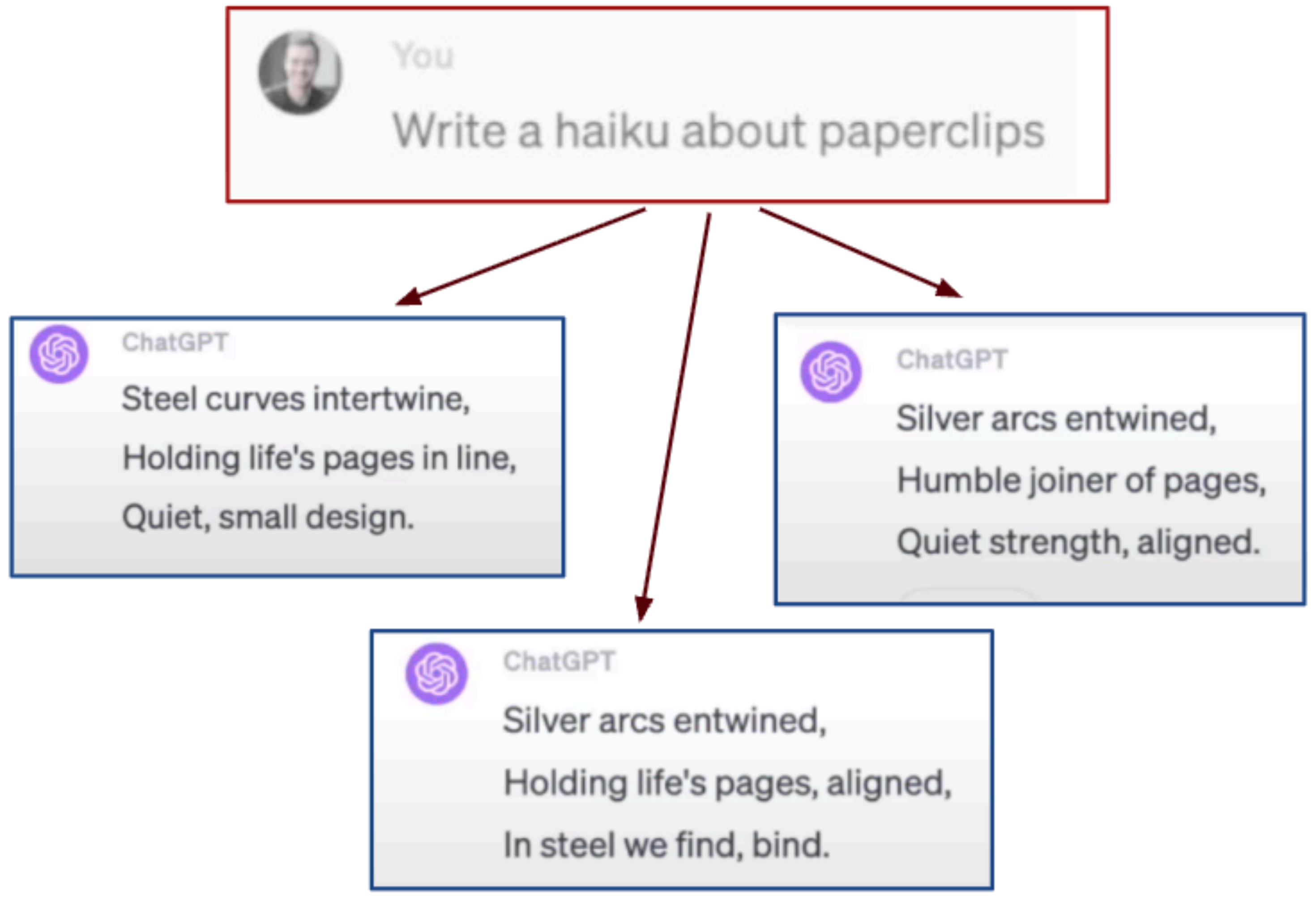
<USER>
Can you write a short introduction about the relevance of the term "monopsony" in economics? Please use examples related to potential monopsonies in the labour market and cite relevant research.

<ASSISTANT>
"Monopsony" refers to a market structure where there is only one buyer for a particular good or service. In economics, this term is particularly relevant in the labor market, where a monopsony employer has significant power over the wages and working conditions of their employees. The presence of a monopsony can result in lower wages and reduced employment opportunities for workers, as the employer has little incentive to increase wages or provide better working conditions…

SL

3 Reinforcement Learning from Human Feedback

Teach model to produce output closer to what humans like.



You
Write a haiku about paperclips

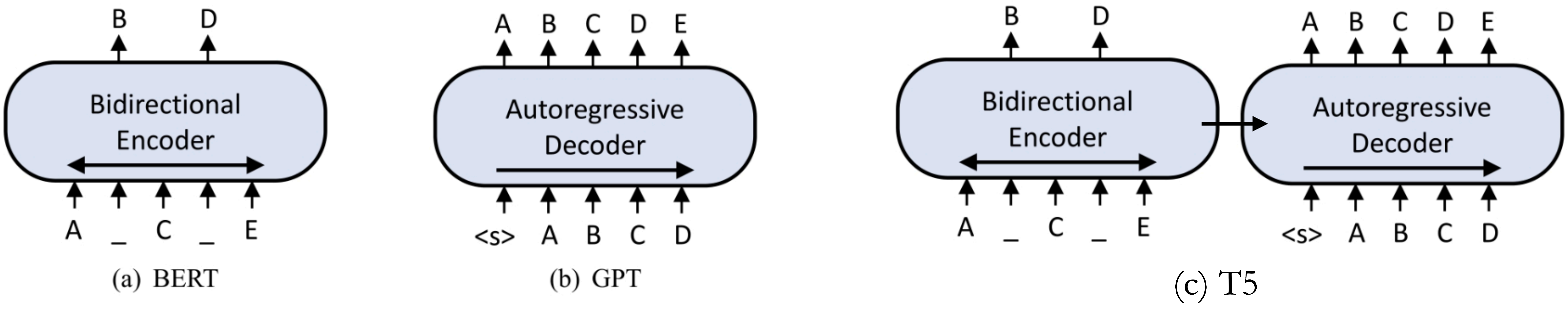
ChatGPT
Steel curves intertwine,
Holding life's pages in line,
Quiet, small design.

ChatGPT
Silver arcs entwined,
Humble joiner of pages,
Quiet strength, aligned.

ChatGPT
Silver arcs entwined,
Holding life's pages, aligned,
In steel we find, bind.

RL

Language model architectures



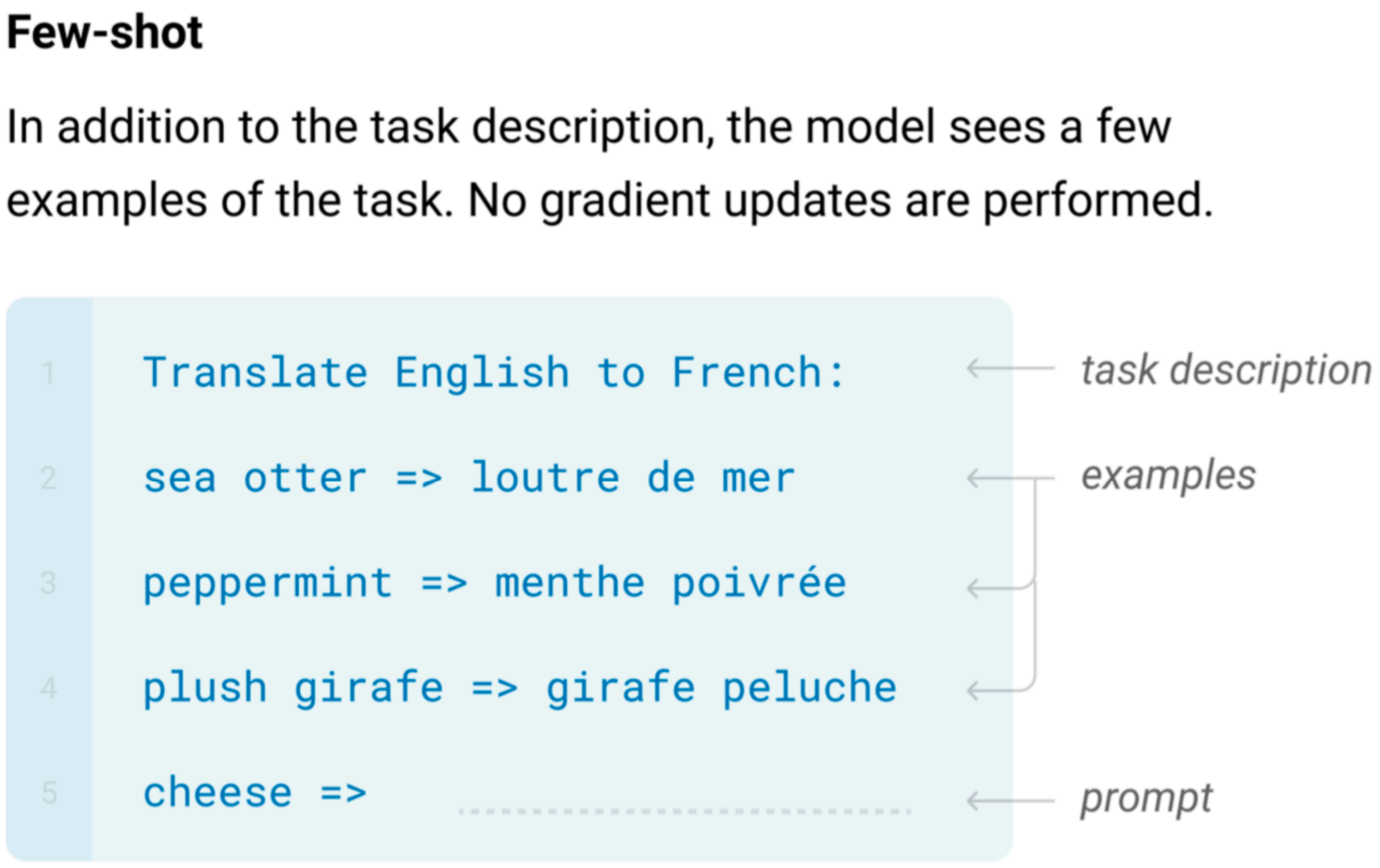
(a) BERT (b) GPT (c) T5

Take 2min to write down (laptop / phone / etc.)
what you’ve learned.
Write down what they mean in your own words.

Share what you’ve written down with your neighbor and discuss.

GPT-3: "Language models are few-shot learners"

more on this later



Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
Translate English to French:
sea otter => loutre de mer
peppermint => menthe poivrée
plush girafe => girafe peluche
cheese =>
```

task description
examples
prompt

One emergent capability of large language models is *in-context learning*.

Here, the "task" is defined within the language model's context, and the model *picks up the task* and *solves it* for the given sample both during a single forward pass

Brown et al. Language models are few-shot learners. NeurIPS 2020.

In-context Learning: benefitting from more examples in the input

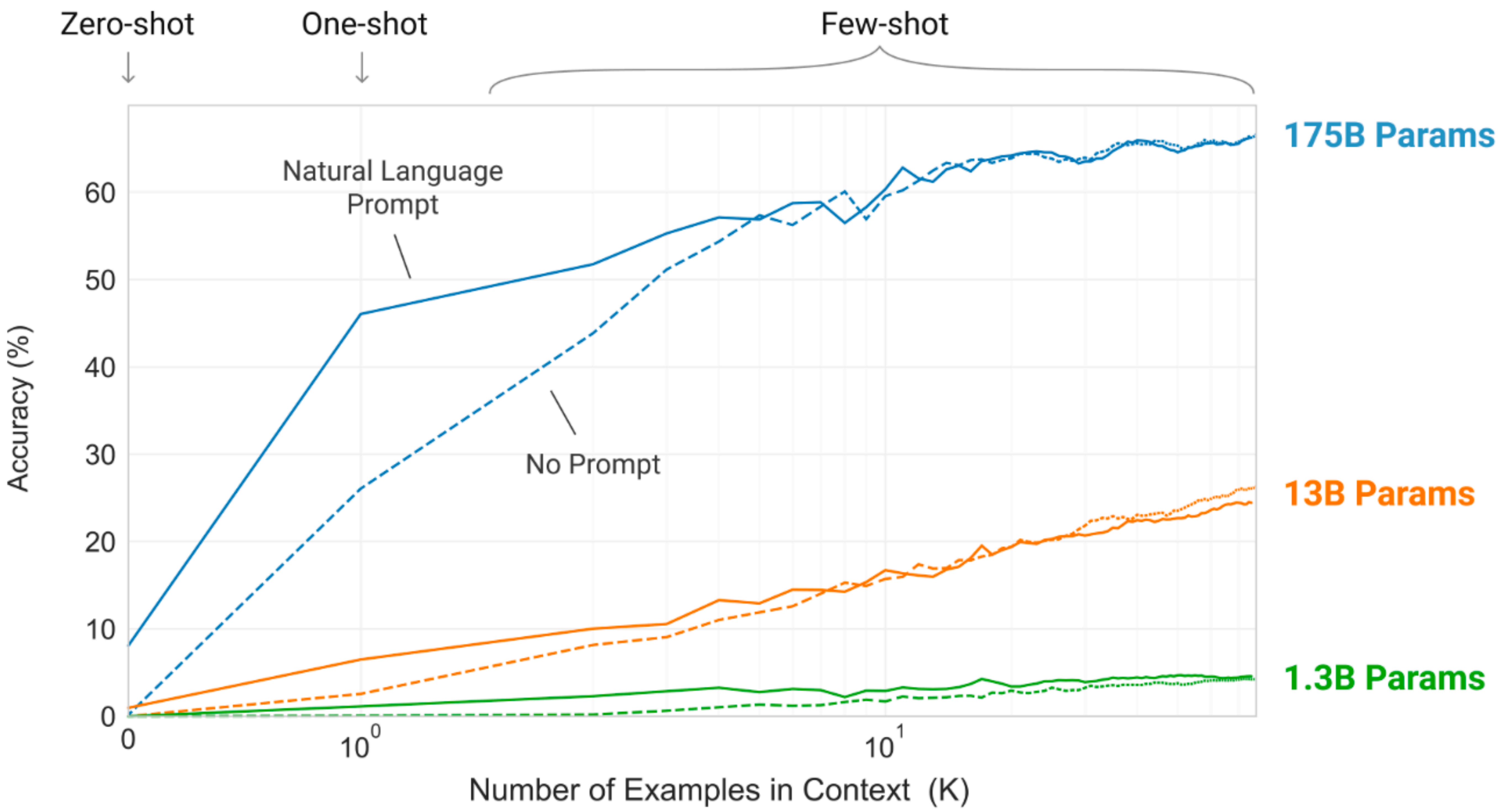


Figure 1.2: Larger models make increasingly efficient use of in-context information. We show in-context learning performance on a simple task requiring the model to remove random symbols from a word, both with and without a natural language task description (see Sec. 3.9.2). The steeper "in-context learning curves" for large models demonstrate improved ability to learn a task from contextual information. We see qualitatively similar behavior across a wide range of tasks.

Brown et al. Language models are few-shot learners. NeurIPS 2020.

Multi-modal Learning

+ captions/ thoughts?

What modalities does Deep Learning (mostly) deal with?

- Generally: anything on the internet
- Images
- Text
- Speech audio
- LiDAR points
- 3D models
- ....

Multiple modalities

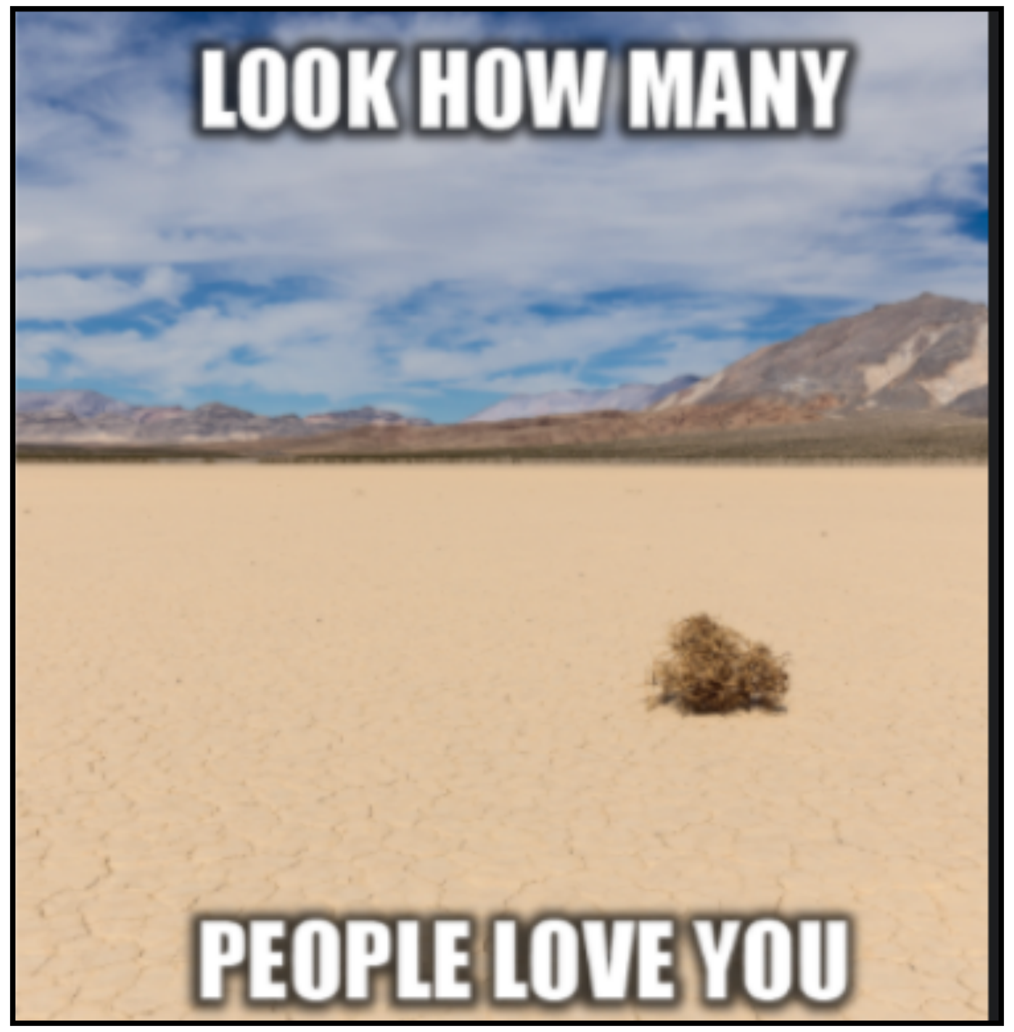
- Videos (RGB frames + audio + audio transcriptions if there's speech)
- Image-text (e.g. images with captions, images with alt text)
- ...

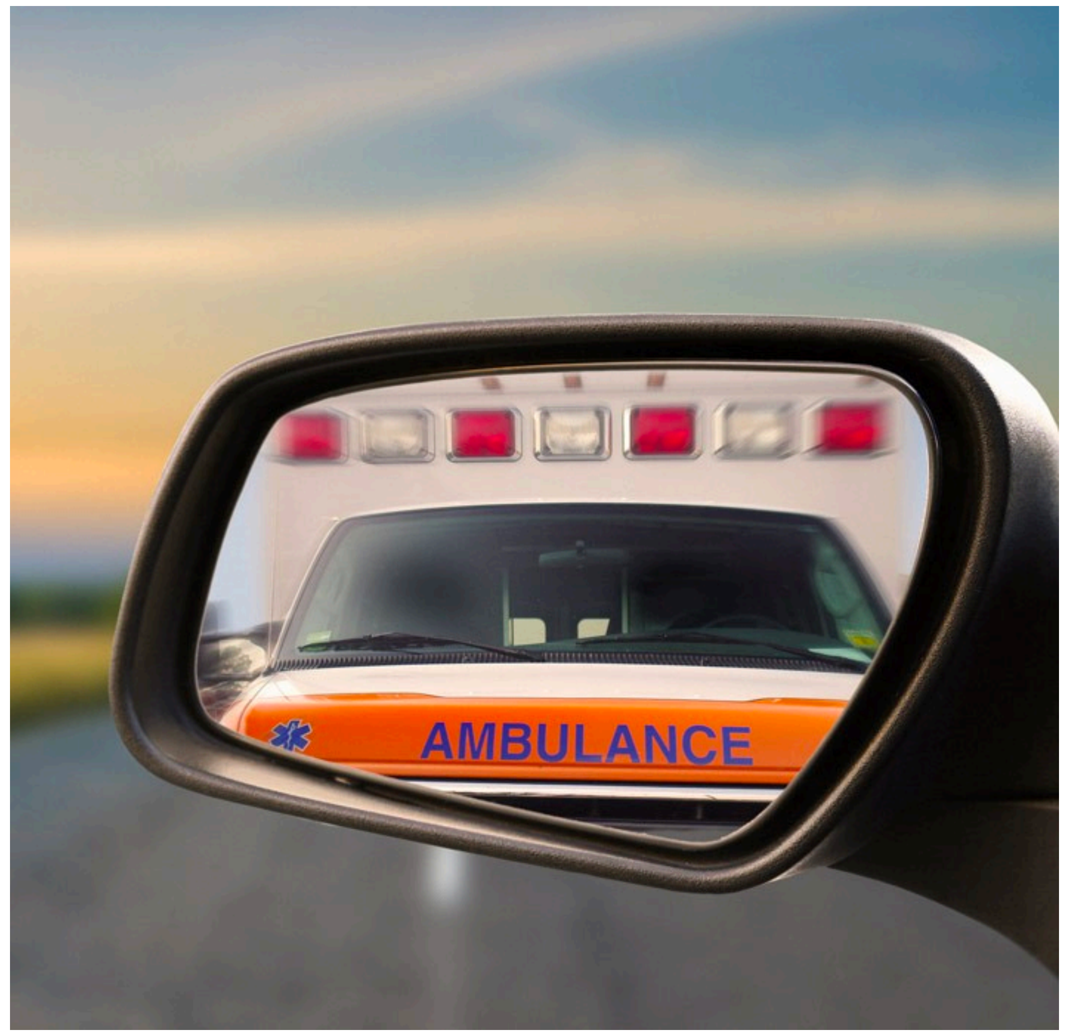
What makes multi-modal learning interesting? e.g. vision-language

Text is like an ”augmentation” / broader description

The man at bat readies to swing at the pitch while the umpire looks on.

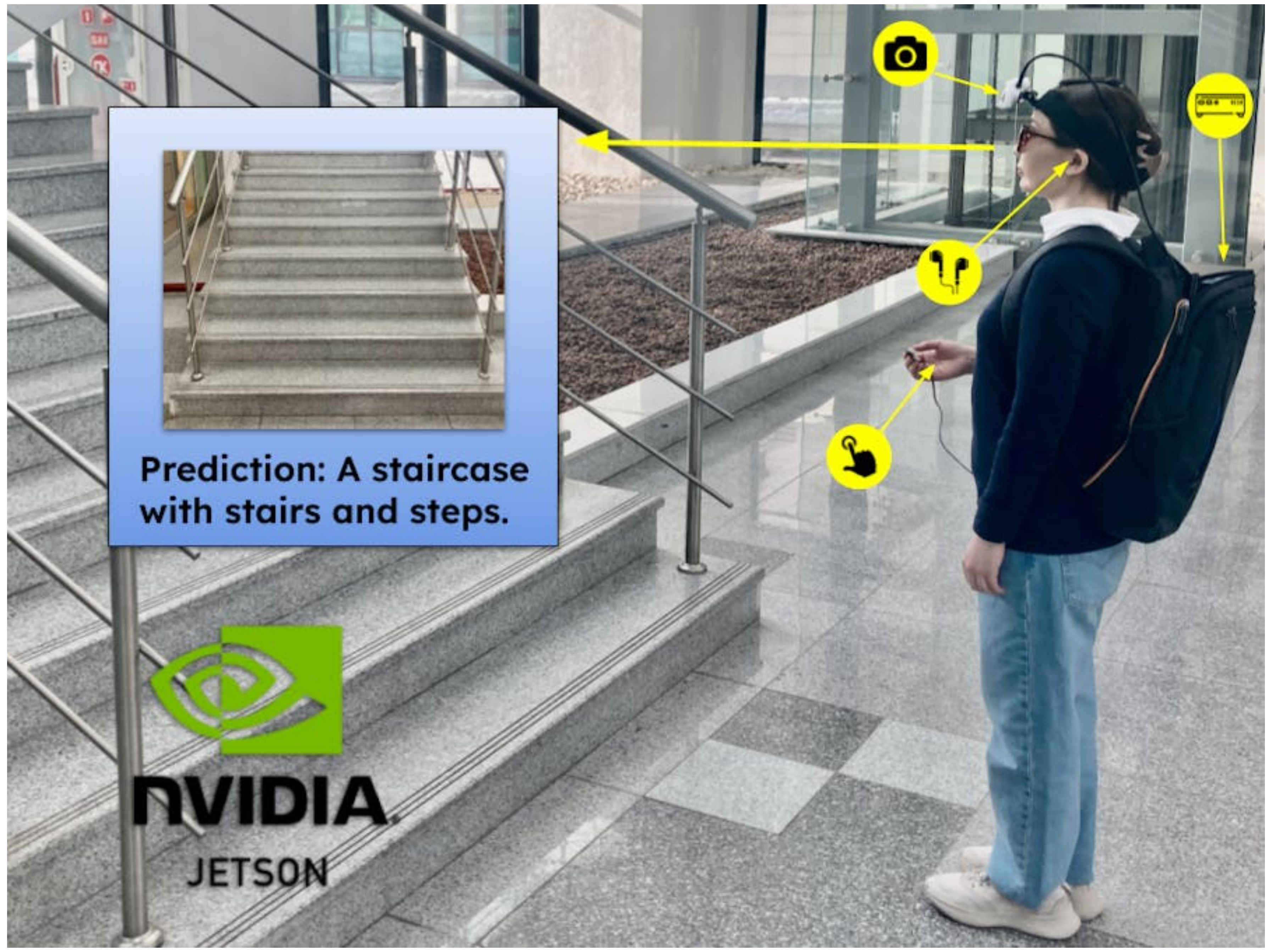
The meaning depends on both modalities (rarer)



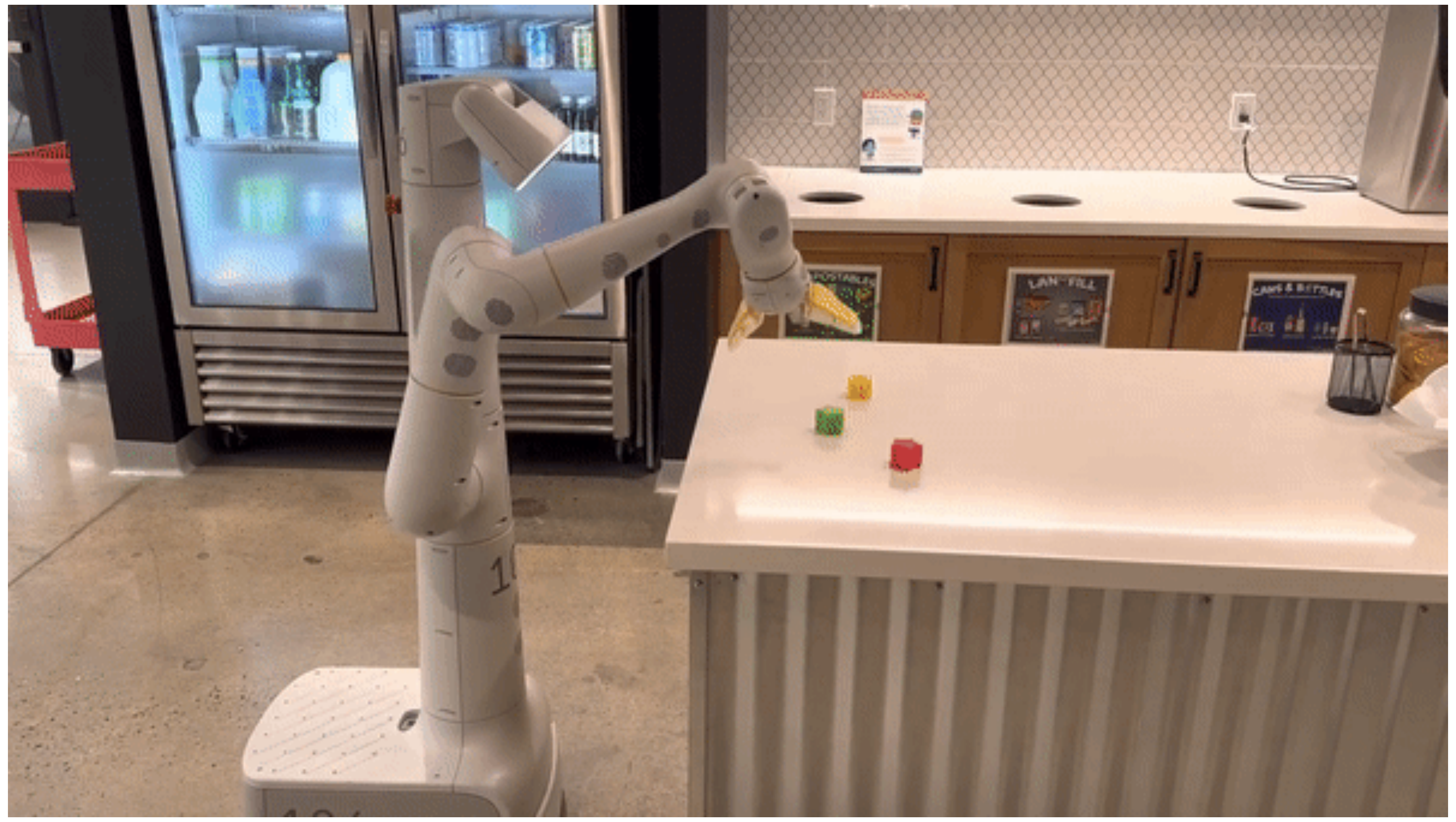


What makes multi-modal learning interesting? e.g. vision-language

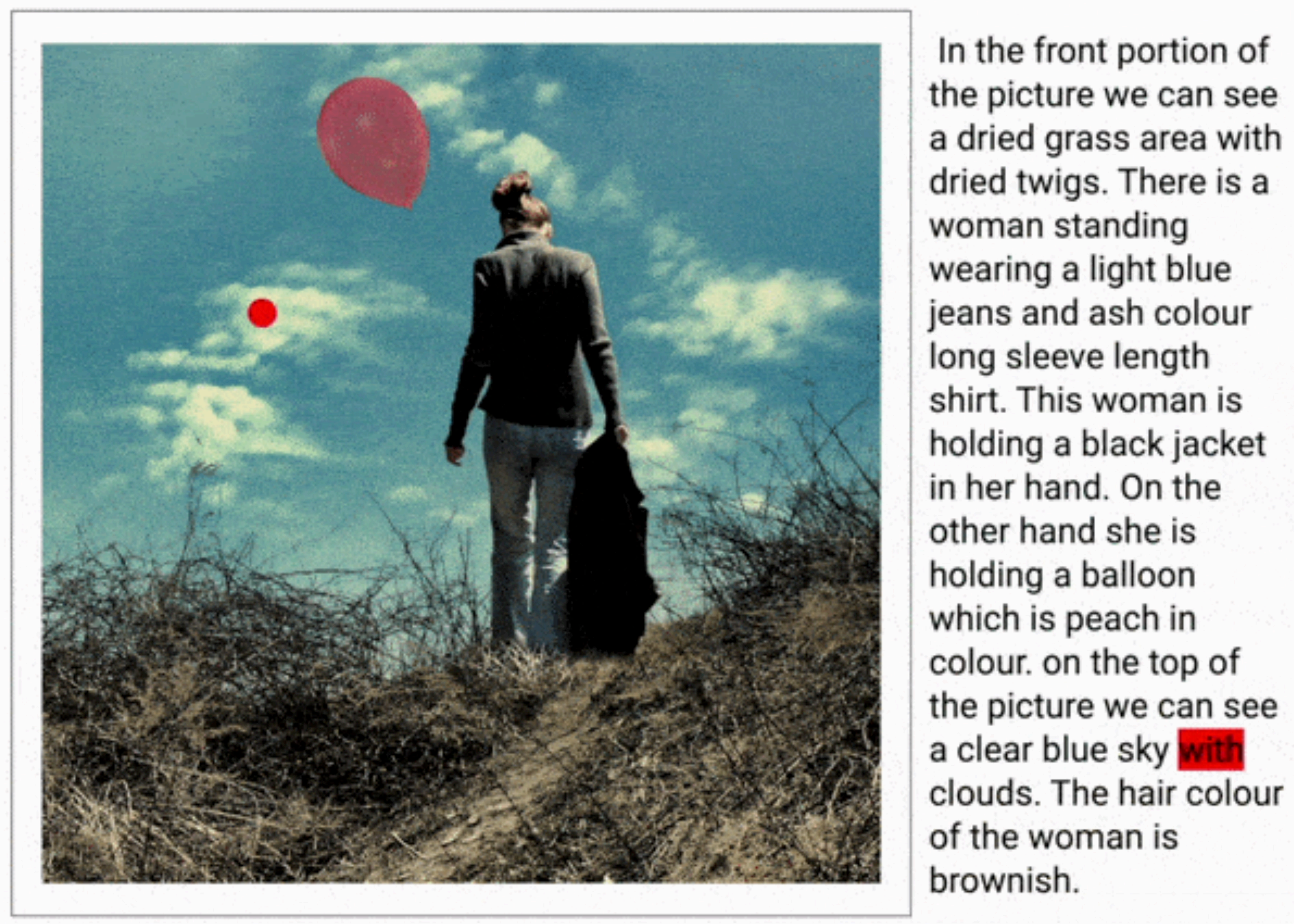
Interfaces: e.g. for visually impaired people



Ground common sense knowledge in real-world



Text can also be very detailed



In the front portion of the picture we can see a dried grass area with dried twigs. There is a woman standing wearing a light blue jeans and ash colour long sleeve length shirt. This woman is holding a black jacket in her hand. On the other hand she is holding a balloon which is peach in colour. on the top of the picture we can see a clear blue sky with clouds. The hair colour of the woman is brownish.

850k images with such descriptions

+audio

+pointer

+(partially): segmentations

Pont-Tuset et al. Connecting Vision and Language with Localized Narratives. ECCV 2020

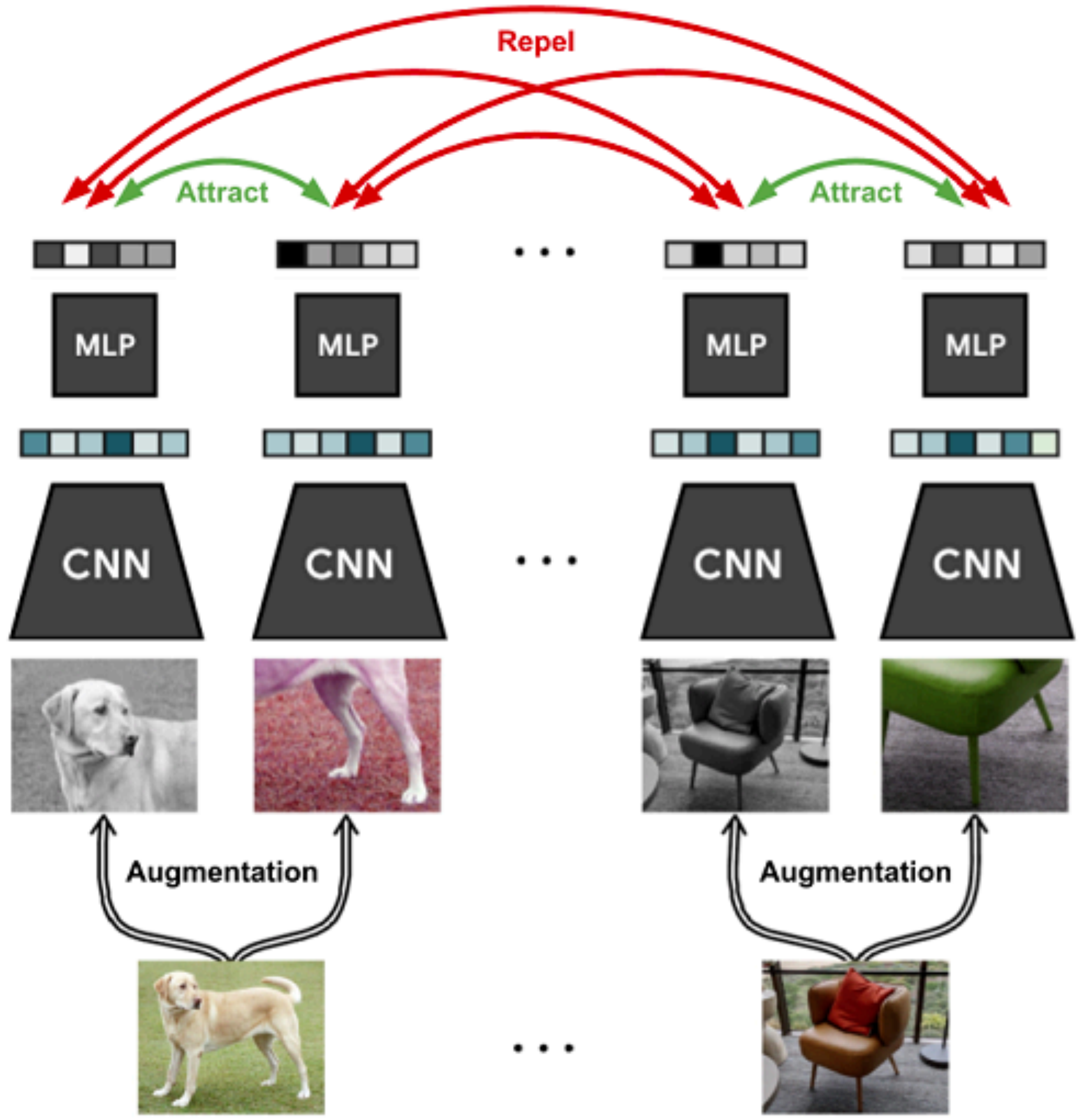
But really: the language part makes it *very* ”generaliseable” or ”general”

Language is an almost universal format for posing and solving tasks

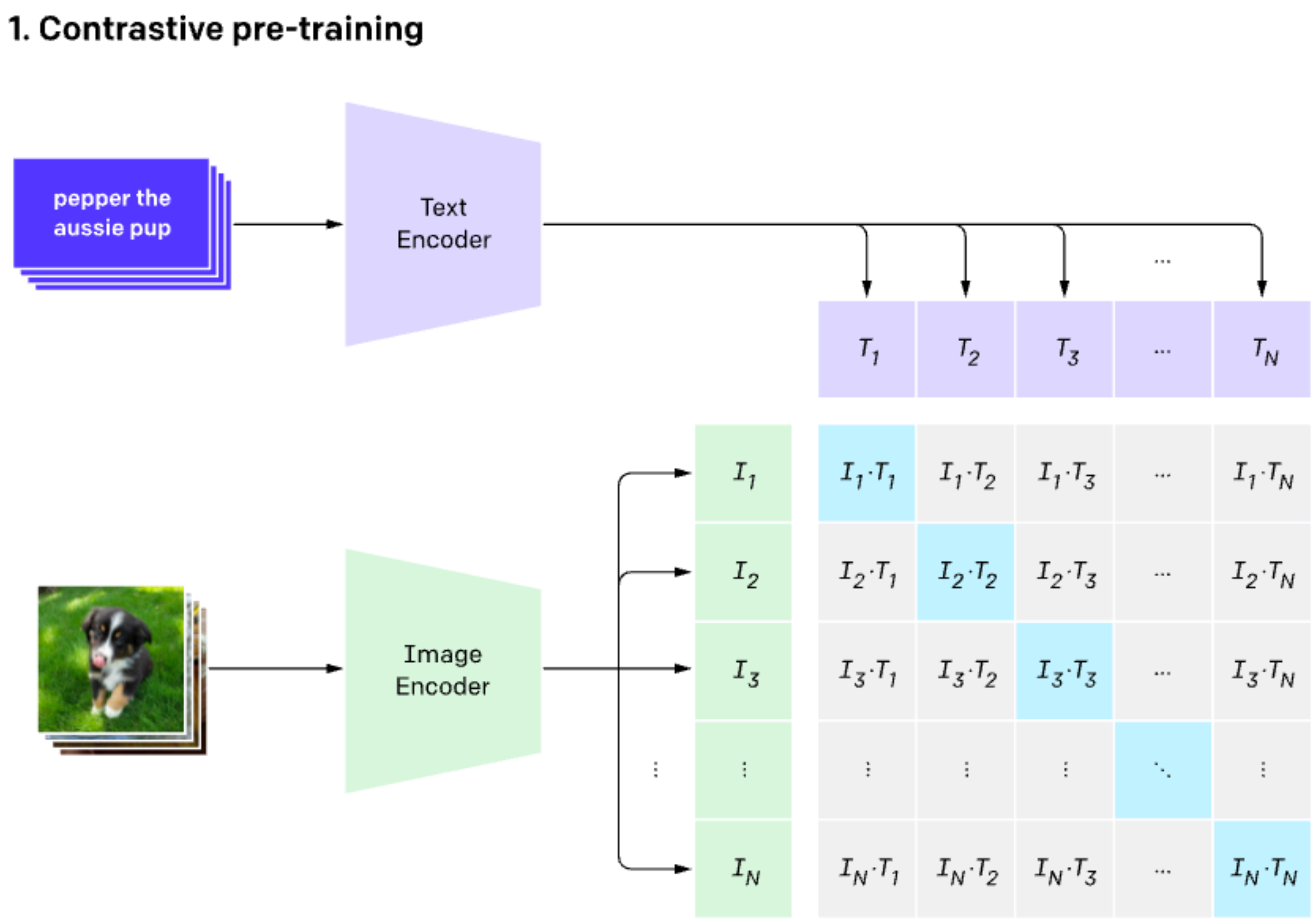
Language further has advantage of being human understandable

Language models are few-shot (in-context) learners

What if we use a caption of an image as its augmentation?



SimCLR



CLIP: instead of augmentation, uses an image caption
(the magic is in the training data)

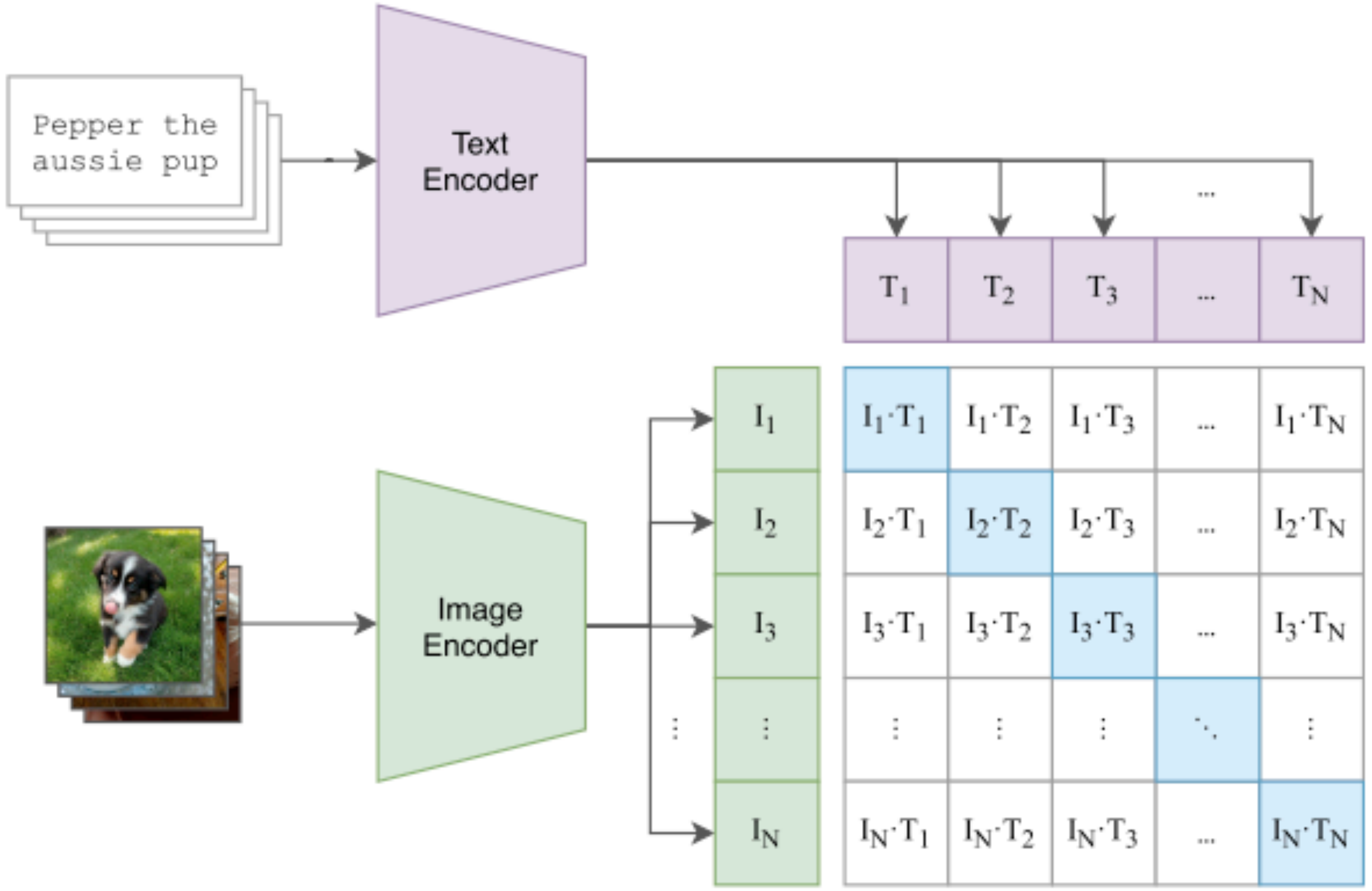
Radford et al. Learning Transferable Visual Models From Natural Language Supervision. ICML 2020
but also: Desai & Johnson VirTex: Learning Visual Representations from Textual Annotations. CVPR 2021 etc.

Multimodal training with CLIP

CLIP or **C**ontrastive **L**anguage-**I**mage **P**re-training[1]

Consists of Image Encoder (CNN/ViT) and Text Encoder (Transformer).

Given a pair (image, caption), CLIP processes each modality with the corresponding encoder – yielding a specific embedding for each.



	T_1	T_2	T_3	...	T_N
I_1	$I_1 \cdot T_1$	$I_1 \cdot T_2$	$I_1 \cdot T_3$	...	$I_1 \cdot T_N$
I_2	$I_2 \cdot T_1$	$I_2 \cdot T_2$	$I_2 \cdot T_3$	...	$I_2 \cdot T_N$
I_3	$I_3 \cdot T_1$	$I_3 \cdot T_2$	$I_3 \cdot T_3$	...	$I_3 \cdot T_N$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
I_N	$I_N \cdot T_1$	$I_N \cdot T_2$	$I_N \cdot T_3$	...	$I_N \cdot T_N$

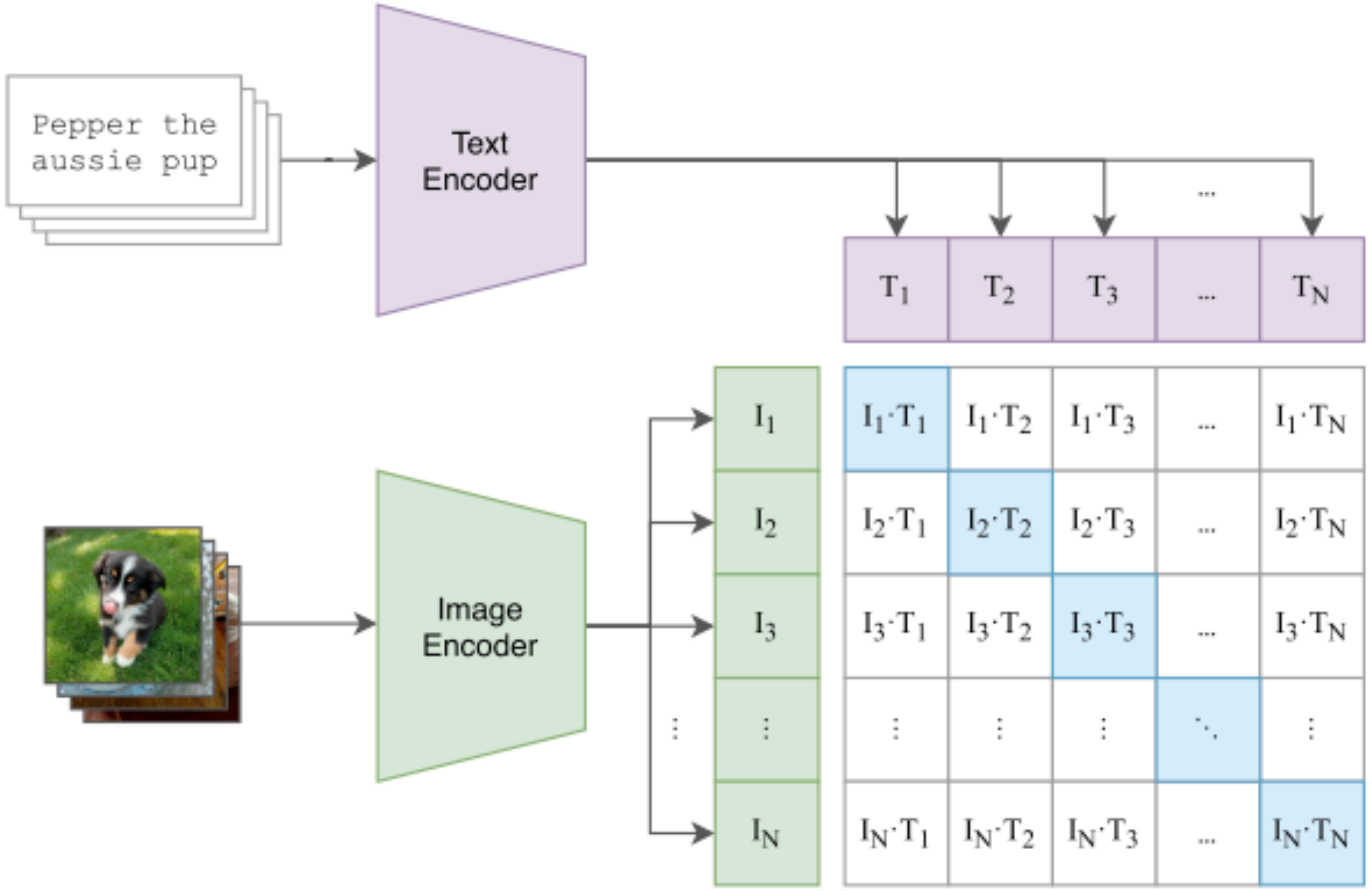
$$I_1 = \text{ImageEncoder}(\text{image}_1);$$
$$I_2 = \text{ImageEncoder}(\text{image}_2)$$
$$\ldots$$
$$T_1 = \text{TextEncoder}(\text{caption}_1);$$
$$T_2 = \text{TextEncoder}(\text{caption}_2)$$
$$\ldots$$

Fundamental AI Lab

Learning Transferable Visual Models From Natural Language Supervision, Radford et al. (2021)

Multimodal training with CLIP

Instead of writing the whole caption, CLIP solves an easier proxy pretraining task of predicting which text as a whole, is paired with which image.


Pepper the aussie pup
Text Encoder
Image Encoder
T_1 T_2 T_3 ... T_N
I_1 I_2 I_3 ⋮ I_N
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ ... $I_1 \cdot T_N$
$I_2 \cdot T_1$ $I_2 \cdot T_2$ $I_2 \cdot T_3$ ... $I_2 \cdot T_N$
$I_3 \cdot T_1$ $I_3 \cdot T_2$ $I_3 \cdot T_3$ ... $I_3 \cdot T_N$
$I_N \cdot T_1$ $I_N \cdot T_2$ $I_N \cdot T_3$ ... $I_N \cdot T_N$

- **Maximize** the cosine similarity of the image and text embeddings of true pairs ($I_1 * T_1$)
- **Minimize** the cosine similarity of the embeddings of incorrect pairs ($I_1 * T_2$)

- Formally: CLIP optimizes a noise-contrastive cross-modal loss.

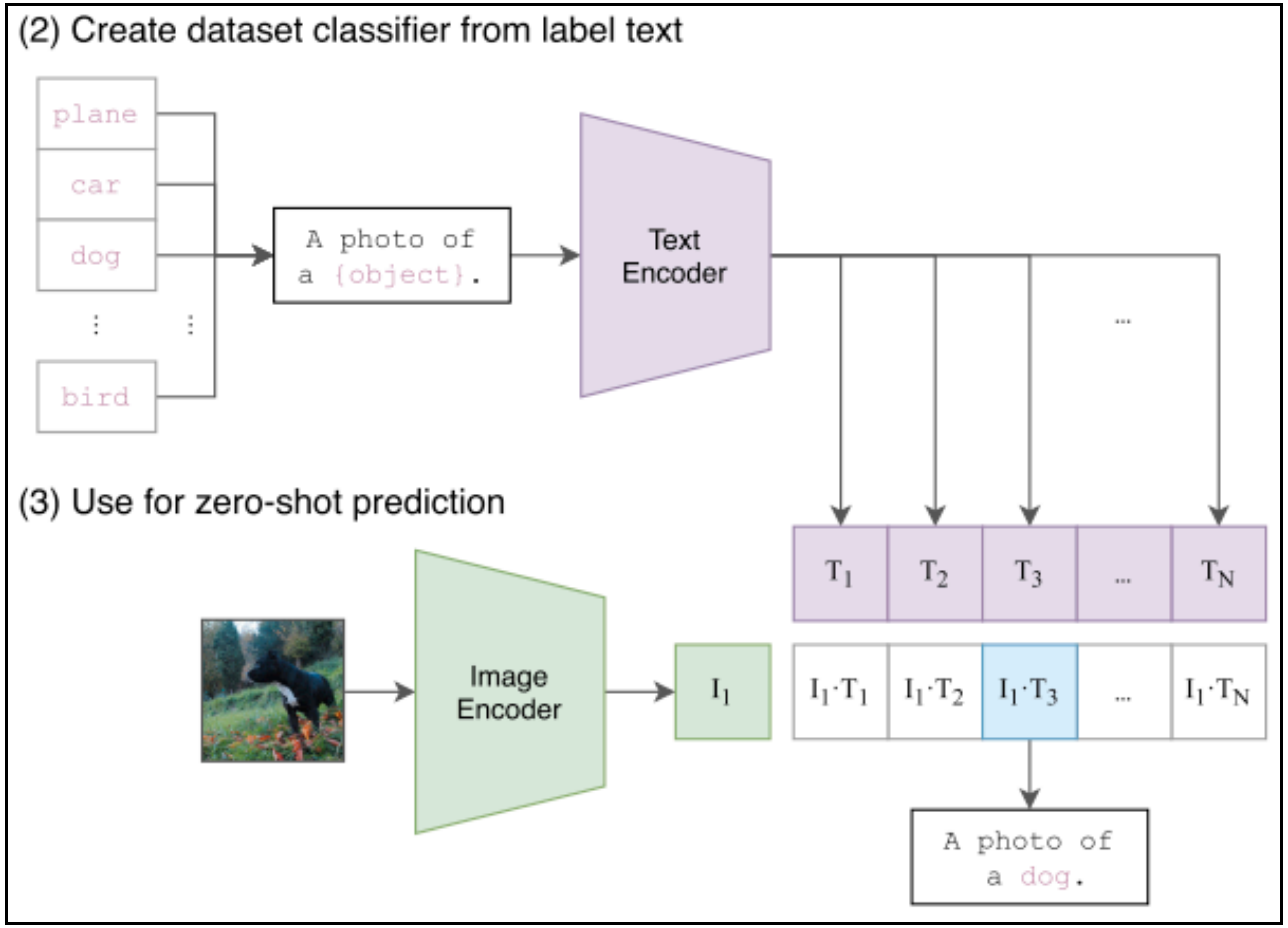
Zero-shot open-vocabulary classification

At inference time: CLIP shows zero-shot classification abilities

Predicting labels which were never observed during training

- First: compute the feature embedding of the image: I1 and the feature embedding of all possible texts T1, T2, T3 ...
- Then: compute the cosine similarity of these embeddings, normalized into a probability distribution via a softmax.

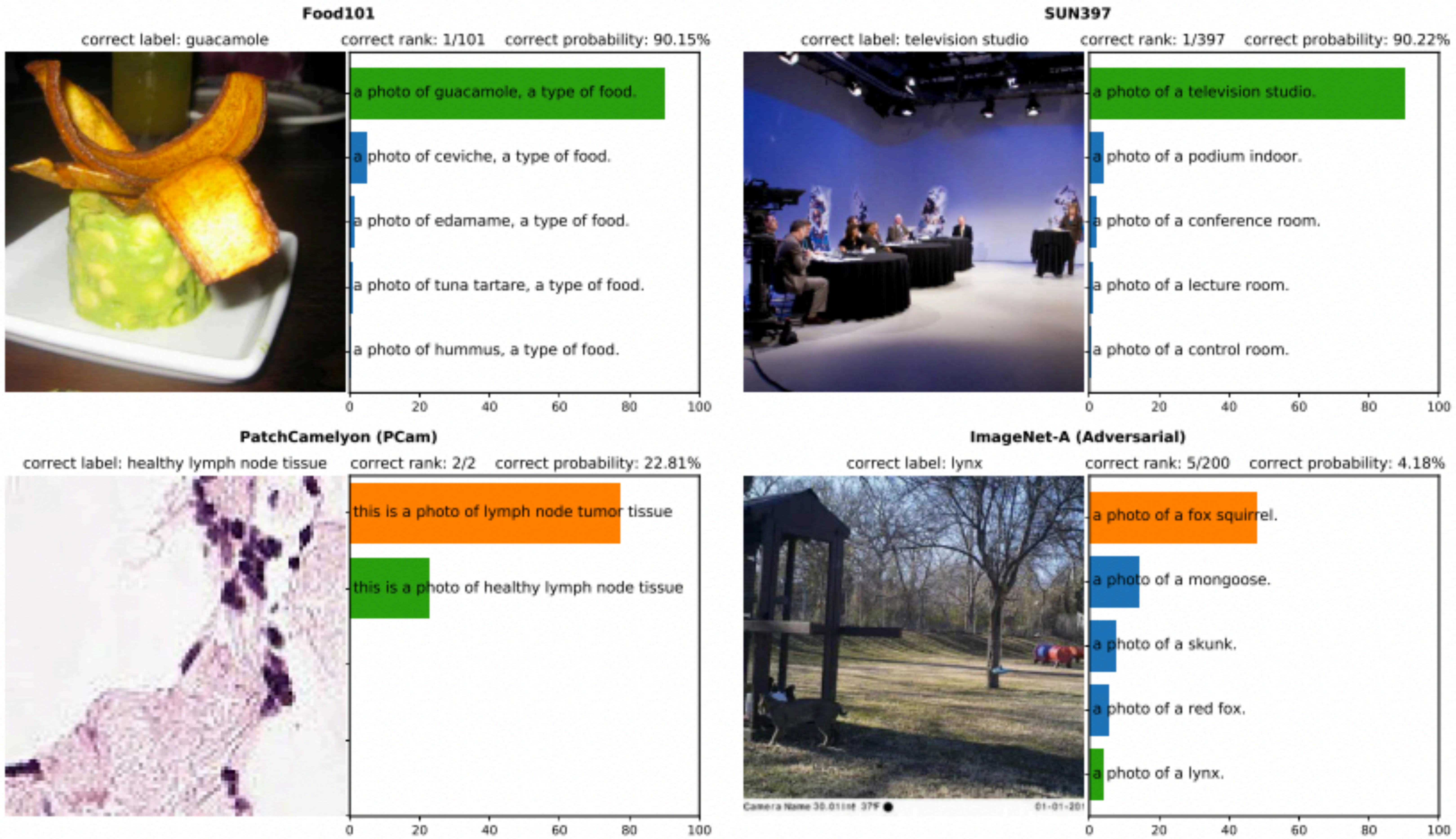
- This gives the most probable (image, text) pair, hence the predicted class.



CLIP: Zero-Shot Examples

Visualization of predictions from CLIP zero-shot classifiers

- The predicted probability of the top 5 classes is shown along with the text used to represent the class.



PS: zero-shot image-to-text not new: e.g. in 2016:

This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the version available on IEEE Xplore.

Learning Visual N-Grams from Web Data

Ang Li* Allan Jabri Armand Joulin Laurens van der Maaten
University of Maryland
College Park, MD 20742, USA
`angli@umiacs.umd.edu`
Facebook AI Research
770 Broadway, New York, NY 10025, USA
`{ajabri,ajoulin,lvdmaaten}@fb.com`

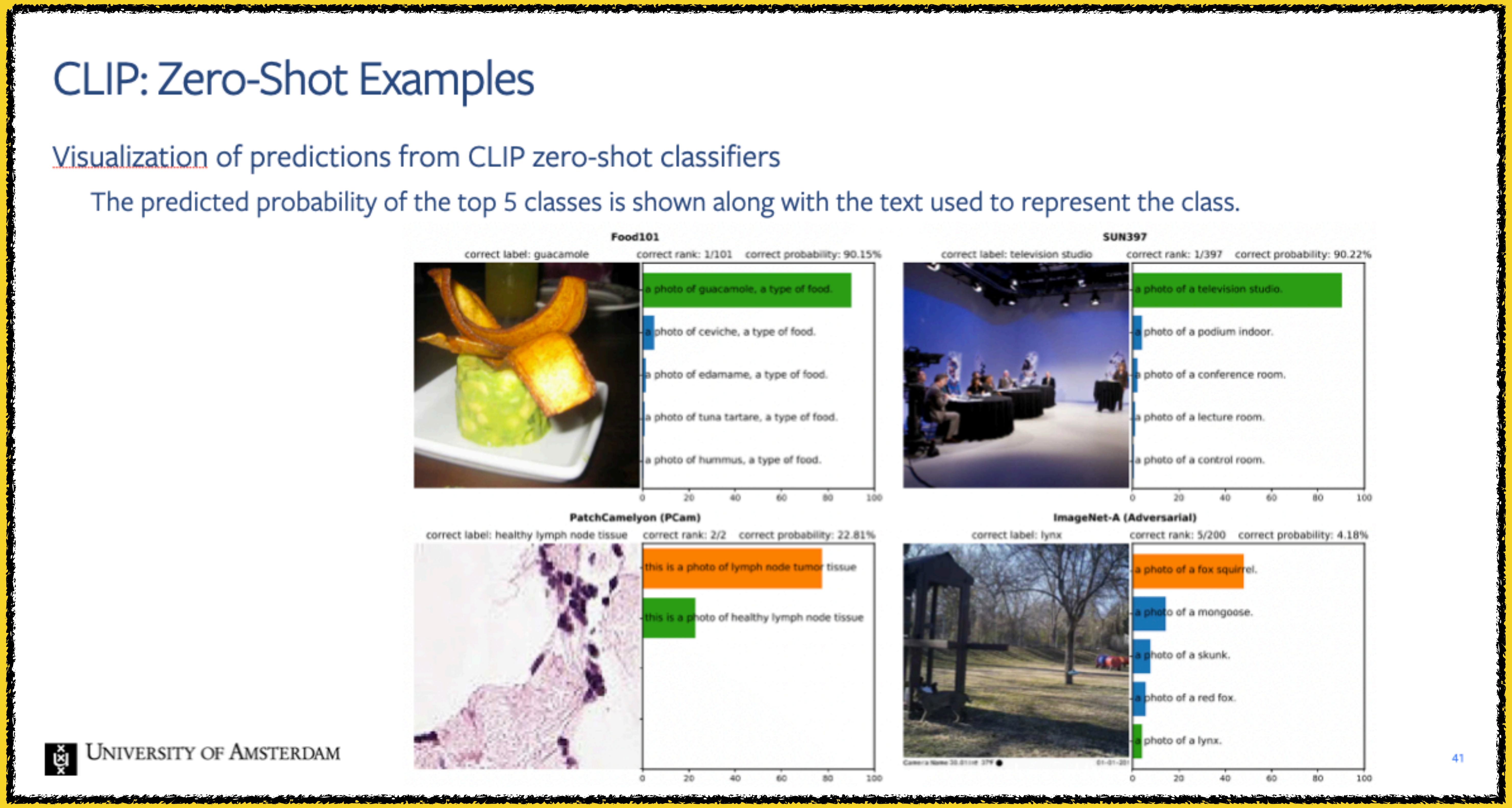
Abstract

Real-world image recognition systems need to recognize tens of thousands of classes that constitute a plethora of visual concepts. The traditional approach of annotating thousands of images per class for training is infeasible in such a scenario, prompting the use of webly supervised data. This paper explores the training of image-recognition systems on large numbers of images and associated user comments, without *using manually labeled images. In particular, we develop visual n-gram models that can predict arbitrary phrases that are relevant to the content of an image. Our visual n-gram models are feed-forward convolutional networks trained using new loss functions that are inspired by n-gram models commonly used in language modeling. We demonstrate the merits of our models in phrase prediction, phrase-based image retrieval, relating images and captions, and zero-shot transfer.*

Predicted n-grams
lights
Burning Man
Mardi Gras
parade in progress

Predicted n-grams
GP
Silverstone Classic
Formula 1
race for the

Predicted n-grams
navy yard
construction on the
Port of San Diego
cargo


Food101
correct label: guacamole
correct rank: 1/101 correct probability: 90.15%
a photo of guacamole, a type of food.
a photo of ceviche, a type of food.
a photo of edamame, a type of food.
a photo of tuna tartare, a type of food.
a photo of hummus, a type of food.
SUN397
correct label: television studio
correct rank: 1/397 correct probability: 90.22%
a photo of a television studio.
a photo of a podium indoor.
a photo of a conference room.
a photo of a lecture room.
a photo of a control room.
PatchCamelyon (PCam)
correct label: healthy lymph node tissue
correct rank: 2/2 correct probability: 22.81%
this is a photo of lymph node tumor tissue
this is a photo of healthy lymph node tissue
ImageNet-A (Adversarial)
correct label: lynx
correct rank: 5/200 correct probability: 4.18%
a photo of a fox squirrel.
a photo of a mongoose.
a photo of a skunk.
a photo of a red fox.
a photo of a lynx.
0
20
40
60
80
100

CLIP: Zero-Shot Examples

Visualization of predictions from CLIP zero-shot classifiers

- The predicted probability of the top 5 classes is shown along with the text used to represent the class.

UNIVERSITY OF AMSTERDAM

41

Quick recap: Why is this use of CLIP called "zero-shot" classification?

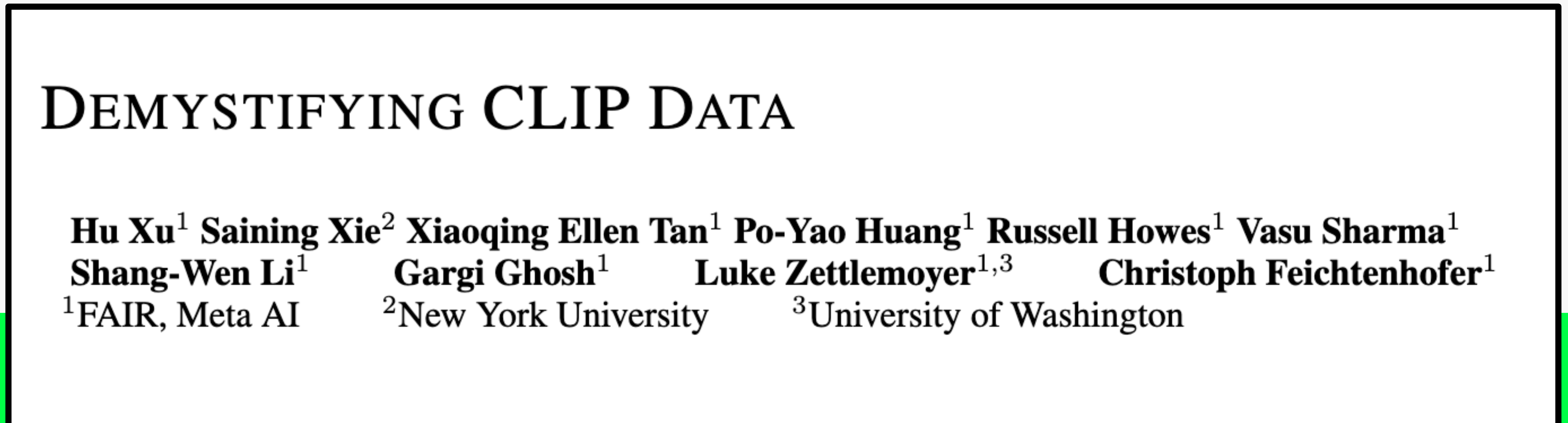
1) Because it does not require new training
2) Because these categories are new to the model
3) Because it requires only one forward pass per images
4) People mostly shouldn't call it zero-shot!

Appendix p.14, Table 11:

	ImageNet	Food-101	CIFAR10	CIFAR100	CUB	SUN397	Cars	Aircraft	DTD	Pets	Caltech-101	Flowers	MNIST	FER-2013	STL-10	EuroSAT	RESISC45	GTSRB	KITTI	Country211	PCAM	UCF101	Kinetics700	CLEVR	HatefulMemes	SST2
MetaCLIP (400M) ViT-L	76.2	90.7	95.5	77.4	75.9	70.5	84.7	40.4	62.0	93.7	94.4	76.4	61.7	46.5	99.3	59.7	71.9	47.5	29.9	30.9	70.1	75.5	57.1	35.1	56.6	65.6
# of cls. w/ non-zero counts	703/998	52/101	10/10	93/100	1/200	193/397	0/196	8/100	40/47	15/37	86/102	61/102	10/10	12/12	10/10	2/10	32/45	1/43	0/4	190/211	1/2	5/101	122/700	8/8	1/2	2/2

Table 11: Measuring task-alignment. First row: MetaCLIP (400M) ViT-L/14 accuracy, second row: number of classes matched in metadata

"Interestingly, there seems to be a correlation with the accuracy and the number of classes matched in the metadata."



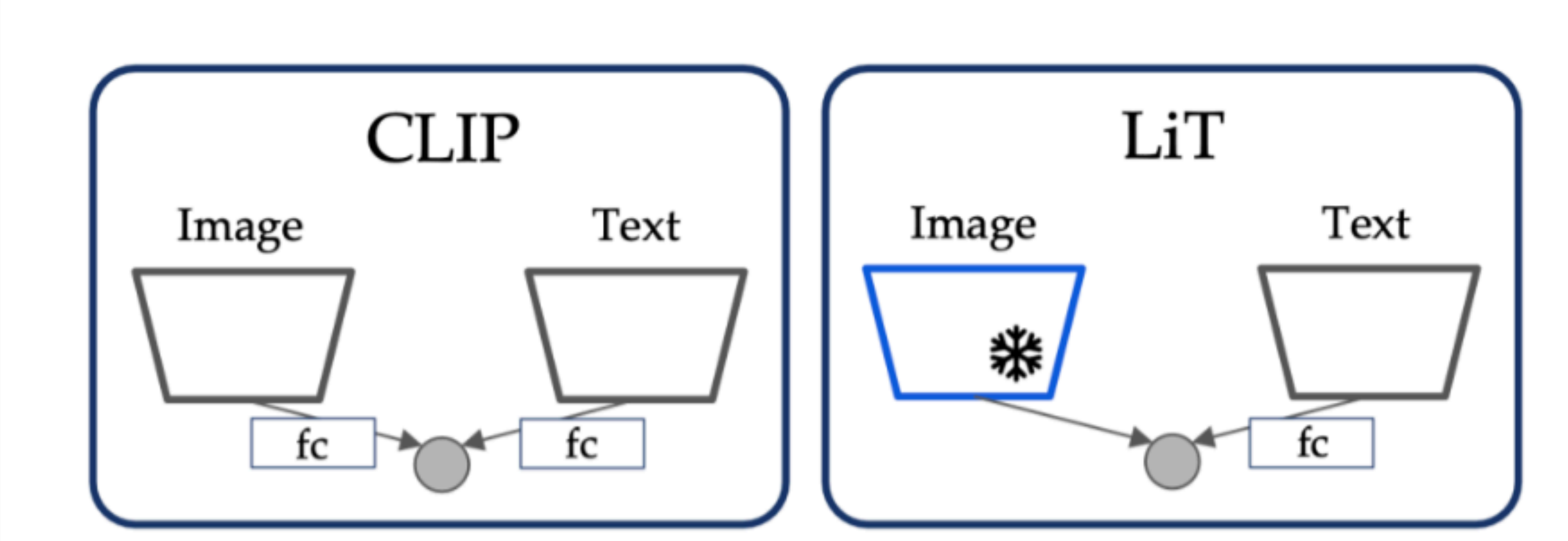
DEMYSTIFYING CLIP DATA

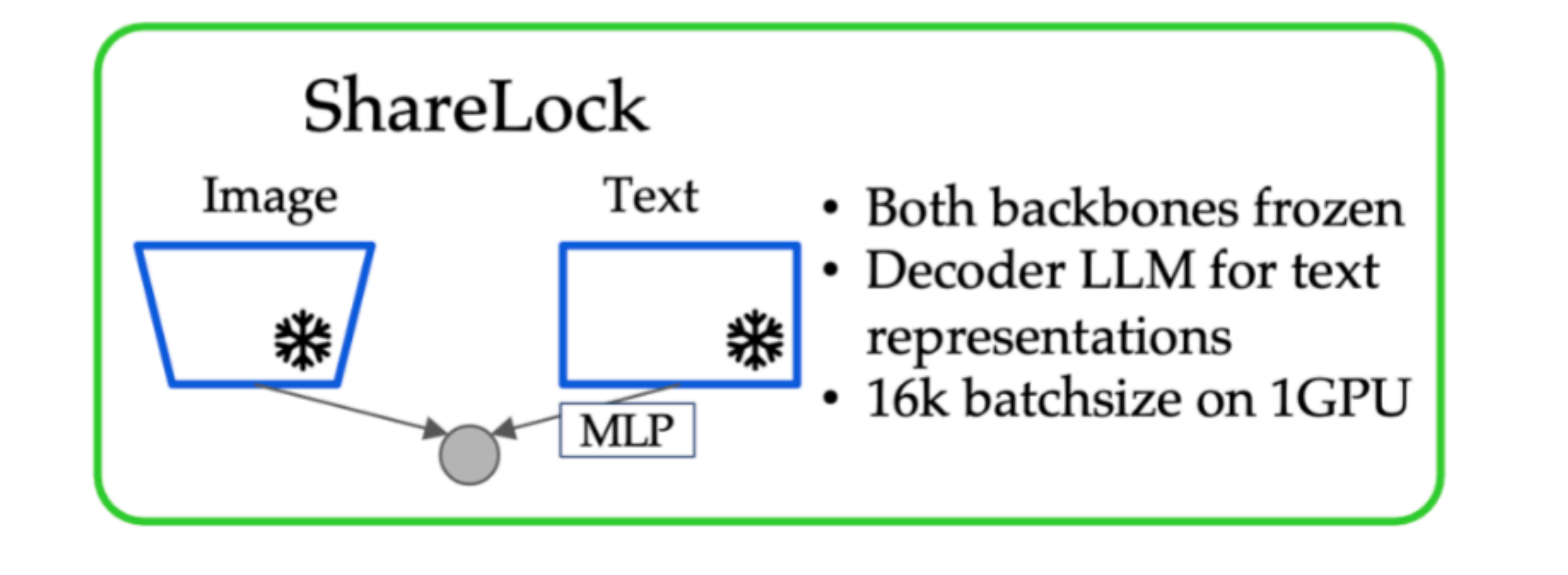
Hu Xu[1] Saining Xie[2] Xiaoqing Ellen Tan[1] Po-Yao Huang[1] Russell Howes[1] Vasu Sharma[1]
Shang-Wen Li[1] Gargi Ghosh[1] Luke Zettlemoyer[1,3] Christoph Feichtenhofer[1]
[1]FAIR, Meta AI [2]New York University [3]University of Washington

CLIP, for the most part, is evaluated within-domain (it's just a big domain)

But surely language features, e.g. from pretrained models should help generalise?

New method: **Shared** Vision-Language-**Locked** Tuning



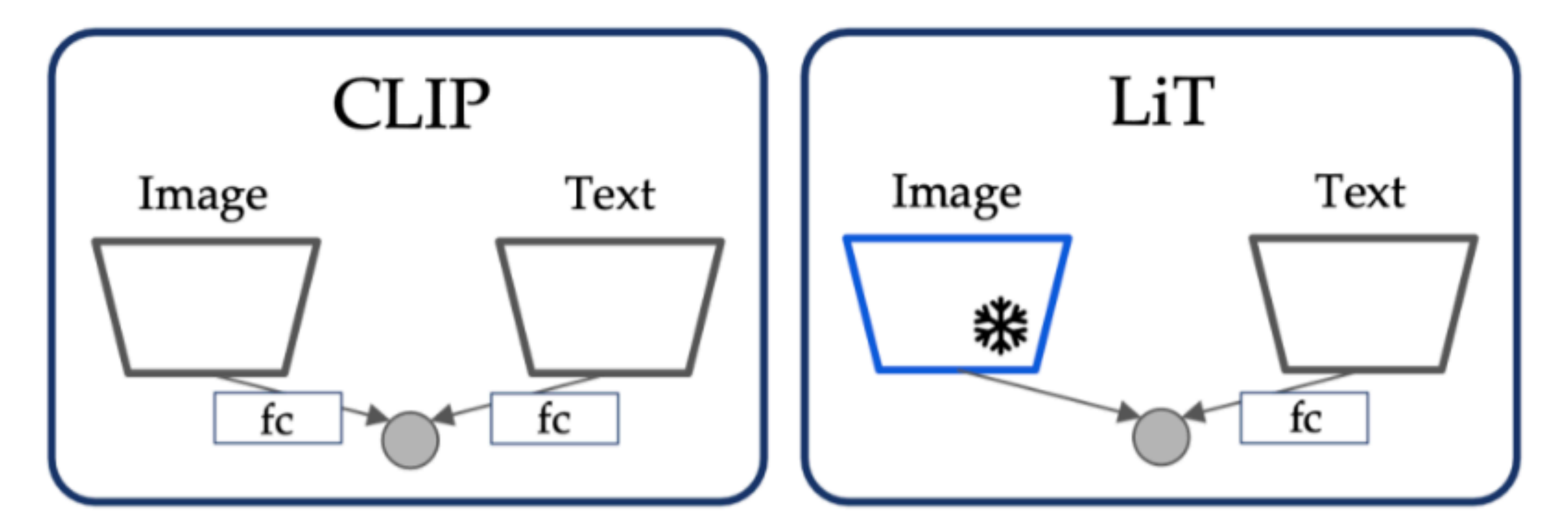


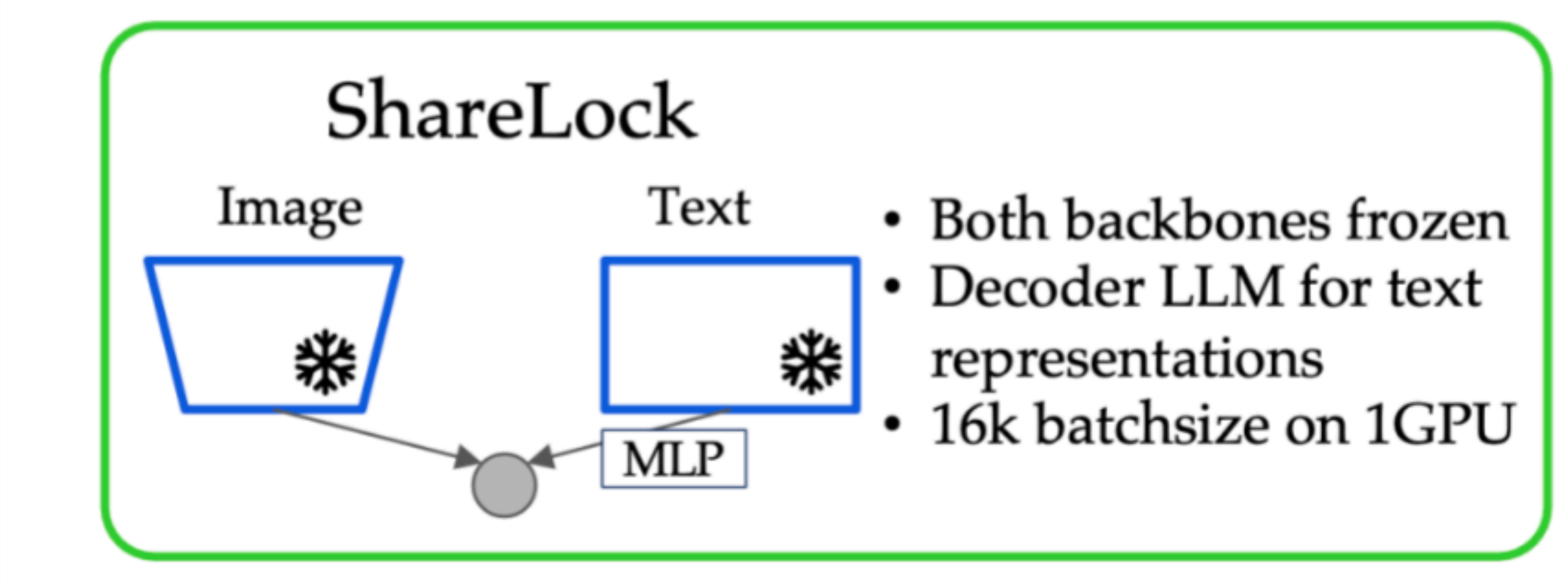
- Both backbones frozen
- Decoder LLM for text representations
- 16k batchsize on 1GPU

Result: CLIP-style model with that only mostly takes frozen representations

Visual Alignment in Text-Only LLMs: New Frontiers in Data Efficiency. ArXiv
Jona Ruthardt, Gertjan J. Burghouts, Serge Belongie, Yuki M. Asano

New method: **Shared** Vision-Language-**Locked** Tuning

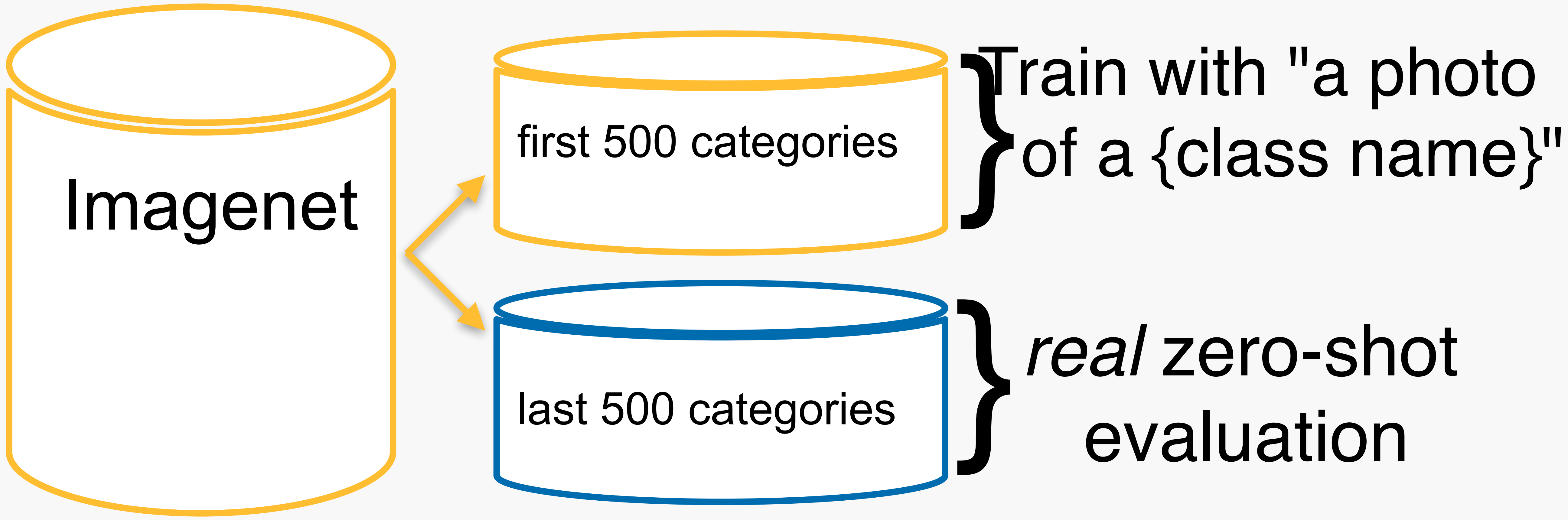




- Both backbones frozen
- Decoder LLM for text representations
- 16k batchsize on 1GPU

Result: CLIP-style model with that only mostly takes frozen representations

New evaluation: Mutually exclusive vision-language dataset splits



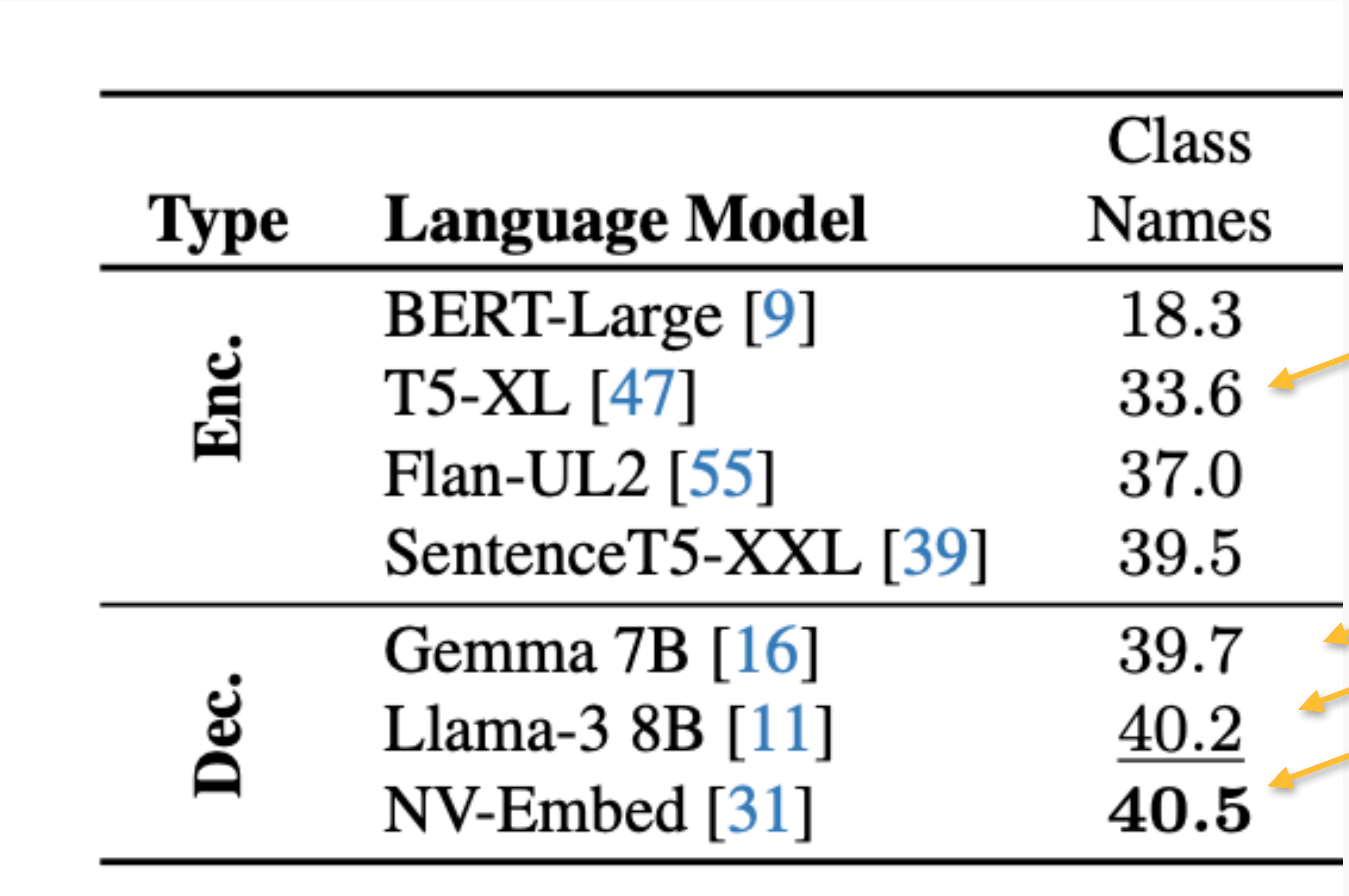
Train with "a photo of a {class name}"

real zero-shot evaluation

Result: Clean measurement of *generalisation ability* from LLM

Visual Alignment in Text-Only LLMs: New Frontiers in Data Efficiency. ArXiv
Jona Ruthardt, Gertjan J. Burghouts, Serge Belongie, Yuki M. Asano

Decoder representations are actually really good.



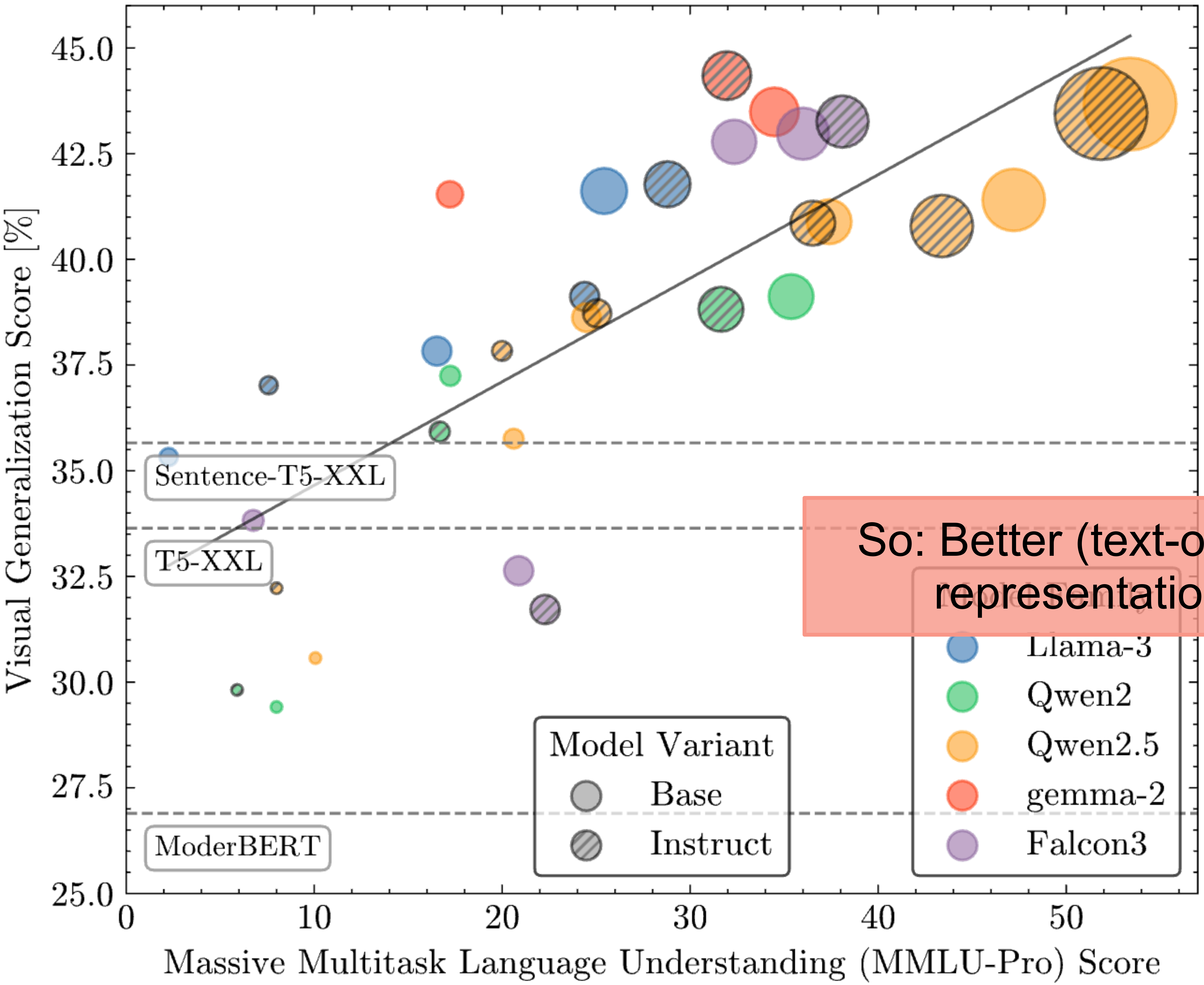
Type	Language Model	Class Names
Enc.	BERT-Large [9]	18.3
	T5-XL [47]	33.6
	Flan-UL2 [55]	37.0
	SentenceT5-XXL [39]	39.5
Dec.	Gemma 7B [16]	39.7
	Llama-3 8B [11]	40.2
	NV-Embed [31]	**40.5**

What people previously used

New billion-scale LLMs

LLMs contain knowledge that helps visual zero-shot classification

Visual Alignment in Text-Only LLMs: New Frontiers in Data Efficiency. ArXiv
Jona Ruthardt, Gertjan J. Burghouts, Serge Belongie, Yuki M. Asano

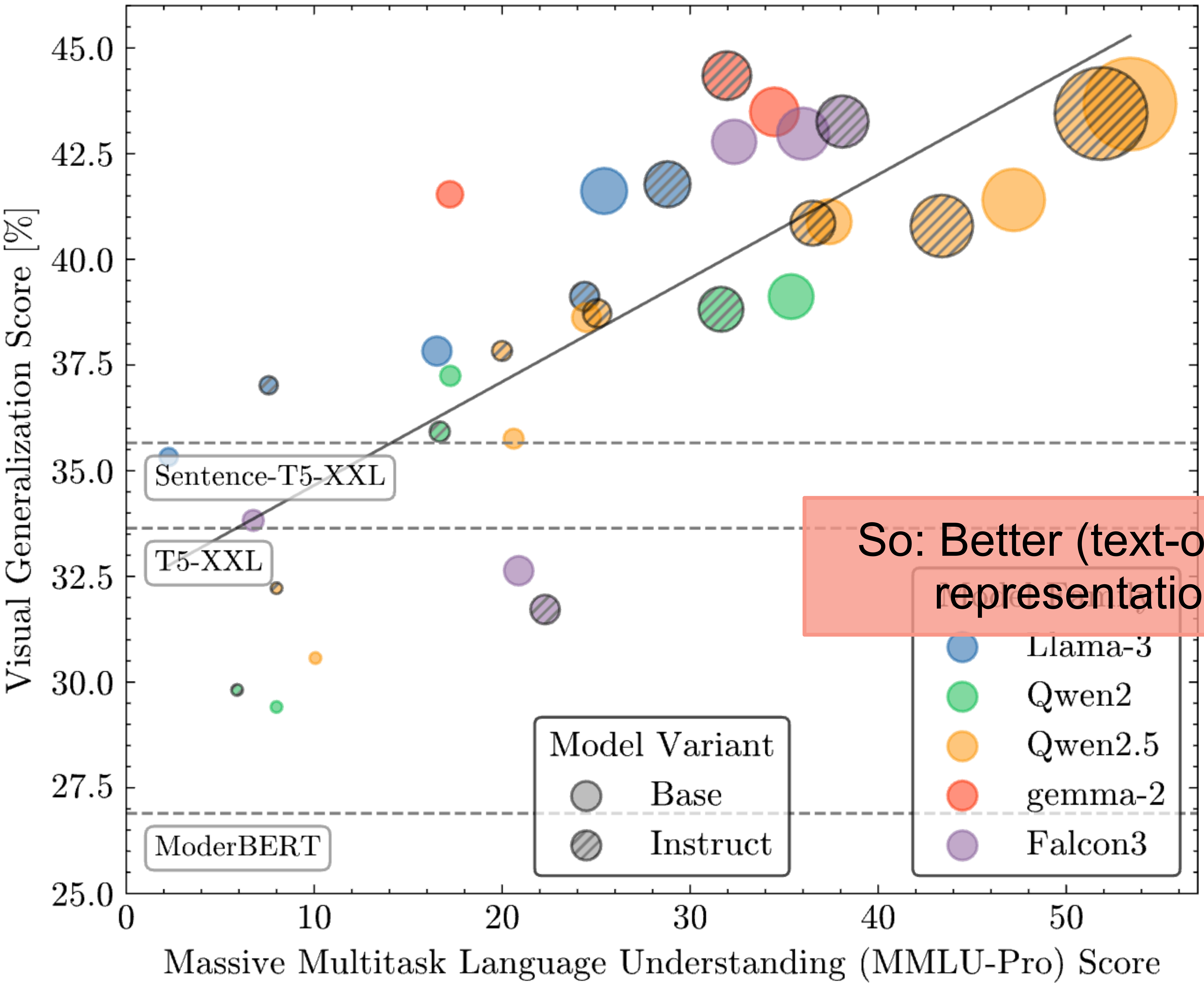


LLM's ShareLock performance correlates with (text-only) MMLU evaluation!

So: Better (text-only!) LLMs have a better representation of the **visual** world

Fundamental AI Lab

UTN



LLM's ShareLock performance correlates with (text-only) MMLU evaluation!

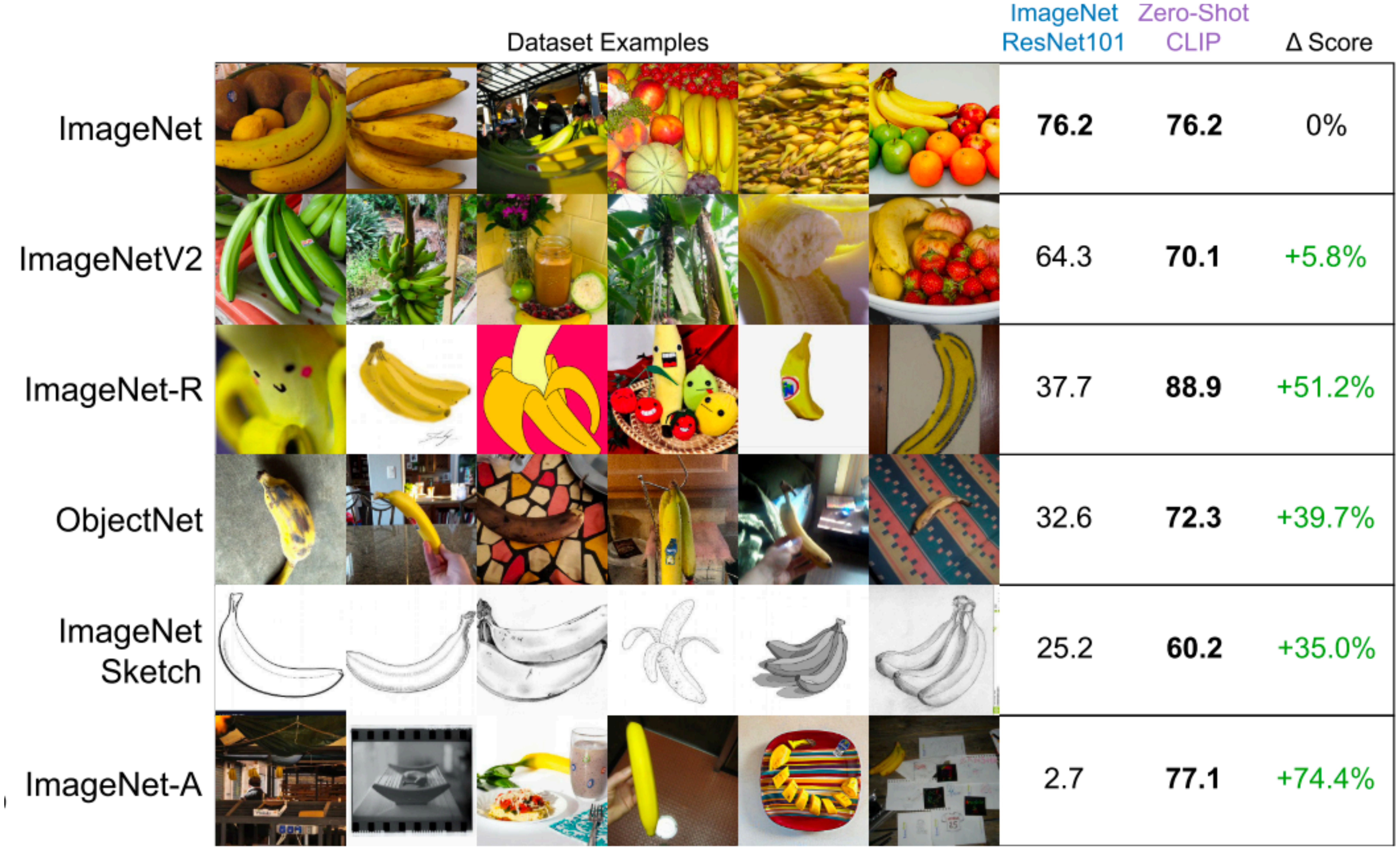
So: Better (text-only!) LLMs have a better representation of the **visual** world

CLIP trained models are much more robust

Zero-shot CLIP models are much more robust.

Reason: it's not the language part, but the data.

Note that no one *paid* for labelling/annotating the data. It was already there, but it does use additional learning signal (i.e. it's not self-supervised learning, but has a similar philosophy)



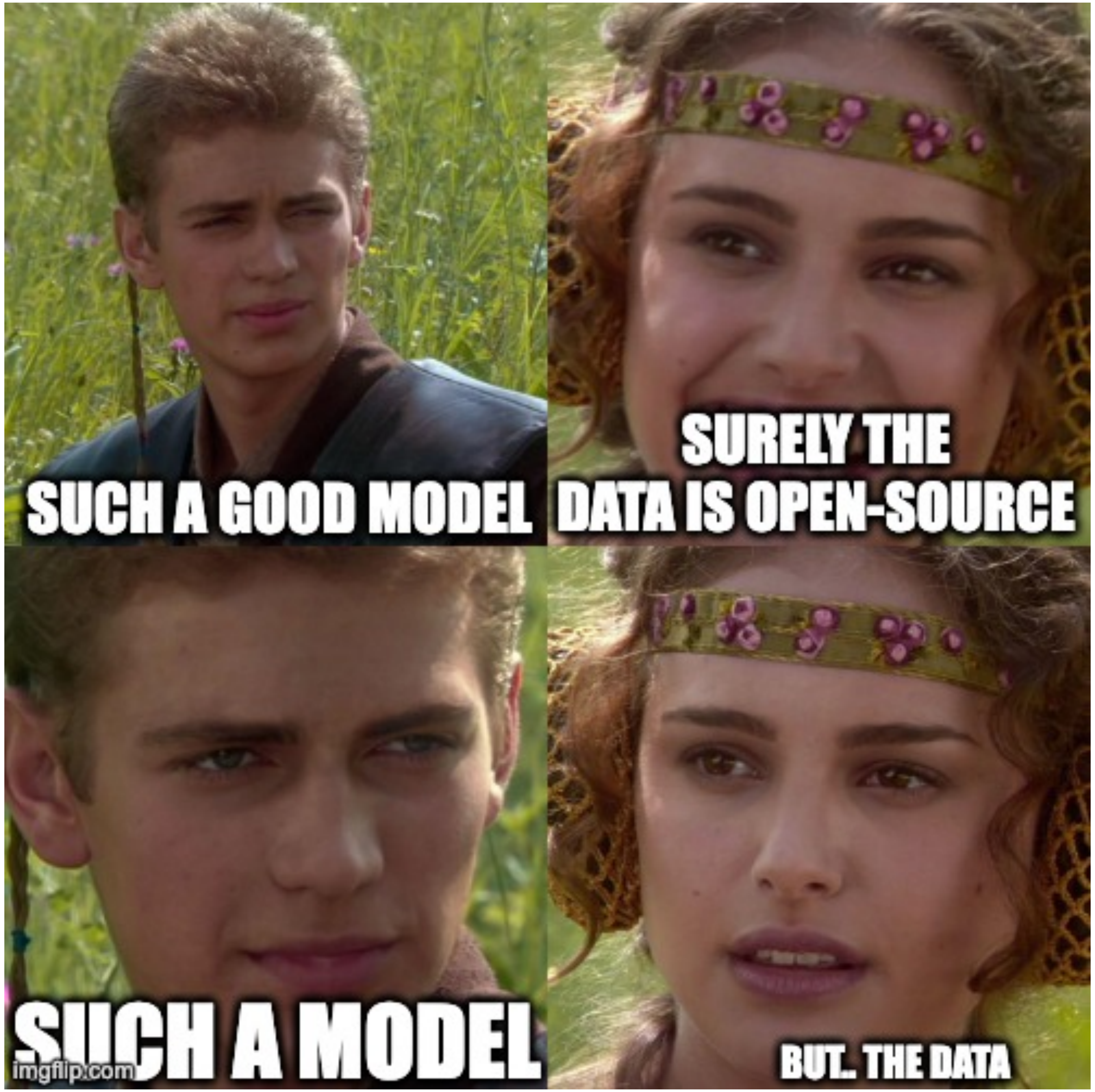
	Dataset Examples	ImageNet ResNet101	Zero-Shot CLIP	Δ Score
ImageNet		**76.2**	**76.2**	0%
ImageNetV2		64.3	**70.1**	+5.8%
ImageNet-R		37.7	**88.9**	+51.2%
ObjectNet		32.6	**72.3**	+39.7%
ImageNet Sketch		25.2	**60.2**	+35.0%
ImageNet-A		2.7	**77.1**	+74.4%

Gadre et al. DataComp: In search of the next generation of multimodal datasets. NeurIPS 2023

CLIP training data

400M proprietary image-caption pairs

.. but there's some open-source alternatives



CLIP pretrained models have been used in a variety of downstream tasks

- Text-conditional image generation models
- Video understanding/classification models
- Dataset cleaning
- …
- **Vision Language Models**

"Vibrant portrait painting of Salvador Dalí with a robotic half."

vibrant portrait painting of Salvador Dalí with a robotic half face

Some further developments of CLIP

Do you really need to train the image-encoder from scratch? No.

Lucas Beyer: also at IAISS'25!

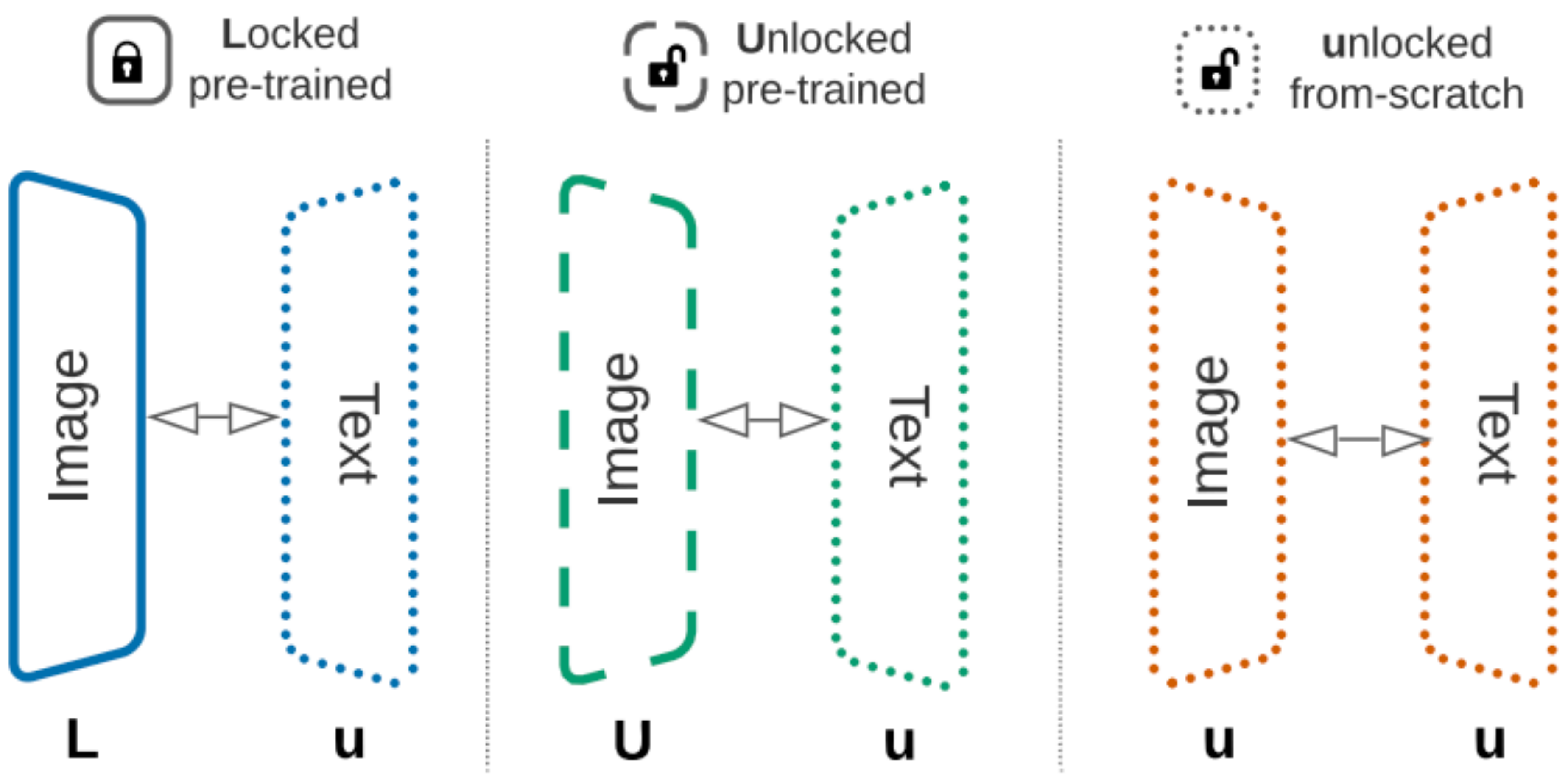


Figure 2. Design choices for contrastive-tuning on image-text data. Two letters are introduced to represent the image tower and text tower setups. `L` stands for locked variables and initialized from a pre-trained model, `U` stands for unlocked and initialized from a pre-trained model, `u` stands for unlocked and randomly initialized. `Lu` is named as "Locked-image Tuning" (LiT).

Method	ImgNet	ImgNet-v2	Cifar100	Pets
`Lu`	70.1	61.7	70.9	88.1
`Uu`	57.2	50.2	62.1	74.8
`uu`	50.6	43.3	47.9	70.3

Locking the image model is better.

Table 3: Zero-shot transfer results on ImageNet (variants).

Model	IN	IN-v2	IN-R	IN-A	ObjNet	ReaL
CLIP	76.2	70.1	88.9	77.2	72.3	-
ALIGN	76.4	70.1	92.2	75.8	72.2	-
BASIC	85.7	80.6	95.7	85.6	78.9	-
CoCa	86.3	80.7	96.5	90.2	82.7	-
LiT-g/14	85.2	79.8	94.9	81.8	82.5	88.6
LiT-e/14	85.4	80.6	96.1	88.0	84.9	88.4
LiT-22B	85.9	80.9	96.0	90.1	87.6	88.6

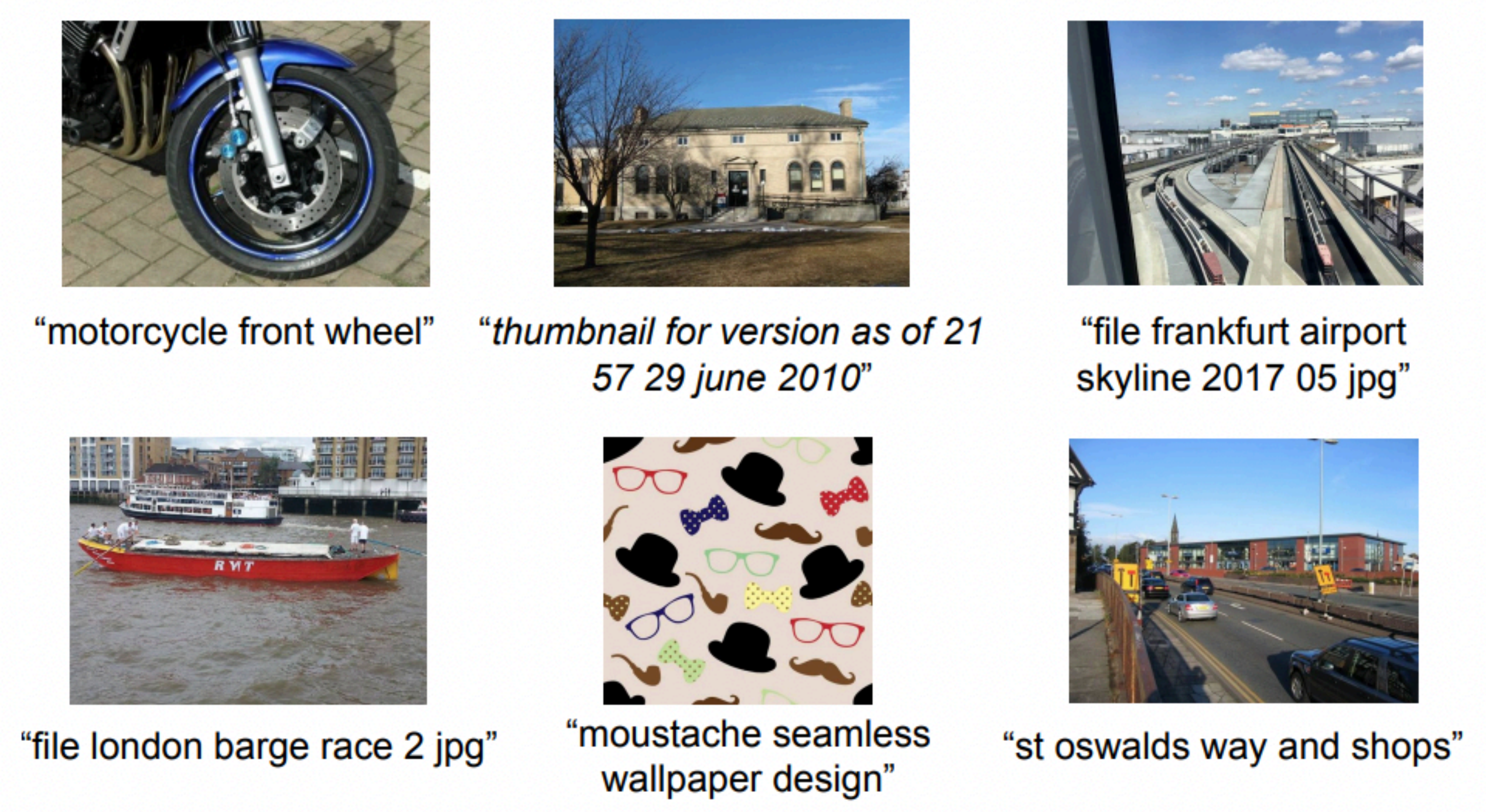
With only requiring one forward pass for getting image embeddings, can combine this with using a 22B parameter ViT

Fundamental AI Lab

UTN

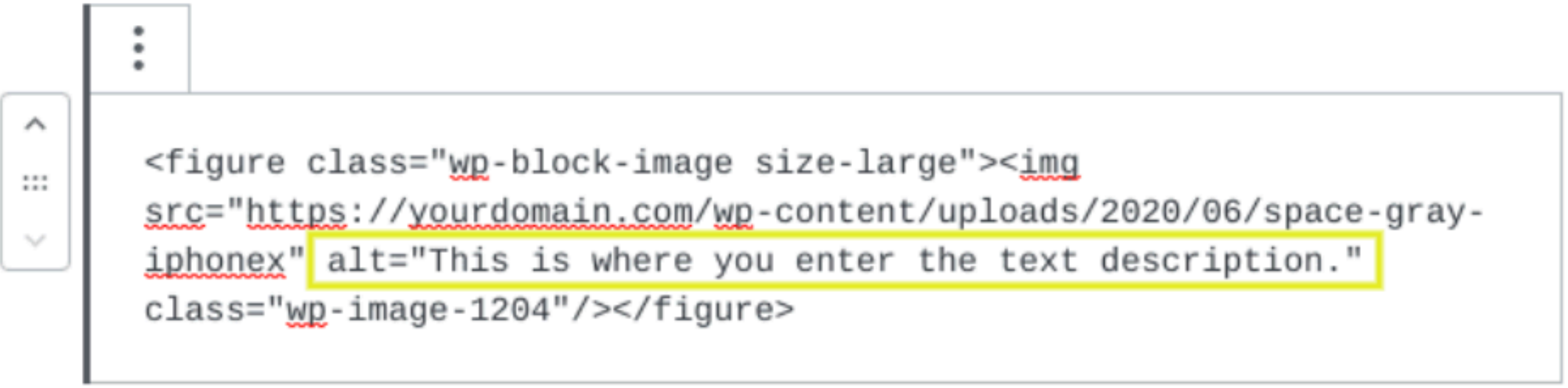
Zhai et al. LiT: Zero-Shot Transfer with Locked-image text Tuning. CVPR 2022
Dheghani et al. Scaling Vision Transformers to 22 Billion Parameters. 2023

Scaling to larger datasets: ALIGN



“motorcycle front wheel” *“thumbnail for version as of 21 57 29 june 2010”* “file frankfurt airport skyline 2017 05 jpg”

“file london barge race 2 jpg” “moustache seamless wallpaper design” “st oswalds way and shops”



```
<figure class="wp-block-image size-large"><img
src="https://yourdomain.com/wp-content/uploads/2020/06/space-gray-
iphonex" alt="This is where you enter the text description."
class="wp-image-1204"/></figure>
```

Their innovation: start with very noisy dataset and:

- Filter based on images:
  - remove small ones, remove ones with >1k captions/alt texts
- Filter based on text:
  - alt-text with >10 occurences are removes (e.g. ”1920x10280”)
  - too short or too long, or too rare
- Result: dataset size ~2B (CLIP: 400M)

We train the model on 1024 Cloud TPUv3 cores with 16 positive pairs on each core. Therefore the total effective batch size is 16384.

Jia et al. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. ICML 2021

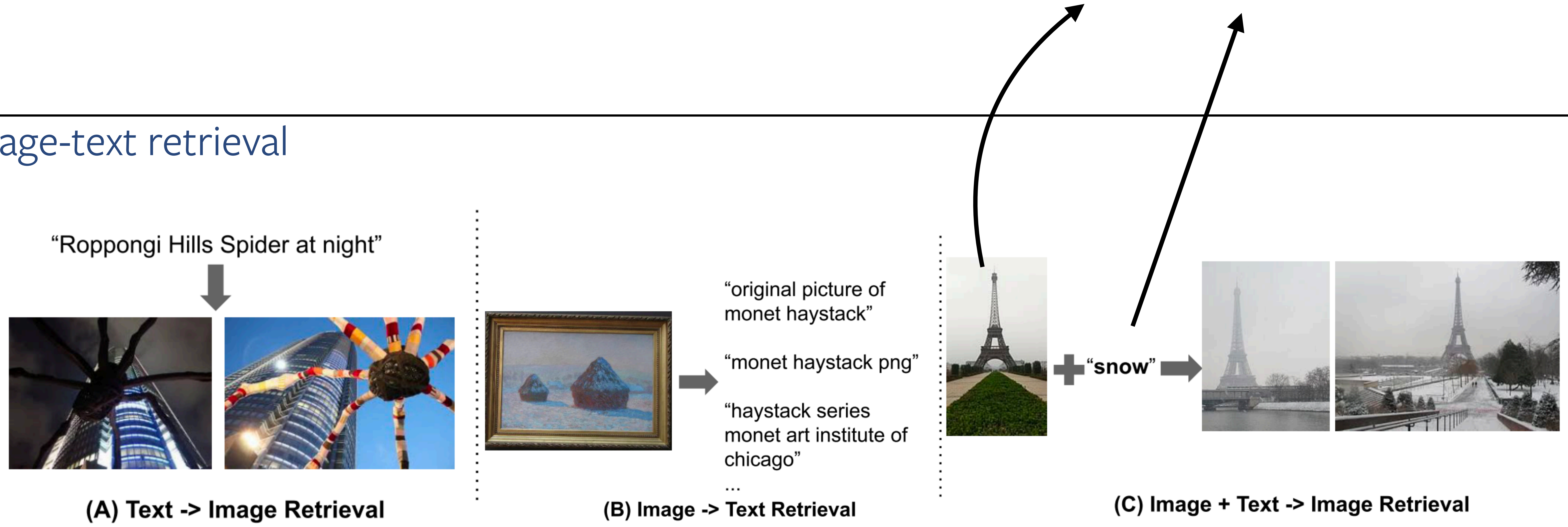
ALIGN paper shows some more multimodal applications

simple addition of two vectors

(C): How?

vec_1 + vec_2 --> find nearest images in database

Image-text retrieval


“Roppongi Hills Spider at night”

(A) Text -> Image Retrieval

“original picture of monet haystack”

“monet haystack png”

“haystack series monet art institute of chicago”

...

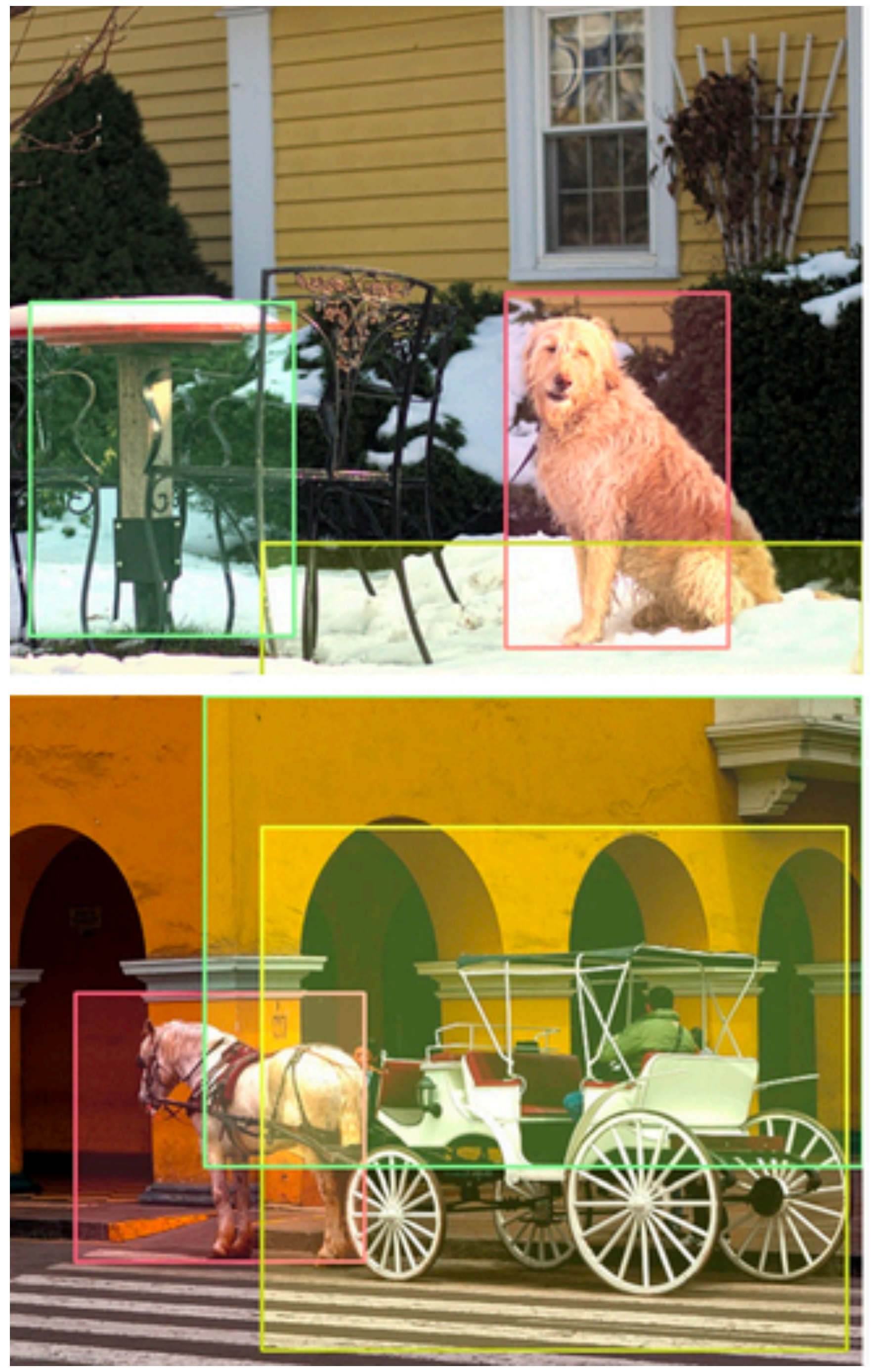
(B) Image -> Text Retrieval

(C) Image + Text -> Image Retrieval

Jia et al. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. ICML 2021

Text-image retrieval tasks/datasets

E.g. MS-COCO



{"**caption**": "a snow covered ground outside of a yellow colored house with a dog tied to an outdoor chair",
"**predict1**": "snow is falling on the outside of a house and a dog is sitting in a chair",
"**predict2**": "a dog is laying in the snow near a table and chairs",
"**keywords**": "snow house dog chair "}

{"**caption**": "a white horse drawn carriage in front of a yellow building",
"**predict1**": "a horse drawn carriage is parked in front of a building",
"**predict2**": "a very pretty horse pulling a fancy carriage",
"**keywords**": "horse carriage building "}

image-to-text

	OURS	A cat laying on the grass playing with a big dog	A cat plays with a dog in the grass	In a grassy field is a dog and a cat who are rubbing noses	The dog and cat are in a field of grass	A small dog standing next to a small kitten

count number of correct captions given a number of retrieved instances (e.g. 5)

text-to-image

A small cat lying in the grass paws at a dogs muzzle	OURS					

count number of correct images given a number of retrieved instances (e.g. 5)

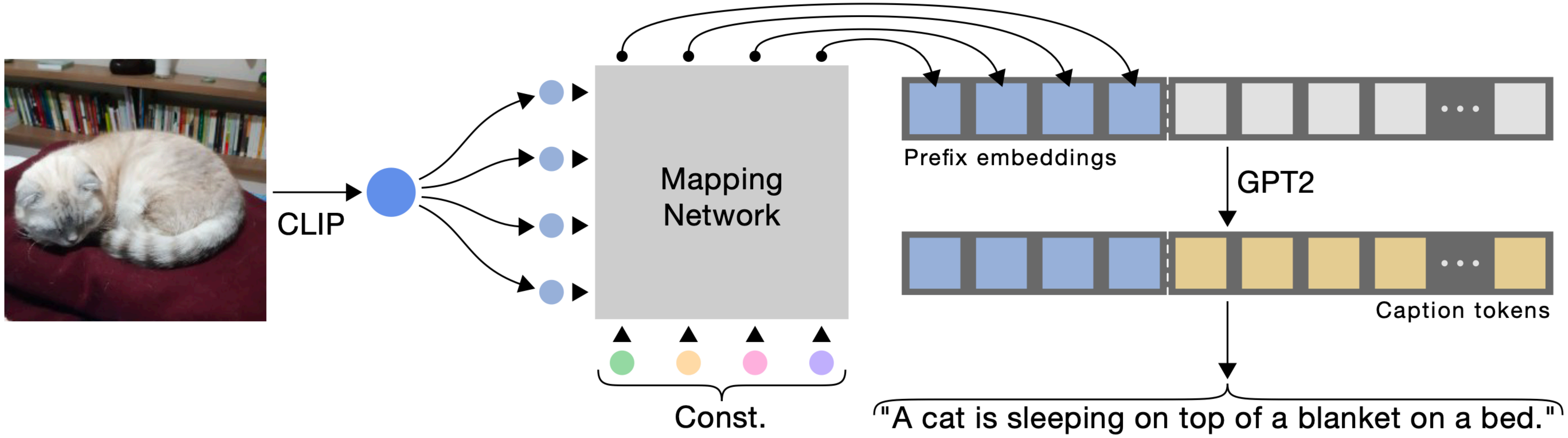
https://www.researchgate.net/figure/Visualization-of-retrieval-on-MSCOCO_fig4_360585864
Lin et al. Microsoft COCO: Common Objects in Context. ECCV 2014

Multimodal text *generative* models (MLLMs)

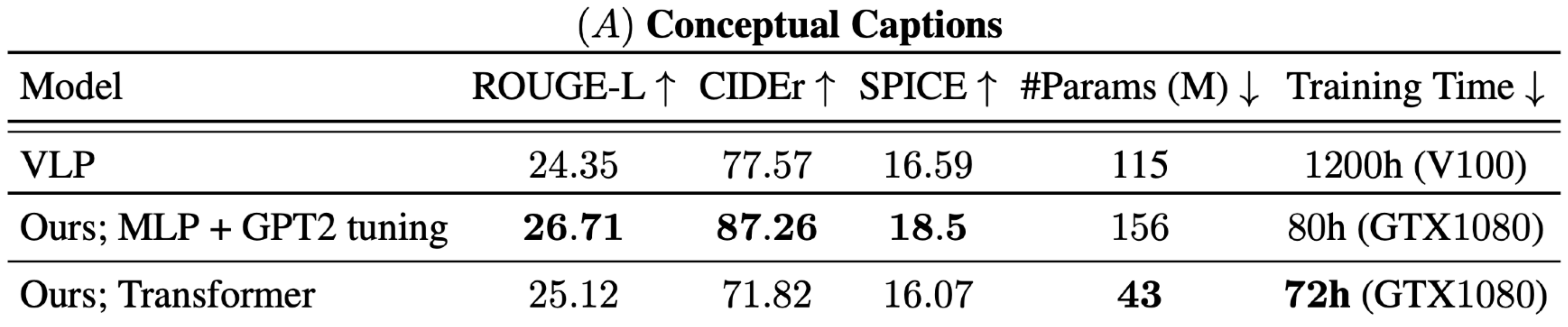
"I still wish we'd gotten a pool, instead of this ridiculous sculpture."

ClipCap: CLIP Prefix for Image Captioning

Now a language **decoder** model!



- Uses CLIP visual encoder, further transforms the visual embedding to match the input-space of GPT-2.
- GPT-2 kept frozen or adapted
- Trained for captioning



(*A*) **Conceptual Captions**

Model	ROUGE-L ↑	CIDEr ↑	SPICE ↑	#Params (M) ↓	Training Time ↓
VLP	24.35	77.57	16.59	115	1200h (V100)
Ours; MLP + GPT2 tuning	**26.71**	**87.26**	**18.5**	156	80h (GTX1080)
Ours; Transformer	25.12	71.82	16.07	**43**	**72h** (GTX1080)

Mokady et al. ClipCap: CLIP Prefix for Image Captioning. 2022

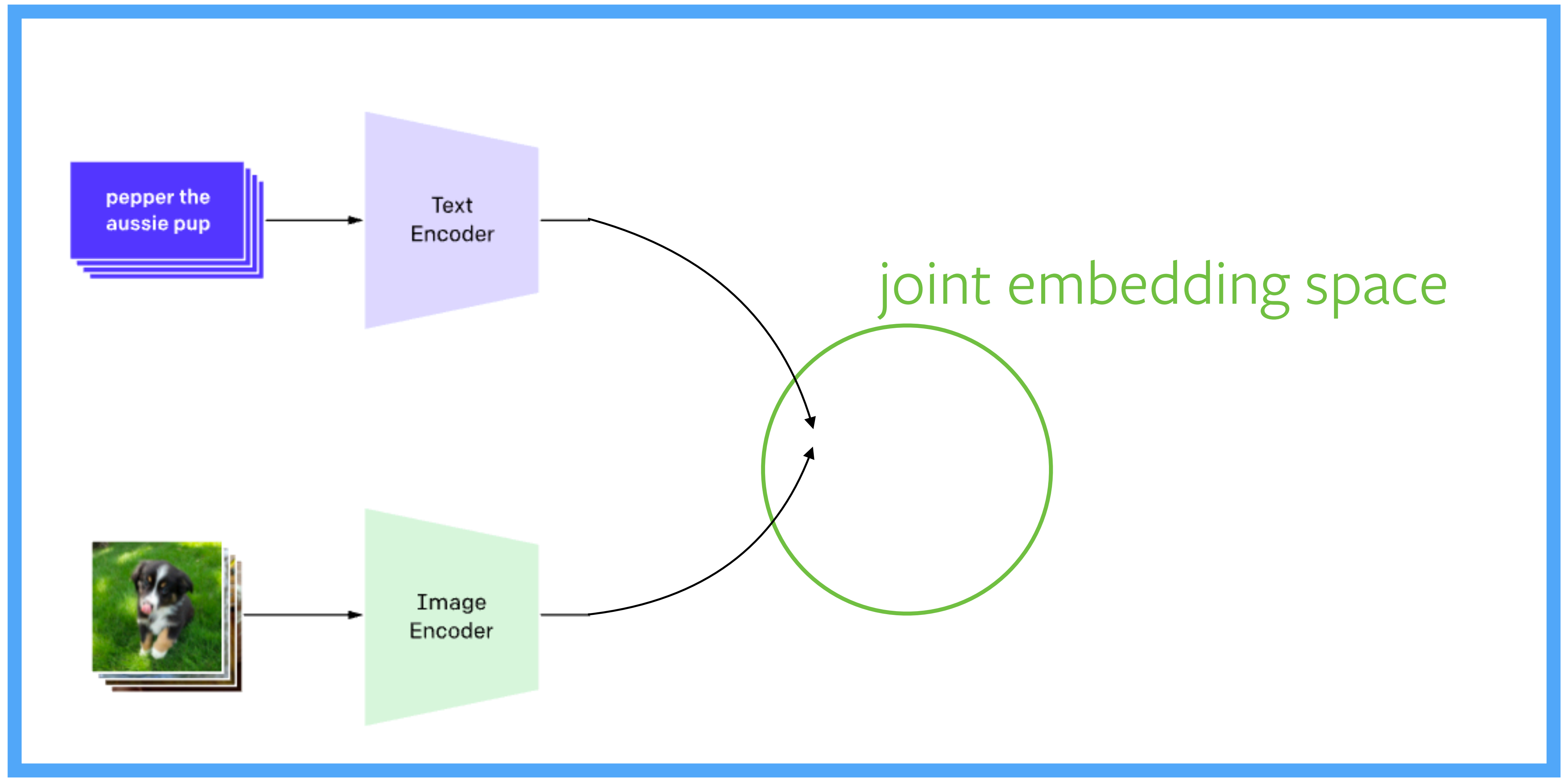
80h on 1 (old)GPU

Some terminology:

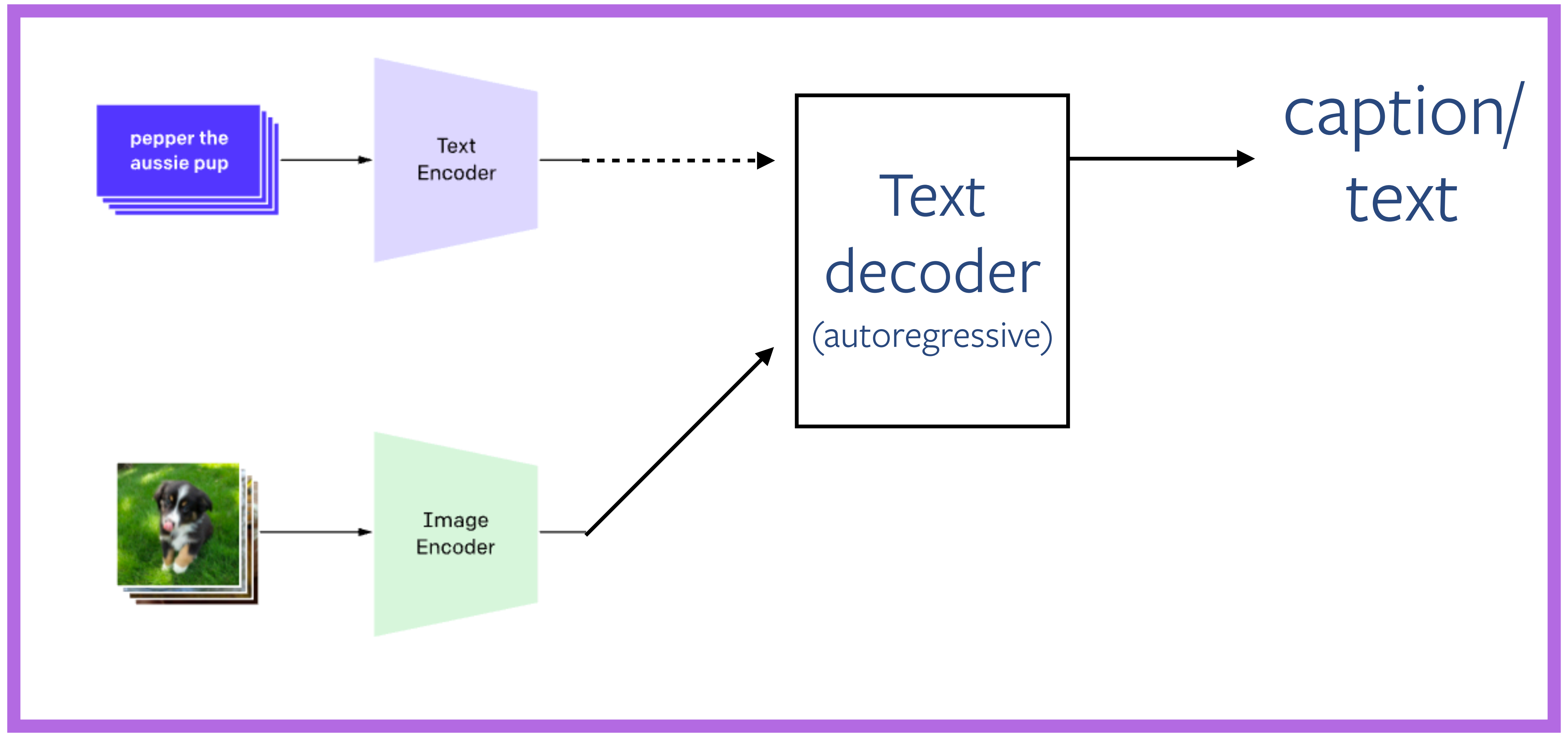
Vision-Language model vs Visual Language model [or: Multimodal LLM (MLLM)]

or

encoder vs decoder architectures



Both modalities mapped into a joint embedding space. Great for cross-modal retrieval, or refined joint-modal retrieval (Eiffel-tower-image+"snow")



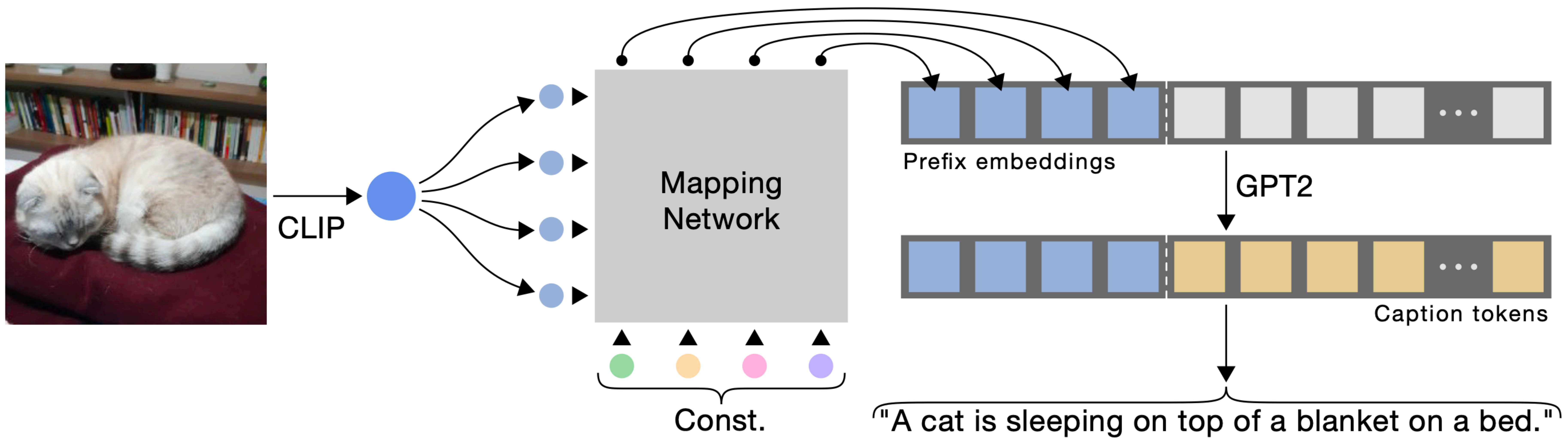
When text decoder is a frozen language model:

Image --> "language space", s.t. decoder can deal with it.

Fundamental AI Lab

UTN

ClipCap: CLIP Prefix for Image Captioning



Question 1: why didn’t they use GPT3?

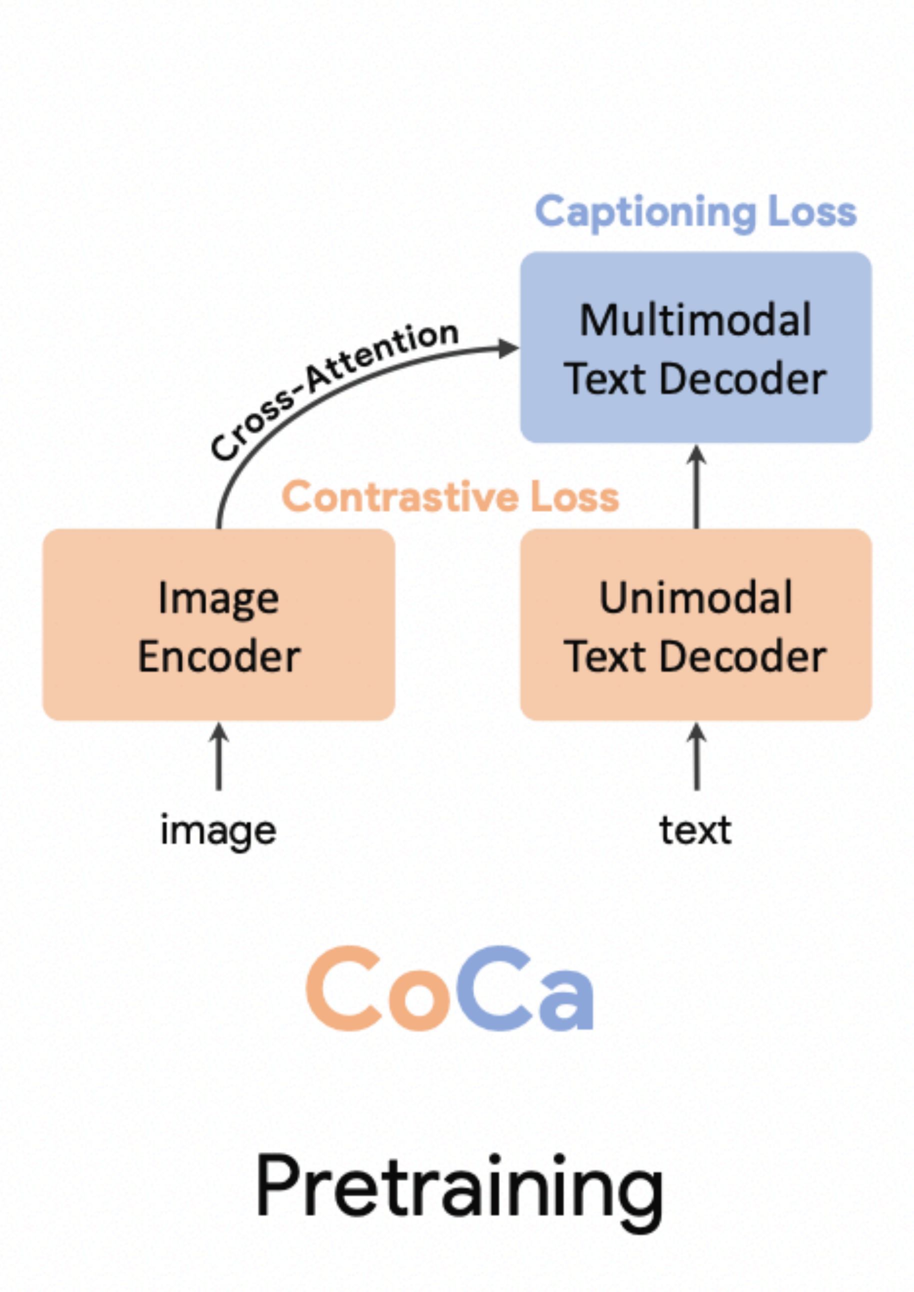
1) The sparse attention in GPT-3 would lead to only looking at parts of the image
2) GPT-2 does the captioning job well enough, so no need for GPT-3
3) The training requires access to the model weights

Question 2: why is the transformer-adaptation (& freezing GPT-2) variant nice?

1) The model learns to better forget what it learned during language-only training
2) The language model can be made very efficient
3) Transformers are faster than fully connected layers
4) The number of parameters doesn’t depend on the number of CLIP’s visual output size

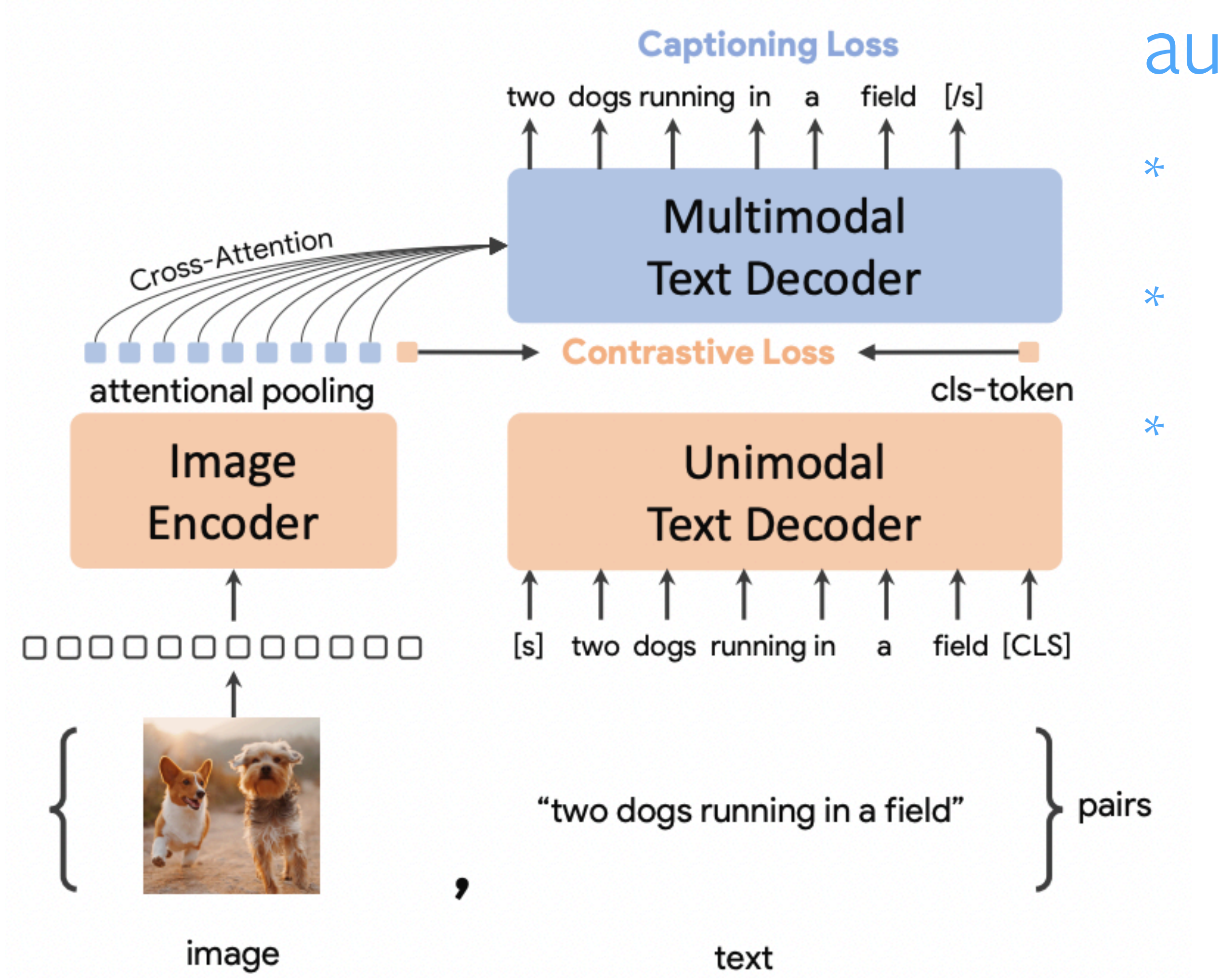
CoCa: Contrastive Captioners are Image-Text Foundation Models

https://colab.research.google.com/github/mlfoundations/open_clip/blob/master/docs/Interacting_with_open_coca.ipynb



Caption generation is autoregressive, starting from a [start] token

How it works:



CLIP-like contrastive aligning of [cls] tokens

discriminative

auto-regressive decoding:

- start with a [start] token
- this needs to get mapped to the first word
- first sampled word (+[start]) needs to get mapped to second etc

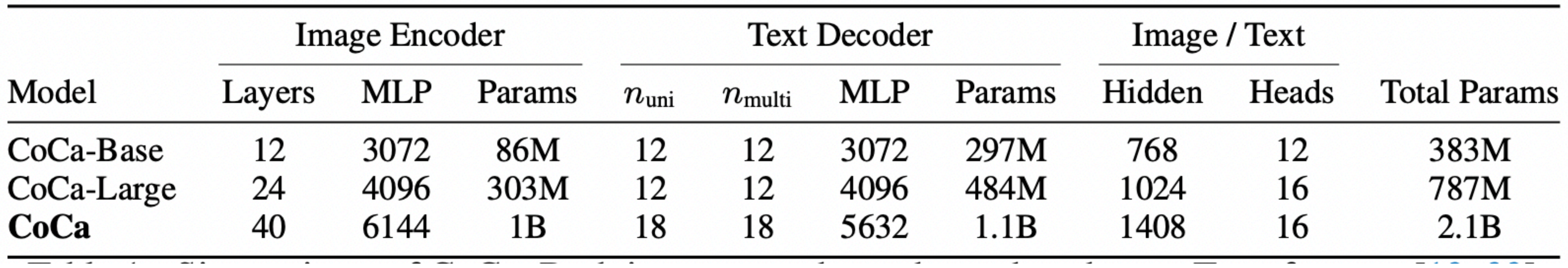
generative

Yu et al. Contrastive Captioners are Image-Text Foundation Models. TMLR 2022

What you can do with CoCa

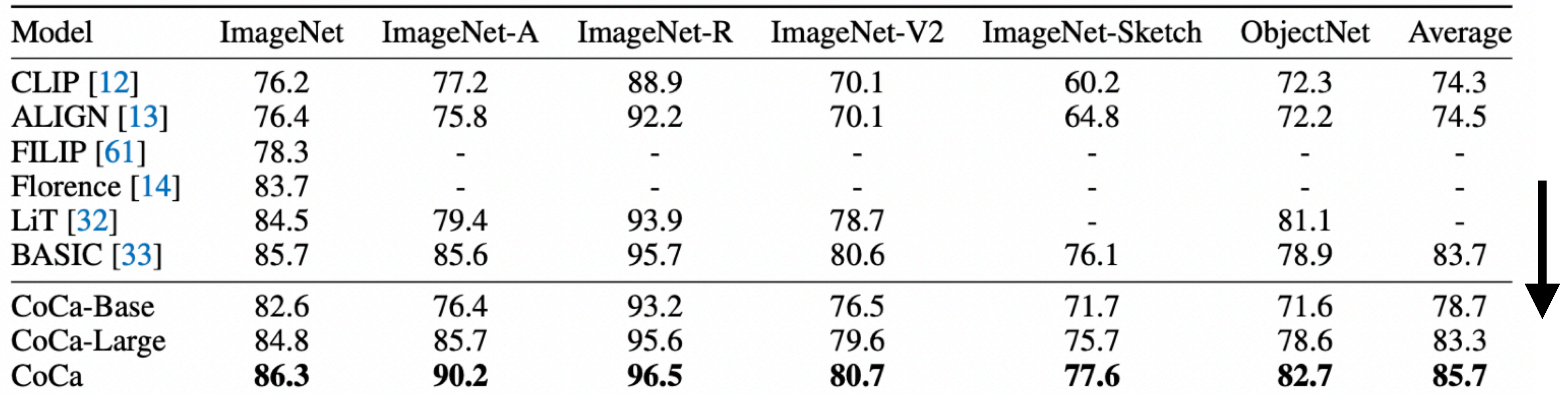
https://colab.research.google.com/github/mlfoundations/open_clip/blob/master/docs/Interacting_with_open_coca.ipynb

Scale it



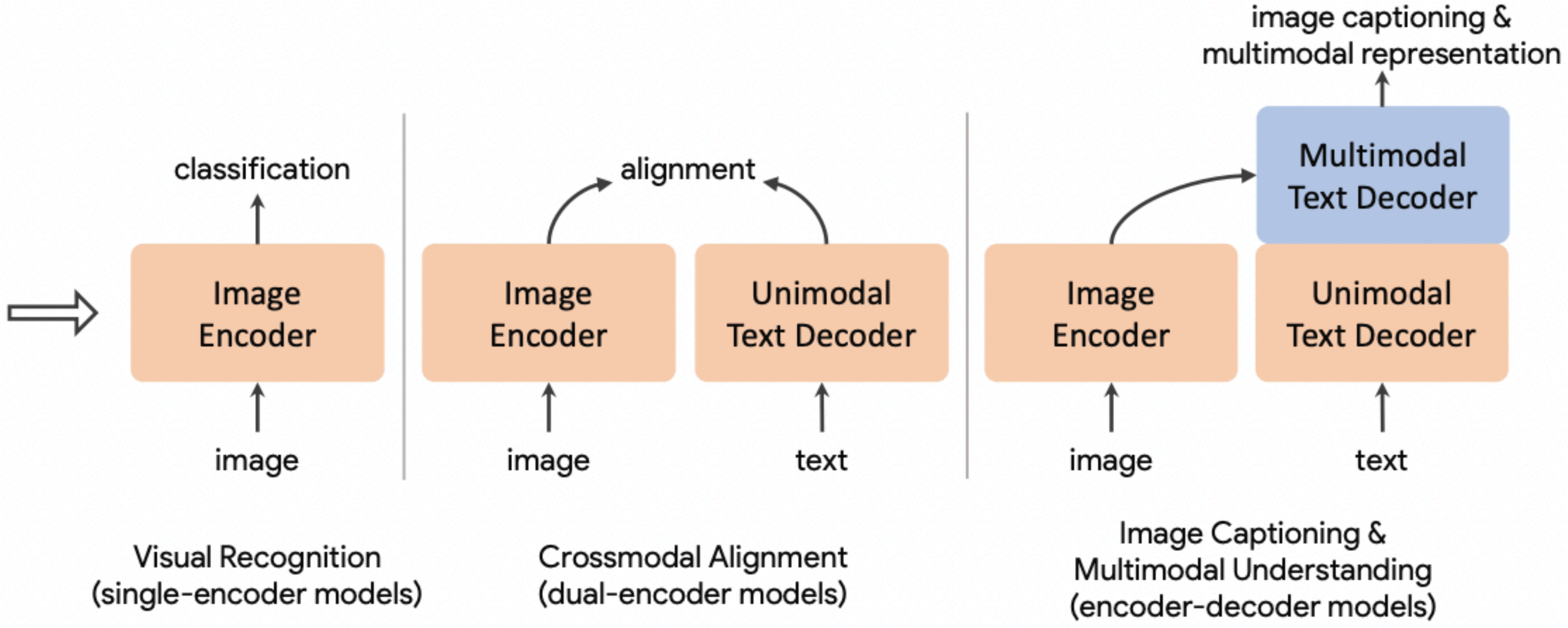
Model	Image Encoder			Text Decoder				Image / Text		Total Params
	Layers	MLP	Params	n_{uni}	n_{multi}	MLP	Params	Hidden	Heads	
CoCa-Base	12	3072	86M	12	12	3072	297M	768	12	383M
CoCa-Large	24	4096	303M	12	12	4096	484M	1024	16	787M
CoCa	40	6144	1B	18	18	5632	1.1B	1408	16	2.1B

Use the visual encoder



Model	ImageNet	ImageNet-A	ImageNet-R	ImageNet-V2	ImageNet-Sketch	ObjectNet	Average
CLIP [12]	76.2	77.2	88.9	70.1	60.2	72.3	74.3
ALIGN [13]	76.4	75.8	92.2	70.1	64.8	72.2	74.5
FILIP [61]	78.3	-	-	-	-	-	-
Florence [14]	83.7	-	-	-	-	-	-
LiT [32]	84.5	79.4	93.9	78.7	-	81.1	-
BASIC [33]	85.7	85.6	95.7	80.6	76.1	78.9	83.7
CoCa-Base	82.6	76.4	93.2	76.5	71.7	71.6	78.7
CoCa-Large	84.8	85.7	95.6	79.6	75.7	78.6	83.3
CoCa	**86.3**	**90.2**	**96.5**	**80.7**	**77.6**	**82.7**	**85.7**

Table 4: Zero-shot image classification results on ImageNet [9], ImageNet-A [64], ImageNet-R [65], ImageNet-V2 [66], ImageNet-Sketch [67] and ObjectNet [68].



Zero-shot, frozen-feature or finetuning

Generate captions

a hand holding a san francisco 49ers football

a row of cannons with the eiffel tower in the background

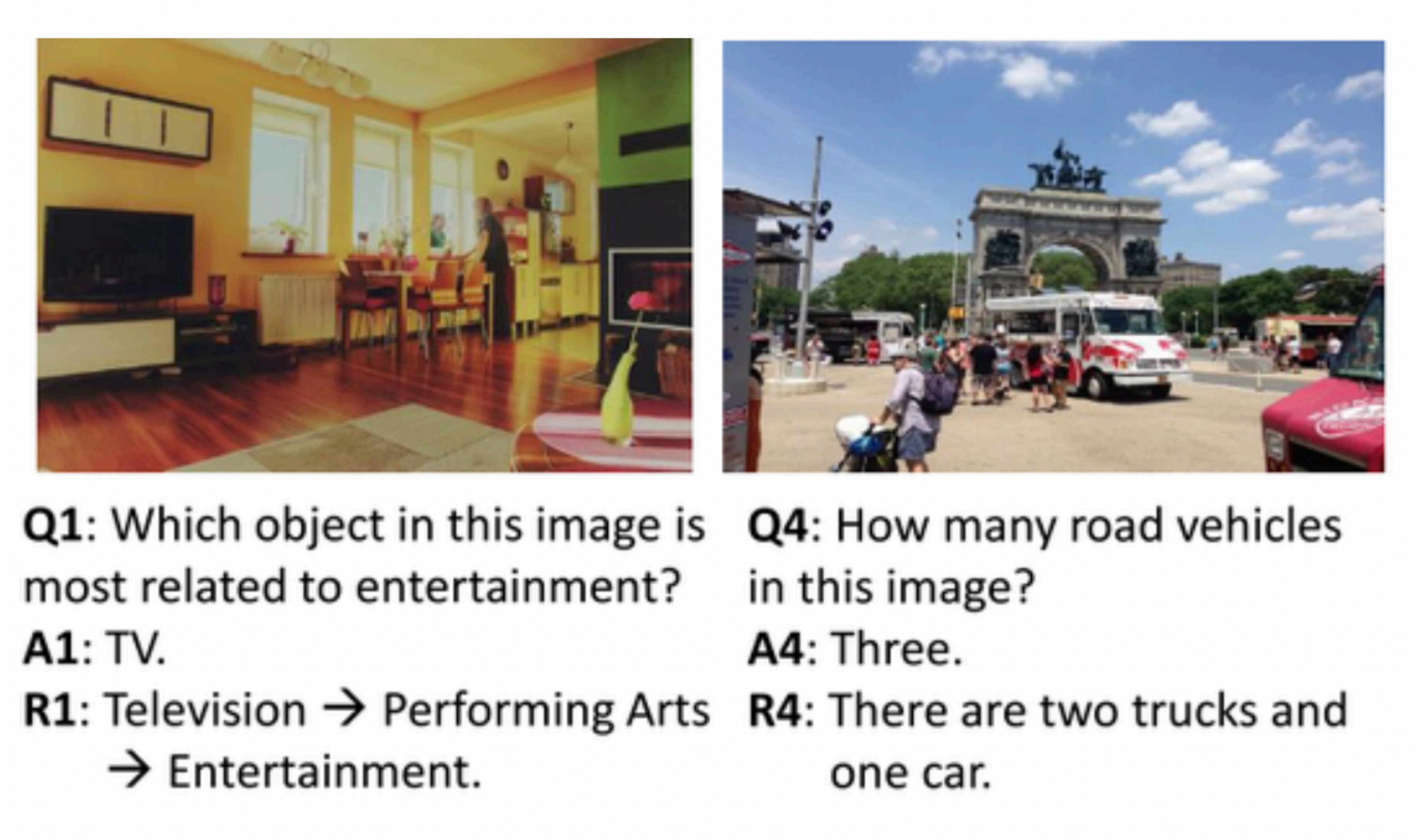
We use the JFT-3B dataset [21] with label names as the paired texts, and the ALIGN dataset [13] with noisy alt-texts.

Pretraining CoCa takes about 5 days on 2,048 CloudTPUv4 chips 😢

Fundamental AI Lab UTN

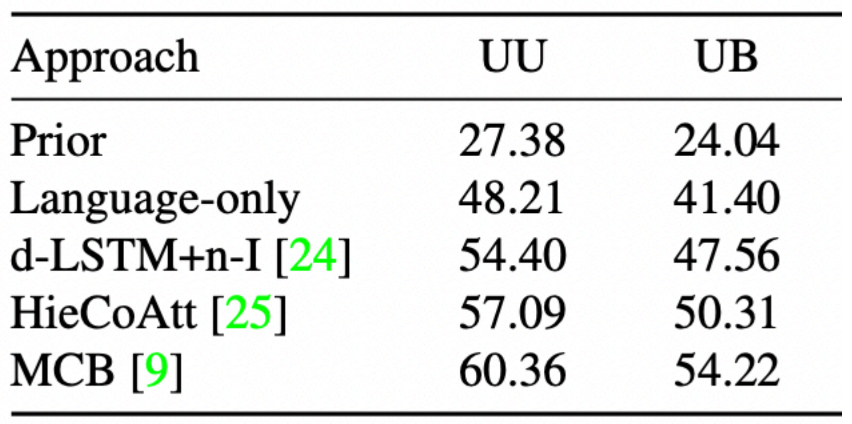
Yu et al. Contrastive Captioners are Image-Text Foundation Models. TMLR 2022

What you can do with MLLMs: Multi-modal understanding, e.g. VQA



Q1: Which object in this image is most related to entertainment?
A1: TV.
R1: Television → Performing Arts → Entertainment.

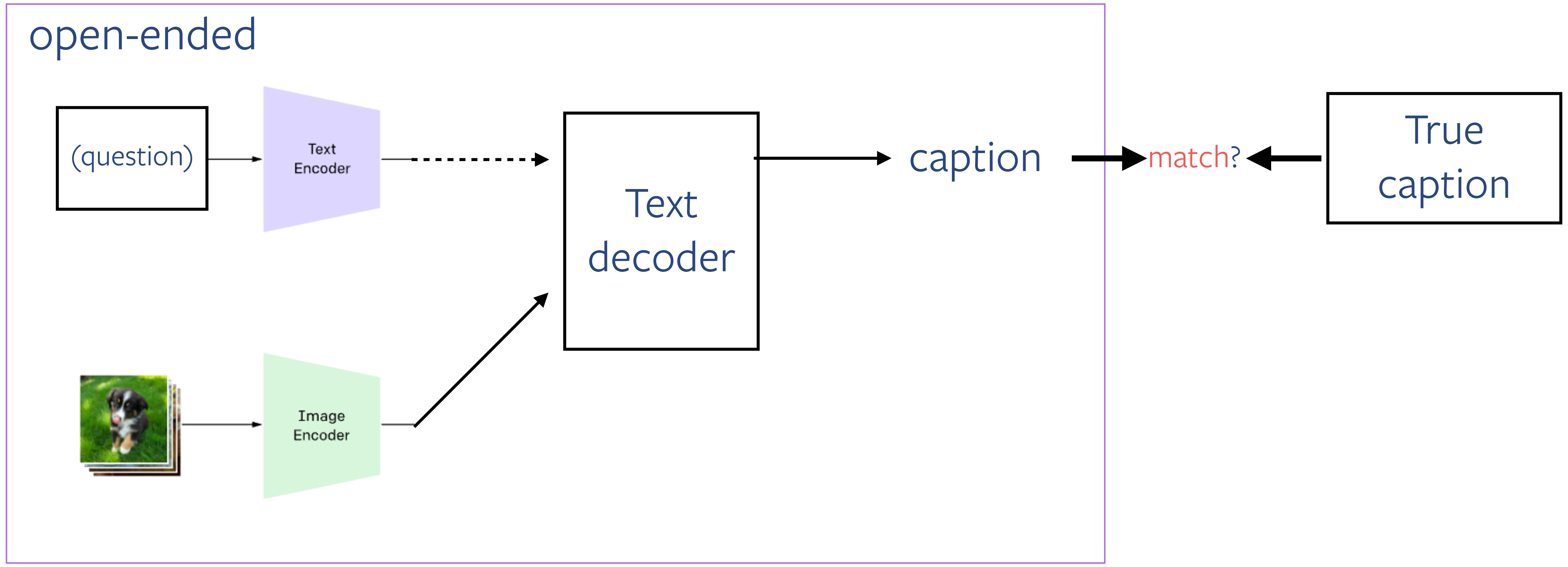
Q4: How many road vehicles in this image?
A4: Three.
R4: There are two trucks and one car.

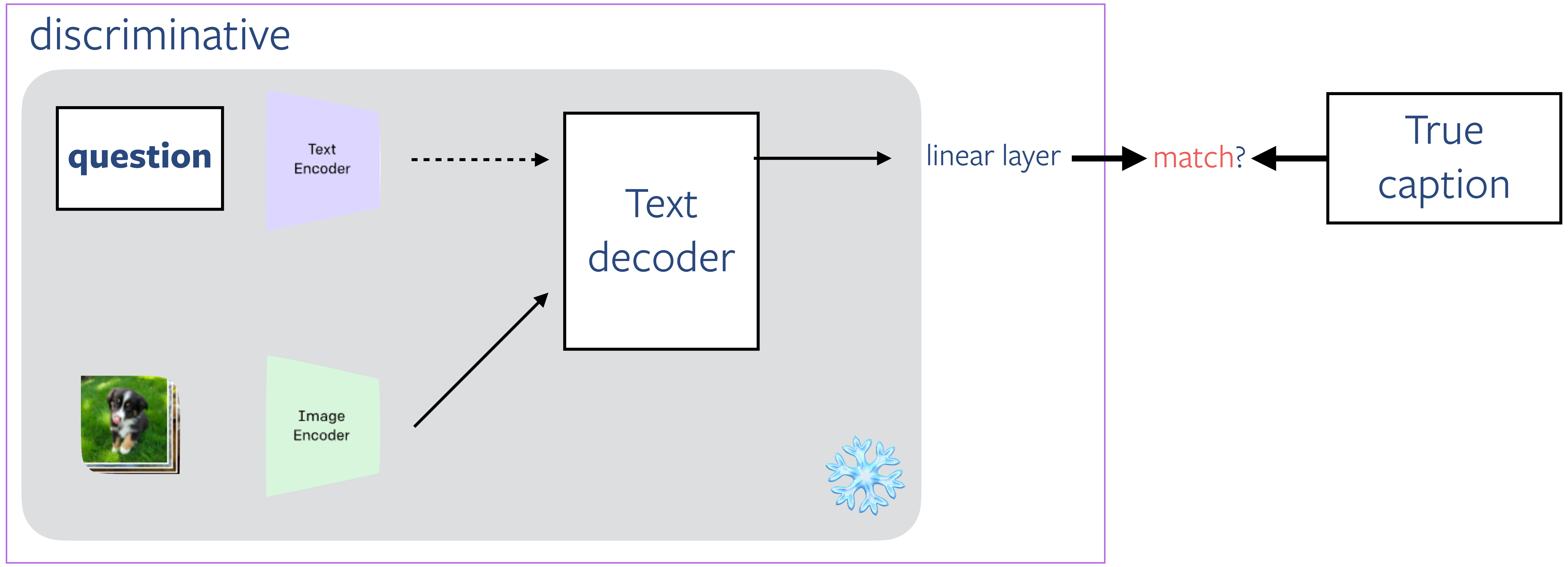


Approach	UU	UB
Prior	27.38	24.04
Language-only	48.21	41.40
d-LSTM+n-I [24]	54.40	47.56
HieCoAtt [25]	57.09	50.31
MCB [9]	60.36	54.22

Note: some questions could be answered without image

--> VQA-v2 (balanced images to each question)

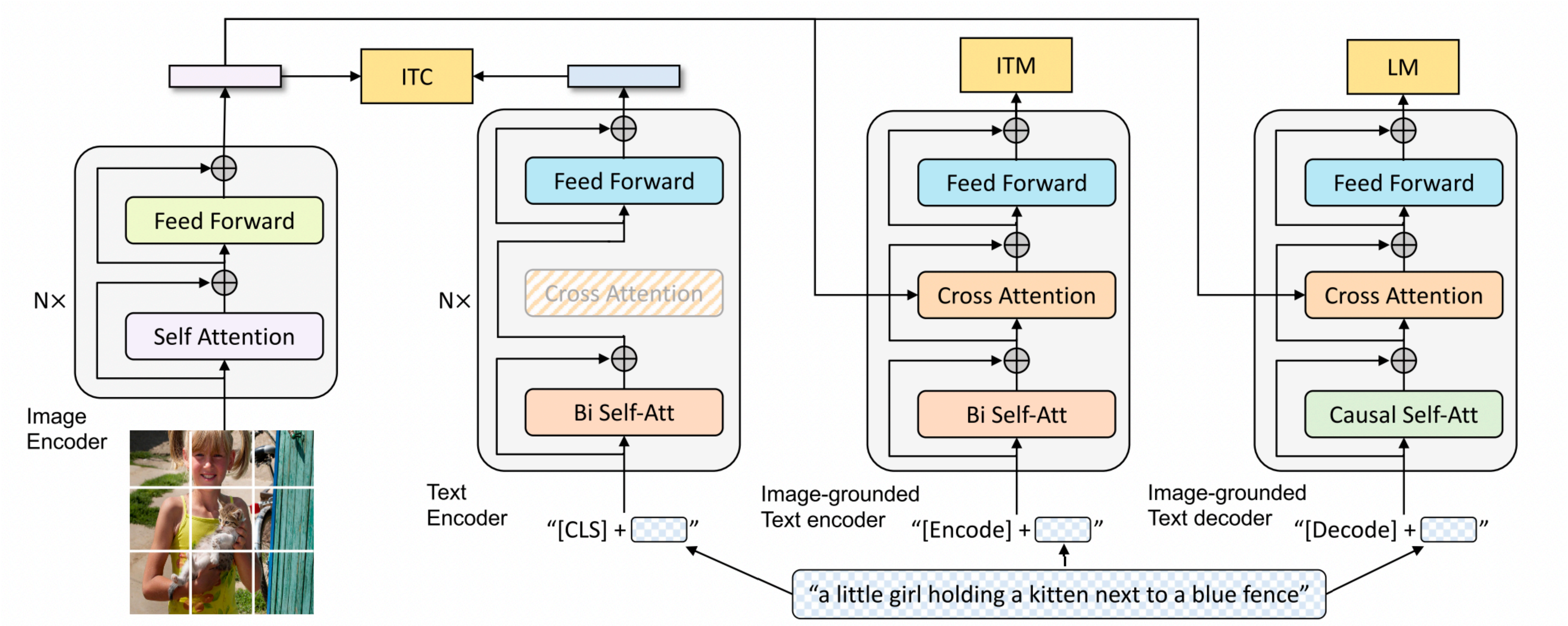




Aggrawal et al. VQA: visual question answering. ICCV 2015
Goyal et al. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. CVPR 2017

BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation

https://github.com/huggingface/notebooks/blob/main/examples/image_captioning_blip.ipynb



+ iterative data filtering and dataset expansion strategy

by using synthetic captions via LM (~text augmentation) as GT

and ITM&ITC model as filtering

diverse captions (sample with some non-zero temperature from your captioning model) help

32 GPUs

ITC: Image-text contrastive learning
ITM: Image-text binary matching (yes?/no?)
LM: autoregressive captioning

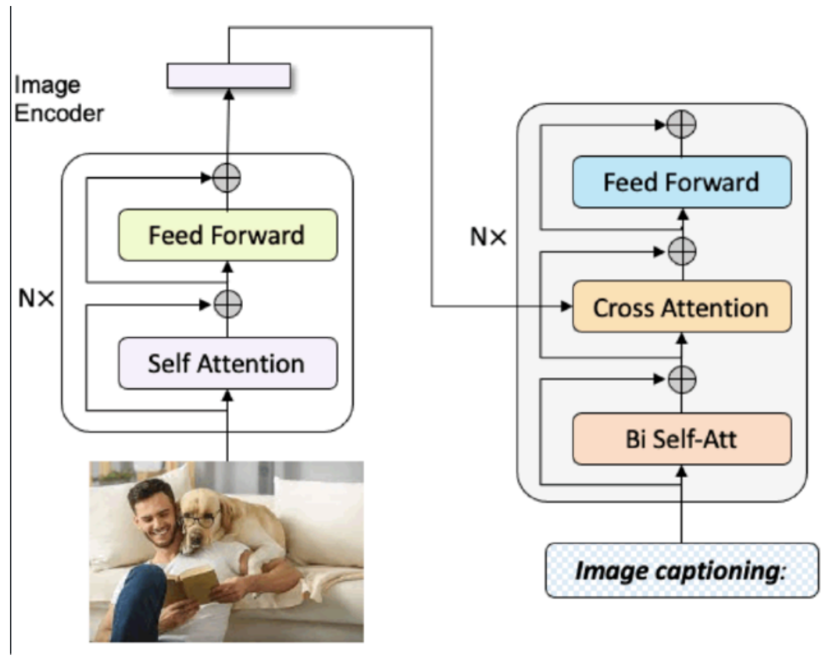
Li et al. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. ICML 2022
https://huggingface.co/docs/transformers/model_doc/blip

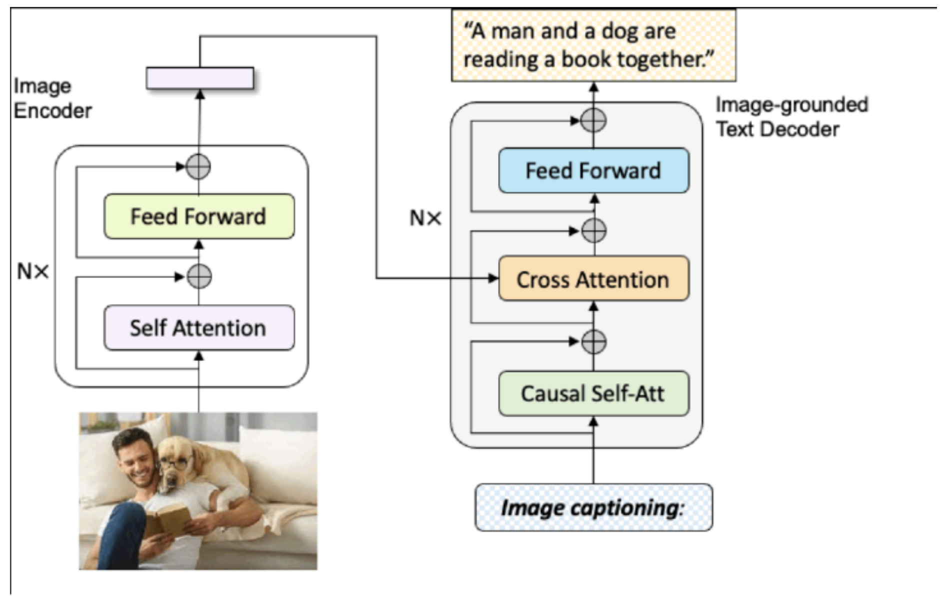
BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation

https://github.com/huggingface/notebooks/blob/main/examples/image_captioning_blip.ipynb

Various usage modes:

image-caption matching, image-captioning





Text & image encoding & text decoder allows for more flexible applications:

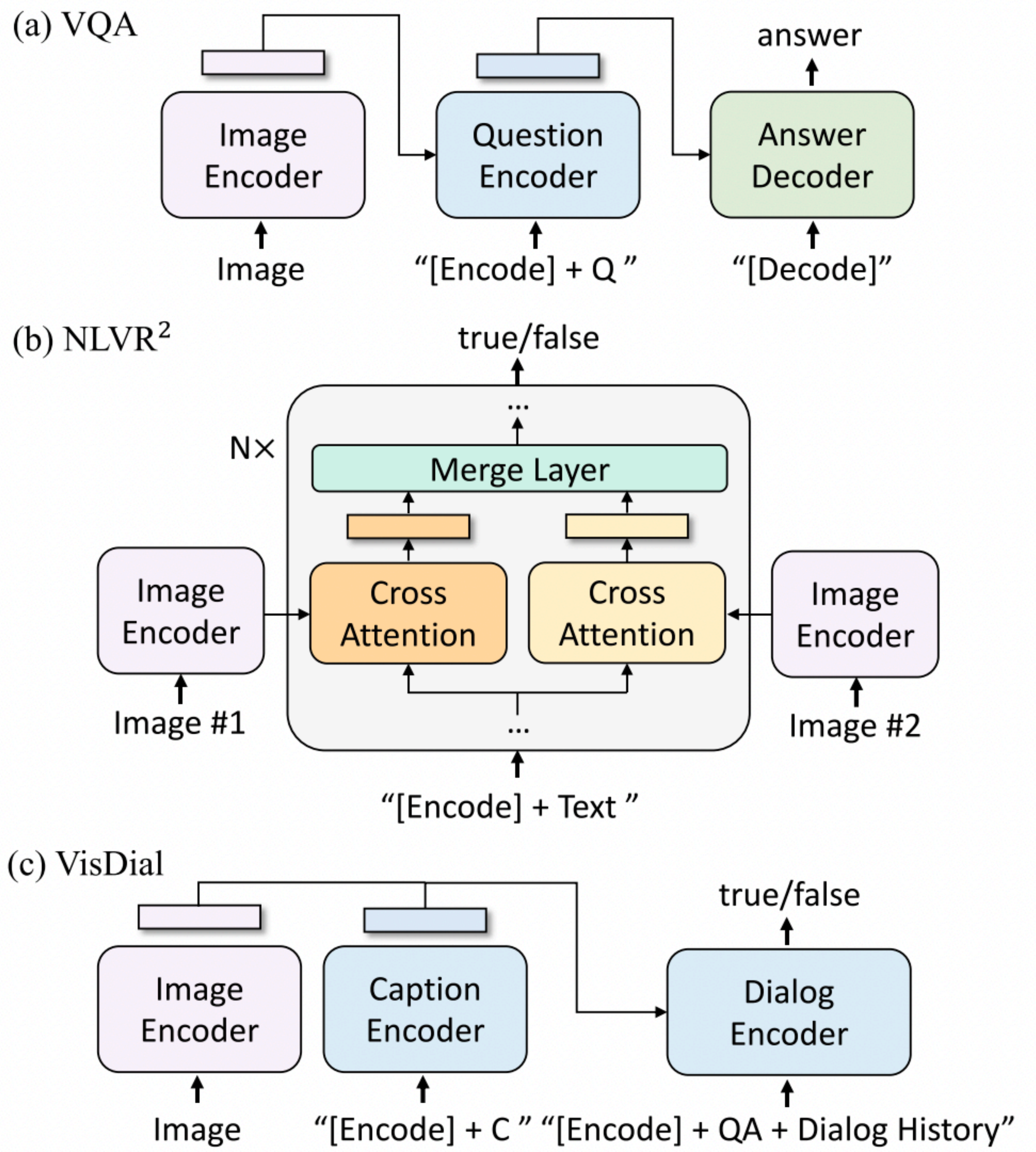
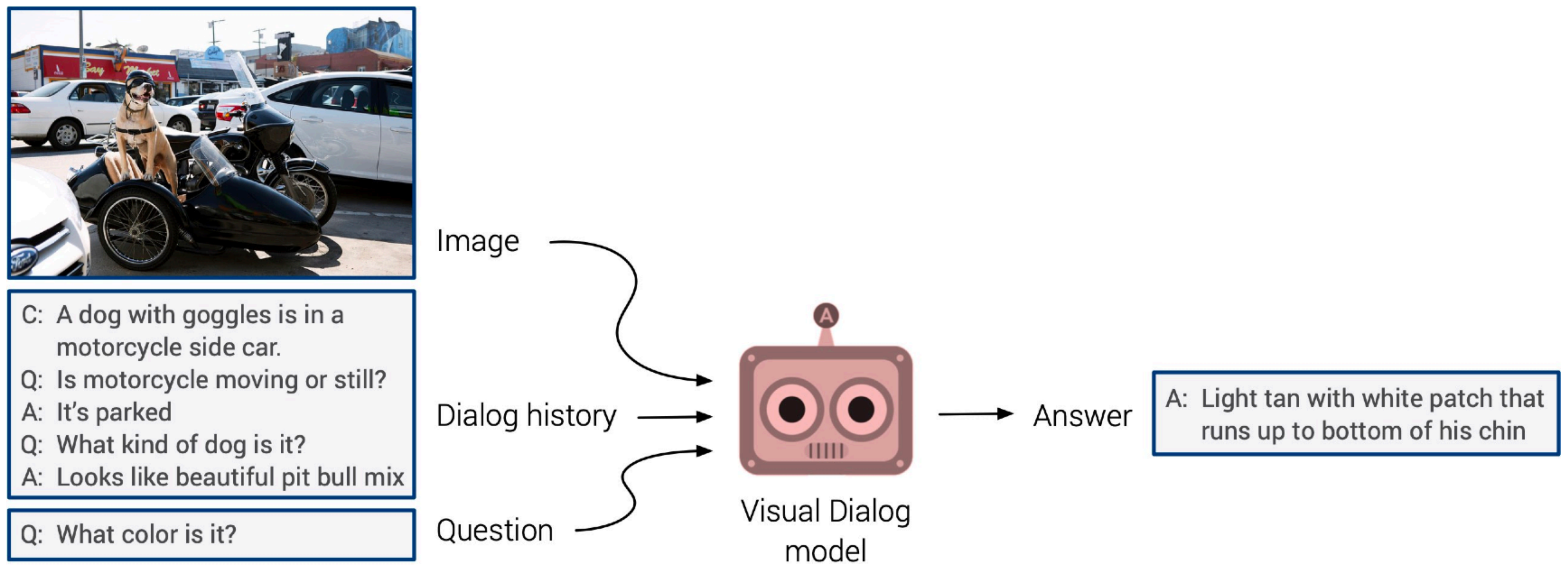


Figure 5. Model architecture for the downstream tasks. Q: question; C: caption; QA: question-answer pair.

Li et al. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. ICML 2022
https://huggingface.co/docs/transformers/model_doc/blip

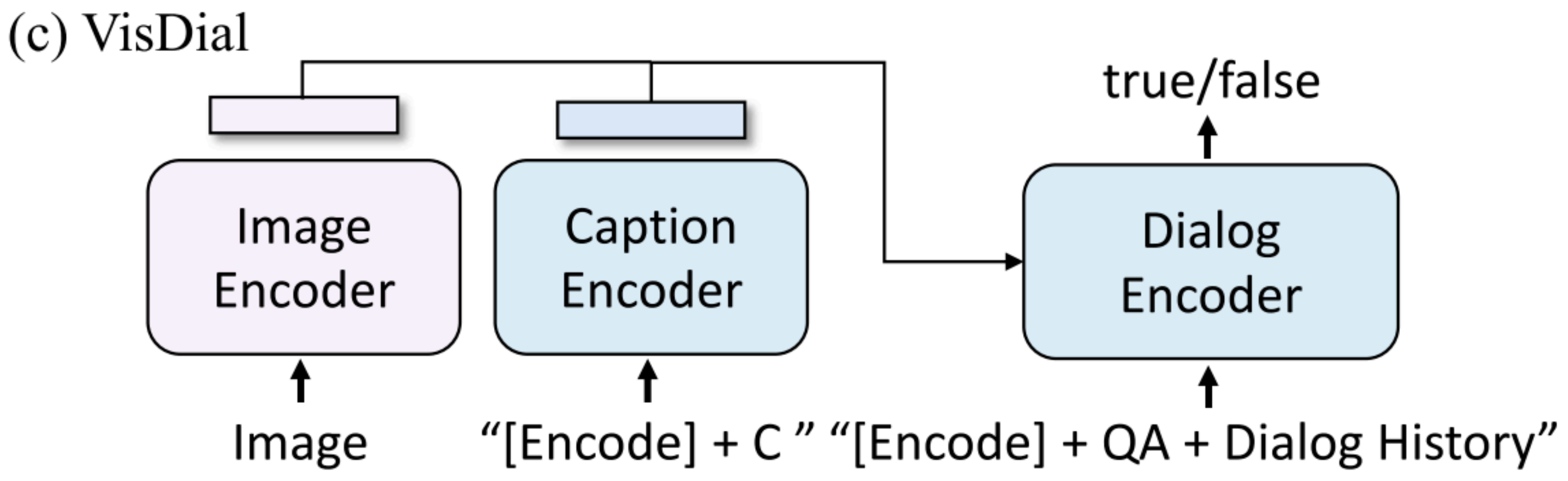
What you can do with visual-language models: Multi-modal understanding, e.g.

VisDial (here: discriminative)


C: A dog with goggles is in a motorcycle side car.
Q: Is motorcycle moving or still?
A: It's parked
Q: What kind of dog is it?
A: Looks like beautiful pit bull mix

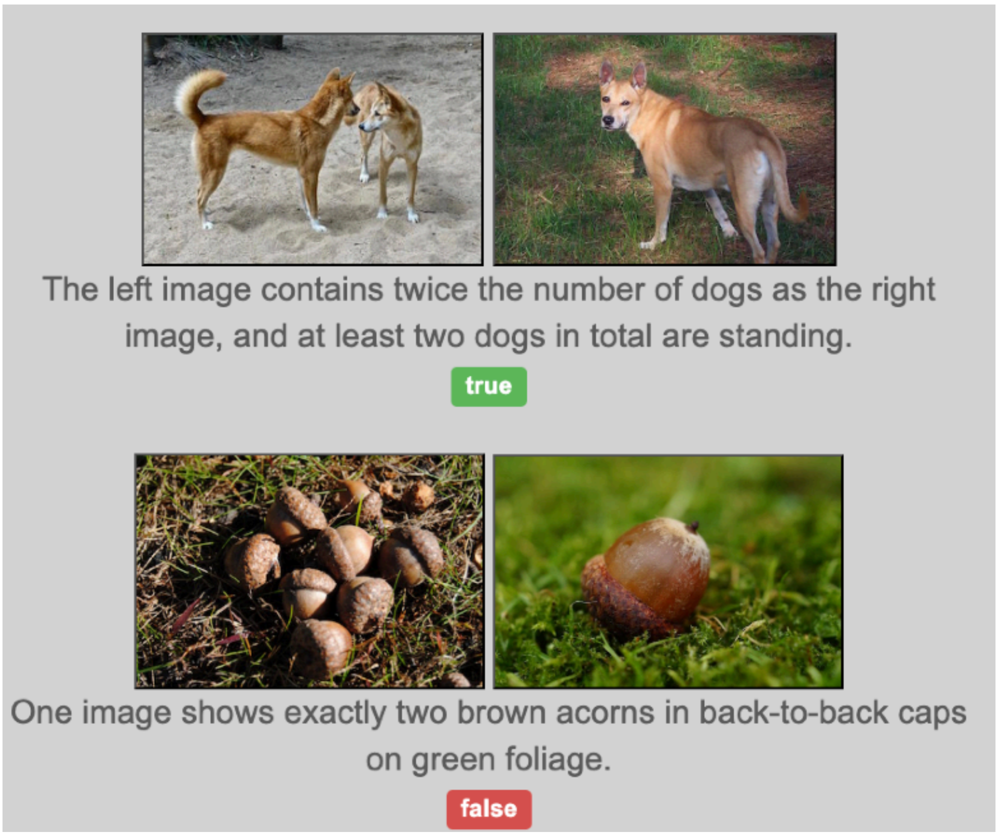
Q: What color is it?

A: Light tan with white patch that runs up to bottom of his chin



NLVR2

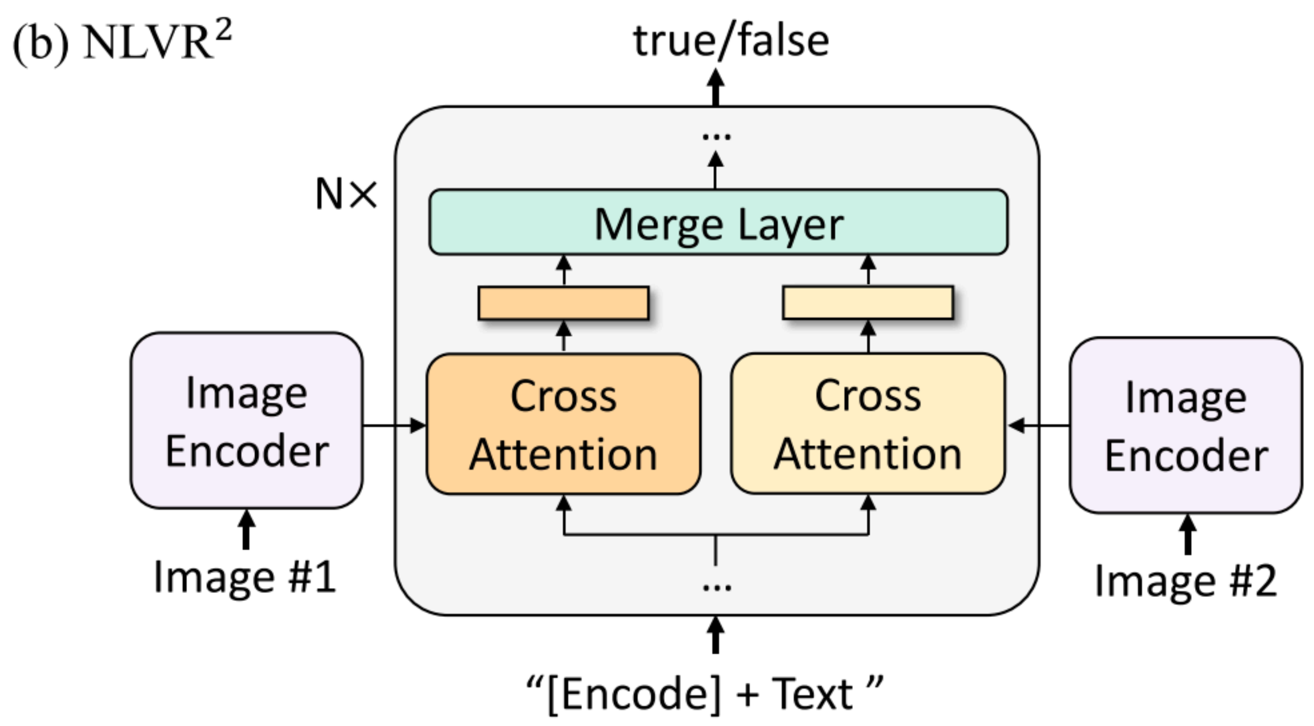
discriminative


The left image contains twice the number of dogs as the right image, and at least two dogs in total are standing.

true

One image shows exactly two brown acorns in back-to-back caps on green foliage.

false



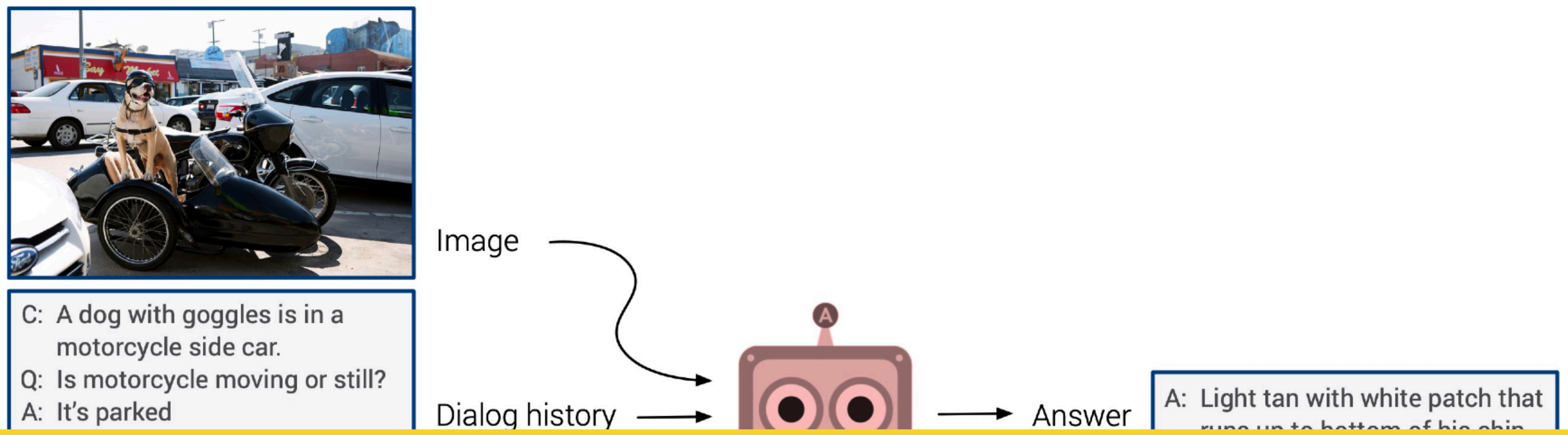
Fundamental AI Lab

UTN

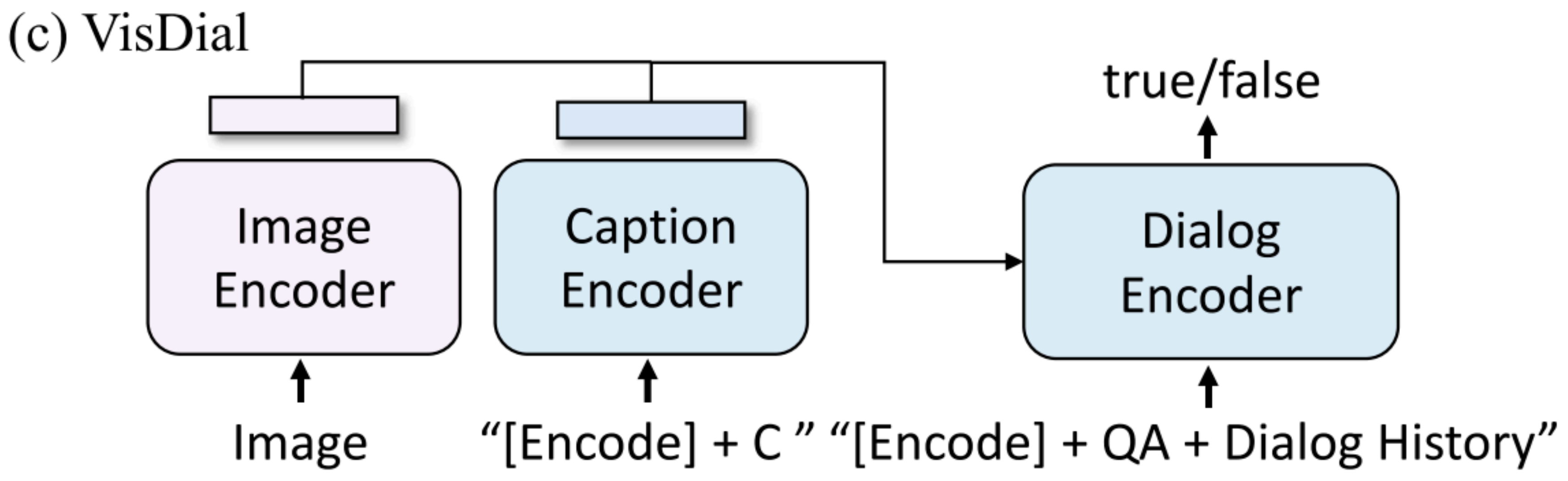
Das et al. Visual dialog. CVPR 2017
Suhr et al. A Corpus of Natural Language for Visual Reasoning. ACL 2017
Suhr et al. A Corpus for Reasoning About Natural Language Grounded in Photographs. ACL 2019

What you can do with visual-language models: Multi-modal understanding, e.g.

VisDial (here: discriminative)

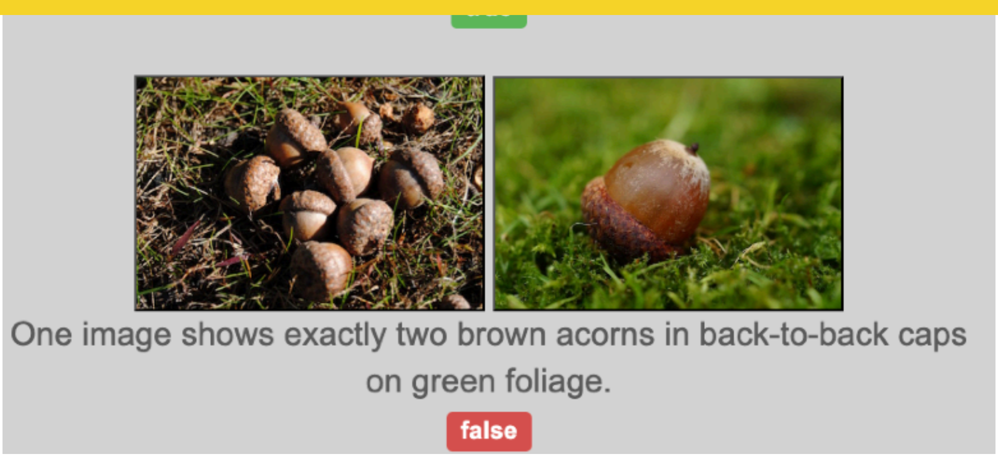

C: A dog with goggles is in a motorcycle side car.
Q: Is motorcycle moving or still?
A: It's parked

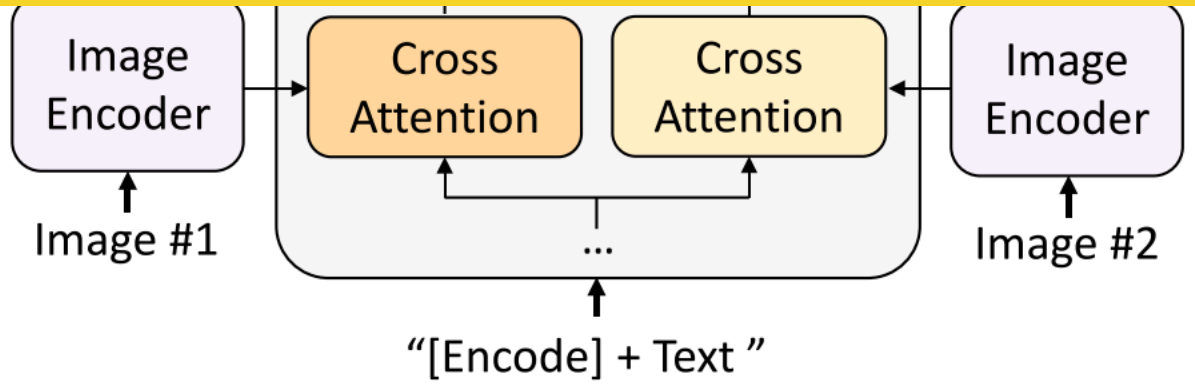
A: Light tan with white patch that



Now-a-days: just solve it in a completely generative way without finetuning:
"Here's two images... answer with 'yes' or 'no' "

discriminative

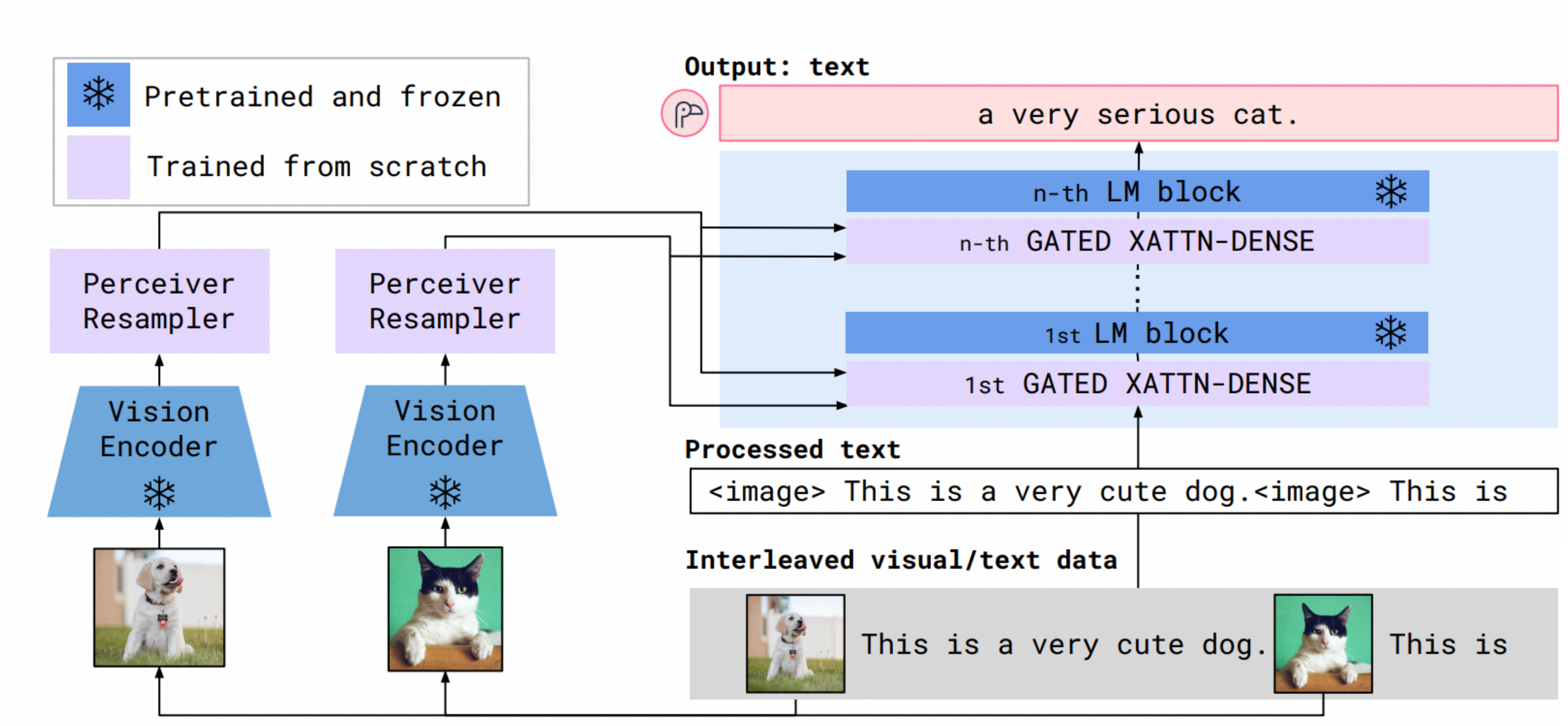

One image shows exactly two brown acorns in back-to-back caps on green foliage.



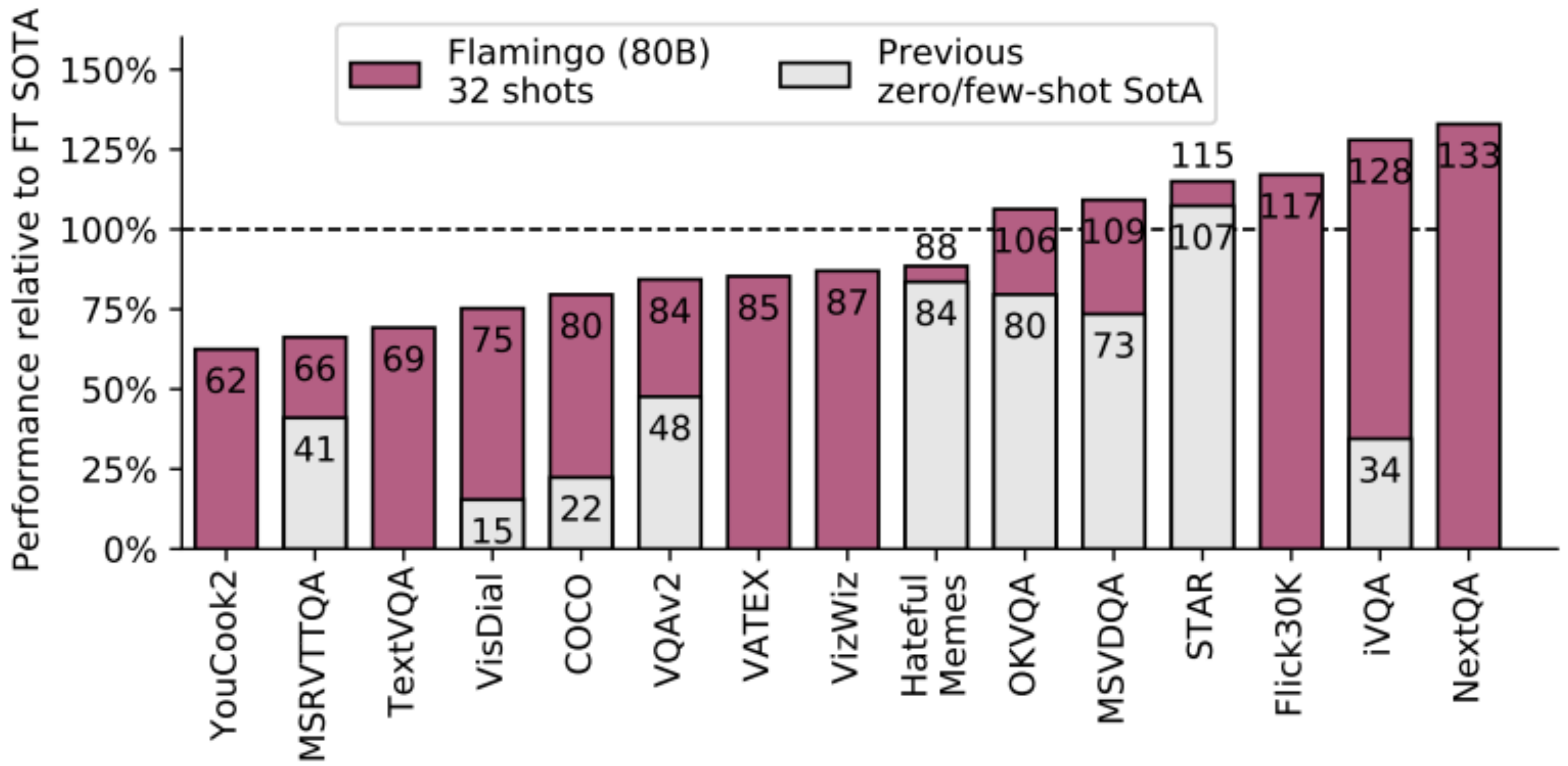
Das et al. Visual dialog. CVPR 2017
Suhr et al. A Corpus of Natural Language for Visual Reasoning. ACL 2017
Suhr et al. A Corpus for Reasoning About Natural Language Grounded in Photographs. ACL 2019

Flamingo: a Visual Language Model for Few-Shot Learning

https://github.com/mlfoundations/open_flamingo

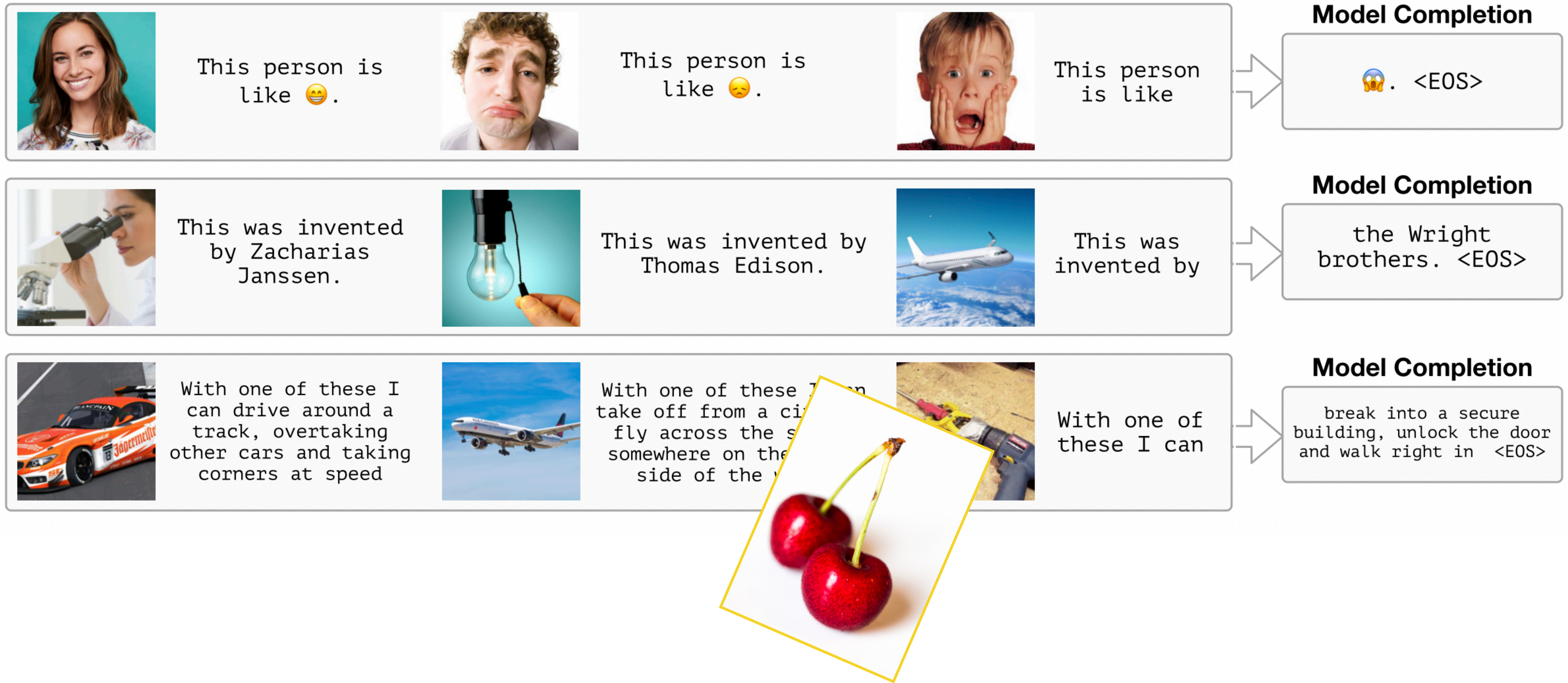


- Uses sota frozen LLM, contrastive pretrained CNN
- Introduces zero-initted learnable attention blocks
- Trained on 43M webpages, each including <=5imgs, plus text + ALIGN's 1.8B text-image pairs + 27M videos
- Uses Perceiver (a transformer) to produced fixed context vision input size
- Very strong performance



Alayrac et al. Flamingo: a Visual Language Model for Few-Shot Learning. NeurIPS 2022

Frozen: Multimodal Few-Shot Learning with Frozen Language Models


This person is like 😁.

This person is like 😞.

This person is like

Model Completion

😱. <EOS>

This was invented by Zacharias Janssen.

This was invented by Thomas Edison.

This was invented by

Model Completion

the Wright brothers. <EOS>

With one of these I can drive around a track, overtaking other cars and taking corners at speed

With one of these I n
take off from a ci
fly across the s
somewhere on the
side of the

With one of these I can

Model Completion

break into a secure building, unlock the door and walk right in <EOS>

Method:

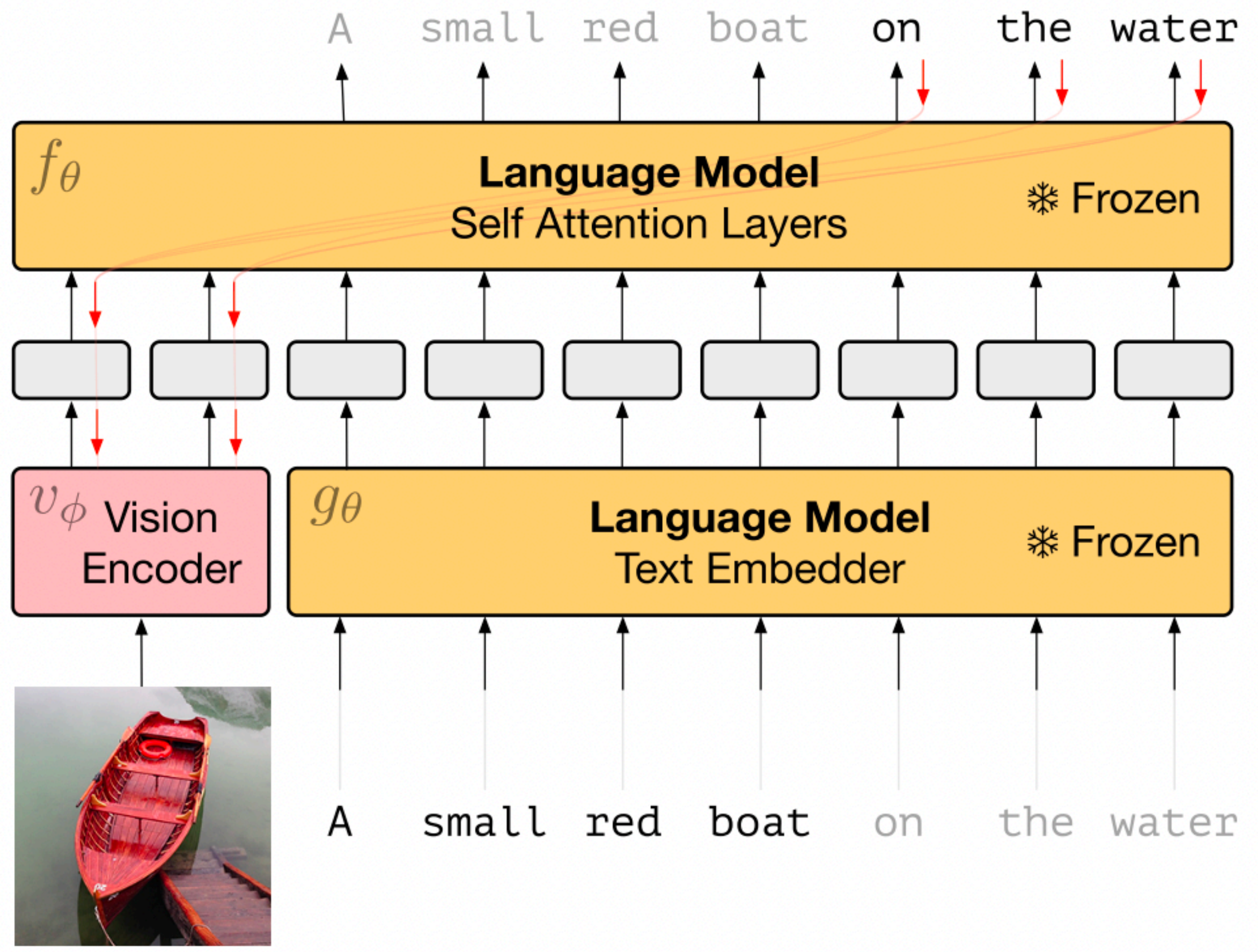


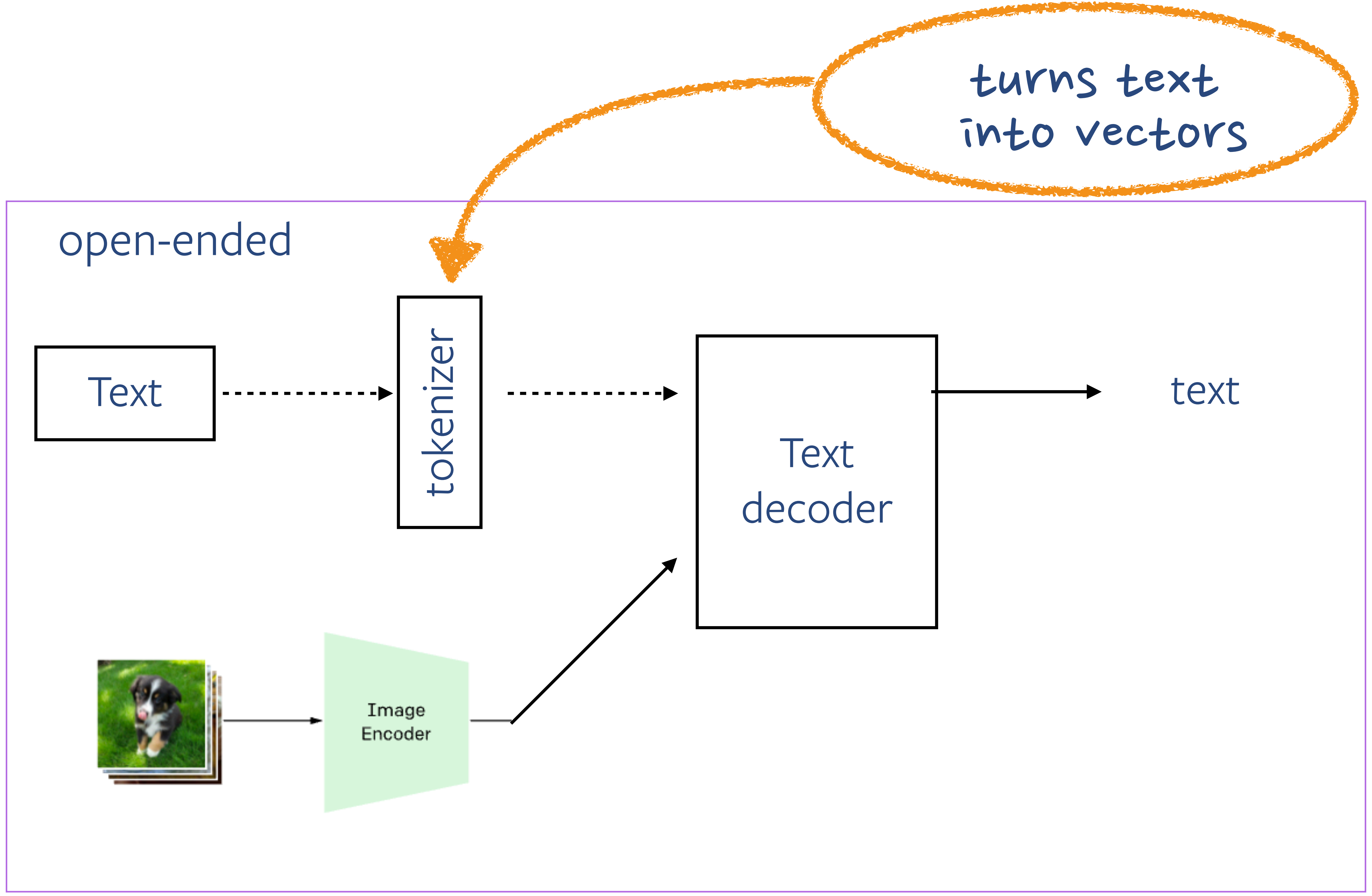
Figure 2: Gradients through a frozen language model's self attention layers are used to train the vision encoder.

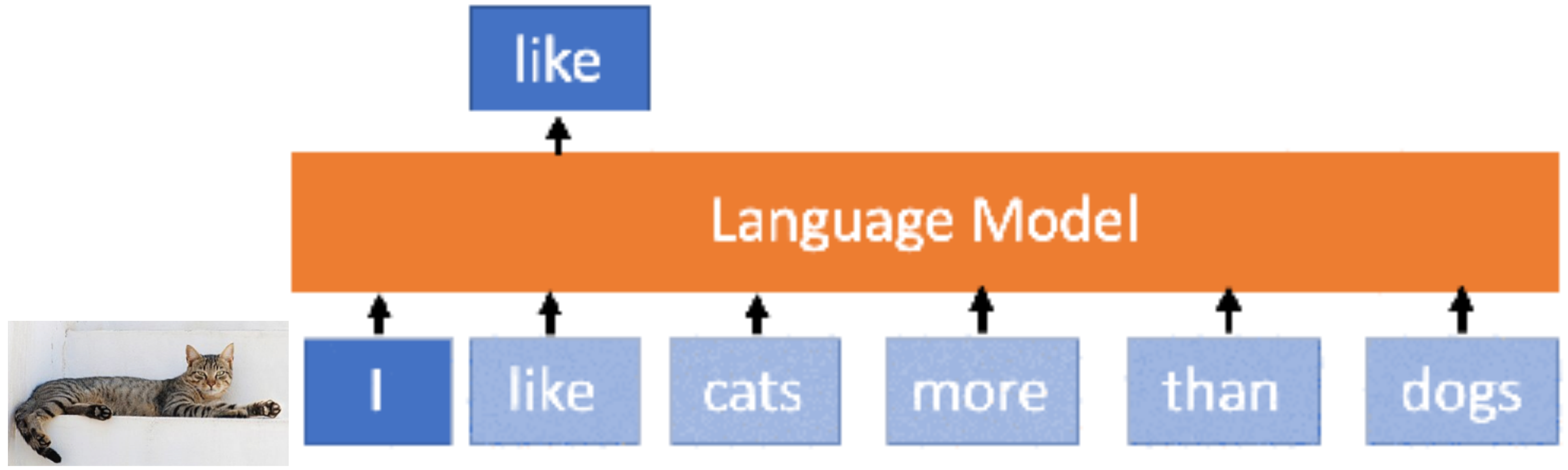
Fundamental AI Lab UTN

Tsimpoukelli et al. Multimodal Few-Shot Learning with Frozen Language Models. NeurIPS 2021.

Llava model family: "the open GPT4V models"

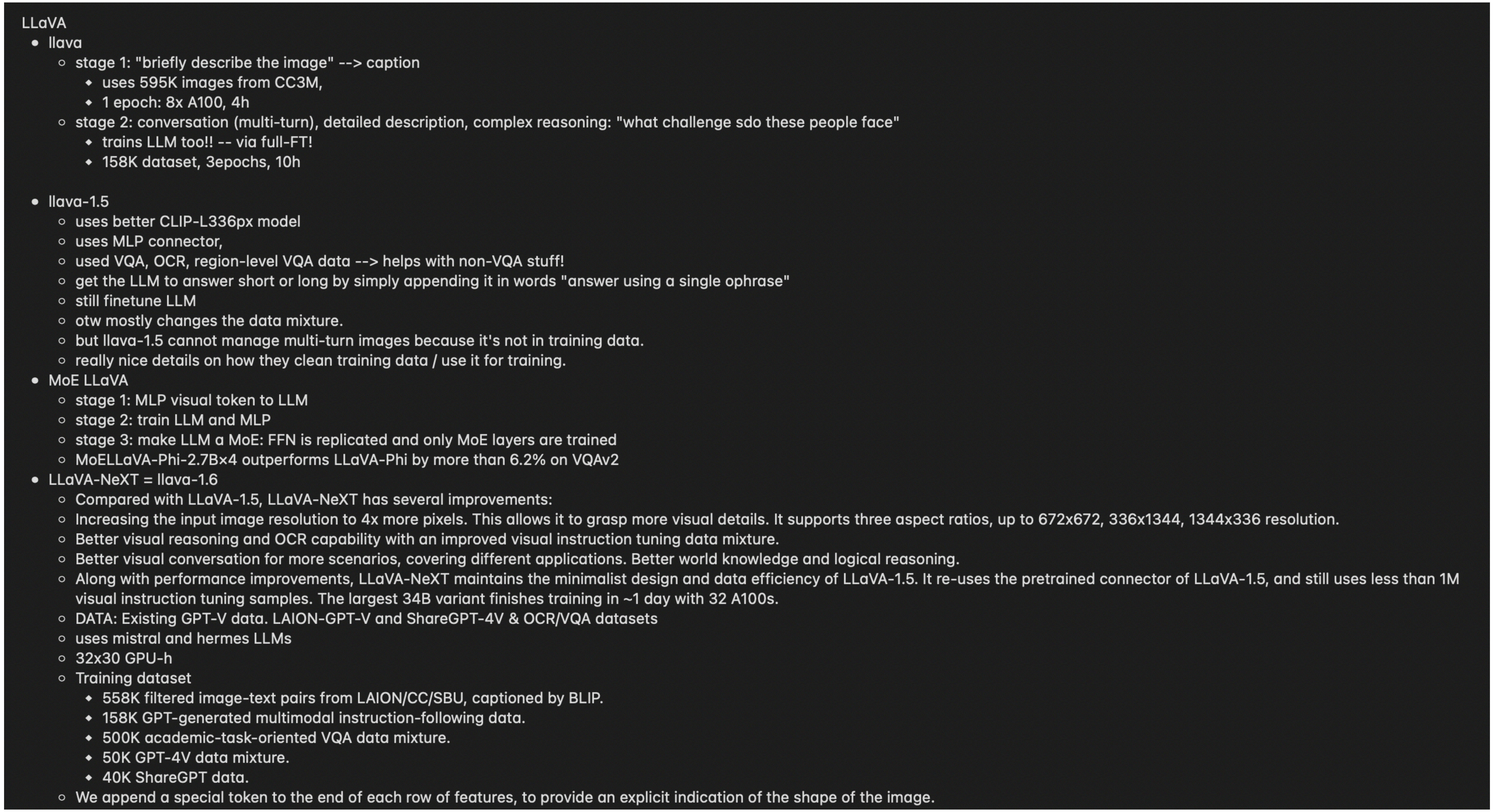
https://llava-vl.github.io/





- Image model stays frozen
- Only a small "projector" network maps image representations to language model input space
- images are treated like "words"
- Llava made multi-stage training popular

Since then…



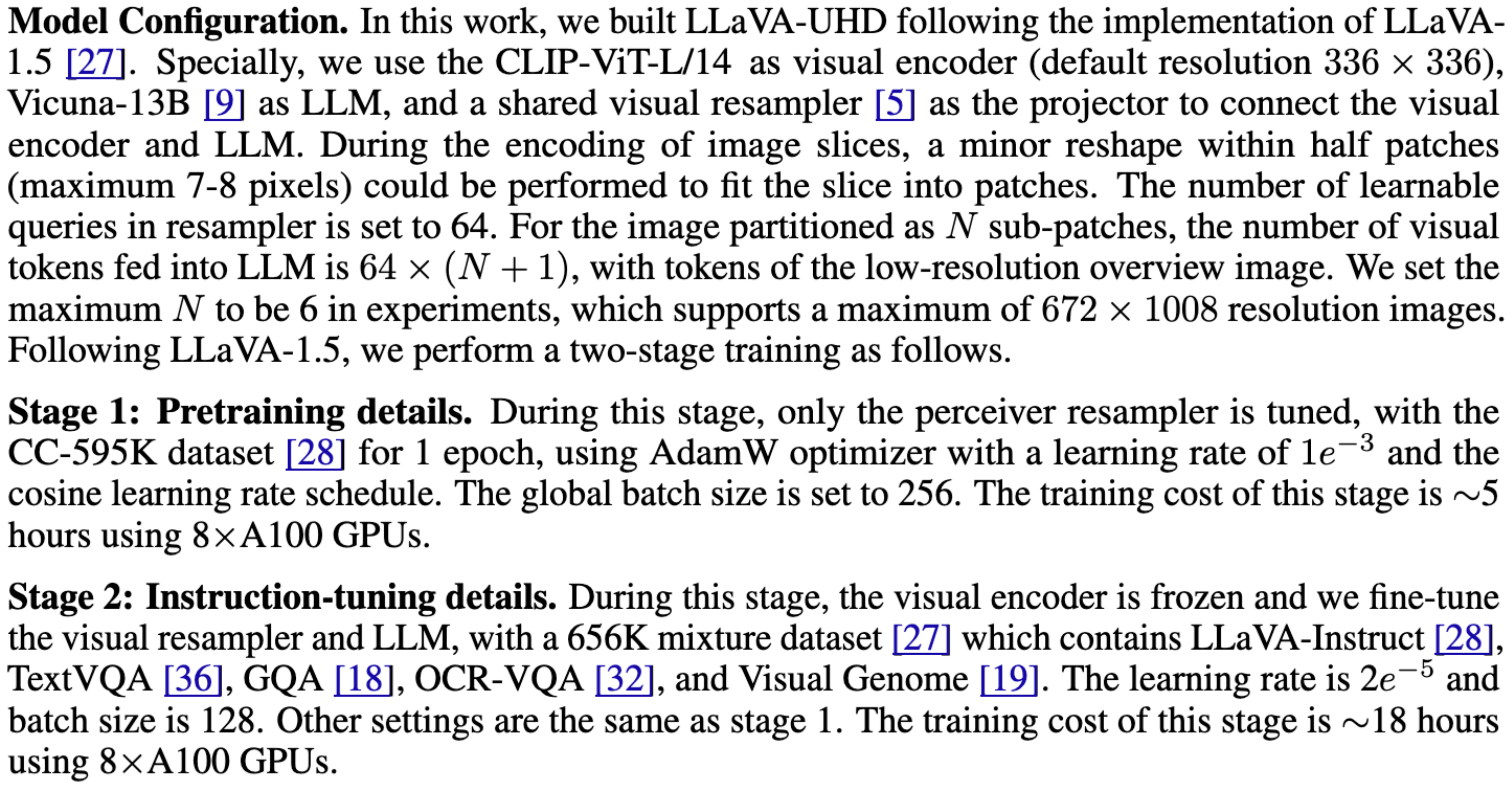
LLaVA

- llava
  - stage 1: "briefly describe the image" --> caption
    - uses 595K images from CC3M,
    - 1 epoch: 8x A100, 4h
  - stage 2: conversation (multi-turn), detailed description, complex reasoning: "what challenge sdo these people face"
    - trains LLM too!! -- via full-FT!
    - 158K dataset, 3epochs, 10h

- llava-1.5
  - uses better CLIP-L336px model
  - uses MLP connector,
  - used VQA, OCR, region-level VQA data --> helps with non-VQA stuff!
  - get the LLM to answer short or long by simply appending it in words "answer using a single ophrase"
  - still finetune LLM
  - otw mostly changes the data mixture.
  - but llava-1.5 cannot manage multi-turn images because it's not in training data.
  - really nice details on how they clean training data / use it for training.
- MoE LLaVA
  - stage 1: MLP visual token to LLM
  - stage 2: train LLM and MLP
  - stage 3: make LLM a MoE: FFN is replicated and only MoE layers are trained
  - MoELLaVA-Phi-2.7B×4 outperforms LLaVA-Phi by more than 6.2% on VQAv2
- LLaVA-NeXT = llava-1.6
  - Compared with LLaVA-1.5, LLaVA-NeXT has several improvements:
  - Increasing the input image resolution to 4x more pixels. This allows it to grasp more visual details. It supports three aspect ratios, up to 672x672, 336x1344, 1344x336 resolution.
  - Better visual reasoning and OCR capability with an improved visual instruction tuning data mixture.
  - Better visual conversation for more scenarios, covering different applications. Better world knowledge and logical reasoning.
  - Along with performance improvements, LLaVA-NeXT maintains the minimalist design and data efficiency of LLaVA-1.5. It re-uses the pretrained connector of LLaVA-1.5, and still uses less than 1M visual instruction tuning samples. The largest 34B variant finishes training in ~1 day with 32 A100s.
  - DATA: Existing GPT-V data. LAION-GPT-V and ShareGPT-4V & OCR/VQA datasets
  - uses mistral and hermes LLMs
  - 32x30 GPU-h
  - Training dataset
    - 558K filtered image-text pairs from LAION/CC/SBU, captioned by BLIP.
    - 158K GPT-generated multimodal instruction-following data.
    - 500K academic-task-oriented VQA data mixture.
    - 50K GPT-4V data mixture.
    - 40K ShareGPT data.
  - We append a special token to the end of each row of features, to provide an explicit indication of the shape of the image.

+ more

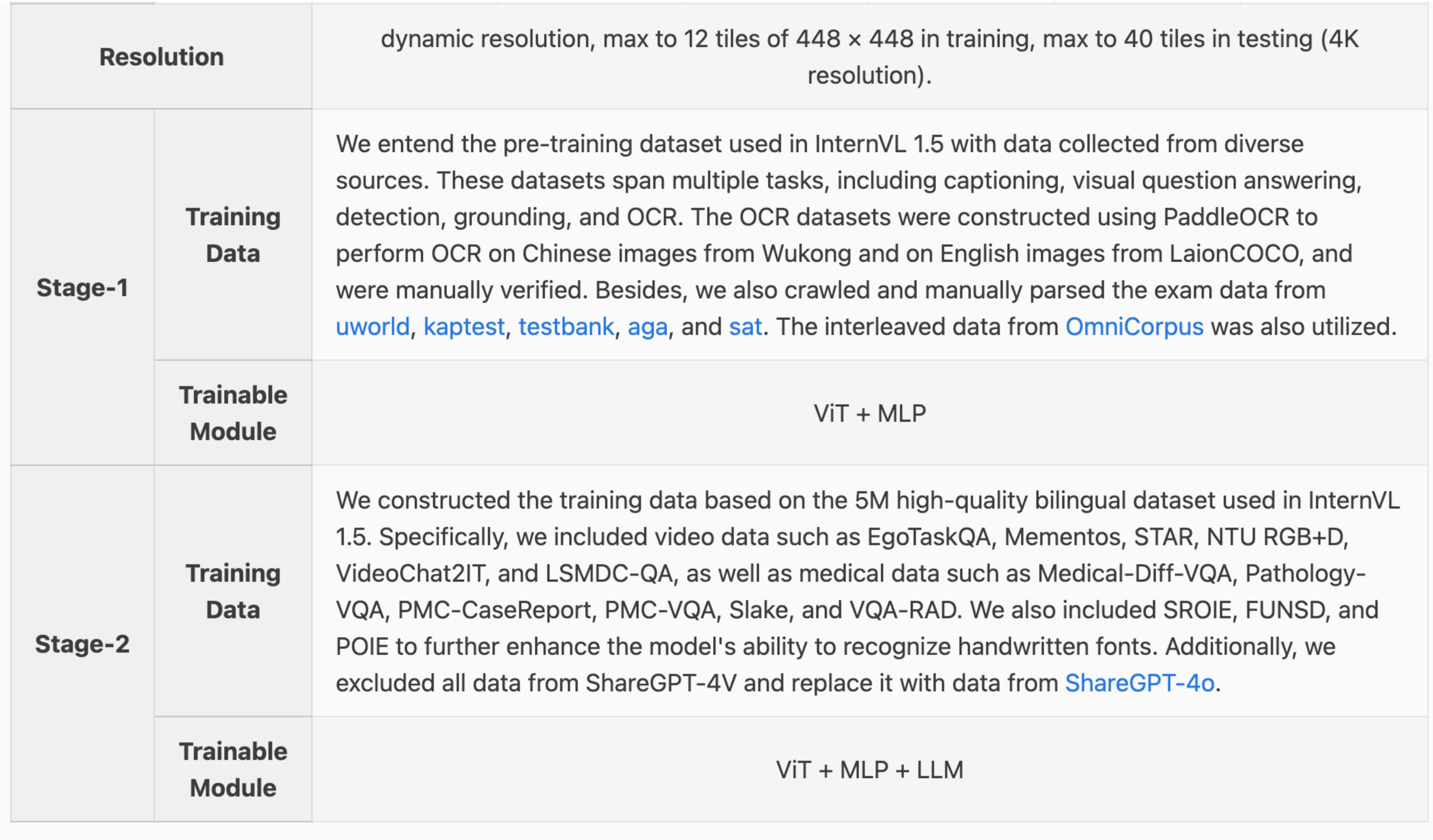
Training data and stages. Two examples



Model Configuration. In this work, we built LLaVA-UHD following the implementation of LLaVA-1.5 [27]. Specially, we use the CLIP-ViT-L/14 as visual encoder (default resolution 336×336), Vicuna-13B [9] as LLM, and a shared visual resampler [5] as the projector to connect the visual encoder and LLM. During the encoding of image slices, a minor reshape within half patches (maximum 7-8 pixels) could be performed to fit the slice into patches. The number of learnable queries in resampler is set to 64. For the image partitioned as N sub-patches, the number of visual tokens fed into LLM is $64 \times (N + 1)$, with tokens of the low-resolution overview image. We set the maximum N to be 6 in experiments, which supports a maximum of 672×1008 resolution images. Following LLaVA-1.5, we perform a two-stage training as follows.

Stage 1: Pretraining details. During this stage, only the perceiver resampler is tuned, with the CC-595K dataset [28] for 1 epoch, using AdamW optimizer with a learning rate of $1e^{-3}$ and the cosine learning rate schedule. The global batch size is set to 256. The training cost of this stage is $\sim$5 hours using 8×A100 GPUs.

Stage 2: Instruction-tuning details. During this stage, the visual encoder is frozen and we fine-tune the visual resampler and LLM, with a 656K mixture dataset [27] which contains LLaVA-Instruct [28], TextVQA [36], GQA [18], OCR-VQA [32], and Visual Genome [19]. The learning rate is $2e^{-5}$ and batch size is 128. Other settings are the same as stage 1. The training cost of this stage is $\sim$18 hours using 8×A100 GPUs.



Resolution		dynamic resolution, max to 12 tiles of 448 × 448 in training, max to 40 tiles in testing (4K resolution).
Stage-1	**Training Data**	We entend the pre-training dataset used in InternVL 1.5 with data collected from diverse sources. These datasets span multiple tasks, including captioning, visual question answering, detection, grounding, and OCR. The OCR datasets were constructed using PaddleOCR to perform OCR on Chinese images from Wukong and on English images from LaionCOCO, and were manually verified. Besides, we also crawled and manually parsed the exam data from uworld, kaptest, testbank, aga, and sat. The interleaved data from OmniCorpus was also utilized.
	Trainable Module	ViT + MLP
Stage-2	**Training Data**	We constructed the training data based on the 5M high-quality bilingual dataset used in InternVL 1.5. Specifically, we included video data such as EgoTaskQA, Mementos, STAR, NTU RGB+D, VideoChat2IT, and LSMDC-QA, as well as medical data such as Medical-Diff-VQA, Pathology-VQA, PMC-CaseReport, PMC-VQA, Slake, and VQA-RAD. We also included SROIE, FUNSD, and POIE to further enhance the model's ability to recognize handwritten fonts. Additionally, we excluded all data from ShareGPT-4V and replace it with data from ShareGPT-4o.
	Trainable Module	ViT + MLP + LLM

Xu et al. LaVA-UHD: an LMM Perceiving Any Aspect Ratio and High-Resolution Images. 2024
Chen et al. How Far Are We to GPT-4V? Closing the Gap to Commercial Multimodal Models with Open-Source Suites. 2024

Current basic principles

- ”Align” modalities via captioning task + limited training (e.g. only projector)
- Then include more complex tasks, e.g. VQA, object-loc etc. + train LM too
- Add high-resolution training stage, e.g. with OCR/ document tasks + adapt visual model
- Convert training data into ”chat-like” format
- Keep everything as general as possible by sticking to language outputs
  - e.g. object-localisation: ”where’s the dog? --> It’s at [25,50,70,120].”
- Better datasets matter immensely. Using synthetic data from GPT4 is therefore popular

Current basic principles

- ”Align” modalities via captioning task + limited training (e.g. only projector)
- Then include more complex tasks, e.g. VQA, object-loc etc. + train LM too
- Add high-resolution training stage, e.g. with OCR/ document tasks + adapt visual model
- Convert training data into ”chat-like” format
- Keep everything as general as possible by sticking to language outputs
  - e.g. object-localisation: ”where’s the dog? --> It’s at [25,50,70,120].”
- Better datasets matter immensely. Using synthetic data from GPT4 is therefore popular

Fundamental AI Lab UTN

Example of text & image generative LLM:
CM3: A Causal Masked Multimodal Model of the Internet

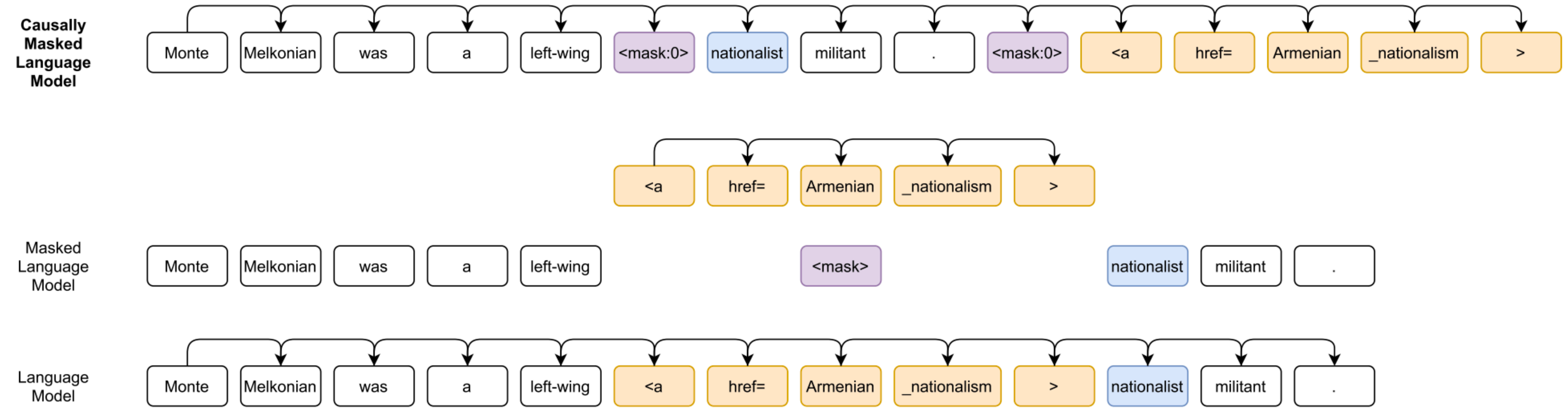


Figure 1: A visual representation of various language modeling objectives as well as our proposed causal language modeling objective with a single mask ($n = 1$). Given the left-to-right nature of causal language models (bottom row) we would not be able to generate the Wikipedia entity link highlighted in orange.

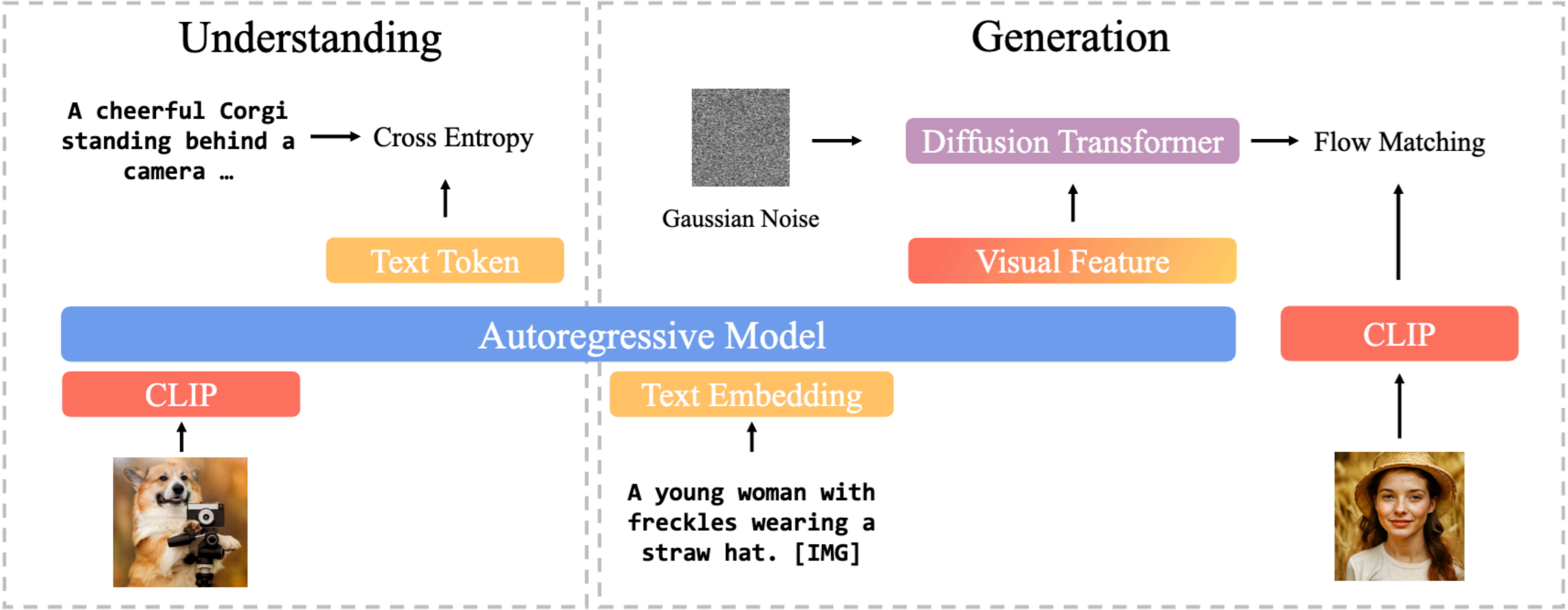
- trained on 1TB of webpages with images
- images encoded as VQ-VAE-GAN tokens
- It can therefore create new images, e.g. by prompting with *<img src="*

`<img src="`

Aghajanyan et al. CM3: A Causal Masked Multimodal Model of the Internet. 2022

384 A100 GPU for 24 days

(GPT4o?) / Metamorph / BLIP3o



Fundamental AI Lab UTN

Chen, Xu, Pan, Hu, Qin, Goldstein, Huang, Zhou, Xie, Savarese, Xue, Xiong, Xu. BLIP3-o: A Family of Fully Open Unified Multimodal Models—Architecture, Training and Dataset. 2025

Recap

Single-modal self-supervised pretraining methods (DINOv2/v3, Franca, MAE, SimCLR, GPT)

Multi-modal pretraining (CLIP, ALIGN, CoCa)

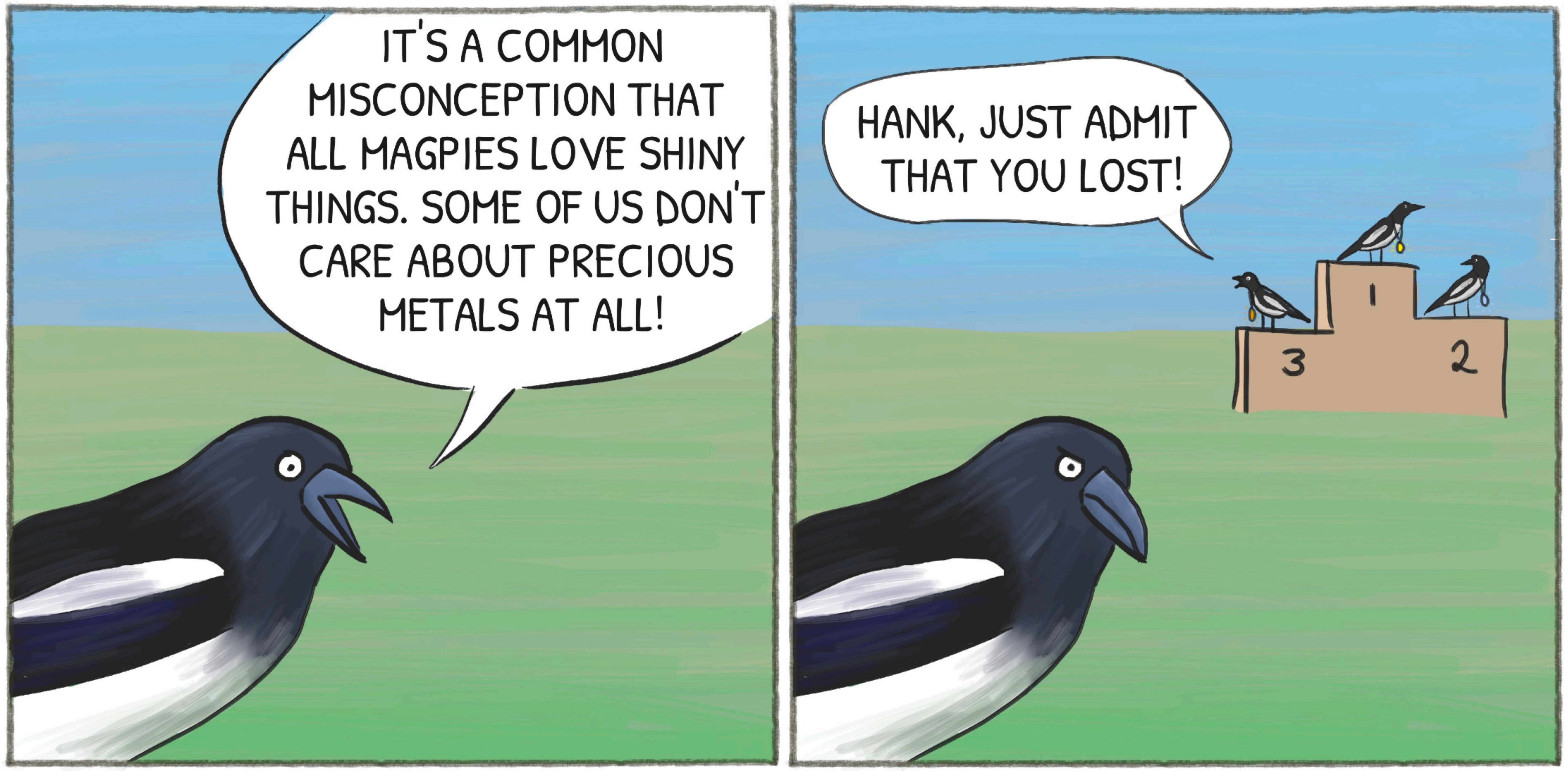
Beyond contrastive (BLIP, ClipCap)

Multimodal Large Language Models (Frozen, Flamingo, Llava, BLIP3o)

Tasks (VQA, VisDial)

Multimodal
In-context learning

towards more useful systems


IT'S A COMMON MISCONCEPTION THAT ALL MAGPIES LOVE SHINY THINGS. SOME OF US DON'T CARE ABOUT PRECIOUS METALS AT ALL!
HANK, JUST ADMIT THAT YOU LOST!
1
2
3

BONUSCONTEXT.COM

context matters

© BONUS CONTEXT

Vision-language in-context learning (ICL)

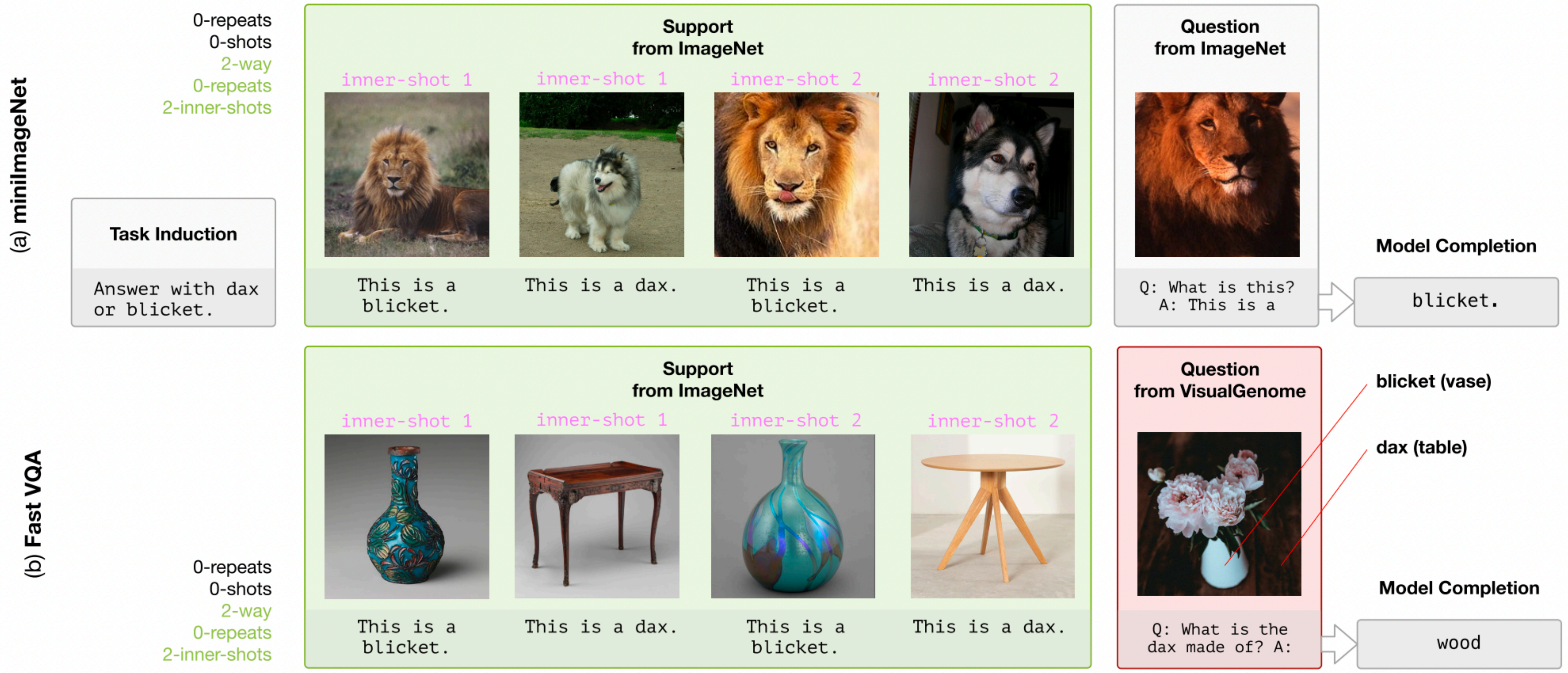


Figure 4: Examples of (a) the Open-Ended miniImageNet evaluation (b) the Fast VQA evaluation.

- Here, ICL is short for something like "open-ended vision-language few-shot evaluation"
- Open-ended: it needs to infer what it's supposed to do & what the answer style is.
- Vision-language: it needs to process both the image & the text info
- Few-shot: few-shot samples "support set" are provided as input, along with the test sample
- "fast-binding": text & image are associated within the single forward pass

Tsimpoukelli et al. Multimodal Few-Shot Learning with Frozen Language Models. NeurIPS 2021.

In-context learning in vision-language models is cool

- Especially because models like Frozen, Flamingo, FROMAGe weren't explicitly trained for in-context learning
- But Flamingo and CM3 were trained with websites,
  - so samples that resemble in-context learning might be frequent
  - but the same is true for LLMs
- So these VL models obtain a significant (and useful part) of their ability from the language models

--> studying language models (and related papers) useful!

Quiz: turn to your neighbour and briefly explain the core idea behind in-context learning

Food for thought:

What are the core principles and ideas?

What is the difference to zero-shot learning?

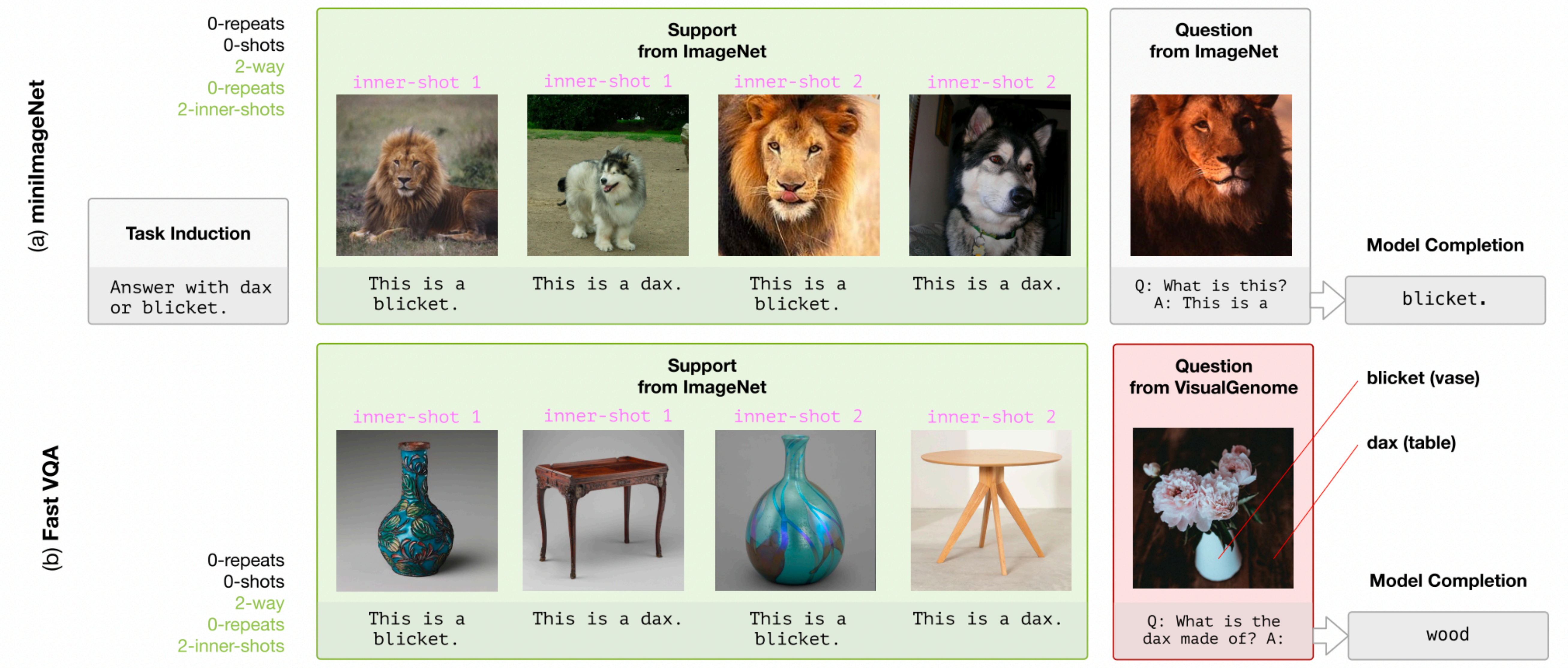
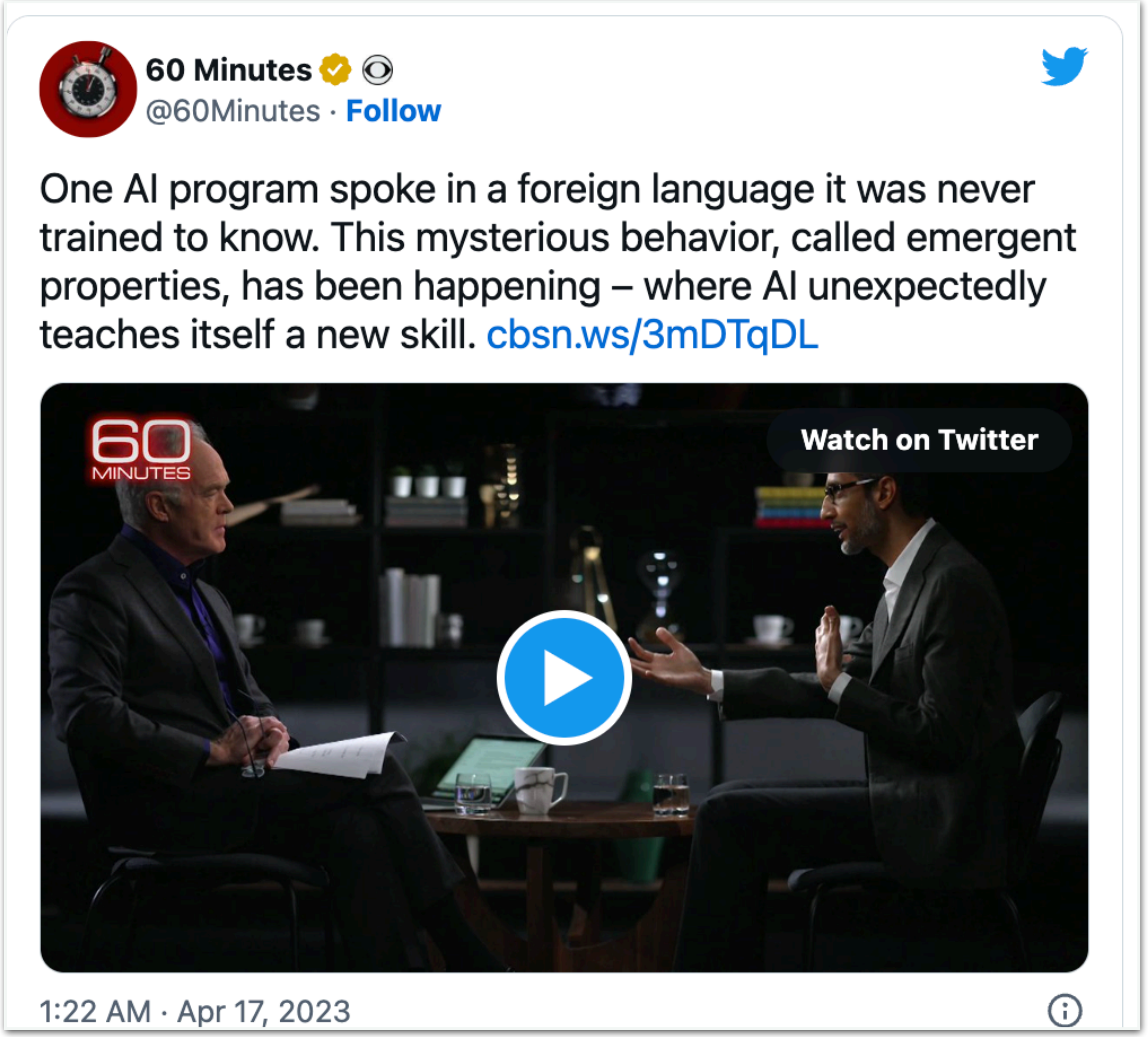


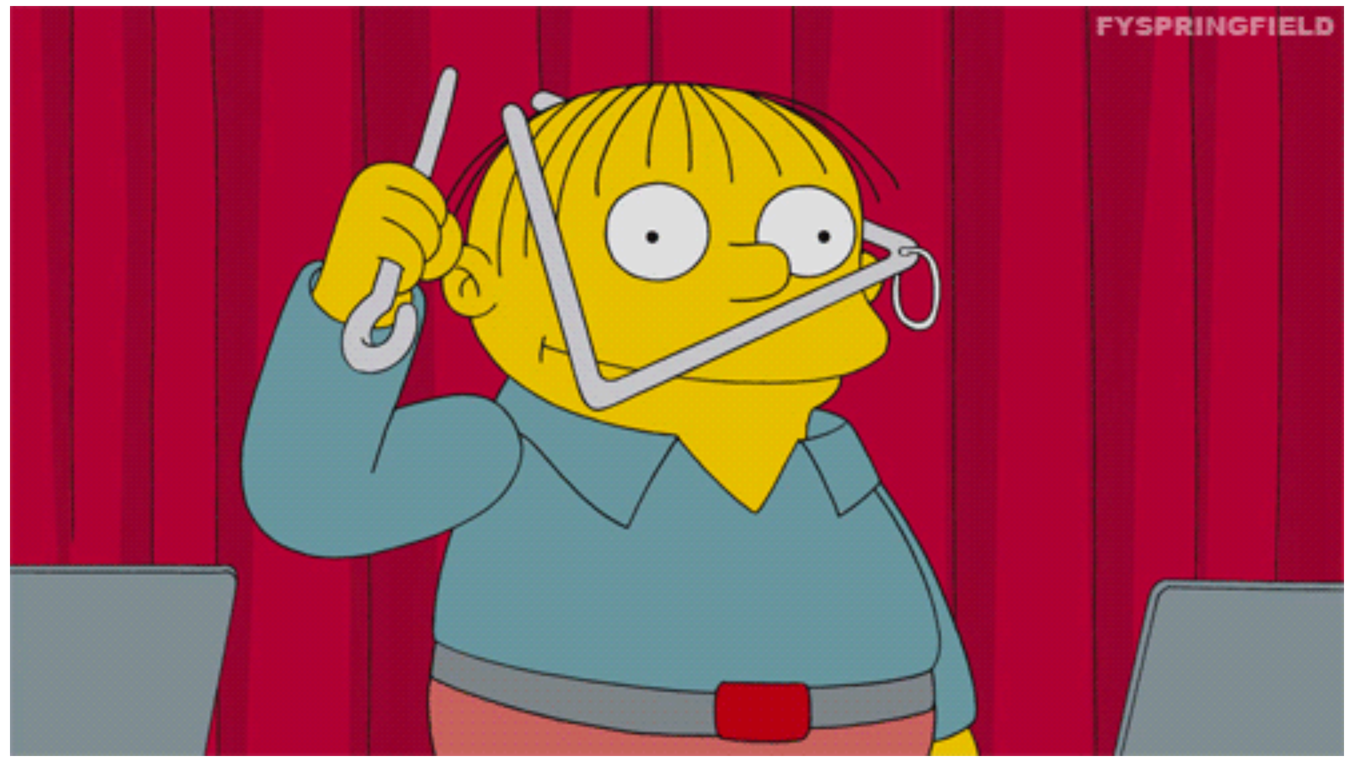
Figure 4: Examples of (a) the Open-Ended miniImageNet evaluation (b) the Fast VQA evaluation.

Careful about AI hype + news


60 Minutes @60Minutes · Follow

One AI program spoke in a foreign language it was never trained to know. This mysterious behavior, called emergent properties, has been happening – where AI unexpectedly teaches itself a new skill. cbsn.ws/3mDTqDL

1:22 AM · Apr 17, 2023



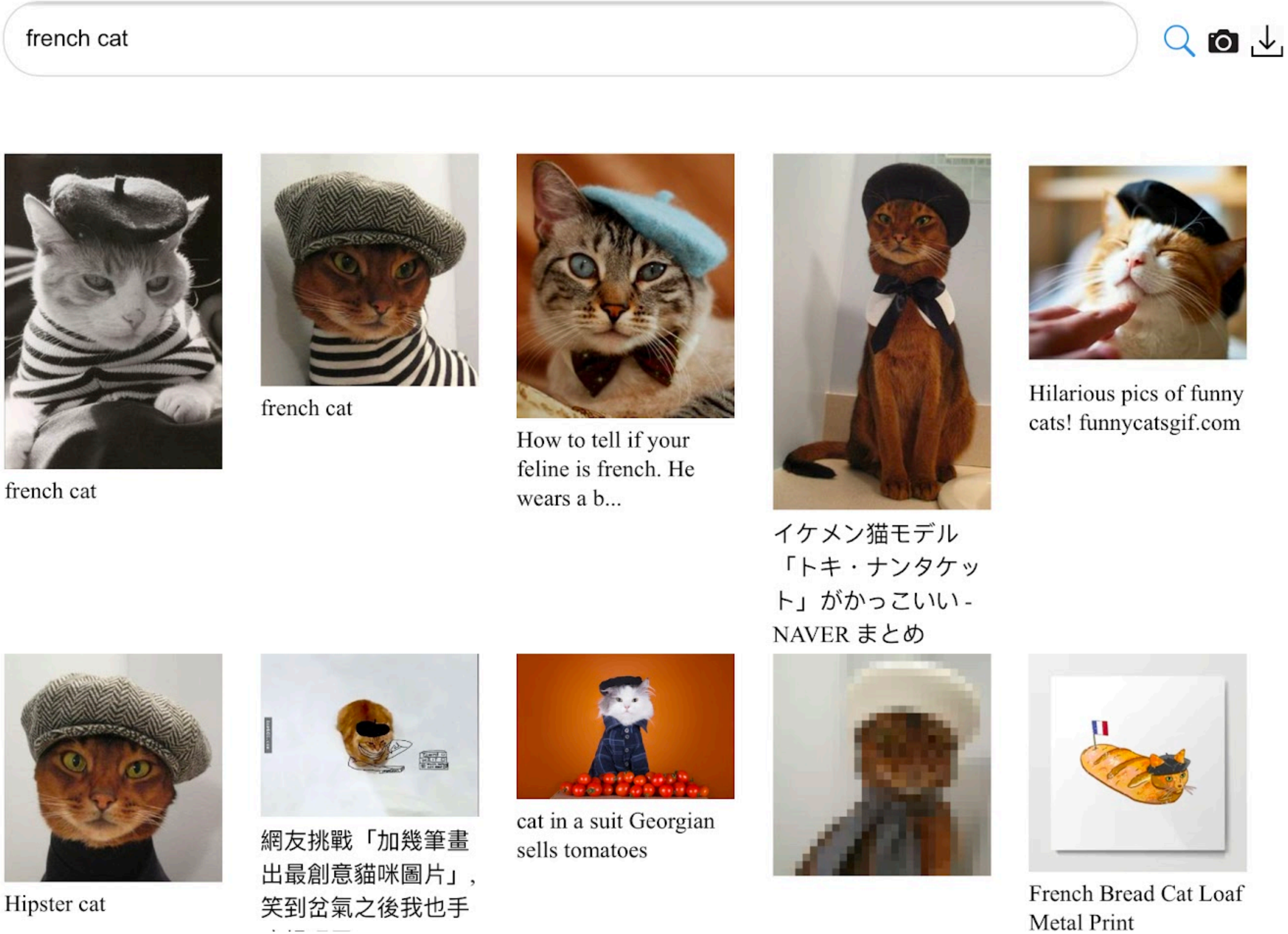
- Easy to complain about this sort of stuff (see also DL1)
- But why does this keep happening?

My two cents:

- Deep learning in industry is increasingly a marketing battle
- Moreover, companies do not have much incentive to really document/ analyse their training data

more on this: https://twitter.com/mmitchell_ai/status/1648029417497853953

Vision-Language Datasets



french cat

french cat

french cat

How to tell if your feline is french. He wears a b...

イケメン猫モデル「トキ・ナンタケット」がかっこいい - NAVER まとめ

Hilarious pics of funny cats! funnycatsgif.com

Hipster cat

網友挑戰「加幾筆畫出最創意貓咪圖片」，笑到岔氣之後我也手

cat in a suit Georgian sells tomatoes

French Bread Cat Loaf Metal Print

LAION: Large-scale Artificial Intelligence Open Networ

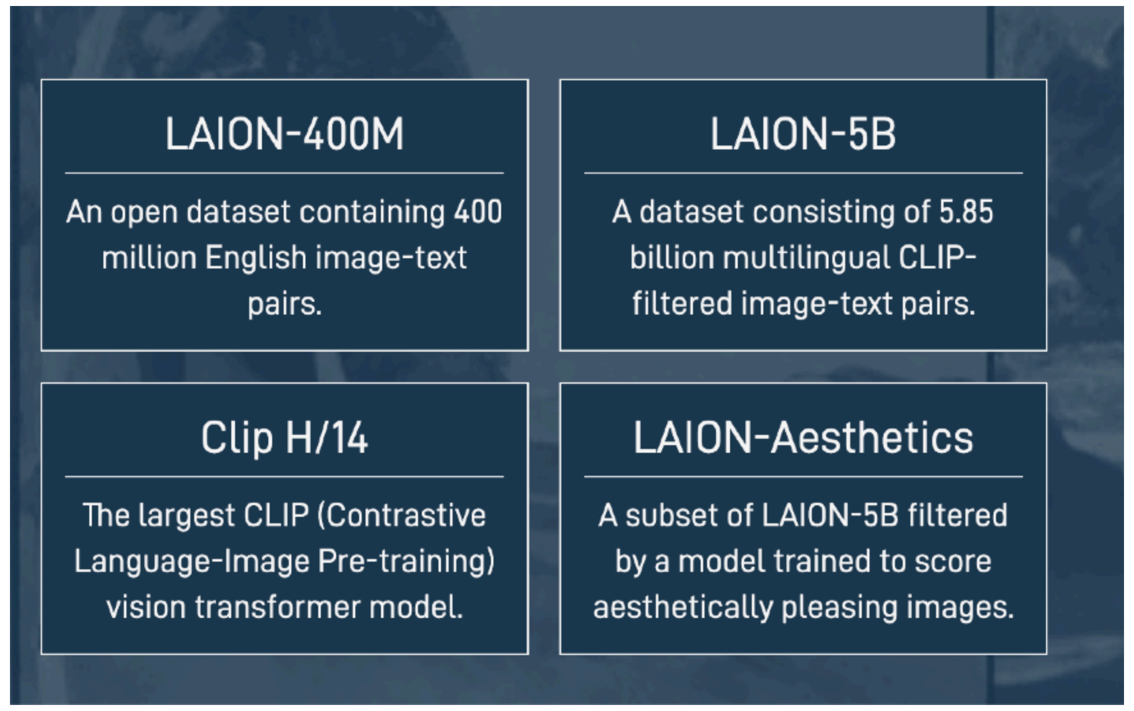
had to be taken down because of highly illegal/ harmful content

Use "dump of internet": Common Crawl

CLIP-based filtering ~90% removed, yielding ~6 billion

Further filtering of NSFW, watermarked images

Training dataset for generative models like Stable Diffusion

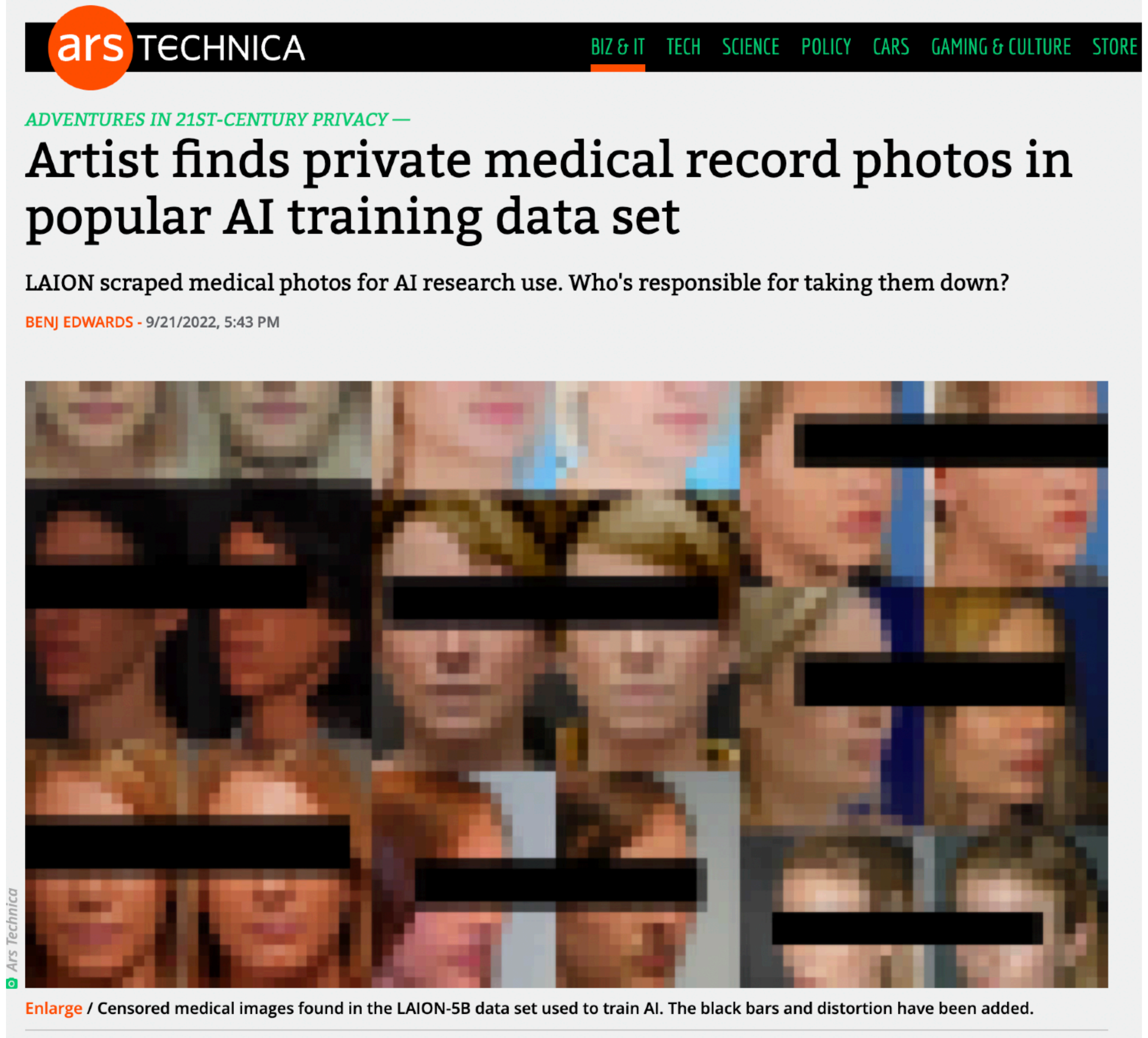


LAION-400M
An open dataset containing 400 million English image-text pairs.

LAION-5B
A dataset consisting of 5.85 billion multilingual CLIP-filtered image-text pairs.

Clip H/14
The largest CLIP (Contrastive Language-Image Pre-training) vision transformer model.

LAION-Aesthetics
A subset of LAION-5B filtered by a model trained to score aesthetically pleasing images.



ADVENTURES IN 21ST-CENTURY PRIVACY —

Artist finds private medical record photos in popular AI training data set

LAION scraped medical photos for AI research use. Who's responsible for taking them down?

BENJ EDWARDS - 9/21/2022, 5:43 PM

Enlarge / Censored medical images found in the LAION-5B data set used to train AI. The black bars and distortion have been added.

Schumann et al. LAION-5B: An open large-scale dataset for training next generation image-text models. NeurIPS-Data 2022
https://arstechnica.com/information-technology/2022/09/artist-finds-private-medical-record-photos-in-popular-ai-training-data-set/

Demo

https://rom1504.github.io/clip-retrieval (doesn’t work atm)

Explore some search terms. What sort of content do you find?

During a break: discuss with your collegues the pros and cons of the dataset.

Conceptual Captions (CC3M, CC12M)

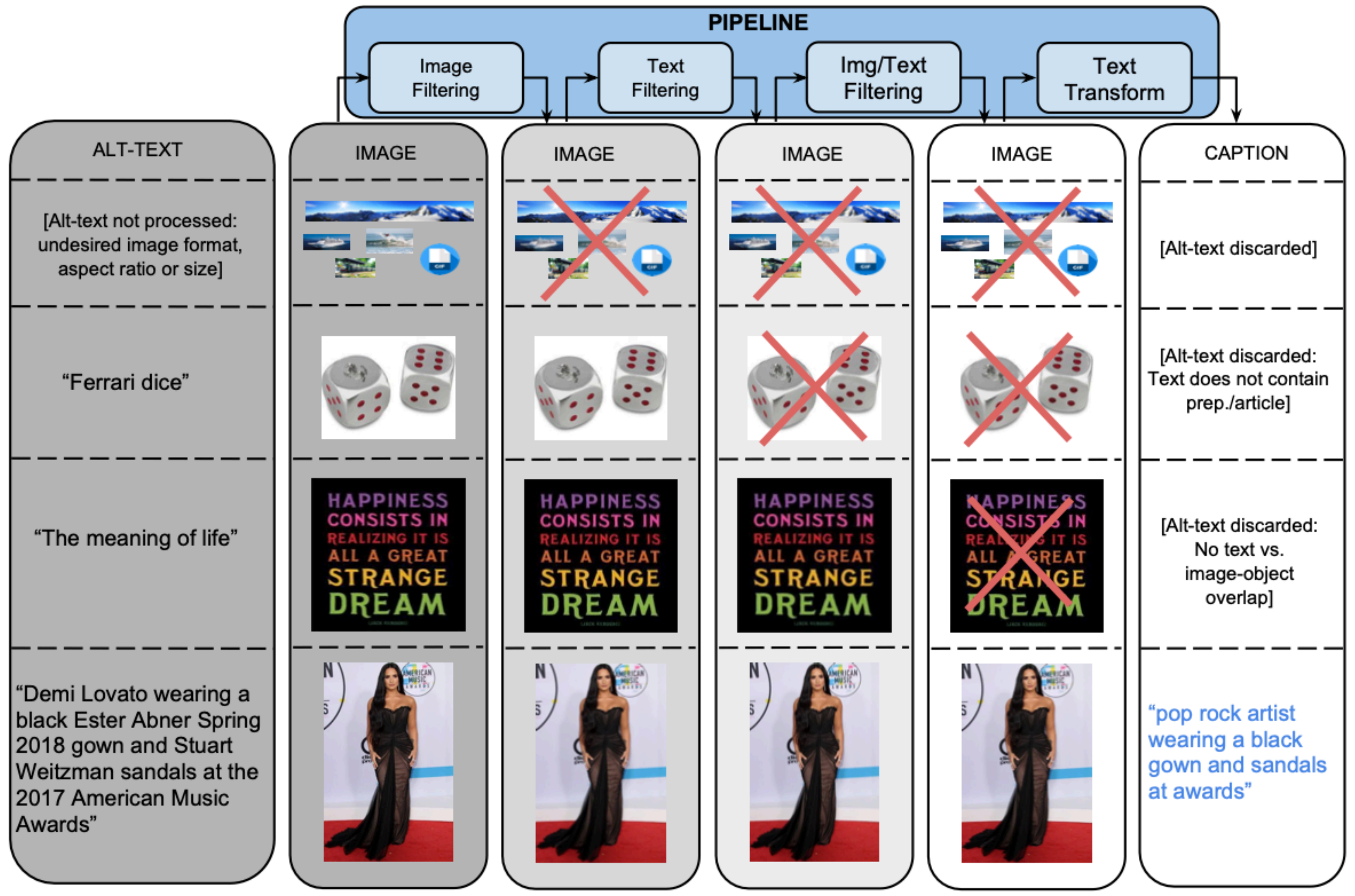

PIPELINE
Image Filtering
Text Filtering
Img/Text Filtering
Text Transform
ALT-TEXT
IMAGE
IMAGE
IMAGE
IMAGE
CAPTION
[Alt-text not processed: undesired image format, aspect ratio or size]
[Alt-text discarded]
"Ferrari dice"
[Alt-text discarded: Text does not contain prep./article]
"The meaning of life"
HAPPINESS CONSISTS IN REALIZING IT IS ALL A GREAT STRANGE DREAM
[Alt-text discarded: No text vs. image-object overlap]
"Demi Lovato wearing a black Ester Abner Spring 2018 gown and Stuart Weitzman sandals at the 2017 American Music Awards"
"pop rock artist wearing a black gown and sandals at awards"

Figure 2: Conceptual Captions pipeline steps with examples and final output.

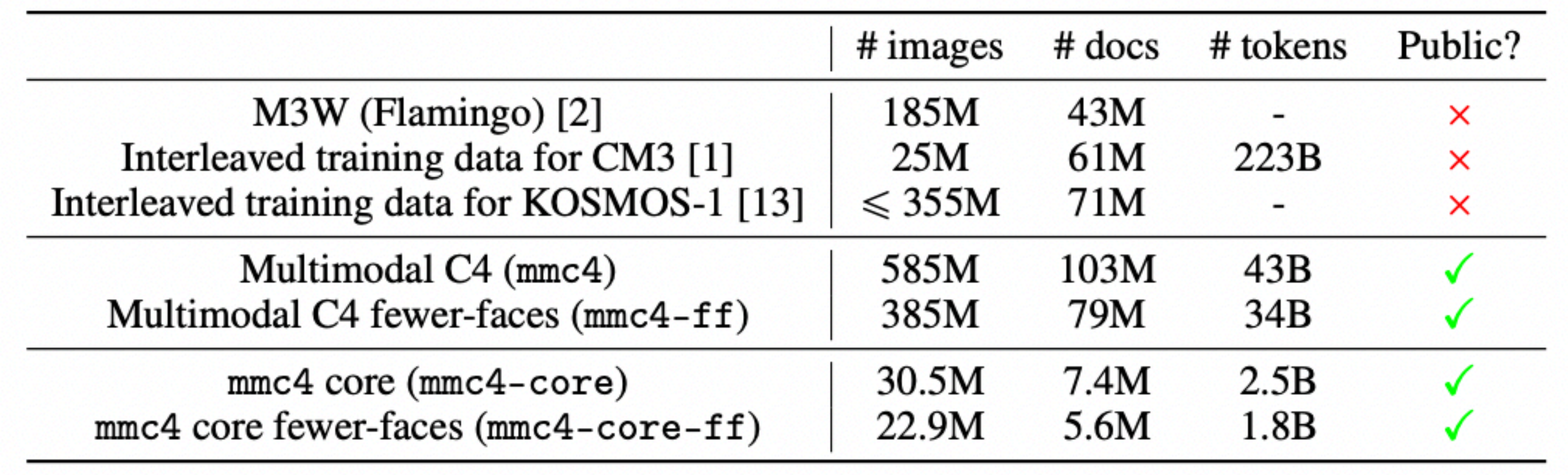
Clean based on: alt-text:

- high unique word ratio covering various POS tags
- remove ones with high rate of token repetition
- Capitalisation is good indicator
- Filter based on NSFW
- ... -> 3% remains
- further filtering with supervised image classifier

Finally: replace with hypernyms (e.g. "actor"), remove locations etc.

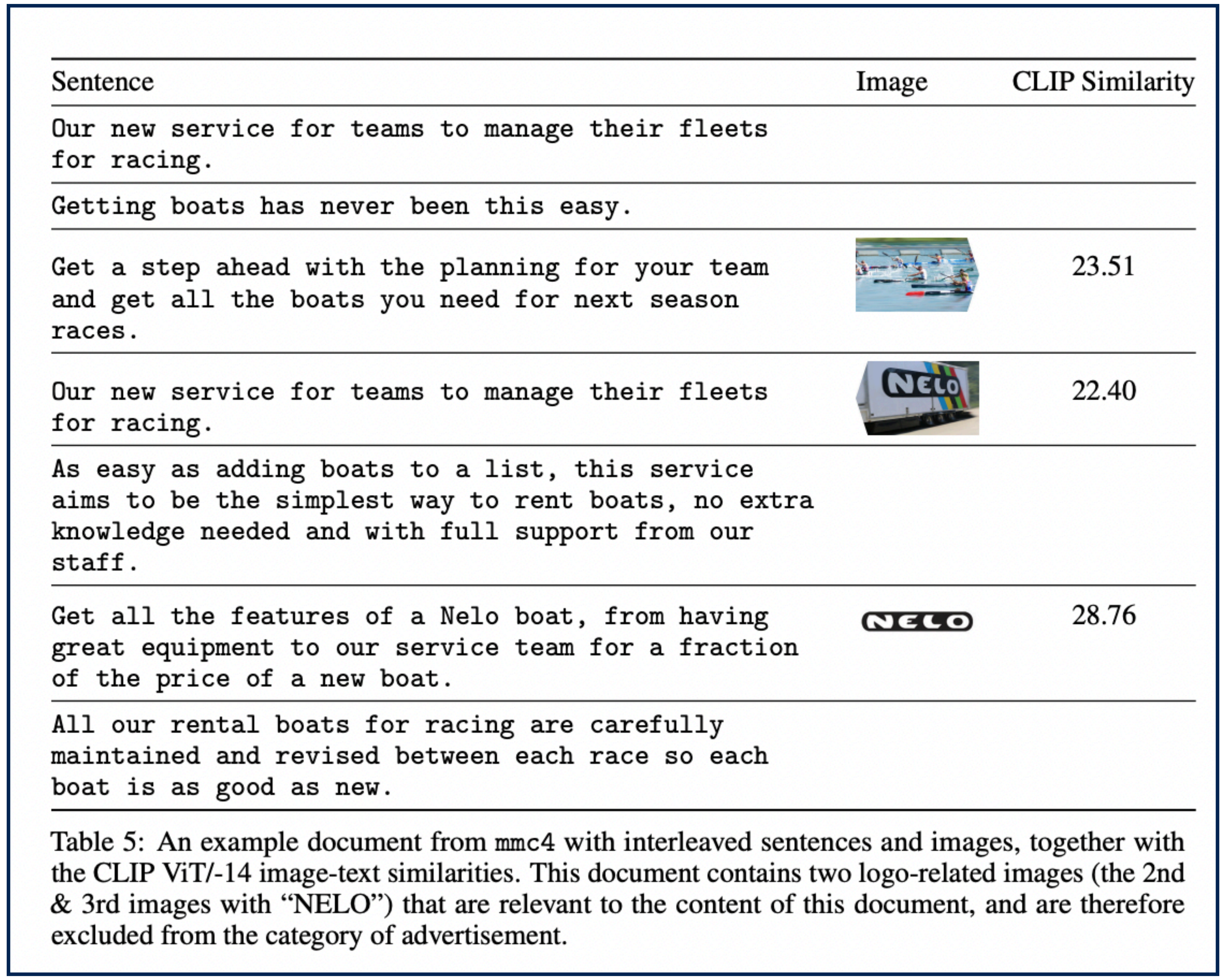
Sharma et al. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. ACL 2016
Changpinyo et al. Conceptual 12M: Pushing Web-Scale Image-Text Pre-Training To Recognize Long-Tail Visual Concepts. CVPR 2021

Multimodal C4: An open, billion-scale corpus of images interleaved with text.



	# images	# docs	# tokens	Public?
M3W (Flamingo) [2]	185M	43M	-	✗
Interleaved training data for CM3 [1]	25M	61M	223B	✗
Interleaved training data for KOSMOS-1 [13]	$\leqslant$ 355M	71M	-	✗
Multimodal C4 (`mmc4`)	585M	103M	43B	✓
Multimodal C4 fewer-faces (`mmc4-ff`)	385M	79M	34B	✓
mmc4 core (`mmc4-core`)	30.5M	7.4M	2.5B	✓
mmc4 core fewer-faces (`mmc4-core-ff`)	22.9M	5.6M	1.8B	✓

- Large dataset
- Several manual and CLIP based filters



Sentence	Image	CLIP Similarity
`Our new service for teams to manage their fleets for racing.`		
`Getting boats has never been this easy.`		
`Get a step ahead with the planning for your team and get all the boats you need for next season races.`		23.51
`Our new service for teams to manage their fleets for racing.`		22.40
`As easy as adding boats to a list, this service aims to be the simplest way to rent boats, no extra knowledge needed and with full support from our staff.`		
`Get all the features of a Nelo boat, from having great equipment to our service team for a fraction of the price of a new boat.`		28.76
`All our rental boats for racing are carefully maintained and revised between each race so each boat is as good as new.`		

Table 5: An example document from `mmc4` with interleaved sentences and images, together with the CLIP ViT/-14 image-text similarities. This document contains two logo-related images (the 2nd & 3rd images with "NELO") that are relevant to the content of this document, and are therefore excluded from the category of advertisement.

https://github.com/allenai/mmc4 ; April 14th 2023

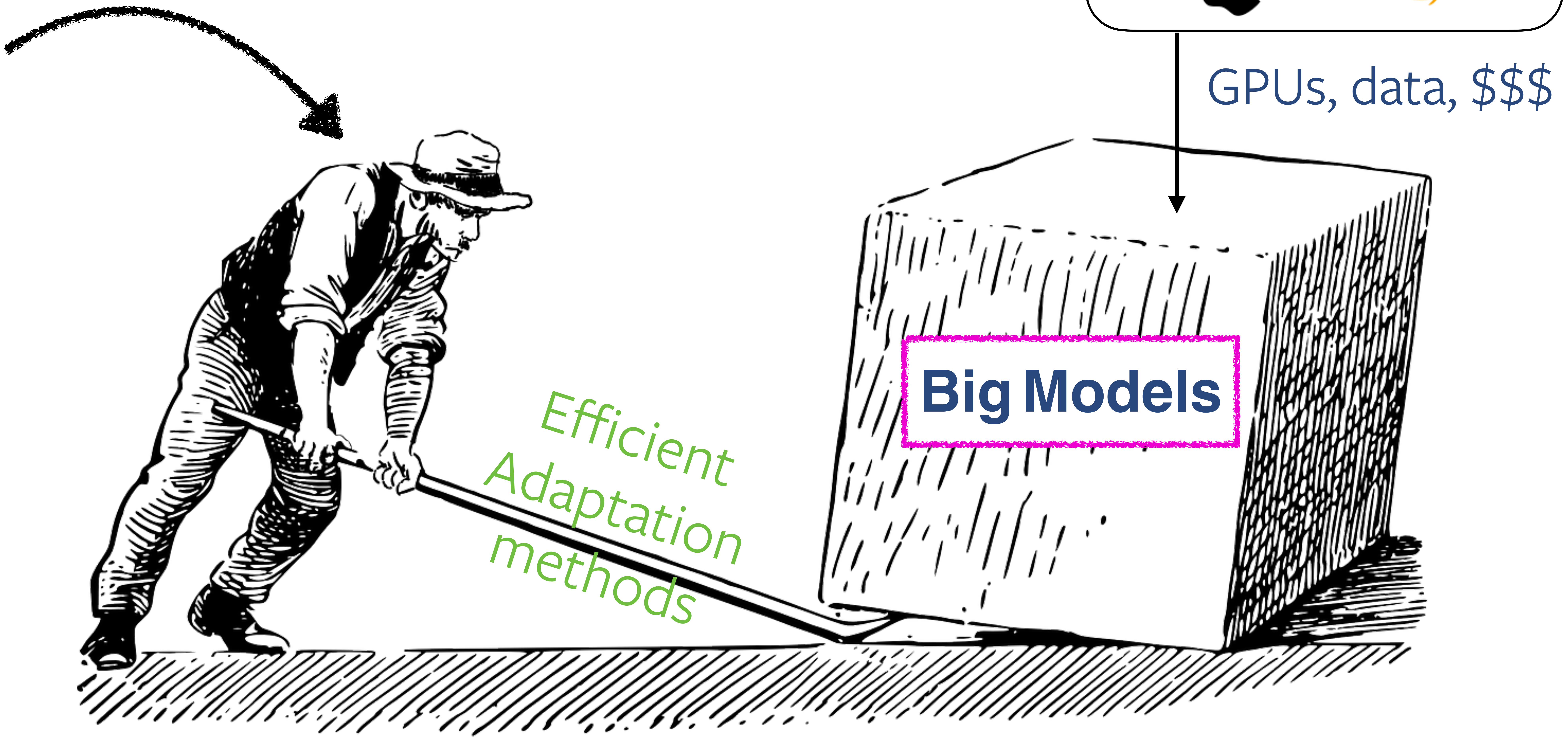
Brief note about text-to-image models (see Vicky's lecture!)

Just a high-level summary:

They use the embeddings of a language model to generate an image.
The more and the better the data, the better.
Bigger models give better results, especially because diffusion models scale well.

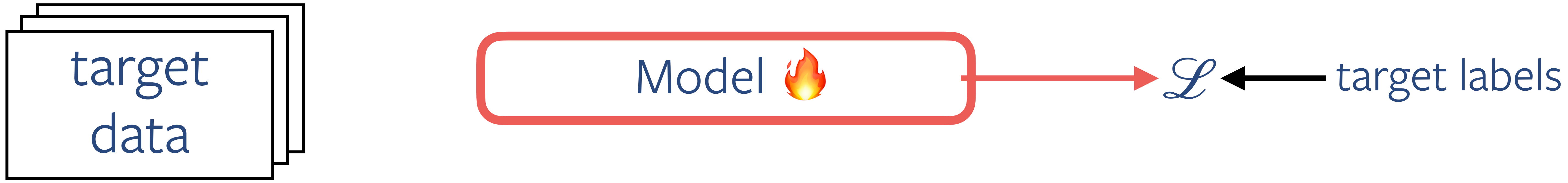
So you've trained your Language/Vision/Vision-Language model...

Parameter-efficient Adaptation Methods (PEFT)



Main ways of adapting models (1/2)

Full-finetuning

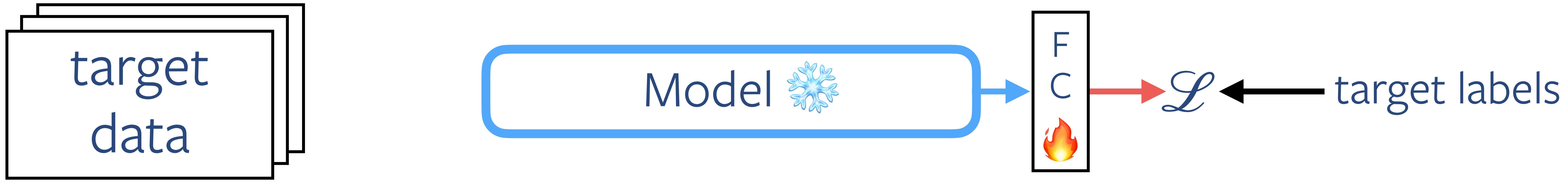


target data

Model 🔥

$\mathscr{L}$ ← target labels

Limited-finetuning (e.g. linear probing)



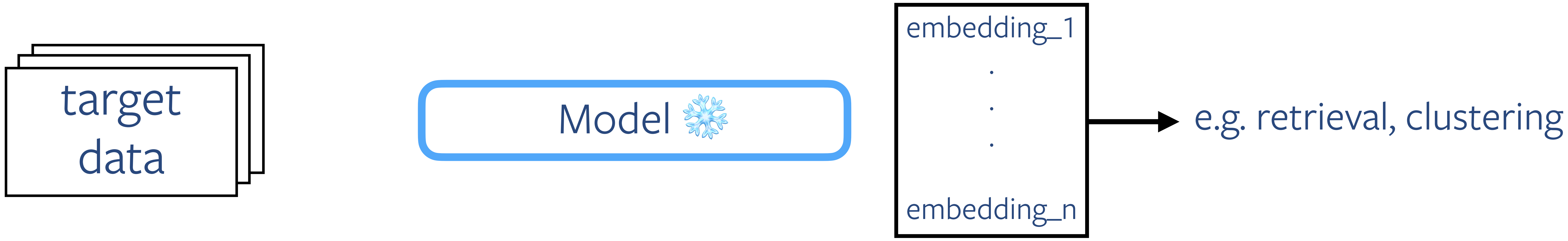
target data

Model ❄️

FC 🔥

$\mathscr{L}$ ← target labels

No-finetuning (e.g. used for retrieving similar instances)



target data

Model ❄️

embedding_1

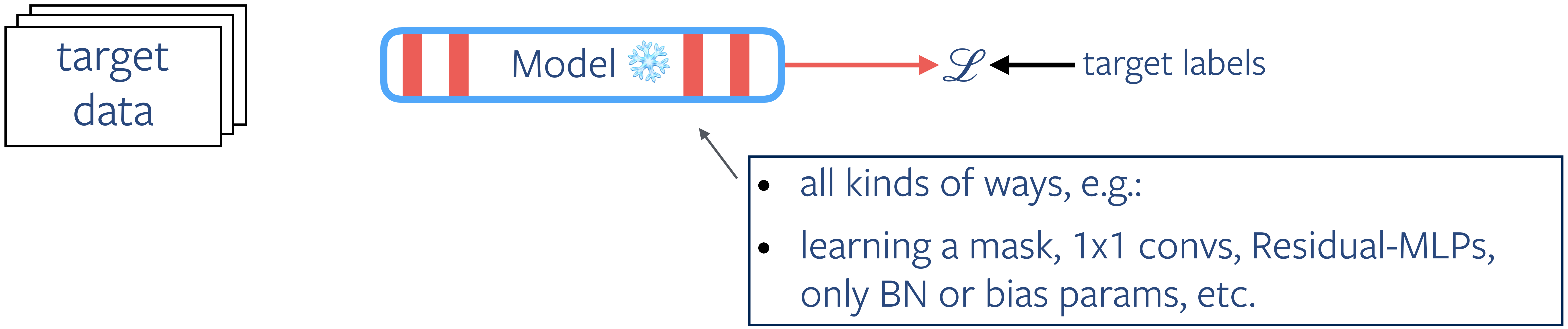
.
.
.

embedding_n

→ e.g. retrieval, clustering

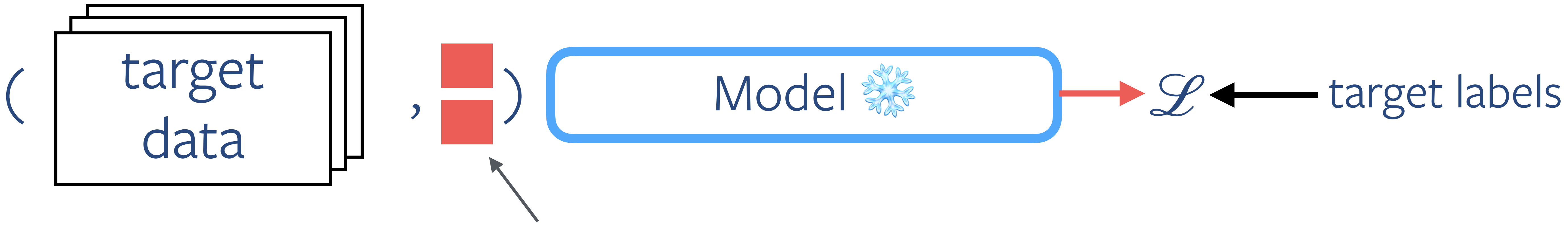
Main ways of adapting models (2/2)

Adapters



- all kinds of ways, e.g.:
- learning a mask, 1x1 convs, Residual-MLPs, only BN or bias params, etc.

Prompt/prefix learning

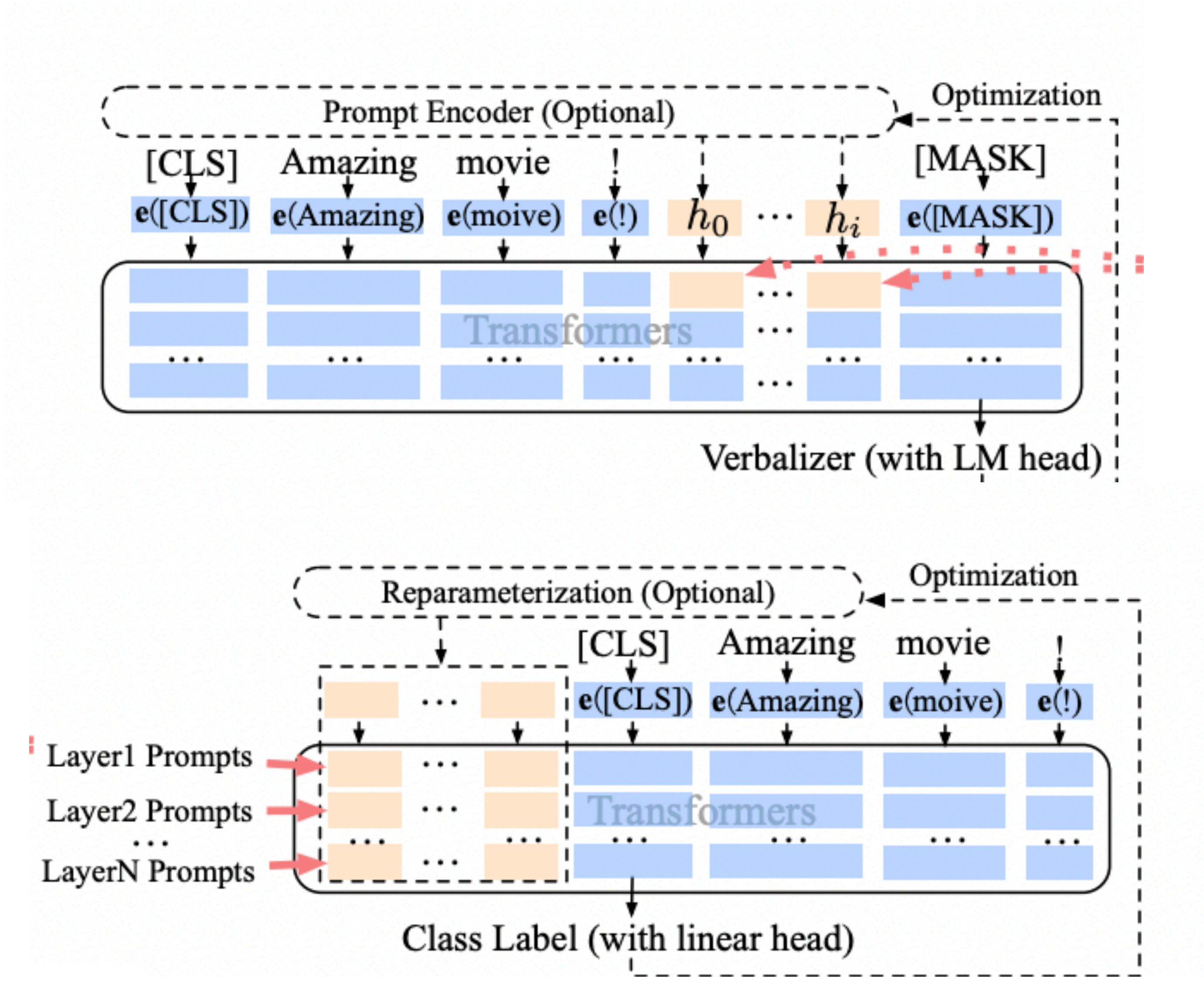


learnable, additional inputs

- similar to prompt manual engineering [like "step-by-step" or "trending on artstation"]

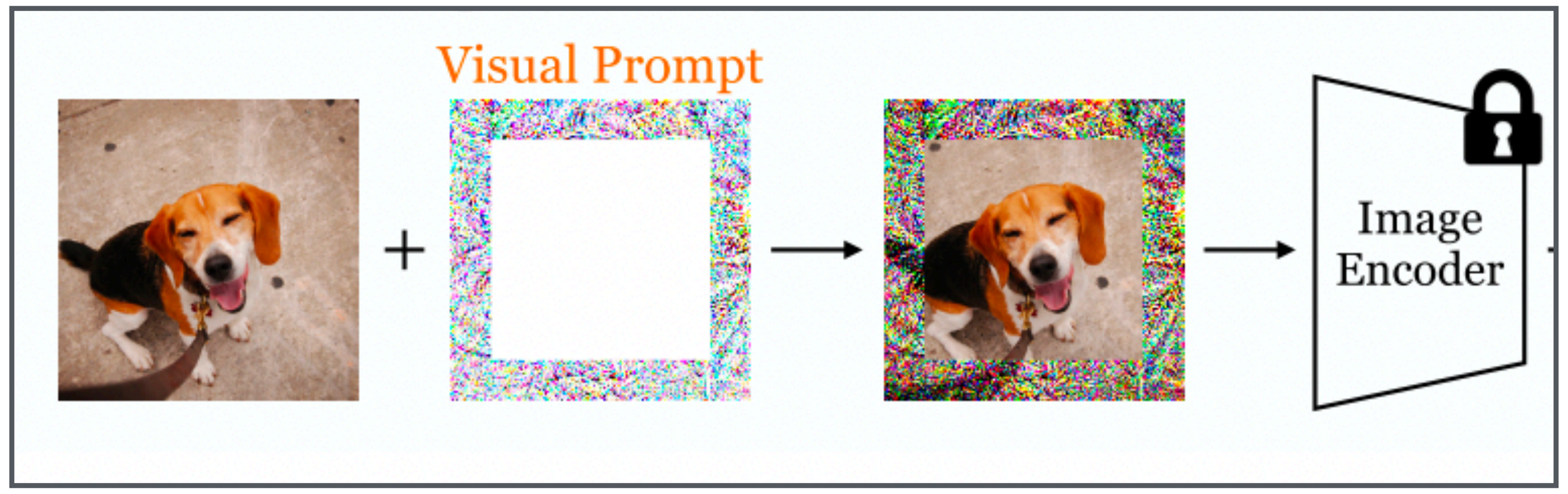
Rebuffi et al. Learning multiple visual domains with residual adapters. NeurIPS 2017, Houlsby et al. Parameter-Efficient Transfer Learning for NLP. ICML 2019
Elsayed et al. Adversarial reprogramming of neural networks. ICLR 2019
Li et al. Prefix-tuning: Optimizing continuous prompts for generation. ACL 2021

Prompt learning: per task

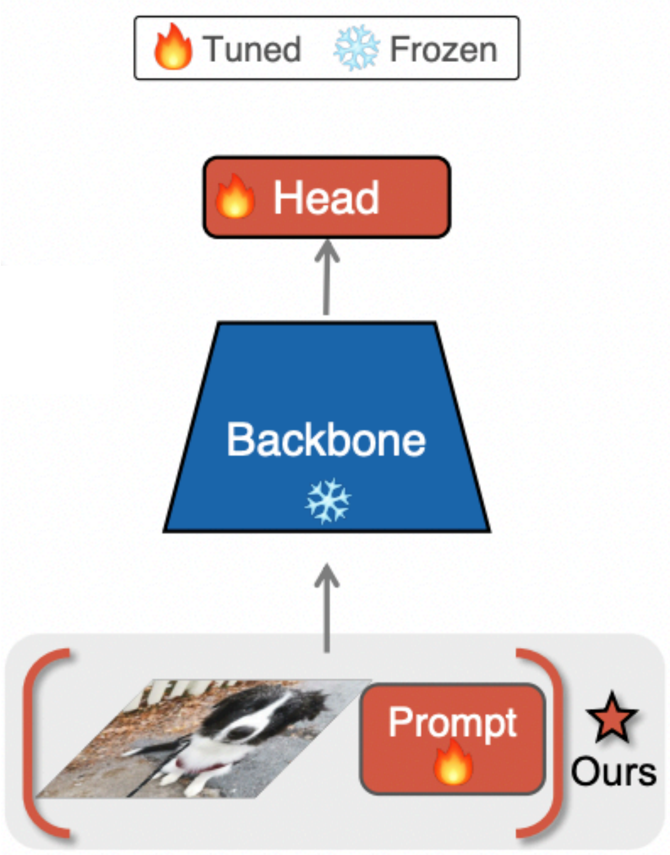


Li et al. Prefix-tuning: Optimizing continuous prompts for generation. ACL 2021

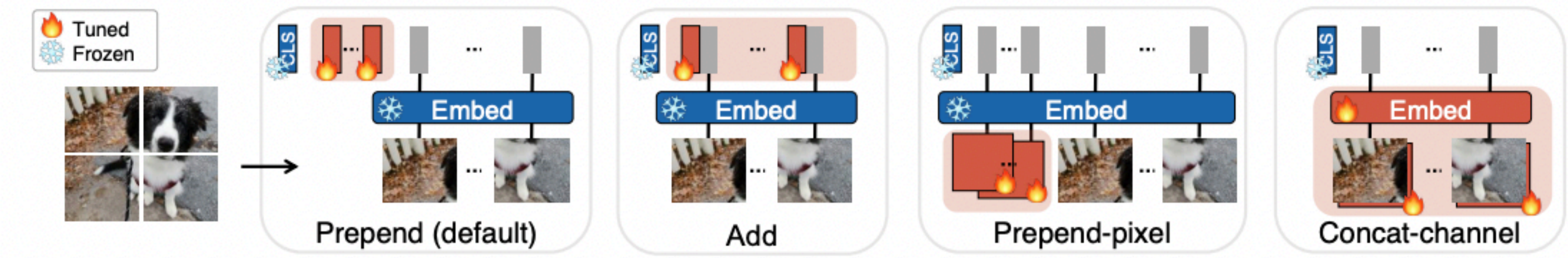
- prefixes are just learnable vectors
- 🧐: are reparameterised as an MLP that gets a fixed input (”more stable”)
- Extend this: ”deep prompt tuning”
- but increases memory (bc. of attn)



- Works also for CNNs
- Strictly input-only

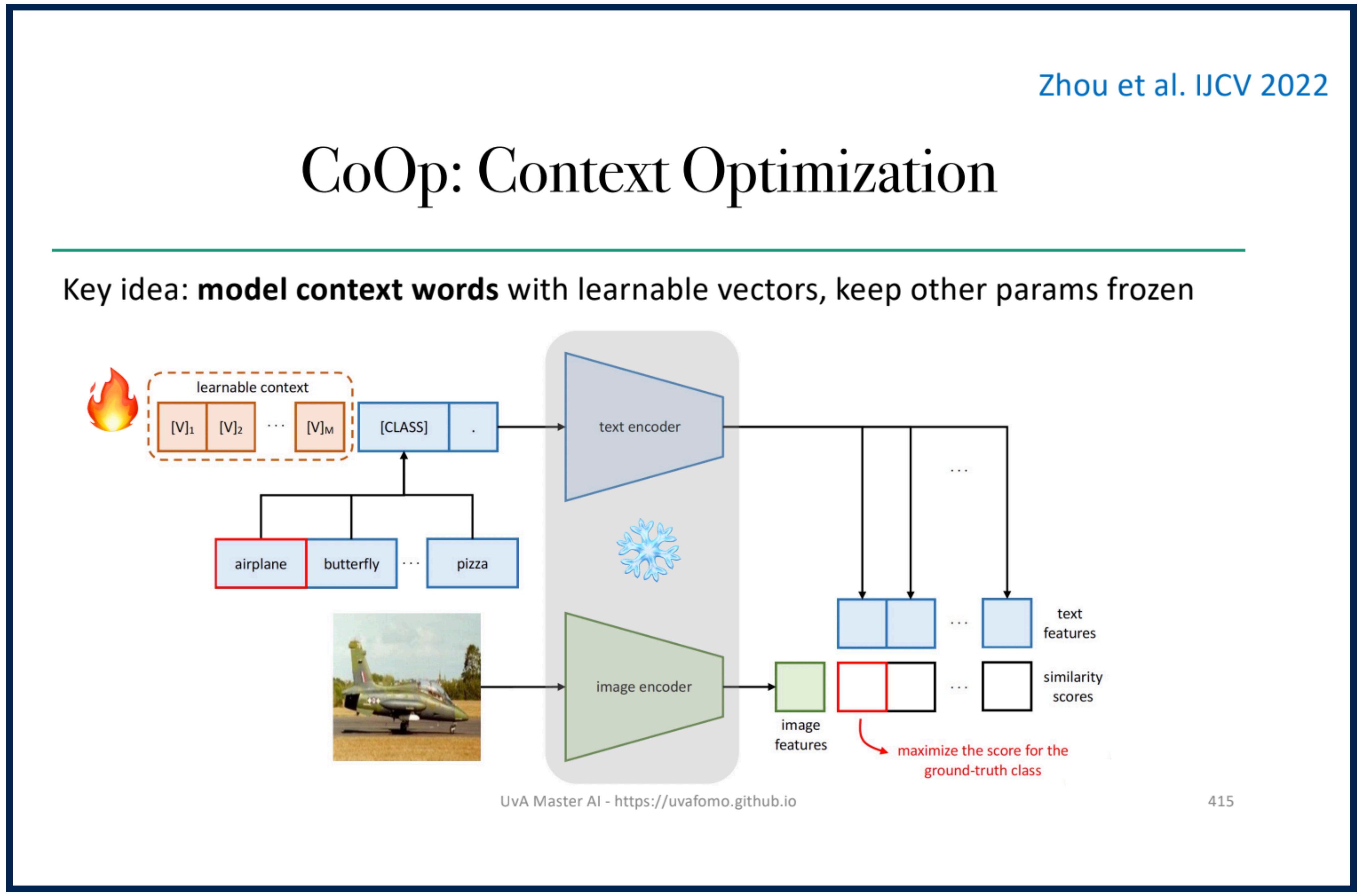


- Actually also trains linear layer on top
- 👇 explore various ways of prompting inputs for visual inputs



Li et al. Prefix-tuning: Optimizing continuous prompts for generation. ACL 2021
Liu et al. P-Tuning: Prompt Tuning Can Be Comparable to Fine-tuning Across Scales and Tasks. ACL 2022
Bhang et al. Exploring Visual Prompts for Adapting Large-Scale Models. 2022
Zhou et al. Learning to Prompt for Vision-Language Models.IJCV 2021l, Zhou et al. Conditional Prompt Learning for Vision-Language Models. CVPR 2022
Jia et al. Visual Prompt Tuning.. ECCV 2022

Visual prompting: change the embedding, despite keeping encoder frozen



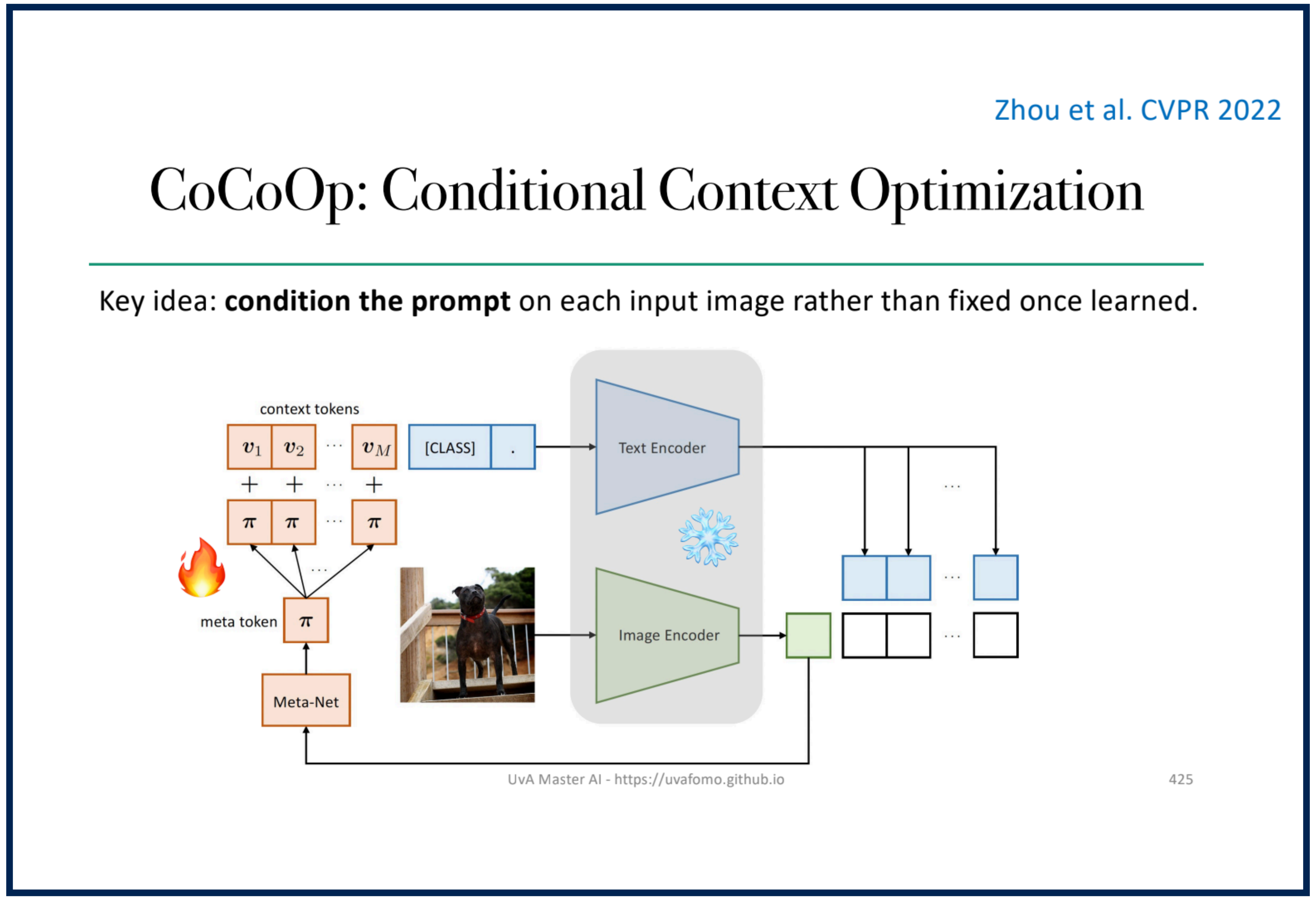
Zhou et al. IJCV 2022

CoOp: Context Optimization

Key idea: **model context words** with learnable vectors, keep other params frozen

UvA Master AI - https://uvafomo.github.io 415

- Learnable, vector-version of "this is an [image/ photograph/illustration] of…"



Zhou et al. CVPR 2022

CoCoOp: Conditional Context Optimization

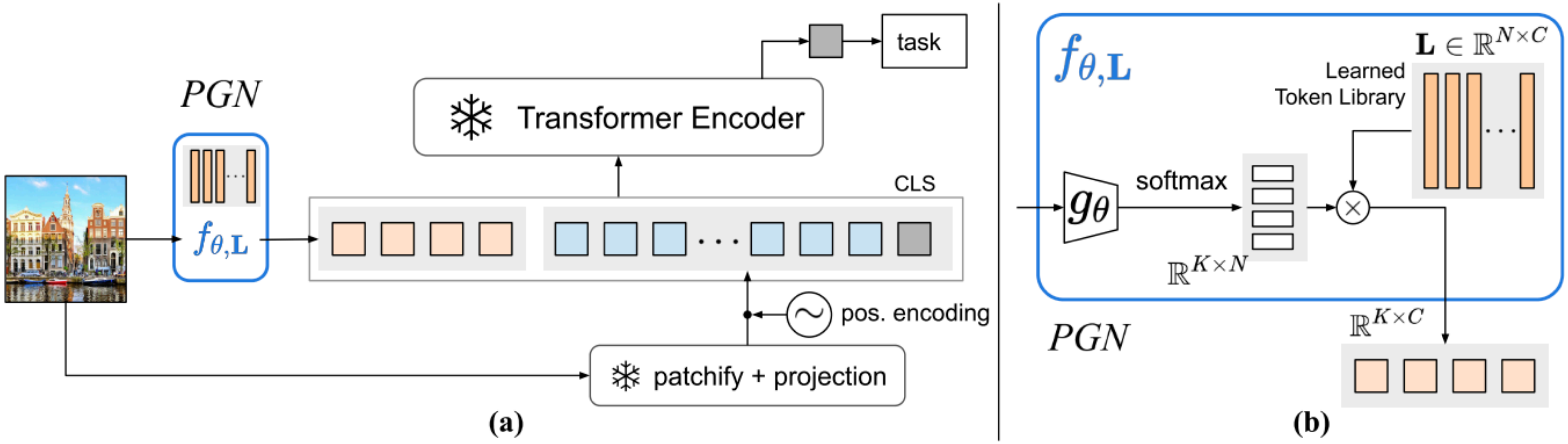
Key idea: **condition the prompt** on each input image rather than fixed once learned.

UvA Master AI - https://uvafomo.github.io 425

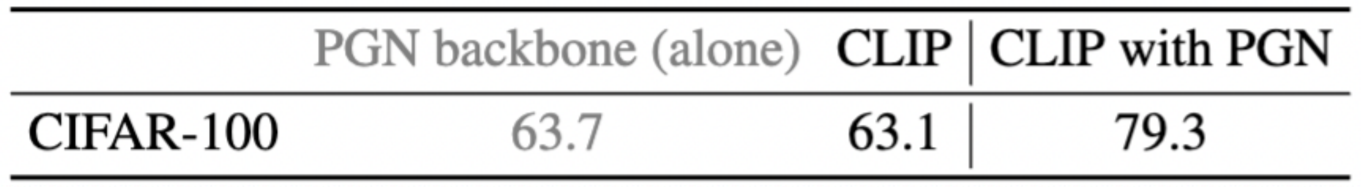
- Condition this additionally on image

Fundamental AI Lab UTN

Prompt learning per datum, *in input-space only*

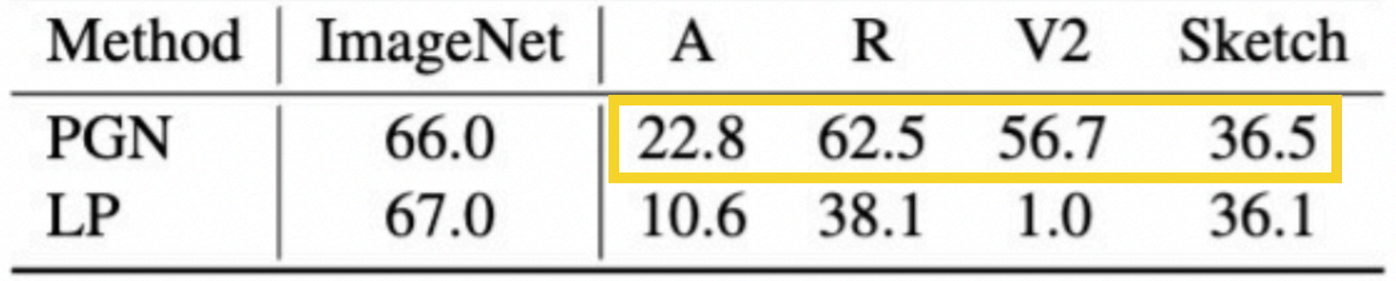


- Learn a input-to-prompt mini-network
- Generate prompts from a set of learnable prompts
- Prompts (learned in space after first conv1), can be made to be input-only (convs are linear operation!)



	PGN backbone (alone)	CLIP	CLIP with PGN
CIFAR-100	63.7	63.1	79.3

PGN learns what's missing in CLIP



Method	ImageNet	A	R	V2	Sketch
PGN	66.0	22.8	62.5	56.7	36.5
LP	67.0	10.6	38.1	1.0	36.1

More robust compared to linear probing (LP)

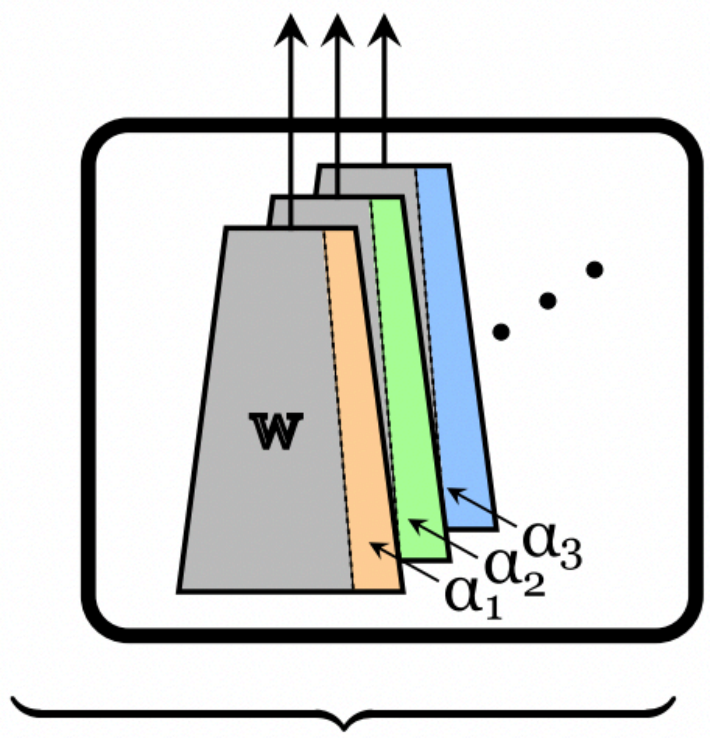
Loedeman et al. Prompt Generation Networks for Efficient Adaptation of Frozen Vision Transformers. 2022

Adapters: any modification ”in the middle” of NNs

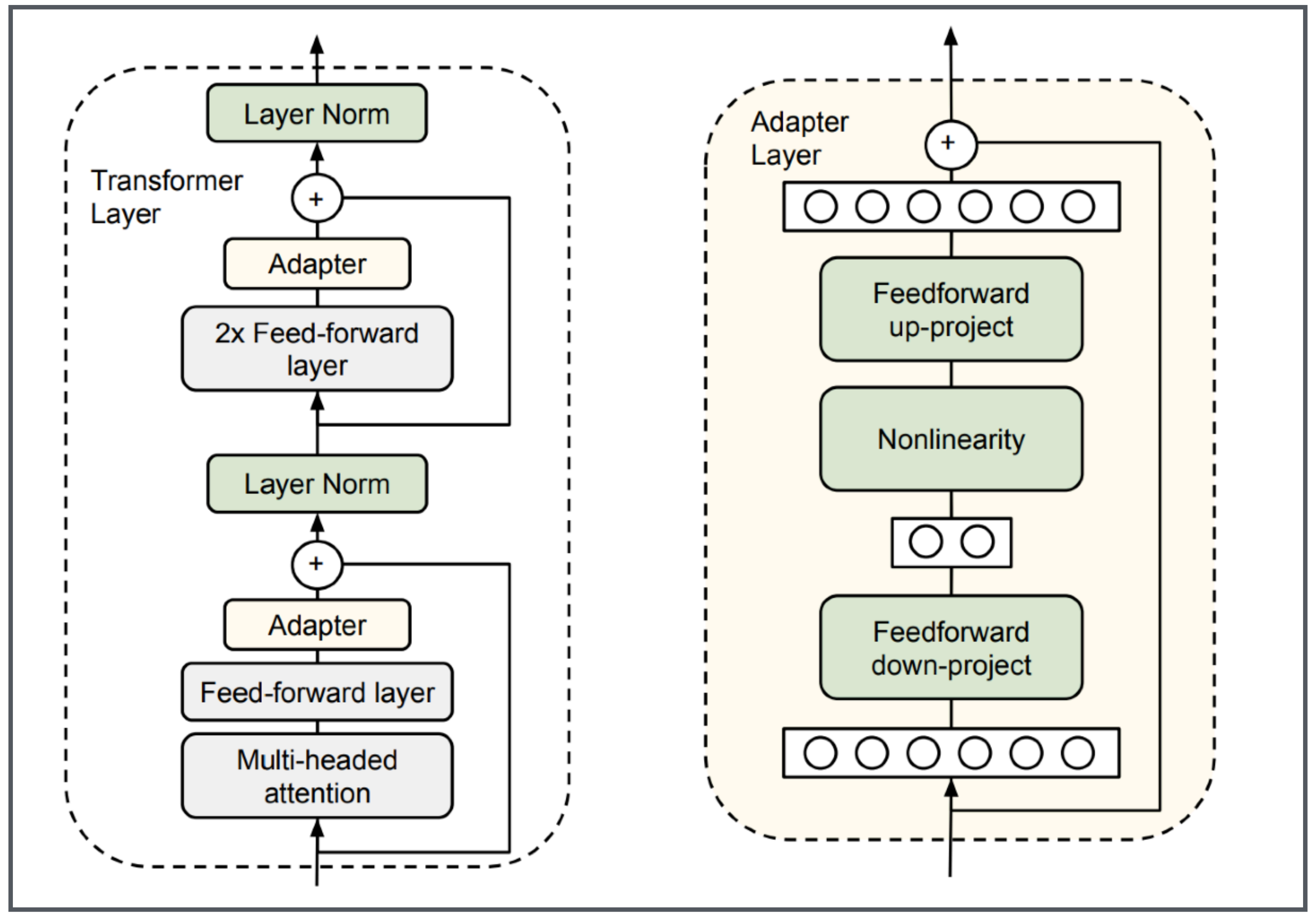
Simplest form: residual adapters

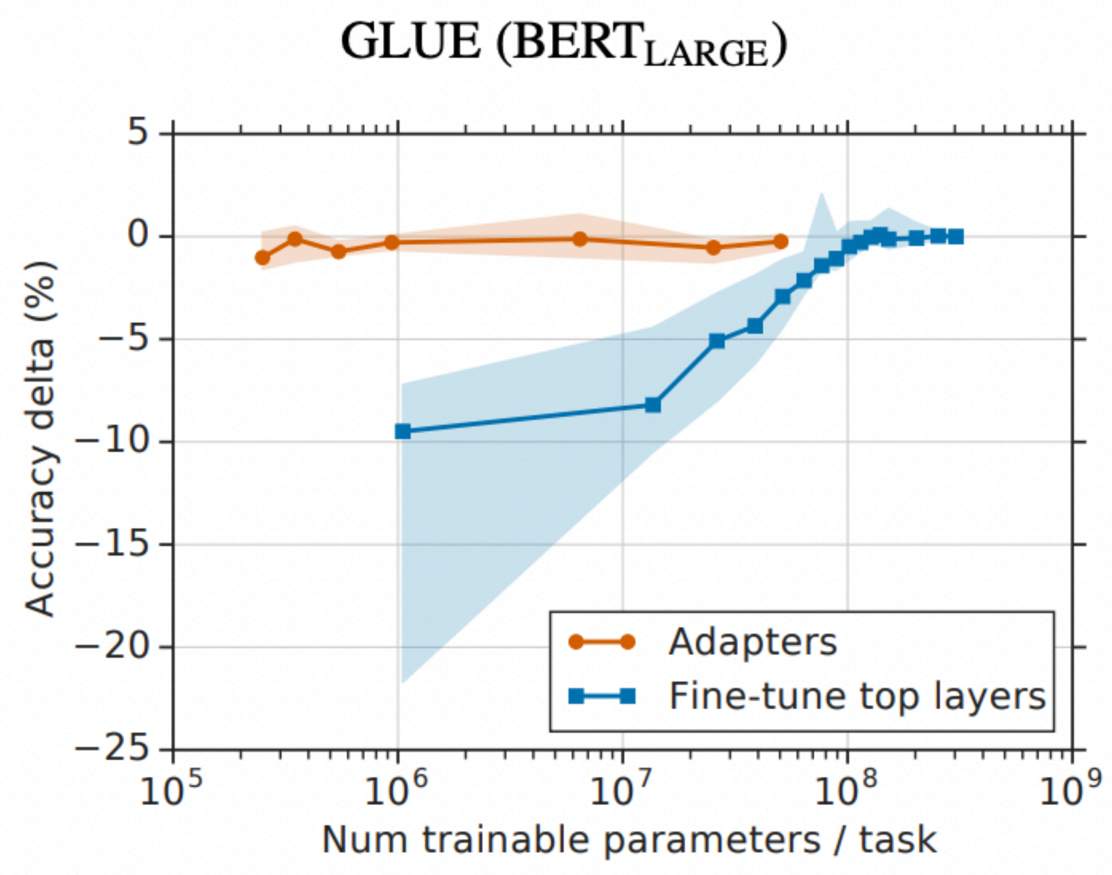
$$g(x;\alpha) = x + \alpha * x.$$

limit α to e.g. 1x1 conv



(a) universal parametric family



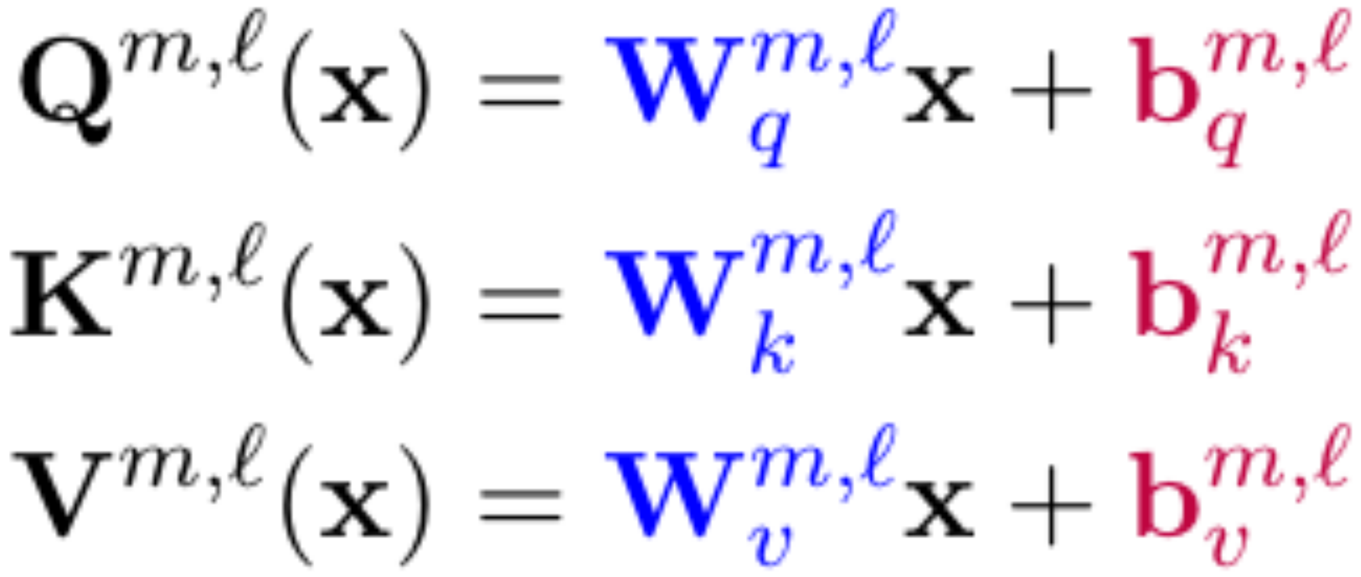


- (-) makes computation graph more complex; adds inference time
- (+) doesn’t require much memory to store
- (+) very expressive/performant and fast to learn

Rebuffi et al. Learning multiple visual domains with residual adapters. NeurIPS 2017, Houlsby et al. Parameter-Efficient Transfer Learning for NLP. ICML 2019

Fine-tuning only the bias terms / Norm layers

BitFit



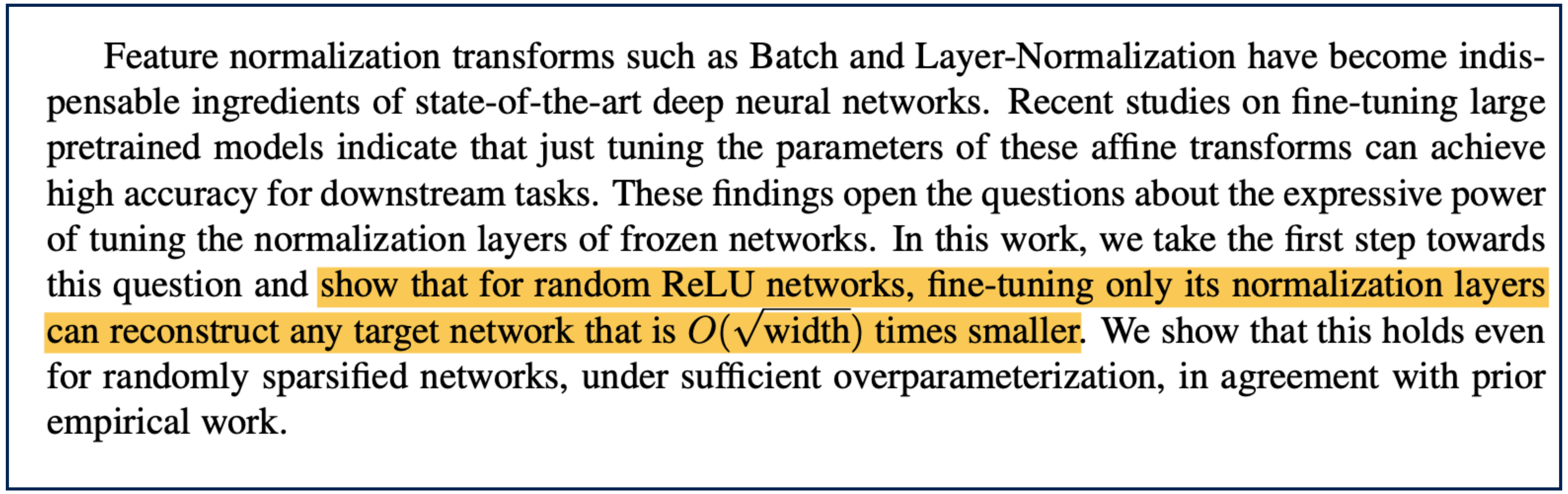
$$\mathbf{Q}^{m,\ell}(\mathbf{x}) = \mathbf{W}_q^{m,\ell}\mathbf{x} + \mathbf{b}_q^{m,\ell}$$

$$\mathbf{K}^{m,\ell}(\mathbf{x}) = \mathbf{W}_k^{m,\ell}\mathbf{x} + \mathbf{b}_k^{m,\ell}$$

$$\mathbf{V}^{m,\ell}(\mathbf{x}) = \mathbf{W}_v^{m,\ell}\mathbf{x} + \mathbf{b}_v^{m,\ell}$$

frozen, trainable

- Learn only the bias terms
- Learning vectors instead of matrices -> efficient



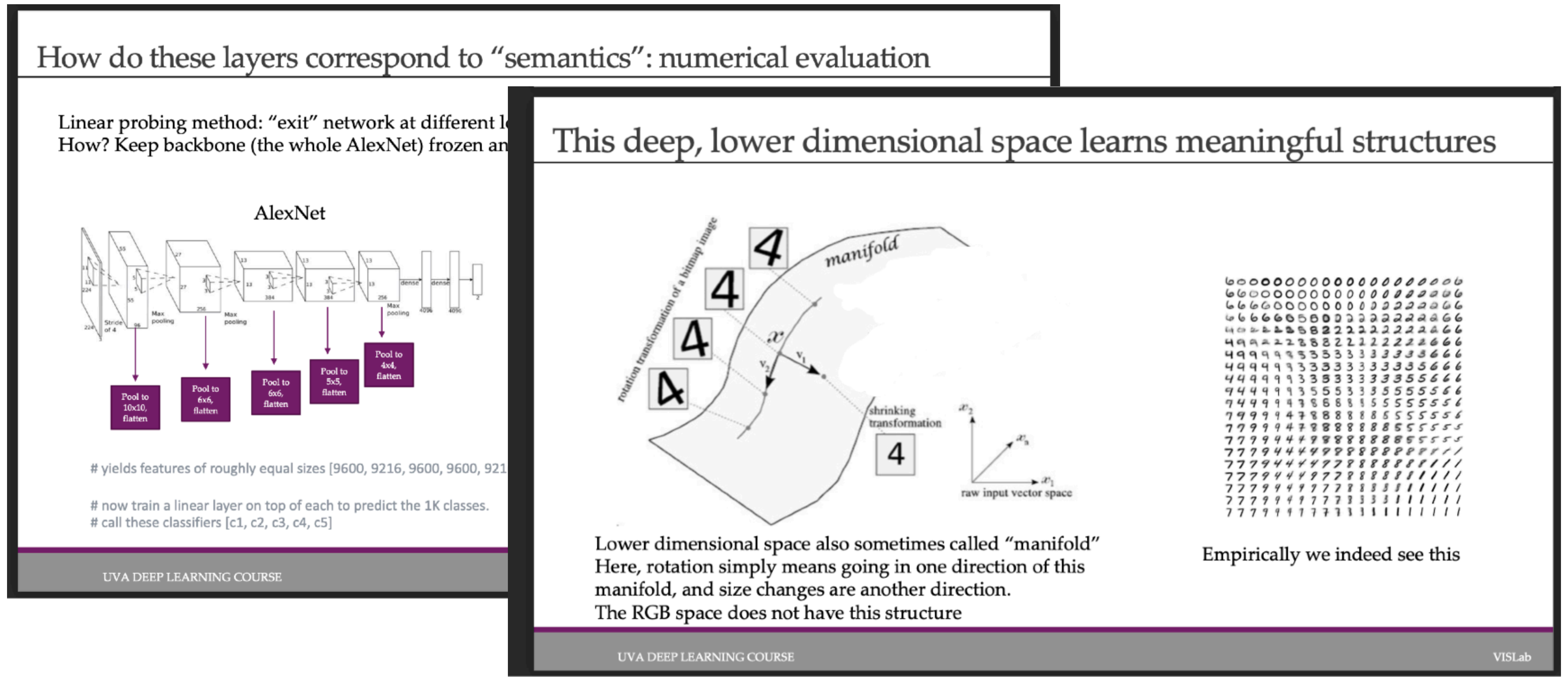
Feature normalization transforms such as Batch and Layer-Normalization have become indispensable ingredients of state-of-the-art deep neural networks. Recent studies on fine-tuning large pretrained models indicate that just tuning the parameters of these affine transforms can achieve high accuracy for downstream tasks. These findings open the questions about the expressive power of tuning the normalization layers of frozen networks. In this work, we take the first step towards this question and show that for random ReLU networks, fine-tuning only its normalization layers can reconstruct any target network that is $O(\sqrt{\text{width}})$ times smaller. We show that this holds even for randomly sparsified networks, under sufficient overparameterization, in agreement with prior empirical work.

- Learn only the LayerNorm / BatchNorm parameters
- Show that it is quite expressive in theory (and practice)

BitFit: Simple Parameter-efficient Fine-tuning for Transformer-based Masked Language-models. Ben-Zaken et al. ACL 2022
The Expressive Power of Tuning Only the Normalization Layers. Giannou et al. PMLR 2023

LoRA: adapting matrix multiplies in efficiently / "a generalisation of full-finetuning"

$$\boxed{\operatorname{rank}(AB) \leq \min(\operatorname{rank}(A), \operatorname{rank}(B))}$$



How do these layers correspond to "semantics": numerical evaluation

Linear probing method: "exit" network at different l
How? Keep backbone (the whole AlexNet) frozen an

AlexNet

yields features of roughly equal sizes [9600, 9216, 9600, 9600, 921

now train a linear layer on top of each to predict the 1K classes.
call these classifiers [c1, c2, c3, c4, c5]

UVA DEEP LEARNING COURSE

This deep, lower dimensional space learns meaningful structures

Lower dimensional space also sometimes called "manifold"
Here, rotation simply means going in one direction of this manifold, and size changes are another direction.
The RGB space does not have this structure

Empirically we indeed see this

UVA DEEP LEARNING COURSE VISLab

Remember from DL1:

- real data ~ lies on lower dimensional manifold,
- DNNs map from RGB space gradually to more semantic space.

Normal fully connected layer:

$$h = W_0 x$$

LoRA adapted:

$$h = W_0 x + \Delta W x = W_0 x + \underset{\uparrow}{BA}x$$

BA is low-rank matrix.

"Low-rank"
--> think of it as outer-product of few vectors

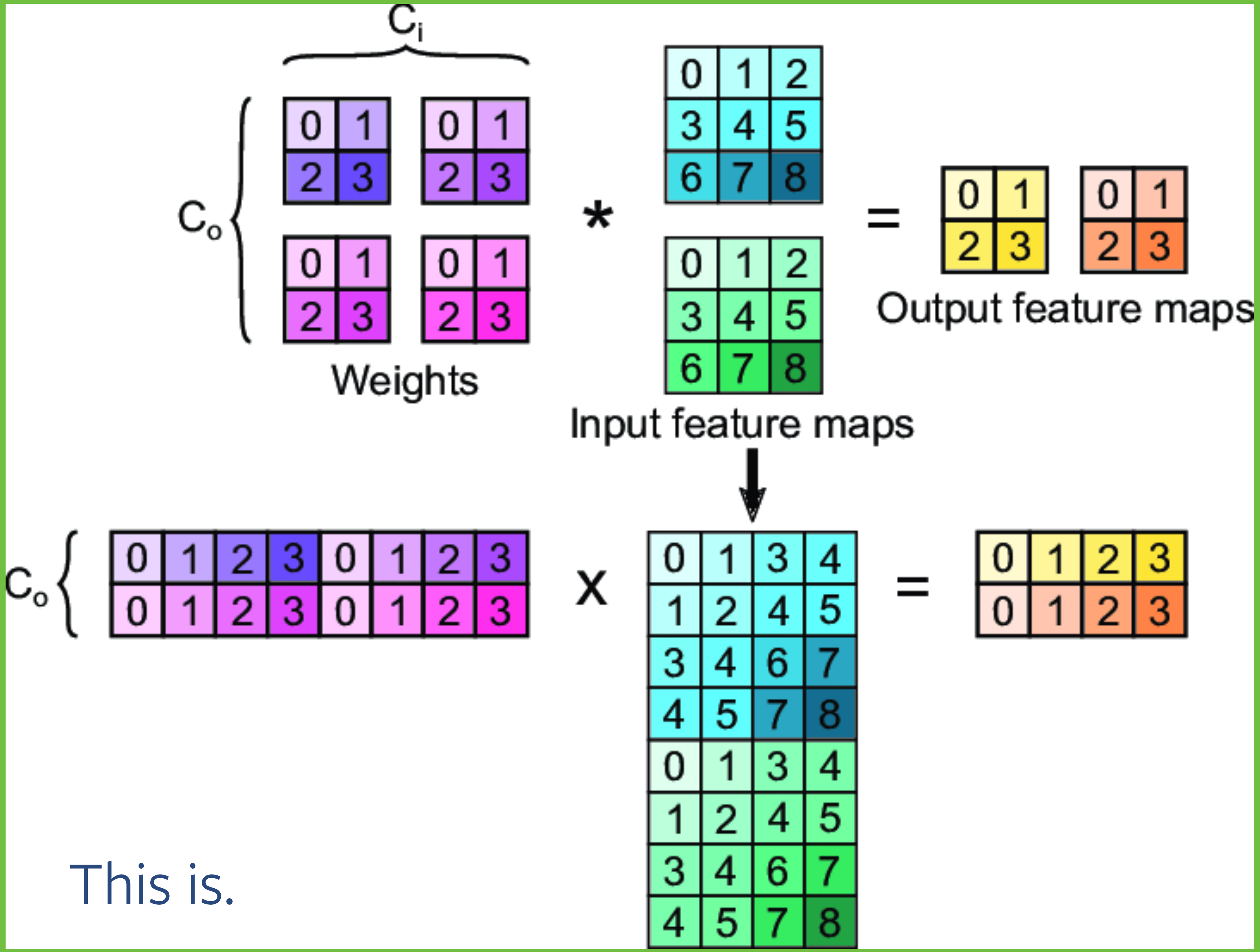
- (-) not as expressive as adapters
- (+) linear op, so after training can be fused with original weights --> same speed

Hu et al. LoRA: Low-Rank Adaptation of Large Language Models. ICLR 2022
Aghajanyan et al. Intrinsic Dimensionality Explains the Effectiveness of Language Model Fine-Tuning. ACL 2021

Think: how would you apply LoRA on a convolutional network like a ResNet/U-Net?

$$\begin{bmatrix} x & y & x & 0 & 0 & 0 \\ 0 & x & y & x & 0 & 0 \\ 0 & 0 & x & y & x & 0 \\ 0 & 0 & 0 & x & y & x \end{bmatrix}$$

This is not the right way to think about it.



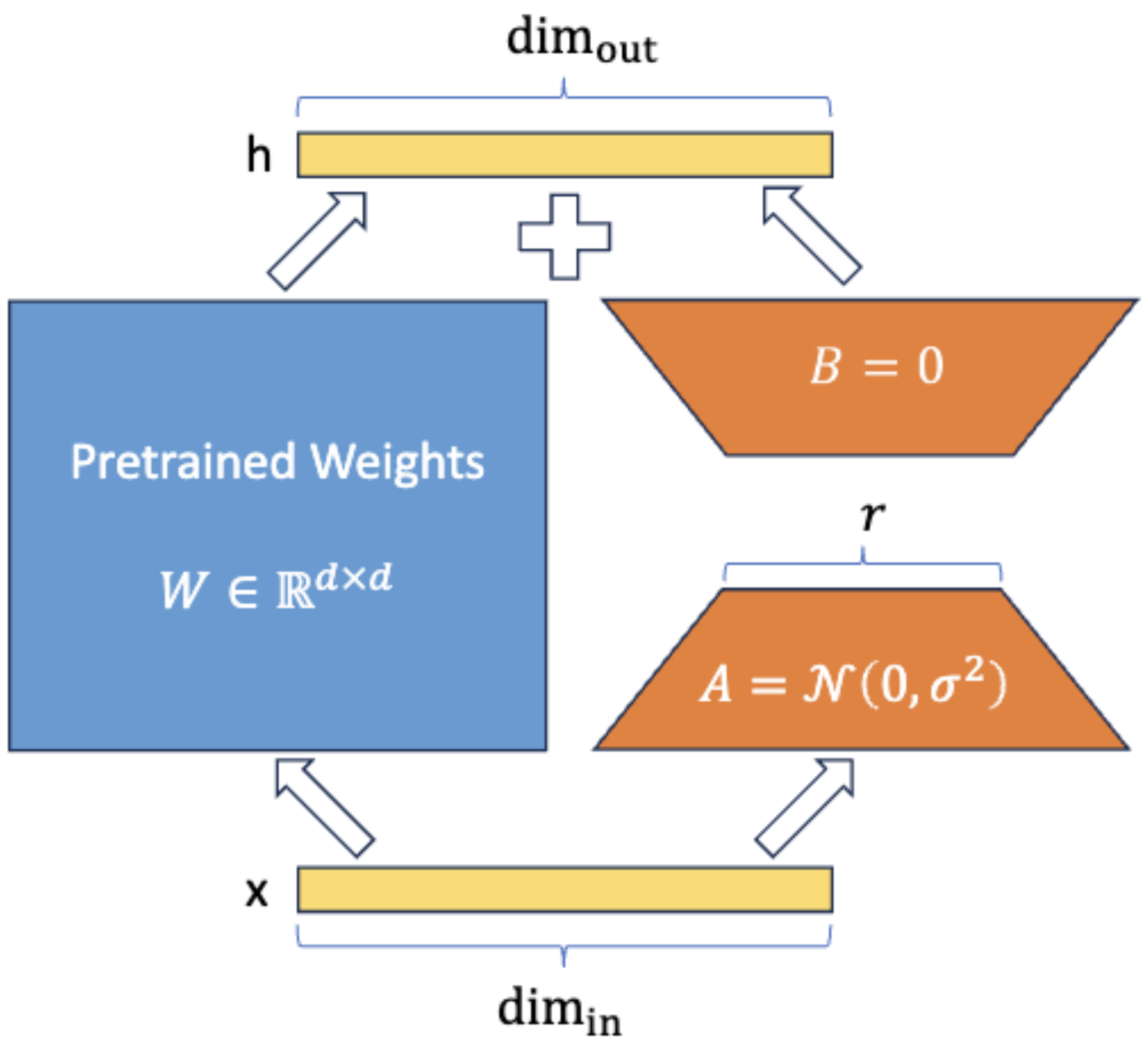
This is.

VeRA: Vector-based Random Matrix Adaptation

DAWID J. KOPICZKO, TIJMEN BLANKEVOORT, YUKI M. ASANO

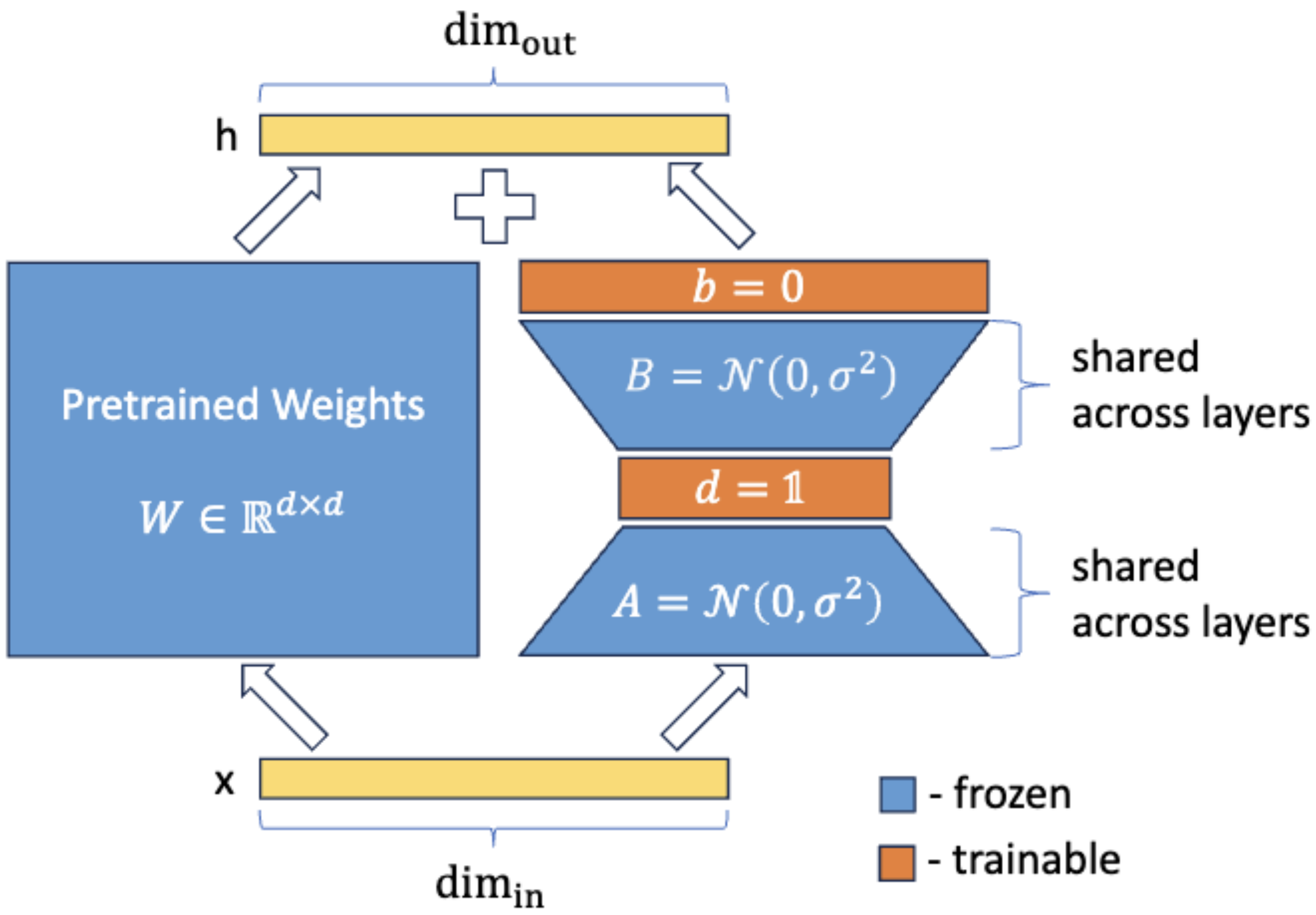
ICLR’24

We make LoRA more efficient



dim_{out}
h
Pretrained Weights
$W \in \mathbb{R}^{d \times d}$
$B = 0$
r
$A = \mathcal{N}(0, \sigma^2)$
x
dim_{in}

Low-Rank Adaptation (LoRA)



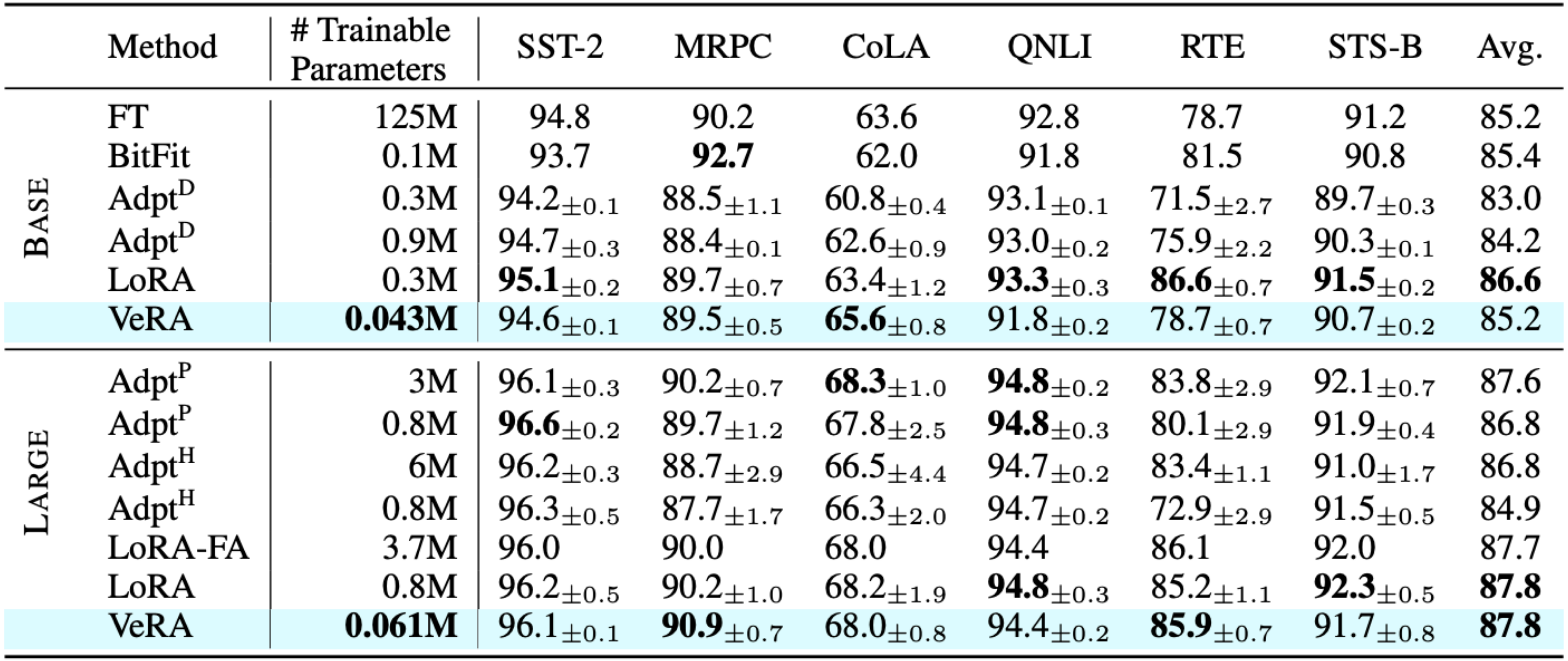
Vector-based Random Matrix Adaptation (VeRA)

W_new = W_old + AB,
where A,B are low-rank
learned per-layer

W_new = W_old + AdBb,
where A,B are random & frozen, same across layers;
d,b are learned vectors

VeRA: Vector-based Random Matrix Adaptation. Kopiczko et al. ICLR 2024

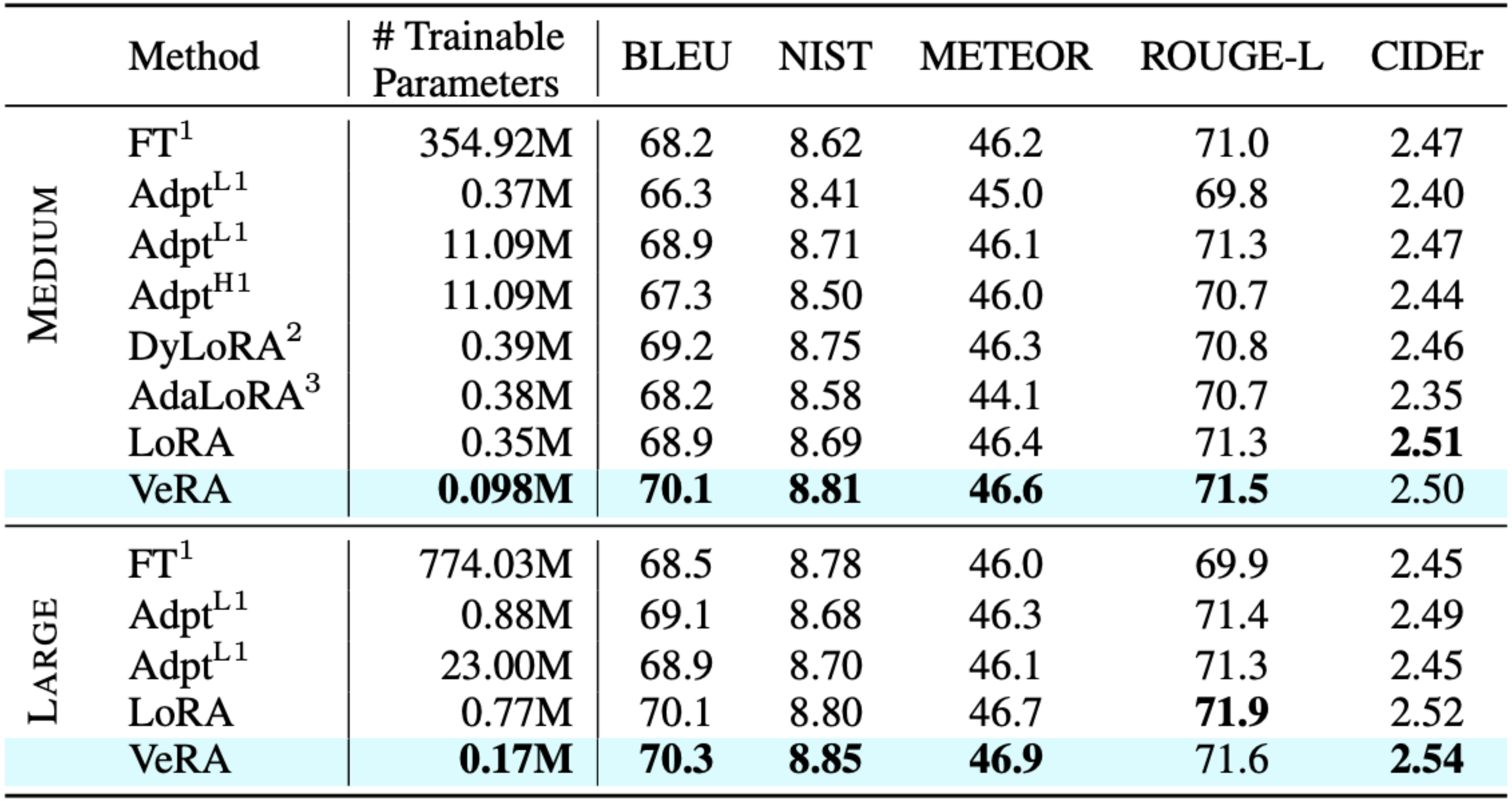
Results on GLUE with RoBERTa



	Method	# Trainable Parameters	SST-2	MRPC	CoLA	QNLI	RTE	STS-B	Avg.
BASE	FT	125M	94.8	90.2	63.6	92.8	78.7	91.2	85.2
	BitFit	0.1M	93.7	**92.7**	62.0	91.8	81.5	90.8	85.4
	Adpt^{D}	0.3M	94.2$_{\pm0.1}$	88.5$_{\pm1.1}$	60.8$_{\pm0.4}$	93.1$_{\pm0.1}$	71.5$_{\pm2.7}$	89.7$_{\pm0.3}$	83.0
	Adpt^{D}	0.9M	94.7$_{\pm0.3}$	88.4$_{\pm0.1}$	62.6$_{\pm0.9}$	93.0$_{\pm0.2}$	75.9$_{\pm2.2}$	90.3$_{\pm0.1}$	84.2
	LoRA	0.3M	**95.1**$_{\pm0.2}$	89.7$_{\pm0.7}$	63.4$_{\pm1.2}$	**93.3**$_{\pm0.3}$	**86.6**$_{\pm0.7}$	**91.5**$_{\pm0.2}$	**86.6**
	VeRA	**0.043M**	94.6$_{\pm0.1}$	89.5$_{\pm0.5}$	**65.6**$_{\pm0.8}$	91.8$_{\pm0.2}$	78.7$_{\pm0.7}$	90.7$_{\pm0.2}$	85.2
LARGE	Adpt^{P}	3M	96.1$_{\pm0.3}$	90.2$_{\pm0.7}$	**68.3**$_{\pm1.0}$	**94.8**$_{\pm0.2}$	83.8$_{\pm2.9}$	92.1$_{\pm0.7}$	87.6
	Adpt^{P}	0.8M	**96.6**$_{\pm0.2}$	89.7$_{\pm1.2}$	67.8$_{\pm2.5}$	**94.8**$_{\pm0.3}$	80.1$_{\pm2.9}$	91.9$_{\pm0.4}$	86.8
	Adpt^{H}	6M	96.2$_{\pm0.3}$	88.7$_{\pm2.9}$	66.5$_{\pm4.4}$	94.7$_{\pm0.2}$	83.4$_{\pm1.1}$	91.0$_{\pm1.7}$	86.8
	Adpt^{H}	0.8M	96.3$_{\pm0.5}$	87.7$_{\pm1.7}$	66.3$_{\pm2.0}$	94.7$_{\pm0.2}$	72.9$_{\pm2.9}$	91.5$_{\pm0.5}$	84.9
	LoRA-FA	3.7M	96.0	90.0	68.0	94.4	86.1	92.0	87.7
	LoRA	0.8M	96.2$_{\pm0.5}$	90.2$_{\pm1.0}$	68.2$_{\pm1.9}$	**94.8**$_{\pm0.3}$	85.2$_{\pm1.1}$	**92.3**$_{\pm0.5}$	**87.8**
	VeRA	**0.061M**	96.1$_{\pm0.1}$	**90.9**$_{\pm0.7}$	68.0$_{\pm0.8}$	94.4$_{\pm0.2}$	**85.9**$_{\pm0.7}$	91.7$_{\pm0.8}$	**87.8**

VeRA: Vector-based Random Matrix Adaptation. Kopiczko et al. ICLR 2024

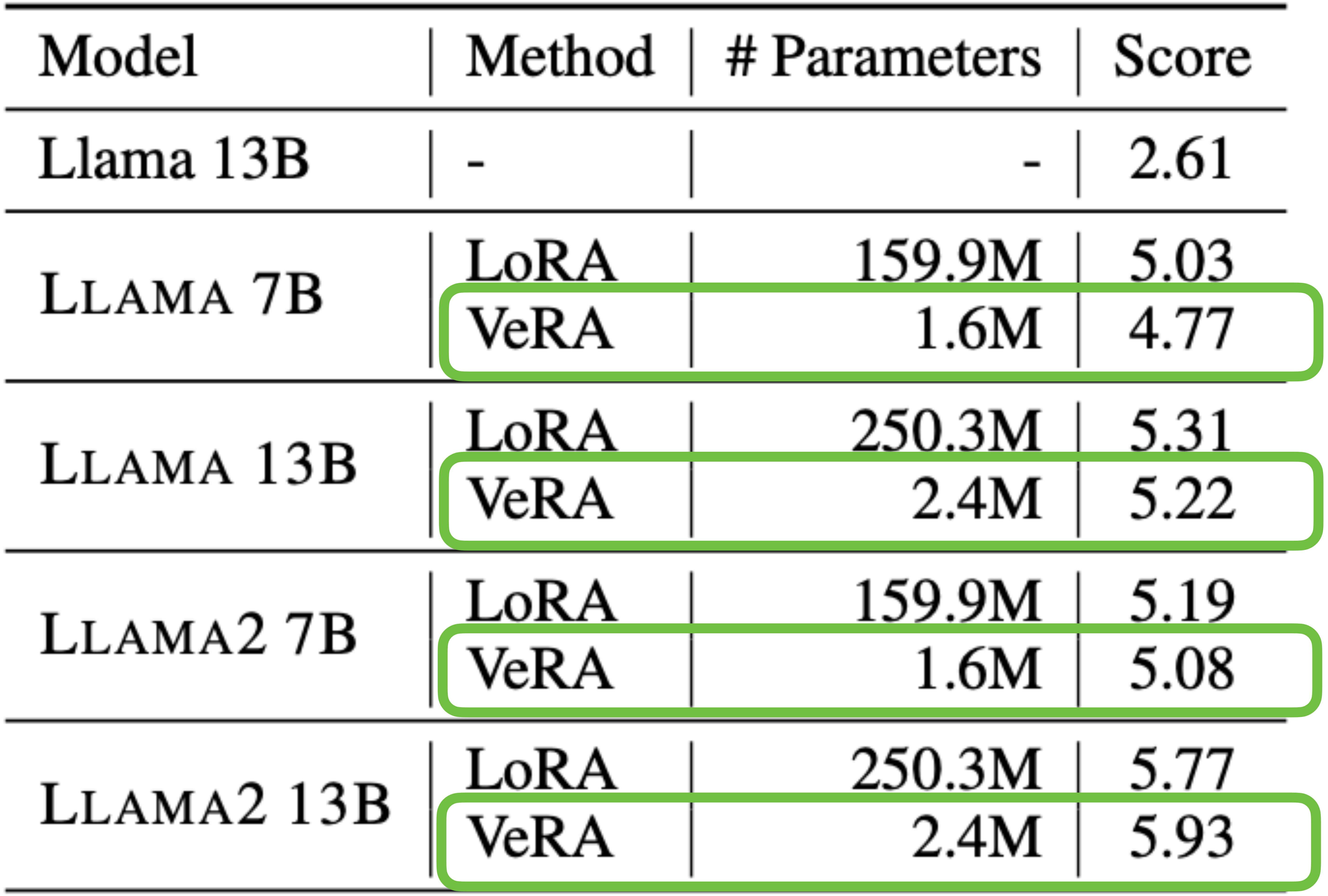
Results on E2E benchmark with GPT2



	Method	# Trainable Parameters	BLEU	NIST	METEOR	ROUGE-L	CIDEr
MEDIUM	FT^1	354.92M	68.2	8.62	46.2	71.0	2.47
	Adpt^{L1}	0.37M	66.3	8.41	45.0	69.8	2.40
	Adpt^{L1}	11.09M	68.9	8.71	46.1	71.3	2.47
	Adpt^{H1}	11.09M	67.3	8.50	46.0	70.7	2.44
	DyLoRA^2	0.39M	69.2	8.75	46.3	70.8	2.46
	AdaLoRA^3	0.38M	68.2	8.58	44.1	70.7	2.35
	LoRA	0.35M	68.9	8.69	46.4	71.3	**2.51**
	VeRA	**0.098M**	**70.1**	**8.81**	**46.6**	**71.5**	2.50
LARGE	FT^1	774.03M	68.5	8.78	46.0	69.9	2.45
	Adpt^{L1}	0.88M	69.1	8.68	46.3	71.4	2.49
	Adpt^{L1}	23.00M	68.9	8.70	46.1	71.3	2.45
	LoRA	0.77M	70.1	8.80	46.7	**71.9**	2.52
	VeRA	**0.17M**	**70.3**	**8.85**	**46.9**	71.6	**2.54**

VeRA: Vector-based Random Matrix Adaptation. Kopiczko et al. ICLR 2024

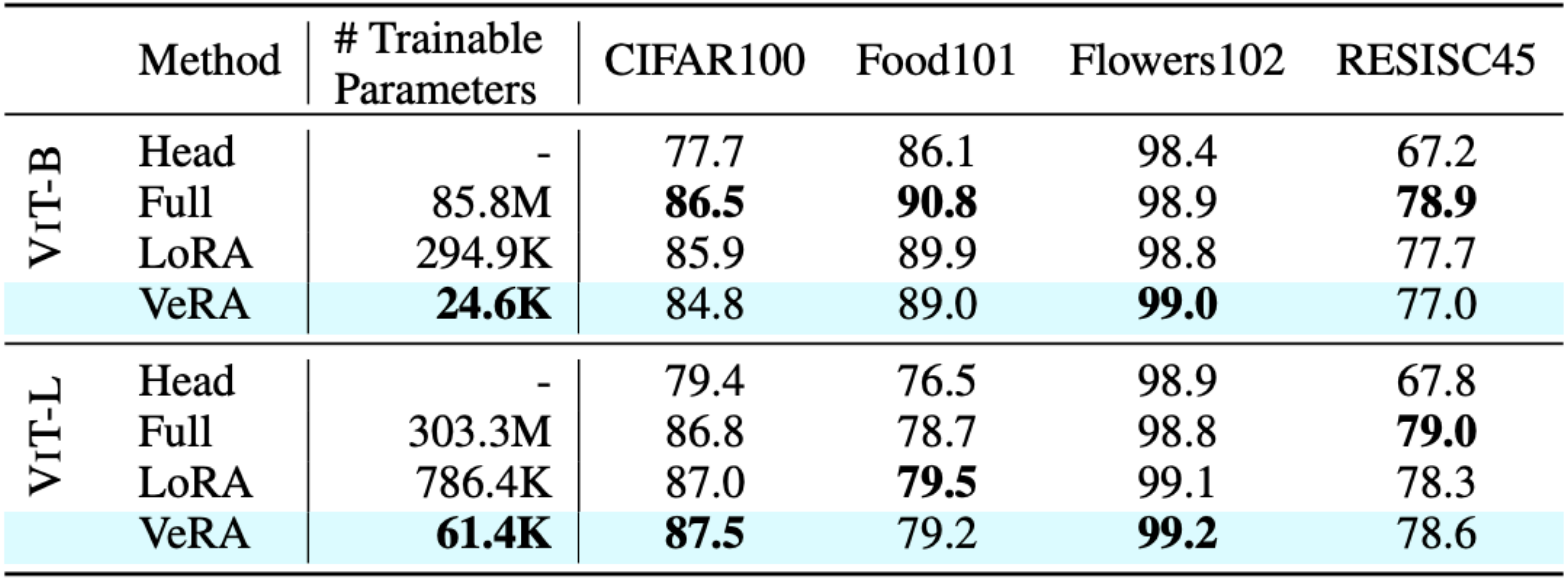
Instruction tuning: better than LoRA with 100x less parameters



Model	Method	# Parameters	Score
Llama 13B	-	-	2.61
LLAMA 7B	LoRA	159.9M	5.03
	VeRA	1.6M	4.77
LLAMA 13B	LoRA	250.3M	5.31
	VeRA	2.4M	5.22
LLAMA2 7B	LoRA	159.9M	5.19
	VeRA	1.6M	5.08
LLAMA2 13B	LoRA	250.3M	5.77
	VeRA	2.4M	5.93

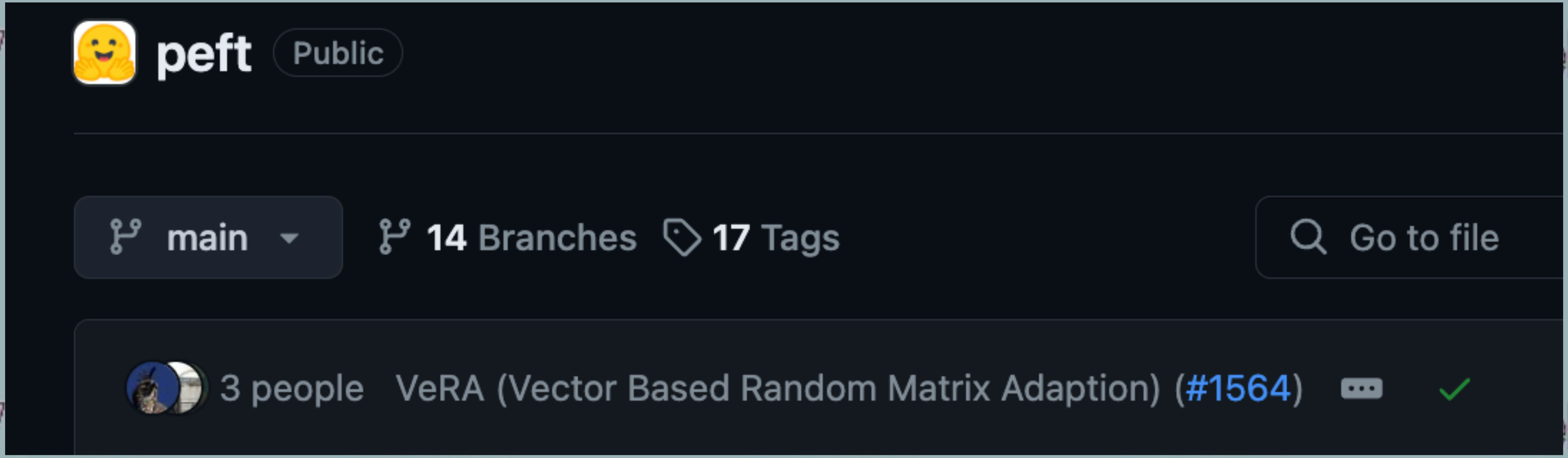
VeRA: Vector-based Random Matrix Adaptation. Kopiczko et al. ICLR 2024

Works also on Image Classification with pretrained ViT



	Method	# Trainable Parameters	CIFAR100	Food101	Flowers102	RESISC45
ViT-B	Head	-	77.7	86.1	98.4	67.2
	Full	85.8M	**86.5**	**90.8**	98.9	**78.9**
	LoRA	294.9K	85.9	89.9	98.8	77.7
	VeRA	**24.6K**	84.8	89.0	**99.0**	77.0
ViT-L	Head	-	79.4	76.5	98.9	67.8
	Full	303.3M	86.8	78.7	98.8	**79.0**
	LoRA	786.4K	87.0	**79.5**	99.1	78.3
	VeRA	**61.4K**	**87.5**	79.2	**99.2**	78.6

VeRA: Vector-based Random Matrix Adaptation. Kopiczko et al. ICLR 2024

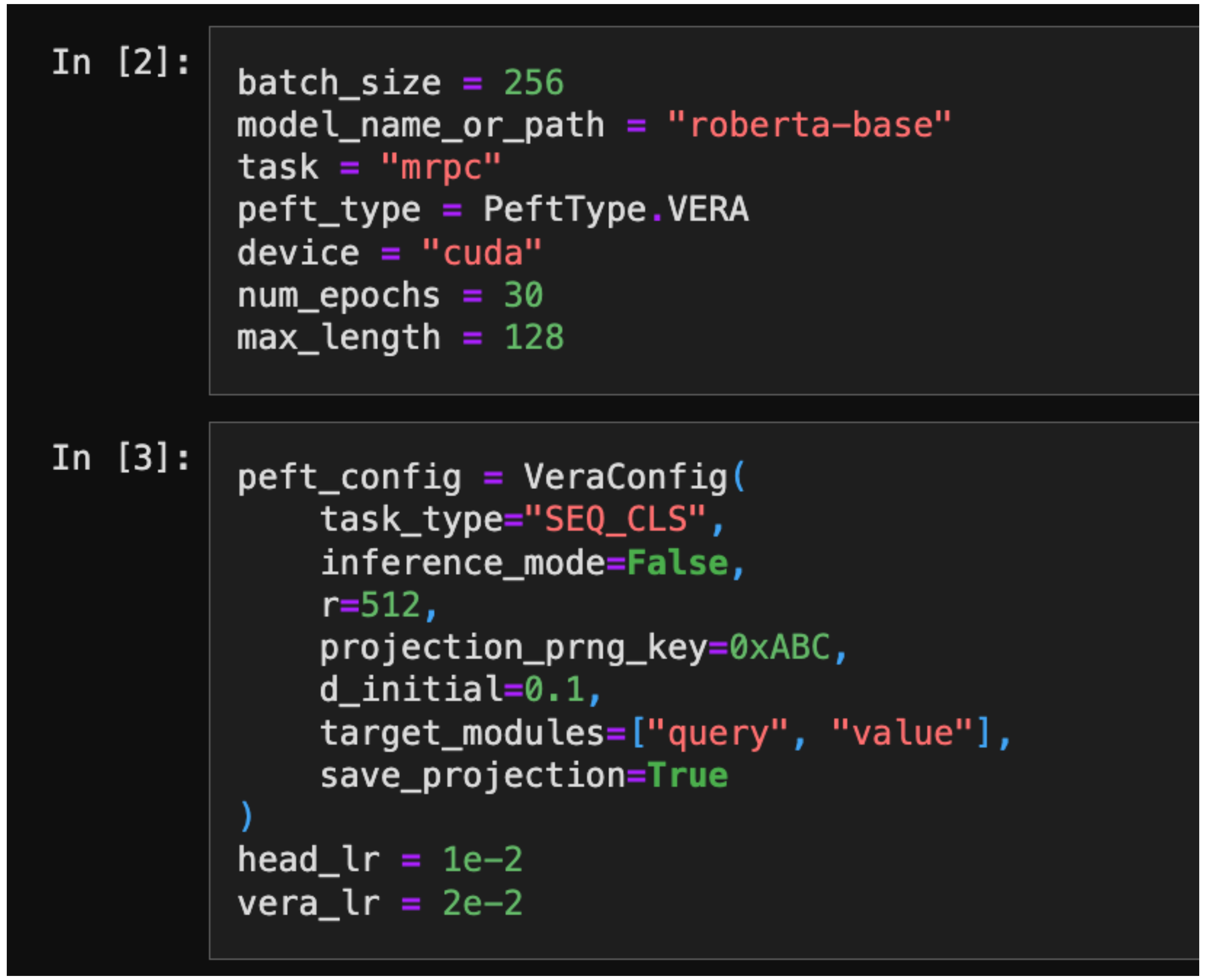


3 people VeRA (Vector Based Random Matrix Adaption) (#1564)

now on HuggingFace!

https://github.com/huggingface/peft

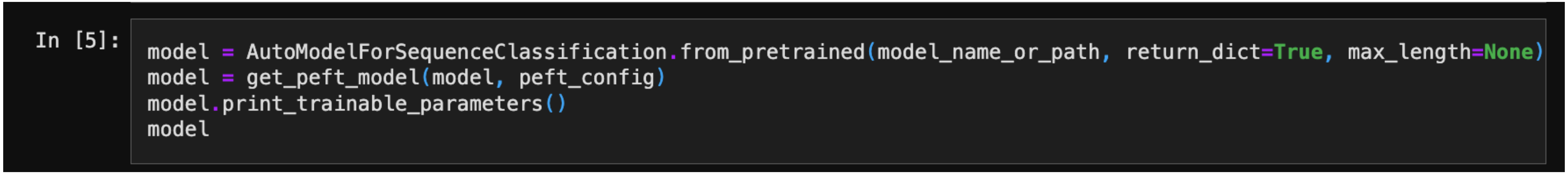
🤗 PEFT: https://github.com/huggingface/peft



```
In [2]:
batch_size = 256
model_name_or_path = "roberta-base"
task = "mrpc"
peft_type = PeftType.VERA
device = "cuda"
num_epochs = 30
max_length = 128
```

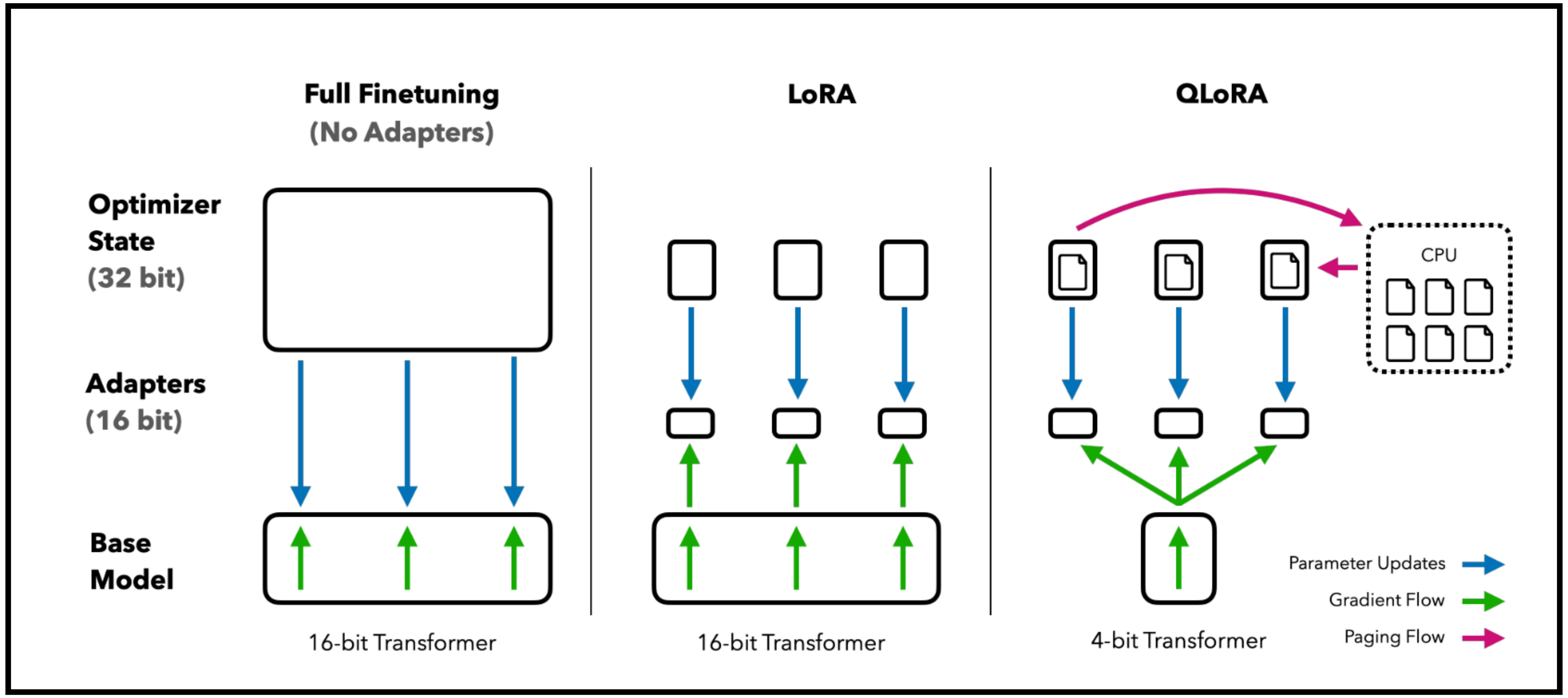
```
In [3]:
peft_config = VeraConfig(
    task_type="SEQ_CLS",
    inference_mode=False,
    r=512,
    projection_prng_key=0xABC,
    d_initial=0.1,
    target_modules=["query", "value"],
    save_projection=True
)
head_lr = 1e-2
vera_lr = 2e-2
```

Super easy to use!
For vision or NLP transformers.



```
In [5]:
model = AutoModelForSequenceClassification.from_pretrained(model_name_or_path, return_dict=True, max_length=None)
model = get_peft_model(model, peft_config)
model.print_trainable_parameters()
model
```

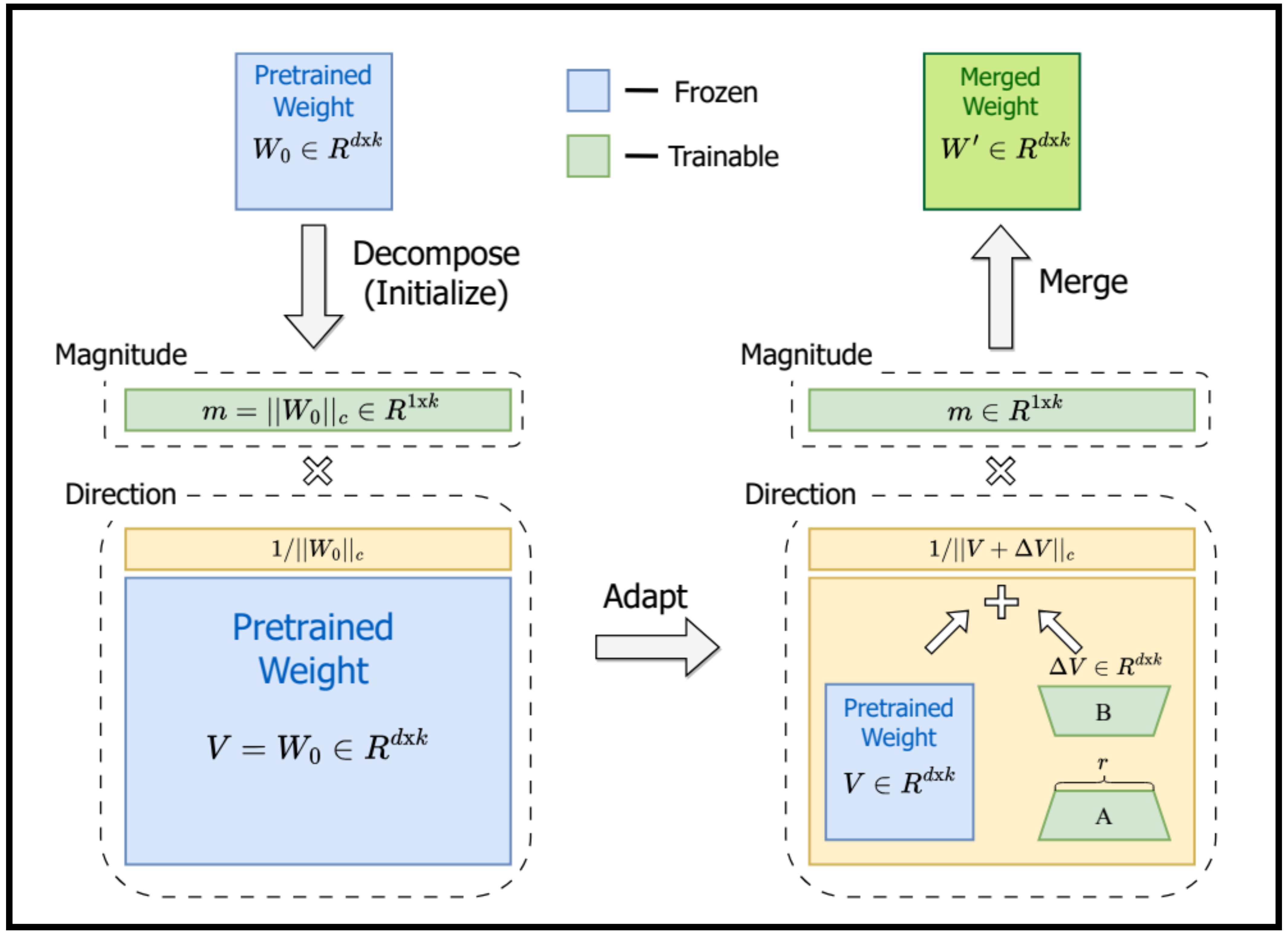
QLoRA



- Better 4-bit datatype
- Double quantisation: quantise the quantisation constants

QLORA: Efficient Finetuning of Quantized LLMs. Dettmers et al. NeurIPS 2023

DoRA: Weight-Decomposed Low-Rank Adaptation



- Adapt the direction, not the magnitude
- See also weight-norm (2016)

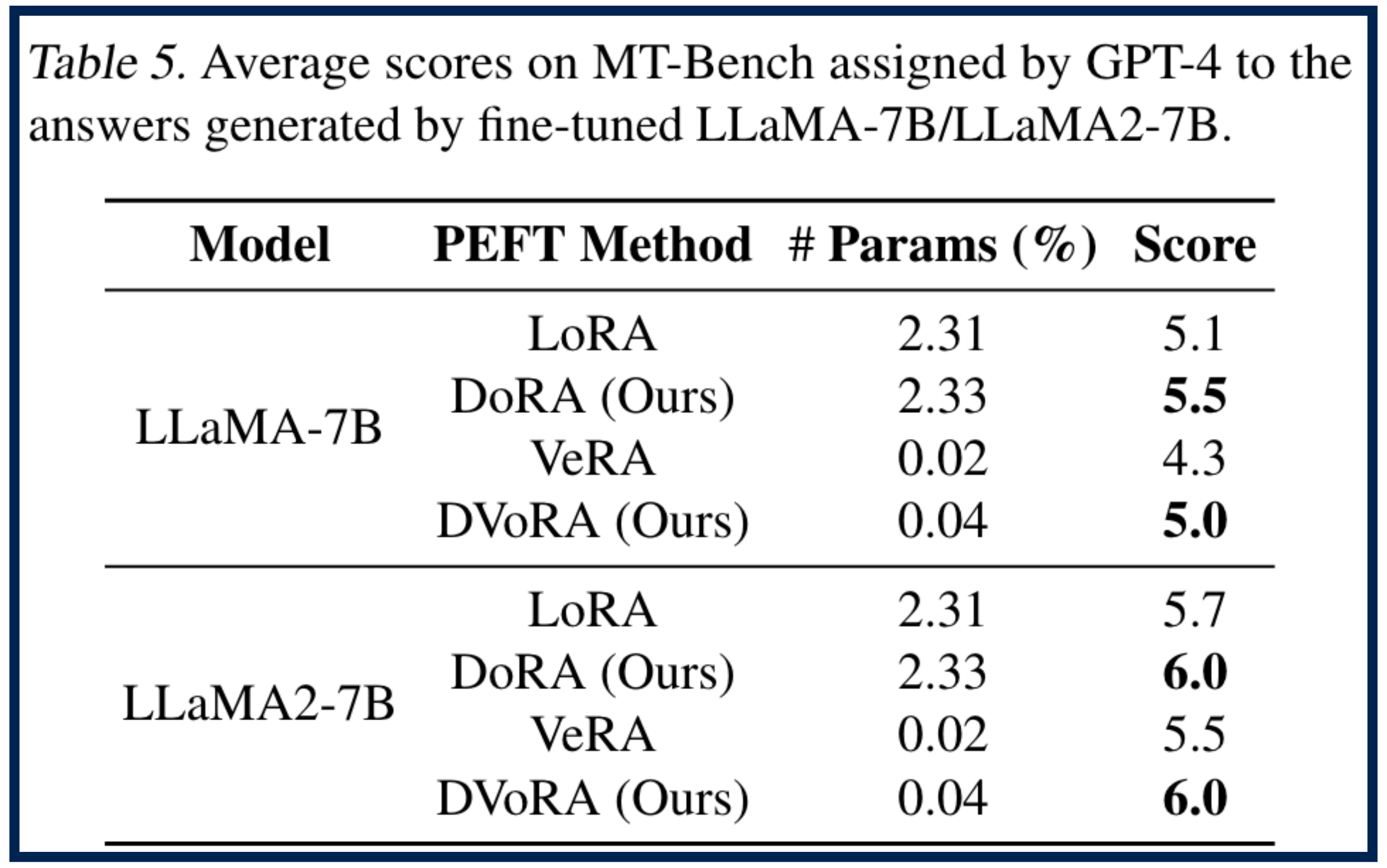


Table 5. Average scores on MT-Bench assigned by GPT-4 to the answers generated by fine-tuned LLaMA-7B/LLaMA2-7B.

Model	PEFT Method	# Params (%)	Score
LLaMA-7B	LoRA	2.31	5.1
	DoRA (Ours)	2.33	**5.5**
	VeRA	0.02	4.3
	DVoRA (Ours)	0.04	**5.0**
LLaMA2-7B	LoRA	2.31	5.7
	DoRA (Ours)	2.33	**6.0**
	VeRA	0.02	5.5
	DVoRA (Ours)	0.04	**6.0**

- Combinable with VeRA

DoRA: Weight-Decomposed Low-Rank Adaptation. Liu et al. 2024
Weight Normalization: A Simple Reparameterization to Accelerate Training of Deep Neural Networks. Salimans & Kingma. NeurIPS 2016

PINs:
Positional Insert unlocks object localisation abilities in VLMs

MICHAEL DORKENWALD, NIMROD BARAZANI, CEES G. M. SNOEK*, YUKI M. ASANO*

CVPR'24

Vision-Language Models are great at many things, but not localisation.

Prompt 1: Provide a bounding box around the cat

Prompt 2: Localise the cat in the image

Dorkenwald, Barazani, Snoek, Asano. PINs: Positional Insert unlocks object localisation abilities in VLMs, CVPR'24.

Our solution: *unlock* localisation abilities in frozen VLMs

VLMs are bad at localising and cannot handle the bbox detection task

But (somewhat noisy) localisation does emerge in some VLMs

Try to **unlock** the forgotten localisation abilities in **frozen VLMs**

Dorkenwald, Barazani, Snoek, Asano. PINs: Positional Insert unlocks object localisation abilities in VLMs, CVPR'24.

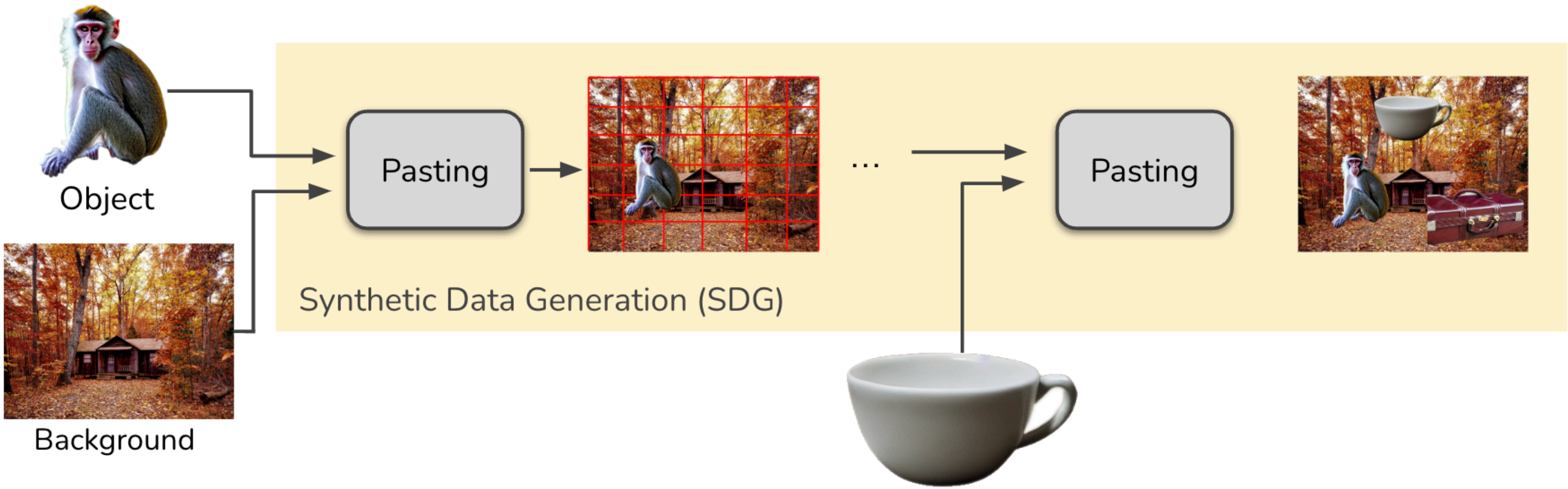
Our approach

frozen VLM, e.g. Flamingo

Positional Insert (PIN) module

Synthetic, unlabeled data

The data



- Because we paste the object, we know it's location
- By pasting multiple objects, we avoid the model focusing on artifacts

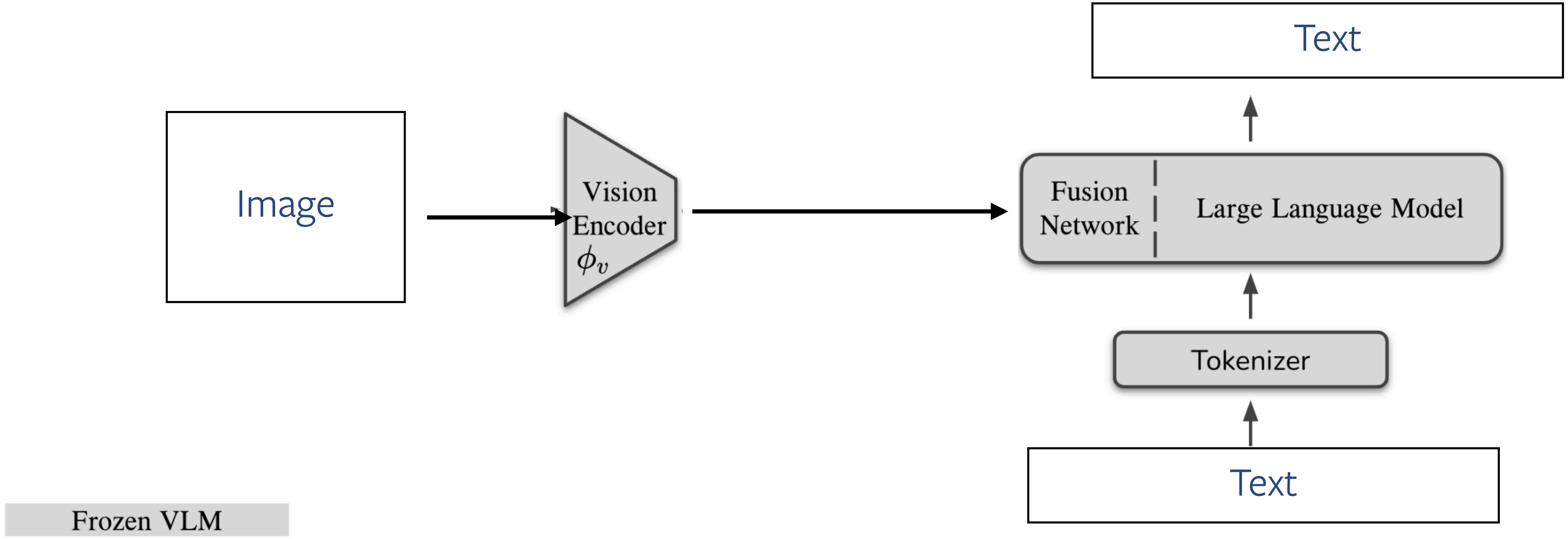
Zhao et al. X-Paste: Revisiting Scalable Copy-Paste for Instance Segmentation using CLIP and StableDiffusion. ICML 2023

Example generated data

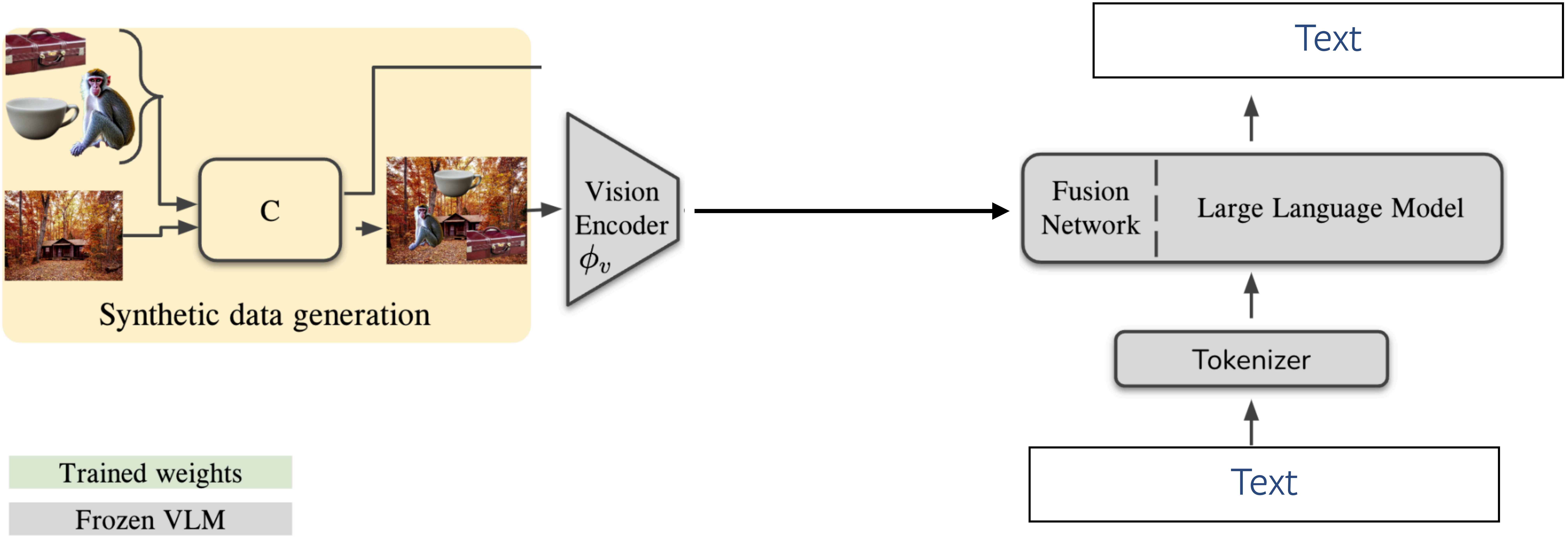
- Note: non-realism is not an issue, as we keep the vision encoder completely frozen
- We only paste objects from categories non-overlapping with our test data
- This means we're in the the zero-shot transfer case

Dorkenwald, Barazani, Snoek, Asano. PINs: Positional Insert unlocks object localisation abilities in VLMs, CVPR'24.

Default Flamingo

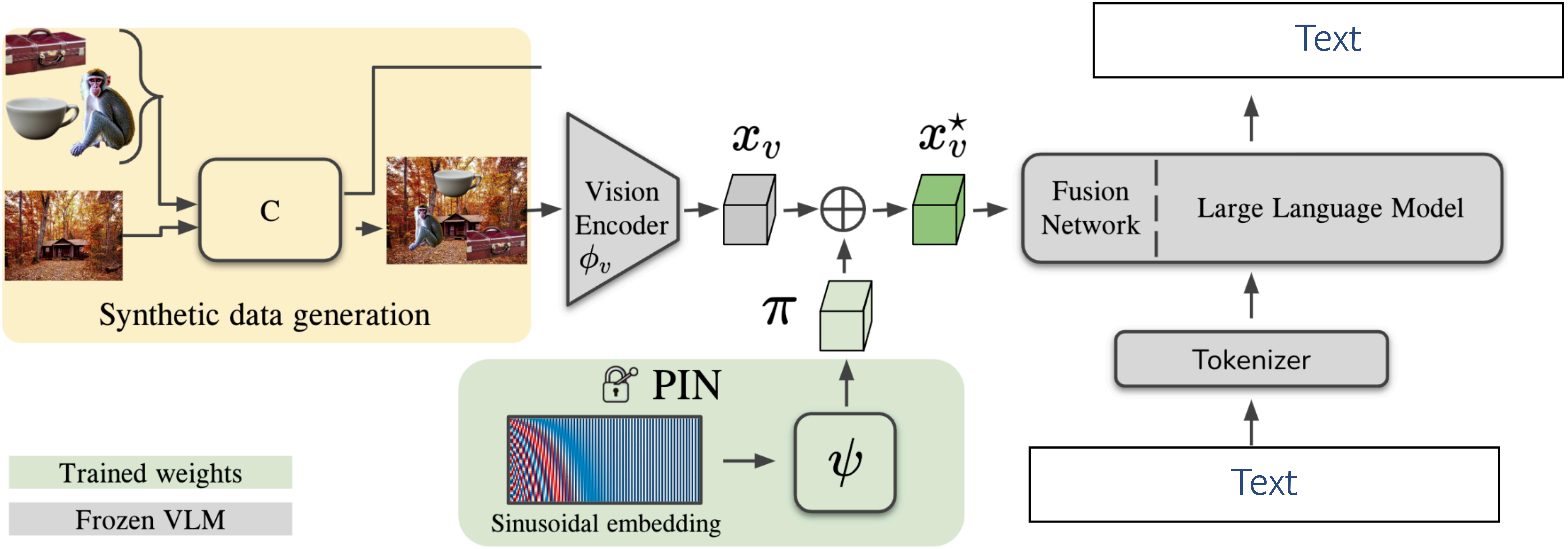


Our method 1: feed the frozen vision encoder synthetic data



Dorkenwald, Barazani, Snoek, Asano. PINs: Positional Insert unlocks object localisation abilities in VLMs, CVPR’24.

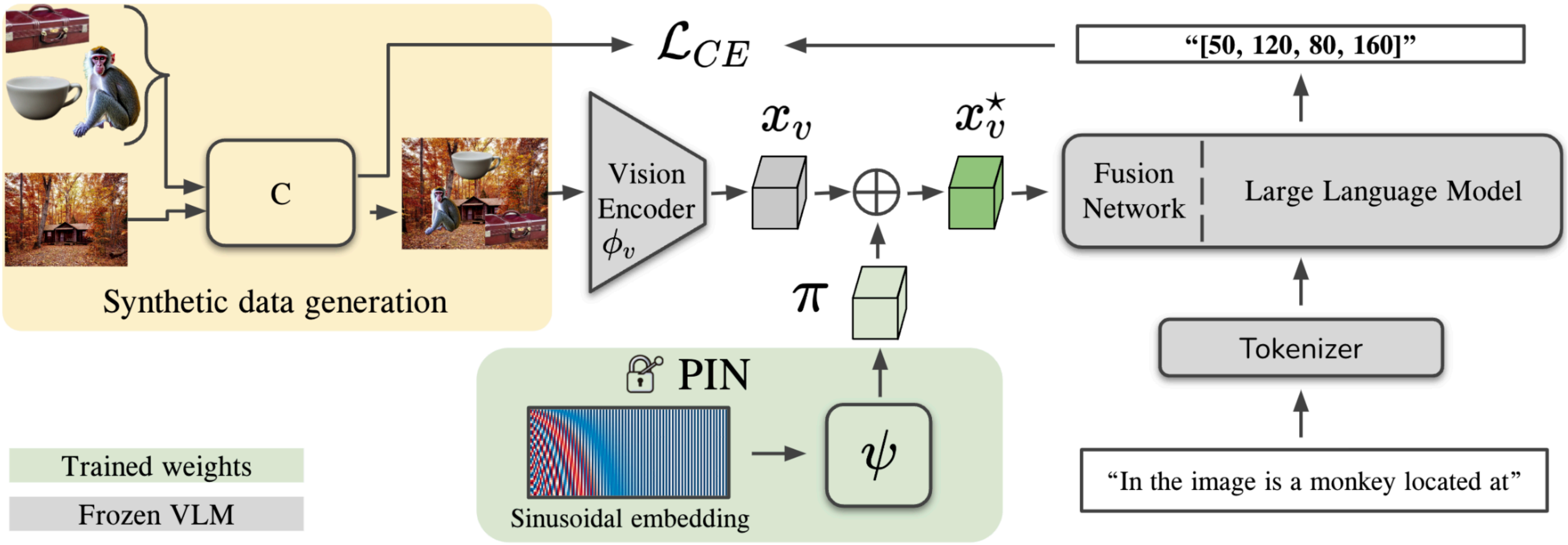
Our method 2: provide VLM spatial learning capacity



Fundamental AI Lab UTN

Dorkenwald, Barazani, Snoek, Asano. PINs: Positional Insert unlocks object localisation abilities in VLMs, CVPR’24.

Our method 3: train using pasted obj locations via next-word prediction



Dorkenwald, Barazani, Snoek, Asano. PINs: Positional Insert unlocks object localisation abilities in VLMs, CVPR’24.

What is the PIN? It's a PEFT method for Vision-Language Models.

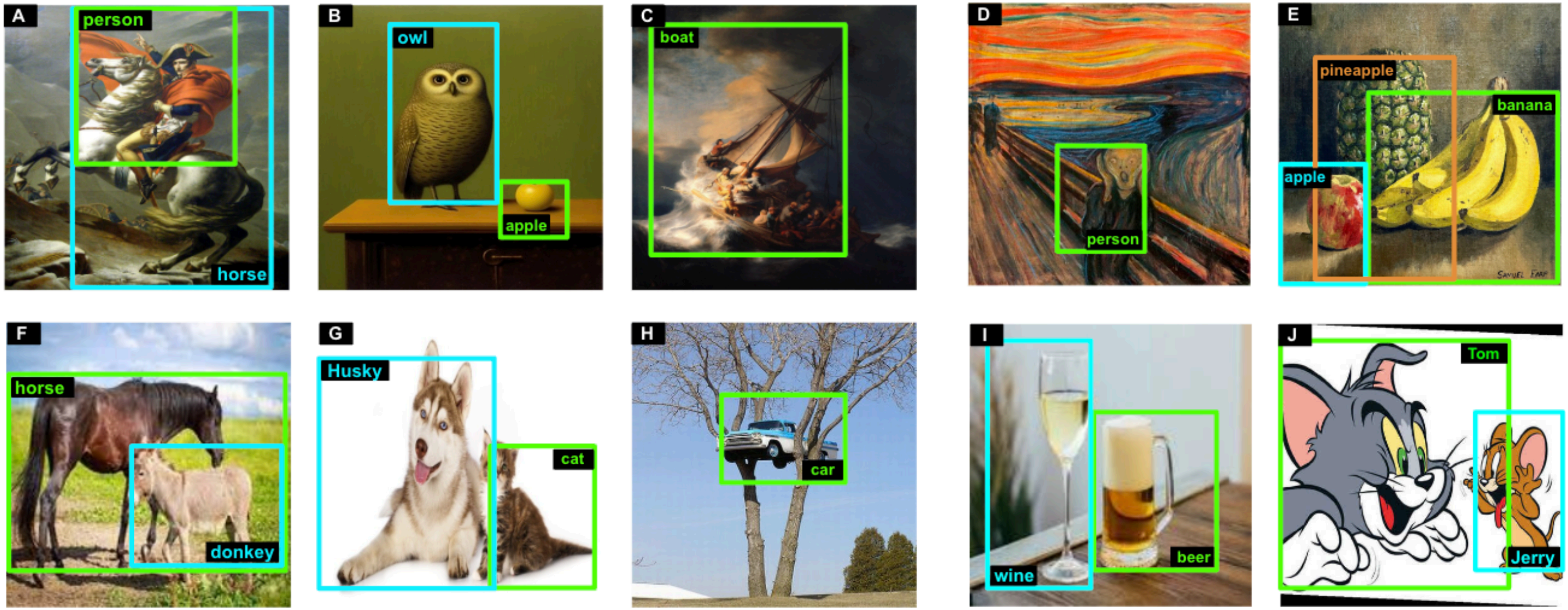
```
pos_encoding = get_sinusoid_encoding_table(n_patches=196, d_hid=64)

MLP = nn.Sequential(
    nn.Linear(64, 512),
    nn.SiLU(),
    nn.LayerNorm(512),
    nn.Linear(512, 768),
    nn.SiLU(),
    nn.LayerNorm(768),
    nn.Linear(768, 1024),
)

PIN = MLP(pos_encoding)
```

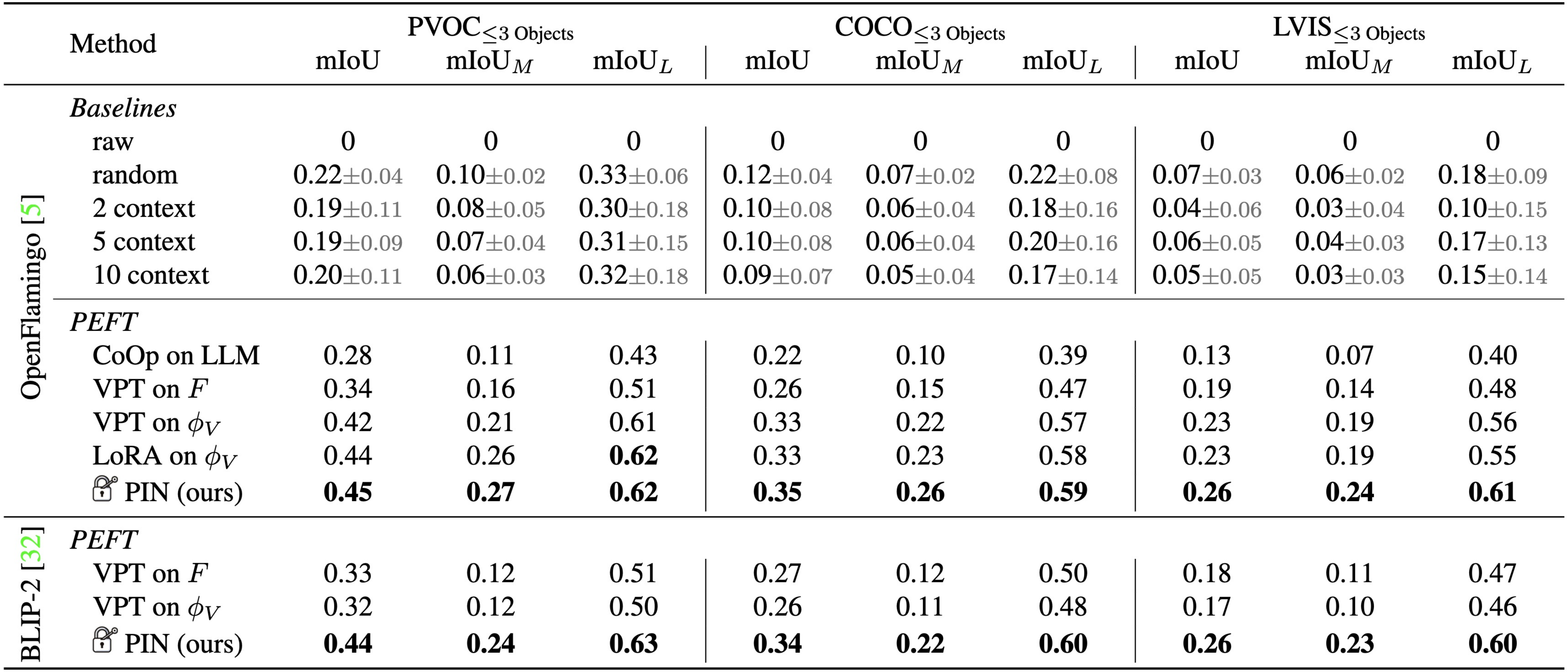
Just 10 LoC!

Results (all on categories not in our pasting-objects)



Dorkenwald, Barazani, Snoek, Asano. PINs: Positional Insert unlocks object localisation abilities in VLMs, CVPR'24.

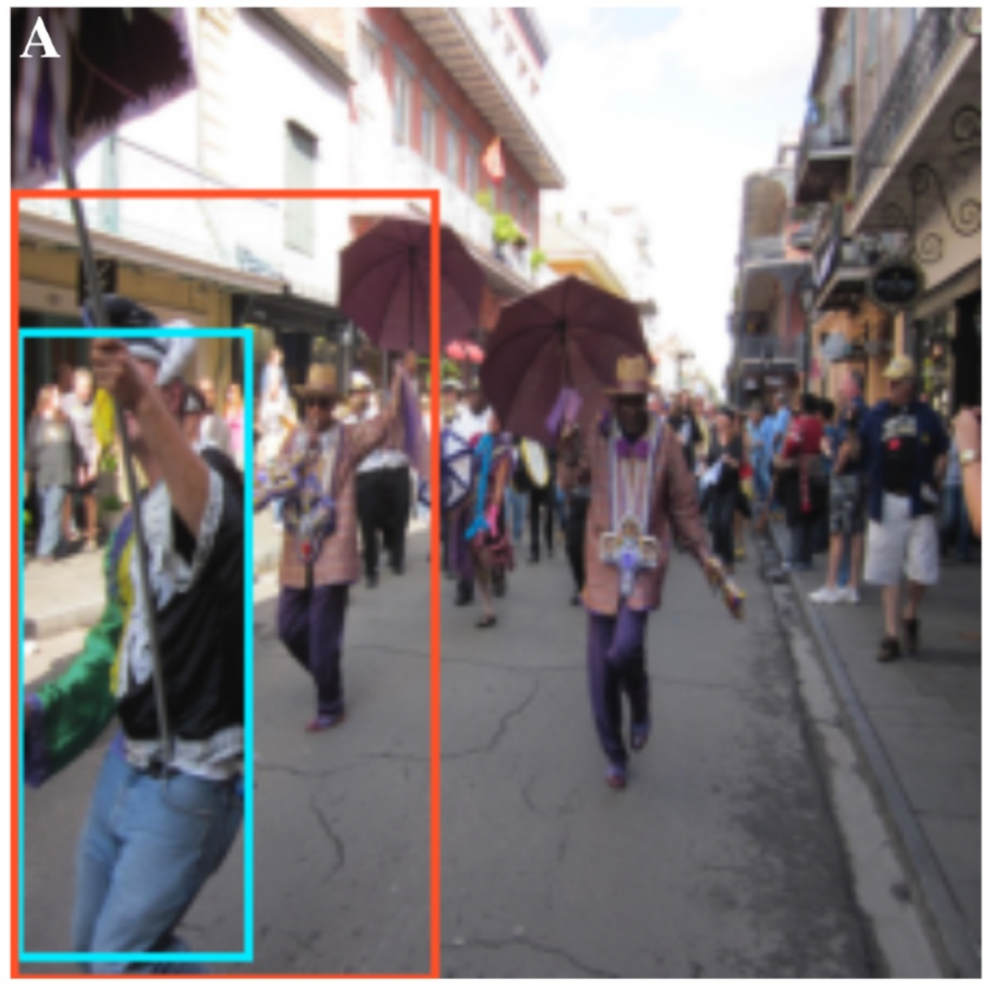
We beat common PEFT methods



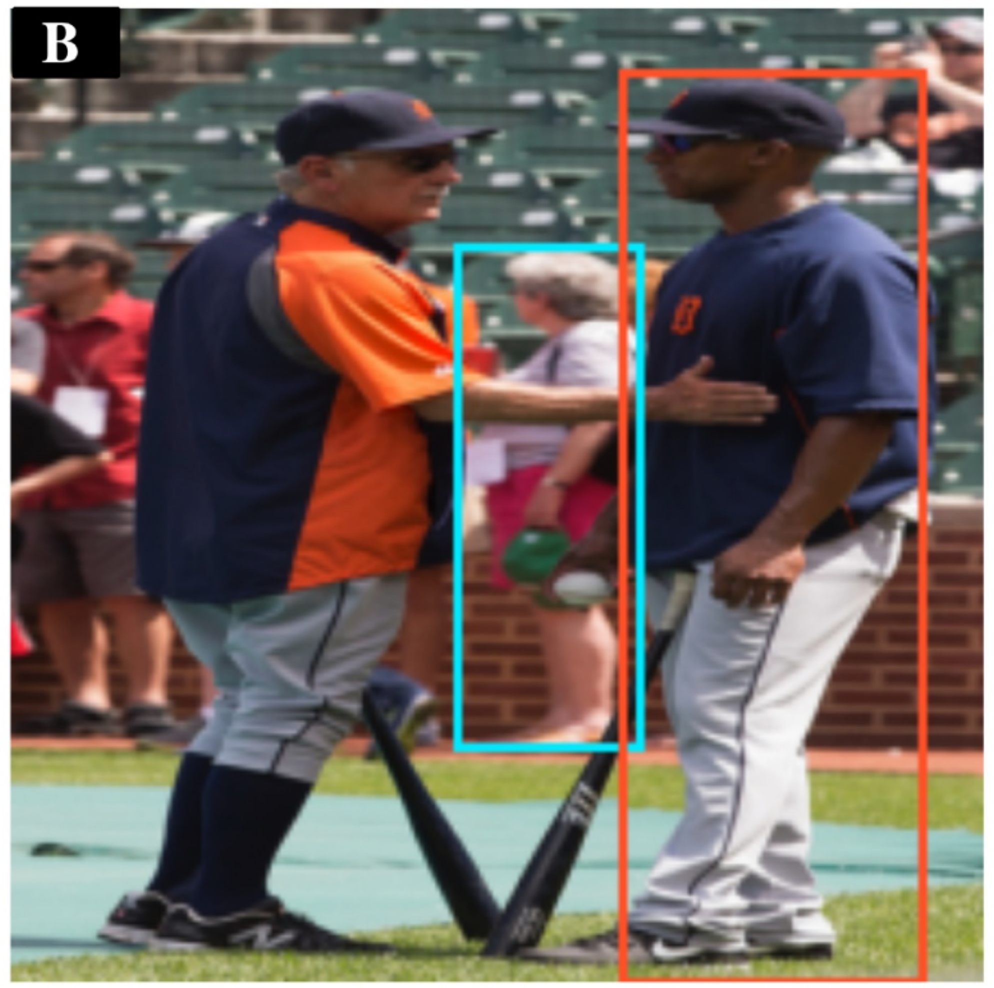
	Method	PVOC$_{\leq 3 \text{ Objects}}$			COCO$_{\leq 3 \text{ Objects}}$			LVIS$_{\leq 3 \text{ Objects}}$		
		mIoU	mIoU$_M$	mIoU$_L$	mIoU	mIoU$_M$	mIoU$_L$	mIoU	mIoU$_M$	mIoU$_L$
OpenFlamingo [5]	*Baselines*									
	raw	0	0	0	0	0	0	0	0	0
	random	0.22±0.04	0.10±0.02	0.33±0.06	0.12±0.04	0.07±0.02	0.22±0.08	0.07±0.03	0.06±0.02	0.18±0.09
	2 context	0.19±0.11	0.08±0.05	0.30±0.18	0.10±0.08	0.06±0.04	0.18±0.16	0.04±0.06	0.03±0.04	0.10±0.15
	5 context	0.19±0.09	0.07±0.04	0.31±0.15	0.10±0.08	0.06±0.04	0.20±0.16	0.06±0.05	0.04±0.03	0.17±0.13
	10 context	0.20±0.11	0.06±0.03	0.32±0.18	0.09±0.07	0.05±0.04	0.17±0.14	0.05±0.05	0.03±0.03	0.15±0.14
	PEFT									
	CoOp on LLM	0.28	0.11	0.43	0.22	0.10	0.39	0.13	0.07	0.40
	VPT on F	0.34	0.16	0.51	0.26	0.15	0.47	0.19	0.14	0.48
	VPT on ϕ_V	0.42	0.21	0.61	0.33	0.22	0.57	0.23	0.19	0.56
	LoRA on ϕ_V	0.44	0.26	**0.62**	0.33	0.23	0.58	0.23	0.19	0.55
	🔓 PIN (ours)	**0.45**	**0.27**	**0.62**	**0.35**	**0.26**	**0.59**	**0.26**	**0.24**	**0.61**
BLIP-2 [32]	*PEFT*									
	VPT on F	0.33	0.12	0.51	0.27	0.12	0.50	0.18	0.11	0.47
	VPT on ϕ_V	0.32	0.12	0.50	0.26	0.11	0.48	0.17	0.10	0.46
	🔓 PIN (ours)	**0.44**	**0.24**	**0.63**	**0.34**	**0.22**	**0.60**	**0.26**	**0.23**	**0.60**

Dorkenwald, Barazani, Snoek, Asano. PINs: Positional Insert unlocks object localisation abilities in VLMs, CVPR'24.

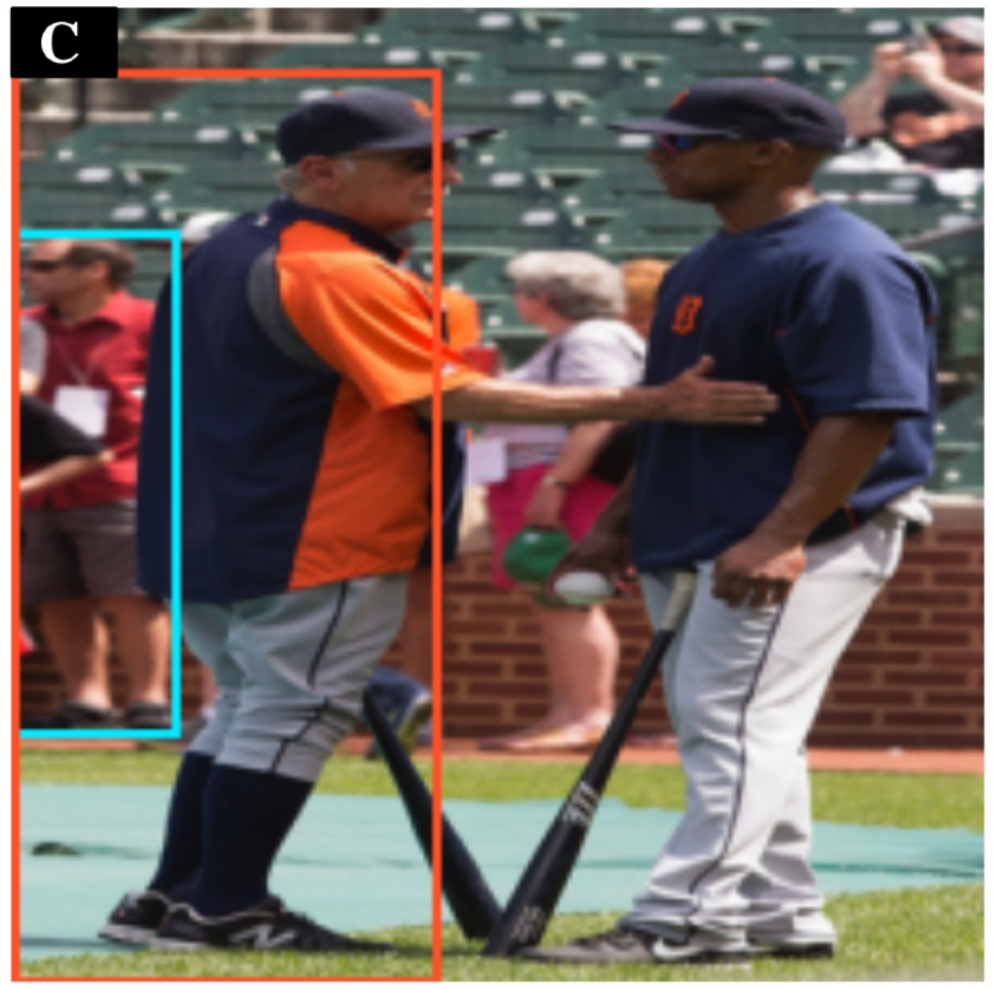
With slight modification, can work on RefCOCO.



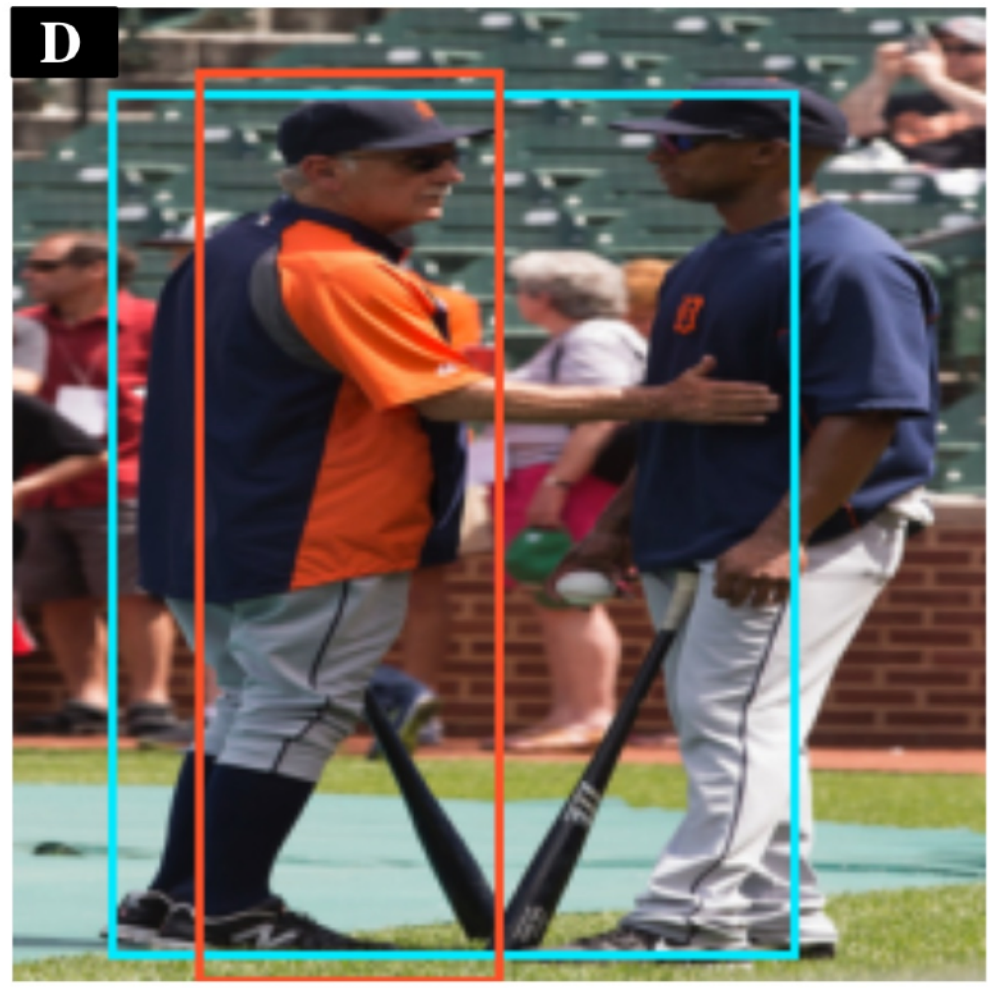
“Left black shirt”



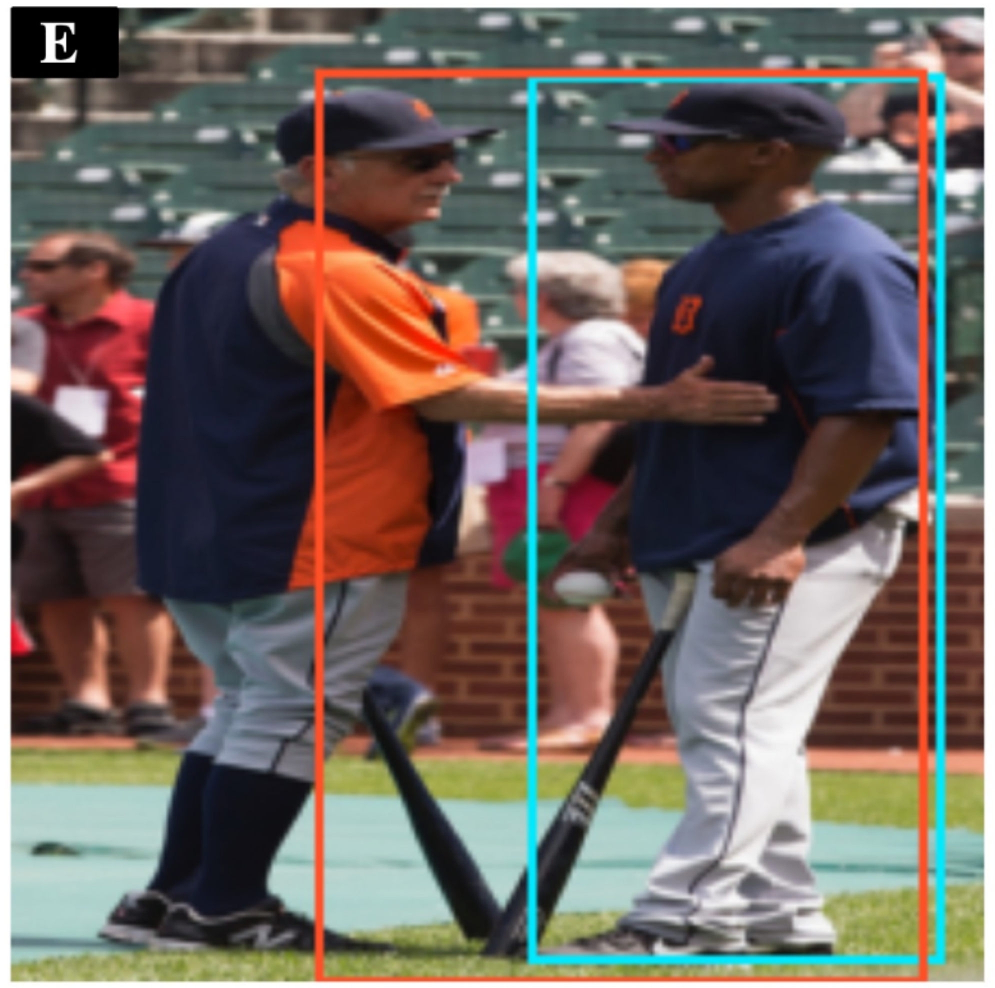
“Old lady in between the players”



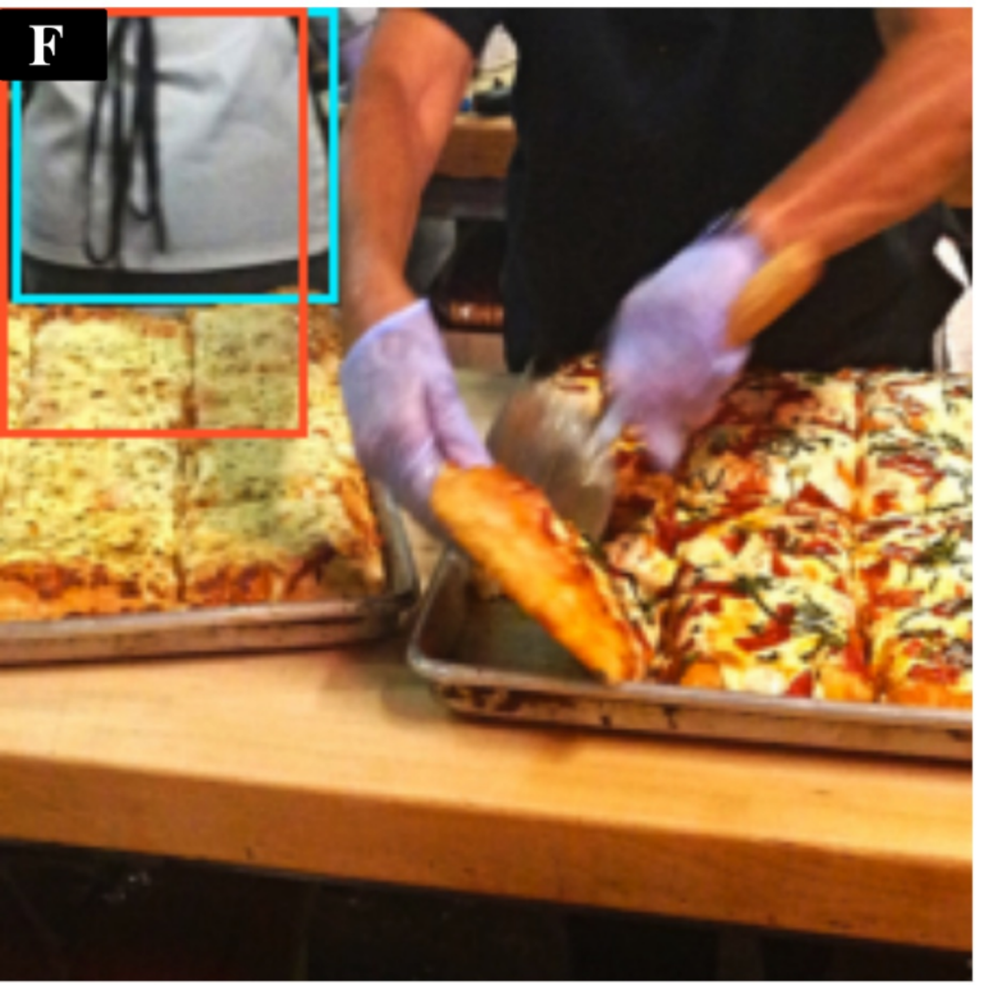
“A guy in red on left”



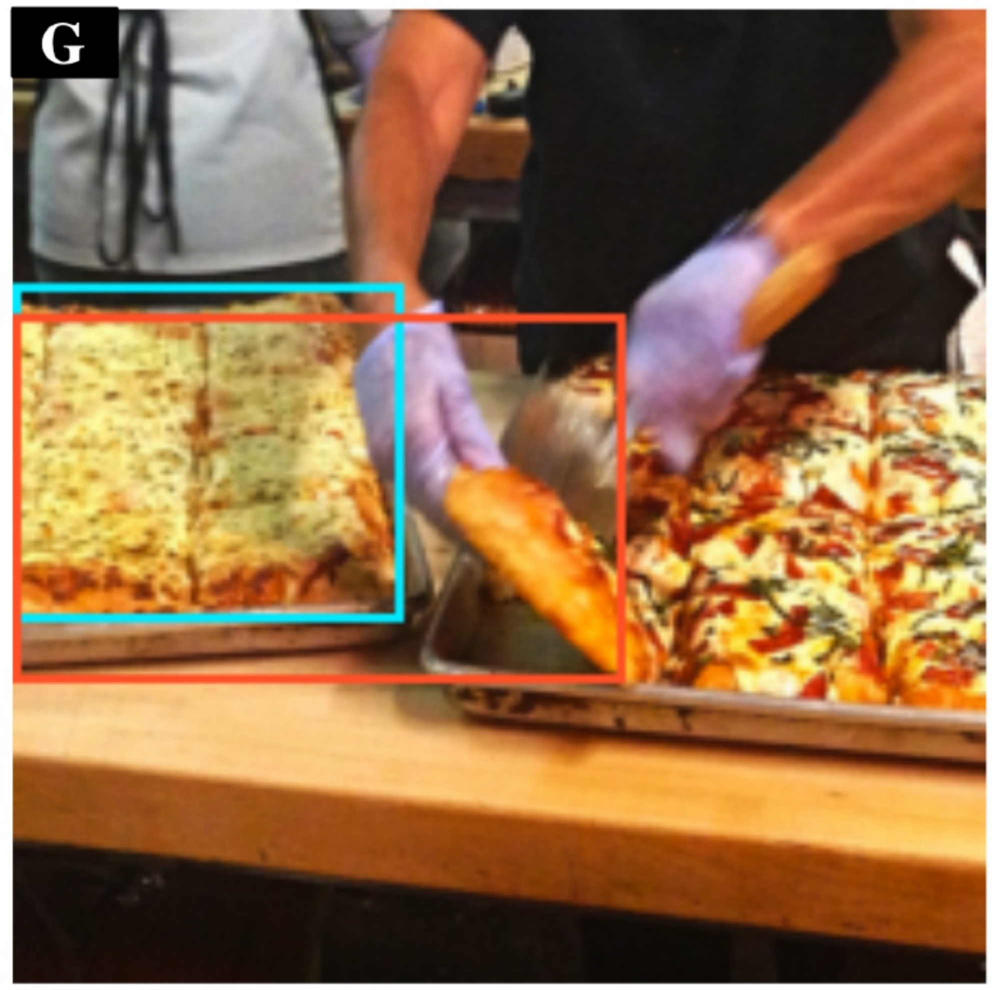
“Guy in orange”



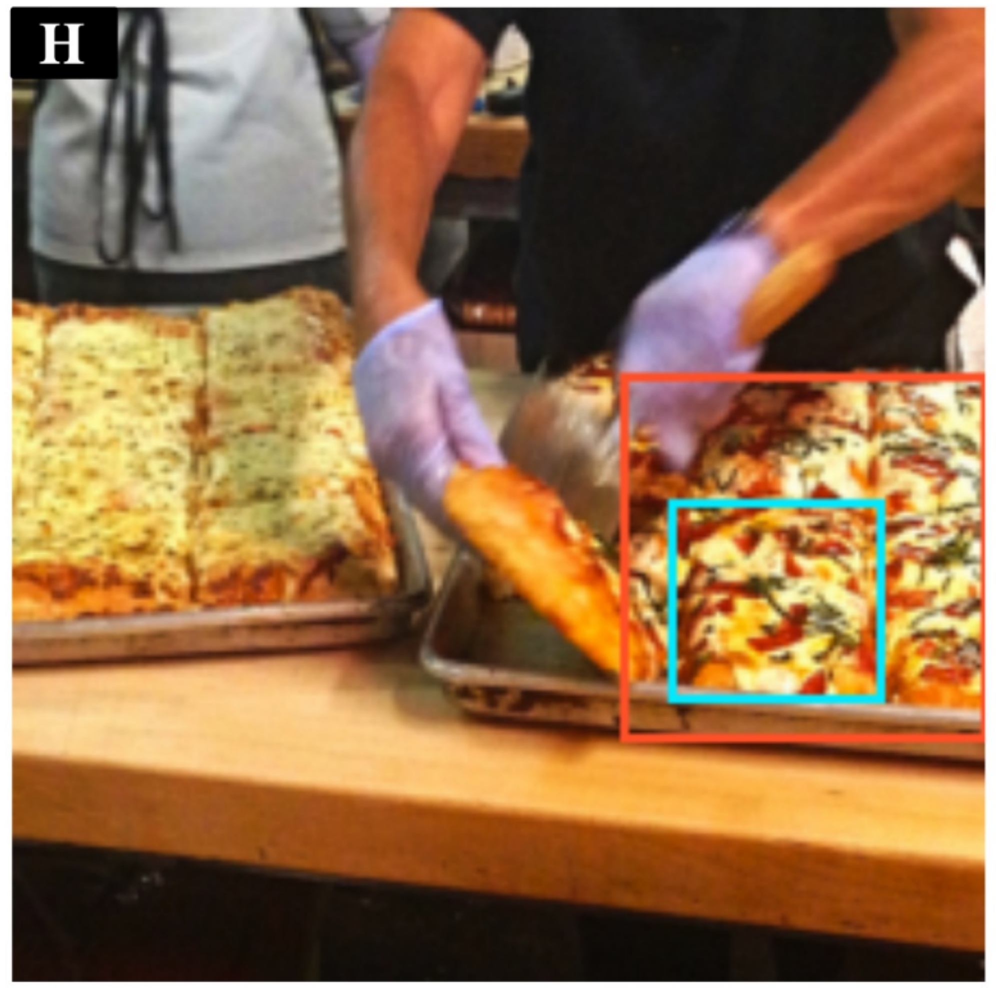
“Right player”



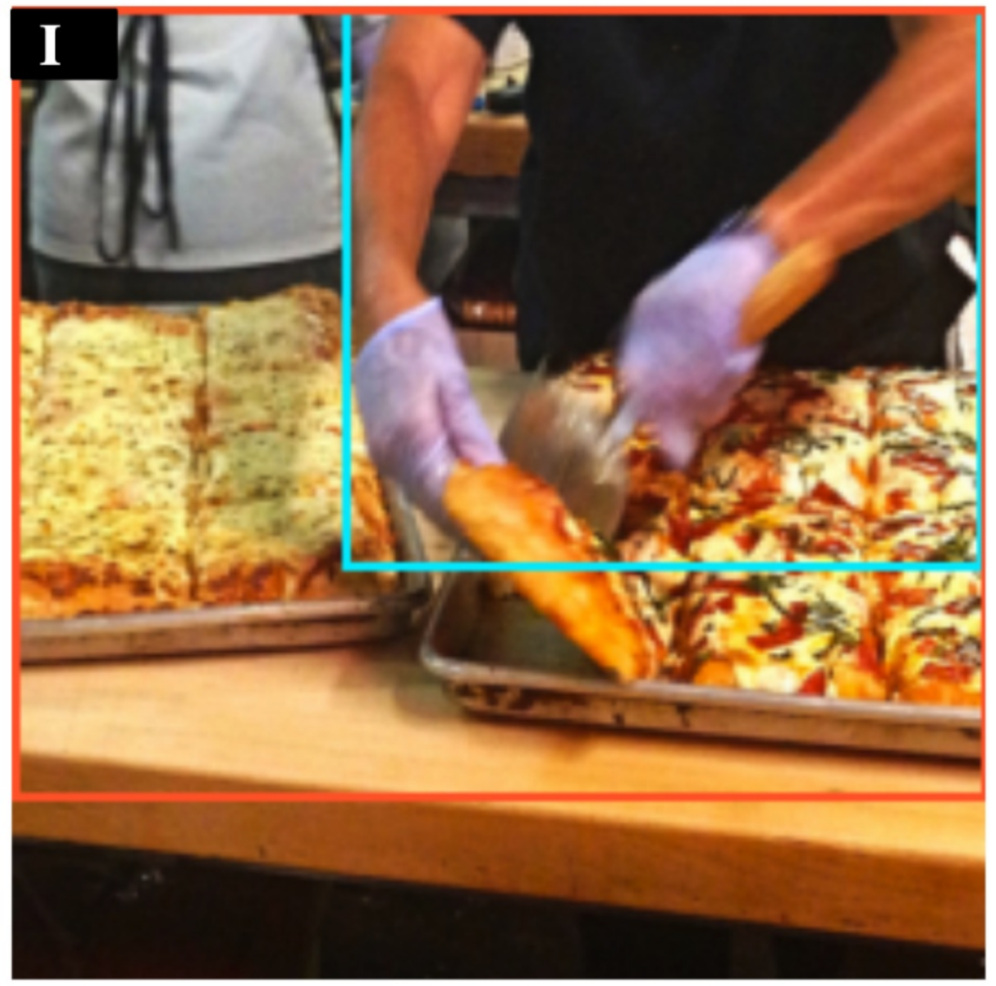
“Top left apron strings”



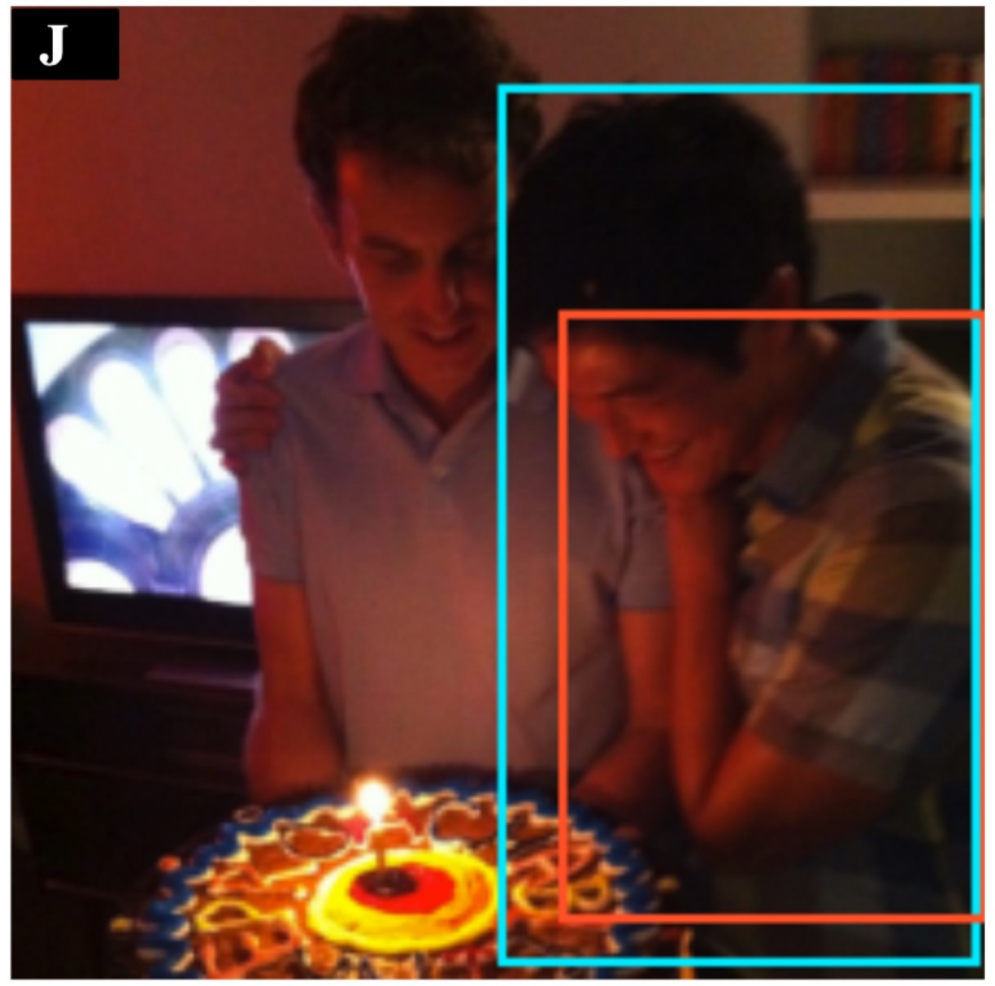
“Pizza squares left”



“Pizza right front piece in middle”



“A man black”

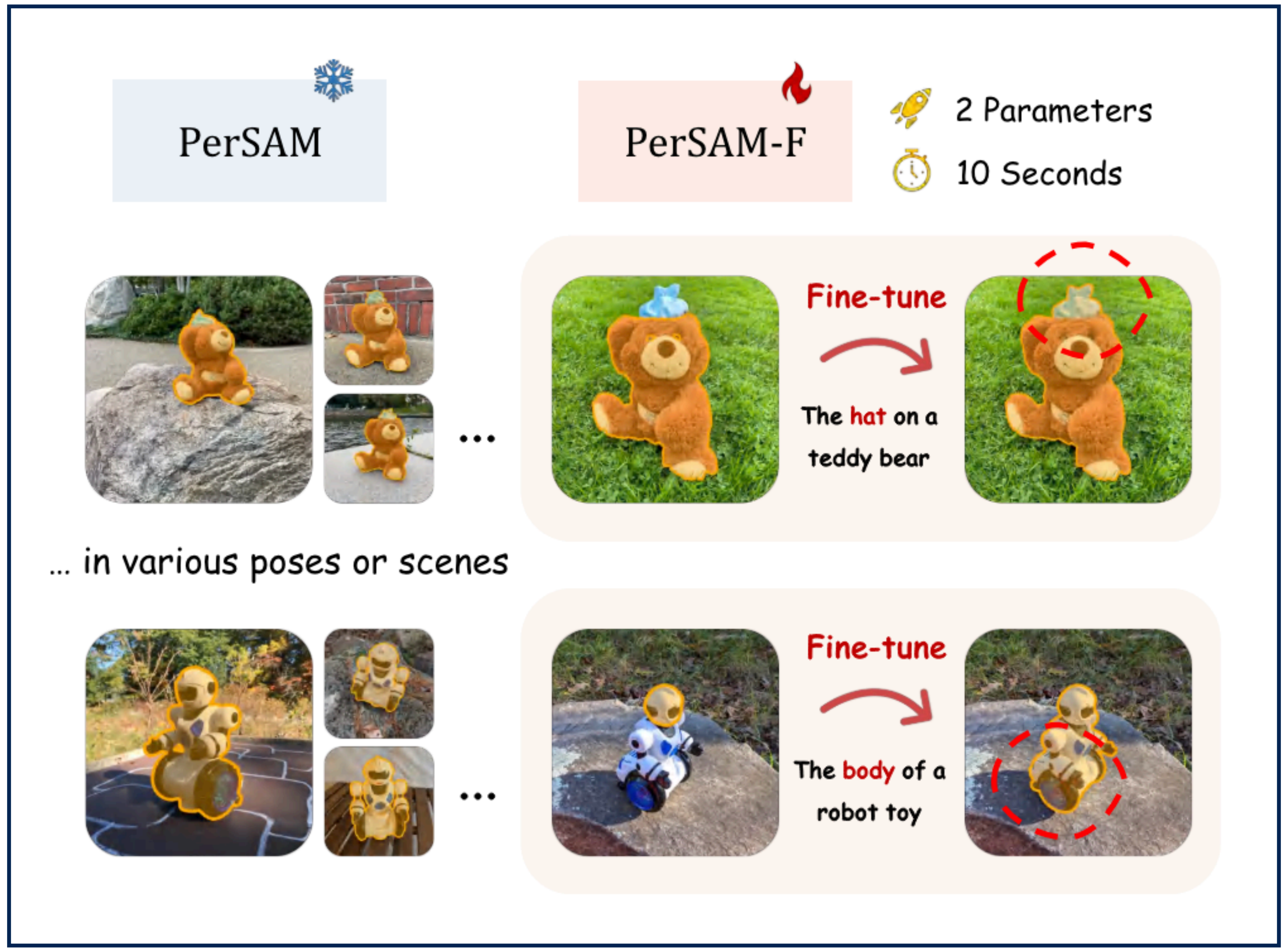


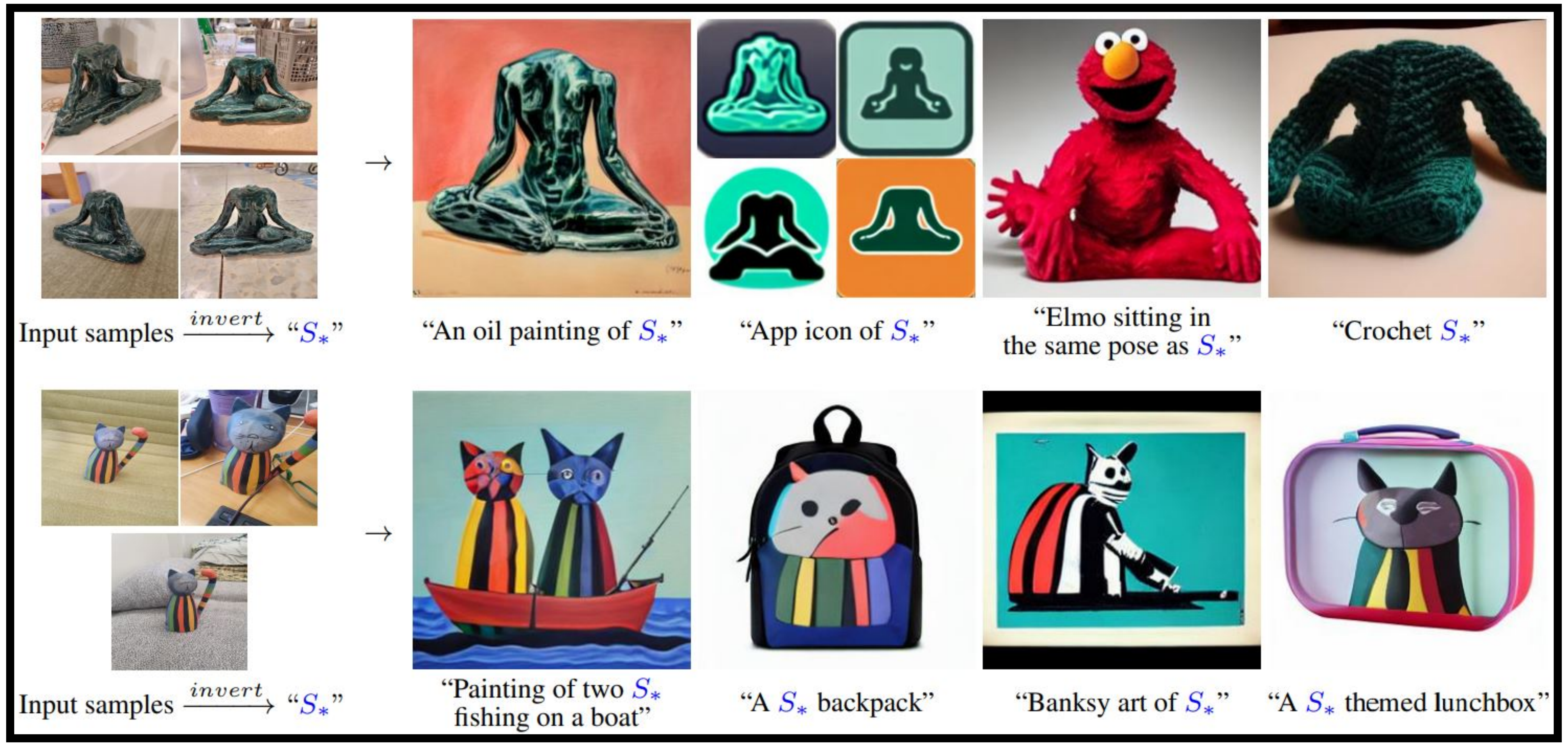
“A right person”

🟥 Predictions 🟦 Ground Truth

Topics / related ideas not covered in this lecture

- text-inversion / DreamBooth
- personalised SAM
- Early-fusion models
- Video LLMs




Input samples $\xrightarrow{invert}$ "S_*"
"An oil painting of S_*"
"App icon of S_*"
"Elmo sitting in the same pose as S_*"
"Crochet S_*"
Input samples $\xrightarrow{invert}$ "S_*"
"Painting of two S_* fishing on a boat"
"A S_* backpack"
"Banksy art of S_*"
"A S_* themed lunchbox"

An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. Gal et al. ICLR 2023
DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. Ruiz et al. ICCV 2023
Personalize Segment Anything Model with One Shot. Zhang et al. ICLR 2024

Take 2 minutes to write down what you've learned so far in this lecture, also include what you find hard to understand.

Next, turn to your neighbor and share notes for 4min.

Recap

Single-modal self-supervised pretraining methods (MAE, DINOv2, Franca, SimCLR, GPT)

Multi-modal pretraining (CLIP, ALIGN, CoCa)

Beyond contrastive (BLIP, ClipCap)

Multimodal Large Language Models (Frozen, Flamingo, Llava, BLIP3o)

Tasks (VQA, VisDial)

Multimodal few-shot learning

Pretraining Datasets (CC3M, LAION, ..)

Text-conditional image generative models

Large Model Adaptation methods (Promt learning, LoRA, Adapters, VeRA, PIN)

X: @y_m_asano
email: yuki.asano@utn.de

Fundamental
AI Lab

UTN