

Foundation Models and Prompts

Danna Gurari

University of Colorado Boulder
Fall 2024



Review

- Last lecture:
 - Multimodal applications
 - Image captioning dataset challenges
 - Image captioning algorithms
 - Visual question answering dataset challenges
 - Discussion
- Assignments (Canvas):
 - Reading assignment was due earlier today
 - Project outline due on Wednesday
 - Reading assignment due in one week
- Questions?

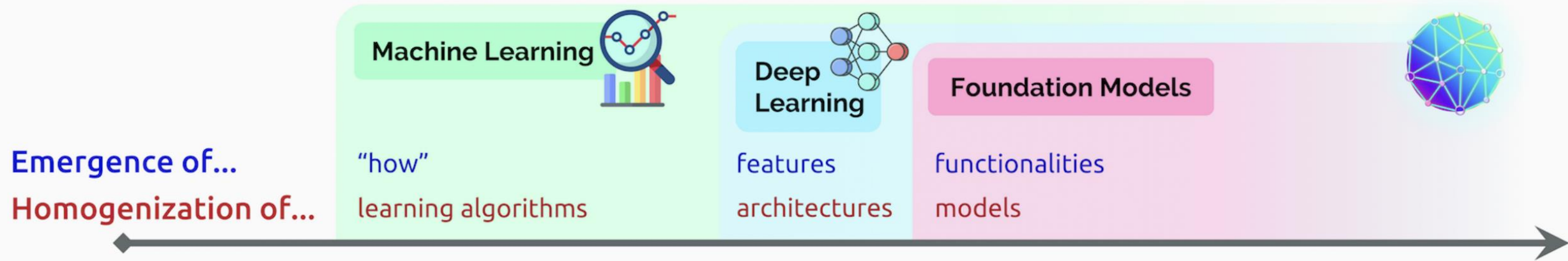
Today's Topics

- Foundation Models
- Textual Prompting & Zero-shot Learning
- Visual Prompting & In-context Few-shot Learning
- Prompt Tuning
- Discussion (chosen by YOU 😊)

Today's Topics

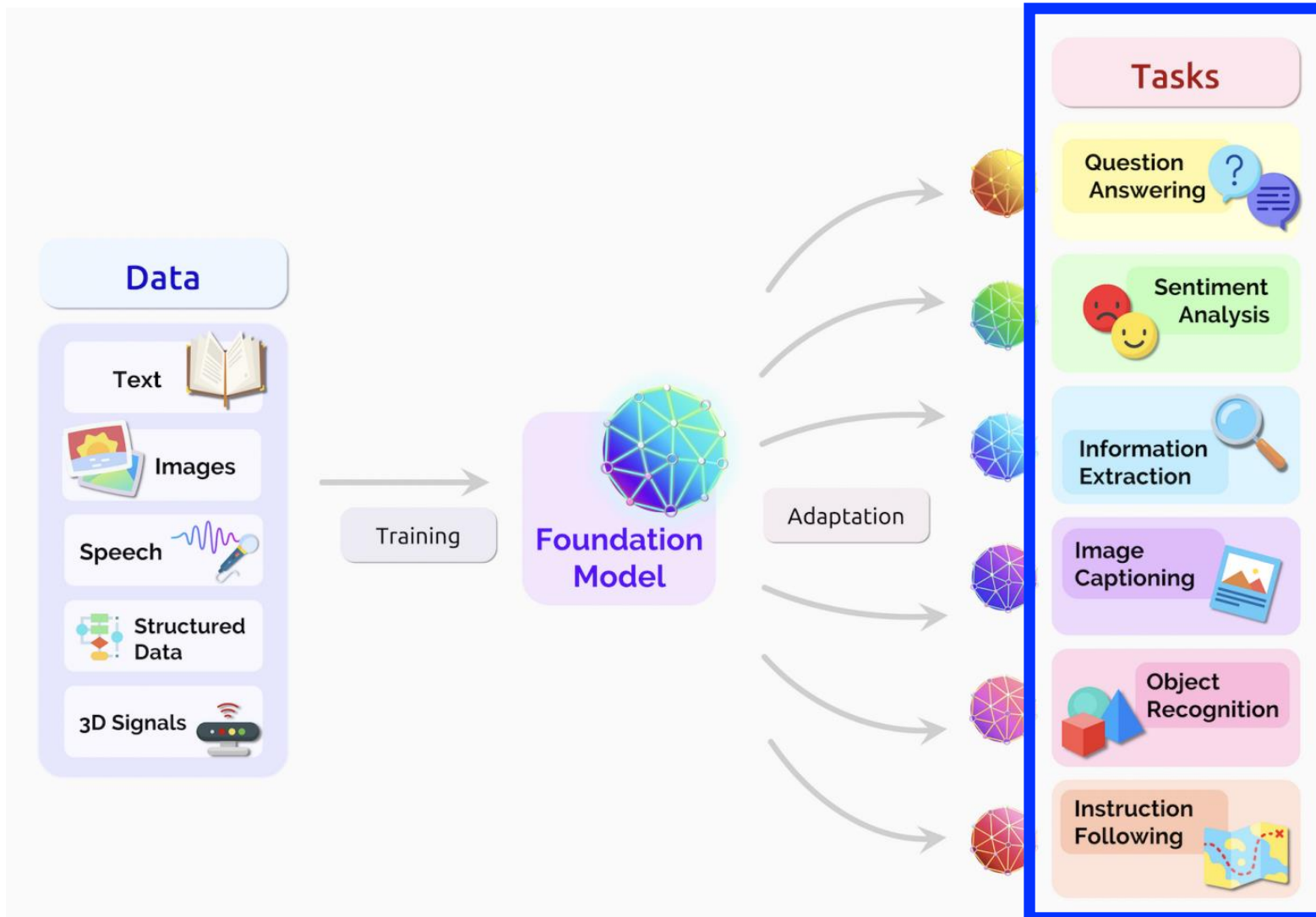
- Foundation Models
- Textual Prompting & Zero-shot Learning
- Visual Prompting & In-context Few-shot Learning
- Prompt Tuning
- Discussion (chosen by YOU 😊)

Definition of “Foundation Model”



Coined in 2021, it references the recent paradigm shift to develop a single model that can implicitly support many downstream tasks.

Foundation Models: Development Pipeline



Evaluate with modern benchmark datasets for many:

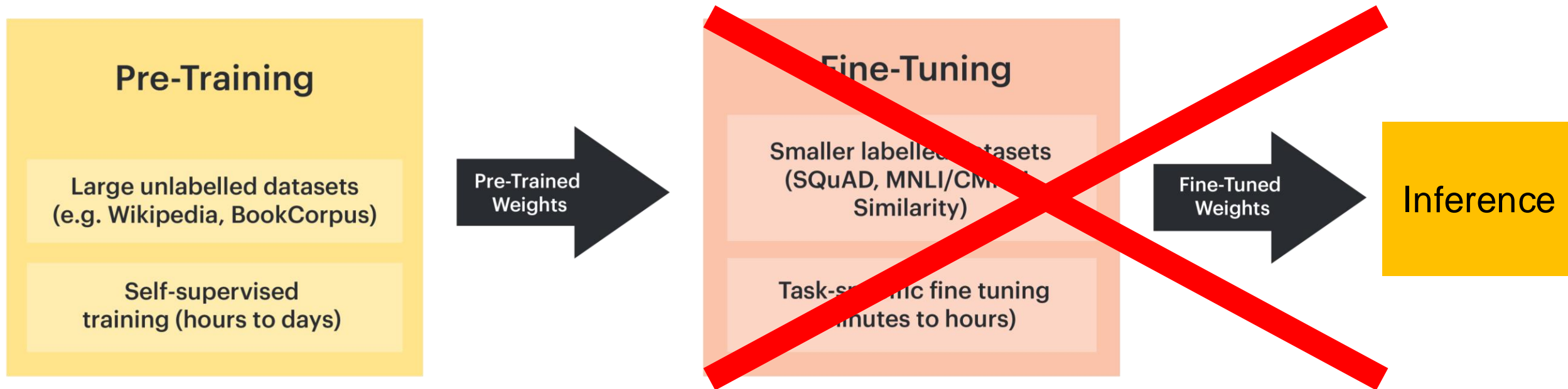
1. **Different tasks** (e.g., object recognition, scene classification)
2. **Different distributions of the same task** (e.g., ImageNet versus data from blind people)

Foundation Models: Why Now?

Availability of key ingredients:

1. Transformer model architecture
2. Lots more training data by using Internet data
3. Sufficient hardware with modern GPUs

Foundation Model Novelty

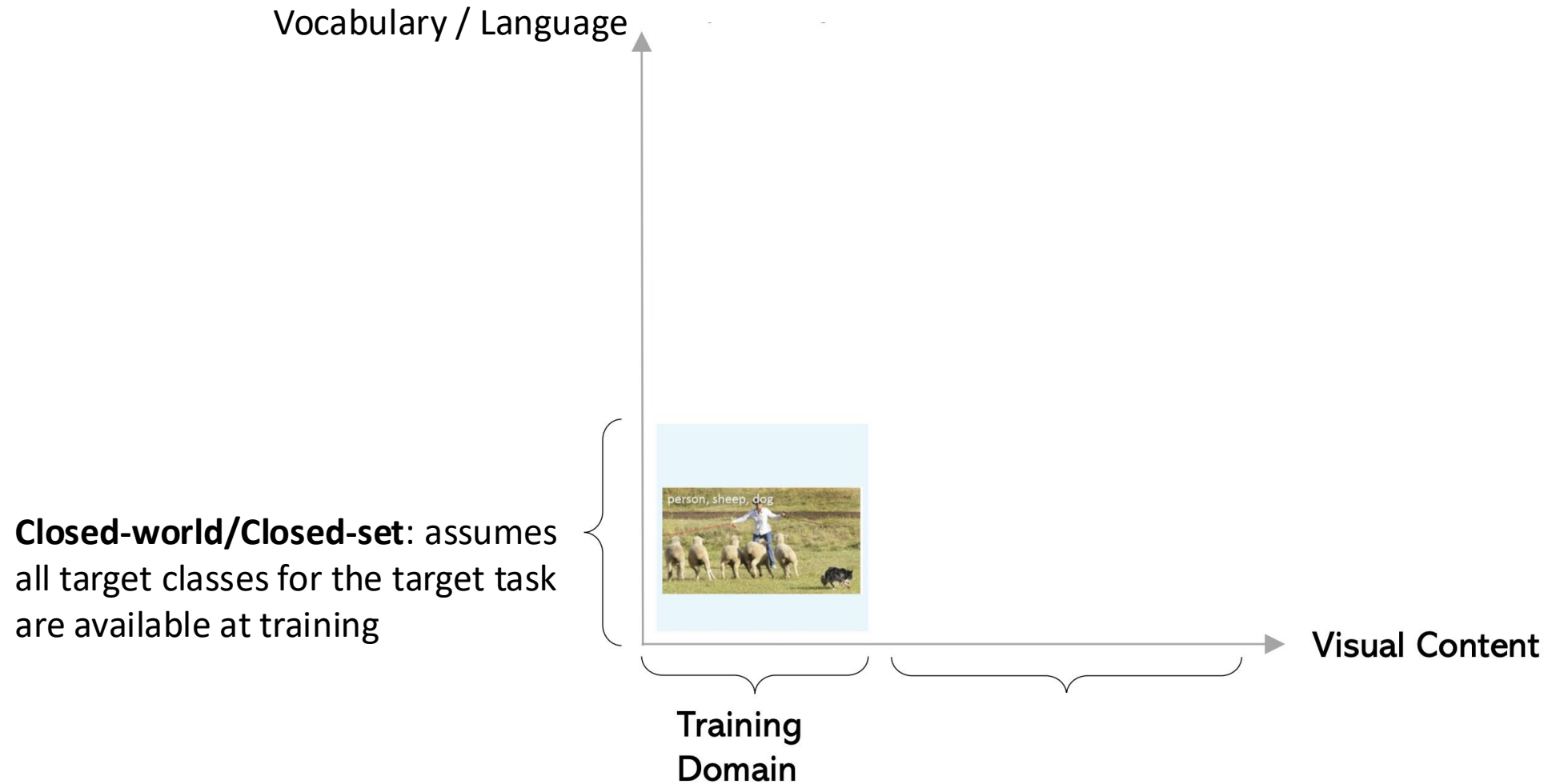


New emergent behavior discovered around 2018 (in NLP) that a foundation model can be used ***as is*** for many downstream tasks with ***prompting!***

Foundation Model Novelty

As a Result, Foundation Models Can Generalize
Beyond The Closed-World Setting with Limited/No Training Data

Beyond Closed-World Setting



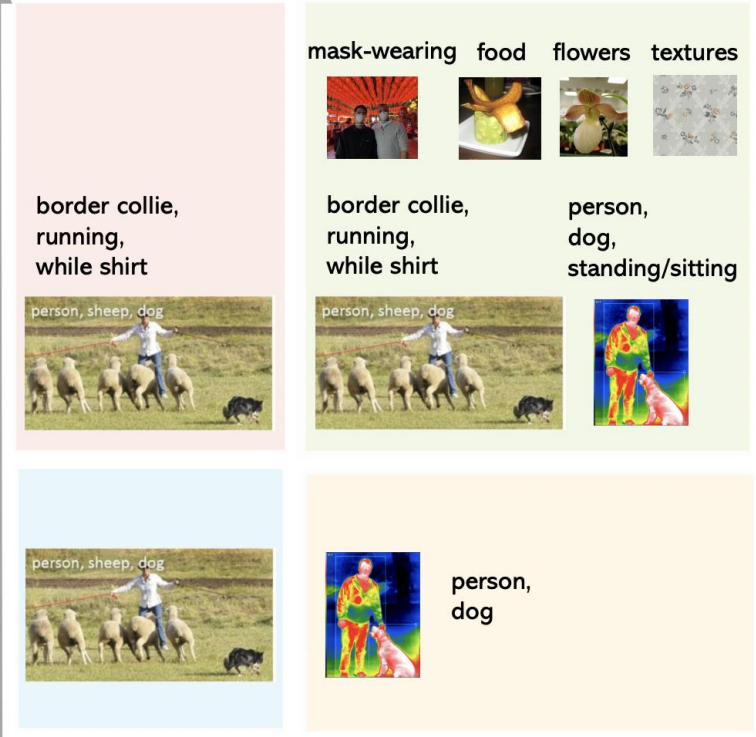
Beyond Closed-World Setting

Vocabulary / Language

Open vocabulary/Zero-shot:
 generalize to a new task with *no labeled training data for the target task* (open vocab permits same-category annotations for other tasks, such as captions for classification)

Open world/In the wild for different tasks (e.g., detection):
 succeed for all categories, whether seen or not seen during training

Closed-world/Closed-set: assumes all target classes for the target task are available at training



Training Domain

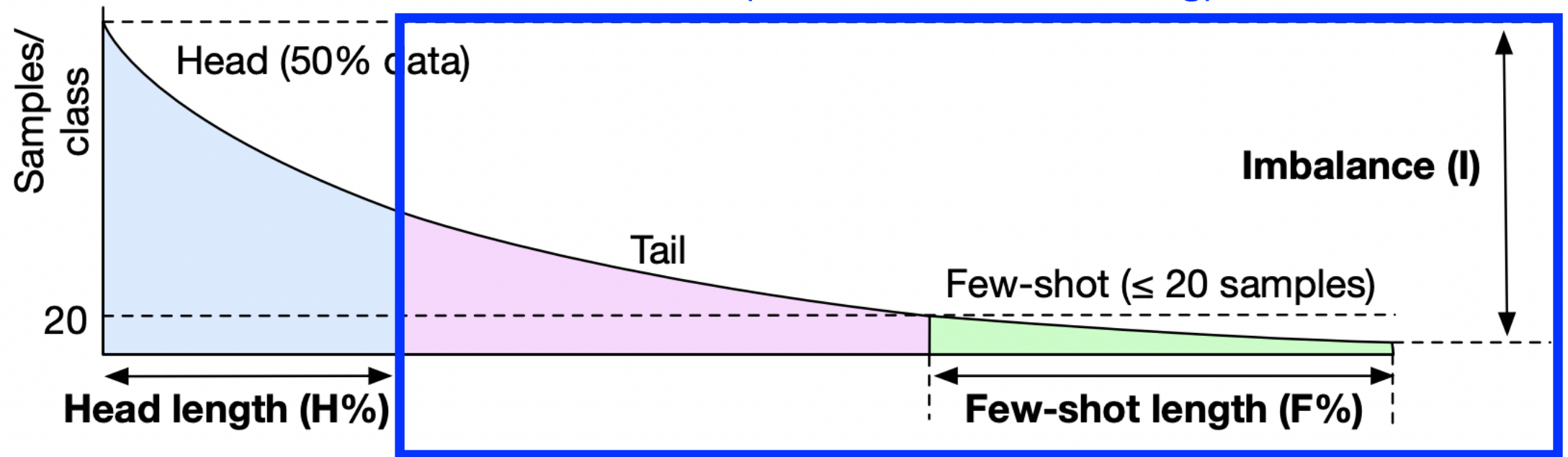
Out-of-domain/Robustness testing:
 same content observed differently

Visual Content

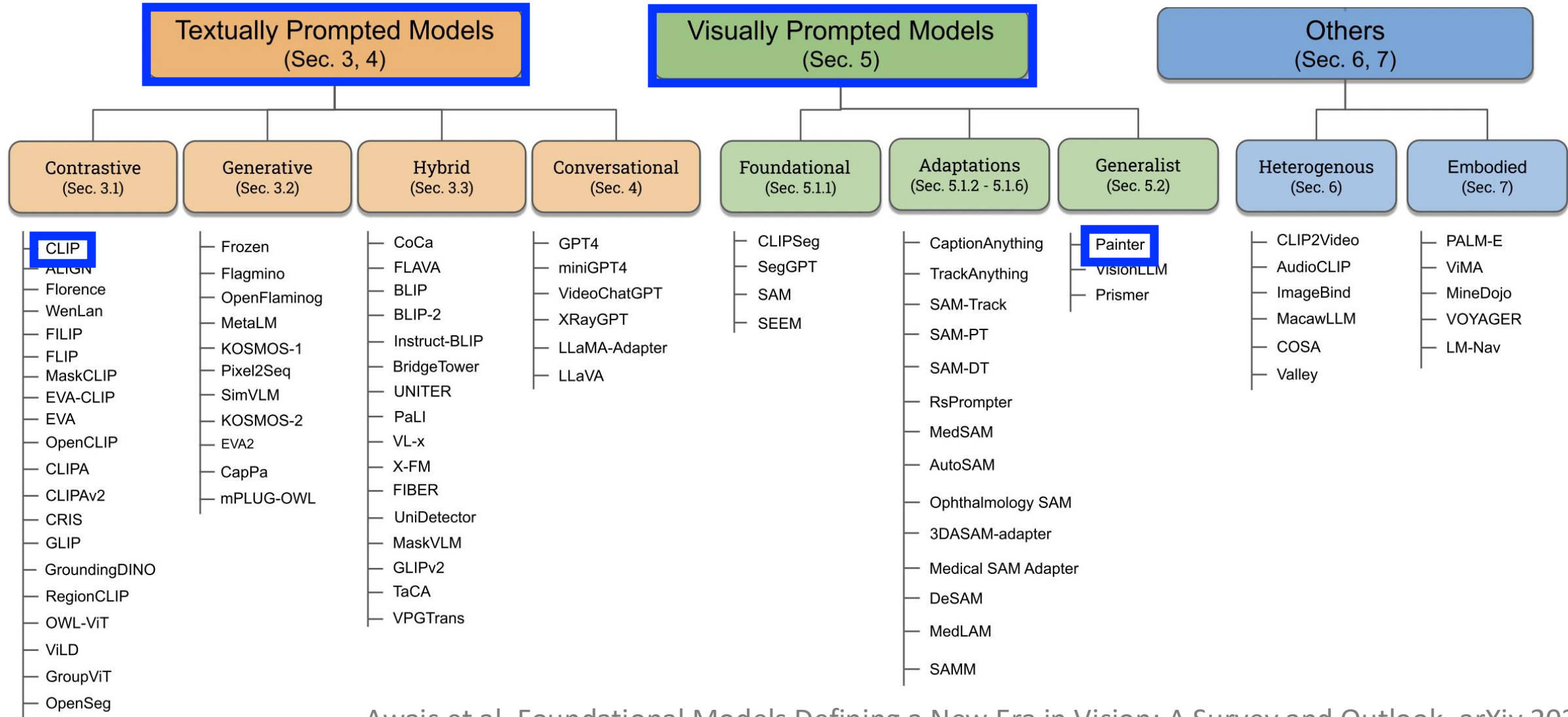
Open set classification/Out-of-distribution detection:
 predict whether a sample is drawn from the distribution observed at training time

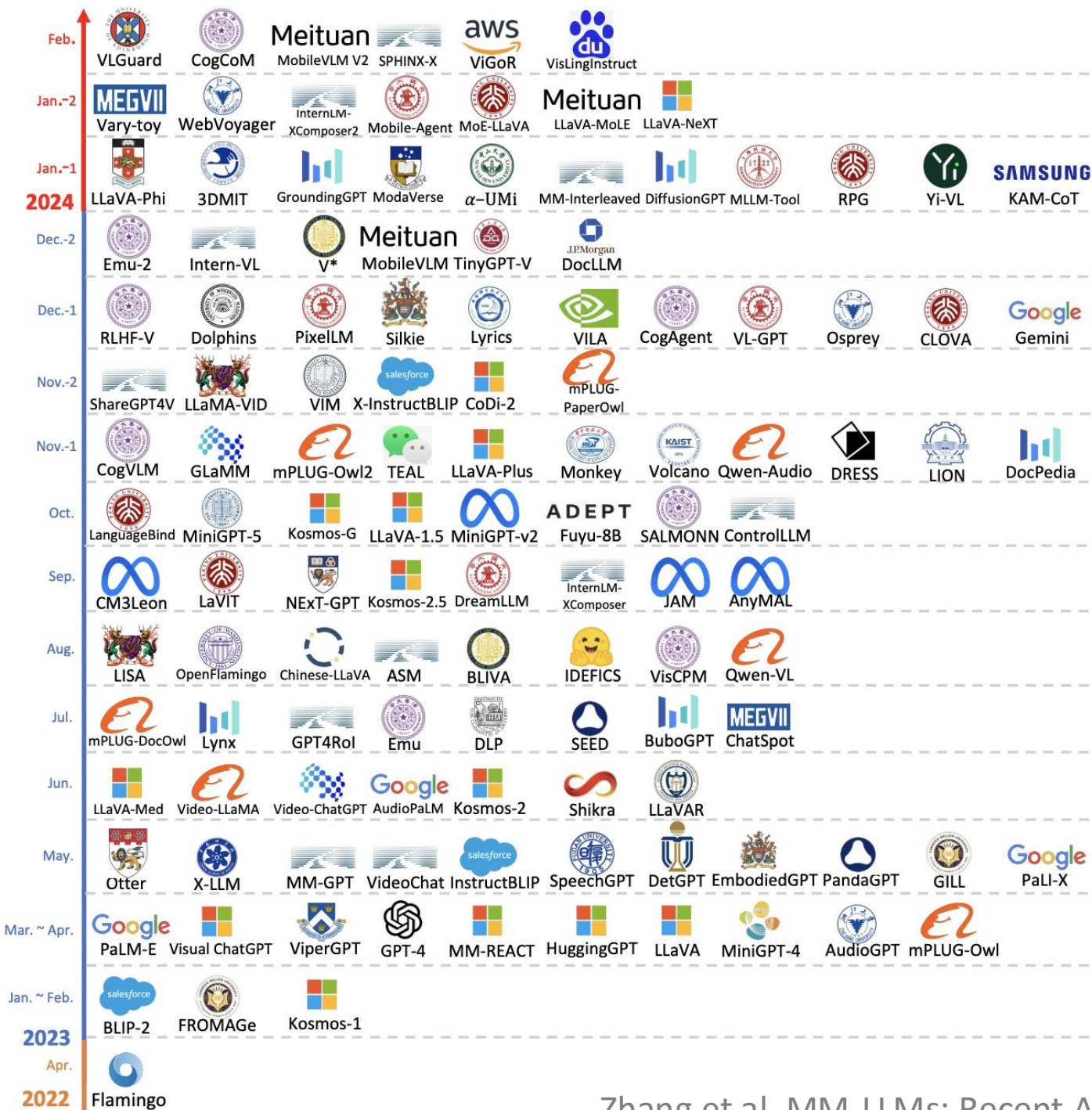
Beyond Large Amounts of Training Data

Learning with Limited/No Labeled Training Data
(Zero/Few-Shot Learning)



Prompting Visual Foundation Models





Model developers in academia & industry!

What Are Risks of Using Foundation Models?

- e.g.,
 - Any biases/limitations trickle to all downstream models
 - Current status quo is computationally expensive models (and so models that are bad for environment)

Today's Topics

- Foundation Models
- Textual Prompting & Zero-shot Learning
- Visual Prompting & In-context Few-shot Learning
- Prompt Tuning
- Discussion (chosen by YOU 😊)

Foundation Models: What's New?

Key ingredients identified:

1. Transformer model architecture
2. Lots more training data by using Internet data
3. Sufficient hardware with modern GPUs


Curating Image-Text Pairs from [Internet](#); e.g.,

1. Image-Text Pair Collection

- Source: Wikipedia, given its high quality (editorially reviewed), large size (~124M pages), and diversity (279 languages)

- Extracted ~150 million image-text pairs

For Each Image, Multiple Texts Extracted:

 **WIKIPEDIA**
The Free Encyclopedia

Berimbau

Article [Talk](#) Read [Edit](#) [View history](#) [Tools](#)

From Wikipedia, the free encyclopedia

The **berimbau** (Portuguese pronunciation: [berĩˈbaw], borrowed from Kimbundu *mbirimbau*^[1]) is a traditional Angolan [musical bow](#) that is commonly used in [Brazil](#).^[2]

It consists of a single-stringed bow attached to a gourd resonator and is played with a stick and a coin or stone to create different tones and rhythms.

The berimbau was used in many parts of Africa and Brazil during the 19th century to accompany chants and storytelling.^[3] It is part of the [candomblé](#) tradition, later incorporated into the [Afro-Brazilian](#) art [capoeira](#). Until the mid-20th century, it was used almost exclusively within the black community, but after the popularization of capoeira, it gain wider popularity.


Today, berimbau is used in various genres of popular music.

History [\[edit\]](#)

Berimbau is an adaptation of African gourde [musical bows](#), as no [Indigenous Brazilian](#) or [European](#) people use musical bows.^{[2][5]} According to the musicologist [Gerard Kubik](#), the *berimbau* and the "southwest Angolan variety called *mbulumbumba* are identical in construction and playing technique, as well as in tuning and in a number of basic patterns played."^[6] The assimilation of this Angolan instrument is evident also in other [Bantu](#) terms used for musical bow in [Brazilian Portuguese](#), including *urucungo*, and *madimba lungungu*.

In 1859, French journalist [Charles Ribeyrolles](#) described free practices of [African slaves](#) on

Berimbau



Angola musical bow (1922), known as berimbau in Brazil.

 **WIKIMEDIA COMMONS**

File:History of Inventions USNM 41 Angola M

From Wikimedia Commons, the free media repository

File [File history](#) [File u](#)



[Download](#)
all sizes

[Use this file](#)
on the web

[Use this file](#)
on a wiki

[Email a link](#)
to this file

[Information](#)
about reusing

ANGOLA MUSICAL BOW.

(1) Wikipedia description with (2) associated alt-text and (3) attribution on Wikimedia page

Curating Image-Text Pairs from **Internet**; e.g.,

1. Image-Text Pair Collection

- Source: Wikipedia, given its high quality (editorially reviewed), large size (~124M pages), and diversity (279 languages)
- Extracted ~150 million image-text pairs

2. Filtering

- Removed images with “generic” or meaningless text (e.g., maps), unsuitable licenses, questionable content (e.g., pornography, violence), and width or height < 100 pixels
- Only kept example in top 100 languages

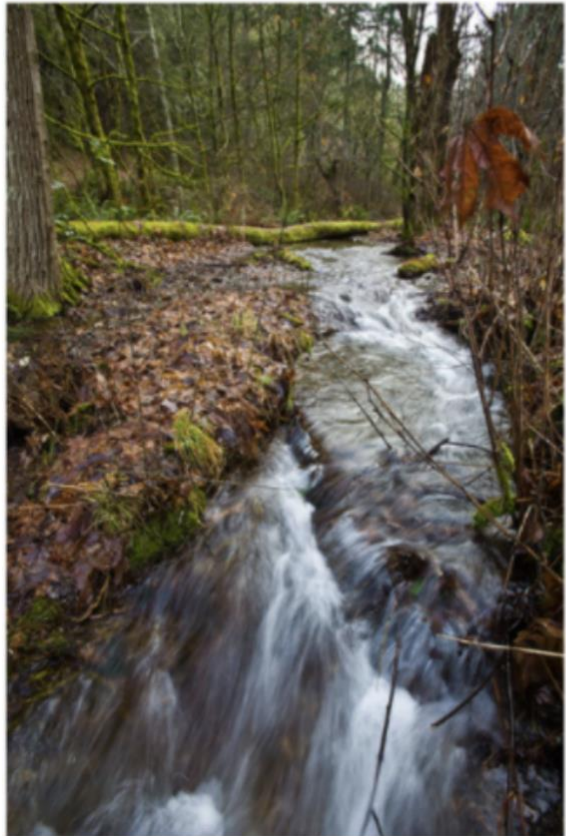
3. Human Quality Validation

- Crowdsourced ratings for nearly 4,400 examples
- Majority vote label used from 3 independent ratings
- Examples were in English (~3,000), German (300), French (300), Spanish (300), Russian (300), Chinese (300), & Hindi (100)

Task: Given an image, descriptions and a title, answer the given questions

More instructions on how to complete the task are available in this [guidelines doc](#)

Title: Sequelitchew Creek



Text Description 1

Sequalitchew Creek, lower canyon

Does Text 1 describe the above image well?

Yes Maybe No

Text Description 2

Sequalitchew Creek, lower canyon

Does Text 2 describe the above image well?

Yes Maybe No

Combined Text Description

Text1: Sequelitchew Creek, lower canyon
Text2: Sequelitchew Creek, lower canyon
Extra: Sequelitchew-Creek-lower-canyon.jpg Sequelitchew Creek, located in Fort Lewis, Washington, was the location of the original Fort Nisqually trading

Does Text1 + Text2 + Extra descriptions combined as a whole describe the above image well?

Yes Maybe No

Submit

- Results from first two questions suggested both reference and attribution texts are high-quality
- No major difference found across different languages

Curating Image-Text Pairs from [Internet](#); e.g.,

Dataset	Images	Text	Languages
Flickr30K [39]	32K	158K	< 8
SBU Captions [24]	~1M	~1M	1
MS-COCO [21]	~330K	~1.5M	< 4
CC [5]	~3.3M	~3.3M	1
WIT	11.5M	37.6M	108

WIT has 37.6 million (image, text) pairs describing 11.5 million unique images spanning 108 languages (each with 12K+ examples)

Foundation Model: CLIP

Key ingredients:

1. Transformer model architecture
2. Lots more training data by using Internet data
3. Sufficient hardware with modern GPUs

Why CLIP?

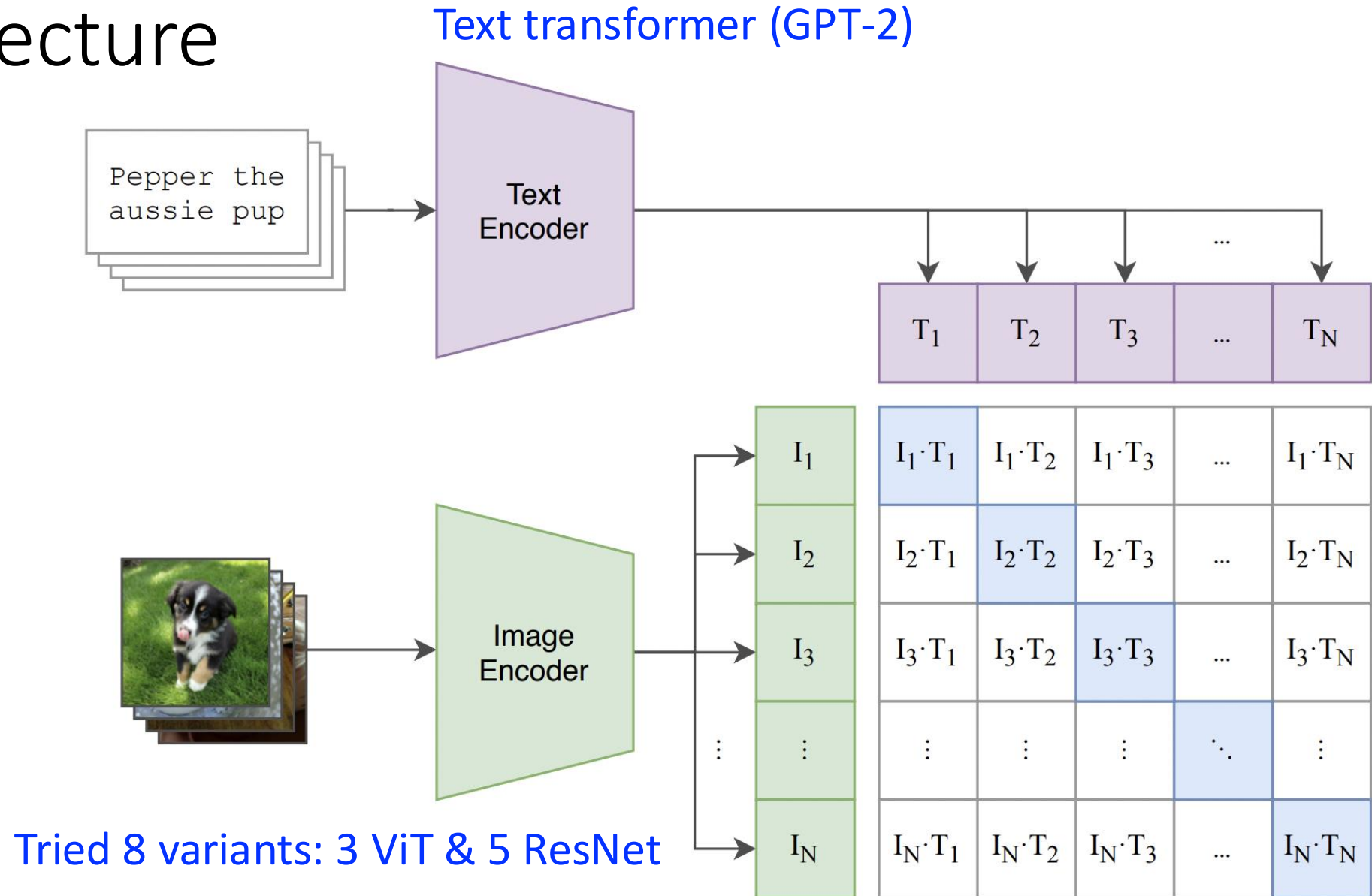
Named after the proposed technique: **C**ontrastive **L**anguage **I**mage **P**re-training

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

CLIP Model: Novelty

- Train image analysis models with natural language supervision using the **vast amounts of publicly available data on the Internet**

CLIP Architecture



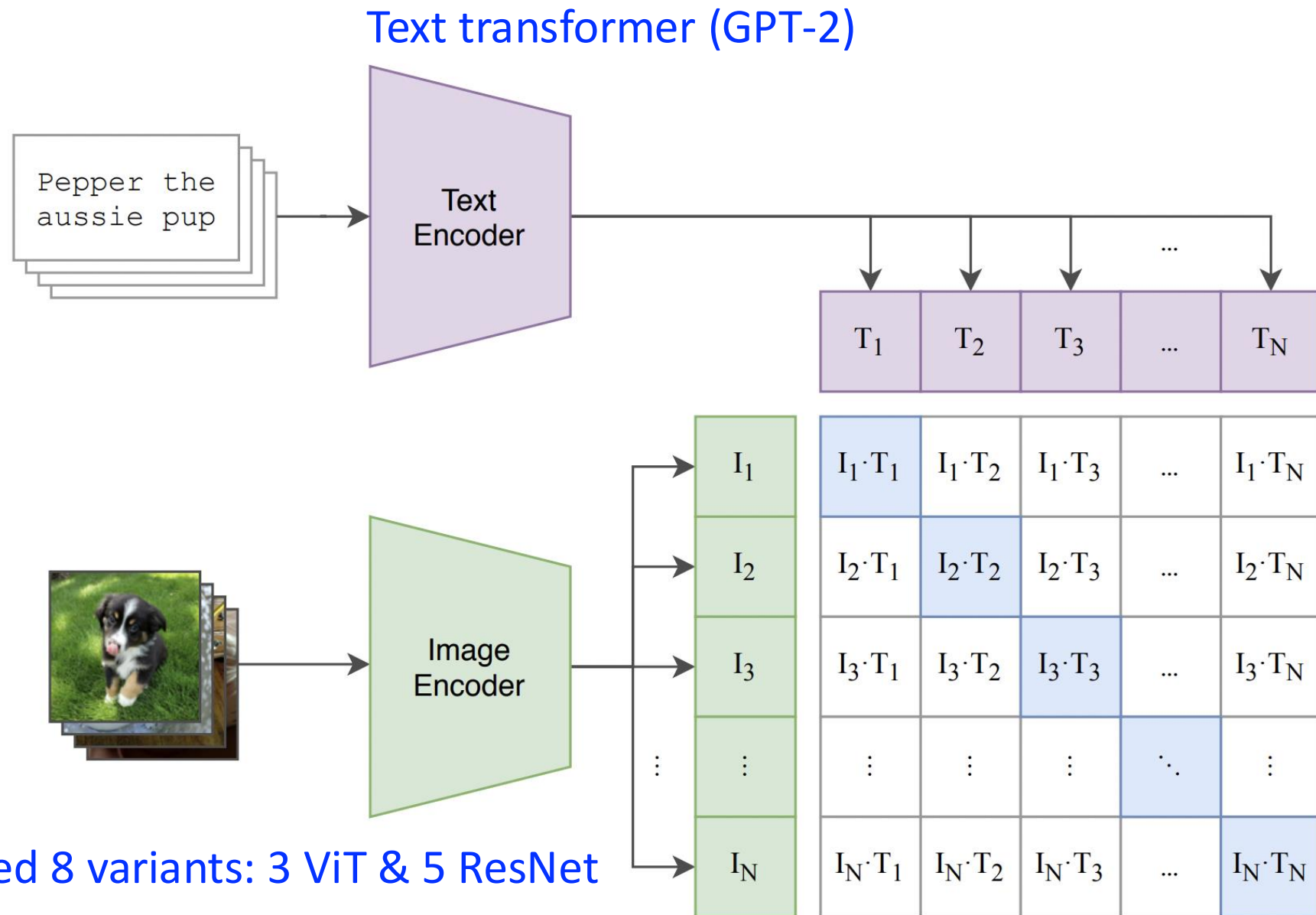
CLIP Training

Task: predict which image-text pairs match using 400 million image-text pairs from Internet containing any of 500,000 queries (e.g., words occurring 100+ times in English version of Wikipedia and all WordNet synonyms)

- Largest ResNet model took 18 days to train on 592 V100 GPUs and largest ViT took 12 days on 256 V100 GPUs

- Experiments run with largest (“best”) ViT model

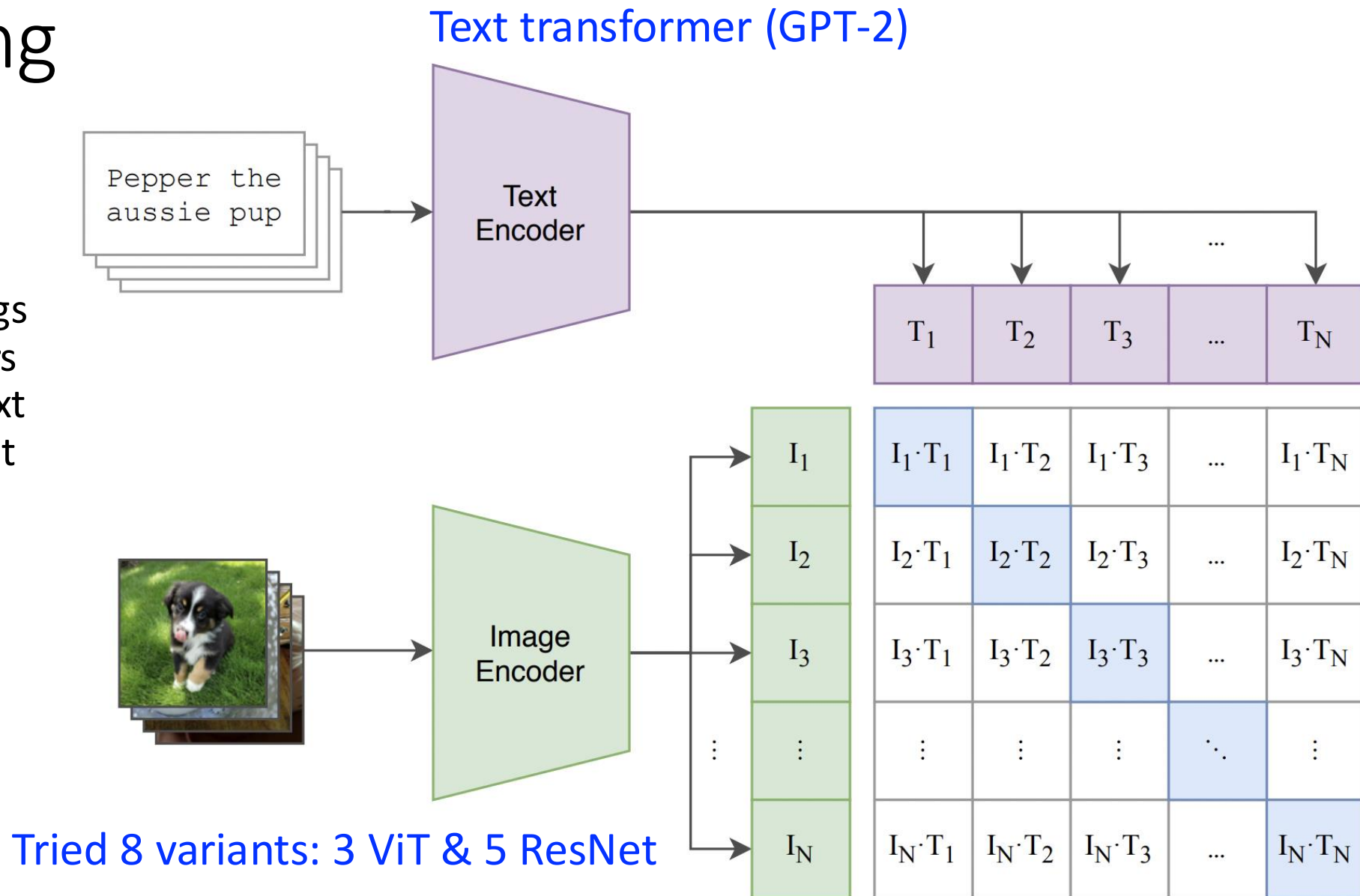
Tried 8 variants: 3 ViT & 5 ResNet



CLIP Training

- Learns feature embeddings for image and text encoders that push correct image-text pairs together and incorrect image-text pairs apart.

- Learns nouns, verbs, adjectives, and more!

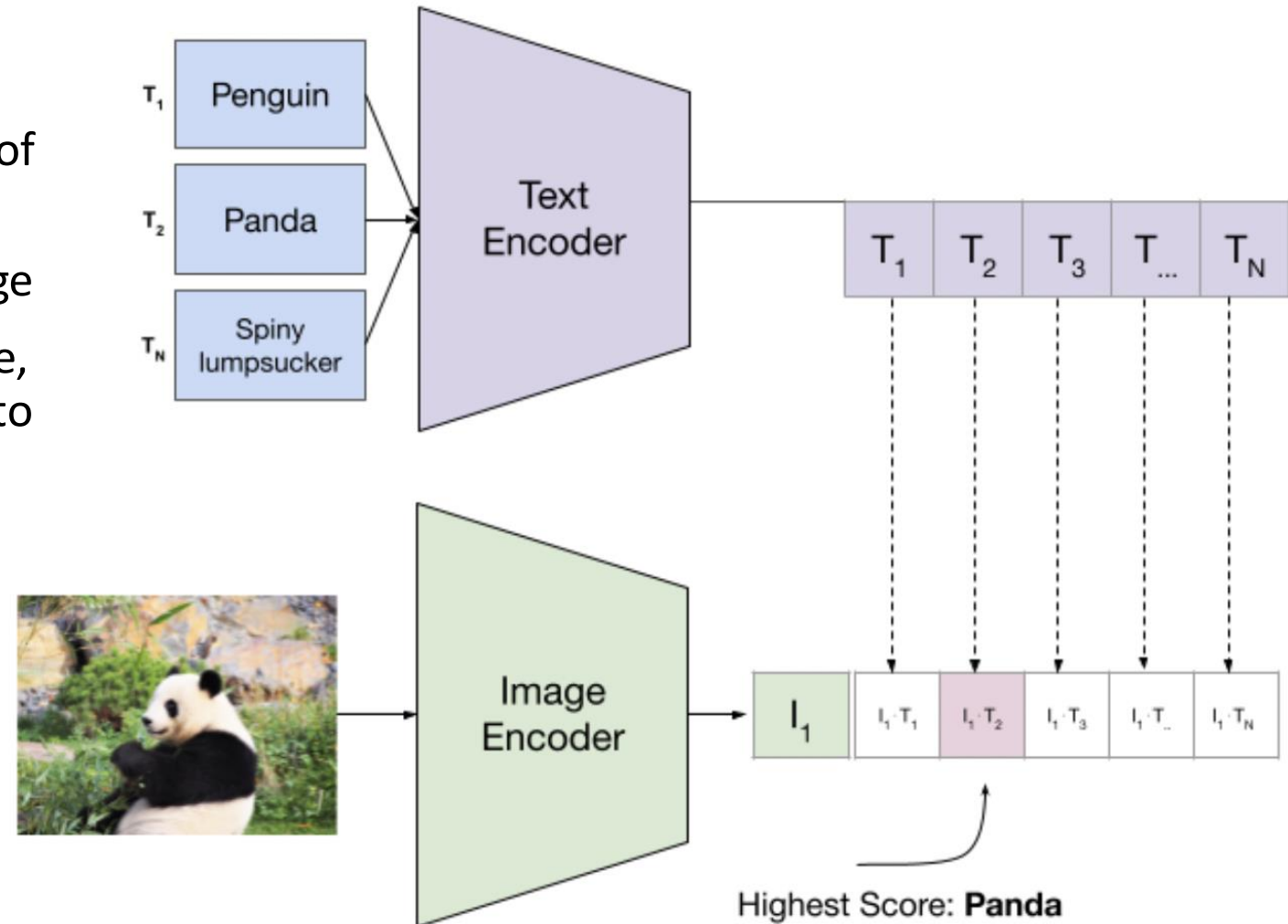


Zero-Shot Performance
Evaluated on Over 30 Datasets

CLIP Inference

e.g., zero-shot classification:

1. Compute feature embedding for names of all classes in the dataset by its encoder
2. Compute feature embedding of the image
3. Compute cosine similarity of each (image, text) pair embedding followed by softmax to identify most probable match



CLIP Inference

Prompts “engineered” to mimic that training data often had sentences (instead of words):

- classification: “A photo of a {label}”
- fine-grained classification: “A photo of a {label}, a type of pet/food/aircraft/etc”
- satellite image classification: “A satellite photo of a {label}”
- ensembles: “A photo of a big/small/etc {label}”

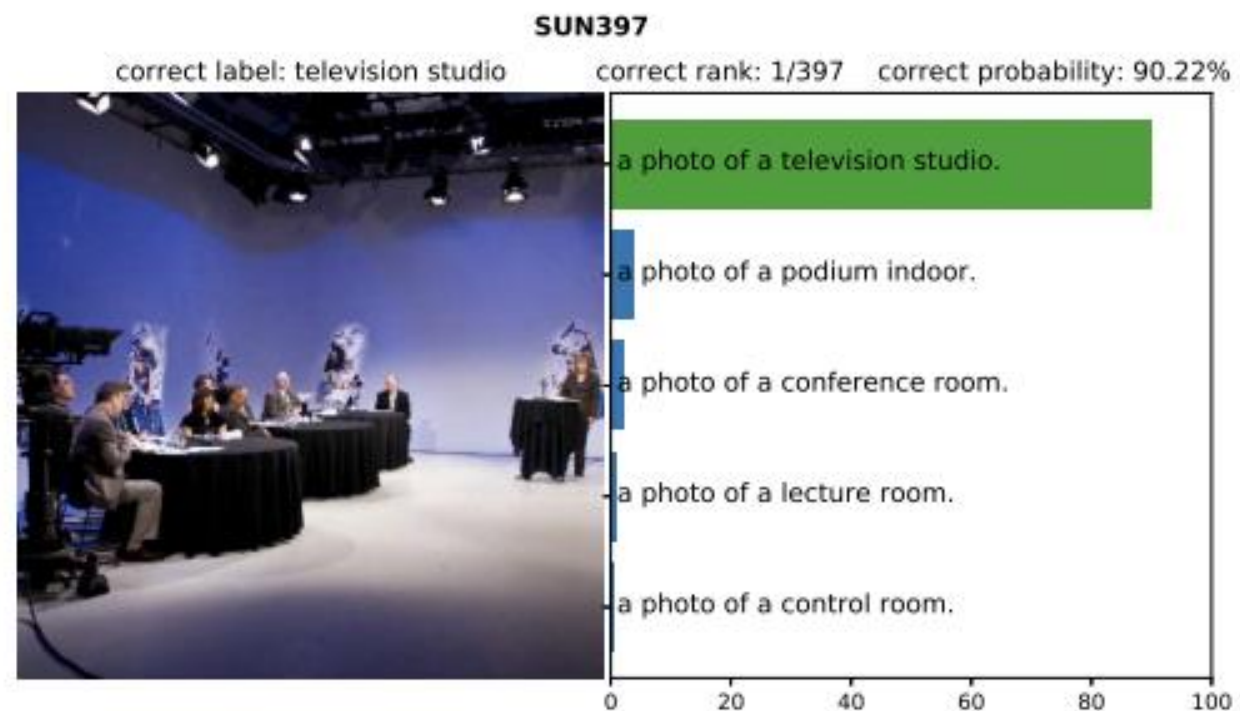
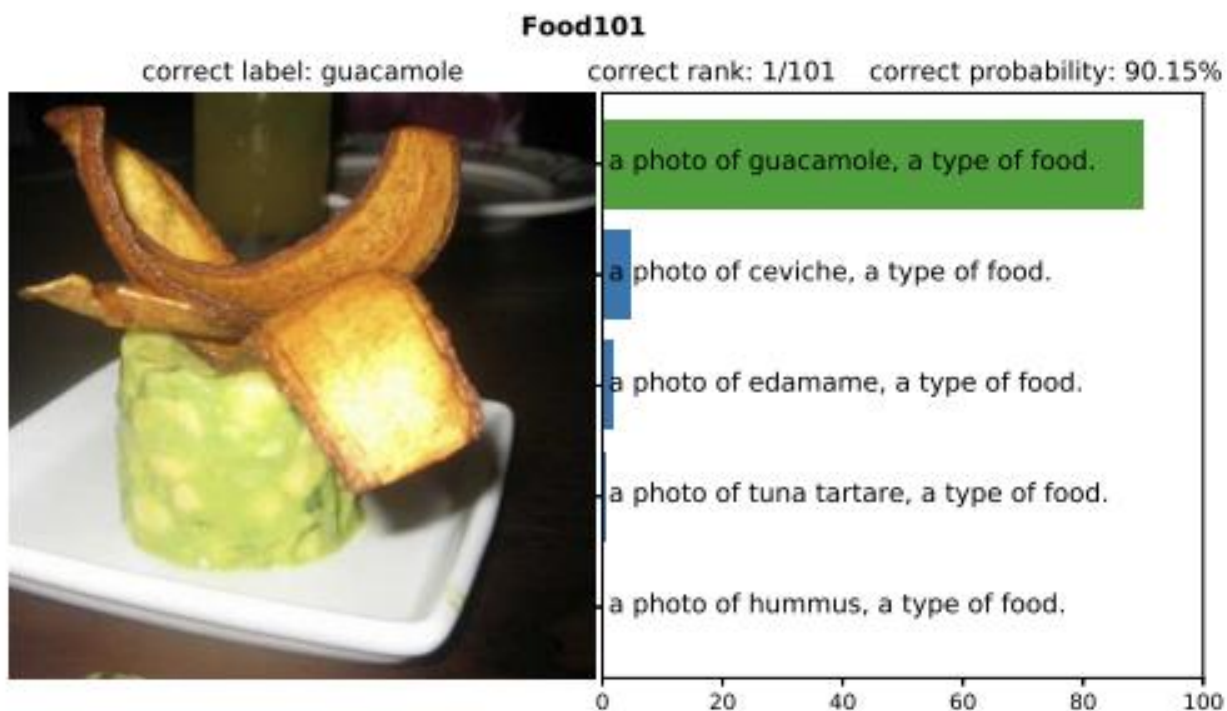
CLIP Evaluation

Subset of datasets shown here:

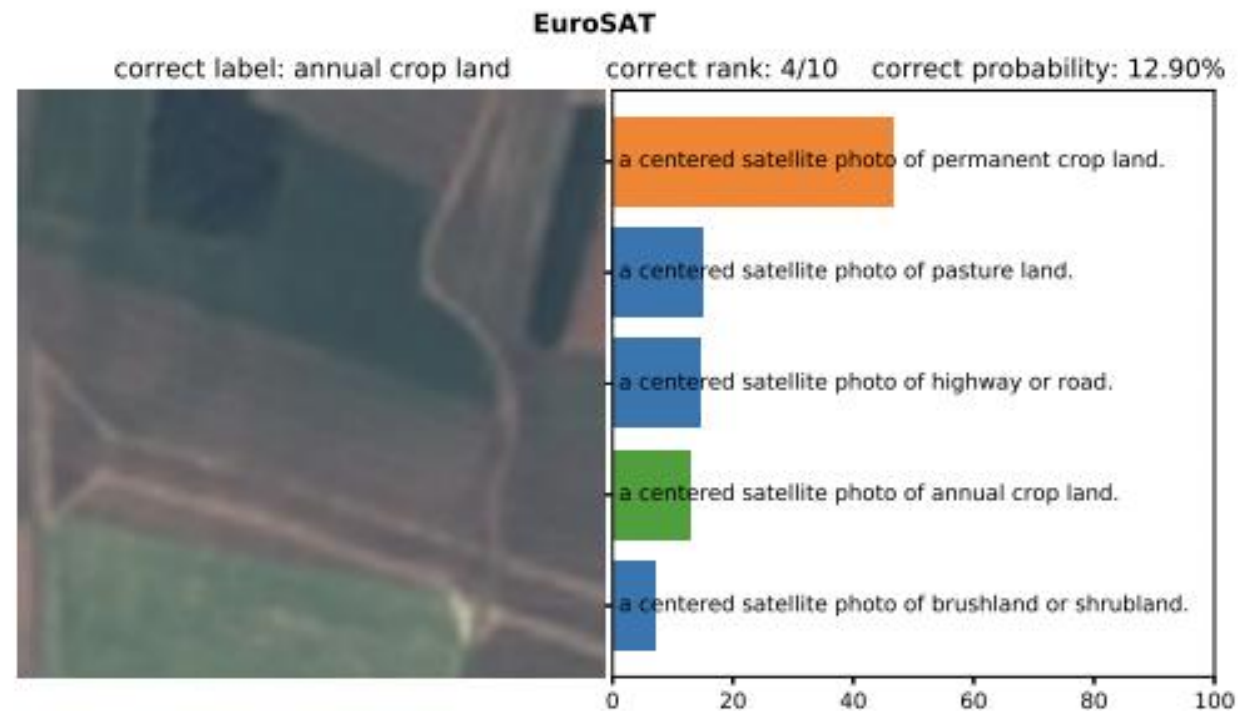
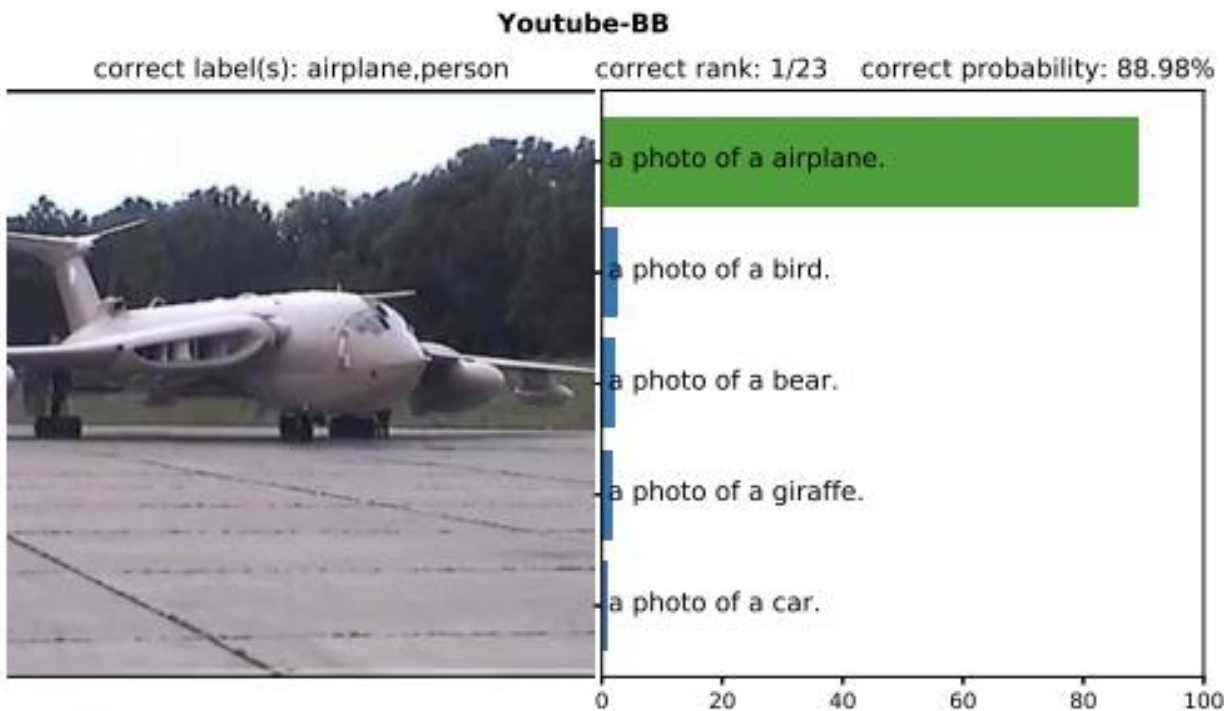
Classification evaluation spanned fine-grained classification (e.g., food, bird, aircraft, and car categories), distribution shifts for ImageNet categories (e.g., corrupted images), and more

Dataset	Classes	Train size	Test size	Evaluation metric
Food-101	102	75,750	25,250	accuracy
CIFAR-10	10	50,000	10,000	accuracy
CIFAR-100	100	50,000	10,000	accuracy
Birdsnap	500	42,283	2,149	accuracy
SUN397	397	19,850	19,850	accuracy
Stanford Cars	196	8,144	8,041	accuracy
FGVC Aircraft	100	6,667	3,333	mean per class
Pascal VOC 2007 Classification	20	5,011	4,952	11-point mAP
Describable Textures	47	3,760	1,880	accuracy
Oxford-IIIT Pets	37	3,680	3,669	mean per class
Caltech-101	102	3,060	6,085	mean-per-class
Oxford Flowers 102	102	2,040	6,149	mean per class
MNIST	10	60,000	10,000	accuracy
Facial Emotion Recognition 2013	8	32,140	3,574	accuracy
STL-10	10	1000	8000	accuracy
EuroSAT	10	10,000	5,000	accuracy
RESISC45	45	3,150	25,200	accuracy
GTSRB	43	26,640	12,630	accuracy
KITTI	4	6,770	711	accuracy
Country211	211	43,200	21,100	accuracy
PatchCamelyon	2	294,912	32,768	accuracy
UCF101	101	9,537	1,794	accuracy
Kinetics700	700	494,801	31,669	mean(top1, top5)
CLEVR Counts	8	2,000	500	accuracy
Hateful Memes	2	8,500	500	ROC AUC
Rendered SST2	2	7,792	1,821	accuracy
ImageNet	1000	1,281,167	50,000	accuracy

CLIP: Qualitative Results



CLIP: Qualitative Results

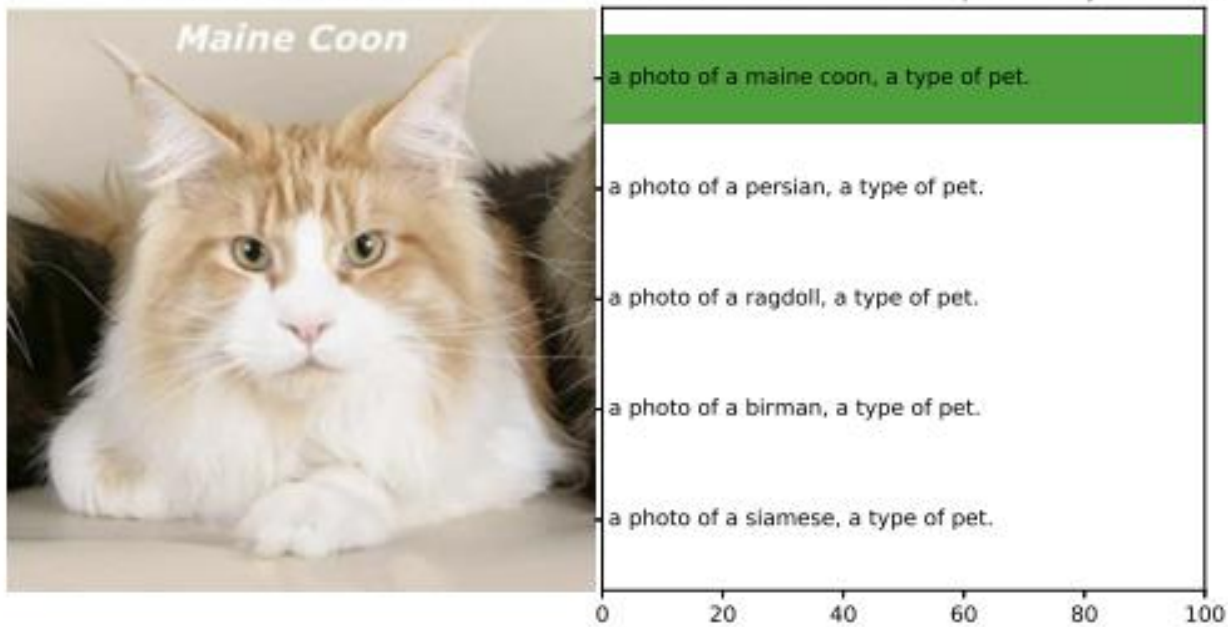


CLIP: Qualitative Results

Oxford-IIIT Pets

correct label: Maine Coon

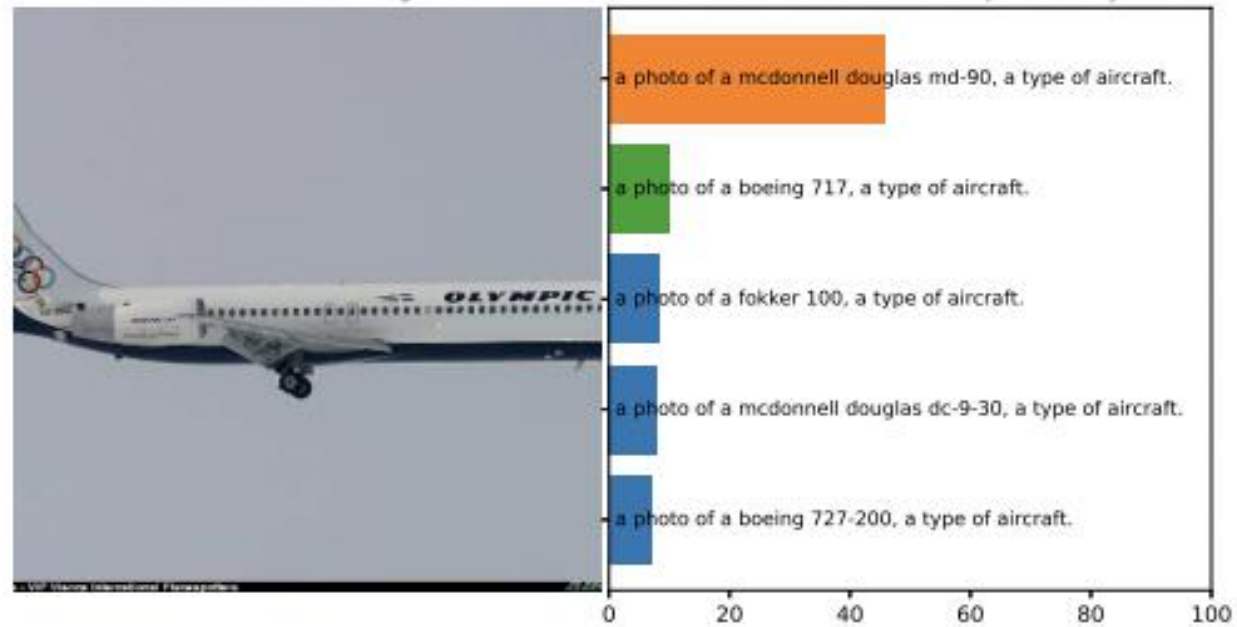
correct rank: 1/37 correct probability: 99.99%



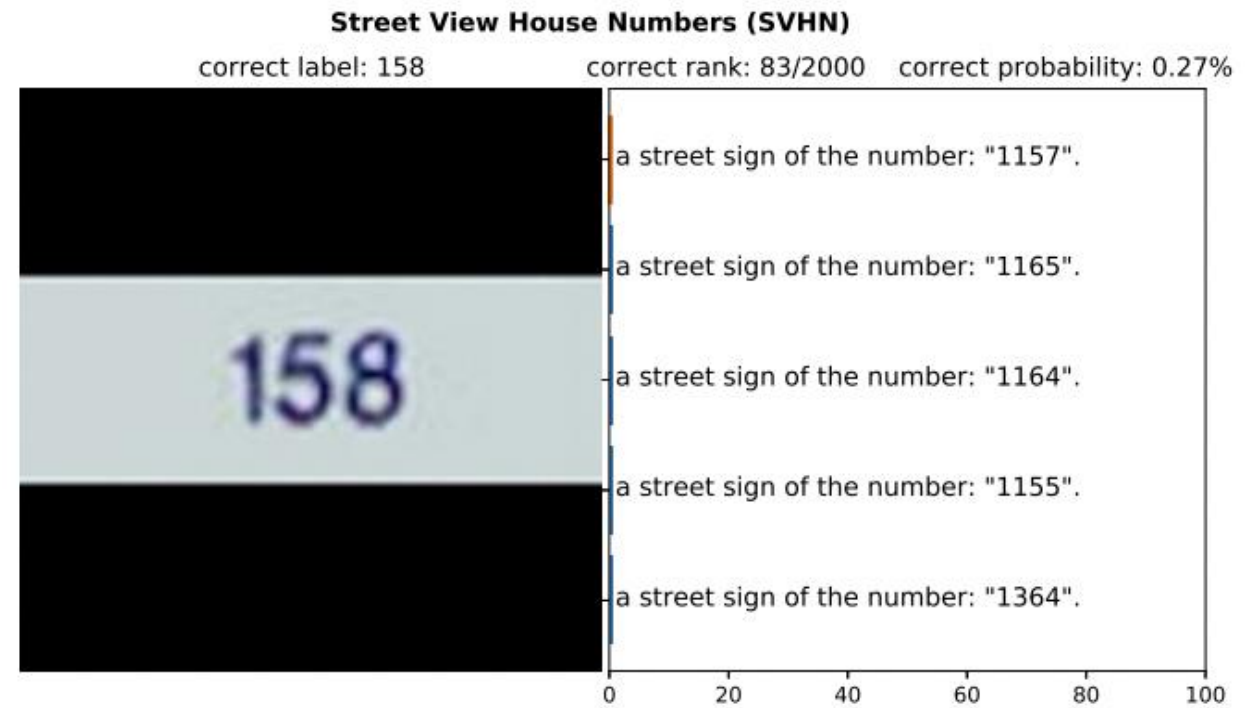
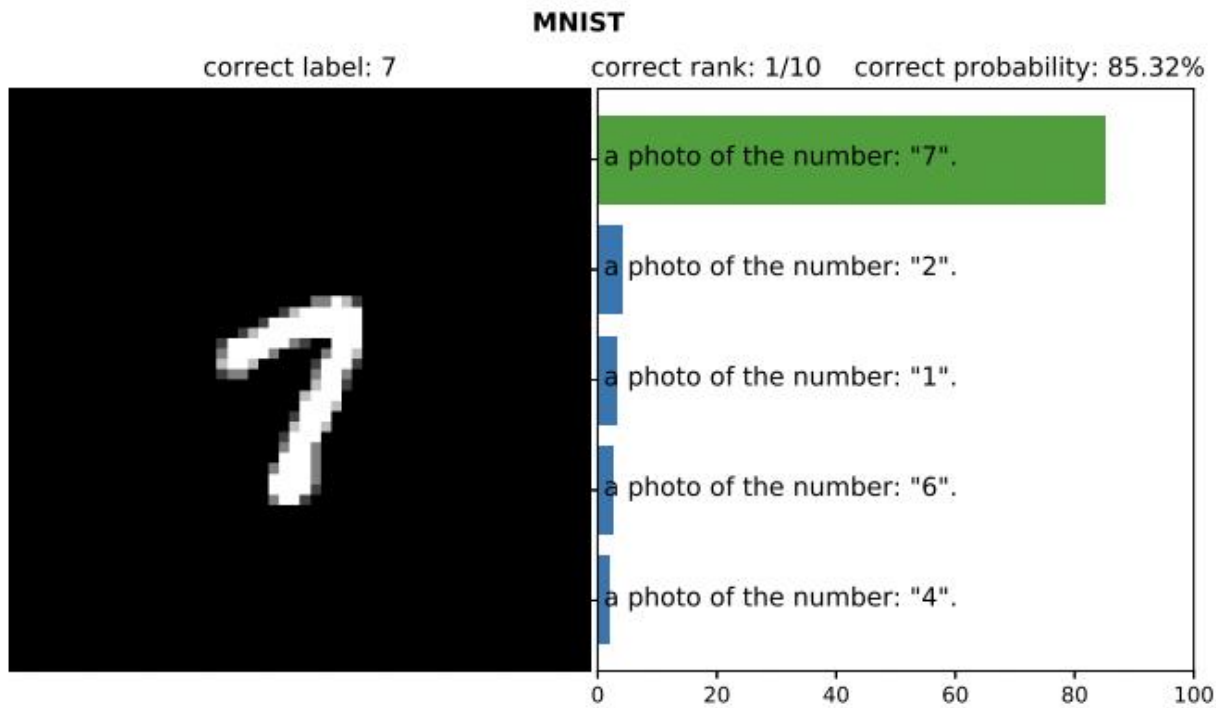
FGVC Aircraft

correct label: Boeing 717

correct rank: 2/100 correct probability: 9.91%



CLIP: Qualitative Results



CLIP: Qualitative Results

Hateful Memes

correct label: meme

correct rank: 1/2 correct probability: 99.20%

coffee isn't helping



get the jumper cables

a meme.

a hatespeech meme.

0 20 40 60 80 100

German Traffic Sign Recognition Benchmark (GTSRB)

correct label: red and white triangle with exclamation mark warning

correct rank: 1/43 correct probability: 45.75%



a zoomed in photo of a "red and white triangle with exclamation mark warning" traffic sign.

a zoomed in photo of a "red and white triangle with black right curve approaching warning" traffic sign.

a zoomed in photo of a "red and white triangle car skidding / slipping warning" traffic sign.

a zoomed in photo of a "red and white triangle rough / bumpy road warning" traffic sign.

a zoomed in photo of a "red and white triangle with black left curve approaching warning" traffic sign.

0 20 40 60 80 100

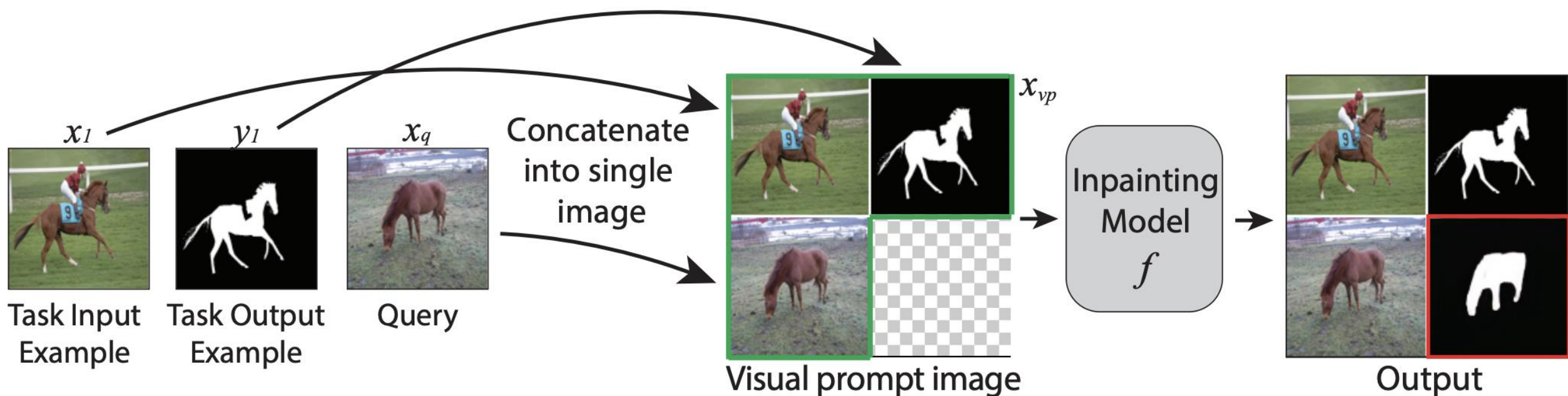
Today's Topics

- Foundation Models
- Textual Prompting & Zero-shot Learning
- **Visual Prompting & In-context Few-shot Learning**
- Prompt Tuning
- Discussion (chosen by YOU 😊)

Motivation

- Goal: Define general-purpose prompts based on **images rather than text**
- Observation: foundation models achieved better performance for NLP tasks when provided “in-context” examples
 - i.e., [Task description, Examples, Prompt]
 - e.g., “Translate English to Spanish. Computer -> Computadora. Vision ->
- Idea: Use in-context few-shot learning for image-based prompts

Novel Idea: Image Inpainting



Designed to adapt to any “image-to-image translation” task by using the model as is (e.g., no fine-tuning required)

Idea

Image inpainting for prompting introduced in 2022 by Bar et al.



Edge detection



Colorization



Inpainting



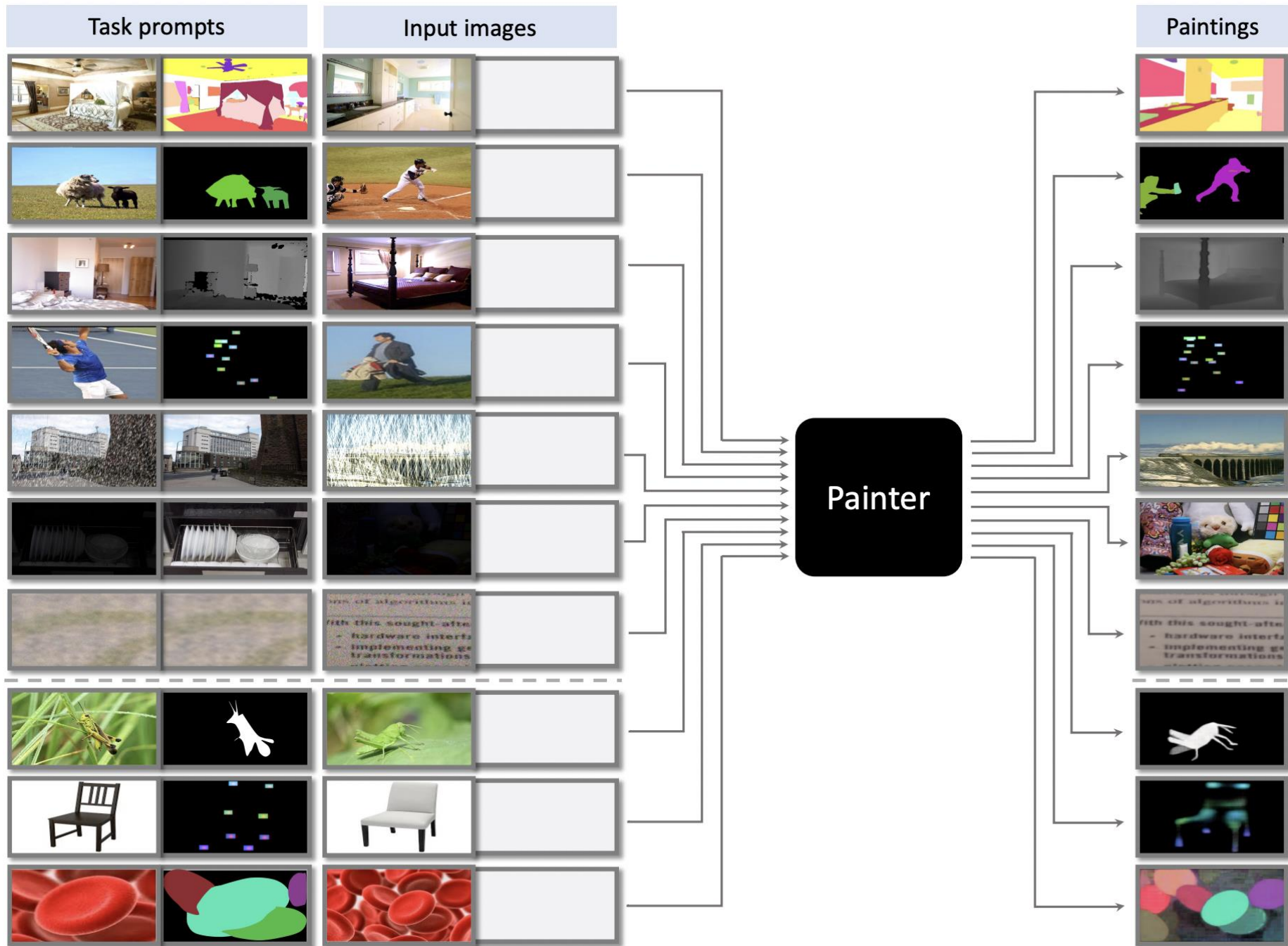
Segmentation



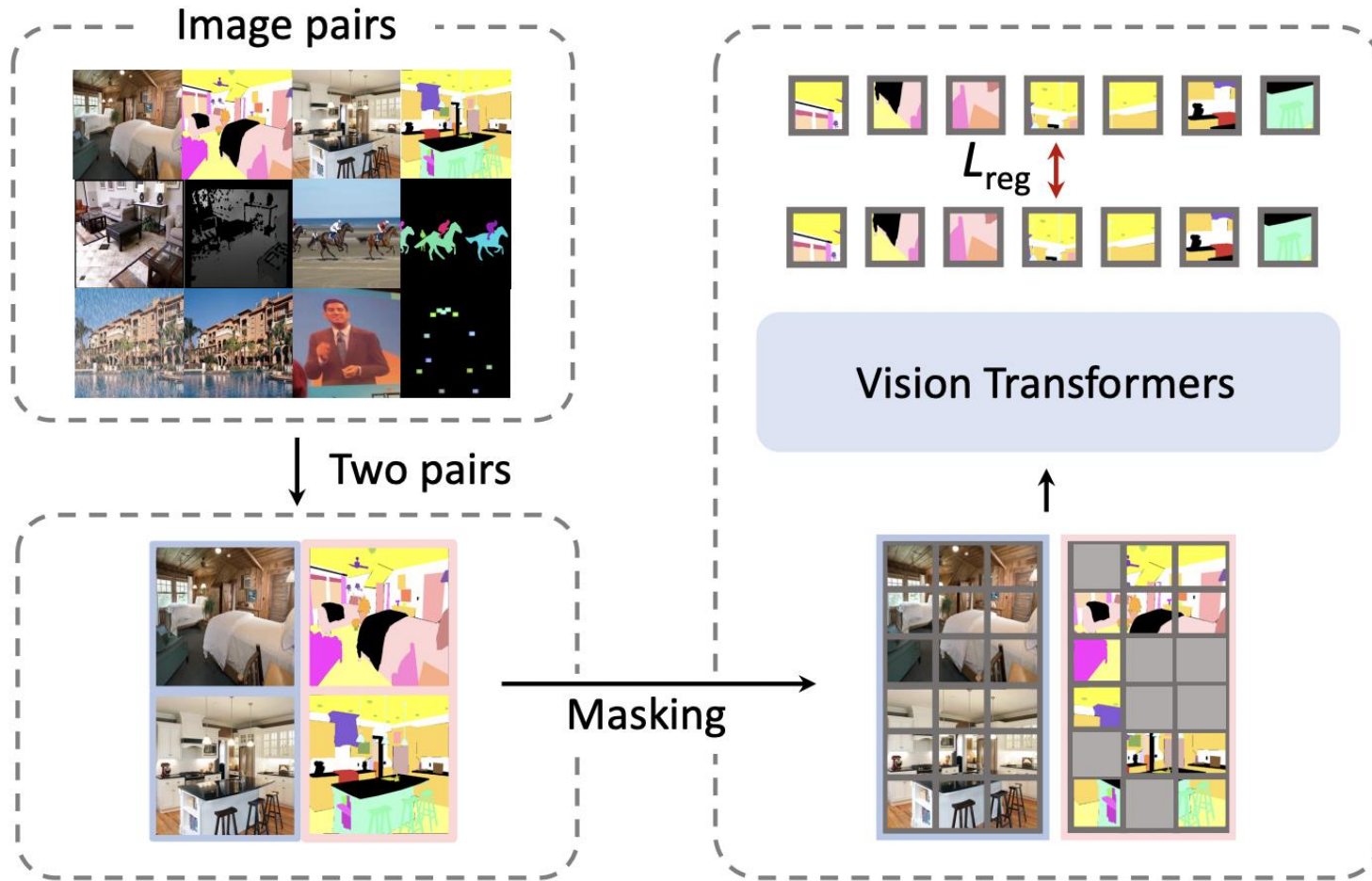
Style transfer

Idea

Idea extended in 2023 by Wang et al. on standard vision benchmark datasets



Training: Masked Image Modeling



Uses self-supervised learning such that the model predict values in masked out patches

Uses standard vision benchmarks for each evaluated task

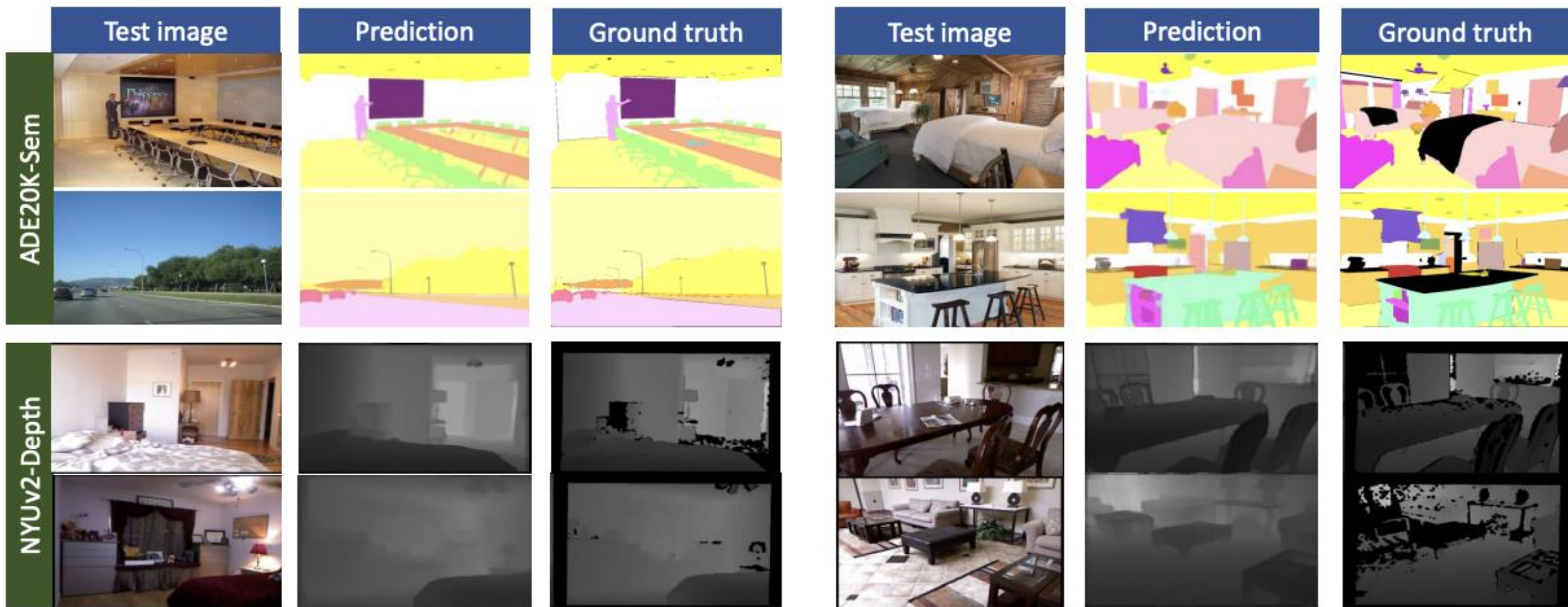
□ An image □ A GT image □ A patch ■ A masked patch

Experimental Results

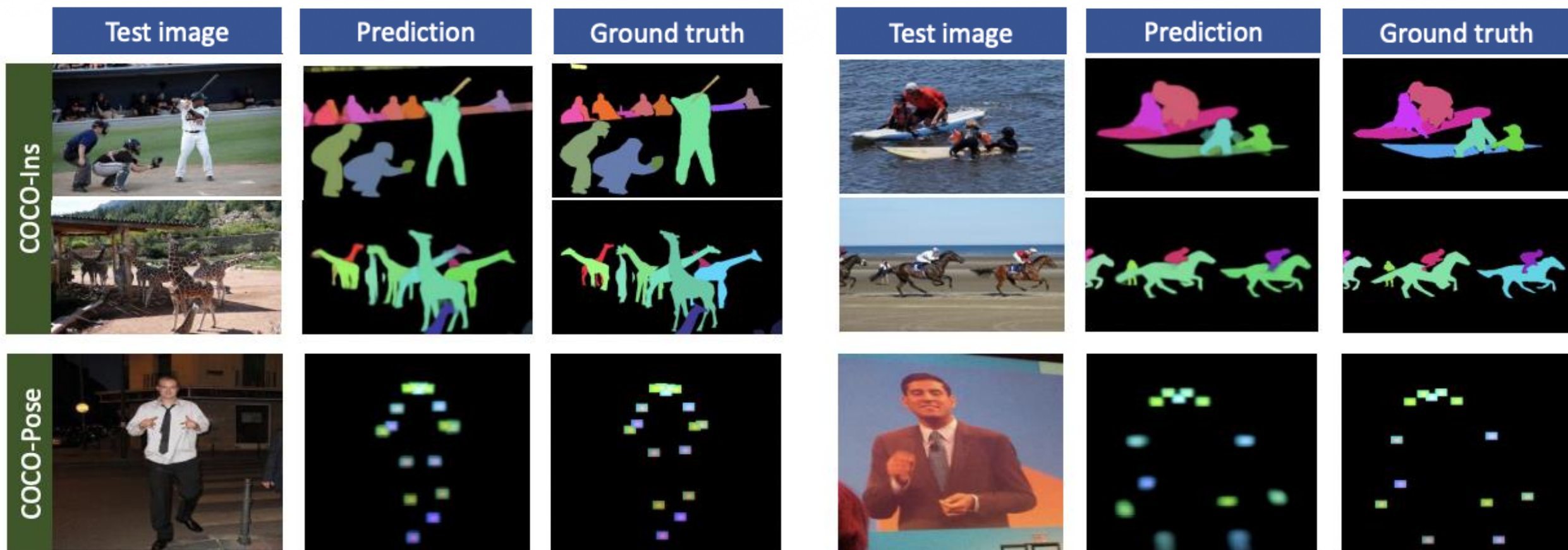
(Used for prompt the best performing example pair per task from all examples in the training dataset)

Model achieves state-of-the-art performance on depth estimation for NYUv2 dataset and outperforms other generalist models on several more tasks.

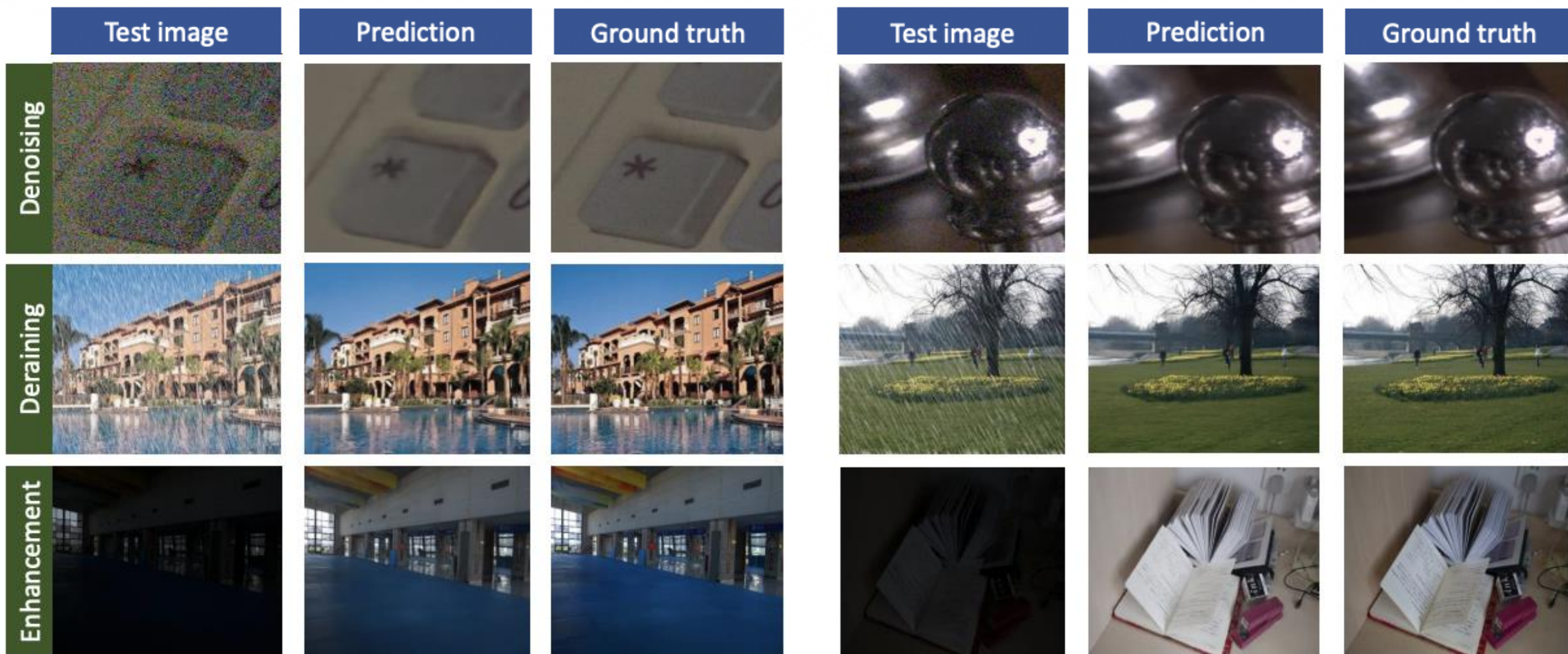
Qualitative Results: In-Domain Results



Qualitative Results: In-Domain Results

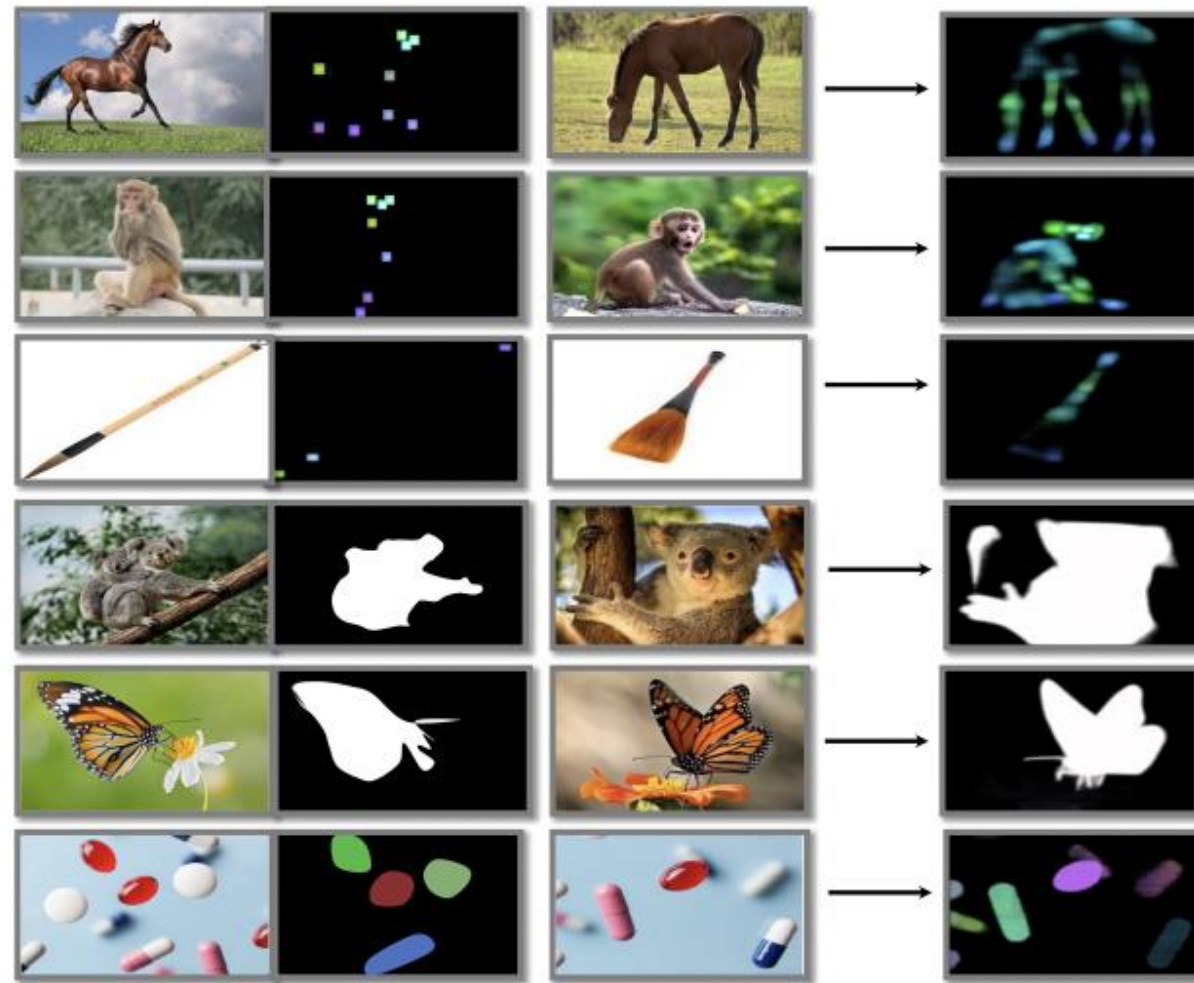


Qualitative Results: In-Domain Results



Qualitative Results: Open-Vocabulary Results (i.e., Categories Not Seen at Training)

In-context examples, prompts, and predictions for keypoint detection, object segmentation, and instance segmentation



When Might One Choose A Visual Prompt Versus a Textual Prompt?

- e.g.,
 - Greater equity for different languages as non-English languages often are poorly supported if at all
 - Empowering people appropriately based on their (dis)abilities: e.g., blind and deaf users

Today's Topics

- Foundation Models
- Textual Prompting & Zero-shot Learning
- Visual Prompting & In-context Few-shot Learning
- **Prompt Tuning**
- Discussion (chosen by YOU 😊)

Motivation

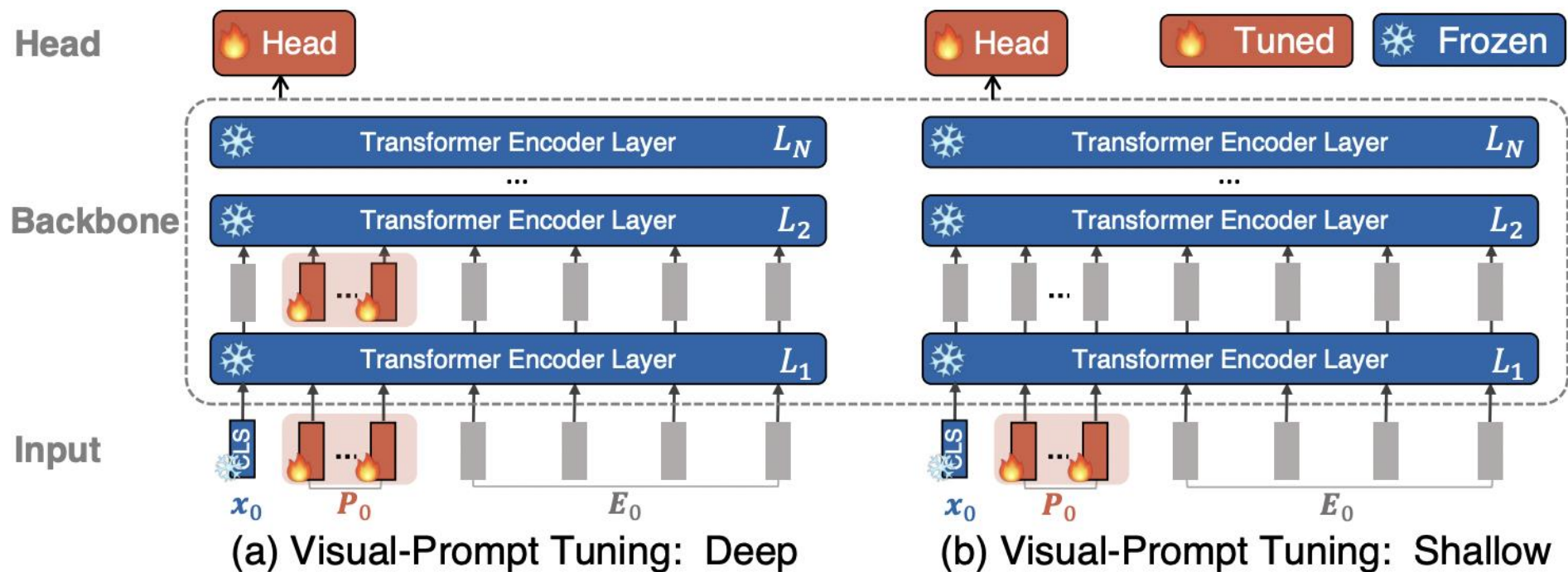
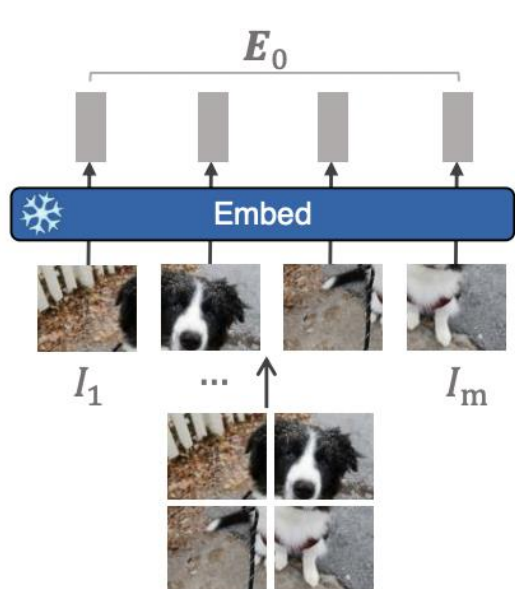
The image shows a screenshot of the PromptAttack marketplace website. The main header features the PromptAttack logo, a search bar with the text "Search Prompts, @authors or #tags", and navigation links for "Marketplace", "Login", and "Register". The main content area has a dark blue background with the text "Unleash the power of Artificial Intelligence" and "Prompt Attack Your #1 Prompt Marketplace". Below this, there is a description: "PromptAttack is a marketplace where you can purchase and sell high-quality prompts that generate optimal stunning results while also reducing your API expenses." At the bottom of the main content area are two buttons: "Find a prompt" and "Sign Up".

Three example prompts are displayed as cards:

- Jagged Cut Out Punk Posters** by @midrun, priced at \$2.99. The image shows a collage of punk-style faces and a train.
- Research Paper Summarizer** by @laxman1986, priced at \$2.99. The image shows a document icon.
- Studio Quality Product Fruit...** by @midrun, priced at \$2.99. The image shows a variety of fresh fruits like grapes, raspberries, and limes.

Manually engineering prompts is challenging to do well (leading to MANY prompt marketplaces)

Idea: Replace Manually-Authored Prompts with Learnable Parameters



Learned prompts adapt frozen model (e.g., no fine-tuning required) to different target tasks

What Are Benefits of Visual Prompt Tuning?

- Typically, little training data is needed because only a limited amount of parameters need to be trained
- Few task-specific parameters need to be learned and stored to support a new task, compared to model fine-tuning
- Prevents overfitting generalizable knowledge and overfitting to the task
- Provides a static knowledge-base

Today's Topics

- Foundation Models
- Textual Prompting & Zero-shot Learning
- Visual Prompting & In-context Few-shot Learning
- Prompt Tuning
- Discussion (chosen by YOU 😊)

Today's Topics

- Foundation Models
- Textual Prompting & Zero-shot Learning
- Visual Prompting & In-context Few-shot Learning
- Prompt Tuning
- Discussion (chosen by YOU 😊)



The End