Visual Foundation Models and Prompts

Danna Gurari
University of Colorado Boulder
Fall 2023

https://home.cs.colorado.edu/~DrG/Courses/RecentAdvancesInComputerVision/AboutCourse.html

Review

- Last lecture:
  - Motivation
  - ViT
  - Swin Transformer
  - Discussion

- Assignments (Canvas):
  - Reading assignment and project proposal due earlier today
  - Reading assignments due next Monday and Wednesday (for student-led lectures)
  - Project outline due after Fall break (overview of expectations on website)

- Questions?

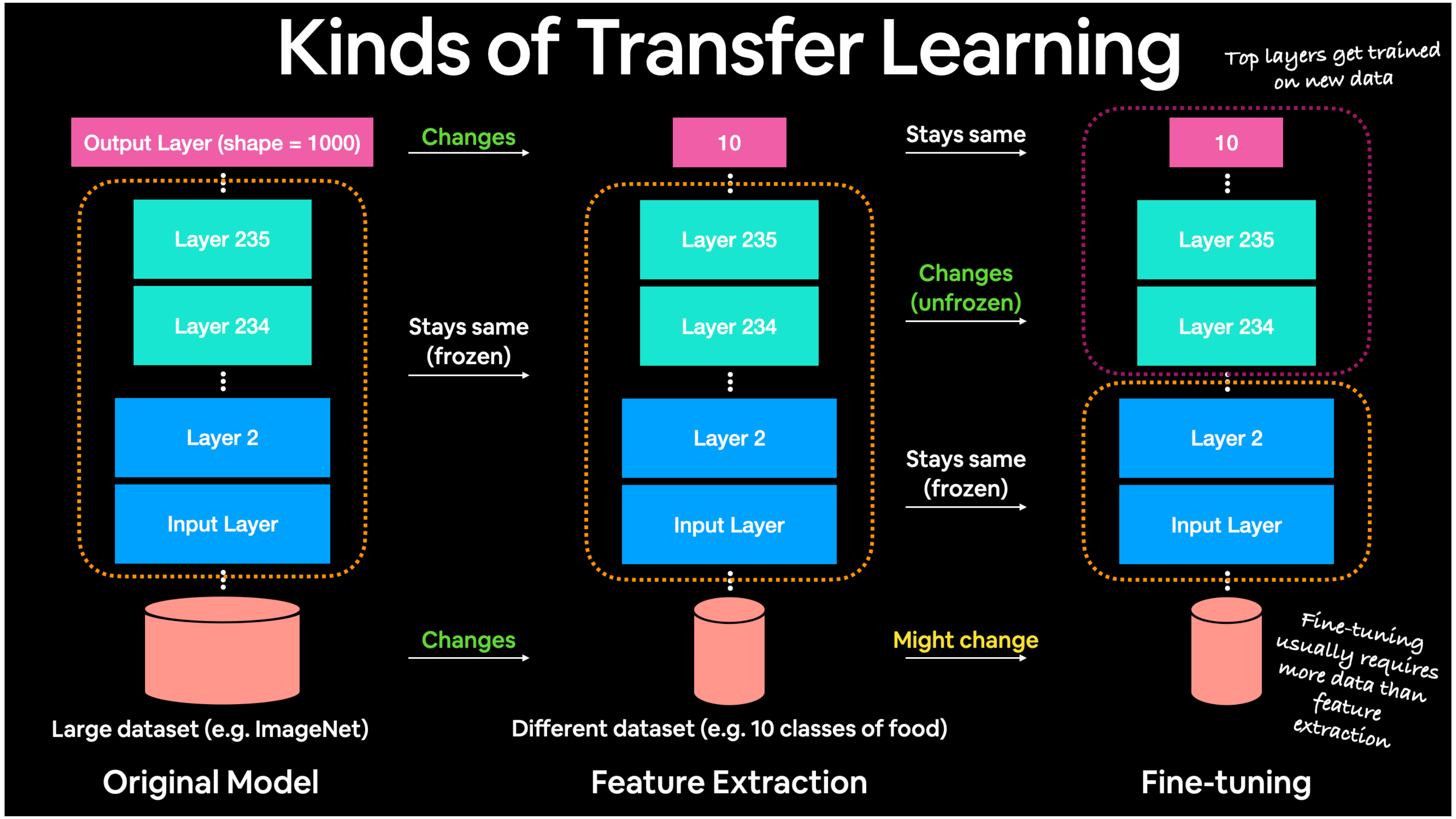
Today’s Topics

- Sampler of Today’s Popular Computer Vision Problems
- Foundation Models
- Textual Prompting & Zero-shot Learning
- Visual Prompting & In-context Few-shot Learning
- Prompt Tuning
- Discussion

Today’s Topics

- Sampler of Today’s Popular Computer Vision Problems
- Foundation Models
- Textual Prompting & Zero-shot Learning
- Visual Prompting & In-context Few-shot Learning
- Prompt Tuning
- Discussion

What We Learned Works Over Past Decade

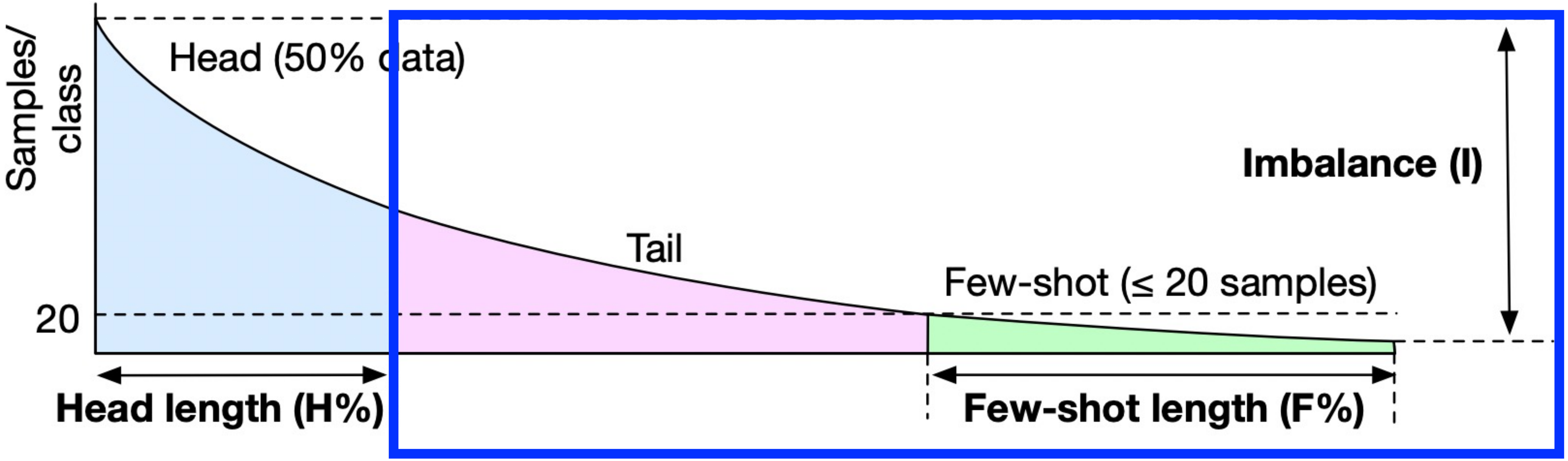


Can achieve strong performance with lots of labeled data for target task (aka closed world problems) when training from scratch or fine-tuning

https://dev.mrdbourke.com/tensorflow-deep-learning/04_transfer_learning_in_tensorflow_part_1_feature_extraction/

Open Problems: Beyond Big Data

Learning with Limited Amounts of Labeled Training Data
(e.g., Few-Shot Learning)

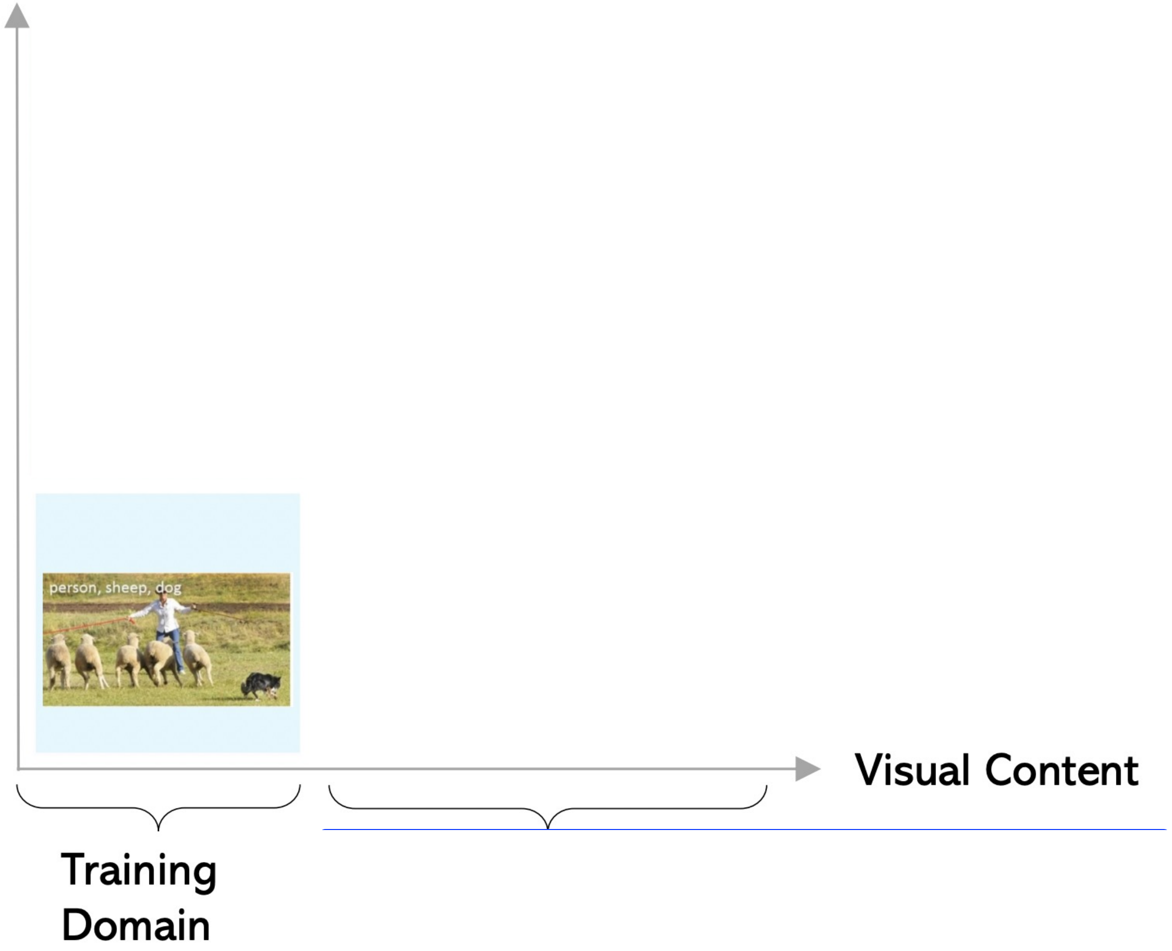


Perrett et al. Use Your Head: Improving Long-Tail Video Recognition. CVPR 2023.

Open Problems: Beyond Closed-World Setting

Vocabulary / Language

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



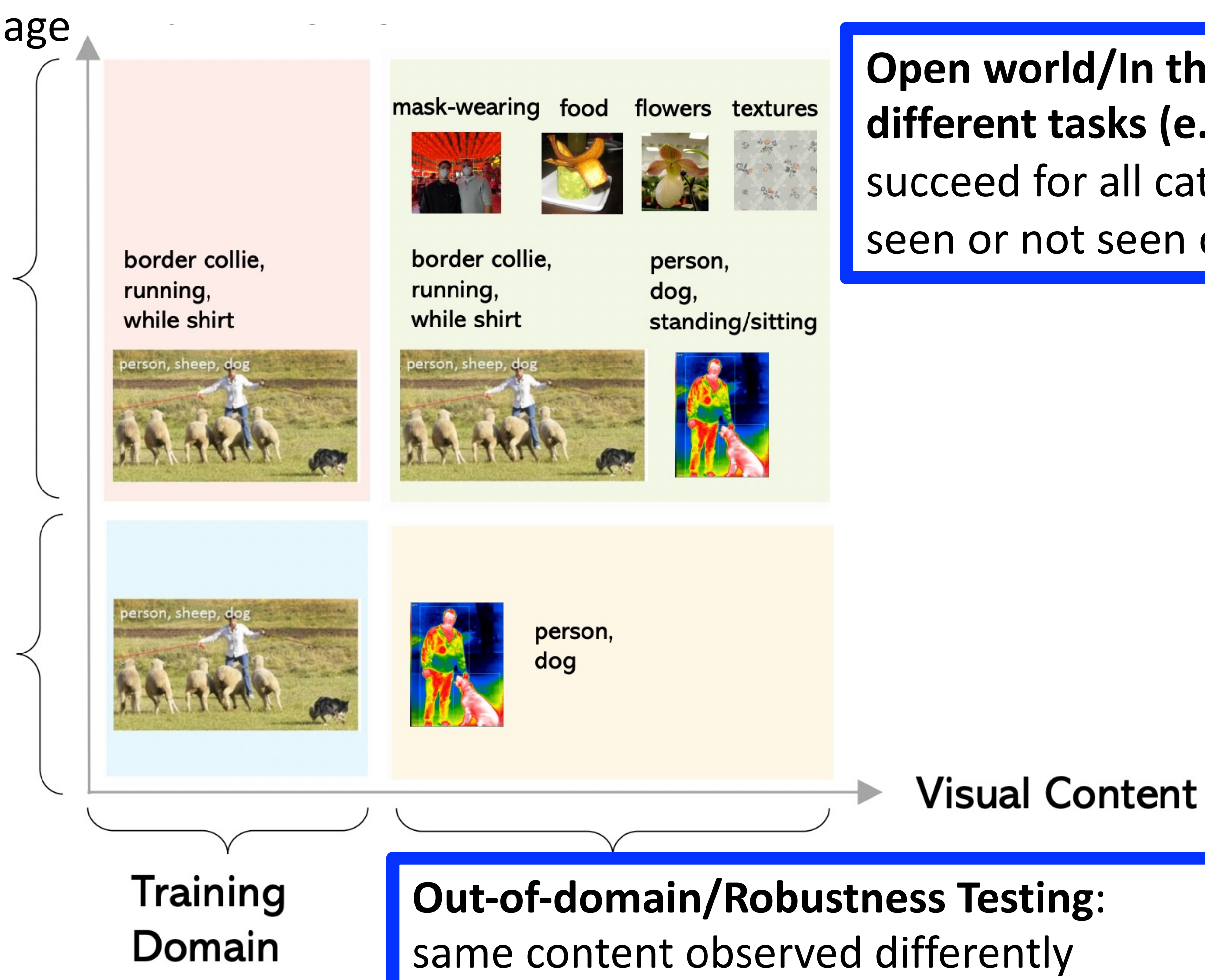
https://arxiv.org/pdf/2210.09263.pdf

Open Problems: Beyond Closed-World Setting

Open vocabulary/Zero-shot: generalize to task with no labeled training data for the target task (e.g., novel categories)

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 are available at training



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

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

https://arxiv.org/pdf/2210.09263.pdf

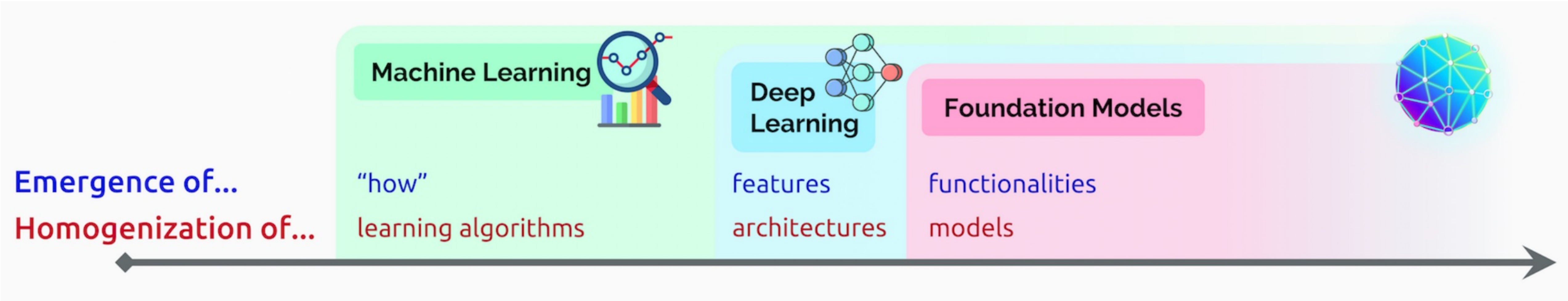
New Paradigm:

Current Findings Suggest Foundation Models Generalize Well With Limited Training Data and Beyond Closed World Tasks

Today’s Topics

- Sampler of Today’s Popular Computer Vision Problems
- Foundation Models
- Textual Prompting & Zero-shot Learning
- Visual Prompting & In-context Few-shot Learning
- Prompt Tuning
- Discussion

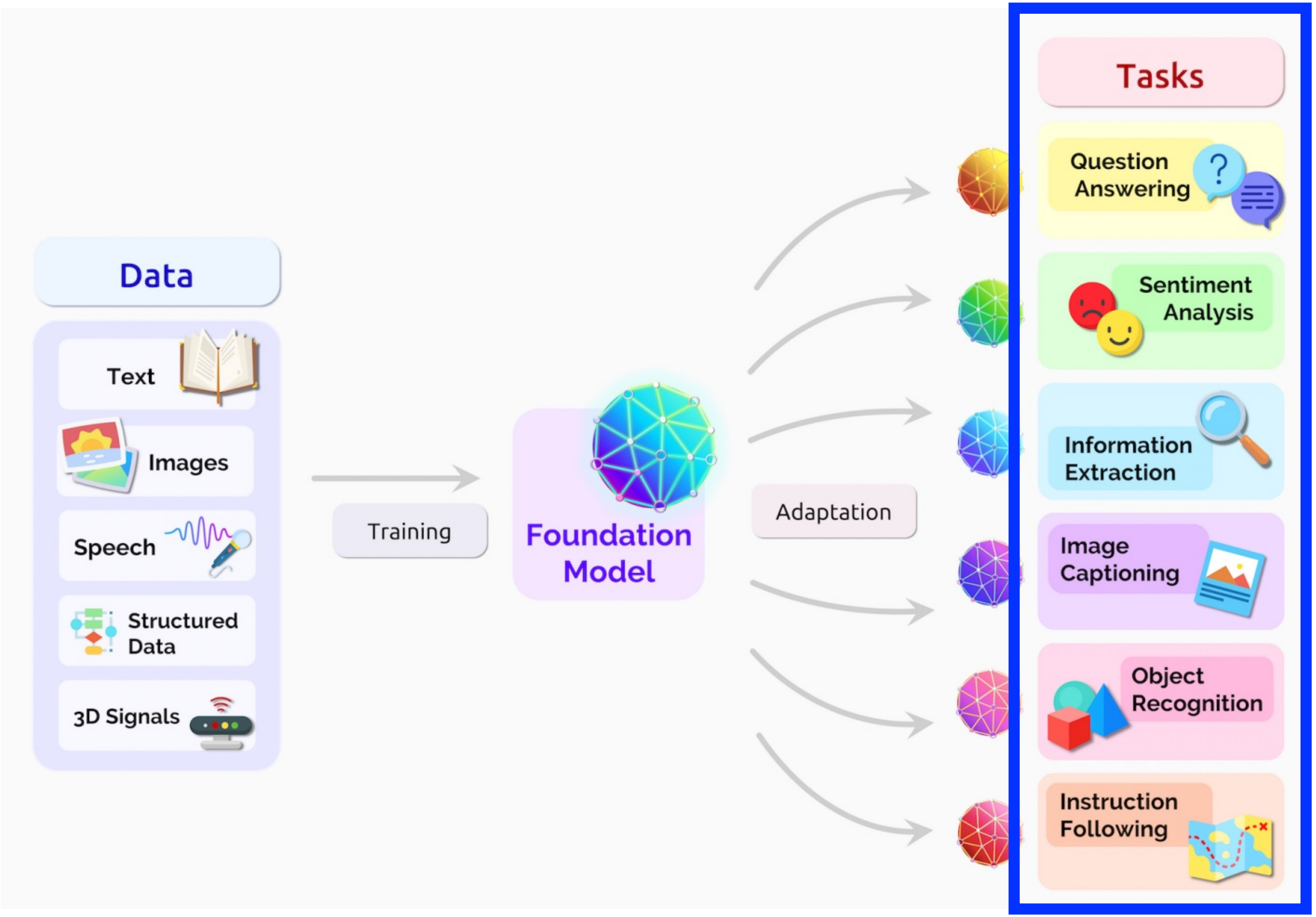
Definition of "Foundation Model"



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

Bommasani et al. On the Opportunities and Risks of Foundation Models. arXiv 2021.

Foundation Models: Training to Evaluation



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)

Bommasani et al. On the Opportunities and Risks of Foundation Models. arXiv 2021.

Foundation Models: Why Now?

Key ingredients identified:

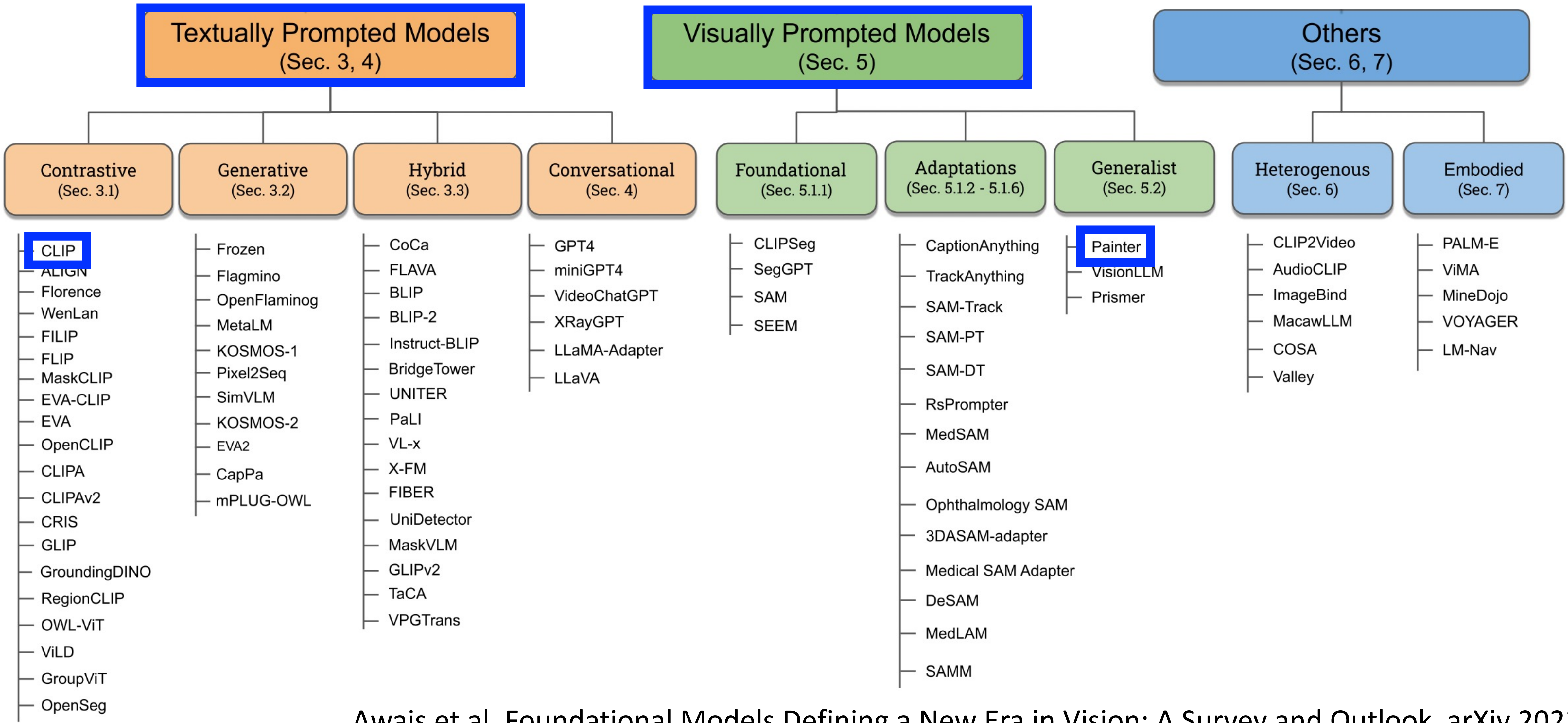
1. Transformer model architecture

2. Lots more training data by using Internet data

3. Sufficient hardware with modern GPUs

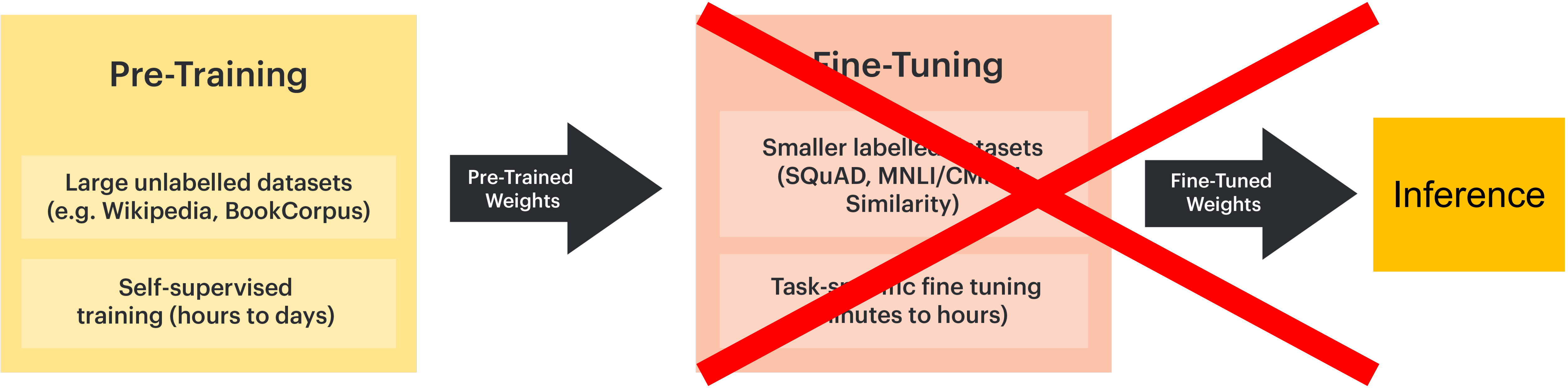
Bommasani et al. On the Opportunities and Risks of Foundation Models. arXiv 2021.

Foundation Models in Computer Vision



Awais et al. Foundational Models Defining a New Era in Vision: A Survey and Outlook. arXiv 2023.

Beyond Pretraining and Fine-Tuning Paradigm



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

https://docs.graphcore.ai/projects/bert-training/en/latest/bert.html

Today’s Topics

- Sampler of Today’s Popular Computer Vision Problems
- Foundation Models
- Textual Prompting & Zero-shot Learning
- Visual Prompting & In-context Few-shot Learning
- Prompt Tuning
- Discussion

Foundation Models: What's New?

Key ingredients identified:

1. Transformer model architecture
2. $\boxed{\text{Lots more training data by using Internet data}}$
3. Sufficient hardware with modern GPUs

Bommasani et al. On the Opportunities and Risks of Foundation Models. arXiv 2021.

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

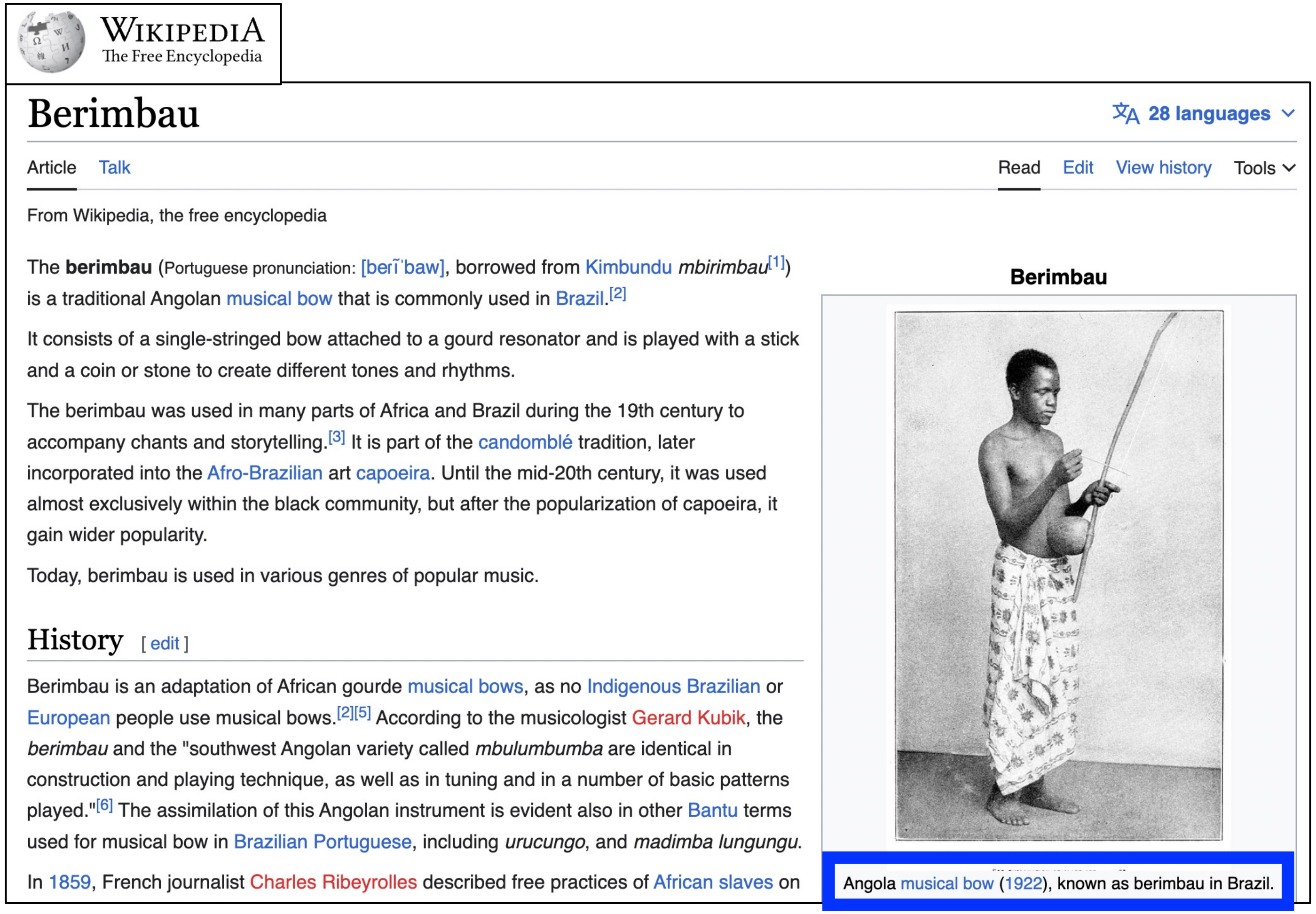
1. Image-Text Pair Collection

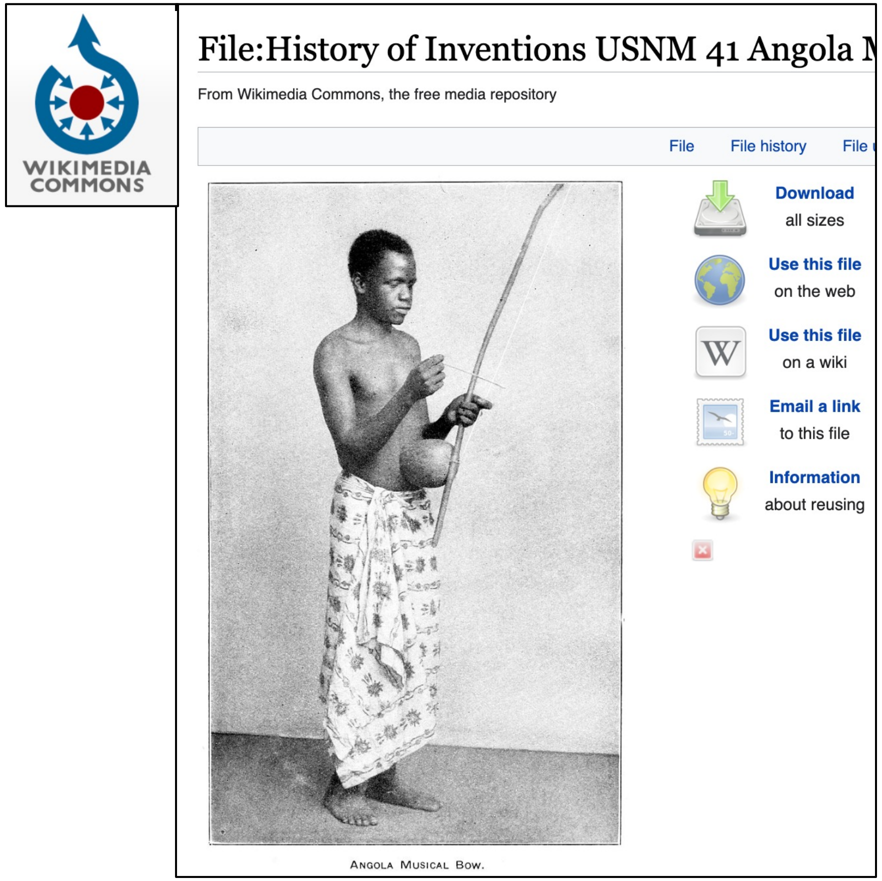
- Source: Wikipedia, given its high quality (editorially reviewed), large size (~124M pages), and diversity (279 languages)

- Extracted ~150 million image-text pairs

Srinivasan et al. WIT: Wikipedia-based Image Text Dataset for Multimodal Multilingual Machine Learning. SIGIR 2021.

For Each Image, Multiple Texts Extracted:





(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

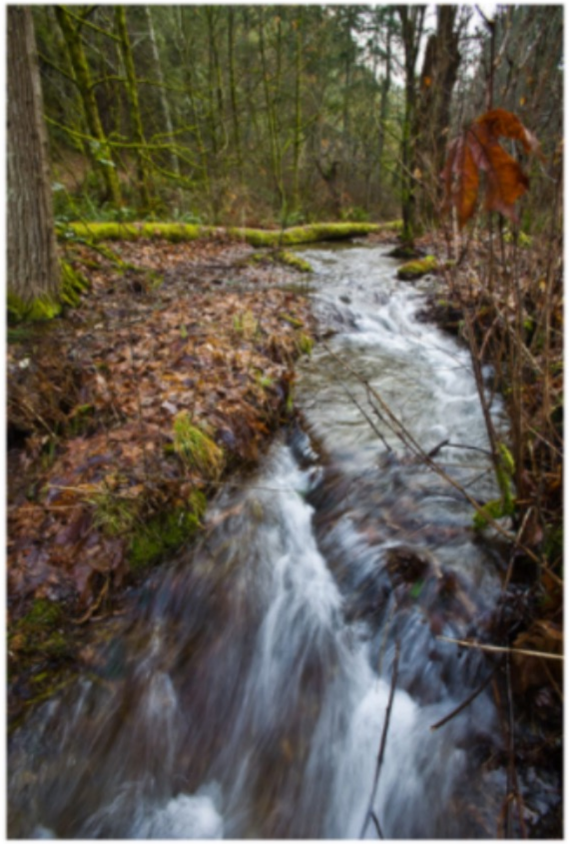
- Examples were in English (~3,000), German (300), French (300), Spanish (300), Russian (300), Chinese (300), & Hindi (100)

Srinivasan et al. WIT: Wikipedia-based Image Text Dataset for Multimodal Multilingual Machine Learning. SIGIR 2021.

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: Sequalitchew 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: Sequalitchew Creek, lower canyon Text2: Sequalitchew Creek, lower canyon Extra: Sequalitchew-Creek-lower-canyon.jpg Sequalitchew 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

Srinivasan et al. WIT: Wikipedia-based Image Text Dataset for Multimodal Multilingual Machine Learning. SIGIR 2021.

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)

Srinivasan et al. WIT: Wikipedia-based Image Text Dataset for Multimodal Multilingual Machine Learning. SIGIR 2021.

Foundation Model: CLIP

Key ingredients:

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

Bommasani et al. On the Opportunities and Risks of Foundation Models. arXiv 2021.

Why CLIP?

Named after the proposed technique: Contrastive Language Image Pre-training

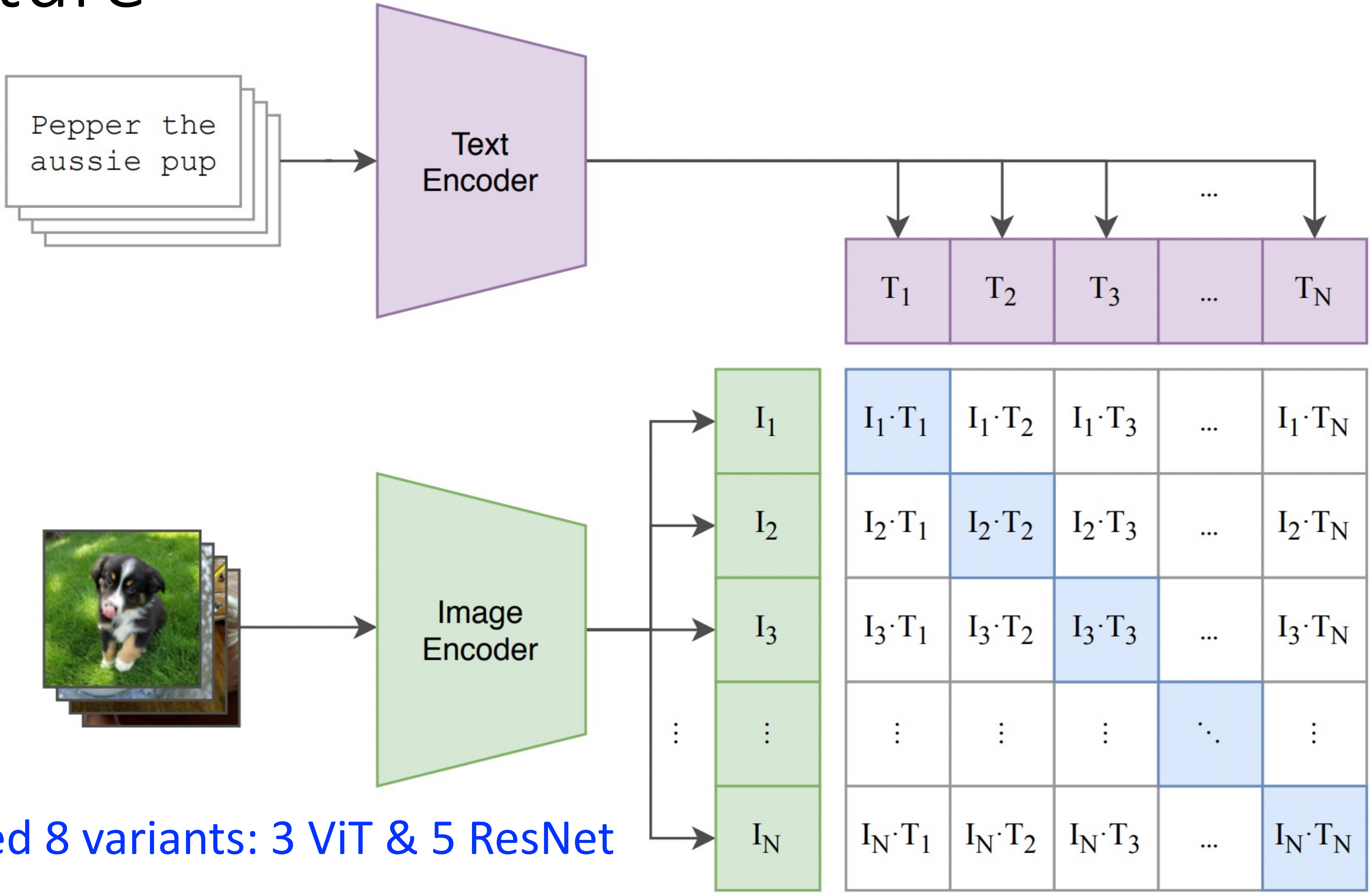
Radford et al. Learning Transferable Visual Models From Natural Language Supervision. ICML 2021.

CLIP Model: Novelty

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

CLIP Architecture

Text transformer (GPT-2)



Tried 8 variants: 3 ViT & 5 ResNet

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

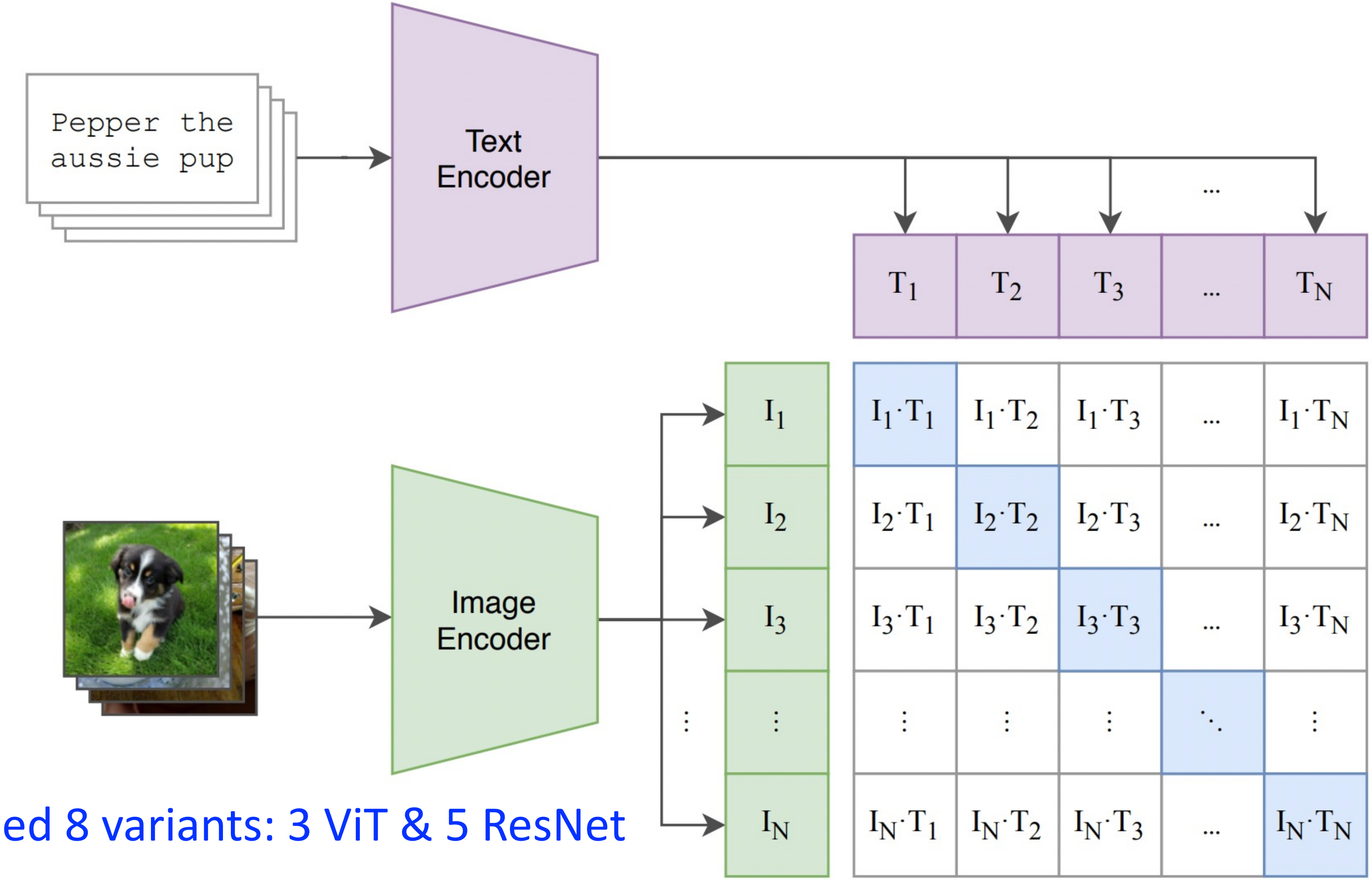
CLIP Training

Text transformer (GPT-2)

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

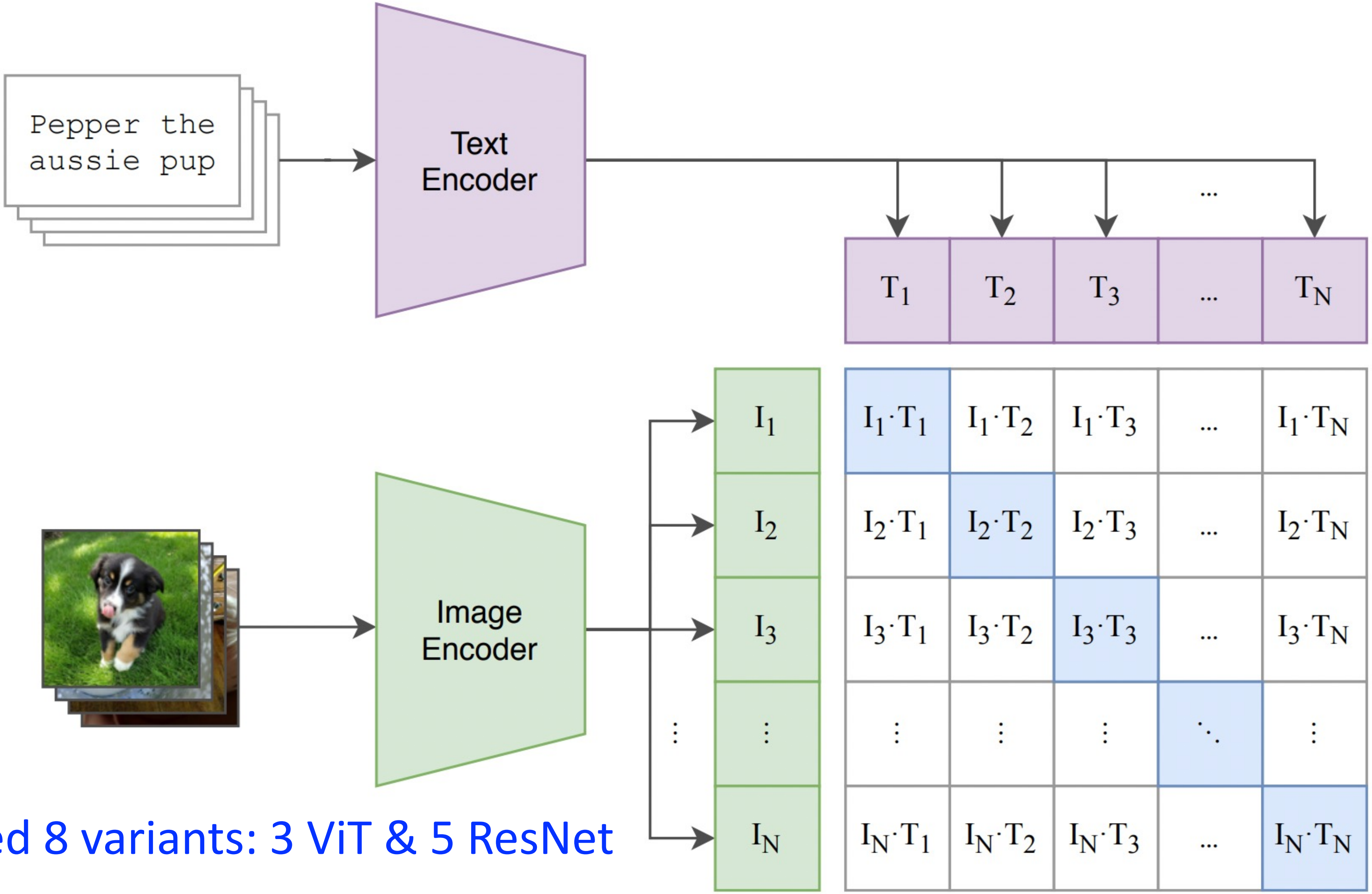
Radford et al. Learning Transferable Visual Models From Natural Language Supervision. ICML 2021.

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!

Text transformer (GPT-2)



Tried 8 variants: 3 ViT & 5 ResNet

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

Zero-Shot Performance Evaluated on Over 30 Datasets

CLIP Evaluation

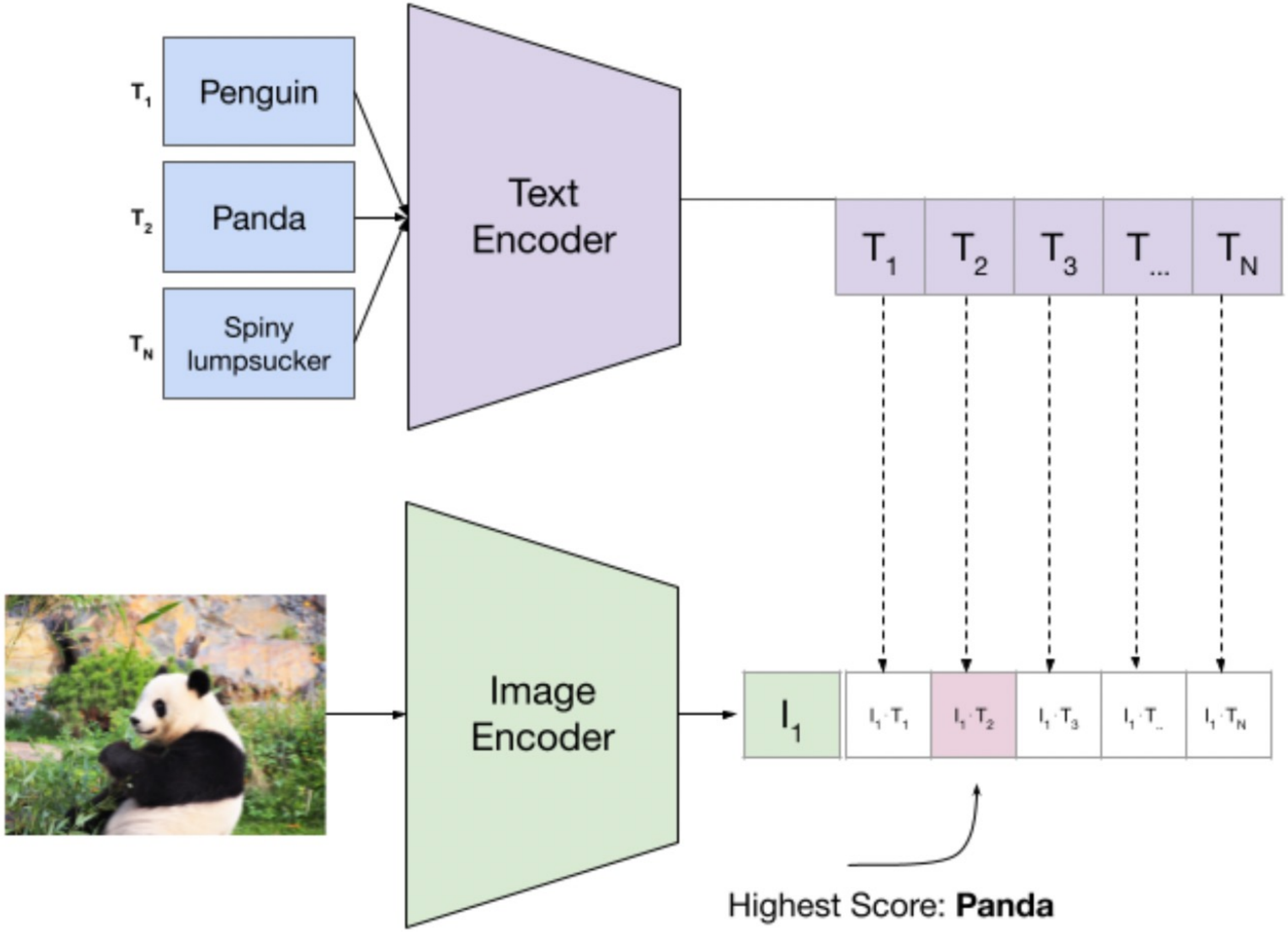
Subset of datasets shown here:

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 Inference

e.g., zero-shot classification: configure representations for all candidate labels (e.g., animal species) using the pretrained encoder and then predict category of image contents based on cosine similarity to category candidates



https://towardsdatascience.com/understanding-zero-shot-learning-making-ml-more-human-4653ac35ccab

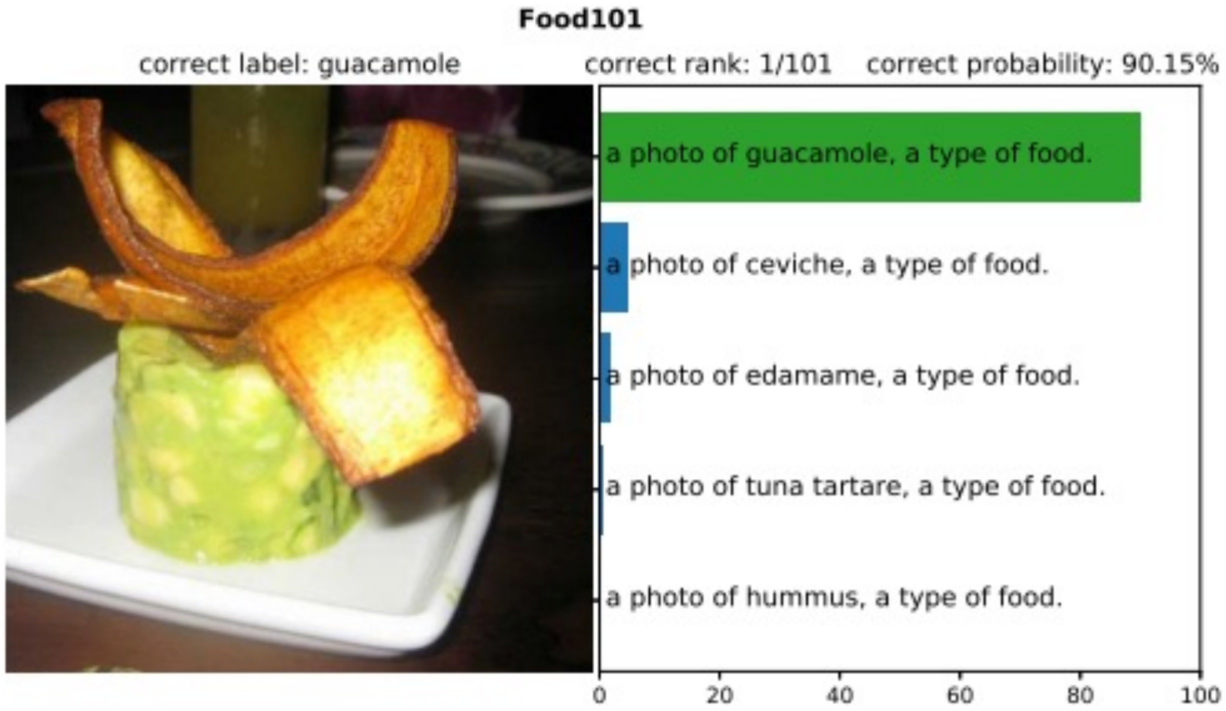
CLIP Inference

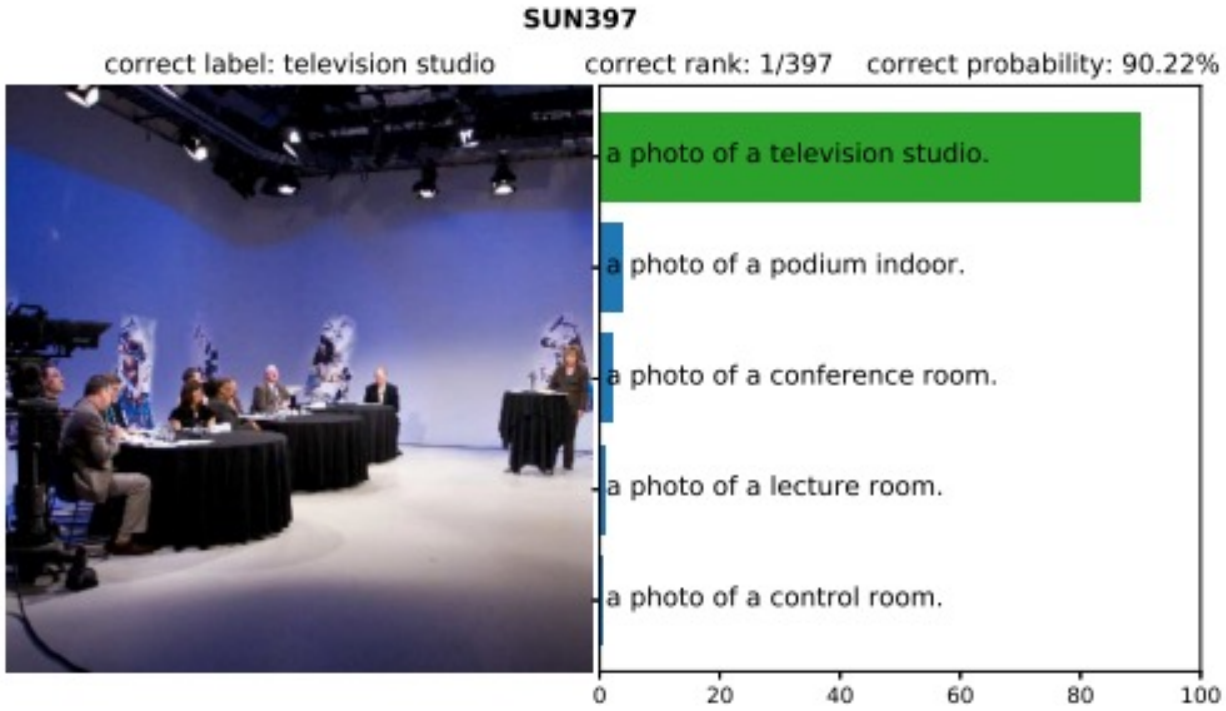
Prompts “engineered” that mimic training data (by being a sentence):

- 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}”

https://towardsdatascience.com/understanding-zero-shot-learning-making-ml-more-human-4653ac35ccab

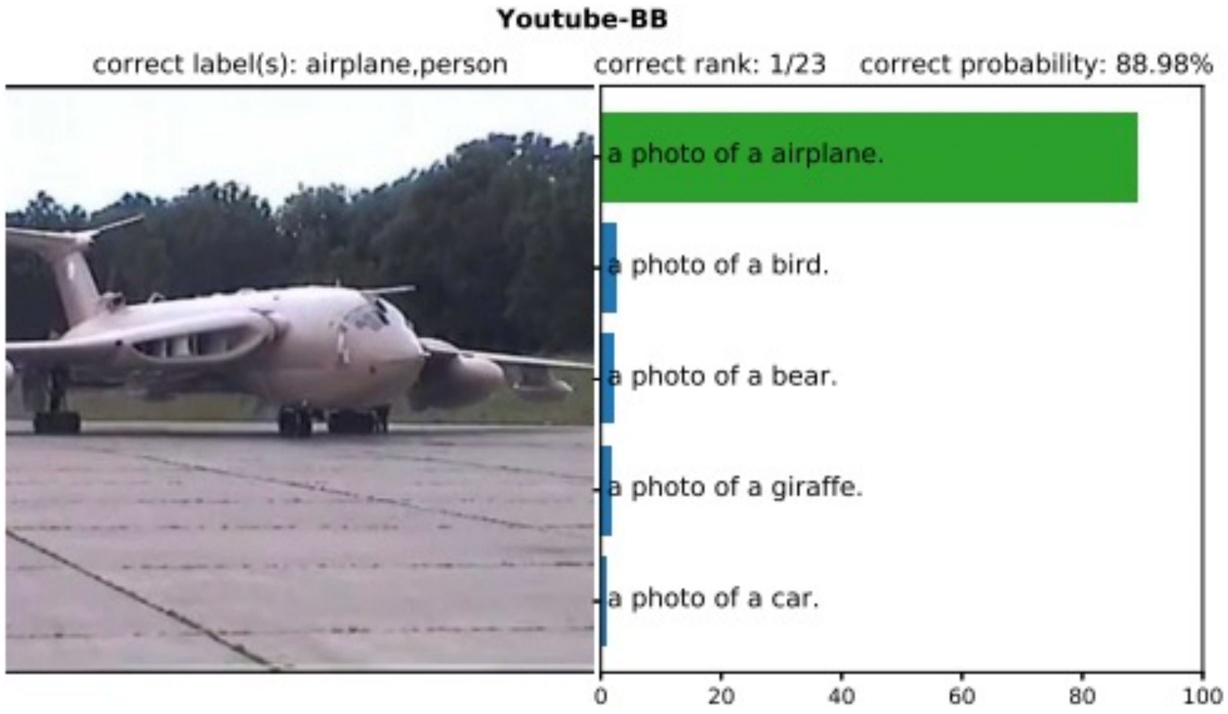
CLIP: Qualitative Results

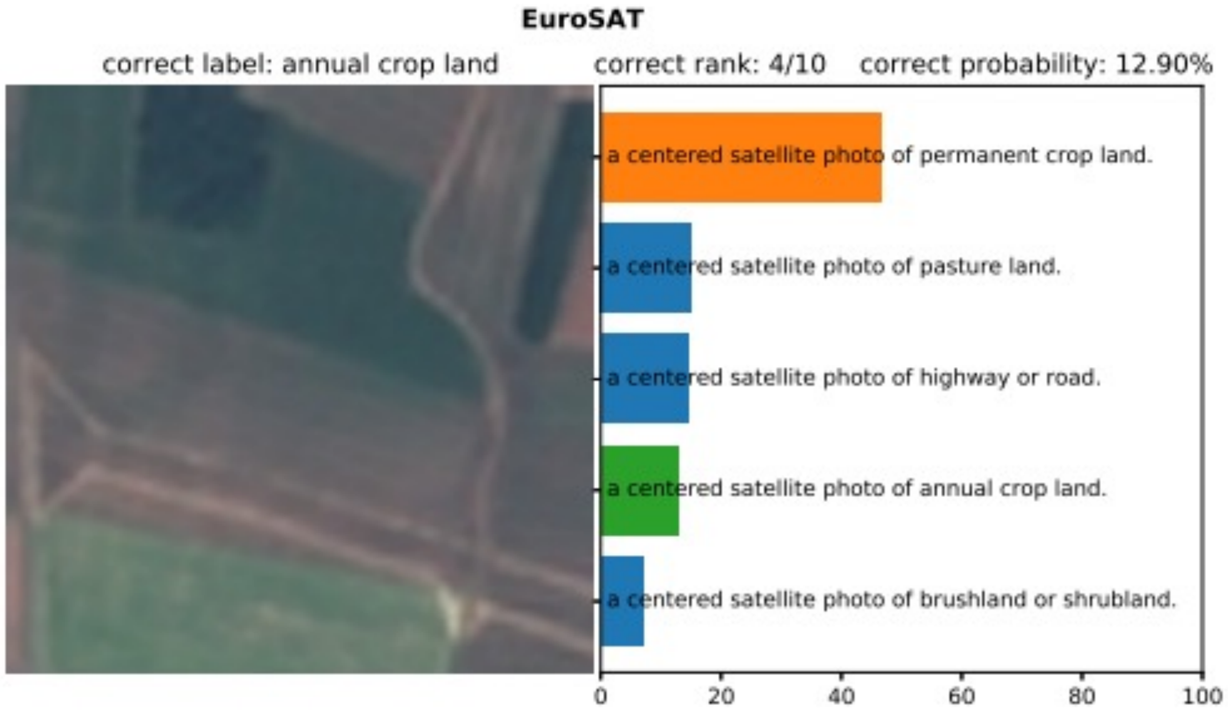




Radford et al. Learning Transferable Visual Models From Natural Language Supervision. ICML 2021.

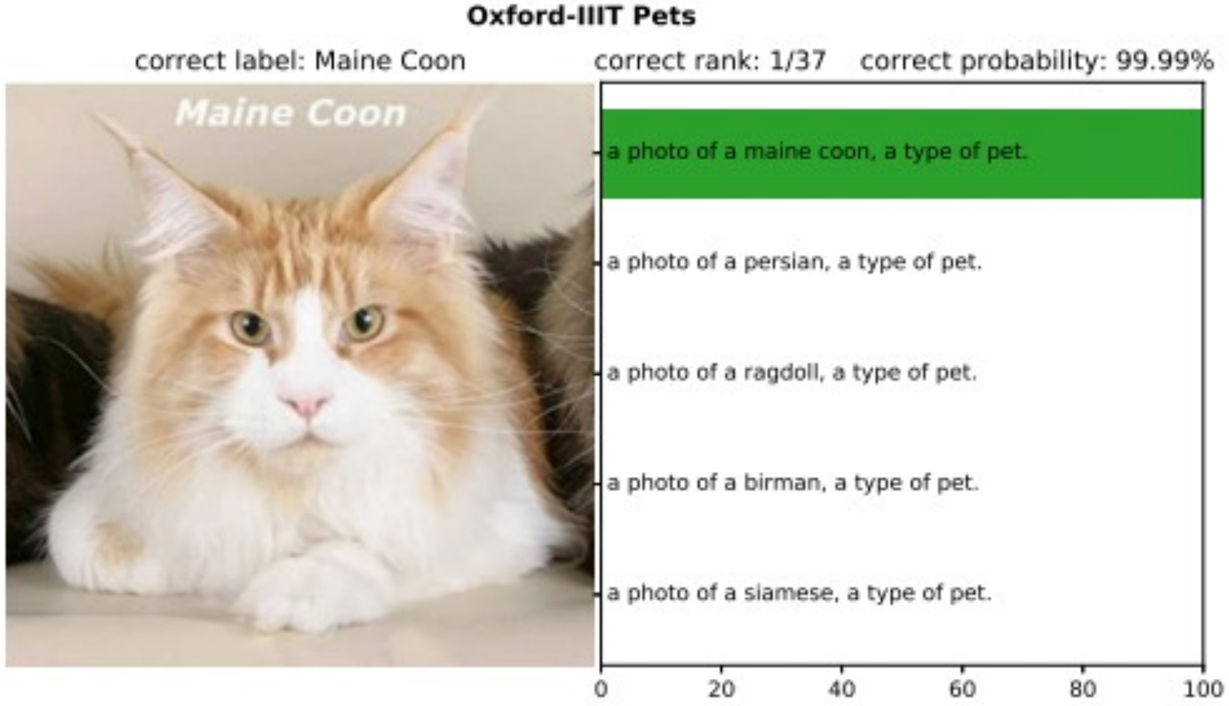
CLIP: Qualitative Results

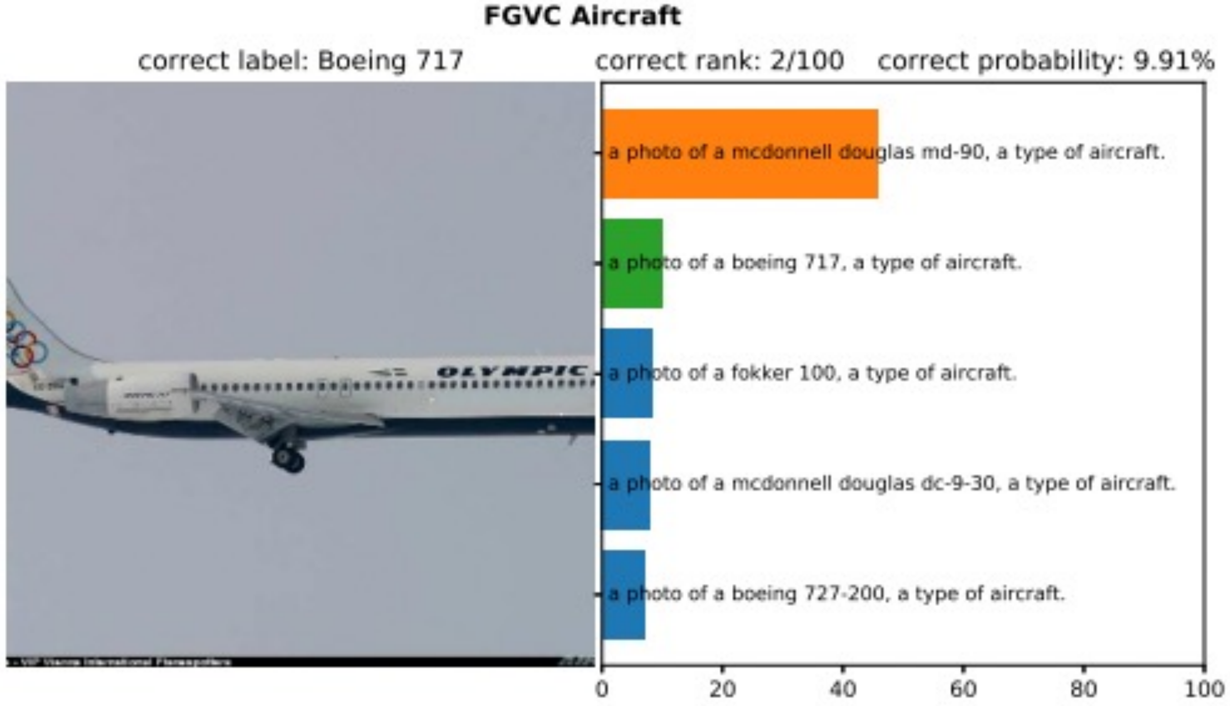




Radford et al. Learning Transferable Visual Models From Natural Language Supervision. ICML 2021.

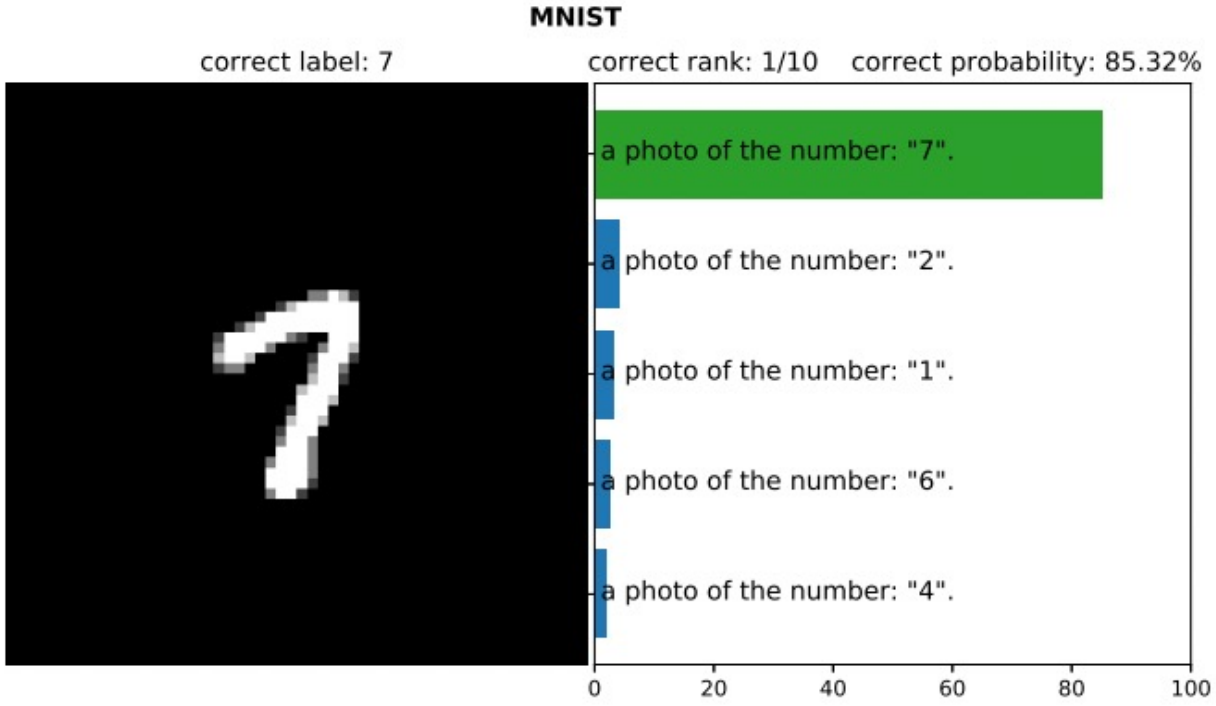
CLIP: Qualitative Results

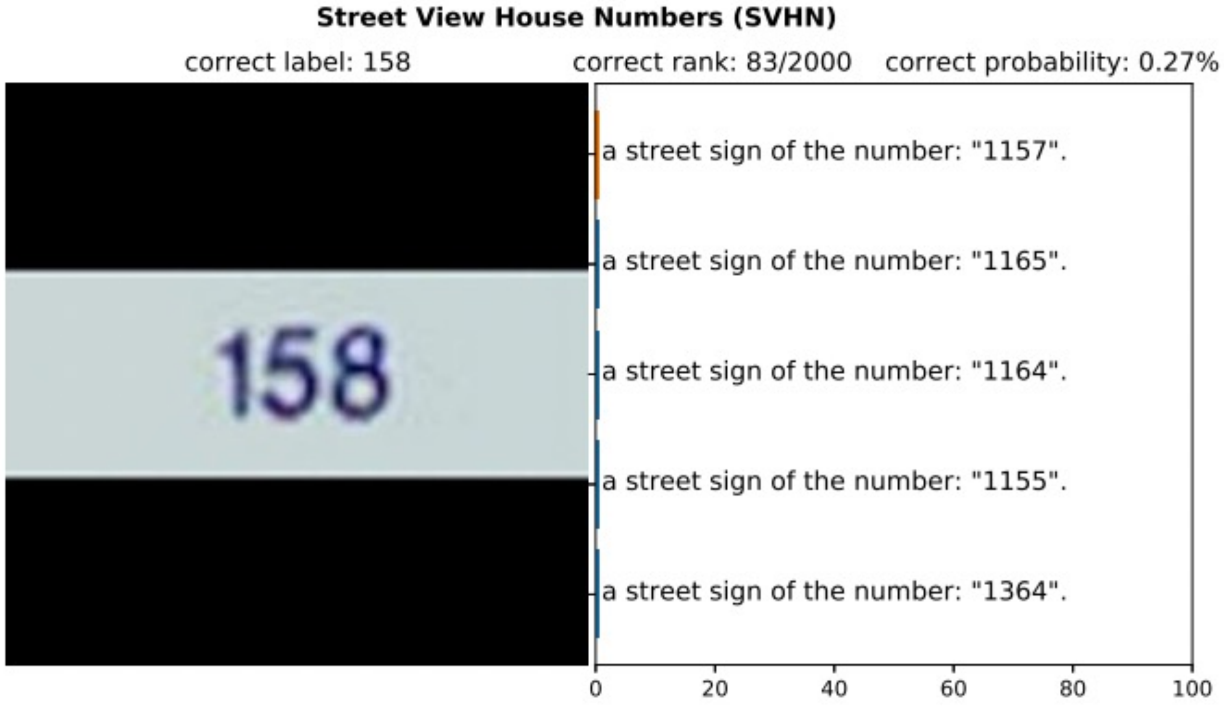




Radford et al. Learning Transferable Visual Models From Natural Language Supervision. ICML 2021.

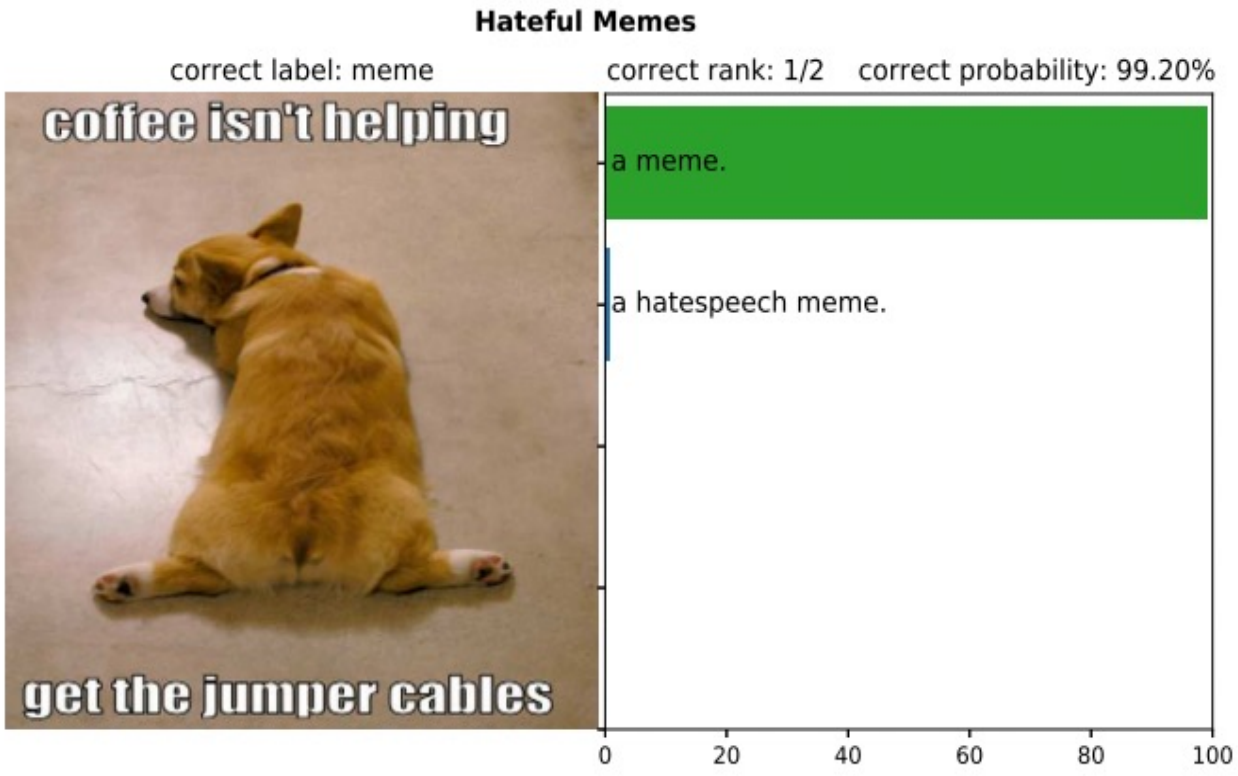
CLIP: Qualitative Results

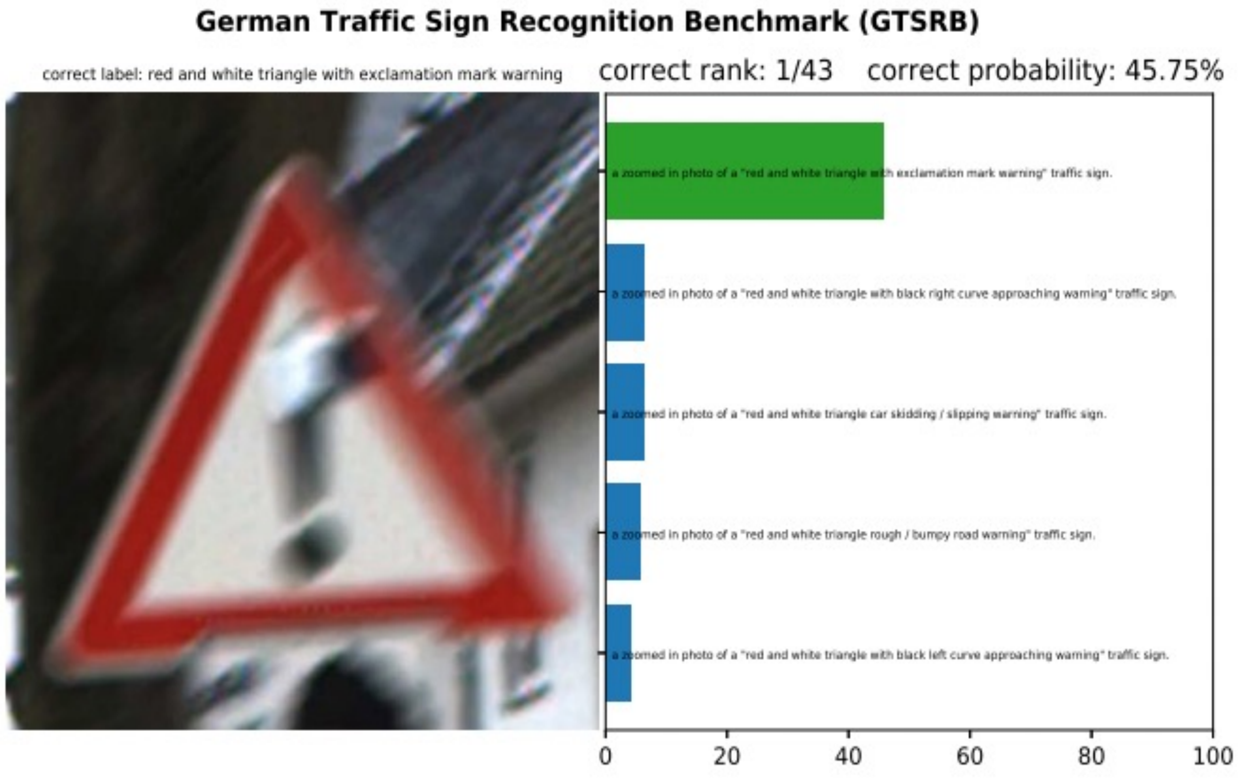




Radford et al. Learning Transferable Visual Models From Natural Language Supervision. ICML 2021.

CLIP: Qualitative Results





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

Today’s Topics

- Sampler of Today’s Popular Computer Vision Problems
- Foundation Models
- Textual Prompting & Zero-shot Learning
- Visual Prompting & In-context Few-shot Learning
- Prompt Tuning
- Discussion

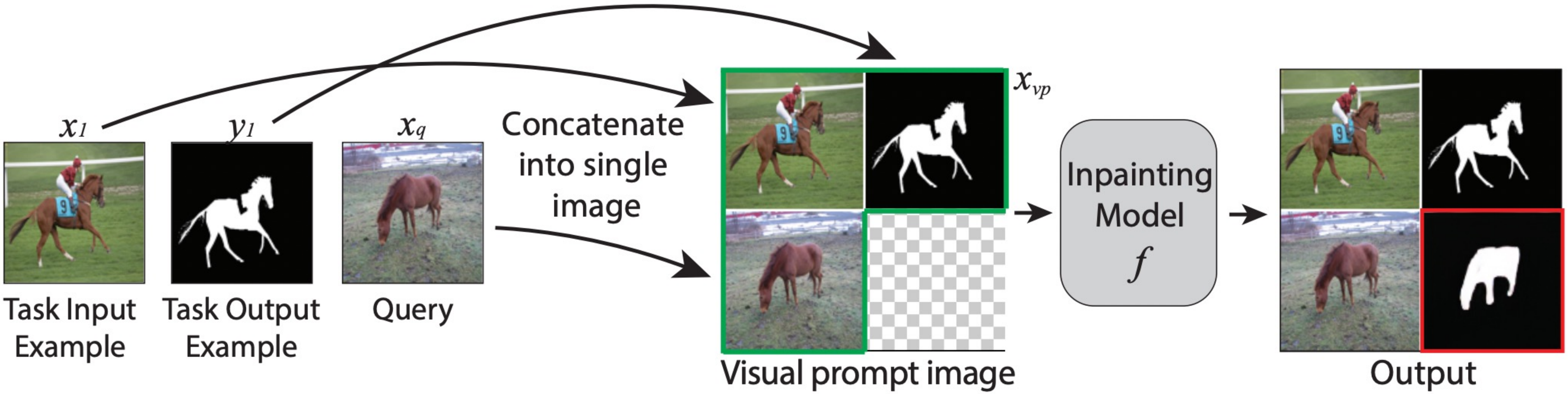
Motivation

- 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

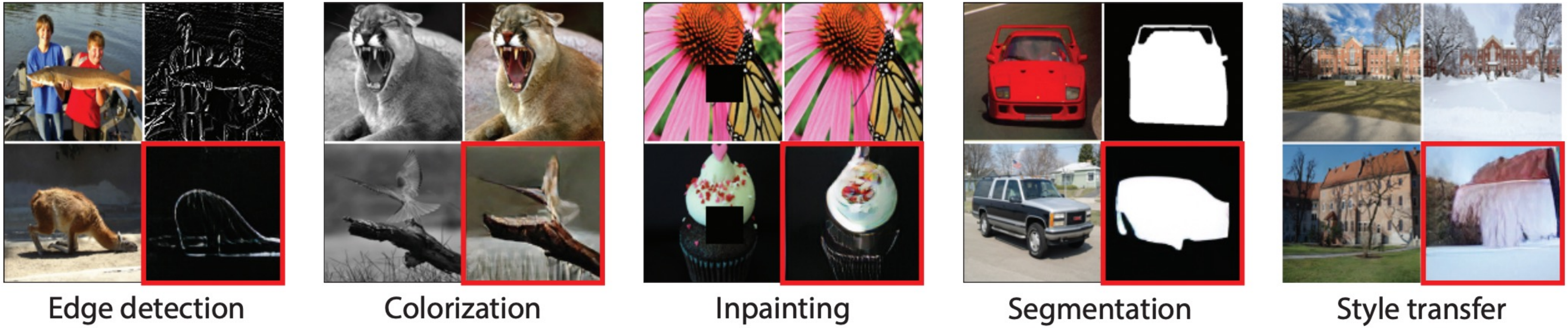


Designed to adapt to any "image-to-image translation" task
by using the model as is (e.g., no fine-tuning required)

Bar et al. Visual Prompting via Image Inpainting. Neurips 2022.

Idea

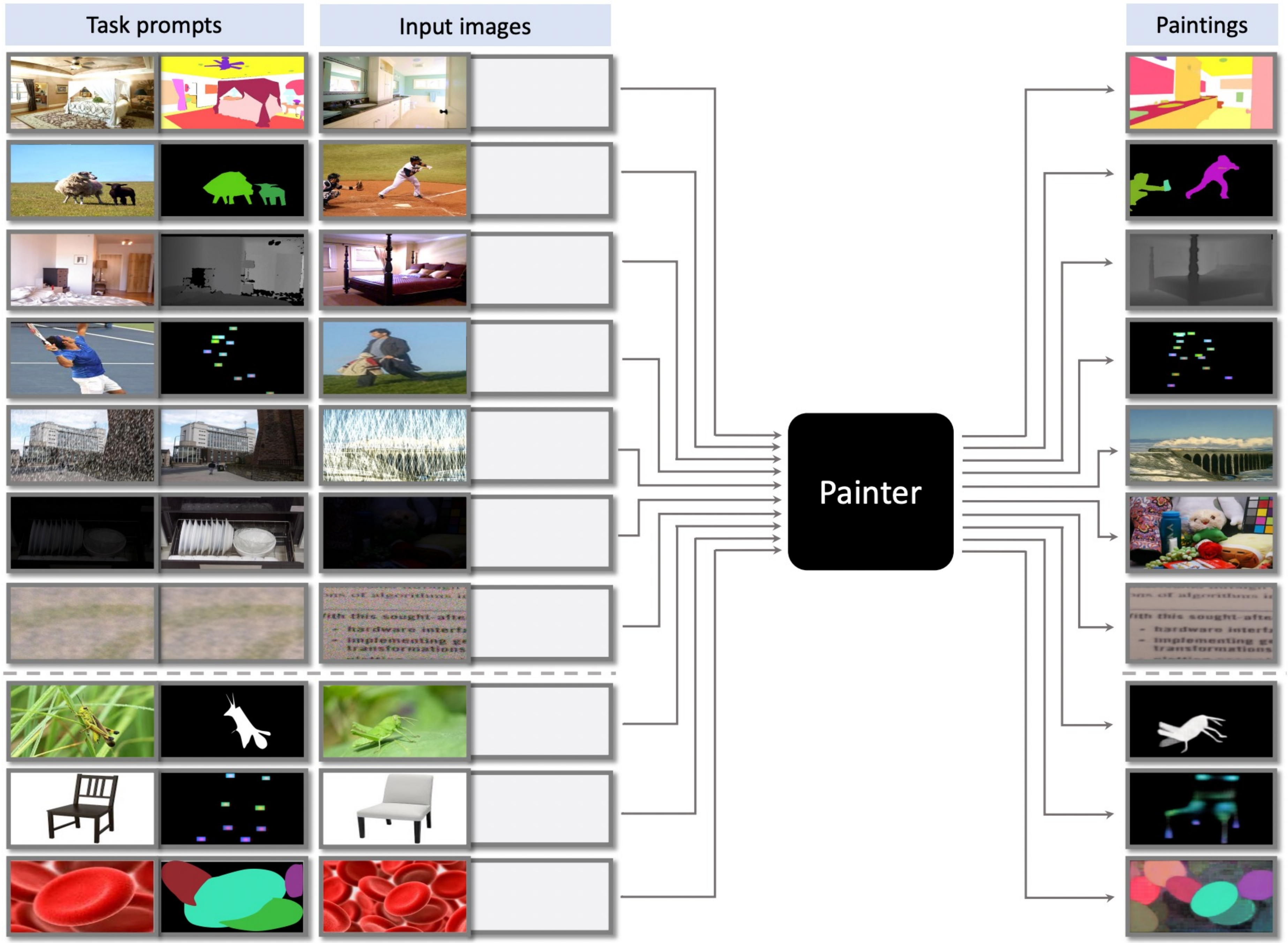
Image inpainting for prompting introduced in 2022 by Bar et al.



Bar et al. Visual Prompting via Image Inpainting. Neurips 2022.

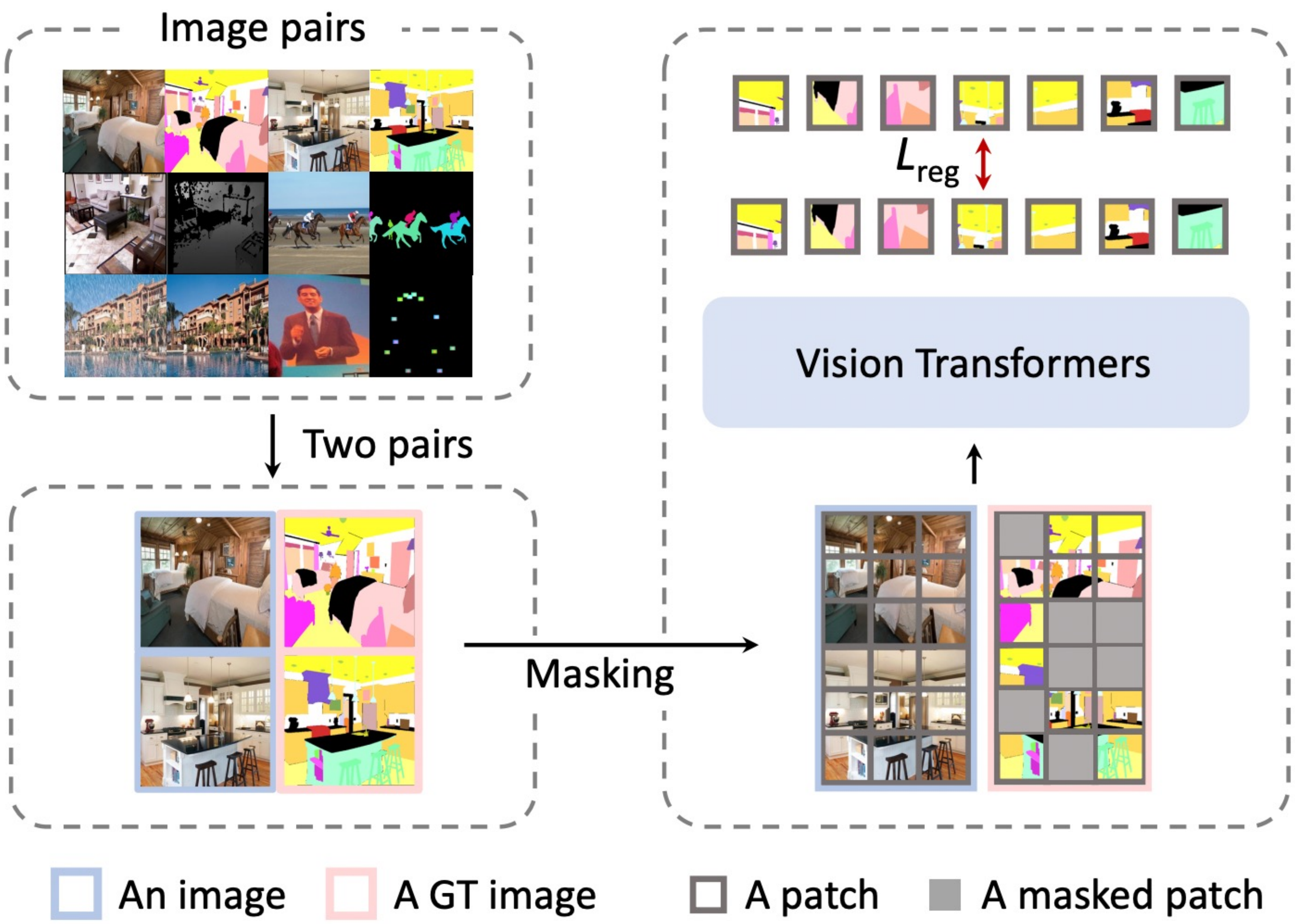
Idea

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



Wang et al. Images Speak in Images: A Generalist Painter for In-Context Visual Learning. CVPR 2023.

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

Wang et al. Images Speak in Images: A Generalist Painter for In-Context Visual Learning. CVPR 2023.

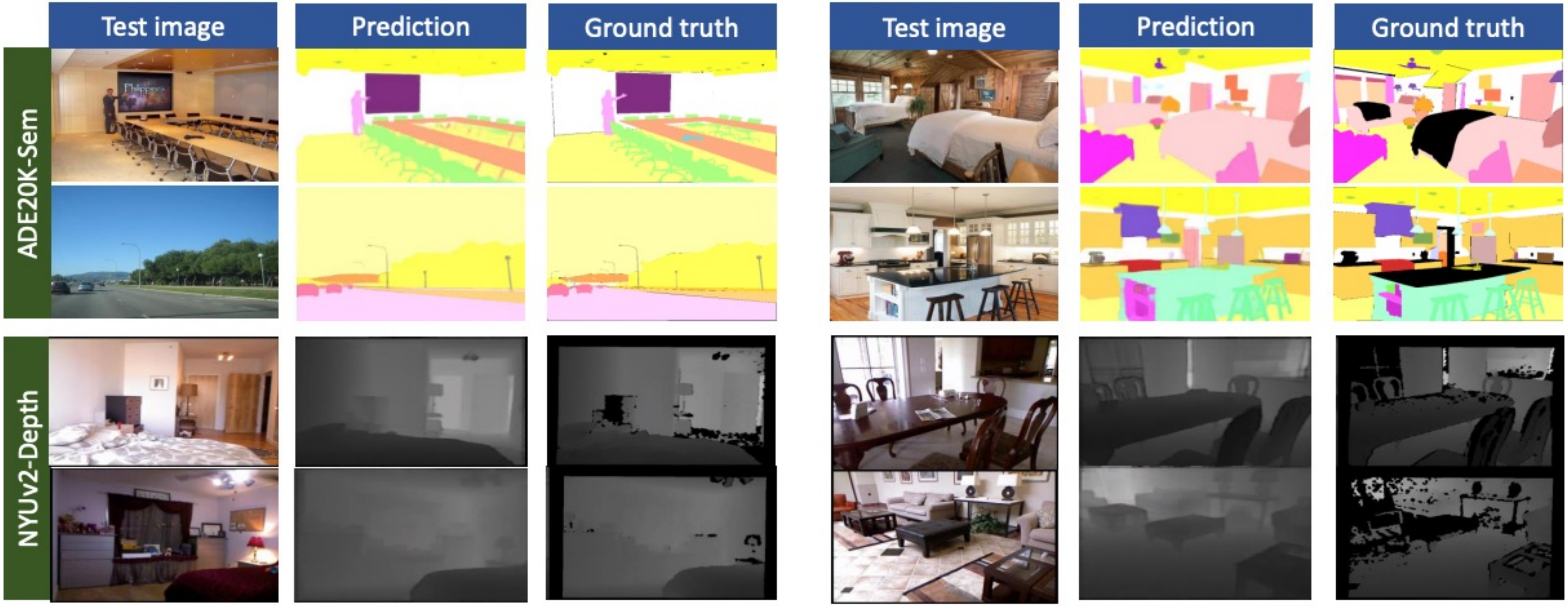
Experimental Results

(Used as Prompt the best performing example-per 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.

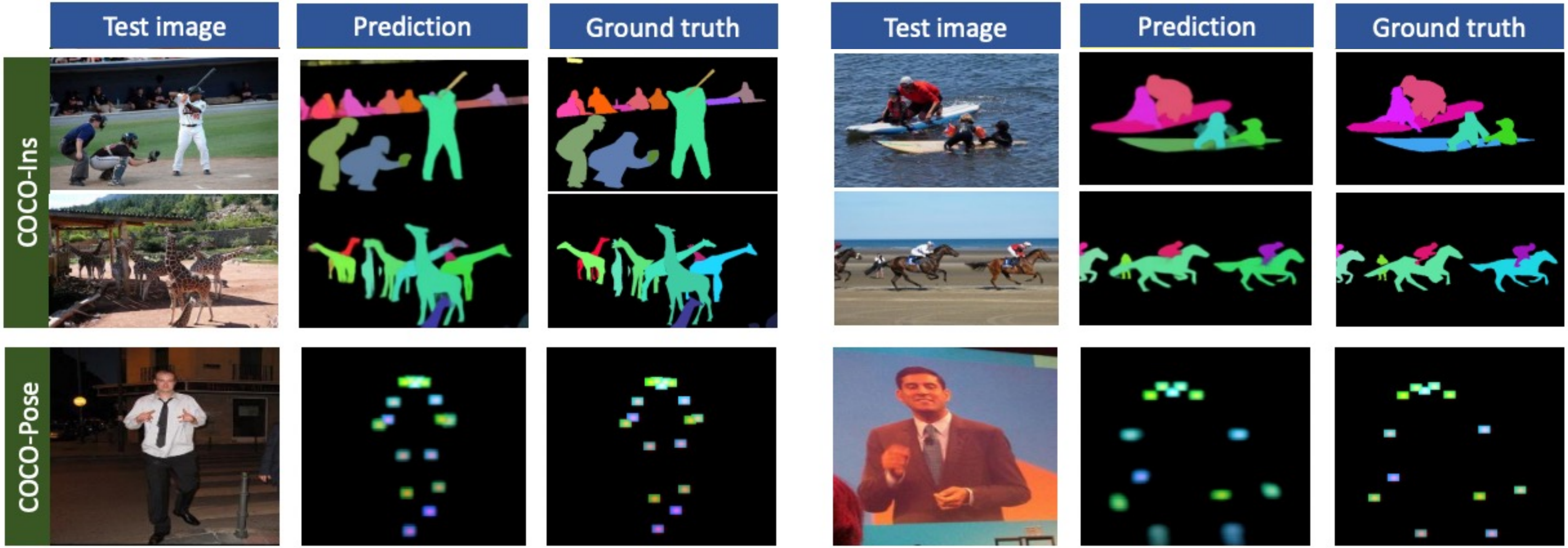
Wang et al. Images Speak in Images: A Generalist Painter for In-Context Visual Learning. CVPR 2023.

Qualitative Results: In-Domain Results



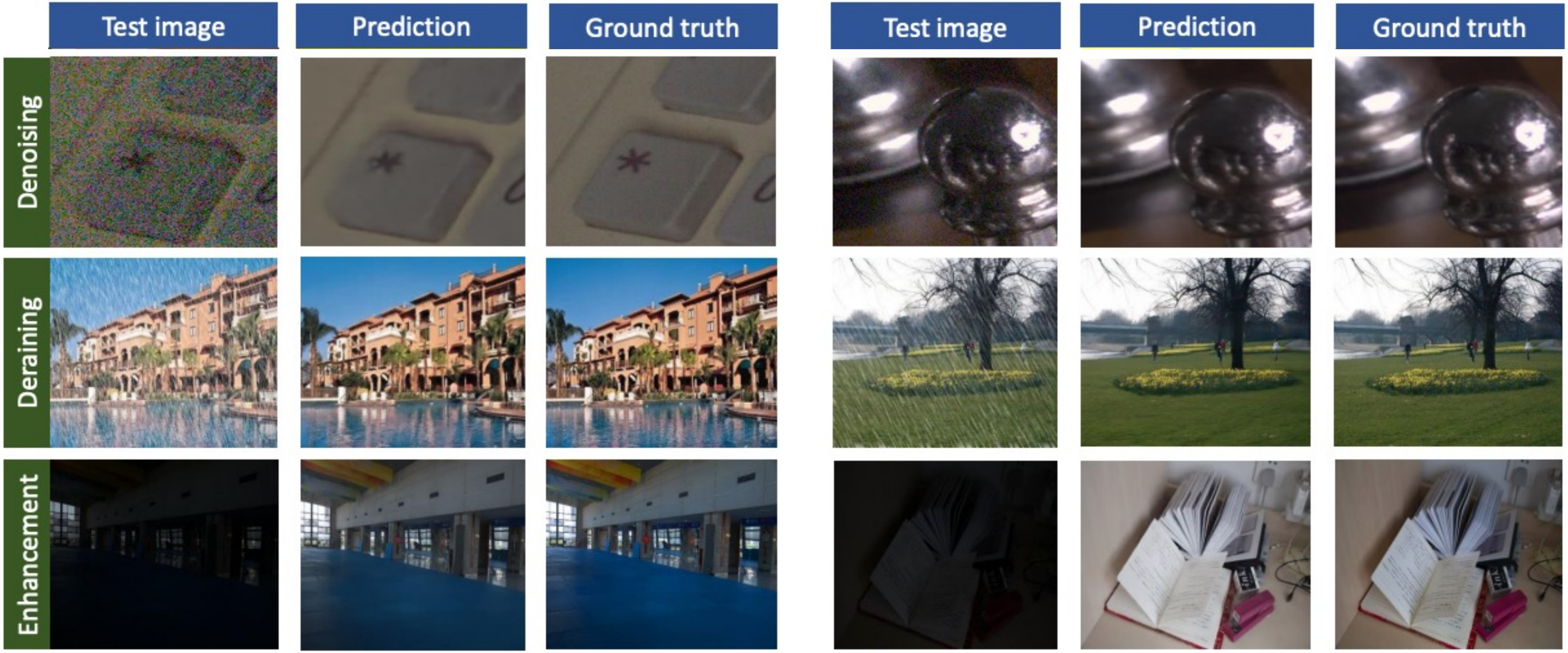
Wang et al. Images Speak in Images: A Generalist Painter for In-Context Visual Learning. CVPR 2023.

Qualitative Results: In-Domain Results



Wang et al. Images Speak in Images: A Generalist Painter for In-Context Visual Learning. CVPR 2023.

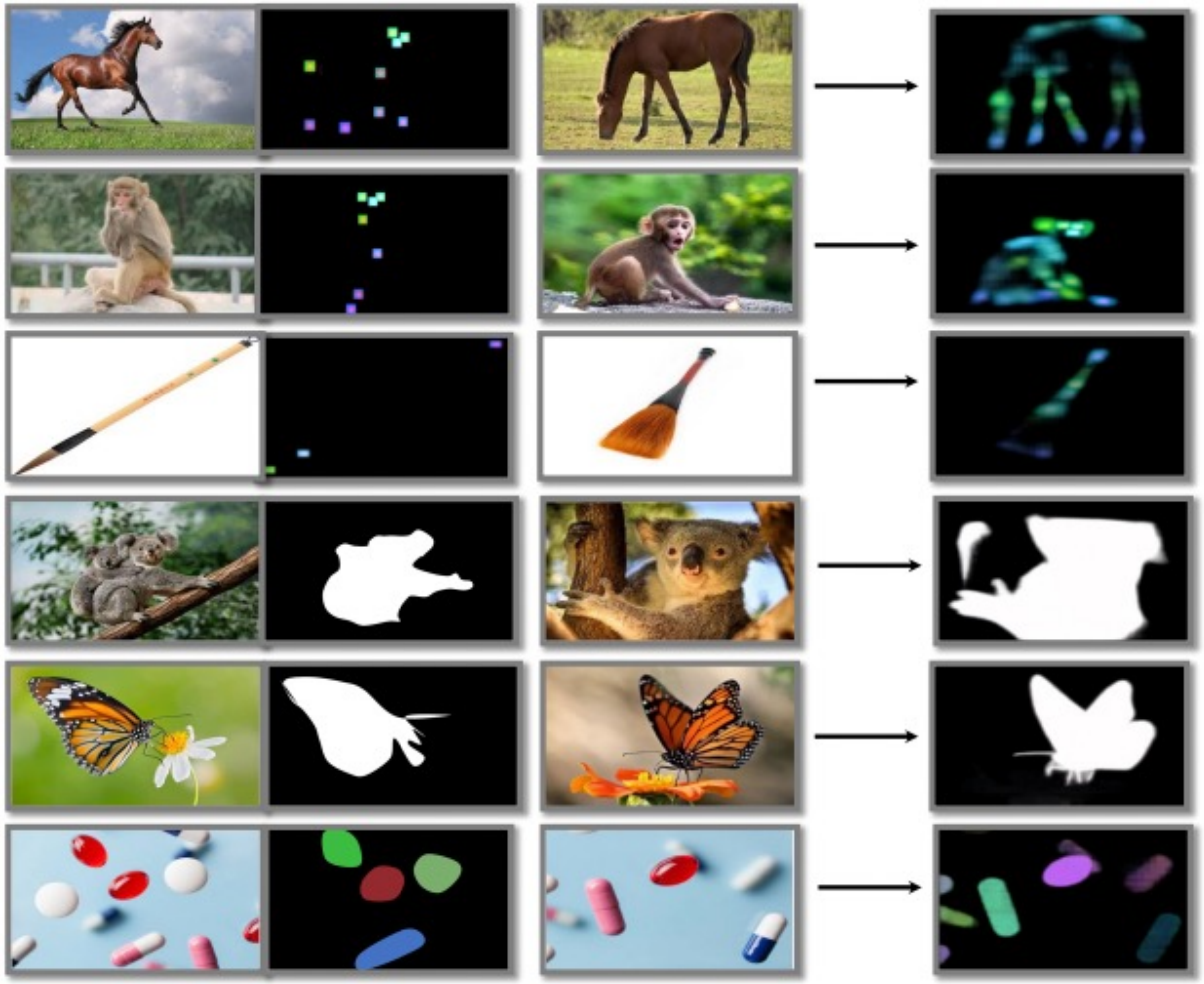
Qualitative Results: In-Domain Results



Wang et al. Images Speak in Images: A Generalist Painter for In-Context Visual Learning. CVPR 2023.

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

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

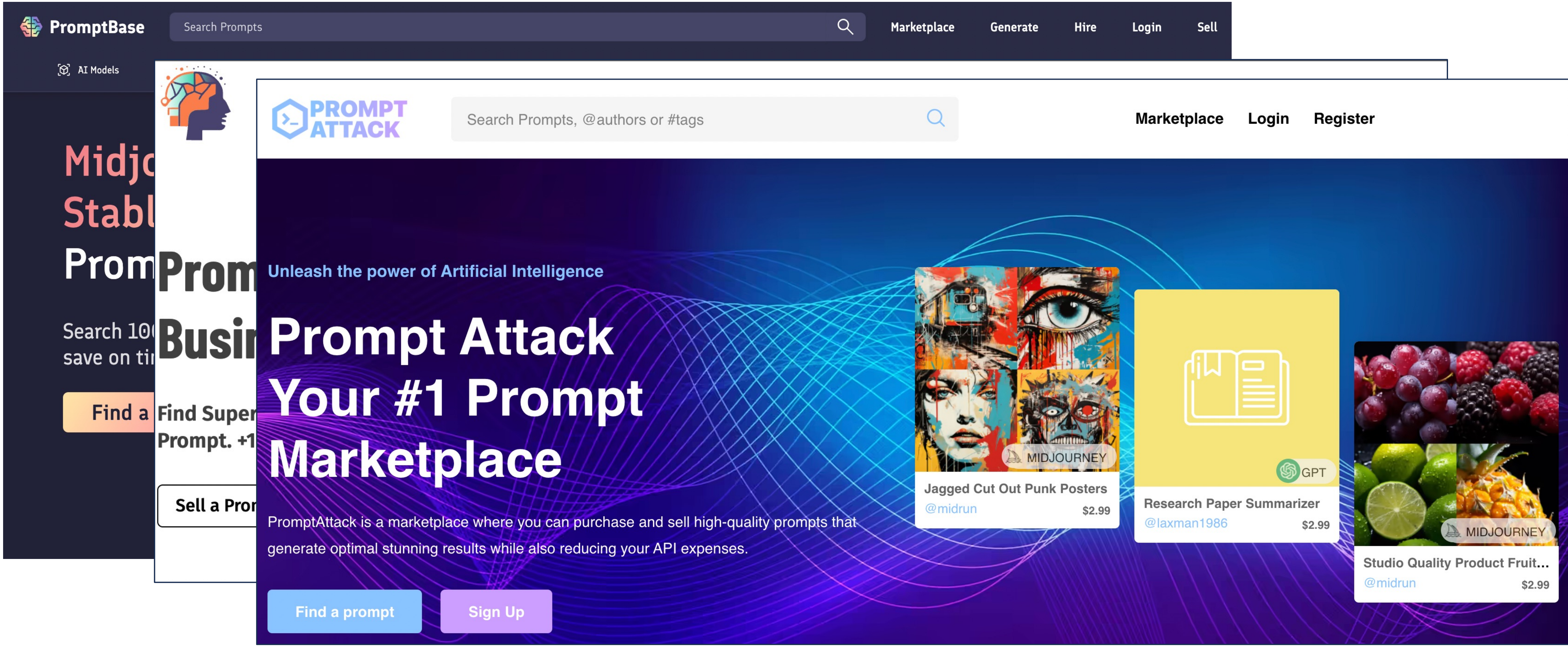


Wang et al. Images Speak in Images: A Generalist Painter for In-Context Visual Learning. CVPR 2023.

Today’s Topics

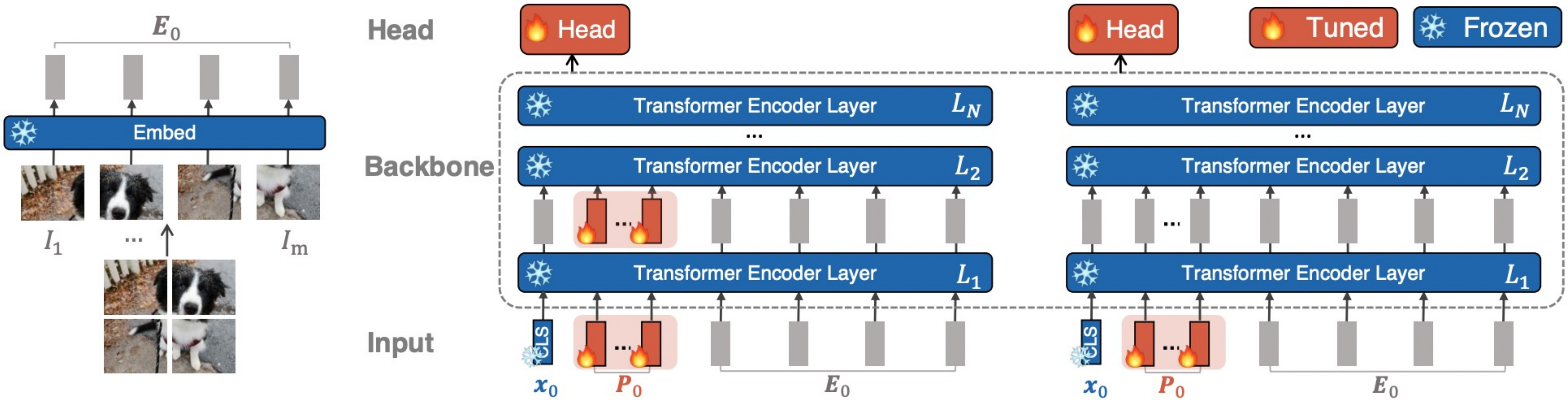
- Sampler of Today’s Popular Computer Vision Problems
- Foundation Models
- Textual Prompting & Zero-shot Learning
- Visual Prompting & In-context Few-shot Learning
- Prompt Tuning
- Discussion

Motivation



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

Idea: Replace Manually-Authored Prompts with Learnable Parameters



(a) Visual-Prompt Tuning: Deep (b) Visual-Prompt Tuning: Shallow

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

Jia et al. Visual Prompt Tuning. ECCV 2022.

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

Jia et al. Visual Prompt Tuning. ECCV 2022.

Today’s Topics

- Sampler of Today’s Popular Computer Vision Problems
- Foundation Models
- Textual Prompting & Zero-shot Learning
- Visual Prompting & In-context Few-shot Learning
- Prompt Tuning
- Discussion

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

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

- Sampler of Today's Popular Computer Vision Problems
- Foundation Models
- Textual Prompting & Zero-shot Learning
- Visual Prompting & In-context Few-shot Learning
- Prompt Tuning
- Discussion


The End