Large Vision & Language Models

Fei Pan

Research Fellow, Computer Science Engineering

University of Michigan

What Are Large Vision & Language Models?

Wrench

Multifunctional Tool

Source: United AI

Large Vision & Language Models

CLIP

GPT-4

SAM

Groundeding DINO

Classification

Segmentation

Classification Task — CLIP & GPT-4

CLIP: Connecting text and images

Source: OpenAI

Correct Predictions by CLIP

Food101

guacamole (90.1%) Ranked **1** out of 101 labels

- ✓ 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.

Youtube-BB

airplane, person (89.0%) Ranked 1 out of 23 labels

- ✓ a photo of a **airplane**.
- × a photo of a **bird**.
- × a photo of a **bear**.
- × a photo of a **giraffe**.
- × a photo of a **car**.

Wrong Predictions by CLIP

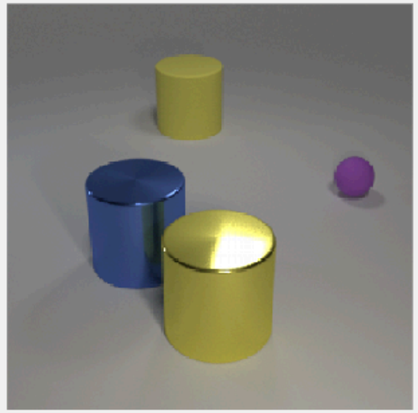
ImageNet-A (Adversarial)

lynx (47.9%) Ranked 5 out of 200 labels

- × 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**.

CLEVR Count

4 (75.0%) Ranked 2 out of 8 labels



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

Source: OpenAI

Correct Predictions by CLIP

SUN397

television studio (90.2%) Ranked 1 out of 397 labels

✓ 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.**

UCF101

Volleyball Spiking (99.3%) Ranked 1 out of 101 labels

✓ a photo of a person **volleyball spiking.**

× a photo of a person **jump rope.**

× a photo of a person **long jump.**

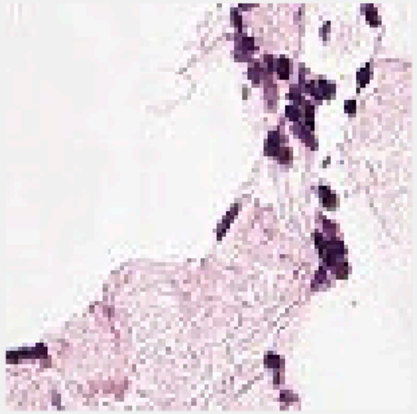
× a photo of a person **soccer penalty.**

× a photo of a person **table tennis shot.**

Wrong Predictions by CLIP

PatchCamelyon (PCam)

healthy lymph node tissue (77.2%) Ranked 2 out of 2 labels



× this is a photo of **lymph node tumor tissue**

✓ this is a photo of **healthy lymph node tissue**

Facial Emotion Recognition 2013 (FER2013)

angry (49.0%) Ranked 5 out of 7 labels

× a photo of a **happy** looking face.

× a photo of a **neutral** looking face.

× a photo of a **surprised** looking face.

× a photo of a **fearful** looking face.

✓ a photo of a **angry** looking face.

"A Photo of {Label}…."

Various Datasets & Objects

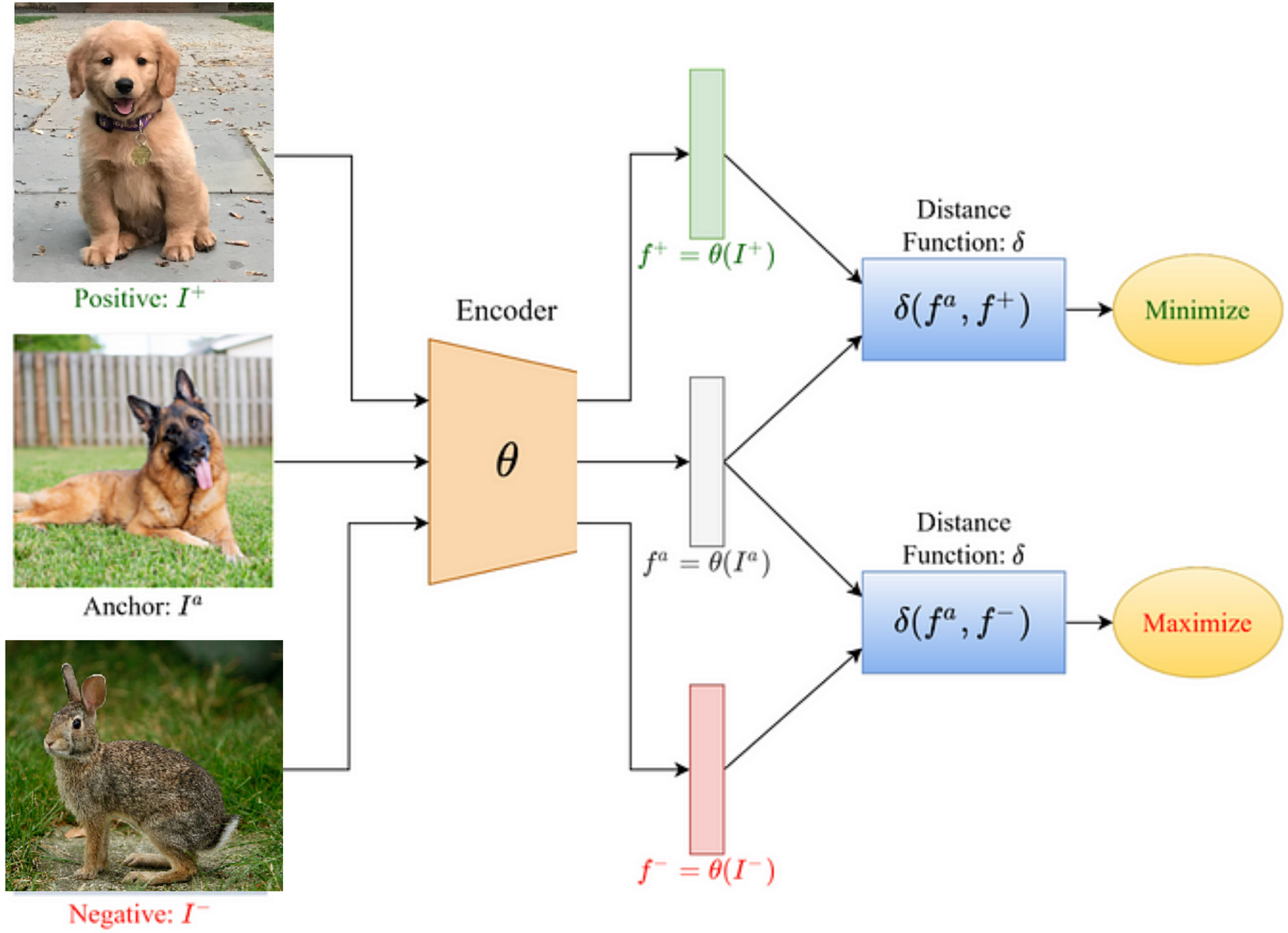
Source: OpenAI

What is CLIP

- Stands for Contrastive Language-Image Pre-training
- Requires huge amounts of images and their corresponding captions
- Is a strong zero-shot classifier

What is Contrastive Learning?

Contrastive learning is a self-supervised learning technique to train a model by pulling together similar pairs and pushing apart dissimilar ones.



Zero-shot learning is to let a model recognize correctly a class it has **never encountered** before during training (unknown class).

Training Classes

Tiger

Wolf

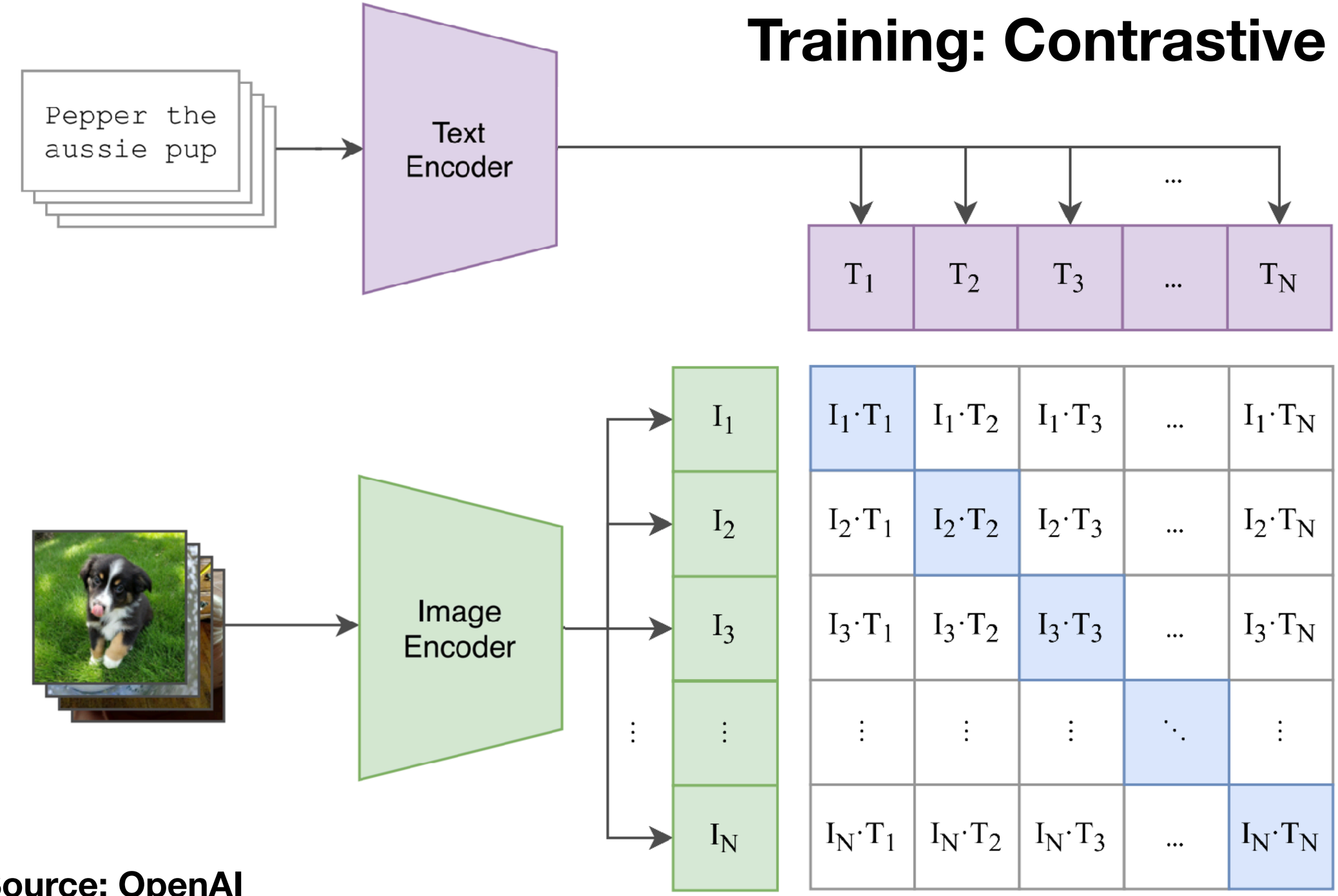
Raccoon

Leopard

Unknown Classes

Panda

Training: Contrastive Learning

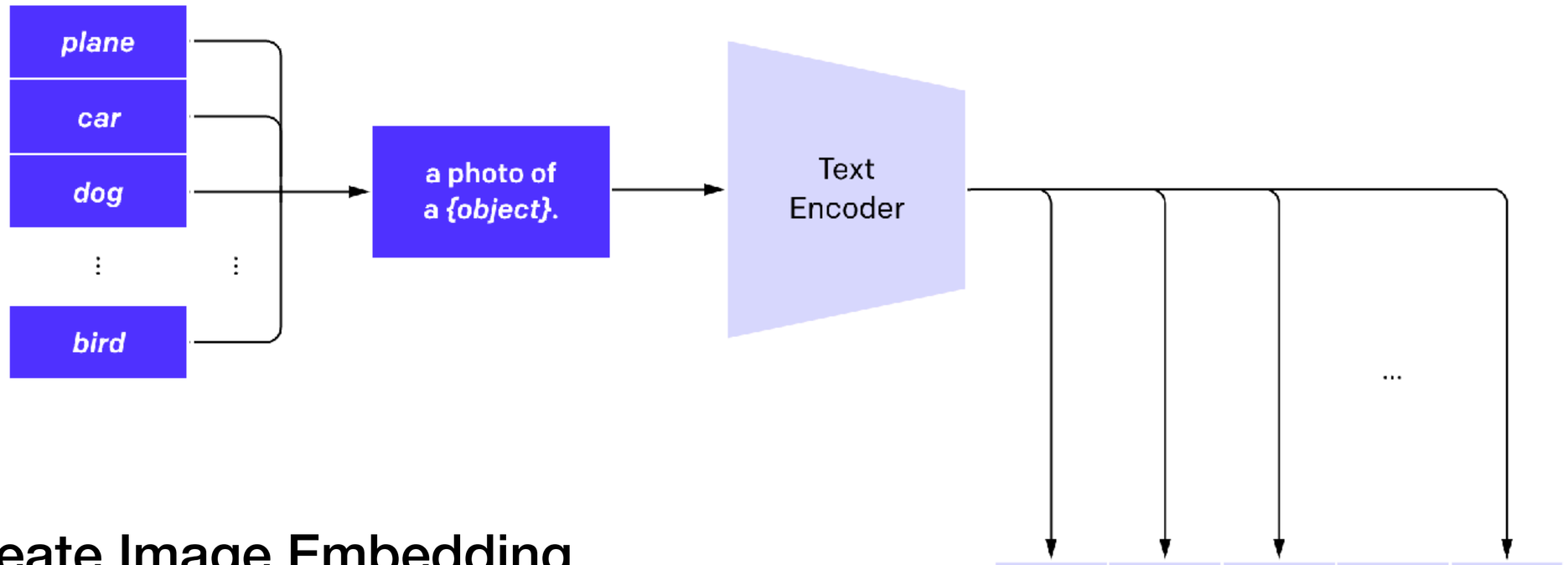


Contrastive Loss

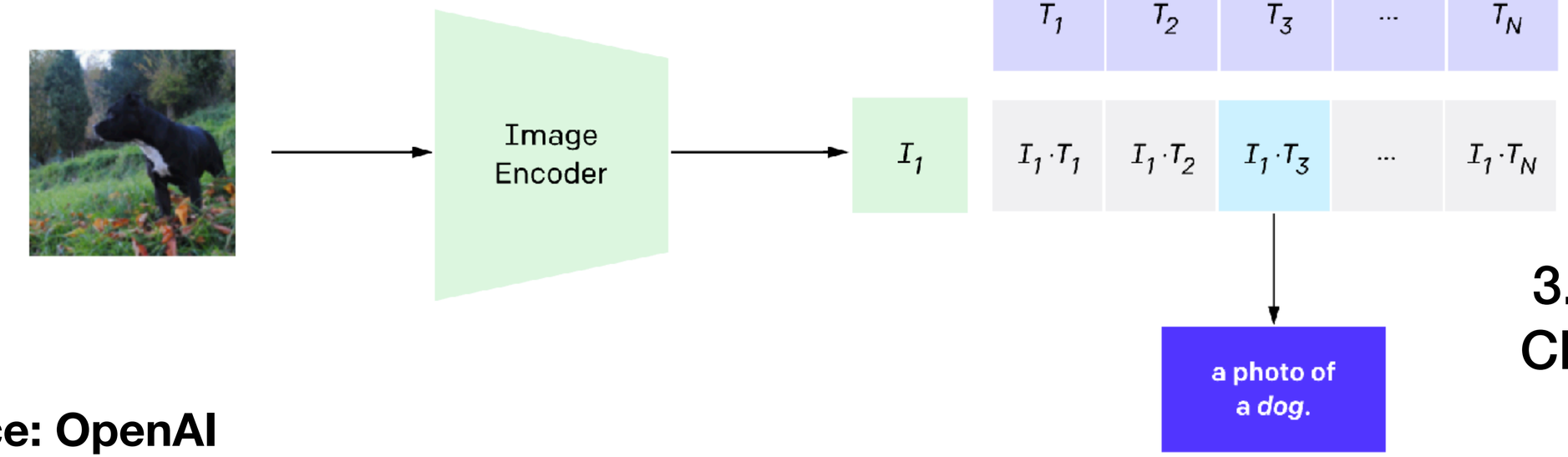
Source: OpenAI

Testing

1. Create Text Embeddings for Class Names



2. Create Image Embedding



3. Zero-Shot Classification

Source: OpenAI

CLIP Training Details

- Existing datasets are “small”
  - Visual Genome and MSCOCO ~ 100,000
  - YFCC100M “high quality” images ~ 15M
- CLIP is trained on 400M <Image, Text> pairs constructed from a variety of publicly available sources
- CLIP requires very little data augmentations. Only applied a random square crop
- Temperature parameter (scalar) for SoftMax logits is learned automatically
- The text encoder is adopted from transformer
- The image encoder is either a modified ConvNet or visual transformer

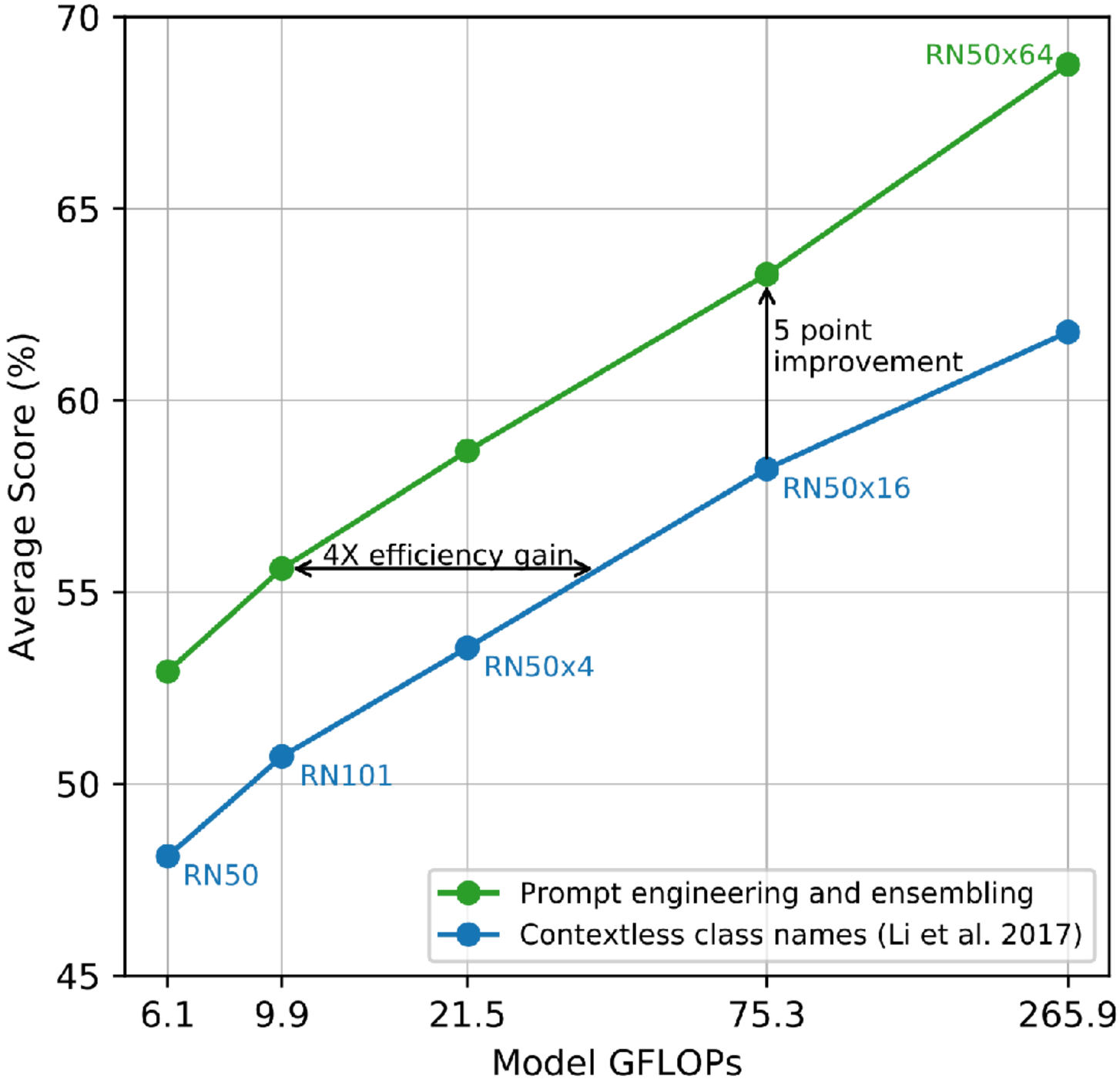
Prompt Engineering

- Why making “a photo of {label}” for CLIP model instead of directly giving a class name/ vocabulary to it?
  - A performance increase of 1.3% using this prompt.

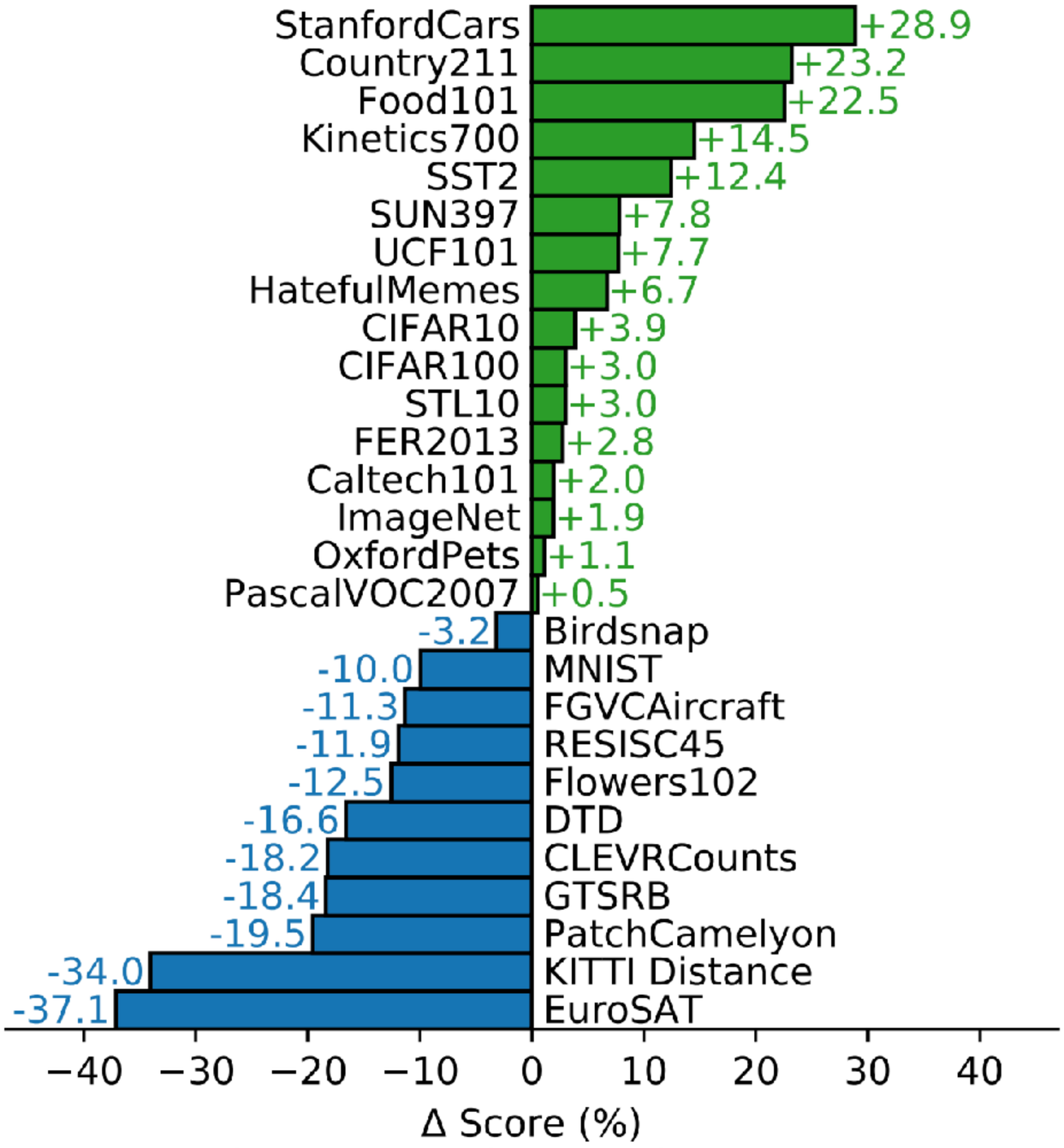
- For domain specific tasks/datasets (Oxford-IIIT Pets), slightly more specific prompts work better
  - “A photo of {label}, a type of pet”

- The ensemble of prompts also work well
  - “A photo of a big {label}” & “A photo of a small {label}”

Prompt engineering and ensembling obtain extra performance gain compared to contextless class names.



Source: OpenAI

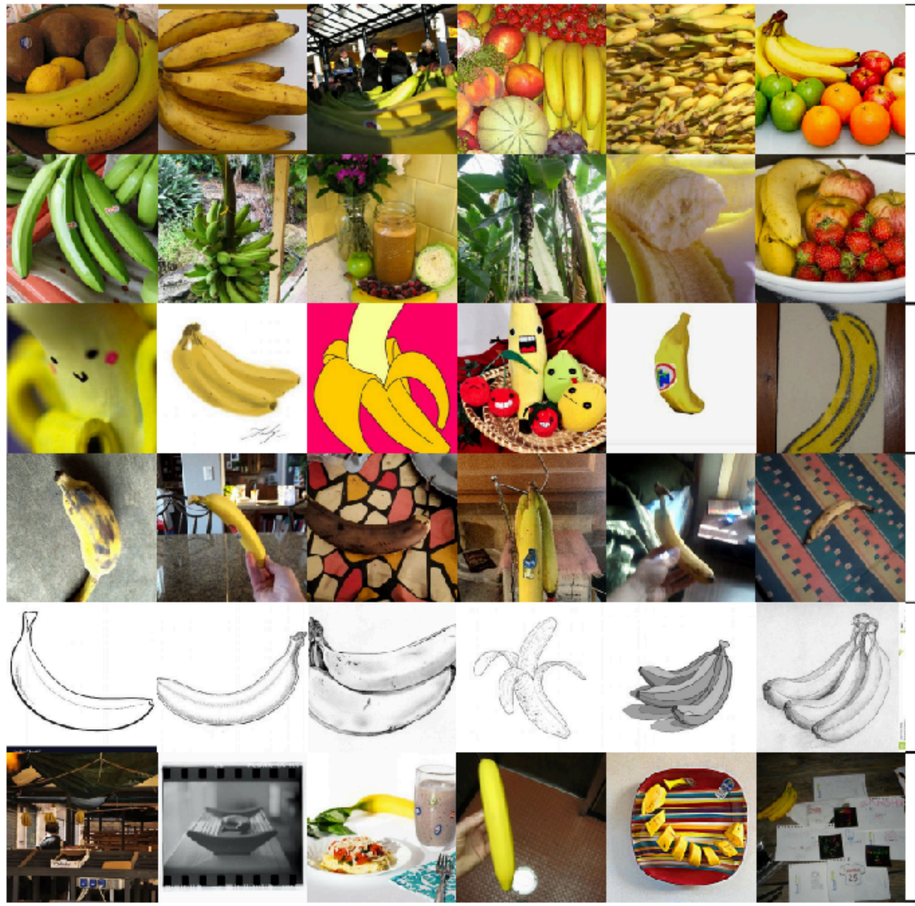


Zero-Shot CLIP vs. Linear Probe on ResNet50

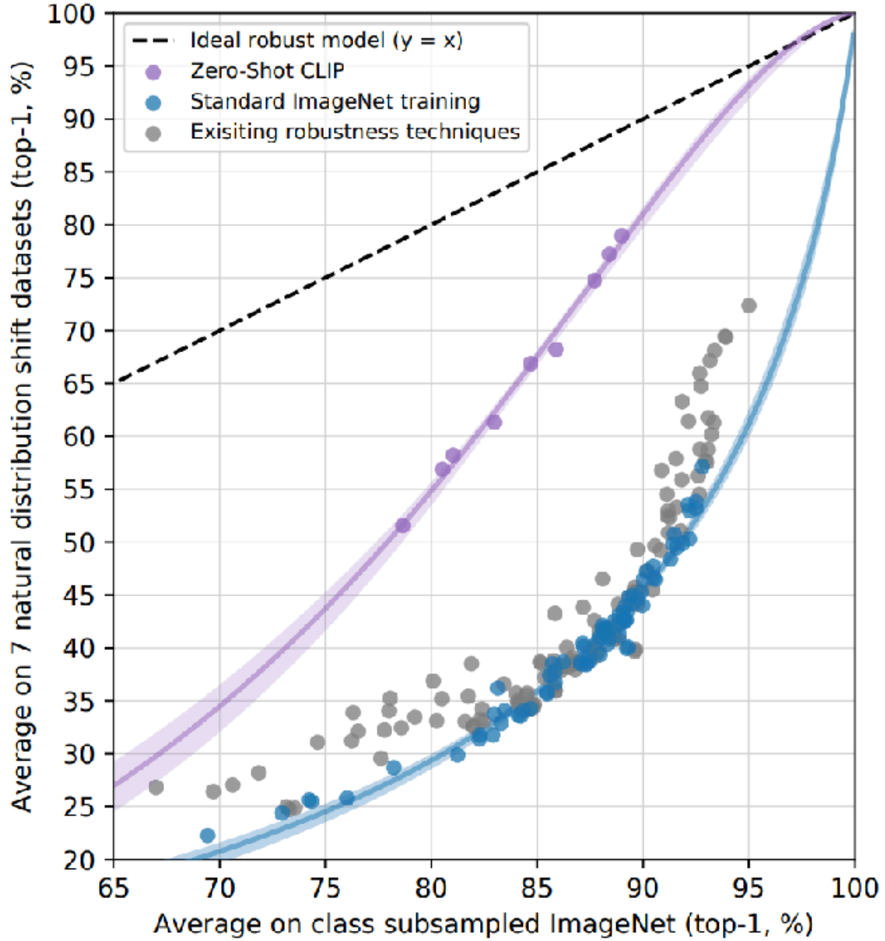
CLIP is competitive w/ fully supervised baseline for zero-shot classification

Source: OpenAI

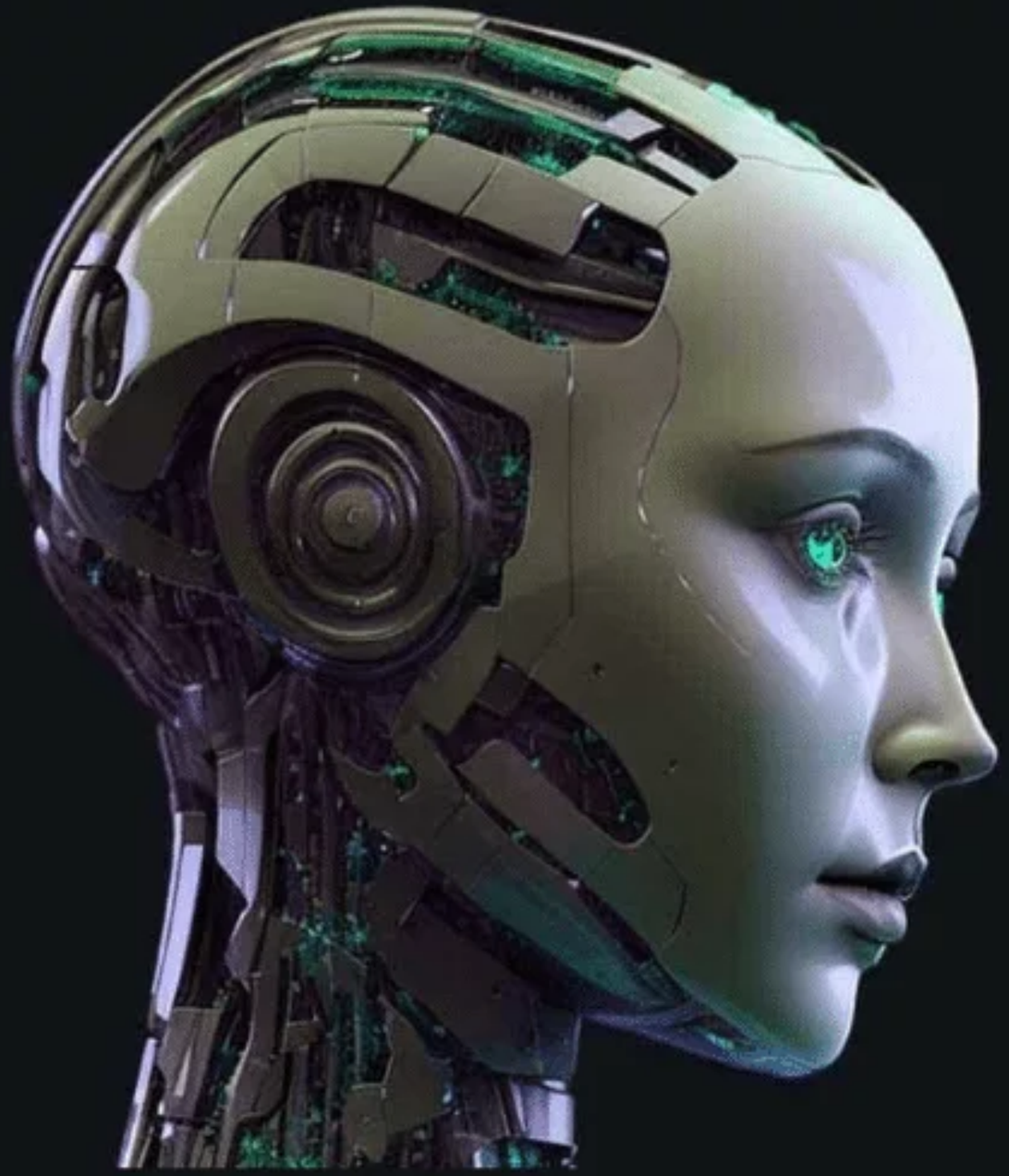
CLIP is much more robust to distribution shift than supervised models pre-trained on ImageNet



	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%

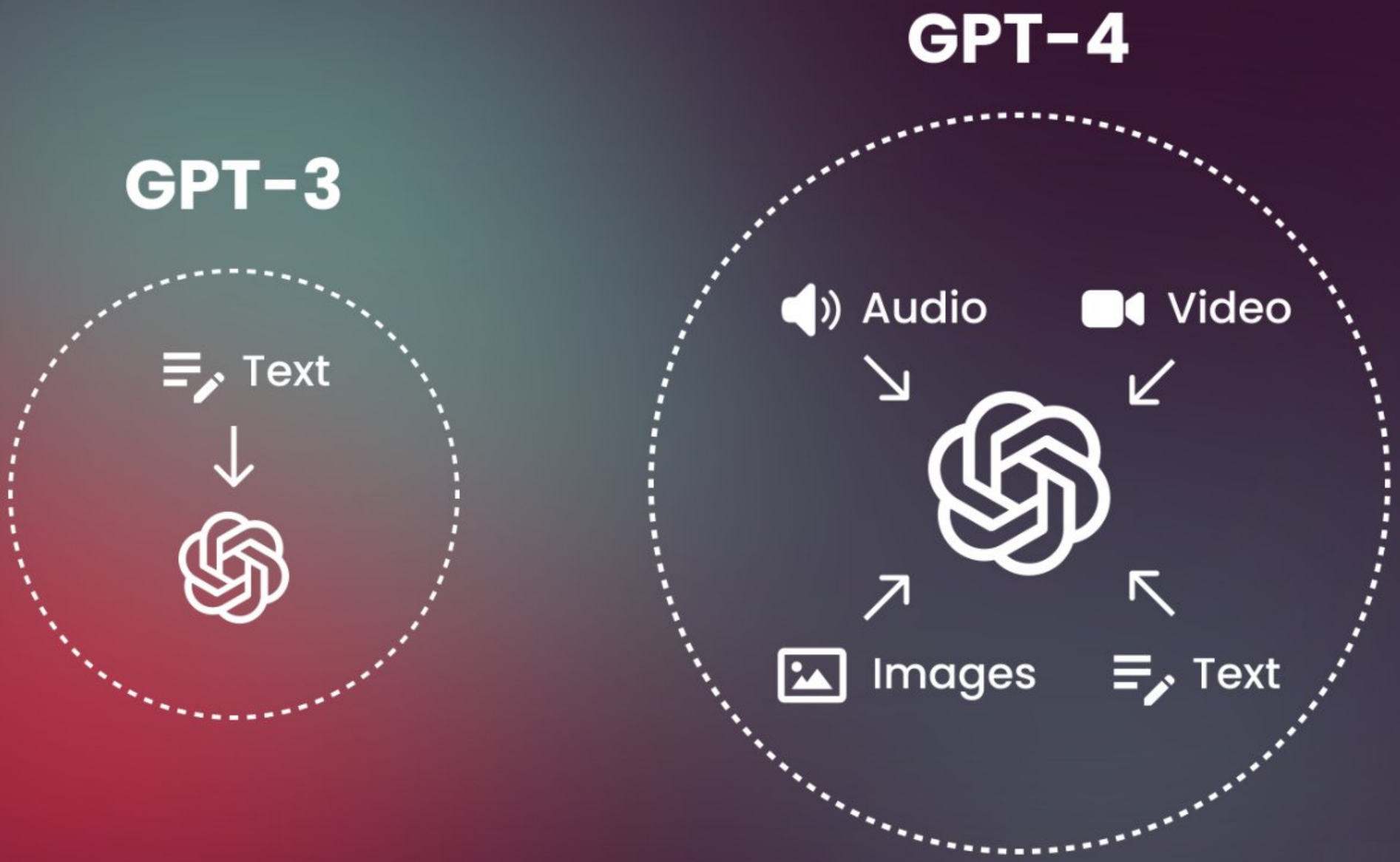


Source: OpenAI



OpenAI

GPT-4

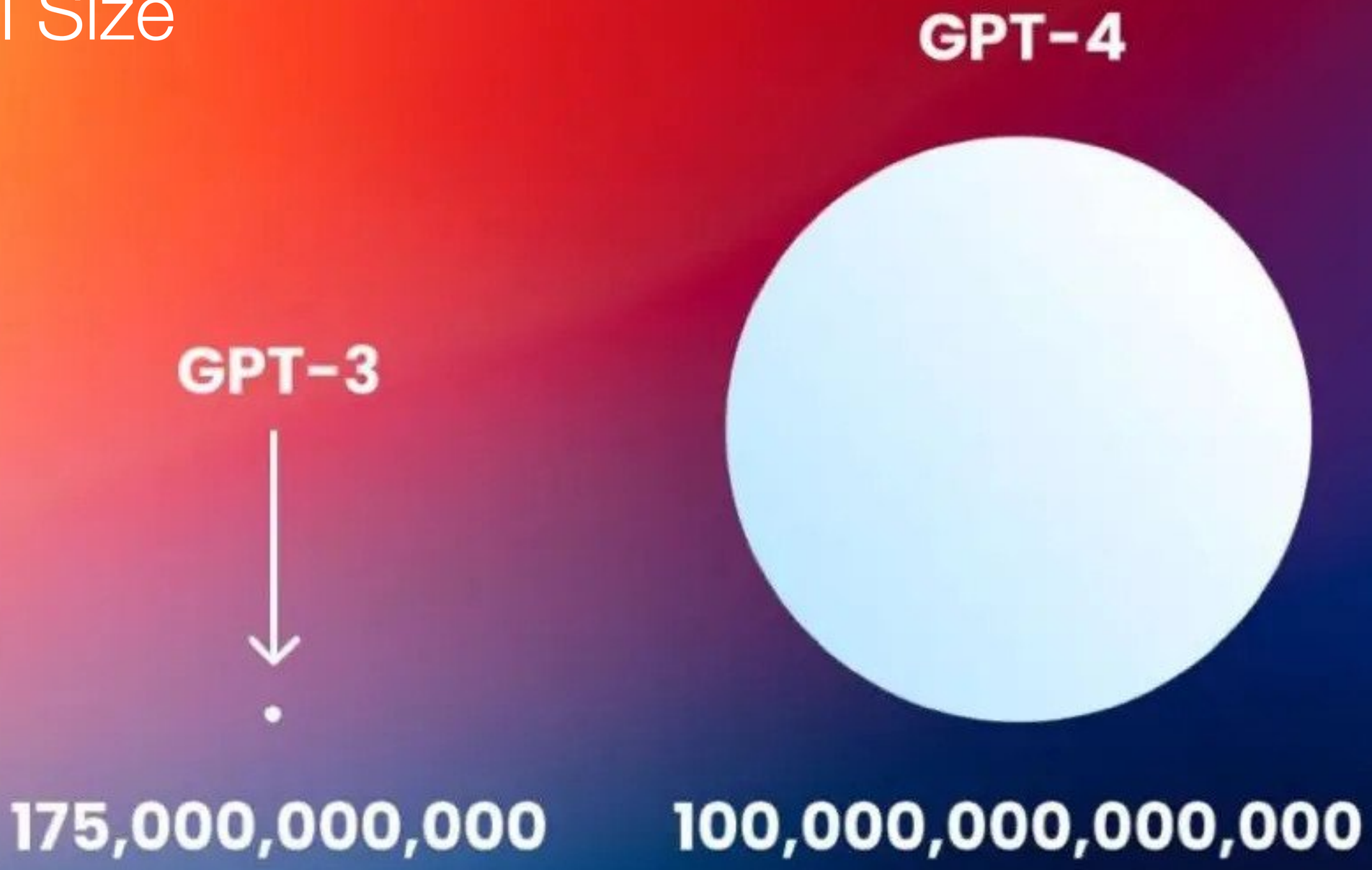


Source: X

What would happen if the strings were cut?

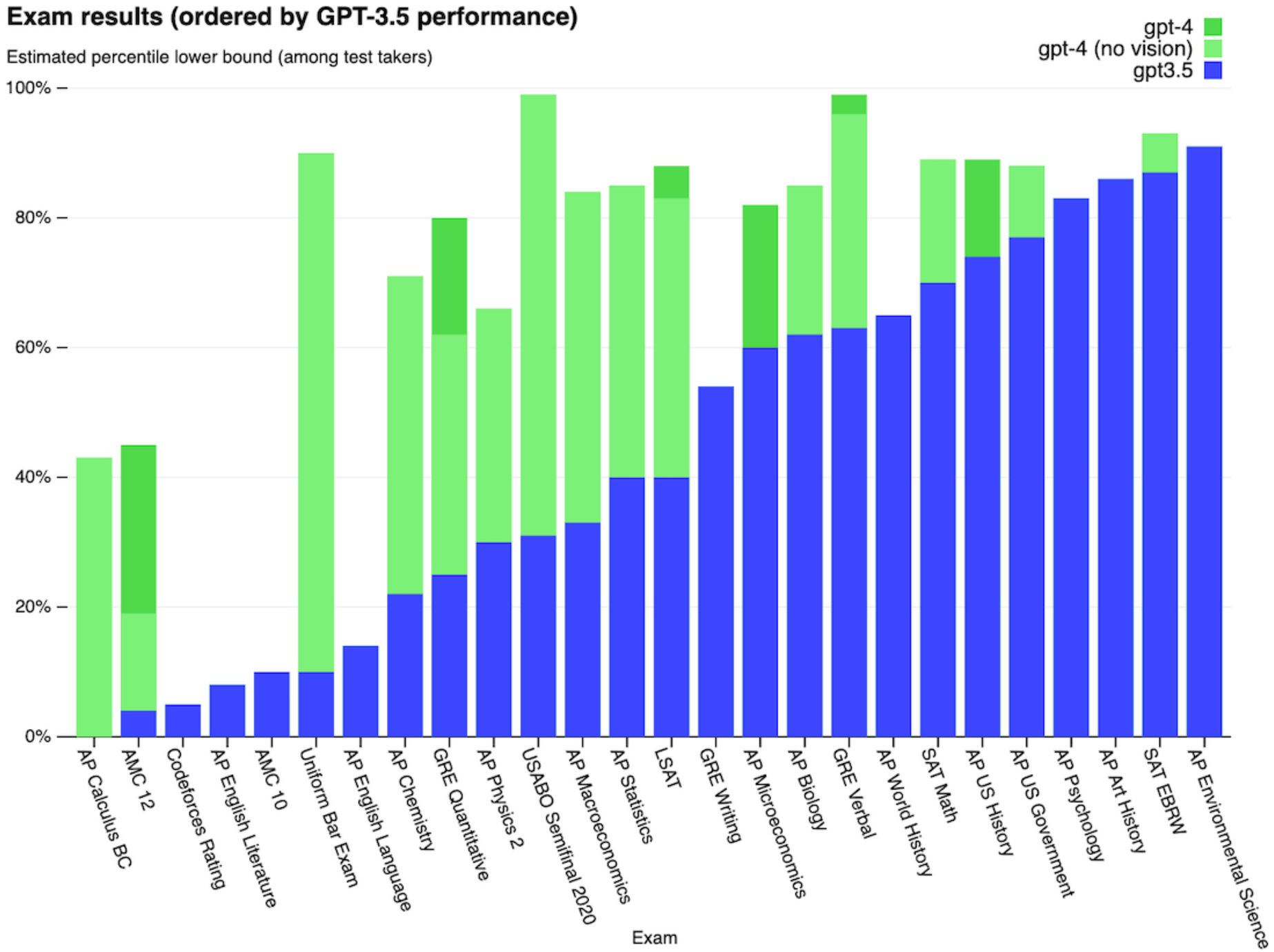
Source: OpenAI

Model Size



Source: X

Passing Exams



Source: OpenAI

What is funny about this photo?

Source: hmmm (reddit)

User What is funny about this image? Describe it panel by panel.

Source: hmmm (Reddit)

Understand jokes

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.

Source: https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg

What is unusual about this photo?

Understand unusual behavior

User What is unusual about this image?

Source: https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg

GPT-4 The unusual thing about this image is that a man is ironing clothes on an ironing board attached to the roof of a moving taxi.

Source: OpenAI

GPT-4 Classification:

Prompt Engineering

can you tell me which roof type of this house among hip, gable, and flat (choose only one type) and explain the reasons.

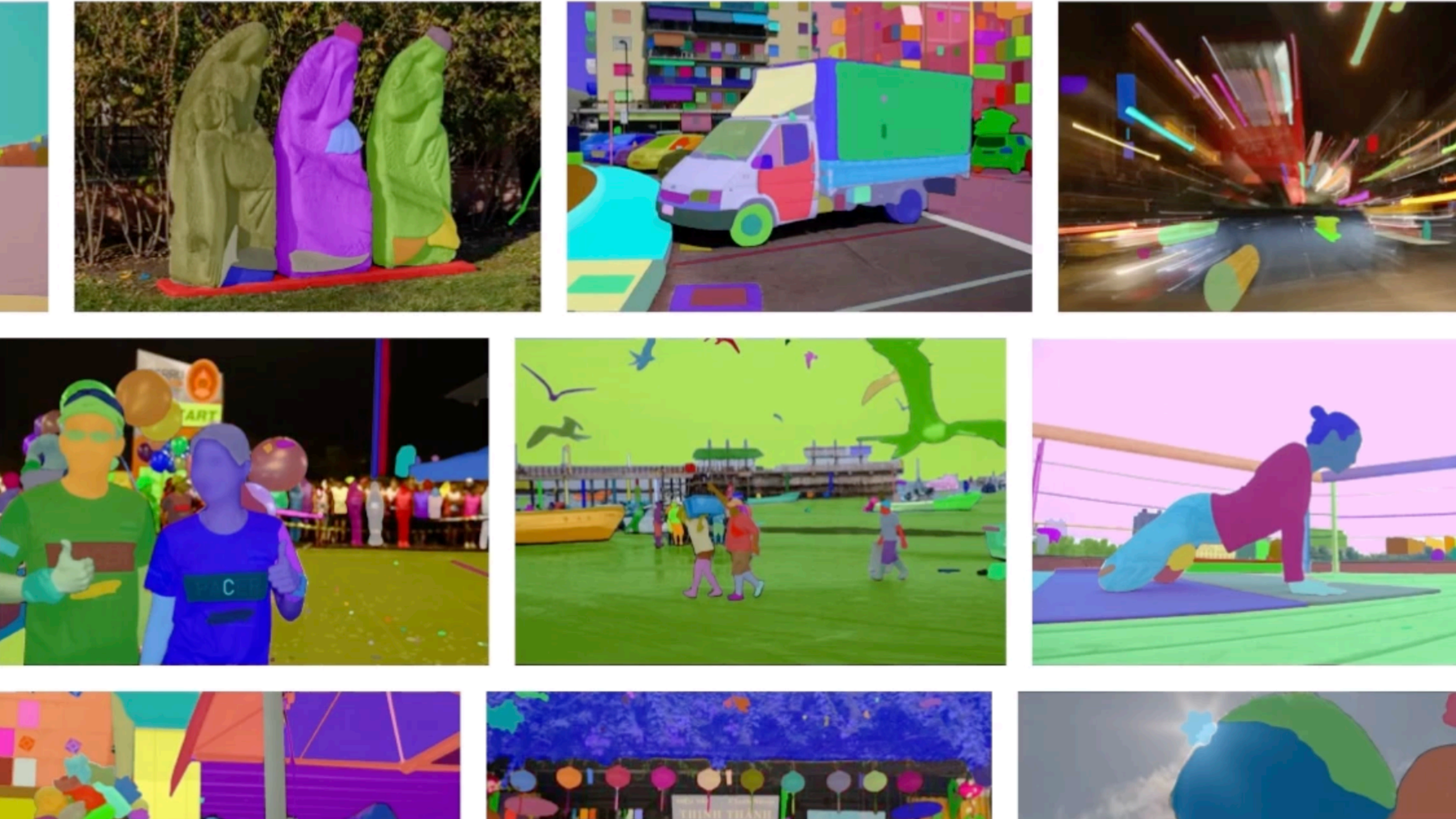
The roof type of the house in the image is a flat roof. This is identified by the following characteristics:

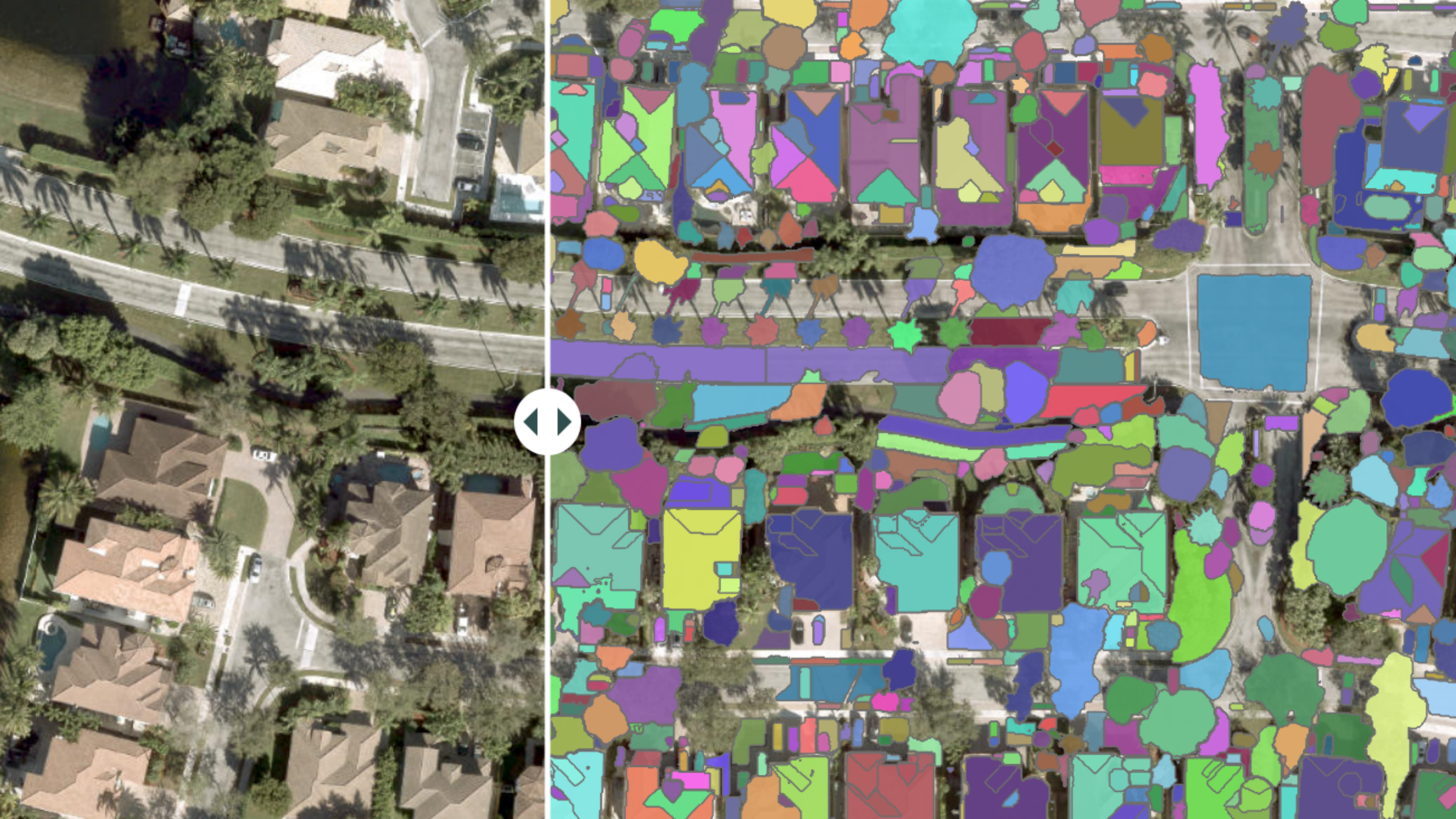
- The roof appears to have no visible slope; it looks horizontal or nearly horizontal.
- There are no ridges, hips, gables, or valleys as one would find in hip or gable roofs.
- The roof's outline is a simple rectangle with no changes in geometry that would suggest a slope.

Flat roofs are common in arid climates and on commercial or industrial buildings, but they can also be found on residential structures. They are known for a modern appearance and can be designed to accommodate green roofs, rooftop patios, or for the installation of solar panels. However, they require careful waterproofing to prevent leaks.

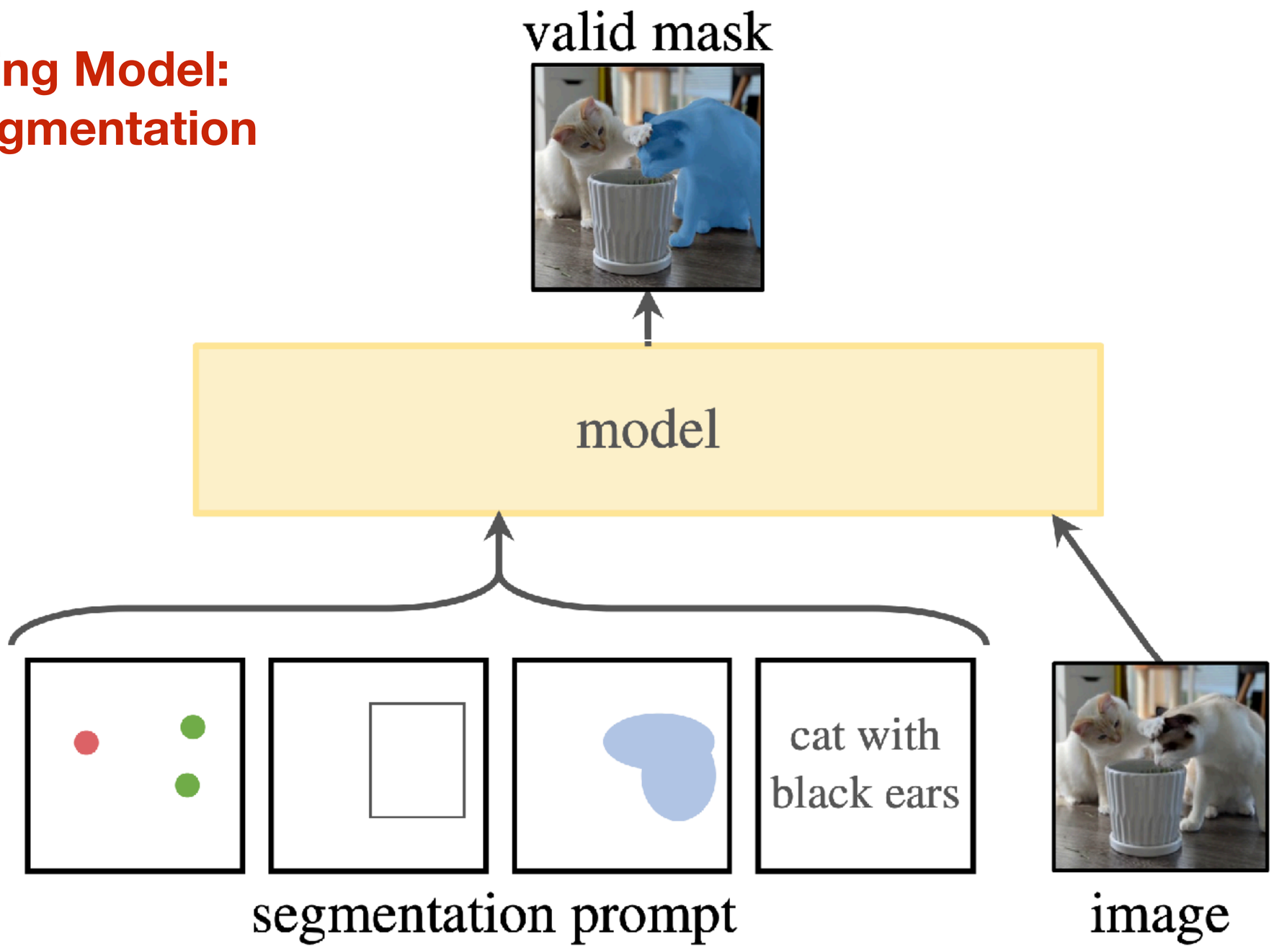
Segmentation Task — SAM & GroundingDINO

Segment Anything Model (SAM)



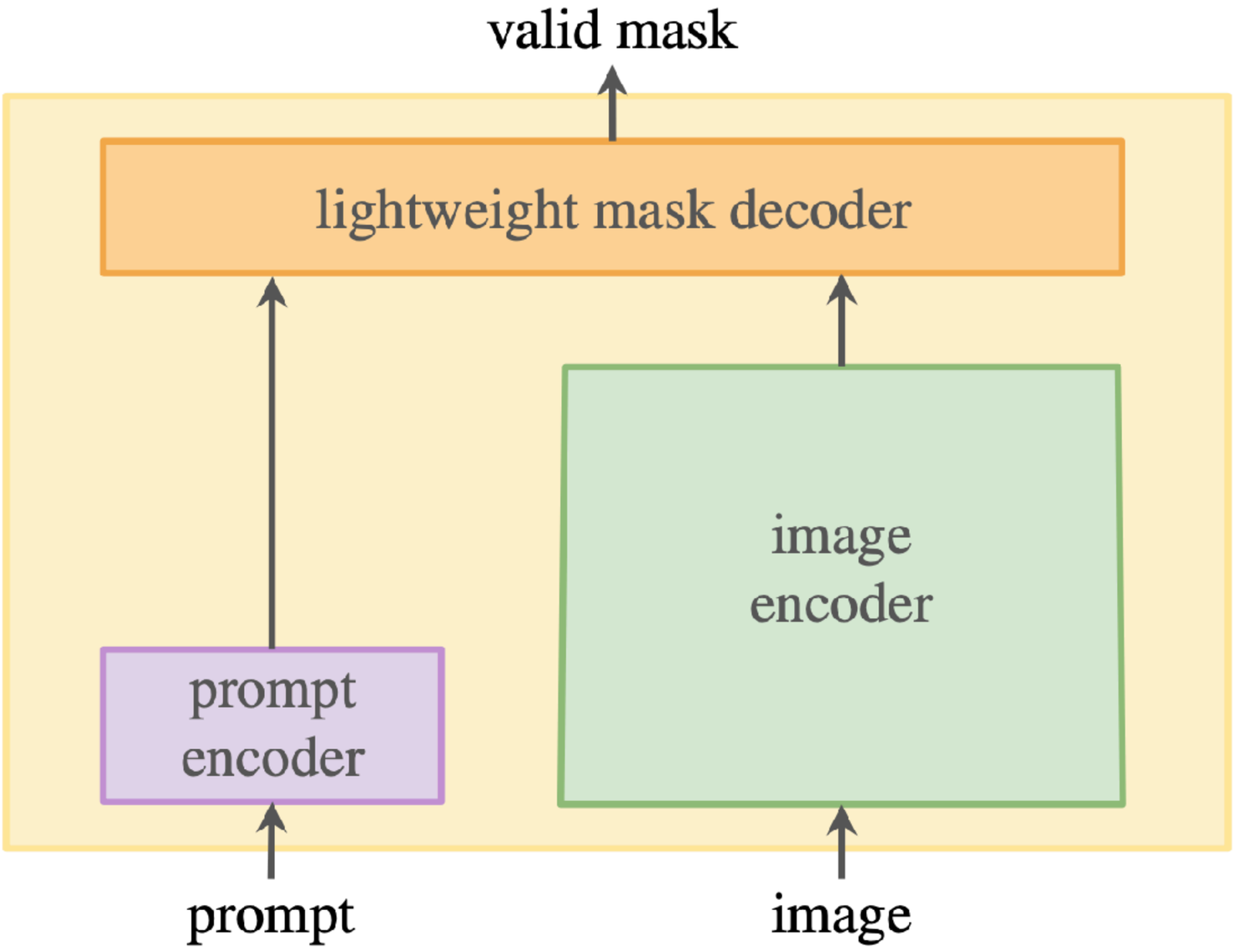


Segment Anything Model: A Promptable Segmentation



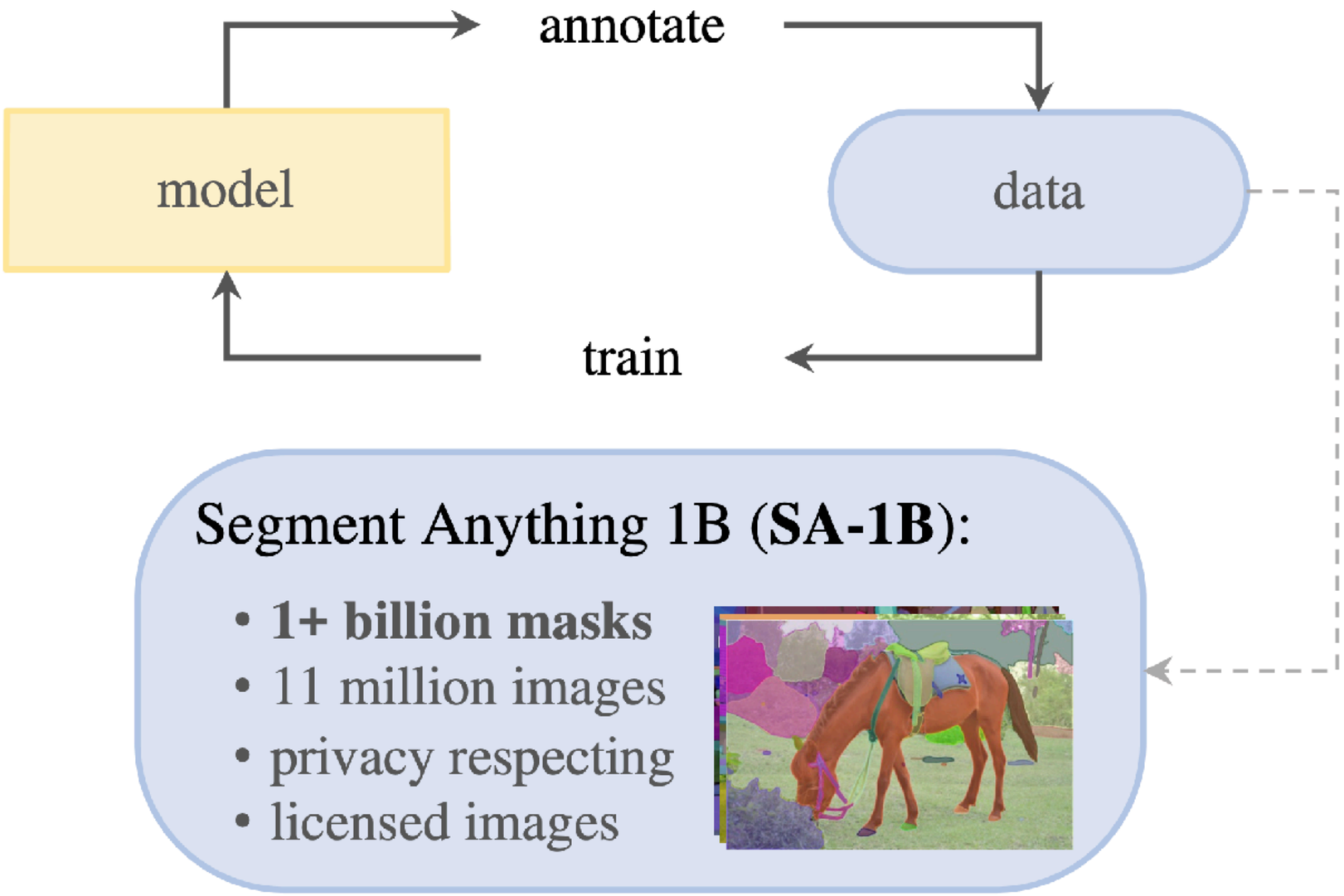
Source: SAM

Model Architecture



Source: SAM

Huge Amount of Training Data



Source: SAM

SA-1B dataset

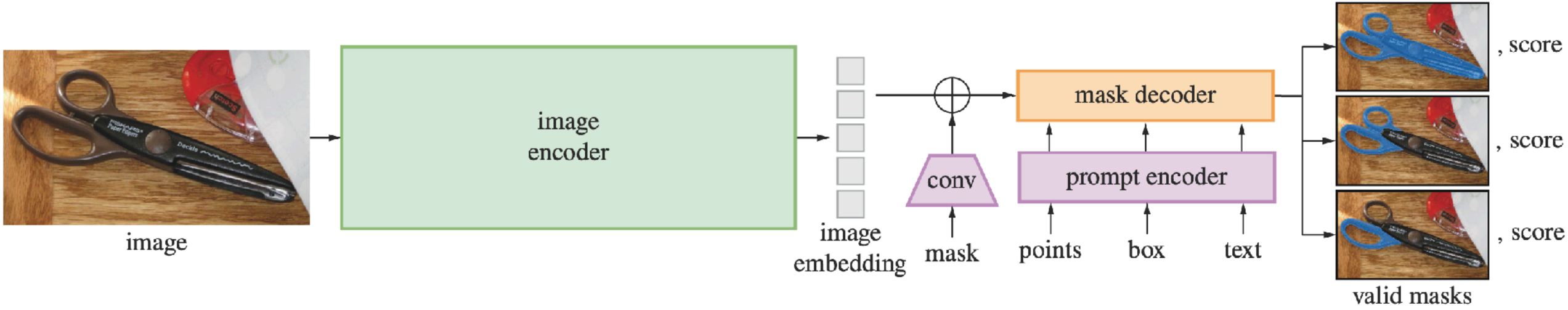
11M images

1B masks

100～ masks/img

Source: SAM

Segment Anything Model (SAM) Overview

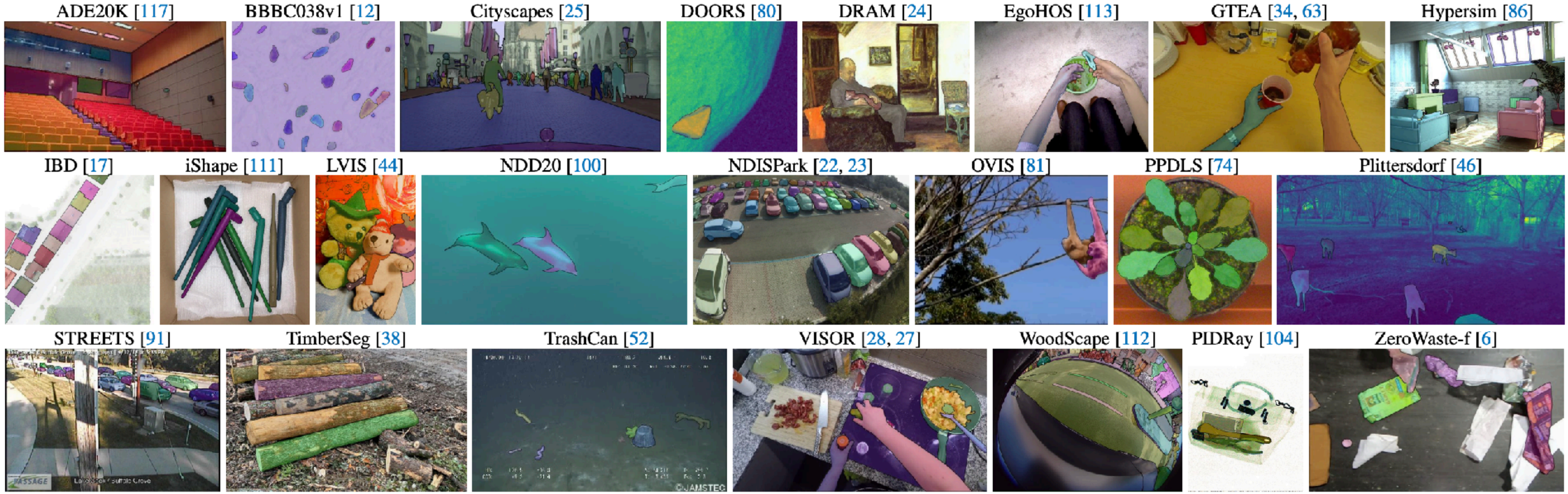


Source: SAM

Architecture Details

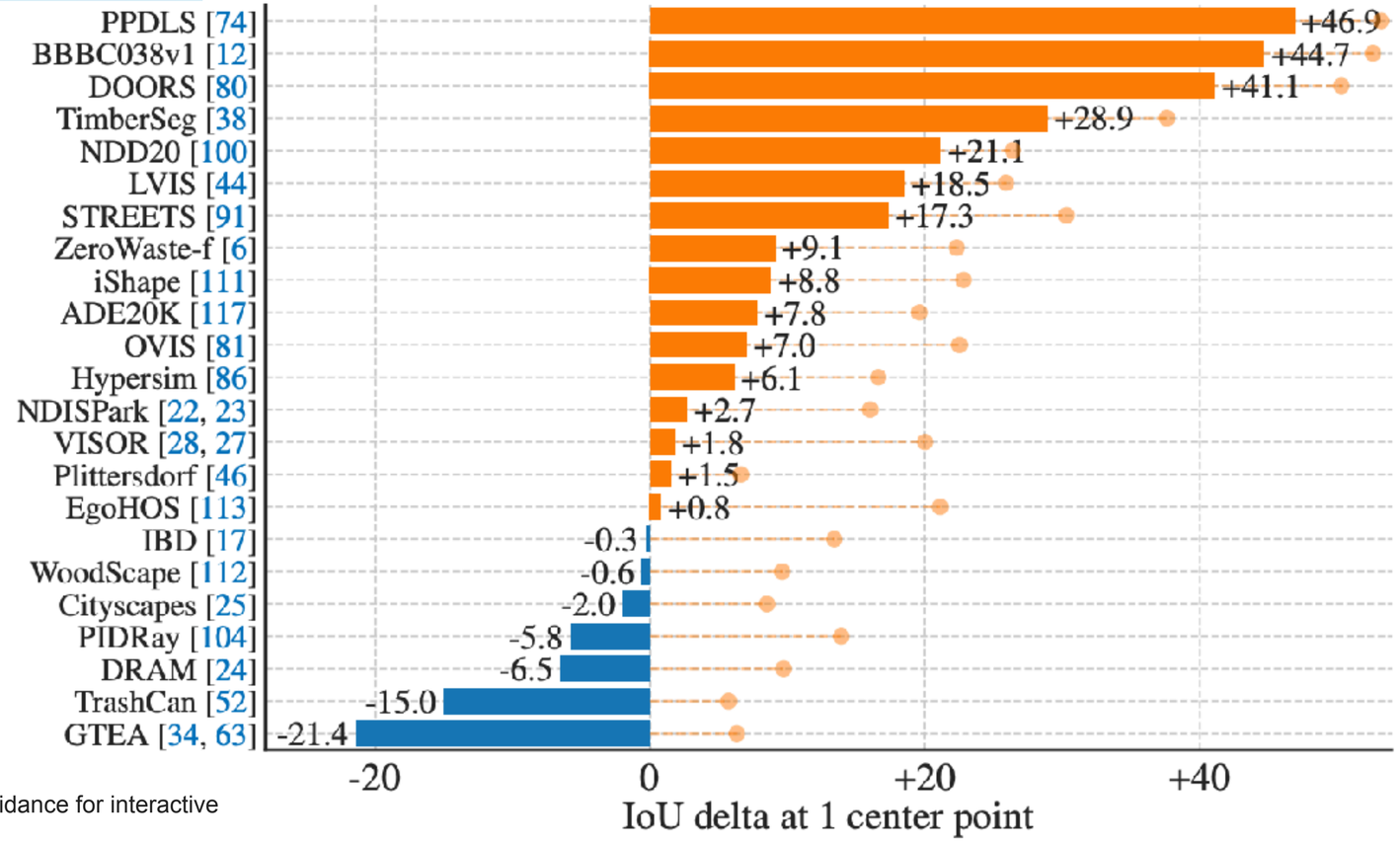
- Image encoder: adopted from MAE pre-trained ViT network
- Prompt encoder: two sets of prompts are considered: sparse prompts (boxes, points, texts) and dense prompts (masks)
  - For sparse prompts, boxes & points are embedded using positional embeddings, texts using CLIP text encoder. Dense prompts such as masks are embedded using conv layers.
- Mask decoder: map the image embedding and prompt embeddings
- Resolving ambiguity: let model predict 3 output masks for a single prompt
- Loss: focal loss and dice loss for supervised training

SAM's Zero-Shot Segmentation over 23 Diverse Datasets



Source: SAM

Mean IoU of SAM and RITM (the strongest single point segmenter)



RITM: Sofiiuk, Konstantin, Ilya A. Petrov, and Anton Konushin. "Reviving iterative training with mask guidance for interactive segmentation." *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2022.

SAM is a zero-shot edge detector

Source: SAM

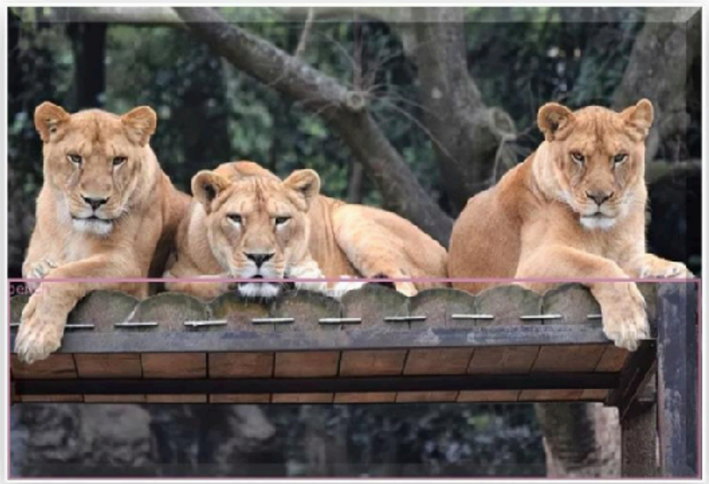
On BSDS500 Dataset

Grounding DINO: Grounded Pre-Training for Open-Set Object Detection

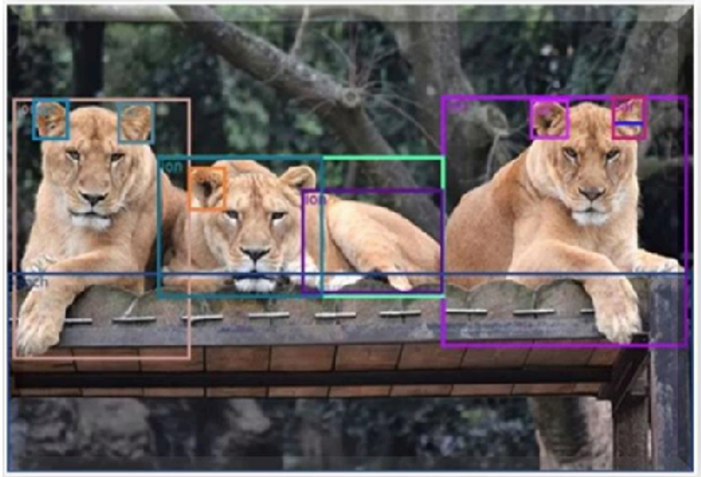
Close-Set Detection

Open-Set Detection

Trained on a fixed set of class



bench



ear, lion, bench

Adding new classes are expensive & time consuming

person

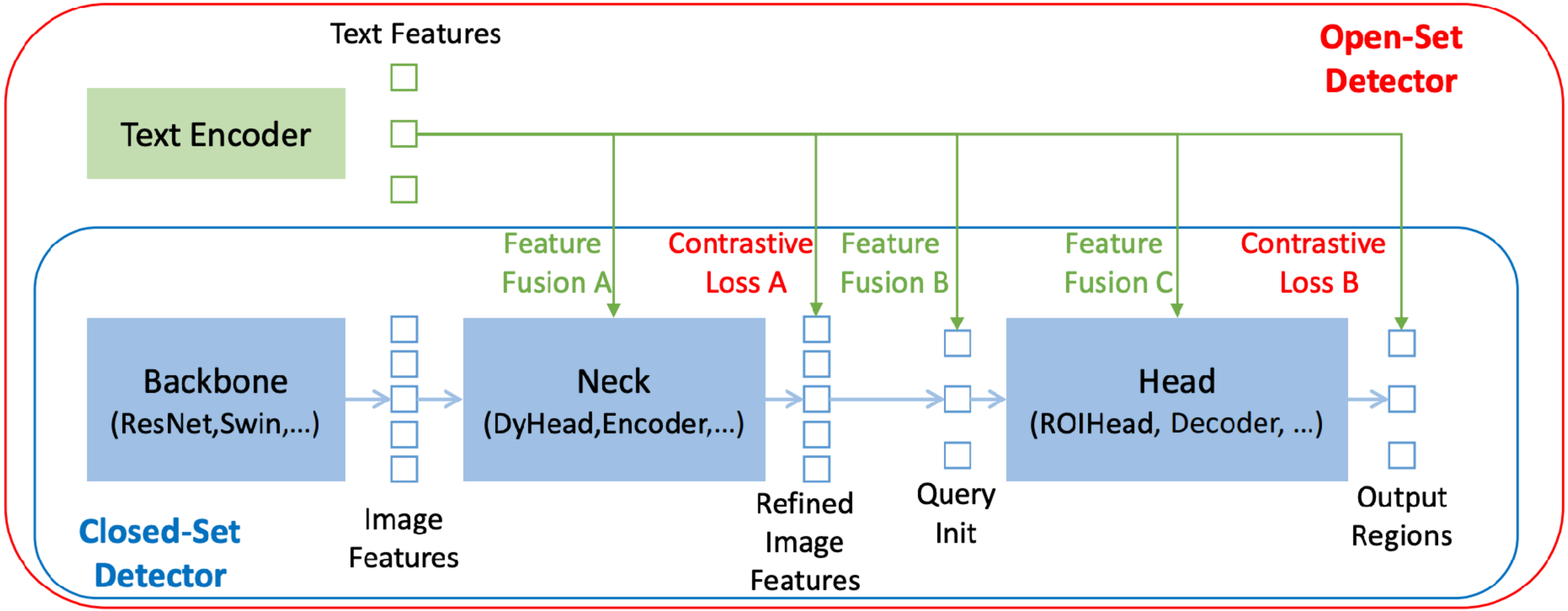
Standard Object Detection

worldcup

Zero-Shot Transfer to Novel Categories

Source: Grounding DINO

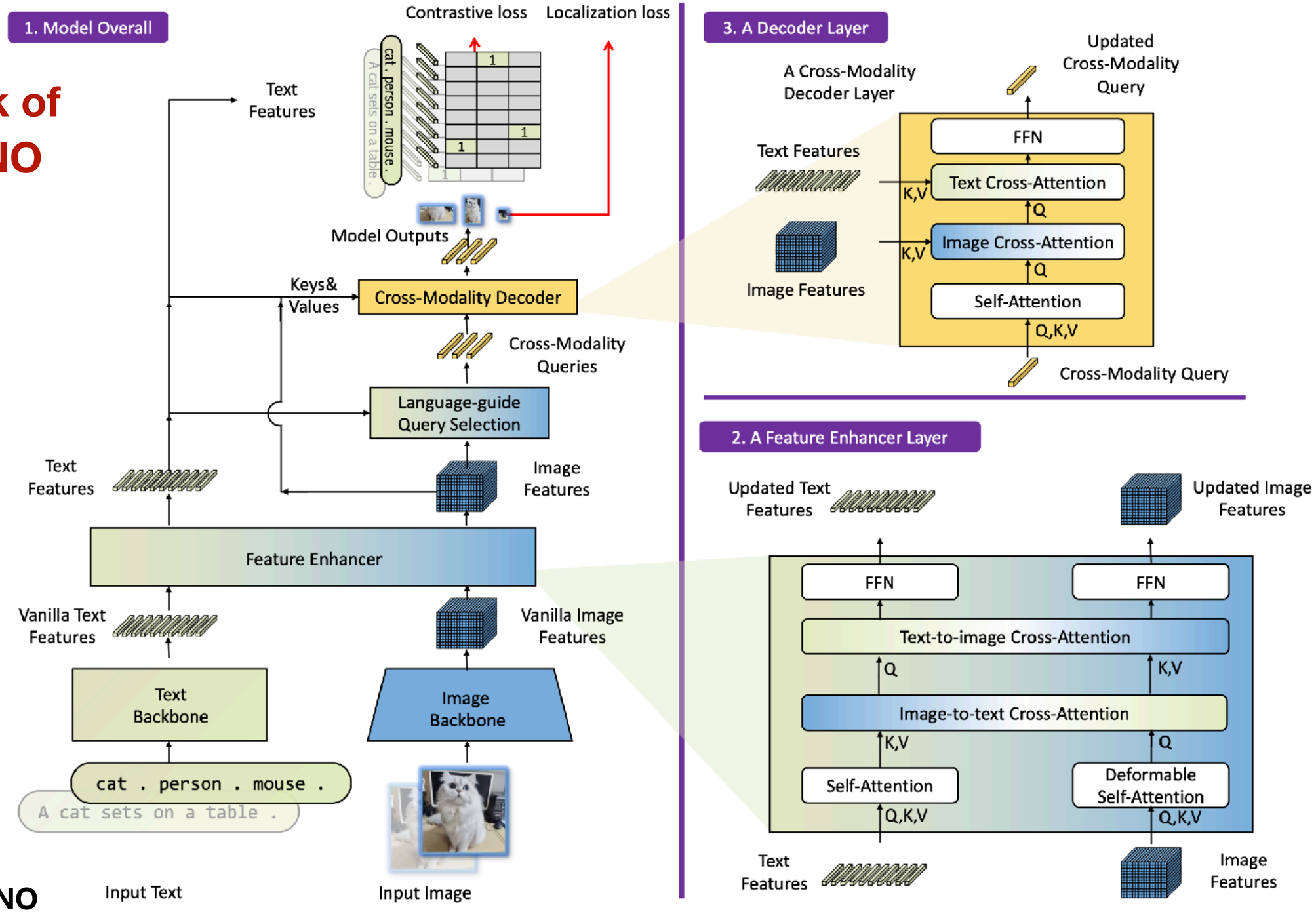
Extending closed-set detector to open-set detector



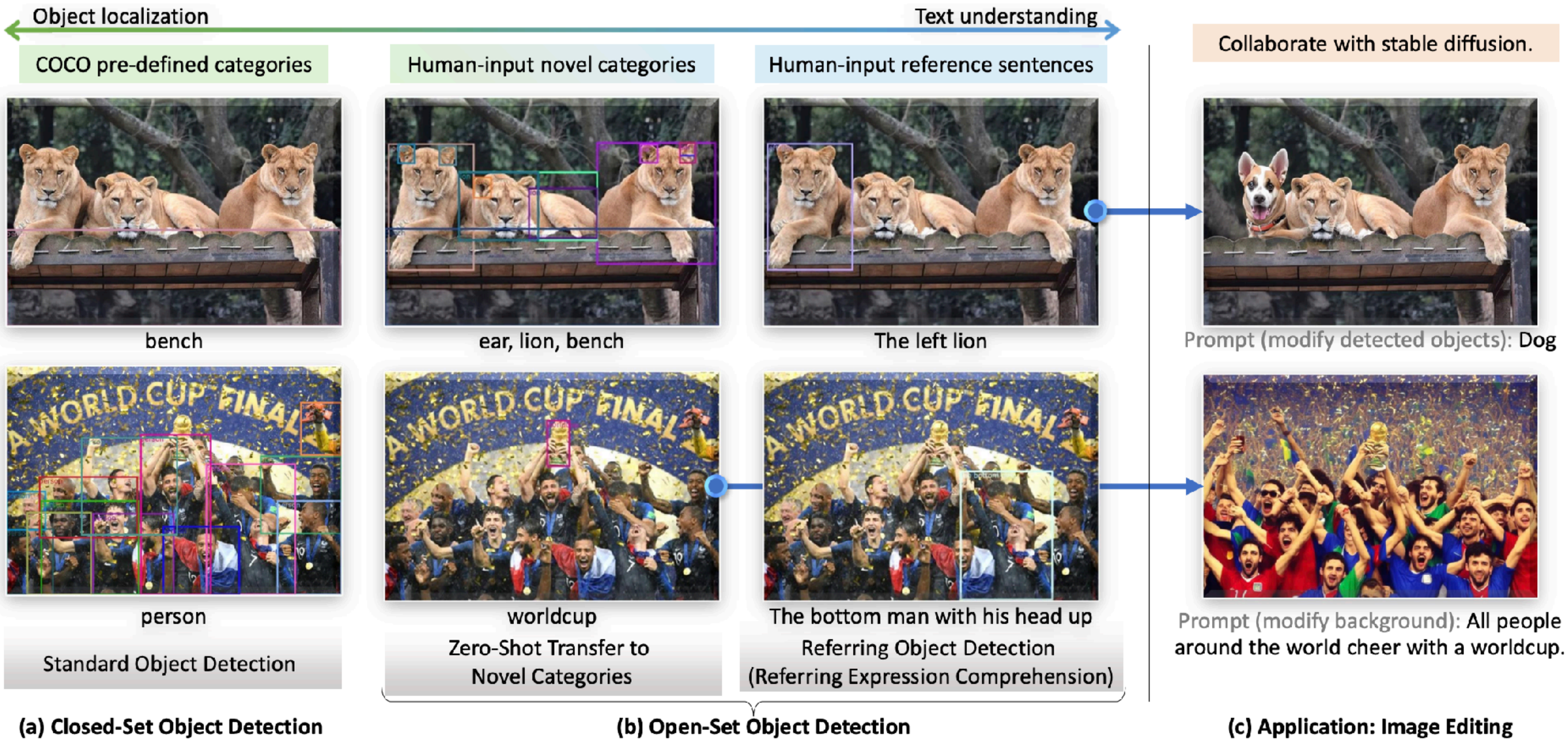
Source: Grounding DINO

The Framework of Grounding DINO

Large-ScaleLabeled Data: Image, Captions, Bounding Boxes

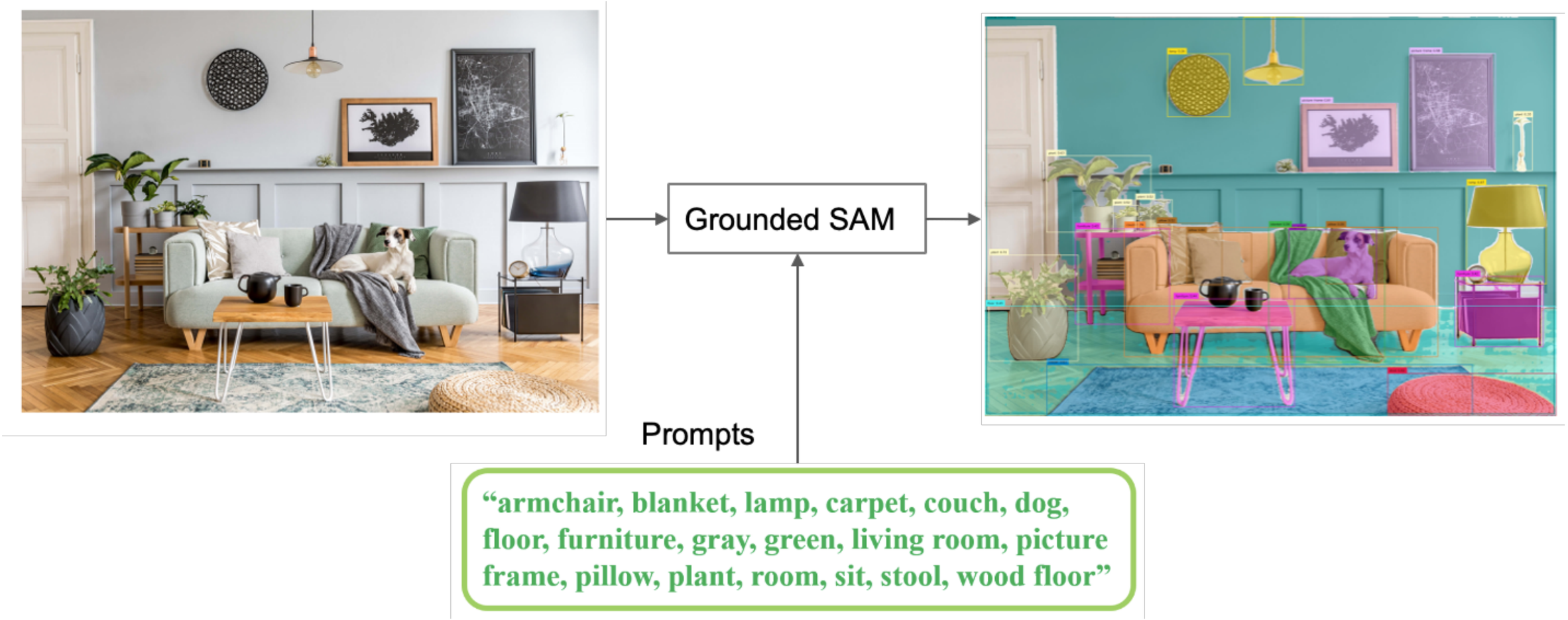


Source: Grounding DINO



Source: Grounding DINO

GroundingDINO + SAM = Grounded SAM



Source: GroundedSAM

The Workflow of Grounded SAM

Prompts: facade, roof, window, door, terrain, tree, road

Input Image

Grounding DINO

Object Detection

SAM

Object Segmentation

trees, road, terrain, windows, door, facade, roof