



Visual Intelligence <> SPG Sharing

September 1, 2023

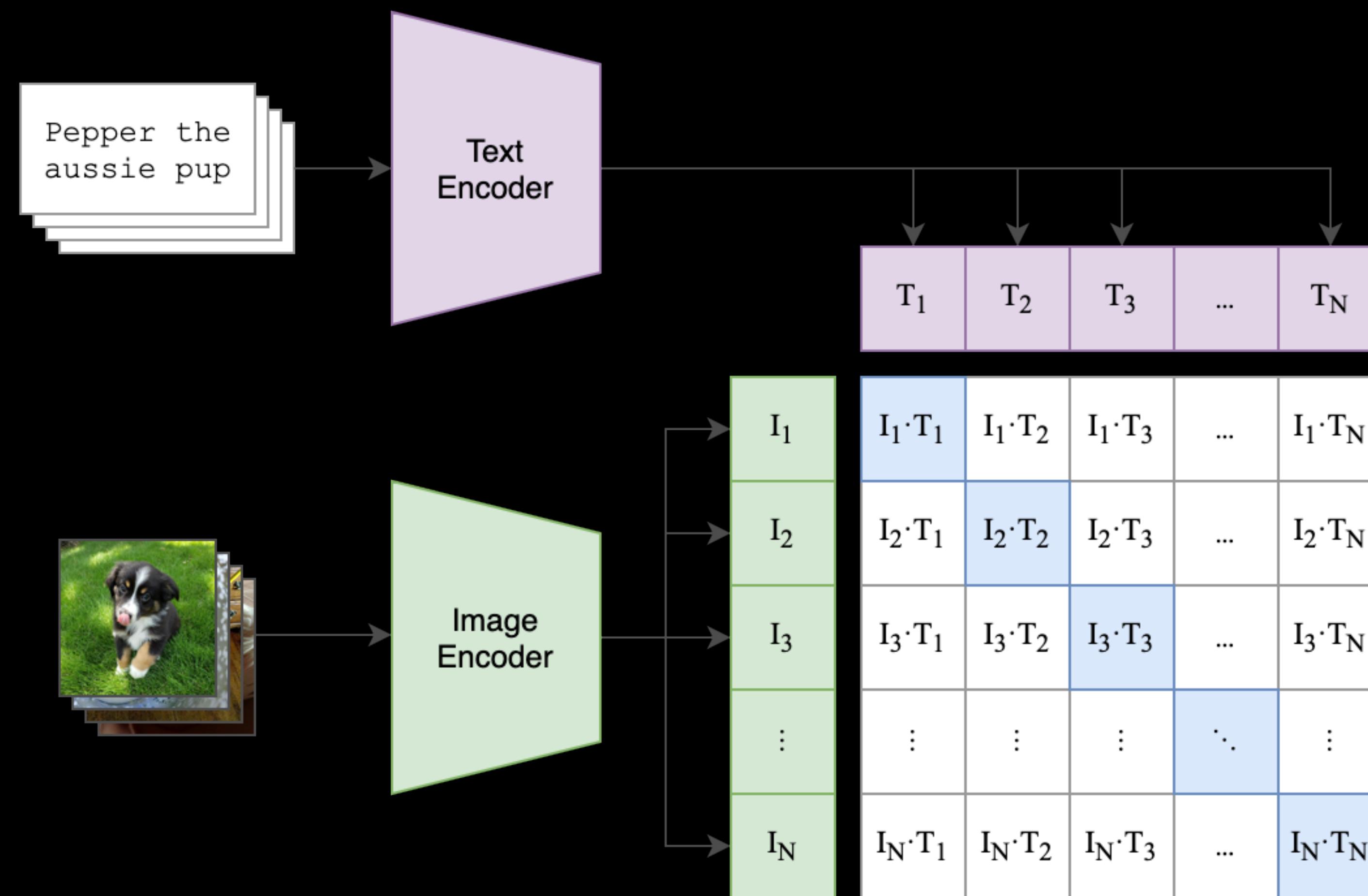
CLIP

Connecting text and images

Eng DRI: Chen Chen, Bowen Zhang

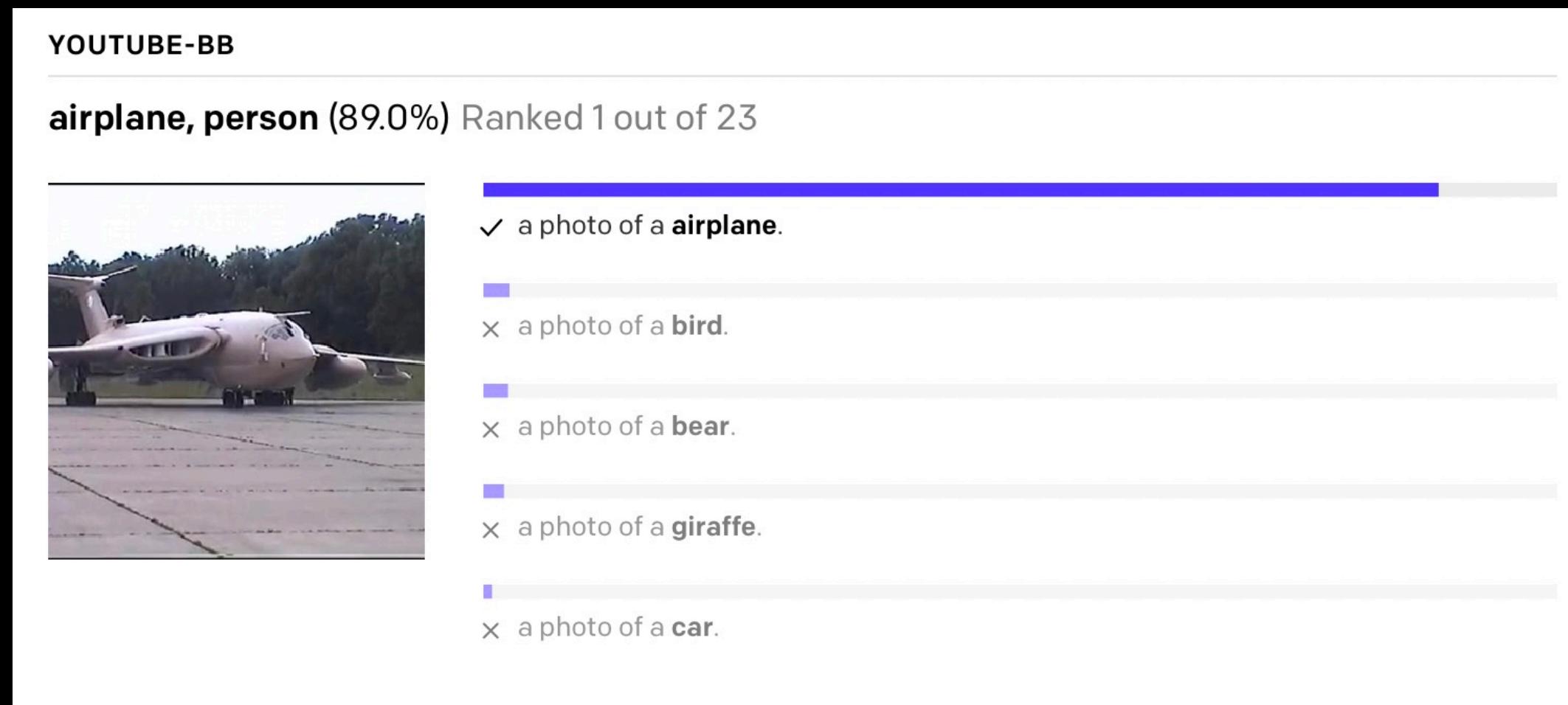
What is CLIP?

- Contrastive Language-Image Pre-Training
- Text / Image representations in the same latent space



Applications

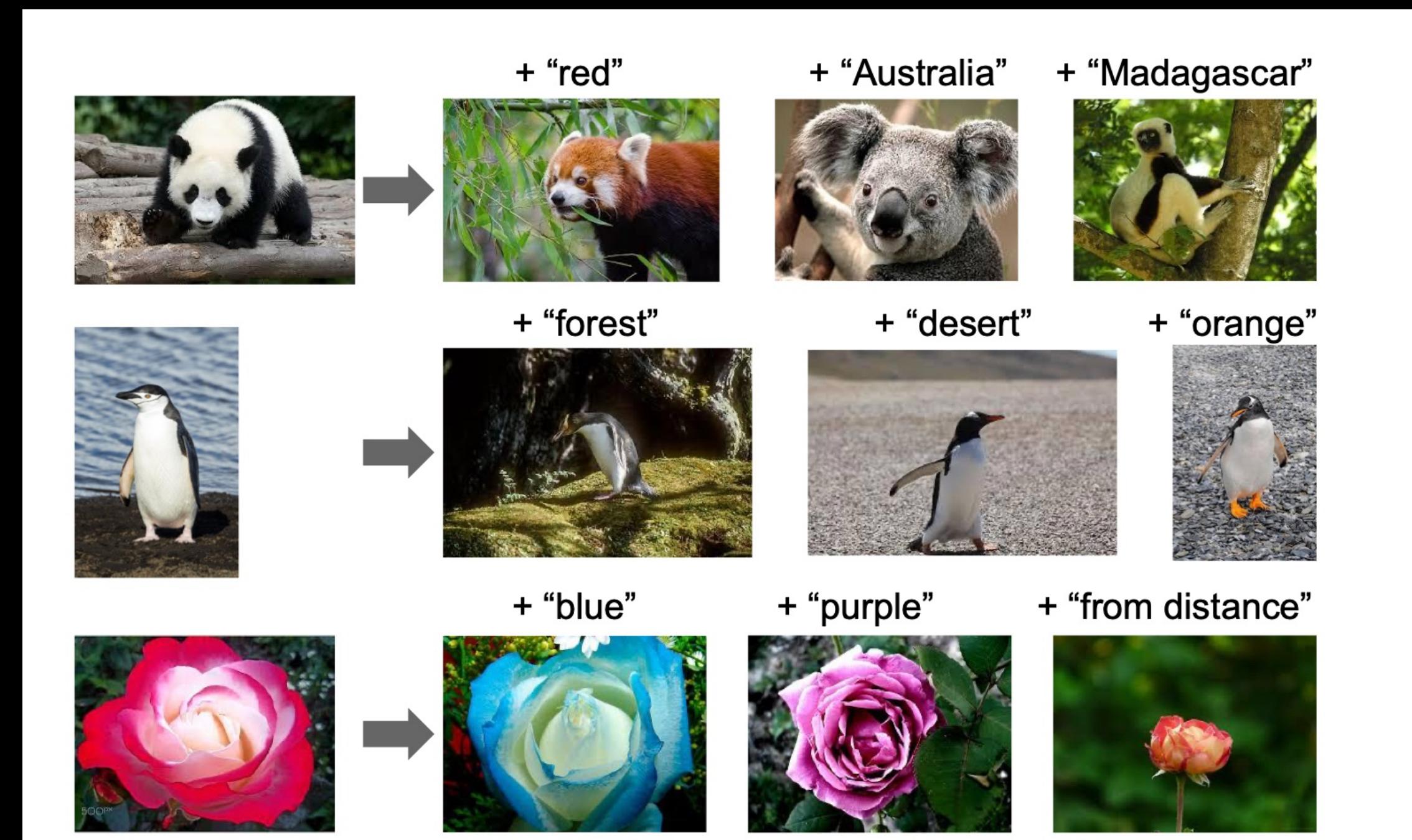
- Zero-shot Classification



- Text / Image Search



- Multimodal Search



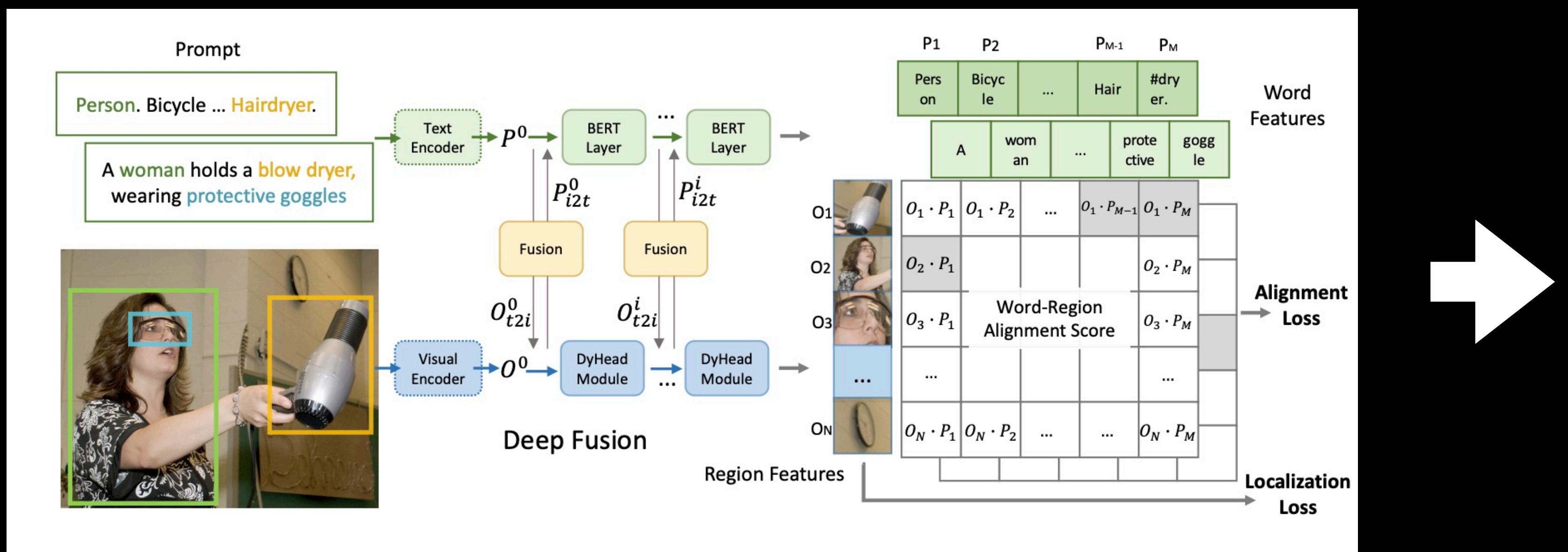
GLIP

Object detection, bounding boxes, and object classes

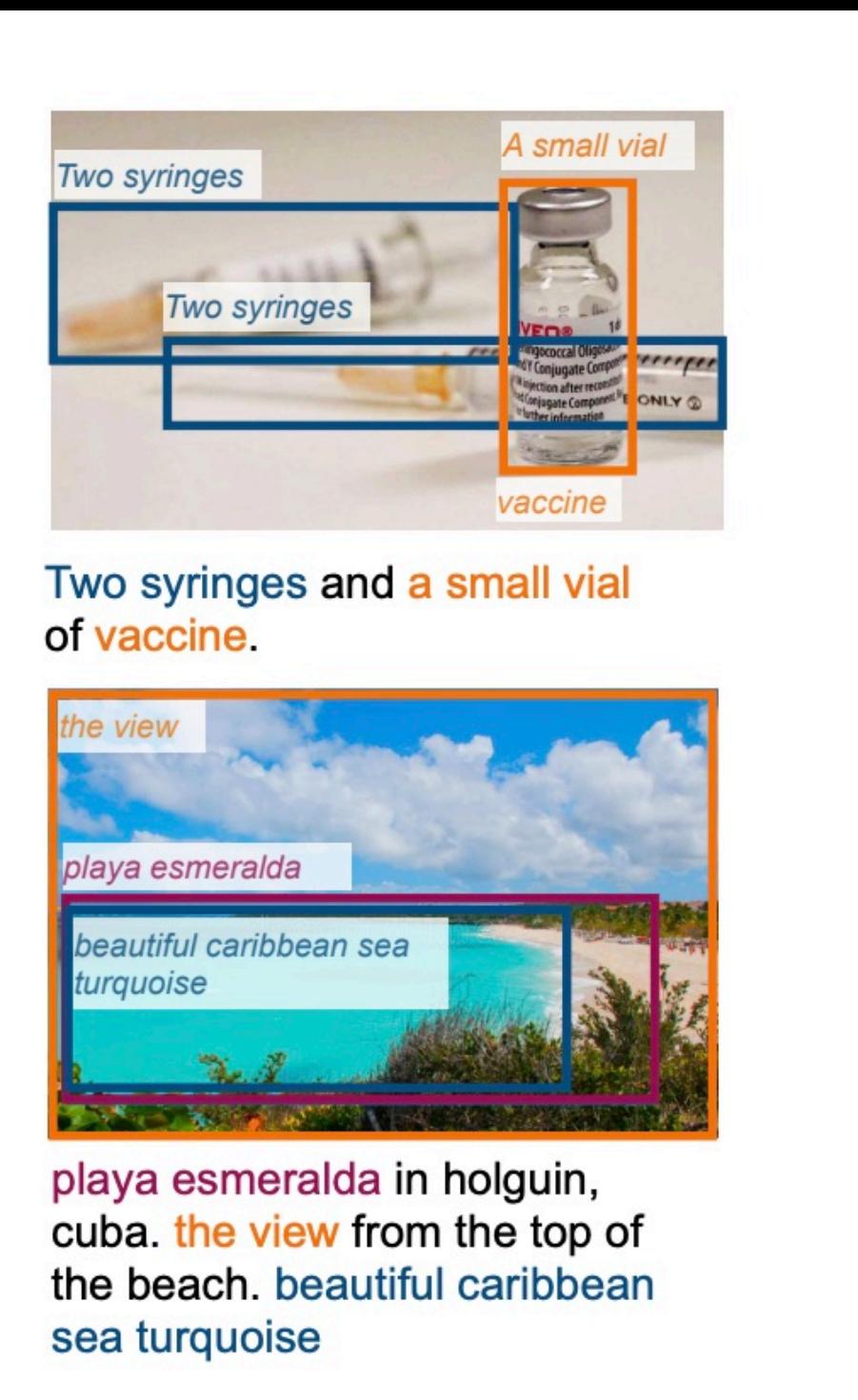
Eng DRI: Haotian (AIML) Zhang

GLIPv2: Zero-shot object localization

- Grounded Language Image Pre-training:



Inference



The Zero-Shot Segmentation Model uses the previous detection boxes as the prompts for Segment Anything (SAM) and generate the corresponding segmentation masks.

image

tasks

Zero-Shot Detection Zero-Shot Segmentation

text

eggs

Clear

Submit

image

name entities

['eggs']

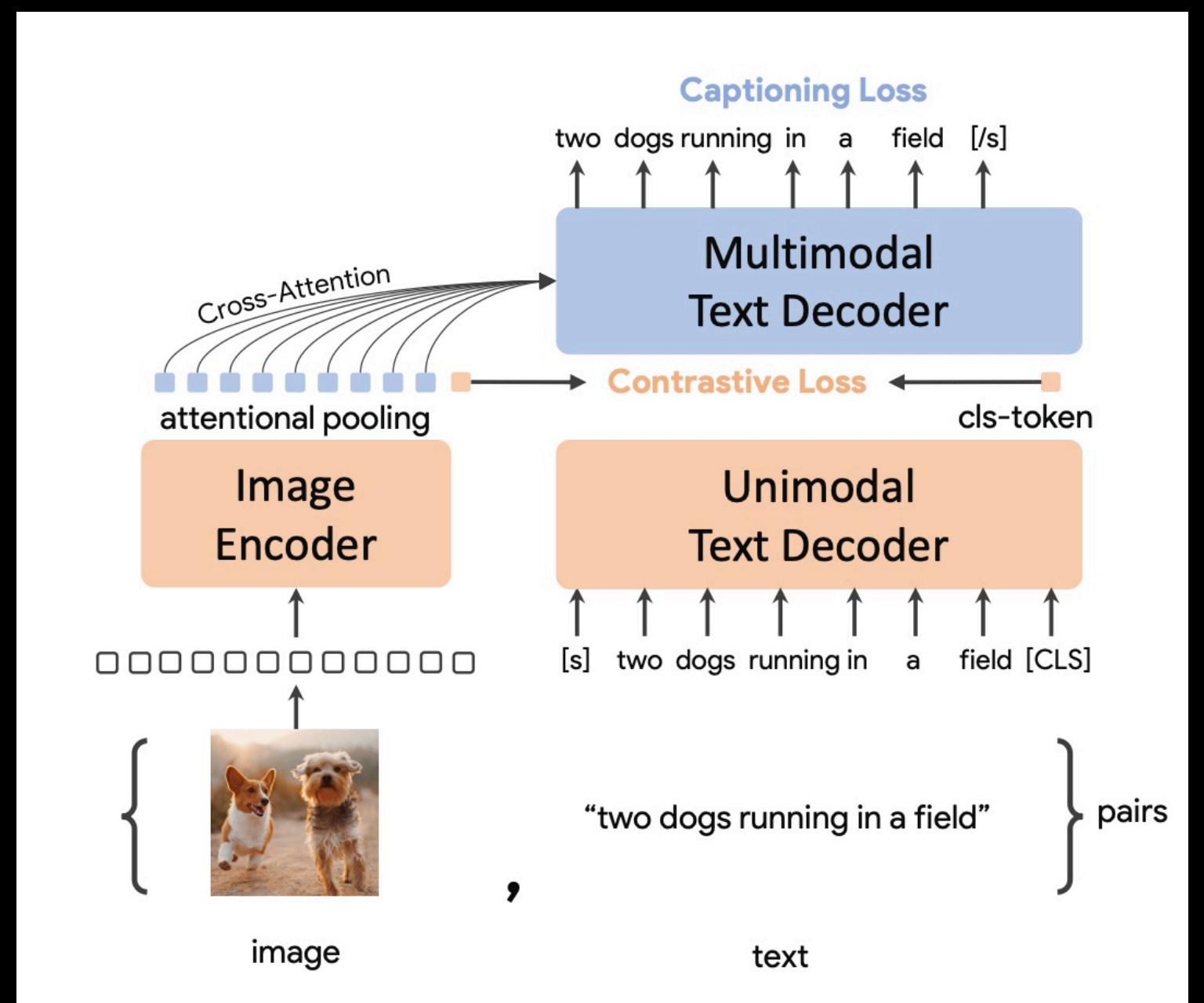
CoCa

High quality contrastive image captioner

Eng DRI: Yinfai Yang, Zhe Gan

Coca Image-Text Model

- Contrastive Captioner (CoCa)
 - Joint train captioning loss with contrastive loss.
 - Trained on billions of noisy image-text pairs.
 - Pre-trained and COCO Fine-tuned B-16 / L-14 models available.
 - Easy to scale to billions of parameters model.
 - Code available in [AJAX](#), active develop in progress.



Ref:



EntityNet

Large taxonomy image classification (scene, things, activities, etc.)

Eng DRI: Rajaditya Mukherjee

EntityNet

Large Taxonomy Content Understanding Model

+2x (up to 5x) increase in visual concepts compared to SydRo baseline

Robot
Figurine
Glass

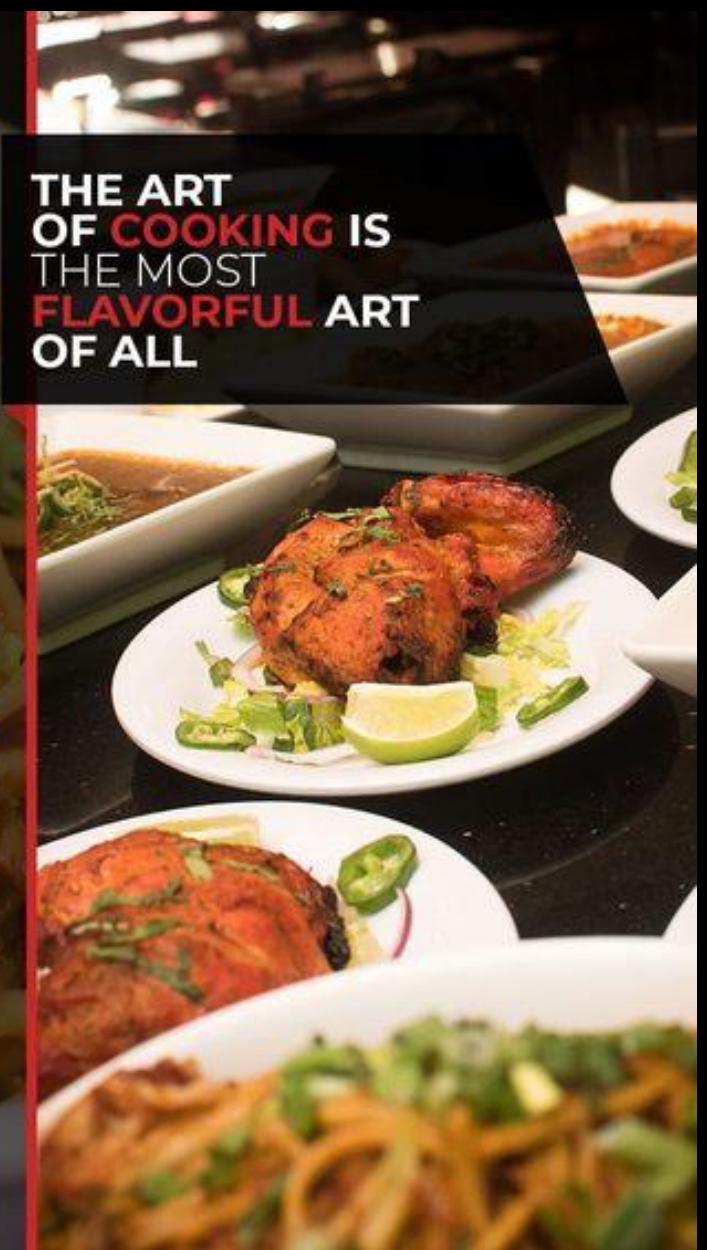
Iron Man
Superhero
Toy
Avengers



Printed Page
Bowl
Logo

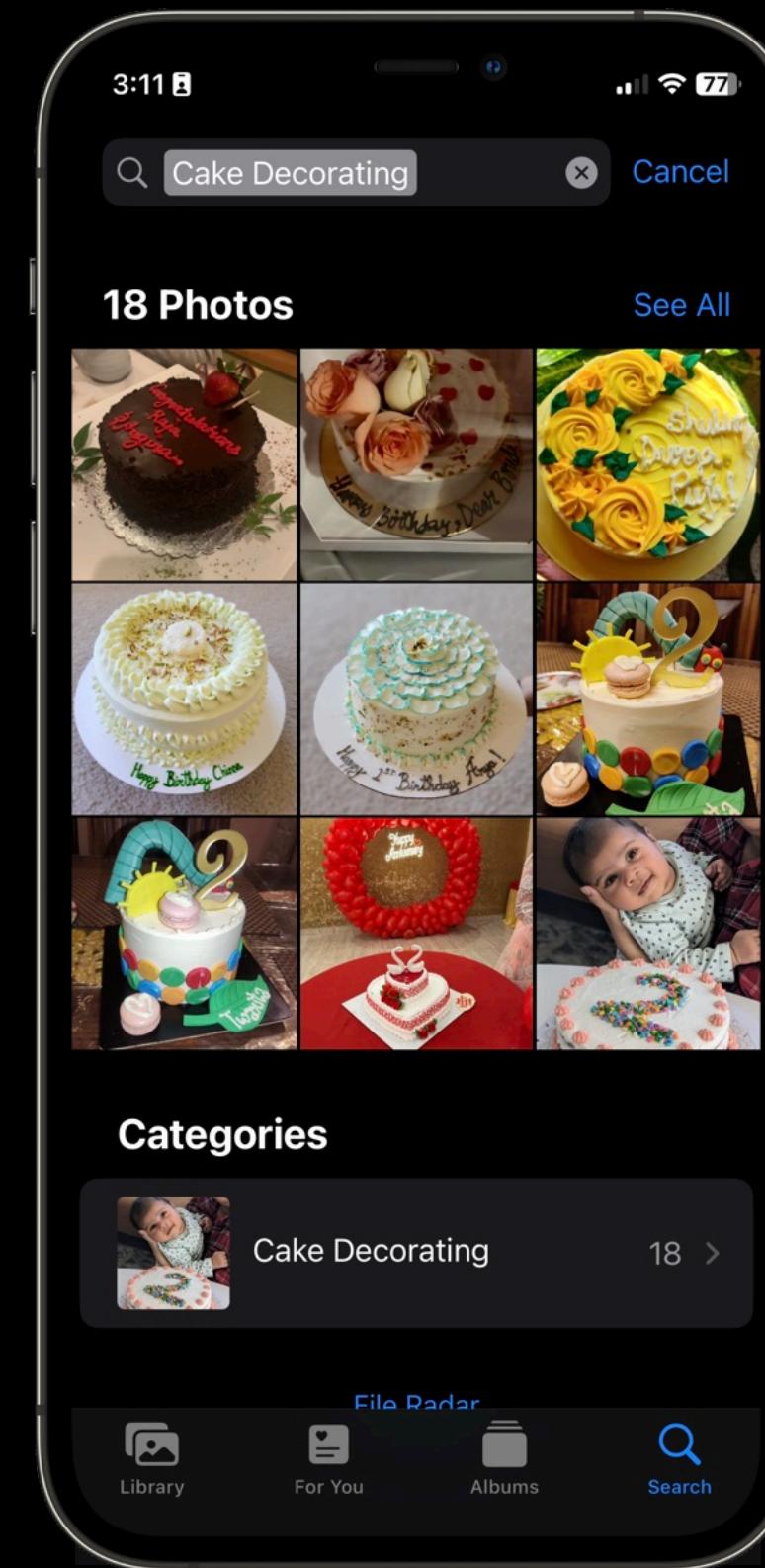
Menu
Fast Food

see more menus @ zmenu.com	
VEGETARIAN DELIGHT	
AALOO PALAK	\$7.99
AALOO GOBI	\$7.99
PALAK PANEER	\$8.99
NAVRAATAN KORMA	\$8.99
AALOO MUTTON	\$8.99
BHINDHI MASALA	\$8.99
CHOLAY BHATUREY	\$8.99
CHANA MASALA	\$7.99
PANEER TIKKA MASALA	\$8.99
MUTTER PANEER	\$8.99
DAAL FRY	\$7.99
PANEER CHILI	\$8.99
DAAL MAKHNI	\$8.99
MIX VEG KARahi	\$8.99
BEVERAGES	
FRESH LIME SODA	\$3.49
MANGO LASSI	\$2.99
SWEET LASSI	\$2.99
SALT LASSI	\$2.99
FRESH TENDER COCONUT	\$3.99
BOTTLED WATER	\$1.25
SODA	\$1.49
COCONUT JUICE CAN	\$2.49
PERIER	\$2.49
DESSERTS	
GULAB JAMUN (2 PCS)	\$1.99
KHEER	\$3.99
GAJAR HALWA	\$3.99
KARIMI SPECIAL	
BEEF NIHARI	\$10.99
BEEF HALEEM	\$10.99
CHICKEN SAMOSAS (6 PCS)	\$5.99
BEEF SAMOSAS (6 PCS)	\$5.99
NAAN	
PLAIN NAAN	\$1.99
BUTTER NAAN	\$1.99
GARLIC NAAN	\$2.99
TANDOORI ROTI	\$2.49
ONION KULCHA	\$4.49
ALOO NAAN	\$4.49
PESTO NAAN	\$3.49
KEEMA NAAN	\$5.99

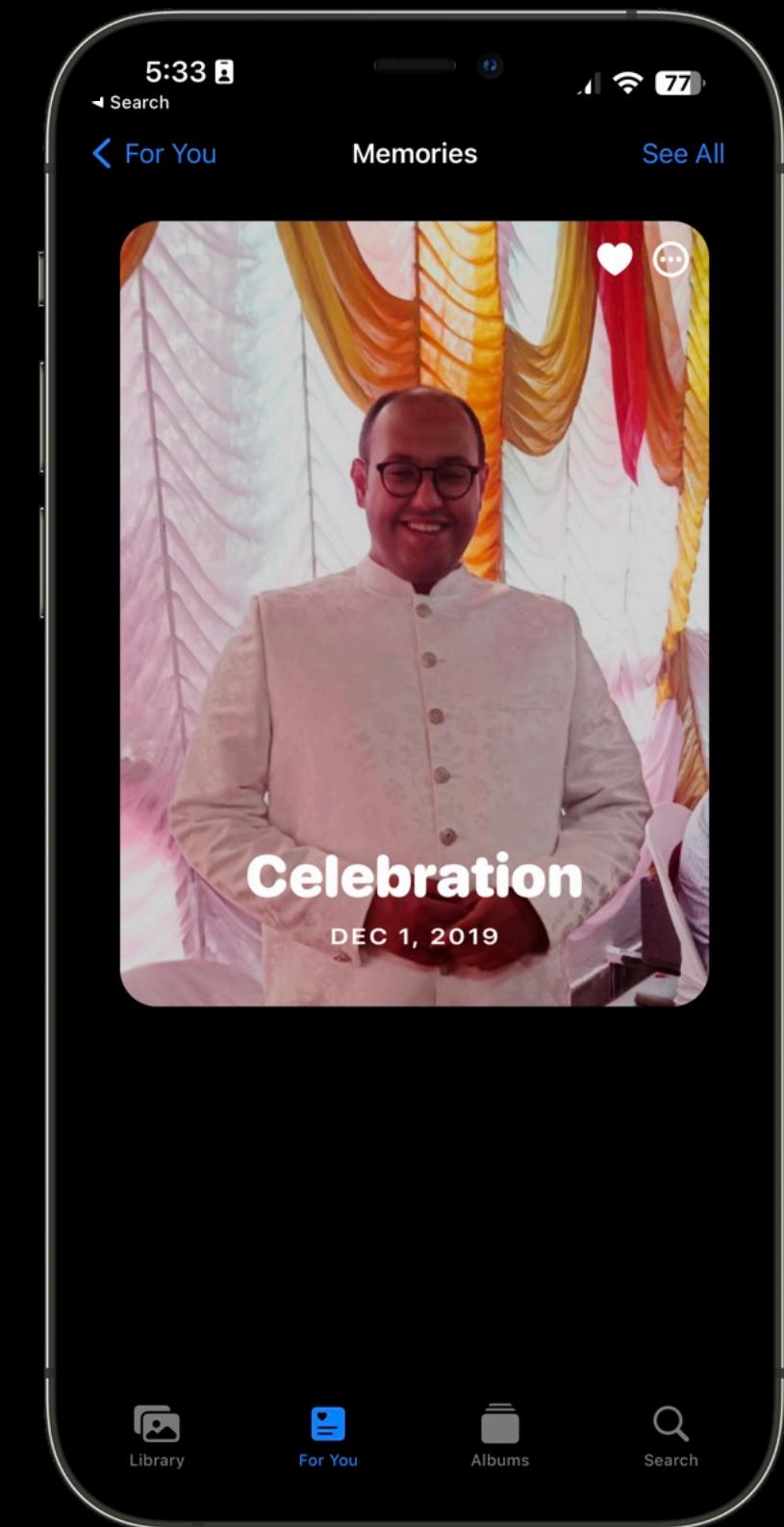


EntityNet Experiences

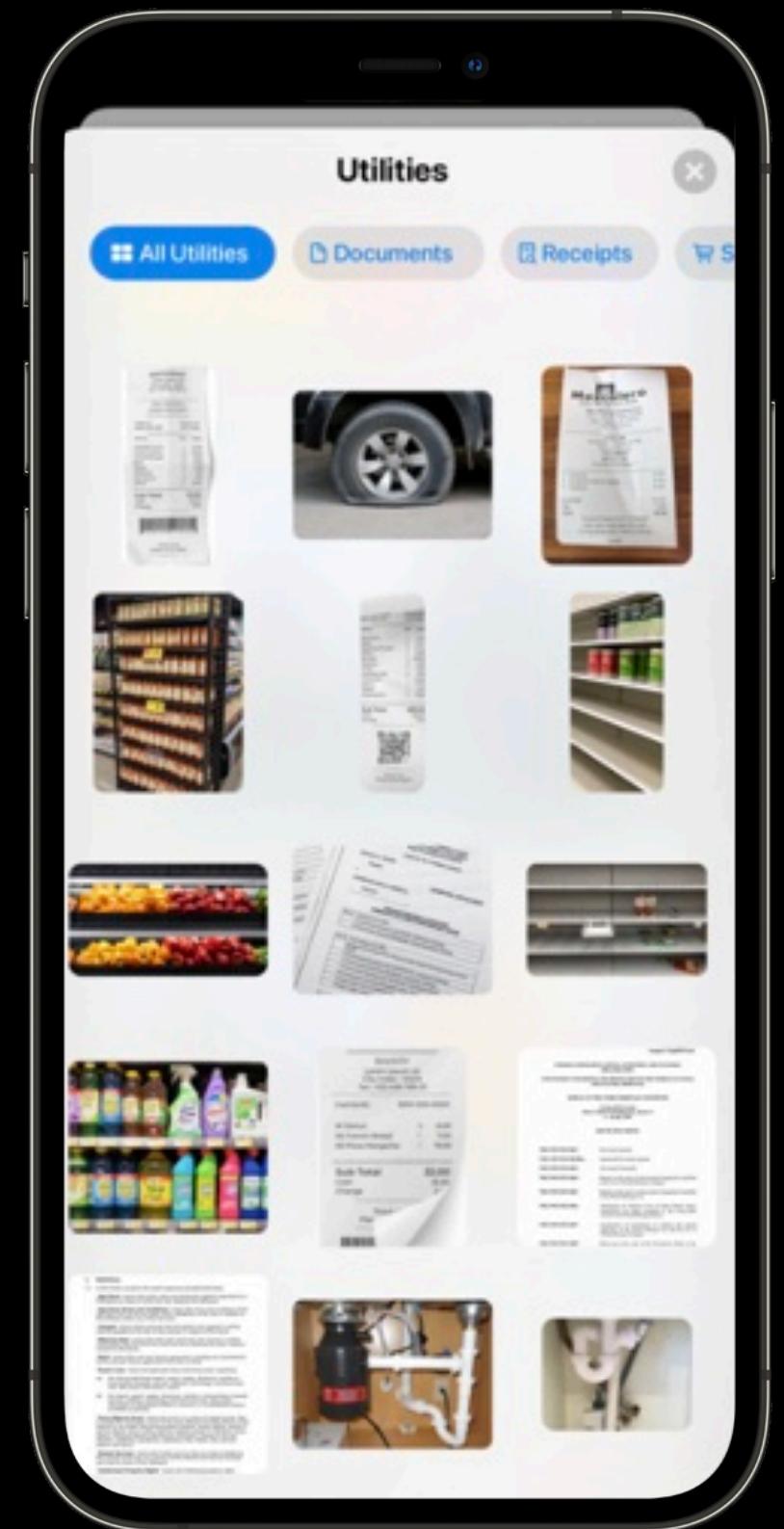
Empowering Search Experiences across Diverse Product Experiences



Enhanced on-device Asset Search



Memories and Storytelling



Fine Grained Collections*

*Upcoming Feature

MOFI

Generic image embedding model

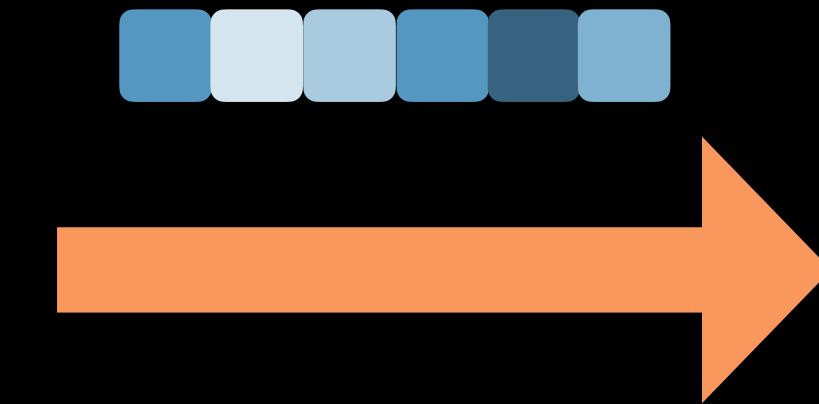
Eng DRI: Wentao Wu, Aleksei Timofeev

MOFI Learning Image Representations from Noisy Entity Annotated Images

- Goal: Generic image embedding for image2image retrieval



Query Image

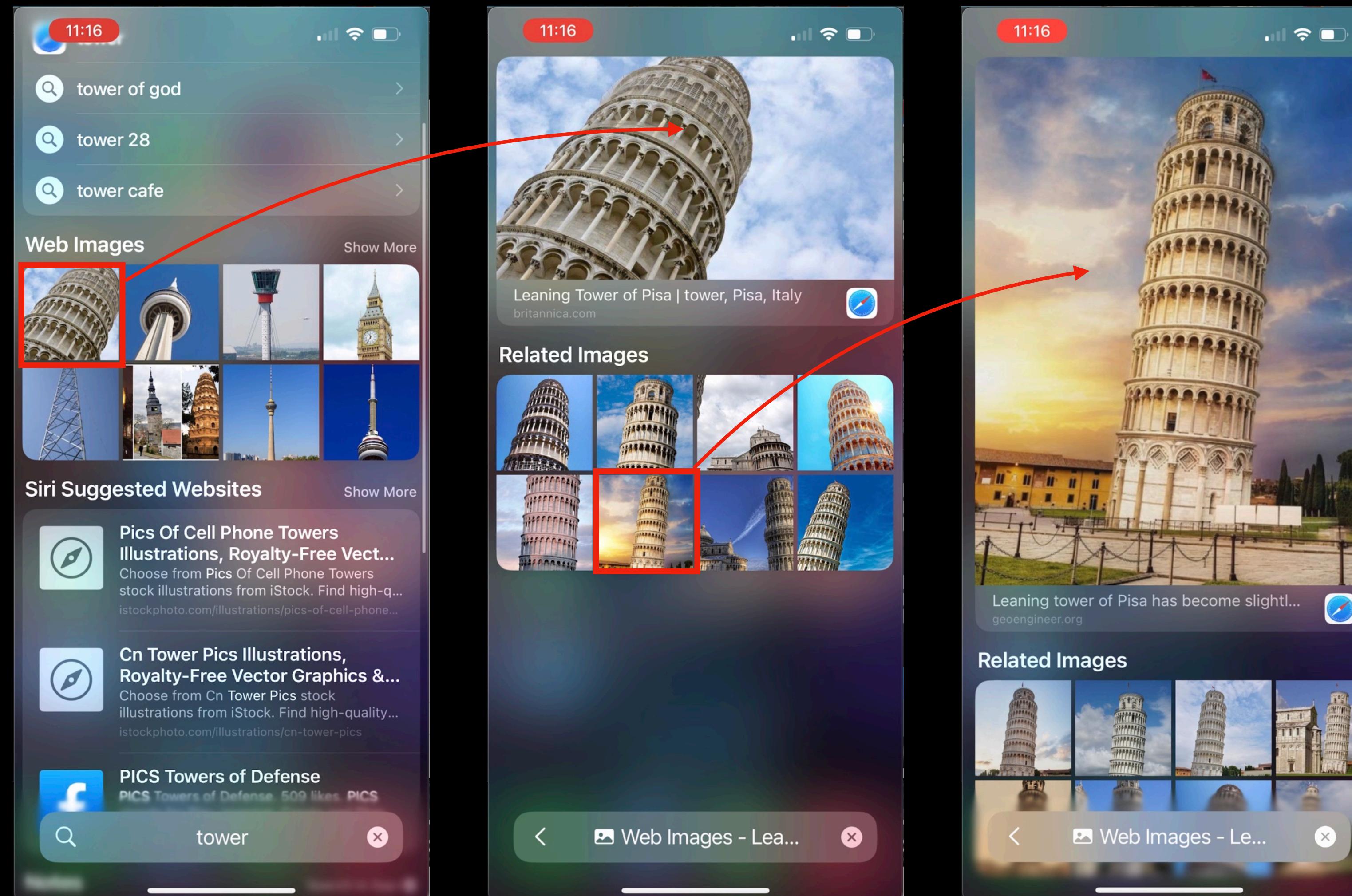


Similar Image



Related Images

Overview



Visual Lookup Improvement



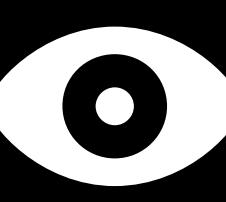
Photos



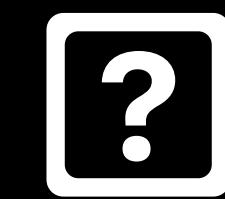
Safari



Camera Roll



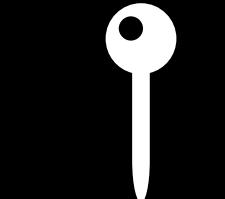
QuickLook



Pets



Nature



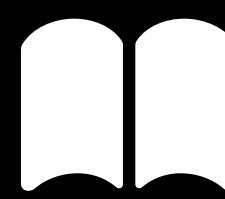
Landmark



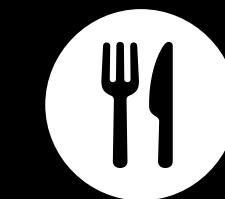
Art



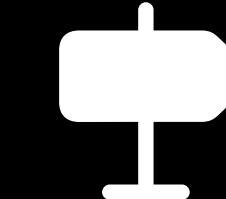
Albums



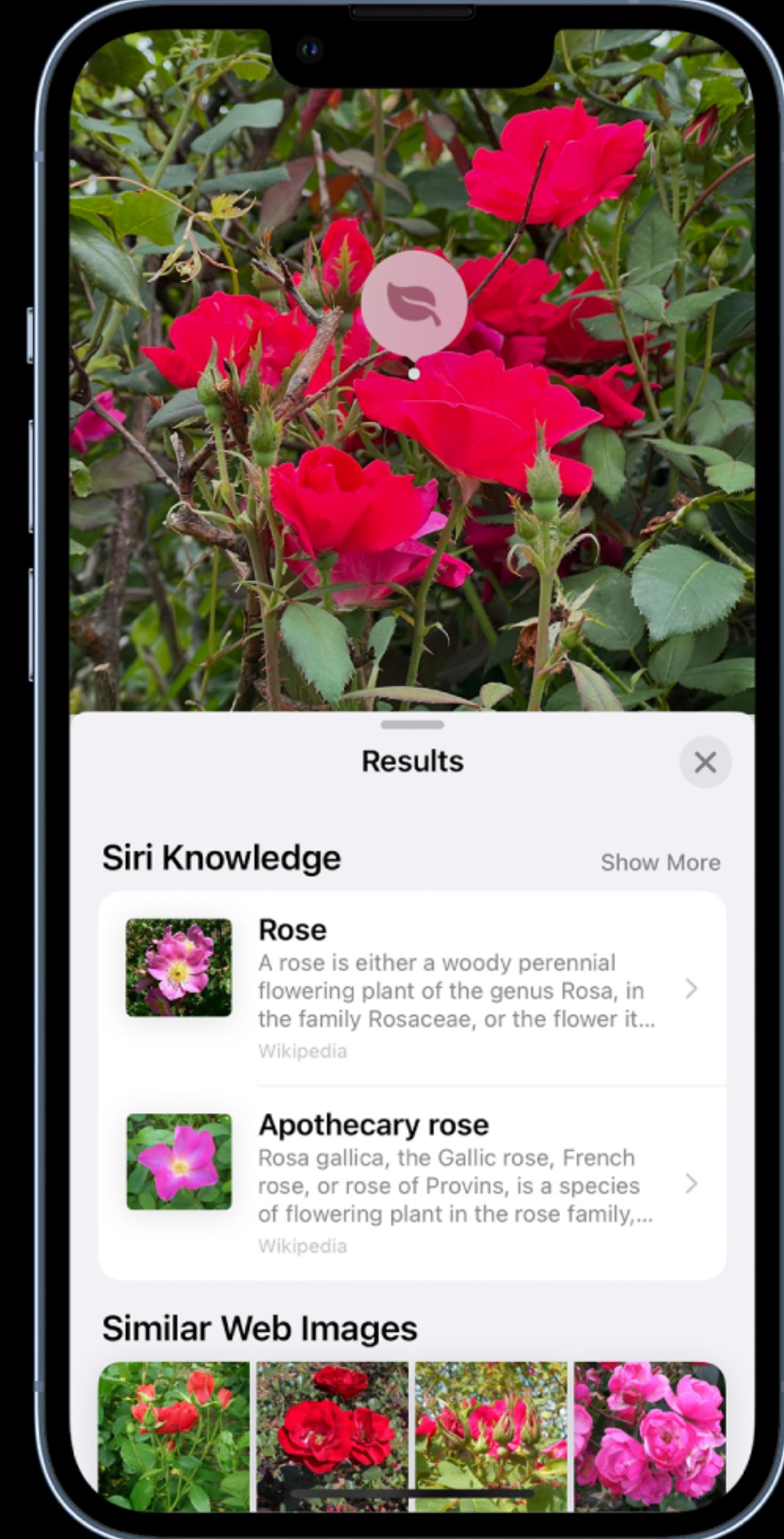
Books



Food



Signs & Symbols



Credit: James Byun

Apple Confidential—Internal Use Only

MFM

mGPT

Enhancing LLM with multimodal understanding (image to start)

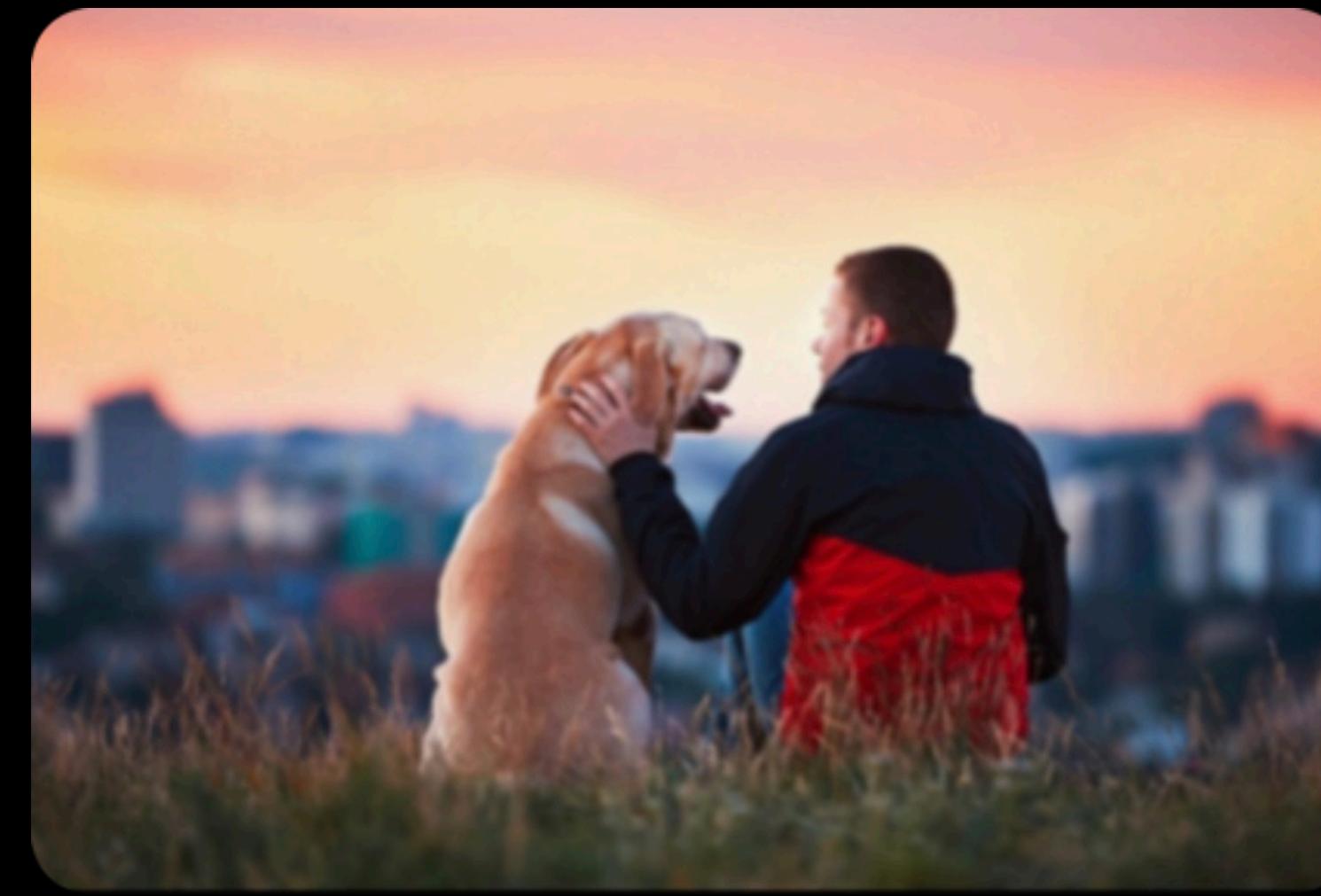
Eng DRI: Yinfei Yang

Creative Writing

Photo Inspired



Help me write an
advertisement about this



Can you write a beautiful
poem about this?

Storytelling

Photo Inspired



Tell me a scary story based on these pictures and send it to my mom

Image Analysis and Description

Short Form - Captioning

Instruction: What would be an appropriate summary for this image?
Response (Ours): An old Panera Bread sign is outside on a street.



Image Analysis and Description

Long Form Image Description

Instruction: Write an exhaustive depiction of the given image

Response1 (Ours): The image shows a woman sitting down at a table with a partially eaten donut in her hands. She is wearing glasses and a purple sweater. The table is set with various utensils, including a fork, a knife, and a spoon. There are several people present in the background, and it appears they are either dining or engaged in conversation. A handbag can be seen placed near the middle of the scene, and another handbag is located closer to the top right corner. In addition to the utensils, there is a clock on the wall in the background.



Image Analysis and Description

Emotive Captions (SIML)



Happy/Joyful: A smile with raised cheeks and/or crinkled eyes.



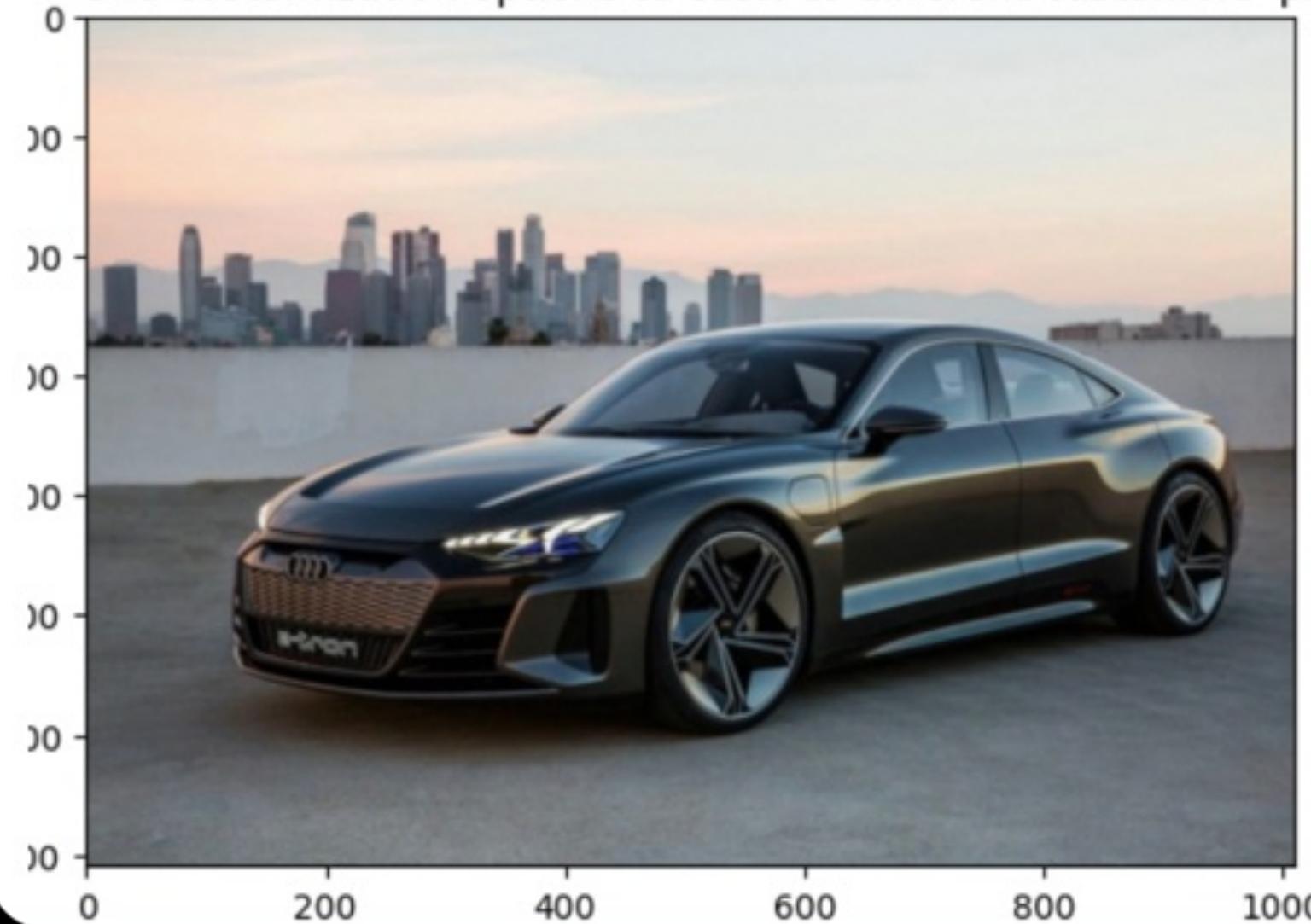
Sad: A frown with downturned corners of the mouth and/or tears.

Visual Knowledge Acquisition and Assistant

Visual Question Answering with Reasoning

Instruction: Explain the advantages of this product.

Response (Ours): The advantages of this product are that it is made of a lightweight material, making it more fuel-efficient and environmentally friendly. It can be more easily transported, as the lightweight material reduces the overall weight and reduces the stress on the vehicle during transit. Additionally, the black Audi's design and looks may resemble that of a sports car, which can generate excitement and passion among customers. It can offer various performance features and customization options to cater to different customers' preferences, resulting in increased satisfaction and brand loyalty.



Visual Knowledge Acquisition and Assistant

Visual Question Answering with Reasoning



Example from Otter/MIMIC-IT

Multimodal Progression

Input

- Text
- Image
- Video
- Audio
- 3D

Output

- Text
- Image
- Video
- Audio
- 3D

Core Technologies Summary

Available Now:

- CLIP

- Multi-modal vision and language model, cross-modality retrieval and many downstream tasks
- Eng DRI: Chen Chen, Bowen Zhang

- CoCa

- Contrastive Captioner, vision-language encoder-decoder model, image captioning, cross modal retrieval, VQA, etc.
- Eng DRI: Yinfai Yang, Zhe Gan

- GLIP

- Detect multiple objects in an image and provide object classes and bounding boxes
- Eng DRI: Chen Chen, Bowen Zhang

- ObjectDetector

- Detect multiple objects in an image and provide object classes and bounding boxes
- Eng DRI: Marcin Eichner, Michael Emmersberger

- EntityNet

- Large taxonomy image classification (scene, things, activities, etc.)
- Eng DRI: Rajaditya Mukherjee

Coming soon:

- MOFI

- Generic image embedding model to improve image-to-image lookup and expand domain coverage in Visual Lookup
- Eng DRI: Wentao Wu

- mGPT

- General purpose multimodal foundation models with strong text, image, audio and video capabilities
- Eng DRI: Yinfai Yang

- Image Generative Models (2D/3D)

- Text to image generation and editing, 3D object/scene construction and editing
- Eng DRI: Ohil Manyam



TM and © 2023 Apple Inc. All rights reserved.