



Florence: A New Computer Vision Foundation Model

Lu Yuan

Microsoft Cloud & AI

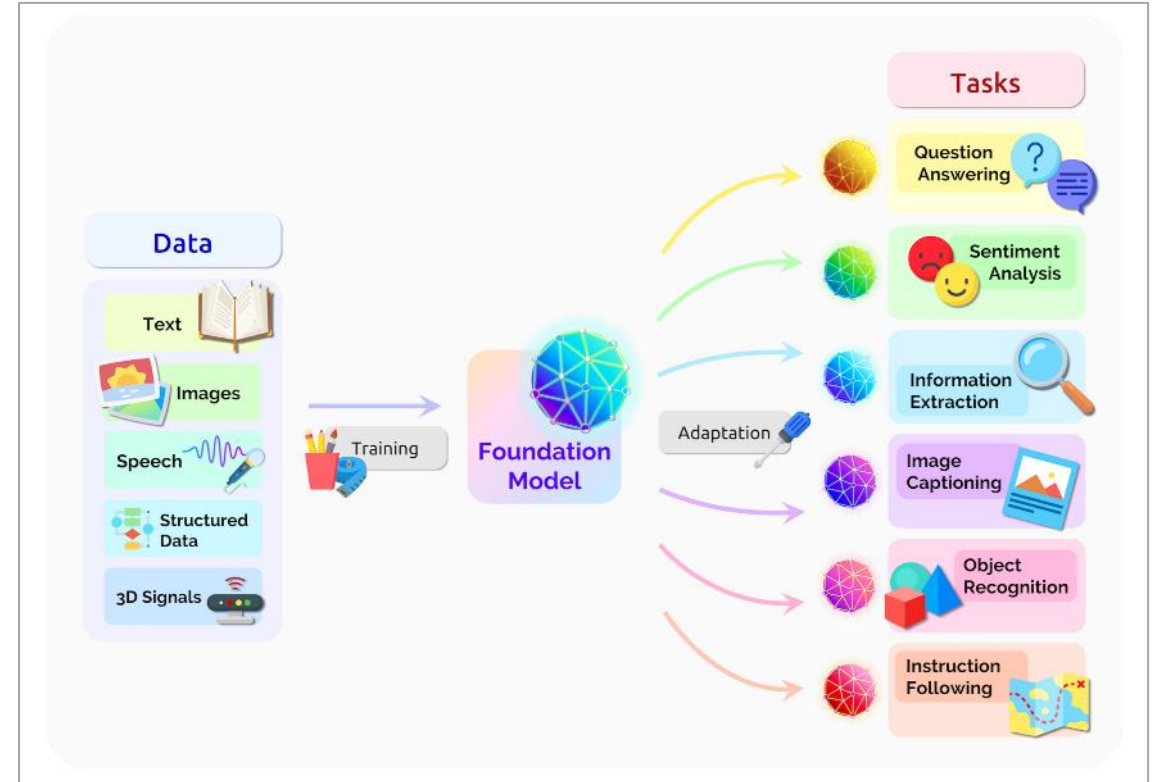
The Era of Foundation Models

A foundation model can centralize the information from all the data from various modalities.

This one model can then be adapted to a wide range of downstream tasks.

Existing Foundation Models:

- GPT-3
- CLIP
- Florence
- Flamingo
- CoCa
- PaLI



R. Bommasani et. al., On the Opportunities and Risks of Foundation Models, CRFM Stanford, 2021

A Glimpse of Diverse Computer Vision Tasks

Image Classification

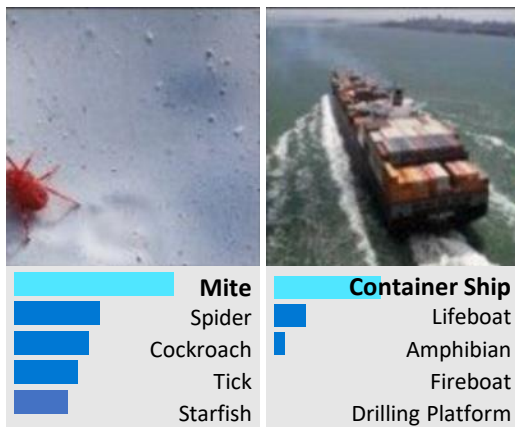
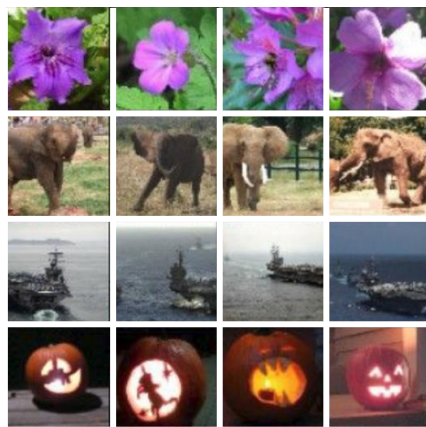


Image Retrieval



Object Detection

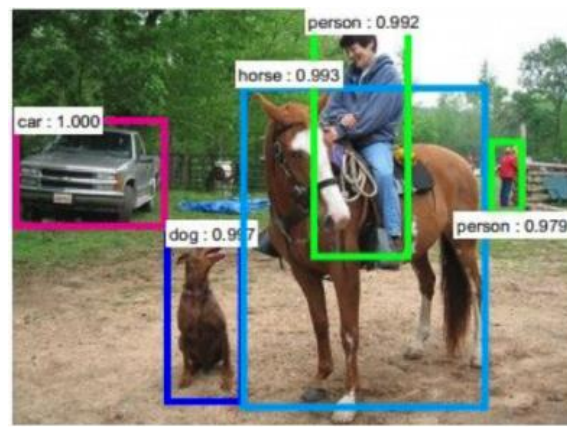
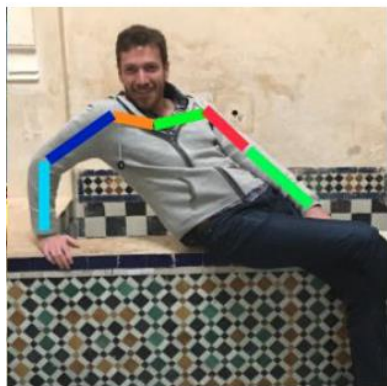


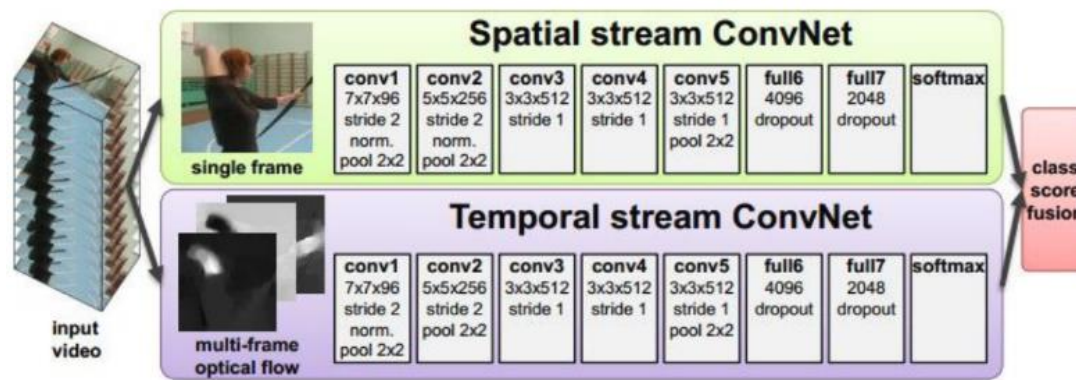
Image Segmentation



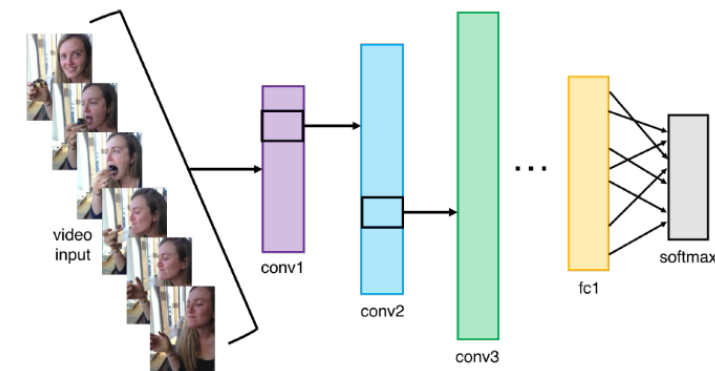
Pose Estimation



Video Classification

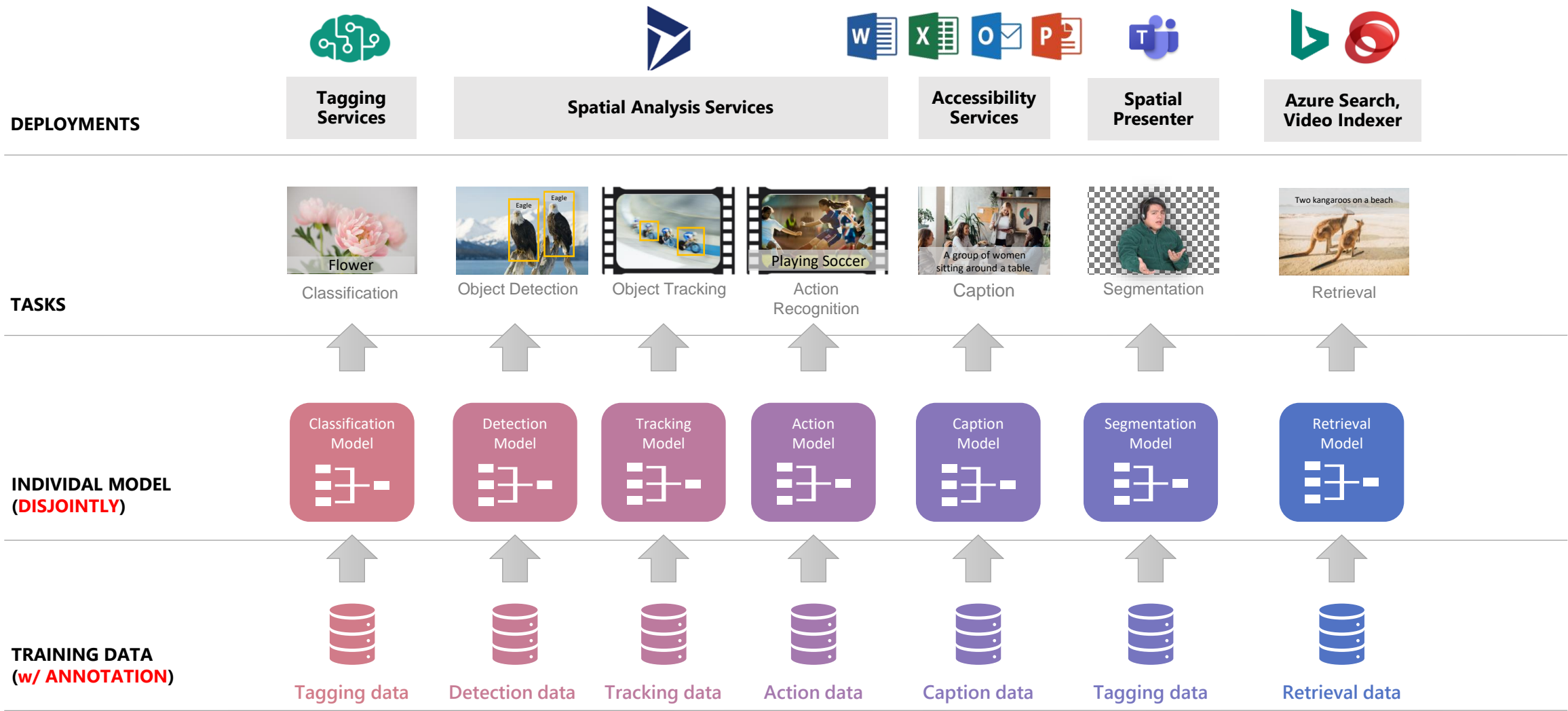


Activity Recognition



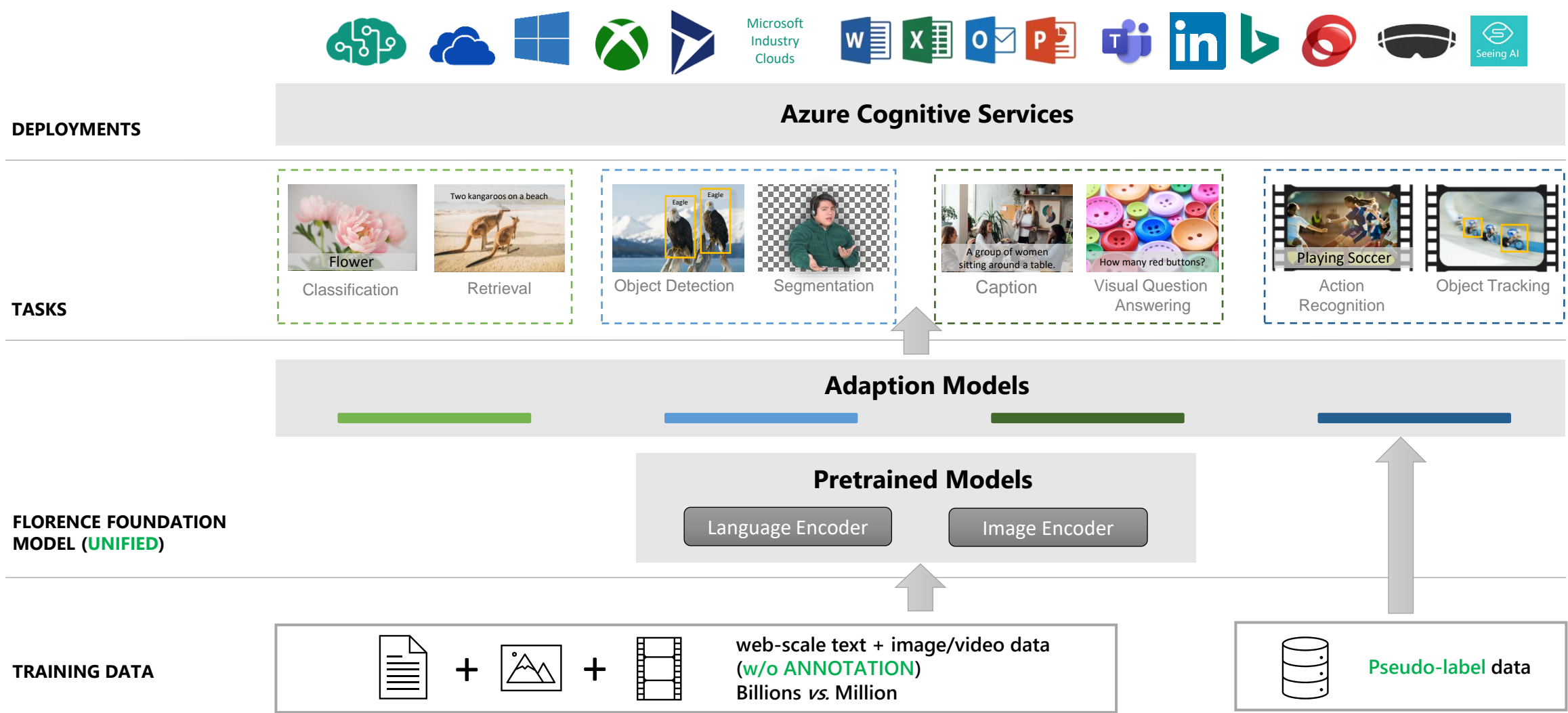
Before Florence: High cost & slow deployment

Each service is trained disjointly



After Florence: Low cost & fast deployment

Unified vision services



Florence: A New Foundation Model for Computer Vision

Florence unified **space**, **time** and **modalities** in computer vision under one pre-training + adapter framework

Training data

900M

Image encoder

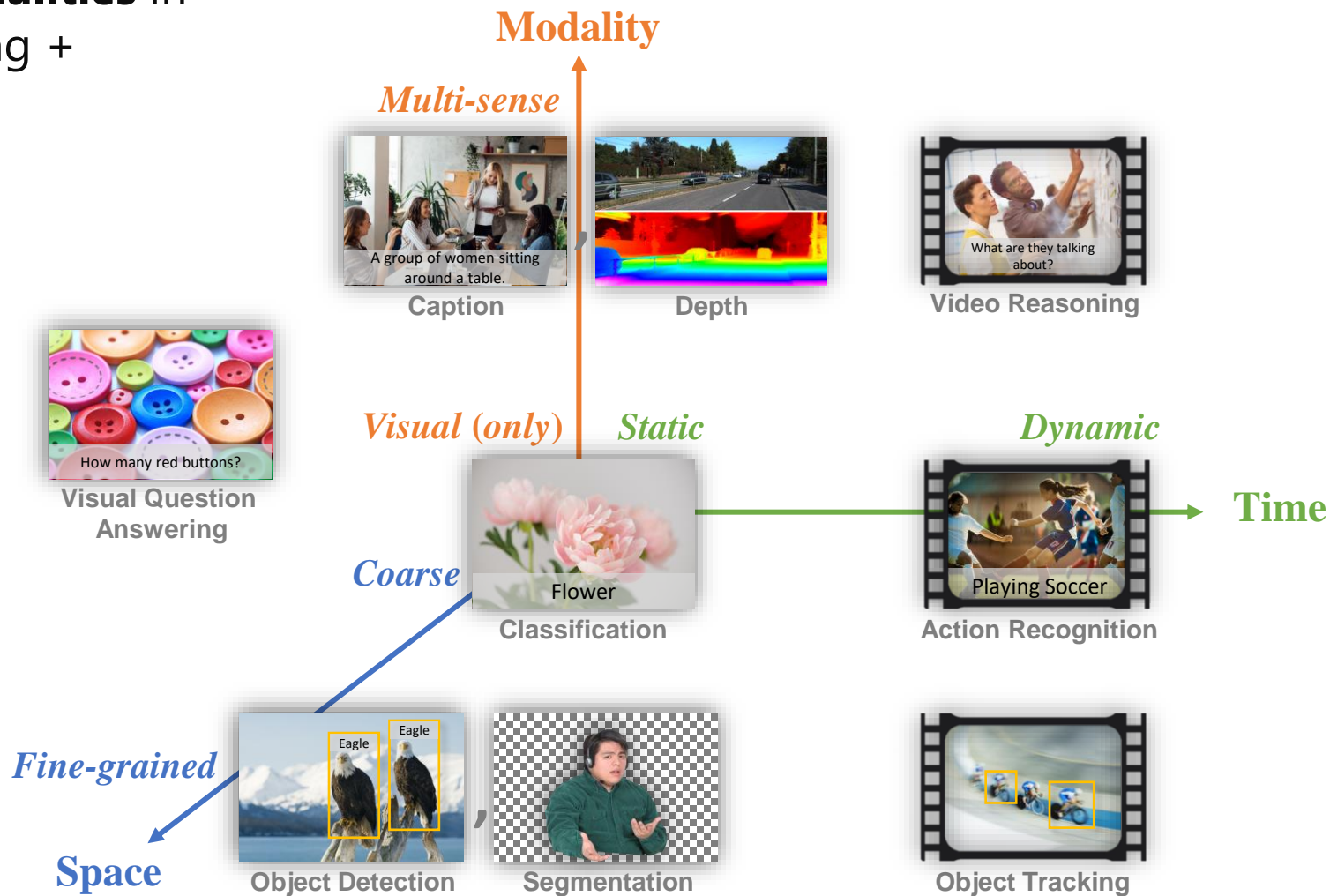
CoSwin (637M parameters)

Text encoder

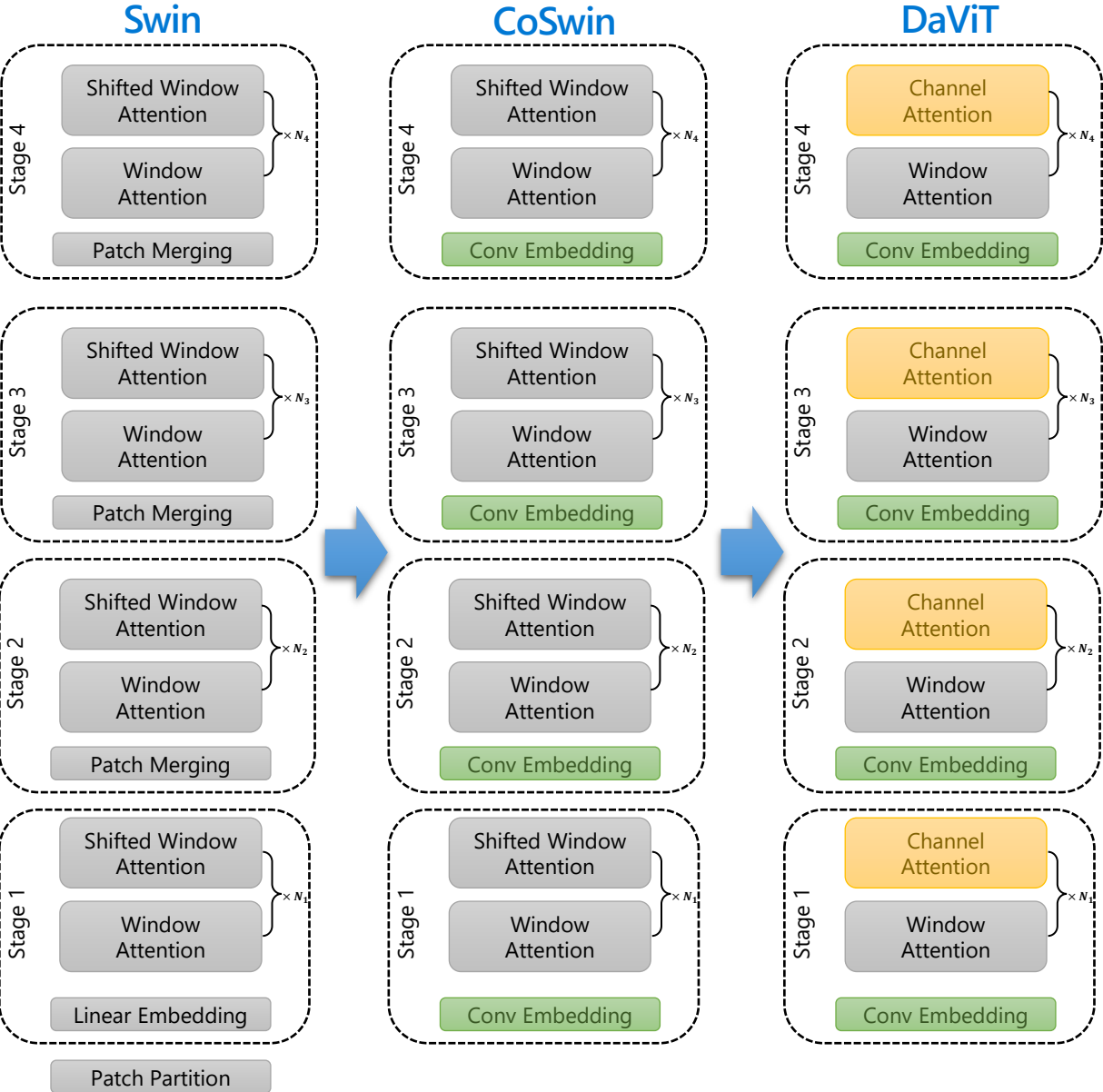
Transformers (256M parameters)

Compute resource

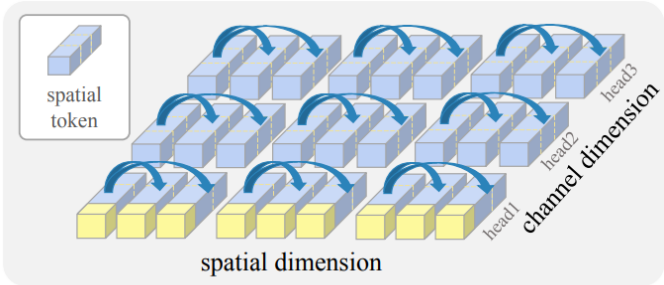
NVIDIA-A100 x 512, 14 days



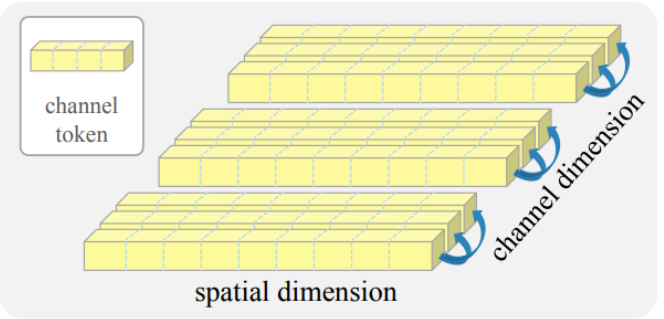
Vision Encoder: Hierarchical Transformer



Dual attention Vision Transformer: Enjoy the efficiency of local attention, meanwhile have the ability of global interaction.



(a) Spatial Window Multihead Self-attention



(b) Channel Group Self-attention

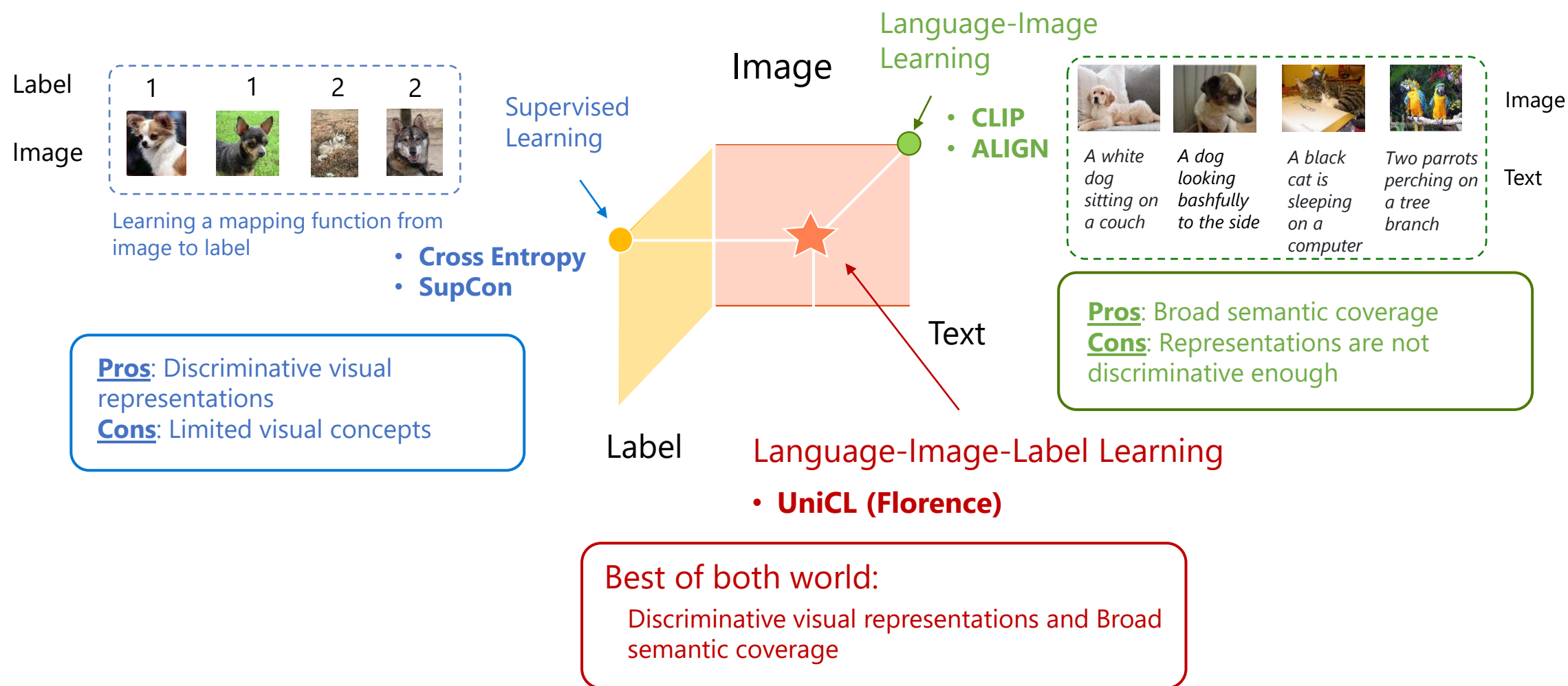
transpose
tokenization

Model*	ImageNet-1k
Swin-T	81.3
CoSwin-T	81.7
DaViT-T	82.8

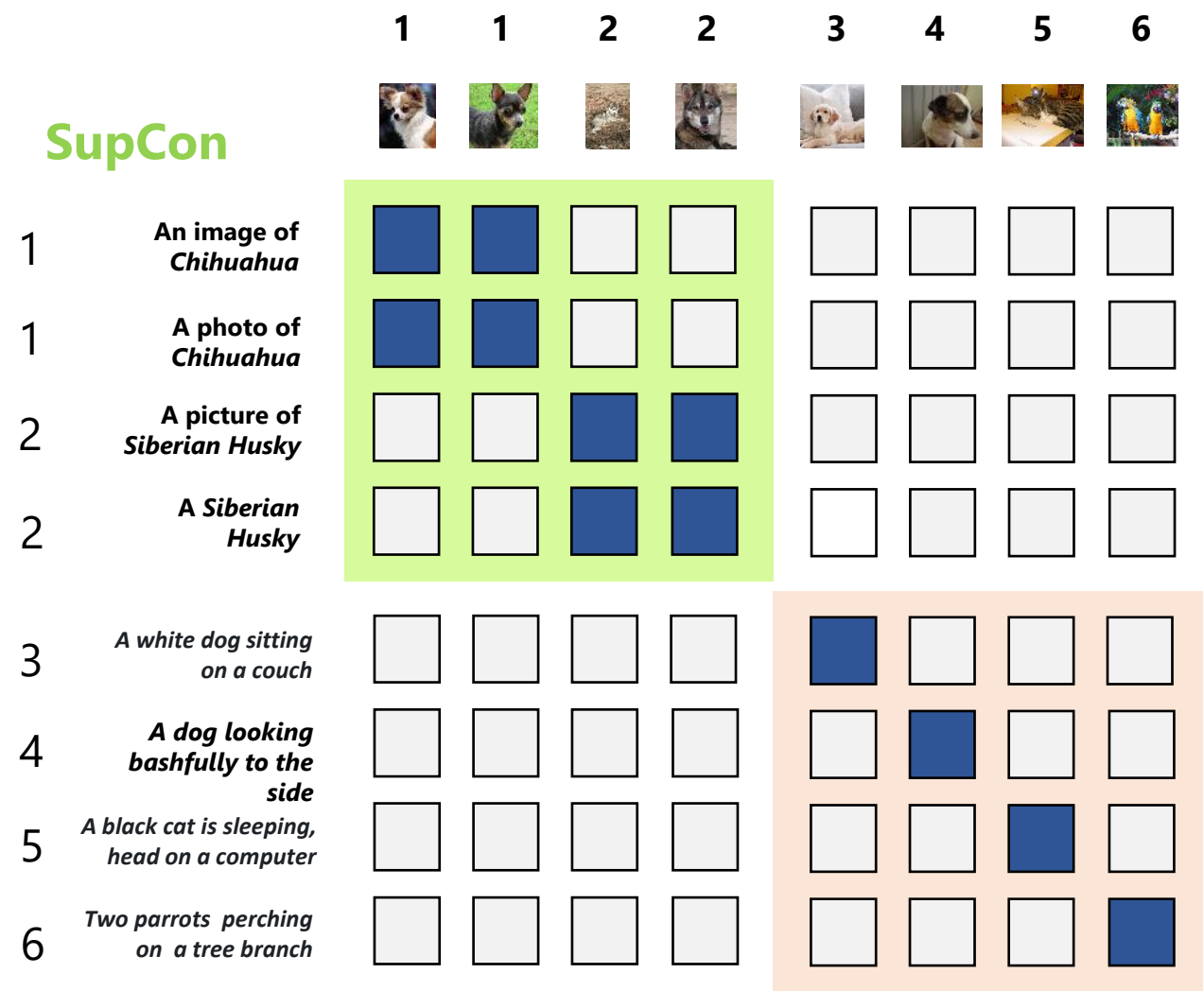
*Models trained on ImageNet-1k

[1] Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. (ICCV 2021)
[2] Florence: A New Foundation Model for Computer Vision. (arXiv 2111.11432)
[3] DaViT: Dual Attention Vision Transformers. (ECCV 2022)

Unified Contrastive Learning in Image-Text-Label Space (UniCL)



Unified Contrastive Learning in Image-Text-Label Space (UniCL)



UniCL

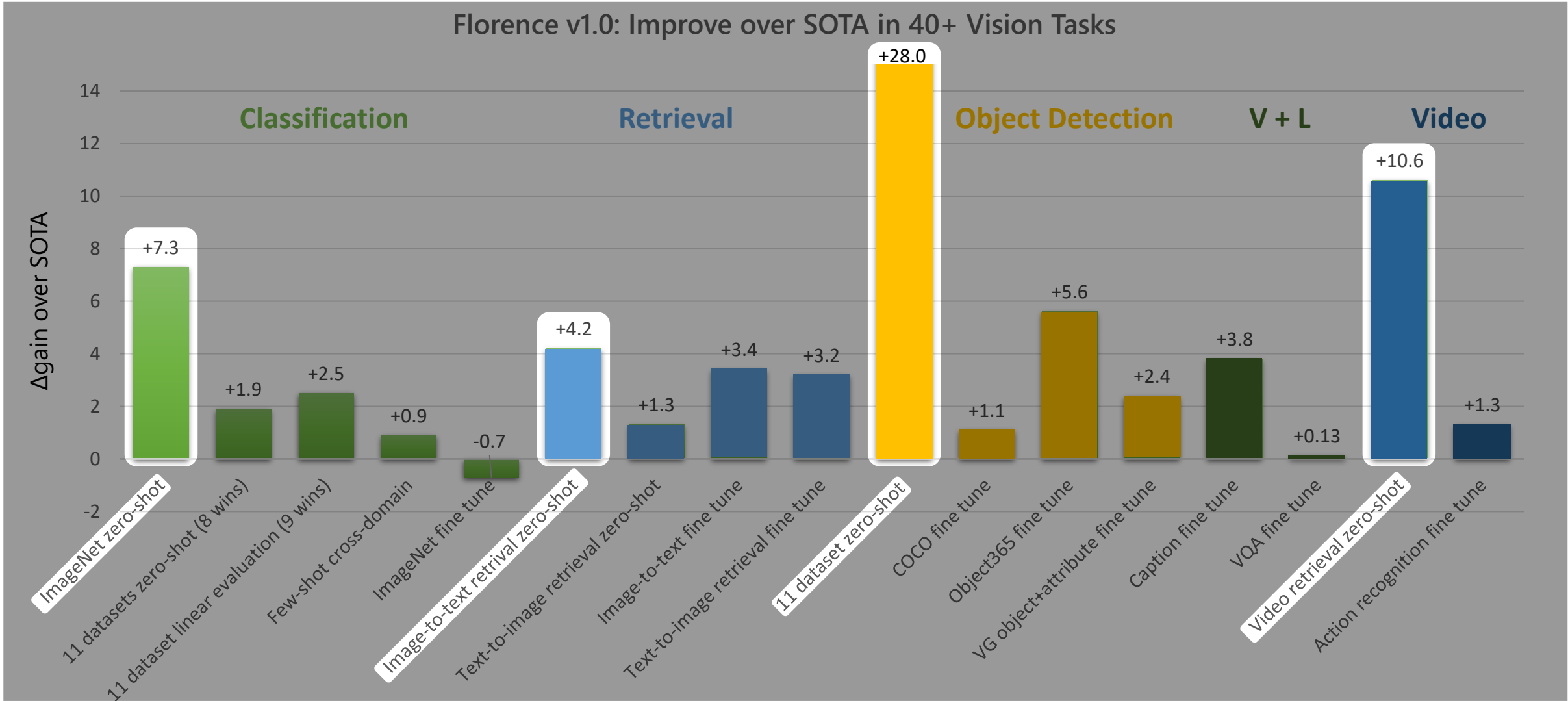
CLIP

Method	Zero-Shot on IN-1K
CLIP (400M data, 500M params)	76.2
Florence-UniCL (900M data, 900M params)	83.7

Florence:

1st Foundation Model to Demonstrate Quality Leap in Multiple CV Tasks

Florence v1.0: Improve over SOTA in 40+ Vision Tasks



Florence: A New Foundation Model for Computer Vision. (arXiv 2111.11432. Florence v1.0 released on 11/5/2021)

Florence Encoder + Text Decoder Adaptor (GIT)

Achieved SOTA results on 12 image/video captioning and QA tasks

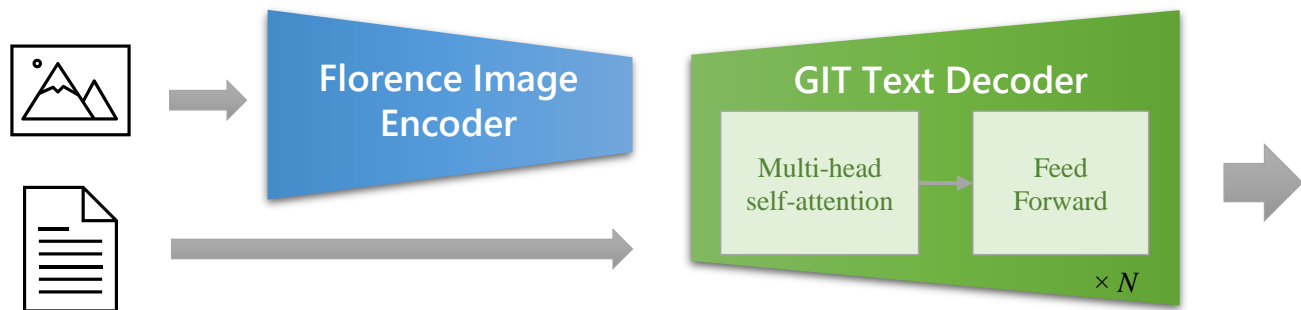
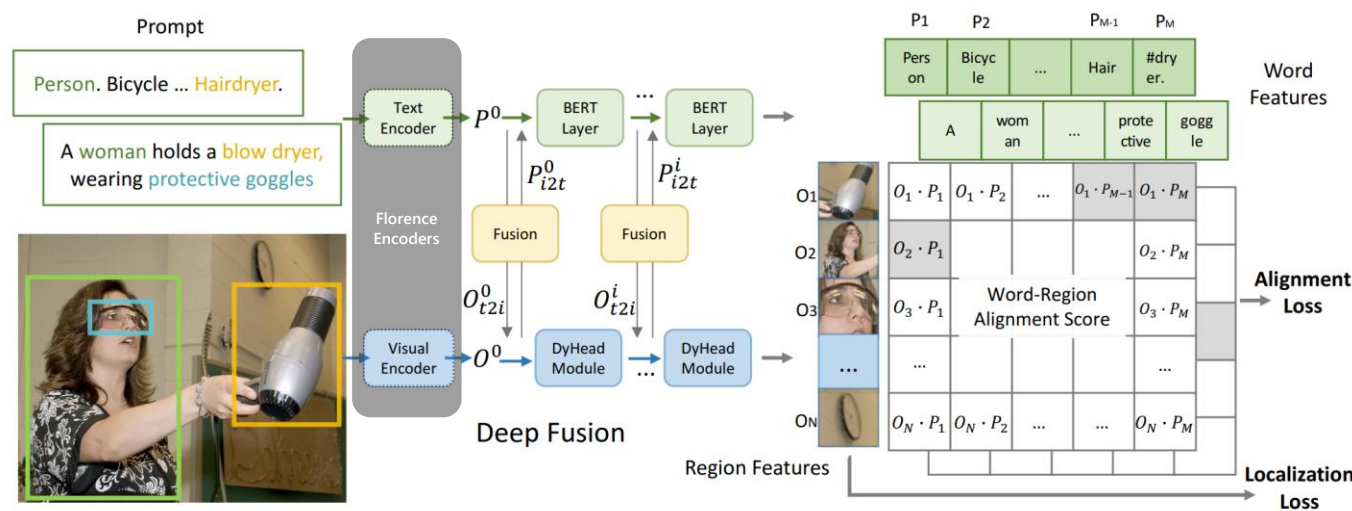


	Image captioning				Image QA			Video captioning			Video QA	
	COCO*	nocaps*	VizWiz*	TextCaps*	ST-VQA*	VizWiz*	OCR-VQA	MSVD	MSRVTT	VATEX*	MSVD-QA	TGIF-Frame
Prior SOTA	138.7 [111]	120.6 [106]	94.1 [21]	109.7 [104]	59.7 [104]	65.4 [2]	64.1 [27]	120.6 [58]	60 [78]	86.5 [86]	48.3 [89]	69.5 [109]
GIT (ours)	148.8	123.0	114.4	138.2	69.6	67.5	68.1	180.2	73.9	93.8	56.8	72.8
Δ	+10.1	+3.7	+20.3	+28.5	+9.9	+2.1	+4.0	+59.6	+13.9	+7.3	+8.5	+3.3

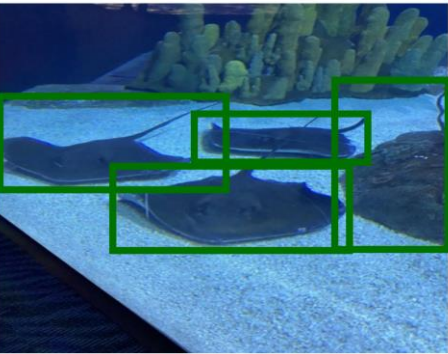
Florence Encoders + Object Detection Adaptor (GLIP)

Achieved SOTA results on zero-shot ODinW

<https://computer-vision-in-the-wild.github.io/eccv-2022/>



Grounding examples:



Grounded Language-Image Pre-training (CVPR 2022)
GLIPv2: Unifying Localization and Vision-Language Understanding (NeurIPS 2022)

Object Detection in the Wild

Organized by: CVinW_2022
Starts on: Dec 31, 2021 4:00:00 PM
Ends on: May 31, 2022 4:59:59 PM

Overview Evaluation Phases Participate Leaderboard

Leaderboard

Please submit your results to the most relevant Phase (See "Evaluation" page for details).

Phase: Zero-Shot, Split: Test Split

Order by metric

B - Baseline * - Private V - Verified

Rank	Participant team	Average Score (t)	Median Score (t)	AerialMaritimeDrone_large (t)	AerialMaritimeDrone_tiled (t)	AmericanSignLanguageLetters_ (t)
1	Florence (FL-1.5-D5)	25.8	14.3	21.1	17.6	1.4
2	DetCLIP-team (DetCLIP)	24.9	18.3	18.3	20.6	4.8
3	GLIPv2_team (GLIPv2-T)	22.3	8.9	14.9	15.9	2.3
4	OmLab (OmDet)	19.7	8.9	10.8	17.8	8.1
5	ODinW_Team (GLIP-T) B	19.6	5.1	13.7	12.6	2.5
6	FIBER (FIBER)	19.5	10.4	17.0	17.4	0.0
7	Google Research (OWL-viT L/14 @ 672)	18.8	9.8	10.5	20.0	2.0

Florence: Pushing Open-World Perception Toward Cognition

- **Open-World Recognition**
 - ❑ Millions of tags
 - ❑ Open-vocabulary search
 - ❑ Object discovery
- **Self-evolving Learning**
- **Leveraging External Knowledge:**
Descriptive, Explainable, Predictive
 - ❑ Story telling
 - ❑ Open question and answer
 - ❑ Video narrator



Florence: Open-world Recognition

Recognized object categories: 20k → millions ...

Species



American white ibis



sunflower hearts



shamu show



roebuck deer

Landmark



Mt rainier Washington



Griffith observatory



BMW headquarter



Snoqualmie ridge

Logo



Microsoft



Honda Logo



usps tracking



Starbucks

Products



capri sun fruit punch case



cambells well yes minestrone with kale soup



barefoot contessa cookbook



dove sensitive skin beauty bar

Celebrity



jean reno



chalize theron



dwade



elon musk

Movie



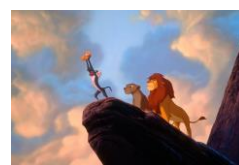
the return of the jedi



on strange tides, pirates of the caribbean

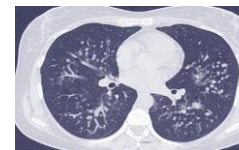


avengers trails



the lion king movie

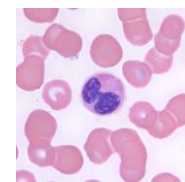
Medical



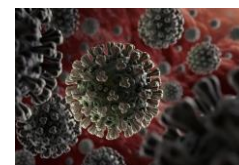
chest CT



abdominal organs



monoblasts



virus

Artworks



romanian glassware



wooden statue

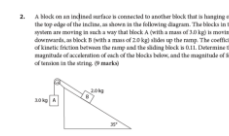


along the river during the qingming festival



irises painting

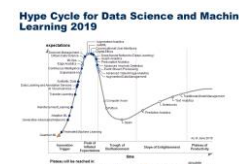
Documents



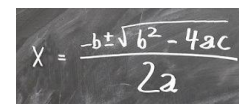
free body diagram



dock receipt



gartner hype curve



ecuaciones algebraicas



Florence: Open-world Recognition

Recognized object categories: 20k → millions ...

Species



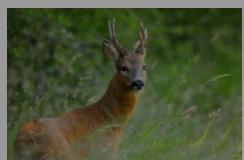
American white ibis



sunflower hearts

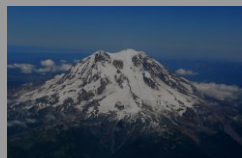


shamu show

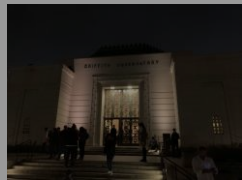


roe buck deer

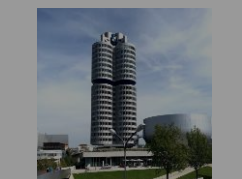
Landmark



Mt rainier Washington



Griffith observatory

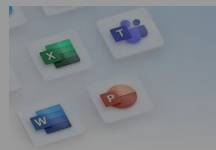


BMW headquarter



Snoqualmie ridge

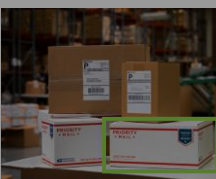
Logo



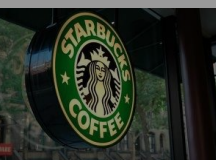
Microsoft



Honda Logo



usps tracking



Starbucks

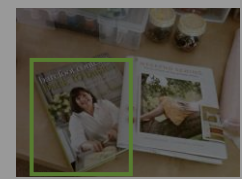
Products



capri sun fruit punch case



cambells well yes minestrone with kale soup

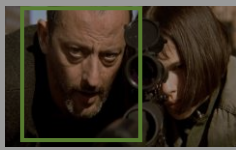


barefoot contessa cookbook



dove sensitive skin beauty bar

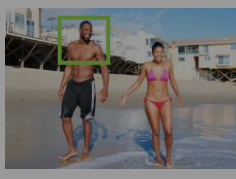
Celebrity



jean reno



chalize theron



dwade

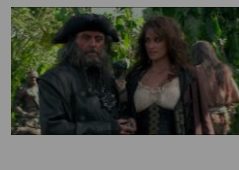


elon musk

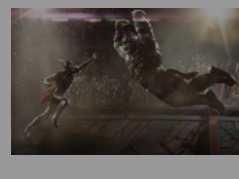
Movie



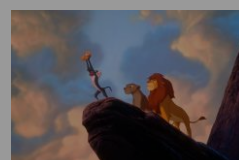
the return of the jedi



on strange tides, pirates of the caribbean

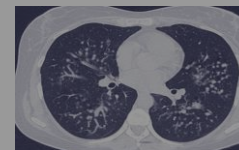


avengers trails

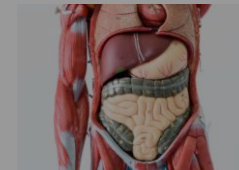


the lion king movie

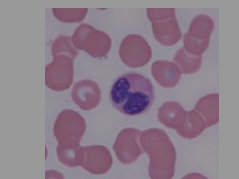
Medical



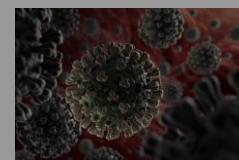
chest CT



abdominal organs



monoblasts



virus

Artworks



romanian glassware



wooden statue

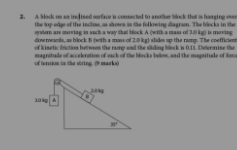


along the river during the qingming festival



irises painting

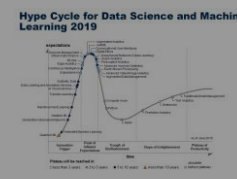
Documents



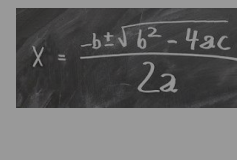
free body diagram



dock receipt



gartner hype curve



ecuaciones algebraicas

Florence: Open-world Recognition

Recognized object categories: 20k → millions ...

Species



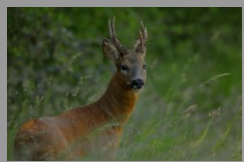
American white ibis



sunflower hearts



shamu show

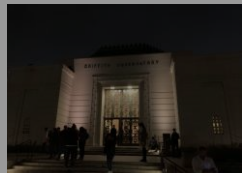


roebuck deer

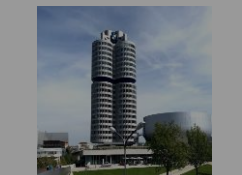
Landmark



Mt rainier Washington



Griffith observatory

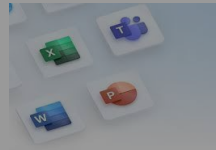


BMW headquarter



Snoqualmie ridge

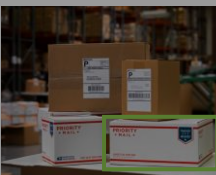
Logo



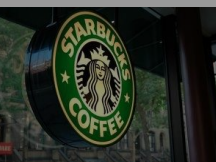
Microsoft



Honda Logo



usps tracking



Starbucks

Products



capri sun fruit punch case



cambells well yes minestrone with kale soup

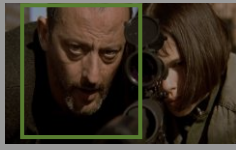


barefoot contessa cookbook



dove sensitive skin beauty bar

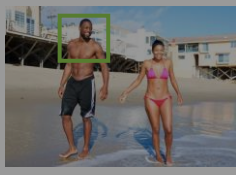
Celebrity



jean reno



chalize theron



dwade

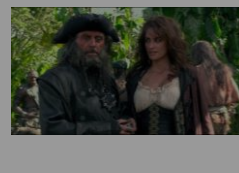


elon musk

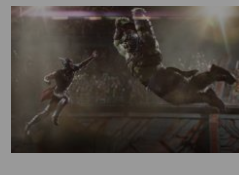
Movie



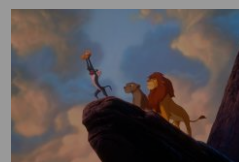
the return of the jedi



on strange tides, pirates of the caribbean

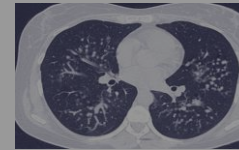


avengers trails



the lion king movie

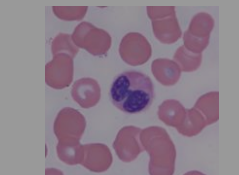
Medical



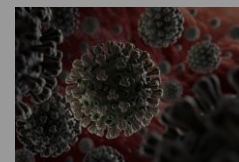
chest CT



abdominal organs



monoblasts

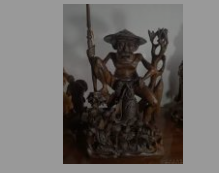


virus

Artworks



romanian glassware



wooden statue

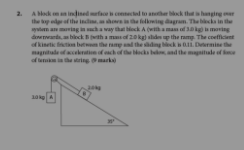


along the river during the qingming festival



irises painting

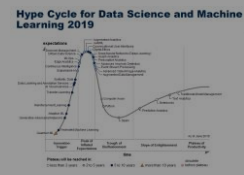
Documents



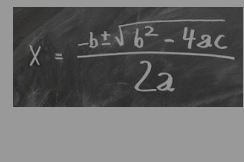
free body diagram



dock receipt



gartner hype curve



ecuaciones algebraicas

Florence: Open-world Recognition

Recognized object categories: 20k → millions ...

Species

Landmark

Logo

Products

Celebrity


Movie

Medical


Artworks

Documents

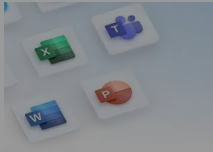
»




American white ibis



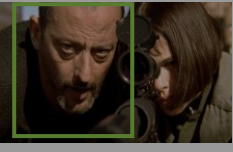
Mt rainier Washington




Microsoft



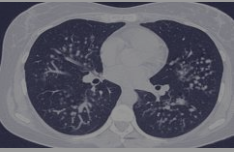
capri sun fruit punch case




jean reno



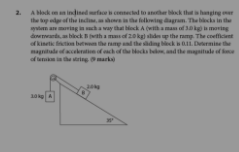
the return of the jedi




chest CT



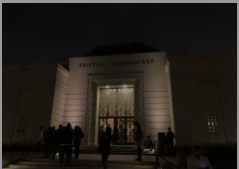
romanian glassware




free body diagram




sunflower hearts




Griffith observatory



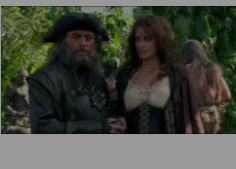
Honda Logo




cambells well yes minestrone with kale soup




chalize theron




on strange tides, pirates of the caribbean




abdominal organs




wooden statue




dock receipt




shamu show




BMW headquarter



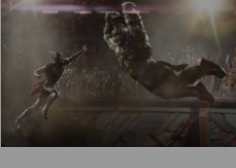
usps tracking



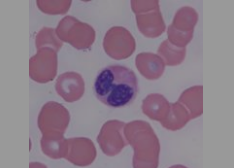
barefoot contessa cookbook




dwade



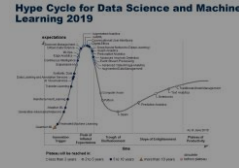
avengers trails




monoblasts




along the river during the qingming festival




gartner hype curve




roebuck deer




Snoqualmie ridge




Starbucks



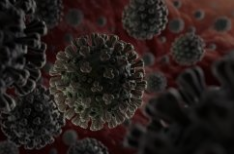
dove sensitive skin beauty bar




elon musk



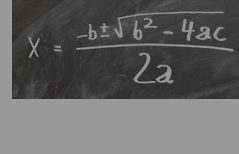
the lion king movie



virus



irises painting



ecuaciones algebraicas

Florence: Semantics

Text-image retrieval



Search:

“Microsoft”

Associate:

“Windows”, “etc.” (Microsoft products) to the query words

Object discovery



Search:

“Game without age restriction”

Search:

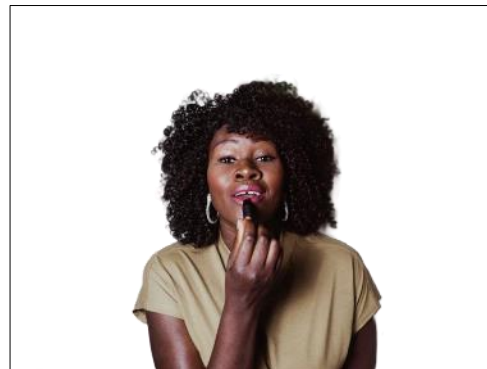
“Game for kids at least 10 years old”

Search:

“Game for teen”

State-of-the-Art Human Matting powered by Florence

Trained on 2M human matting data using pre-trained Florence visual encoder



Expanding from Human to Broader Categories

Florence pre-training empowers zero-shot segmentation ability



Sunflowers



Bambootula Spider



Red Grape



iPhone XR

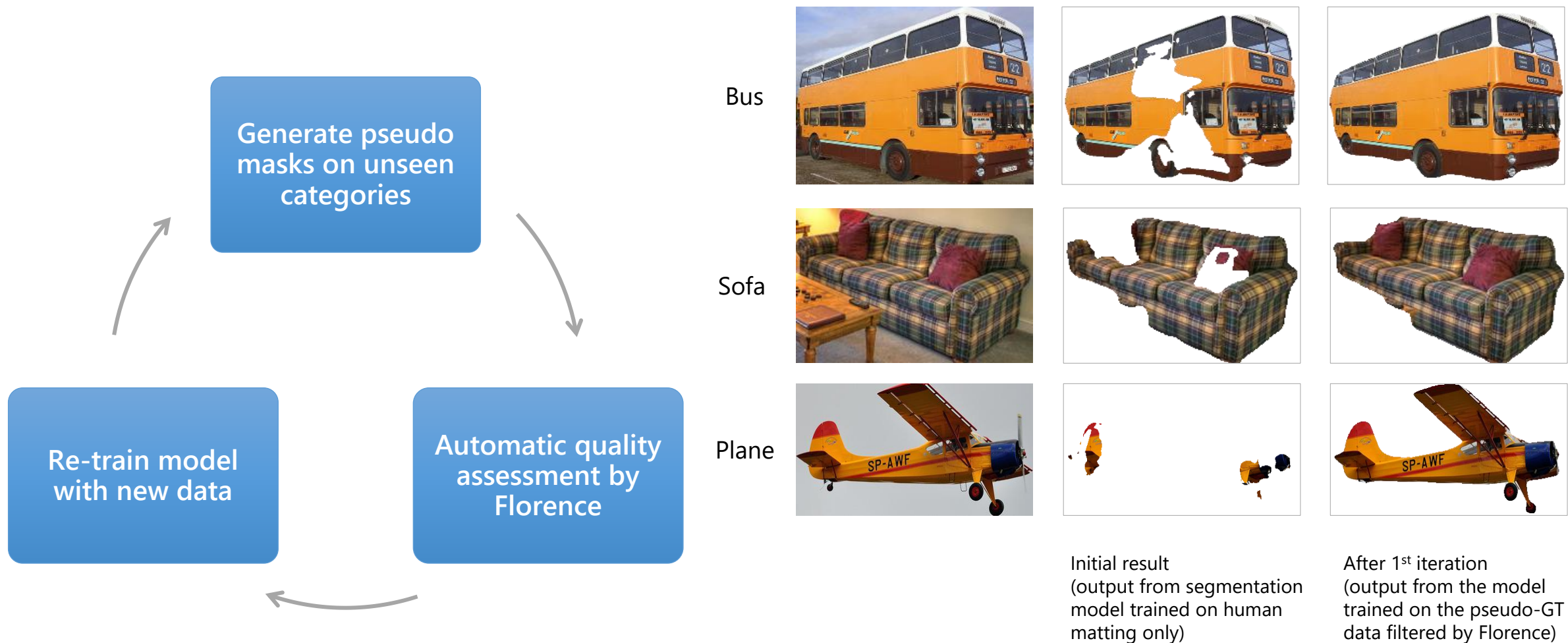


Veggie Burgers

Output from a Florence segmentation model only trained on human matting data

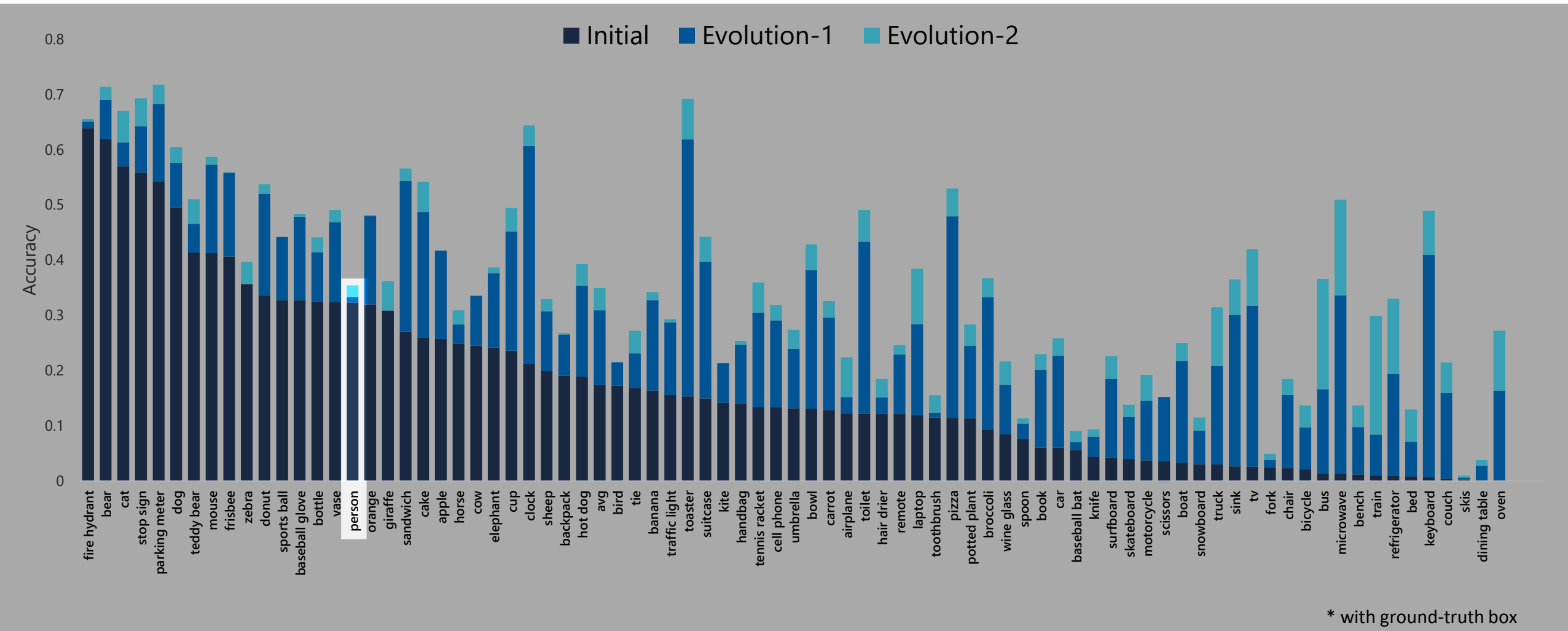
A Self-Evolving Learning System for Segmentation

Explicitly expands the segmentation ability to unseen categories



Florence segmentation self-evolves

Evaluated on COCO instance segmentation



* with ground-truth box

Florence Creative AI Capability: Story-Telling



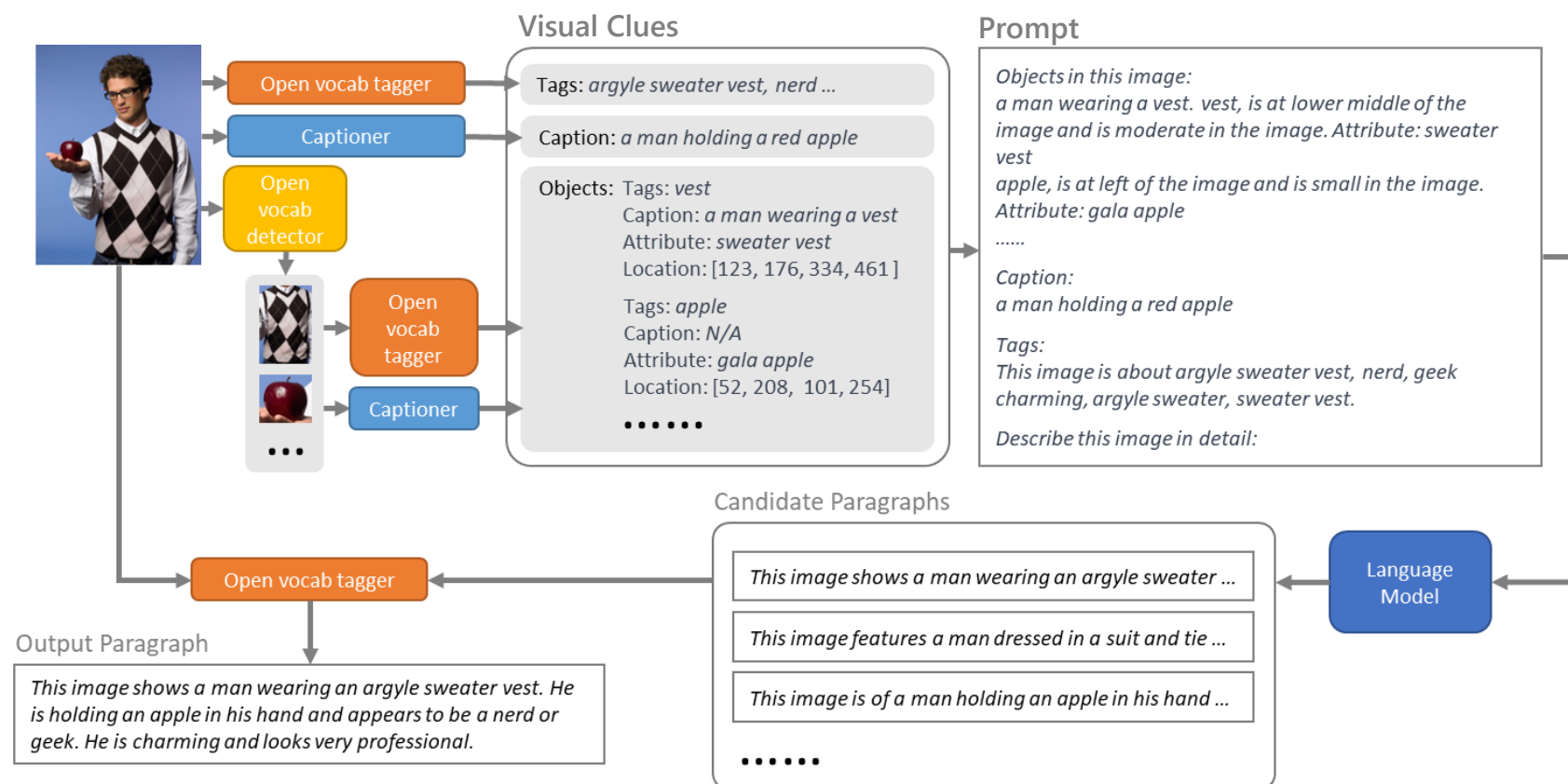
"The image is of a cover of The Economist magazine. The headline is "March of the machines". The cover features a drawing of the Maschinenmensch, or machine man, in the center. The background is art deco in style. This image represents the rise of robots and artificial intelligence, as they become increasingly important in our society."

Florence Creative AI Capability: Story-Telling



"The image is of a cover of The Economist magazine. The image features a roller coaster in the clouds. The roller coaster is red and blue and has people on it. Above the roller coaster are the words " When the ride ends". The roller coaster in the clouds is a metaphor for the economy. It's a fun, exciting ride that everyone loves until it crashes down to earth, causing economic loss and recession. A market crash is the final nail in the coffin, leaving people reeling in its wake."

Story-Telling: Florence + GPT-3



Visual clues:

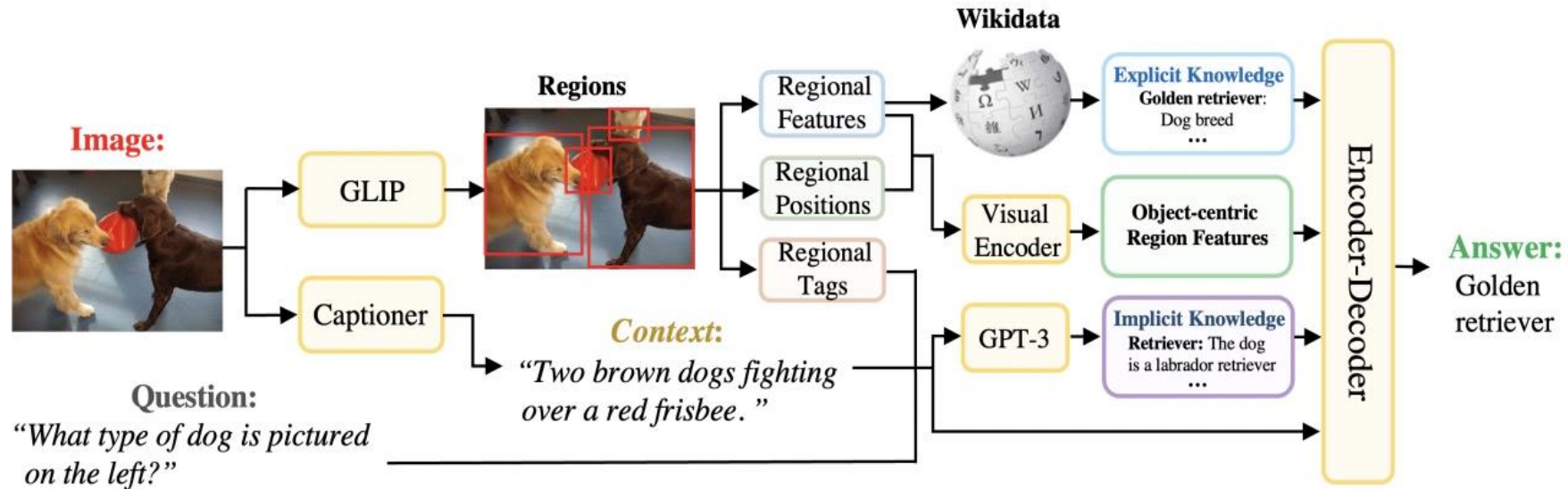
- A holistic and composable representation of the image.
- A natural bridge between vision and language foundation models.
- Interpretable, not only for humans, but also for machines.

Bridging with Explicit Structured Textual clue:

- Easy implementation with no extra training.
- Premium language quality.
- Versatile to different applications.

Florence: Knowledge-based Visual Question & Answer

SOTA in OK-VQA: leverage regional representations for retrieving external knowledge



Previous SOTA

- Retrieve various of external knowledge in a simple **sliding window manner**
- Use question + retrieved knowledge to answer the question (**no visual features!**)
- Local visual features are important in retrieving external knowledge
 - **Regional features** to retrieve external knowledge, e.g., from Wikidata
 - **Regional descriptions** to obtain implicit knowledge, e.g., using GPT-3
- The final answering model should look at the image thoroughly
 - Extended language encoder-decoder model to **incorporate the regional features and region coordinates**.

Florence: Knowledge-based Visual Question & Answer



Q: What is on this sandwich?


C: A man eating a sandwich. sandwich, snack food, food, person, outdoor

A: Cheese

GT: ['Cheese', 'Cheese', 'Cheese', 'Cheese', 'Cheese', 'Cheese', 'Cheese', 'Cheese', 'Omelet', 'Omelet']

Acc.: 1.0

Object Regions



Explicit Knowledge

Shelpek:
Kazakh flatbread, using butter, milk and sugar

...

Cheeseburger:
Hamburger topped with cheese

Implicit Knowledge

Cheese: The cheese is the most important part of the sandwich

...

Cheddar: The cheese is the main ingredient in the sandwich



Q: What breed of dog is this?


C: A brown dog laying on a couch with blankets. mammal, wall, dog, sofa, floor

A: Terrier

GT: ['Terrier', 'Terrier', 'Terrier', 'Terrier', 'Crossbreed', 'Crossbreed', 'Pit bull', 'Pit bull', 'Shepard', 'Shepard']

Acc.: 1.0

Object Regions



Explicit Knowledge

Brazilian Terrier:
Dog breed

...

Comforter:
Type of bedcover, often not as thick as a duvet

Implicit Knowledge

Poodle: The dog is a poodle

...

Terrier: The dog is a terrier

Integrative Multi-modality: Video Narrator

Automatically generate the narration of the video and its neural synthesized speech



Q & A

Thanks!