JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

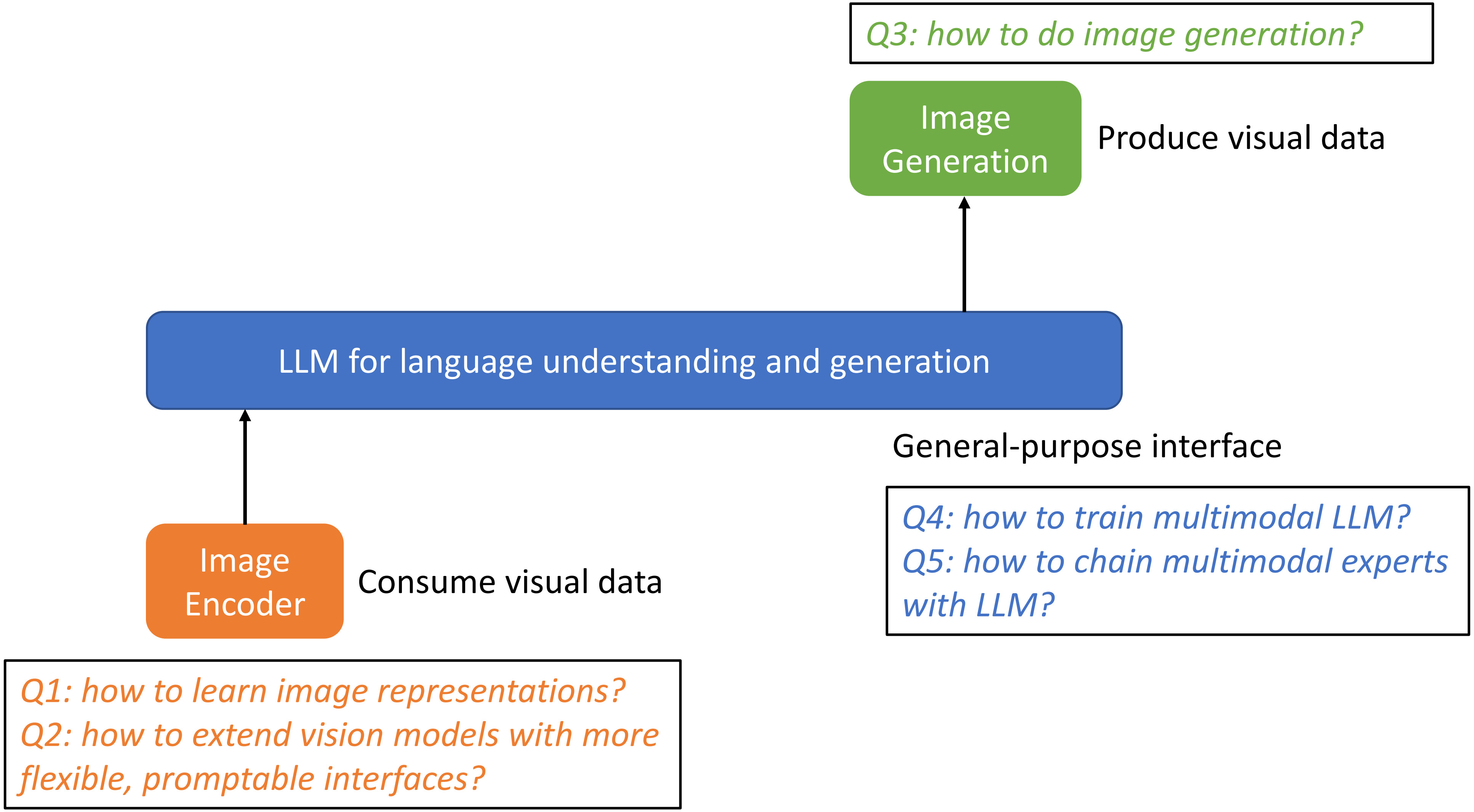
Microsoft

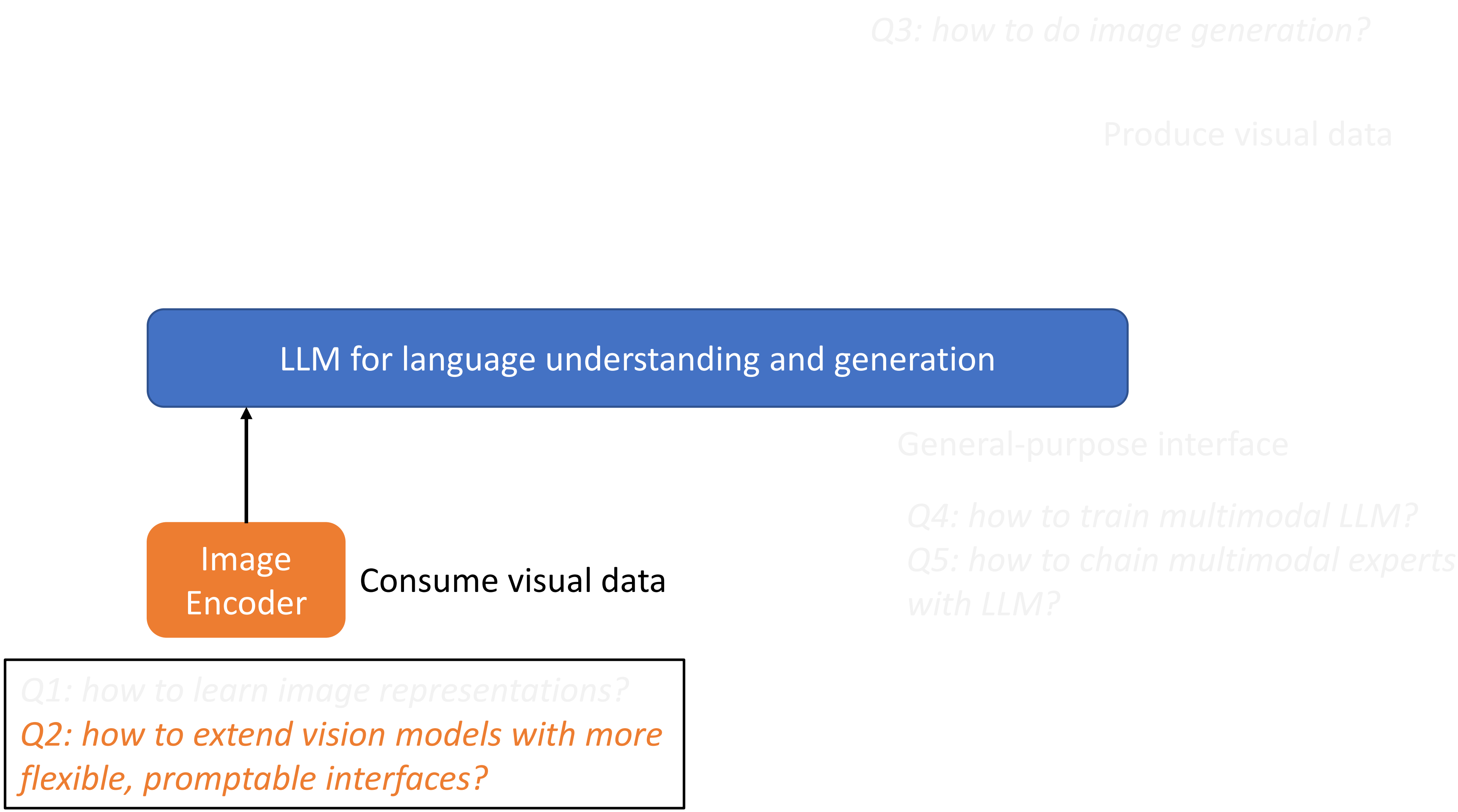
From Specialist to Generalist: Towards General Vision Understanding Interface

Jianwei Yang

Microsoft Research

06/19/2023

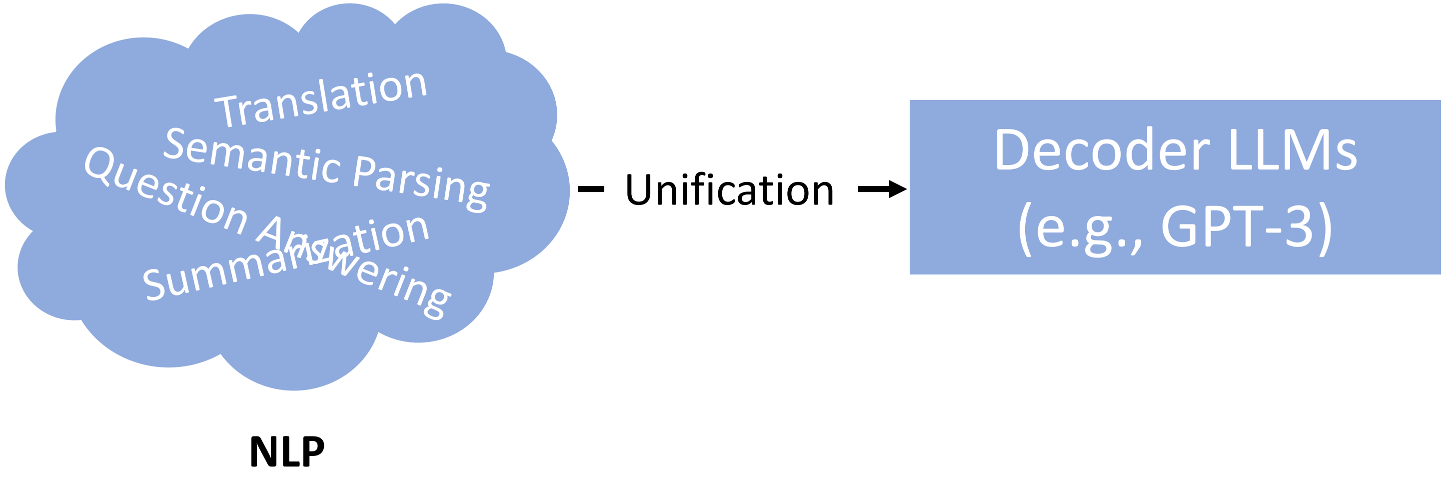

Q3: how to do image generation?
Image Generation
Produce visual data
LLM for language understanding and generation
General-purpose interface
Q4: how to train multimodal LLM?
Q5: how to chain multimodal experts with LLM?
Image Encoder
Consume visual data
Q1: how to learn image representations?
Q2: how to extend vision models with more flexible, promptable interfaces?


Q3: how to do image generation?
Produce visual data
LLM for language understanding and generation
General-purpose interface
Q4: how to train multimodal LLM?
Q5: how to chain multimodal experts with LLM?
Image Encoder
Consume visual data
Q1: how to learn image representations?
Q2: how to extend vision models with more flexible, promptable interfaces?

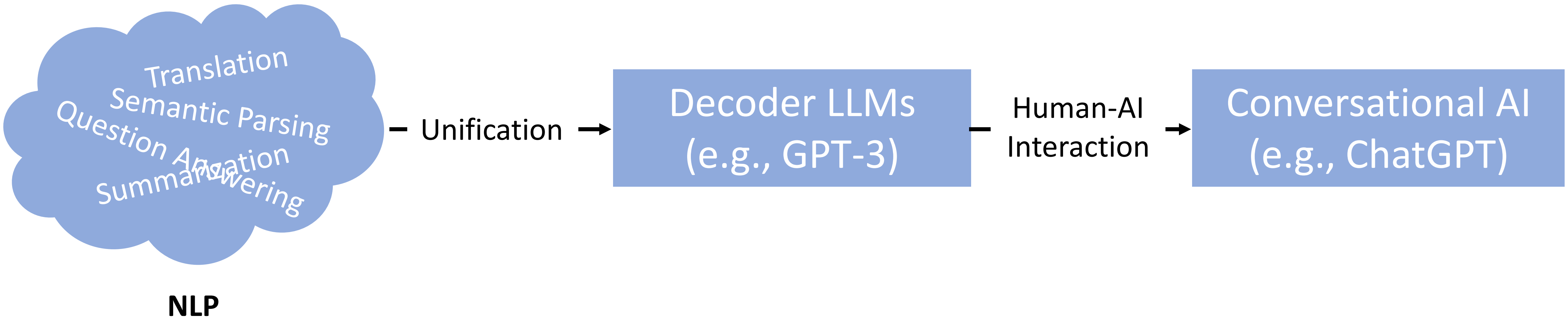
A Lesson from LLMs

NLP

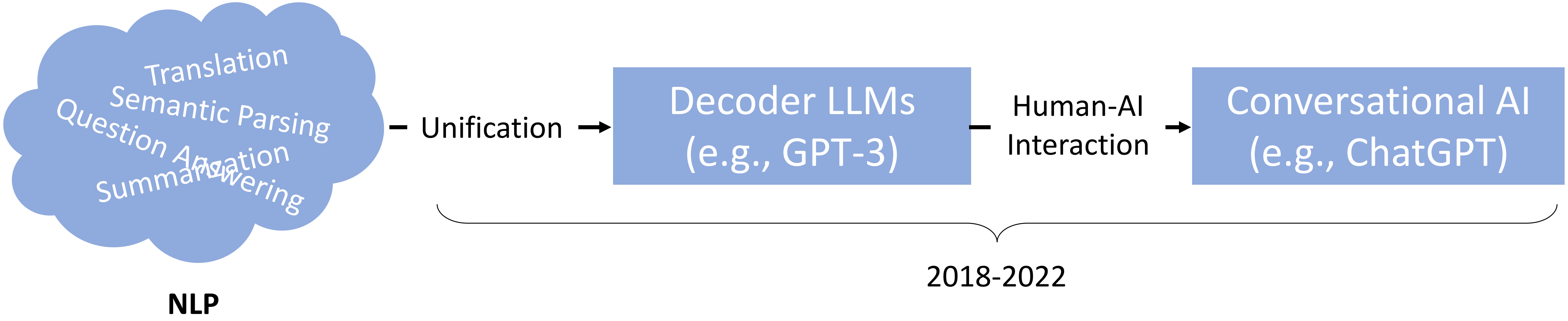
A Lesson from LLMs



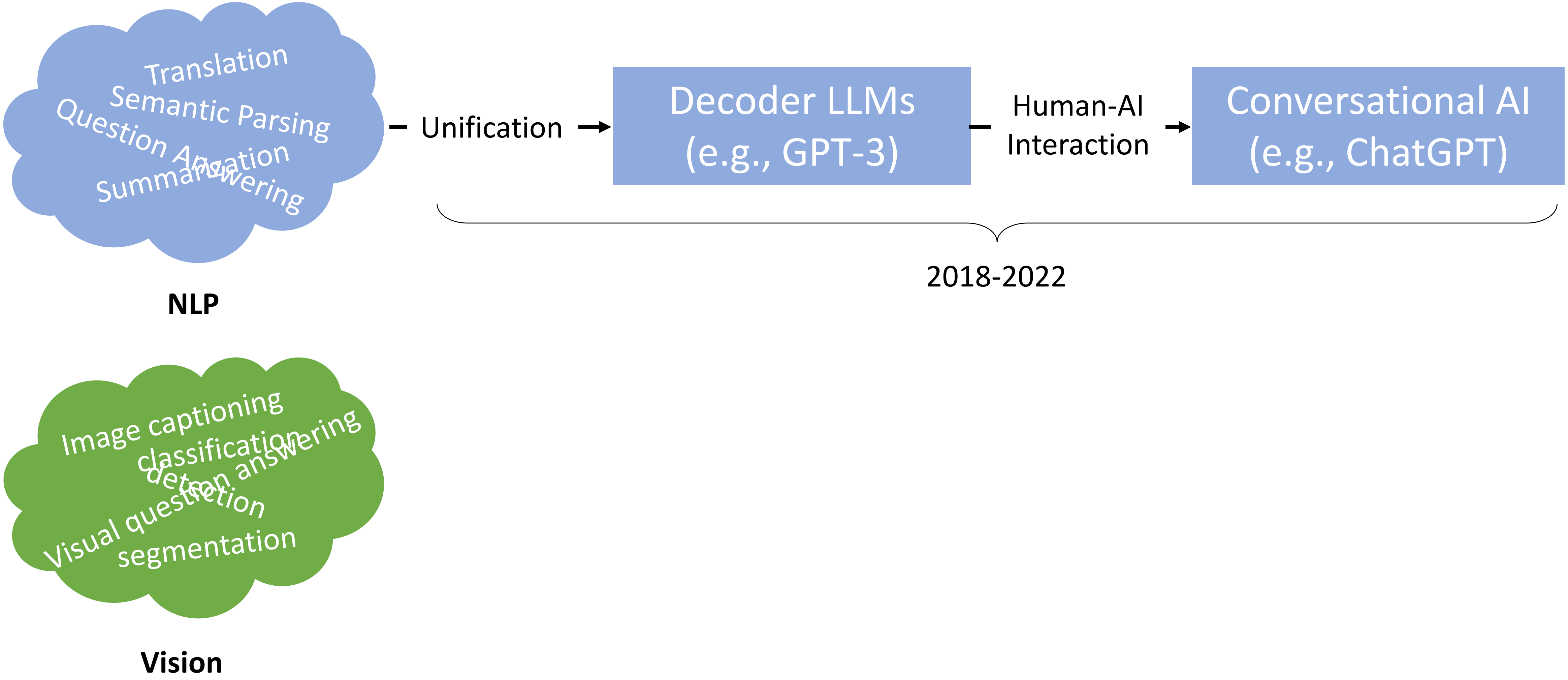
A Lesson from LLMs



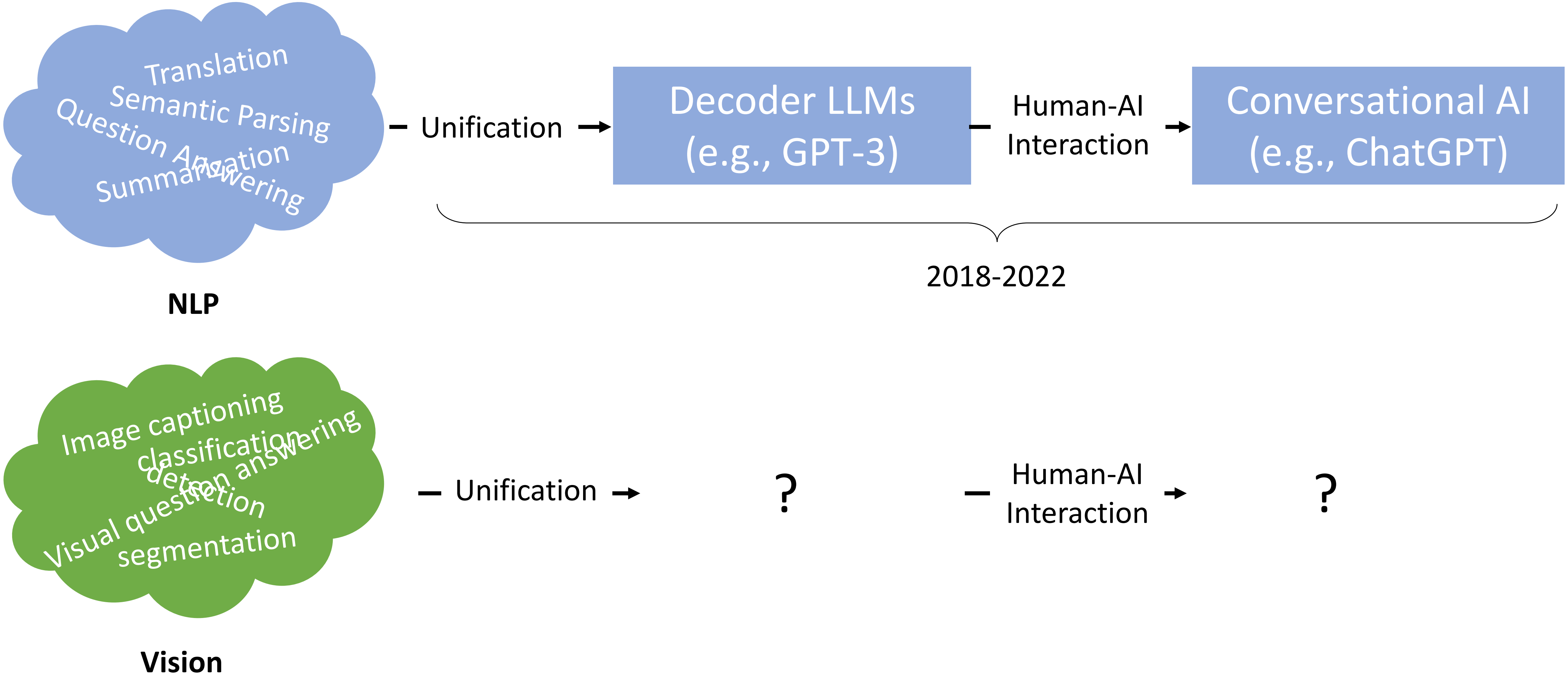
A Lesson from LLMs



A Lesson from LLMs



A Lesson from LLMs



Unique Challenges in Vision: Modeling

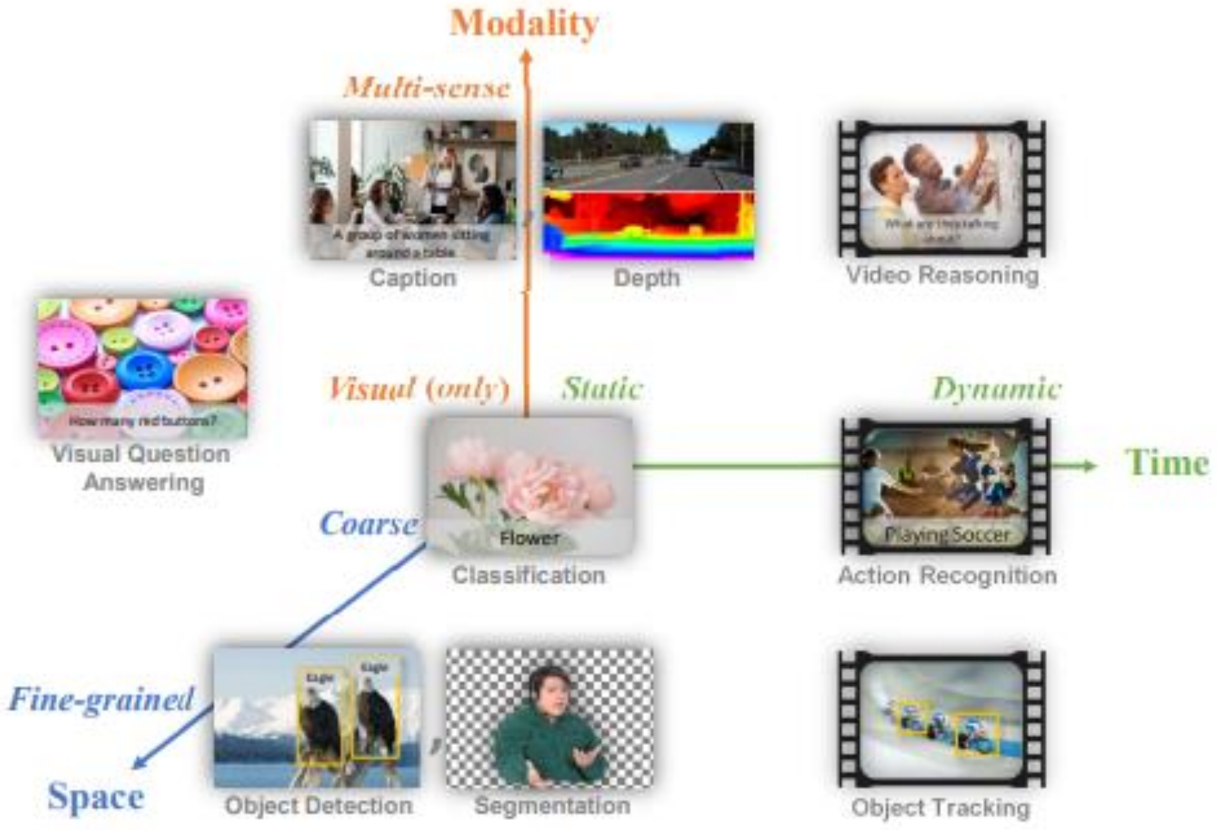


Image Source: Project Florence

Unique Challenges in Vision: Modeling

a) Different types of inputs:

Temporality: static image, video sequence

Multi-modality: w/text, w/audio, etc.

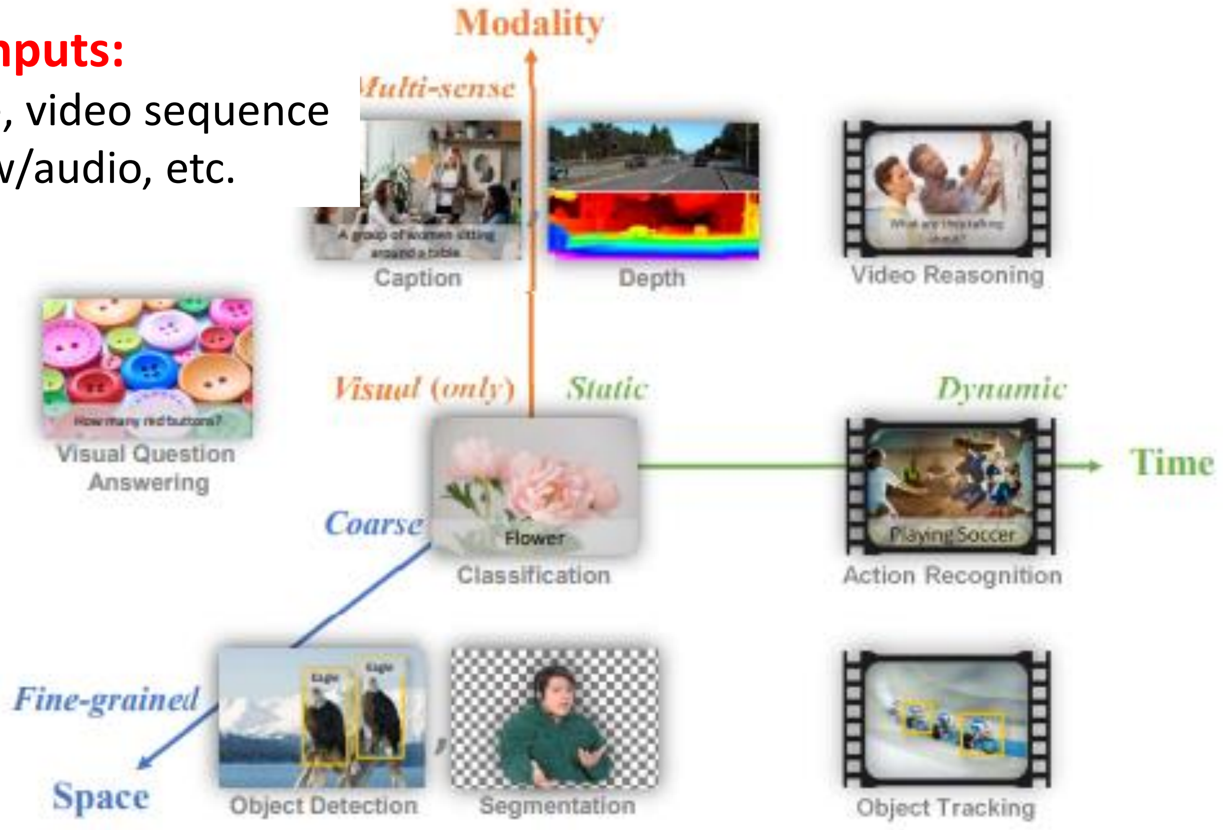


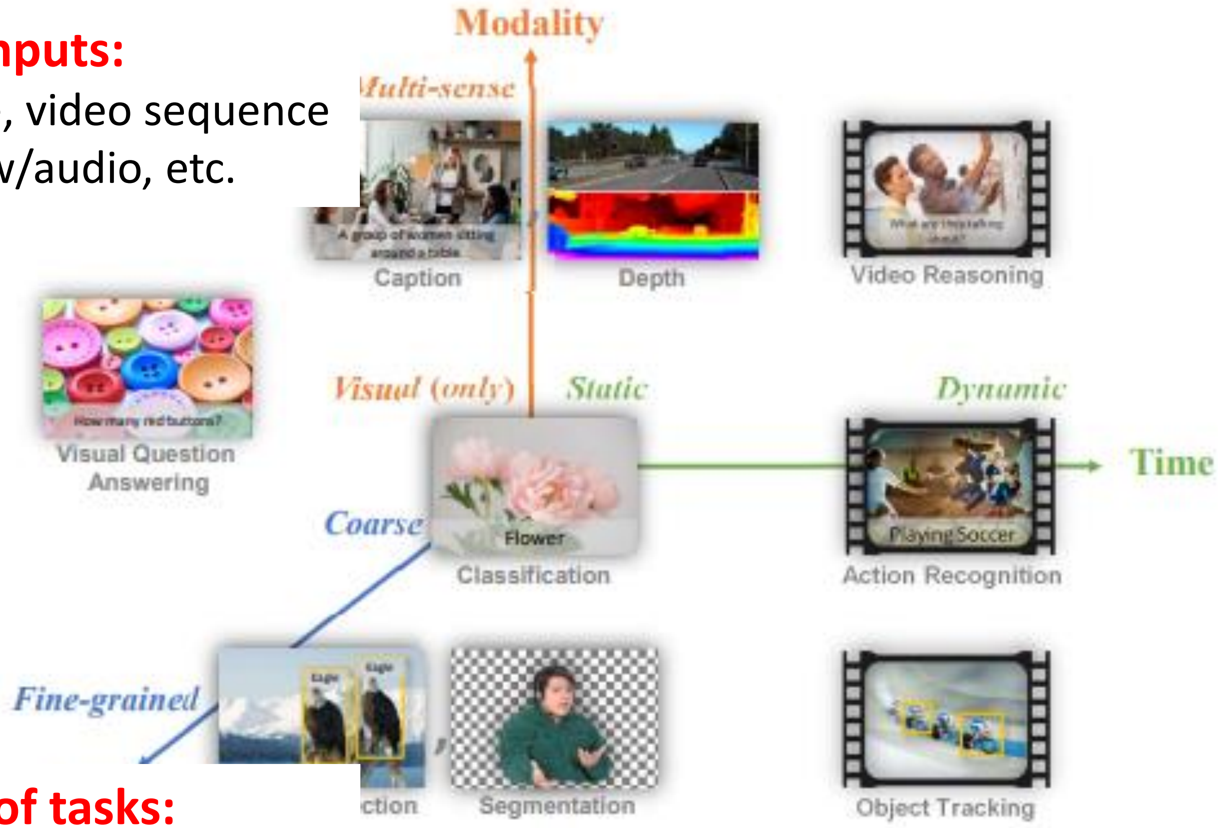
Image Source: Project Florence

Unique Challenges in Vision: Modeling

a) Different types of inputs:

Temporality: static image, video sequence

Multi-modality: w/text, w/audio, etc.



b) Different granularities of tasks:

Image-level: classification, captioning, etc.

Region-level: object detection, grounding, etc.

Pixel-level: segmentation, depth, SR, etc.

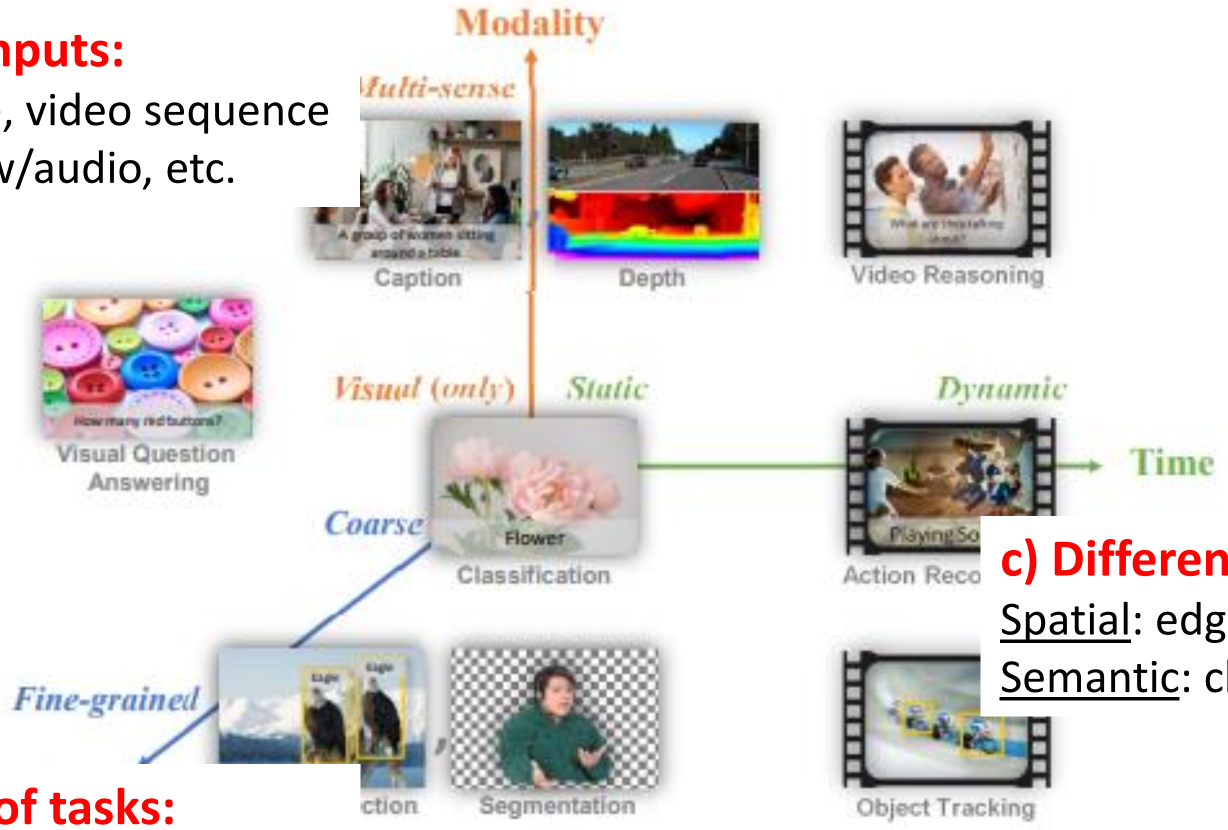
Image Source: Project Florence

Unique Challenges in Vision: Modeling

a) Different types of inputs:

<u>Temporality</u>: static image, video sequence

<u>Multi-modality</u>: w/text, w/audio, etc.



b) Different granularities of tasks:

<u>Image-level</u>: classification, captioning, etc.

<u>Region-level</u>: object detection, grounding, etc.

<u>Pixel-level</u>: segmentation, depth, SR, etc.

c) Different types of outputs:

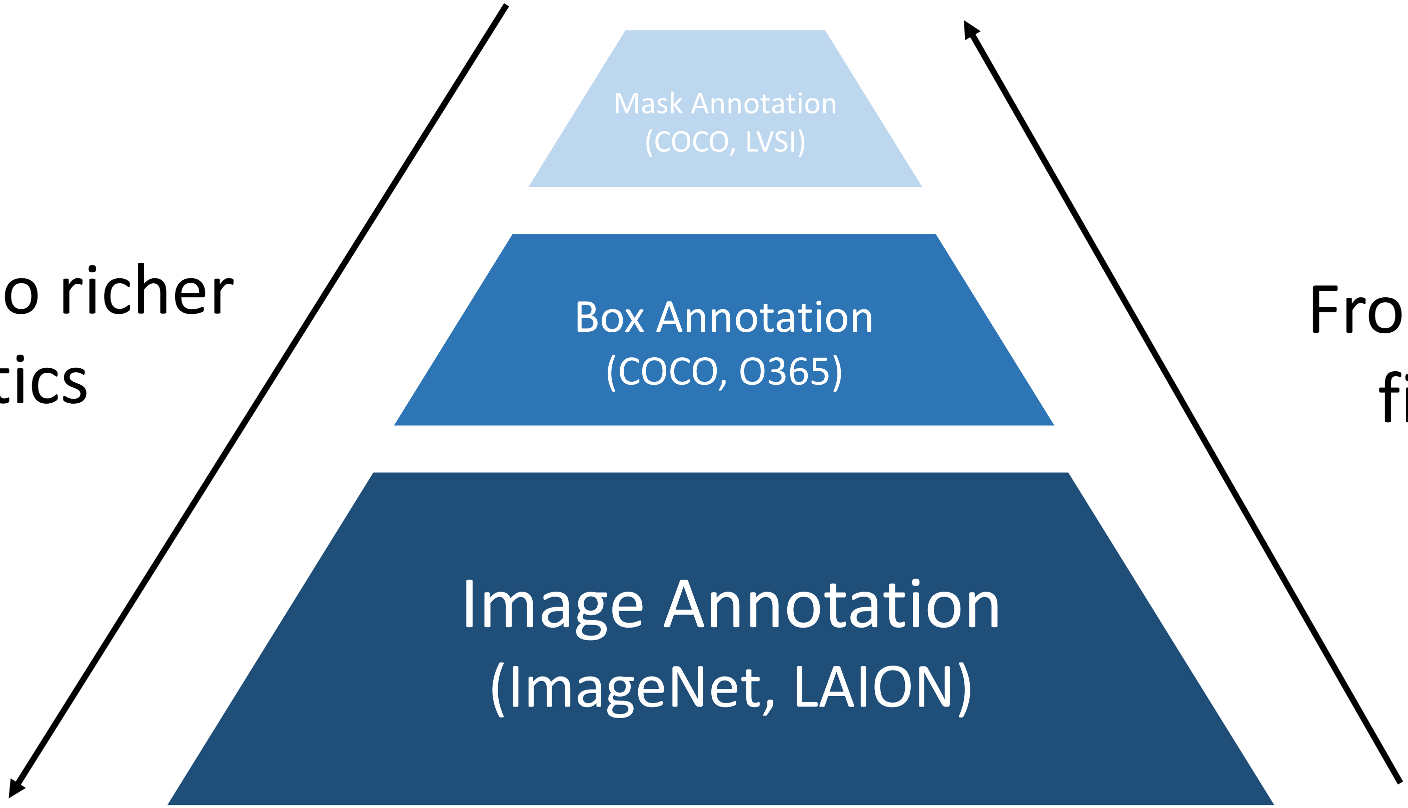
<u>Spatial</u>: edges, boxes, masks, etc.

<u>Semantic</u>: class labels, descriptions, etc.

Image Source: Project Florence

Unique Challenges in Vision: Data

From poor to richer semantics



From coarse to finer grain

Scales differ significantly across different types of annotations

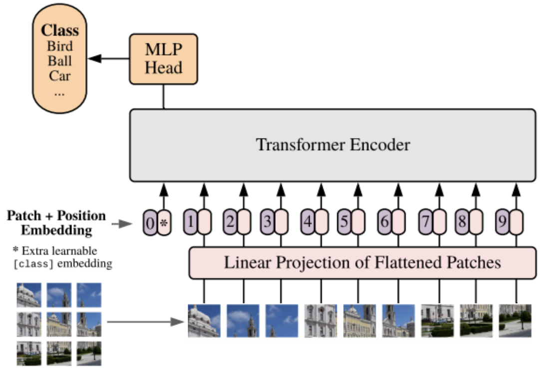
Clear Attempts towards General Vision

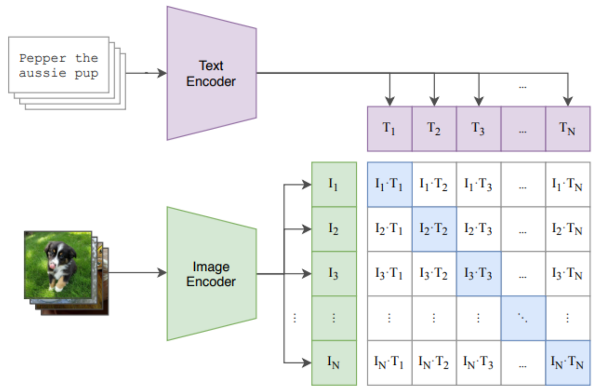
Clear Attempts towards General Vision

Closed-set Classification → Open-world Recognition

AlexNet[1], *ResNet*[2], *ViT*[3]

CLIP[4], *ALIGN*[5], *FLORENCE*[6]





[1] Krizhevsky et al. "Imagenet classification with deep convolutional neural networks.“. *NeurIPS* 2012
[2] He et al. "Deep residual learning for image recognition." *CVPR* 2016.
[3] Dosovitskiy et al. "An image is worth 16x16 words: Transformers for image recognition at scale.“ *ICLR 2021*.
[4] Radford et al. Learning transferable visual models from natural language supervision, *ICML* 2021
[5] Jia et al. "Scaling up visual and vision-language representation learning with noisy text supervision." *ICML* 2021.
[6] Yuan et al. "Florence: A new foundation model for computer vision." *arXiv 2021*.

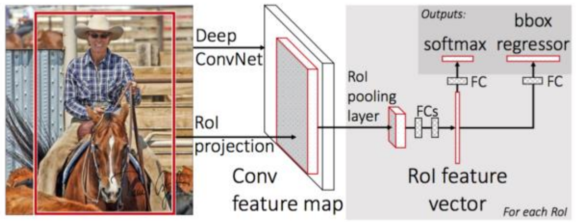
Clear Attempts towards General Vision

Closed-set Classification → Open-world Recognition

Specialist Models → Generalist Models

Detection[1], Segmentation[2], VQA[3]

Pixel2Seqv2[4], UniTAB[5], OFA[6], Unified-IO[7], X-Decoder[8]



[1] Girshick. "Fast r-cnn." *CVPR* 2015.
[2] He et al. "Mask r-cnn." *ICCV* 2017.
[3] Antol et al. "Vqa: Visual question answering." *ICCV* 2015.
[4] Chen et al. "A unified sequence interface for vision tasks." *NeurIPS 2022*.
[5] Yang et al. "Unitab: Unifying text and box outputs for grounded vision-language modeling." *ECCV 2022*.
[6] Wang et al. "Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework." *ICML* 2022.
[7] Lu et al. "Unified-io: A unified model for vision, language, and multi-modal tasks." *ICLR* 2022.
[8] Zou et al. "Generalized decoding for pixel, image, and language." *CVPR* 2023.

Clear Attempts towards General Vision

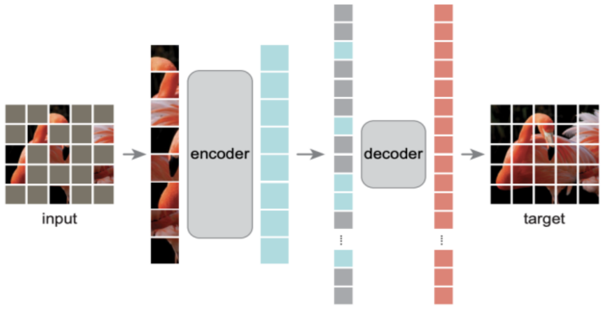
Closed-set Classification → Open-world Recognition

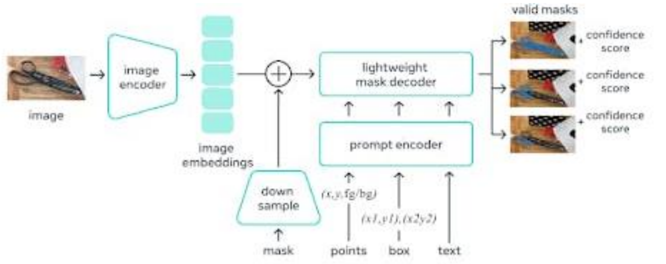
Specialist Models → Generalist Models

Representation Learning → Promptable Interface

BEIT[1], MAE[2], DINO[3]

SAM[4], SegGPT[5], SEEM[6]





[1] Bao et al. BEiT: BERT Pre-Training of Image Transformers, ICLR 2022.
[2] He et al. "Masked autoencoders are scalable vision learners." *CVPR* 2022..
[3] Caron et al. "Emerging properties in self-supervised vision transformers." *ICCV* 2021.
[4] Kirillov et al. "Segment anything." *arXiv* 2023.
[5] Wang et al. "Seggpt: Segmenting everything in context." *arXiv* 2023.
[6] Zou et al. "Segment everything everywhere all at once." *arXiv* 2023.

Clear Attempts towards General Vision

Closed-set Classification → Open-world Recognition

Specialist Models → Generalist Models

Representation Learning → Promptable Interface

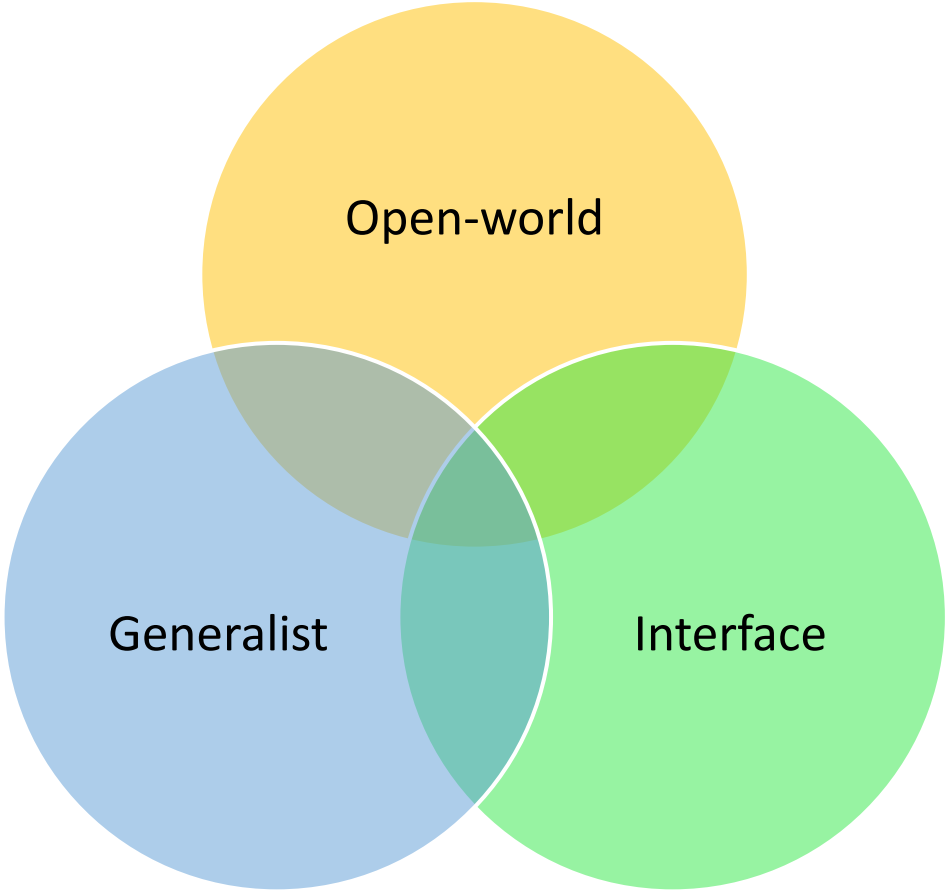
Clear Attempts towards General Vision

Open-world Recognition

Generalist Models

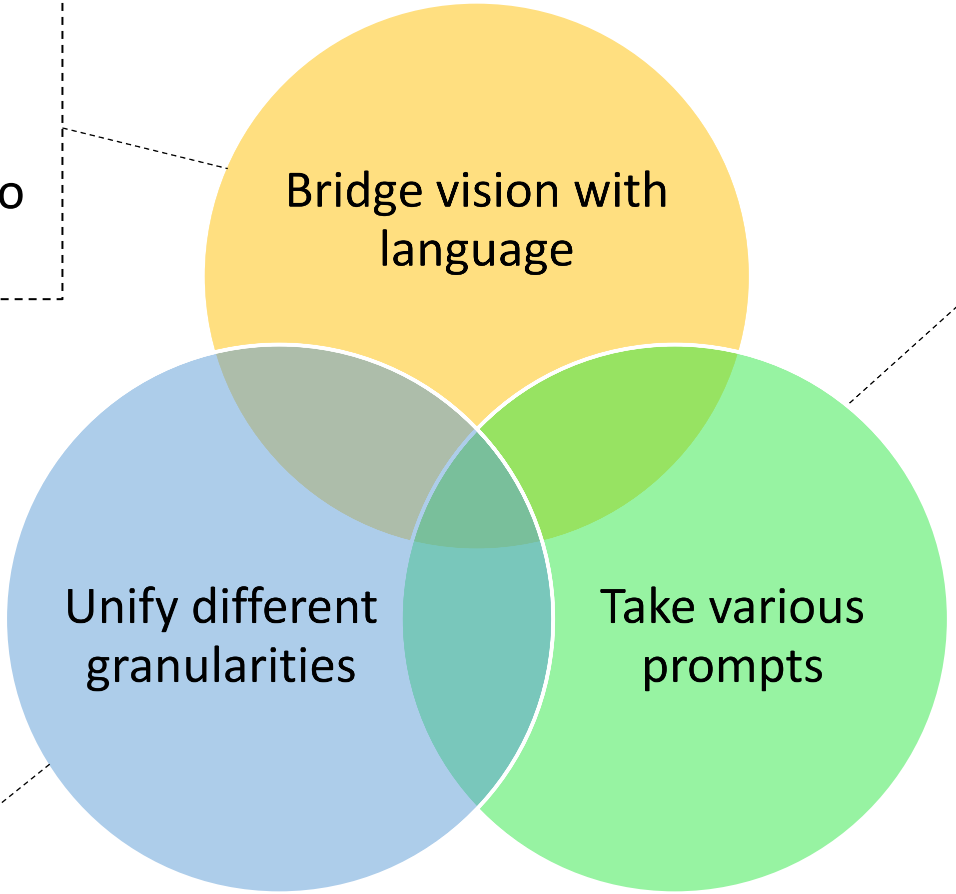
Promptable Interface

In this talk



In this talk

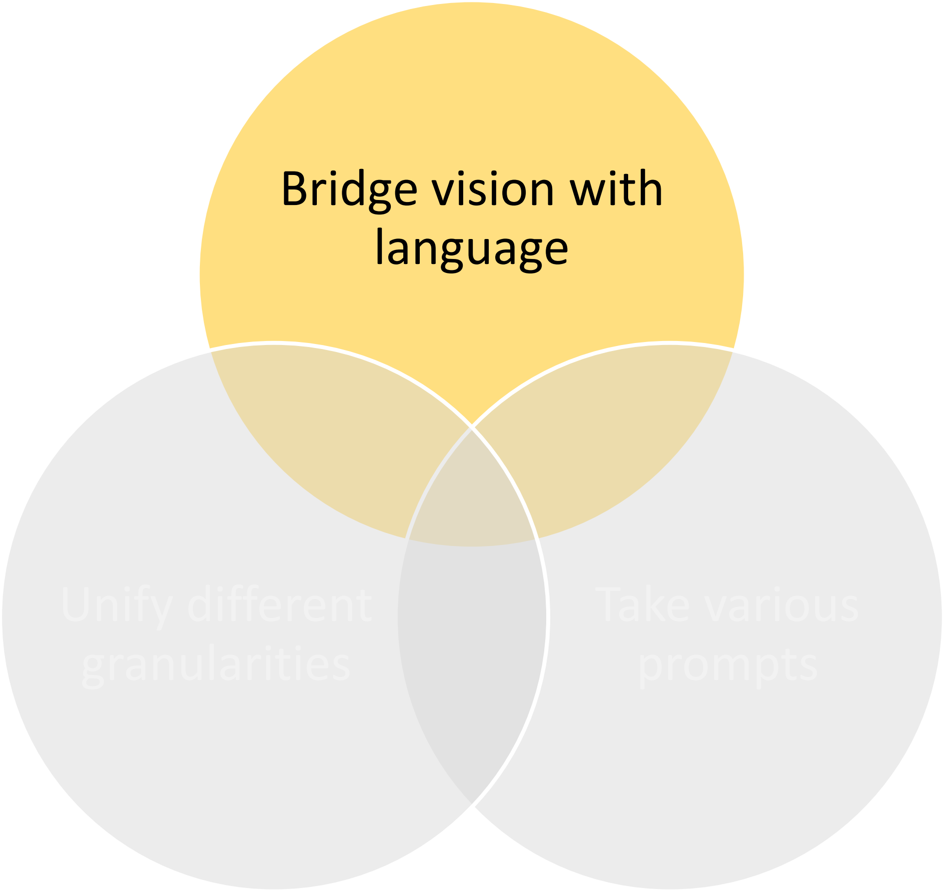
Intuition: language as the common space to share information
Benefit: Zero-shot transfer to novel vocabularies



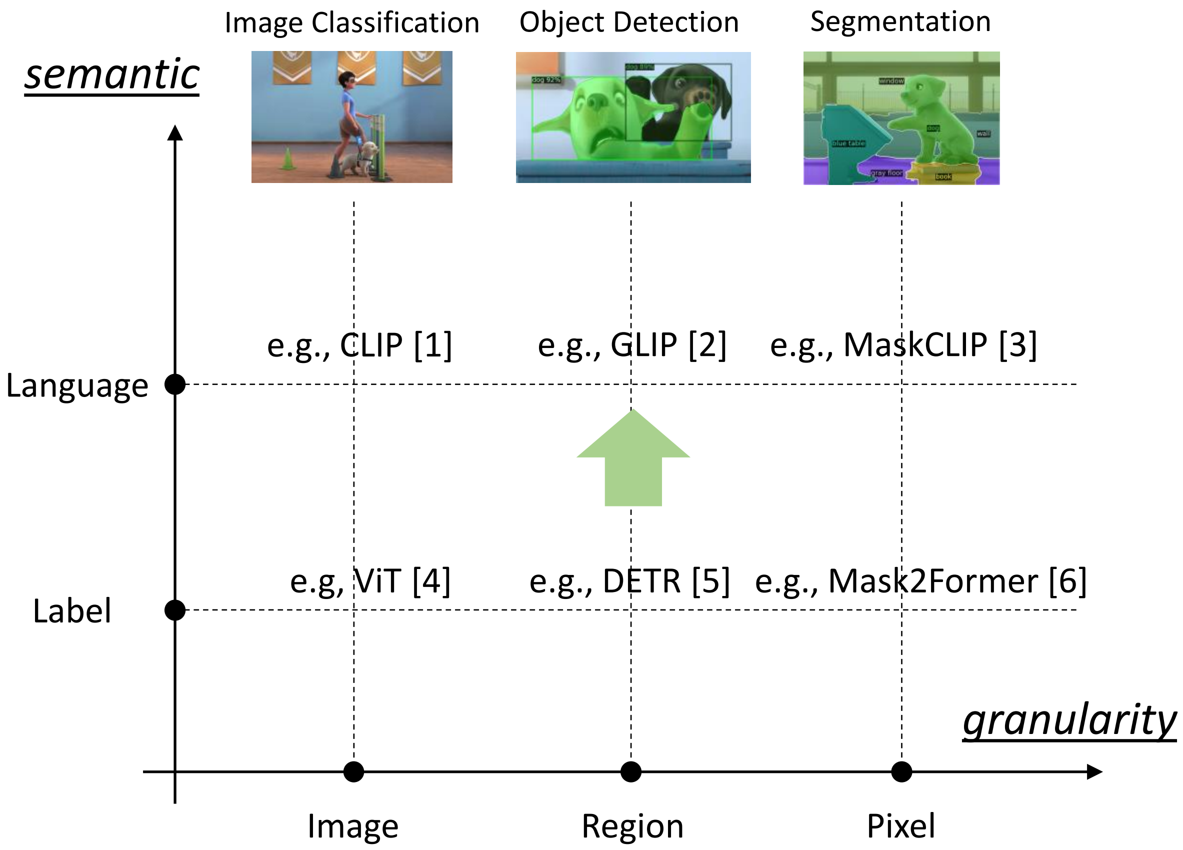
Intuition: language, spatial prompts and beyond
Benefit: Reduce the ambiguity of expressing human intents

Intuition: vision is multi-task, multi-granularity
Benefit: Build synergy across task granularities

I. Bridge Vision with Language

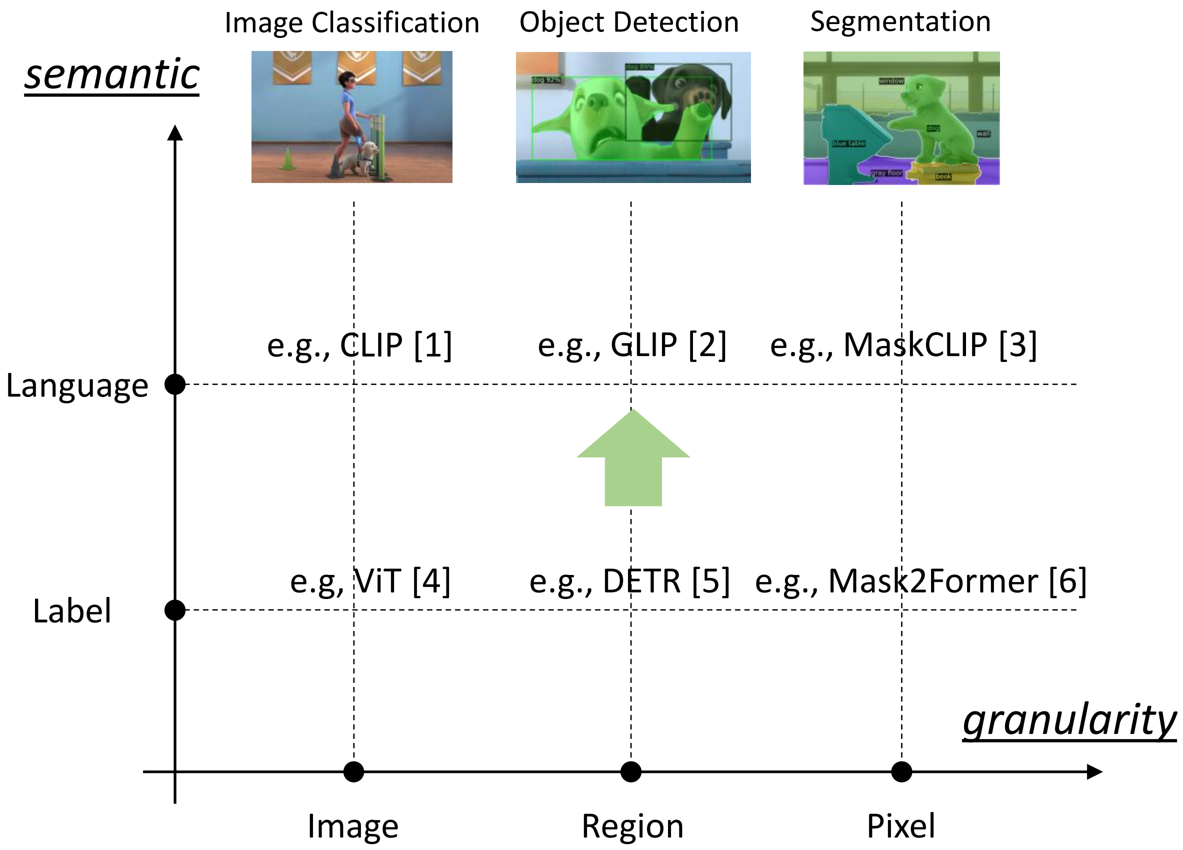


Bridge Vision with Language



[1] Radford et al. "Learning transferable visual models from natural language supervision." ICML, PMLR, 2021
[2] Li et al. "Grounded language-image pre-training." CVPR, 2022
[3] Zhou et al. "Extract Free Dense Labels from CLIP." ECCV, 2022
[4] Dosovitskiy et al. “An image is worth 16x16 words: Transformers for image recognition at scale.” *ICLR, 2021*
[5] Carion et al. "End-to-end object detection with transformers." *ECCV, 2020*
[6] Cheng et al. "Masked-attention mask transformer for universal image segmentation." *CVPR. 2022*

Bridge Vision with Language

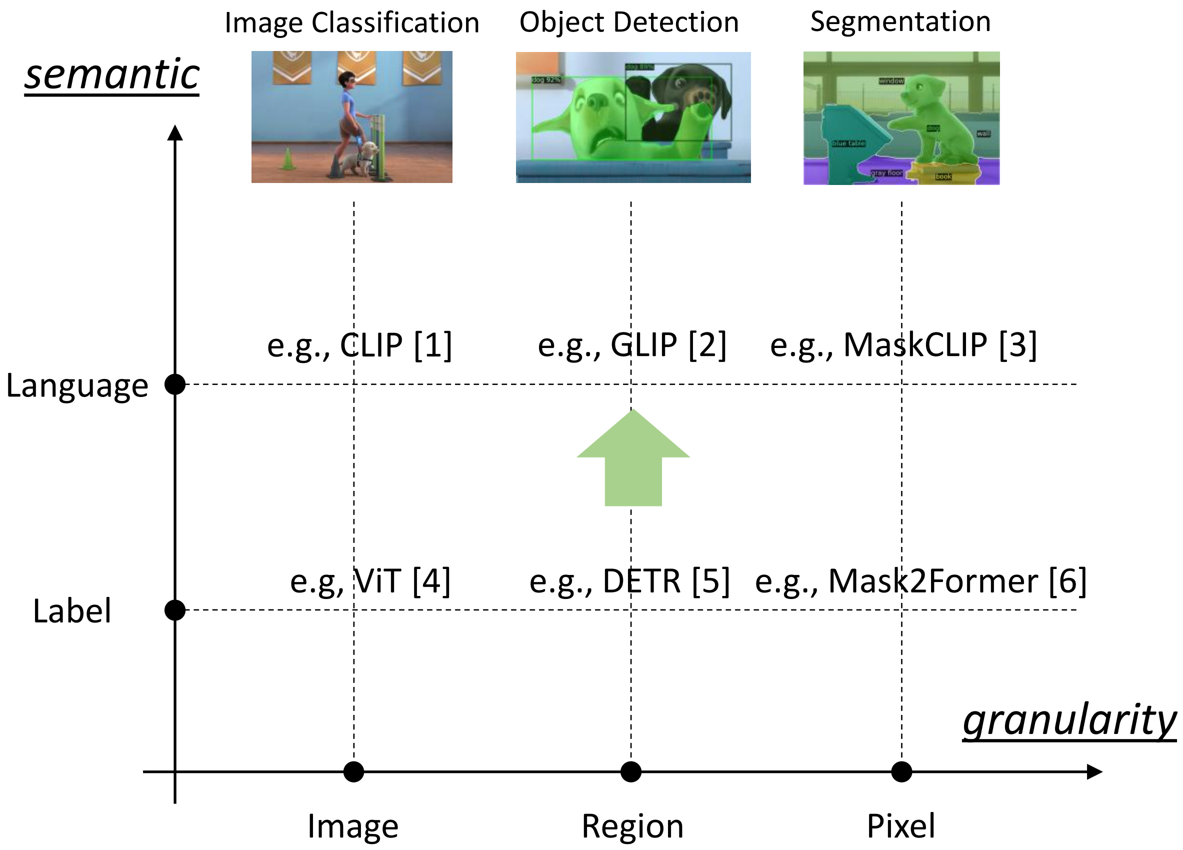


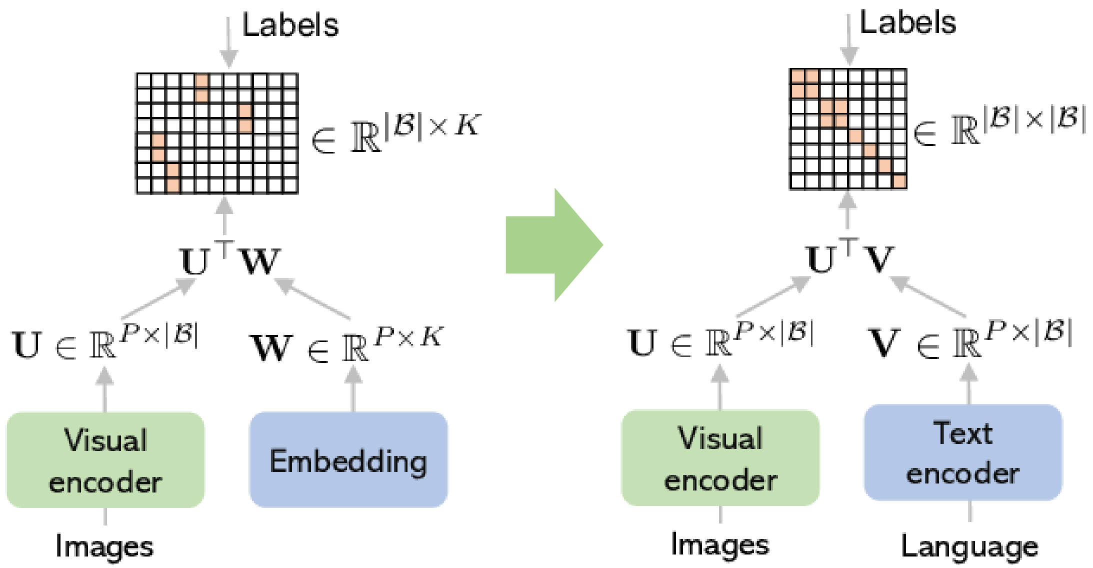
(a) Converting labels to language is agnostic to granularity

(b) Coarse-grained knowledge can be transferred to fine-grained tasks

[1] Radford et al. "Learning transferable visual models from natural language supervision." ICML, PMLR, 2021
[2] Li et al. "Grounded language-image pre-training." CVPR, 2022
[3] Zhou et al. "Extract Free Dense Labels from CLIP." ECCV, 2022
[4] Dosovitskiy et al. "An image is worth 16x16 words: Transformers for image recognition at scale." *ICLR, 2021*
[5] Carion et al. "End-to-end object detection with transformers." *ECCV, 2020*
[6] Cheng et al. "Masked-attention mask transformer for universal image segmentation." *CVPR. 2022*

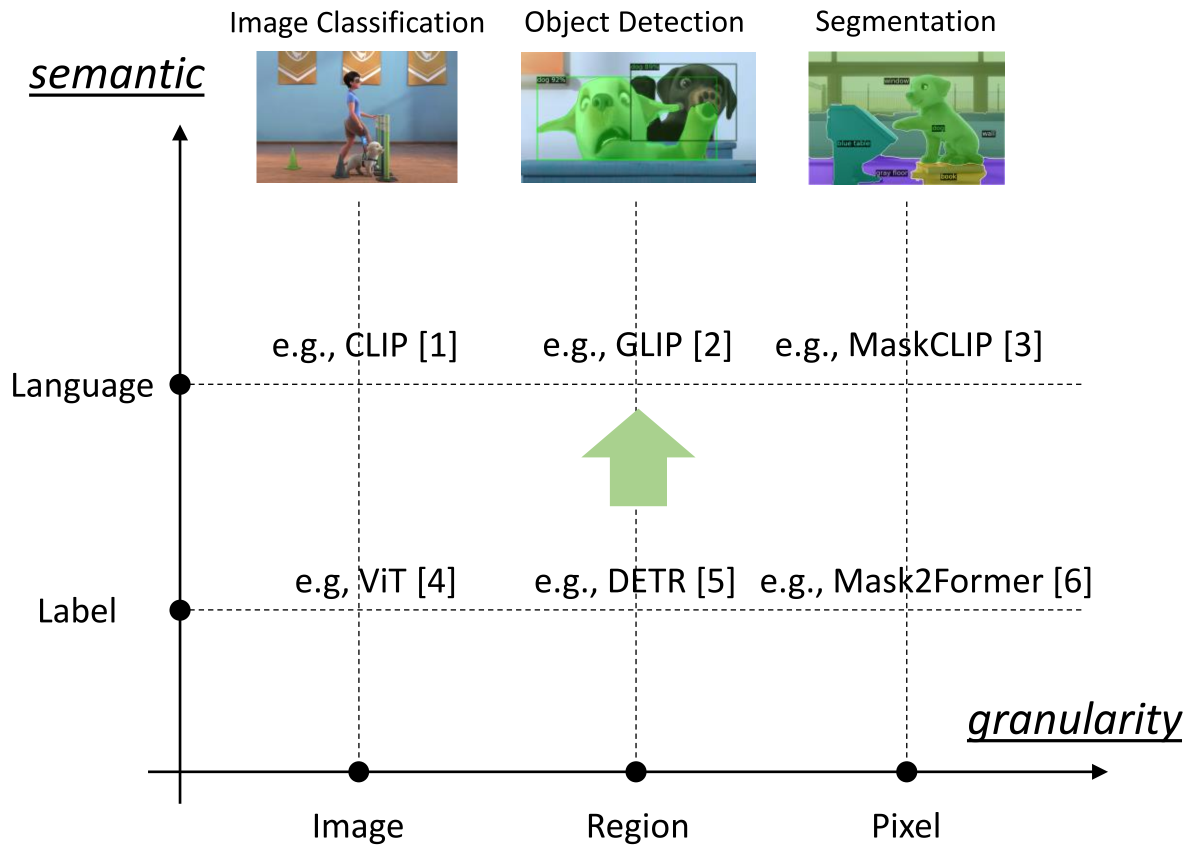
Bridge Vision with Language

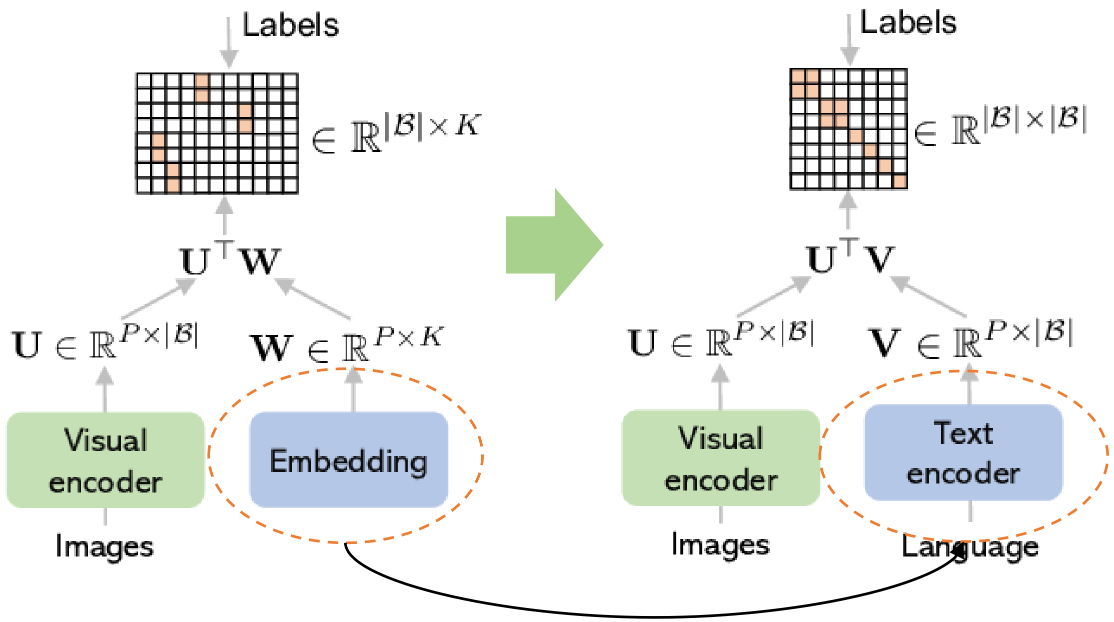




[1] Radford et al. "Learning transferable visual models from natural language supervision." ICML, PMLR, 2021
[2] Li et al. "Grounded language-image pre-training." CVPR, 2022
[3] Zhou et al. "Extract Free Dense Labels from CLIP." ECCV, 2022
[4] Dosovitskiy et al. “An image is worth 16x16 words: Transformers for image recognition at scale.” *ICLR, 2021*
[5] Carion et al. "End-to-end object detection with transformers." *ECCV, 2020*
[6] Cheng et al. "Masked-attention mask transformer for universal image segmentation." *CVPR. 2022*

Bridge Vision with Language

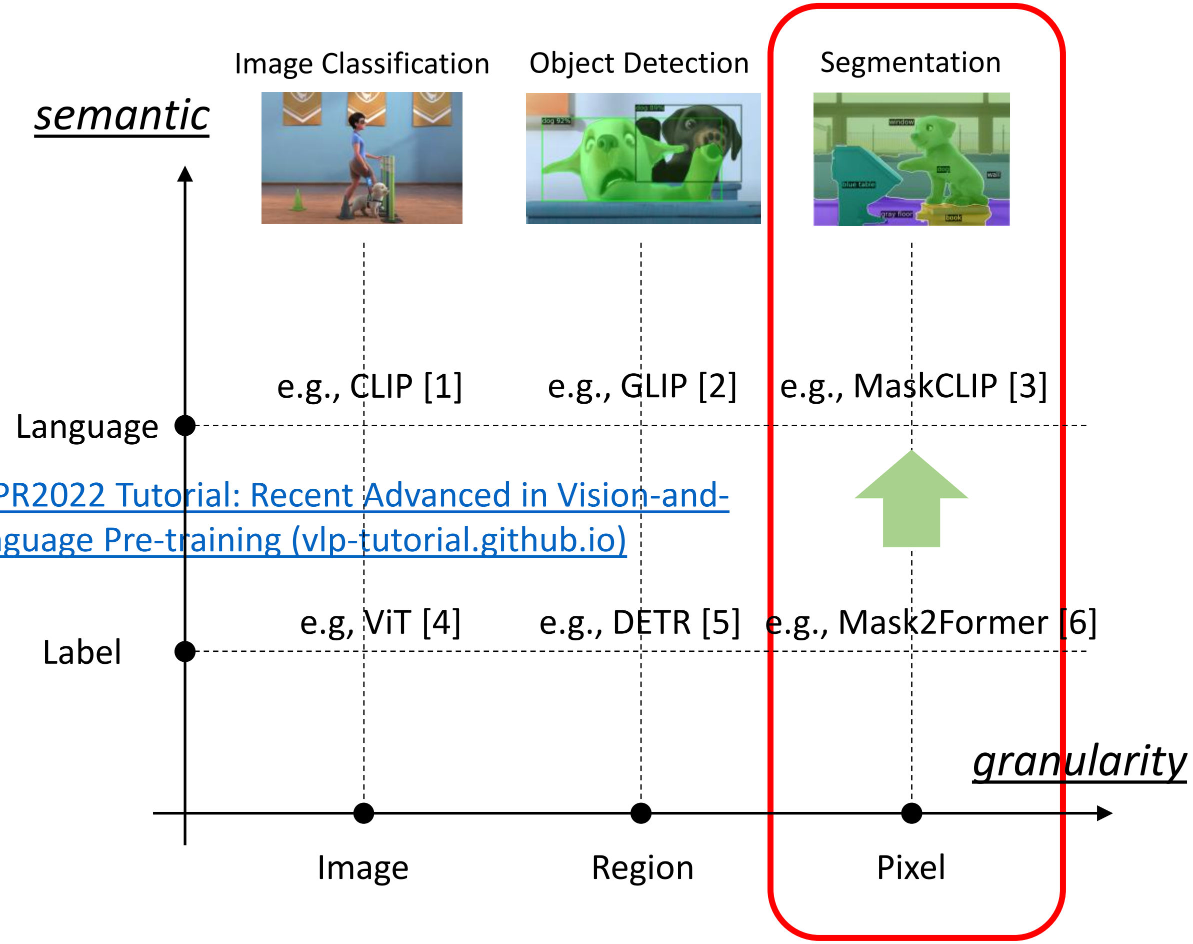




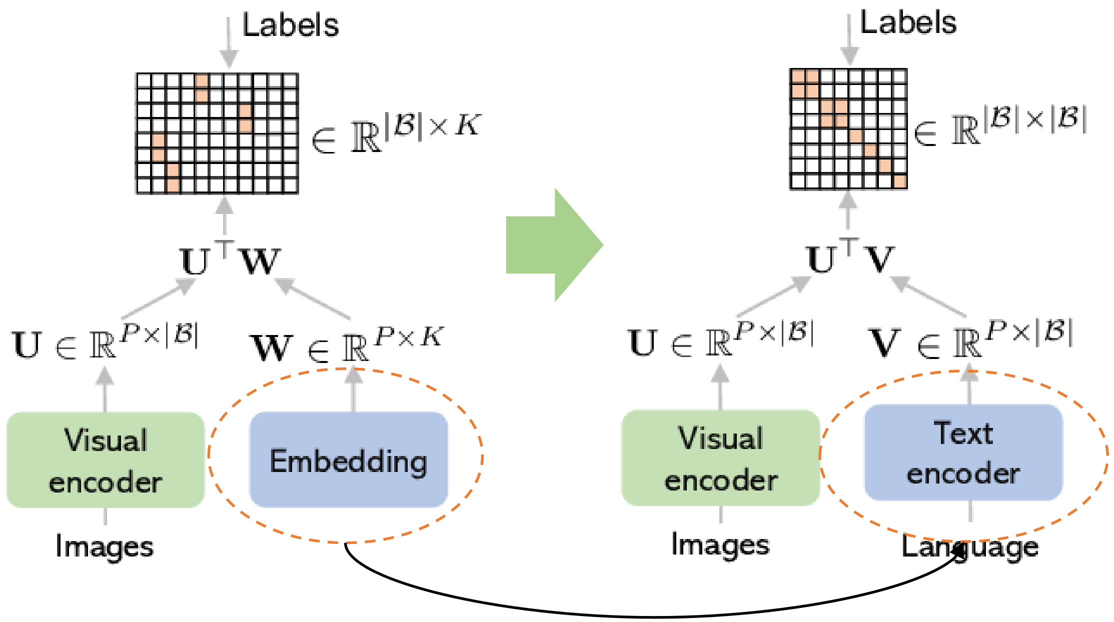
Replace labels with concept names, and use text encoder to encode all concepts as they are language tokens

[1] Radford et al. "Learning transferable visual models from natural language supervision." ICML, PMLR, 2021
[2] Li et al. "Grounded language-image pre-training." CVPR, 2022
[3] Zhou et al. "Extract Free Dense Labels from CLIP." ECCV, 2022
[4] Dosovitskiy et al. "An image is worth 16x16 words: Transformers for image recognition at scale." *ICLR, 2021*
[5] Carion et al. "End-to-end object detection with transformers." *ECCV, 2020*
[6] Cheng et al. "Masked-attention mask transformer for universal image segmentation." *CVPR. 2022*

Bridge Vision with Language



[CVPR2022 Tutorial: Recent Advanced in Vision-and-Language Pre-training (vlp-tutorial.github.io)](vlp-tutorial.github.io)



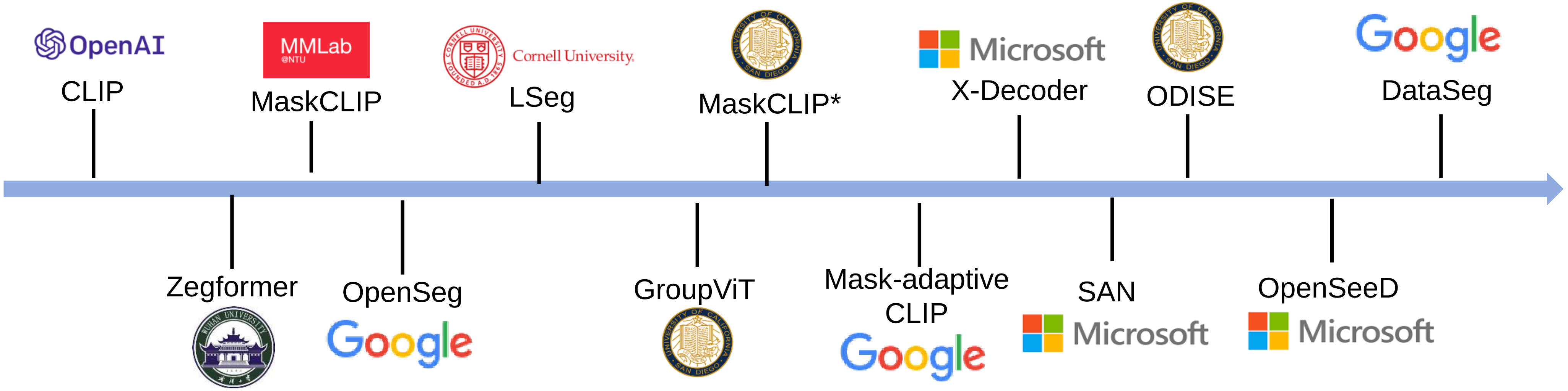
Replace labels with concept names, and use text encoder to encode all concepts as they are language tokens

[1] Radford et al. "Learning transferable visual models from natural language supervision." ICML, PMLR, 2021
[2] Li et al. "Grounded language-image pre-training." CVPR, 2022
[3] Zhou et al. "Extract Free Dense Labels from CLIP." ECCV, 2022

[4] Dosovitskiy et al. “An image is worth 16x16 words: Transformers for image recognition at scale.” *ICLR, 2021*
[5] Carion et al. "End-to-end object detection with transformers." *ECCV, 2020*
[6] Cheng et al. "Masked-attention mask transformer for universal image segmentation." *CVPR. 2022*

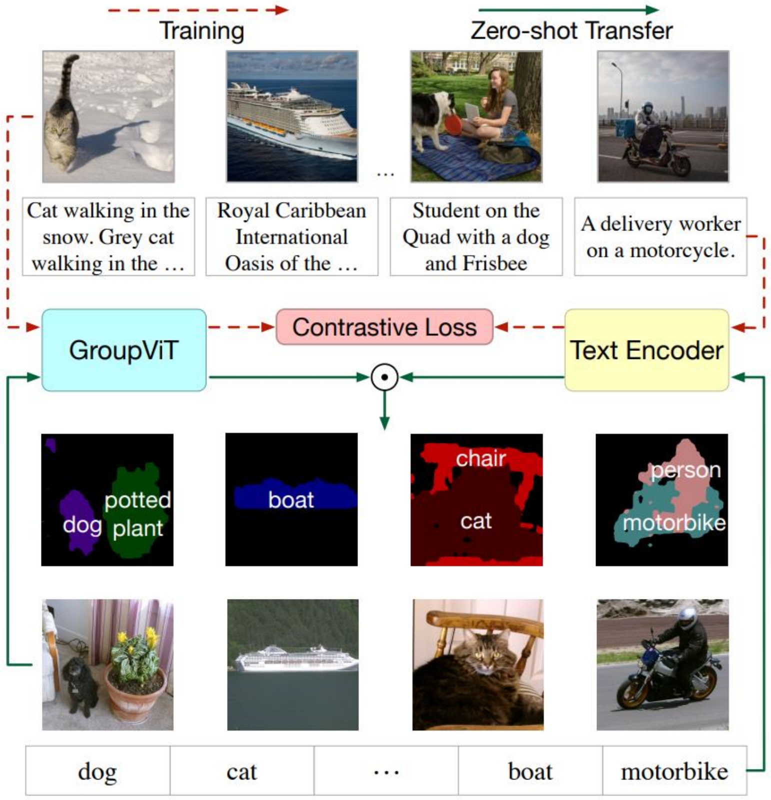
Bridge Vision with Language for Segmentation

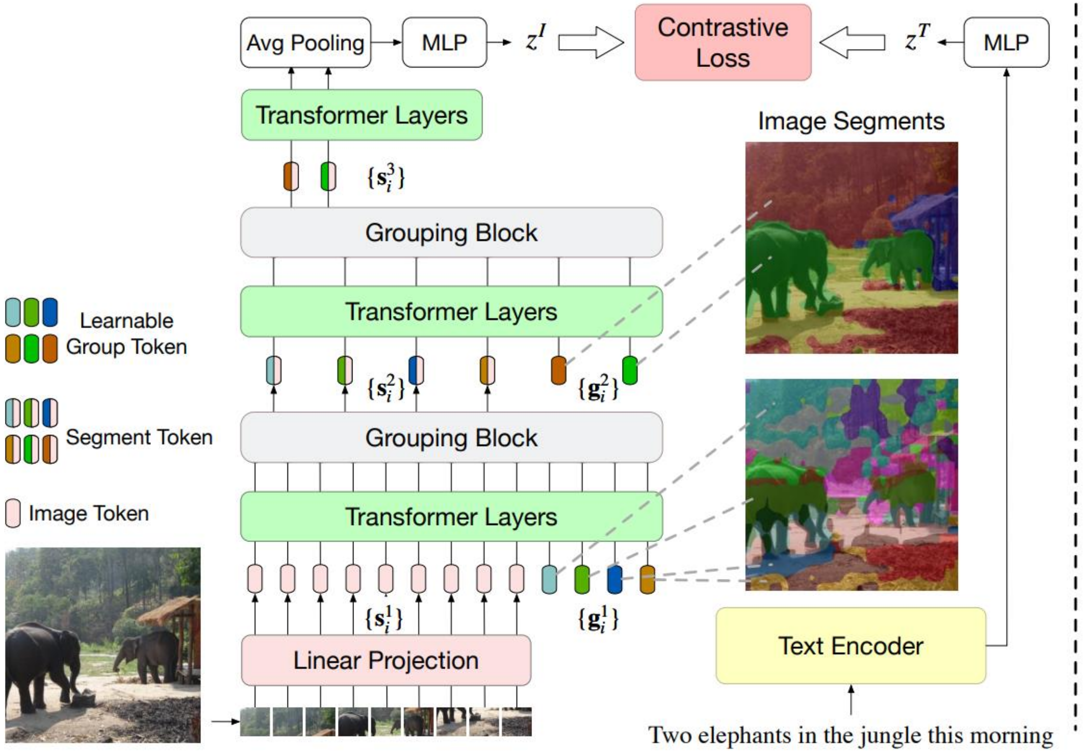
- Segmentation tasks:
  - Generic segmentation (semantic/instance/panoptic segmentation)
  - Referring segmentation (segment image with specific text phrase)
- Methodologies:
  - Initialize from CLIP *v.s.* train from scratch
  - Weakly supervised training *v.s.* supervised training
  - Two-stage *v.s.* end-to-end training



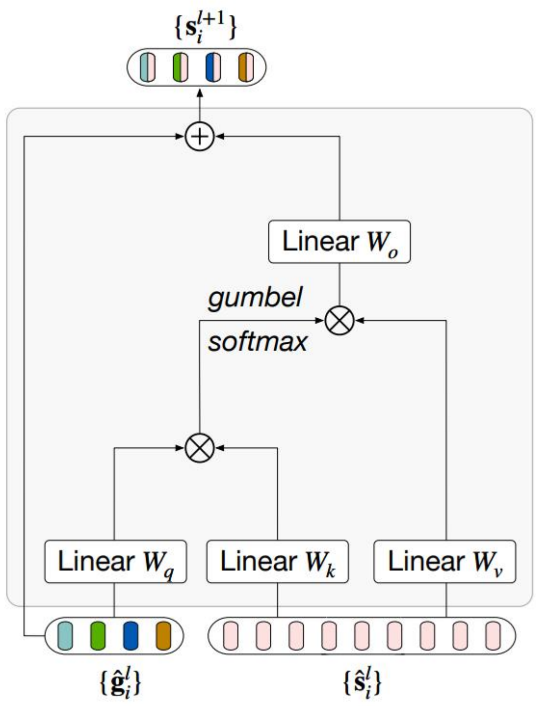
Bridge Vision with Language for Segmentation

- GroupViT: Learn to group semantic similar regions by learning from image-text pairs from scratch:
  - Bottom-up grouping using a novel grouping block
  - Top-down image-text supervision for visual-semantic alignment





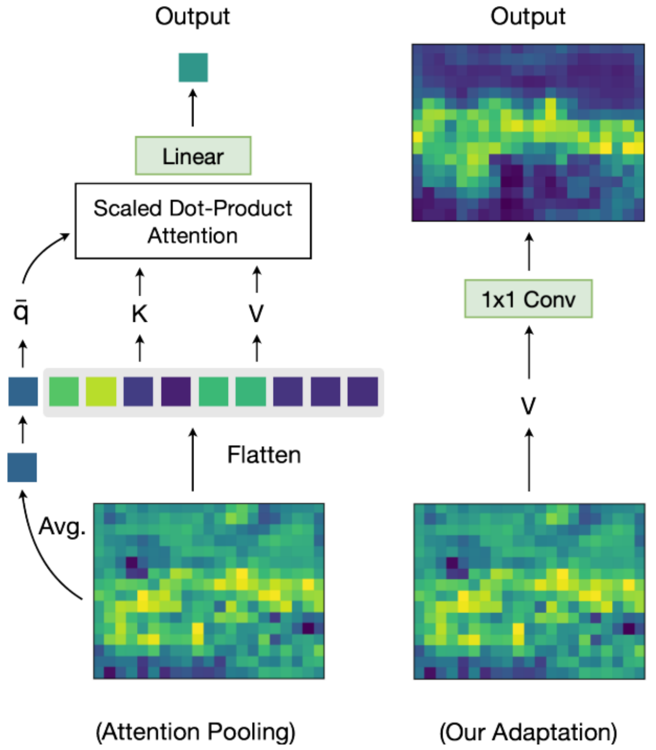
(a) GroupViT Architecture and Training Pipeline

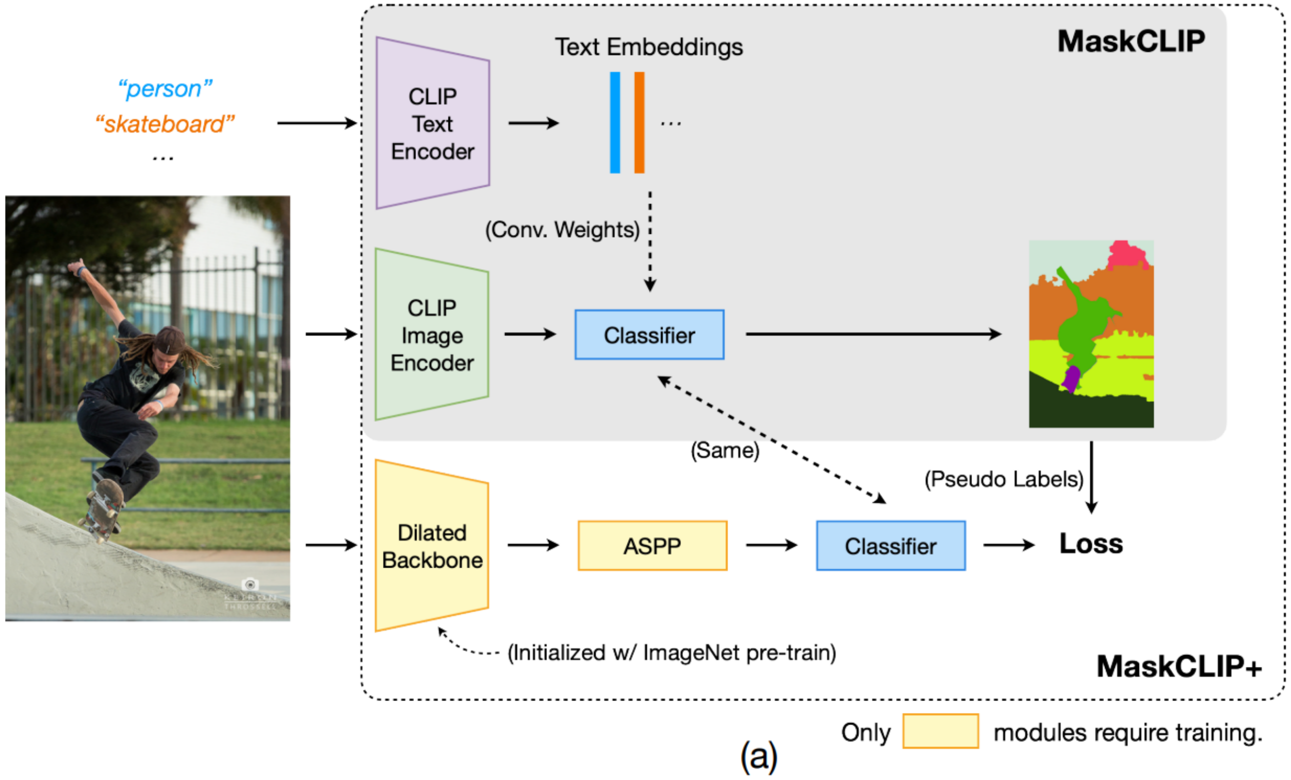


(b) Grouping Block

Bridge Vision with Language for Segmentation

- MaskCLIP: Extract free dense label from CLIP
  - Change attention pooling to a new adaptation strategy
  - Pseudo-label masks using CLIP as the teacher model

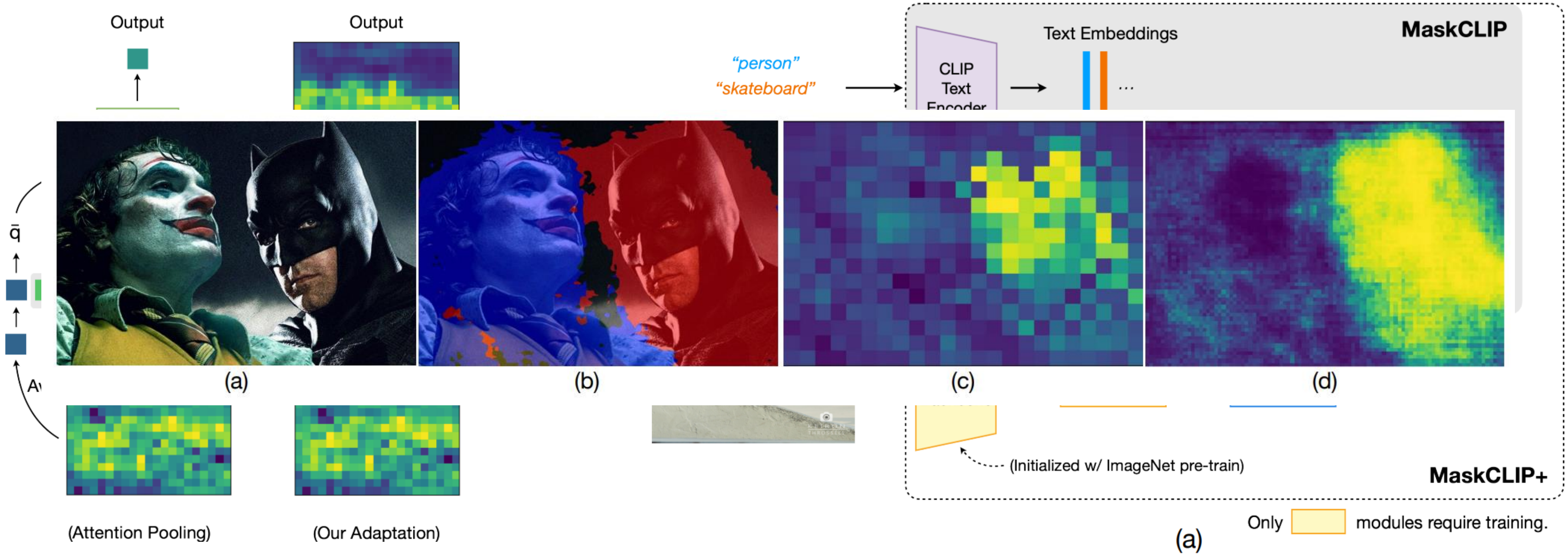




[1] Zhou et al. "Extract Free Dense Labels from CLIP." ECCV, 2022

Bridge Vision with Language for Segmentation

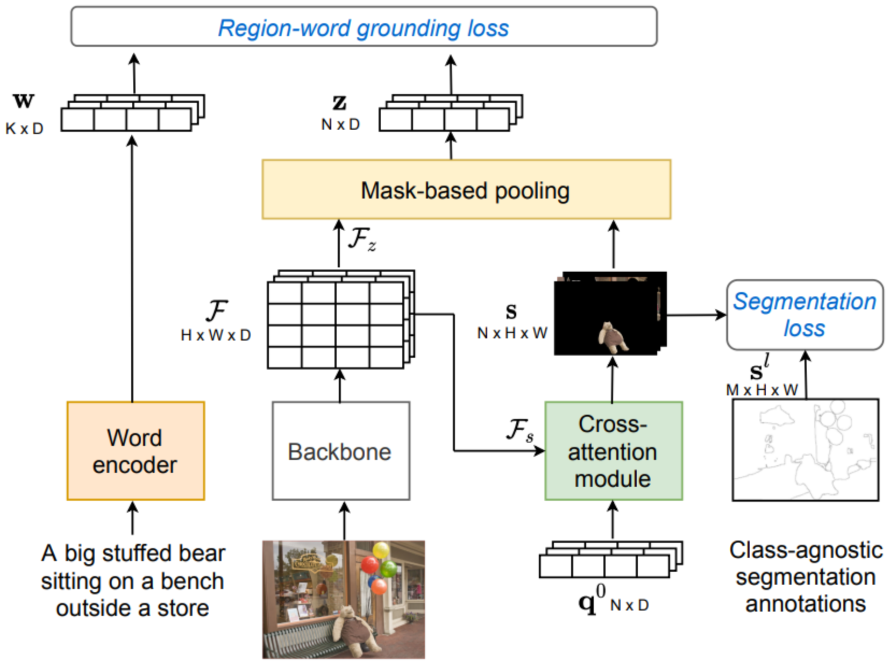
- MaskCLIP: Extract free dense label from CLIP
  - Change attention pooling to a new adaptation strategy
  - Pseudo-label masks using CLIP as the teacher model

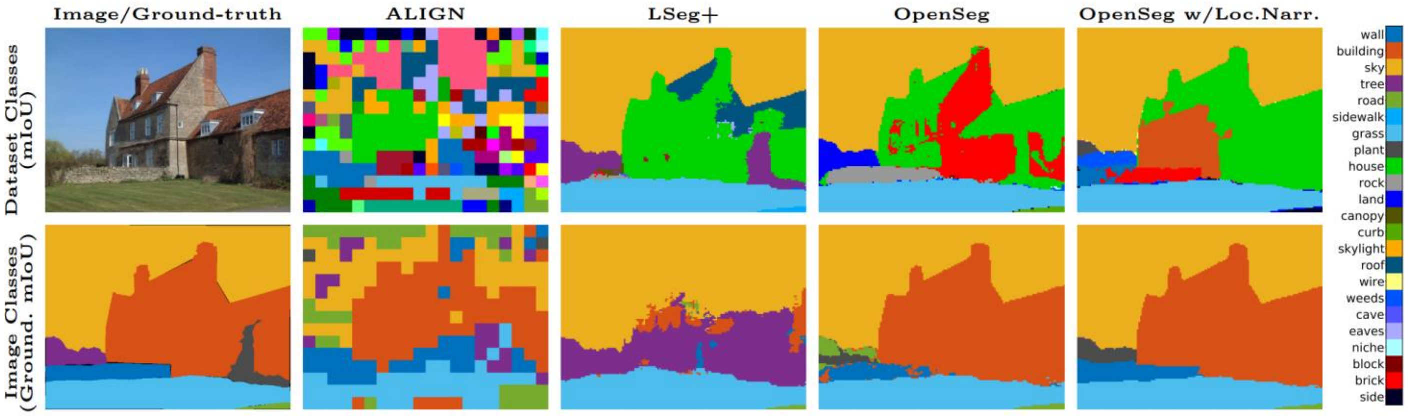


[1] Zhou et al. "Extract Free Dense Labels from CLIP." ECCV, 2022

Bridge Vision with Language for Segmentation

- OpenSeg: Weakly supervised learning by enforcing fine-grained alignment between textual features and mask-pooled features.
  - Learn from image-text pairs and local narrations.
  - A pretrained mask proposal network is used.

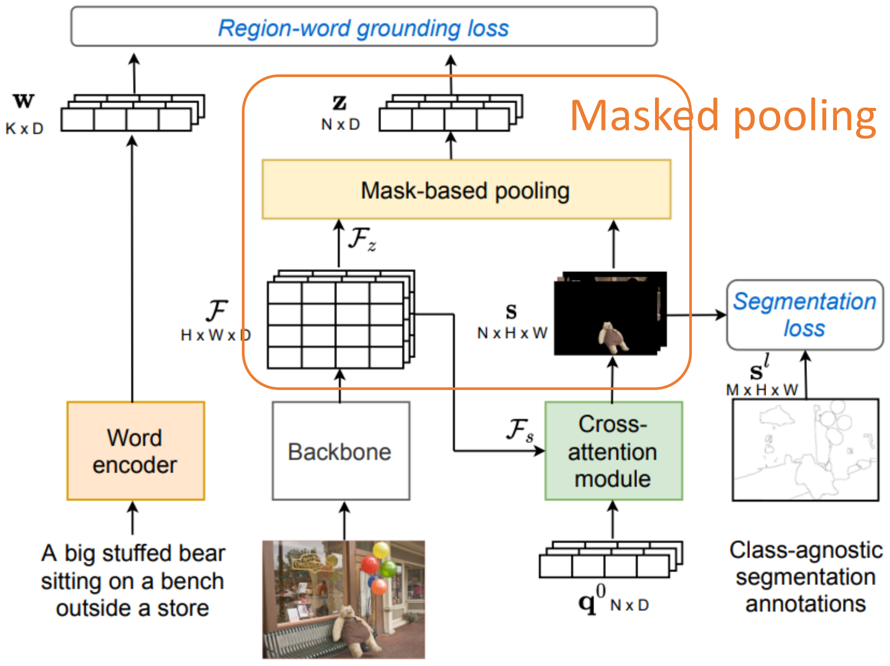


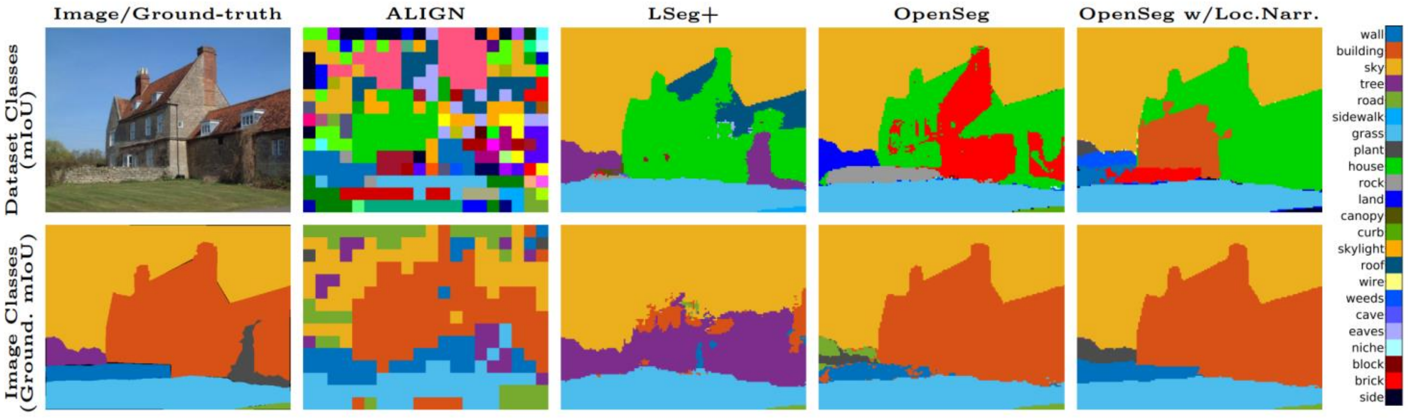


	COCO Train			mIoU					Grounding mIoU				
	label	mask	cap.	A-847	PC-459	A-150	PC-59	COCO	A-847	PC-459	A-150	PC-59	COCO
ALIGN	✗	✗	✗	4.8	3.6	9.7	18.5	15.6	17.8	21.8	25.7	34.2	28.2
ALIGN w/proposal	✗	✓	✗	5.8	4.8	12.9	22.4	17.9	17.3	19.7	25.3	32.0	23.6
LSeg+	✓	✓	✗	3.8	7.8	18.0	**46.5**	55.1	10.5	17.1	30.8	56.7	60.8
OpenSeg	✗	✓	✓	6.3	9.0	21.1	42.1	36.1	21.8	32.1	41.0	57.2	48.2
OpenSeg w/L. Narr.	✗	✓	✓	**6.8**	**11.2**	**24.8**	45.9	38.1	**25.4**	**39.0**	**45.5**	**61.5**	48.2

Bridge Vision with Language for Segmentation

- OpenSeg: Weakly supervised learning by enforcing fine-grained alignment between textual features and mask-pooled features.
  - Learn from image-text pairs and local narrations.
  - A pretrained mask proposal network is used.

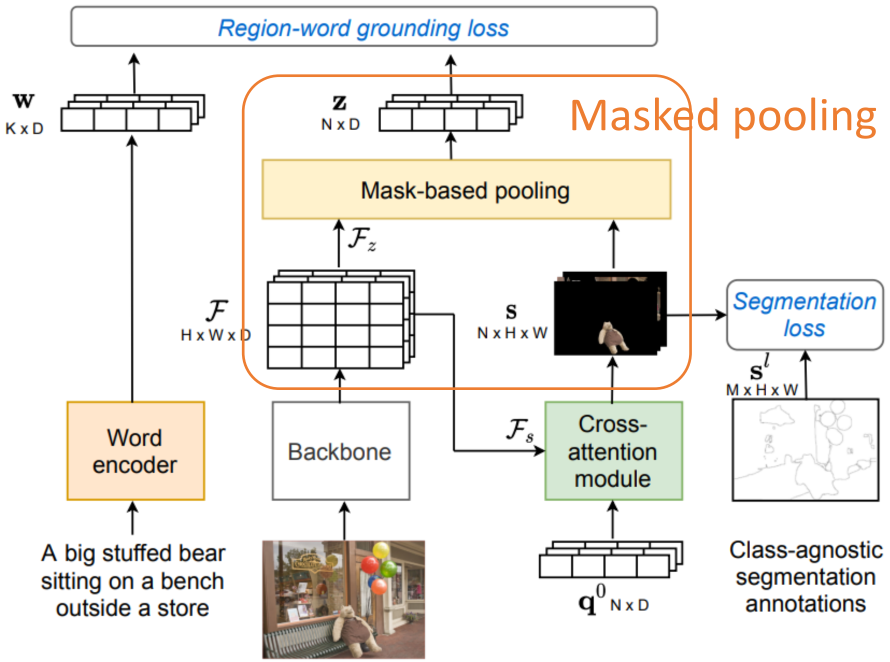


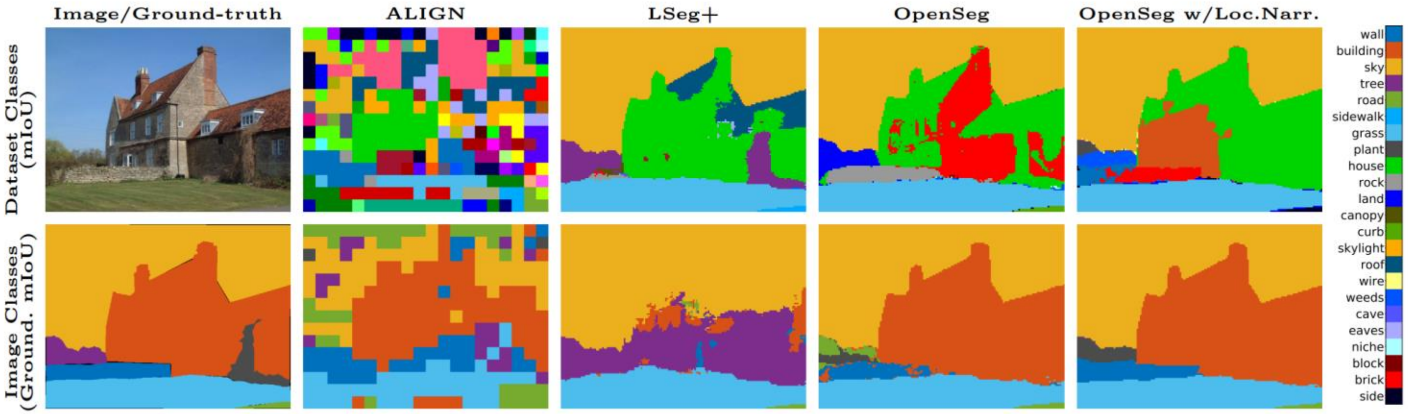


	COCO Train			mIoU					Grounding mIoU				
	label	mask	cap.	A-847	PC-459	A-150	PC-59	COCO	A-847	PC-459	A-150	PC-59	COCO
ALIGN	✗	✗	✗	4.8	3.6	9.7	18.5	15.6	17.8	21.8	25.7	34.2	28.2
ALIGN w/proposal	✗	✓	✗	5.8	4.8	12.9	22.4	17.9	17.3	19.7	25.3	32.0	23.6
LSeg+	✓	✓	✗	3.8	7.8	18.0	**46.5**	55.1	10.5	17.1	30.8	56.7	60.8
OpenSeg	✗	✓	✓	6.3	9.0	21.1	42.1	36.1	21.8	32.1	41.0	57.2	48.2
OpenSeg w/L. Narr.	✗	✓	✓	**6.8**	**11.2**	**24.8**	45.9	38.1	**25.4**	**39.0**	**45.5**	**61.5**	48.2

Bridge Vision with Language for Segmentation

- OpenSeg: Weakly supervised learning by enforcing fine-grained alignment between textual features and mask-pooled features.
  - Learn from image-text pairs and local narrations.
  - A pretrained mask proposal network is used.



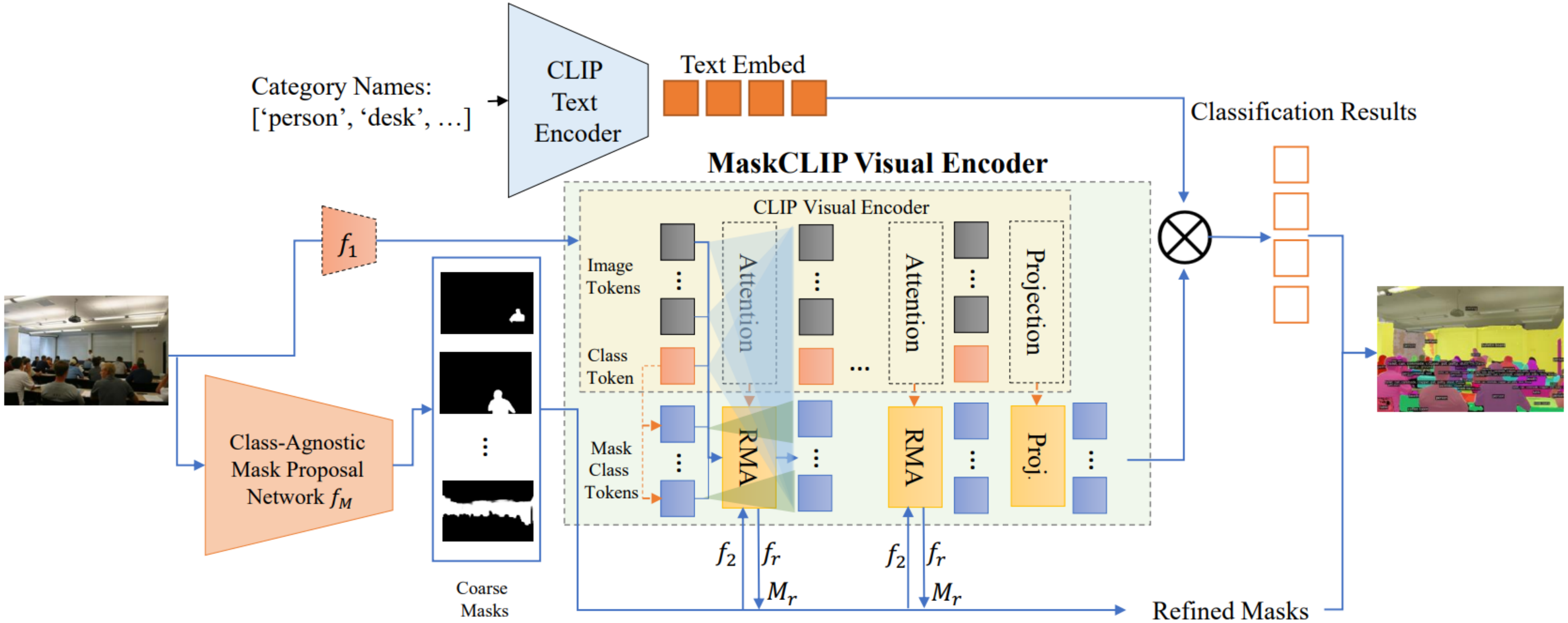


	COCO Train			mIoU					Grounding mIoU				
	label	mask	cap.	A-847	PC-459	A-150	PC-59	COCO	A-847	PC-459	A-150	PC-59	COCO
ALIGN	✗	✗	✗	4.8	3.6	9.7	18.5	15.6	17.8	21.8	25.7	34.2	28.2
ALIGN w/proposal	✗	✓	✗	5.8	4.8	12.9	22.4	17.9	17.3	19.7	25.3	32.0	23.6
LSeg+	✓	✓	✗	3.8	7.8	18.0	**46.5**	55.1	10.5	17.1	30.8	56.7	60.8
OpenSeg	✗	✓	✓	6.3	9.0	21.1	42.1	36.1	21.8	32.1	41.0	57.2	48.2
OpenSeg w/L. Narr.	✗	✓	✓	**6.8**	**11.2**	**24.8**	45.9	38.1	**25.4**	**39.0**	**45.5**	**61.5**	48.2

Image-text pairs helps, and local narrations further improve the performance

Bridge Vision with Language for Segmentation

- MaskCLIP (UCSD): Supervised training for panoptic segmentation with COCO using CLIP as the initialization
  - Two-stage training: 1) mask proposal network training; 2) CLIP model adaptation

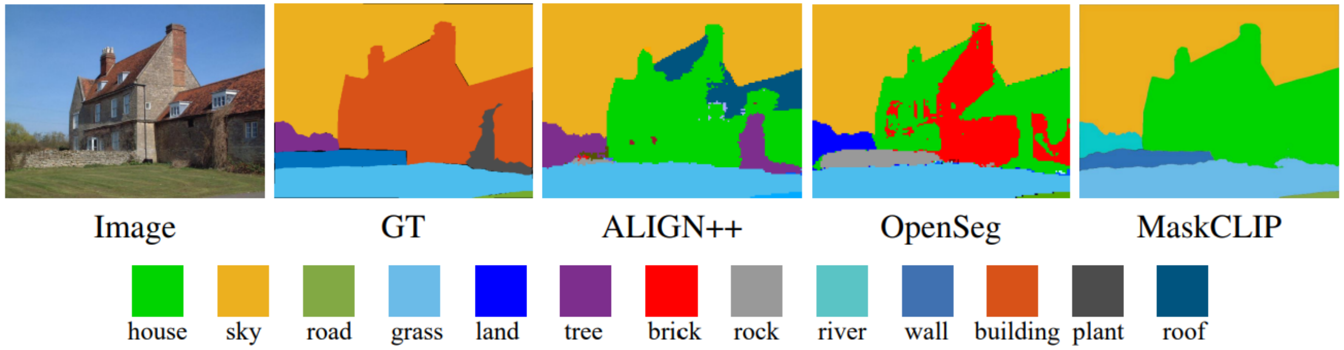


Bridge Vision with Language for Segmentation

- MaskCLIP (UCSD): Supervised training for panoptic segmentation with COCO using CLIP as the initialization
  - Two-stage training: 1) mask proposal network training; 2) CLIP model adaptation

CLIP baseline works and mask proposals help slightly

Method	COCO Training Data	A-150 ↑	A-847 ↑	P-459 ↑	P-59 ↑
ALIGN (Jia et al., 2021)	None	10.7	4.1	3.7	15.7
ALIGN w/ proposals (Jia et al., 2021)	Masks	12.9	5.8	4.8	22.4
LSeg+ (Li et al., 2022a)	Masks + Labels	18.0	3.8	7.8	46.5
OpenSeg (Ghiasi et al., 2022)	Masks + Captions	21.1	6.3	9.0	42.1
SimSeg (Xu et al., 2022)	Masks + Labels	20.5	7.0	-	**47.7**
CLIP Baseline	Masks	13.8	5.2	5.2	25.3
MaskCLIP w/o RMA	Masks	14.9	5.6	5.3	26.1
MaskCLIP (MaskRCNN)	Masks + Labels	22.4	6.8	9.1	41.3
MaskCLIP	Masks + Labels	**23.7**	**8.2**	**10.0**	45.9

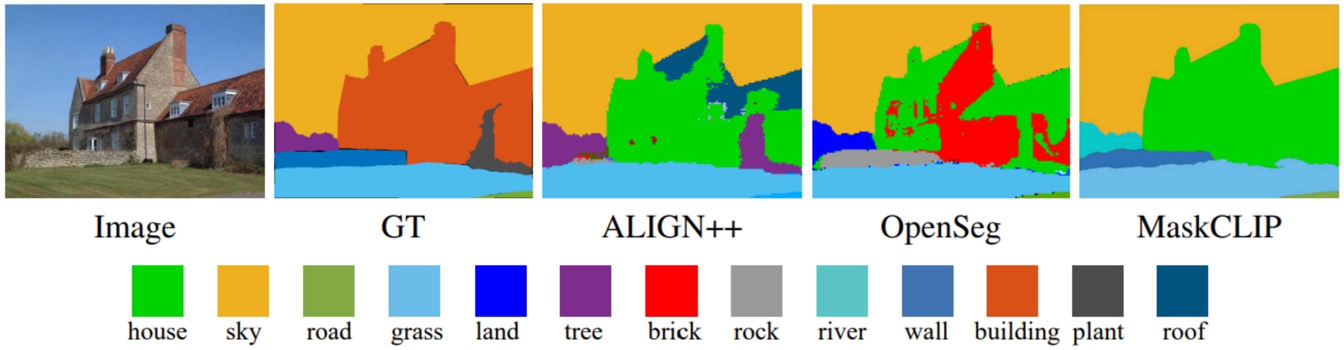


Bridge Vision with Language for Segmentation

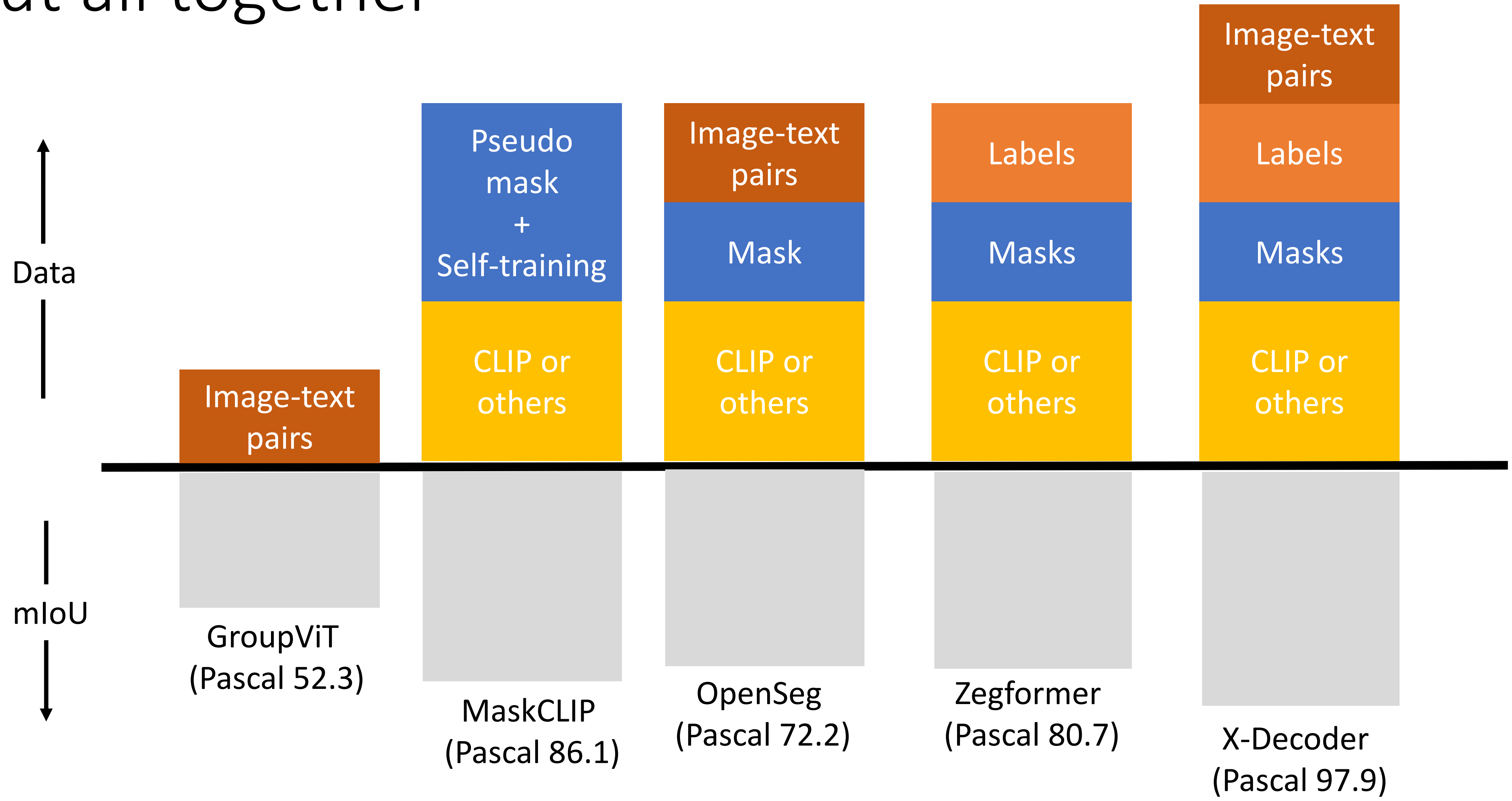
- MaskCLIP (UCSD): Supervised training for panoptic segmentation with COCO using CLIP as the initialization
  - Two-stage training: 1) mask proposal network training; 2) CLIP model adaptation

Method	COCO Training Data	A-150 ↑	A-847 ↑	P-459 ↑	P-59 ↑
ALIGN (Jia et al., 2021)	None	10.7	4.1	3.7	15.7
ALIGN w/ proposals (Jia et al., 2021)	Masks	12.9	5.8	4.8	22.4
LSeg+ (Li et al., 2022a)	Masks + Labels	18.0	3.8	7.8	46.5
OpenSeg (Ghiasi et al., 2022)	Masks + Captions	21.1	6.3	9.0	42.1
SimSeg (Xu et al., 2022)	Masks + Labels	20.5	7.0	-	**47.7**
CLIP Baseline	Masks	13.8	5.2	5.2	25.3
MaskCLIP w/o RMA	Masks	14.9	5.6	5.3	26.1
MaskCLIP (MaskRCNN)	Masks + Labels	22.4	6.8	9.1	41.3
MaskCLIP	Masks + Labels	**23.7**	**8.2**	**10.0**	45.9

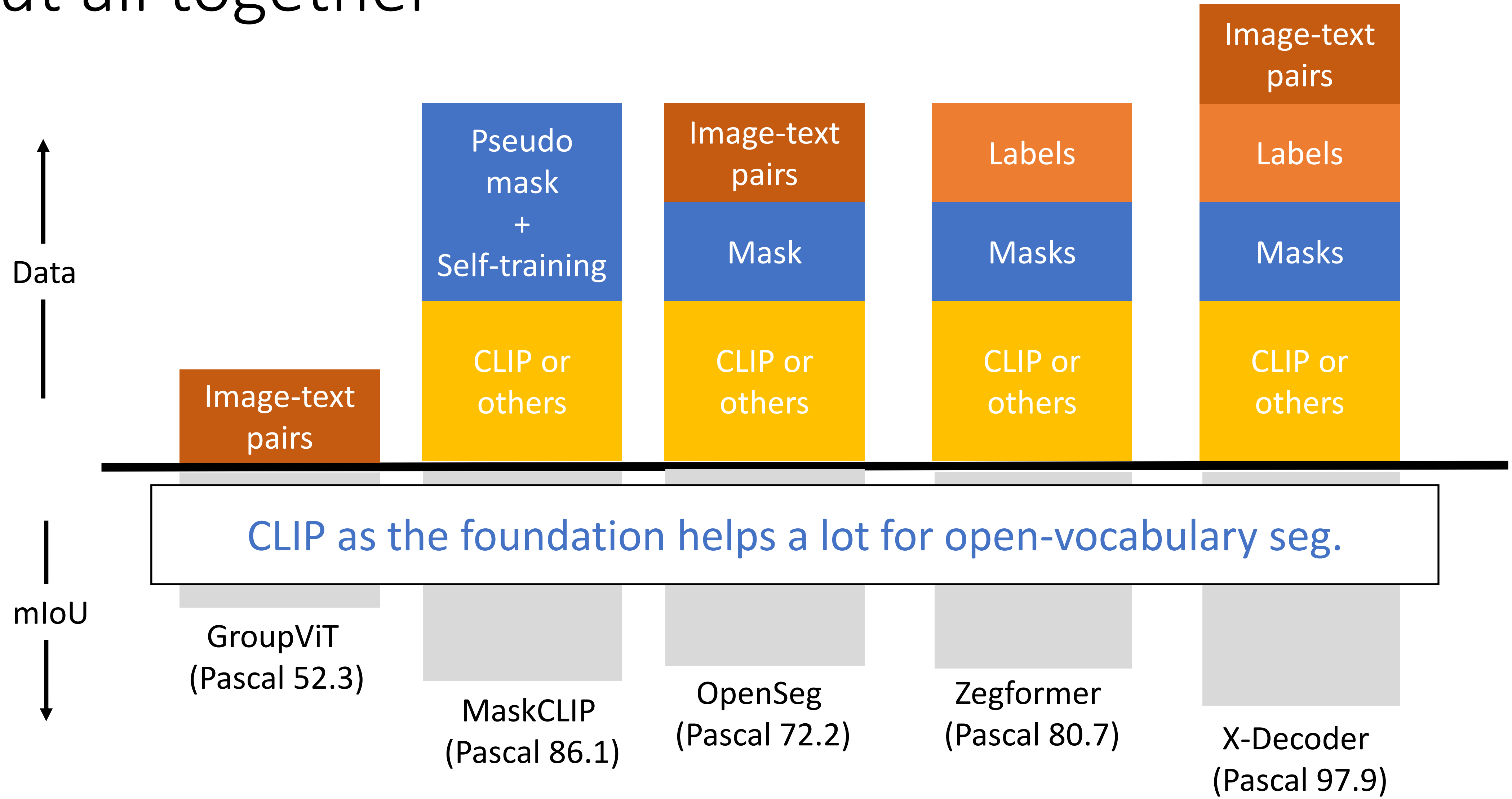
Label information significantly boost open-vocabulary performance.



Put all together

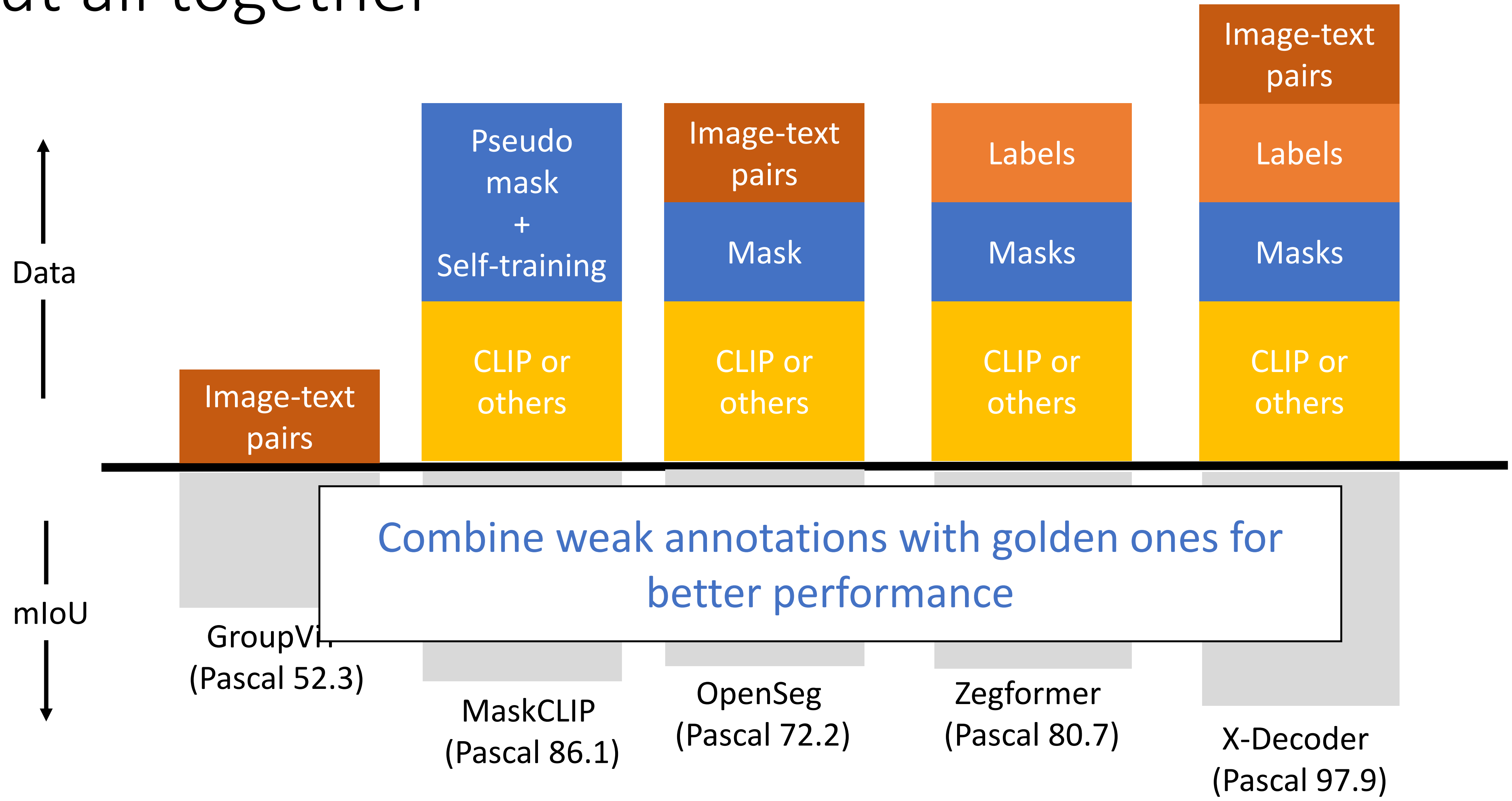


Put all together

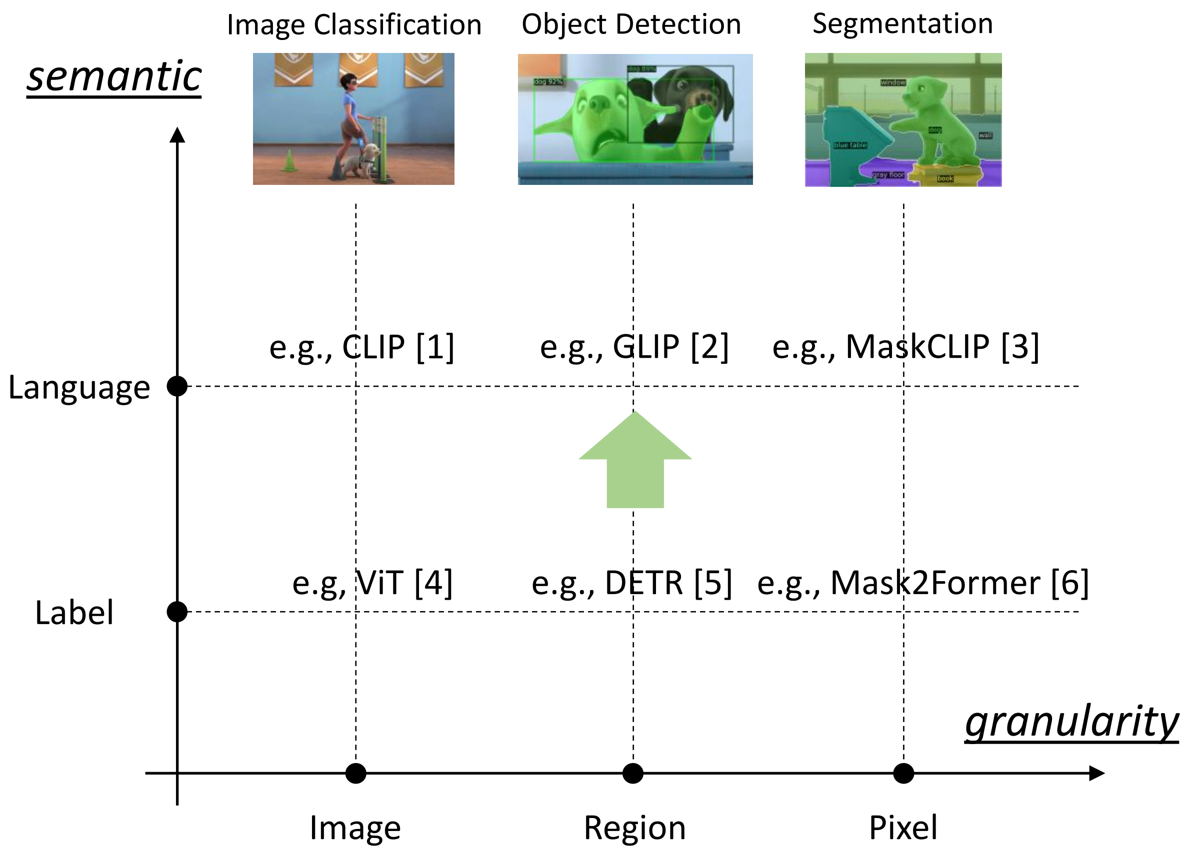


CLIP as the foundation helps a lot for open-vocabulary seg.

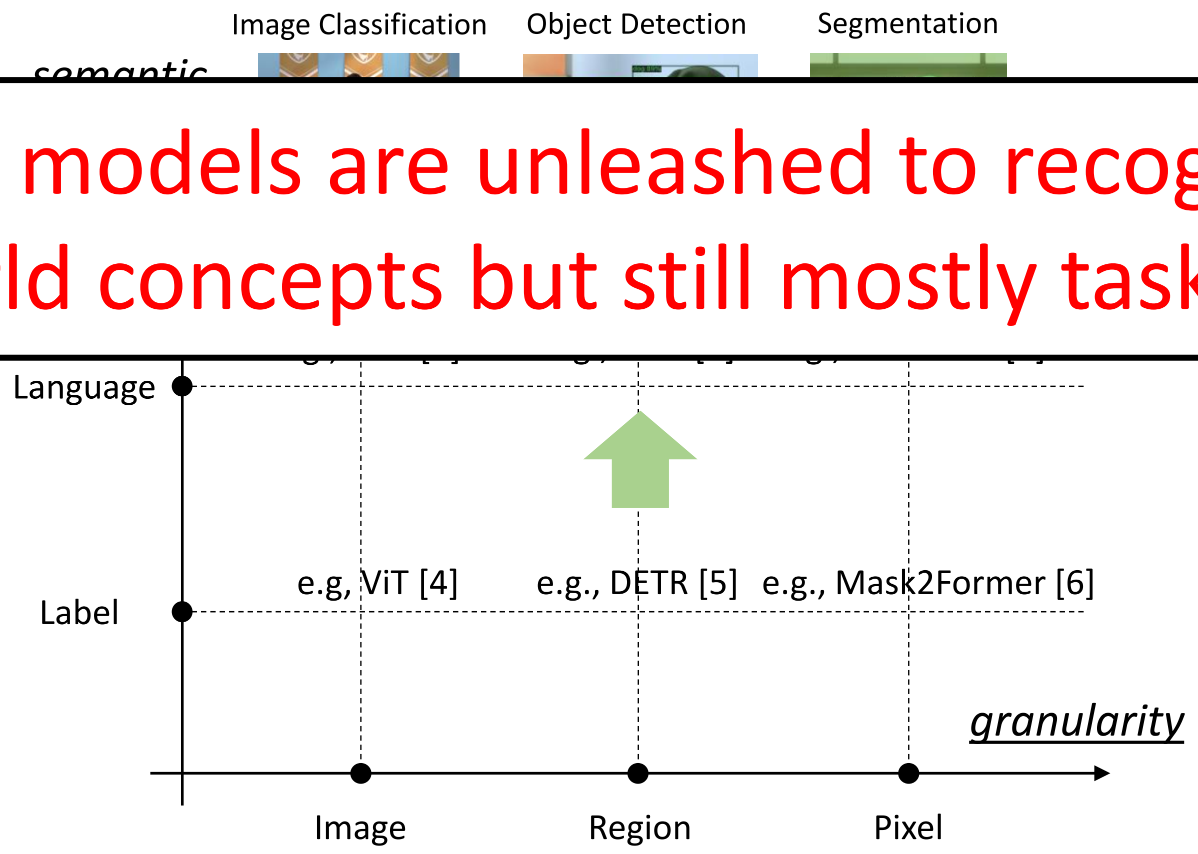
Put all together



Bridge Vision with Language for Core Vision

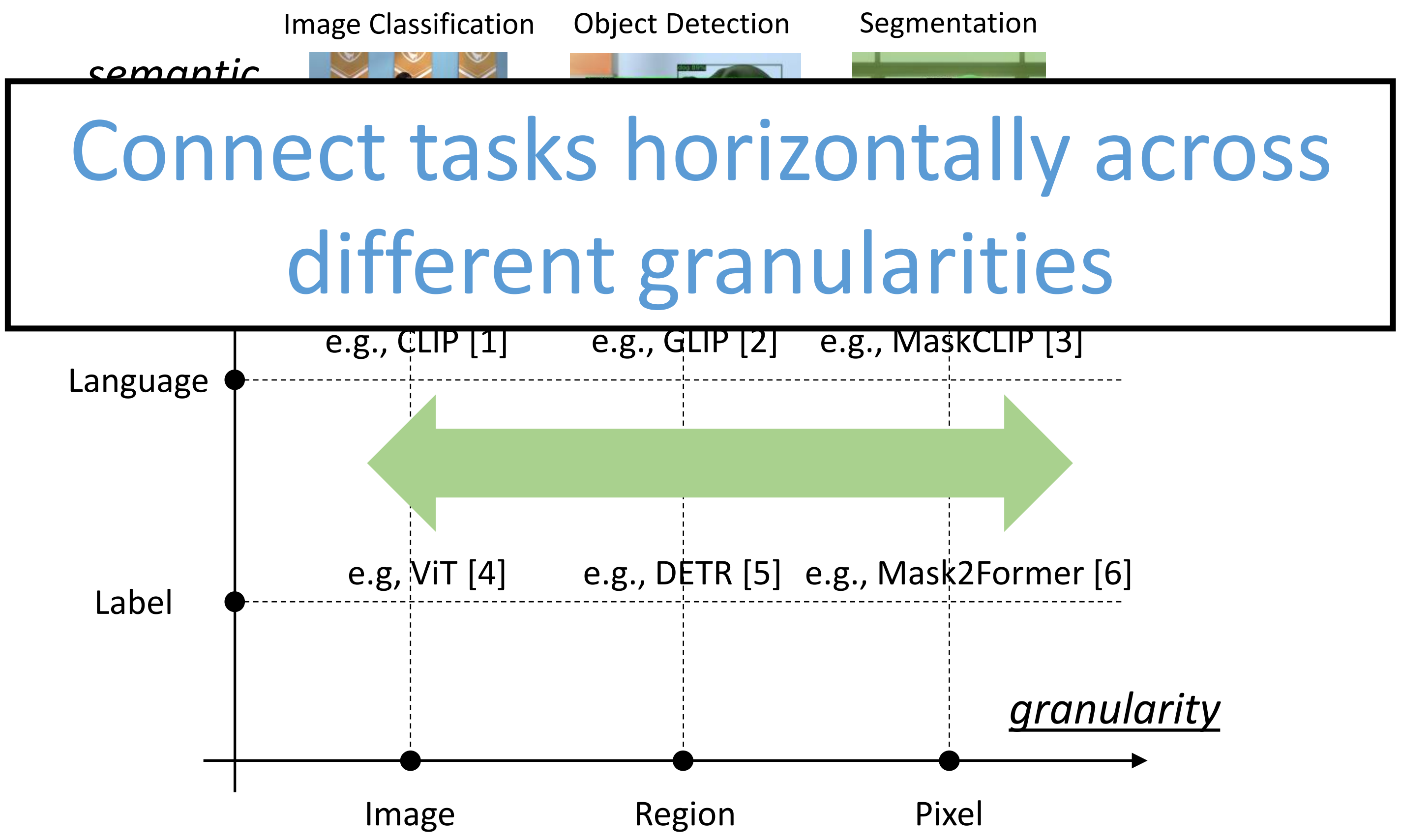


Bridge Vision with Language for Core Vision

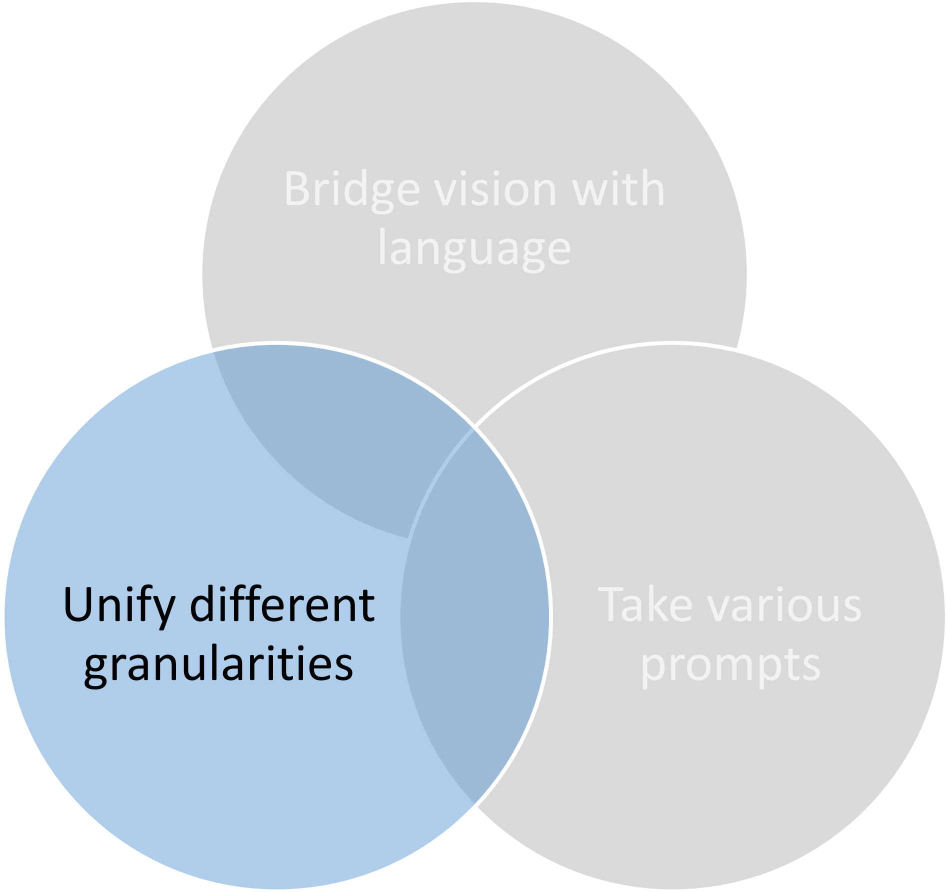


These models are unleashed to recognize open-world concepts but still mostly task-specific

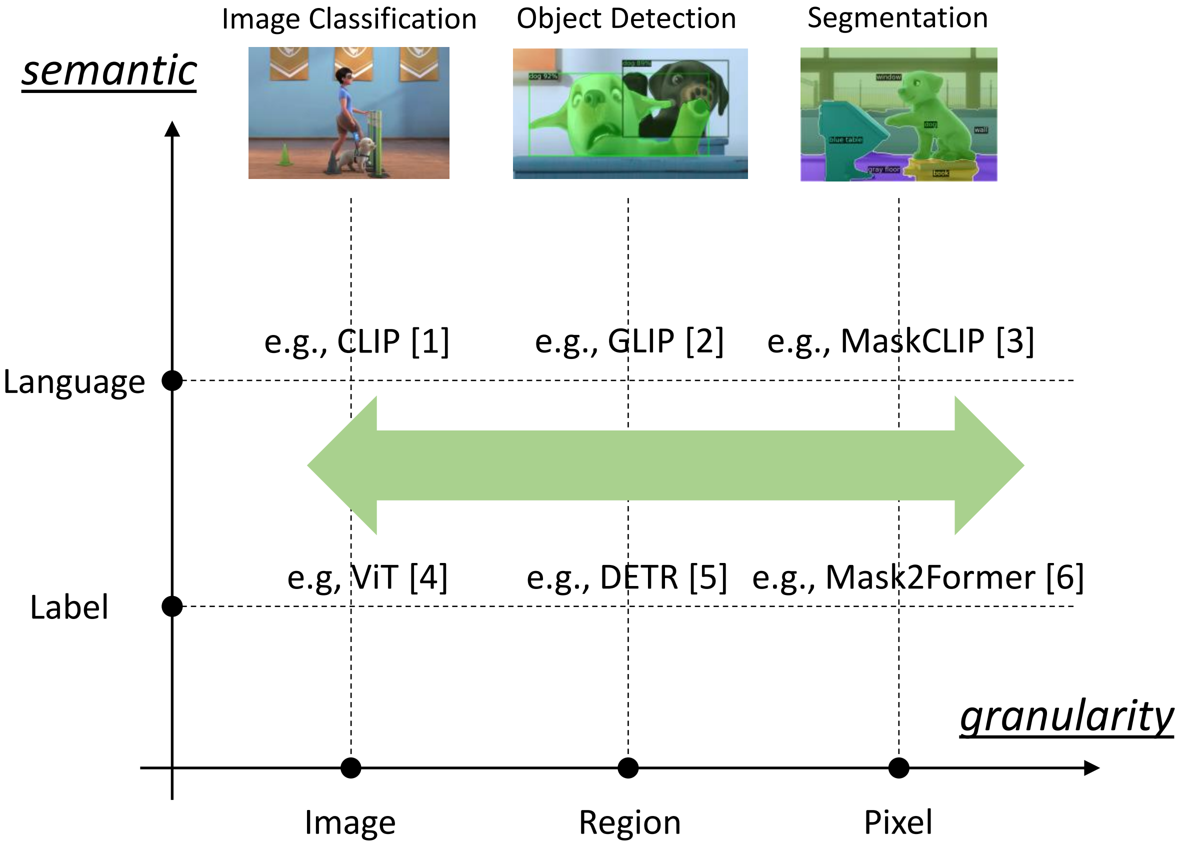
Bridge Vision with Language for Core Vision

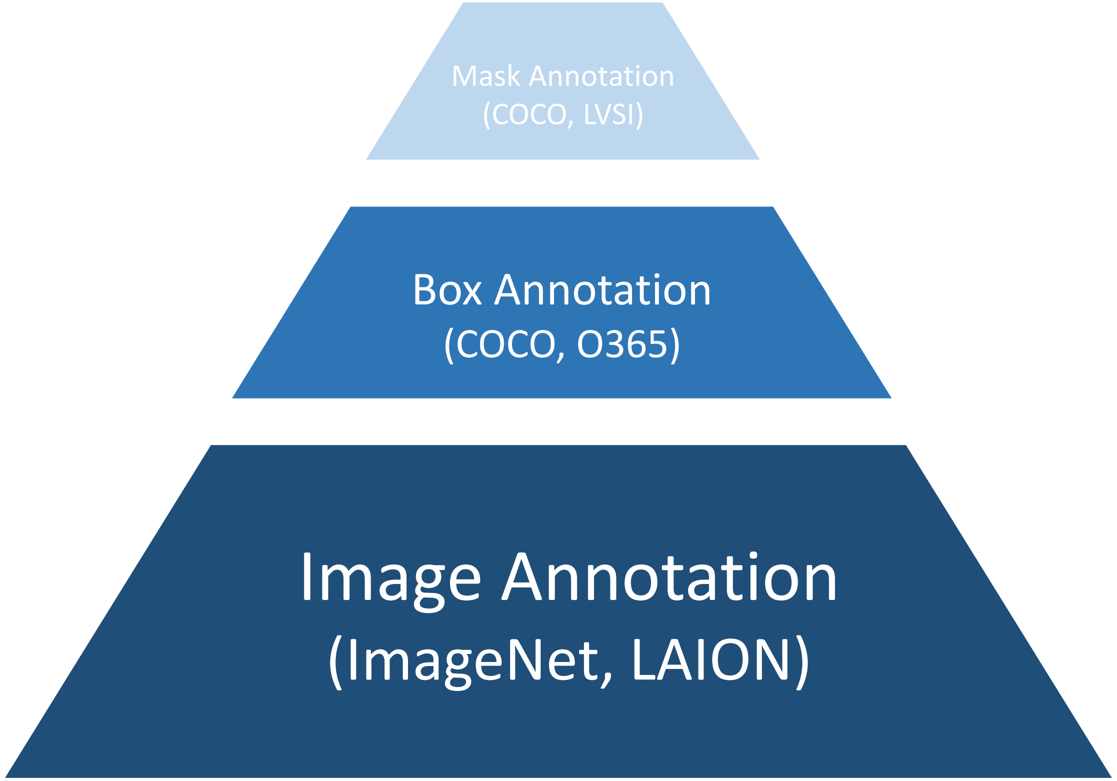


II. Unify Different Granularities

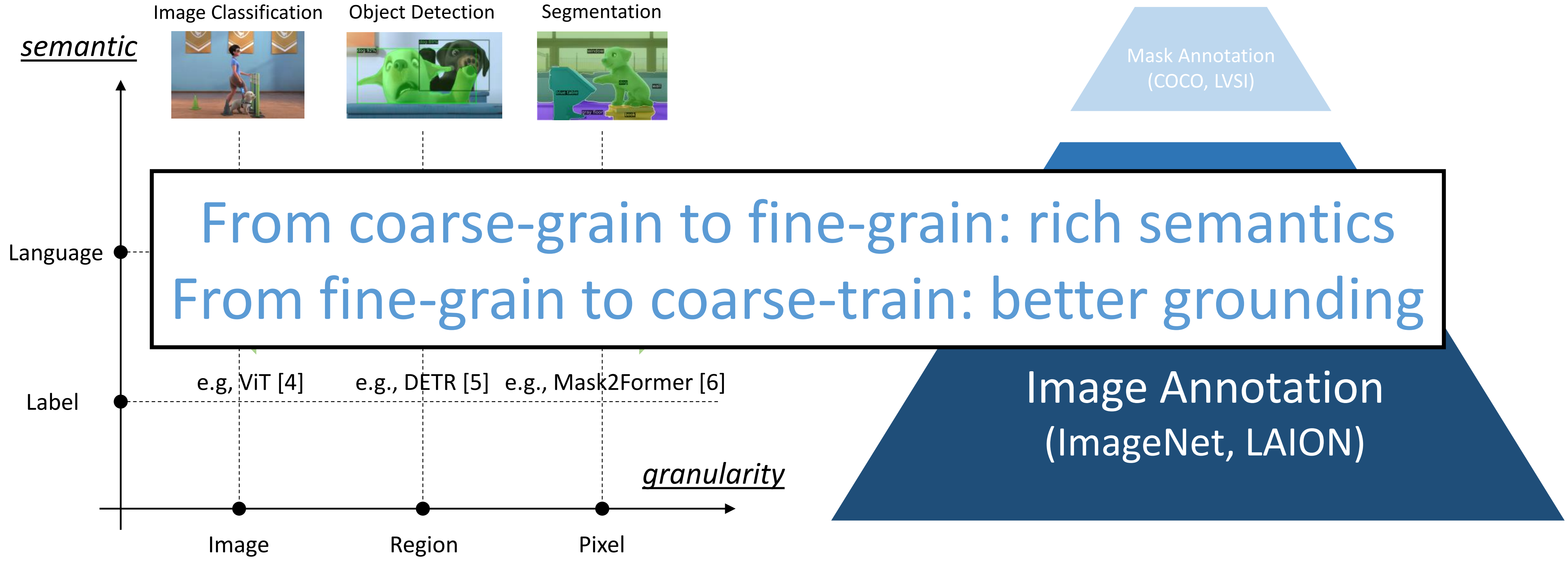


Unify Different Granularities





Unify Different Granularities



From coarse-grain to fine-grain: rich semantics

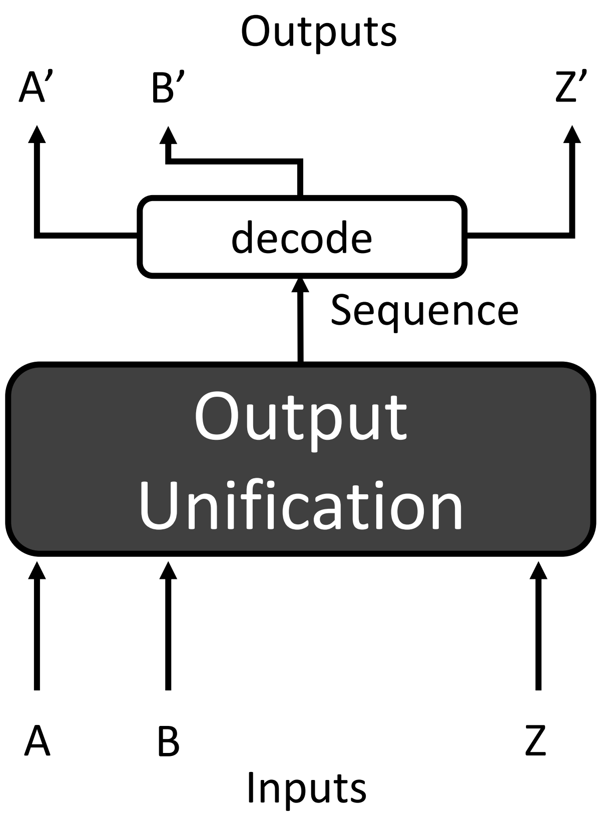
From fine-grain to coarse-train: better grounding

Unify Different Granularities

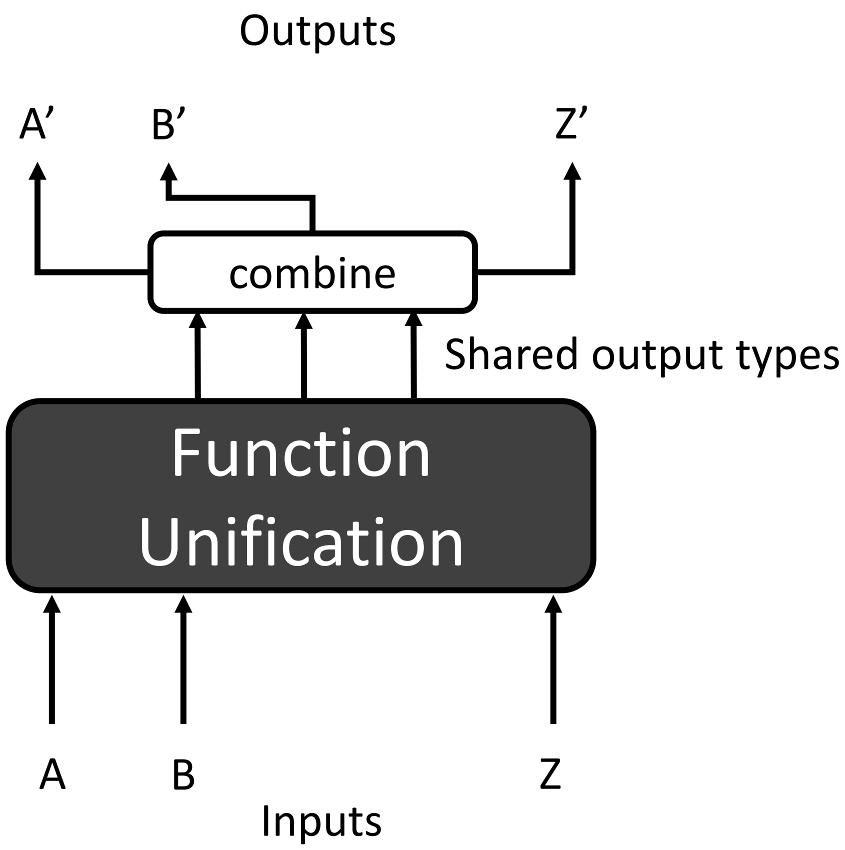
- Tasks we are considering:
  - Image-level: image recognition, image-text retrieval, image captioning, visual question answering, etc.
  - Region-level: object detection, dense caption, phrase grounding, etc.
  - Pixel-level: generic segmentation, referring segmentation, etc.

- Two types of unifications:
  - Output unification: convert all outputs into sequence.
  - Functionality unification: share the commons maximally but with respect to the differences.

Unify Different Granularities



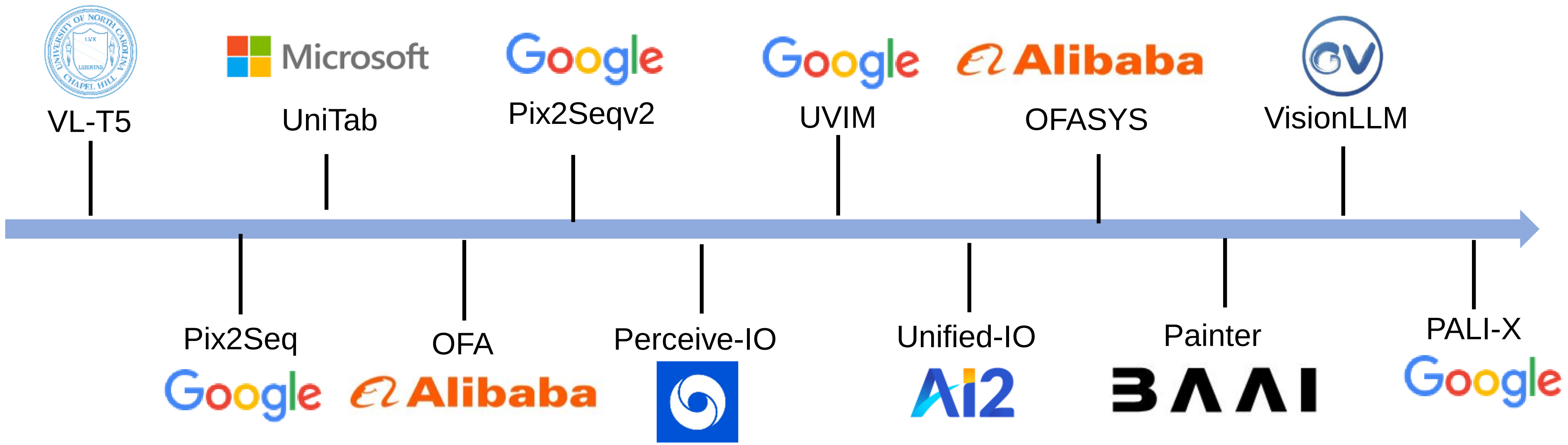
Convert all outputs into sequence and decode to corresponding outputs



Predict shared output types and combine one or more to produce the final outputs

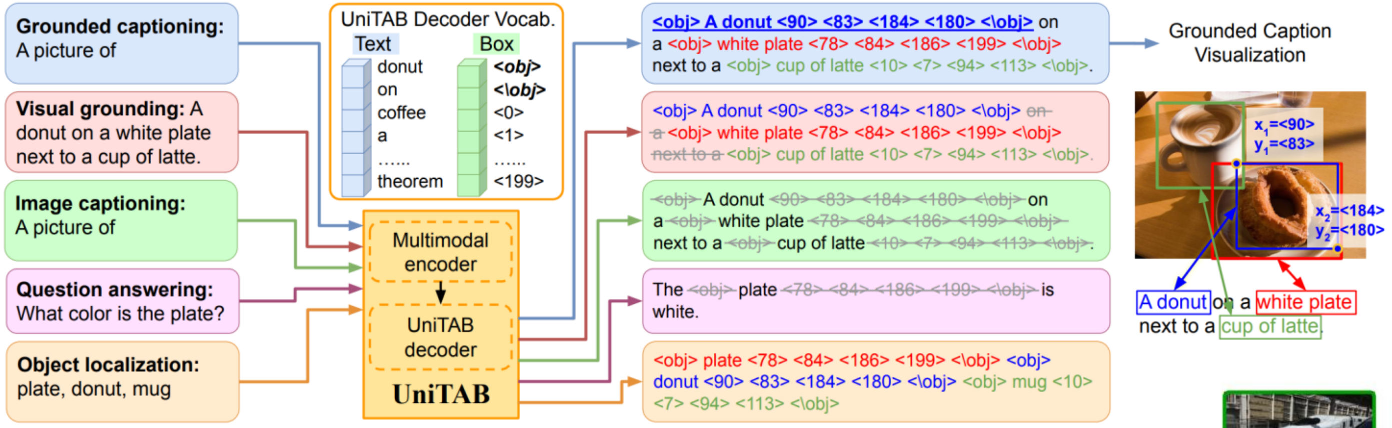
Outputs Unification

- Convert both inputs and outputs into sequences:
    - Inputs: Text as it is or add some prefixes; Image into a sequence of tokens (not necessarily)
    - Outputs: Boxes: a sequence of coordinates (top left + bottom right); Masks: a sequence of polygon coordinates encompassing mask; Key points: a sequence of coordinates.



Outputs Unification

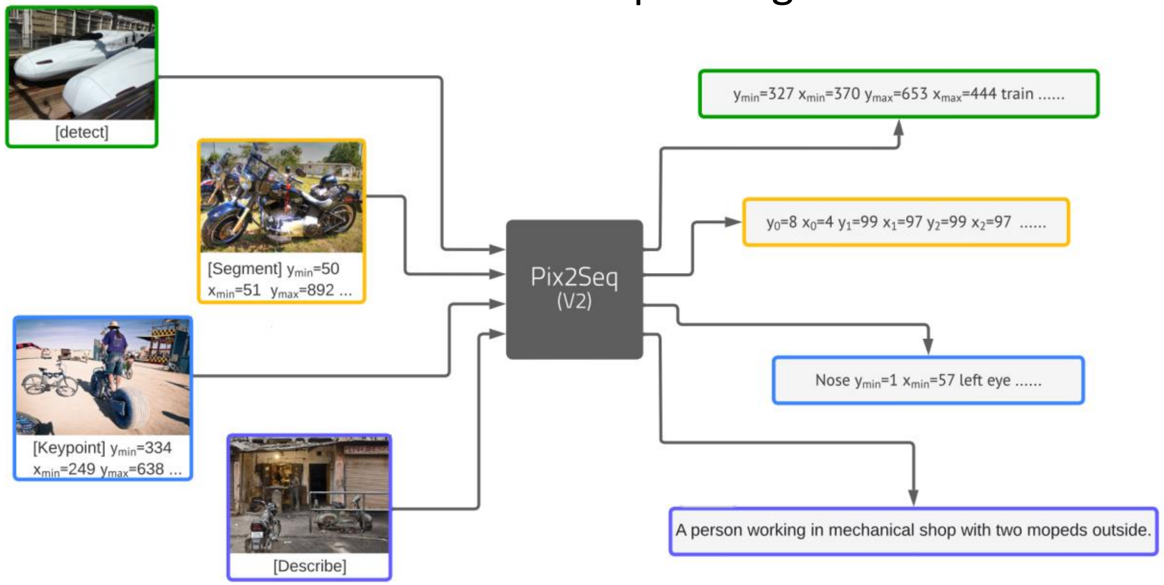
- UniTab and Pix2Seqv2: Unify text and box outputs with no specific modules



Method	Caption Eval. B@4	M	C	S	Grounding Eval. $F1_{all}$	$F1_{loc}$
NBT [49]	27.1	21.7	57.5	15.6	-	-
GVD [86]	27.3	22.5	62.3	16.5	7.55	22.2
Cyclical [50]	26.8	22.4	61.1	16.8	8.44	22.78
POS-SCAN [88]	30.1†	22.6†	69.3†	16.8†	7.17	17.49
Chen *et al.* [9]	27.2	22.5	62.5	16.5	7.91	21.54
UniTAB	**30.1**	**23.7**	**69.7**	**17.4**	**12.95**	**34.79**

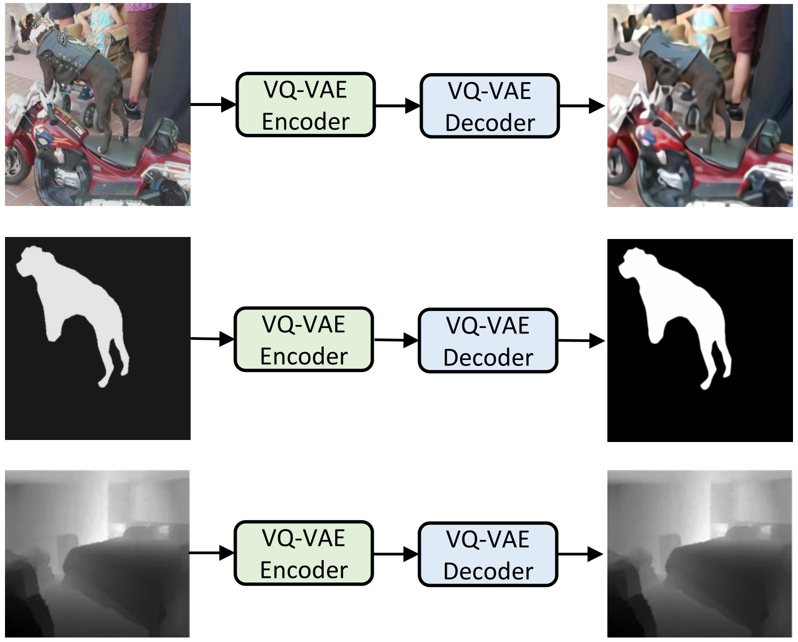
Grounded Captioning Evaluation

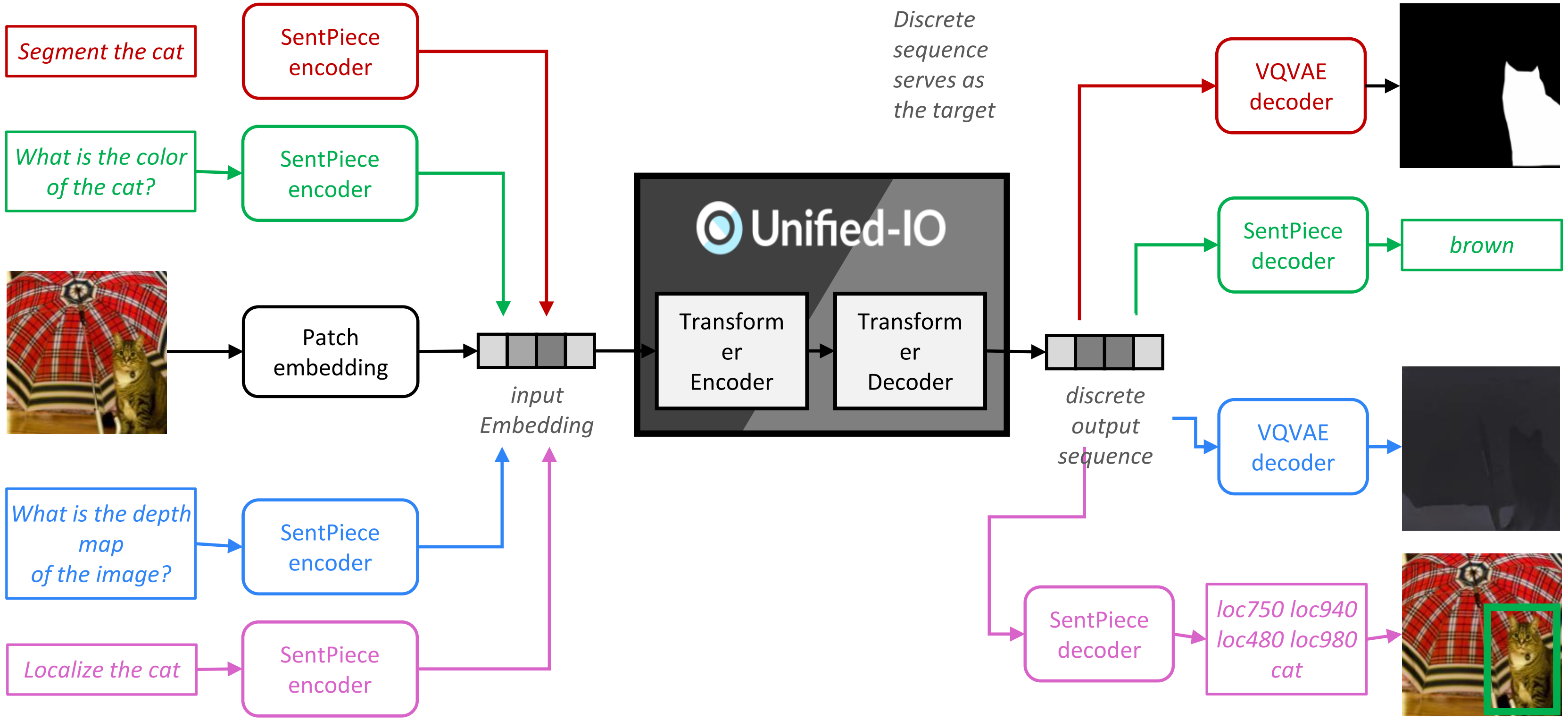
- Common vocabulary: text and coordinates are both tokenized and put into the same vocabulary
- Task prefix: requires a task prefix to determine which task the model is coping with



Outputs Unification

- Unified-IO: unify a wide range of understanding tasks including segmentation
  - Output Quantization: VQVAE for different types of tasks, such as mask, depth, image. (shared by UViM and OFA to some extent)
  - Two-stage pretraining: 1) pretraining VQVAE; 2) jointly pretraining on multiple tasks in a seq-to-seq manner



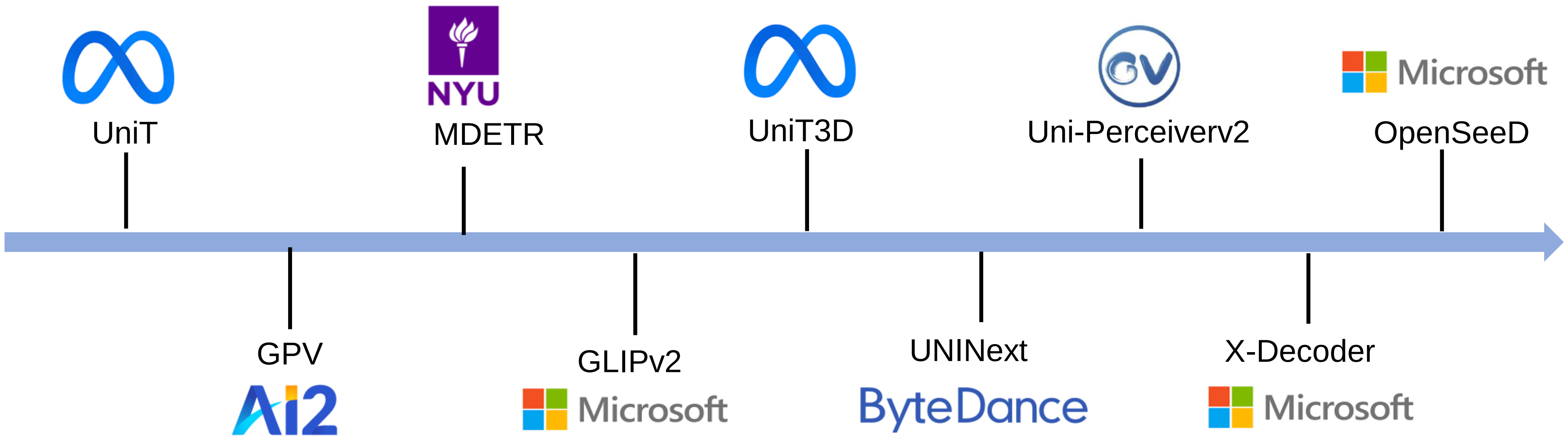


Outputs Unification

- Other works like VisionLLM use LLM as the output interface
- It unifies a wide range of vision tasks so that an encoder-decoder can be trained end-to-end
- It also:
  - needs task-specific decoder to decode the sequence to final outputs:
    - E.g., extract coordinates and translate into a box, convert polygon/color map into mask
  - might be hard to interpret the interactions across different tasks of different granularities
  - may not be able to build a strong cross-task synergy as we expect

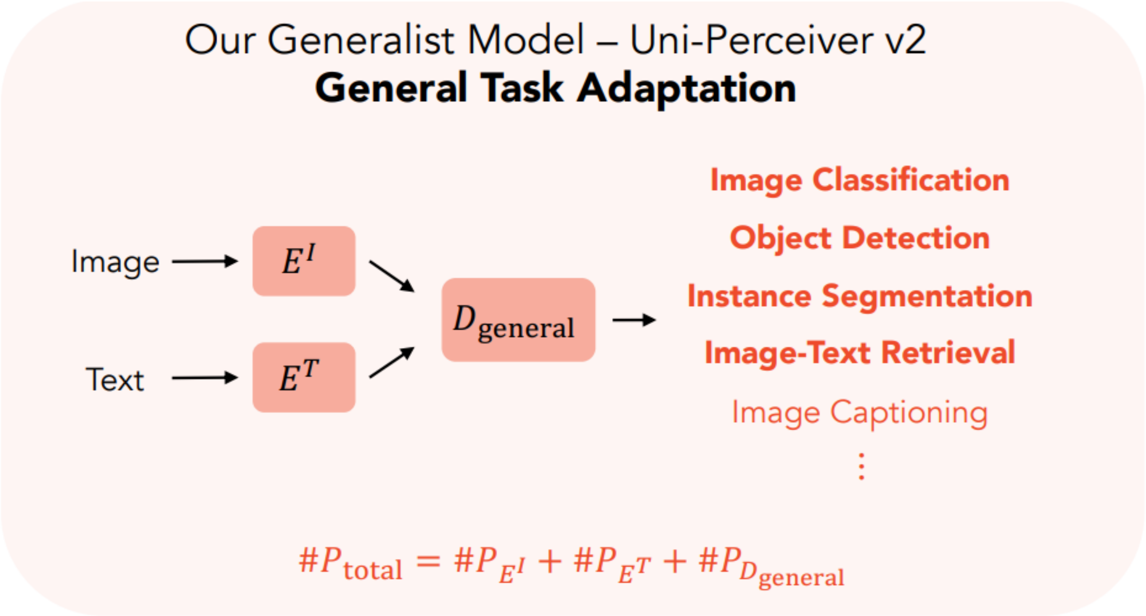
Functionality Unification

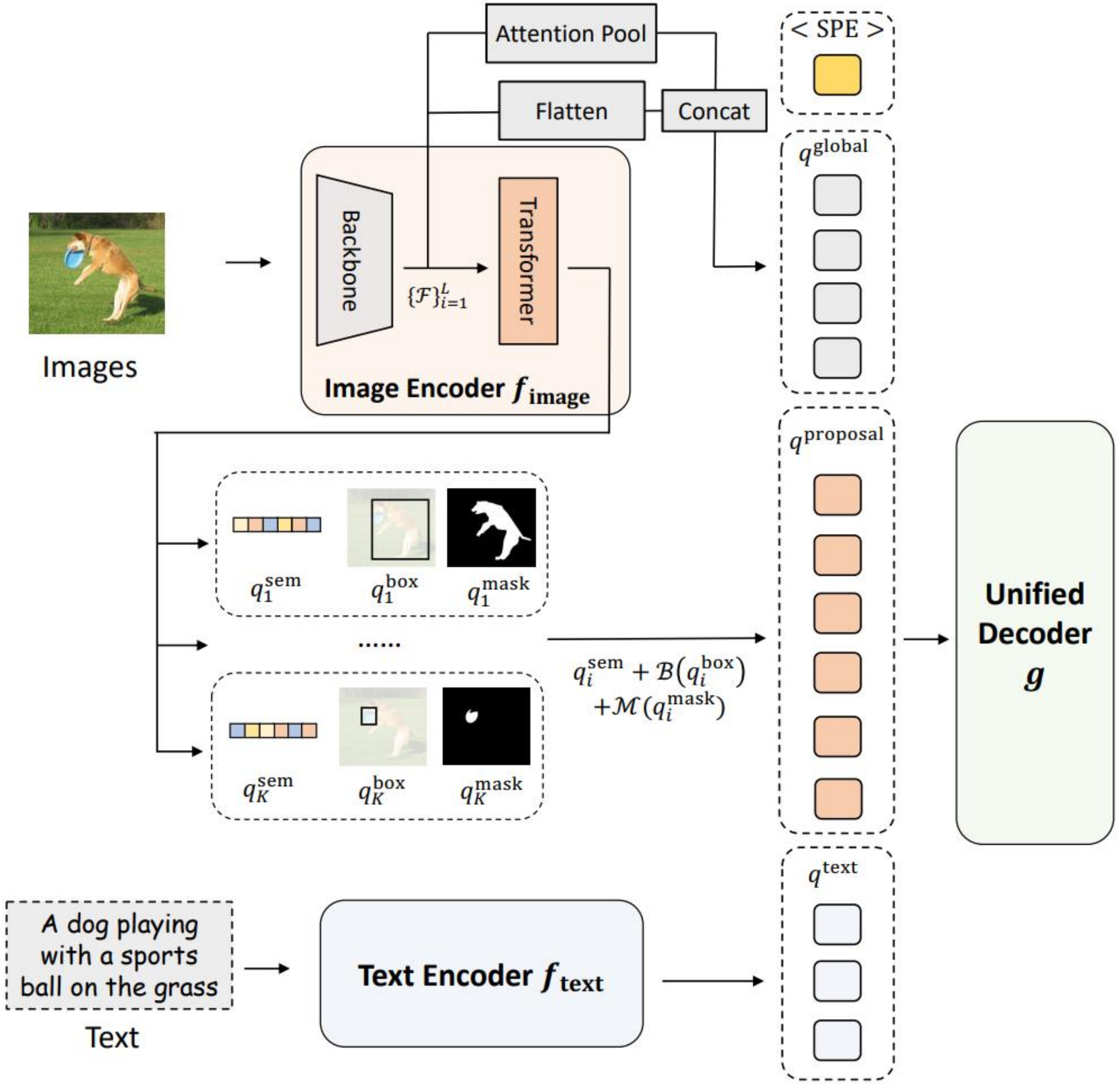
- Vision tasks are not fully isolated:
  - Box outputs: shared by generic object detection, phrase grounding, regional captioning
  - Mask outputs: shared by instance/semantic/panoptic segmentation, referring segmentation, exemplar-based segmentation, etc.
  - Semantic outputs: shared by image classification, image captioning, regional captioning, detection, segmentation, visual question answering, image-text retrieval, etc.



Functionality Unification

- UniPerceiver-v2: a unified decoder is exploited for many vision understanding tasks



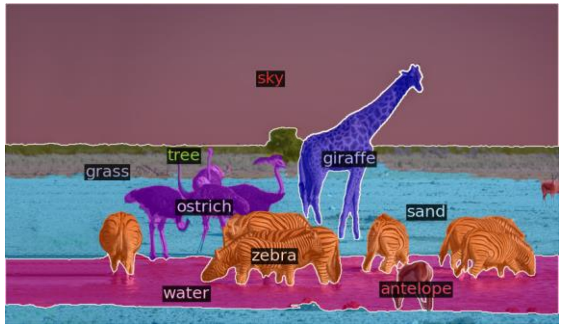


Functionality Unification

- X-Decoder: Generalized Decoding for Pixels, Images, and Language

Functionality Unification

- X-Decoder: Generalized Decoding for Pixels, Images, and Language

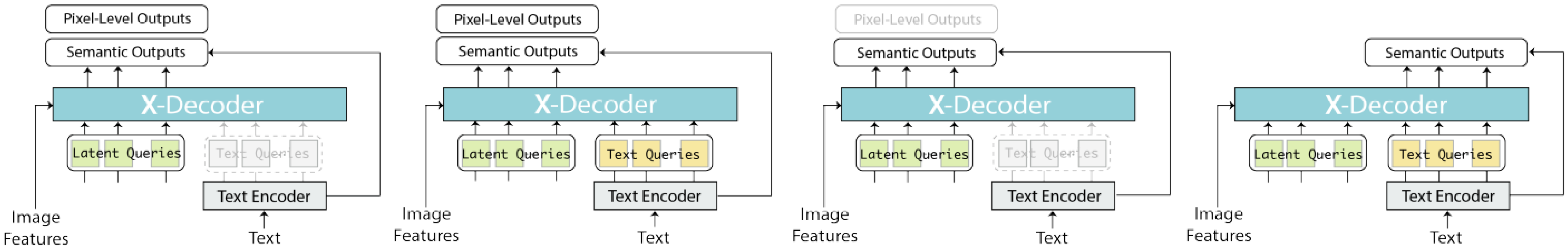


Query:Zebra,antelope,giraffe,ostrich,sky,water,grass,sand,tree

Query: Owl on the left

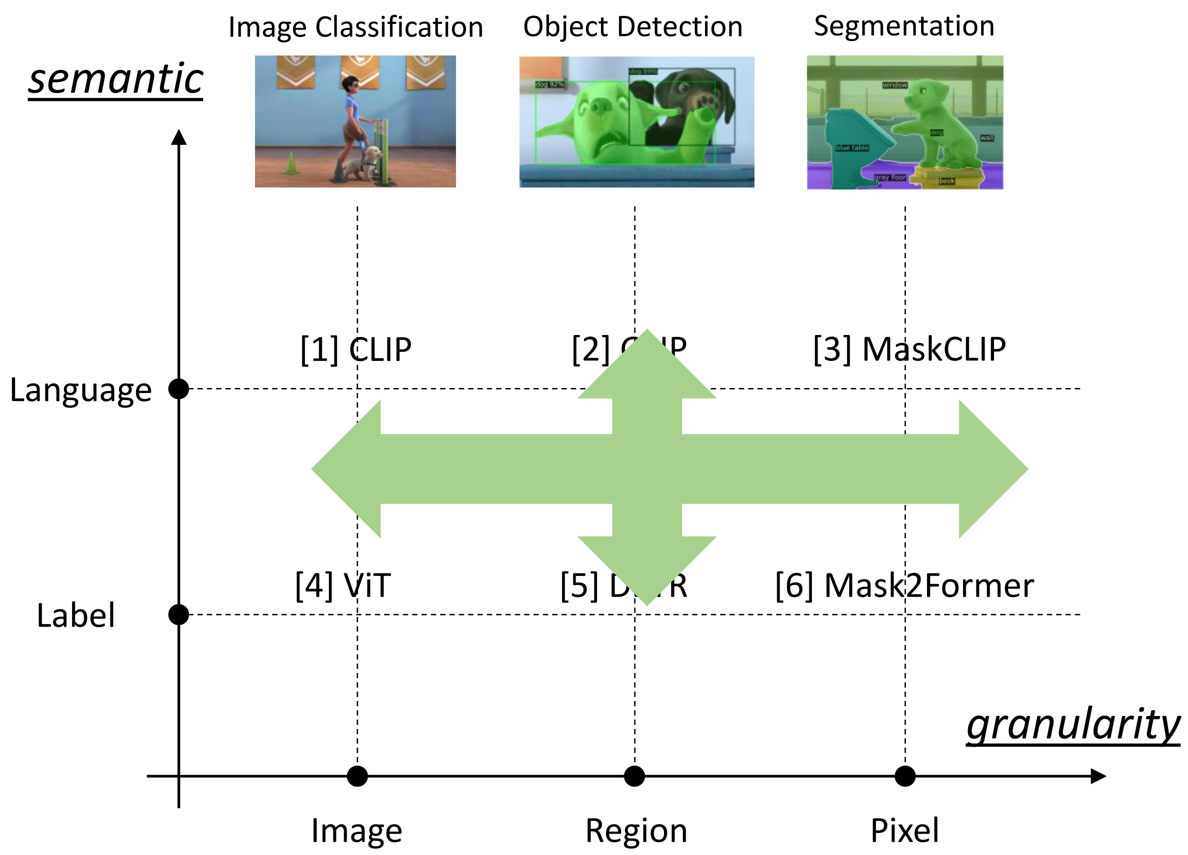
Query: The tangerine on the plate.

Cap: river in the mountains near the town



(a) Generic Segmentation (b) Referring Segmentation (c) Image-Text Retrieval (d) Image Captioning/VQA

Unify Different Granularities



Computer Vision in the Wild

ICinW Image Classification in the Wild

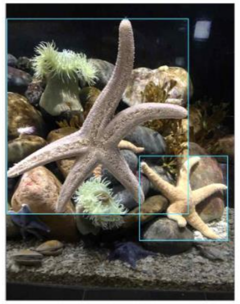
Example of knowledge sources

- Concept name: risotto
- Def_wik: An Italian savoury dish made with rice and other ingredients
- Def_wn: rice cooked with broth and sprinkled with grated cheese
- Path_wn: [risotto, dish, nutriment, food, substance, matter, physical_entity, entity]
- GPT3: ["A rice dish made with arborio rice and typically served with meat or fish.", "A rice dish made by stirring rice into a simmering broth]

Task preview

ODinW Object Detection in the Wild

Example of knowledge sources



- Concept name: starfish
- Def_wik: Any of various asteroids or other echinoderms (not in fact fish) with usually five arms, many of which eat bivalves or corals by everting their stomach.
- Def_wn: echinoderms characterized by five arms extending from a central disk
- Path_wn: [starfish, echinoderm, invertebrate, animal, organism, living_thing, whole, object, physical_entity, entity]
- GPT3: A marine animal of class Asteroidea, typically having a central disk and five arms.

Task preview

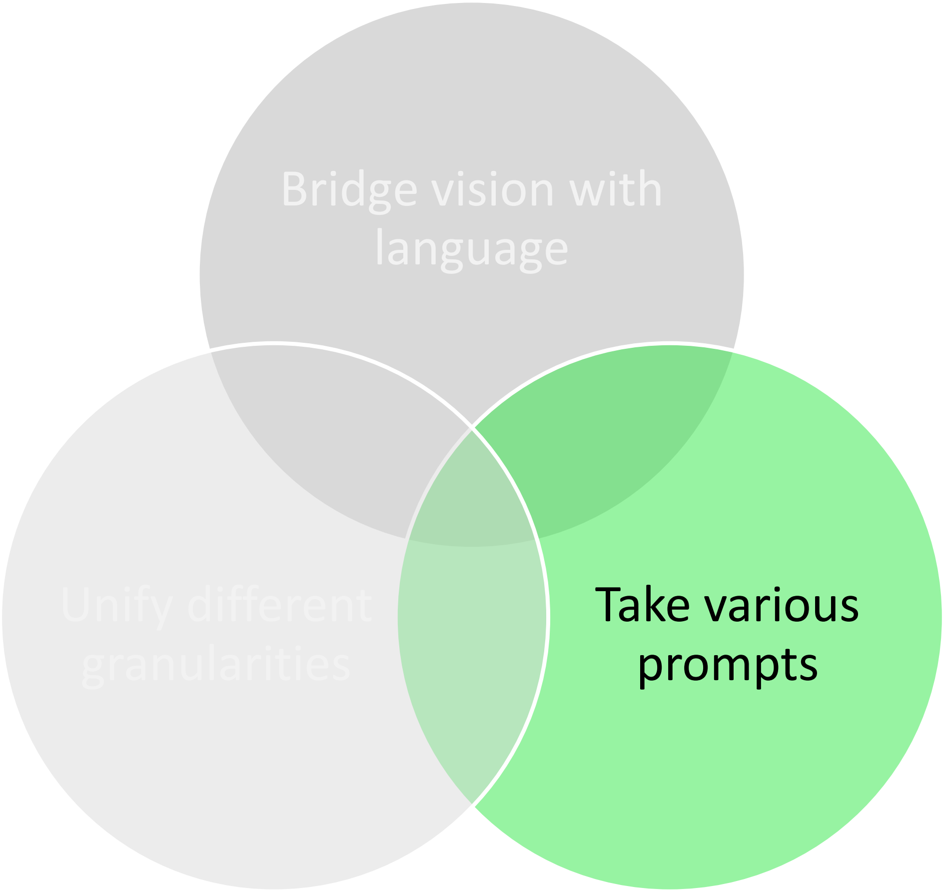
SGinW Segmentation in the Wild

Exemplar images in SGinW Benchmark

Task preview

2nd Workshop on Computer Vision in the Wild, East Ballroom B, June 19th full day

Promptable Interface



How to Enable Vision Model to “Chat”

Decoder LLMs (e.g., GPT) — Human-AI Interaction → Conversational AI (e.g., ChatGPT)

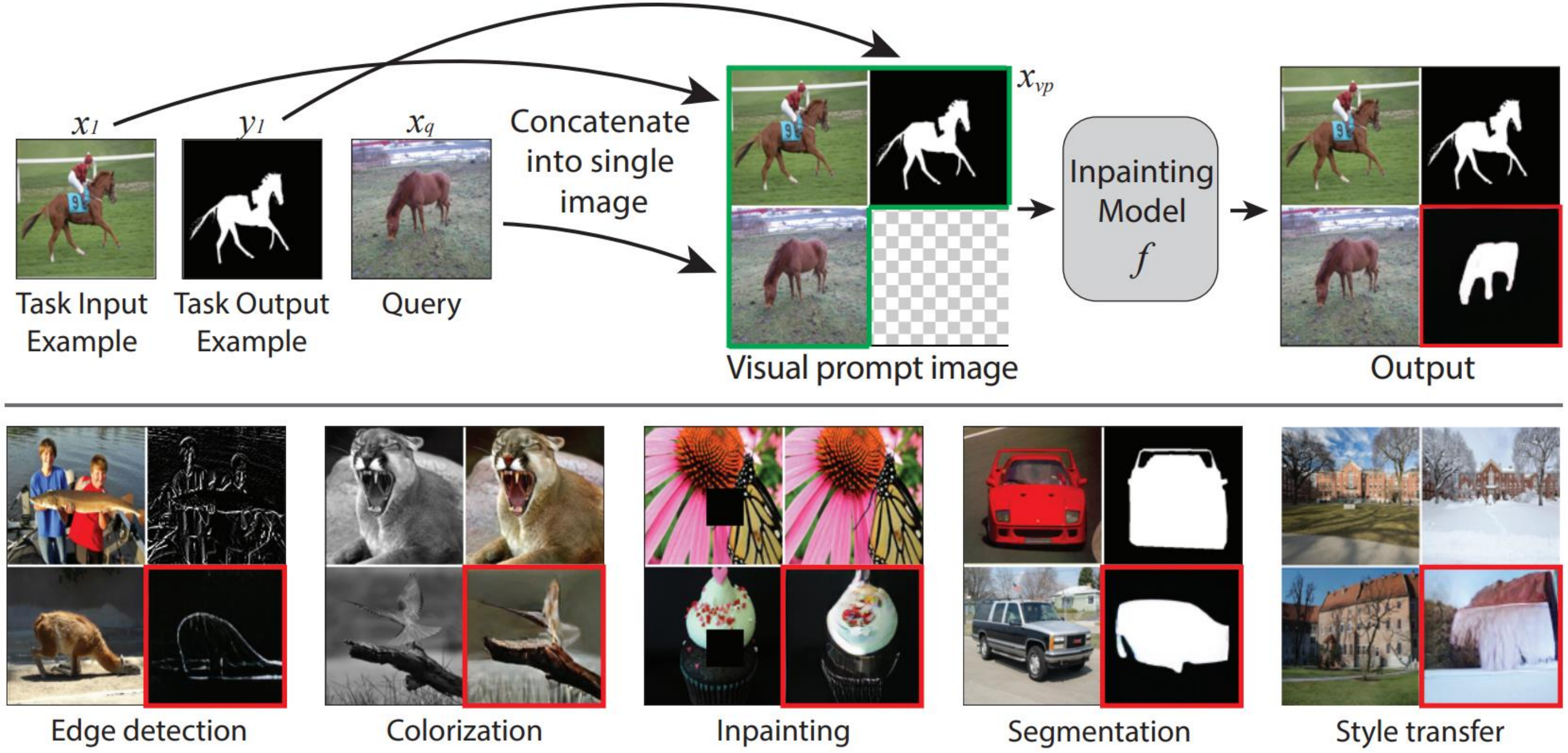
Generalist Vision Models — Human-AI Interaction → ?

How to Enable Vision Model to “Chat”

- We need to build a promptable interface with two important properties:
  - Promptable for in-context learning: Instead of finetuning the model parameters, simply providing some contexts will make the model precit
  - Interactive for user-friendly interface: multi-round of interaction between human and AI is important to finish complicated tasks.

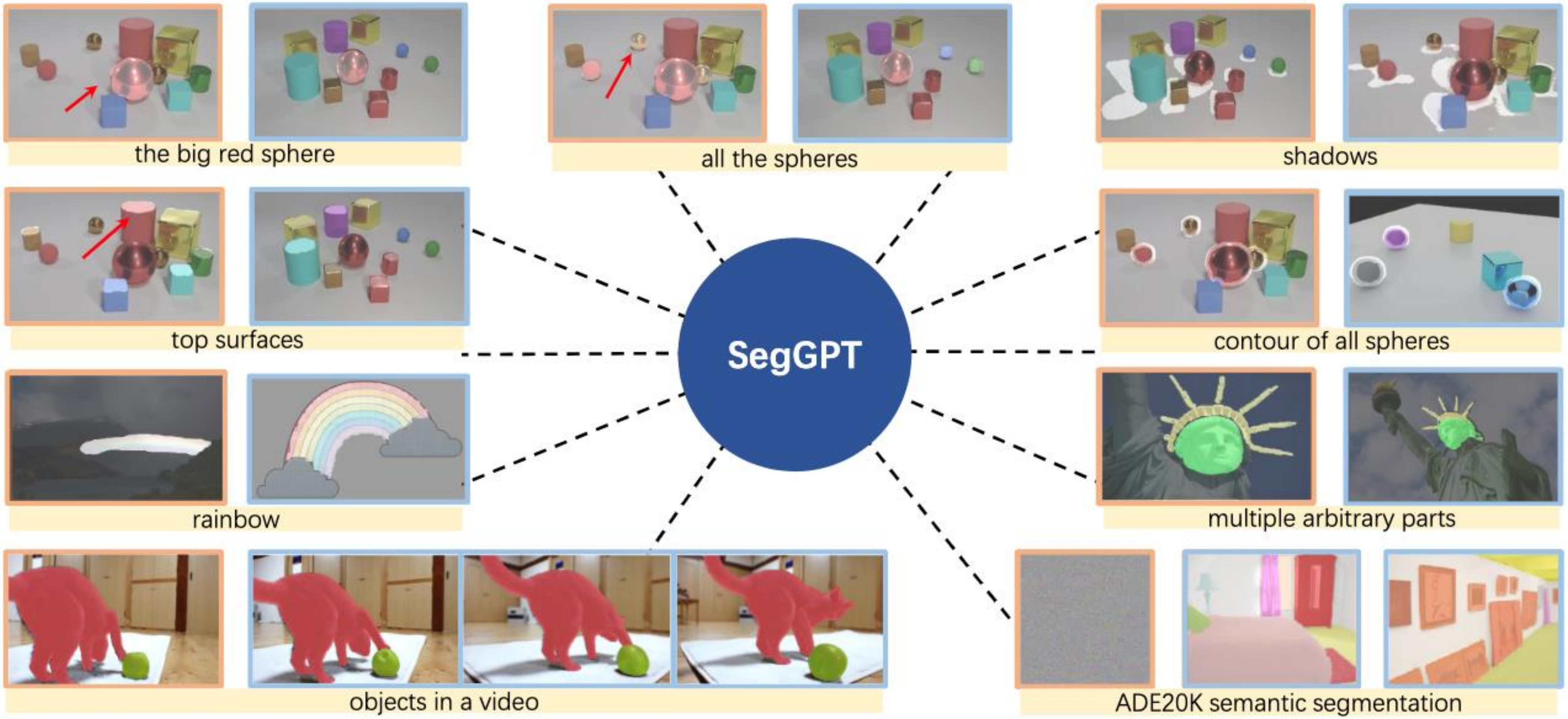
In-Context Learning for Vision

- Visual Prompting via Image Inpainting:
  - Concatenate in-context sample with query into a single image
  - Ask model to inpaint the missed part of the image grid



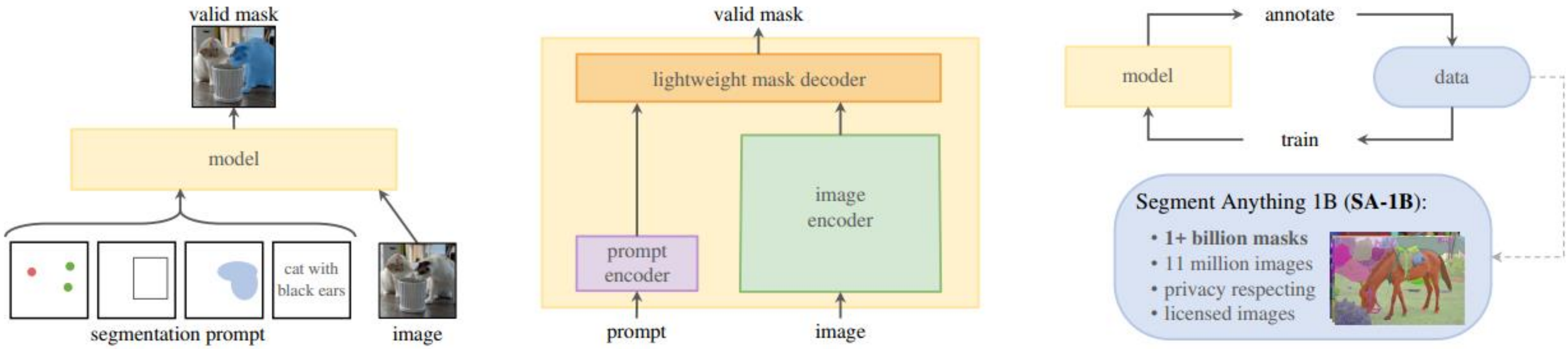
In-Context Learning for Vision

- SegGPT: Segment Everything as in-context learning

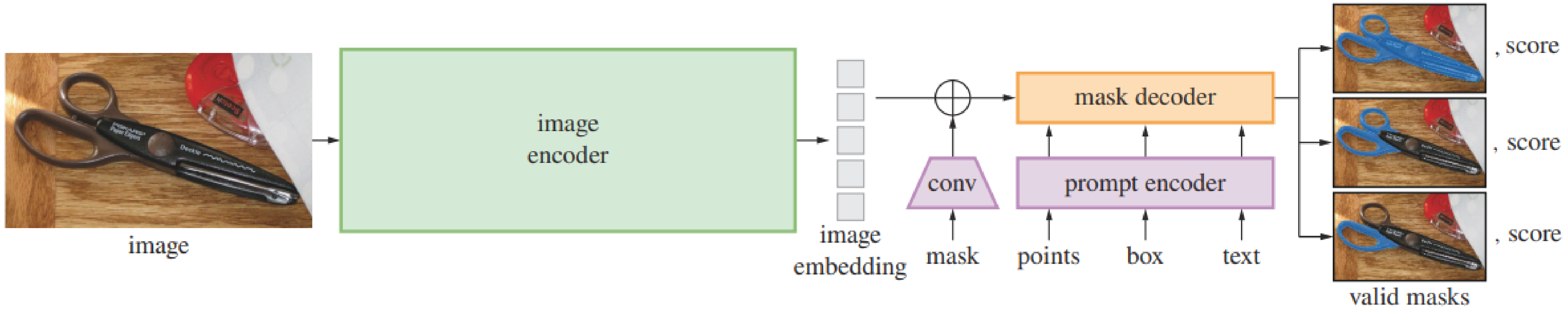


Interactive Interface for Vision

- SAM: Segment Anything
  - Promptable segmentation



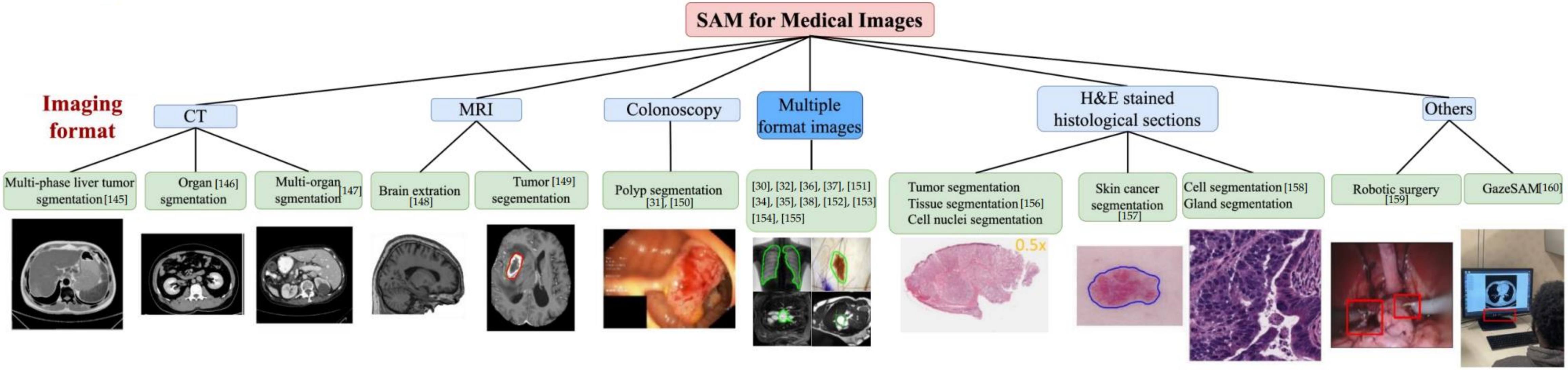
(a) **Task**: promptable segmentation (b) **Model**: Segment Anything Model (**SAM**) (c) **Data**: data engine (top) & dataset (bottom)



Interactive Interface for Vision

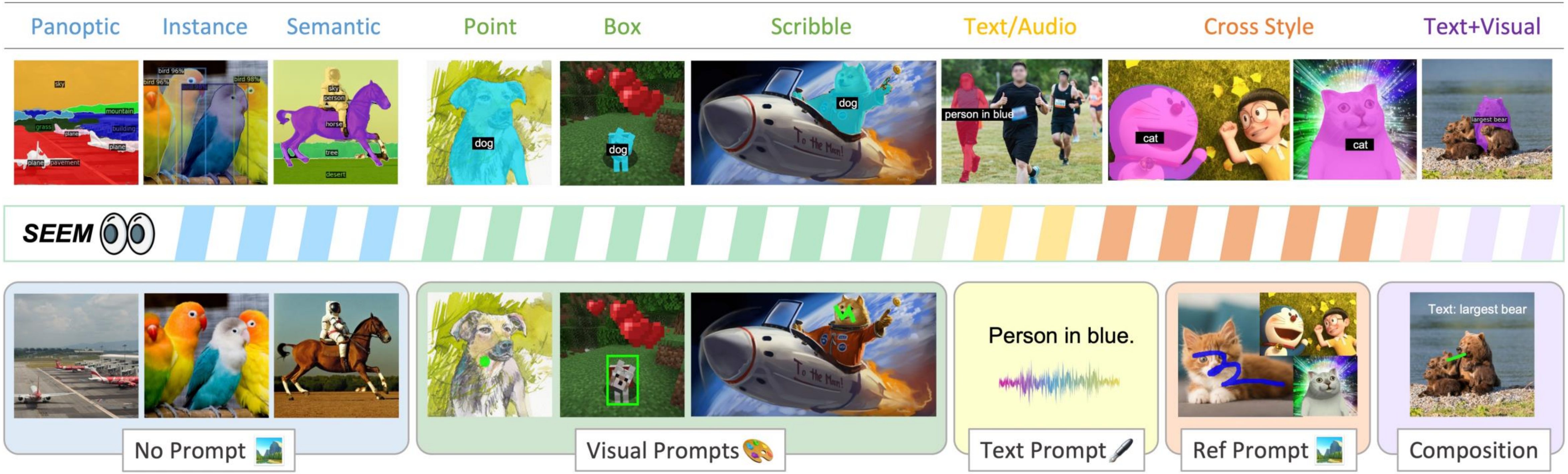
- SAM: Segment Anything

Text Prompt: Bench — Grounded-SAM Output — Stable-Diffusion Inpainting A Sofa, high quality, detailed



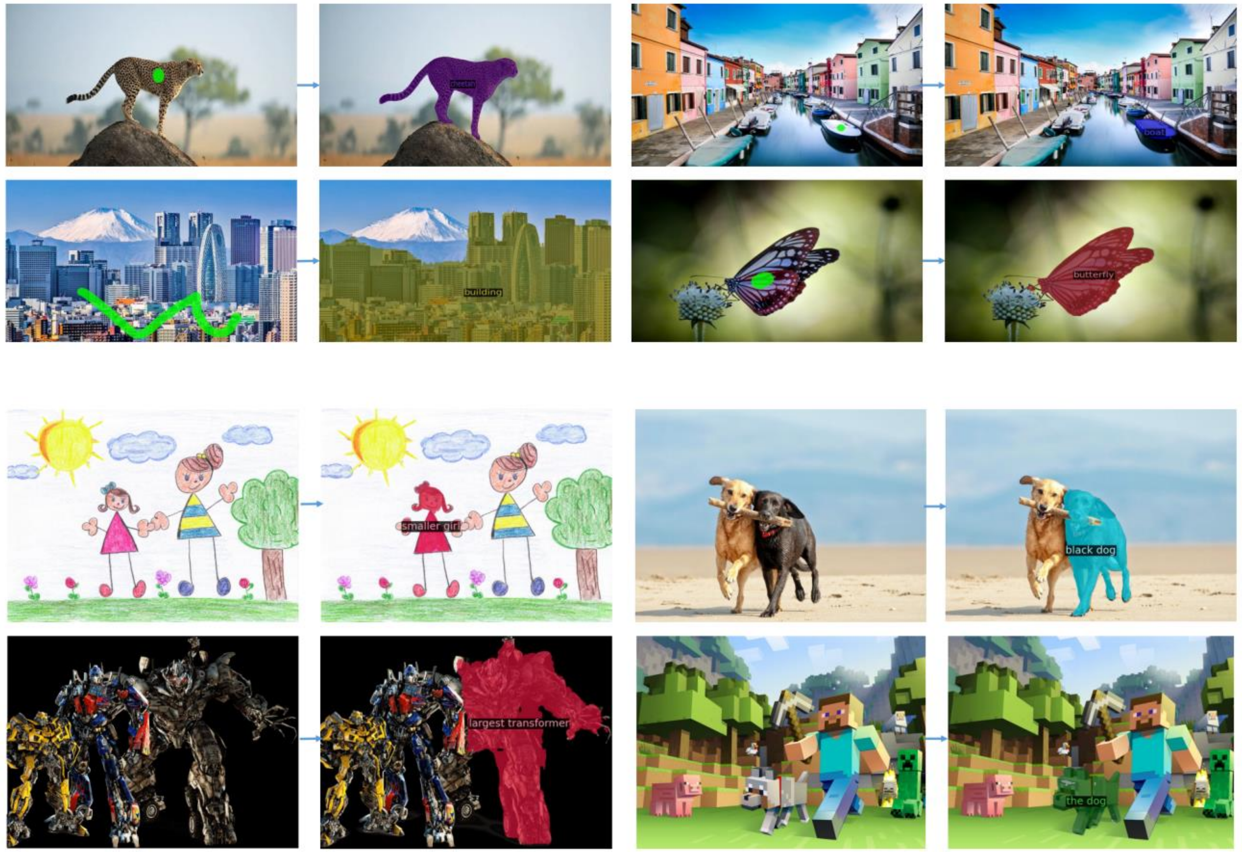
Interactive Interface for Vision

- SEEM: Segment Everything Everywhere all at Once



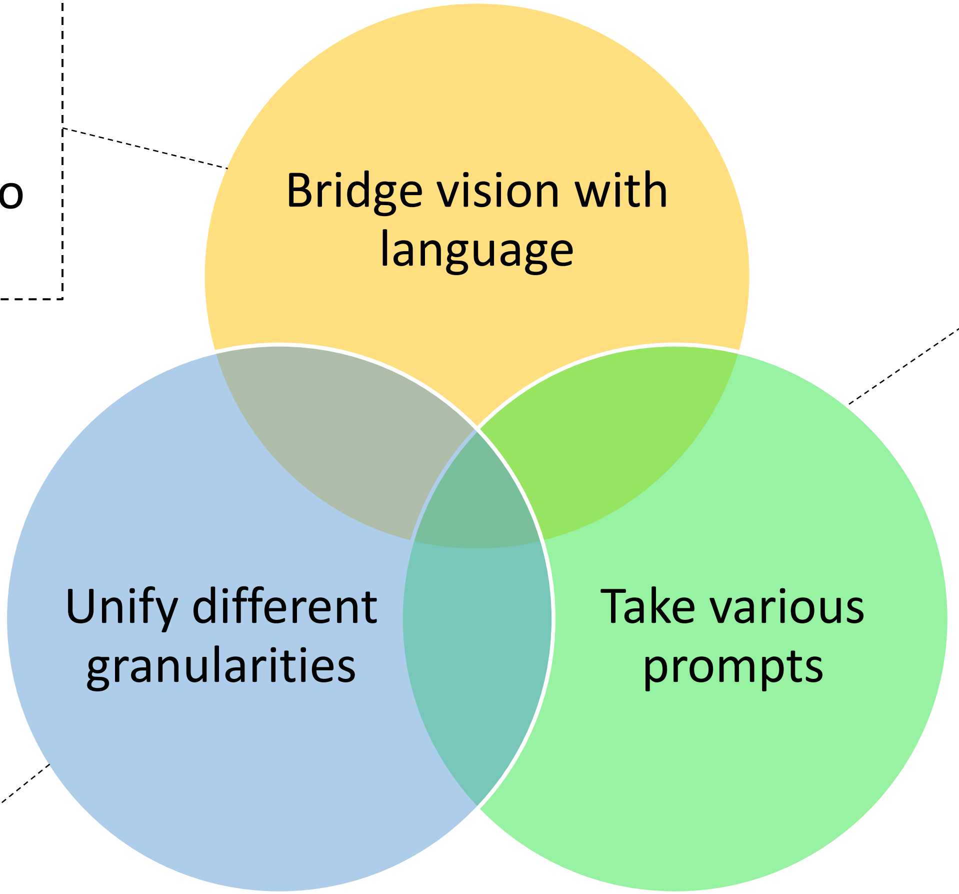
Interactive Interface for Vision

- SEEM: Segment Everything Everywhere all at Once



A quick recap

Intuition: Human use language as the common space to share information
Benefit: Zero-shot transfer to novel vocabularies



Intuition: Human uses both language, spatial prompts and beyond for vision.
Benefit: Reduce the ambiguity of expressing human intents

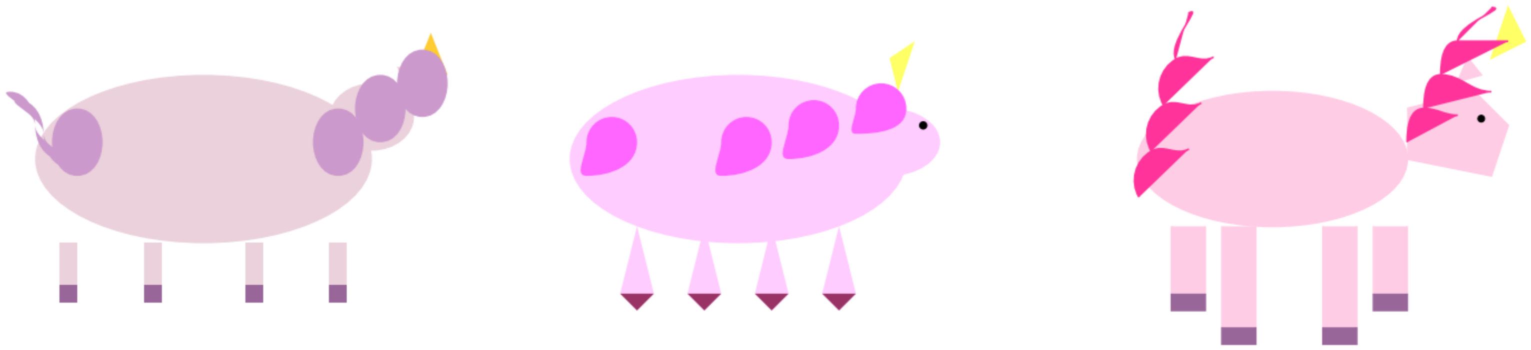
Intuition: Human vision is for multi-task, multi-granularity
Benefit: Build synergy across task granularities

Sparks of Artificial General Intelligence (AGI)

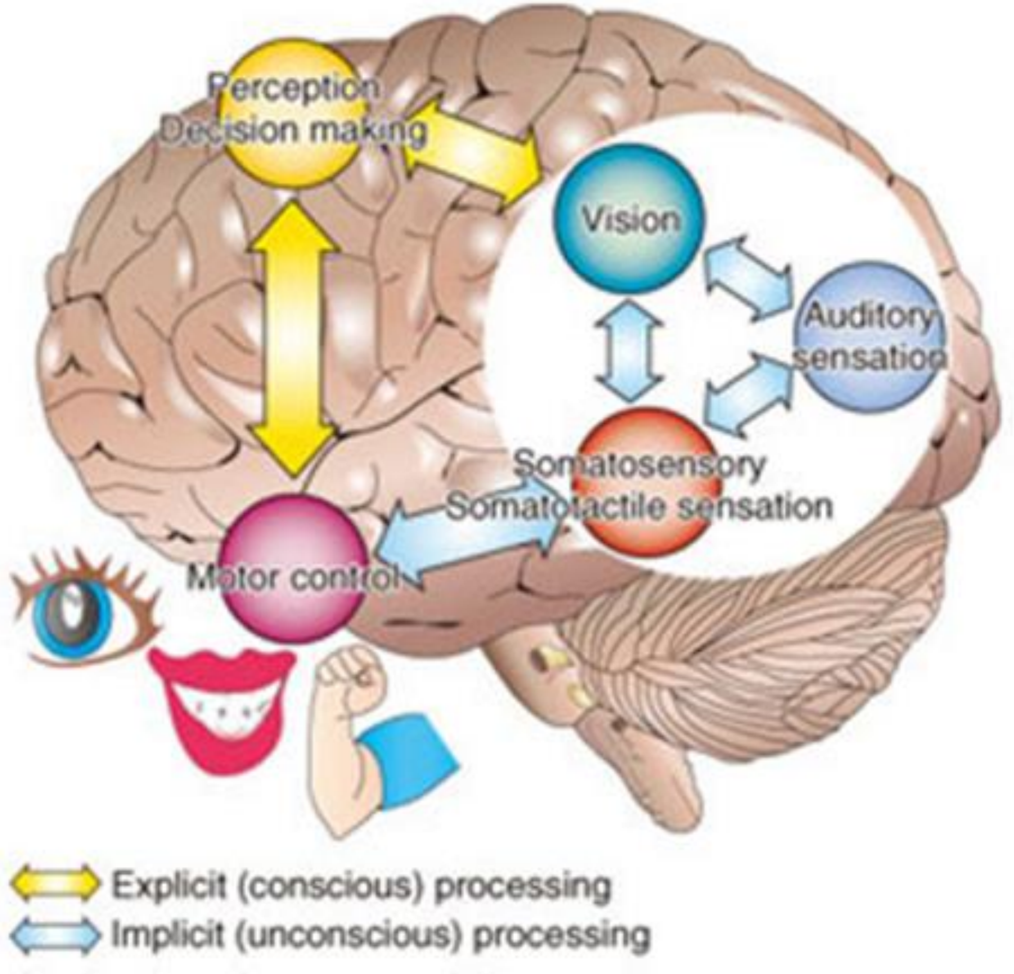
Sparks of Artificial General Intelligence: Early experiments with GPT-4

Sébastien Bubeck Varun Chandrasekaran Ronen Eldan Johannes Gehrke
Eric Horvitz Ece Kamar Peter Lee Yin Tat Lee Yuanzhi Li Scott Lundberg
Harsha Nori Hamid Palangi Marco Tulio Ribeiro Yi Zhang

Microsoft Research

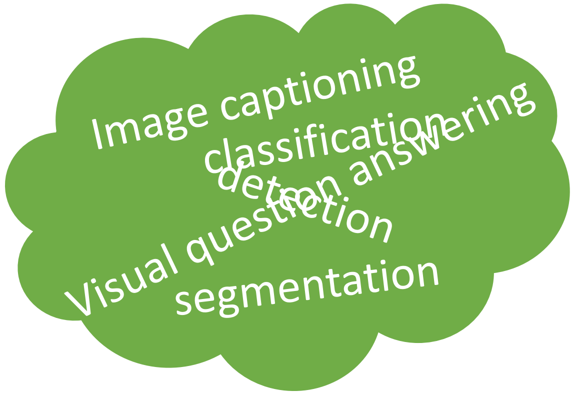


Artificial General Intelligence (AGI)



- Natural Language Processing
- Computer Vision
- Auditory sensation - Speech
- Motor control - Action
- …

Drawing dots for generalist vision to



Vision

– Unification → …… – Human-AI Interaction → ……

We are fortunate to have a lot of imagination space!!!

Enable an intimate cooperation with LLMs for physic world task

Give GPT, ChatGPT, BioGPT the eyes!

Empower more grounded image/video manipulation

Let DALLE-1/2 not only imaging things but grounding to the realistic!

Achieve multi-sensory general intelligent agent!

A real agent that can see, talk, act!

Thanks for your attention!