AIM 2: Artificial Intelligence in Medicine II

Harvard - BMIF 203 and BMI 702, Spring 2025

Lecture 11: Combining image and text modalities in AI (CLIP), Vision and vision-language pre-training, A general vision interface in LLMs, Multimodal LLMs

For the Study of Natural & Artificial Intelligence at Harvard University

Marinka Zitnik
marinka@hms.harvard.edu

Evolution of modeling paradigm

Task-specific Modeling

Training on *small-scale, well-annotated* data

Models are developed with a task-specific approach to learning

Atelectasis
Cardiomegaly
Consolidation
Edema
Effusion
Emphysema
Fibrosis
Hernia
Infiltration
Mass
Nodule
Pleural Thickening
Pneumonia
Pneumothorax

Specialized models are designed for every new task and every new dataset

Pneumothorax or not?

Stroke or not?

Evolution of modeling paradigm

Task-specific Modeling

Training on small-scale, well-annotated data

Early "Foundation" Models

Pre-training on *large-scale, noisy* data

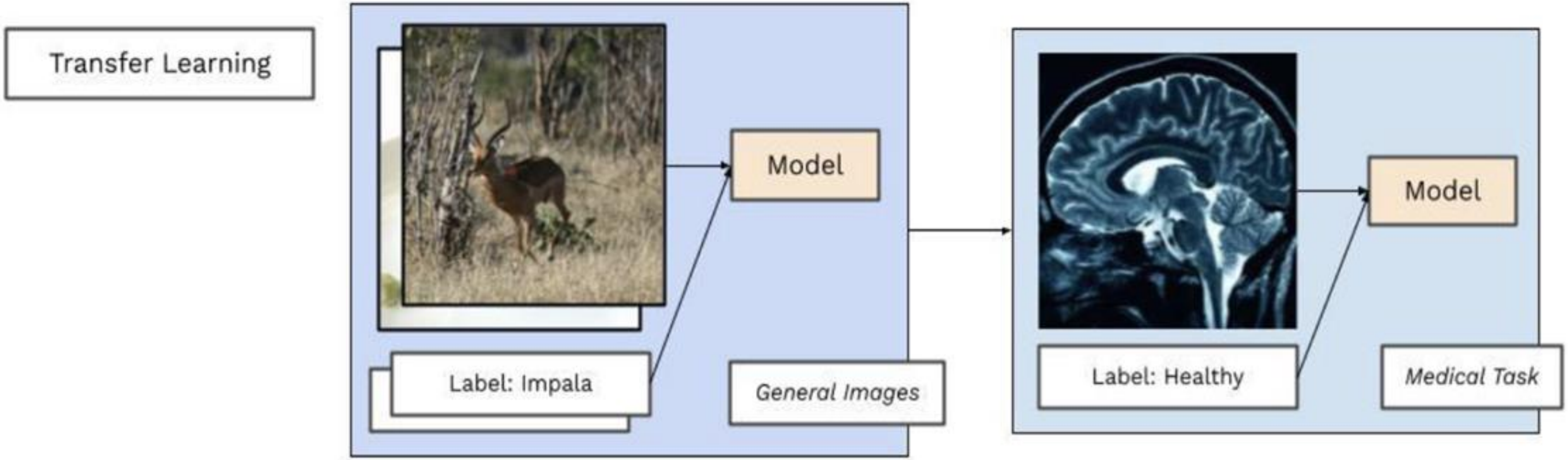
Task-specific finetuning on small-scale, well-annotated data

NLP: BERT, RoBERTa, T5, …
VL: UNITER, OSCAR, VinVL,…

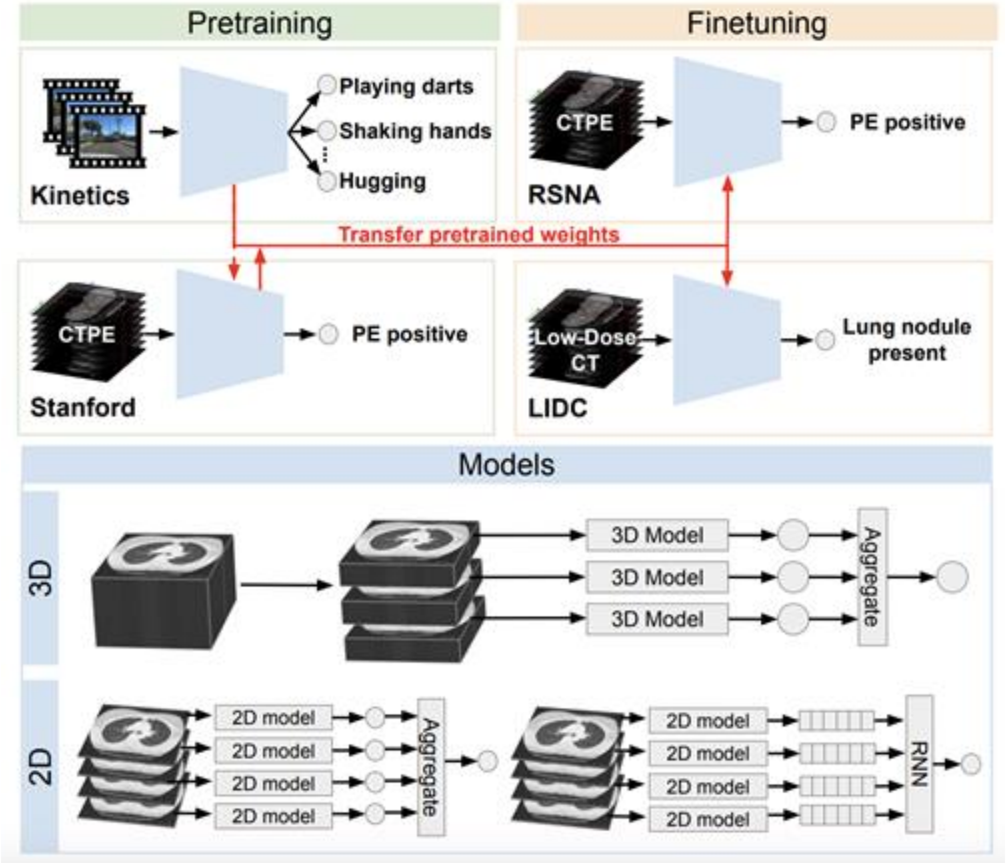
Foundation models

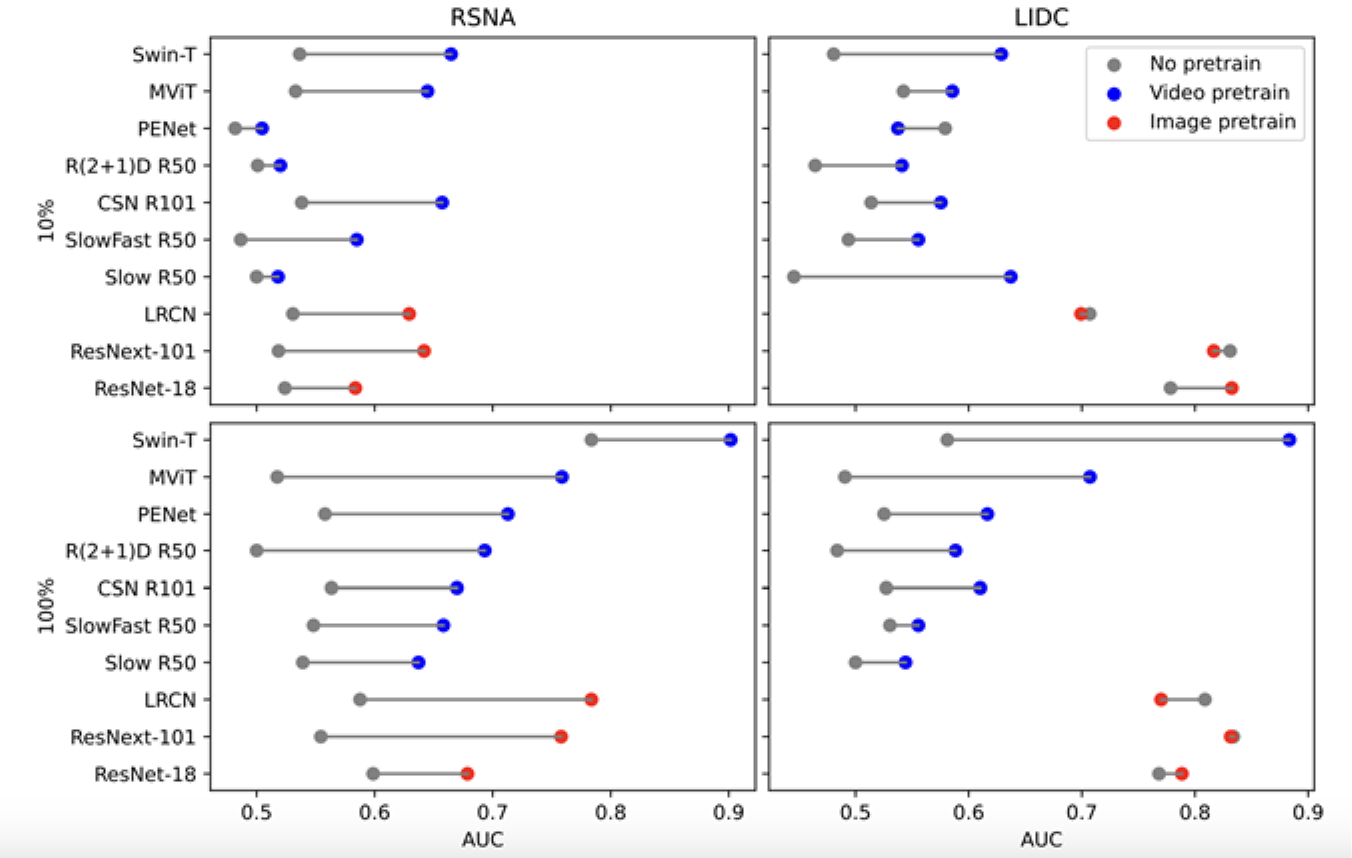
- These are pre-trained AI models that serve as a starting point for developing more specific AI models
- Foundation models are trained on large amounts of data, and can be fine-tuned for specific applications, such as detecting lesions or segmenting anatomical structures

Finetuning general models on a well-annotated, small-scale medical dataset



Finetuning general models on many annotated, small-scale medical datasets





Video Pretraining Advances 3D Deep Learning on Chest CT Tasks. *arXiv preprint arXiv:2304.00546*.

Evolution of modeling paradigm

Task-specific Modeling

Training on small-scale, well-annotated data

Early "Foundation" Models

Pre-training on large-scale, noisy data

Task-specific finetuning on small-scale, well-annotated data

NLP: BERT, RoBERTa, T5, …
VL: UNITER, OSCAR, VinVL,…

Nowadays: *Generalist Modeling*

Pre-training on *XX..XLarge-scale*, noisy data

Zero-shot or In-context Few-shot with a few examples as demonstration

LLMs: GPT3, PaLM, LLaMa, …
LMMs: Flamingo, PaLM-E, GPT-4, …

Evolution of modeling paradigm

Task-specific Modeling

Training on small-scale, well-annotated data

Instruction-following Models

Pre-training on large-scale, noisy data

Instruction tuning on small-scale, pseudo-labeling data

NLP: Chat-GPT, Alpaca, Vicuna, …
VL: LLaVa, MiniGPT4, Otter, …

Generalist Modeling

Pre-training on XX..XLarge-scale, noisy data

Zero-shot or In-context Few-shot with A few examples as demonstration

LLMs: GPT3, PaLM, LLaMa, …
LMMs: Flamingo, PaLM-E, GPT-4, …

( , a dog is running through the grass)

Image Generation

Produce visual data

LLMs and models for image understanding and generation

Part 3: *How to make an LLM that can see and chat?*

Image Encoder

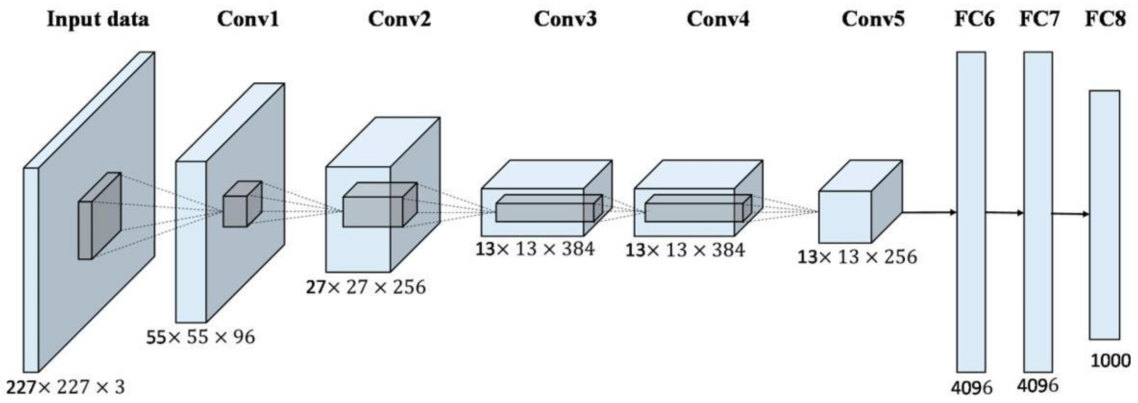
Consume visual data

Part 1: *How to learn image representations?*

Part 2: *How to extend vision models with more flexible, promptable interfaces?*

Part 1: Vision and Vision-Language Pre-training

Supervised Learning



Contrastive Language-Image Pre-training

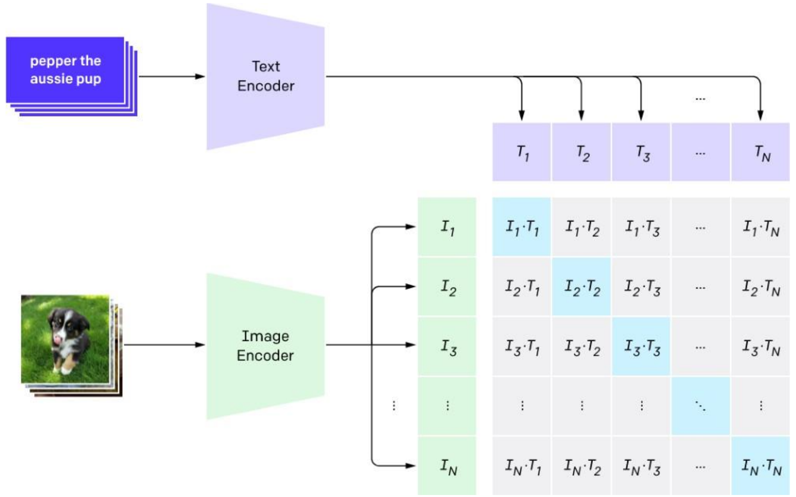
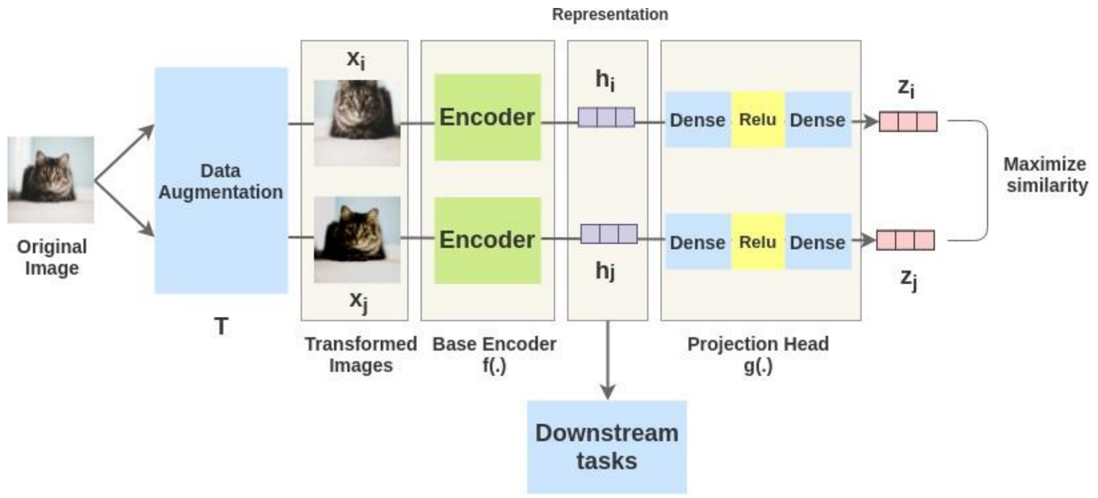
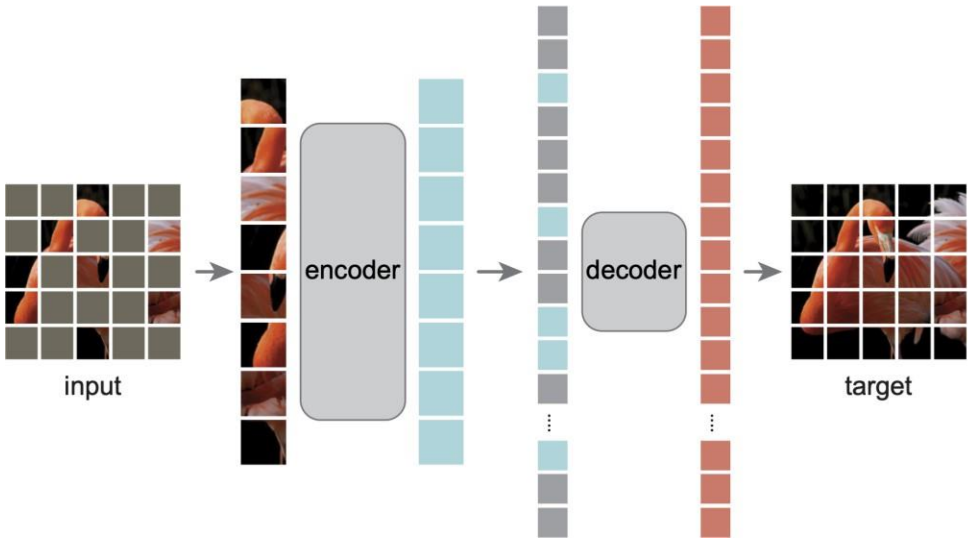


Image-only (Non-)Contrastive Learning



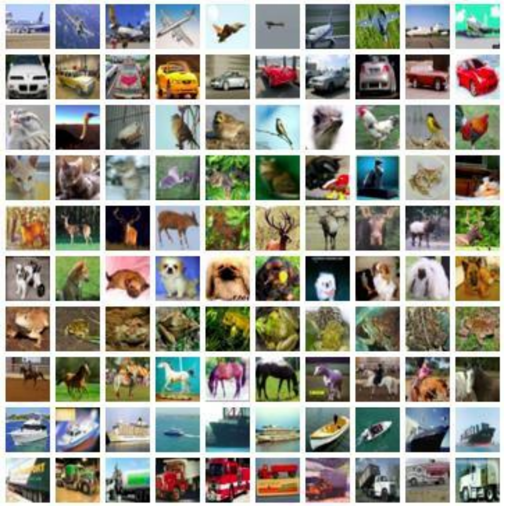
Masked Image Modeling (MIM)



Supervised learning

- Mapping an image to a *discrete label* which is associated to a visual concept
- Human annotation is expensive, and the labels can be limited
- *Private* datasets created by industrial labs:
  - JFT-300M, JFT-3B[1], IG-3.6B[2] (called weakly-supervised pre-training in this case)
  - Noisy weak supervision, can be very powerful for learning universal image embeddings

MNIST



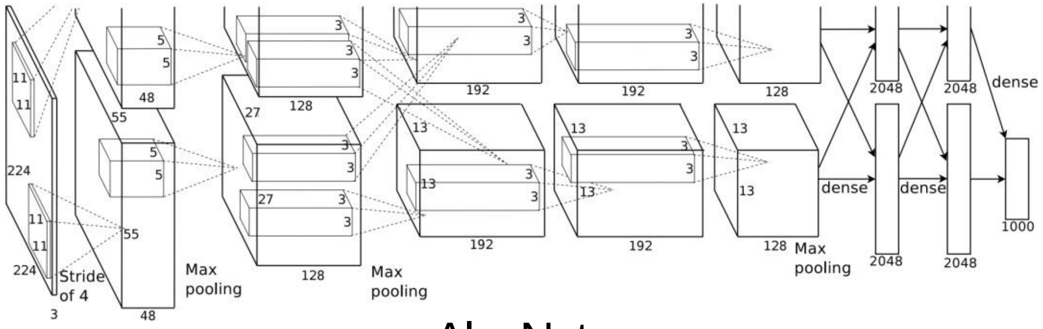
CIFAR-10

ImageNet

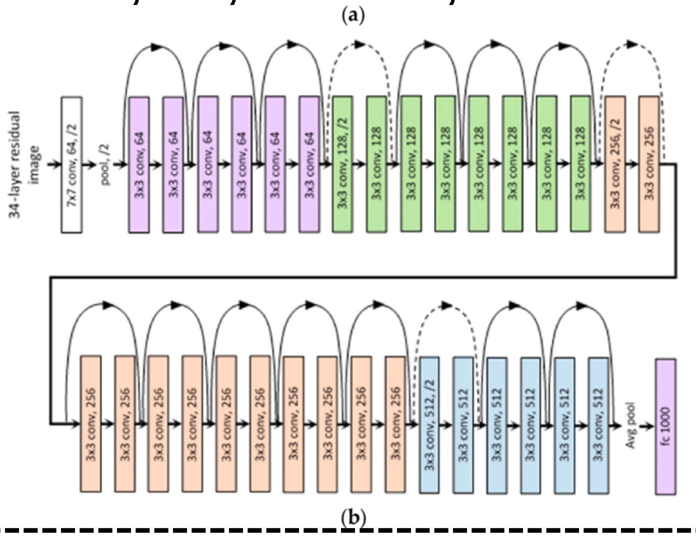
1 Scaling vision transformers, CVPR 2022
2 Revisiting weakly supervised pre-training of visual perception models, CVPR 2022

Supervised learning

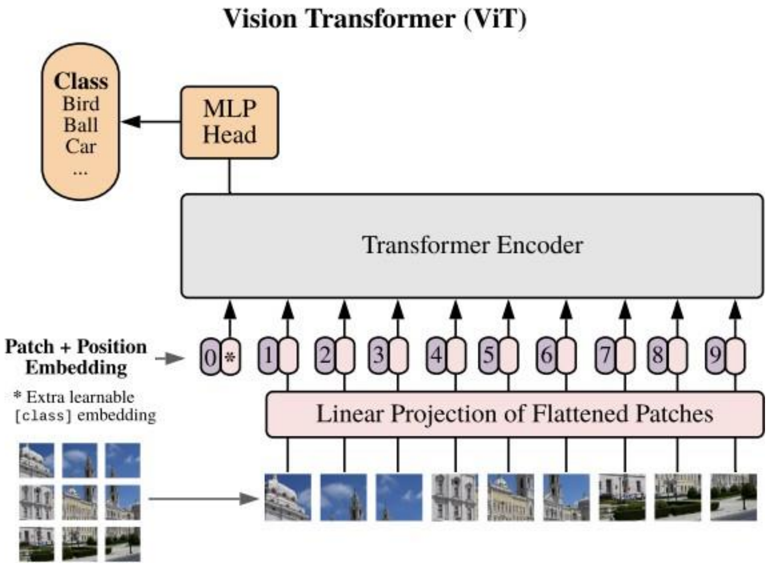
- Powered architectures ranging from AlexNet, ResNet, ViT, to Swin, and all the modern vision backbones



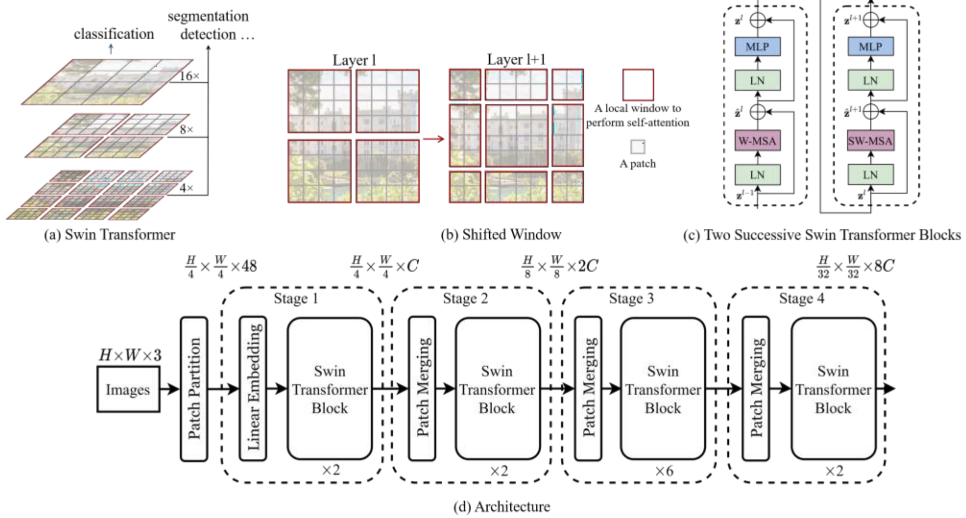
AlexNet



ResNet

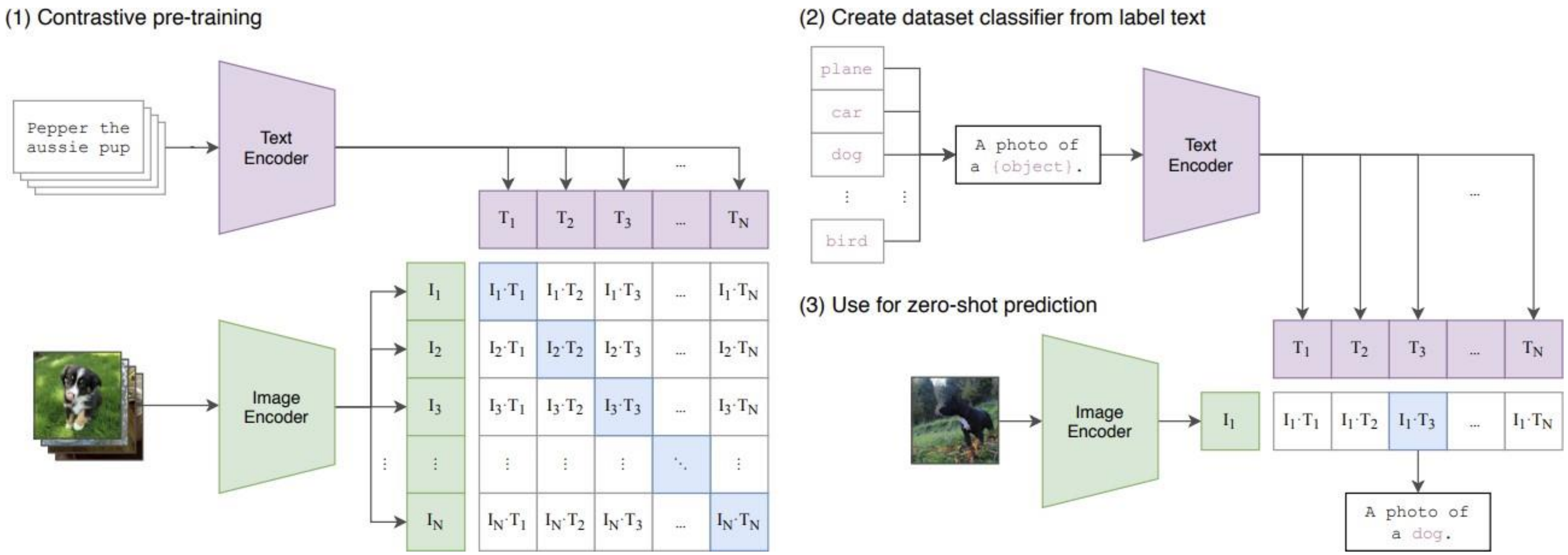


Swin



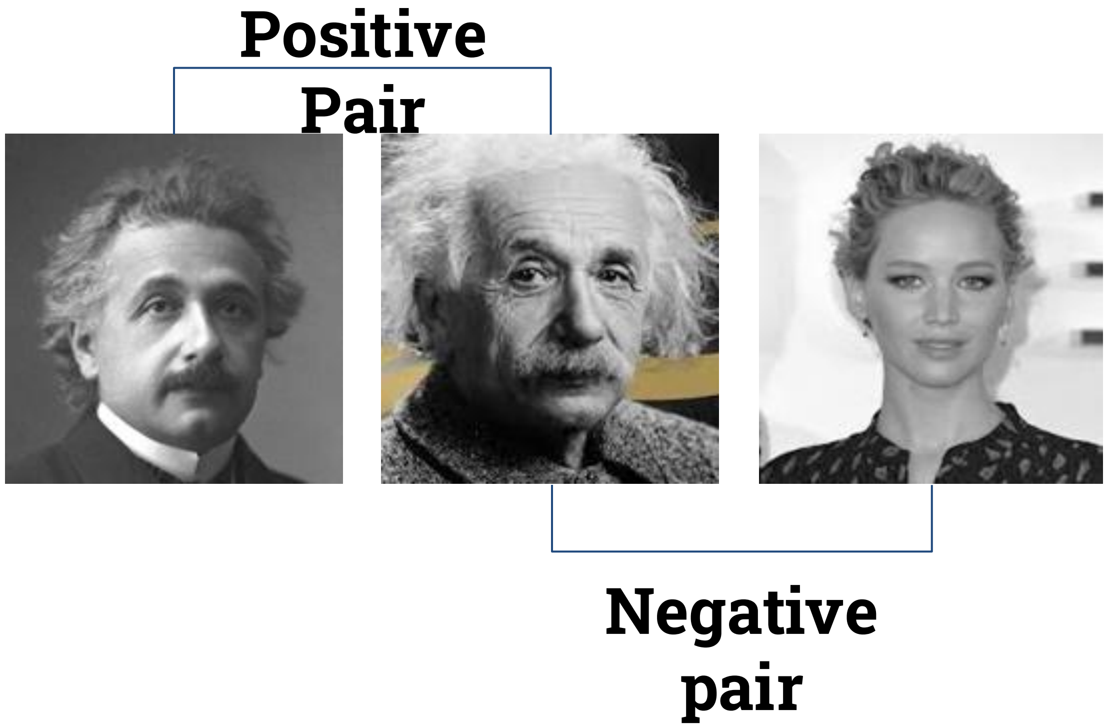
Contrastive language-image pre-training

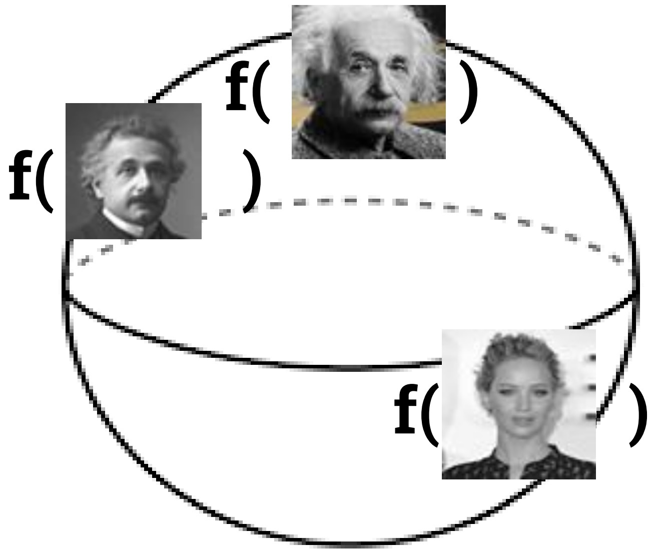
- Learning image representations from web-scale noisy text supervision
  - Training: simple *contrastive* learning, and the beauty lies in large-scale pre-training
  - Downstream: *zero-shot* image classification and image-text retrieval
    - Image classification can be reformatted as a retrieval task via considering the semantics behind label names



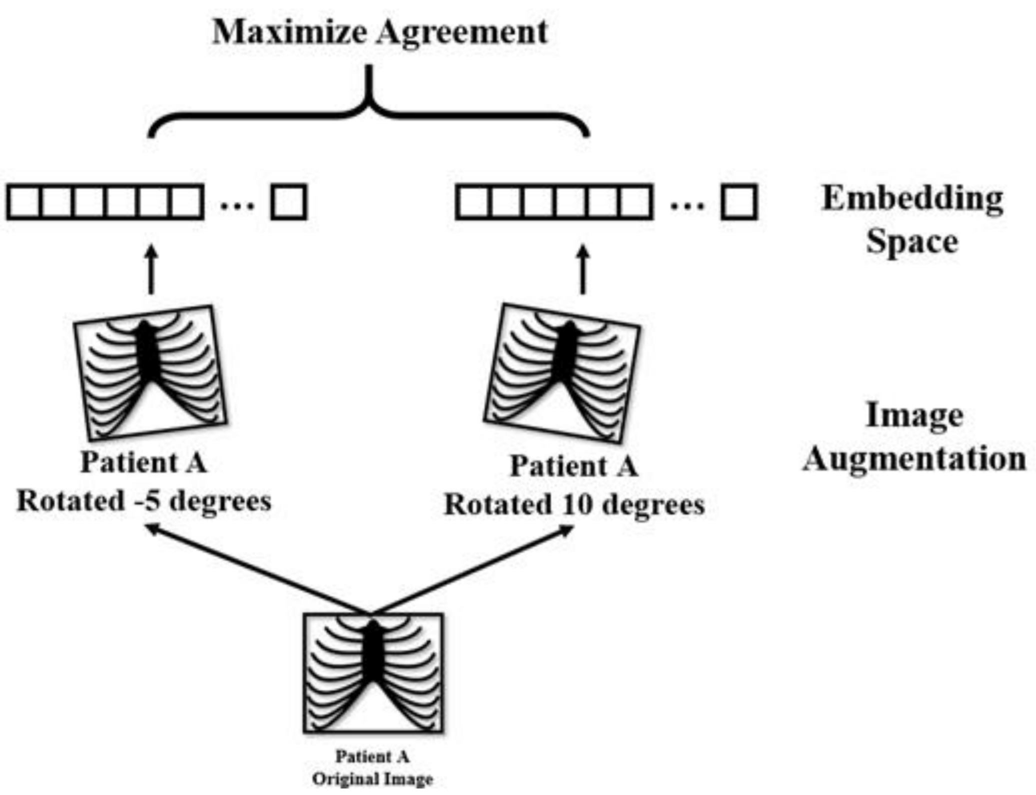
1 Learning transferable visual models from natural language supervision, ICML 2021
2 Scaling up visual and vision-language representation learning with noisy text supervision, ICML 2021

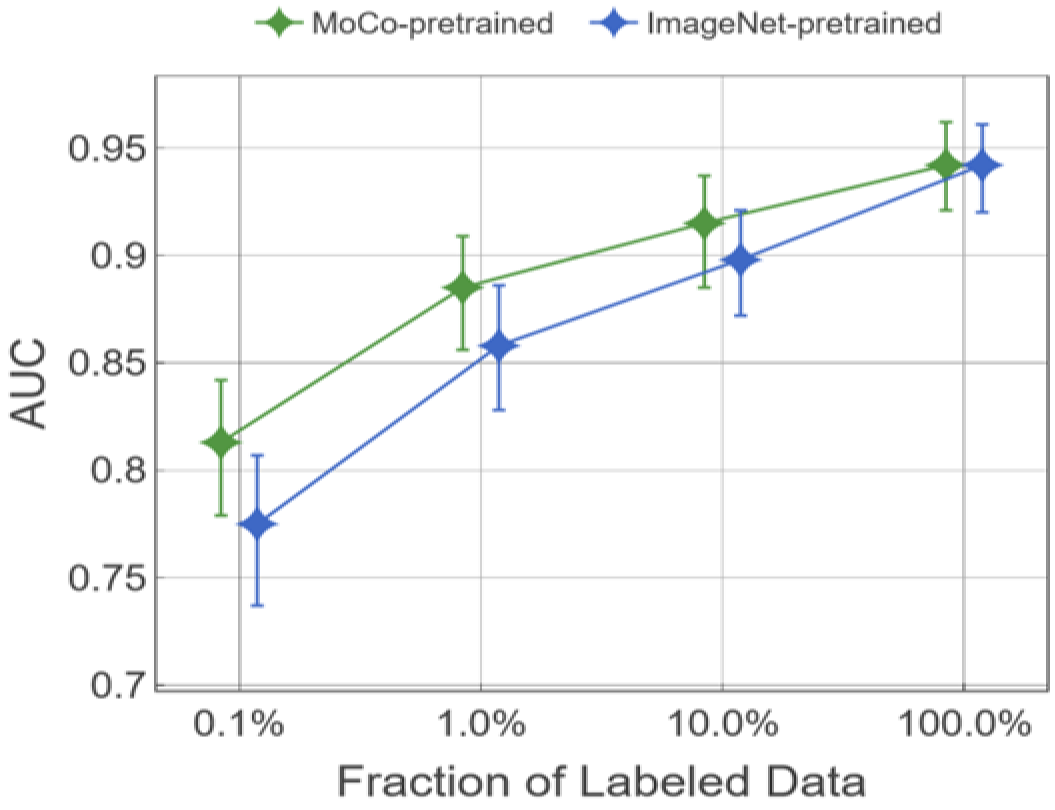
Contrastive pre-training makes similar samples represented more closely while pushing different samples far away





Contrastive pre-training using image augmentations can lead to label-efficient learning

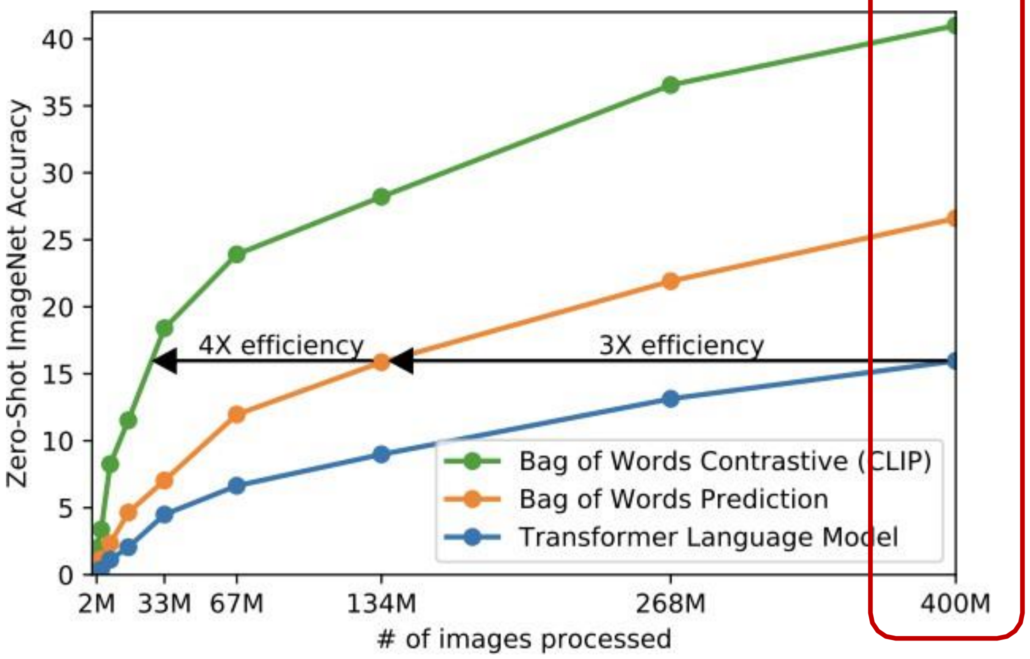




Moco pretraining improves representation and transferability of chest x-ray models. in Medical Imaging with Deep Learning 728–744 (PMLR, 2021)

Contrastive language-image pre-training

- The idea is simple, and can be dated back to a long while ago
  - In the large-scale pre-training era: CLIP[1] and ALIGN[2]
  - *Data scale* matters: Models are frequently trained with billions of image-text pairs
  - *Batch size* matters: 32k by default; *Model size* matters



Language is a stronger form of supervision than classical closed-set labels. Language provides rich information for supervision. Therefore, *scaling*, which can involve increasing capacity (model scaling) and increasing information (data scaling), is essential for attaining good results in language-supervised training.

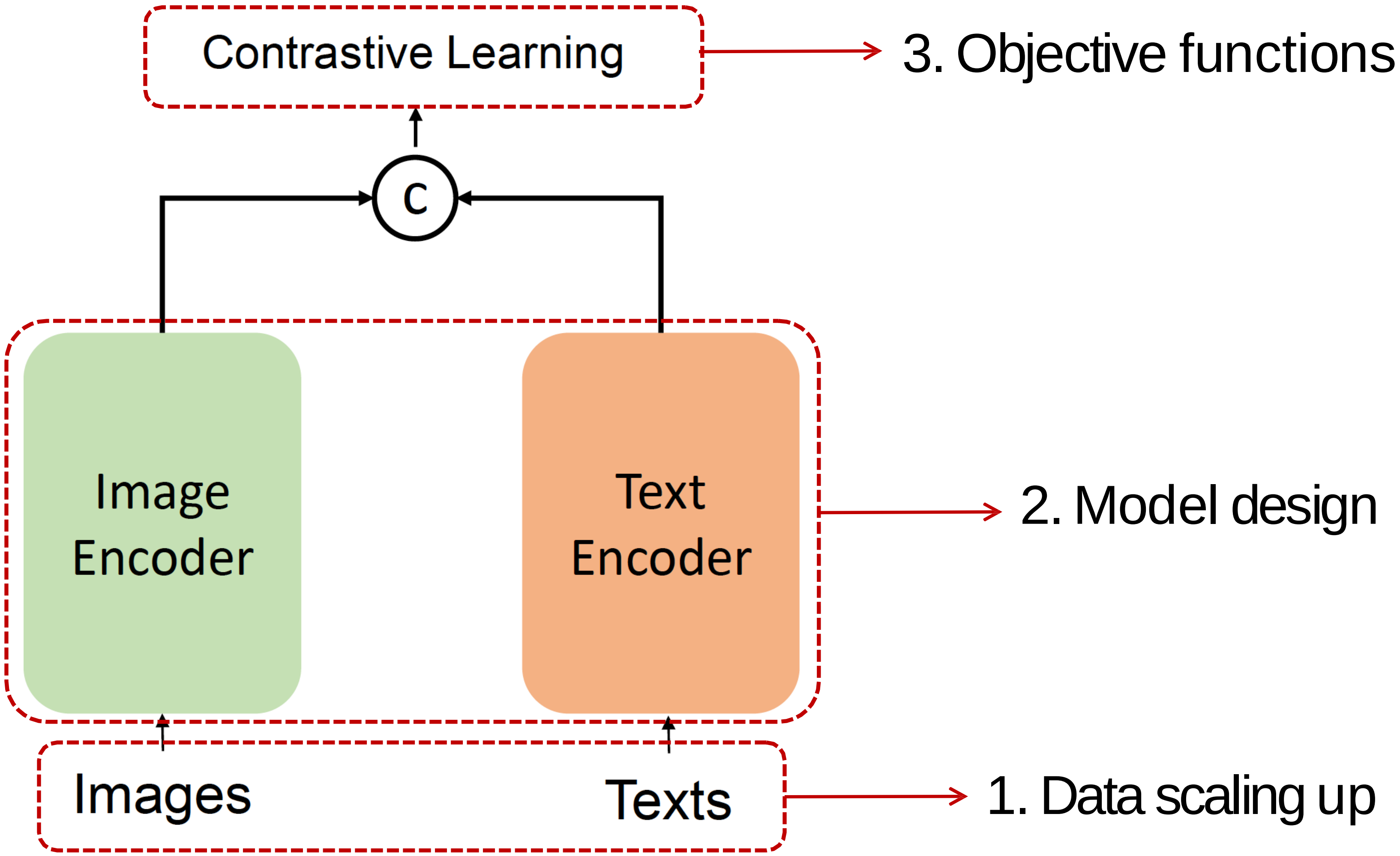
CLIP [52] is an outstanding example of "*simple algorithms that scale well*". The simple design of CLIP allows it to be relatively easily executed at substantially larger scales and achieve big leaps compared to preceding methods. Our method largely maintains the simplicity of CLIP

Quote from the FLIP paper

1 Learning transferable visual models from natural language supervision, ICML 2021
2 Scaling up visual and vision-language representation learning with noisy text supervision, ICML 2021

How to improve CLIP

- Since the birth of CLIP, tons of follow-up works and applications

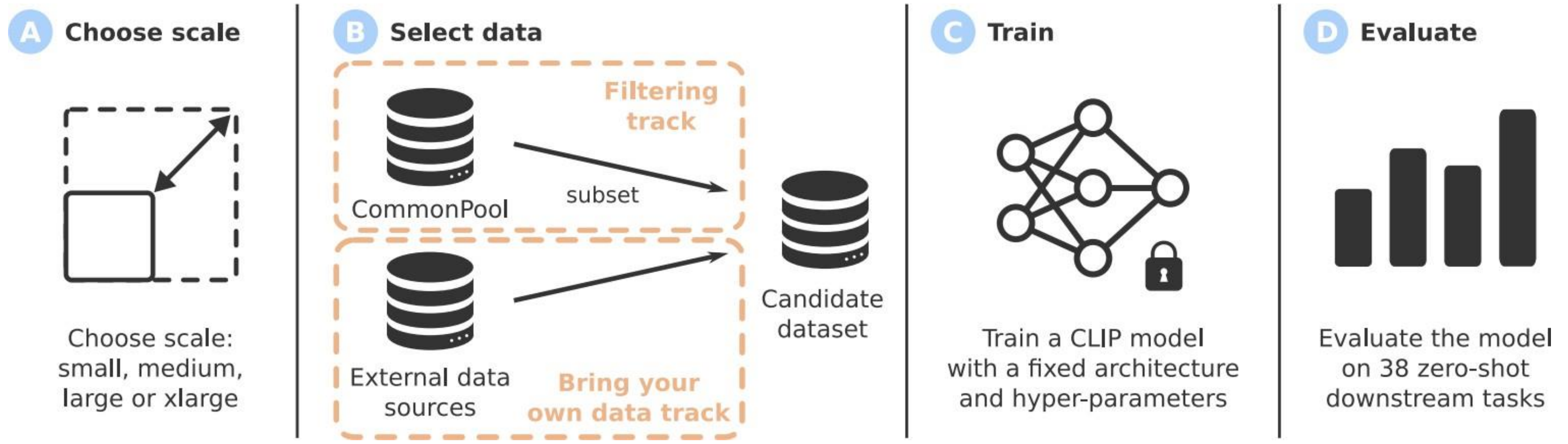


Data scaling up

- Reproducible scaling laws for CLIP training
  - Open large-scale LAION-2B dataset
  - Pre-training OpenCLIP across various scales

	Data	Arch.	ImageNet	VTAB+	COCO
CLIP [55]	WIT-400M	L/14	75.5	55.8	61.1
Ours	LAION-2B	L/14	75.2	54.6	71.1
Ours	LAION-2B	H/14	78.0	56.4	73.4

- DataComp: We know scale matters, how to further scale it up
  - In search of the next-generation image-text datasets
  - Instead of fixing the dataset, and designing different algorithms, the authors propose to fix the CLIP training method, but select the datasets instead



1 Reproducible scaling laws for contrastive language-image learning, CVPR 2023
2 Datacomp: In search of the next generation of multimodal datasets, 2023

Model design: Vision-centric approach

- FLIP: Scaling CLIP training via masking
  - Training: still use CLIP loss, without incorporating the MIM loss
  - Trick: randomly masking out image patches with a high masking ratio, and only encoding the visible patches
  - Results: turns out this does not hurt performance, but improves training efficiency

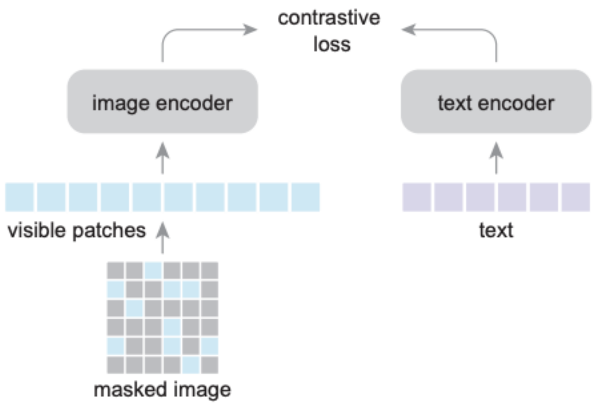
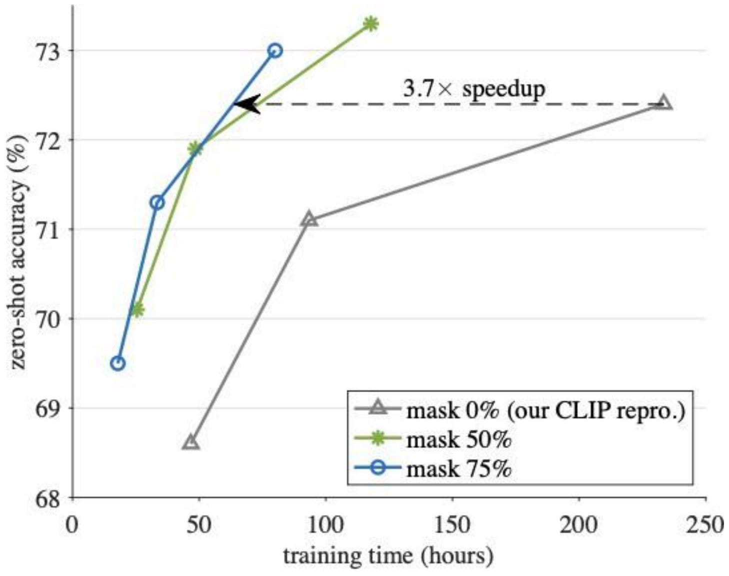


Figure 2. **Our FLIP architecture**. Following CLIP [52], we perform contrastive learning on pairs of image and text samples. We randomly mask out image patches with a high masking ratio and encode only the visible patches. We do not perform reconstruction

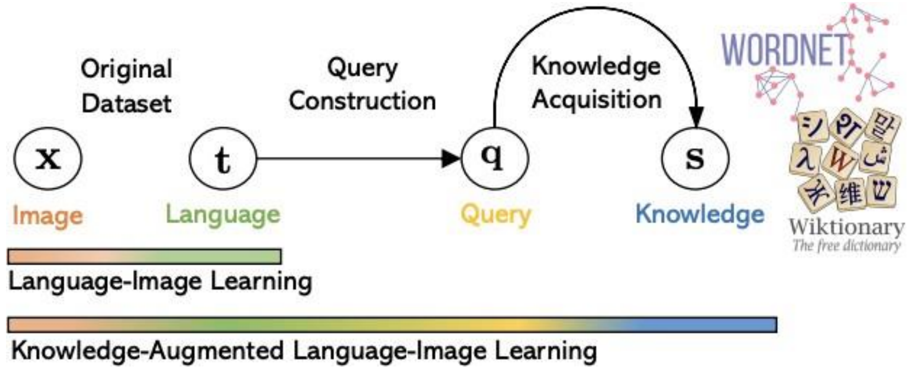


[1] Scaling language-image pre-training via masking, CVPR 2023

Model design: Language-centric approach

- K-Lite: External knowledge
  - The Wiki definition of entities (or, the so-called knowledge) can be naturally used together with the original alt-text for contrastive pre-training

Figure 1: Motivating examples: knowledge explains the content of the rare dish concepts.



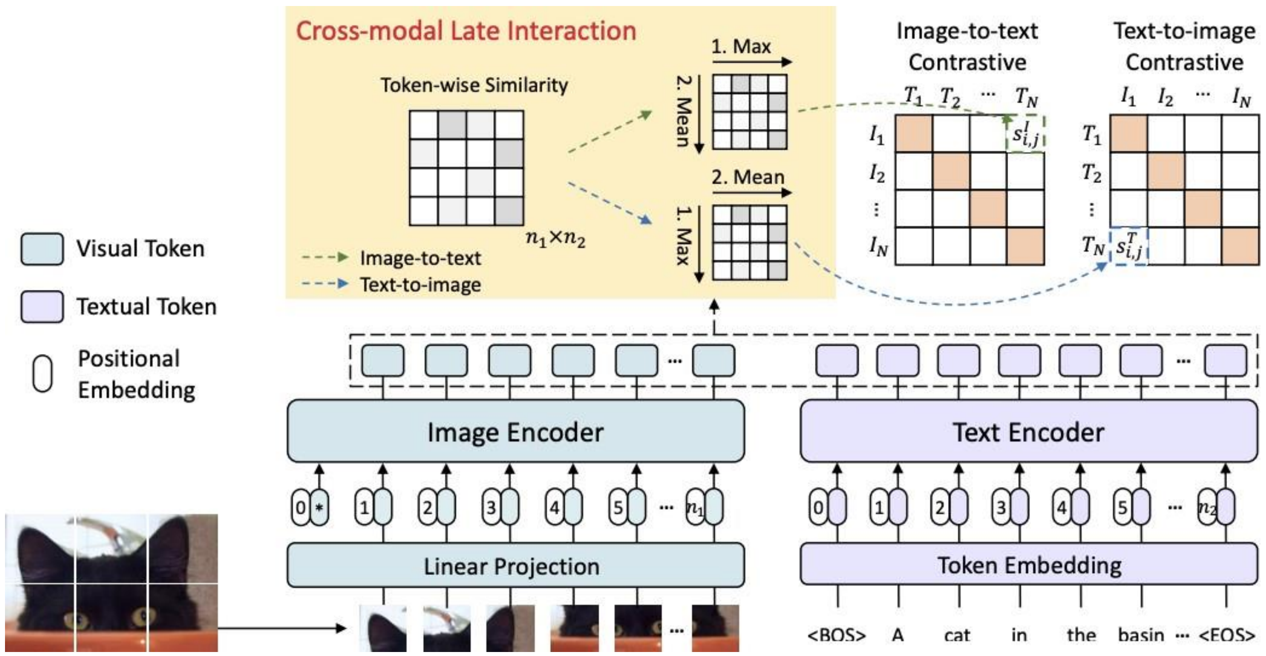
Enriching alt-text with entity descriptions enhances performance.

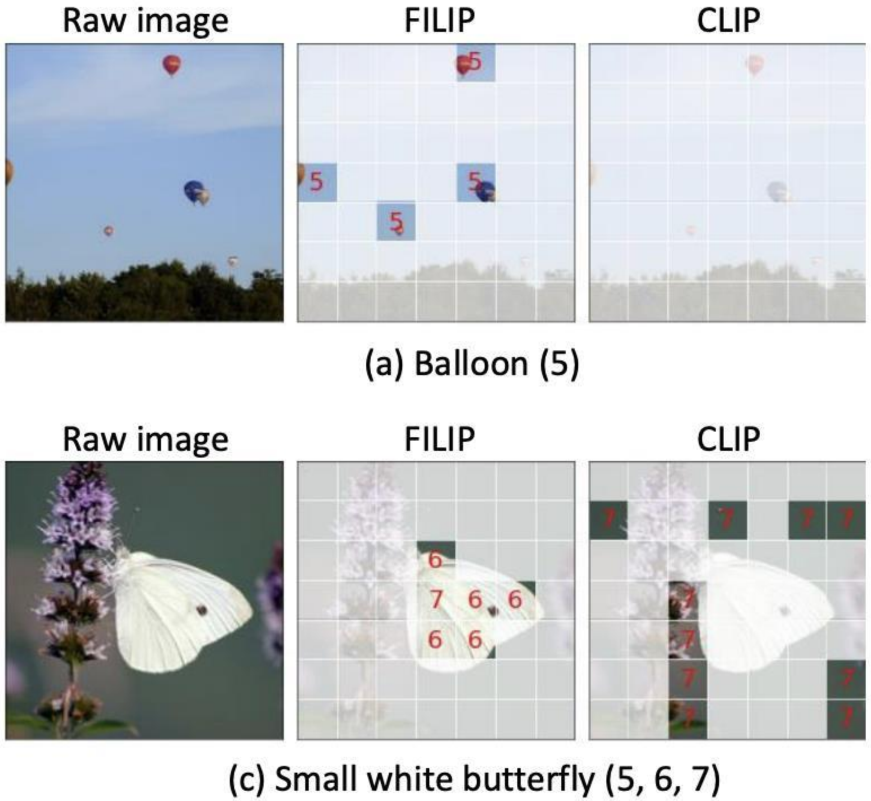
Training Data: Dataset	Training Data: # Samples	Method	ImageNet-1K: Zero-shot	ICinW (20 datasets): Zero-shot	ICinW (20 datasets): Linear Probing	ICinW (20 datasets): Fine-tuning
ImageNet-21K	13M (full)	UniCL	28.16	27.15	53.07 ± 4.15	55.96 ± 2.50
	13M (full)	K-LITE	**30.23**	**33.44**	**53.92** ± 1.05	**57.81** ± 1.48
YFCC-14M + ImageNet-21K	14M (half)	UniCL	34.43	34.30	53.50 ± 2.22	56.45 ± 2.48
	14M (half)	K-LITE	36.67	36.50	49.48 ± 2.23	55.88 ± 1.64
	14M (half)	K-LITE◇	42.36	36.50	54.28 ± 3.66	52.11 ± 4.90
	27M (full)	UniCL	43.06	35.99	55.96 ± 3.38	58.25 ± 2.98
	27M (full)	K-LITE	**45.67**	**38.89**	**57.06** ± 1.48	58.24 ± 2.36
GCC-15M + ImageNet-21K	15M (half)	UniCL	41.64	36.31	53.86 ± 2.73	59.04 ± 3.13
	15M (half)	K-LITE	44.26	39.53	55.91 ± 2.53	58.20 ± 3.39
	15M (half)	K-LITE◇	47.30	40.32	57.38 ± 2.70	60.72 ± 2.29
	28M (full)	UniCL	46.83	38.90	57.92 ± 3.31	60.99 ± 2.74
	28M (full)	K-LITE	**48.76**	**41.34**	**58.56** ± 3.12	**63.39** ± 1.74

[1] K-lite: Learning transferable visual models with external knowledge, NeurIPS 2022

Objective function: Fine-grained modeling

- FILIP: Fine-grained supervision
  - Still dual encoder, not a fusion encoder
  - But compute the loss by first computing the token-wise similarity, and then aggregating the matrix by max pooling
  - Learns word-patch alignment that is good for visualization





[1] FILIP: Fine-grained Interactive Language-Image Pre-Training, ICLR 2022

Can CLIP be combined with other approaches?

Contrastive Language-Image Pre-training + Supervised Learning = ?

Contrastive Language-Image Pre-training + Image-Only (Non-) Contrastive Learning = ?

Contrastive Language-Image Pre-training + Masked Image Modeling = ?

Image-Only (Non-) Contrastive Learning + Masked Image Modeling = ?

Can CLIP be combined with other approaches?

Contrastive Language-Image Pre-training + Supervised Learning = ?

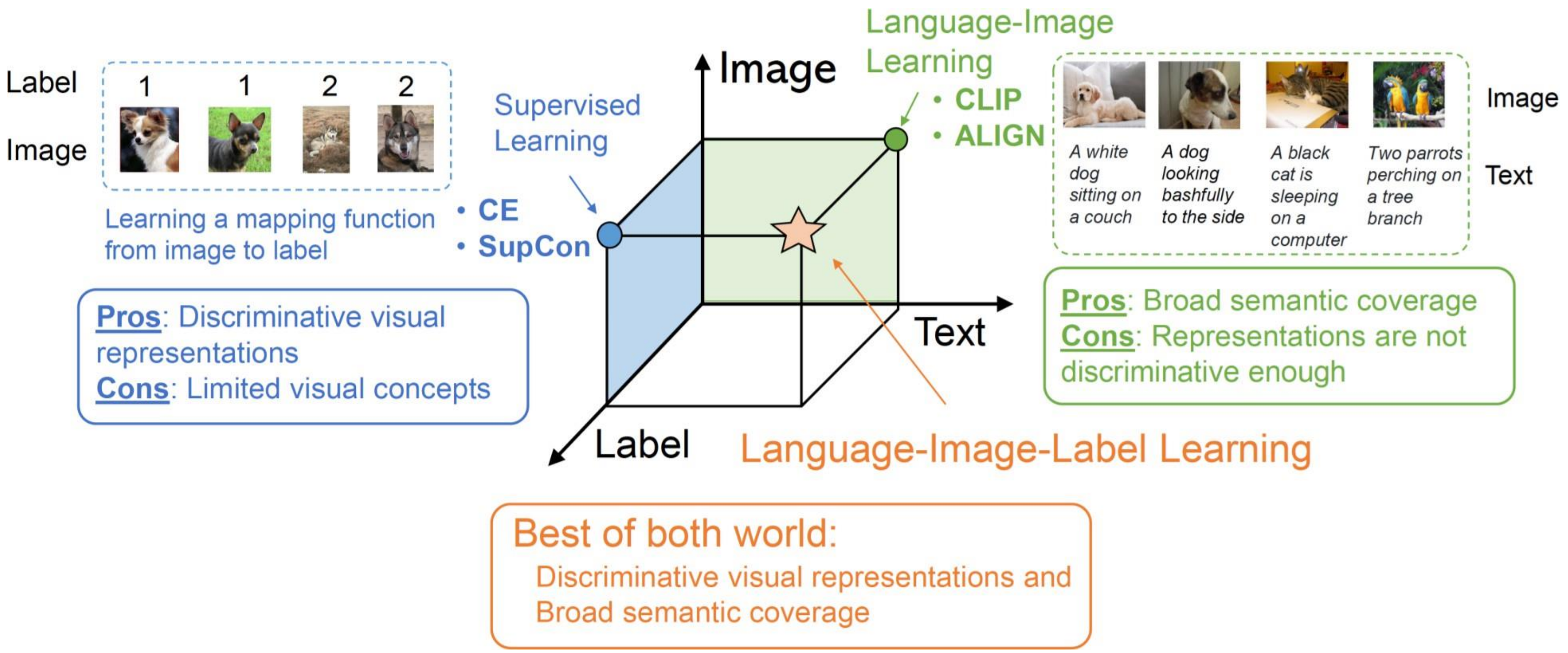
Contrastive Language-Image Pre-training + Image-Only (Non-) Contrastive Learning = ?

Contrastive Language-Image Pre-training + Masked Image Modeling = ?

Image-Only (Non-) Contrastive Learning + Masked Image Modeling = ?

Noisy label + text supervision

- UniCL: Image-text-label space
  - A principled way to use image-label and image-text data together
  - A scaled-up version is the Florence model



1 Unified contrastive learning in image-text-label space, CVPR 2022
2 Florence: A new foundation model for computer vision, 2021

Can CLIP be combined with other approaches?

Contrastive Language-Image Pre-training + Supervised Learning = ? 

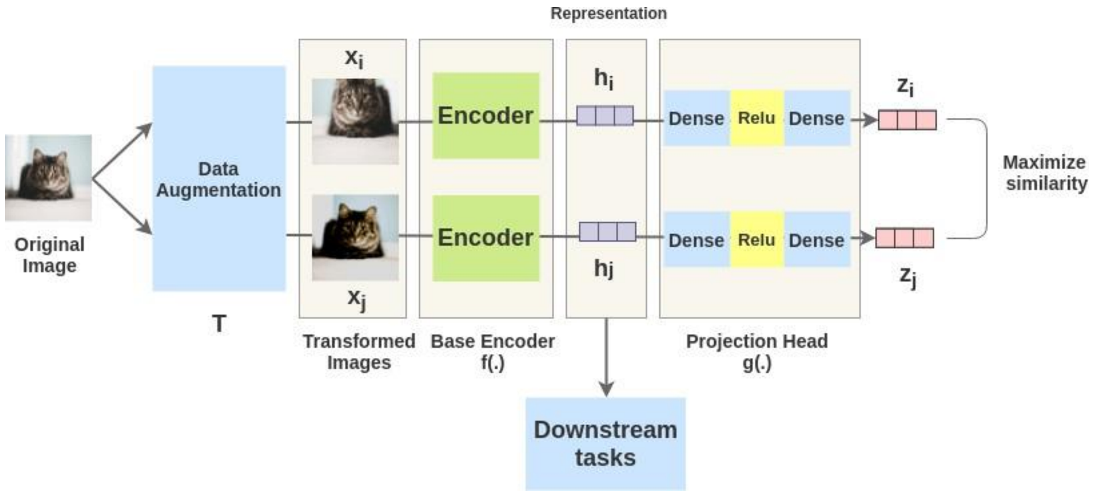
Contrastive Language-Image Pre-training + Image-Only (Non-) Contrastive Learning = ?

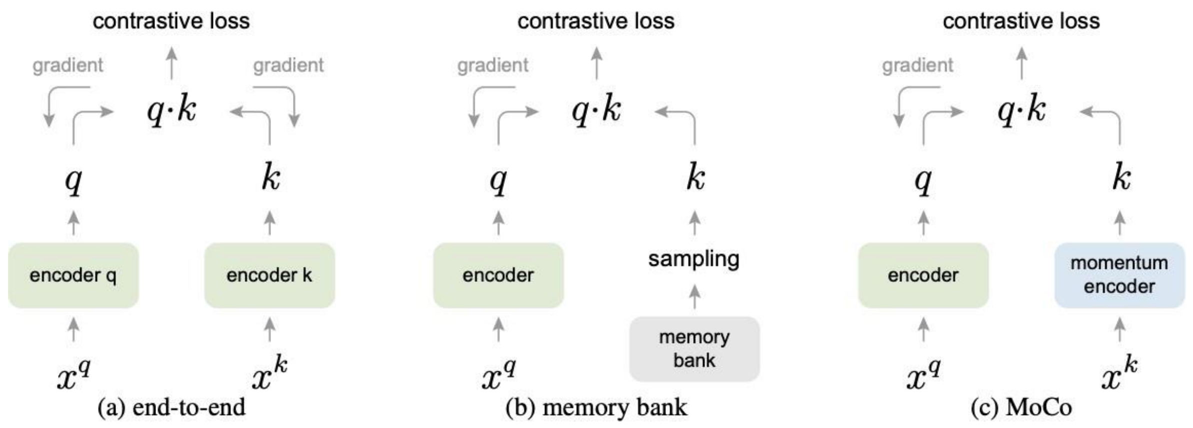
Contrastive Language-Image Pre-training + Masked Image Modeling = ?

Image-Only (Non-) Contrastive Learning + Masked Image Modeling = ?

Image-only (non-)contrastive learning

- SimCLR: A Simple Framework of Contrastive Learning of Visual Representations
  - Given one image, two separate data augmentations are applied
  - A base encoder is followed by a project head, which is trained to maximize agreement using a contrastive loss (i.e., they are from the same image or not)
  - The project head is thrown away for downstream tasks
  - Nicely connected to mutual information maximization
  - A caveat of these line of methods is the requirement of large batch size or memory bank





1 A Simple Framework for Contrastive Learning of Visual Representations, ICML 2020
2 Momentum Contrast for Unsupervised Visual Representation Learning, CVPR 2020

Image-only (non-)contrastive learning

- Recent SSL methods relieve the dependency on negative samples
  - The use of negatives can be replaced by asymmetric architectures (BYOL, SimSiam), dimension de-correlation (Barlow twins), and clustering (SWaV, DINO), etc.

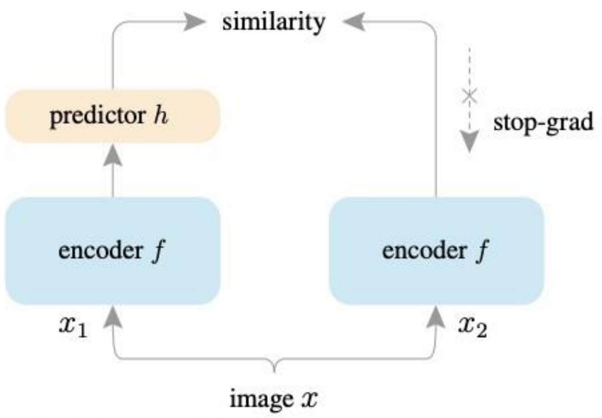


Figure 1. **SimSiam architecture**. Two augmented views of one image are processed by the same encoder network f (a backbone plus a projection MLP). Then a prediction MLP h is applied on one side, and a stop-gradient operation is applied on the other side. The model maximizes the similarity between both sides. It uses neither negative pairs nor a momentum encoder.

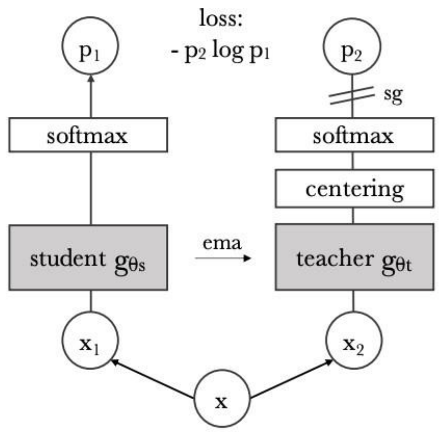
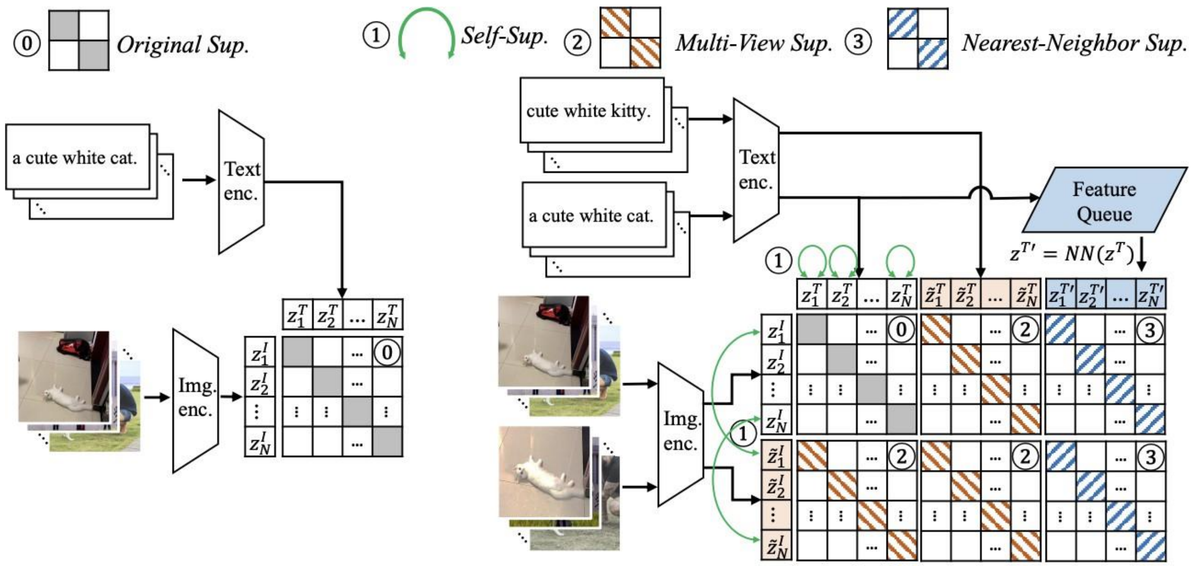


Figure 2: **Self-distillation with no labels.** We illustrate DINO in the case of one single pair of views (x_1, x_2) for simplicity. The model passes two different random transformations of an input image to the student and teacher networks. Both networks have the same architecture but different parameters. The output of the teacher network is centered with a mean computed over the batch. Each networks outputs a K dimensional feature that is normalized with a temperature softmax over the feature dimension. Their similarity is then measured with a cross-entropy loss. We apply a stop-gradient (sg) operator on the teacher to propagate gradients only through the student. The teacher parameters are updated with an exponential moving average (ema) of the student parameters.

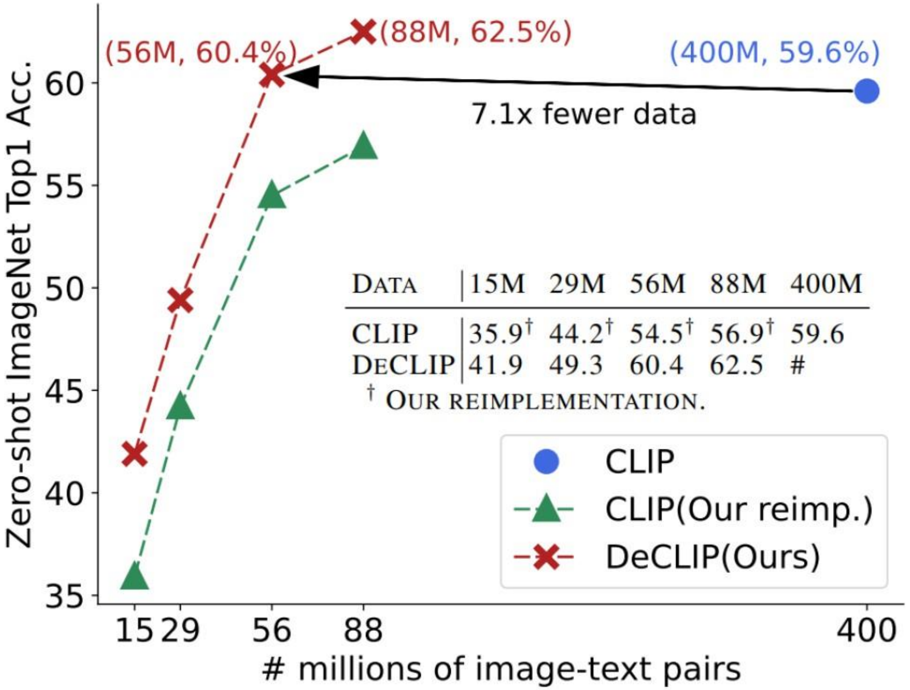
1 Bootstrap your own latent-a new approach to self-supervised learning, NeurIPS 2020
2 Exploring simple siamese representation learning, CVPR 2021
3 Variance-invariance-covariance regularization for self-supervised learning, ICLR 2022
4 Barlow twins: Self-supervised learning via ´ redundancy reduction, ICML 2021
5 Unsupervised learning of visual features by contrasting cluster assignments, NeurIPS 2020
6 Emerging properties in self-supervised vision transformers, ICCV 2021

How to combine CLIP with image-only SSL?

- DeCLIP: supervision exists everywhere
  - Self-supervised learning on each modality: Image (SimSam), Text (MLM)
  - Multi-view supervision and Nearest-neighbor supervision



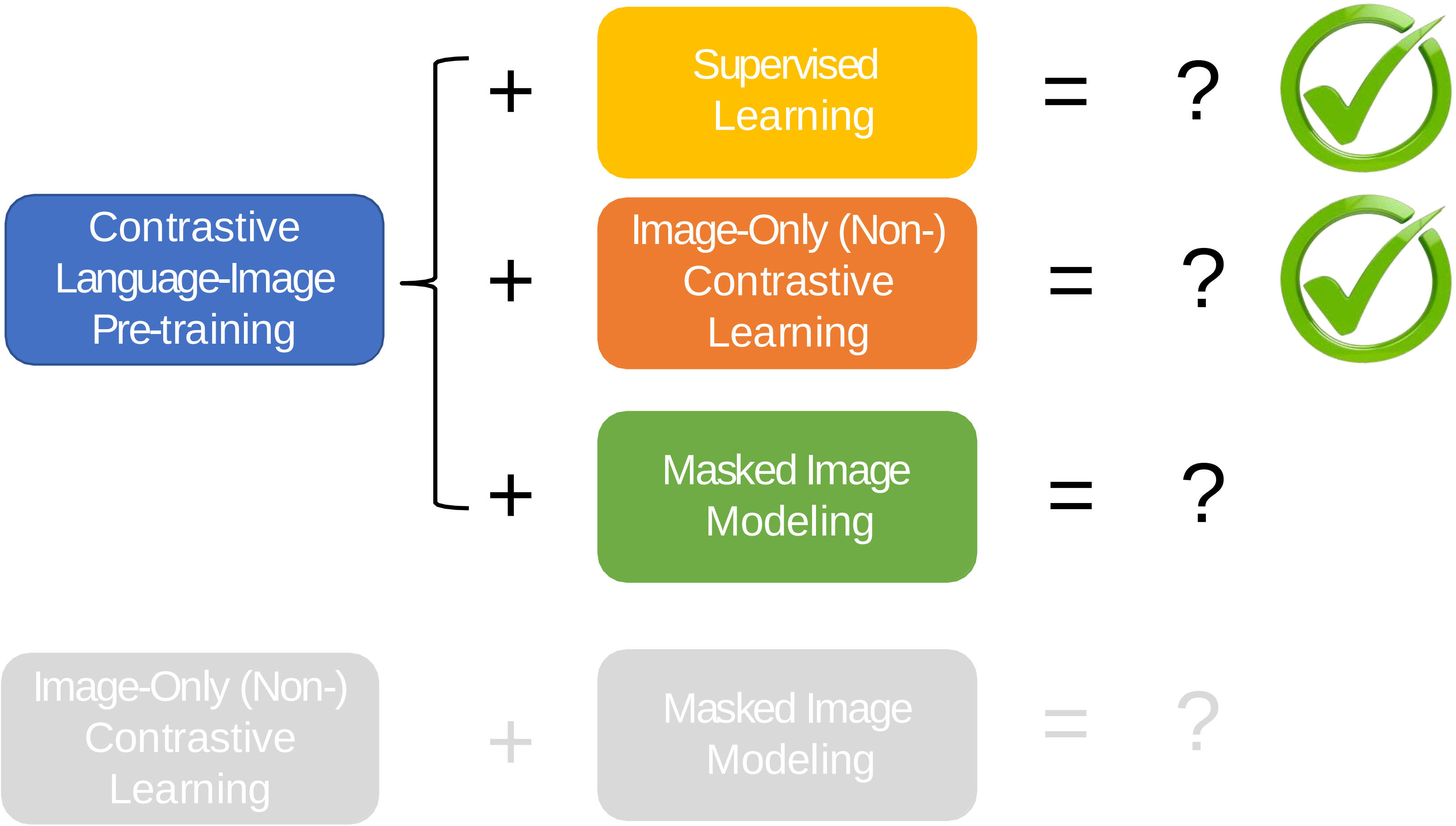
(a) CLIP & ALIGN (b) DeCLIP



Combining vision-language and self-supervised learning improves data efficiency significantly

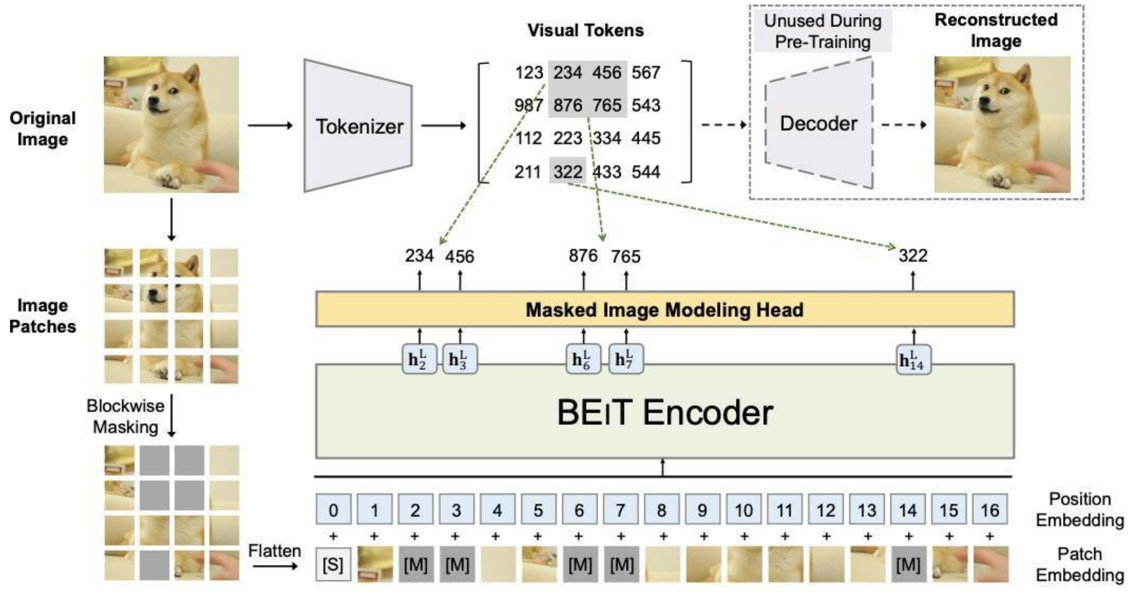
[1] Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm, ICLR 2022

Can CLIP be combined with other approaches?

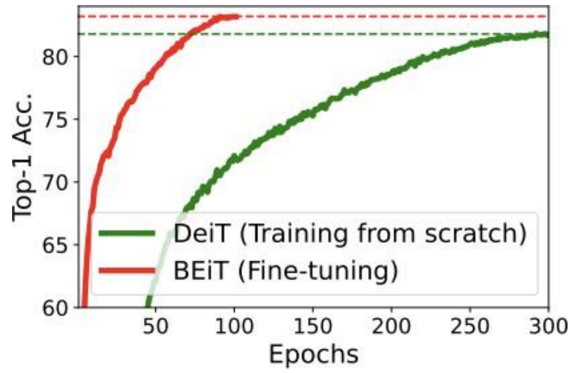


Masked image modeling

- BEiT: BERT Pre-Training of Image Transformers
  - Before pre-training, learn an "image tokenizer" via VQ-VAE/GAN, where an image is tokenized into discrete visual tokens
    - Similar approaches have been used for image generation, such as DALLE, Parti.
  - Randomly masking image patches, pre-train the model to predict masked visual tokens
  - Can be understood as knowledge distillation between the image tokenizer and the BEiT encoder, but the latter only sees partial of the image

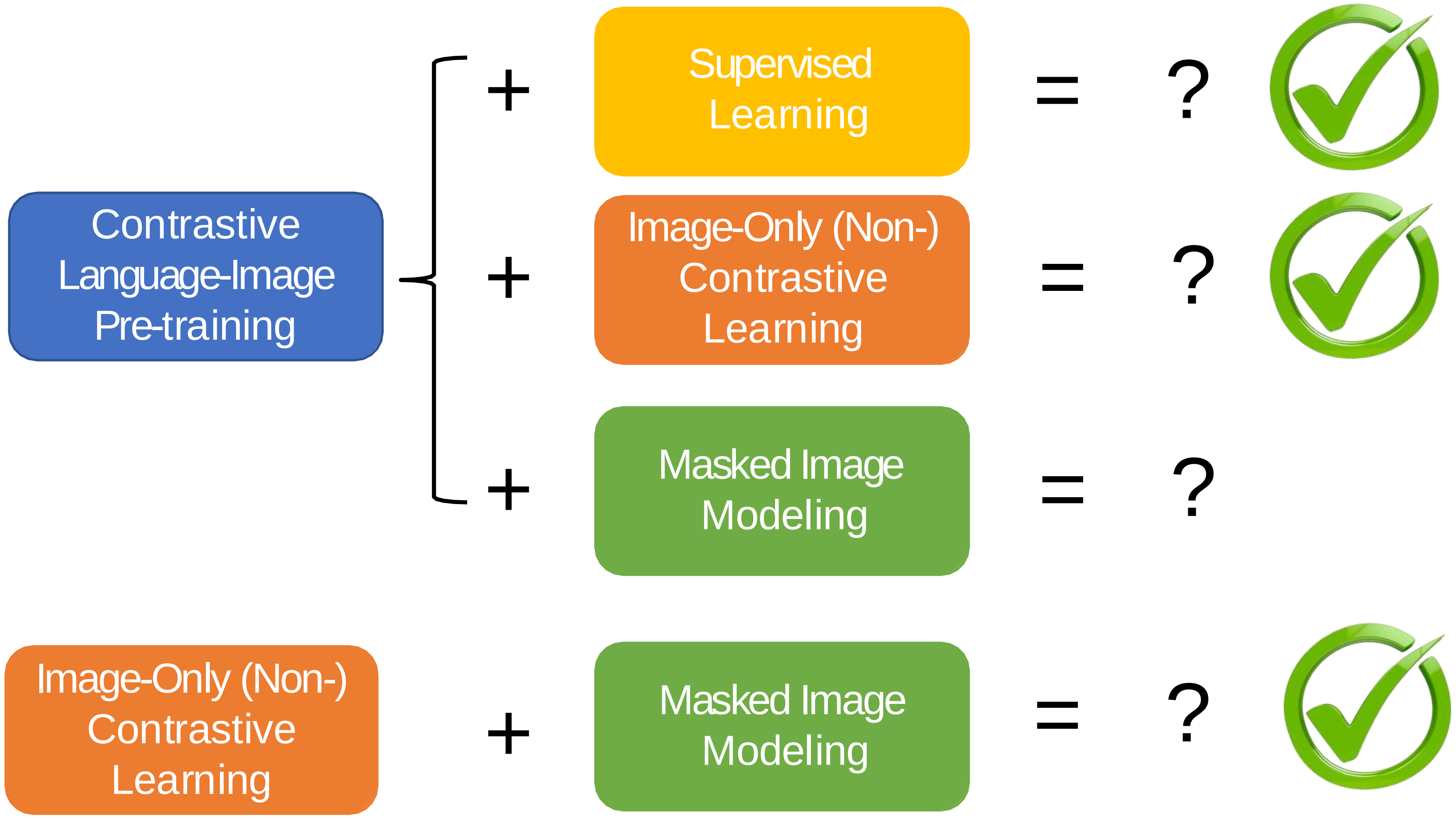


Strong model finetuning performance



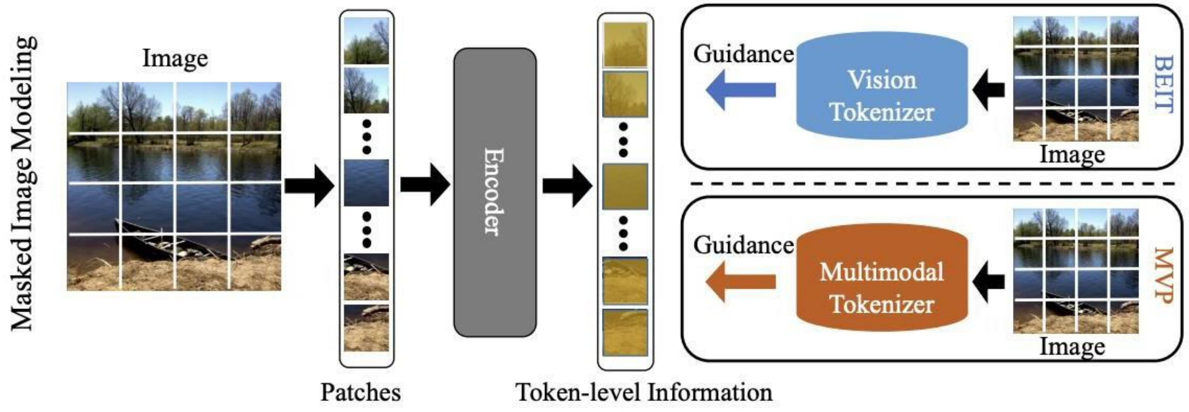
1 BEiT: BERT Pre-Training of Image Transformers, ICLR 2022
2 iBOT: Image BERT Pre-Training with Online Tokenizer, ICLR 2022

Can CLIP be combined with other approaches?

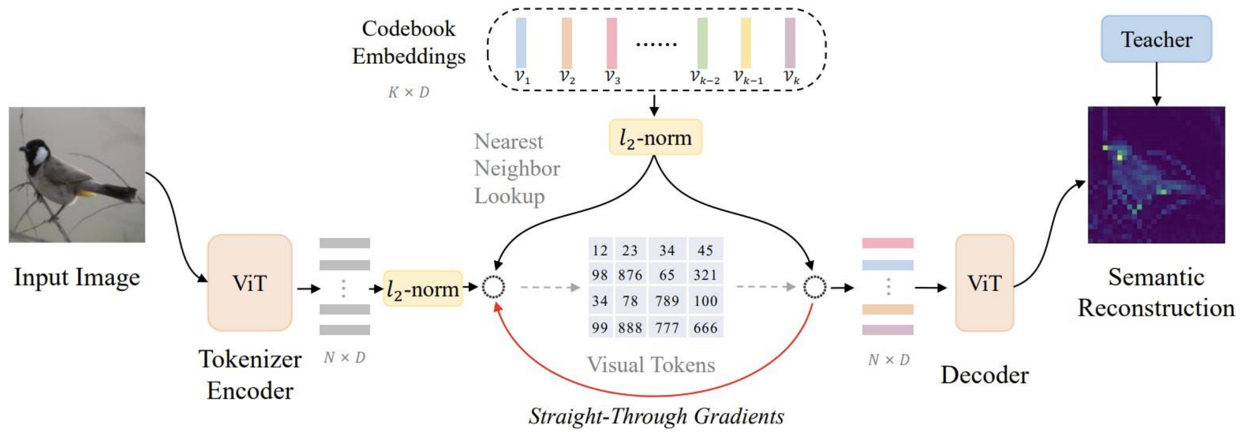


Shallow interaction of CLIP and MIM

- Turns out image features extracted from CLIP are a good target for MIM training
  - Captures the semantics that is missing in MIM training



Approach 1 (MVP): regress CLIP features

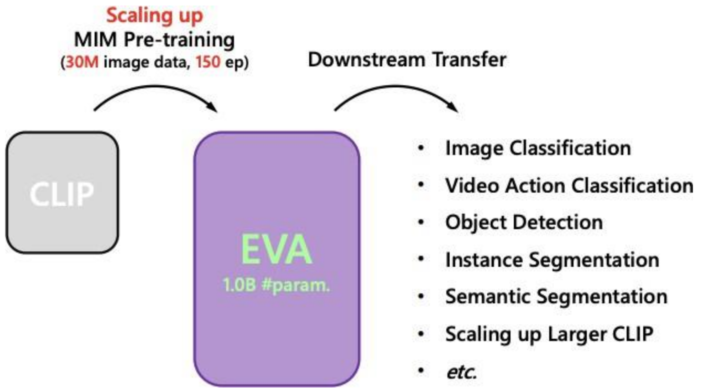


Approach 2 (BEiT v2): compress the information inside CLIP features into the visual tokens, then perform regular BEiT training

1 MVP: Multimodality-guided Visual Pre-training, ECCV 2022
2 BEiT v2: Masked Image Modeling with Vector-Quantized Visual Tokenizers, 2022

Shallow interaction of CLIP and MIM

- This approach is further popularized by the EVA series of work



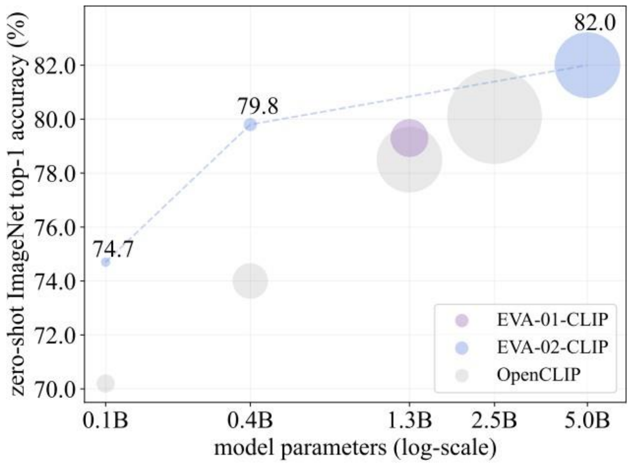


Figure 1: **Summary of CLIP models' ImageNet-1K zero-shot classification performance.** The diameter of each circle corresponds to forward GFLOPs x the number of training samples.

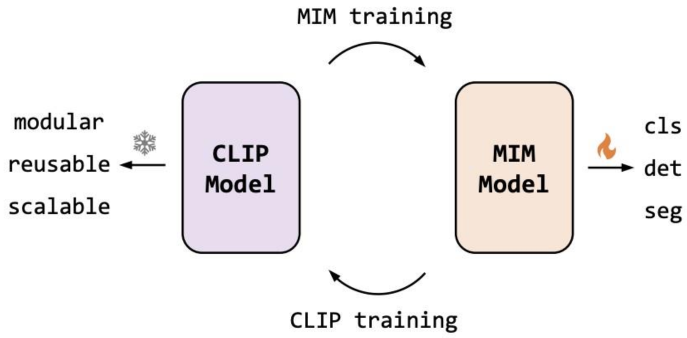
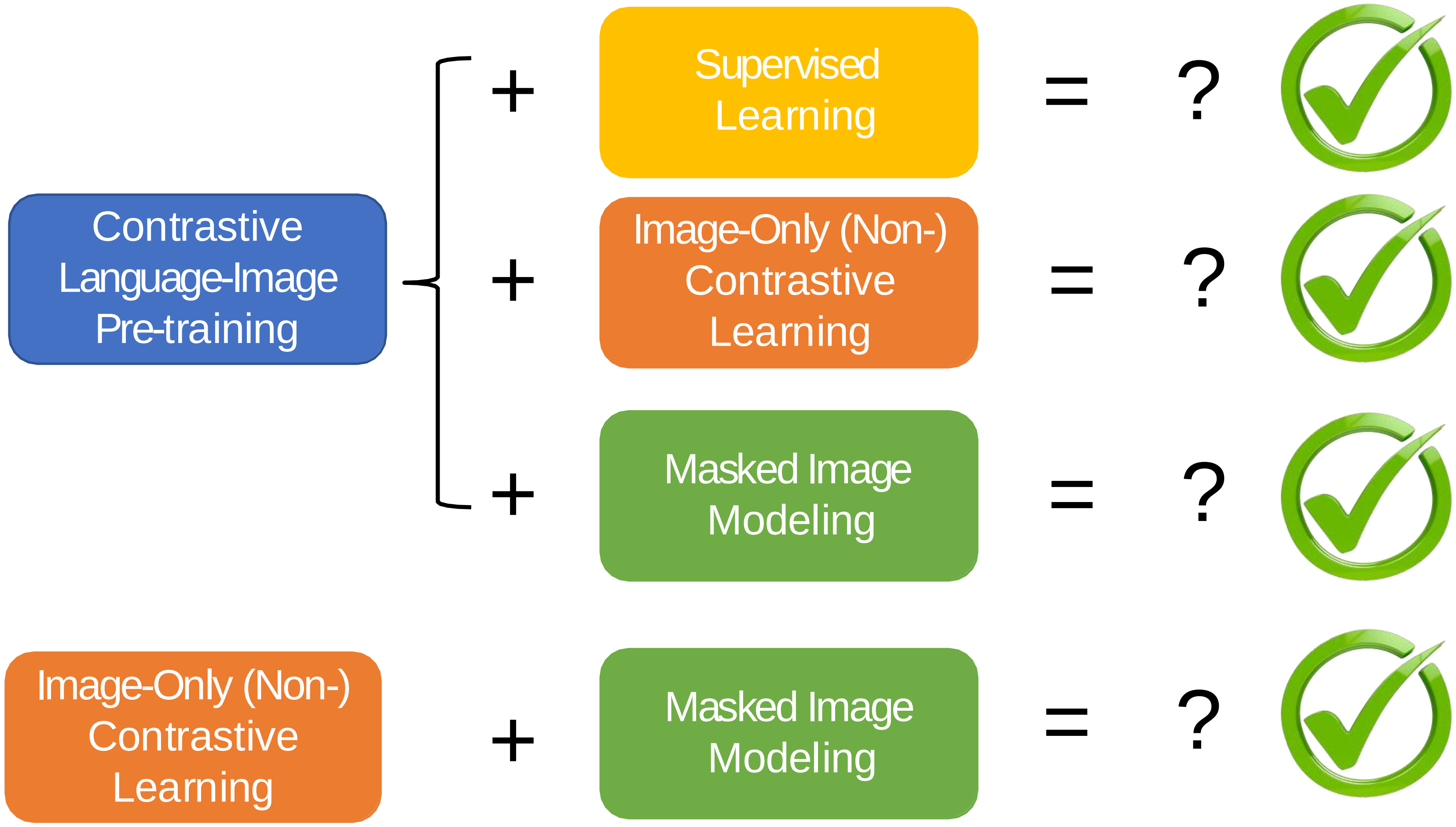


Figure 3: **Alternate learning of MIM and CLIP representations.** Starting with a off-the-shelf CLIP(*e.g.*, OpenAI CLIP [95]), alternate training of the pure MIM visual representations as well as vision-language CLIP representations can improve both MIM and CLIP performances in a bootstrapped manner. The MIM representations can be used to fine-tune various downstream tasks while the (frozen) CLIP representations enable modular, reusable and scalable next-gen model design.

1 EVA: Exploring the Limits of Masked Visual Representation Learning at Scale, CVPR 2023
2 EVA-CLIP: Improved Training Techniques for CLIP at Scale, 2023
3 EVA-02: A Visual Representation for Neon Genesis, 2023.

Can CLIP be combined with other approaches?



( , a dog is running through the grass)

Image Generation

Produce visual data

LLMs and models for image understanding and generation

Part 3: *How to make an LLM that can see and chat?*

Image Encoder

Consume visual data

Part 1: *How to learn image representations?*

Part 2: *How to extend vision models with more flexible, promptable interfaces?*

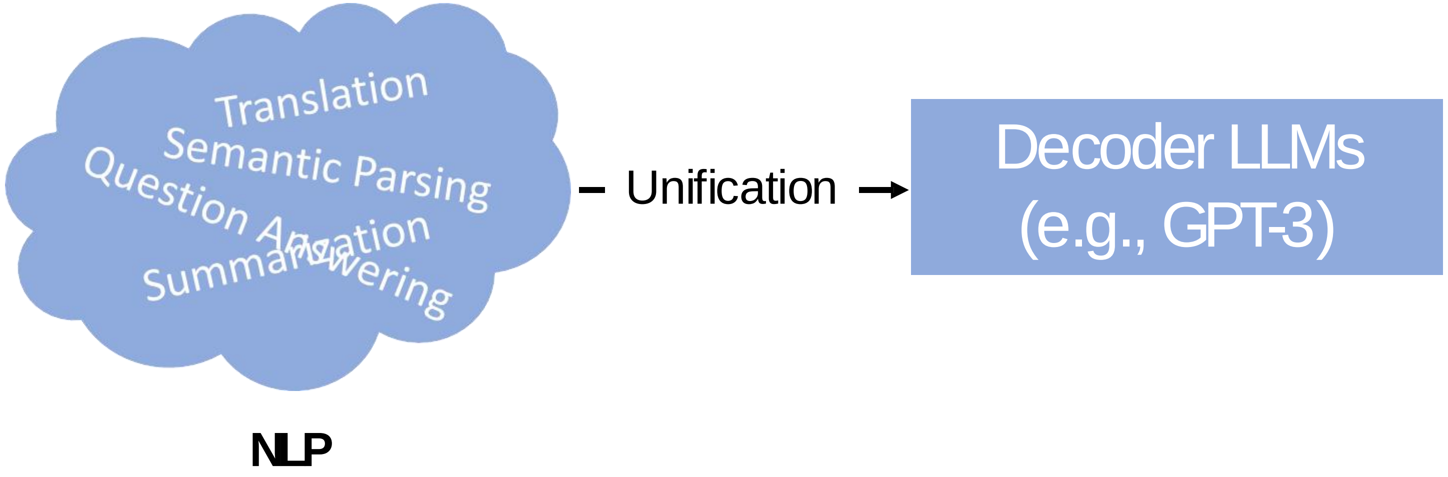
Part 2: Towards Generic Vision Interface

How to design vision interface that is interactive and promptable?

Lessons from LLMs

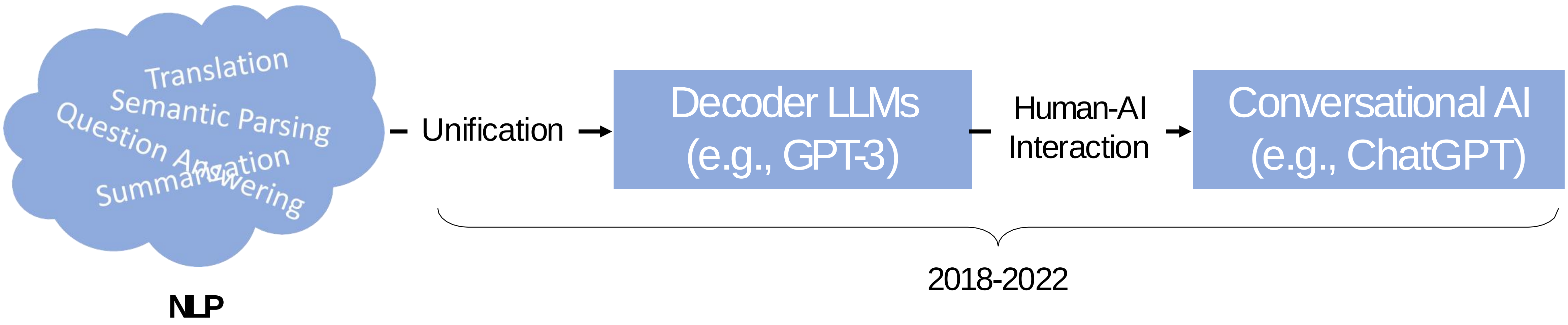
NLP

Lessons from LLMs



Lessons from LLMs

Lessons from LLMs

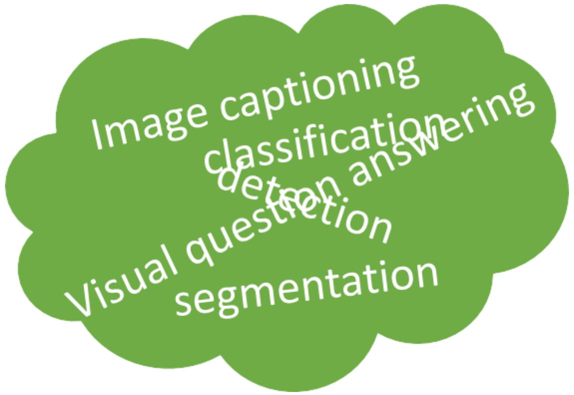


Lessons from LLMs

NLP

– Unification → **Decoder LLMs (e.g., GPT-3)** – Human-AI Interaction → **Conversational AI (e.g., ChatGPT)**

$\underbrace{\hspace{10em}}_{\text{2018-2022}}$



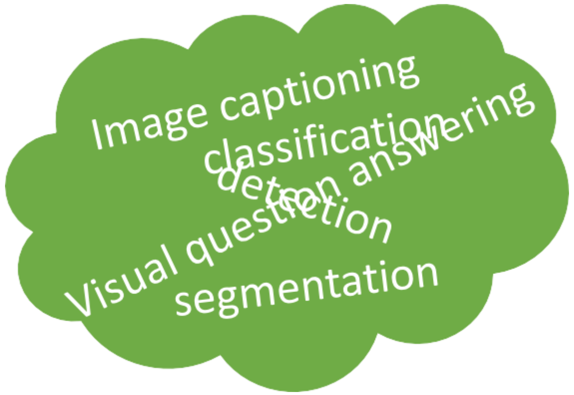
Vision

Lessons from LLMs

NLP

– Unification → Decoder LLMs (e.g., GPT-3) – Human-AI Interaction → Conversational AI (e.g., ChatGPT)

$\underbrace{\hspace{10em}}_{\text{2018-2022}}$



Vision

– Unification → ? – Human-AI Interaction → ?

Unique Challenges in Vision: Modeling

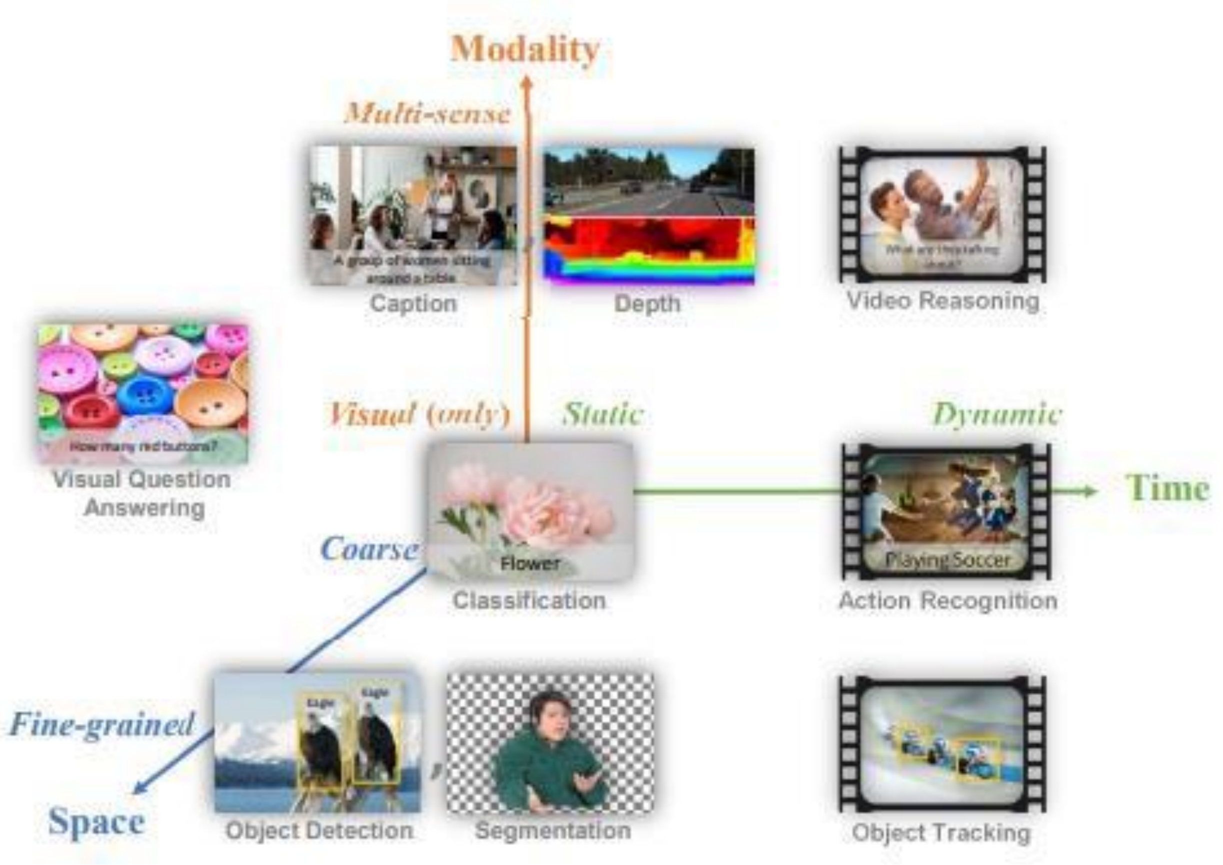


Image Source: Project Florence

Unique Challenges in Vision: Modeling

a) Different types of inputs:

Temporality: static image, video sequence

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

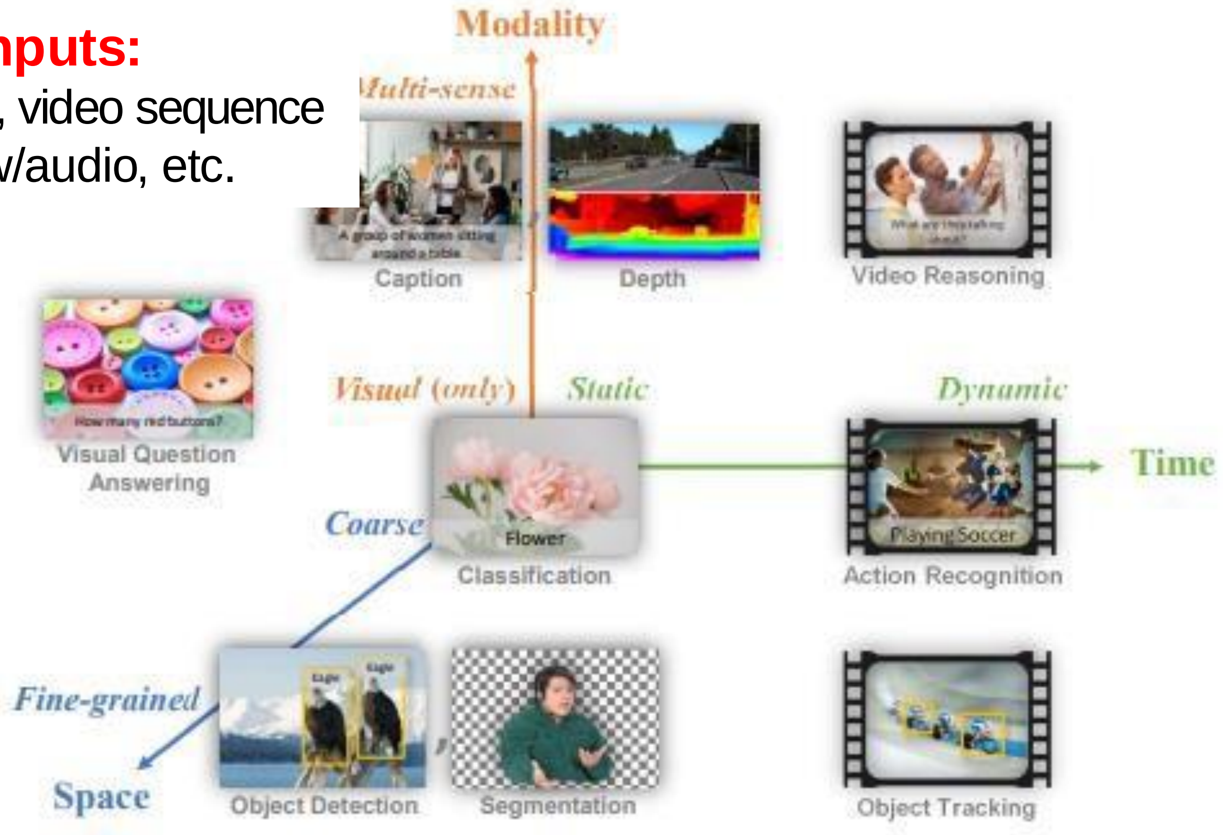
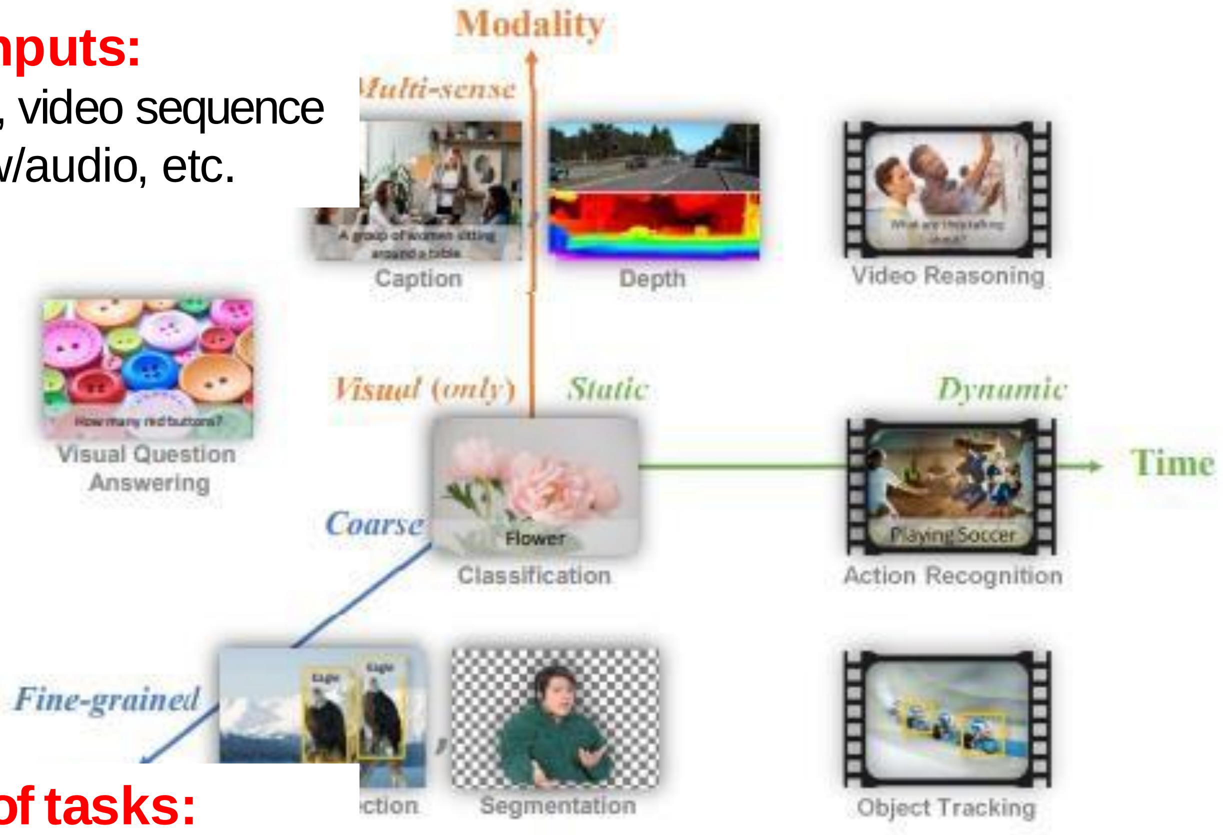


Image Source: Project Florence

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:

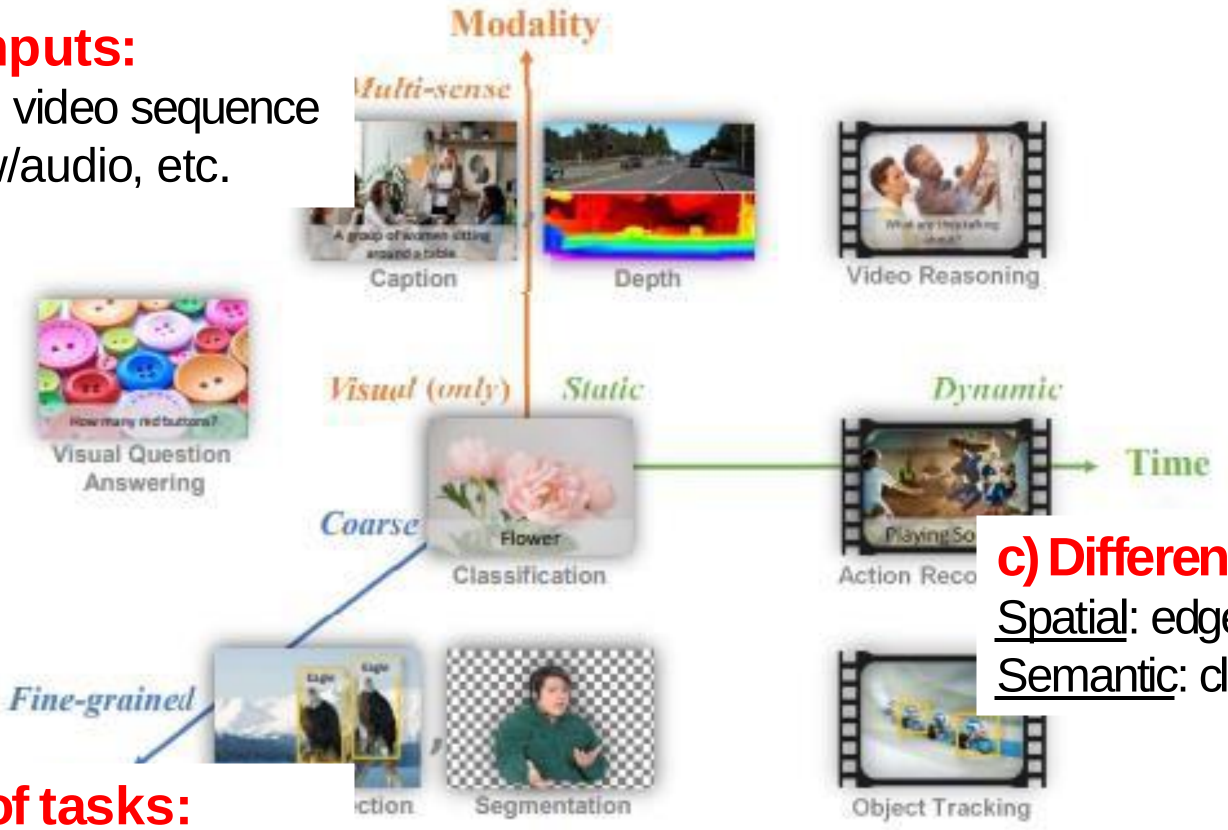
Image-level: classification, captioning, etc.
Region-level: object detection, grounding, etc.
Pixel-level: segmentation, depth, SR, etc.

Image Source: Project Florence

Unique Challenges in Vision: Modeling

a) Different types of inputs:

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.

c) Different types of outputs:

Spatial: edges, boxes, masks, etc.
Semantic: class labels, descriptions, etc.

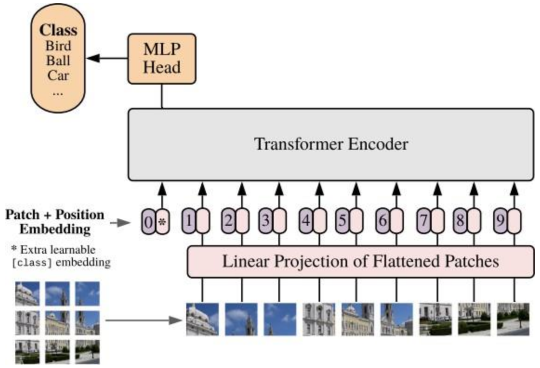
Image Source: Project Florence

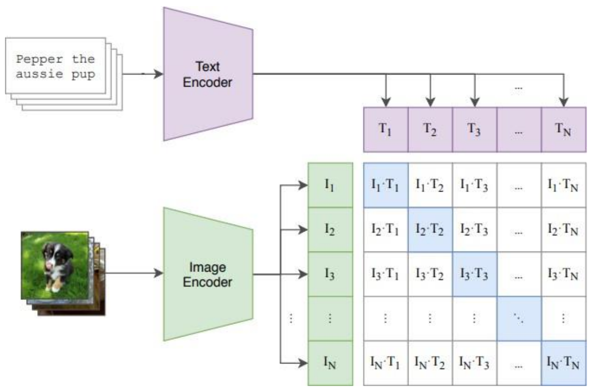
Attempts towards General Vision

Closed-set Classification → Open-world Recognition

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

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





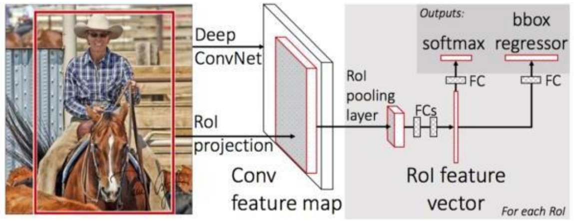
1 Krizhevsky et al. "Imagenet classification with deep convolutional neural networks.“. *NeurIPS* 2012
2 He et al. "Deep residual learning for image recognition." *CVPR* 2016.
3 Dosovitskiy et al. "An image is worth 16x16 words: Transformers for image recognition at scale.“ *ICLR 2021*.
4 Radford et 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*.

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.

Attempts towards General Vision

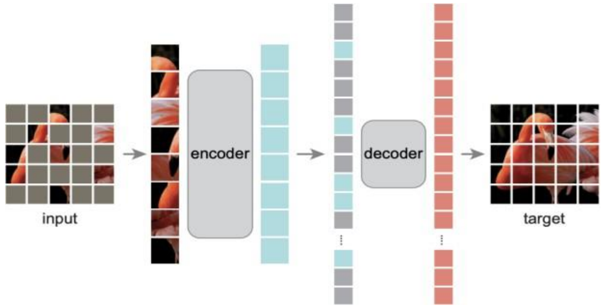
Closed-set Classification → Open-world Recognition

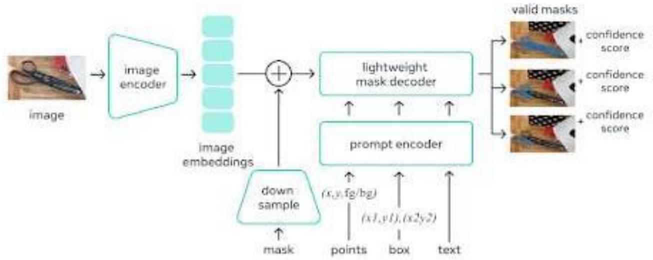
Specialist Models → Generalist Models

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.

Attempts towards General Vision

Closed-set Classification → Open-world Recognition

Specialist Models → Generalist Models

Representation Learning → Promptable Interface

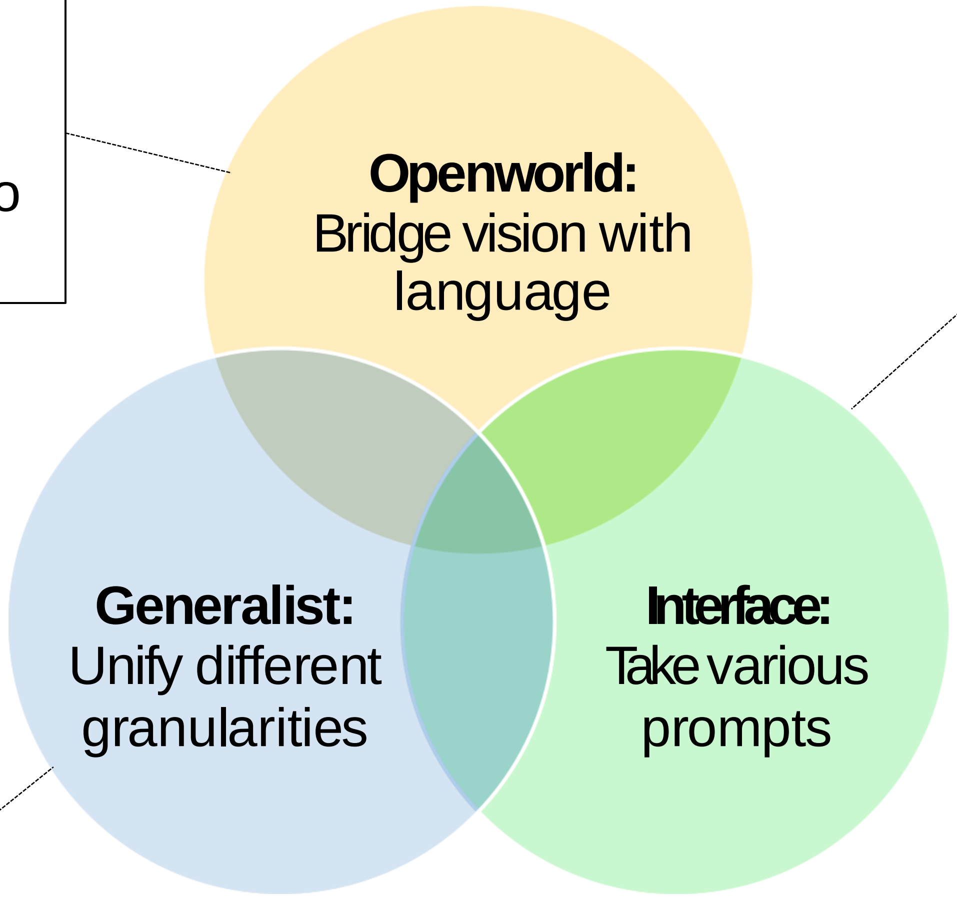
Attempts towards General Vision

Open-world Recognition

Generalist Models

Promptable Interface

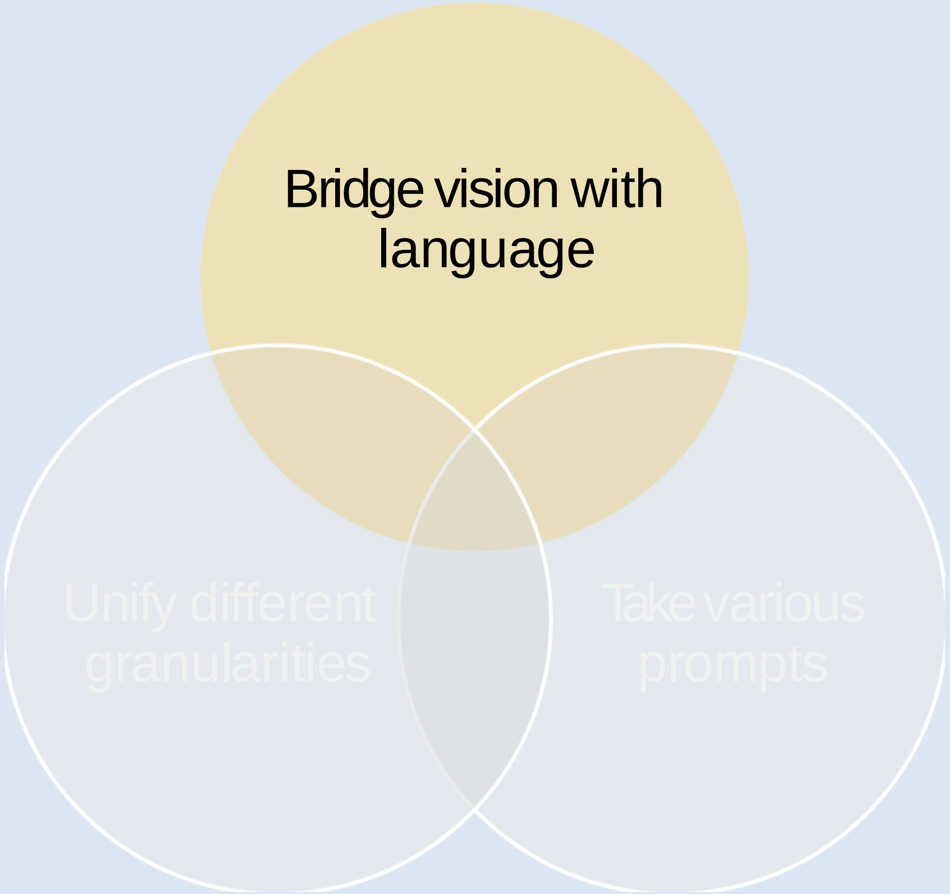
Intuition: language as the common space to share information
Benefit: Zero-shot transfer to novel vocabularies



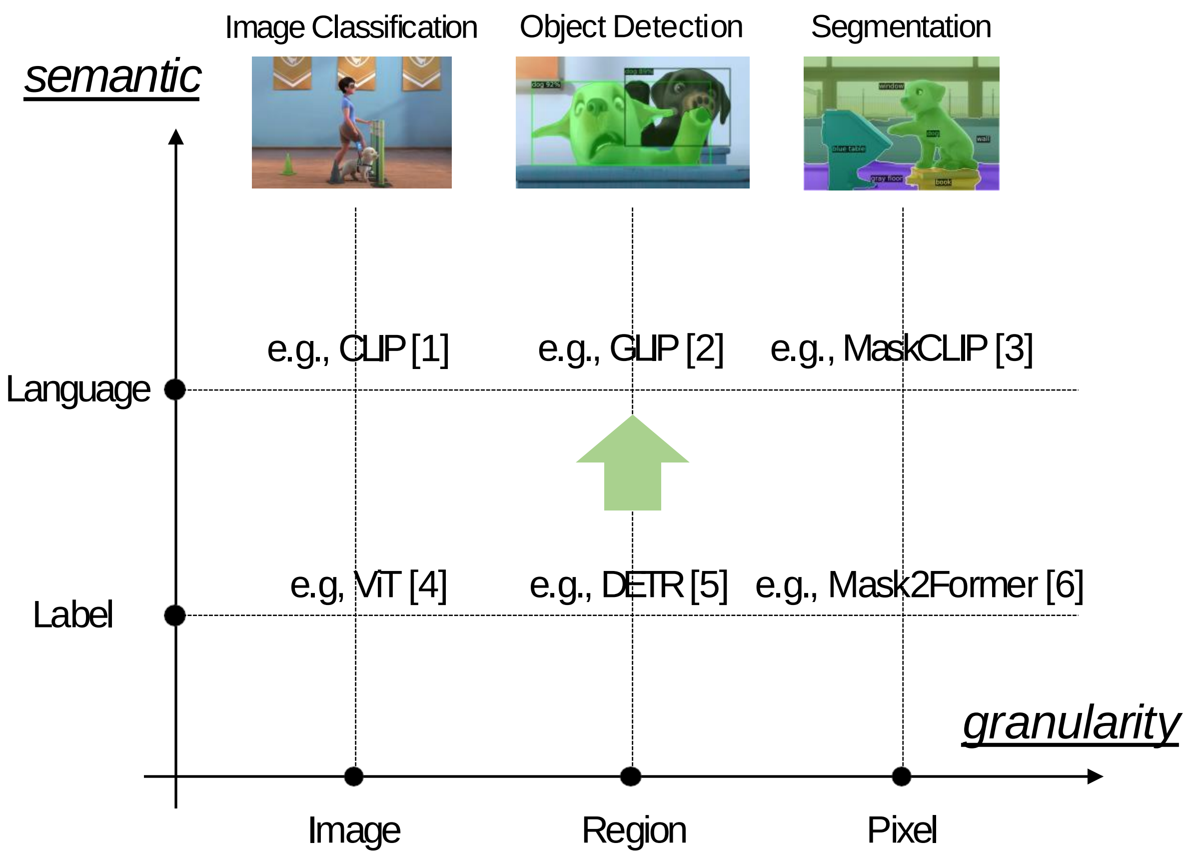
Intuition: language, spatial prompts and beyond
Benefit: Reduce the ambiguity of expressing human intents

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

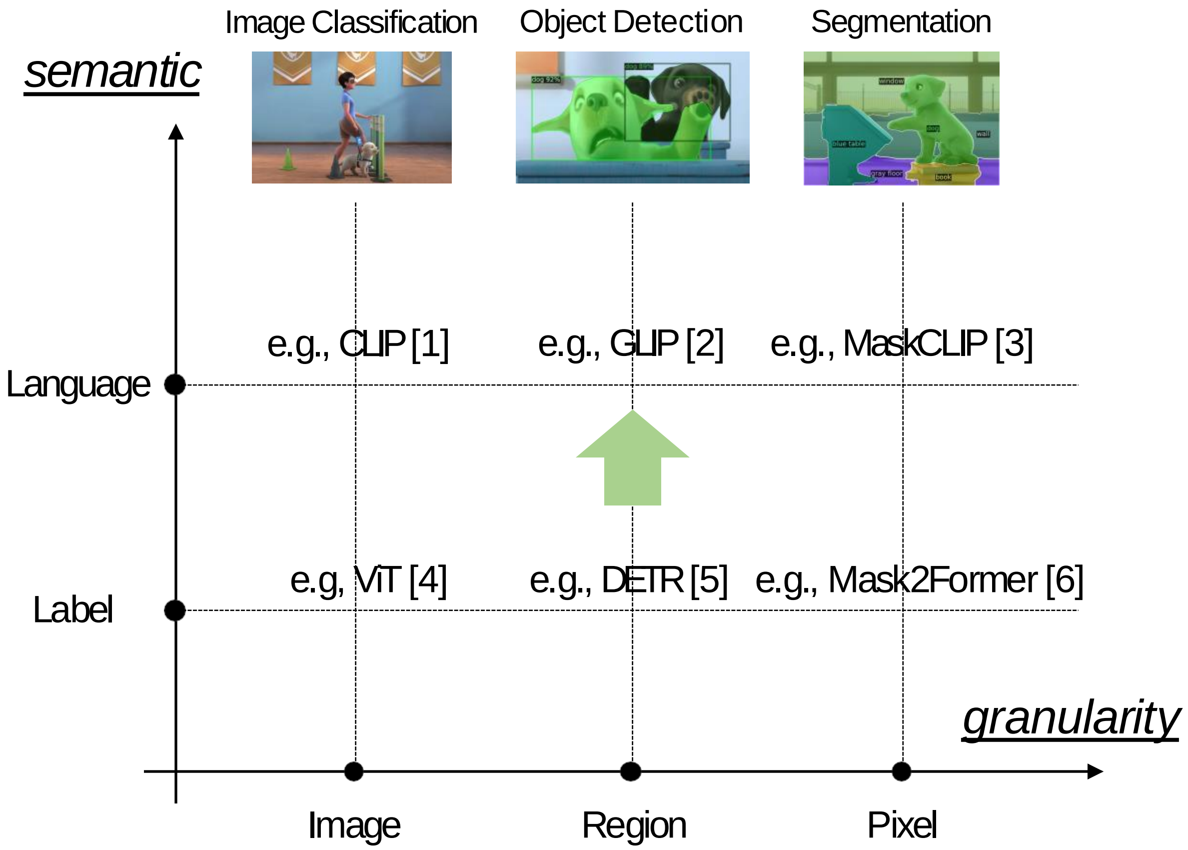
3 Zhou et al. "Extract Free Dense Labels from CLIP." ECCV, 2022

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

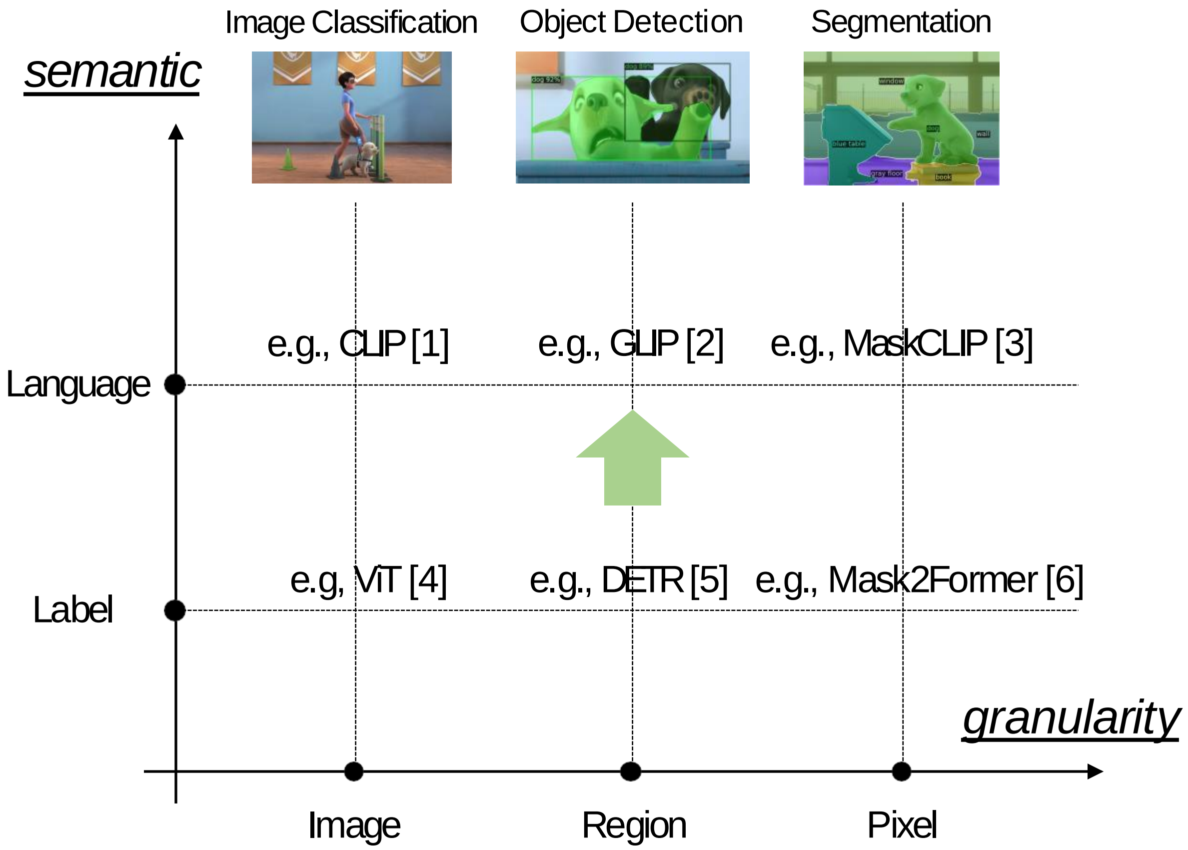


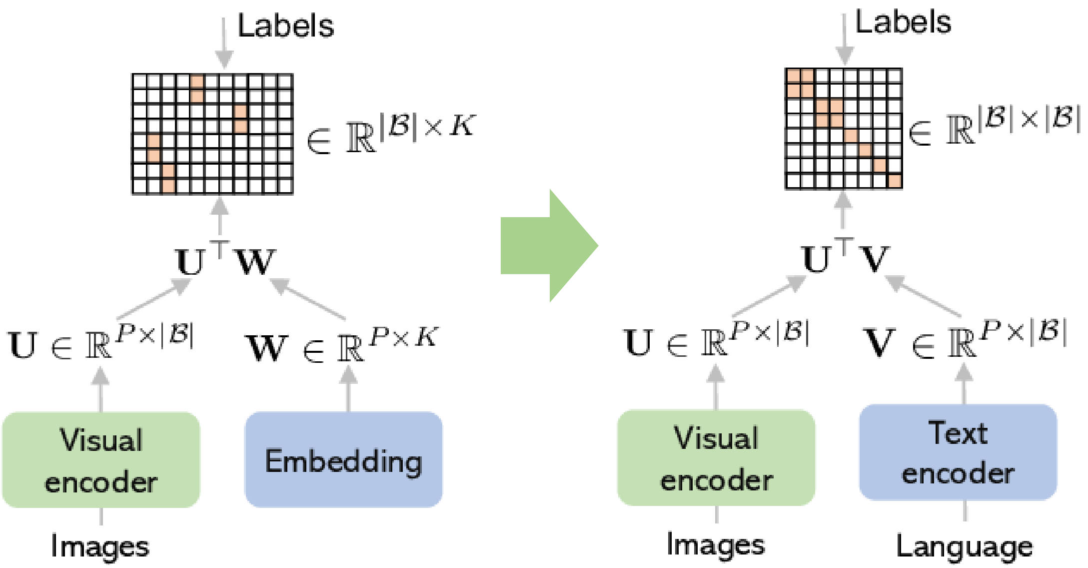
(a) Converting labels to language is agnostic to granularity

(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*

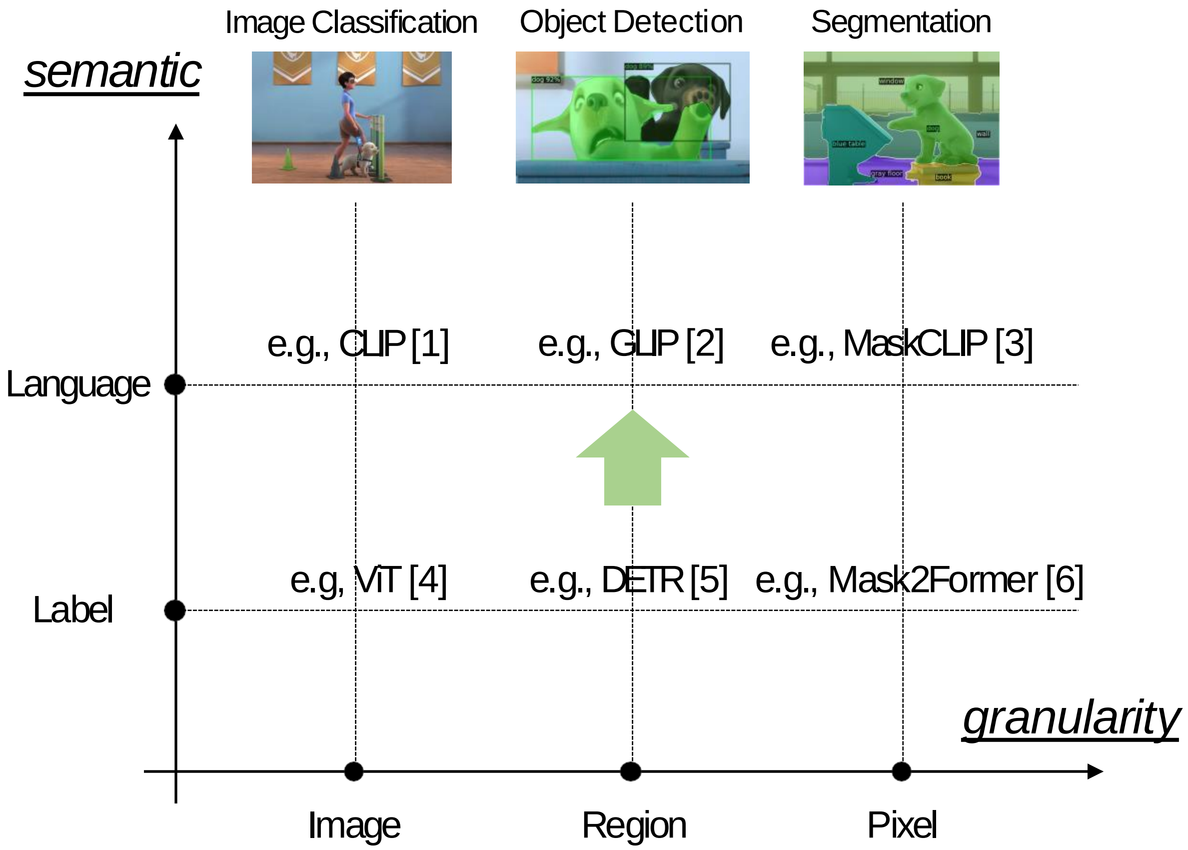
Bridge Vision with Language

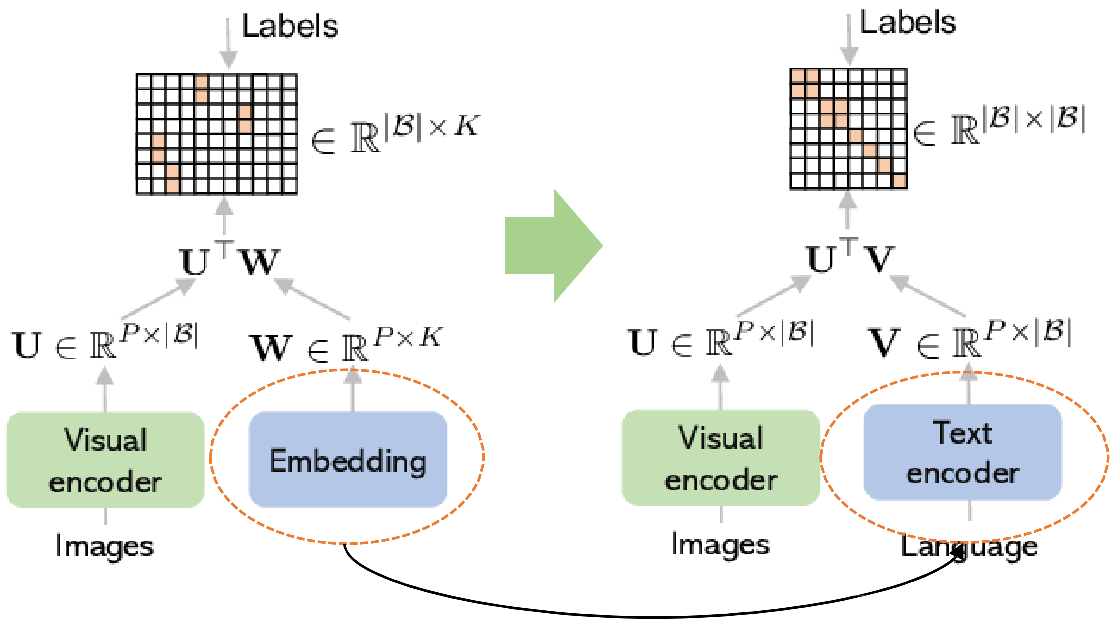




1 Radford et al. "Learning transferable visual models from natural language supervision." ICML, PMLR, 2021
2 Li et al. "Grounded language-image pre-training." CVPR, 2022
3 Zhou et al. "Extract Free Dense Labels from CLIP." ECCV, 2022
4 Dosovitskiy et al. "An image is worth 16x16 words: Transformers for 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



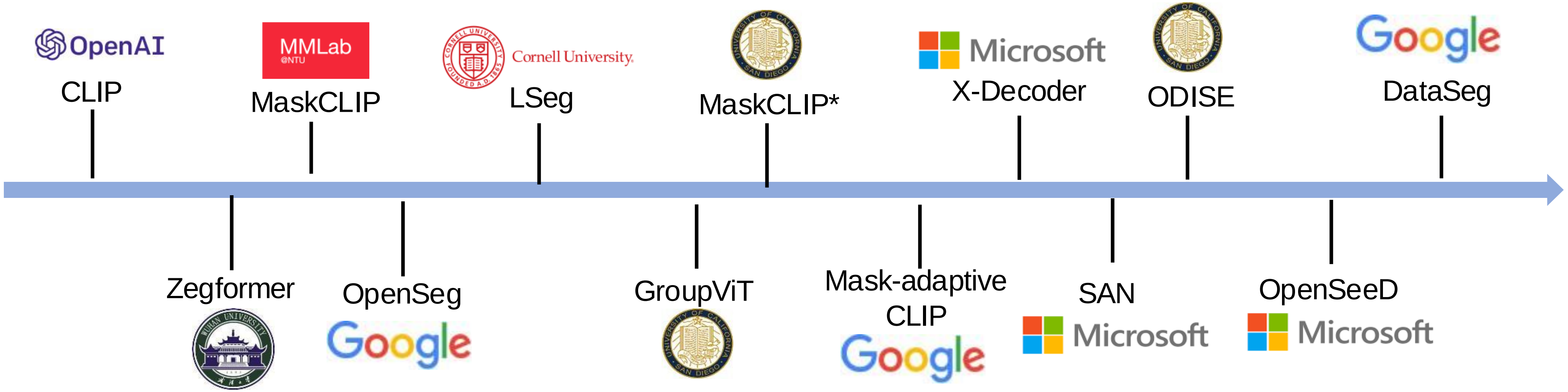


Replace labels with concept names, and use text encoder to encode all concepts as they are language tokens

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*

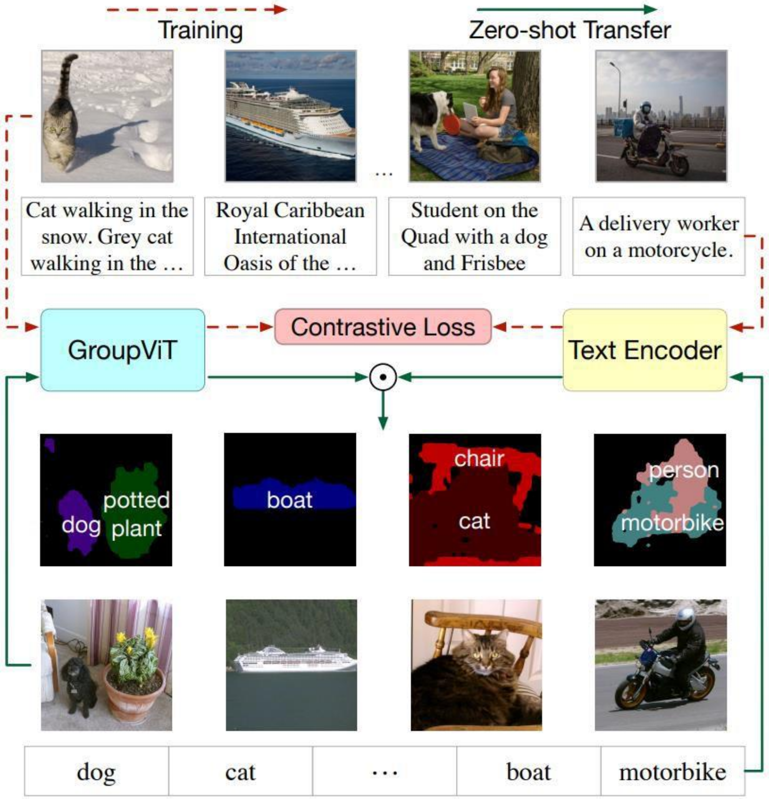
Bridge Vision with Language for Segmentation

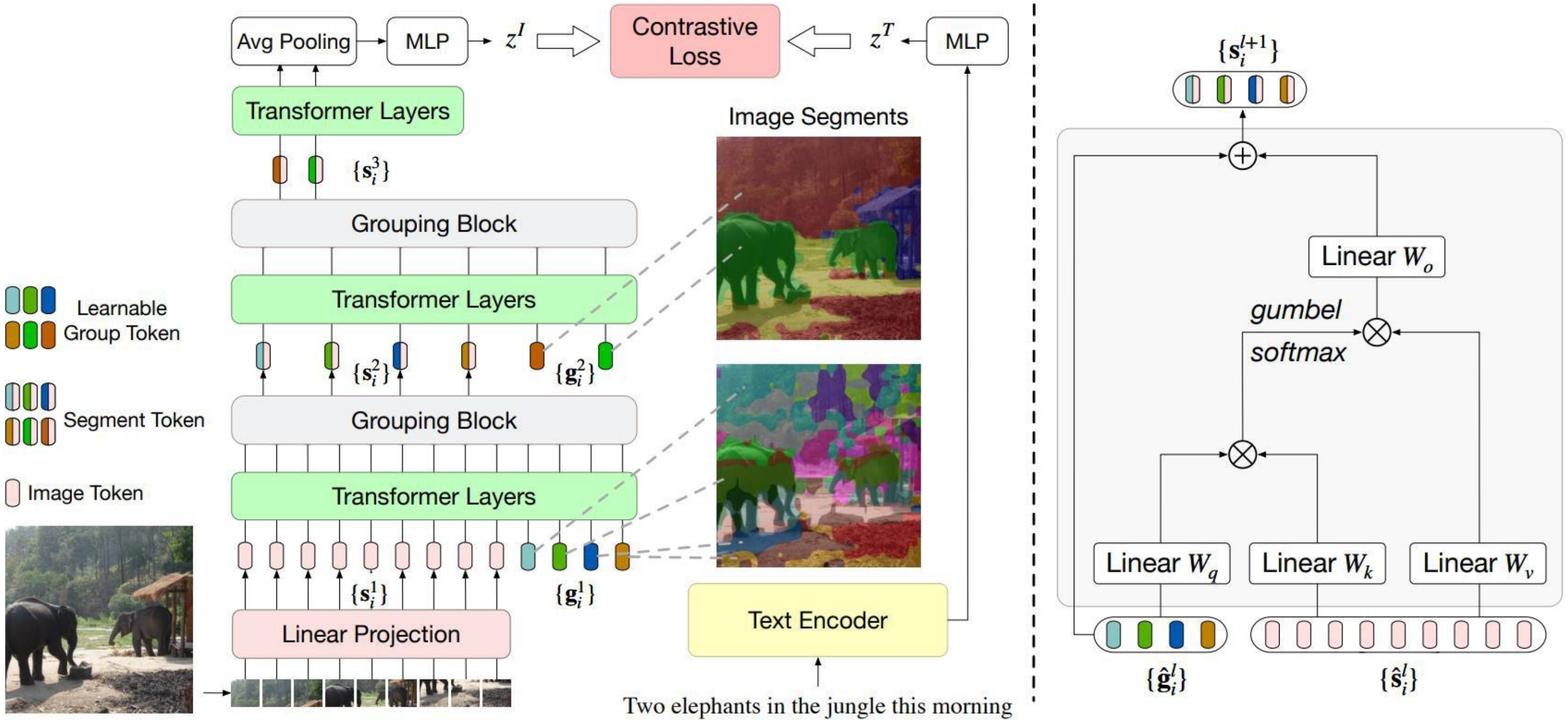
- Segmentation tasks:
  - Generic segmentation (semantic/instance/panoptic segmentation)
  - Referring segmentation (segment image with specific text phrase)
- Methodologies:
  - Initialize from CLIP *v.s.* train from scratch
  - Weakly supervised training *v.s.* supervised training
  - Two-stage *v.s.*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



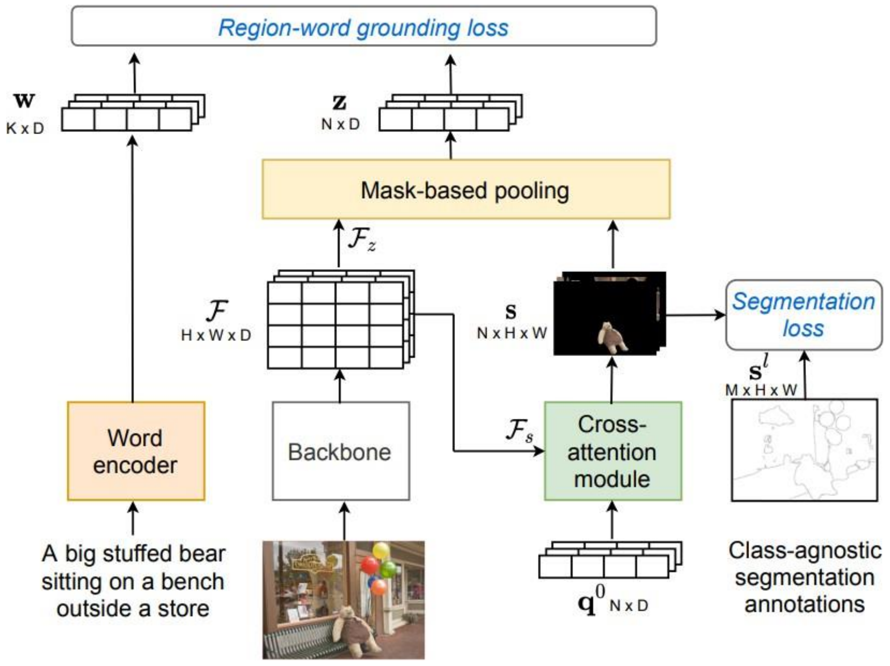


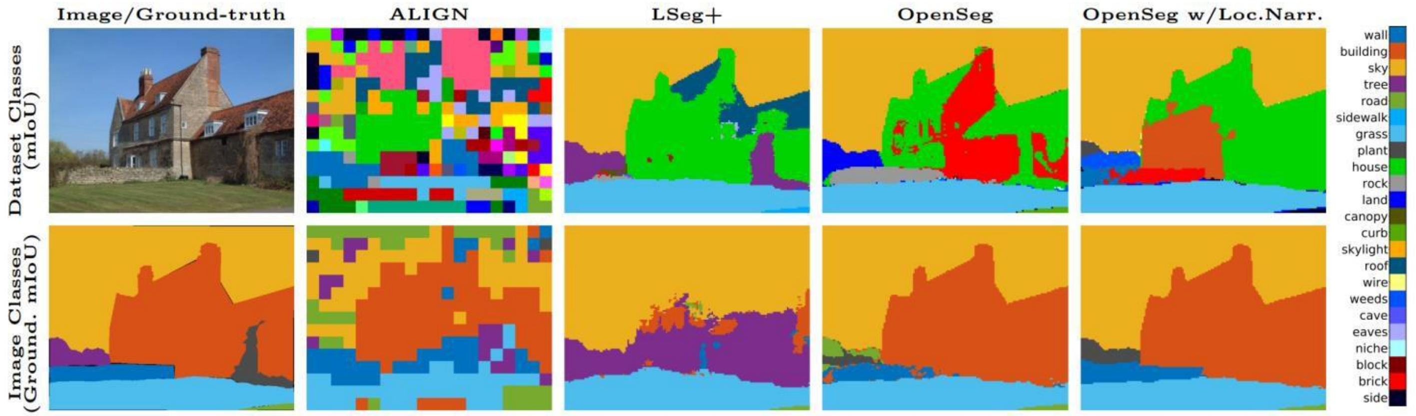
(a) GroupViT Architecture and Training Pipeline

(b) Grouping Block

Bridge Vision with Language for Segmentation

- 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.

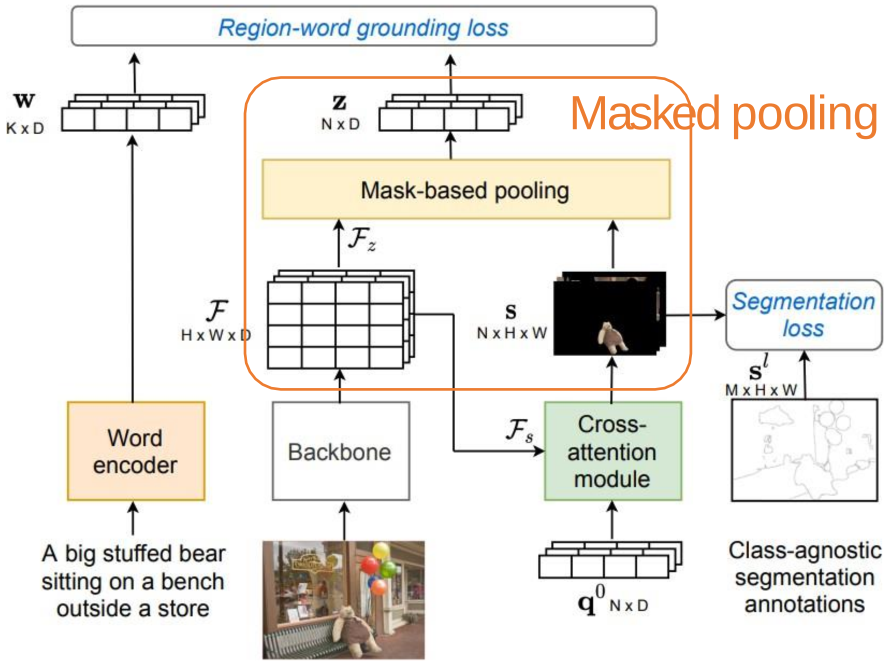


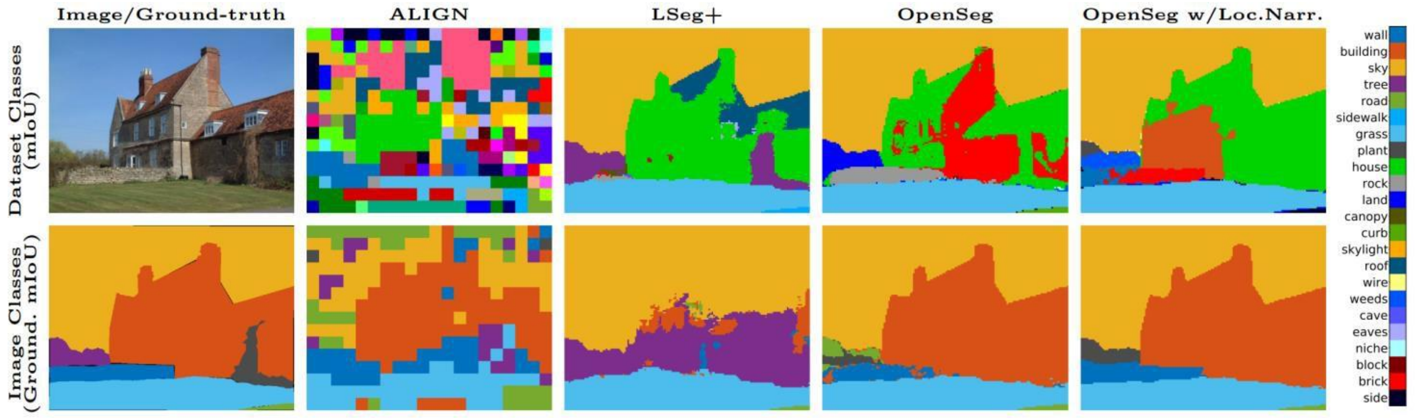


	COCO Train			mIoU					Grounding mIoU				
	label	mask	cap.	A-847	PC-459	A-150	PC-59	COCO	A-847	PC-459	A-150	PC-59	COCO
ALIGN	✗	✗	✗	4.8	3.6	9.7	18.5	15.6	17.8	21.8	25.7	34.2	28.2
ALIGN w/proposal	✗	✓	✗	5.8	4.8	12.9	22.4	17.9	17.3	19.7	25.3	32.0	23.6
LSeg+	✓	✓	✗	3.8	7.8	18.0	**46.5**	55.1	10.5	17.1	30.8	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.

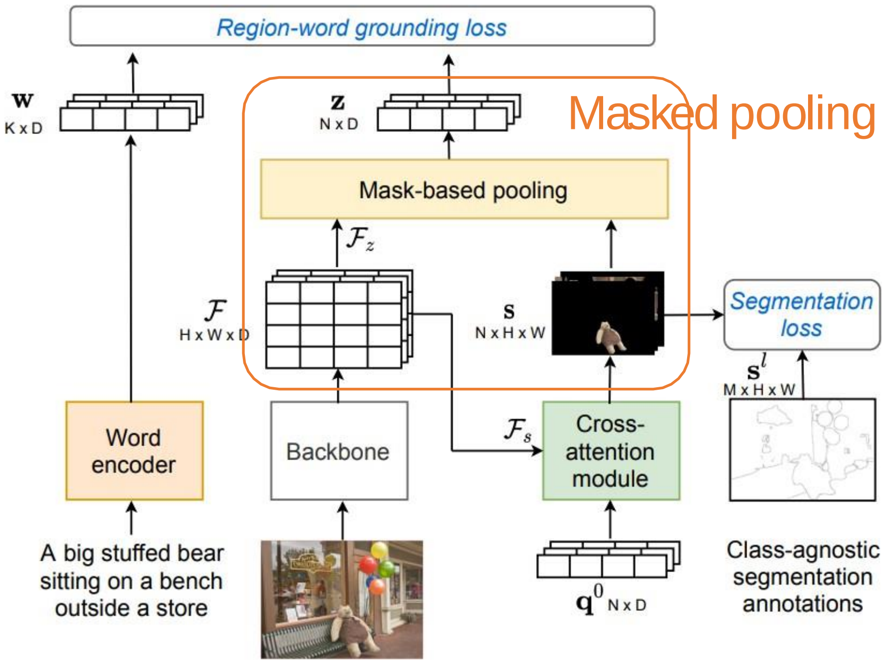


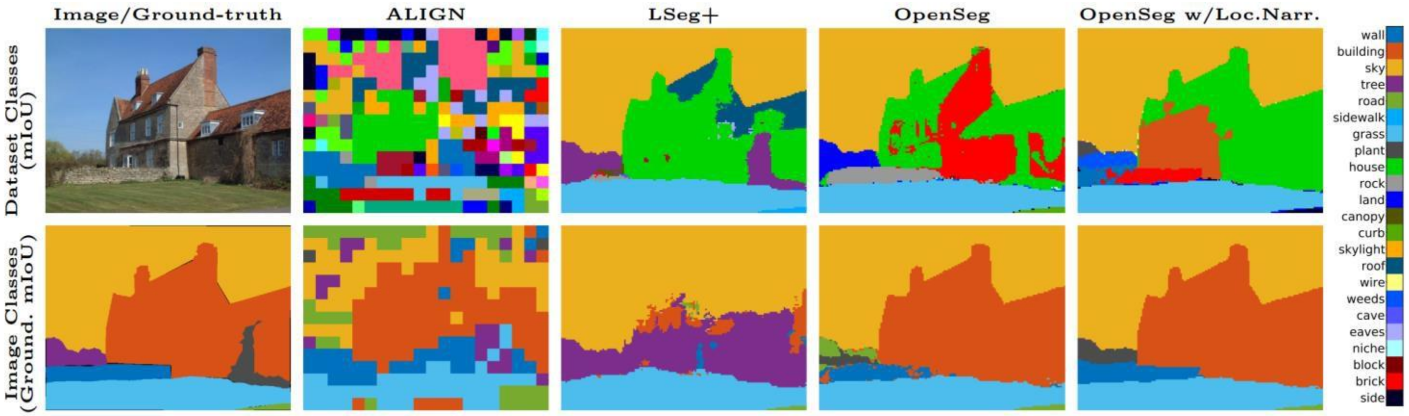


	COCO Train			mIoU					Grounding mIoU				
	label	mask	cap.	A-847	PC-459	A-150	PC-59	COCO	A-847	PC-459	A-150	PC-59	COCO
ALIGN	✗	✗	✗	4.8	3.6	9.7	18.5	15.6	17.8	21.8	25.7	34.2	28.2
ALIGN w/proposal	✗	✓	✗	5.8	4.8	12.9	22.4	17.9	17.3	19.7	25.3	32.0	23.6
LSeg+	✓	✓	✗	3.8	7.8	18.0	**46.5**	55.1	10.5	17.1	30.8	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.

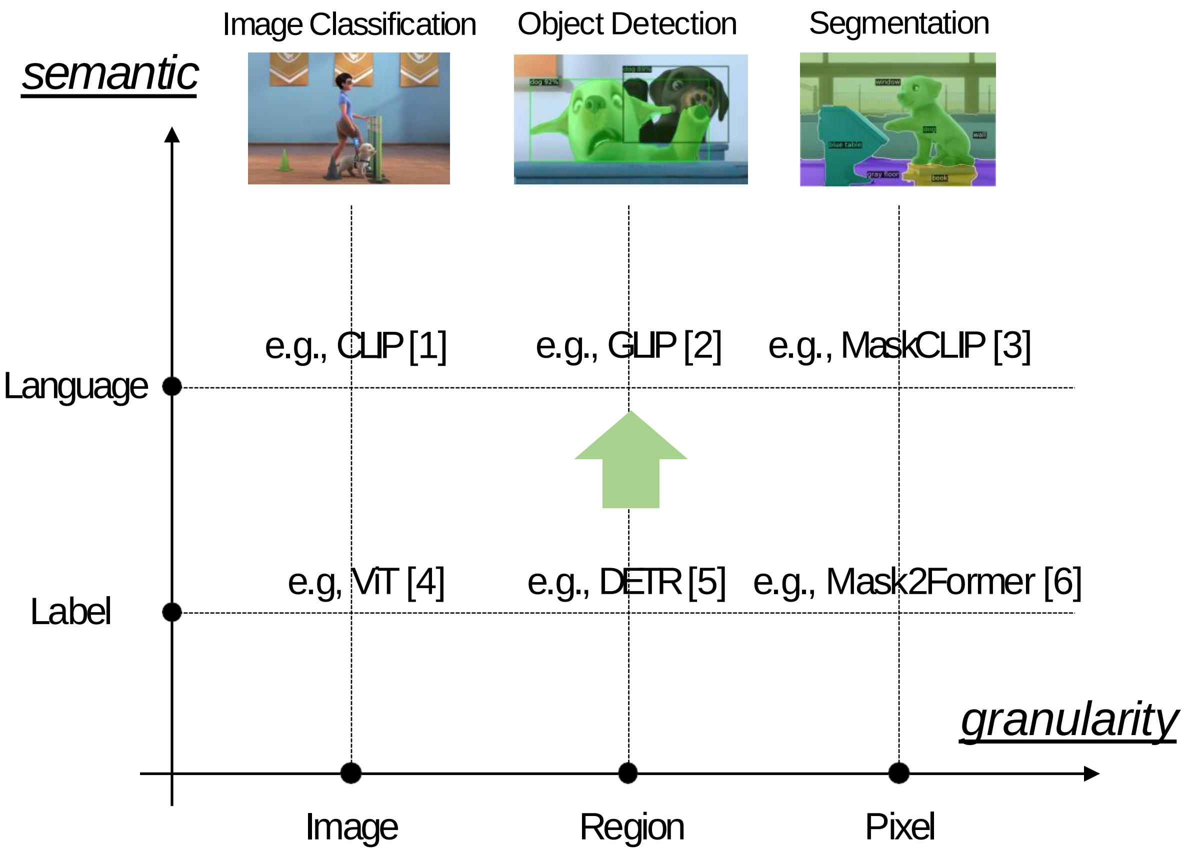




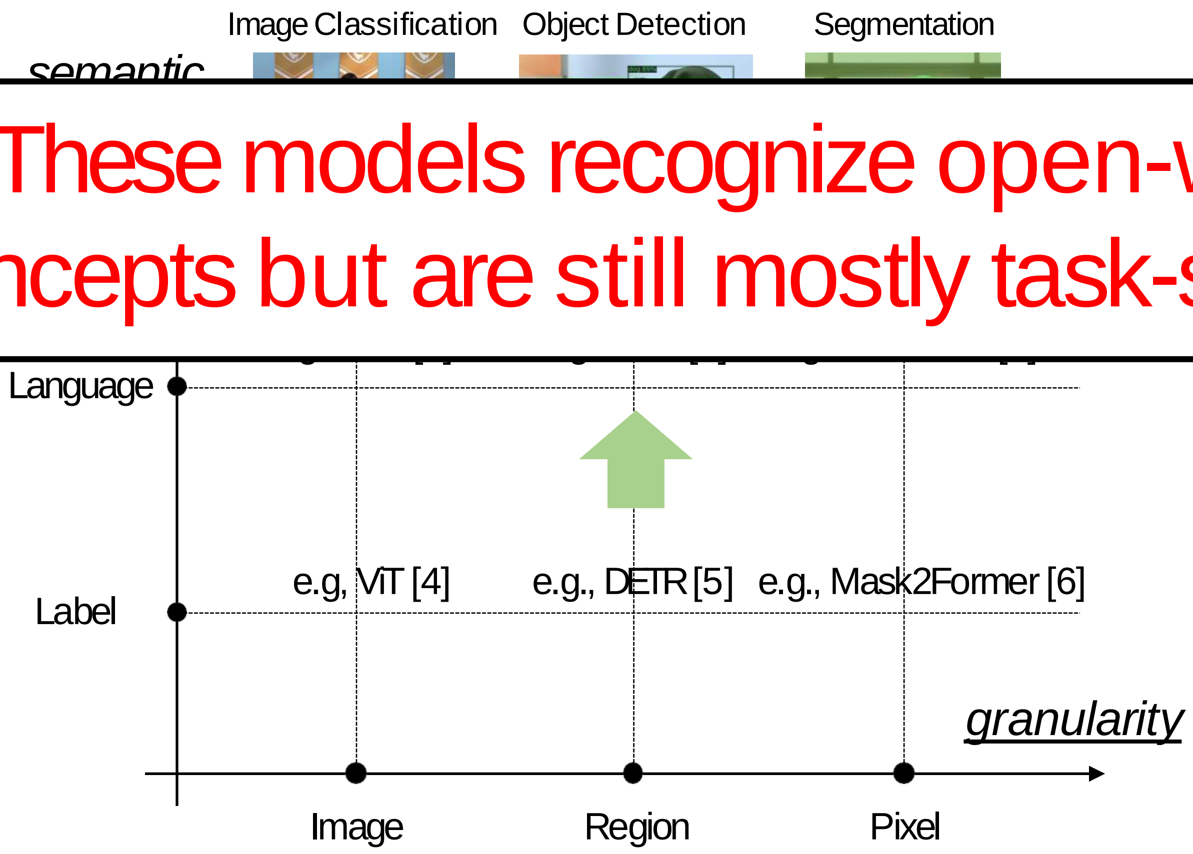
	COCO Train			mIoU					Grounding mIoU				
	label	mask	cap.	A-847	PC-459	A-150	PC-59	COCO	A-847	PC-459	A-150	PC-59	COCO
ALIGN	✗	✗	✗	4.8	3.6	9.7	18.5	15.6	17.8	21.8	25.7	34.2	28.2
ALIGN w/proposal	✗	✓	✗	5.8	4.8	12.9	22.4	17.9	17.3	19.7	25.3	32.0	23.6
LSeg+	✓	✓	✗	3.8	7.8	18.0	**46.5**	55.1	10.5	17.1	30.8	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 Core Vision

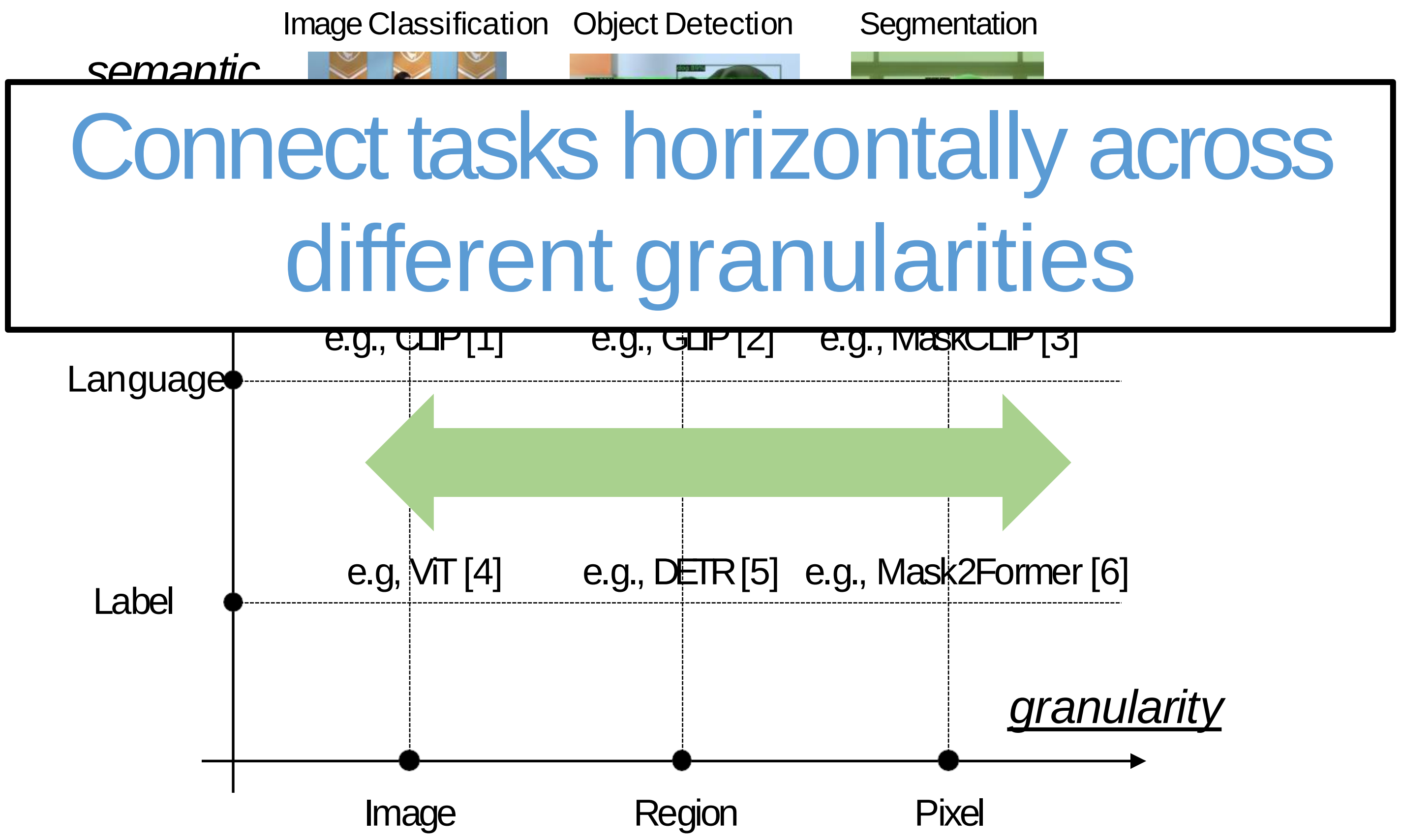


Bridge Vision with Language for Core Vision

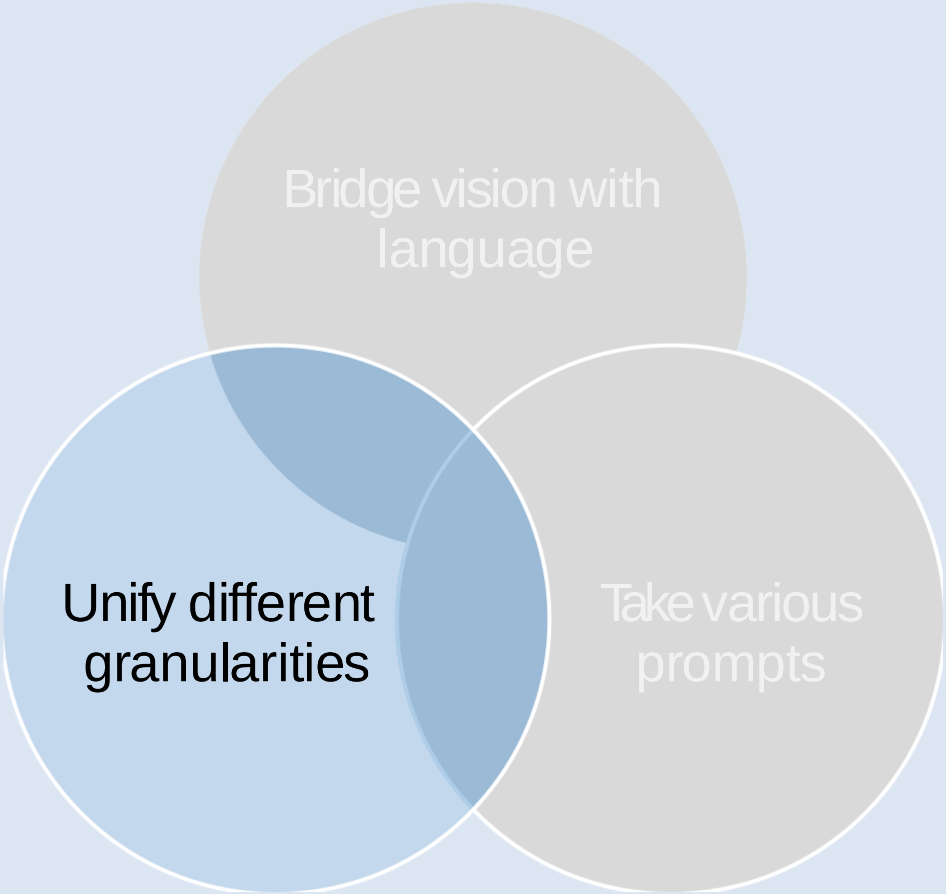


These models recognize open-world concepts but are still mostly task-specific

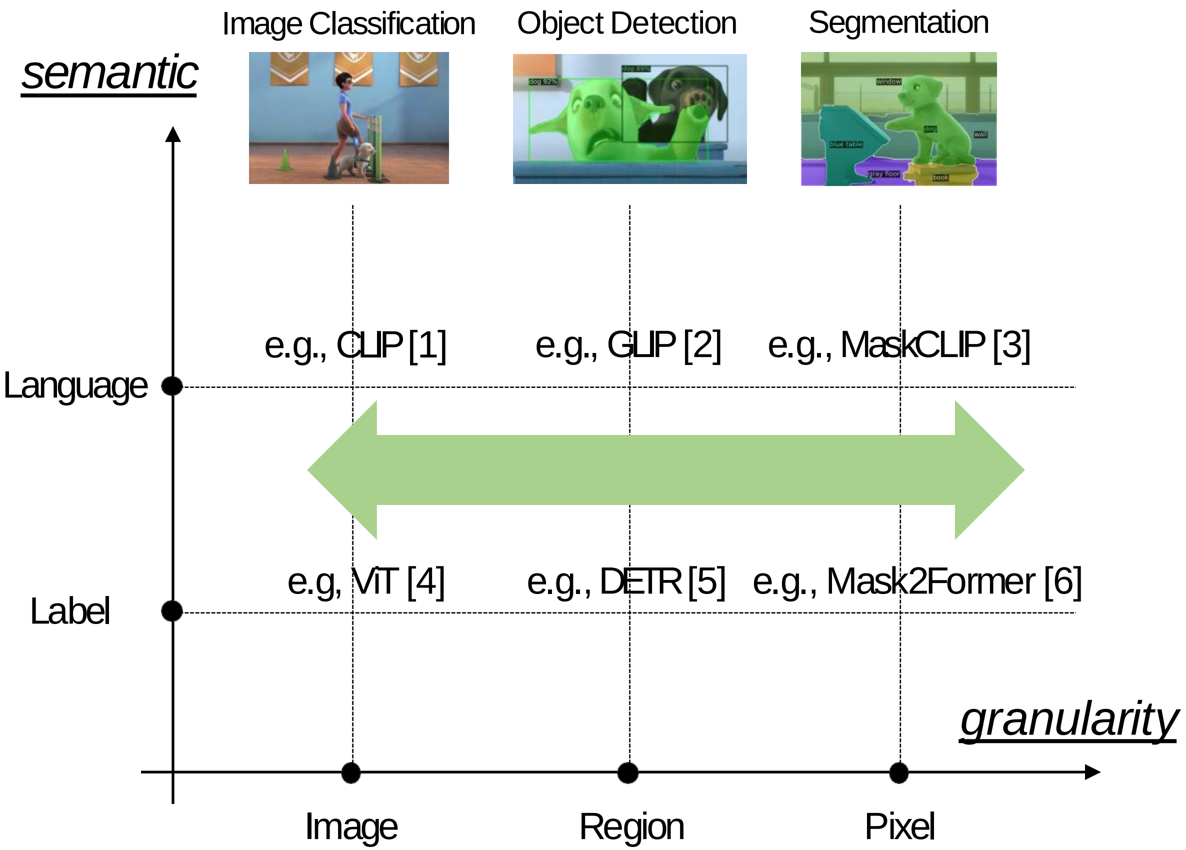
Bridge Vision with Language for Core Vision

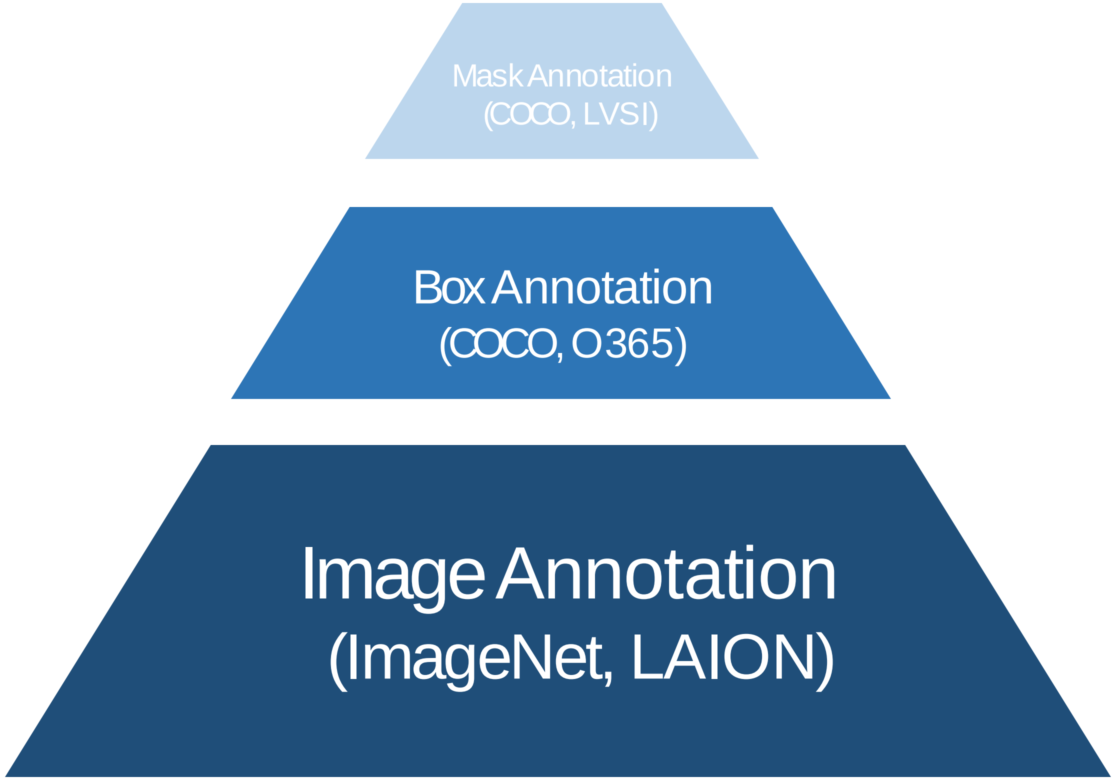


II. Unify Different Granularities

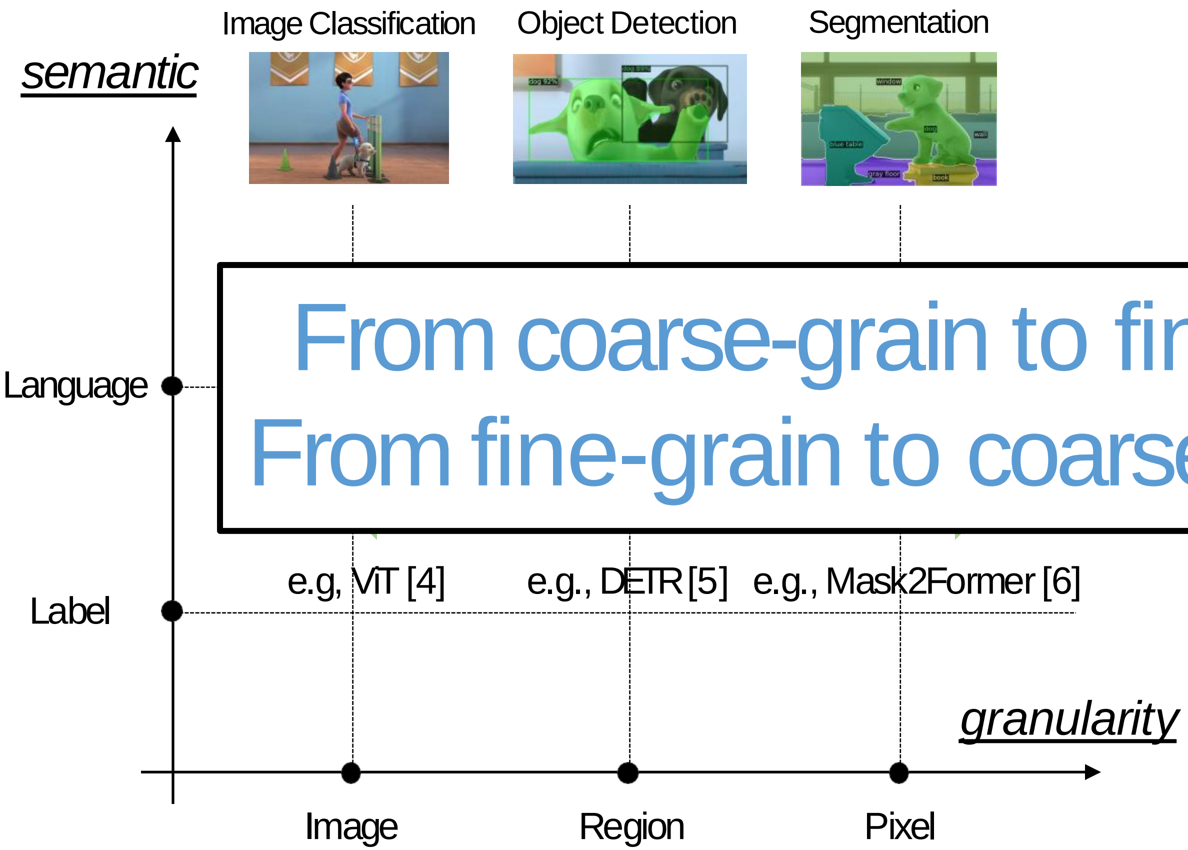


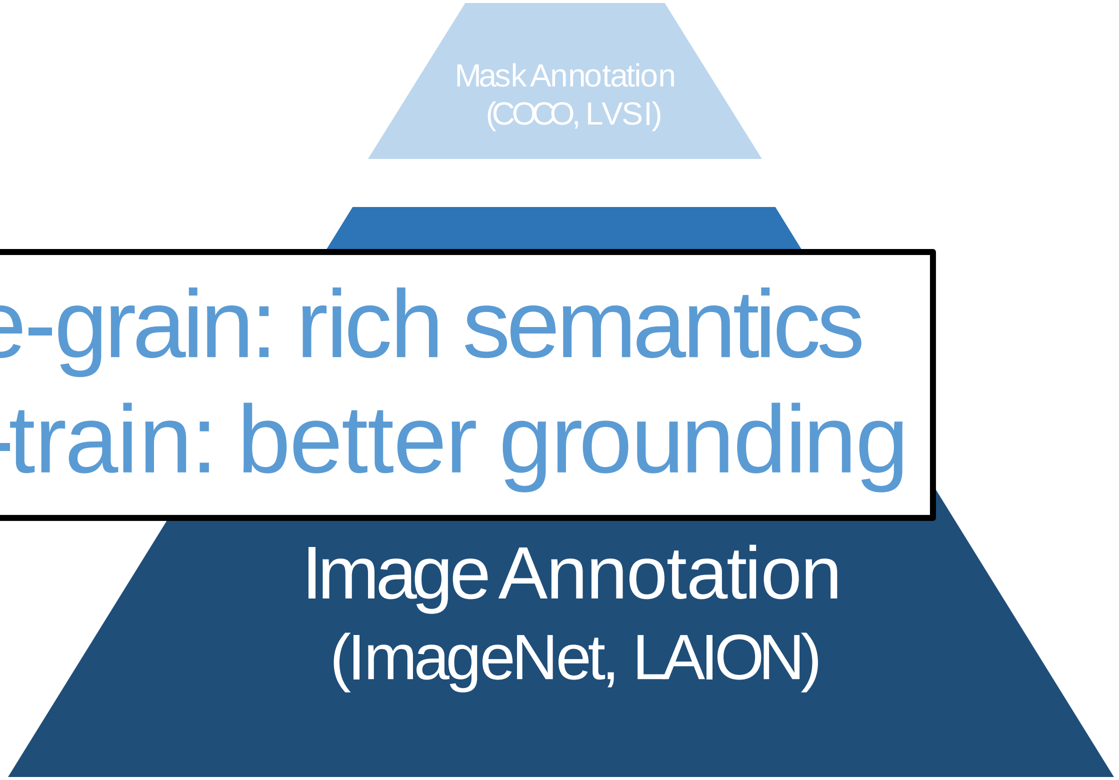
Unify Different Granularities





Unify Different Granularities





From coarse-grain to fine-grain: rich semantics

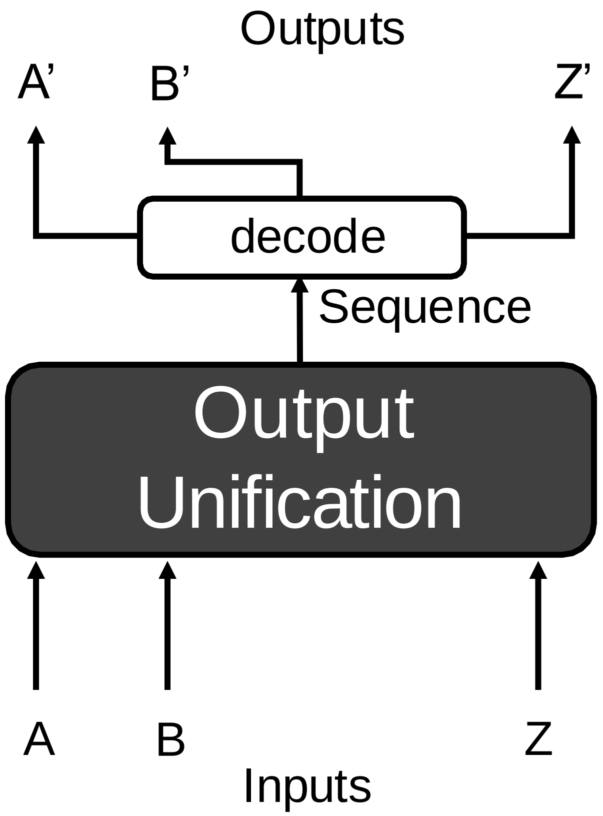
From fine-grain to coarse-train: better grounding

Unify Different Granularities

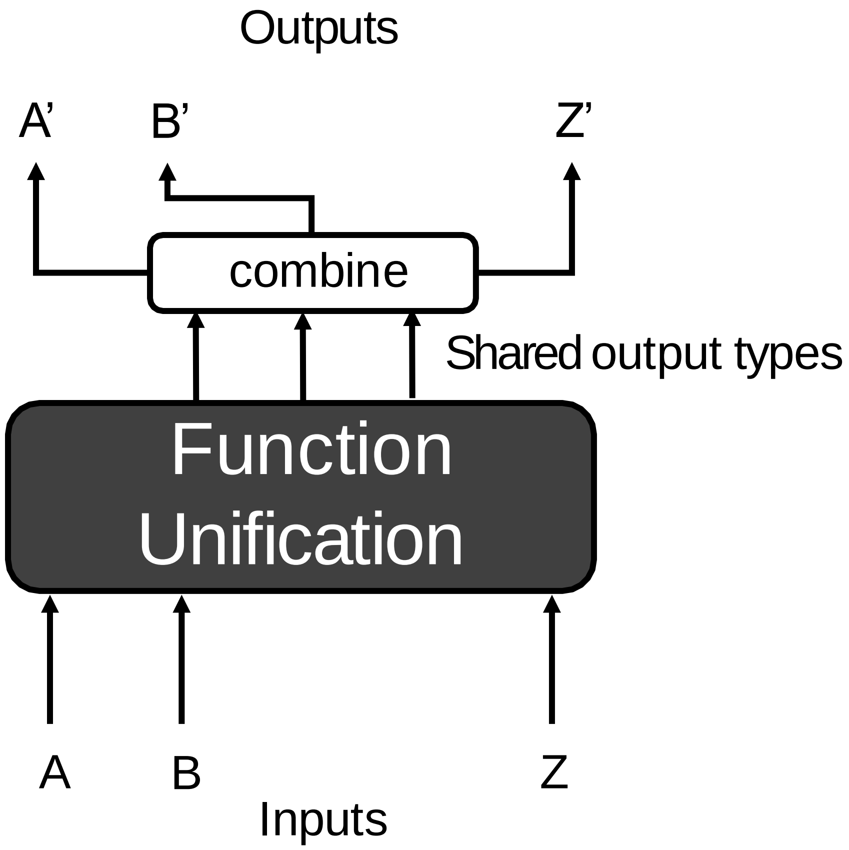
- Tasks we are considering:
  - Image-level: image recognition, image-text retrieval, image captioning, visual question answering, etc.
  - Region-level: object detection, dense caption, phrase grounding, etc.
  - Pixel-level: generic segmentation, referring segmentation, etc.

- 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



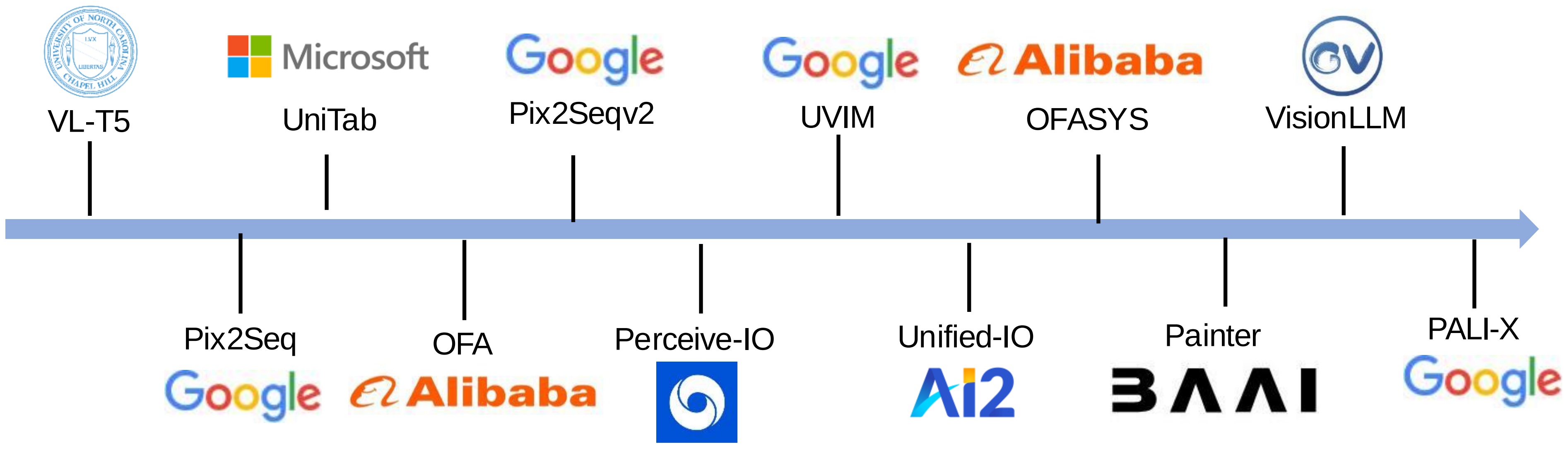
Convert all outputs into sequence and decode to corresponding outputs



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

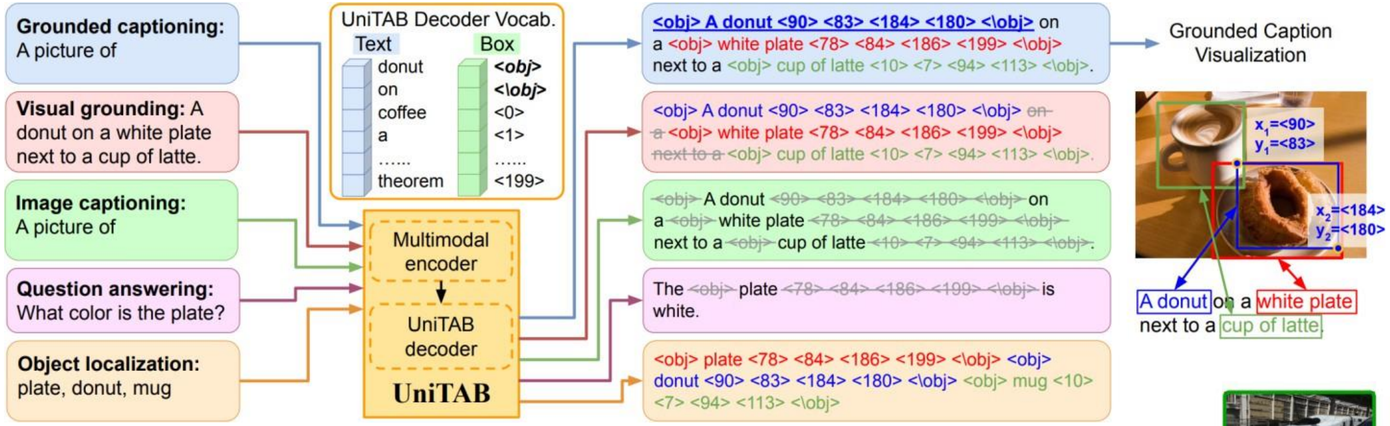
Outputs Unification

- 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

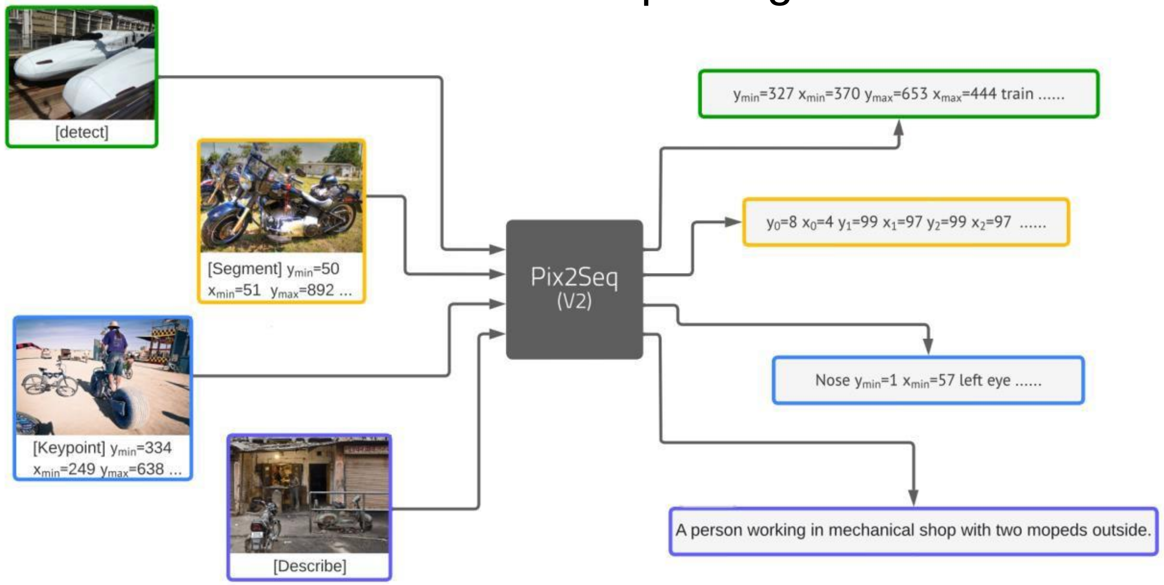
- UniTab and Pix2Seqv2: Unify text and box outputs with no specific modules



Method	Caption Eval.				Grounding Eval.	
	B@4	M	C	S	$F1_{all}$	$F1_{loc}$
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**

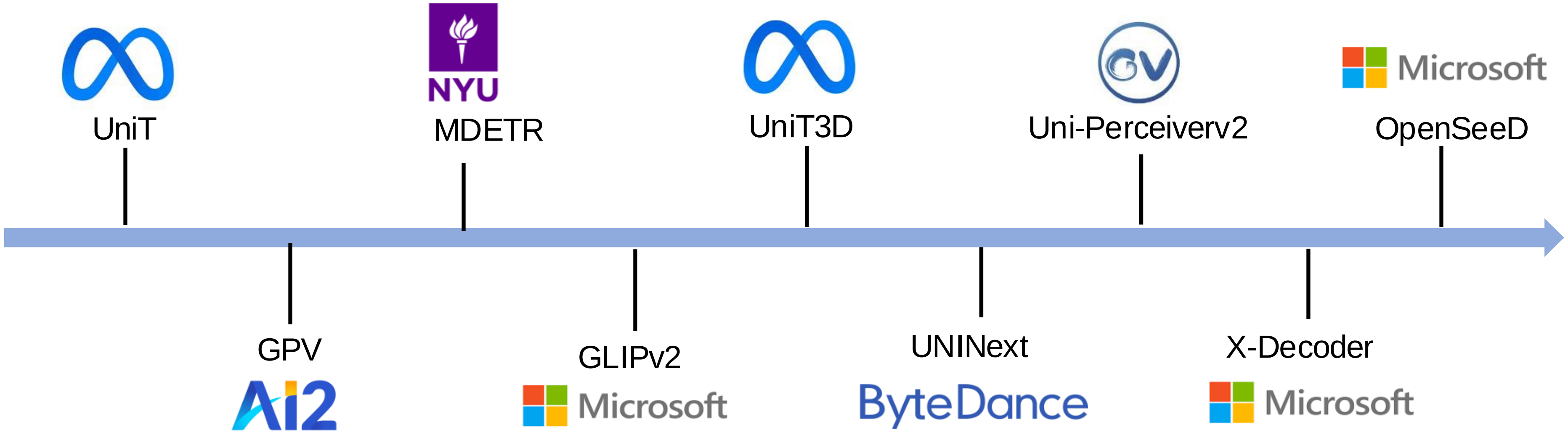
Grounded Captioning Evaluation

- 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



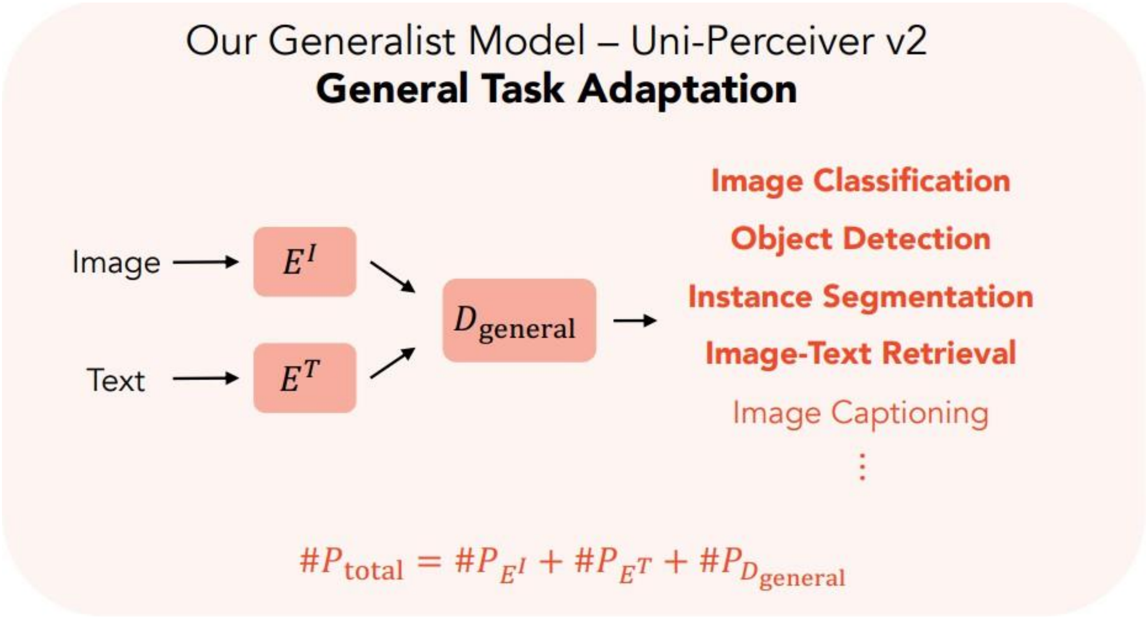
Functionality Unification

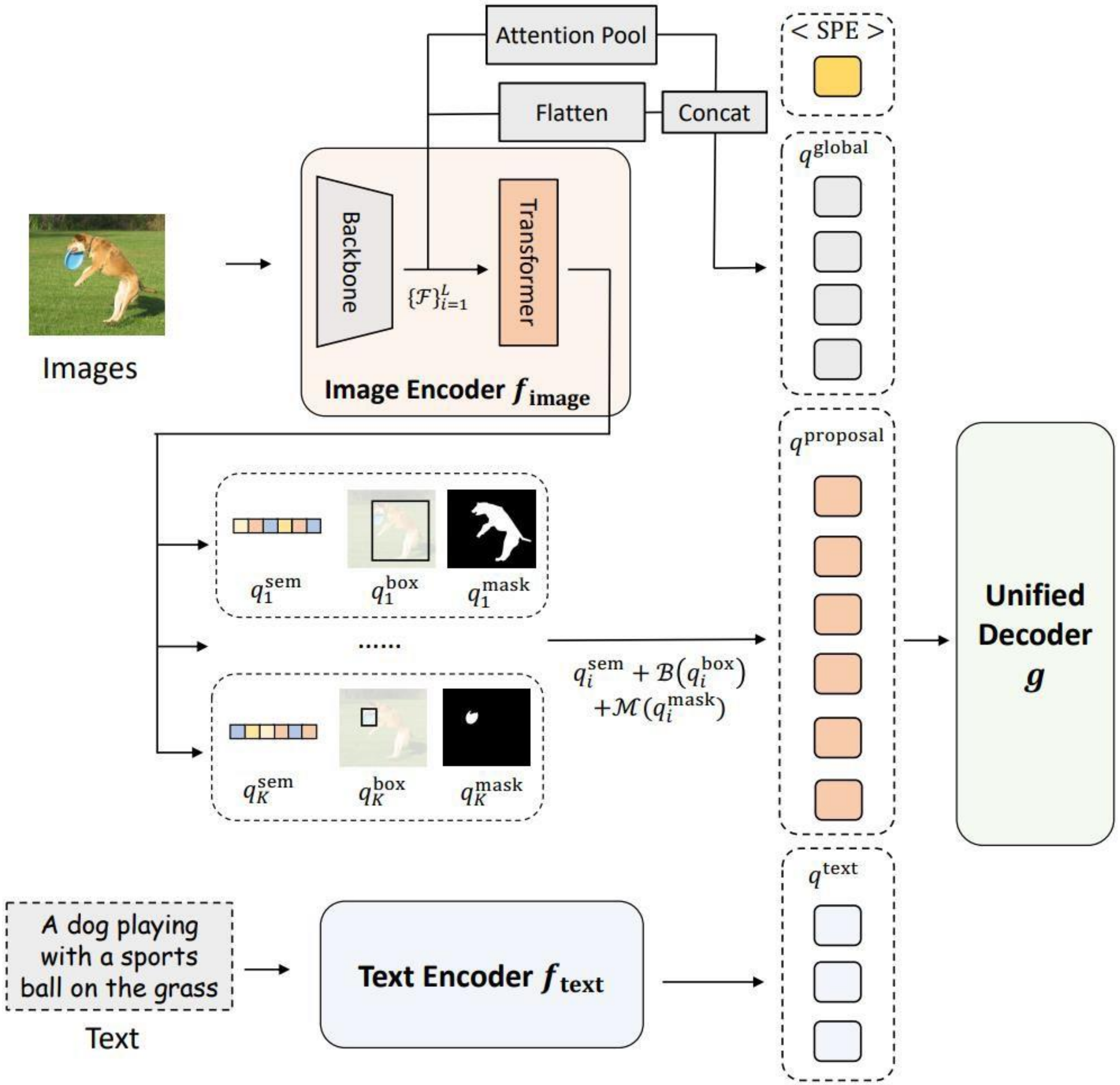
- Vision tasks are not fully isolated:
  - Box outputs: shared by generic object detection, phrase grounding, regional captioning
  - Mask outputs: shared by instance/semantic/panoptic segmentation, referring segmentation, exemplar-based segmentation, etc.
  - Semantic outputs: shared by image classification,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



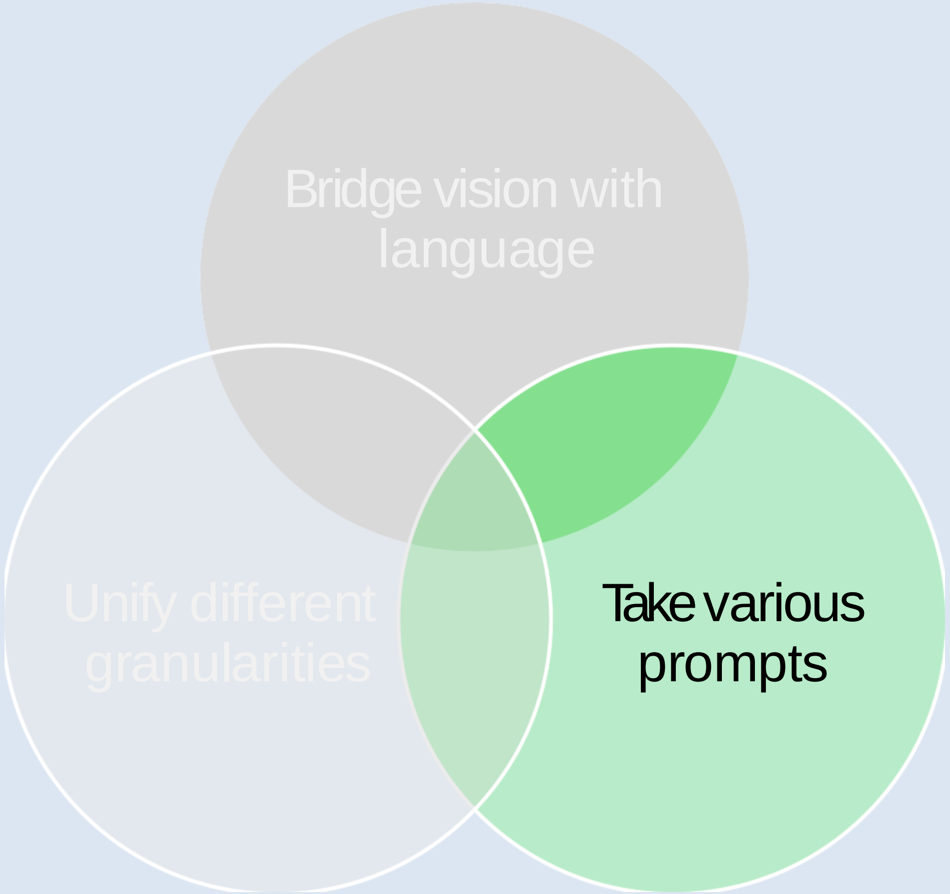
Functionality Unification

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





III. Promptable Interface



How to Enable Vision Model to “Chat”

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

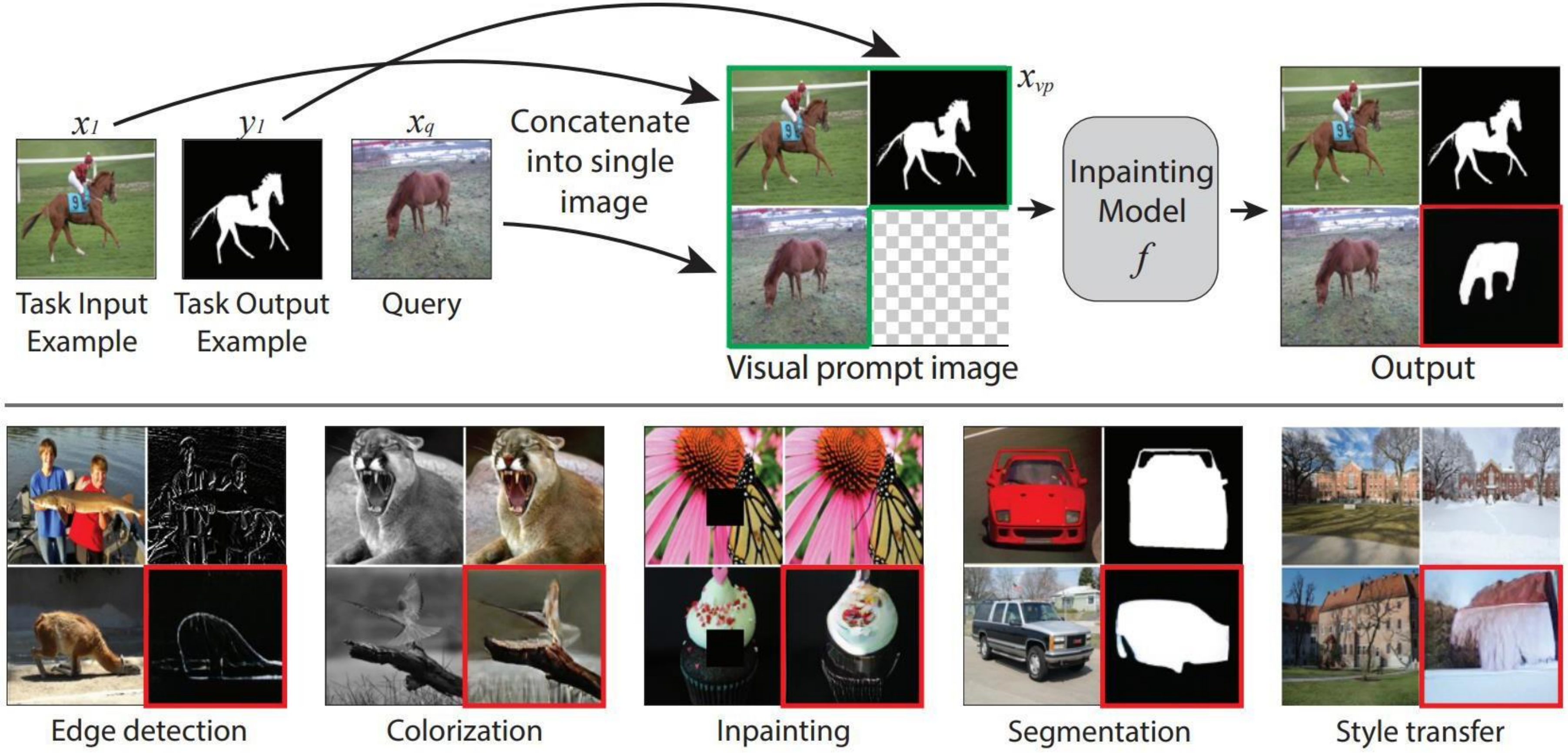
Generalist Vision Models — Human-AI Interaction → ?

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 predict
  - Interactive for user-friendly interface: multi-round of interaction between human and AI is important to finish complicated tasks

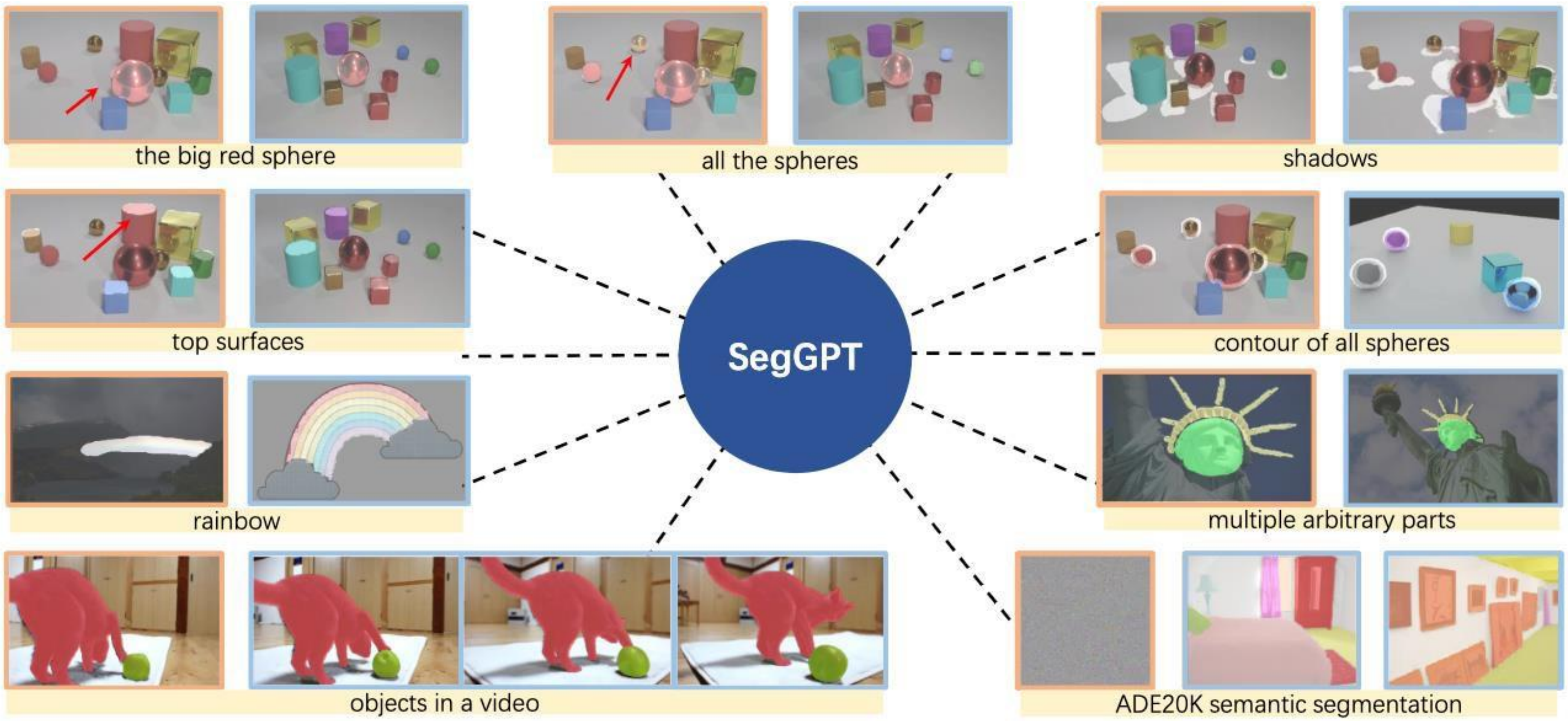
In-Context Learning for Vision

- 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



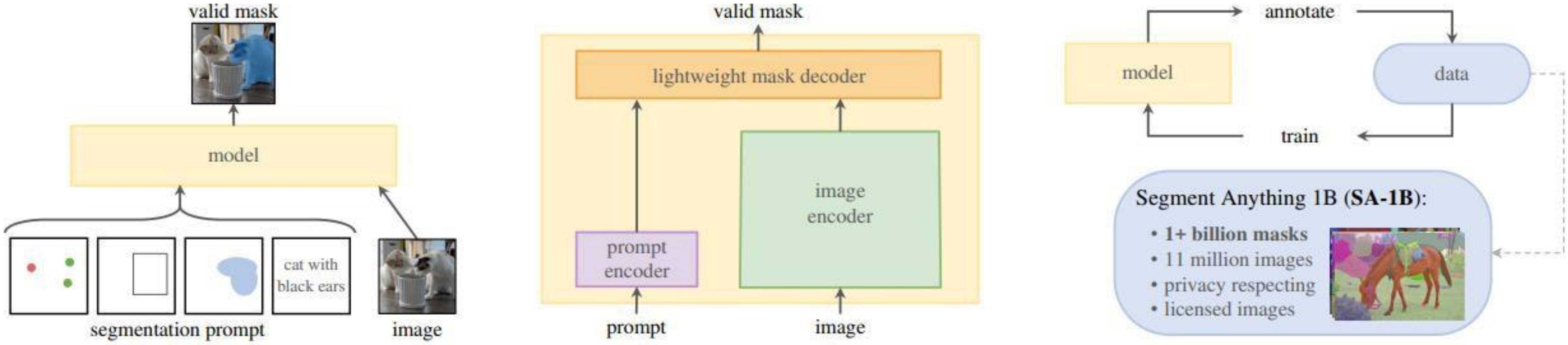
In-Context Learning for Vision

- SegGPT: Segment Everything as in-context learning

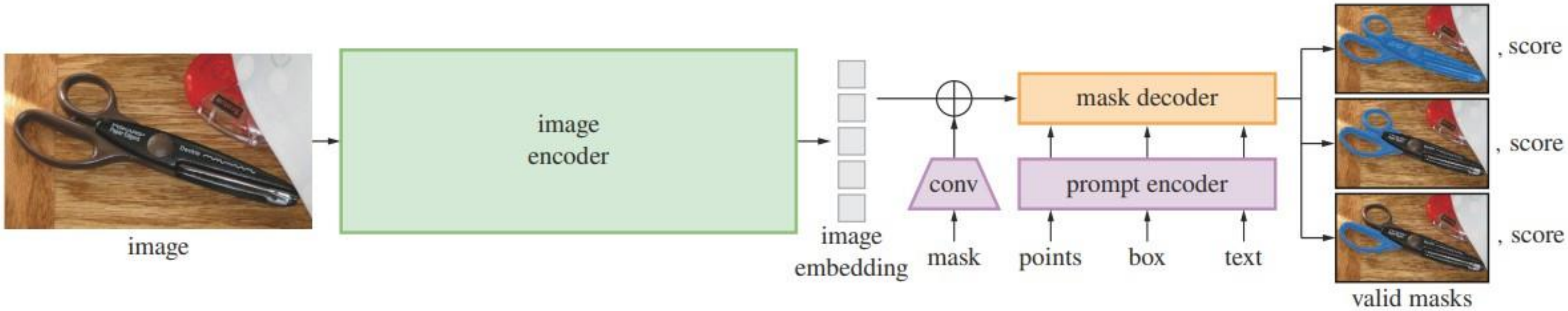


Interactive Interface for Vision

- SAM: Segment Anything
  - Promptable segmentation



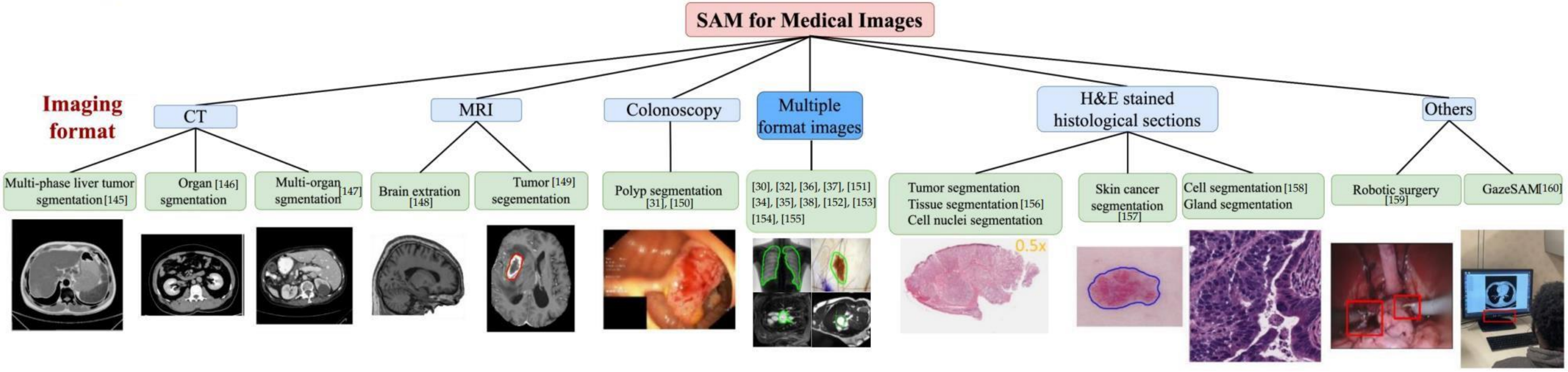
(a) **Task**: promptable segmentation (b) **Model**: Segment Anything Model (**SAM**) (c) **Data**: data engine (top) & dataset (bottom)



Interactive Interface for Vision

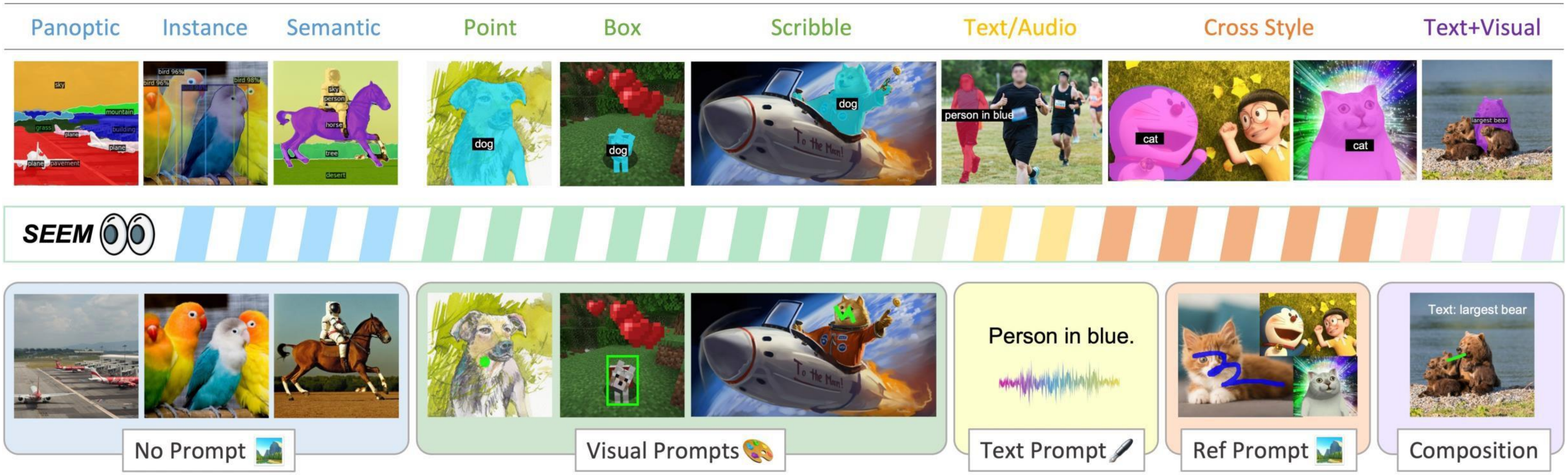
- SAM: Segment Anything

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

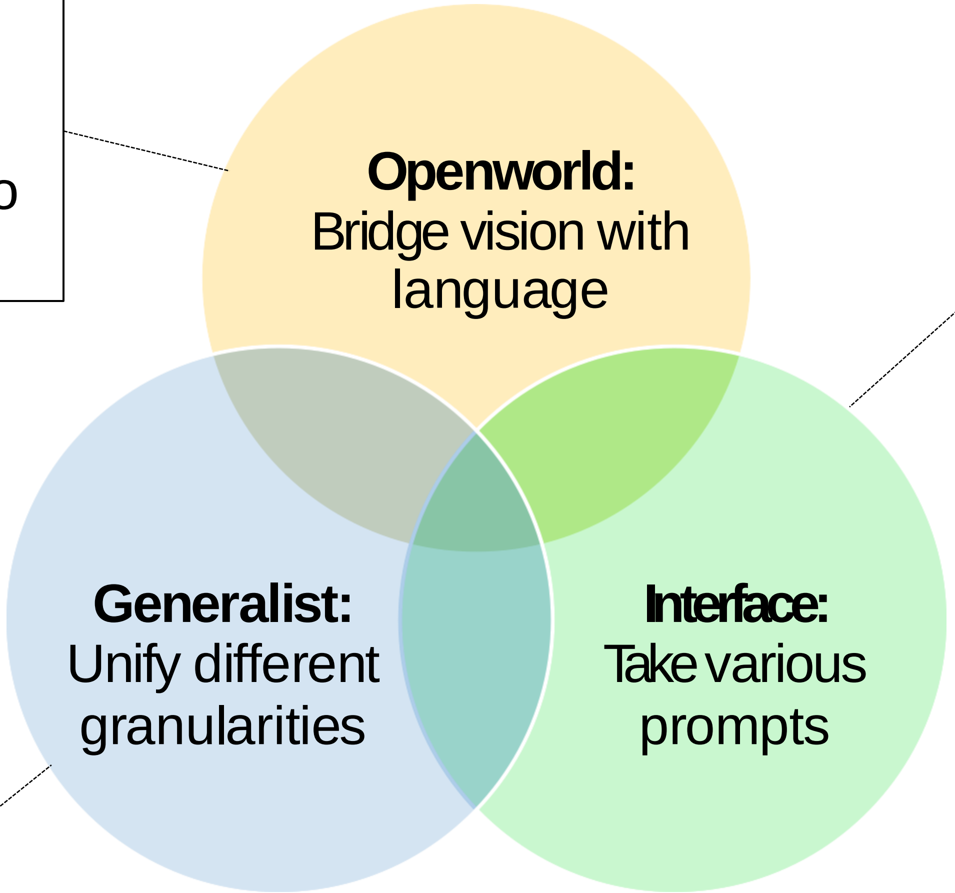


Interactive Interface for Vision

- SEEM: Segment Everything Everywhere all at Once



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

( , a dog is running through the grass)

Image Generation

Produce visual data

LLMs and models for image understanding and generation

Part 3: *How to make an LLM that can see and chat?*

Image Encoder

Consume visual data

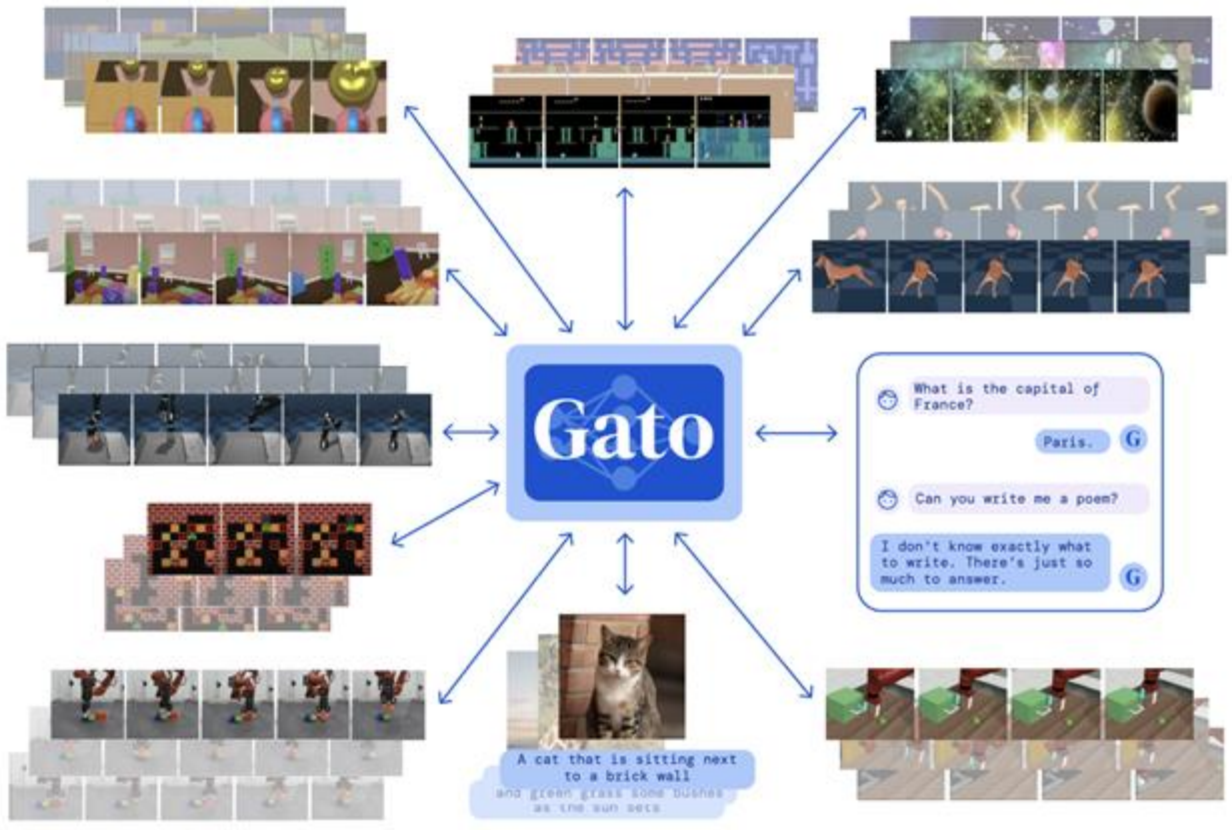
Part 1: *How to learn image representations?*

Part 2: *How to extend vision models with more flexible, promptable interfaces?*

Part 3: Multimodal LLMs

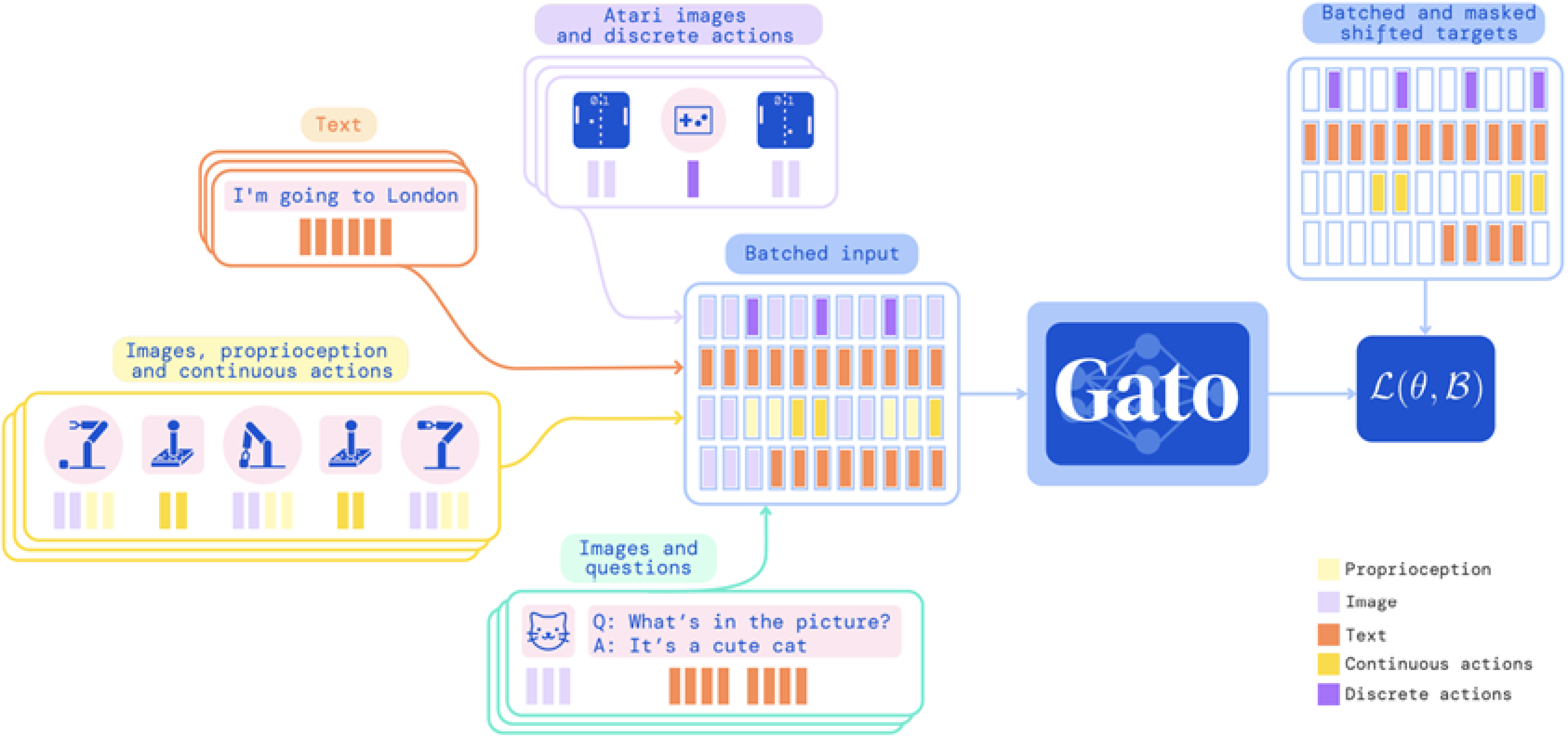
How to make an LLM that can see and chat?

Gato network with the same weights can play Atari, caption images, chat, stack blocks with a real robot arm and much more



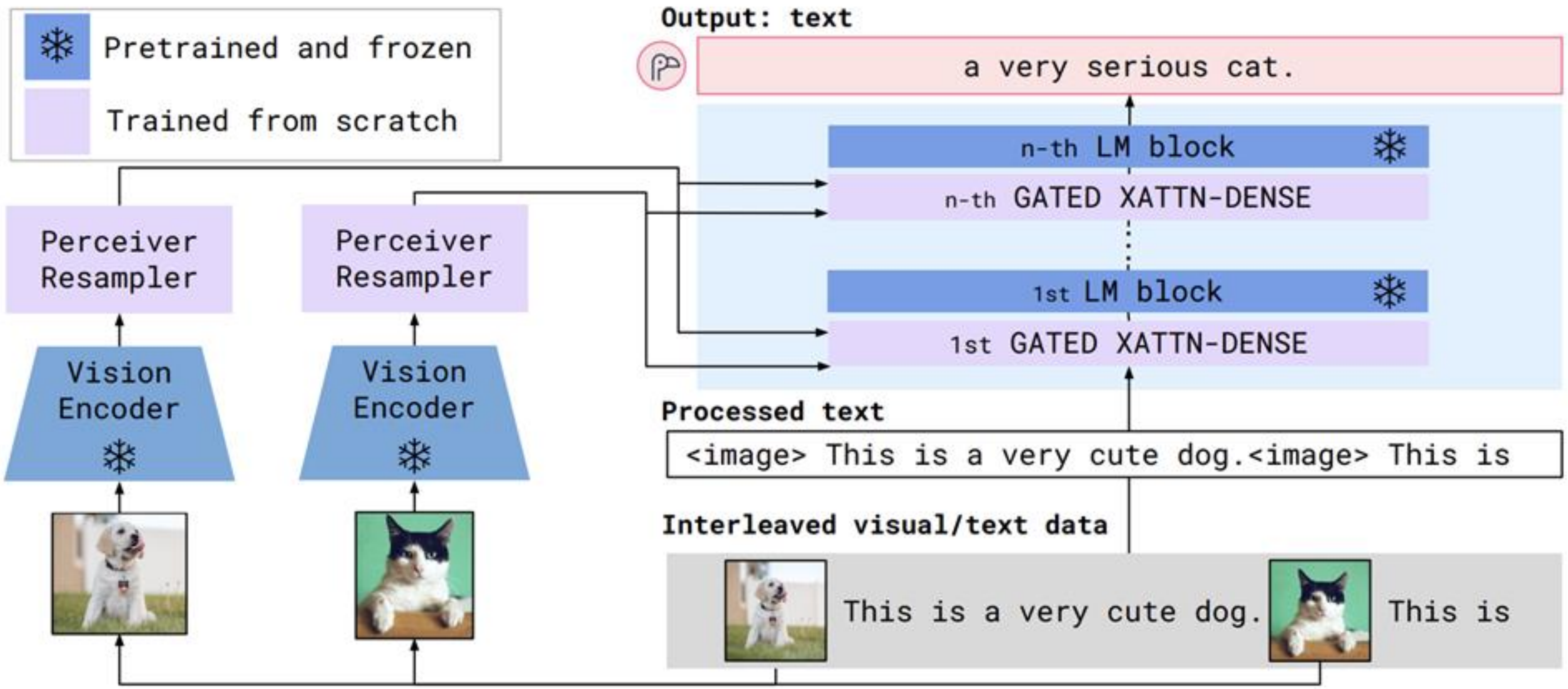
Reed, S. et al. A generalist agent. In Transactions on Machine Learning Research (2022).

Data from different tasks and modalities is serialized into a flat sequence of tokens, batched, and processed by a transformer neural network akin to a large language model.



Reed, S. et al. A generalist agent. In Transactions on Machine Learning Research (2022).

Flamingo is a visual language model that take as input visual data interleaved with text and produce free-form text as output

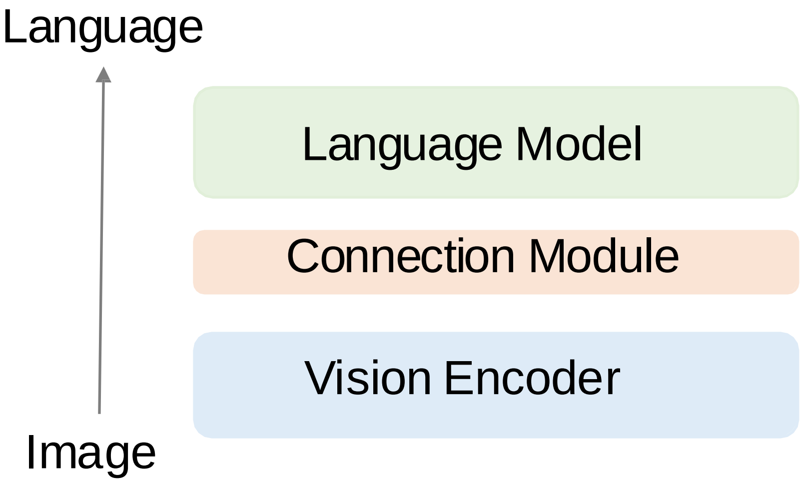


Alayrac, J.-B. et al. Flamingo: a Visual Language Model for few-shot learning. In Advances in Neural Information Processing Systems (eds Oh, A. H. et al.) 35, 23716–23736 (2022).

Large Multimodal Models: Image-to-Text Generative Models

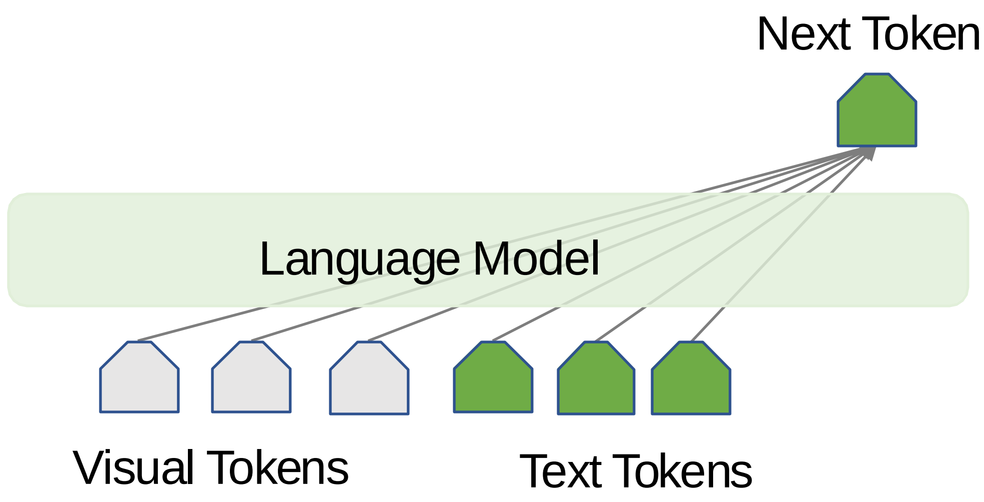
- ❑ Model Architectures
  - (Pre-trained) Image Encoder and Language Models
  - Trainable modules to connect to two modalities

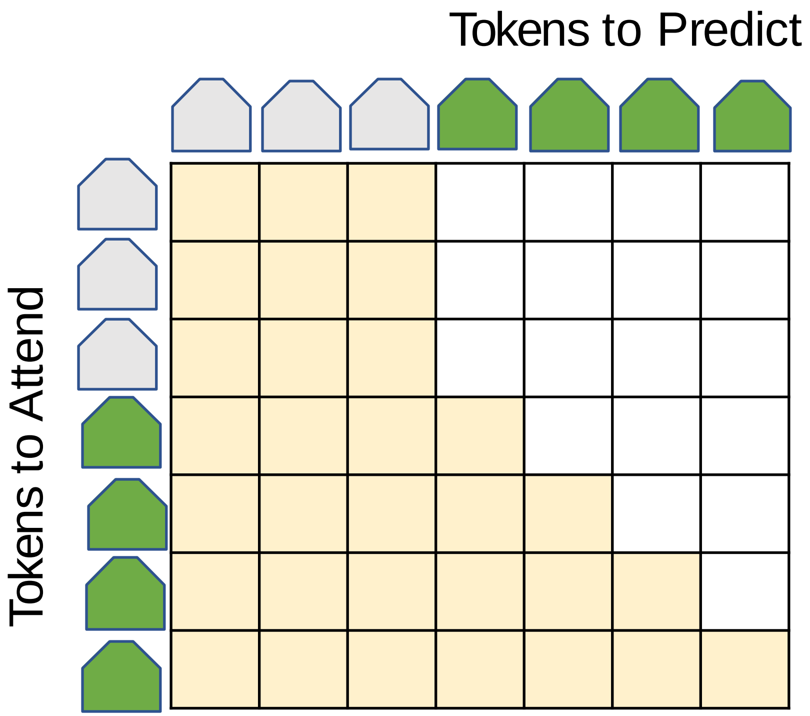
A dog lying on the grass next to a frisbee



Large Multimodal Models: Image-to-Text Generative Models

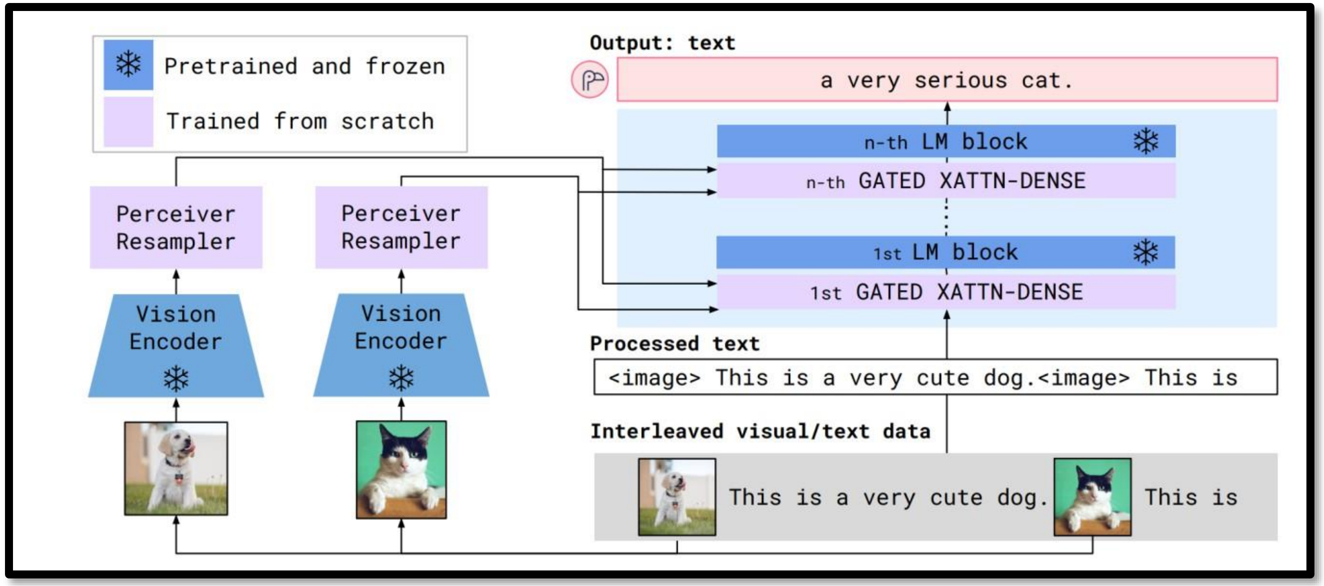
- Training Objective
  - Cross-Attended Image-to-Text Generation
  - Autoregressive loss on **language output**





Example 2: LMM with Interleaved Image-Text Data

- Flamingo:

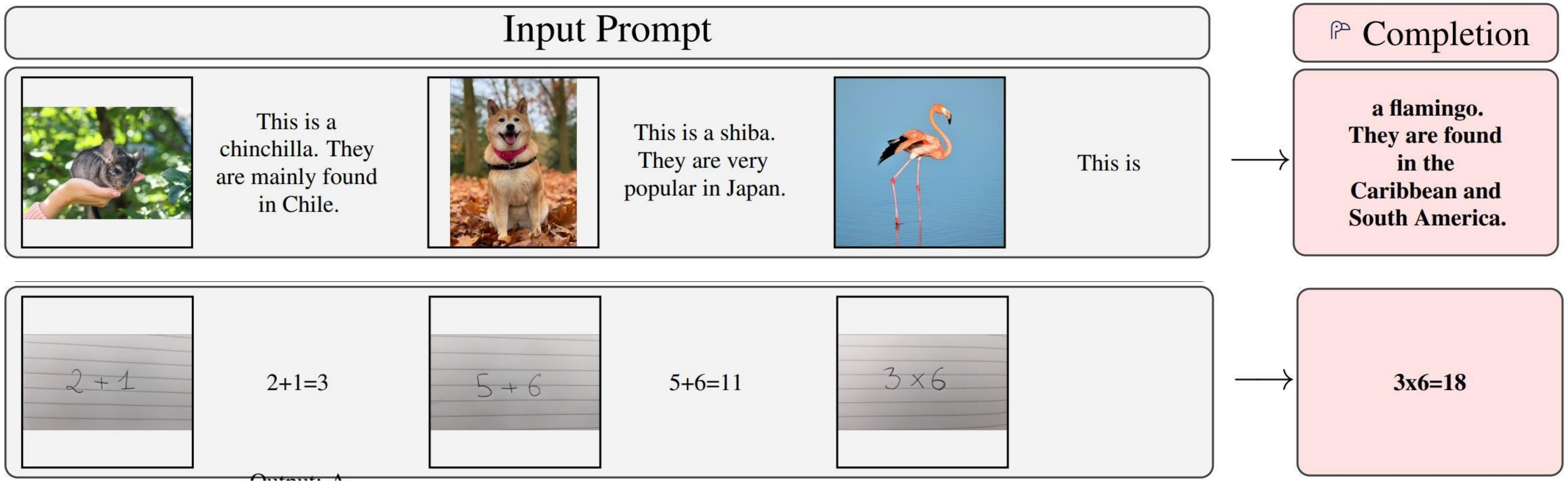


Language Model	Pre-trained: 70B Chinchilla
Connection Module	Perceiver Resampler Gated Cross-attention + Dense
Vision Encoder	Pre-trained: Nonrmalizer-Free ResNet (NFNet)

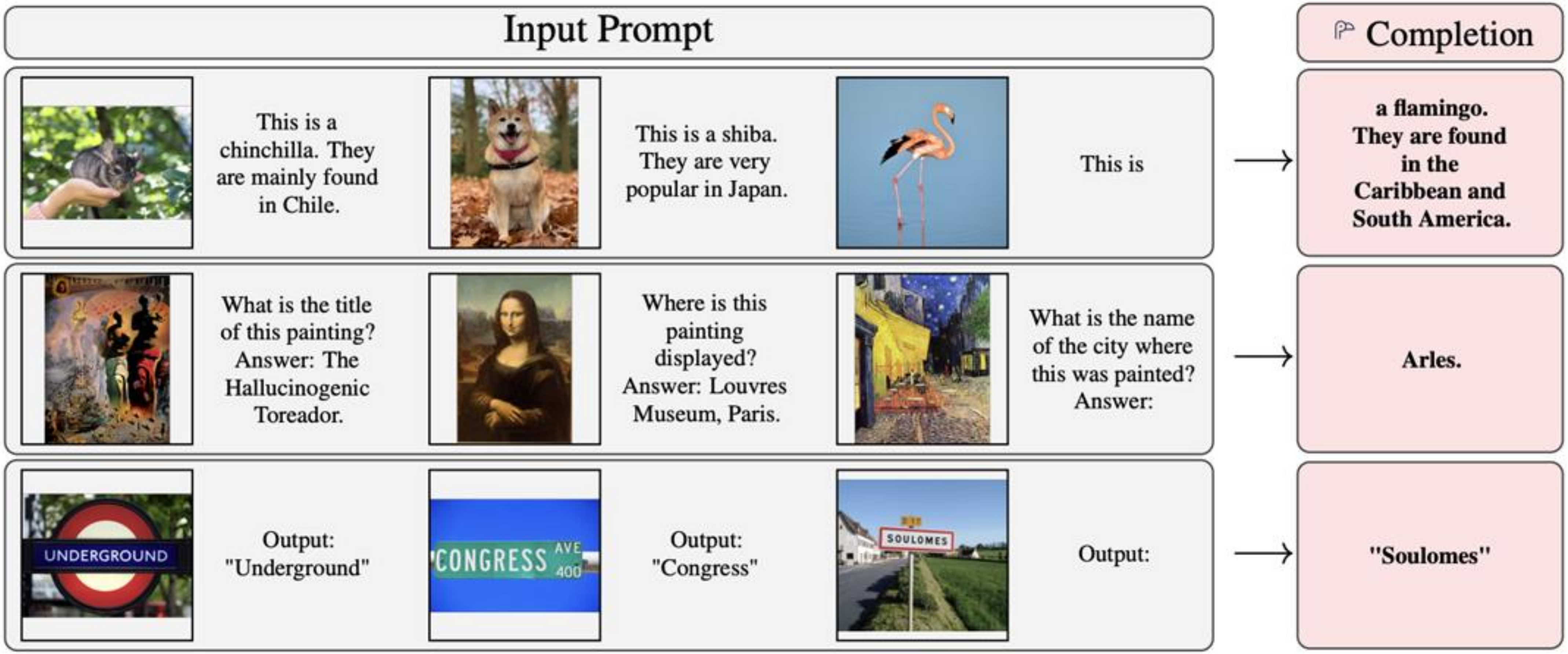
Example 2: LMM with Interleaved Image-Text Data

- Flamingo: Multimodal In-Context-Learning

Emerging Property

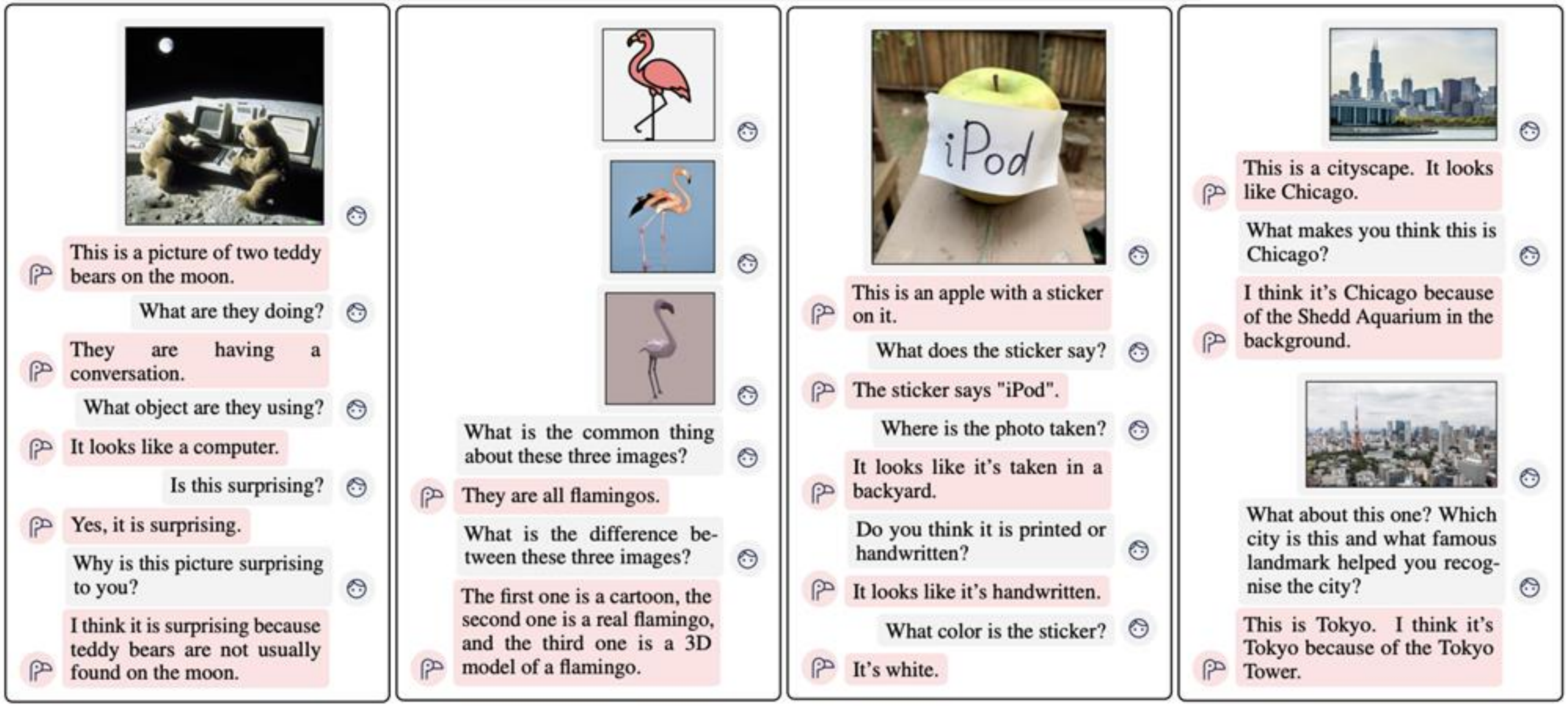


Flamingo rapidly adapts to various image/video understanding tasks with few-shot prompting



Alayrac, J.-B. et al. Flamingo: a Visual Language Model for few-shot learning. In Advances in Neural Information Processing Systems (eds Oh, A. H. et al.) 35, 23716–23736 (2022).

Flamingo is also capable of multi-image visual dialogue without further training



Alayrac, J.-B. et al. Flamingo: a Visual Language Model for few-shot learning. In Advances in Neural Information Processing Systems (eds Oh, A. H. et al.) 35, 23716–23736 (2022).

OpenAI

MultiModal GPT-4

- Model Details: Unknown
- Capability: Strong zero-shot visual understanding & reasoning on many user-oriented tasks in the wild

- How can we build Multimodal GPT-4 like models?

GPT-4 visual input example, Extreme Ironing:	
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.

GPT-4 visual input example, Chicken Nugget Map:	
User	Can you explain this meme?
	
GPT-4	This meme is a joke that combines two unrelated things: pictures of the earth from space and chicken nuggets. The text of the meme suggests that the image below is a beautiful picture of the earth from space. However, the image is actually of chicken nuggets arranged to vaguely resemble a map of the world. The humor in this meme comes from the unexpected juxtaposition of the text and the image. The text sets up an expectation of a majestic image of the earth, but the image is actually something mundane and silly.

GPT-4 Technical Report, OpenAI

( , a dog is running through the grass)

Image Generation

Produce visual data

LLMs and models for image understanding and generation

Part 3: *How to make an LLM that can see and chat?*

Image Encoder

Consume visual data

Part 1: *How to learn image representations?*

Part 2: *How to extend vision models with more flexible, promptable interfaces?*