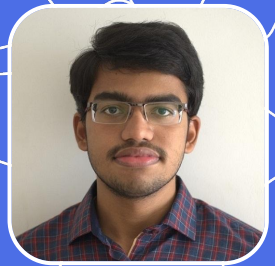

MULTIMODAL AI
SESSION #1
21st June 2024

About Us

Henry Vo
@_lowkeyboi

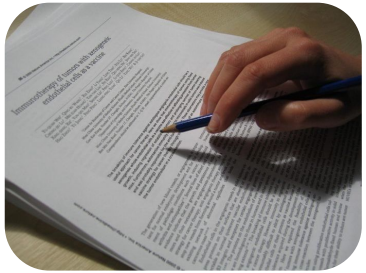


Surya Guthikonda
@suryaguthikonda

What to expect?

#multimodal-ml - Channel
@multimodal - Role

NEWSLETTER



Monthly Newsletter
- Industry
- Research
- Learning Resources

First Monday

Paper Reading Session

Guest Speaker Session

Live Coding Session

Second and Fourth Friday 10 AM ET

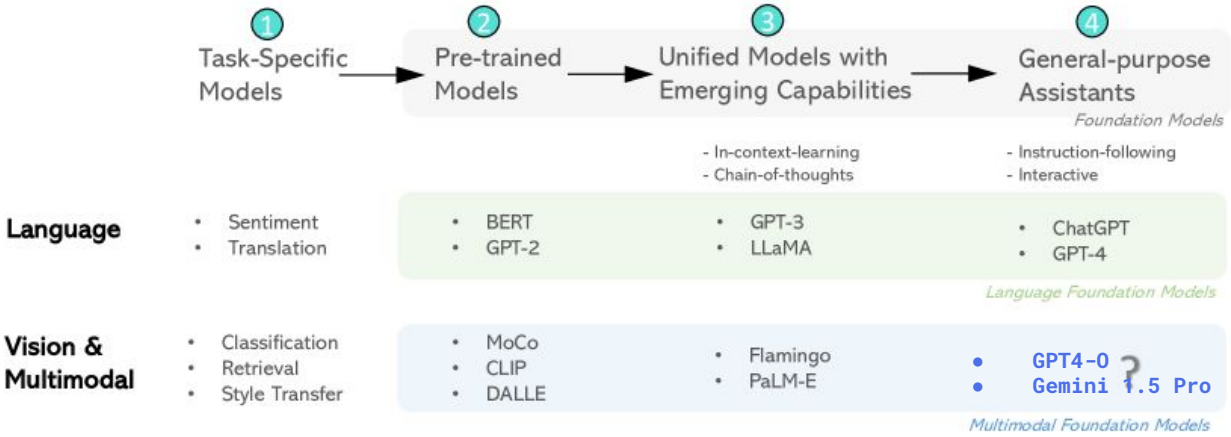
And More...

Multimodal Foundation Models: From Specialists to General-Purpose Assistants

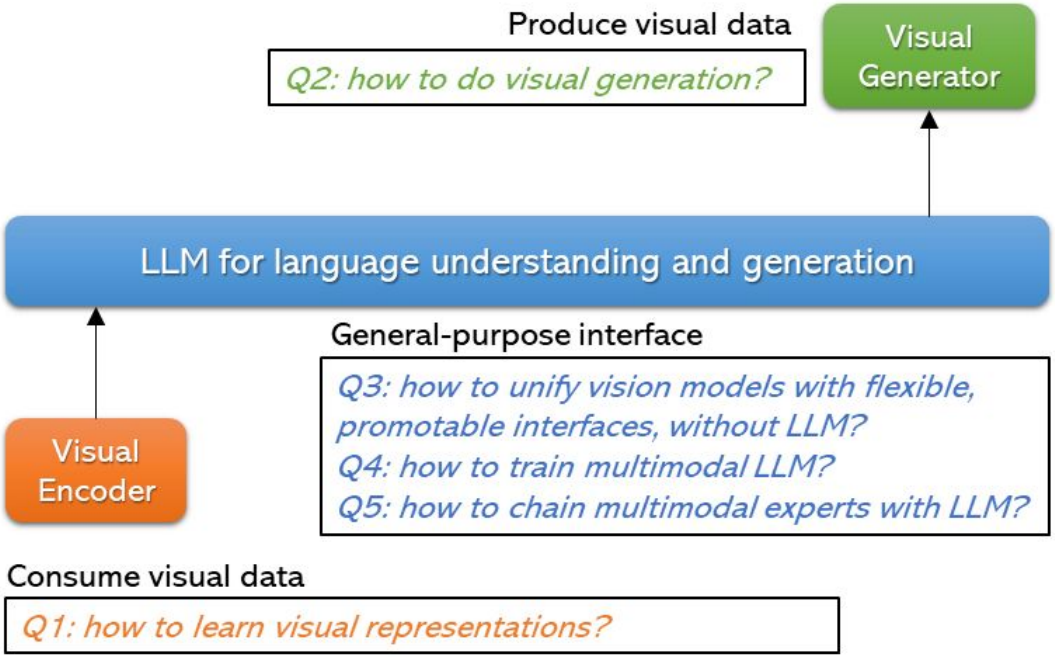
Source:CVPR 2023

Introduction

- General Purpose Assistant excelling in Vision has been a long-standing goal.
- Visual Prompts (user uploaded images, human-drawn clicks,sketches and masks)



Aim of Multimodal Models



Specialists Vs General-Purpose Assistants

Specialists

Visual Understanding Models

Visual Generation Models

General-Purpose Assistants

Unified Architecture - Different Tasks

Human Instruction Following

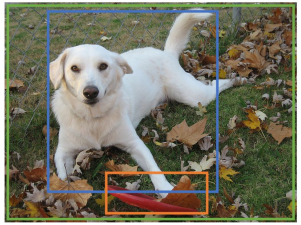
VISUAL UNDERSTANDING

Vision Tasks

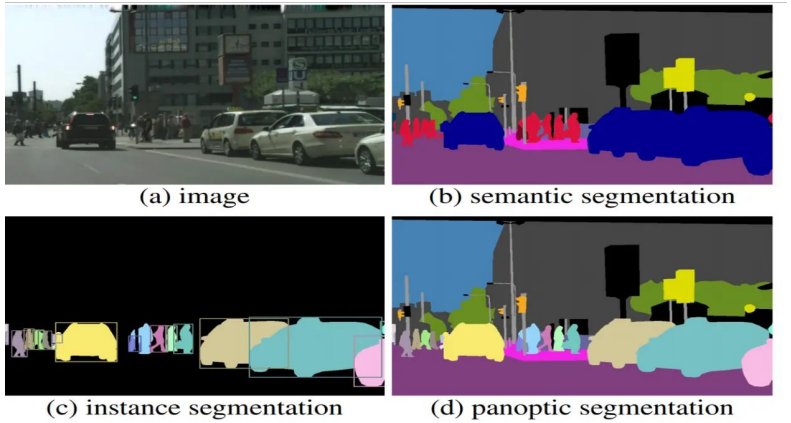
Image Level
- Image Classification
- Image-Text Retrieval

Region Level
- Object Detection
- Phrase Grounding

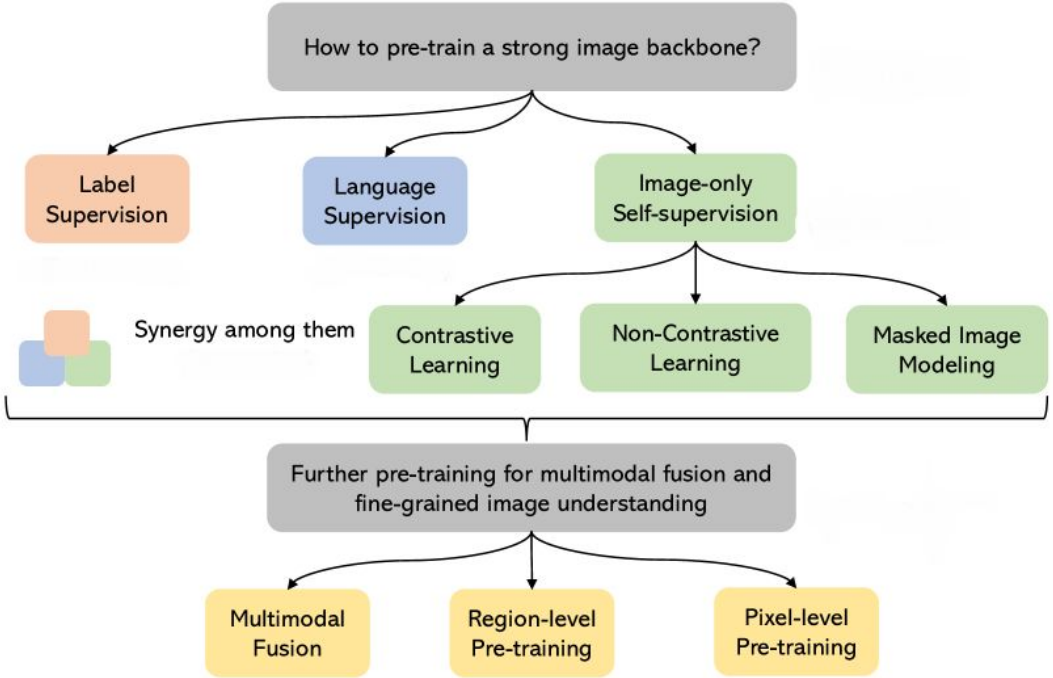
Pixel Level
- Semantic Segmentation
- Instance Segmentation
- Panoptic Segmentation



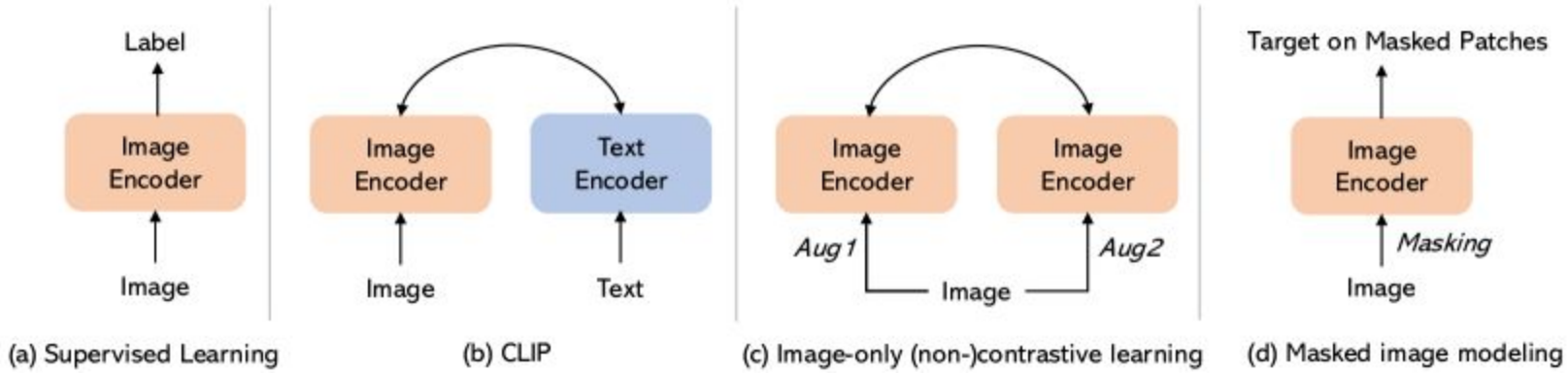
A dog is lying on the grass next to a frisbee.



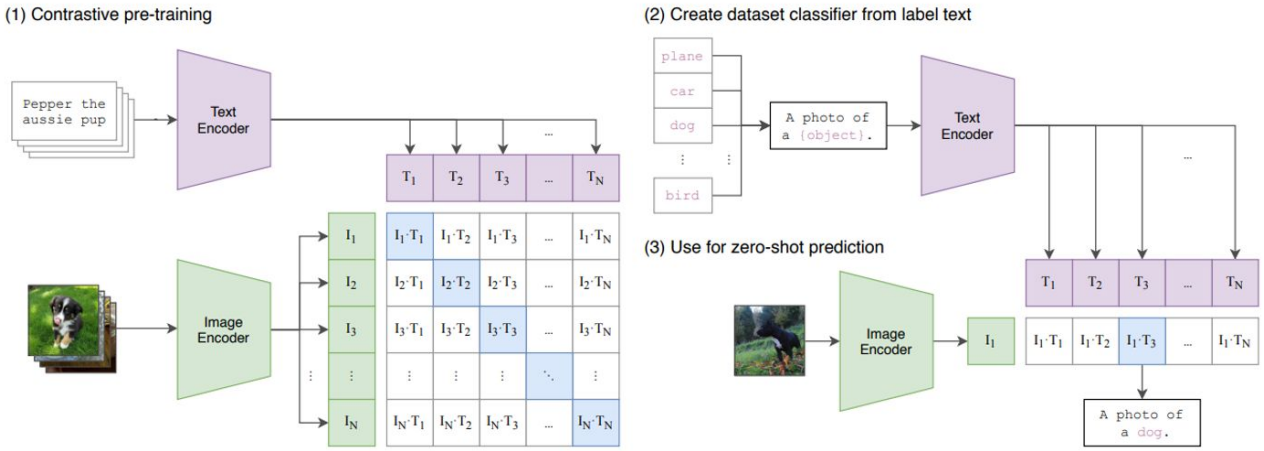
Overview


How to pre-train a strong image backbone?
Label Supervision
Language Supervision
Image-only Self-supervision
Synergy among them
Contrastive Learning
Non-Contrastive Learning
Masked Image Modeling
Further pre-training for multimodal fusion and fine-grained image understanding
Multimodal Fusion
Region-level Pre-training
Pixel-level Pre-training

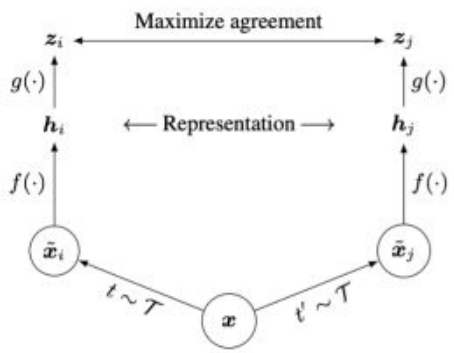
Overview


Label
Image Encoder
Image
Image Encoder
Text Encoder
Image
Text
Image Encoder
Image Encoder
Aug 1
Aug 2
Image
Target on Masked Patches
Image Encoder
Masking
Image
(a) Supervised Learning
(b) CLIP
(c) Image-only (non-)contrastive learning
(d) Masked image modeling

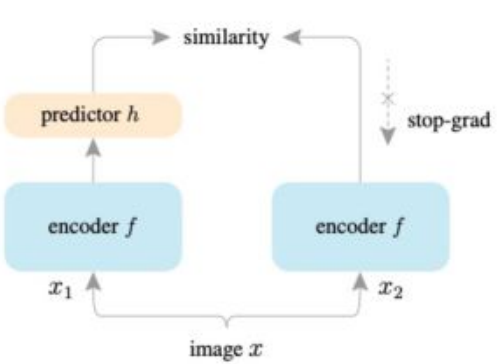
Contrastive Image-Language Pre-training (CLIP)



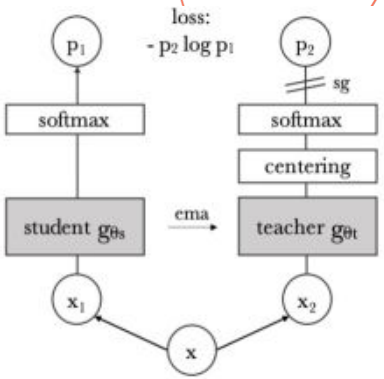
(Non)Contrastive Learning



(a) SimCLR

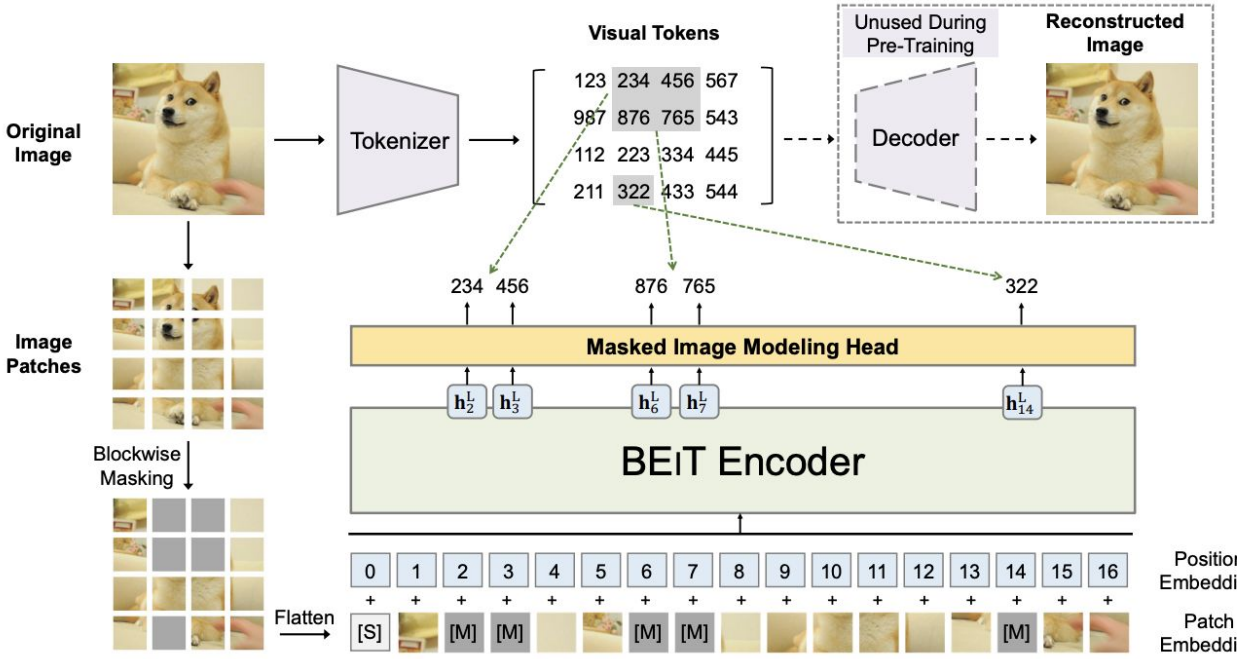


(b) SimSiam



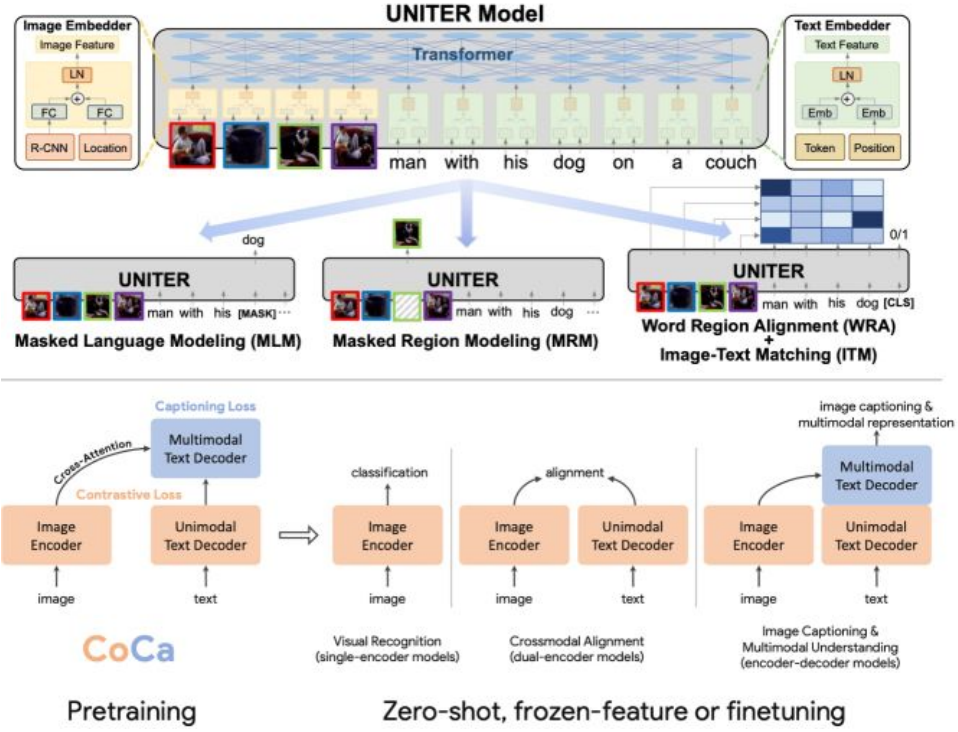
(c) DINO

Masked Image Modelling



Bidirectional Encoder representation from Image Transformers

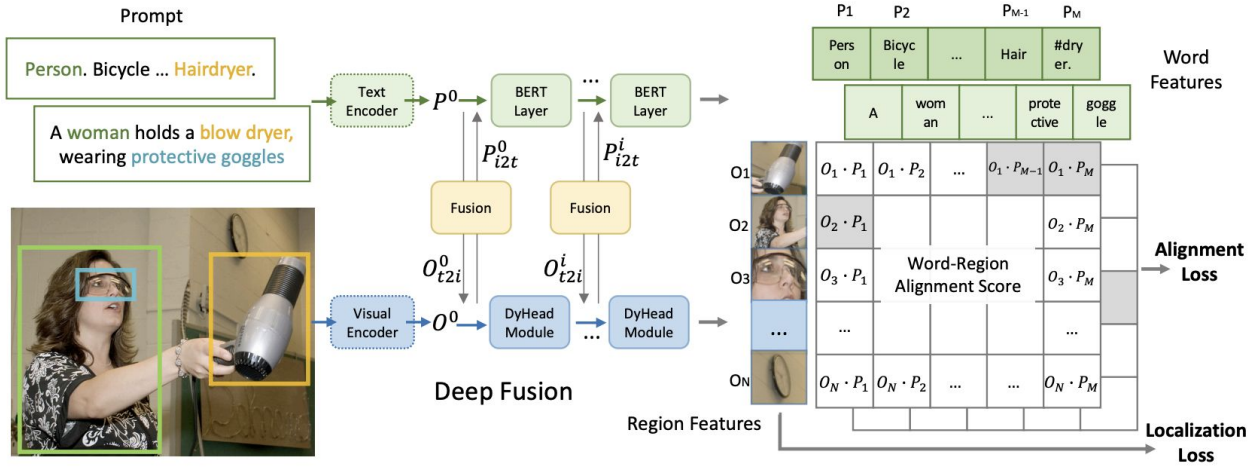
Multimodal Fusion



UNiversal Image-TExt Representation Learning

Contrastive Captioner

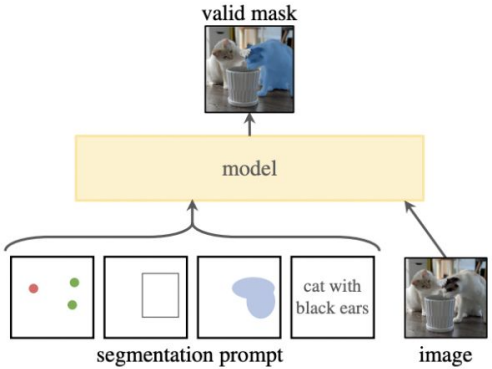
Region Level Pre-training (GLIP)



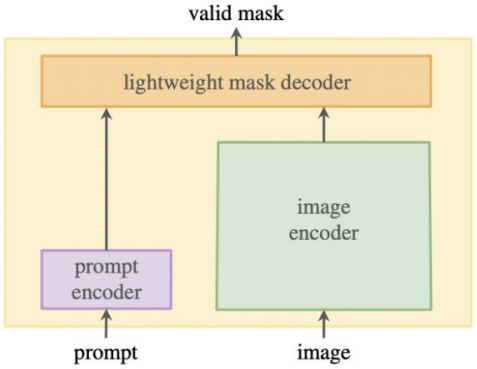
Grounded Language Image Pretraining

Unifies Object Detection and Phrase Grounding

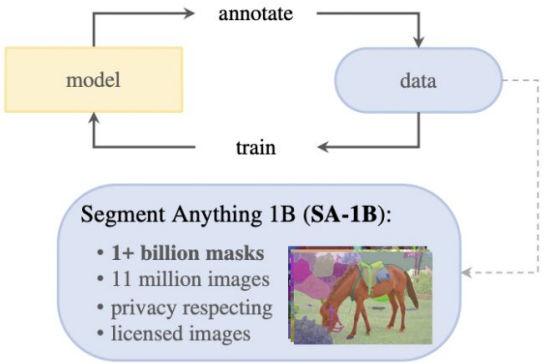
Pixel Level Pre-training (SAM)



(a) **Task**: promptable segmentation

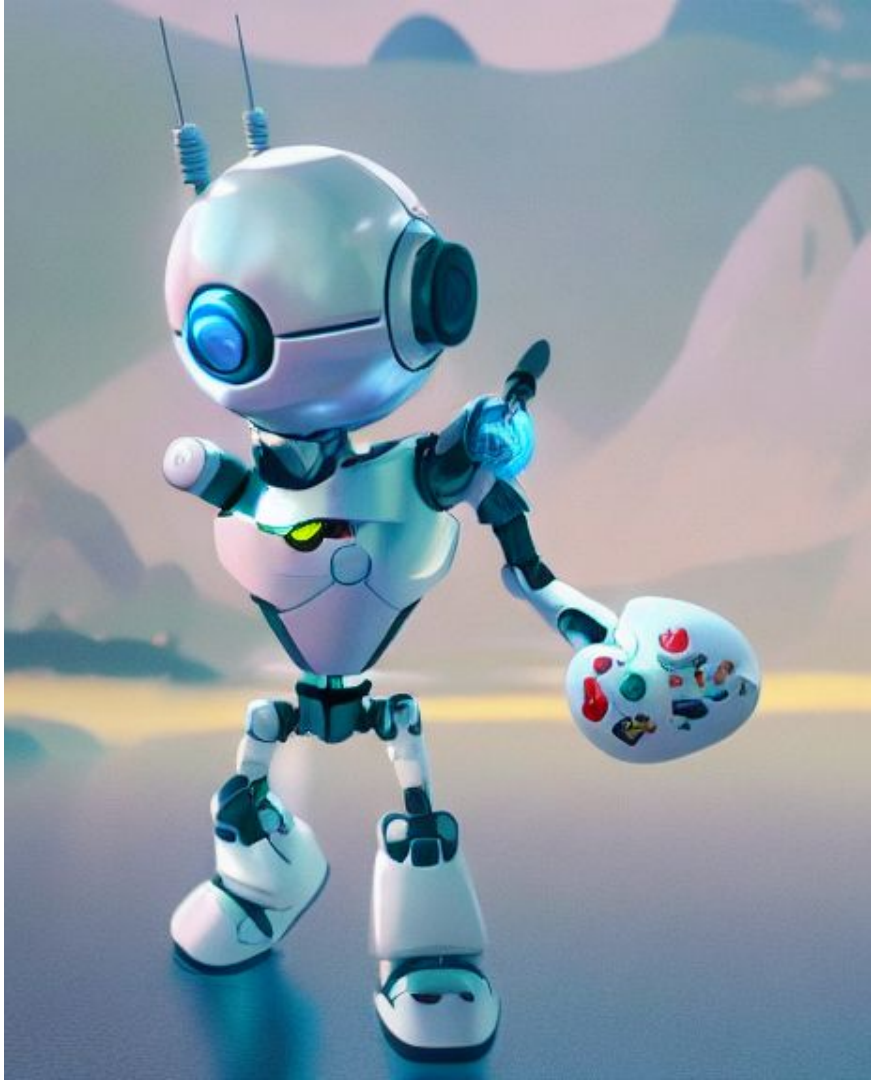


(b) **Model**: Segment Anything Model (**SAM**)

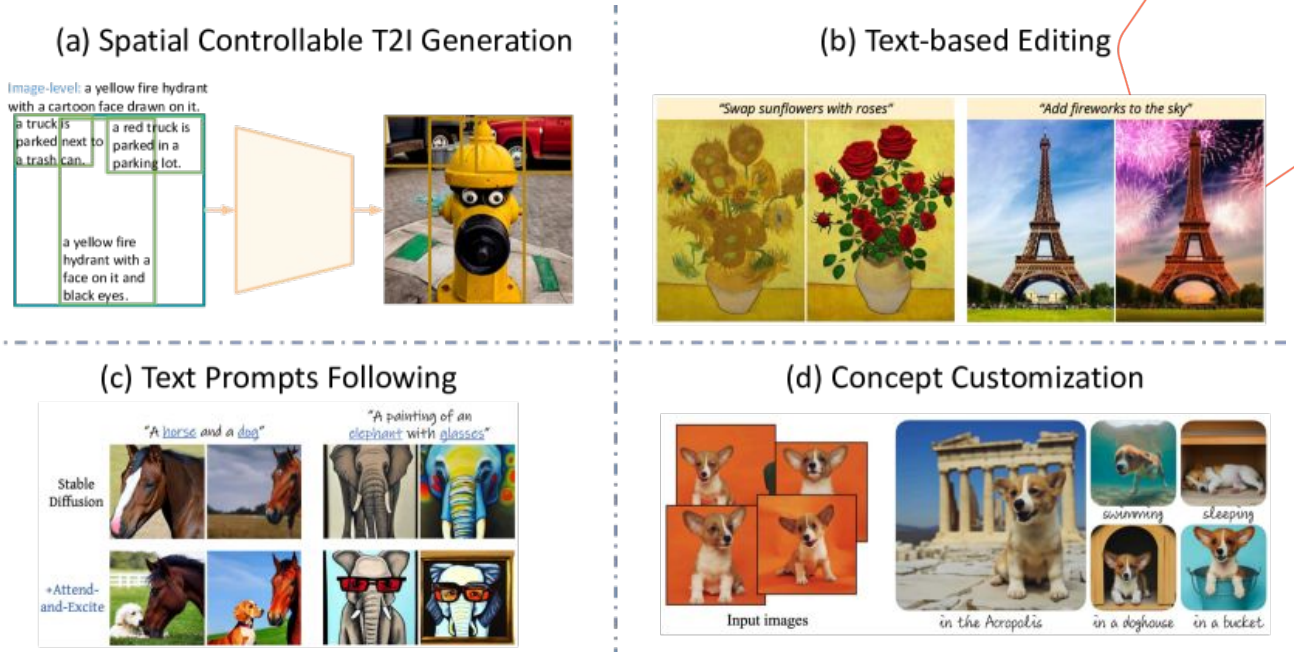


(c) **Data**: data engine (top) & dataset (bottom)

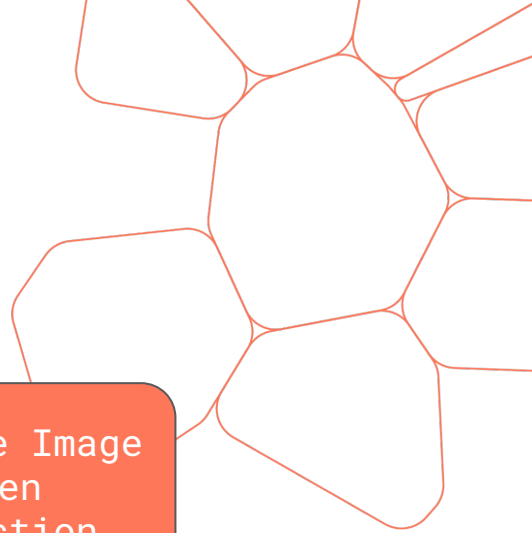
VISUAL GENERATION



Human Alignment in Visual Generation



Text To Image Generation



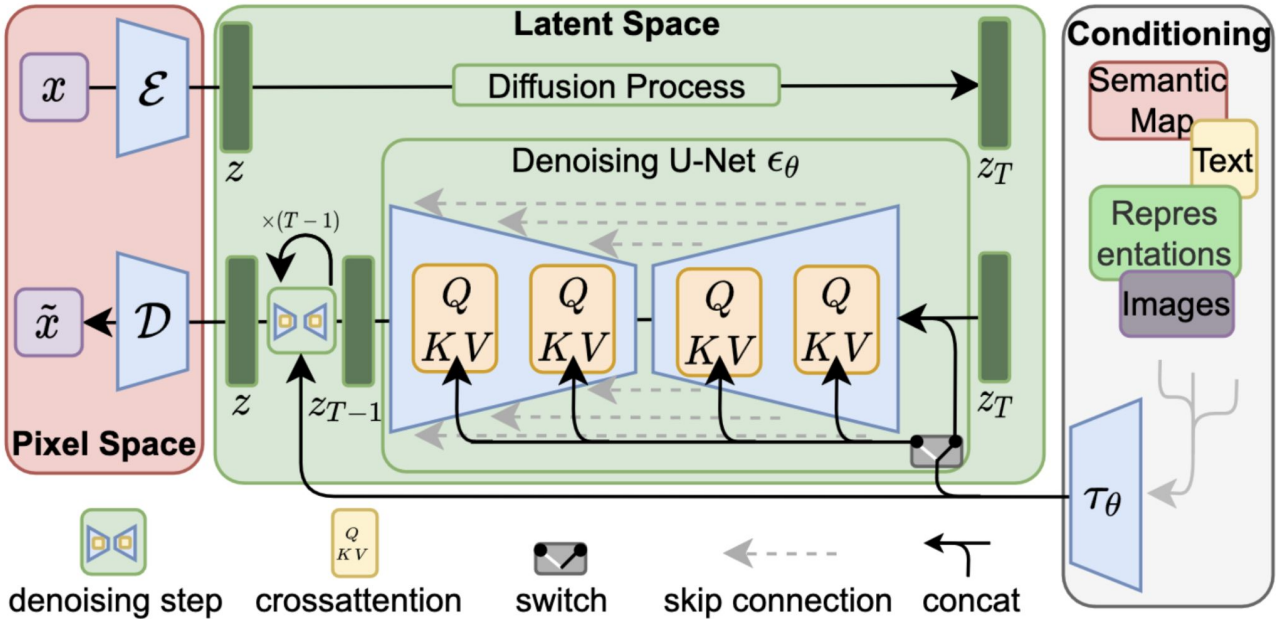
Generative Adversarial Networks

Variational Autoencoders

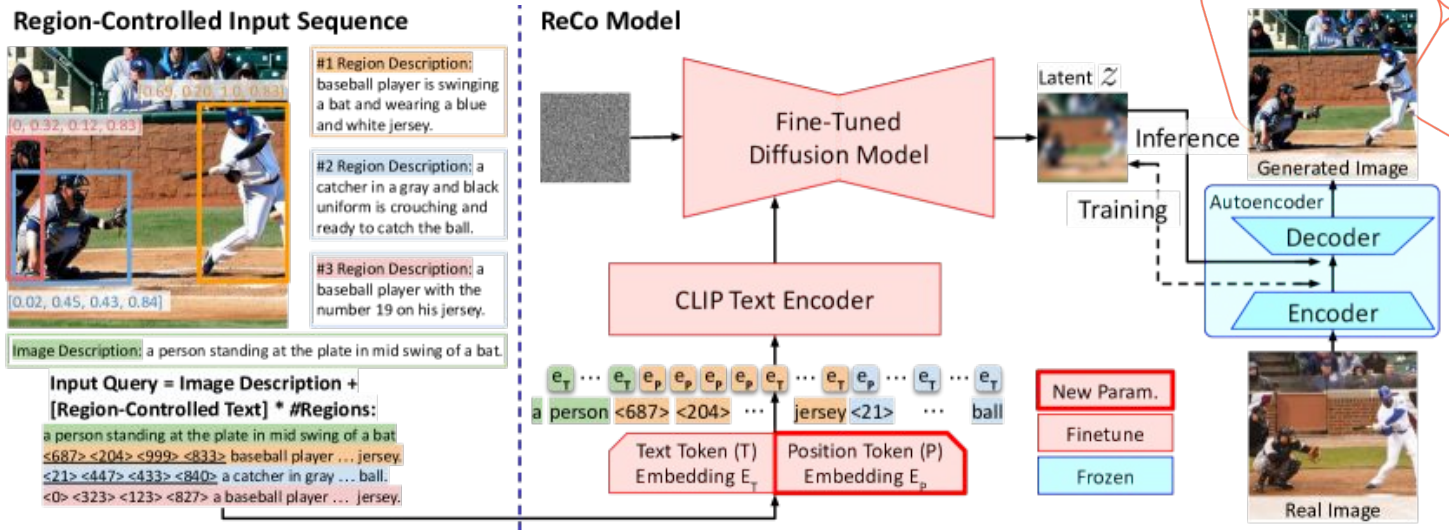
Discrete Image Token Prediction

Diffusion Model

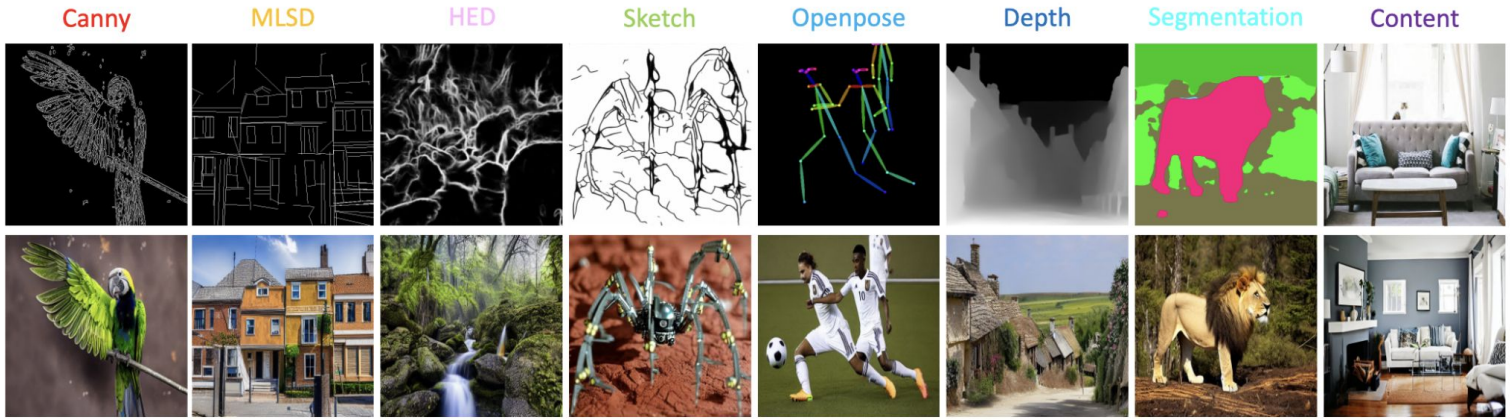
Stable Diffusion



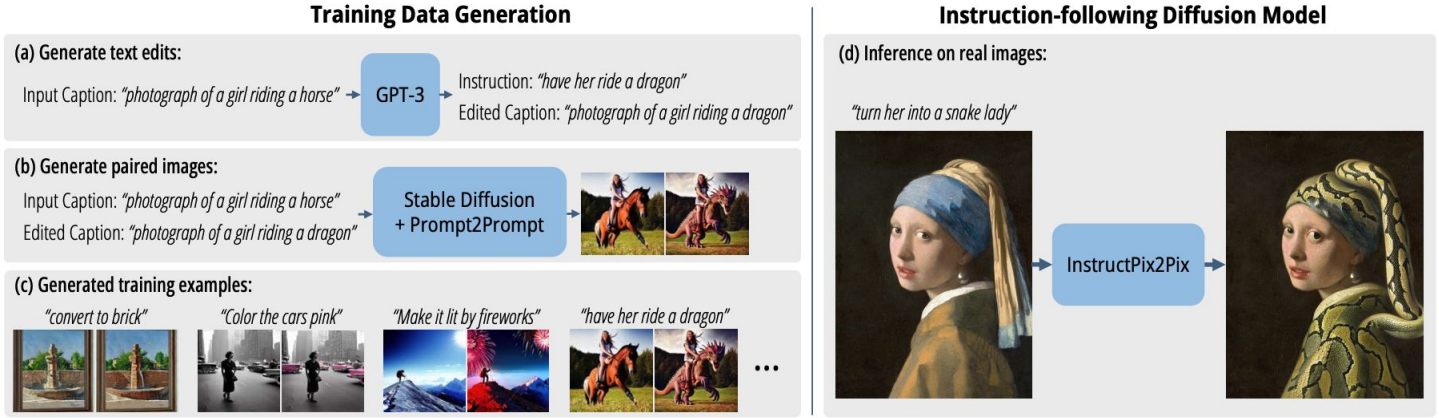
Region-controlled T2I Generation



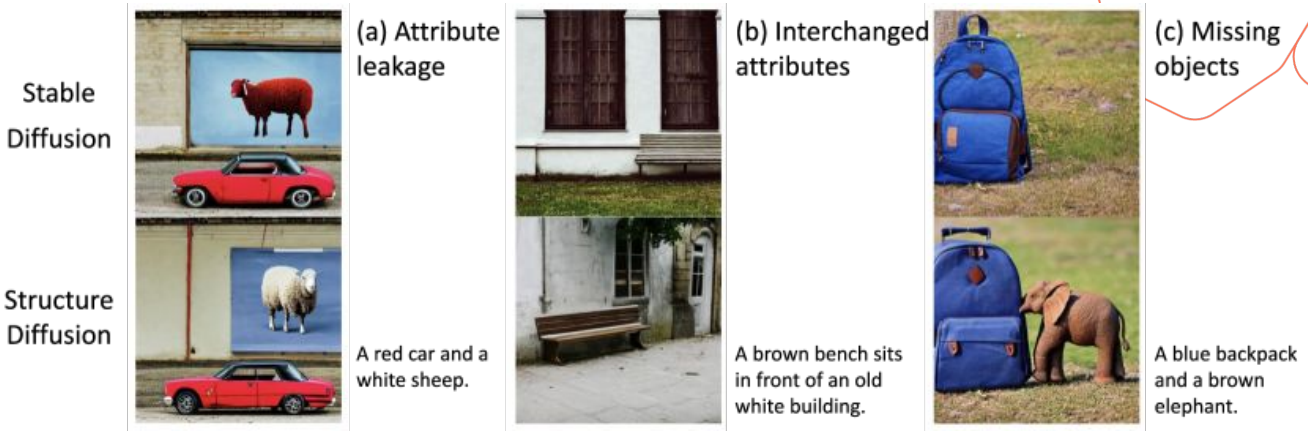
Dense Conditions T2I Generation



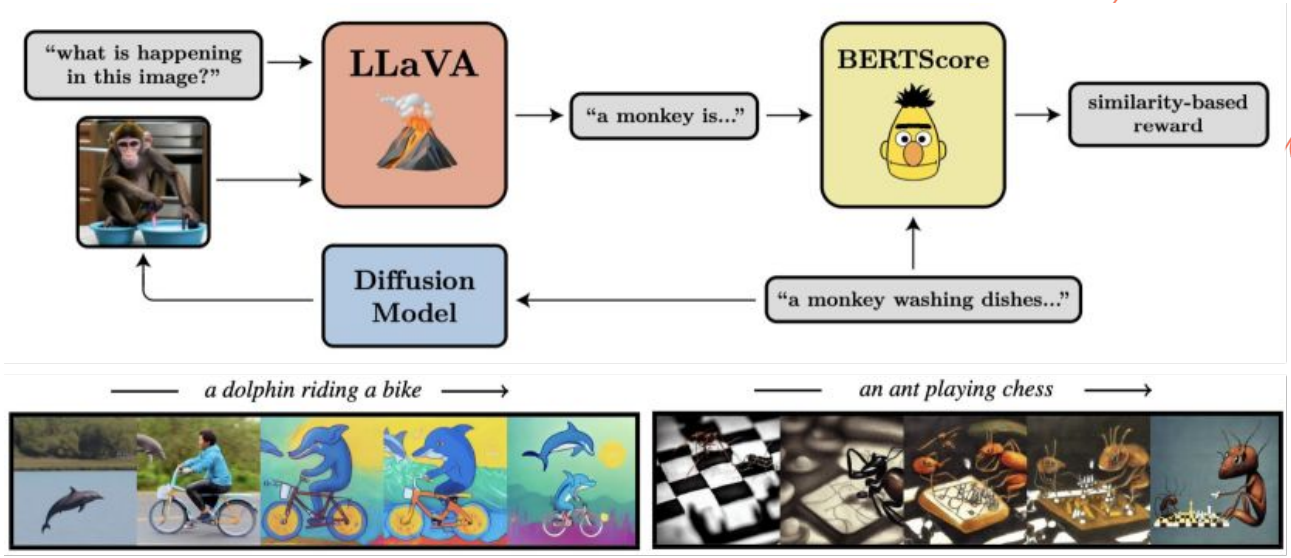
Text Instruction Editing



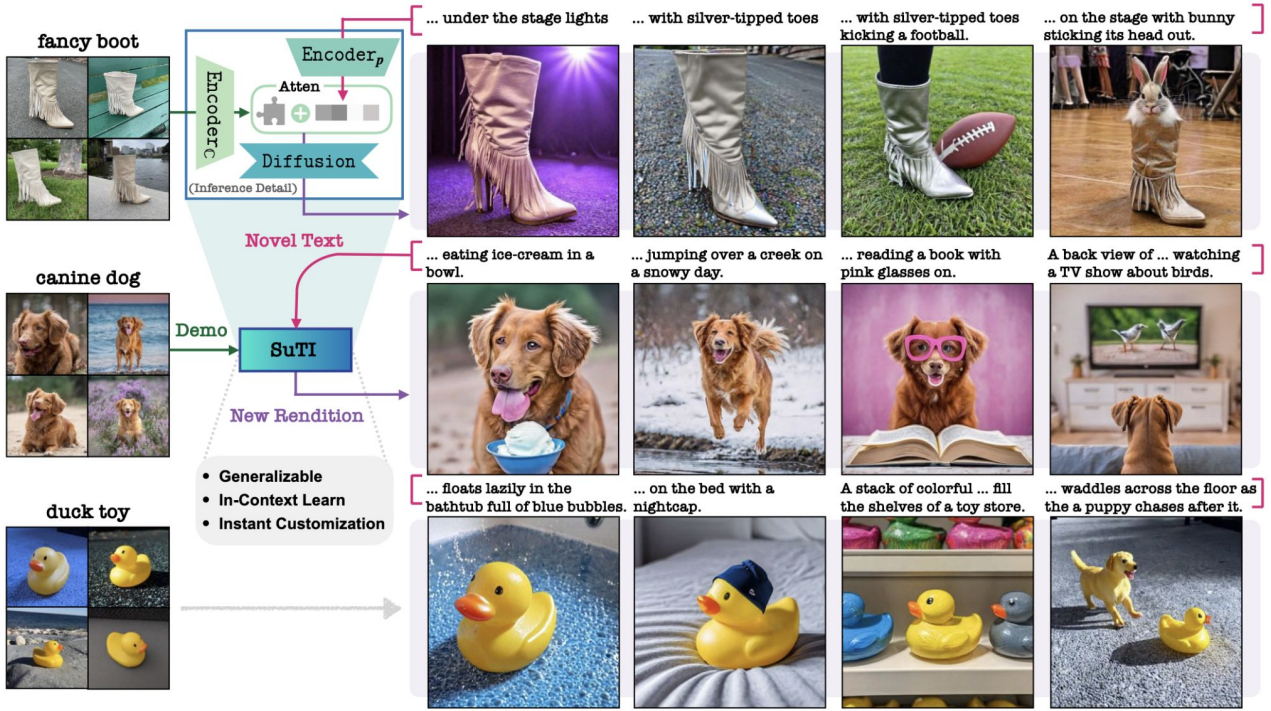
Failure of Vanilla T2I Models



Alignment Tuning



Concept Customization

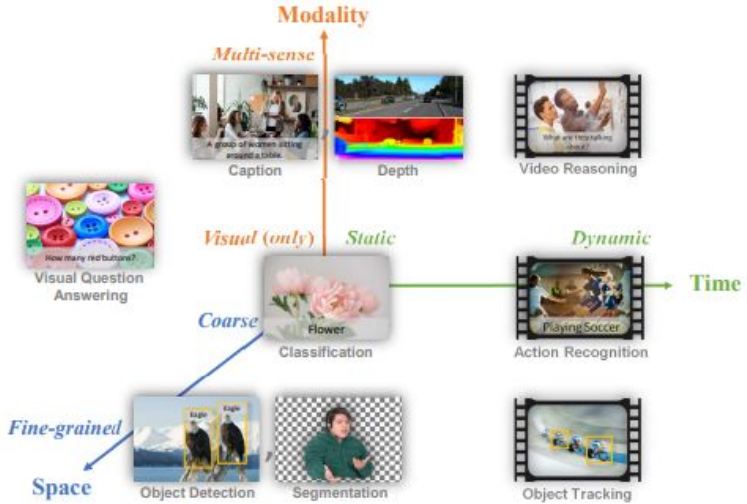


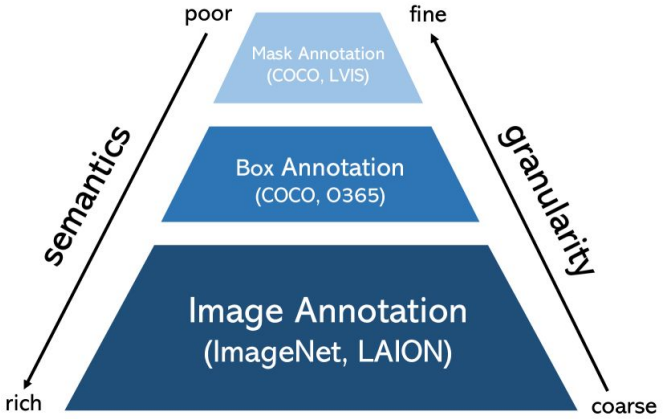
Subject Driven
Text to Image
Generation

UNIFIED VISION MODELS

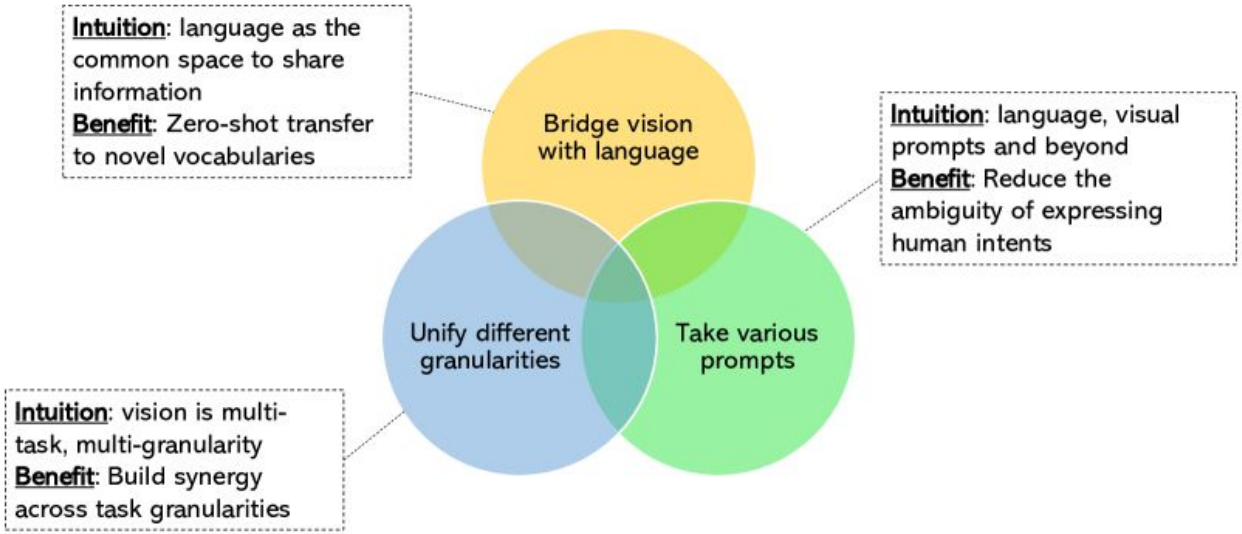
Challenges of Unification

- Varied Inputs and Outputs
- Different Granularities
- Cost of Annotation/Collection





Towards Unified Vision Model



From Closed Set to Open-Set Models

Model Initialization

CLIP Initialized

CLIP Augmented

Model Design

Two-stage models

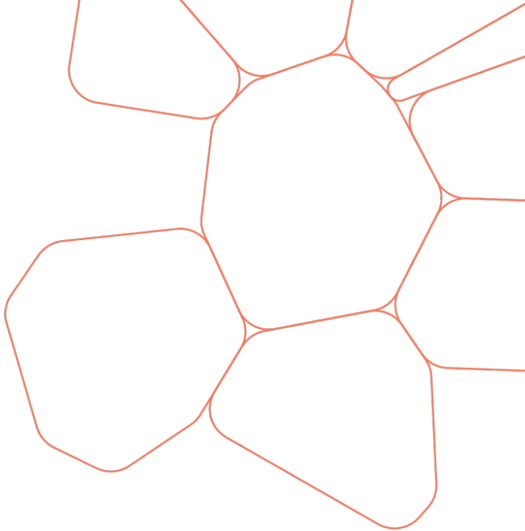
End-to-end models

Model pre-training

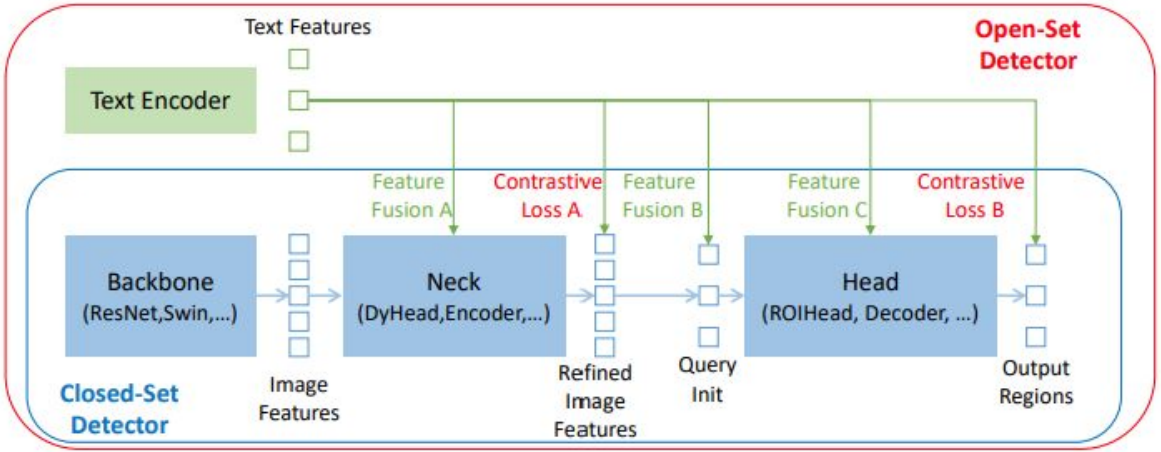
Supervised

Semi-Supervised

Weakly Supervised

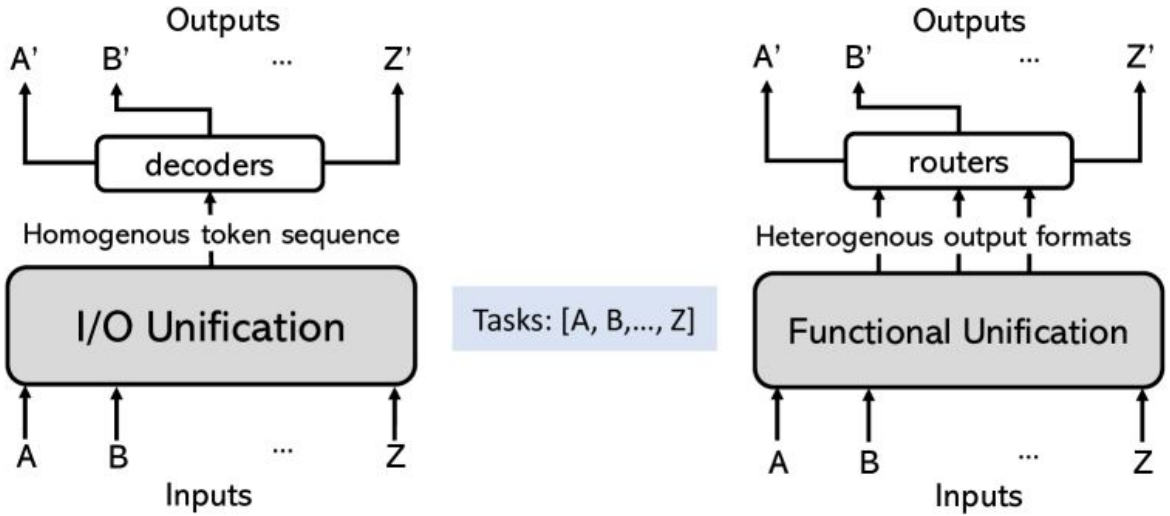


Object Detection Grounding

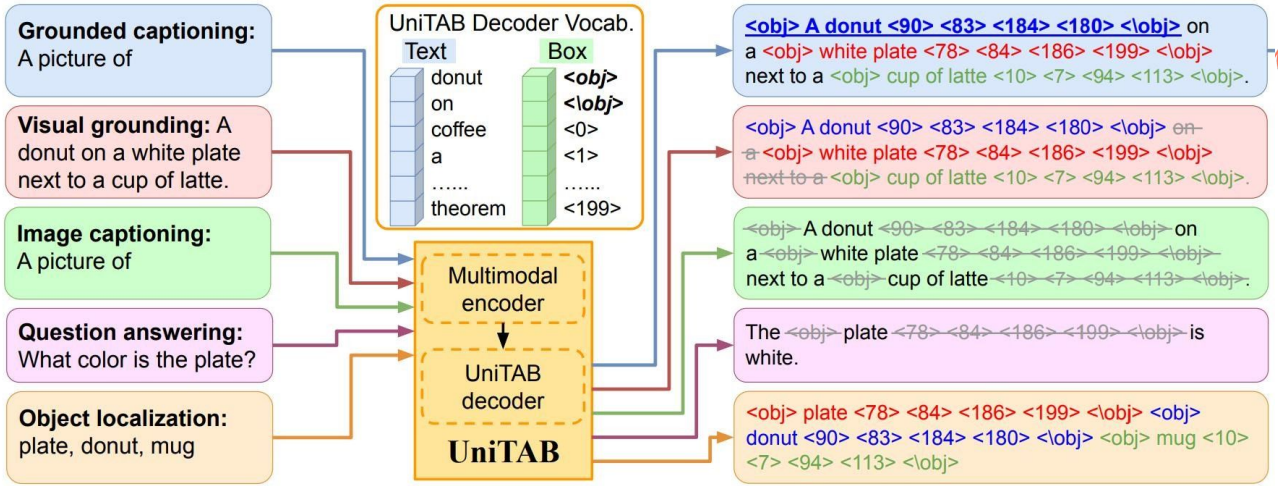


Grounding-DINO

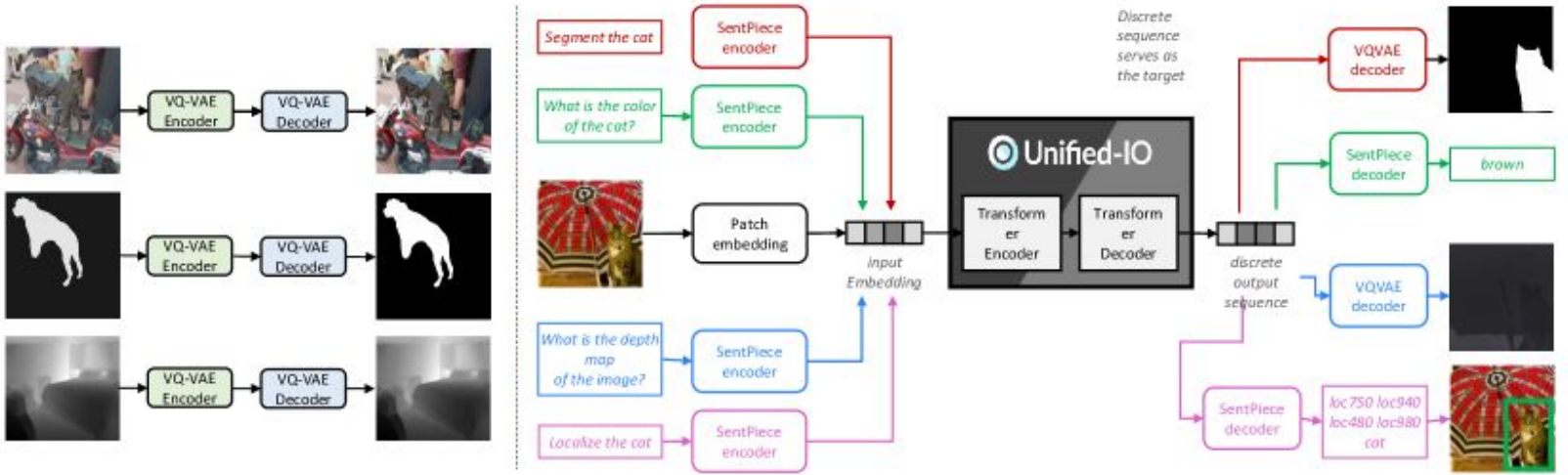
Generic Models



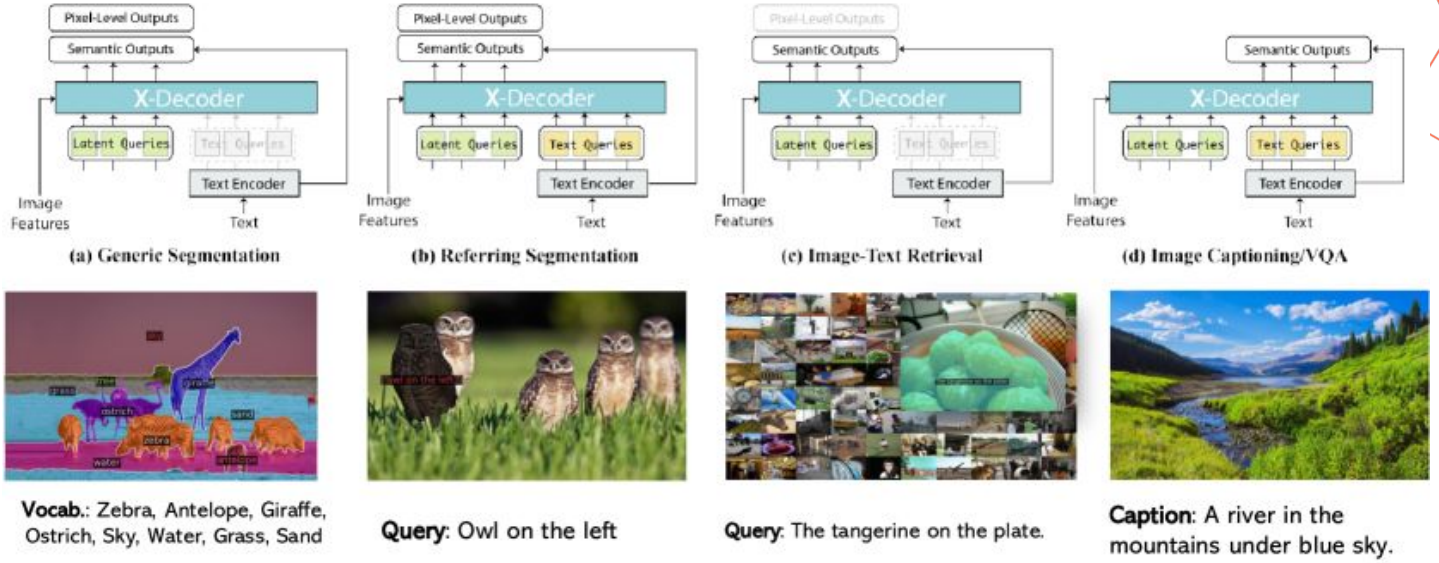
I/O - Sparse and Discrete Outputs



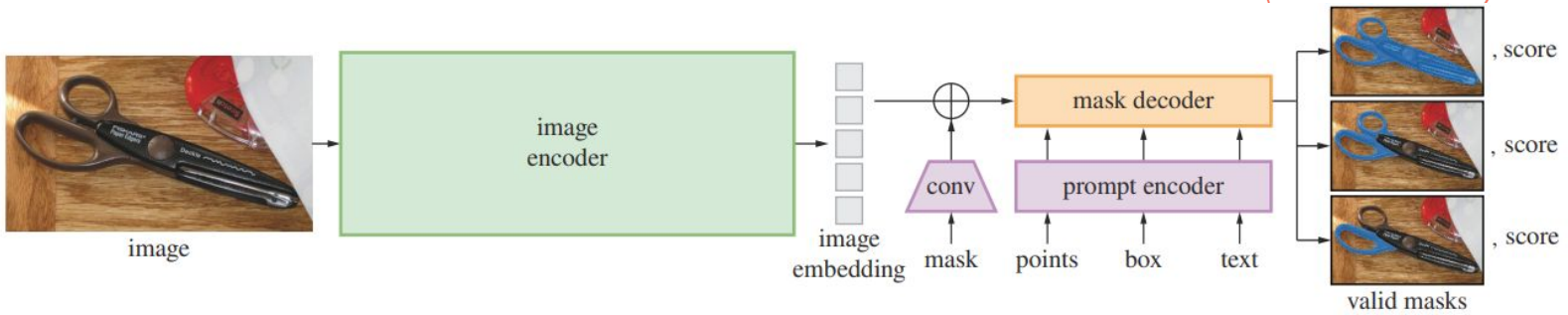
I/O - Dense & Continuous Outputs



Functionality - Unified Learning

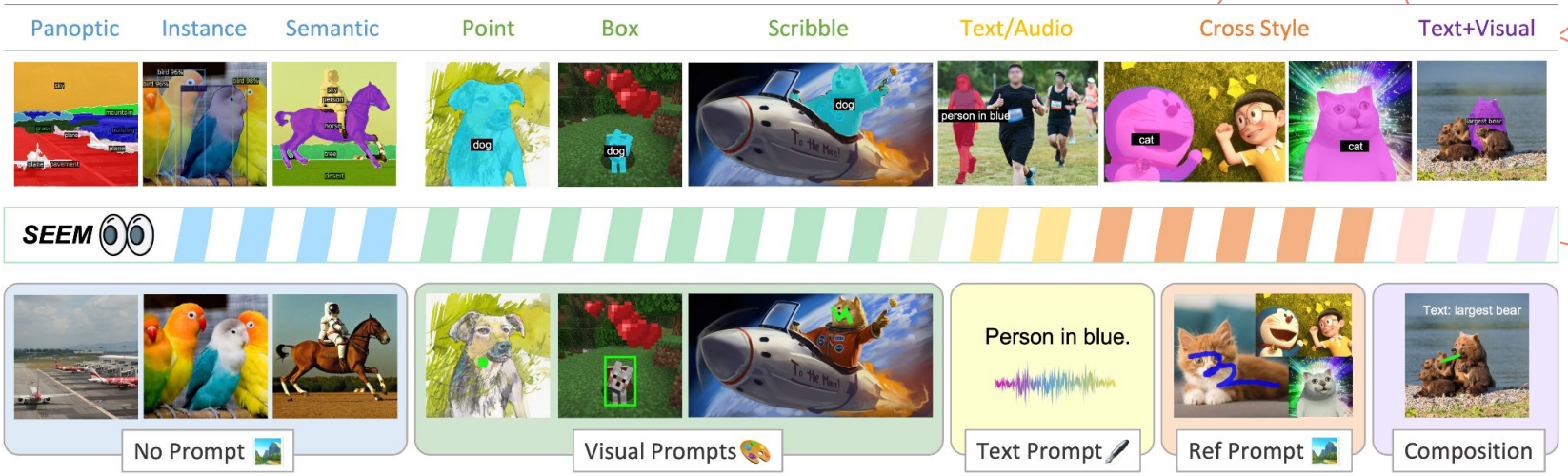


Spatial Prompting



SAM (Segment Anything Model)

Visual Prompting



SEEM (Segment Everything Everywhere All at once)

In-context Prompting

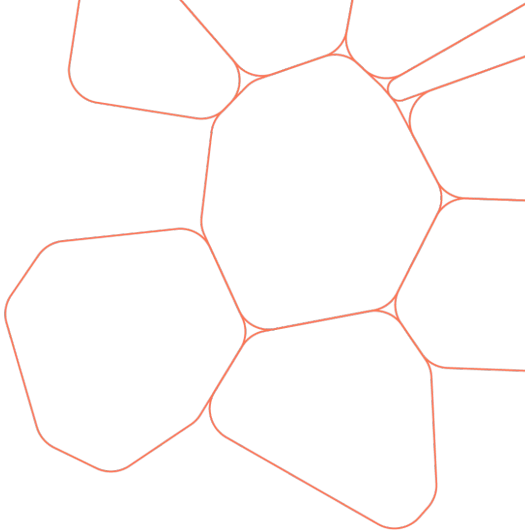
HummingBird (Nearest Neighbour Retrieval)

Open Questions

- Computer Vision in the wild
- Scaling law in Vision
- Vision-centric or language-centric models

Next Week

- Large MultiModal Models: Training with LLMs
- Multimodal Agents: Chaining with Tools
- Conclusion and Research Trends
- What’s Changed Since the Paper Published?




THANK YOU :)