IMEN574/IDSC574 Programming for Data Science

18. **Vision Models II**

Spring 2026

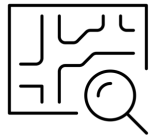
Instructor: **Sangwoo Mo**

Dense Perception

Abstract Perception
“what” (high-level, conceptual)



Dense Perception
“where” (low-level, spatial)



Visual

Semantic

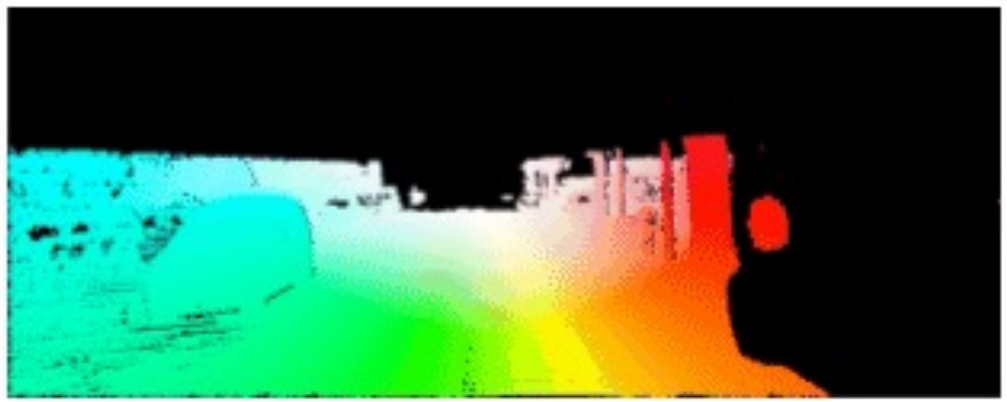
Self-Supervised Learning (e.g., DINO)

Dense Perception

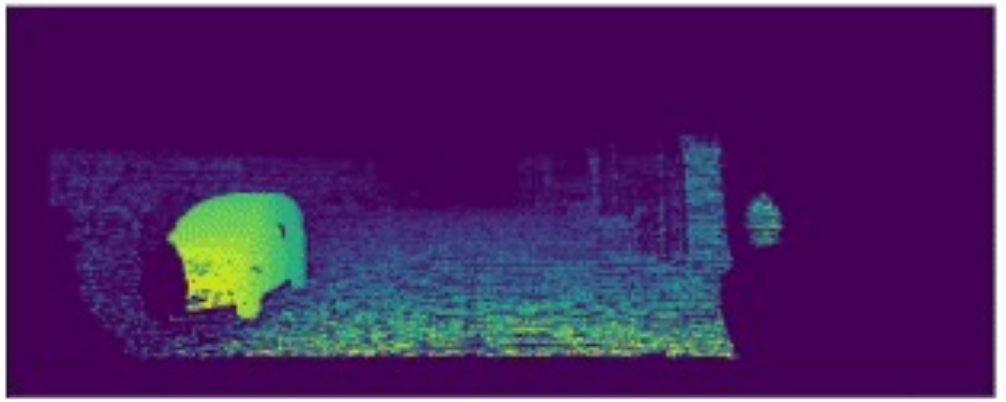


segmentation





optical flow



depth

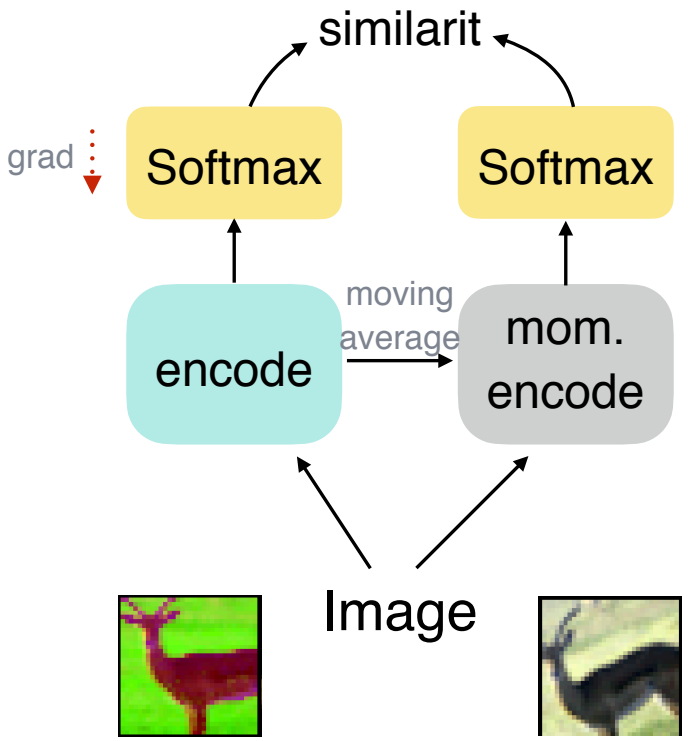
1. Spatialization of Abstract Perception



Figure 3: High-resolution dense features. We visualize the cosine similarity maps obtained with DINOv3 output features between the patches marked with a red cross and all other patches. Input image at 4096×4096. *Please zoom in*, do you agree with DINOv3?

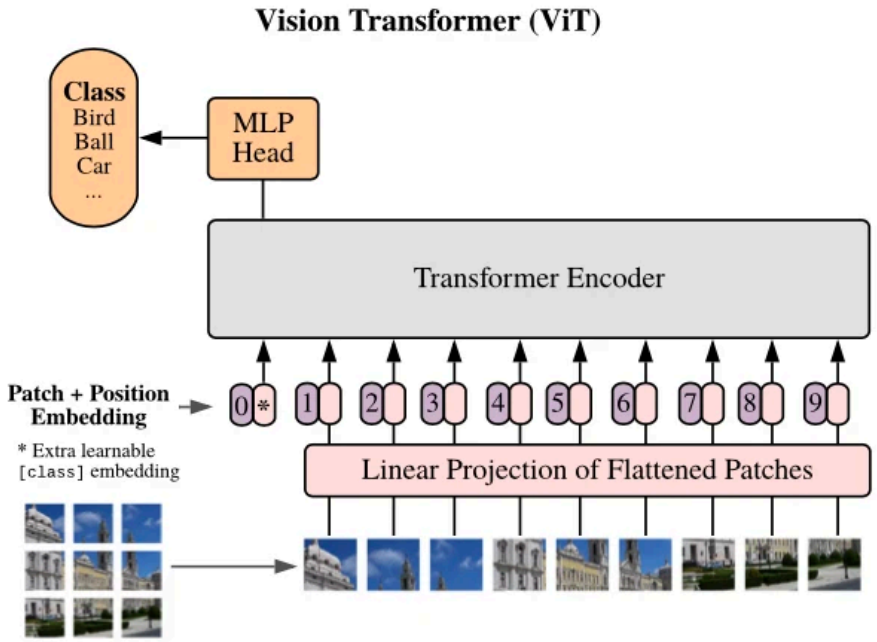
Grounding DINO

DINO - Main idea



Grounding DINO

Type of encoder - Vision Transformer



Vision Transformer - Dosovitskiy et al., 2020

Grounding DINO

No pooling!



Feature maps in CNN

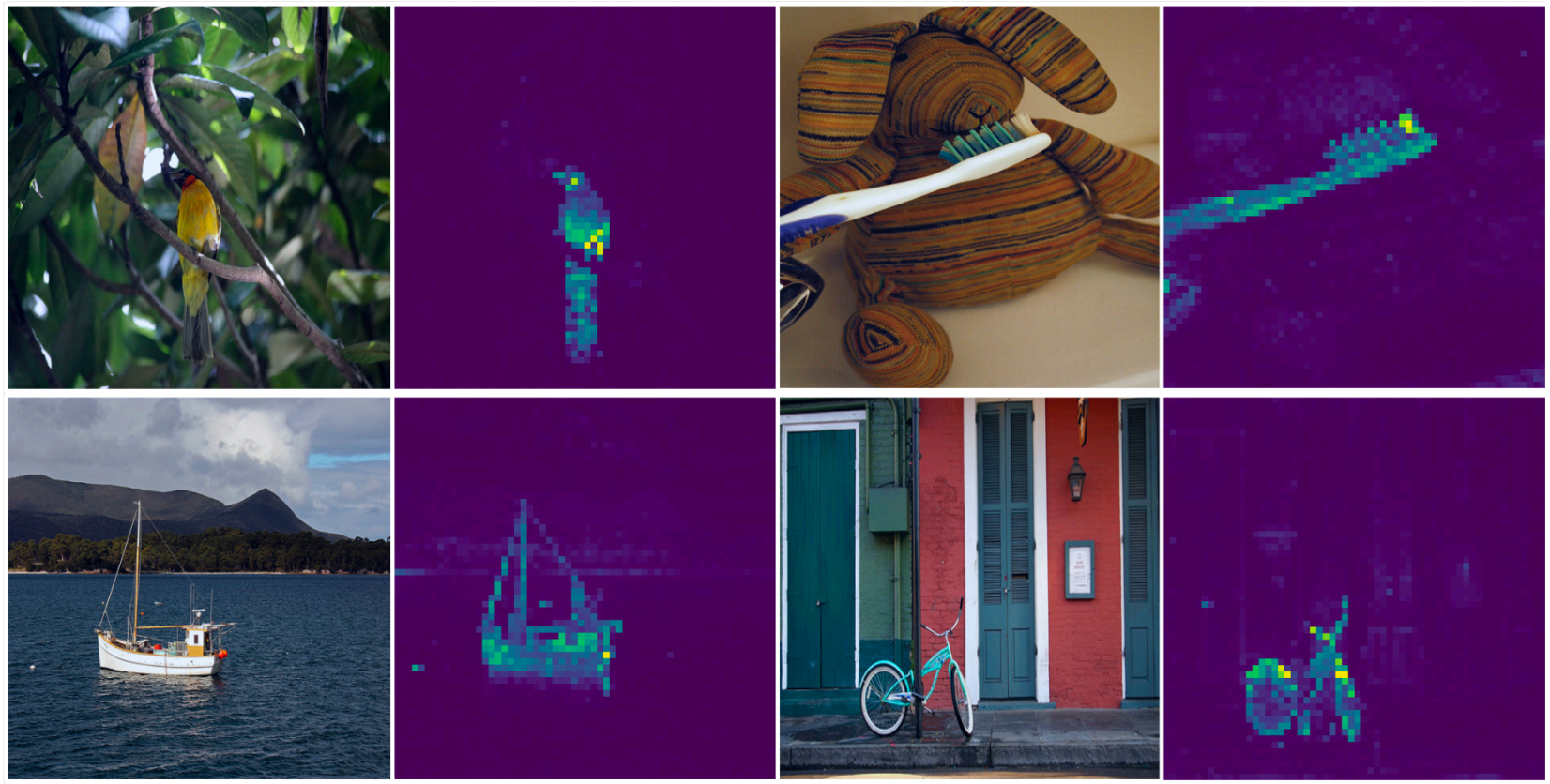


Feature maps in ViT

Source: https://towardsdatascience.com/vision-transformers-or-convolutional-neural-networks-both-de1a2c3c62e

Grounding DINO

Segmentation emerges!



Visualize the “CLS” token attention.

Note that the CLS token or the network are not supervised

Grounding DINO

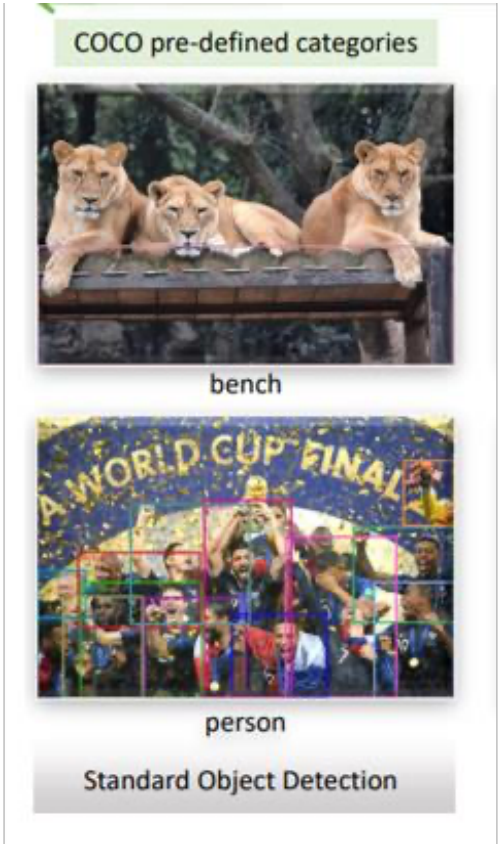
Segmentation across different heads



Grounding DINO

Closed-set Detection

Detect objects of pre-defined classes



Open-set Detection

Detect objects that were not present in training data



Grounding DINO

Introduction

- Strong Open-set object detector which combines transformer-based detector DINO with grounded pre-training
- Introduces language for unseen object generalization.
- Detects arbitrary objects with human inputs (Ex: category or referring expressions)

Grounding DINO

Grounding DINO

Closed-set Detection + Language → Open-set Detection

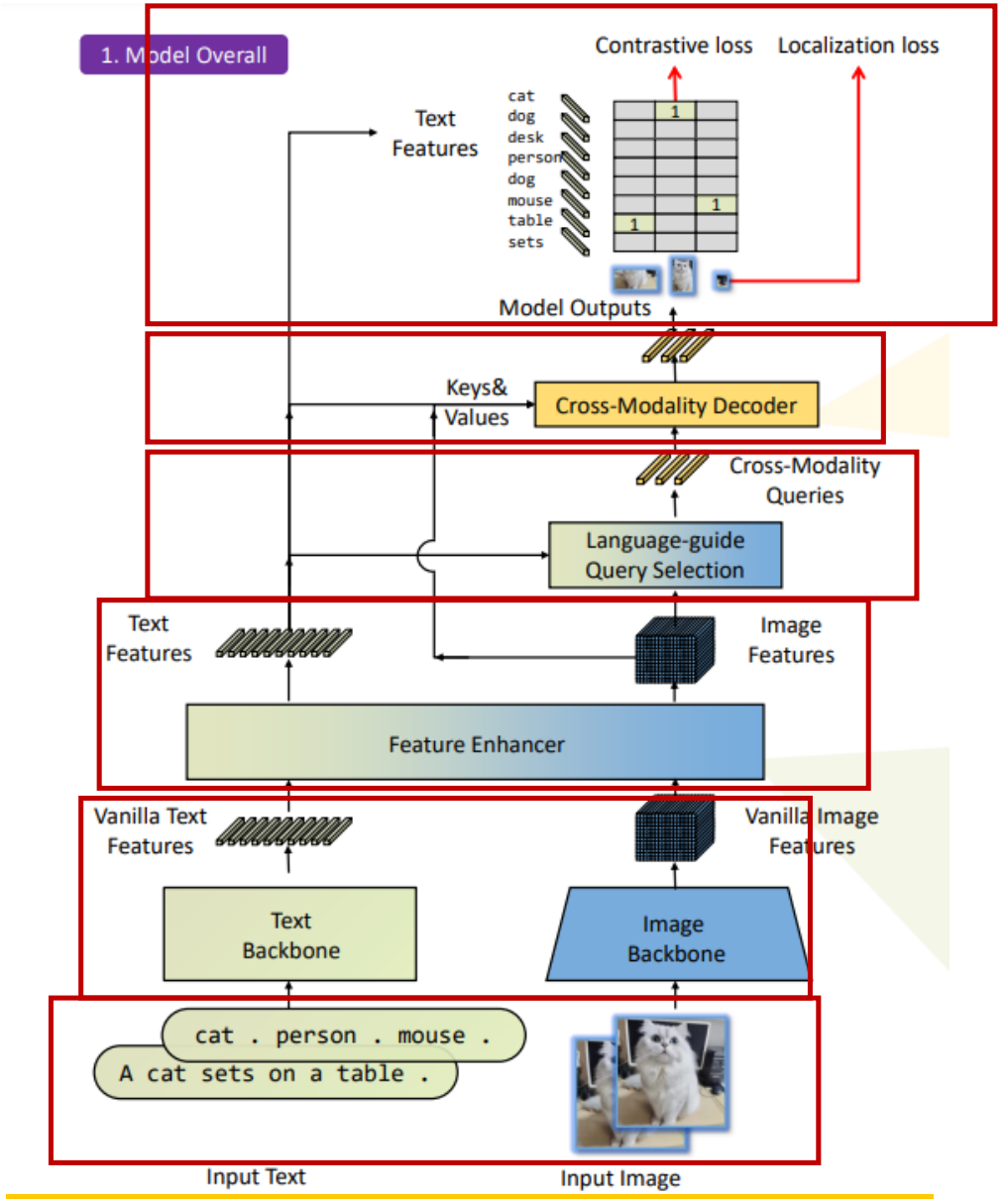
Transformer Based Detector (DINO) + Grounded pre training → Grounding DINO

DINO - DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection

DINO - Emerging Properties in Self-Supervised Vision Transformers

Grounding DINO

- Dual-encoder-single-decoder architecture
- Key Components :
  - Image Backbone
  - Text Backbone
  - Feature Enhancer
  - Language-Guided Query Selection
  - Cross-Modality Decoder



2. Dense Supervision (e.g., SAM)



Carion et al. SAM 3: Segment Anything with Concepts. arXiv 2025.

2. Dense Supervision (e.g., SAM)

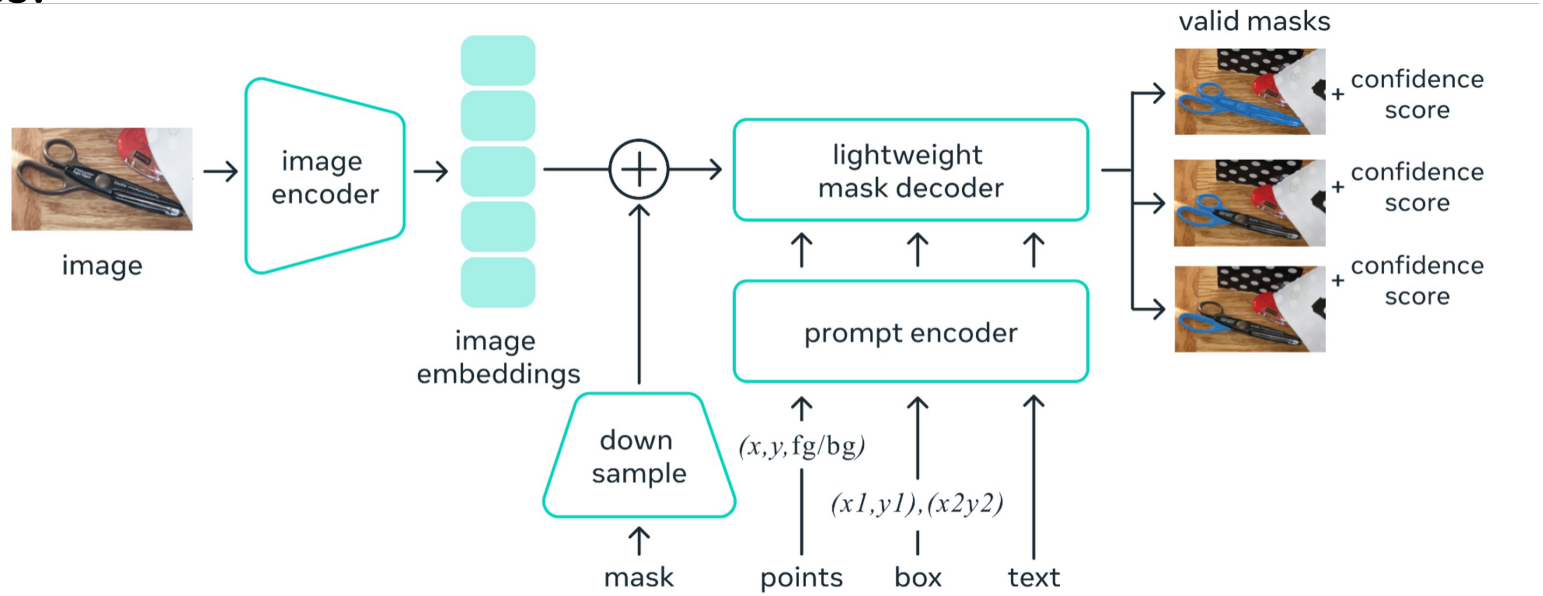


Lin et al. Depth Anything 3: Recovering the Visual Space from Any Views. arXiv 2025.

Segment Anything

Model: Segment Anything Model (SAM)

- A heavyweight ***image encoder*** outputs an image embedding.
- A lightweight ***prompt encoder*** efficiently queries the image embedding.
- A lightweight ***mask decoder*** produces object masks and confidence scores.

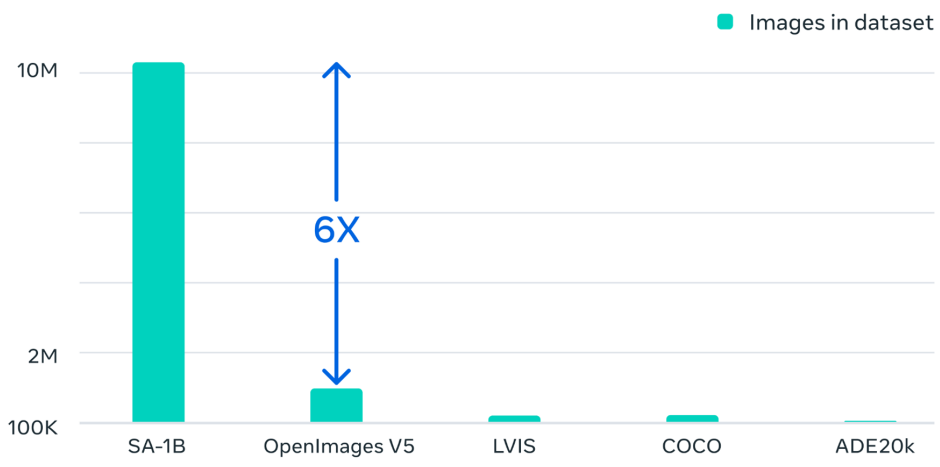


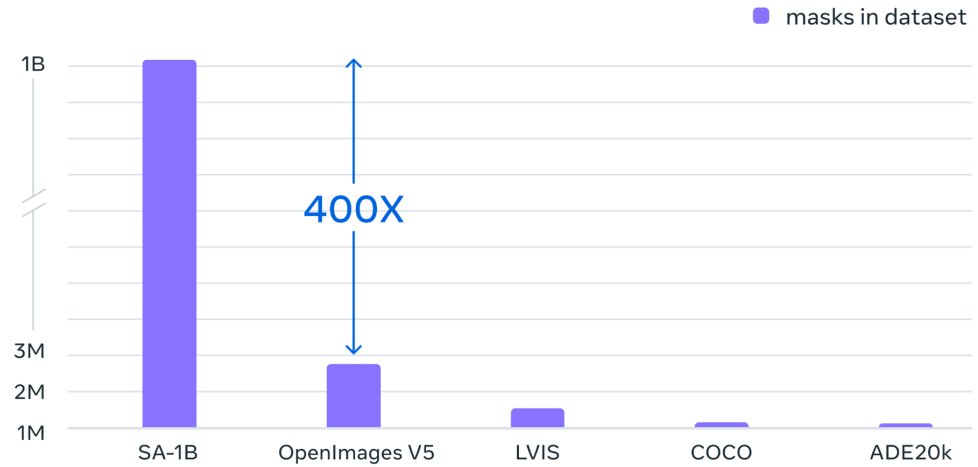
Slide Credits: Qin Liu

Segment Anything

Dataset: SA-1B

- Built with a SAM model in the loop
- 11M images with 1.1B segmentation masks
- 400x more masks than any prior segmentation dataset

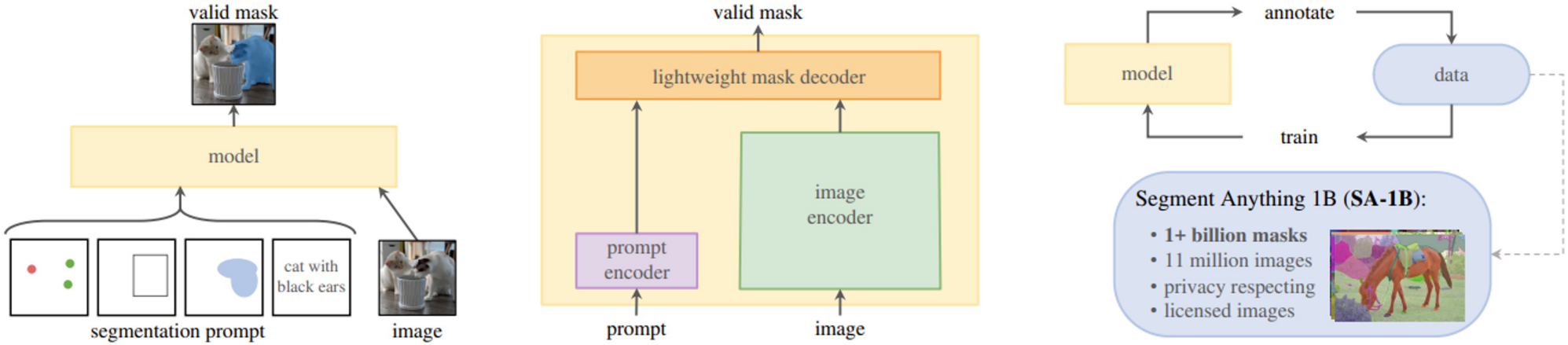




Slide Credits: Qin Liu

Segment Anything

SAM is built with three interconnected components: A task, an model, and a data engine.



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

Slide Credits: Qin Liu

Grounded SAM

Grounded-Segment-Anything



Text Prompt: Bench



Grounded-SAM Output



Stable-Diffusion Inpainting
A Sofa, high quality, detailed

Grounded SAM

Background Matting: The World is Your Green Screen

Soumyadip Sengupta
Vivek Jayaram
Brian Curless
Steve Seitz
Ira Kemelmacher-Shlizerman

University of Washington

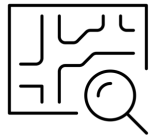


Original Video



Replaced Background

Progress So Far – Part II

	Abstract Perception “what” (high-level, conceptual) 	**Dense Perception** “where” (low-level, spatial) 
Visual	Self-Supervised Learning (e.g., DINO)	1. Spatialization of Abstract Perception 2. Dense Supervision (e.g., SAM)
Semantic		