SAM a great Semantic Segmentations

LLM to generate the reward function

Hantao Zhou

Universitaet Hamburg

January 21, 2024

Table of Contents

Multi-modal processing structure

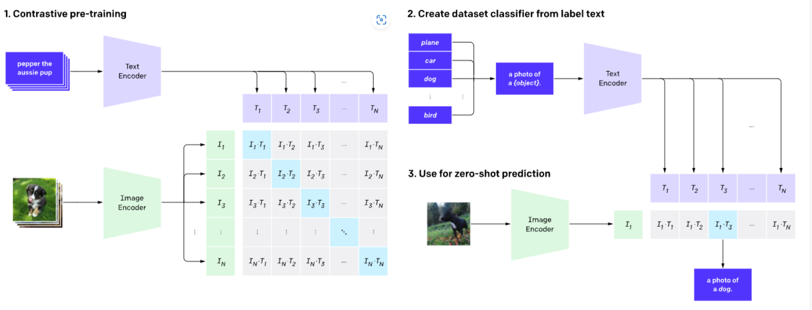
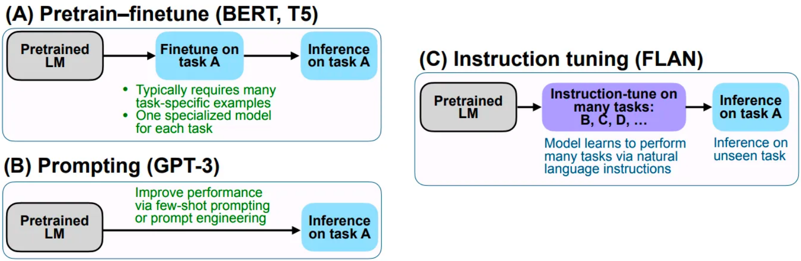


Figure: Multi-modal processing structure

- CLIP

Prompt-based Techniques



1. Instruct Tunning
2. Prompting

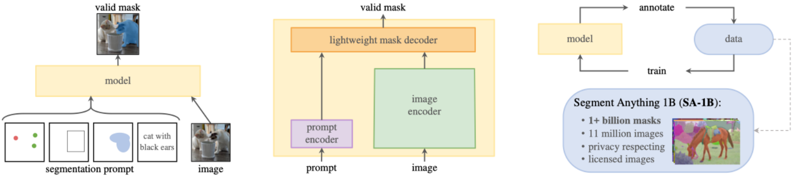
Motivation

1. to build a good big-modal based image model
2. to harness the capability of zero-shot

Scientific Questions

1. What task will enable zero-shot generalization?
2. What is the corresponding model architecture?
3. What data can power this task and model?

General Methods



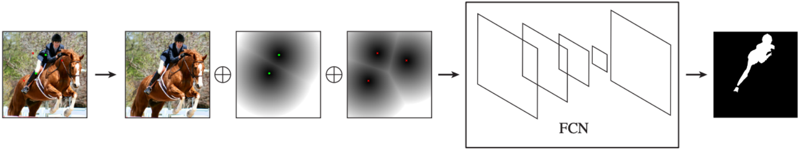
(a) **Task**: promptable segmentation (b) **Model**: Segment Anything Model (**SAM**) (c) **Data**: data engine (top) & dataset (bottom)

1. Promptable Segmentations
2. Encoder-Decoder Architecture
3. Data Engine with Dataset

Task

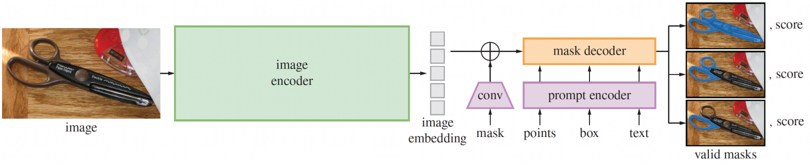
1. Translating the idea of Prompting to the task of semantic segmentation
2. Generate mask for any prompt
3. Leads ot a natural pre-training algorithm

Pretrain



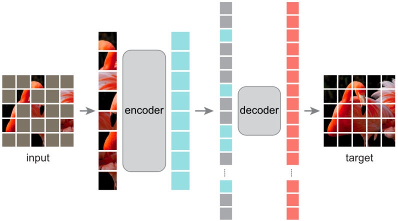
1. Provide with positive and negative clicks
2. Present the answer of correct mask
3. Unlike the classic interactive semantic segmentation, the annotator can provide the mask for any prompt

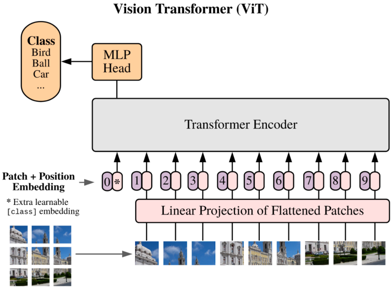
Model Architecture



1. Image Encoder
2. Prompt Encoder
3. Mask Decoder
4. Resolving Ambiguity

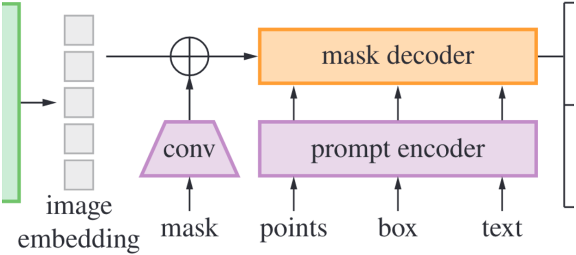
Image Encoder





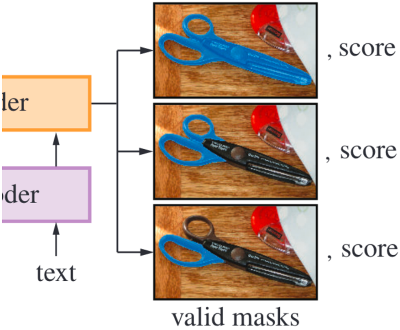
1. MAE
2. ViT

prompt Encoder / Decoder



1. Prompt of Dense and Sparse
2. masks / points, boxes, text
3. Mask encoder map the image embedding, mask and prompts to the result mask

Resolving Ambiguity

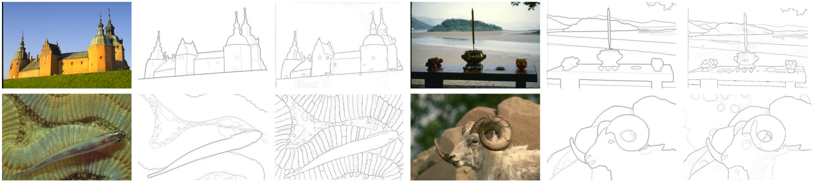


1. Three mask is usually sufficient for representing
2. add estimated IoU

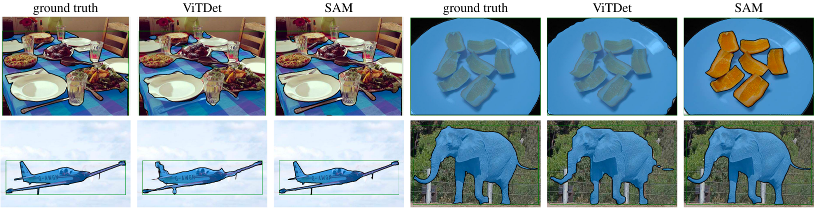
Training

1. Assisted-manual stage
   1.1 like classic interactive semantic segmentation
   1.2 have mechanism for solving granularity problem
   1.3 annotations are based on the models' output
2. Semi-automatic stage
   2.1 Aims to increase the diversity of masks in order timprove the model's generalization ability
   2.2 Ask the annotators to provide different masks
3. Fully-automatic stage

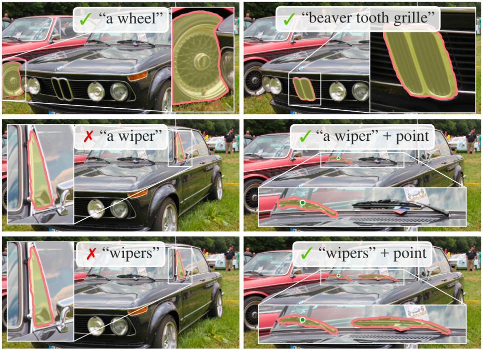
Edge Detection



Instance Segmentation



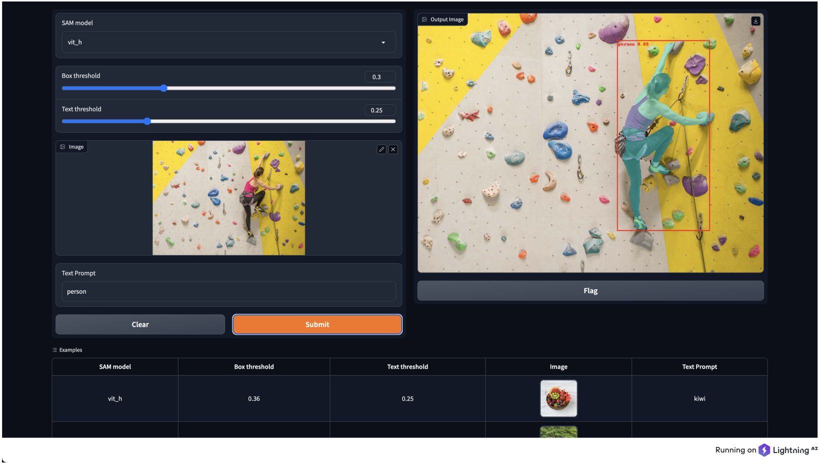
Text-to-Mask


✓ "a wheel"
✓ "beaver tooth grille"
✗ "a wiper"
✓ "a wiper" + point
✗ "wipers"
✓ "wipers" + point

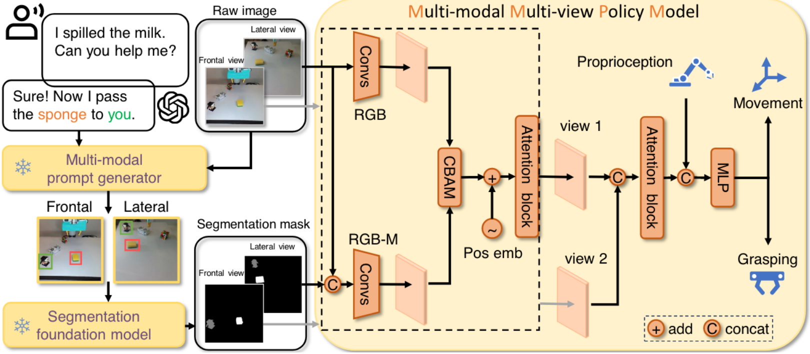
LLAVA-PLUS


Visual Generation
Visual Interaction
Input Image
Conditional Gen.
Editing
Point to multi-level Seg
Box to Seg
Stroke to Seg
2024 16-MONTH CALENDAR
floor-wood
The image depicts a wooden pier extending into a large body of water, with a bench placed in the middle of the pier. The scene is set against a backdrop of trees and mountains, creating a tranquil and picturesque environment.
Lake
bridge
The text "2024 16-MONTH CALENDAR" is detected, suggesting that this calendar is a 16-month calendar for the year 2024.
The lake in question is Lost Lake, which is located in Whistler, British Columbia, Canada.
This image is paw-some!
Who says cats can't be outdoorsy?
#naturelover
LLaVA
Object Seg and Det
Semantic Seg
OCR
Search
Social Media Post
Visual Understanding
External Knowledge
Composition

Language Segment-Anything



Transfer



Method	Seen	Unseen	New background	More distractors	Average
Ours	82.5	80.0	65.0	75.0	**75.625**
-replace mask with bbox	50.0	40.0	25.0	30.0	36.25
-w/o tracking	70.0	50.0	55.0	70.0	61.25
-single view	65.0	80.0	20.0	70.0	58.75
-RGB-M only	85.0	70.0	50.0	70.0	68.75

Grasp Anything

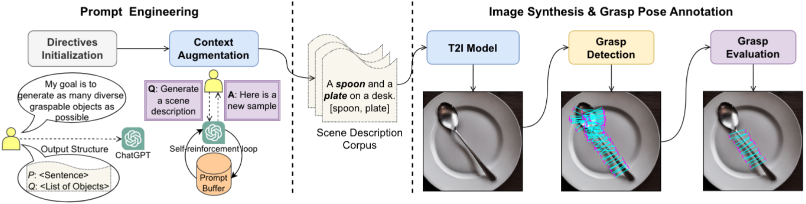
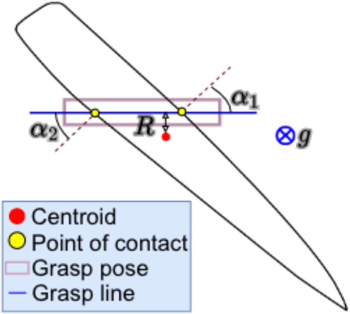
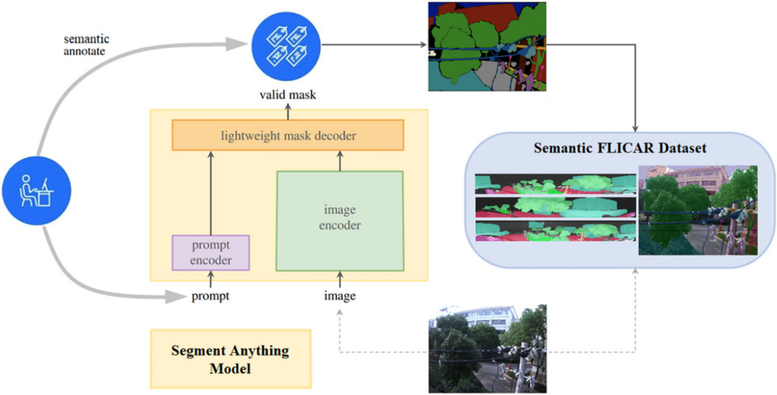


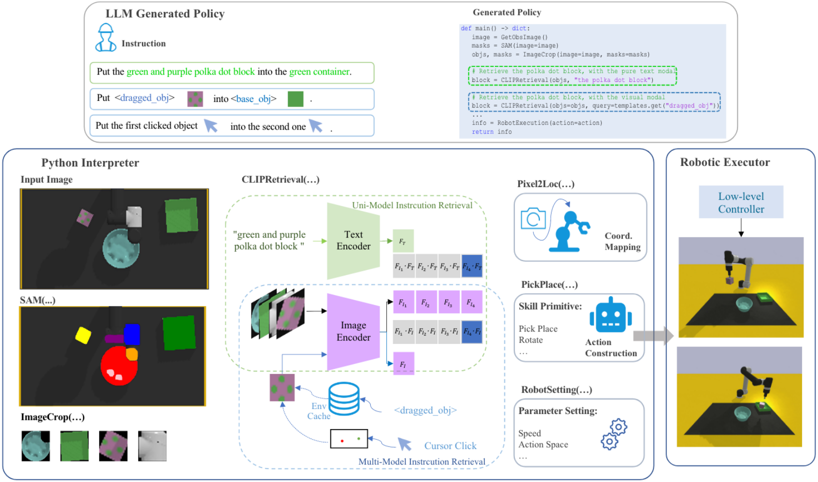
Fig. 2. **Dataset creation pipeline.**



FLICAR


semantic
annotate
valid mask
lightweight mask decoder
image
encoder
prompt
encoder
prompt
image
Segment Anything
Model
Semantic FLICAR Dataset

Instruct2Act



Agriculture Robots

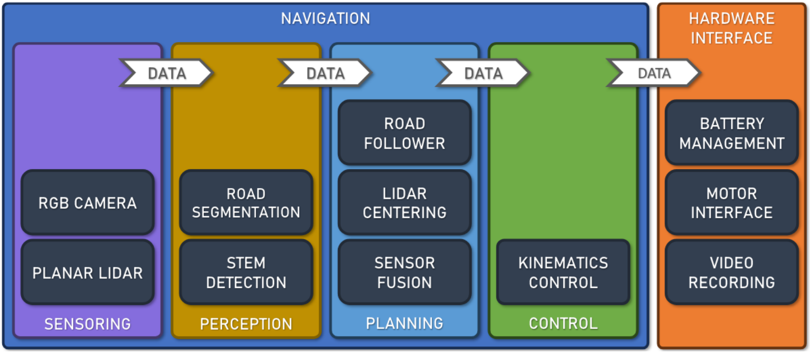


Fig. 1: Overview of the robot platform architecture showing its components and relations

OVIR-3D

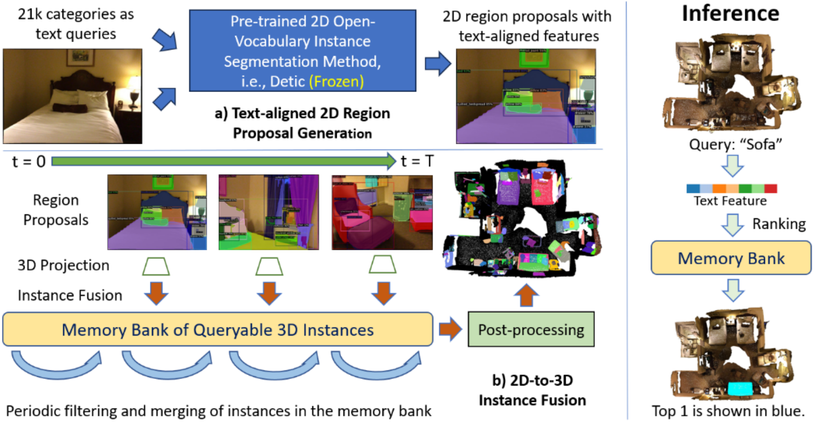


Figure 2: **Pipeline of the proposed method.**

Takeaways

1. A good semantic segmentation model
2. Encoporating human interaction like Prompting can give more possibliities
3. An existing experiment pattern can achieve great result when combined with new emerging techniques