Methods of image grounding and LLM-based robotic manipulation

a CogRob project

Juan Lopez, s5156750
Thijs Lukkien, s3978389

Topic:

LLMs for robot instructions

1.
The idea

> “ Compare two methods of image grounding
> against a ground-truth method, for LLM implementations via natural language instructions ”

2.
The background

LLM → import modules

LLM → rewards
ML → actions

LLM → general knowledge
LLM → world knowledge
LLM → planning
LLM → verification
ML → task learning

2022

ProgPrompt: Generating Situated Robot Task Plans using Large Language Models

PaLM-E: An Embodied Multimodal Language Model

VLM → Encode vision
LLM → Encode text

Language to Rewards for Robotic Skill Synthesis

Embodied Task Planning with Large Language Models

NN → Classification
LLM → Planning
LLM →

Scaling Up and Distilling Down: Language-Guided Robot Skill Acquisition

RoboGPT: an intelligent agent of making embodied long-term decisions for daily instruction tasks

LLM → planning
LLM → module use
[...]

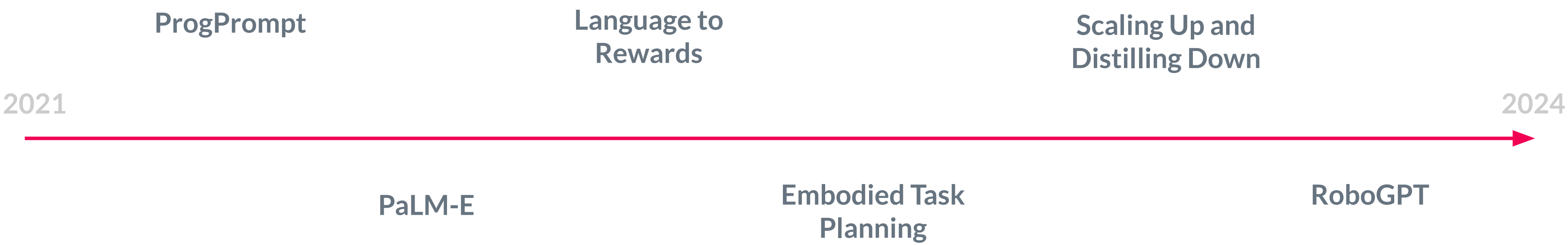
En

2024

LLM → import modules

LLM → rewards
ML → actions

LLM → general knowledge
LLM → world knowledge
LLM → planning
LLM → verification
ML → task learning



VLM → Encode vision
LLM → Encode text

NN → Classification
LLM → Planning
LLM →

LLM → planning
LLM → module use
[...]

LLM → import modules

LLM → rewards
ML → actions

LLM → general knowledge
LLM → world knowledge
LLM → planning
LLM → verification
ML → task learning

ProgPrompt:
Generating Situated Robot
Task Plans using Large
Language M

PaLM-
An Embodied
Multimodal
Language Model

**Language to
Rewards**

with Large Languag
Models

Scaling Up and

ntelligent agent
odied long-term decisions
r daily instruction tasks

VLM → Encode vision
LLM → Encode text

NN → Classification
LLM → Planning
LLM →

LLM → planning
LLM → module use

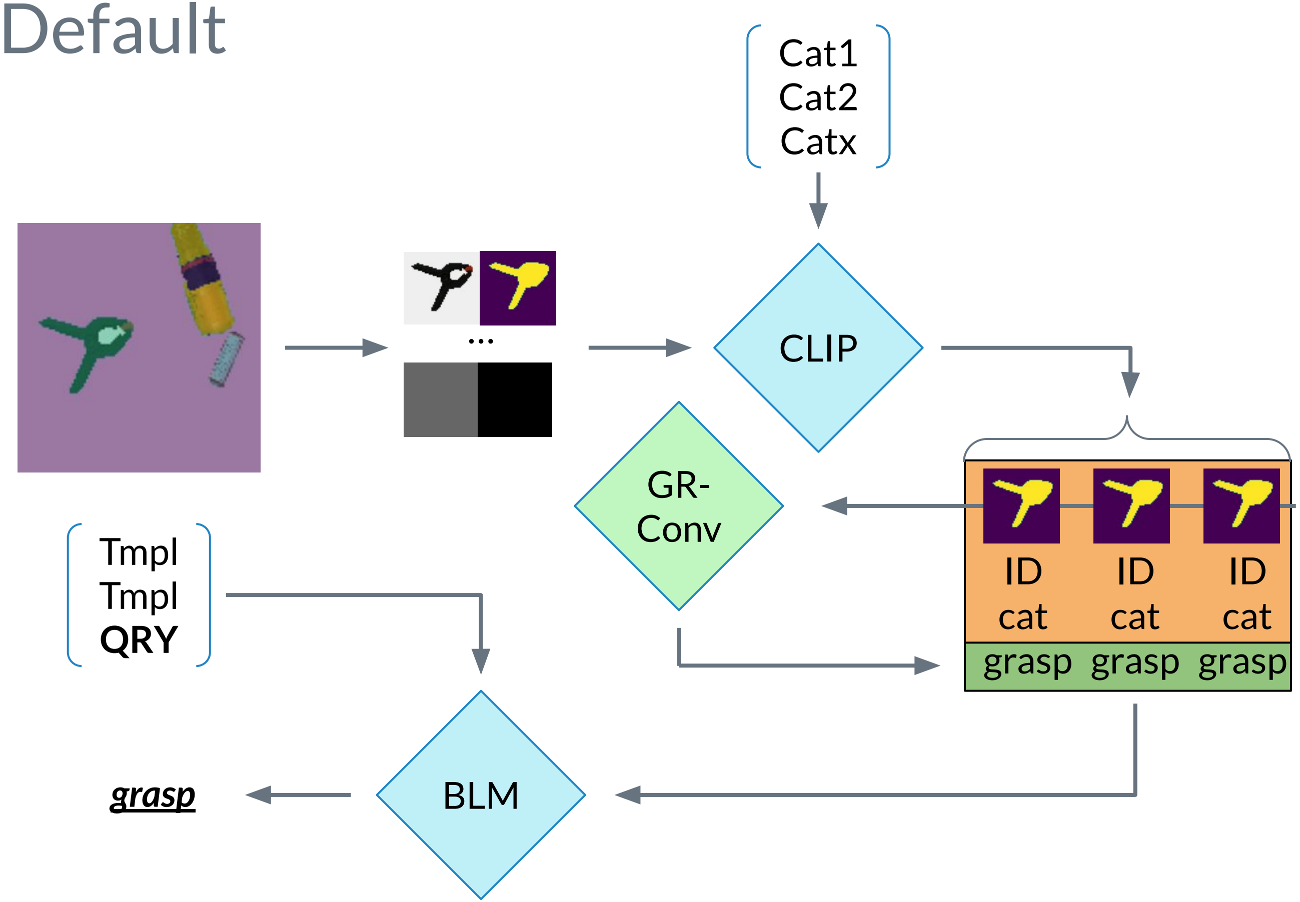
Model Zoo

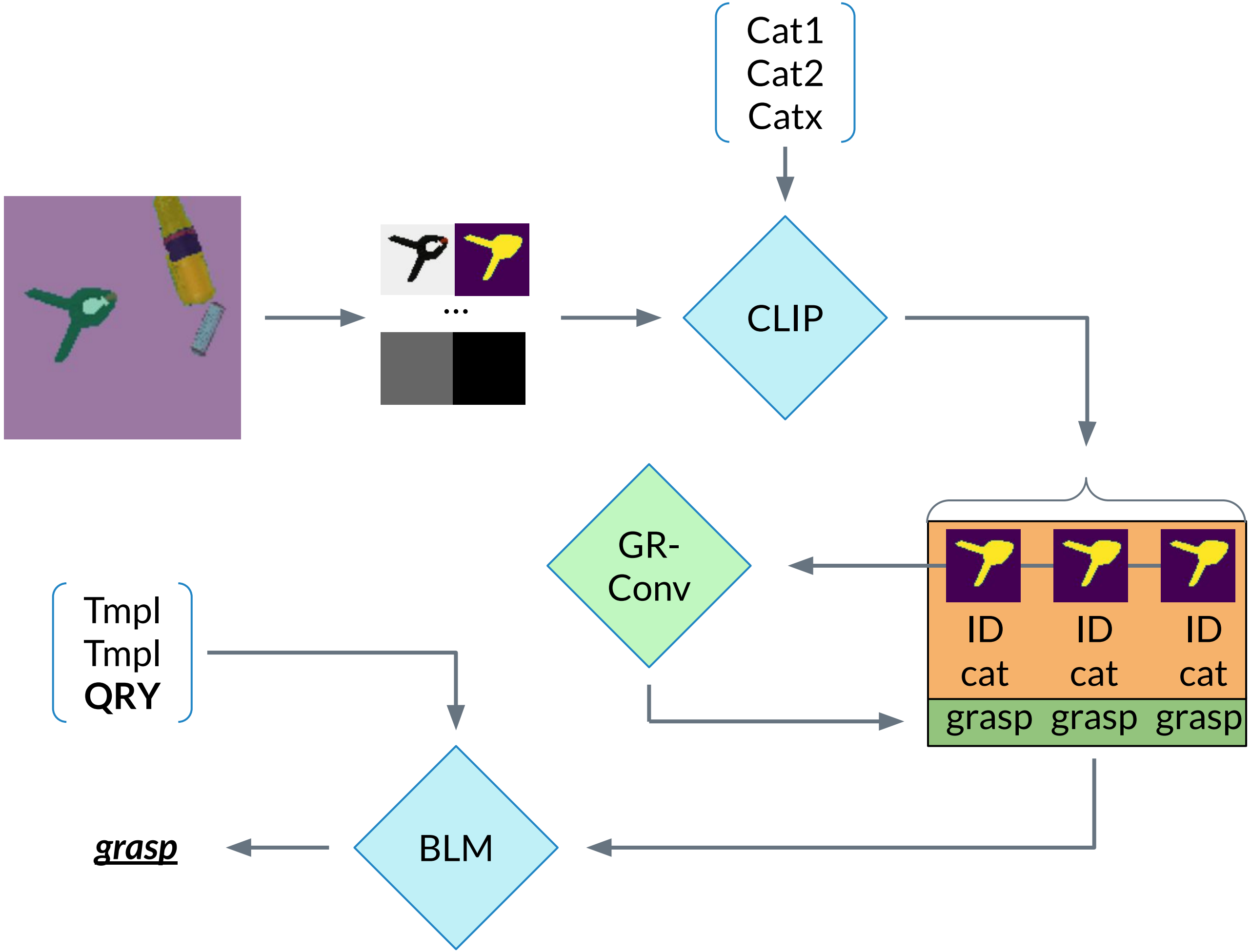
BLOOM	- LLM → text
GR-CONVNET	- CNN → grasps
SAM	- MMTM → Segmentation maps
CLIP	- MMLM → text
KOSMOS2	- MMLM → grounded text + referred bbox images

Additional Considerations

- Only pretrained models, no fine-tuning
- Compare their few-shots/zero-shots capabilities.
- Grasp success depends on
  - Segmentation quality
  - Classification performance
  - Grasp pose estimator

Default


Cat1
Cat2
Catx
...
CLIP
GR-
Conv
ID
cat
ID
cat
ID
cat
grasp
grasp
grasp
Tmpl
Tmpl
QRY
BLM
grasp



Cat1
Cat2
Catx

…

CLIP

GR-Conv

ID
cat
grasp

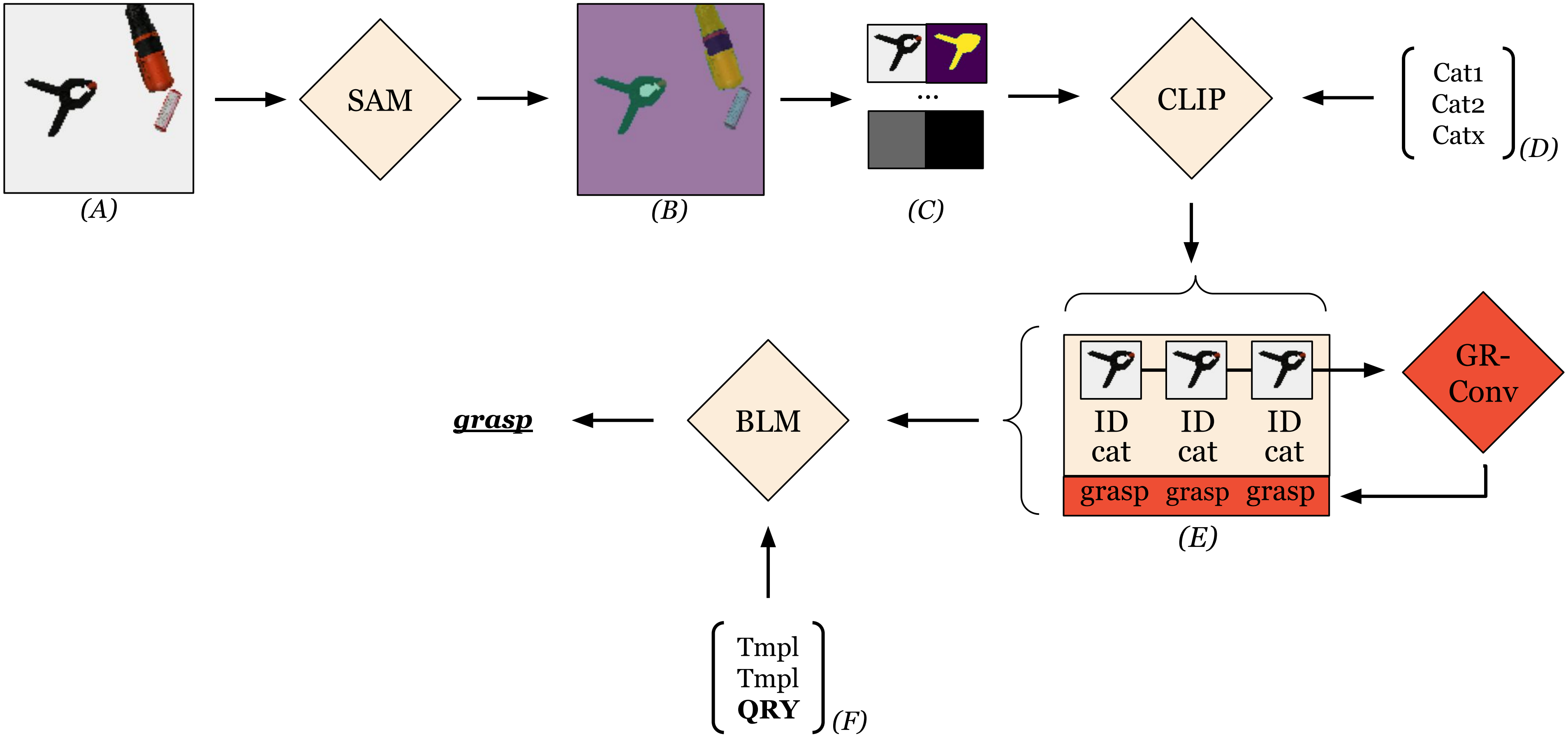
ID
cat
grasp

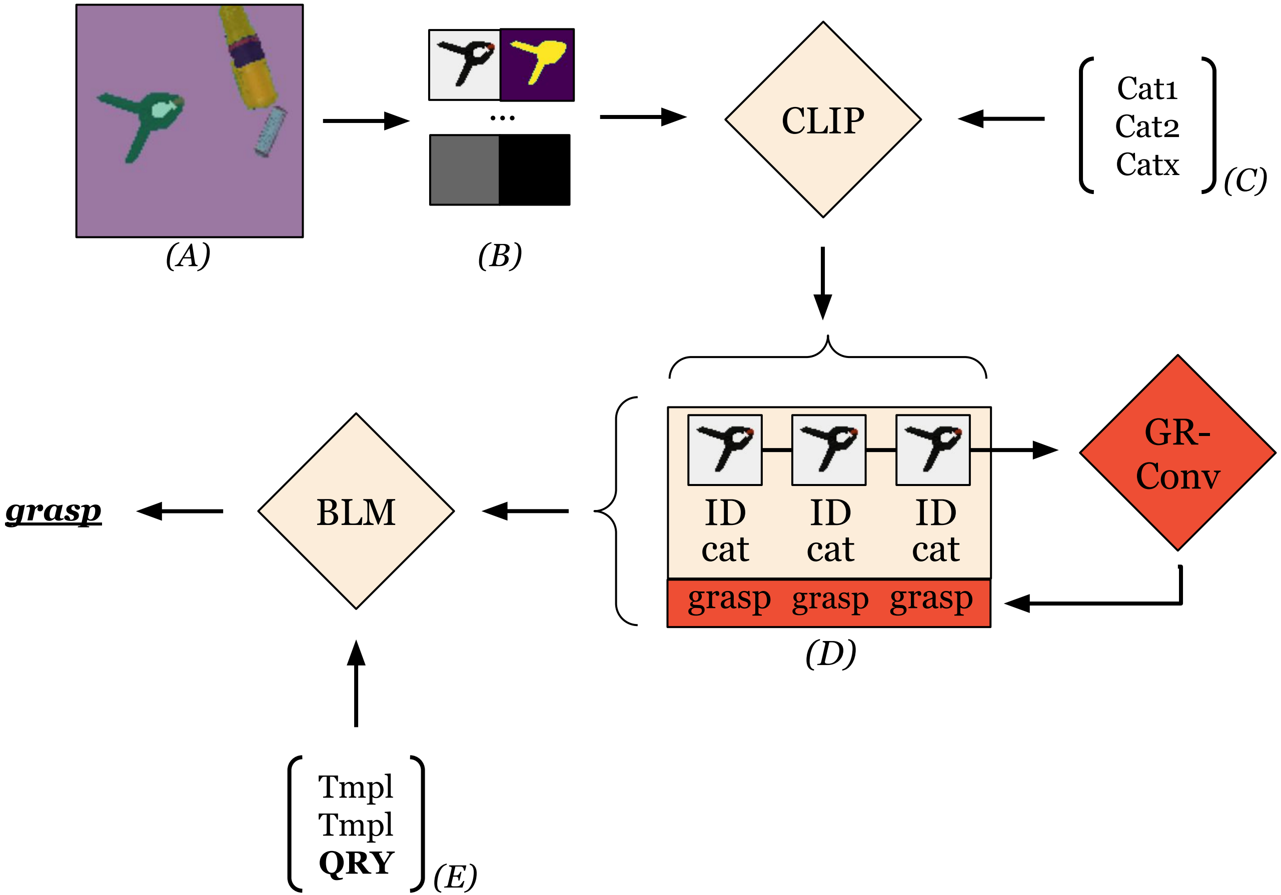
ID
cat
grasp

Tmpl
Tmpl
QRY

BLM

grasp


(A)
SAM
(B)
...
(C)
CLIP
Cat1
Cat2
Catx
(D)
GR-Conv
ID cat
ID cat
ID cat
grasp grasp grasp
(E)
BLM
grasp
Tmpl
Tmpl
QRY
(F)



(A)

(B)

CLIP

Cat1
Cat2
Catx

(C)

GR-Conv

ID cat ID cat ID cat

grasp grasp grasp

(D)

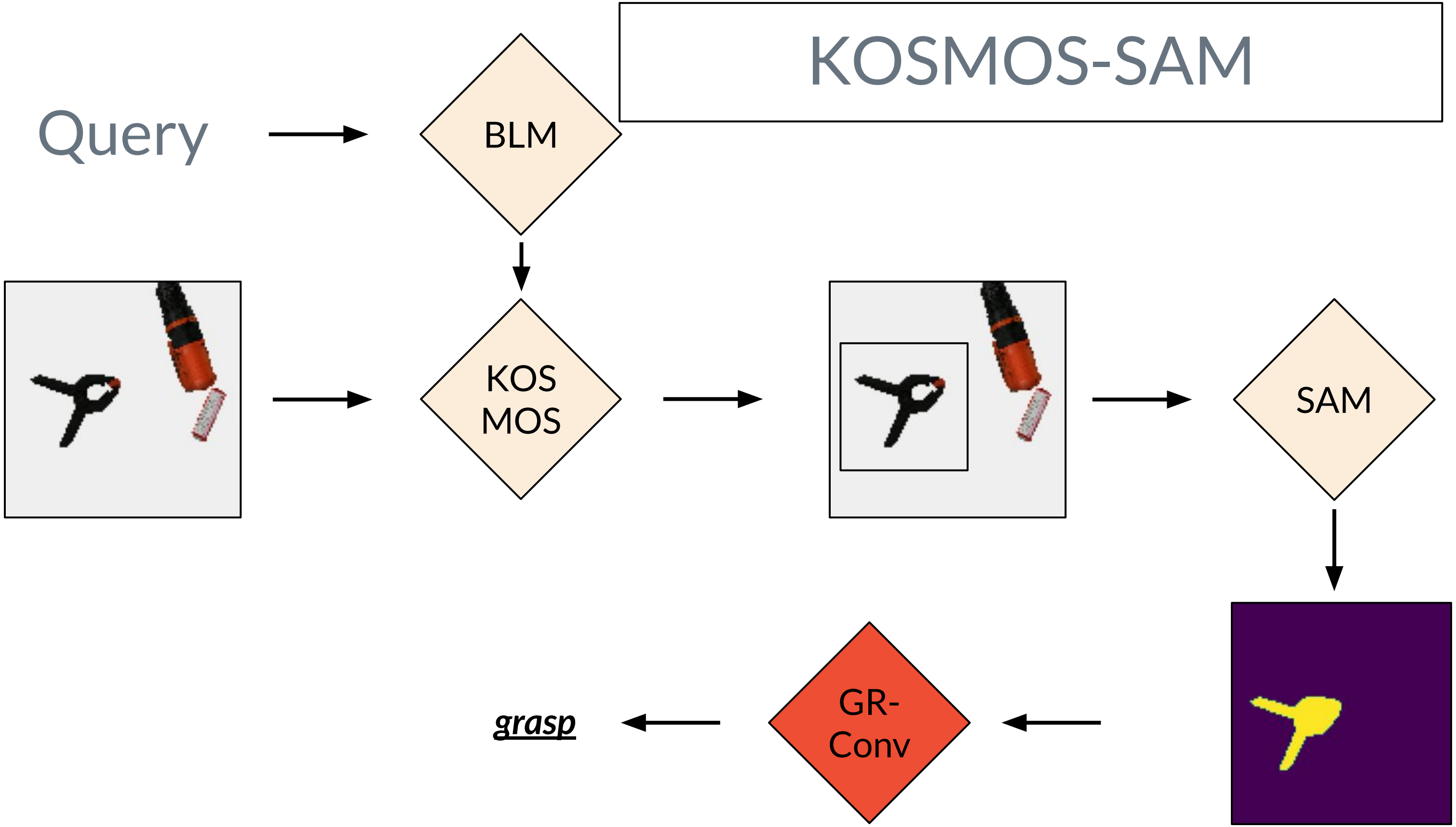
BLM

grasp

Tmpl
Tmpl
QRY

(E)

KOSMOS-SAM



Query

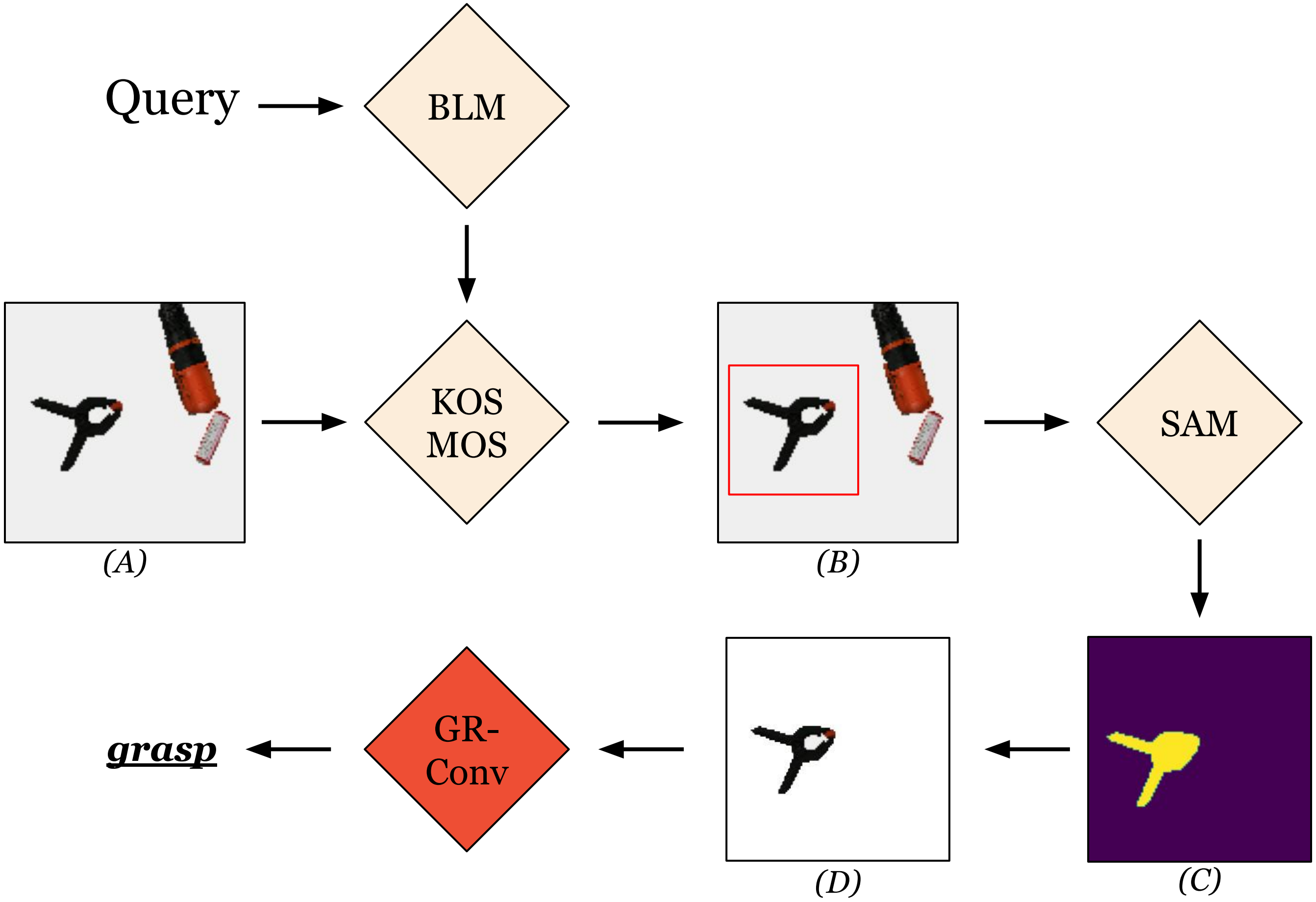
BLM

KOS MOS

SAM

GR-Conv

grasp


Query
BLM
KOS MOS
(A)
(B)
SAM
(C)
(D)
GR-Conv
grasp

3.
Evaluation

Experimental Setup

3 experimental methods:

- Default (Ground Truth Segmentation and CLIP)
- SAM (automatic) - CLIP
- KOSMOS - SAM (bounding box)

- New Collab adaptation

Experimental Setup

- 4 objects per run, 20 runs
- 1 dynamically generated query per object
- 80 queries per pipeline

What we measure

- Overall success of the query completion
- Successful grasp
- Successful delivery
- Successful Object Detection

4.

Results

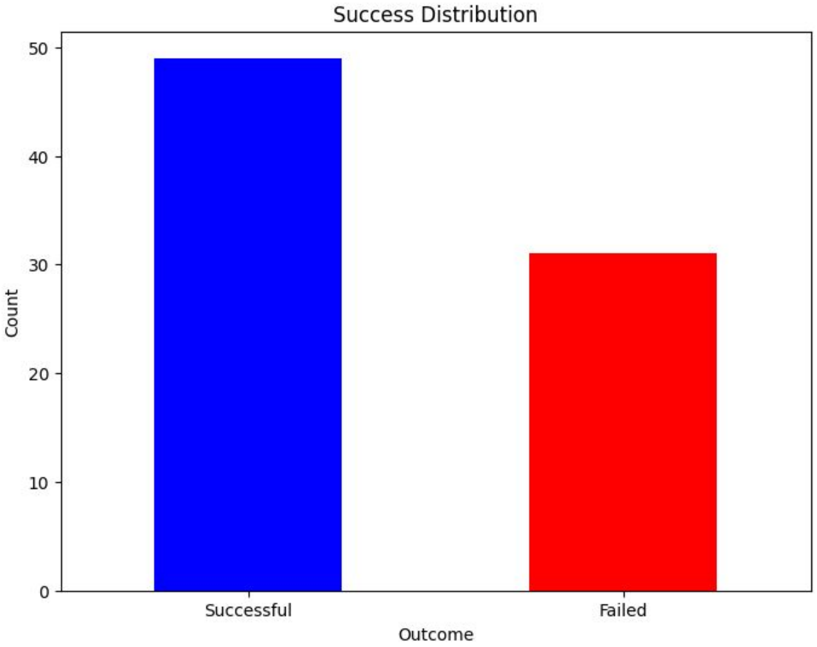
Preliminary results

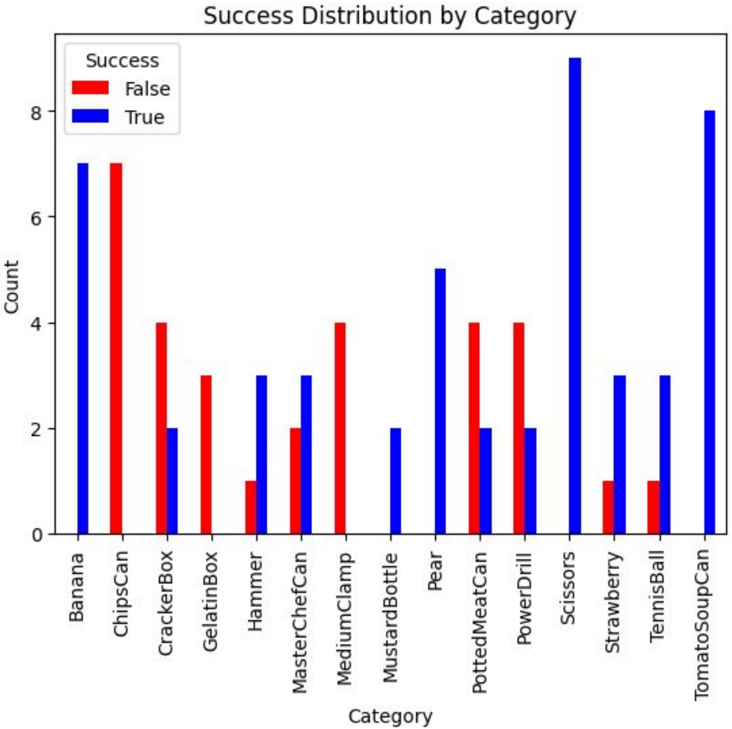
Kosmos-SAM > SAM-CLIP

Kosmos-SAM > ground-CLIP

CLIP-classification

ground-CLIP


Success Distribution
Count
0
10
20
30
40
50
Successful
Failed
Outcome


Success Distribution by Category
Success
🟥 False
🟦 True
Count
0
2
4
6
8
Banana
ChipsCan
CrackerBox
GelatinBox
Hammer
MasterChefCan
MediumClamp
MustardBottle
Pear
PottedMeatCan
PowerDrill
Scissors
Strawberry
TennisBall
TomatoSoupCan
Category

5.
Conclusion

Conclusion

Modality / model-applicability

6.

Discussion

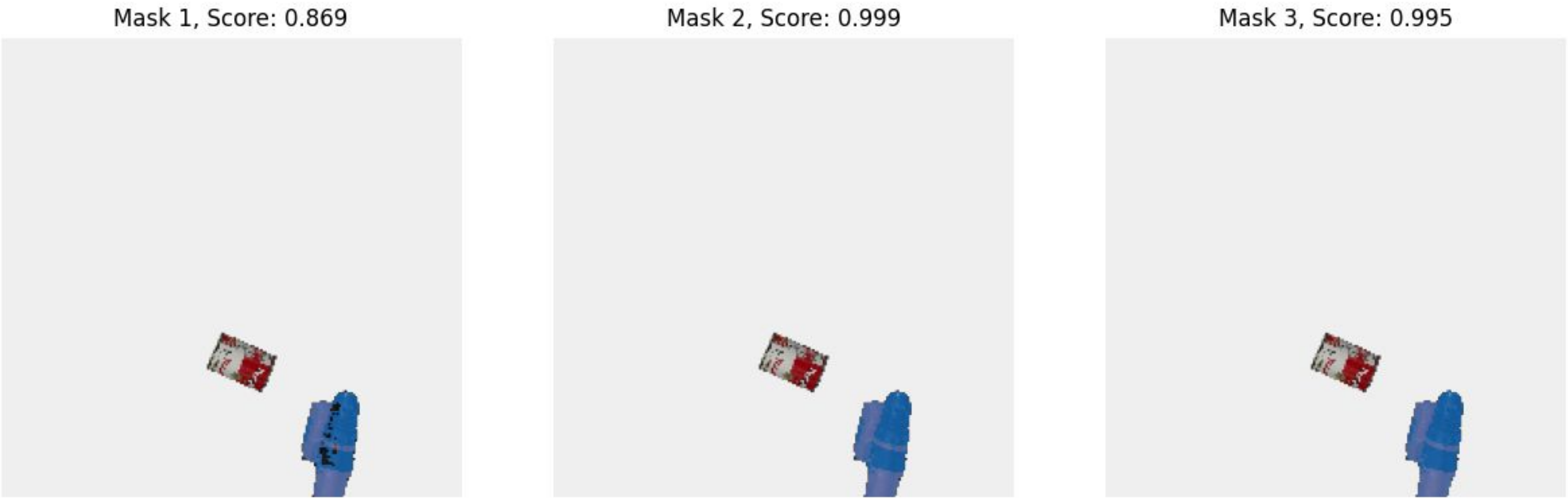
Discussion

- CLIP and BLOOM performance (poor)
- Manual tuning
- Cascading errors : pinpointing difficulties
  - Where does low performance come from in the pipeline? Was it at the start or at the end?
- Potential for simplification
- NextChat - LLM that also creates the segmentation mask text grounding (Mixture of clip and SAM in a single model)

“Questions?”

~ Dhali ~

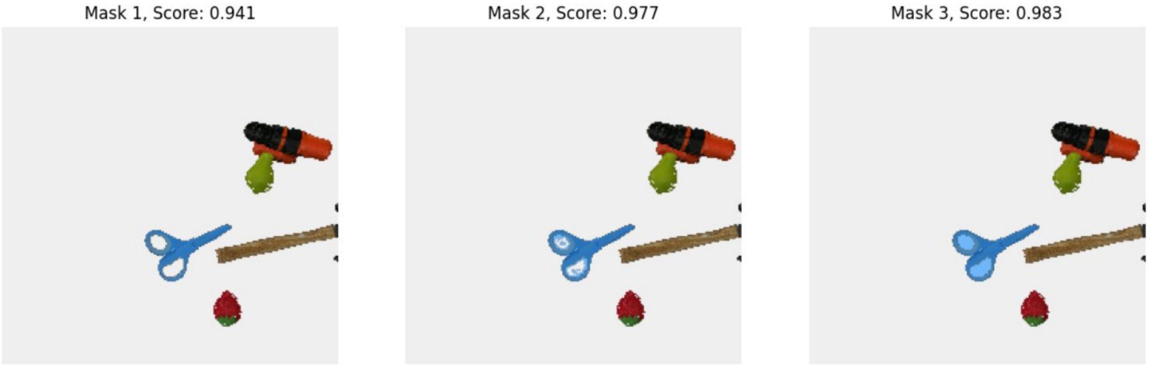
Extra material

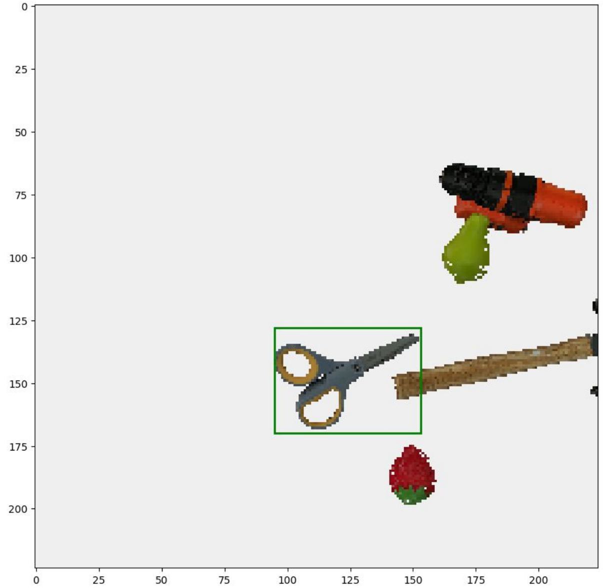


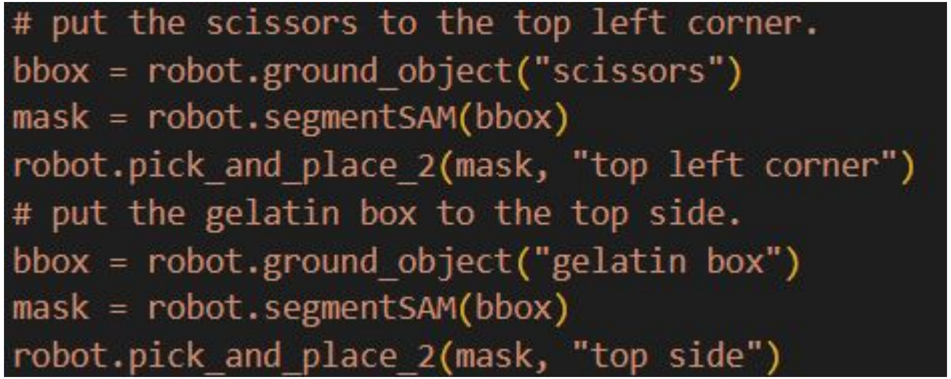
Mask 1, Score: 0.869

Mask 2, Score: 0.999

Mask 3, Score: 0.995

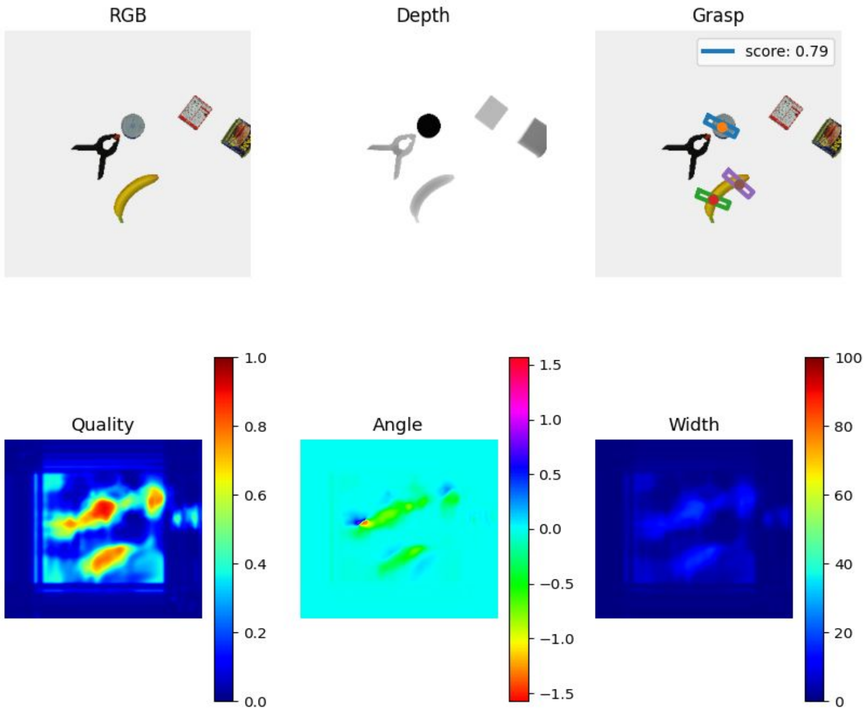






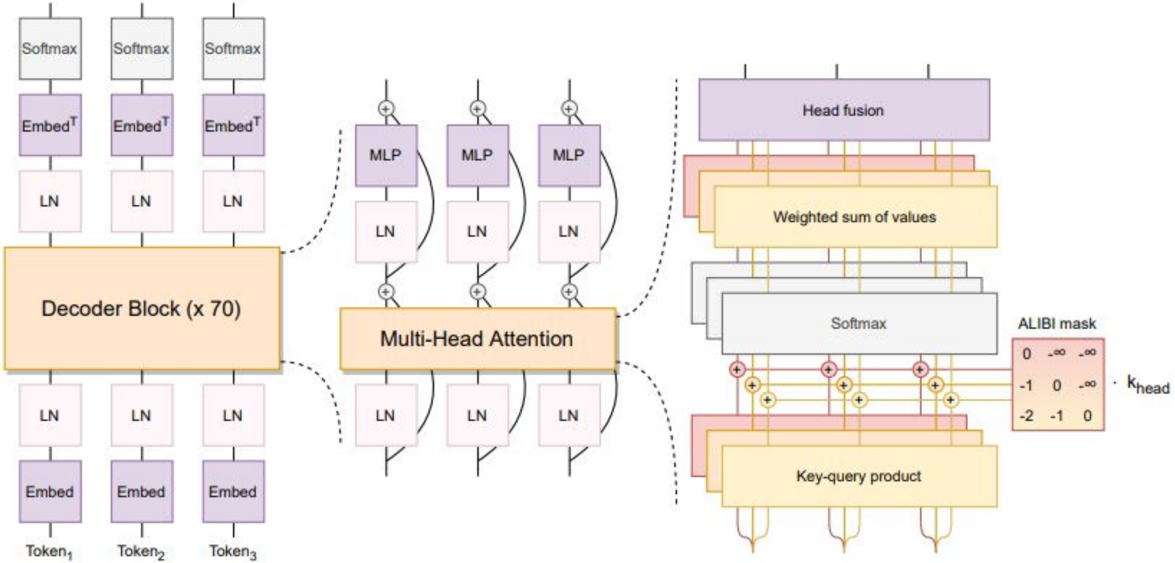
```
# put the scissors to the top left corner.
bbox = robot.ground_object("scissors")
mask = robot.segmentSAM(bbox)
robot.pick_and_place_2(mask, "top left corner")
# put the gelatin box to the top side.
bbox = robot.ground_object("gelatin box")
mask = robot.segmentSAM(bbox)
robot.pick_and_place_2(mask, "top side")
```

GR-CONVNET


RGB
Depth
Grasp
score: 0.79
Quality
1.0
0.8
0.6
0.4
0.2
0.0
Angle
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
Width
100
80
60
40
20
0

BLOOM

Classic LLM similar to GPT-3



SAM

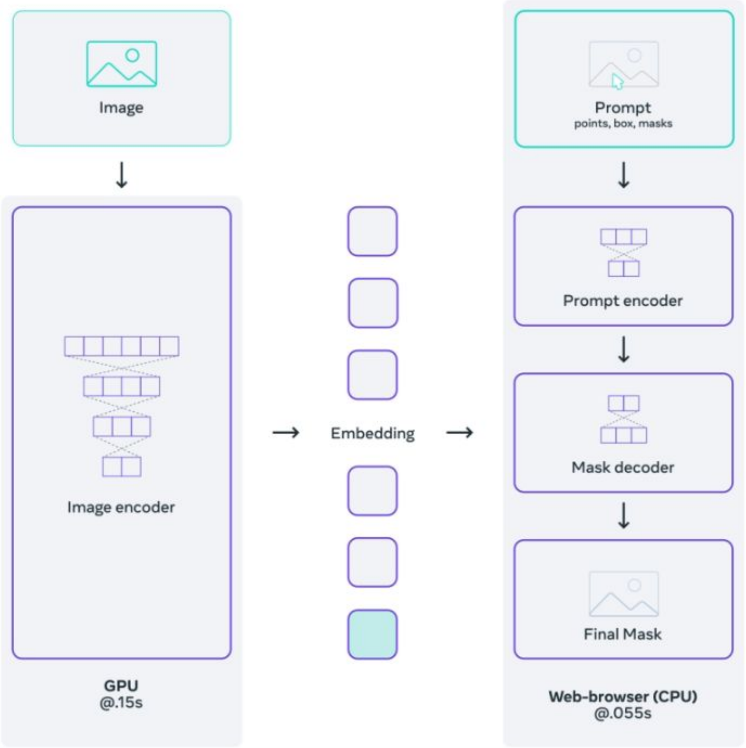
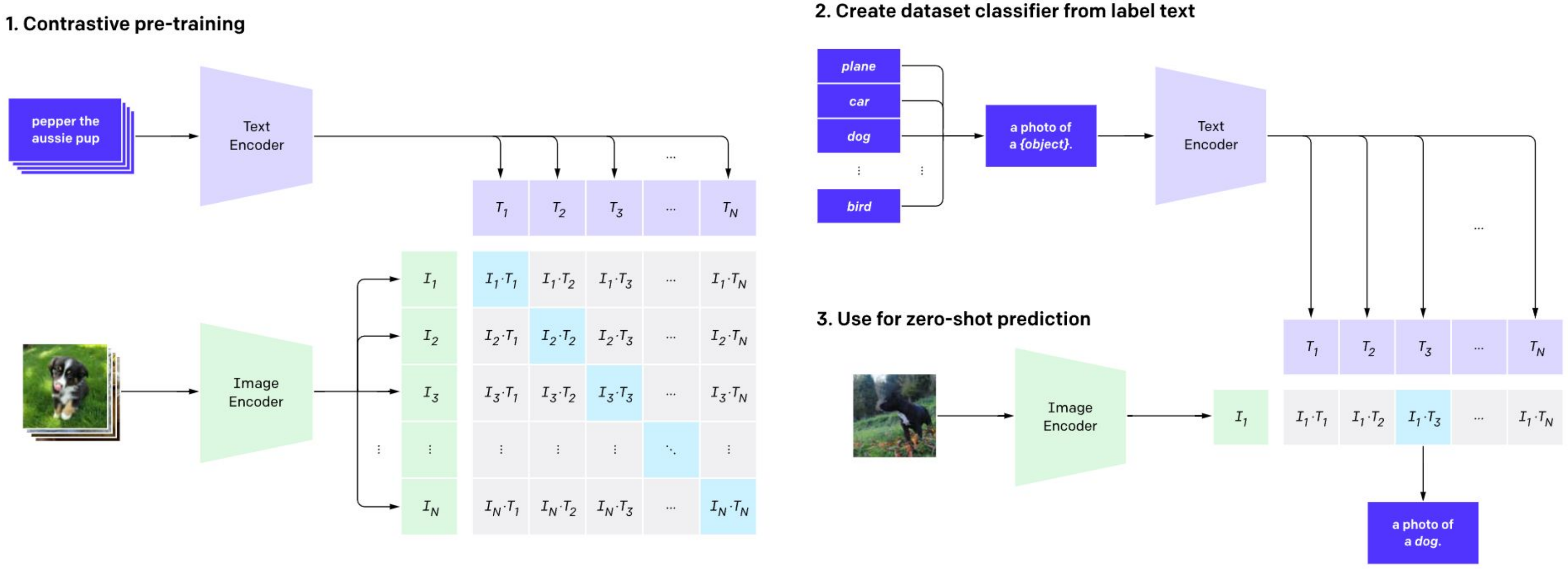


Image
Image encoder
GPU
@.15s
Embedding
Prompt
points, box, masks
Prompt encoder
Mask decoder
Final Mask
Web-browser (CPU)
@.055s

CLIP



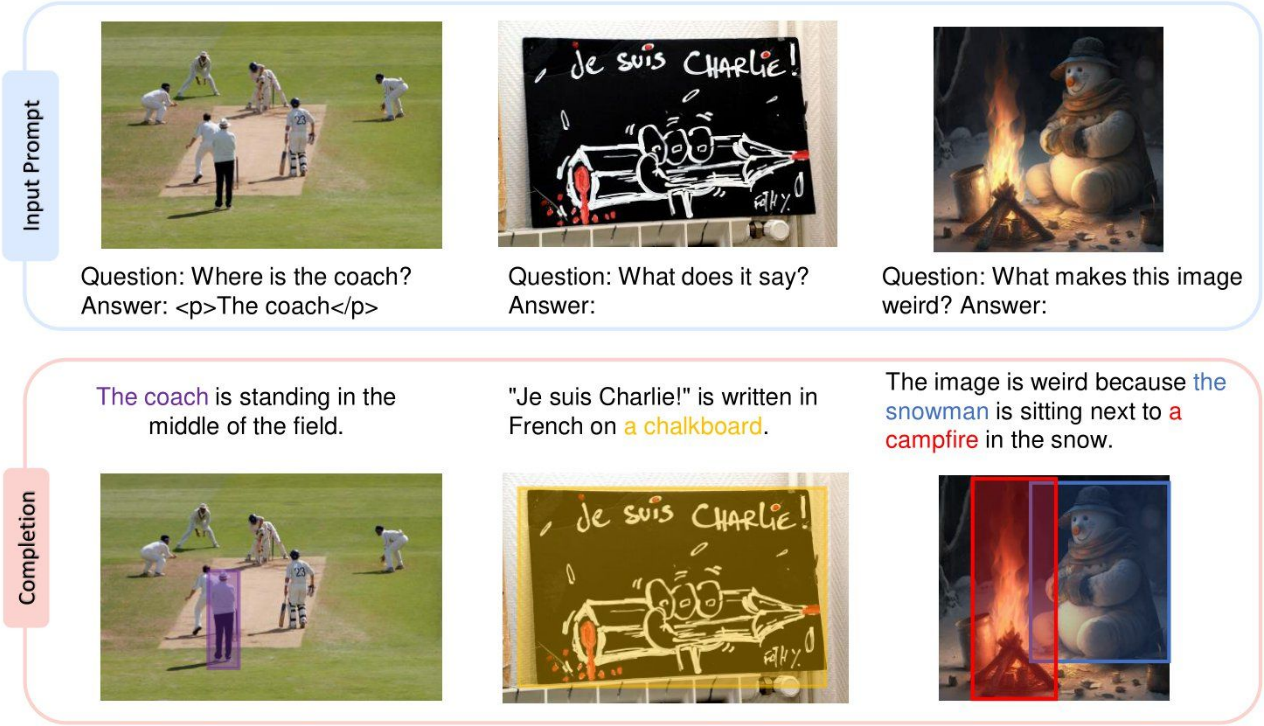
1. Contrastive pre-training

	T_1	T_2	T_3	...	T_N
I_1	$I_1 \cdot T_1$	$I_1 \cdot T_2$	$I_1 \cdot T_3$	...	$I_1 \cdot T_N$
I_2	$I_2 \cdot T_1$	$I_2 \cdot T_2$	$I_2 \cdot T_3$	...	$I_2 \cdot T_N$
I_3	$I_3 \cdot T_1$	$I_3 \cdot T_2$	$I_3 \cdot T_3$	...	$I_3 \cdot T_N$
⋮	⋮	⋮	⋮	⋱	⋮
I_N	$I_N \cdot T_1$	$I_N \cdot T_2$	$I_N \cdot T_3$	...	$I_N \cdot T_N$

2. Create dataset classifier from label text

3. Use for zero-shot prediction

Kosmos-2



Question: Where is the coach?
Answer: <p>The coach</p>

Question: What does it say?
Answer:

Question: What makes this image weird? Answer:

The coach is standing in the middle of the field.

"Je suis Charlie!" is written in French on a chalkboard.

The image is weird because the snowman is sitting next to a campfire in the snow.

(1) Grounded question answering

https://colab.research.google.com/drive/1orQdJ5qkkW44uIlwPrTaKl_BfbwoiydY?usp=sharing

Once you get access to the repo