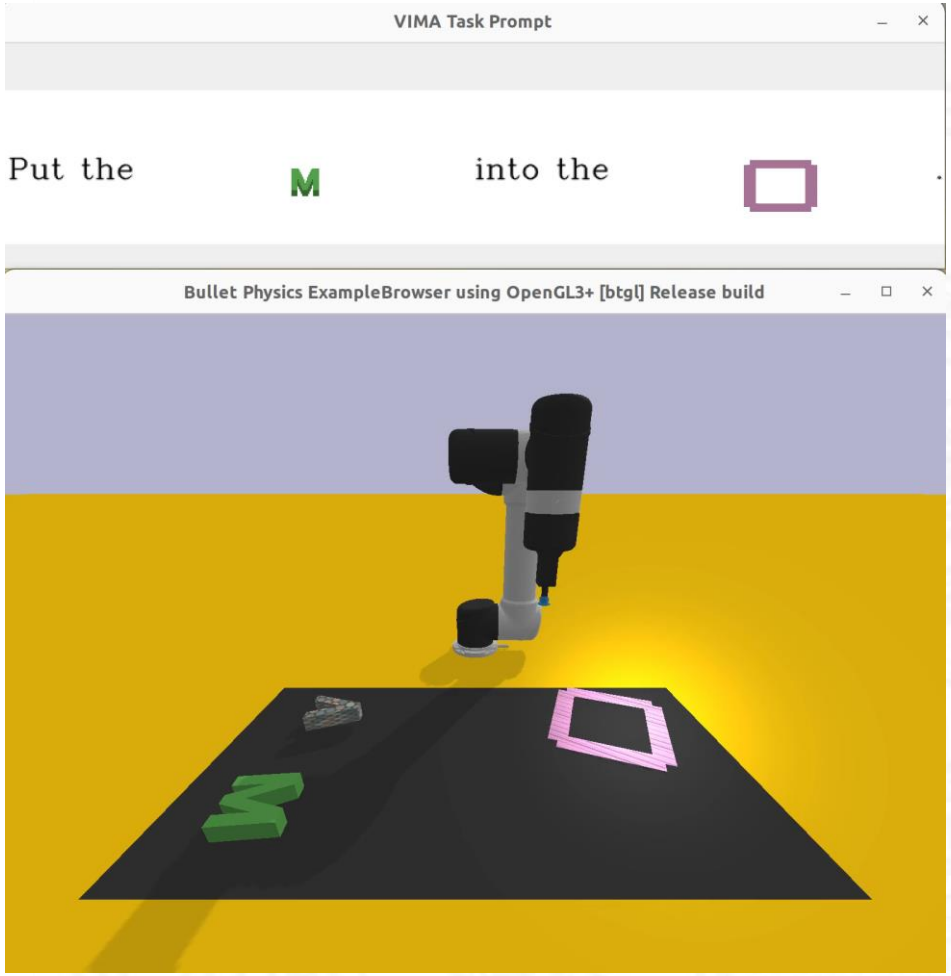
Multi-modal Robotic Perception with VIMA

Kaixin Bai

14.11.2023

VIMA: General Robot Manipulation with Multimodal Prompts

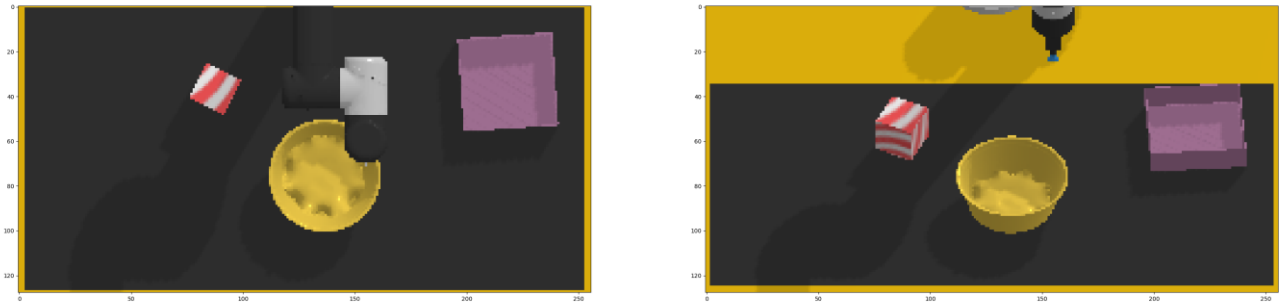


VIMA: General Robot Manipulation with Multimodal Prompts

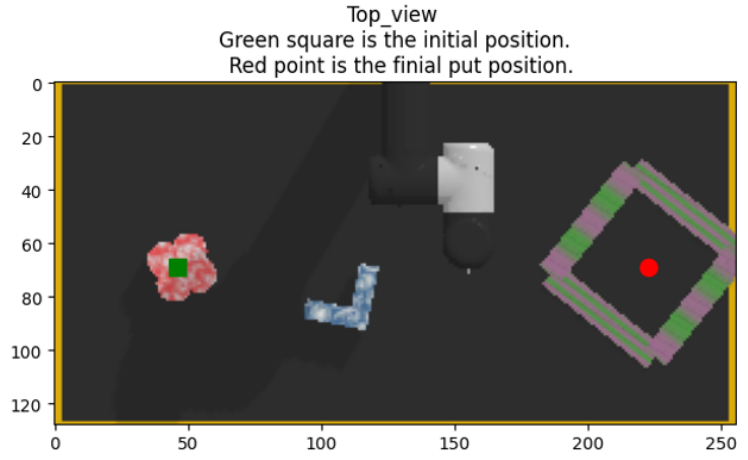
VIMA: General Robot Manipulation with Multimodal Prompts

- Input and output

Input1: multi-modal input



Input2: top view and front view images



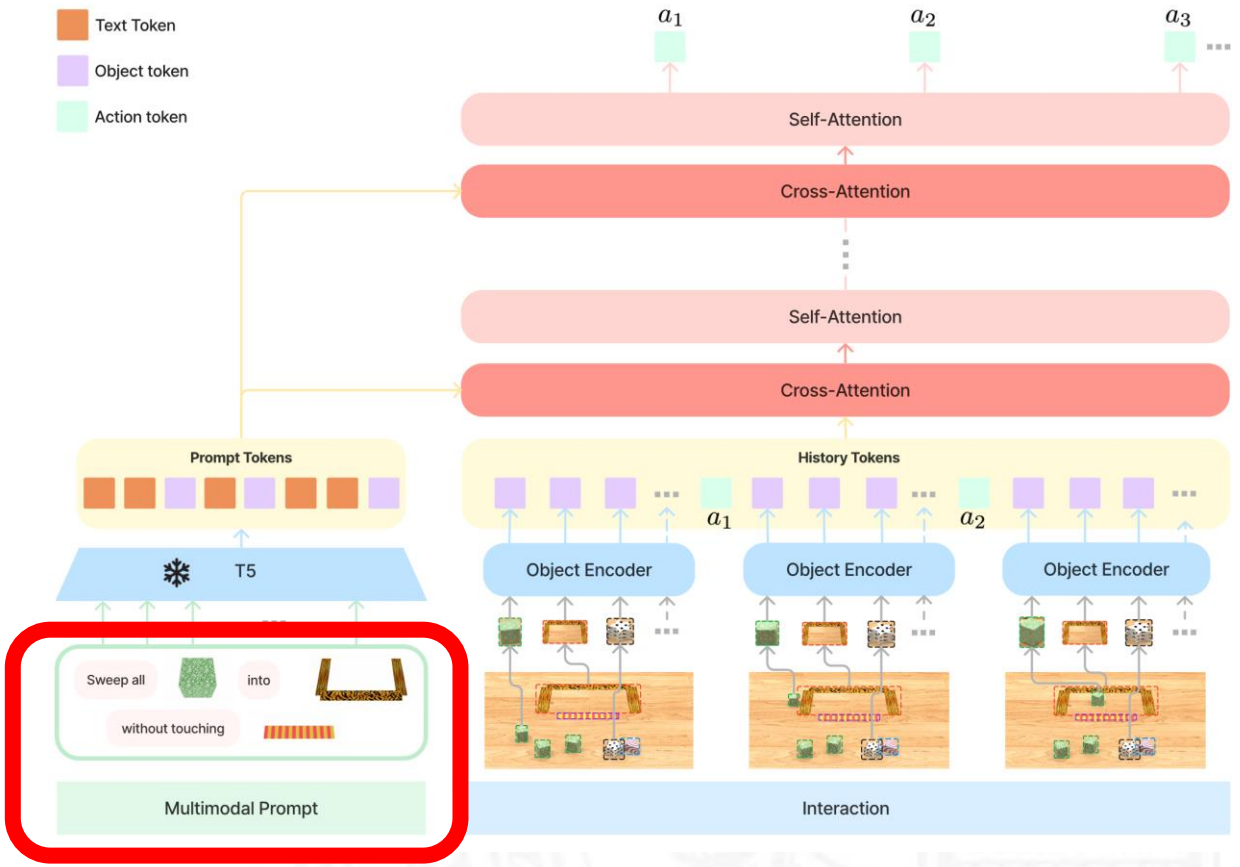
Model output: (Variables:
actions[“pose0_position”],
actions[“pose1_position”])
For initial position [0.50, 0.20]
For final put position [0.5, 0.90]

VIMA: General Robot Manipulation with Multimodal Prompts

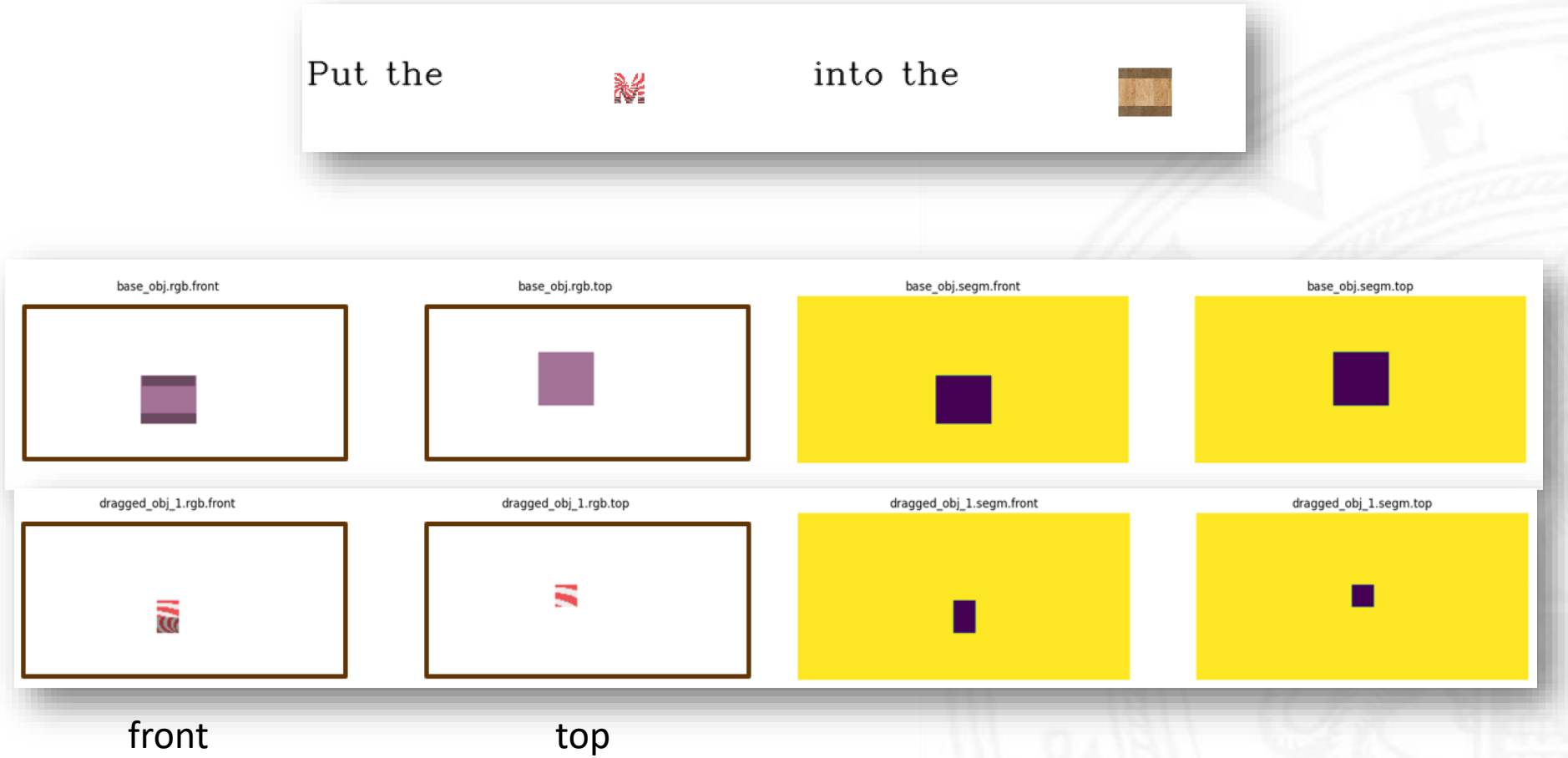
Prompt example:
'Put the {dragged_obj_1} into the {base_obj}.'
Prompt_assets example: Dict

```
prompt = env.prompt
prompt_assets = env.prompt_assets
```

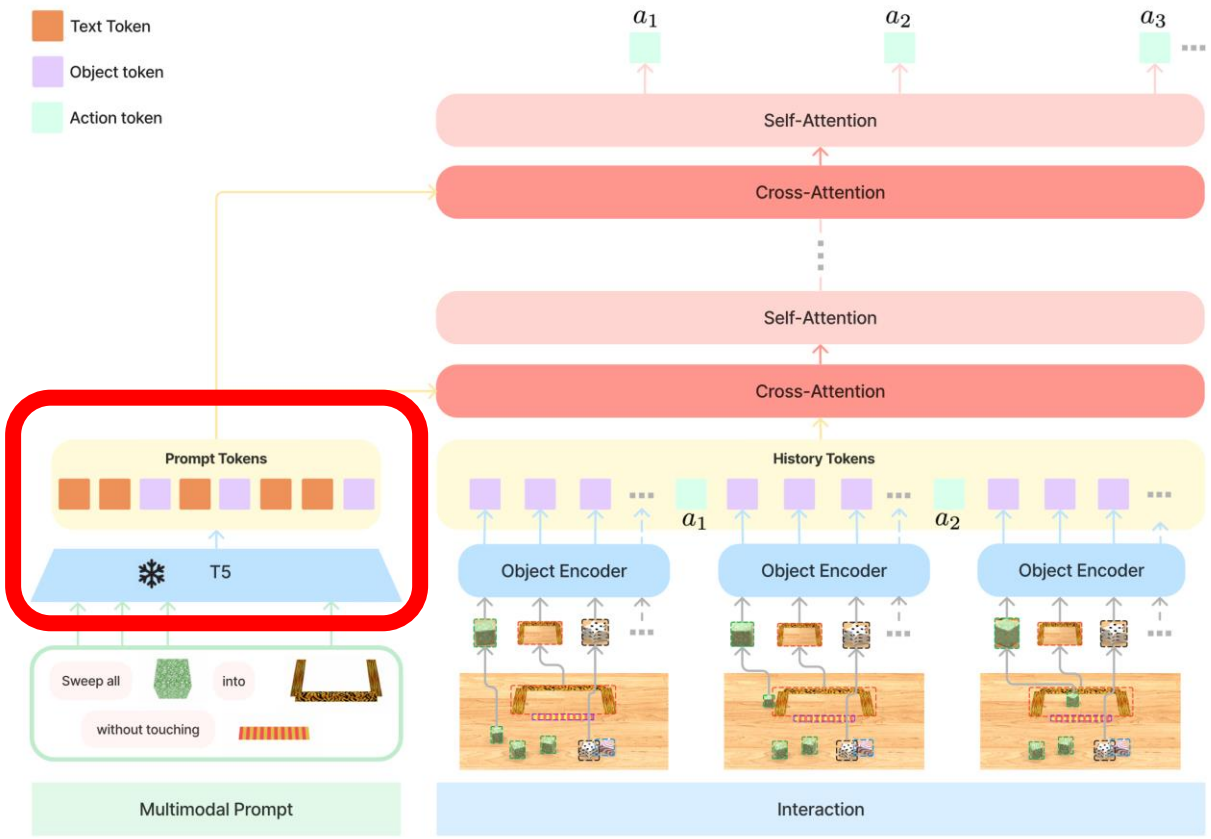
- 'base_obj'
  - 'rgb'
    - 'front': shape(128*256)
    - 'top'
  - 'segm'
    - 'front': shape(128*256)
    - 'top'
    - 'obj_info'
- 'dragged_obj_1': ...

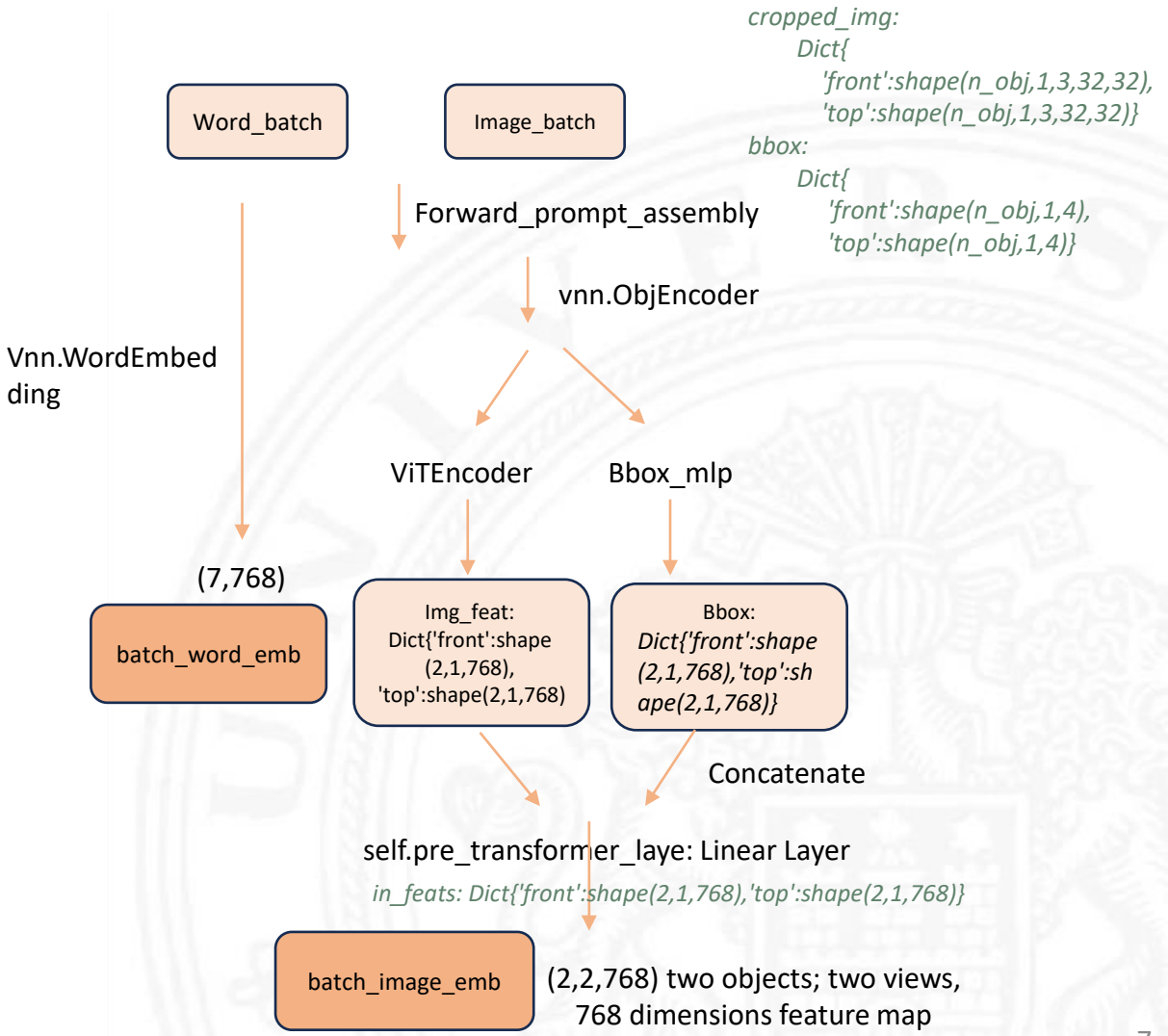


VIMA: General Robot Manipulation with Multimodal Prompts

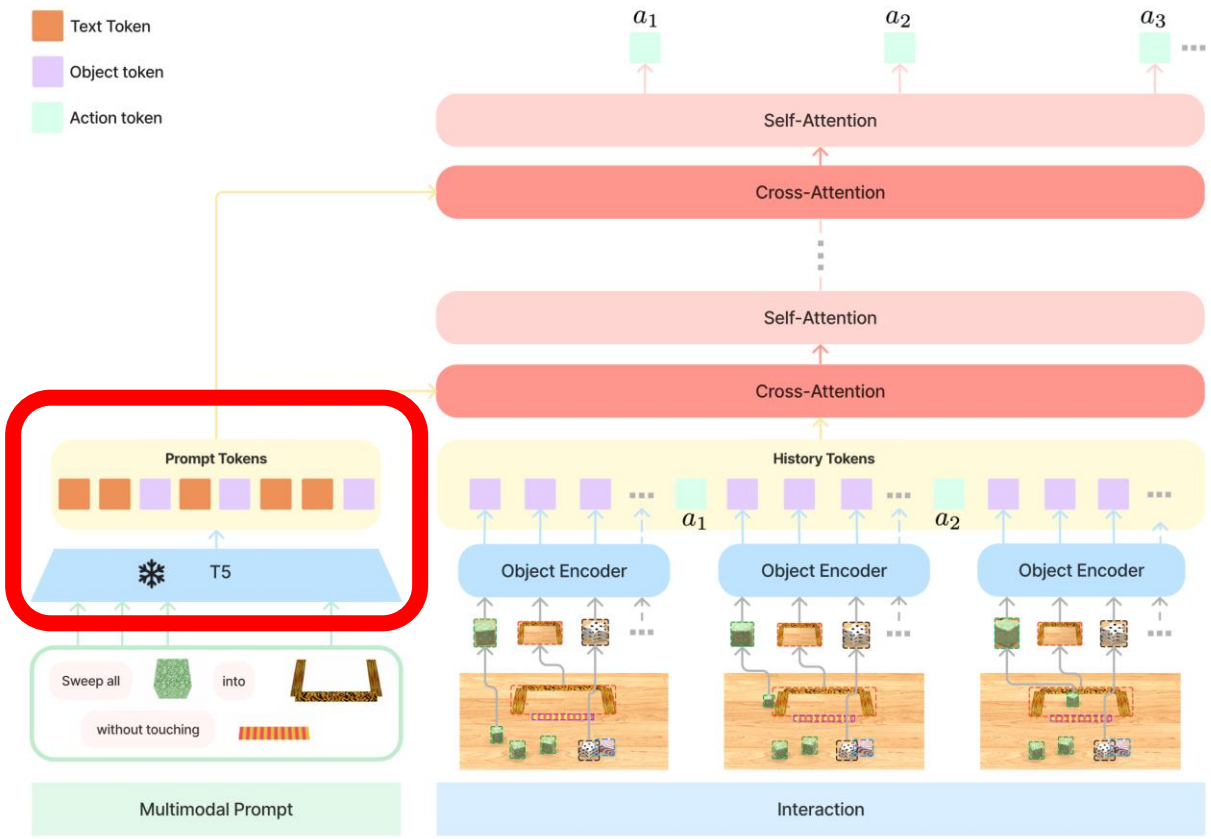


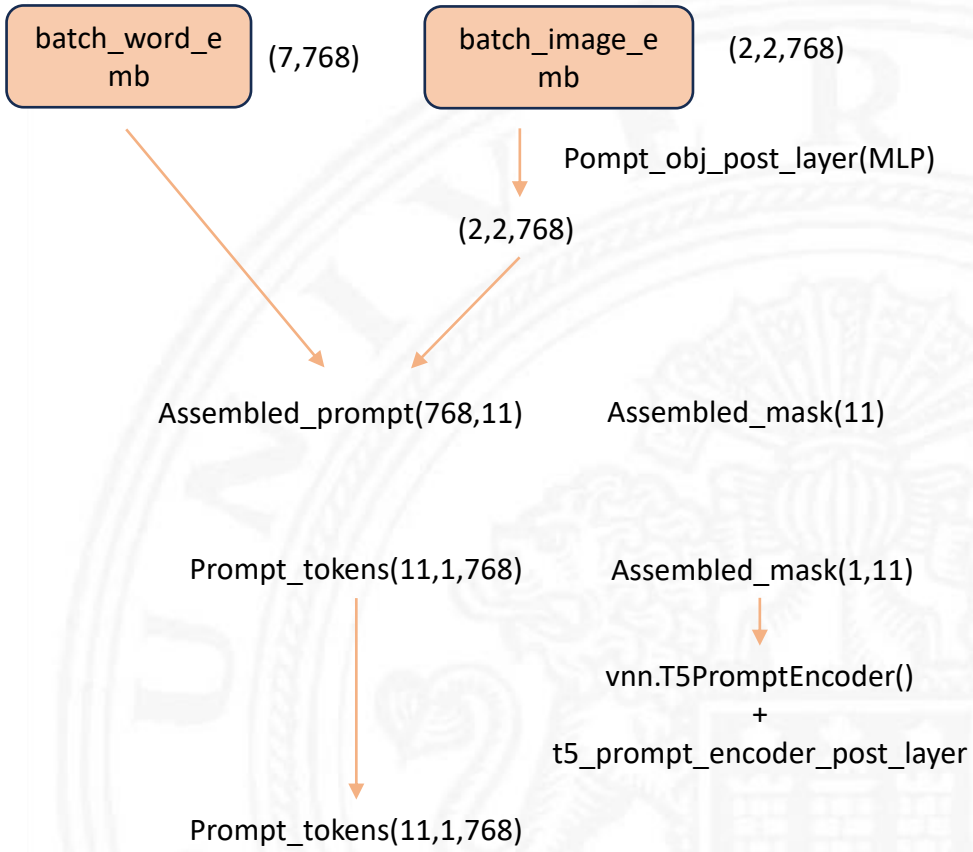
VIMA: General Robot Manipulation with Multimodal Prompts



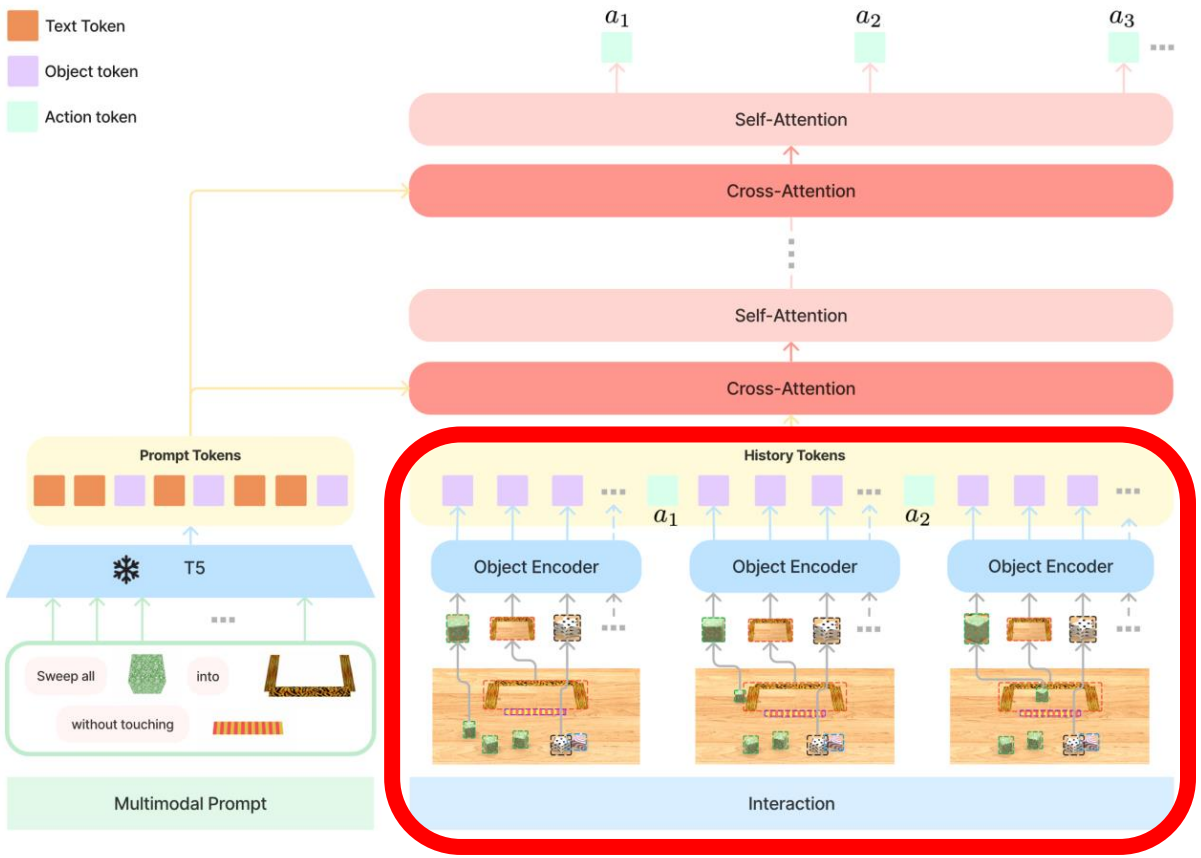


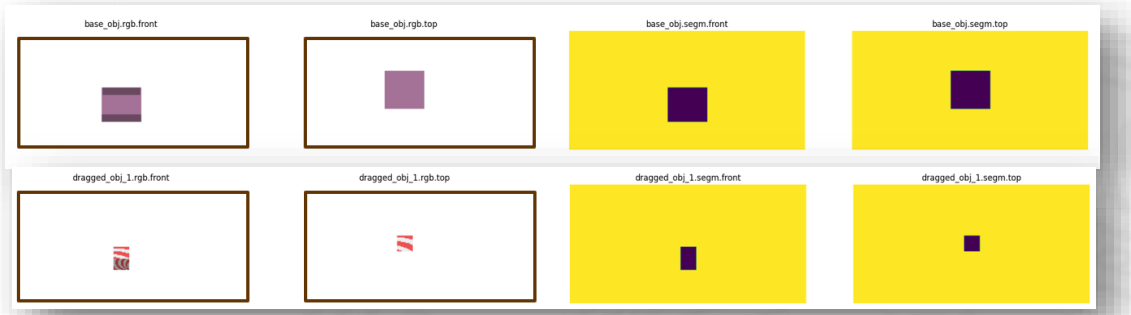
VIMA: General Robot Manipulation with Multimodal Prompts





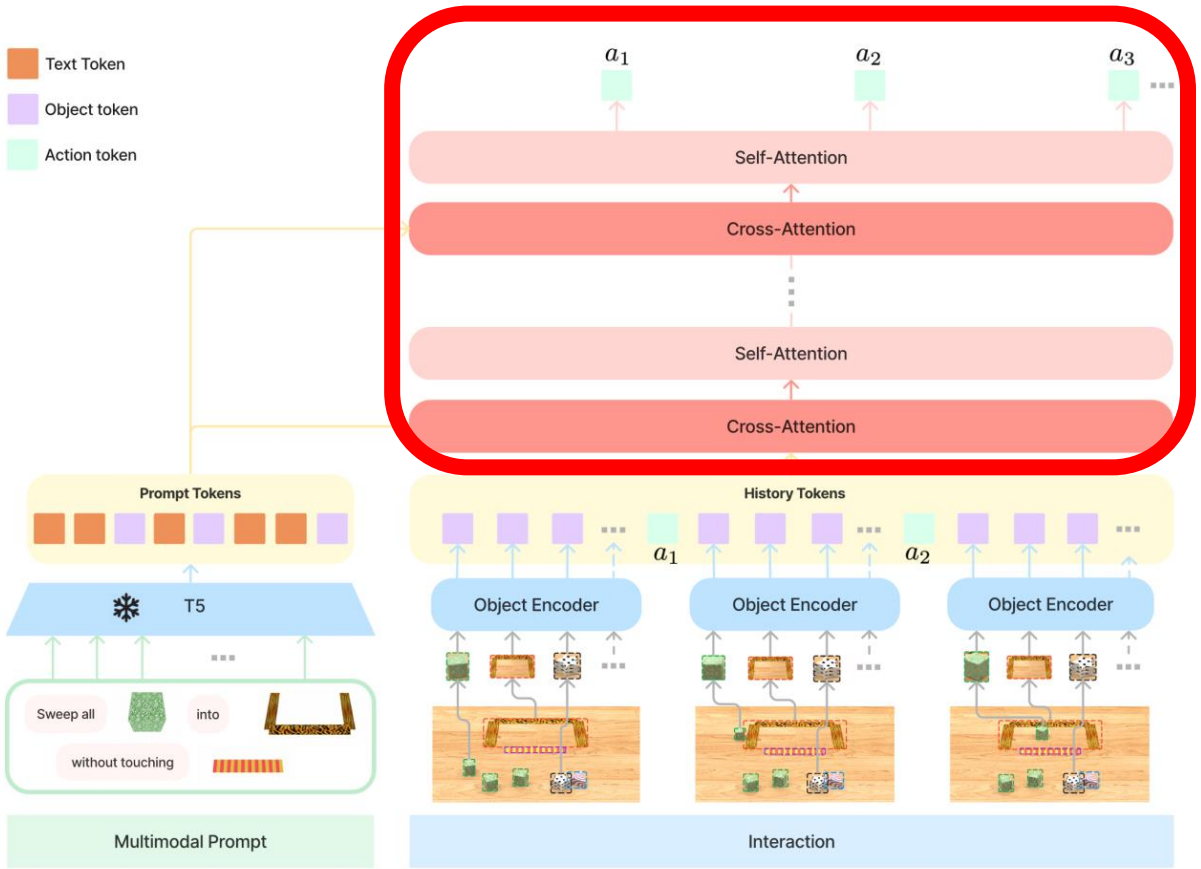
VIMA: General Robot Manipulation with Multimodal Prompts

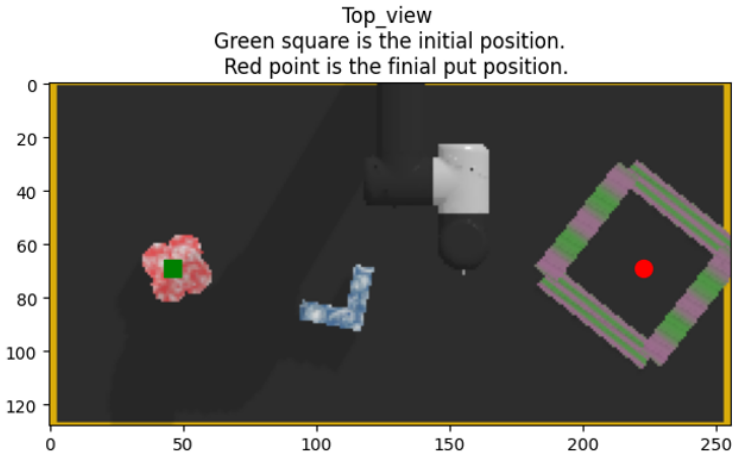




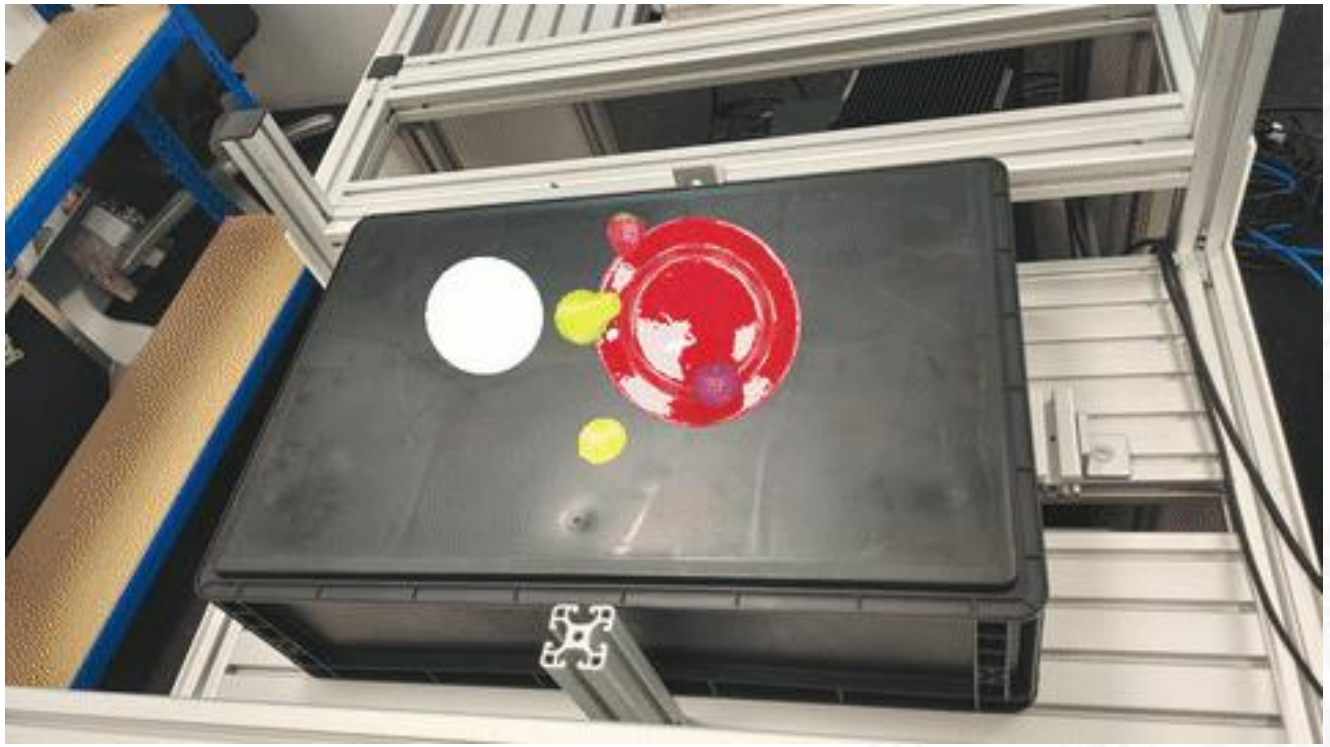
- Object encoder:
  - Cropped_img: ViT
  - Bbox: MLP
  - Fusion layers: MLP

VIMA: General Robot Manipulation with Multimodal Prompts





Model output: (Variables:
actions[“pose0_position”],
actions[“pose1_position”])
For initial position [0.50, 0.20]
For final put position [0.5, 0.90]



Thanks!