



VIMA

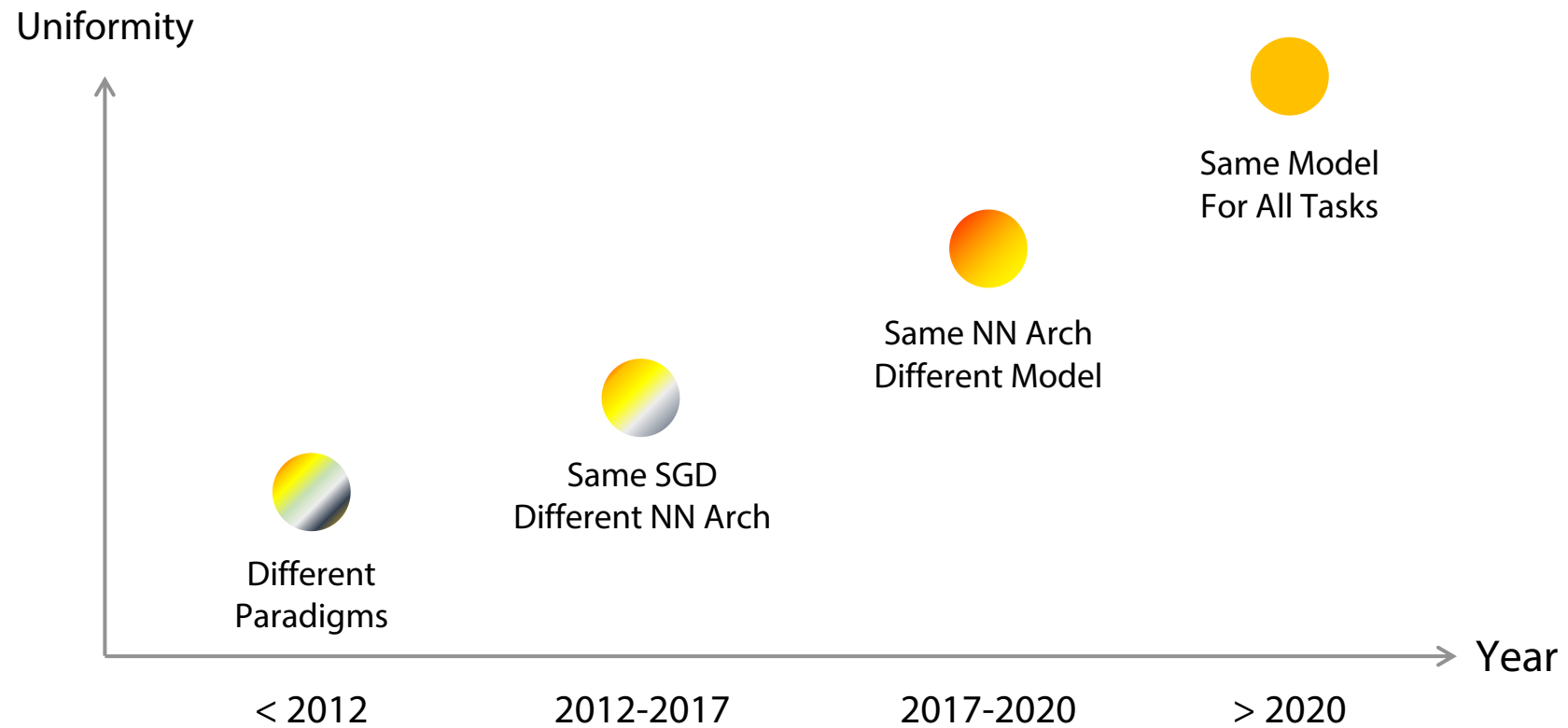
Robot Manipulation with Multimodal Prompts

Yunfan Jiang, Agrim Gupta⁺, Zichen “Charles” Zhang⁺, Guanzhi Wang⁺, Yongqiang Dou, Yanjun Chen,
Li Fei-Fei, Anima Anandkumar, Yuke Zhu[‡], Linxi “Jim” Fan[‡]

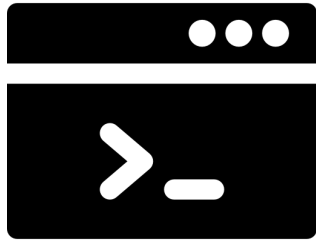


<https://vimalabs.github.io/>

History of AI is a history of unification



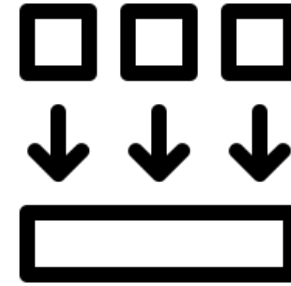
What's so good about it?



A single interface
for all scenarios



A single model
to learn them all



A better way
to **generalize**

Fragmented task specifications and tackling methods

SayCan
 "find a cleaner"
 "go to the trash can"
 "pick up the sponge"
 "try using the vacuum"

"Pick up the green cup"
 Command t
 Control Model R_{t+1} R_t Image I
 once per step once per task

Diagram showing a robot arm and a hierarchical task decomposition tree.



Instruction Following

Goal state Current state

Task Specification
 Geometric
 Language "Move toaster next to coffee maker"
 Image
 Experience

Diagram showing a 3D environment with a path and a task specification table.



Visual Goal: Rearrangement

Diagram showing a Transformer network processing video frames into a Spatial Syntax + MLP, which then feeds into a Control Network and Manipulation Network.



One-shot Video Demonstration

Safe Learning Controller
 Control Policy (Dec. 3.1) Safety Certificate and Filter (Dec. 3.2)
 Prior System Model Prior System Model
 Prior Cost Function Prior Safety Constraints
 Data Buffer & Learning Algorithm

Robot Operating Environment

Start Pose

Safe Learning $Q_{safe}(s_t, a_t) < \epsilon$

Unsafe Learning $Q_{safe}(s_t, a_t) > \epsilon$


Pre-training Phase
 1) Learn Q_{safe} π
 Evaluate π

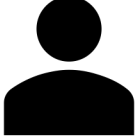
Fine-tuning Phase
 2) Learn $\pi \leftarrow \arg \max_{\pi} J_{\pi, target}$
 s.t. $Q_{safe} < \epsilon$



Constraint Satisfaction


What do we want for robot learning?

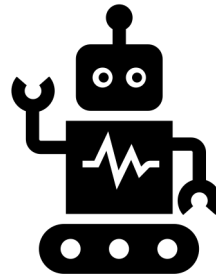
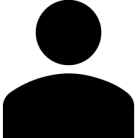
Bring me  from kitchen



Polish the floor like this: 



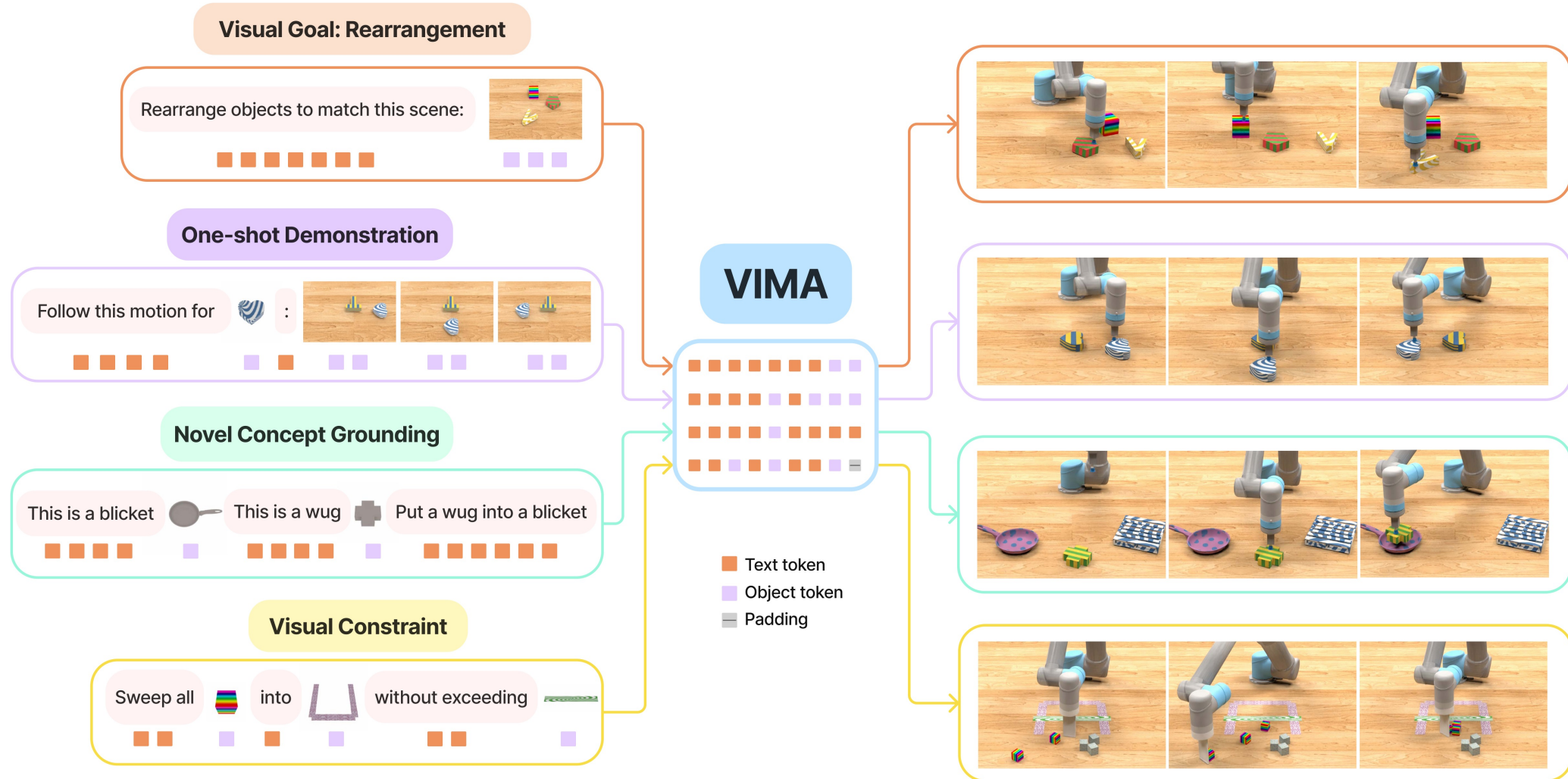
This is a sweeper 
You can use it to clean the table.



Do not enter this room: 



VIMA: Robot Manipulation with Multimodal Prompts



Various task specifications as multimodal prompts

Visual Goal Reaching:

Rearrange objects to match this scene:



Novel Concept Grounding:

This is a blicket



This is a wug



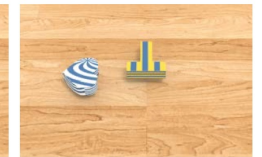
Put a wug into a blicket

One-shot Video Imitation:

Follow this motion for



:

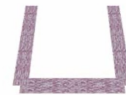


Visual Constraint Satisfaction:

Sweep all



into



without exceeding



...

A novel generalist agent for robot manipulation

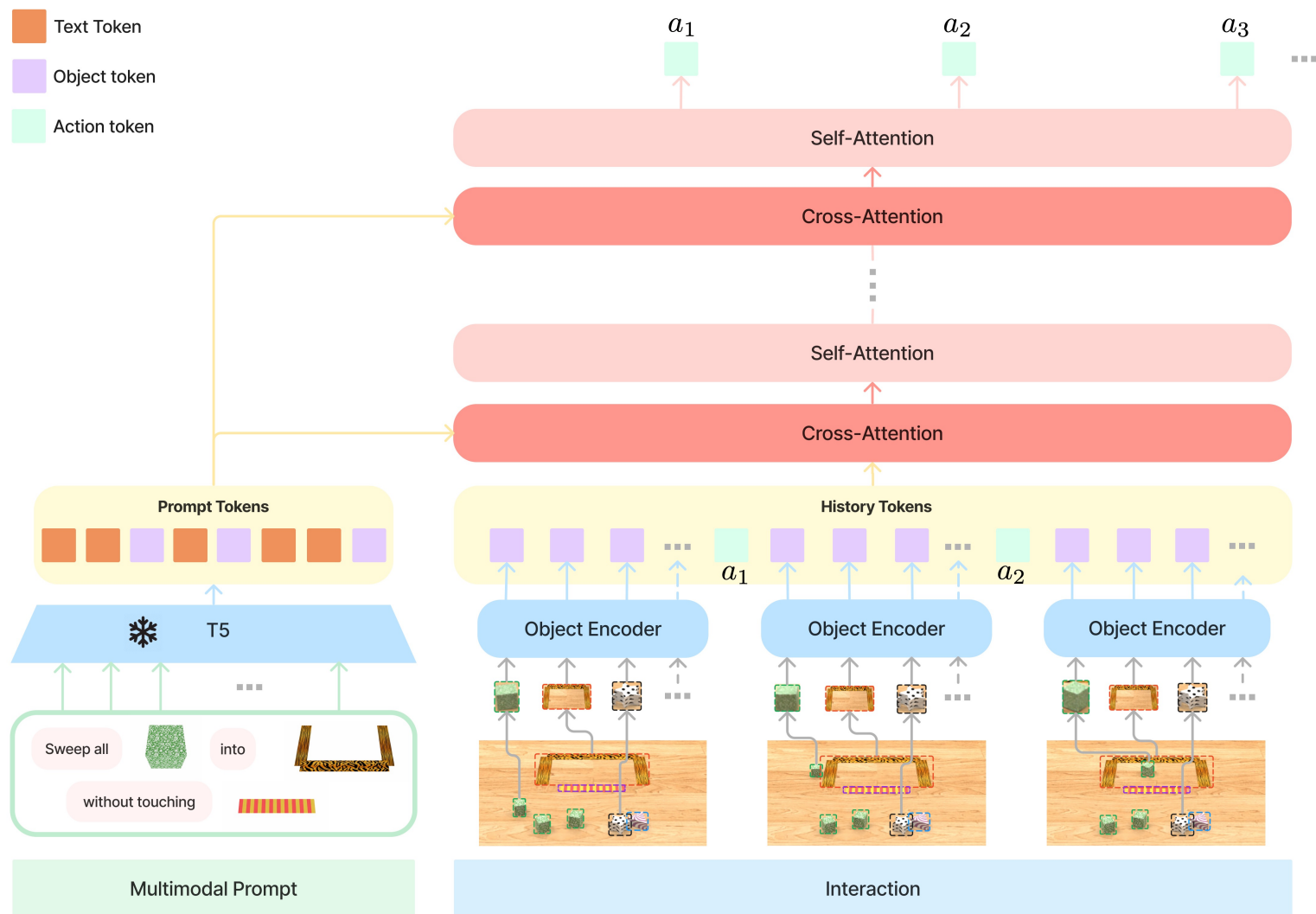
- Transformer encoder-decoder style
- Encode multimodal prompt with a pre-trained LM
- Decode robot arm action one step at a time



VIMA

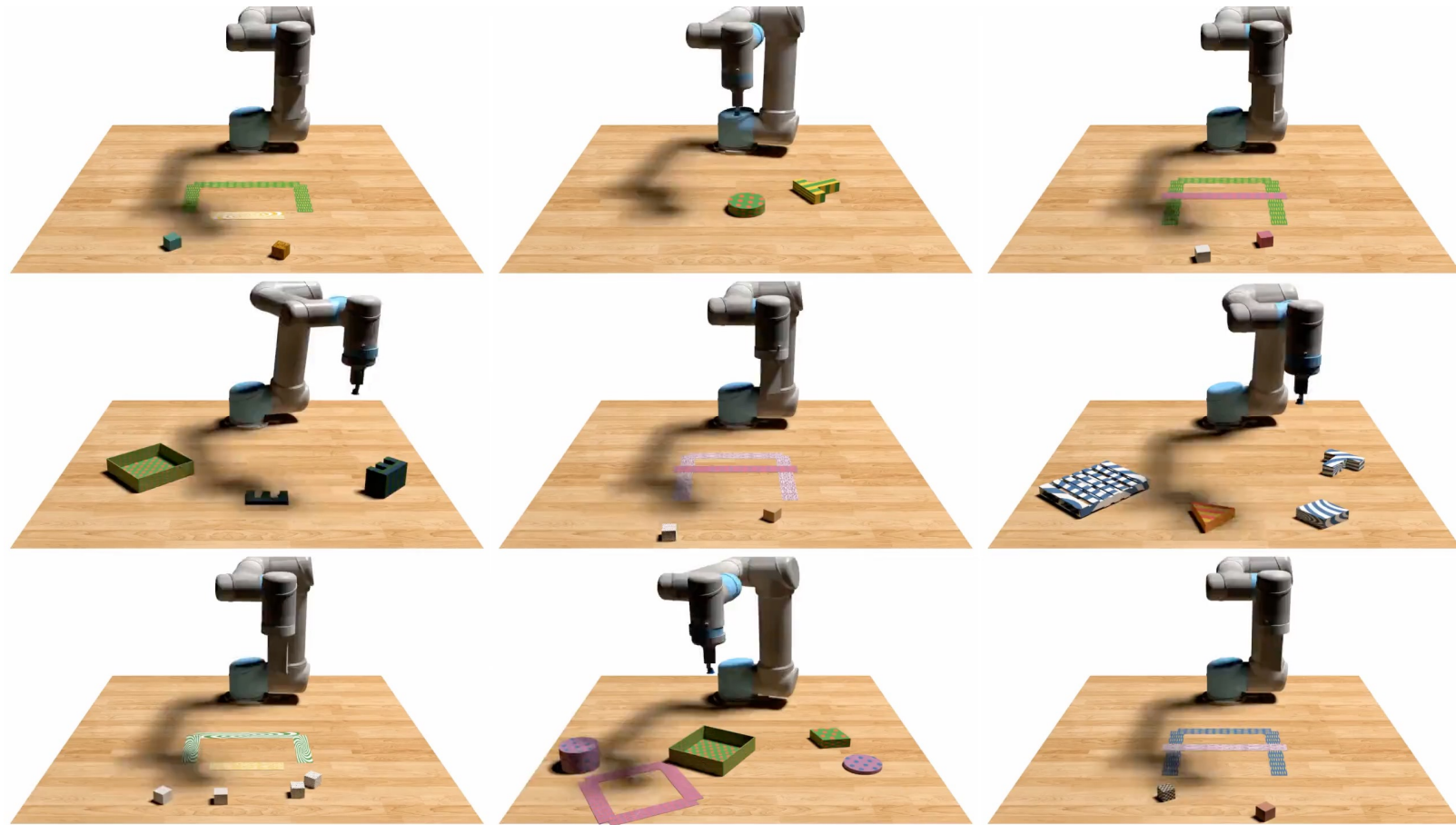
A novel generalist agent for robot manipulation

- Cross-attention to condition history on prompt
- Alternate cross-attention and causal self-attention to decode actions
- Object as tokens



A large-scale benchmark with multimodal prompts


- 17 task templates with multimodal prompts
- All templates are paired with thousands of procedurally generated multimodal prompts
- Scripted oracles to generate expert demonstrations

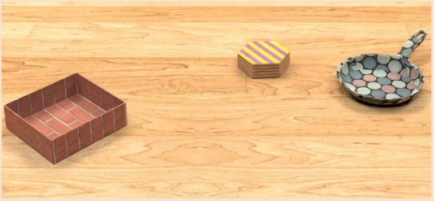


4 levels of generalization

Stronger Generalization

Training

Put the  into the 




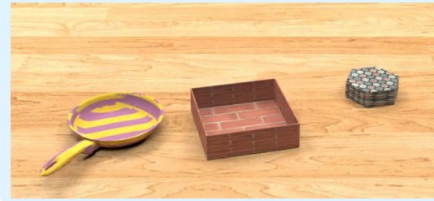
Level 1 Object Placement

Put the  into the 



Level 2 Novel Combination

Put the  into the 



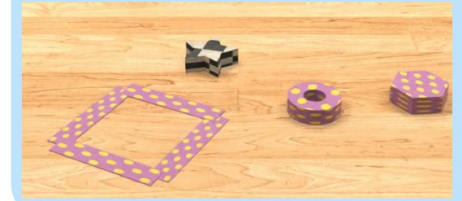
Level 3 Novel Object

Put the  into the 



Level 4 Novel Task

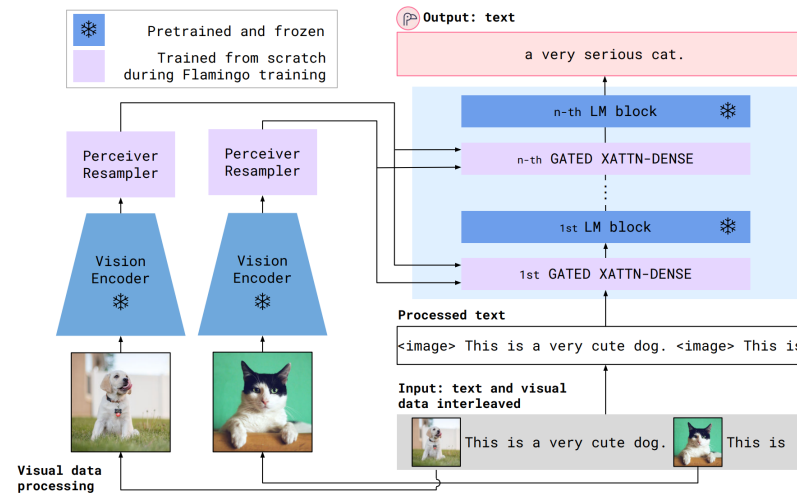
Put all objects with the same texture as  into it



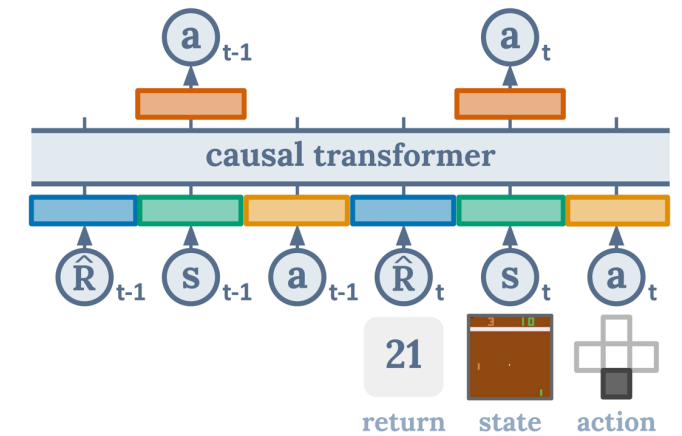
Baselines



VIMA-Gato

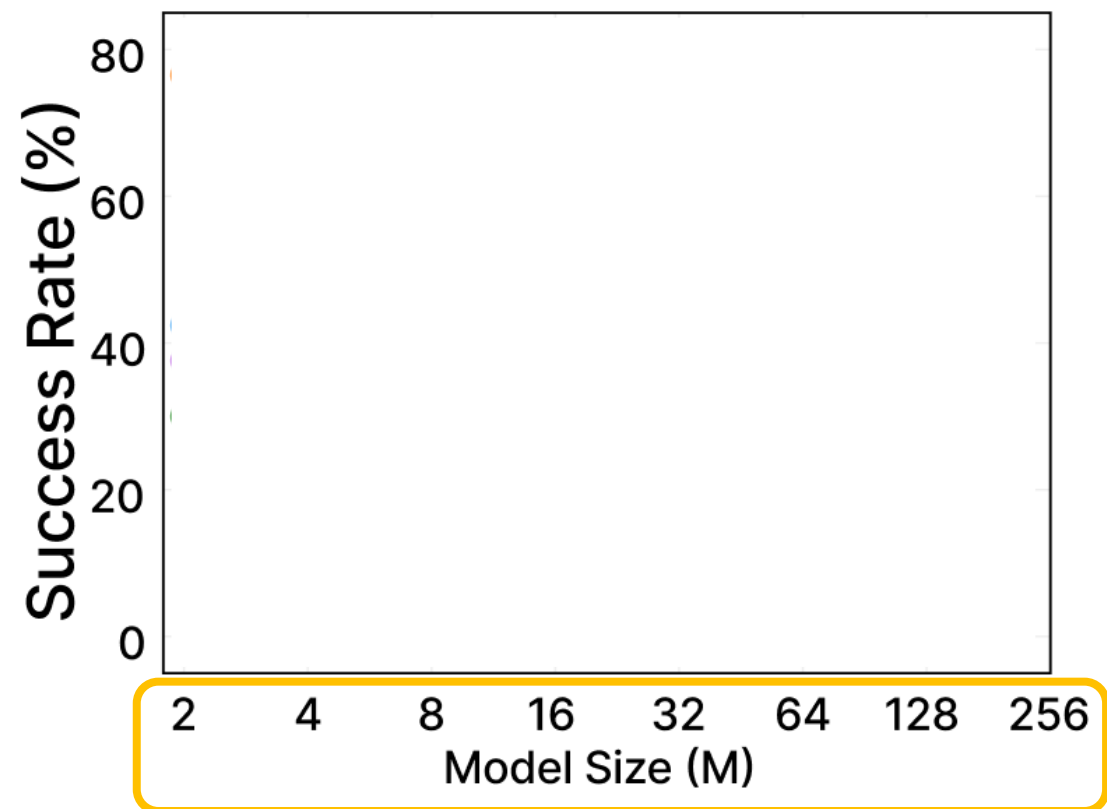


VIMA-Flamingo

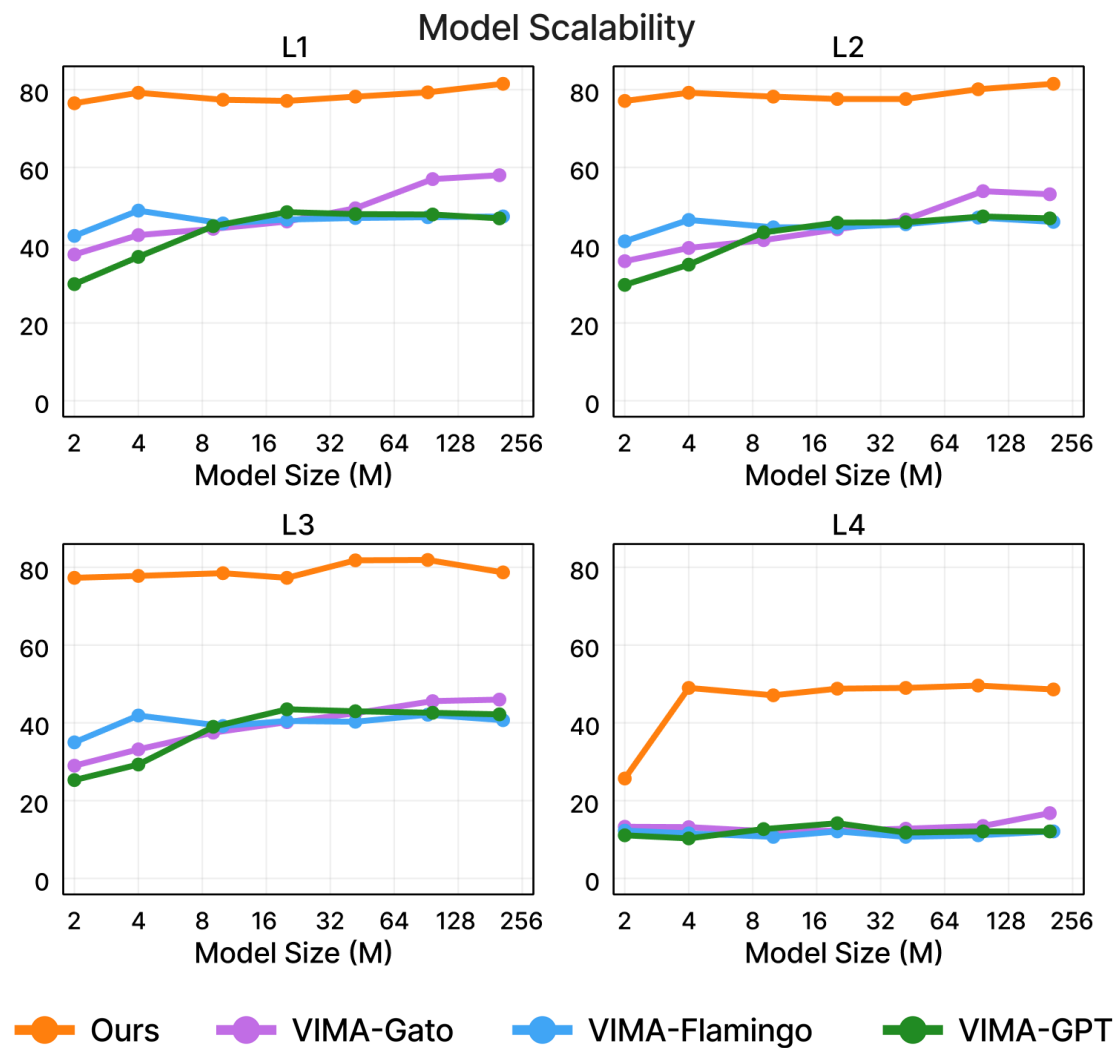


VIMA-GPT

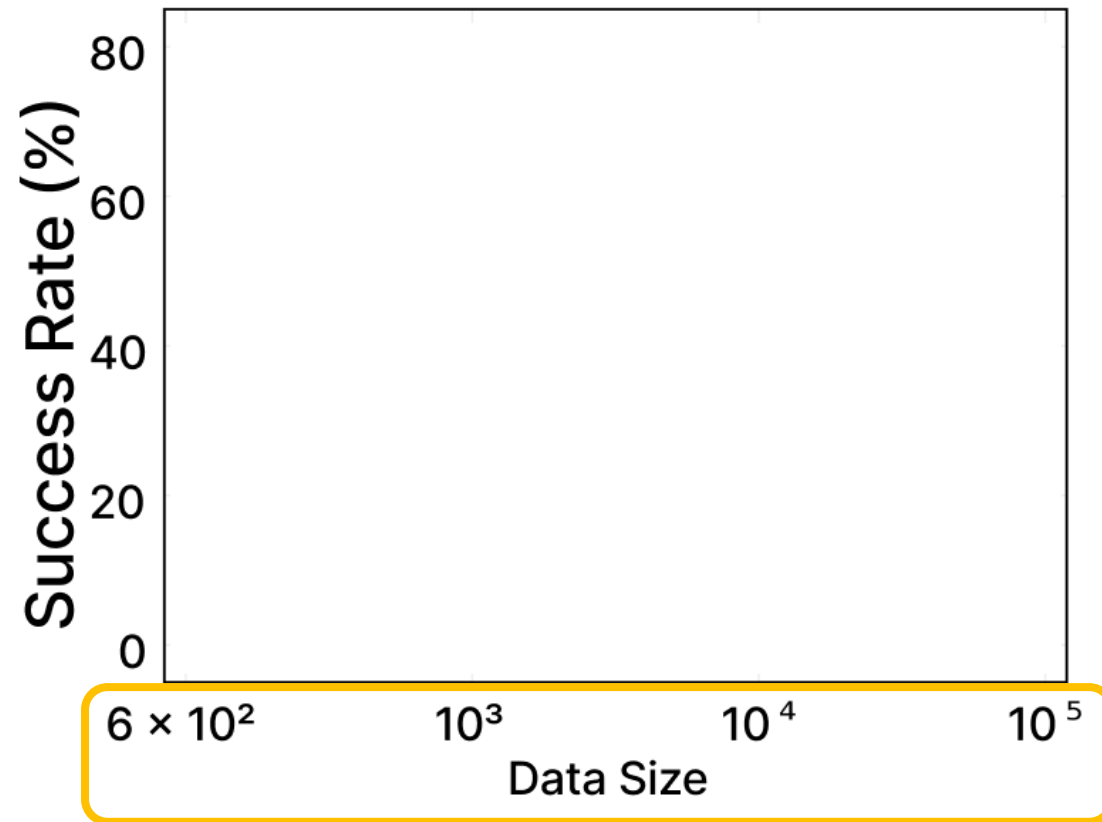
Model scalability



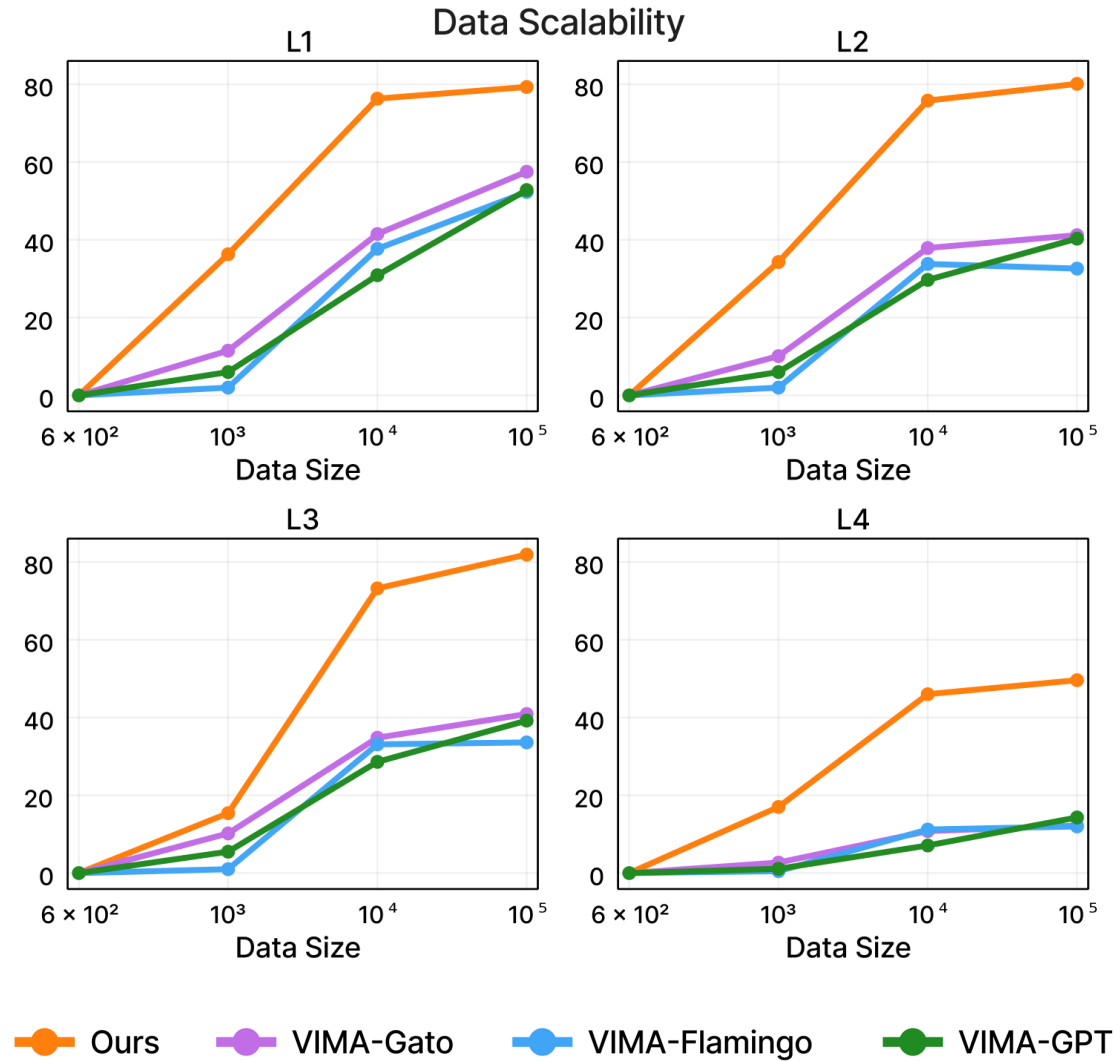
Model scalability



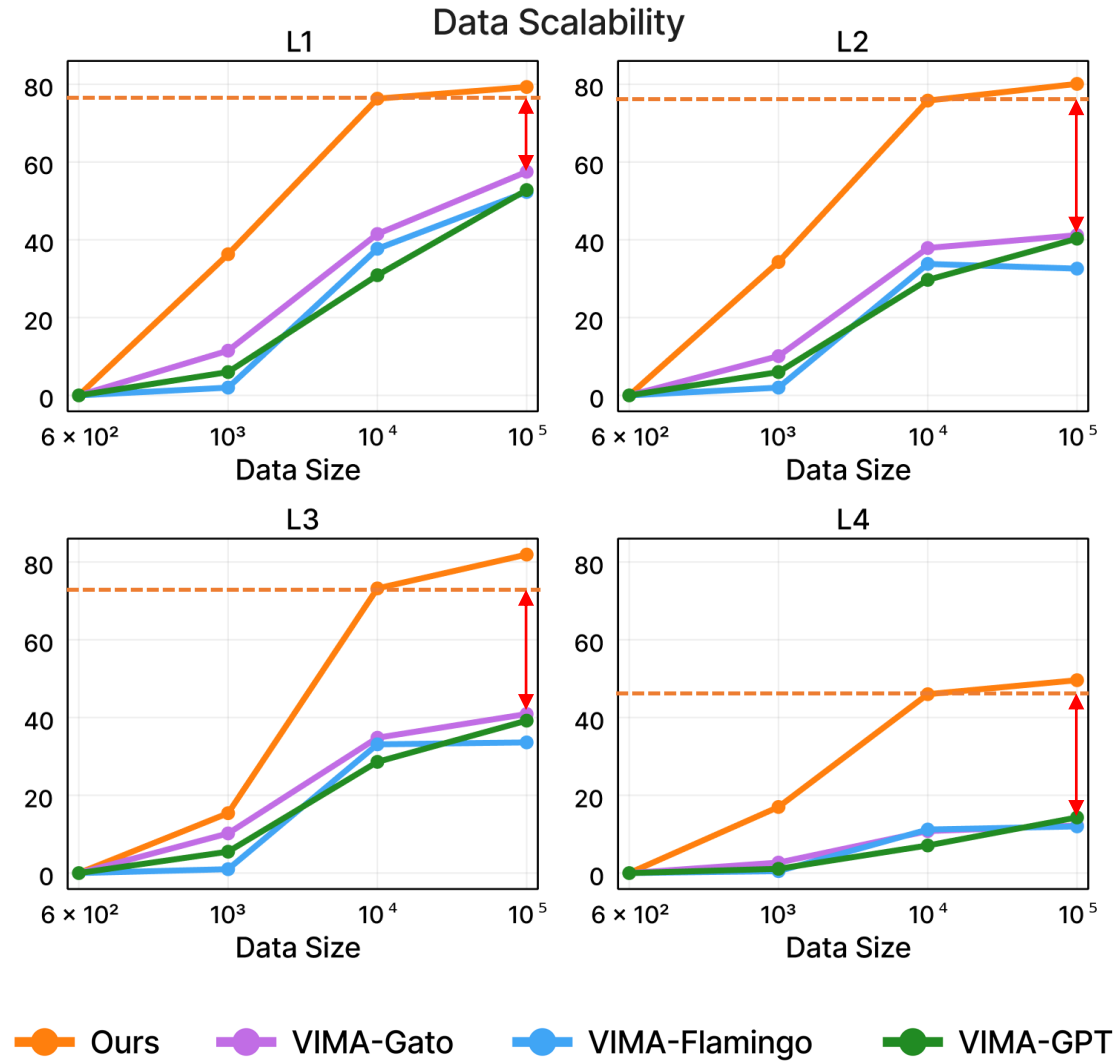
Data scalability



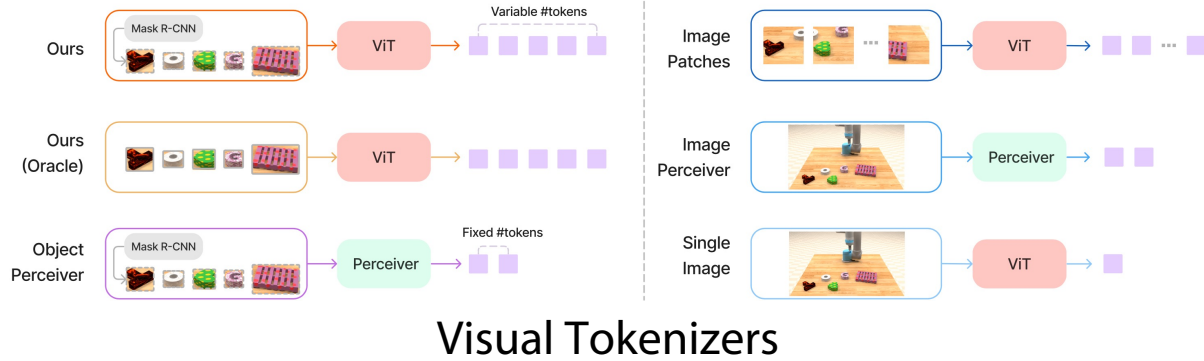
Data scalability



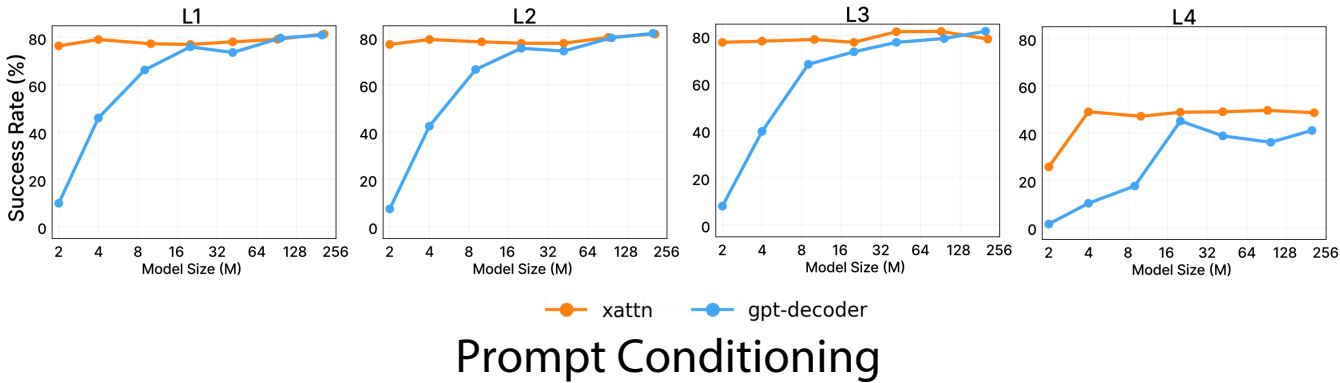
Data scalability



Why our recipe is so effective?



Visual Tokenizers



Prompt Conditioning

Table 13: Performances of our method with differently sized pre-trained T5 prompt encoder. We fix the parameter count of the decision-making part to be 200M.

	t5-small (30M)	t5-base (111M)	t5-large (368M)
L1	78.8	81.5	80.8
L2	79.0	81.5	81.0
L3	80.3	78.7	81.0
L4	49.1	48.6	49.3

Prompt Encoding

Table 14: Evaluation results on tasks with increased amounts of distractors. We fix the parameter count of the decision-making part to be 200M.

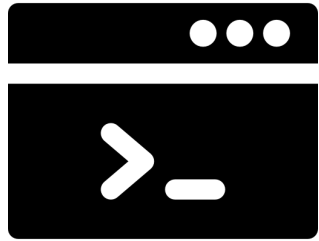
	L1	L2	L3	L4
Original	81.5	81.5	78.7	48.6
More Distractors	78.5	78.6	72.9	47.8
Relevant Performance Decrease (%)	3.6	3.5	7.3	1.6

Table 15: Evaluation results with incomplete and corrupted prompts. We fix the parameter count of the decision-making part to be 200M.

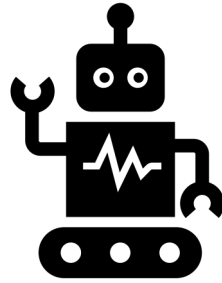
	L1	L2	L3	L4
Original	81.5	81.5	78.7	48.6
Incomplete Prompts	80.8	81.1	77.0	48.0
Corrupted Prompts	78.2	78.1	73.8	45.3
Relevant Performance Decrease w/ Incomplete Prompts (%)	0.8	0.4	2.1	1.2
Relevant Performance Decrease w/ Corrupted Prompts (%)	4.2	4.3	6.6	7.2

Policy Robustness

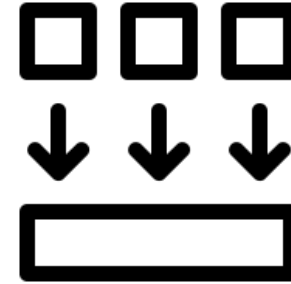
Take-home messages



Multimodal Prompting
for Unification



VIMA: Cross-Attention
+ Object Tokens



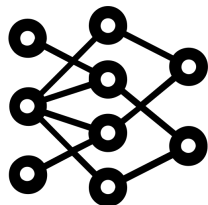
Good Generalization
& Sample-Efficient



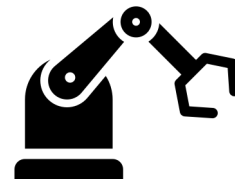
vimalabs.github.io



**Source
Code**



**Pretrained
Models**



**Simulation
Suite**



**Training
Dataset**

Thank you!