UC SANTA BARBARA
amazon
UC SANTA CRUZ

Mastering Robot Manipulation with Multimodal Prompts through Pretraining and Multi-task Fine-tuning

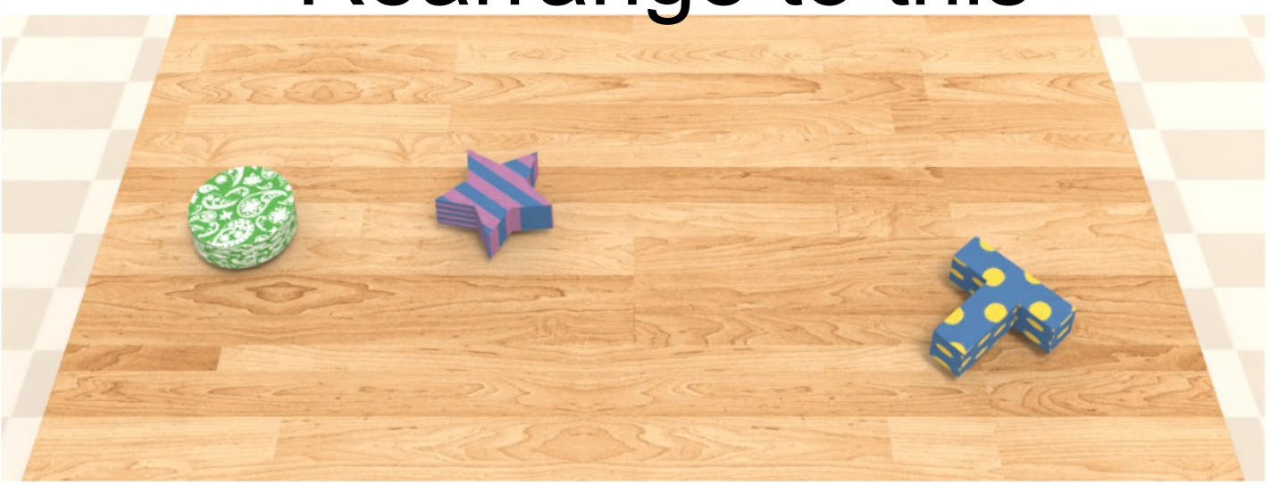
Jiachen Li[1], Qiaozi Gao[2], Michale Johnston[2], Xiaofeng Gao[2], Xuehai He[3], Hangjie Shi[2], Suhaila Shakiah[2], Reza Ghanadan[2], William Yang Wang[1]

ICML International Conference On Machine Learning

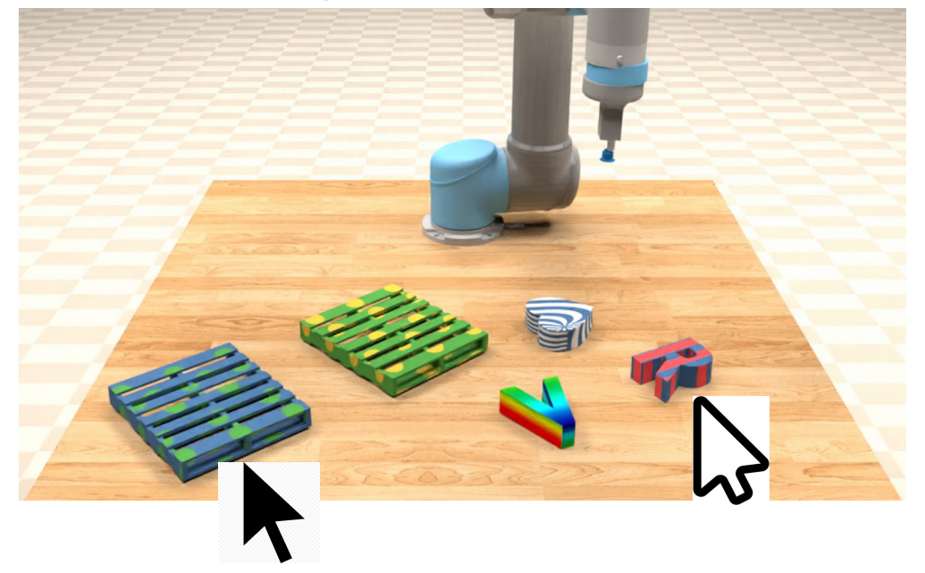
Project Page

Motivation

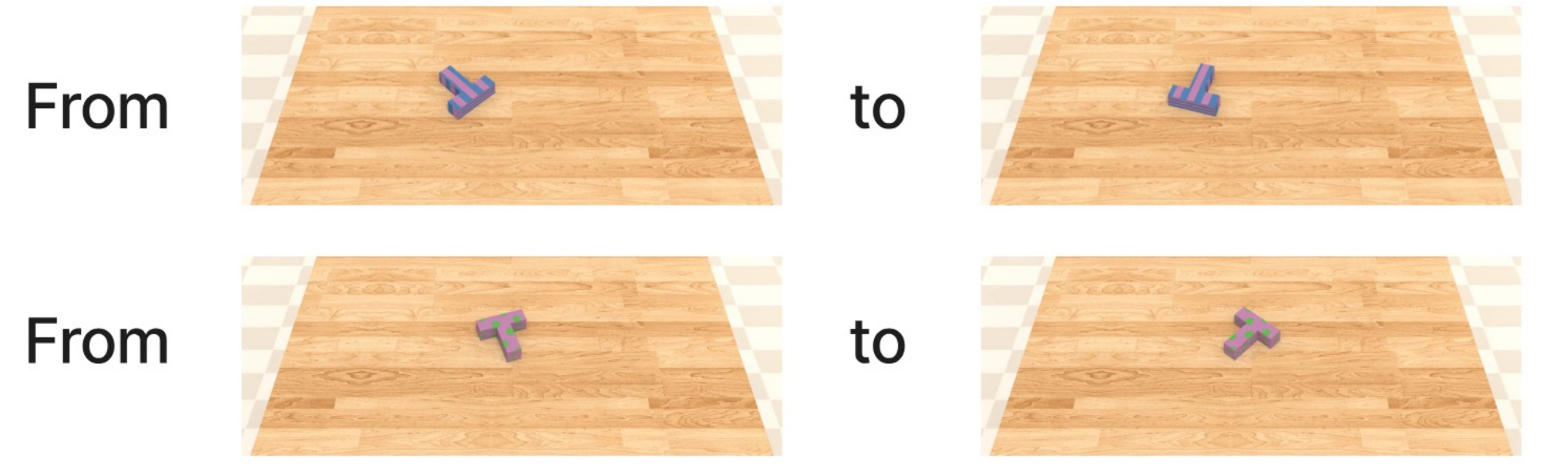
Rearrange to this



Put the object on the object



Twist is defined as rotating object a specific angle. For examples



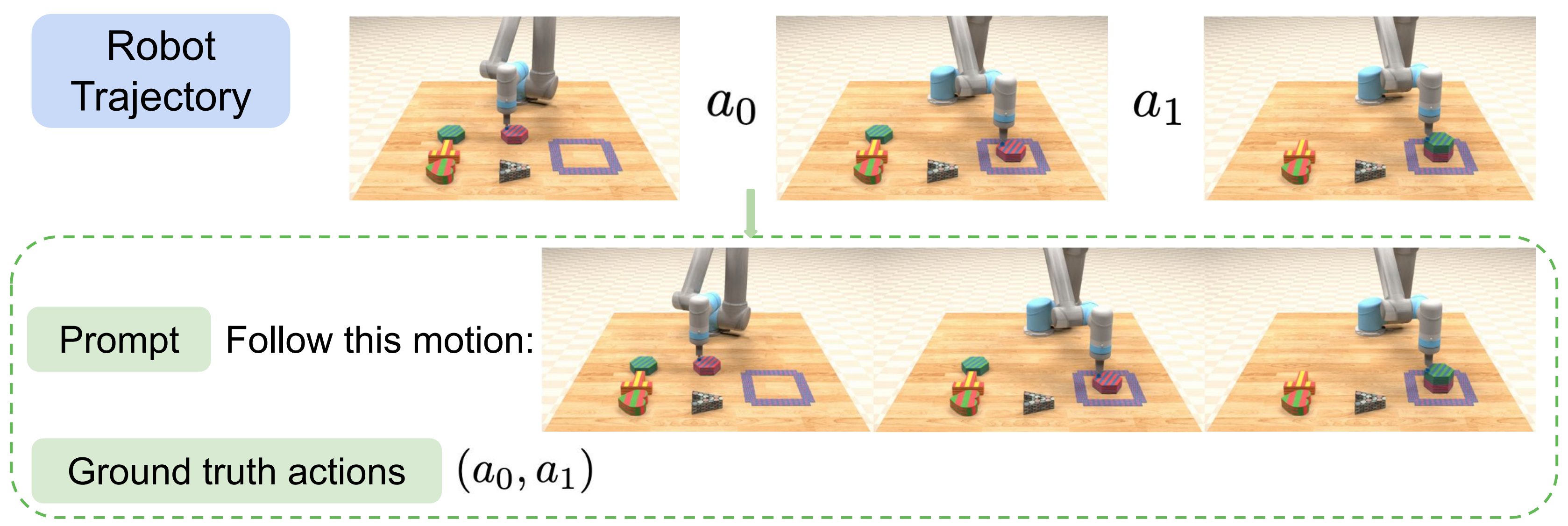
Now twist all yellow paisley objects

- Real-world embodied conversation will be multimodal.
- Pure language might not be expressive and precise enough to describe a task.
- A generalist robot should be able to learn from in-context demonstrations.

Challenges from multimodal prompts

- A robot must understand the underlying transition dynamics suggested by the prompts.
- Imitation Learning falls short in teaching robots to understand inverse dynamics, as **future observations are often masked out** when training to predict actions from the history.

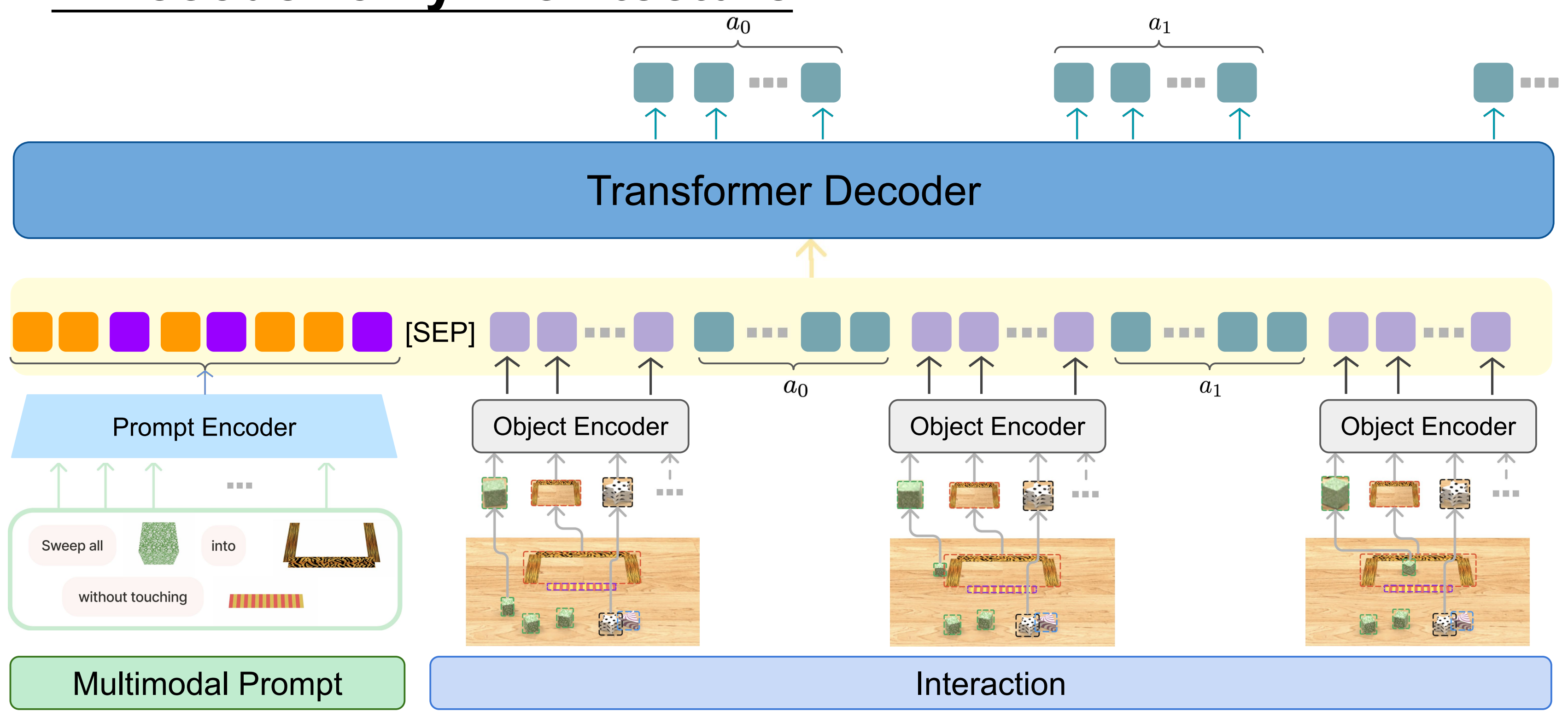
MIDAS: a Two-Stage Training Pipeline



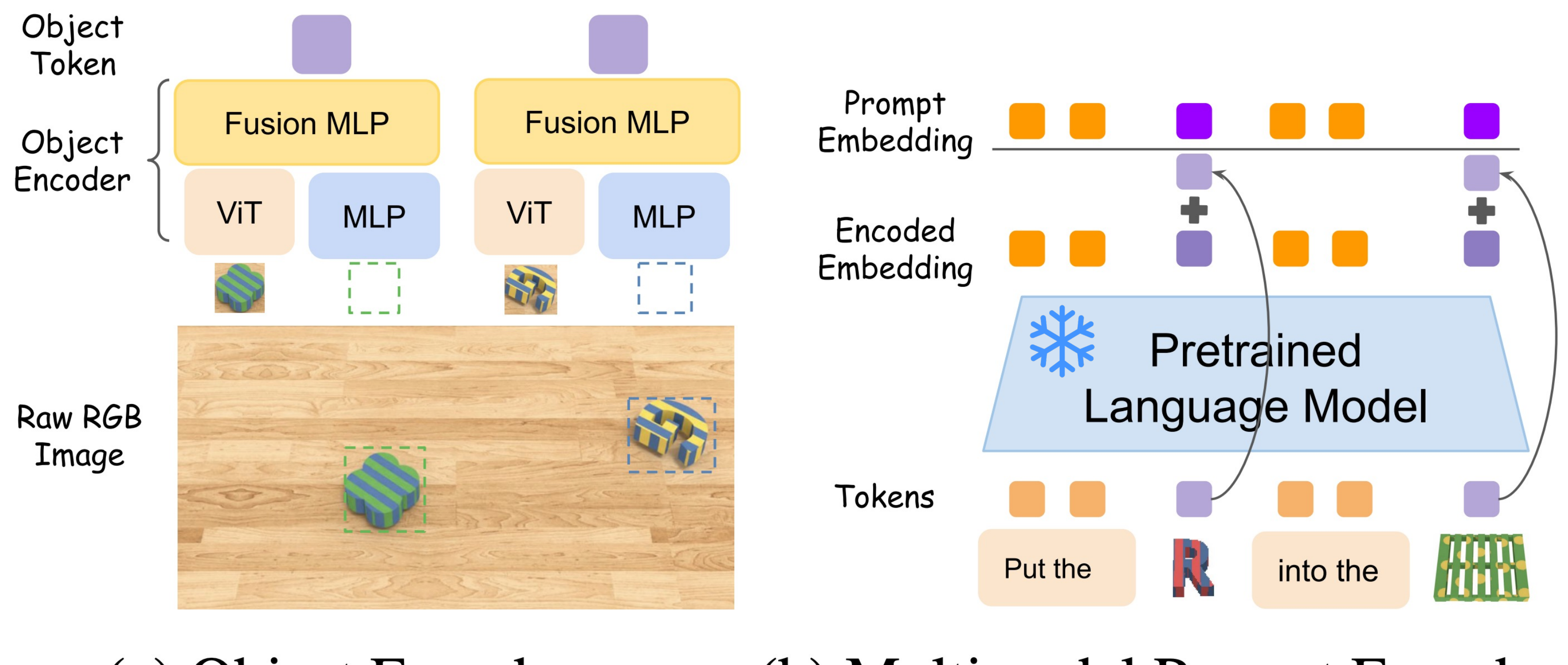
- Given any sequence of robot trajectory, we can always create a motion-following task.
- Our two-stage MIDAS training pipeline: 1) Inverse Dynamic Pretraining + 2) Multitask Finetuning

MIDAS Model Architecture

A Decoder-only Architecture

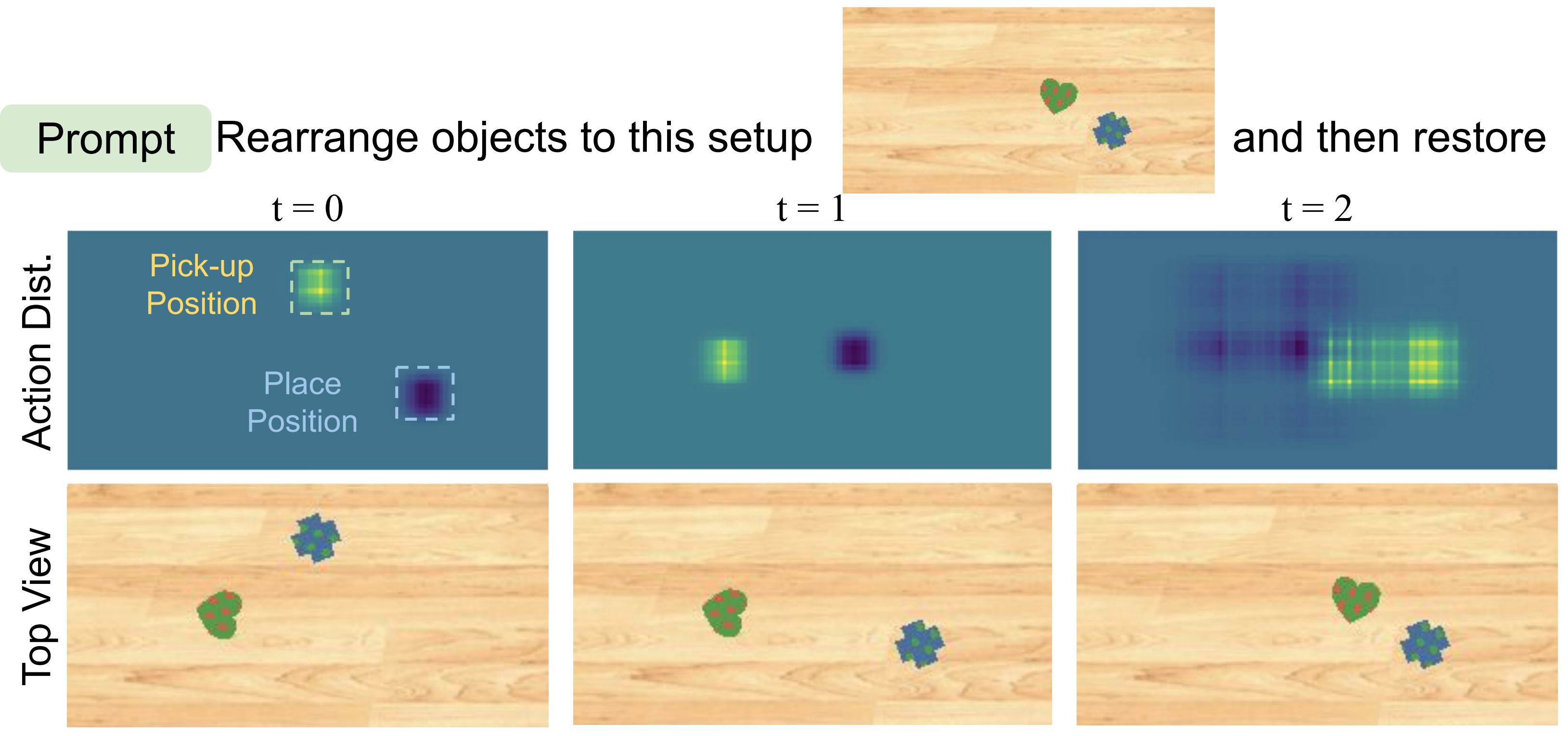


Multimodal Prompt Encoder: Augment pretrained LM with a residual connection to the input object token



(a) Object Encoder (b) Multimodal Prompt Encoder

Model each action dimension as an individual token



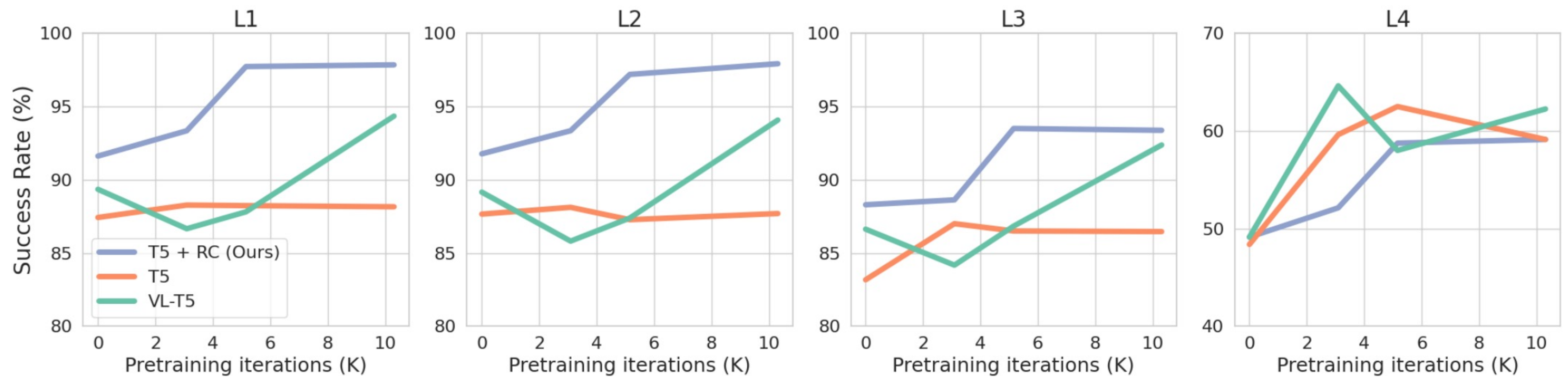
- At t = 2, the robot should move either the heart or the cross block.
- Independently predicting each action dimension can results in task failure.

Main Results on VIMA-BENCH

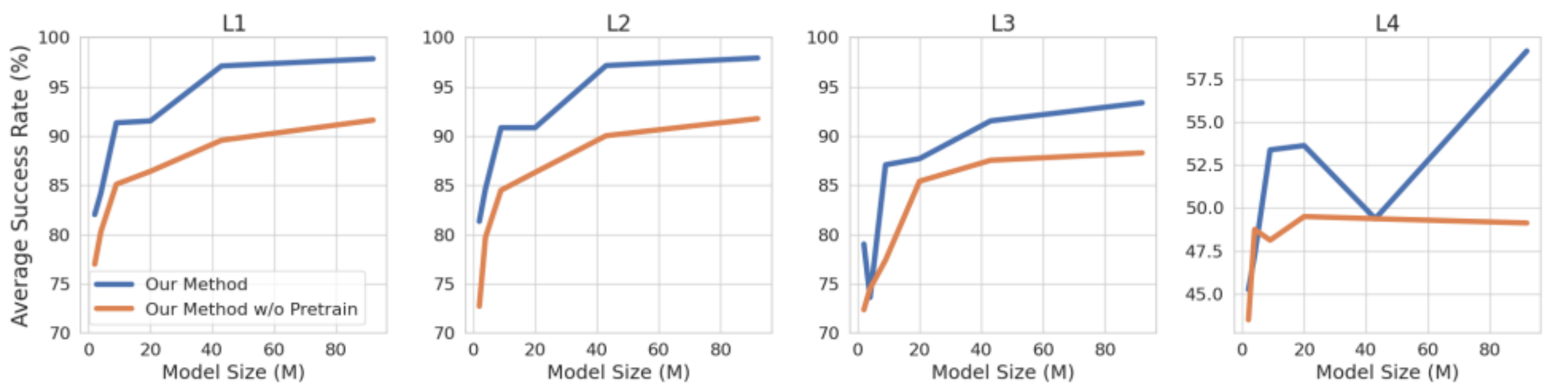
Method	L1					L2					L3					L4	
	Avg	T5	T9	T16	T17	Avg	T5	T9	T16	T17	Avg	T5	T9	T16	T17	Avg	T10
Gato	57.0	44.5	14.0	43.0	1.5	53.9	46.0	10.5	42.0	1.0	45.6	36.0	17.0	41.5	0.0	13.5	0.0
Flamingo	47.2	41.0	3.0	38.0	2.0	47.1	43.0	4.5	40.0	1.0	42.1	36.5	6.0	45.5	0.5	11.1	0.0
GPT	47.9	45.0	8.0	33.0	1.0	47.4	43.0	10.5	34.0	3.0	42.6	32.0	5.0	37.5	0.0	12.1	0.5
VIMA	87.2	65.0	13.5	88.0	77.0	87.0	61.0	12.5	87.5	77.5	84.0	63.0	12.0	**58.5**	78.0	49.6	0.0
Gato OBJ	87.5	62.0	17.0	92.5	80.5	87.5	62.5	16.0	91.5	80.0	84.4	65.5	15.5	46.5	87.5	49.6	0.0
Ours																	
w/o Pretrain	91.6	88.0	20.5	93.0	**98.0**	91.8	87.0	23.5	92.0	**98.0**	88.3	90.0	20.5	50.5	**99.5**	49.1	0.0
w/ Pretrain	**97.8**	**94.0**	**100**	**94.0**	96.5	**97.9**	**96.5**	**100**	**93.0**	96.0	**93.4**	**94.0**	**97.0**	47.0	**98.0**	**59.1**	**41.0**

Ablation Study

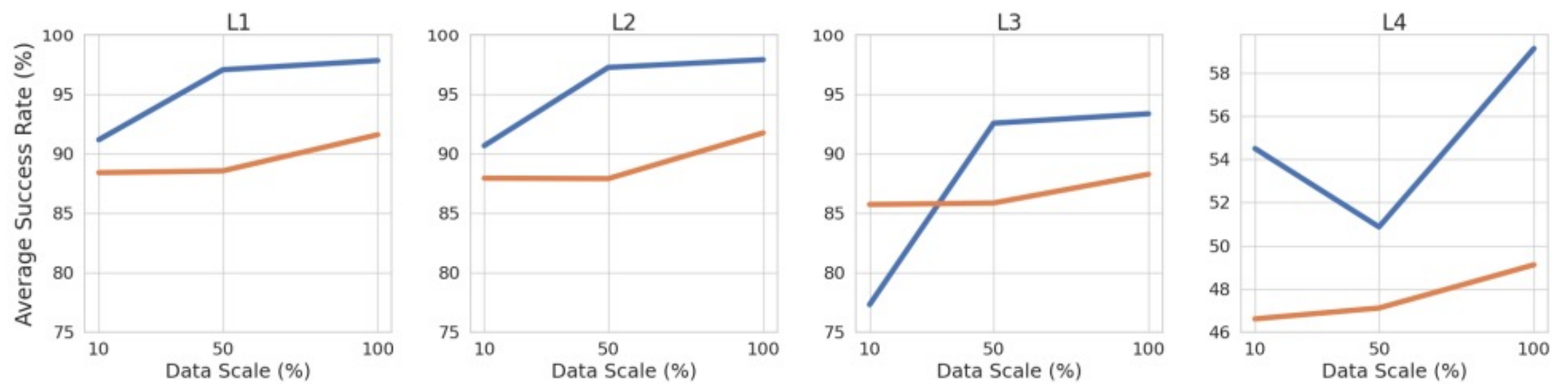
Prompt Encoder



Model Size

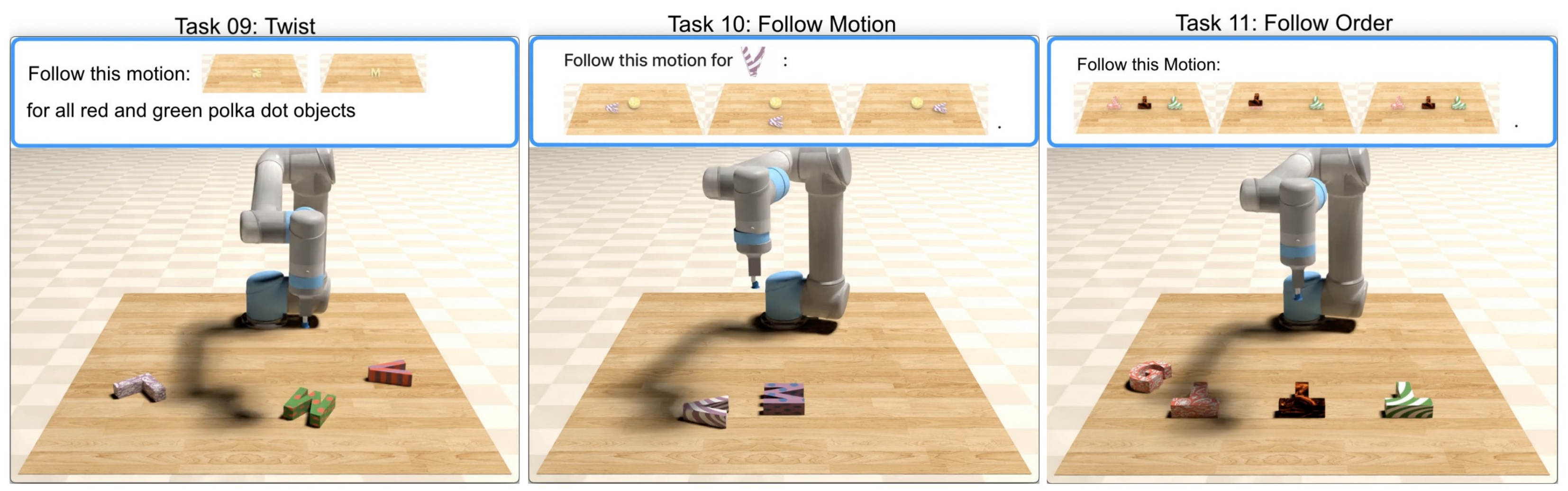


Data Scale



Evaluate In-Context Learning Ability

Modify and **exclude** the tasks below from training



Task	T9: Twist	T10: Follow Motion	T11: Follow Order	Overall
Our Method	**26.5%**	**74.0%**	8.0 %	**36.2%**
Our Method w/o Modified FT	10.0%	43.5 %	**16.5%**	23.3%
Our Method w/ Pretrain Only	8.0 %	2.0%	15.5 %	8.5 %
Our Method w/o Pretrain	1.5 %	0.5 %	0.0%	0.7 %