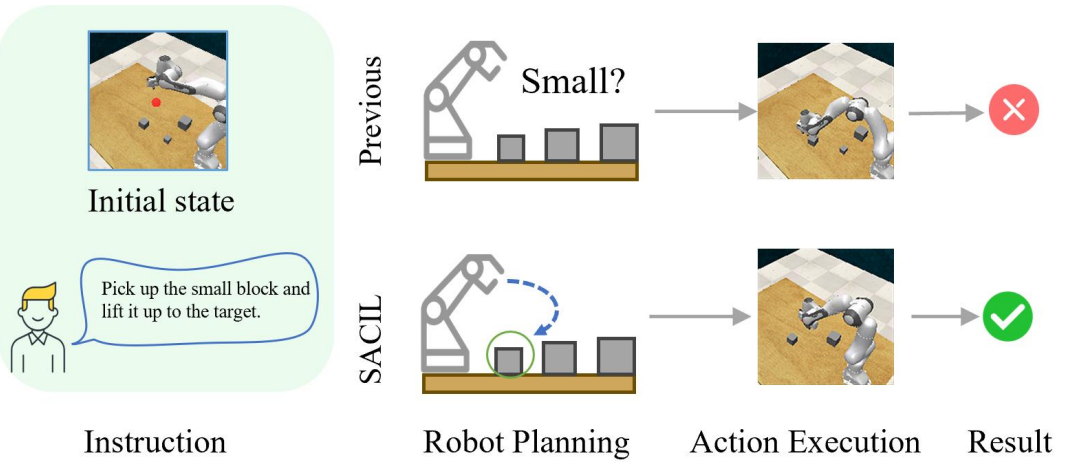
Size-aware Contrastive Imitation Learning for Language-conditioned Multi-task Robotic Manipulation

BMVC 2025

Introduction

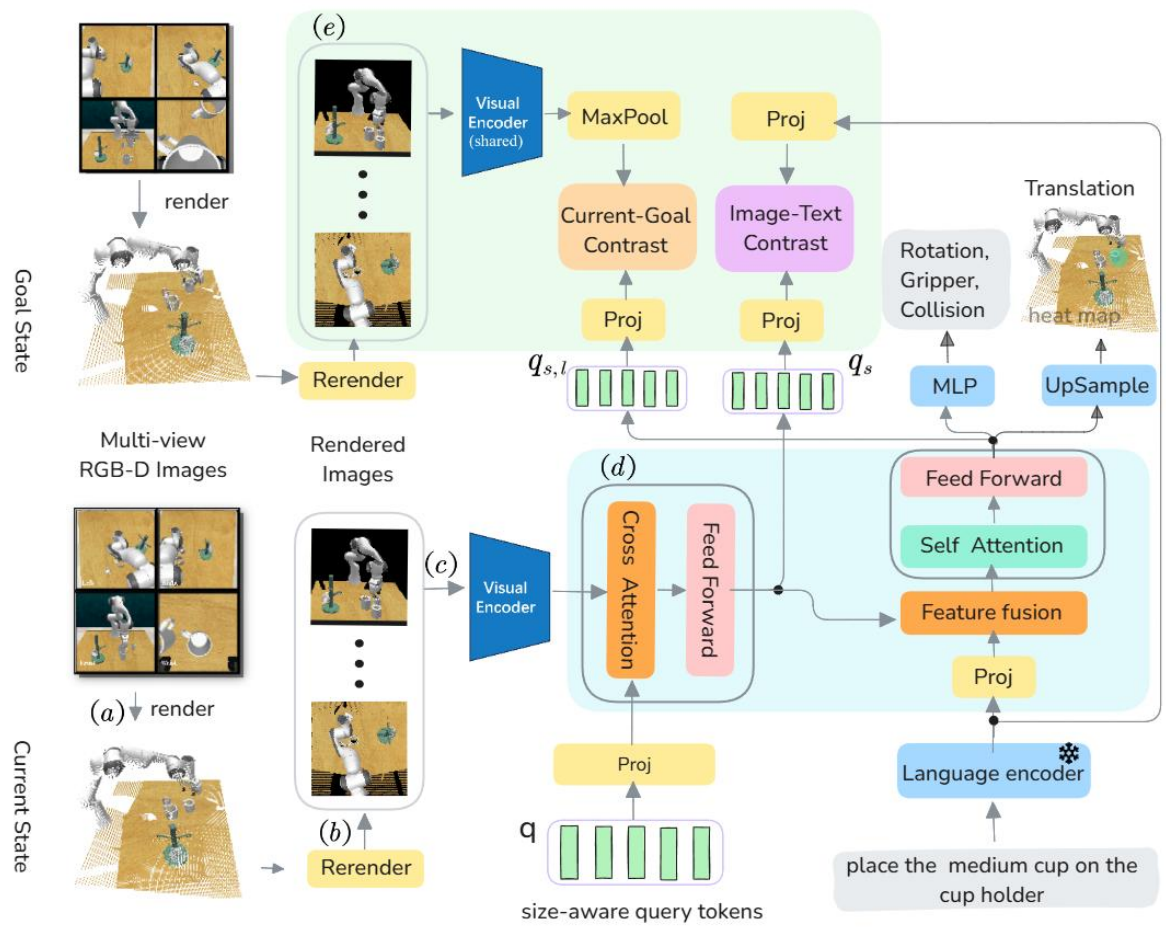
Challenge: Previous approaches have primarily focused on enhancing visual precision and integrating language embeddings, often overlooking the alignment of subtle geometric features with linguistic representations, struggling with tasks that are sensitive to size and spatial constraints.



contributions

- We propose SACIL, a novel contrastive imitation learning framework that explicitly models object scale in language-conditioned manipulation.
- We introduce dual contrastive learning—Image-Text Contrast and Current-Goal Contrast—that jointly align linguistic, visual, and temporal signals for robust multi-step execution.
- We design size-aware query tokens that serve as explicit carriers of scale-related geometric features.
- We construct a benchmark dataset tailored for size-sensitive language-conditioned manipulation, filling the gap in standardized benchmarks for evaluating manipulation tasks that require scale awareness.

Method



Ablation Study

Size-Aware query tokens	place cups	push buttons	put items in drawer	take off peppers	Avg. Suc
✗	12.0	80.0	77.3	5.3	43.7
✓	13.3	85.3	77.3	6.7	45.7

Comparison of size-aware query tokens on four sequential dual-target tasks.

image-text contrast	current-goal contrast	pick and lift	pick up cup	place cup	put in drawer	take off pepper	Avg. Suc
✗	✗	90.7	81.3	10.7	62.7	1.3	49.3
✓	✗	88.0	85.3	14.7	64.0	13.3	53.1
✓	✓	94.7	88.0	16.0	70.7	18.7	57.6

Comparison of different contrastive Imitation Learning methods on five single-targetasks.

Comparision with the SOTA Methods

Models	Avg. Success	Train Time	place cups	push buttons	put items in drawer	take off peppers
HiveFormer[11]	29.4	~11h	0.0	83.3	14.1	**20.0**
PolarNet[3]	2.5	~12h	0.0	0.0	10.0	0.0
Act3D[7]	0.0	~11h	0.0	0.0	0.0	0.0
RVT[10]	40.7	~12h	6.7	**86.7**	61.3	8.0
SACIL	**45.7**	~20h	**13.3**	85.3	**77.3**	6.7

Single-targe tasks performance evaluated RLBench

Models	Avg. Success	Train Time	pick and lift	pick up cup	place cup	put item in drawer	take off pepper
HiveFormer[11]	42.0	~11h	91.7	**90.0**	6.7	20	1.6
PolarNet[3]	35.4	~8h	**96.0**	70.8	0.0	10.0	0.0
Act3D[7]	44.0	~12h	90.7	82.7	**18.7**	9.3	18.7
RVT[10]	49.3	~12h	90.7	81.3	10.7	62.7	1.3
SACIL	**57.6**	~20h	94.7	88.0	16.0	**70.7**	**18.7**

Sequential dual-target tasks evaluated on RLBench

Case Study