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

Jiakai Huang[1]
huangjiakai@m.scnu.edu.cn
Weiping Zheng* [2]
zhengweiping@scnu.edu.cn

[1] School of Artificial Intelligence,
South China Normal University,
Foshan, China

[2] School of Computer Science,
South China Normal University,
Guangzhou, China

Abstract

Translating high-level linguistic instructions into geometrically consistent manipulation actions is a critical challenge in robotics, especially when dealing with objects of diverse geometric properties, such as size and shape. This task demands both linguistic comprehension and precise geometric reasoning from the robot agent. Previous approaches have primarily focused on enhancing visual precision and integrating language embeddings, often overlooking the alignment of subtle geometric features with linguistic representations. In this paper, we propose Size-Aware Contrastive Imitation Learning (SACIL), a novel framework that addresses this gap through two key components: Image-Text Contrast and Current-Goal Contrast. These components ensure the alignment of language embeddings with geometric features and maintain temporal consistency in size-aware reasoning across multi-step tasks. Additionally, we introduce a set of size-aware query tokens to effectively aggregate geometric features. Our experimental results demonstrate that SACIL significantly outperforms state-of-the-art methods, highlighting its potential to enhance size-sensitive reasoning and advance language-conditioned robotic manipulation.

1 Introduction

Humans effortlessly perform a wide range of everyday manipulation tasks—from assembling furniture to organizing kitchenware—by seamlessly integrating an implicit understanding of object attributes such as size, shape, and material with natural language instructions. However, replicating this capability in language-conditioned robotic systems remains a fundamental challenge. Achieving such generalizable manipulation requires not only the ability to parse linguistic commands and perceive complex 3D scenes, but also to ground abstract language concepts into geometrically actionable representations. While recent advances in robotic perception have significantly improved visual feature learning [3, 6, 15], a critical

* denotes the corresponding author.

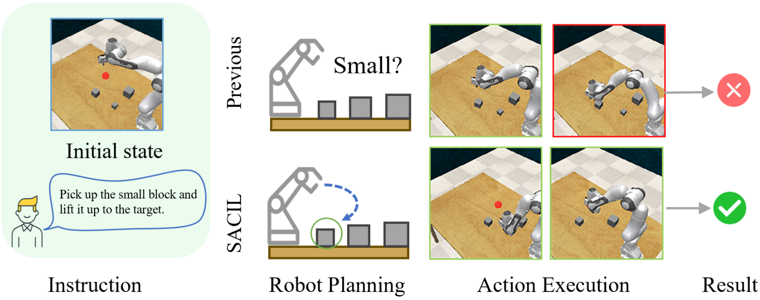

Figure 1: The previous methods (such as Act3D[6]) misinterpret scale semantics. However, our method introduces size-aware contrastive imitation learning to effectively resolve ambiguities.

gap persists: existing methods still struggle to align linguistic descriptors with geometric features, particularly in tasks that are sensitive to size and spatial constraints.

Prior research has predominantly focused on enhancing visual perception through 3D representations, leaveraging voxel grids [11, 13, 31], point clouds [3, 6, 15, 41], or multi-view renderings [7, 9]. For instance, C2FARM[13] dynamically refines voxel grids using attention mechanisms to prioritize task-critical 3D regions. PerAct processes RGB-D inputs into 3D voxel representations and utilizes Percevier [11] Transformer to encode features. Morover, Act3D[6] uses a 3D feature field and coarse-to-fine sampling to compute high-resolution 3D action maps. Despite achieving geometric fidelity and 3D pose prediction, existing methods fail to align linguistic descriptions with object size attributes, leading to prevalent size ambiguity issues. As shown in Figure 1, while conventional approaches struggle with handling small object, our method precisely manipulates it.

To address these limitations, we propose **S**ize-**A**ware **C**ontrastive **I**mitation **L**earning (**SACIL**), a unified framework designed to bridge the gap between language and geometry through dual contrastive alignment. Our key insight is that robust language-conditioned manipulation requires not only spatial grounding of size-related attributes but also temporal consistency across multi-step tasks. SACIL introduces two core components: (1) Image-Text Contrast, which dynamically aligns language embeddings with scale-aware visual features, and (2) Current-Goal Contrast, which promotes consistency between the agent's current observations and its future subgoals in long-horizon sequences. In addition, we design size-aware query tokens that explicitly aggregate geometric cues essential for resolving size ambiguities. Experiments on RLBench demonstrate the effectiveness of SACIL, achieving a 57.6% average success rate on single-target tasks and 45.7% on sequential dual-target tasks, significantly outperforming existing baselines in both short- and long-horizon settings.

Our contributions are summarized as follows: 1) We propose SACIL, a novel contrastive imitation learning framework that explicitly models object scale in language-conditioned manipulation. 2) 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. 3) We design size-aware query tokens that serve as explicit carriers of scale-related geometric features. 4) 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.

2 Related Work

2.1 Language-guided Robotic Manipulation

Language-guided policy learning has gained significant attention in robotics due to its potential for convenient human-robot interaction and skill generalization across diverse tasks [2, 3, 6, 9, 10, 21, 22, 29, 30, 31, 32, 35, 39]. Diverse benchmarks have been curated to benchmark the language-conditioned manipulation[12, 18, 24, 38, 40, 43]. In this work, we adopt RLBench[12] due to its comprehensive coverage of object attributes (e.g., pose, shape, color, size, and category) and challenging task scenarios.

Prior studies have explored vision-language fusion for robotic policy learning, yet critical gaps persist in geometric-language alignment. For example, transformer-based methods like RT-1[1] process multi-modal inputs (visual and linguistic) through a pretrained FiLM EfficientNet model to predict actions. While effective for simple tasks, RT-1's reliance on 2D RGB images inherently limits its capacity for 3D geometric reasoning, particularly for size-sensitive operations. Its successor, RT-2 [44] bridges vision and language by projecting visual tokens into a linguistic embedding space and using LLMs for action generation. However, this linguistic projection risks discarding fine-grained geometric features.

To address 3D perception, recent works adopt RGB-D inputs. Methods such as IFOR[8] have demonstrated potential in simple pick-and-place tasks but remain restricted to tabletop, top-down scenarios, failing to generalize to complex 3D environments. To fully utilize the 3D information in observations, studies like C2F-ARM[13] and PerAct[31] have used 3D voxel grids to represent the robot's workspace.While enabling precise 3D pose prediction through high-resolution voxelization. Despite improved spatial awareness, these methods incur high computational costs and overlooks the alignment between language instructions and visual scale attributes. In contrast, SACIL introduces Image-Text Contrast to linguistic instructions with geometrical feature.

Besides the voxelized features, policy learning based on 3D point cloud representation has gained significant attention[3, 6, 35]. For example, Act3D [6] leverages RGB-D-derived point clouds with CLIP [28]embeddings, employing cross-attention between scene features and action proposals. However, it lacks dynamic adaptation of geometric representations based on real-time feedback, relying on fixed CLIP embeddings and cross-attention mechanisms. Our work follows RVT[9] and Σ-agent[22], which re-render virtual view images from reconstructed 3D point clouds and processes the images using a Transformer network. Different from the above methods, SACIL actively aggregates multi-view geometric cues by size-aware query tokens.

2.2 Contrastive Learning in Sequential Decision-Making

A central challenge in sequential decision-making, encompassing both Reinforcement Learning (RL) and Imitation Learning (IL), is the development of effective representations from high-dimensional inputs like images. Contrastive self-supervised learning[17, 23, 25, 26, 27, 33] has emerged as a powerful technique to address this challenge in both paradigms. Lately, unifying representation and reinforcement learning objectives has become a research focus in RL[4, 14, 16, 27, 34, 36, 42]. For example, MT-Opt [14] unifies these objectives to efficiently acquire diverse robotic skills through shared representations and task impersonation. Similarly, Contrastive UCB [27] integrates contrastive learning directly into online RL to enable provably efficient exploration in MDPs and Markov games. Furthermore, Sta-

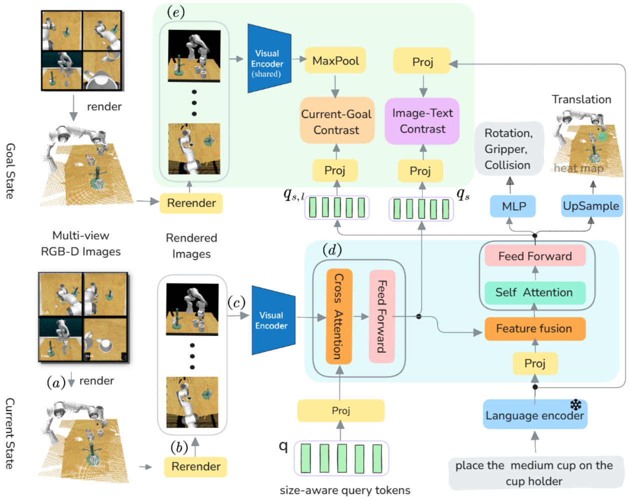

Figure 2: **Pipeline Overview.** (a) The method first takes multi-view RGB-D inputs and generates 3D point clouds. (b) These point clouds are then rendered into orthographic views. (c) A visual encoder is used to encode the orthographic views. (d) The encoded features are processed by the SADQ-Transformer. (e) Finally, dual contrastive learning aligns current-goal states and image-text pairs to predict 7-DoF actions.

ble Contrastive RL[42] demonstrates that a standalone contrastive objective can facilitate effective goal-conditioned RL from offline data with proper design choices. However, these existing contrastive RL methods primarily target reward-driven policy optimization and often overlook the nuanced role of language in grounding geometric features. To address this gap, our work, SACIL introduces dual contrastive imitation learning to align linguistic instructions with geometrically consistent actions.

3 Methods

3.1 Problem Definition

Language-conditioned robotic manipulation requires agents to translate abstract instructions into geometrically consistent actions, a critical challenge when handling objects with diverse properties (e.g., size, shape). Our objective is to train robots to perform manipulation tasks specified by natural language instructions, grounded in geometric and perceptual observations. This problem can be formulated as a language-conditioned Markov decision process(MDP) denoted by $(\mathcal{S}, \mathcal{A}, \mathcal{L}, \mathcal{T}, \gamma)$, with state space $\mathcal{S}$, action space $\mathcal{A}$, instruction space $\mathcal{L}$, transition dynamics $\mathcal{T} : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$, and discount $\gamma \in [0, 1)$. The goal is to learn a policy $\pi(a_t|s_t, l)$ that, given an instruction $l \in \mathcal{L}$, predicts an action $a_t \in \mathcal{A}$ based on the current state $s_t \in \mathcal{S}$ at each time step t.

In this work, the state s_t consists of aligned RGB-D images from four viewpoints: frontal, left shoulder, right shoulder, and wrist. Following RVT [9], we utilize orthographic virtual views from the RGB-D inputs as model inputs. The action space $\mathcal{A}$ is composed of translation $a_{trans} \in \mathbb{R}^3$, rotation $a_{rot} \in \mathbb{R}^{(360/5)\times 3}$, openness $a_{open} \in [0, 1]$ and collision avoidance

$a_{col} \in [0, 1]$.

Under imitation learning settings [31], we optimize the policy via maximum likelihood estimation on expert demonstrations $\mathcal{D} = \{(s_t, a_t, l)\}$. The behavior cloning loss is defined as:

$$\mathcal{L}_{\mathcal{BC}} = -\mathbb{E}_{(s_t, a_t, l) \sim \mathcal{D}} [\log \pi_\theta (a_t | s_t, l)] \tag{1}$$

3.2 Overall Pipeline

The overall pipeline of our framework is illustrated in Figure 2, which integrates multi-view perception, hierarchical feature learning, and dual contrastive learning for language-conditioned robotic manipulation. To fully leverage the 3D information from multi-view RGB-D inputs, we first convert them into point clouds (Figure 2 (a)) and generate five orthographic virtual views through re-rendering (Figure 2 (b)). These views are then processed by the visual encoder with patch-level self-attention (Figure 2 (c)). To dynamically resolve size ambiguities, we initialize size-aware query tokens conditioned on task instructions and fuse them with visual features. These queries iteratively refine geometric cues (e.g., scale, pose) through cross-attention mechanisms in the Size-Aware Dynamic Query Transformer (SADQ-Transformer) (Figure 2 (d)). Simultaneously, an Image-Text Contrastive Loss aligns language embeddings with size-aware visual features, ensuring linguistic-geometric consistency. The SADQ-Transformer aggregates multi-modal features–language tokens, refined query tokens q_s and visual tokens. This fused representation drives a policy network to predict actions. To further enforce geometic coherence in sequential tasks, a Current-Goal Contrastive Loss links the current state-instruction pair (s_t, l) to future subgoals g, mitigating size-related drift during multi-step executions.

These components work in synergy: the query tokens are responsible for aggregating potential scale-related features, while the contrastive loss provides the crucial supervision to ensure these features are correctly aligned with the linguistic size descriptors.

3.3 Visual and Language Encoder

We acquire orthographic virtual views following RVT[9] from 5 cubic viewpoints: the front, left, right, behind, and top. Each view image comprises RGB, depth, and (x, y, z) coordinate channels, where the coordinate channels establish cross-view spatial correspondences. The visual encoder comprises a patch embedding layer and a two-layer self-attention Transformer. We split the images into 20 ×20 patches and leverage a MLP layer to project the embeddings of patch tokens for self-attention. For the self-attention Transformer, each patch token only attends to other tokens within the same virtual view image, which aims to aggregate the information from the same view. The visual encoder is trained from scratch with normalized initialization. The language encoder tokenizes instructions using a frozen CLIP text encoder[28]. To adapt CLIP for size-sensitive language grounding, we introduce a trainable MLP aligning linguistic embeddings to visual scale features via the Image-Text Contrastive loss (Eq.3), which minimizes the distance between language embeddings and size-aware visual features.

3.4 Size-aware Dynamic Query Transformer

Unlike previous methods[9, 22, 31] that rely on static scene representations, we predefine a set of size-aware query tokens q with sematic prior, inspired by the query Transformer[19, 20]. These tokens dynamically adapt to linguistic instructions (e.g., "small cup" or "large item") by aggregating task-related geometric features from multi-view visual input, thereby resolving scale ambiguities. Specially, each query token interacts with visual features through cross-attention mechanisms, focusing on scale-sensitive regions. We set the number of learnable query tokens to five, ensuring comprehensive geometric coverage across all viewpoints.

The optimized queries q_s are then aligned with language embeddings via Image-Text Contrast ($Eq.$ 3), then concatenated visual tokens and language tokens into a unified representation space. This composite feature set undergoes four successive self-attention layers for multimodal feature fusion. During this hierarchical processing phase, q_s interact with both visual patches and linguistic elements, jointly encoding object affordances and spatial constraints in a unified representation space. The final multimodal queries $q_{s,l}$ drive the Current-Goal Contrast ($Eq.$4) learning, ensuring temporal consistency in multi-step tasks. By explicitly leveraging scale information, the model ensures robust action prediction across varying object sizes.

3.5 Size-aware Contrastive Imitation Learning

Existing imitation learning frameworks often fail to align language instructions with geometric-aware representations, especially when object size variations (e.g., "small" vs. "large") introduce ambiguity in task execution. To address this, we propose a size-aware contrastive imitation learning framework that integrates cross-modal alignment and temporal goal reasoning, inspired by priciples from contrastive representation learning[5, 20, 22, 37, 42].
Image-Text Contrast ($\mathbf{s} \leftrightarrow \mathbf{l}$). We formulate the cross-modal similarity metric between visual observation s_t and language instructions l as:

$$f_{\psi,\phi}(s_t,l) = \exp\left(\frac{\psi(s_t)^\top \phi(l)}{\tau}\right) \tag{2}$$

where ψ and ϕ are the state and language instruction representations, and τ a temperature hyperparameter. The bidirectional contrastive objective is defined as:

$$\begin{aligned}\mathcal{L}_{s\leftrightarrow l}(s_t,l) = \frac{1}{N}\sum_{i=1}^{N}\Bigg[&\log\frac{f_{\psi,\phi}(s_t^{(i)+},l^{(i)+})}{f_{\psi,\phi}(s_t^{(i)+},l^{(i)+})+\sum_{l^-\in\mathcal{L}^-} f_{\psi,\phi}(s_t^{(i)+},l^-)}\\ &+\log\frac{f_{\phi,\psi}(l^{(i)+},s_t^{(i)+})}{f_{\phi,\psi}(l^{(i)+},s_t^{(i)+})+\sum_{s_t^-\in\mathcal{S}^-} f_{\phi,\psi}(l^{(i)+},s_t^-)}\Bigg]\end{aligned} \tag{3}$$

Here, $\mathcal{S}^-$ and $\mathcal{L}^-$ respectively denote negative visual states and language instructions sampled from all other trajectories within the same mini-batch. The bidirectional contrastive loss maximizes mutual information between matched (s_t,l) pairs while repelling mismatched negatives (s_t^-,l^-).

Current-Goal Contrast $((\mathbf{s},\mathbf{l}) \leftrightarrow \mathbf{g})$**.** For temporal alignment, we sample future states as goals $g \sim \mathcal{D}$ and compute the joint contrastive loss:

$$\mathcal{L}_{(s,l)\leftrightarrow g}(s_t,l,g) = \frac{1}{N}\sum_{i=1}^{N}\Bigg[\log\frac{f_{\xi,\psi}((s_t^{(i)+},l^{(i)+}),g^{(i)+})}{f_{\xi,\psi}((s_t^{(i)+},l^{(i)+}),g^{(i)+})+\sum\limits_{g^-\in G^-} f_{\xi,\psi}((s_t^{(i)+},l^{(i)+}),g^-)} \\ +\log\frac{f_{\psi,\xi}(g^{(i)+},(s_t^{(i)+},l^{(i)+}))}{f_{\psi,\xi}(g^{(i)+},(s_t^{(i)+},l^{(i)+}))+\sum\limits_{(s_t^{(i)-},l^{(i)-})\in\Omega^-} f_{\psi,\xi}(g^{(i)+},(s_t^{(i)-},l^{(i)-}))}\Bigg] \tag{4}$$

where ξ denotes the cross-modal fusuion representations, G^- contains inter-task goal negatives, and Ω^- includes mismatched vision-language pairs. This contrastive mechanism strengthens the agent's capacity to associate current observations with task-relevant future states.

4 Experiments

4.1 Experiment Setup

Simulation. Our experiments are conducted in the popular RLBench simulated tasks, we utilize CoppeliSm to simulate tasks in RLBench including picking and lifting, picking cup, placing up cup etc. These tasks, all performed by a Franka Panda robot equipped with a parallel gripper, include several variations specified by the associated specification. Visual observations are required from four 128×128 resolution RGB-D cameras at the wrist, left shoulder, right shoulder, and front positions. During the training phase, we use the RLBench training dataset with 100 expert demonstrations for each task.

Baselines. We compare SACIL with the previous state of the arts including: (1) RVT[9], which renders five global orthographic images from input RGB-D images and uses a multi-view transformer for action prediction; and (2) PolerNet[3], which merges 3D point cloud from multi-view cameras and uses multimodel transformers to predict 7-DoF(degrees of freedom) actions. (3) HiveFormer[10], which tokenizes instruction, visual observations and previous actions and uses a multimodal transformer to process these tokens for action prediction. (4) Act3D[6], which computes a scene-level physical 3D feature cloud for 3D action prediction. We re-implement all baselines using public codebases and evaluate their performance on new five single-targe tasks and four sequential dual-target tasks. All baselines are trained on the same set of keyposes extracted from expert demonstrations.

Implementation Details. Following RVT[9], we train our model on cube-viewed re-rendered image from 3D point clouds. Similar to PerAct[31], we implement both translation and rotation data augmentations: 1) Point clouds are random translation within a cubic region of [-0.125, +0.125] along each axis; 2) Random rotations are applied around the vertical z-axis with angular displacements bounded at ±45°. For the training scheme, we train our model for 133K iterations with the batch size of 18. The LAMB optimizer is employed with an initial learning rate of 1.8×10^{-3}. We used cosine learning rate decay with warm-start for 2K steps. The training is conducted on $6 \times NVIDIA$ 4090 GPUs for around 20 hours. For evaluation metric, We use the task success rate to measure performance. For each episode,

Models	Avg. Success	Train Time	pick and lift	pick up cup	place cup	put item in drawer	take off pepper
HiveFormer[10]	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[6]	44.0	~12h	90.7	82.7	**18.7**	9.3	18.7
RVT[9]	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**

Table 1: Single-targe tasks performance evaluated on 25 episodes per task on RLBench, each task requires interacting with one object at a time. The input resolution is 128 ×128. In Train Time, the h represents hours. All other results are success rates, measured in percentage (%).

Models	Avg. Success	Train Time	place cups	push buttons	put items in drawer	take off peppers
HiveFormer[10]	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[6]	0.0	~11h	0.0	0.0	0.0	0.0
RVT[9]	40.7	~12h	6.7	**86.7**	61.3	8.0
SACIL	**45.7**	~20h	**13.3**	85.3	**77.3**	6.7

Table 2: sequential dual-target tasks evaluated on 25 episodes per task on RLBench, each task requires interacting with two objects at a time. We report success rate (%).

100 indicates complete success and 0 indicates failure, with no partial credits. We report the average success rates on 25 episodes for each task and the average success rates on all five single-target tasks and four sequential dual-target tasks separately.

4.2 Comparision with the State-of-the-art Methods

In this section, we compare our method with previous state-of-the-art methods on the RL-Bench tasksuite. The result of these baselines are re-implemented while retaining the previous hyperparameter settings. We evaluate all models three times on the same 25 episodes for each task and report mean results due to the randomness of the sampling-based motion planner. We show quantitative results for the five single-targe tasks in Table 1 and for the four sequential dual-target tasks in Table 2.

On five single-targe tasks, our method achives the best performence with an average success rate of 57.6%, which is the state-of-the-art, outperforming the previous arts by a sizable margin. In particular, our methods achieves a big leaps on long-horizen tasks like put item in draw, take off pepper.

On four sequential dual-target tasks, our method achives the best performence with an average success rate of 45.7%. In contrast, PolarNat and Act3D struggle to maintain consistent performance, failing to preserve geometric consistency across steps. This is likely due to their inability to effectively align language instructions with geometric features and their lack of a robust mechanism for maintaining temporal consistency in size-aware reasoning.

4.3 Ablation Study

Our method introduce an end-to-end contrastive Imitation Learning strategy to tackle the misalignment problem between geometric representations and linguistic features, and avoid path offset. The size-aware query tokens are introduced to aggregate geometric features effectively, which improves the performance of model by 2.0% as shown in Table 3.

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

Table 3: 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

Table 4: Comparison of different contrastive Imitation Learning methods on five single-targe tasks. We evaluated per task 25 episodes on RLBench.

To verify the effectiveness of size-aware contrastive imitation learning, we also conduct an ablation study in Table 4. We first implement a vanilla baseline without any contrastive imitation learning, while the inference process stays the same. By adding the image-text contrast learning, the performance improves by 3.8% compare with the baseline. This gain primarily stems from the alignment enforced between language embeddings and geometric features, which enhances the model's ability to resolve size ambiguity.

Further adding current-goal contrast, we observe that the average success rate increases by 4.5% than the only image-text contrast version. This gain is attributed to current-goal contrast's ability to maintaining size consistency across steps, particularly in long-horizon tasks. After combining image-text contrast and current-goal contrast in our framework, the performance increases from 49.3% to 57.6%, demonstrating the synergistic effect of our contrastive Imitation Learning strategy.

Generalizability of Dual Contrastive Learning. To validate the plug-and-play capability of our dual contrastive learning module, we integrate it into RVT-2[7], a state-of-the-art model leveraging 3D re-rendered images. We select RVT-2 due to its architectural compatibility with SACIL. As shown in Table 5, adding our dual contrastive losses (Image-Text and Current-Goal) improves RVT-2's average success rate by **2.2%** on single-target tasks and **0.3%** on sequential dual-target tasks. This indicates that our framework can enhance existing models reliant on re-rendered 3D representations without requiring architectural overhauls.

4.4 Case Study

We present two examples of generated action sequences in Figure 3, comparing results from RVT[9] and our proposed method. In the top example, the agent is instructed to "Put the large item in the bottom drawer." The RVT agent struggles to complete the task, as it imitates an expert demonstration involving a backward putting motion. In contrast, our method successfully opens the bottom drawer and places the large block inside. This success is pri-

Methods	Avg. Success Single-target tasks	Avg. Success Sequential dual-target tasks
RVT-2[7]	62.1	52.7
SACIL-RVT-2	64.3	53.0
improvement	+2.2	+0.3

Table 5: Multi-task performance of integrating dual contrastive learning into RVT-2.

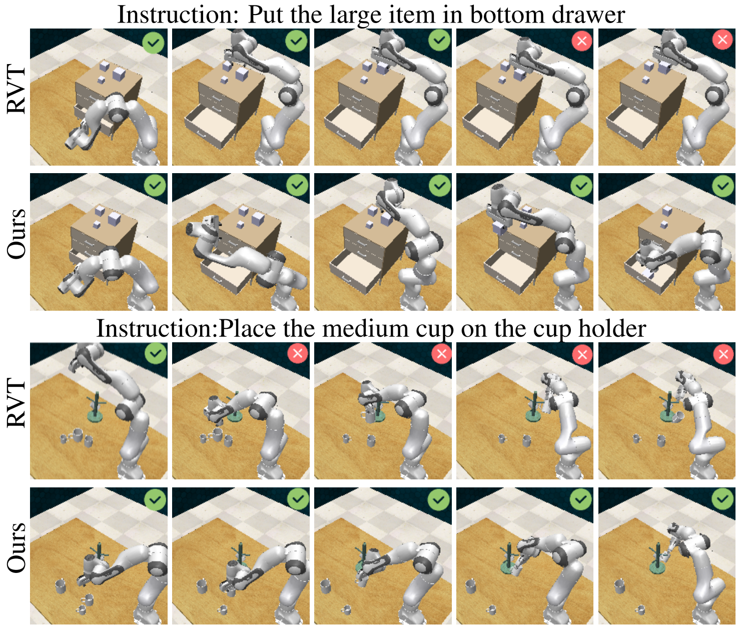

Figure 3: **Case Study.** Red markers indicate significant pose deviations from expert demonstrations, while green markers denote alignment. Our method consistently achieves the intended goal.

marily attributed to our Current-Goal Contrast module, which promotes trajectory coherence in long-horizon tasks.

In the bottom example, the instruction is "Place the medium cup on the cup holder." The RVT agent misinterprets the term "medium", instead manipulating the large cup and ultimately failing to place it correctly. In contrast, our method leverages size-aware query tokens to resolve ambiguities in object size and successfully completes the task. This demonstrates that our method not only understands semantic instructions but also effectively bridges the gap between language understanding and geometric reasoning.

5 Conclusion

In this work, we propose the Size-Aware Contrast Imitation Learning (SACIL), a framework for language-conditioned robotic manipulation that integrates Image-Text Contrast and Current-Goal Contrast to align linguistic instructions with geometrically consistent actions. By incorporating size-aware query tokens, SACIL resolves ambiguities in object scale perception, achieving state-of-the-art results on RLBench with 57.6% (single-target) and 45.7% (sequential dual-target) success rates.

While SACIL demonstrates robust performance in interpreting size-related instructions (e.g., “place the medium cup”) and ensuring trajectory coherence, limitations exist in generalizing to unseen scenarios. Future work will explore multi-modal integration and zero-shot generalization for real-world adaptability. This work advances the development of interpretable language-driven robots for complex human environments.

Acknowledgements

This work was supported in part by the National Key R&D Program of China under Grant 2019YFB1804003.

References

[1] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale.

[2] Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, Ryan Julian, et al. Do as i can, not as i say: Grounding language in robotic affordances. In *Conference on robot learning*, pages 287–318. PMLR, 2023.

[3] Shizhe Chen, Ricardo Garcia Pinel, Cordelia Schmid, and Ivan Laptev. Polarnet: 3d point clouds for language-guided robotic manipulation. In *7th Annual Conference on Robot Learning*.

[4] Benjamin Eysenbach, Ruslan Salakhutdinov, and Sergey Levine. C-learning: Learning to achieve goals via recursive classification. In *International Conference on Learning Representations*.

[5] Benjamin Eysenbach, Tianjun Zhang, Sergey Levine, and Russ R Salakhutdinov. Contrastive learning as goal-conditioned reinforcement learning. *Advances in Neural Information Processing Systems*, 35:35603–35620, 2022.

[6] Theophile Gervet, Zhou Xian, Nikolaos Gkanatsios, and Katerina Fragkiadaki. Act3d: 3d feature field transformers for multi-task robotic manipulation. In *Conference on Robot Learning*, pages 3949–3965. PMLR, 2023.

[7] Ankit Goyal, Valts Blukis, Jie Xu, Yijie Guo, Yu-Wei Chao, and Dieter Fox. Rvt-2: Learning precise manipulation from few demonstrations. In *RSS 2024 Workshop: Data Generation for Robotics*.

[8] Ankit Goyal, Arsalan Mousavian, Chris Paxton, Yu-Wei Chao, Brian Okorn, Jia Deng, and Dieter Fox. Ifor: Iterative flow minimization for robotic object rearrangement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14787–14797, 2022.

[9] Ankit Goyal, Jie Xu, Yijie Guo, Valts Blukis, Yu-Wei Chao, and Dieter Fox. Rvt: Robotic view transformer for 3d object manipulation. In *Conference on Robot Learning*, pages 694–710. PMLR, 2023.

[10] Pierre-Louis Guhur, Shizhe Chen, Ricardo Garcia Pinel, Makarand Tapaswi, Ivan Laptev, and Cordelia Schmid. Instruction-driven history-aware policies for robotic manipulations. In *Conference on Robot Learning*, pages 175–187. PMLR, 2023.

[11] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021.

[12] Stephen James, Zicong Ma, David Rovick Arrojo, and Andrew J Davison. Rlbench: The robot learning benchmark & learning environment. *IEEE Robotics and Automation Letters*, 5(2):3019–3026, 2020.

[13] Stephen James, Kentaro Wada, Tristan Laidlow, and Andrew J Davison. Coarse-to-fine q-attention: Efficient learning for visual robotic manipulation via discretisation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13739–13748, 2022.

[14] Dmitry Kalashnikov, Jacob Varley, Yevgen Chebotar, Benjamin Swanson, Rico Jonschkowski, Chelsea Finn, Sergey Levine, and Karol Hausman. Mt-opt: Continuous multi-task robotic reinforcement learning at scale. *arXiv e-prints*, pages arXiv–2104, 2021.

[15] Tsung-Wei Ke, Nikolaos Gkanatsios, and Katerina Fragkiadaki. 3d diffuser actor: Policy diffusion with 3d scene representations. In *8th Annual Conference on Robot Learning*.

[16] Ksenia Konyushkova, Konrad Zolna, Yusuf Aytar, Alexander Novikov, Scott Reed, Serkan Cabi, and Nando de Freitas. Semi-supervised reward learning for offline reinforcement learning. *arXiv e-prints*, pages arXiv–2012, 2020.

[17] Michael Laskin, Aravind Srinivas, and Pieter Abbeel. Curl: Contrastive unsupervised representations for reinforcement learning. In *International conference on machine learning*, pages 5639–5650. PMLR, 2020.

[18] Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabrael Levine, Michael Lingelbach, Jiankai Sun, et al. Behavior-1k: A benchmark for embodied ai with 1,000 everyday activities and realistic simulation. In *Conference on Robot Learning*, pages 80–93. PMLR, 2023.

[19] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.

[20] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[21] Corey Lynch and Pierre Sermanet. Language conditioned imitation learning over unstructured data.

[22] Teli Ma, Jiaming Zhou, Zifan Wang, Ronghe Qiu, and Junwei Liang. Contrastive imitation learning for language-guided multi-task robotic manipulation. In *8th Annual Conference on Robot Learning*.

[23] Teli Ma, Rong Li, and Junwei Liang. An examination of the compositionality of large generative vision-language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 692–705, 2024.

[24] Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics and Automation Letters*, 7(3):7327–7334, 2022.

[25] Ofir Nachum, Shixiang Gu, Honglak Lee, and Sergey Levine. Near-optimal representation learning for hierarchical reinforcement learning. In *International Conference on Learning Representations*.

[26] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[27] Shuang Qiu, Lingxiao Wang, Chenjia Bai, Zhuoran Yang, and Zhaoran Wang. Contrastive ucb: Provably efficient contrastive self-supervised learning in online reinforcement learning. In *International Conference on Machine Learning*, pages 18168–18210. PMLR, 2022.

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[29] Lin Shao, Toki Migimatsu, Qiang Zhang, Karen Yang, and Jeannette Bohg. Concept2robot: Learning manipulation concepts from instructions and human demonstrations. *The International Journal of Robotics Research*, 40(12-14):1419–1434, 2021.

[30] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Cliport: What and where pathways for robotic manipulation. In *Conference on robot learning*, pages 894–906. PMLR, 2022.

[31] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Perceiver-actor: A multi-task transformer for robotic manipulation. In *Conference on Robot Learning*, pages 785–799. PMLR, 2023.

[32] Simon Stepputtis, Joseph Campbell, Mariano Phielipp, Stefan Lee, Chitta Baral, and Heni Ben Amor. Language-conditioned imitation learning for robot manipulation tasks. *Advances in Neural Information Processing Systems*, 33:13139–13150, 2020.

[33] Adam Stooke, Kimin Lee, Pieter Abbeel, and Michael Laskin. Decoupling representation learning from reinforcement learning. In *International conference on machine learning*, pages 9870–9879. PMLR, 2021.

[34] Ahmed Touati and Yann Ollivier. Learning one representation to optimize all rewards. *Advances in Neural Information Processing Systems*, 34:13–23, 2021.

[35] Zhou Xian and Nikolaos Gkanatsios. Chaineddiffuser: Unifying trajectory diffusion and keypose prediction for robotic manipulation. In *Conference on Robot Learning/Proceedings of Machine Learning Research*. Proceedings of Machine Learning Research, 2023.

[36] Danfei Xu and Misha Denil. Positive-unlabeled reward learning. In *Conference on Robot Learning*, pages 205–219. PMLR, 2021.

[37] R Yang, Y Lu, W Li, H Sun, M Fang, Y Du, X Li, L Han, and C Zhang. Rethinking goal-conditioned supervised learning and its connection to offline rl. In *ICLR 2022-10th International Conference on Learning Representations*, 2022.

[38] Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on robot learning*, pages 1094–1100. PMLR, 2020.

[39] Yanjie Ze, Ge Yan, Yueh-Hua Wu, Annabella Macaluso, Yuying Ge, Jianglong Ye, Nicklas Hansen, Li Erran Li, and Xiaolong Wang. Gnfactor: Multi-task real robot learning with generalizable neural feature fields. In *Conference on Robot Learning*, pages 284–301. PMLR, 2023.

[40] Andy Zeng, Pete Florence, Jonathan Tompson, Stefan Welker, Jonathan Chien, Maria Attarian, Travis Armstrong, Ivan Krasin, Dan Duong, Vikas Sindhwani, et al. Transporter networks: Rearranging the visual world for robotic manipulation. In *Conference on Robot Learning*, pages 726–747. PMLR, 2021.

[41] Tong Zhang, Yingdong Hu, Hanchen Cui, Hang Zhao, and Yang Gao. A universal semantic-geometric representation for robotic manipulation. In *Conference on Robot Learning*, pages 3342–3363. PMLR, 2023.

[42] Chongyi Zheng, Benjamin Eysenbach, Homer Walke, Patrick Yin, Kuan Fang, Ruslan Salakhutdinov, and Sergey Levine. Stabilizing contrastive rl: Techniques for robotic goal reaching from offline data. In *NeurIPS 2023 Workshop on Goal-Conditioned Reinforcement Learning*.

[43] Kaizhi Zheng, Xiaotong Chen, Odest Chadwicke Jenkins, and Xin Wang. Vlmbench: A compositional benchmark for vision-and-language manipulation. *Advances in Neural Information Processing Systems*, 35:665–678, 2022.

[44] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR, 2023.