# Contrastive Representation Regularization for Vision-Language-Action Models


Taeyoung Kim[AB], Jimin Lee[A], Myungkyu Koo[AB], Dongyoung Kim[AB], Kyungmin Lee[A],
Changyeon Kim[A], Younggyo Seo[†C], Jinwoo Shin[†AB]

[A]KAIST, [B]RLWRLD, [C]UC Berkeley, [†]Equal Advising

KAIST AI Kim Jaechul Graduate School · RLWRLD · UC Berkeley

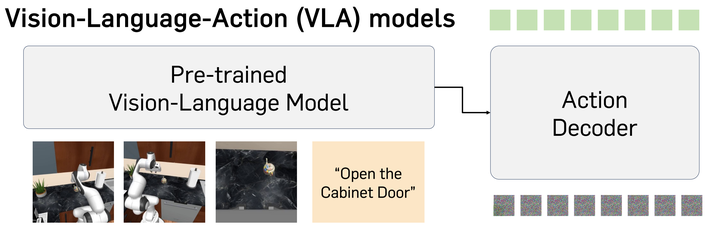


ICML International Conference On Machine Learning


Project Page · Taeyoung's Page

## Motivation

**Vision-Language-Action (VLA) models**

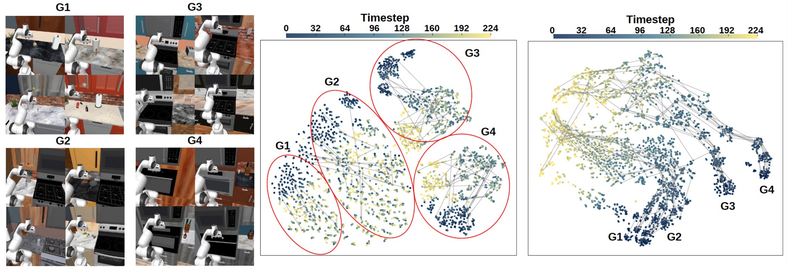



✓ VLMs provide strong conditionings for next action prediction.

✓ However, they are trained on vision-language data, without explicit grounding in robot-control signals.

## Observation & Challenges

✓ We visualize the VLM features of observation and instructions from identical task trajectories, across different environments




Identical task trajectories · VLM features · RS-CL aligned features

① **VLM features are organized by visual appearance**: Even when similar actions are required, features are far apart due to task-agnostic visual factors (e.g., scene layout, large surrounding objects).

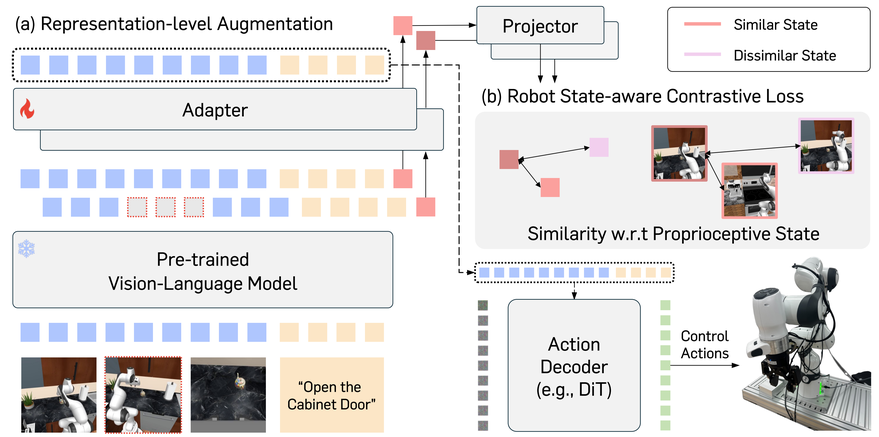
② **Goal:** Ground VLM features in the target robot's control signals, forming smooth representation trajectories aligned with temporal task progress.

✓ **Design requirements**

① Improve action prediction without harming generalization.

② Efficiently adapt to standard VLA training pipelines, without additional training stages or substantial compute overhead.

## Robot State-aware Contrastive Loss (RS-CL)




✓ TL;DR: RS-CL grounds VLM features by pulling together embeddings with similar proprioceptive states.

## Method

1. Incorporating continuous robot states to contrastive learning

- Use proprioceptive state distances as **soft contrastive supervision**.
- Samples with similar proprioceptive states get pulled harder
- Jointly optimize RS-CL with the original action prediction objective.

$$\mathcal{L}_{\text{RS-CL}}(\phi,\psi) = -\sum_{i,j=1}^{B} w_{ij}\,\log\frac{e^{\text{sim}(\mathbf{z}_i,\tilde{\mathbf{z}}_j)/\tau}}{\sum_{k=1}^{B} e^{\text{sim}(\mathbf{z}_i,\tilde{\mathbf{z}}_k)/\tau}},\quad w_{ij} = \frac{e^{-\|\mathbf{q}_i-\mathbf{q}_j\|_2/\beta}}{\sum_{k=1}^{B} e^{-\|\mathbf{q}_i-\mathbf{q}_k\|_2/\beta}},$$

$$\mathcal{L} = \mathcal{L}_{\text{FM}} + \lambda\,\mathcal{L}_{\text{RS-CL}} \qquad \mathbf{q}_i, \mathbf{q}_j : \text{Proprioceptive States}$$

2. Efficient contrastive pair generation

- **Summarization token:** Uses a learnable token to summarize long VLM embeddings into a compact representation.

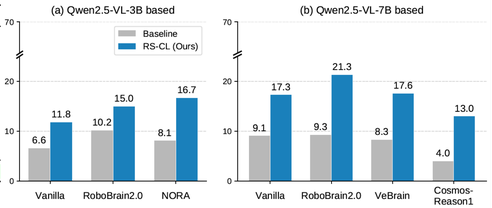
$$[\mathbf{h}, \mathbf{w}] = f_\phi\big(\text{VLM}(\mathbf{O}_t^V, \mathbf{c}) \oplus \mathbf{u}\big),$$

- **View cutoff:** Randomly masks the feature slice corresponding to one observation view.
- Generates augmented pairs at the representation level.

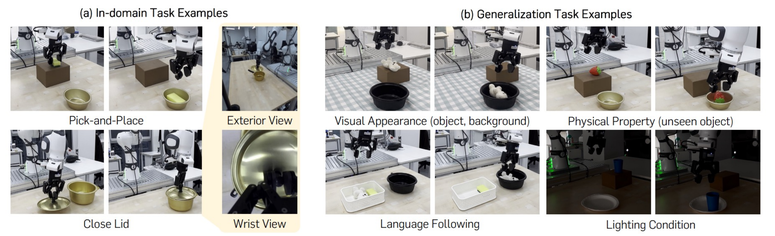
## Experiments

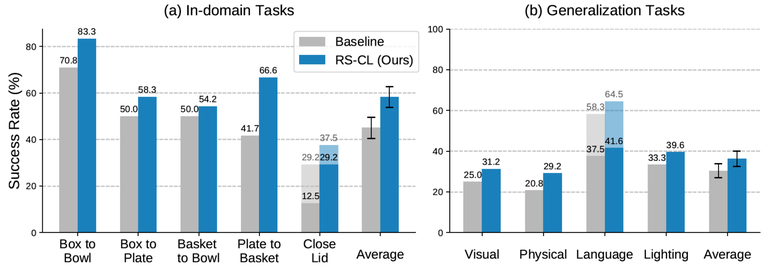
- **Simulation Benchmark Results**

| Method | Avg. |
|---|---|
| GR00T N1 (Bjorck et al., 2025) | 49.6 |
| GR00T N1 + DreamGen (Jang et al., 2025) | 57.6 |
| GR00T N1 + DUST (Won et al., 2025) | 58.5 |
| $\pi_0$ (Black et al., 2025b) | 62.5 |
| $\pi_0$-FAST (Pertsch et al., 2025) | 63.6 |
| GR00T-N1.5 (NVIDIA GEAR, 2025) | 65.7 |
| Video Policy (Liang et al., 2025) | 66.0 |
| FLARE (Zheng et al., 2025) | 66.4 |
| GR00T N1.5 + HAMLET (Koo et al., 2026) | 66.4 |
| **GR00T N1.5 + RS-CL (ours)** | **69.7** |

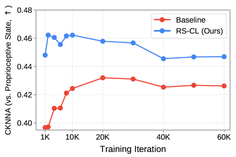



- **Real-world Results**




- **Analysis**




| Representation | Scene Acc.(%) | Task progress Acc.(%) |
|---|---|---|
| Pre-trained VLM embeddings | 99.6 | 1.2 |
| Embeddings trained with RS-CL | 93.6 | 22.9 |
| Ground-truth proprioceptive states | 53.1 | 25.6 |

RS-CL preserves semantics while improving control alignment

| Method | SR (%, ↑) | FLOPs ($\times 10^{12}$, ↓) |
|---|---|---|
| Baseline | 48.2 | 2.58 |
| Multi-view TCN | 50.0 | 7.53 |
| Single-view TCN | 50.3 | 7.53 |
| **RS-CL** | **53.0** | **2.91** |

| Soft-label target | Avg. |
|---|---|
| Baseline (*i.e.*, no regularization) | 65.7 |
| No soft label (*i.e.*, InfoNCE) | 67.3 |
| Next action sequence distance | 66.7 |
| Next single action distance | 66.8 |
| Current state distance | **69.7** |

| Augmentation method | Avg. |
|---|---|
| No augmentation | 65.3 |
| Token cutoff | 66.3 |
| Feature cutoff | 67.5 |
| Span cutoff | 67.3 |
| View cutoff | **69.7** |

| Method | Success rate |
|---|---|
| Baseline | 48.2 |
| **+ RS-CL (EEF pose)** | 53.0 |
| **+ RS-CL (Joint position)** | **53.4** |

| Method | Bimanual Panda + Hands | GR-1 Humanoid |
|---|---|---|
| Baseline | 52.7 | 68.9 |
| **+ RS-CL (Ours)** | **57.3** | **73.1** |

RS-CL works across diverse state configurations