Task Success Prediction for Open-Vocabulary Manipulation Based on Multi-Level Aligned Representations

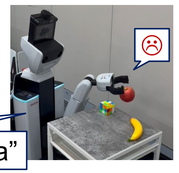
Miyu Goko, Motonari Kambara, Daichi Saito, Seitaro Otsuki, Komei Sugiura (Keio University)

TL;DR

Success prediction prevents subsequent task failure

Method outperforms GPT-4V

"move the apple near by the banana"



Task: Success Prediction for Open-Vocabulary Manipulation

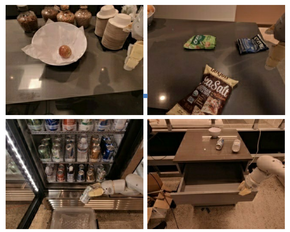
Input

Instruction: "Place a red can on the front right."

Output: "Success"

SP-RT-1 Dataset
- **13,915 samples**
- Based on RT-1 dataset [Brohan+, 22]

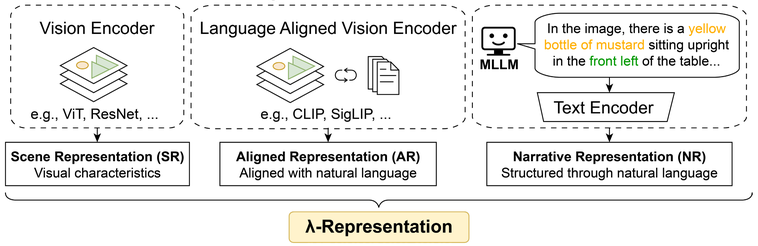
SP-HSR Dataset
- Collected in our lab
- For zero-shot evaluation

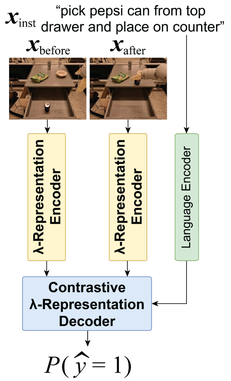


Method: Contrastive λ-Repformer

λ-Representation Encoder

Extracts multi-level aligned representation





Enhances expressiveness by using in parallel

Contrastive λ-Representation Decoder

Aligns difference between image representations with instruction sentence

Qualitative Results

SP-RT-1

"place water bottle upright"

→ ☺ Prediction: "Success"

"move orange near green can"

→ ☺ Prediction: "Failure"

SP-HSR

"place a purple cup on the front right"

→ ☺ Prediction: "Success"

"move the yellow French's bottle"

→ ☺ Prediction: "Failure"

Quantitative Results

Method	Accuracy [%]	
	SP-RT-1	SP-HSR
[LangRob@CoRL22, Xiao+]	71.59	-
GPT-4V (Zero-shot) [Achiam+]	63.90	59
GPT-4V (Few-shot) [Achiam+]	72.14	56
Gemini (Zero-shot) [GeminiTeam+]	67.28	53
Gemini (Few-shot) [GeminiTeam+]	68.44	53
Contrastive λ-Repformer	**80.80**	**60**

+8.66 (SP-RT-1: Gemini (Few-shot) → Contrastive λ-Repformer)

+1 (SP-HSR: GPT-4V (Few-shot) → Contrastive λ-Repformer)

Ablation Studies

- SR had the greatest impact within λ-Representation
- Cross-attention in the Contrastive λ-Representation Decoder captures the differences between images

SR	AR	NR	Acc. [%]
	✓	✓	73.72
✓		✓	79.94
✓	✓		79.70
✓	✓	✓	**80.80**

+7.08 (79.94 → 80.80)

Att. Mechanism	Acc. [%]
Self-Attention	78.88
Cross-Attention	**80.80**

+1.92 (78.88 → 80.80)

Application: Real-Time Video Classification Tasks

Success predictions between frame $t = 0$ and a subsequent frame