Semantic Temporal Abstraction via Vision-Language Model Guidance for Efficient Reinforcement Learning

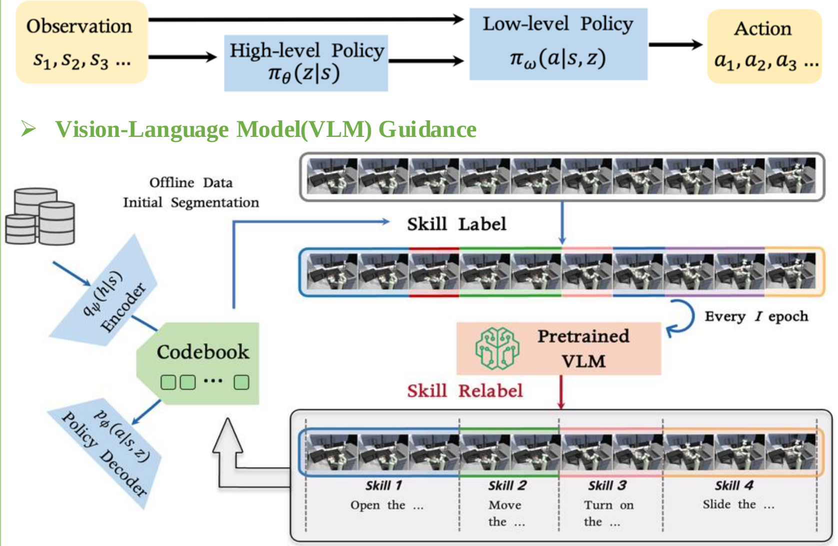
Tian-Shuo Liu[*,1,2], Xu-Hui Liu[*,1], Ruifeng Chen[*,1,2], Lixuan Jin[1], Pengyuan Wang[1,2], Zhilong Zhang[1,2], Yang Yu[1,2]

1. National Key Laboratory for Novel Software Technology, Nanjing University
2. Polixir Technologies

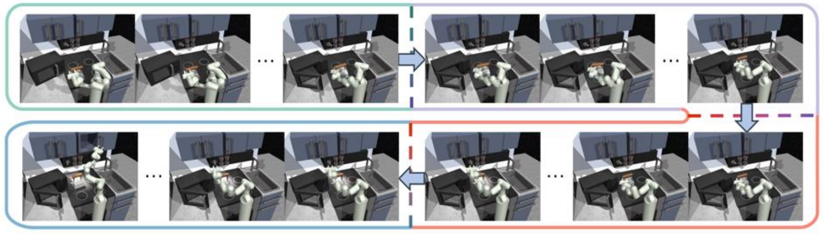
Overview

➢ **In long-horizon tasks, we extract temporally-extended skills for efficient RL.**

➢ **Vision-Language Model(VLM) Guidance**



➢ **Visualization of Segmentation**



Method

- Training the encoder and decoder following:

$$\mathcal{L}_{\psi,\phi} = \hat{\mathbb{E}}_{s,a\sim\mathcal{D}} \left[\|a-\hat{a}\|_2^2 + \beta\|h - \mathbf{sg}(z)\|_2^2 + \|z - \mathbf{sg}(h)\|_2^2 + \|z-e\|_2^2 + \gamma\|\Delta h\|_1 \right]$$

where $\mathbf{sg}(\cdot)$ represents the stop-gradient operation.

- High-level policy learning with Q-update

$$Q(s_t, z_t) \leftarrow r_{t:t+K-1} + \gamma^K Q(s_{t+K}, \underset{z\sim\pi_\theta(z|s_{t+K})}{\arg\max} Q(s_{t+K}, z))$$

- Low-level policy conditioned on primitive skill and observation

$$\min_\omega J(\omega;\mathcal{D}) = \hat{\mathbb{E}}_{\tau\sim\mathcal{D}} \left[-\sum_{t=0}^{H-1} \log \pi_\omega(a_t|s_t, z_t) \right]$$

Theoretical Results

For VanTA

Reduced from $|\mathcal{A}|$

$$v^* - v^{\pi_k} \le O\left(\sqrt{\frac{C \log \frac{|g(|\mathcal{S}|,\boxed{|\mathcal{Z}|})}{\delta}}{(1-\gamma^K)^4 n}} \right) \cdot V_{\max}$$
$$+ O(1) \cdot \sqrt{1 + \hat{Q}_{Var}(\beta)} \frac{H}{K} \sqrt{\frac{\sigma_K^2 \log(|\Pi_\alpha|\delta^{-1})}{n}}$$
$$+ O(R\log(n)) \cdot (1 + \hat{Q}_{Var}(\beta)) \frac{H}{K} \frac{\log(\boxed{|\Pi_\alpha|}\delta^{-1})}{n}$$
$$+ \Delta(\beta, \hat{Q}, \hat{V})$$

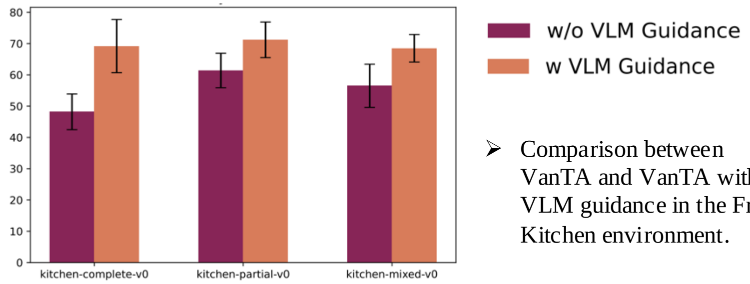
VanTA is more temporally correlated compared with non-VLM guidance method.

Performance

Performance on D4RL benchmark

Task Name	BC	CQL	IQL	RvS	GCSL	WGCSL	LDCQ	VanTA (Ours)
kitchen-complete-v0	65.0	43.8	62.5	50.2	58.6	57.7	52.8	**69.2±8.5**
kitchen-partial-v0	38.0	50.1	46.3	60.3	55.0	59.4	67.8	**71.2±5.7**
kitchen-mixed-v0	51.5	52.4	51.0	51.4	56.2	49.6	62.3	**68.5±4.4**
MiniGrid-DoorKey-6x6-complete-v0	**0.92**	0.67	0.91	0.89	0.82	0.85	0.86	**0.92±0.03**
MiniGrid-DoorKey-6x6-mixed-v0	0.70	0.61	0.72	0.67	0.59	0.79	0.72	**0.80±0.17**
MiniGrid-DoorKey-8x8-complete-v0	0.88	0.43	0.89	**0.94**	0.87	0.76	0.70	0.92±0.07
MiniGrid-DoorKey-8x8-mixed-v0	0.43	0.30	0.47	0.32	0.39	0.44	0.38	**0.51±0.18**
MiniGrid-KeyCorridorS3R3-complete-v0	0.47	0.62	0.72	0.64	0.55	0.57	0.69	**0.81±0.06**
MiniGrid-KeyCorridorS3R3-mixed-v0	0.09	0.23	0.51	0.42	0.21	0.23	0.38	**0.61±0.12**
MiniGrid-RedBlueDoors-6x6-complete-v0	0.80	0.51	**0.90**	0.78	0.73	0.61	0.64	0.85±0.11
MiniGrid-RedBlueDoors-6x6-mixed-v0	0.64	0.47	0.69	0.42	0.72	0.51	0.44	**0.73±0.20**
Crafter-partial	2.69	1.73	2.75	2.11	——	——	0.96	**5.46**

Ablation



➢ Comparison between VanTA and VanTA without VLM guidance in the Franka Kitchen environment.

Summary and Future Work

Summary:

➢ VanTA is a method for extracting discrete, task-relevant and semantic skills.

➢ The grounding of the codebook in semantic knowledge facilitates offline RL.

Future work:

➢ Our initialization technique has significantly reduced query complexity, which we aim to reduce further in future work.