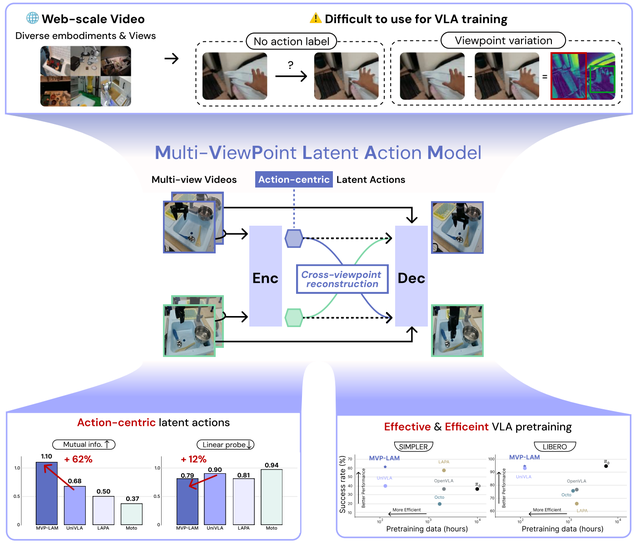
Website

ICML International Conference On Machine Learning

MVP-LAM: Learning Action-Centric Latent Action via Cross-Viewpoint Reconstruction

Jung Min Lee, Dohyeok Lee, Seokhun Ju, Taehyun Cho, Jinwoo Koo, Li Zhao, Sangwoo Hong, and Jungwoo Lee

Takeaway



Motivation

What is latent action?

Latent action = "Representation that encodes frame-to-frame changes"
- Learnable from *in-the-wild videos* without action-label
- *Pseudo-label* of action in VLA or World Model

$$L = \mathbb{E}_{(o,o')\sim D}[\|o' - D_\theta(o, z)\|] \text{ where } z = E_\theta(o, o')$$

Action-centric latent actions

Action-centric latent action="latent action *highly informative to action*"

Frame-to-frame changes = action-driven change + viewpoint-change.
Reducing the effect of viewpoints at the level of latent actions is crucial

$$I(Z_t, A_t) \geq H(Z_t) - I(Z_t; V_t, V_{t+1} | S_t, S_{t+1}) + C$$

Capacity — Leakage to encode viewpoint change

Method

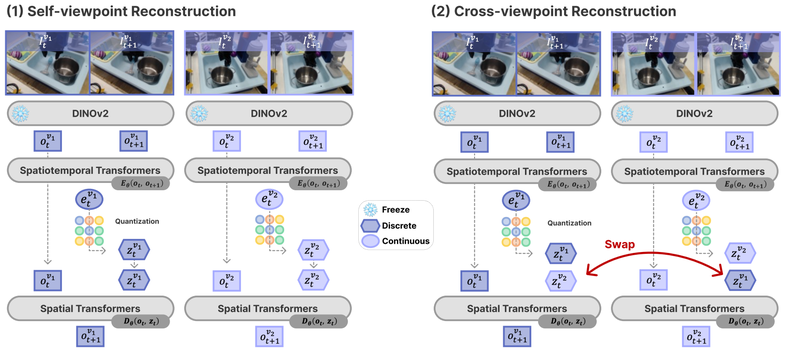
Training MVP-LAM

Cross-viewpoint reconstruction: *latent action in specific viewpoint* is used to *predict future frame in different viewpoint.*

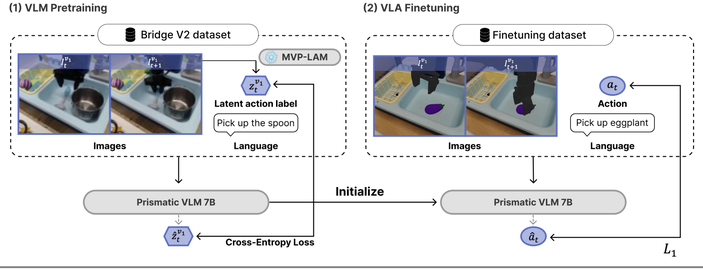
- Prevent latent action from encoding view-specific factor
  =useless to predict in different viewpoint
- Our future work shows that it is equivalent to *canonical correlation analysis (CCA)*

$$L_{\text{self}} = \frac{1}{2} \sum_{v \in \{v_1, v_2\}} \|o_{t+1}^v - D_\theta(o_t^v, z_t^v)\|^2 \quad L_{\text{cross}} = \frac{1}{2} \sum_{\substack{v,\tilde{v} \in \{v_1, v_2\} \\ v \neq \tilde{v}}} \|o_{t+1}^v - D_\theta(o_t^v, z_t^{\tilde{v}})\|^2$$

Architecture

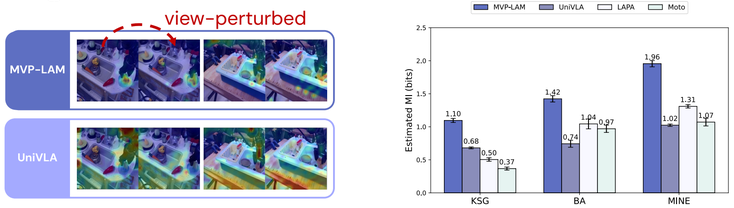


Training VLA with MVP-LAM



Results

Finding 1. MVP-LAM is action-centric

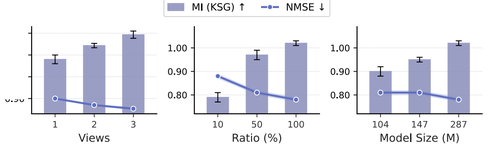


Finding 2. VLA+MVP-LAM outperforms prior VLAs

Method	SIMPLER					LIBERO				
	StackG2Y	Carrot2Plate	Spoon2Towel	Eggplant2Bask	Avg	Spatial	Object	Goal	Long	Avg
Octo-S	8.3	33.3	25.0	12.5	19.8	-	-	-	-	-
Octo-B	0.0	37.5	12.5	20.8	17.7	78.9	85.7	84.6	51.1	75.1
OpenVLA	41.6	50.0	37.5	16.7	36.4	84.7	88.4	79.2	53.7	76.5
π_0 *	37.5	33.3	29.2	45.8	36.5	**96.8**	**98.8**	**95.8**	85.2	**94.2**
LAPA	**54.2**	45.8	**70.8**	58.3	57.3	73.8	74.6	58.8	55.4	65.7
UniVLA	16.7	20.8	54.2	66.7	39.6	95.2	95.4	91.9	87.5	92.5
MVP-LAM	33.3	**66.7**	66.7	**75.0**	**60.4**	96.0	94.6	94.8	**90.8**	94.1

Wrist image + proprio. info + much larger pretraining dataset

Finding 3. MVP-LAM is scalable



Improved action-centricity
- Num. of viewpoints
- Larger dataset
- Larger model size

Finding 4. Multi-view is not enough