Joint-Aligned Latent Action: Towards Scalable VLA Pretraining in the Wild

Hao Luo[1,3], Ye Wang[2,3], Wanpeng Zhang[1,3], Haoqi Yuan[1,3], Yicheng Feng[1,3], Haiweng Xu[1,3], Sipeng Zheng[3], Zongqing Lu[† 1,3]

[1] School of Computer Science, Peking University [2] School of Information, Renmin University of China [3] BeingBeyond

CVPR JUNE 3-7, 2026 DENVER COLORADO

More results & videos are on the Blog page!

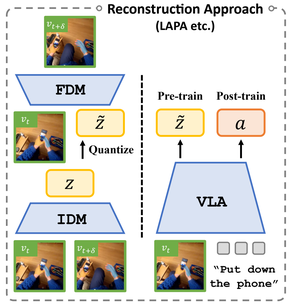
Background

VLA Pre-training with Hybrid Human Videos

- Human demonstrations offer an available substitute for scarce, costly, and embodiment-specific robot data.
- **Lab-collected** videos have accurate hand labels but limited diversity.
- **In-the-wild** videos are diverse but lack reliable action labels.

Reconstruction-based Latent Action Methods

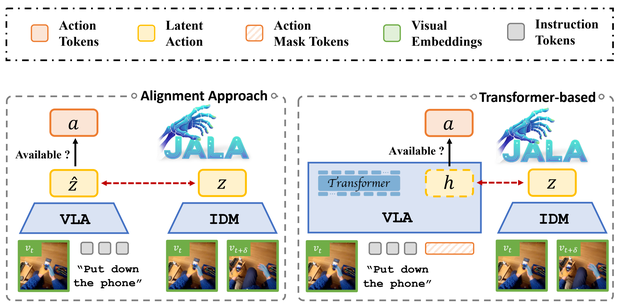
- Pixel reconstruction may overemphasize appearance, background, or camera artifacts rather than **action-relevant dynamics**.
- **Multi-stage modeling → Information mismatch/shift**

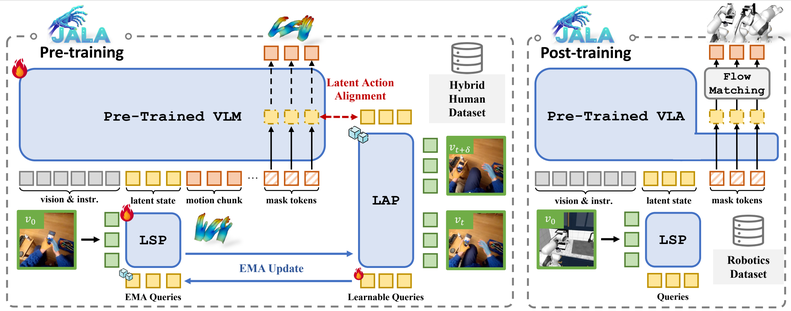


How to learn action-centric representations from hybrid human videos?

Method: Joint Alignment for Latent Action

Our shift: from reconstructing pixels to aligning action-relevant dynamics.





Jointly aligned action representation

- **Action-relevant, predictable** repr. ⟷ **Inverse Dynamics** emb.
- Implementation in VLA: **hidden states** of action prediction

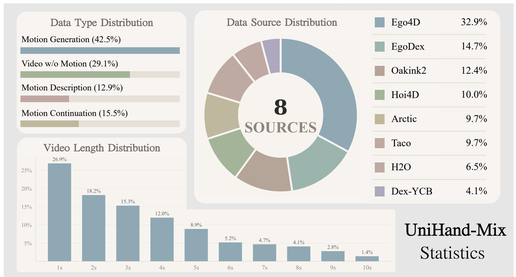
Pre-training on the hybrid dataset

- **Lab-collected**: anchors pretraining with accurate hand-motion supervision.
- **In-the-wild**: expands pretraining with diverse, natural manipulation dynamics.

Hybrid Human Video Dataset

A large-scale human manipulation corpus

- 7.5M+ instruction-video samples
- **2,000+ hours** of human manipulation videos



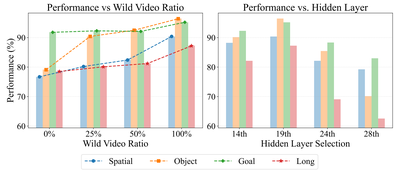
Experimental Results

- JALA improves generalization to in-the-wild human manipulation.
- Better simulation benchmark performance, especially under the few-shot setting.
- Improved dexterous robot manipulation, including long-horizon execution, contact-rich interaction, and fine-grained hand control.

LIBERO

Model	Spatial	Object	Goal	Long	Average
LAPA [19]	83.4	87.6	78.2	68.8	79.5
MolmoAct [63]	87.0	95.4	87.6	77.2	86.6
π0-FAST [32]	96.4	96.8	88.6	60.2	85.5
GR00T N1.5 [3]	94.4	97.6	93.0	90.6	93.9
π0 [2]	**96.8**	**98.8**	95.8	85.2	94.2
UniVLA [39]	95.4	**98.8**	93.6	94.0	95.5
Being-H0 [10]	92.6	96.8	94.0	77.4	90.2
JALA-act	93.4	97.8	94.2	91.8	94.3
JALA w/o dec.	64.6	58.4	61.2	42.2	56.6
LAPA†	87.4	91.2	90.0	65.4	83.5
JALA*	95.2	96.4	97.2	94.0	95.7
JALA-vjepa	95.4	98.0	**98.0**	94.8	96.6
JALA-dino	96.0	98.2	97.4	**96.0**	**96.9**

LIBERO Ablation



Real-World Tasks

Hand motion Prediction

Model	MPJPE ↓		PA-MPJPE ↓		MWTE ↓		MDE ↓	
	Lab	Wild	Lab	Wild	Lab	Wild	Lab	Wild
# Next Token Prediction								
Being-H0	7.61	16.91	1.34	3.81	6.03	14.54	7.16	18.33
Being-H0+dino	7.54	15.14	0.90	2.78	5.85	13.65	6.95	17.17
# Masked Chunk Prediction								
JALA w/o align	7.72	15.73	**0.89**	2.34	6.18	14.02	8.09	16.23
JALA w/o latent	8.26	20.34	1.83	3.94	8.02	17.13	10.17	26.74
JALA-dino	7.16	**11.02**	0.91	**1.12**	**5.77**	**9.79**	7.24	**11.04**
JALA-vjepa	**7.05**	11.54	0.94	1.32	5.85	10.04	**6.73**	11.87

50-shot Performance

Model	RoboCasa Syn.	RoboCasa Human	GR1 Tabletop
GR00T N1.5	20.83	35.17	20.41
LAPA	16.25	22.42	11.42
Being-H0	23.83	31.33	12.91
JALA-act	24.92	32.42	20.25
JALA w/o dec.	14.25	19.33	9.25
LAPA†	20.25	27.33	13.50
JALA*	25.33	33.83	24.50
JALA	**27.58**	**35.42**	**26.33**

Averaged Task Completion Ratio

Model	Put-Three-Obj		Wipe-Board		Water-Plant
	Seen	Unseen	Seen	Unseen	Seen
Being-H0	38.0	16.0	40.0	33.3	36.7
GR00T N1.5	48.0	28.0	56.7	43.3	53.3
JALA w/o align	40.0	32.0	53.3	43.3	56.7
JALA-vjepa	38.0	34.0	66.7	60.0	66.7
JALA-dino	**60.0**	**58.0**	**83.3**	**80.0**	**73.3**