DILL

Disentangling Spurious Correlations in Vision-Language-Action Models via Predicting Domain-Invariant Latent Lookahead

Junghyun Kim[1,2]* Ngseo Kim[2]* ChungWoo Lee[2] Seoyeon Lee[2] Woo-Jeong Baek[5] Adam Zhou[1] Chip Huyen[1] Jun-Ki Lee[2†] Gi-Cheon Kang[3†] Byoung-Tak Zhang[2,4†]

[1] OpenMind [2] Seoul National University [3] Ajou University [4] Tommoro Robotics [5] Hyundai Motor Company

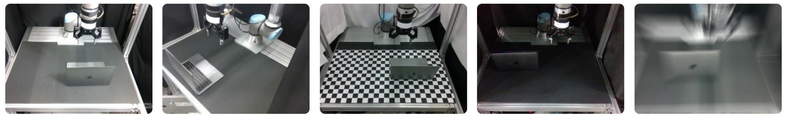
Project page

dill-vla.github.io

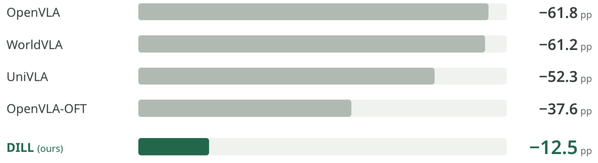
Learn to predict task-relevant futures and rely less on visual shortcuts.

01 Visual Shifts Break VLAs

Same task. Evaluate on different visual conditions.



Success-rate drop on visual shifts (LIBERO+ results)



02 Why VLAs Fail: Shortcut Learning

VLAs learn to choose actions based on incidental visual cues instead of task-relevant information.



π_0: 60% shortcuts in the LIBERO task–view diagnostic.

03 The Key Idea: How can a VLA Learn what matters?

Look to the future.

Future observations reveal how objects move and how a task progresses—rich signals for learning task-relevant features.

But not every future feature matters.

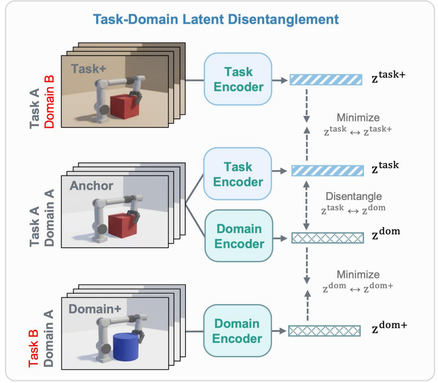
Raw future images and entangled visual latents also preserve the camera view, background, and other incidental details.

> The question is not simply whether a VLA should predict the future, but **what representation of the future it should predict.**

04 DILL: Domain-Invariant Latent Lookahead

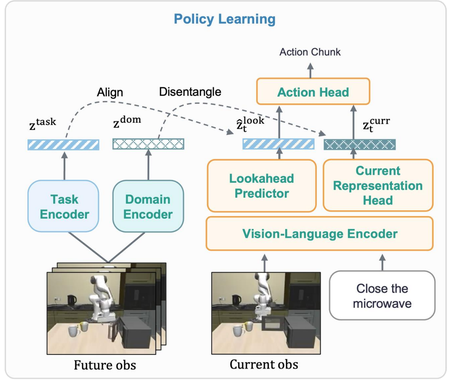
1 Learn to disentangle task from domain

Pretrain a Task-Domain Encoder to separate task-relevant features from incidental domain cues



2 Shape the VLA’s representation

During policy training, predict future task latents and disentangle from domain latents



05 Fewer Shortcuts

When the task–view correlation is reversed:

73% → 5%

Base VLA **Base VLA + DILL**

LIBERO task–view diagnostic

06 Higher Success under Visual Perturbations

Libero+ Evaluation:

Model	Original	Visual perturbations				Average
		Camera	Light	BG	Noise	
OpenVLA [3]	76.5	0.8 ↓75.7	8.1 ↓68.4	34.8 ↓41.7	15.2 ↓61.3	14.7 ↓61.8
WorldVLA [36]	79.1	0.1 ↓79.0	43.7 ↓35.4	17.1 ↓62.0	10.9 ↓68.2	18.0 ↓61.2
UniVLA [37]	95.5	1.8 ↓93.7	69.0 ↓26.5	81.0 ↓14.5	21.2 ↓74.3	43.3 ↓52.3
NORA [38]	87.9	2.2 ↓85.7	45.7 ↓42.2	58.6 ↓29.3	12.8 ↓75.1	29.8 ↓58.1
OpenVLA-OFT [39]	95.3	10.4 ↓84.9	76.8 ↓18.5	93.6 ↓1.7	49.9 ↓45.4	57.7 ↓37.6
Base VLA + SA	82.0	39.3 ↓42.7	56.8 ↓25.2	50.5 ↓31.5	6.3 ↓75.7	38.2 ↓43.8
DILL (Ours)	81.6	68.4 ↓13.2	69.4 ↓12.2	70.0 ↓11.6	68.7 ↓12.9	69.1 ↓12.5

Real-World Evaluation:

Success (%)	Original	Visual Perturbations
Base VLA	84.1	56.2
Base VLA + DILL	**88.9**	**82.9**