LAPA: Latent Action Pretraining from Videos

Best Paper Award (out of 75 accepted papers) 🏅

3rd Workshop on Language and Robot Learning | CoRL 2024 ⚓

Project: latentactionpretraining.github.io

[1]KAIST [2]University of Washington

[3]Microsoft Research [4]NVIDIA [5]Allen Institute for AI

Jishnu P

Reading Group | IRVL

11/15/24

Apart from the CVPR authors, Did anyone get a chance to go over the paper?

29 Pages, arxiv paper

Included main contents here

For experiments results, please refer to the paper

I will try to answer to the best of what I understood😊

Idea: Learn Actions from Videos

HEAD

Large-Scale Robot Datasets

- Expensive to collect ✗
- Requires robot hardware ✗
- Contains robot actions ✓

Idea: Learn Actions from Videos

HEAD

TAIL

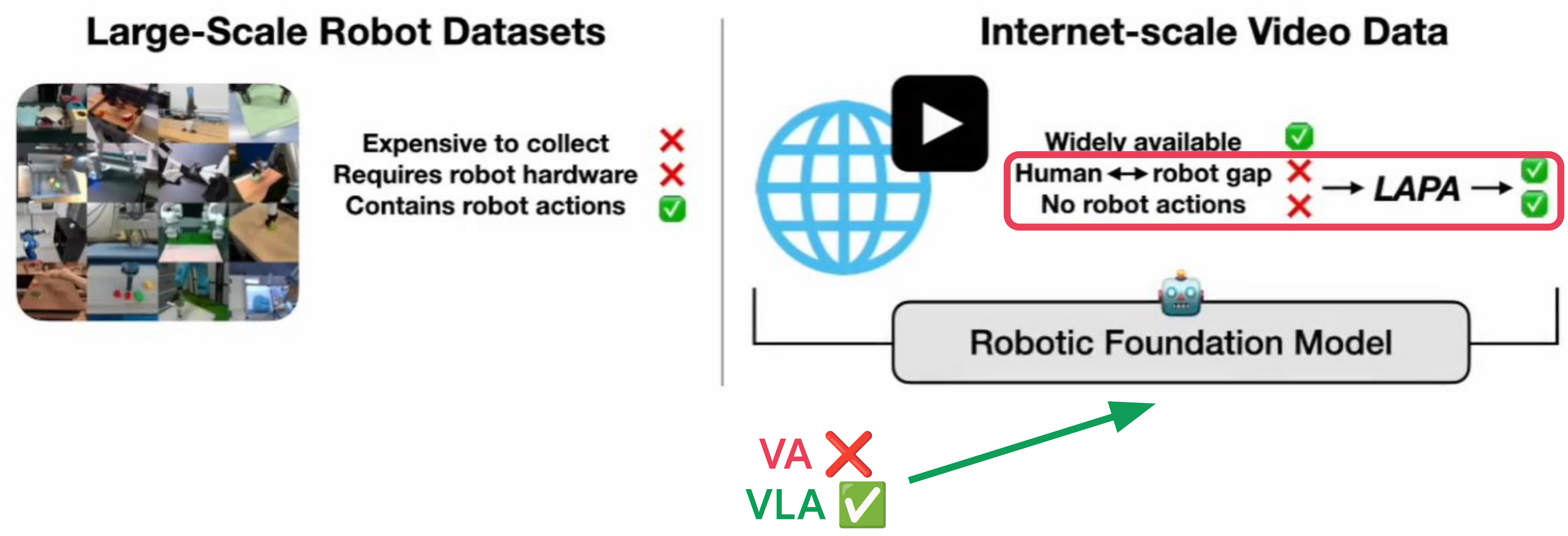
Large-Scale Robot Datasets

- Expensive to collect ❌
- Requires robot hardware ❌
- Contains robot actions ✅

Internet-scale Video Data

- Widely available ✅
- Human ↔ robot gap ❌
- No robot actions ❌

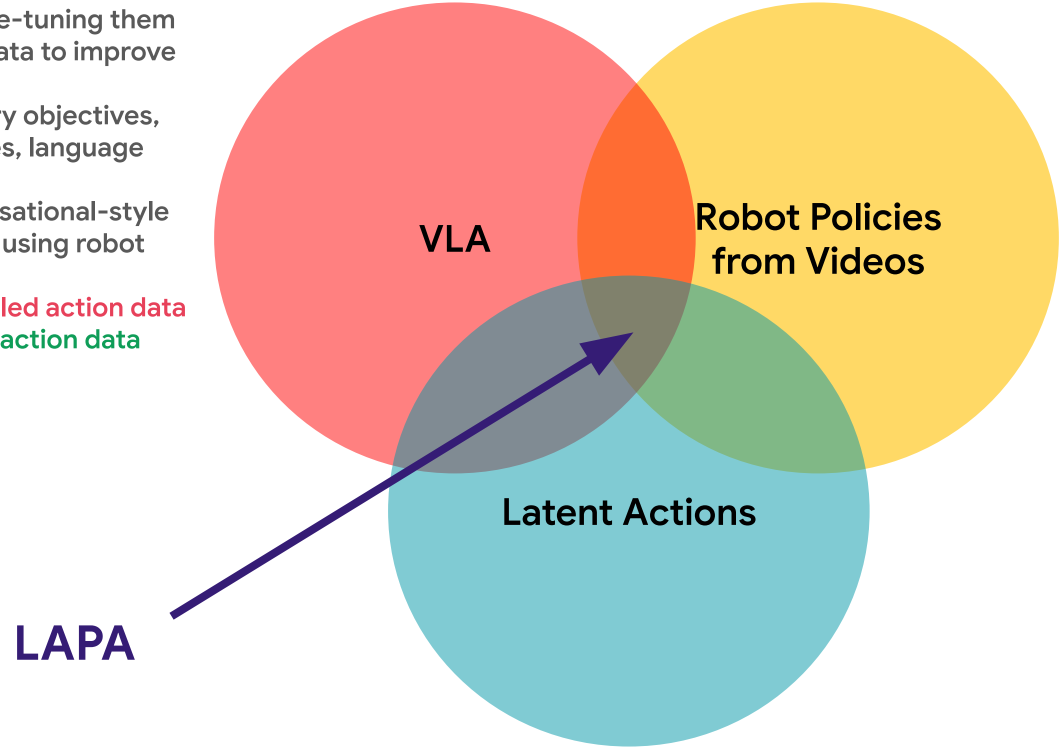
Idea: Learn Actions from Videos



Problem Formulation. Build a generalist robotic foundation model from human motion videos without action labels.

Related Work

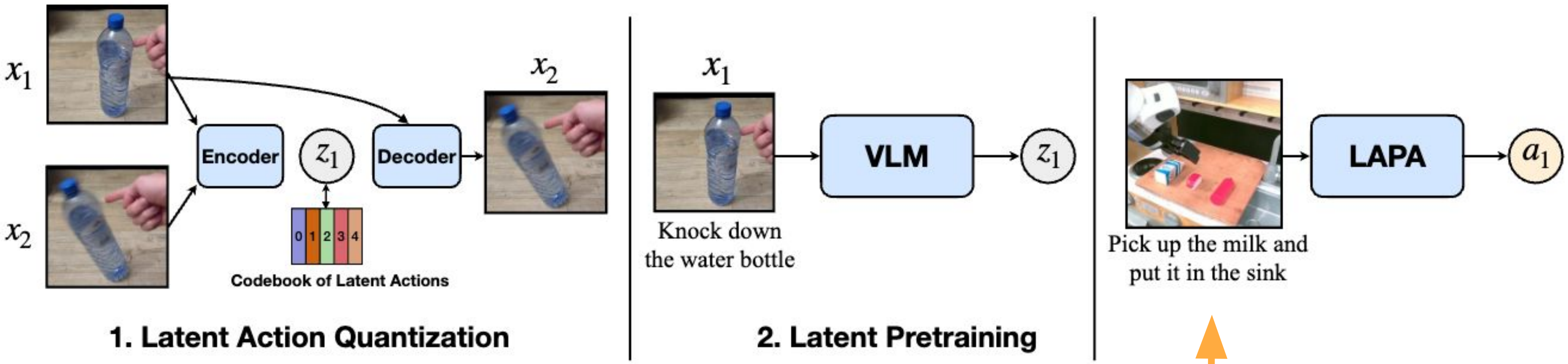
- Extend VLMs by fine-tuning them on robotic action data to improve physical grounding
- Incorporate auxiliary objectives, such as visual traces, language reasoning paths
- Construct a conversational-style instruction dataset using robot trajectory
- Heavily rely on labeled action data
- LAPA doesn't need action data



- Most raw videos do not contain any action labels
- Learn useful visual priors
- Learn robot manipulation policies by retargeting human motions to robot motions. These works rely on off-the-shelf models such as hand pose estimators or motion capture systems to retarget the human motions directly to robot motions.
- These works either learn only task-specific policies or require large in-domain perfectly aligned human-robot data
- Whereas LAPA allows learning the mapping directly from perception to control during pretraining.

Unlike other works that leverage latent actions by converting ground-truth actions into latent to capture better multimodality and task semantics, LAPA derives latent actions directly from observations, not ground-truth actions.

Overview

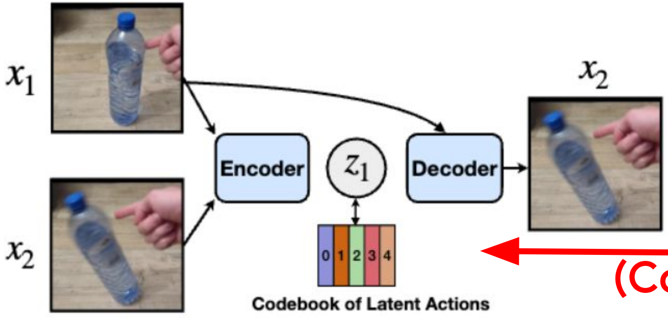


Latent Action Pretraining → **Action Finetuning**

2 Pretraining Stages

Fine-tuning to map the latent actions to real robot actions

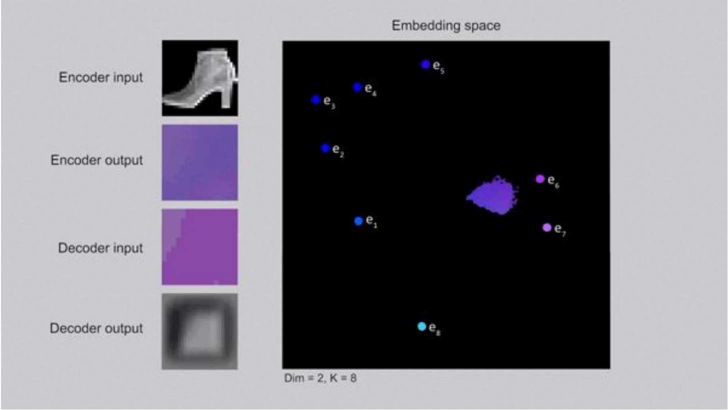
1. Latent Action Quantization



1. Latent Action Quantization

VQ
(Cobebook)

- Use a **VQ**-VAE based objective to **capture the discretized latent delta information between consecutive frames** in a video



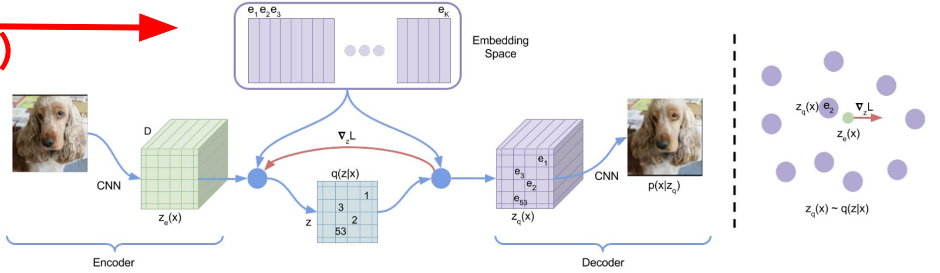


Figure 1: Left: A figure describing the VQ-VAE. Right: Visualisation of the embedding space. The output of the encoder $z(x)$ is mapped to the nearest point e_2. The gradient $\nabla_z L$ (in red) will push the encoder to change its output, which could alter the configuration in the next forward pass.

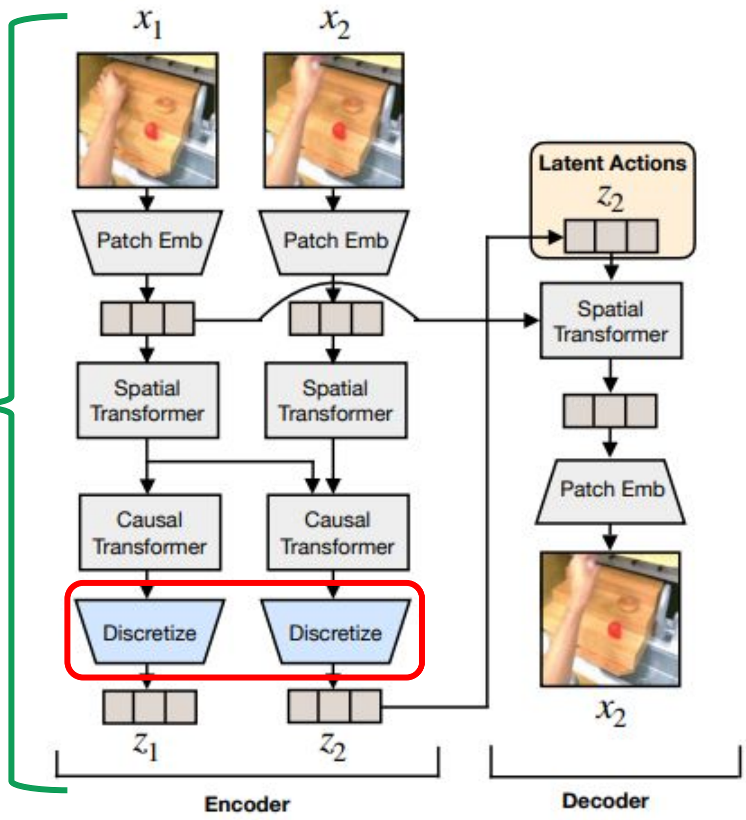
https://www.youtube.com/watch?v=_GD18kRQk0A

For in-depth details: Oord et al. Neural Discrete Representation Learning, NeurIPS 2017

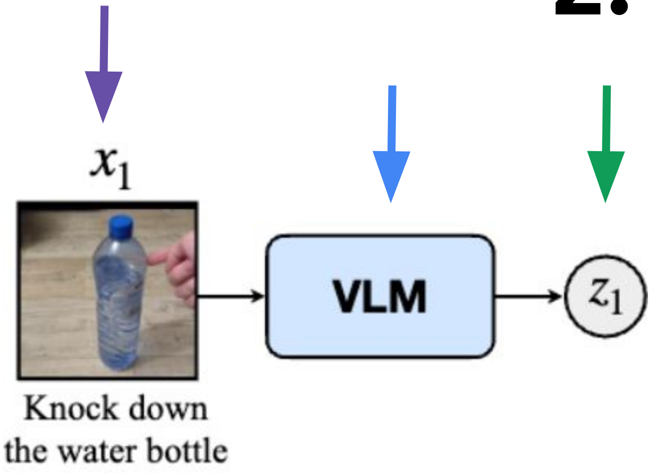
Aim: Learn to tokenize atomic actions without requiring predefined action priors (e.g., end-effector positions, joint positions)

1. Latent Action Quantization (Model)

C-Vi-ViT
Aka
'Phenaki'



2. Latent Pretraining

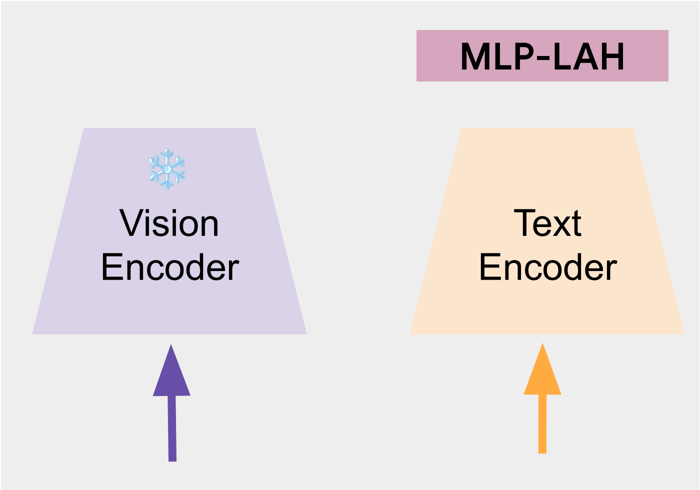


2. Latent Pretraining

- Perform **behavior cloning**
  - by **pretraining a Vision-Language Model**
  - to **predict latent actions derived from the first stage**. GT: ($z_t = f(x_t, x_{t+1})$)
  - based on **video observations** and **task descriptions**

z_t_hat

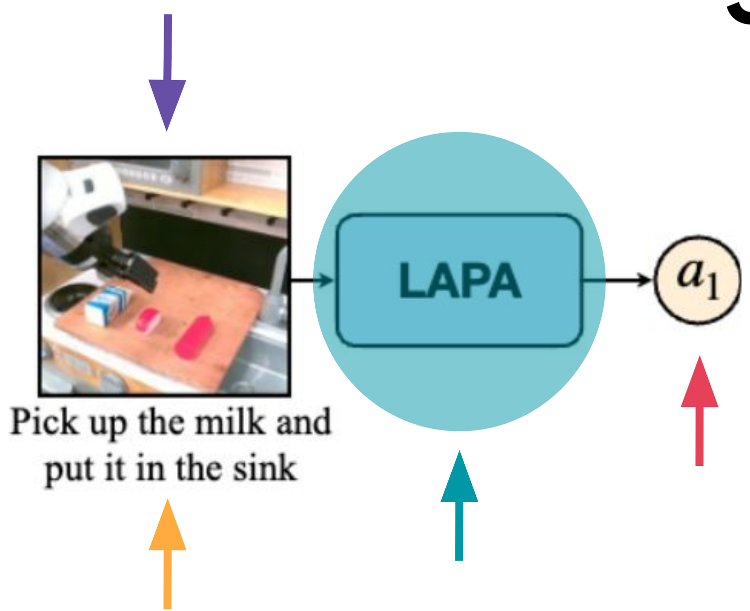
Goal: min || z_t_hat - z_t $||_2$



Instead of using the existing language model head of the VLM, attach a separate latent action head (MLP-LAH) of vocab size |C|.

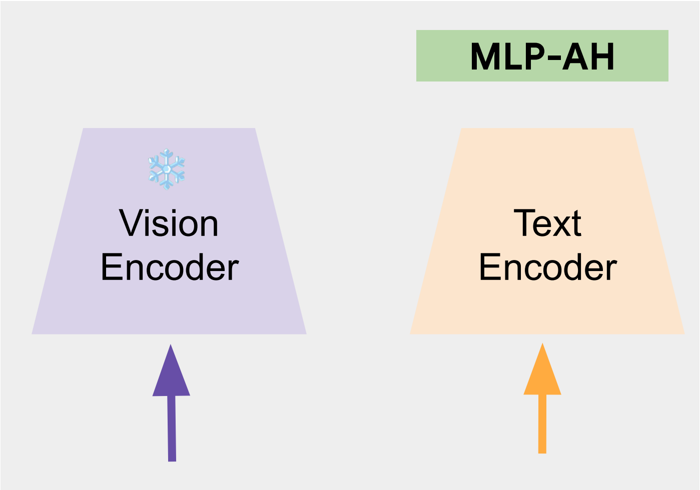
VLM: **7B** Large World Model (LWM-Chat-1M)
https://largeworldmodel.github.io
Applied mechanism is given

3. Finetuning



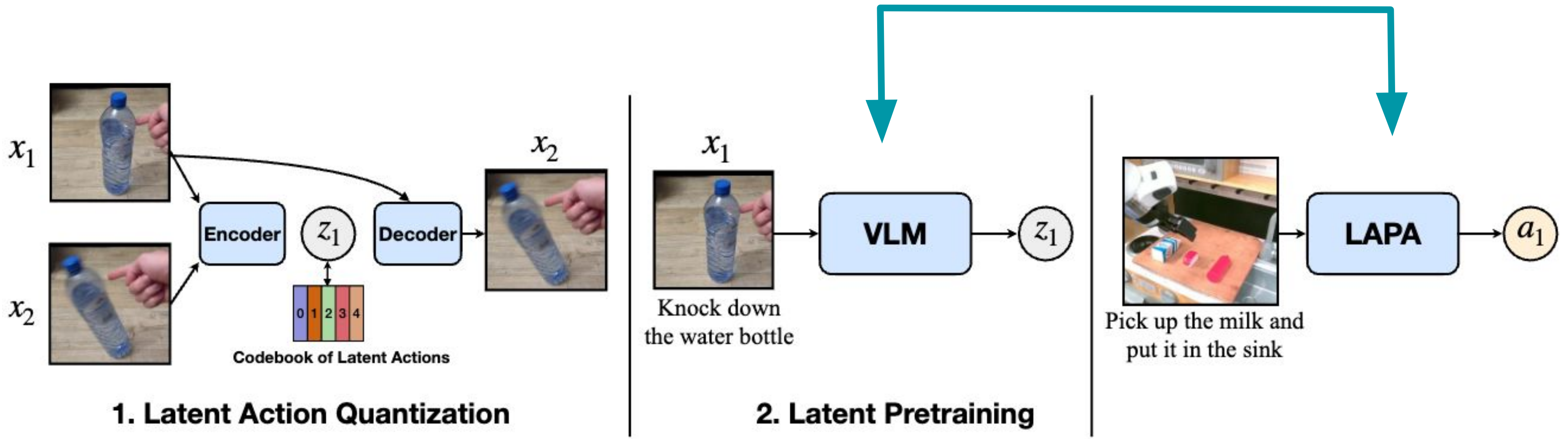
- VLAs pretrained to predict latent actions are not directly executable on real-world robots since latent actions are not actual delta end-effector actions or joint actions.
- To map latent actions to actual robot actions, LAPA is finetuned LAPA on a small set of labeled trajectories that contain ground truth actions (delta end-effector)
  - Fine-tune the model
    - on a **small-scale** robot manipulation dataset with **robot actions**
    - to learn the mapping from the **latent actions** to **robot action**

From authors, "We broadly refer to models having gone through latent pretraining as **LAPA**".



Discard the latent action head (a single MLP layer) and replace it with a new action head (MLP-AH) to generate ground truth actions

LAPA

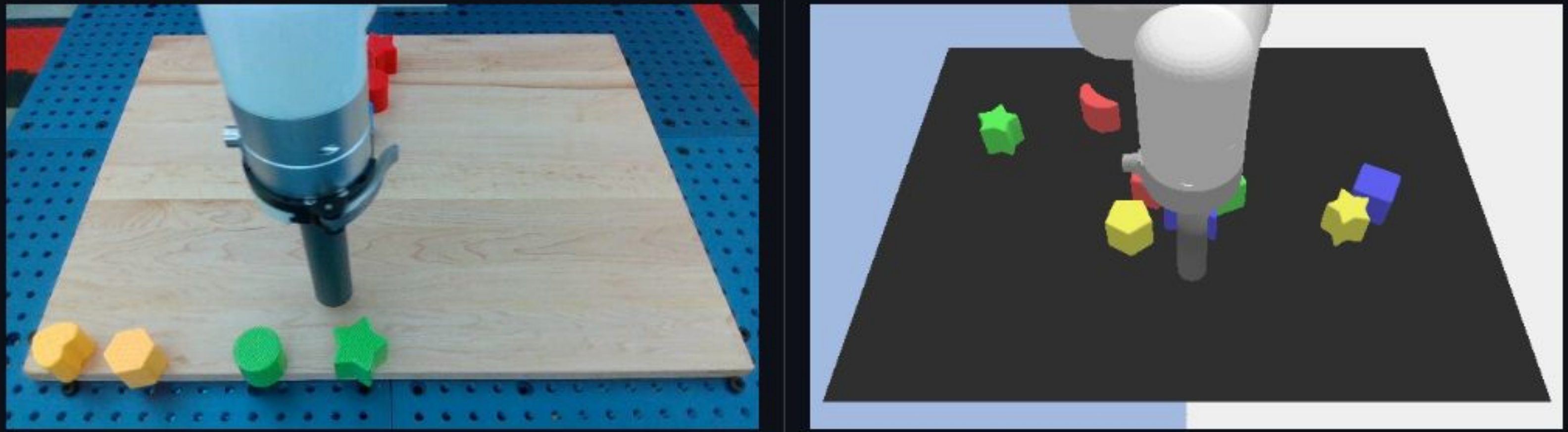


Latent Action Pretraining → **Action Finetuning**

Experiments: Datasets

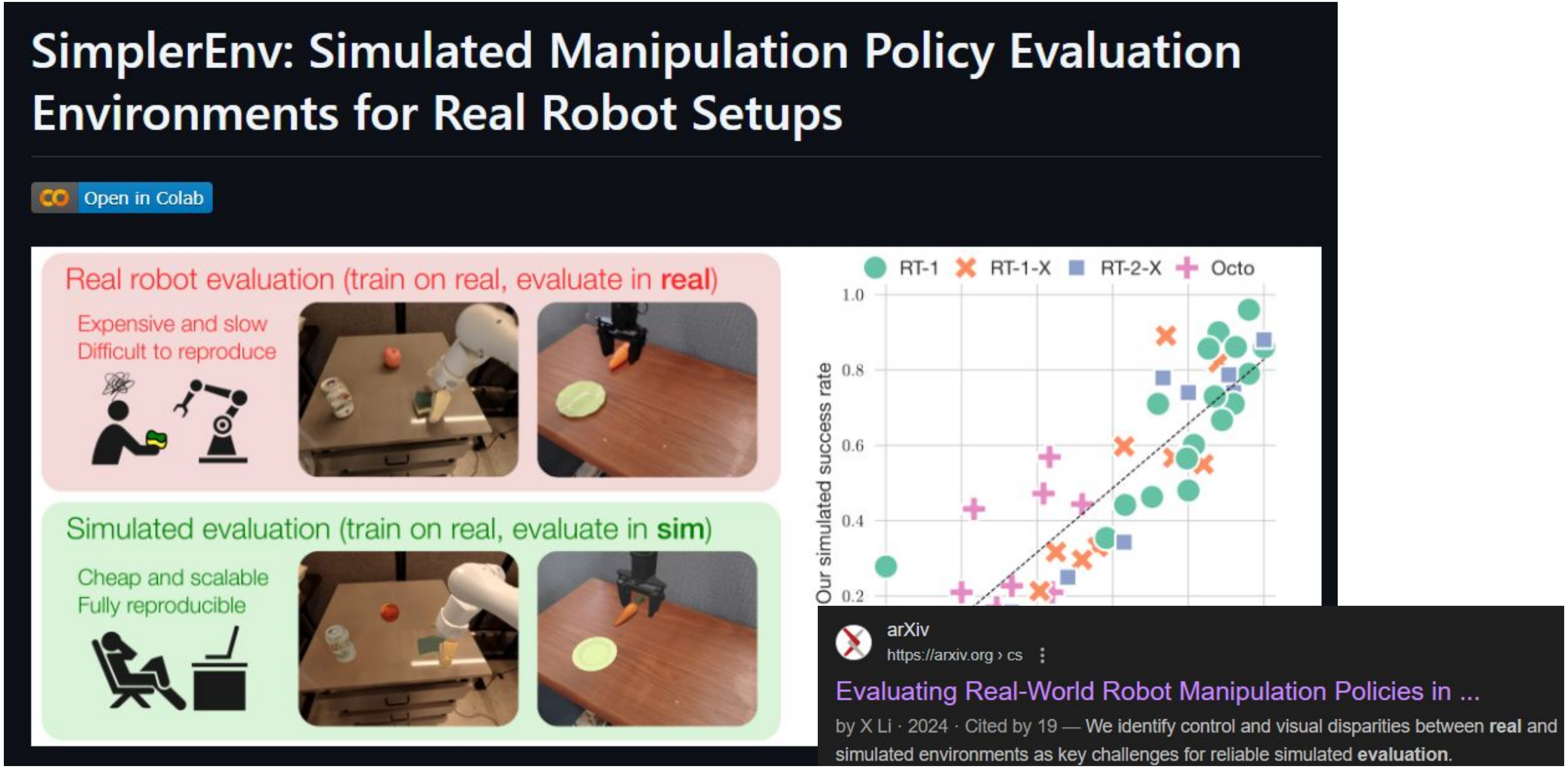
Language Table

Language-Table is a suite of human-collected datasets and a multi-task continuous control benchmark for open vocabulary visuolinguomotor learning.



https://interactive-language.github.io

Experiments: Datasets



https://simpler-env.github.io

Released: 05/24
Citations: 19

Experiments: Datasets

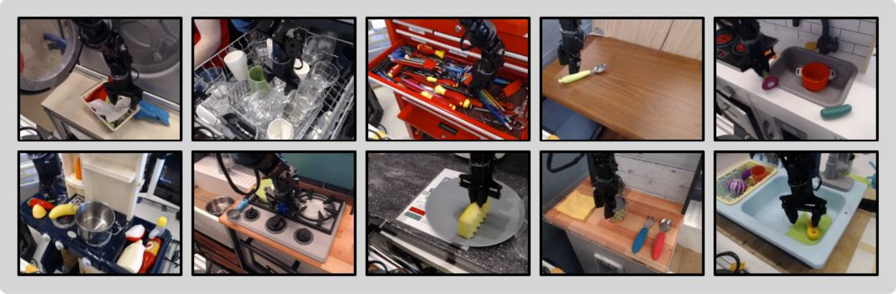
Dataset Composition

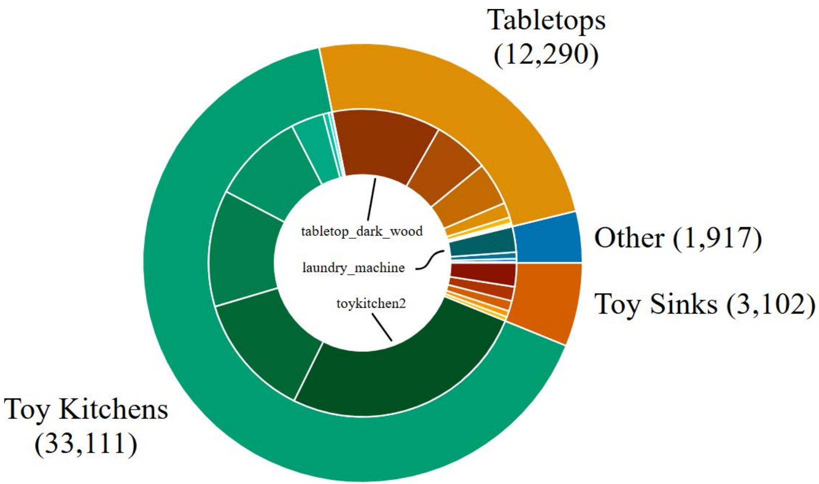
To support broad generalization, we collected data for a wide range of tasks in many environments with variation in objects, camera pose, and workspace positioning. Each trajectory is labeled with a natural langauge instruction corresponding to the task the robot is performing.

- 60,096 trajectories
  - 50,365 teleoperated demonstrations
  - 9,731 rollouts from a scripted pick-and-place policy
- 24 environments
- 13 skills

Environments

The 24 environments in BridgeData V2 are grouped into 4 categories. The majority of the data comes from 7 distinct toy kitchens, which include some combination of sinks, stoves, and microwaves. The remaining environments come from diverse sources, including various tabletops, standalone toy sinks, a toy laundry machine, and more.





https://rail-berkeley.github.io/bridgedata

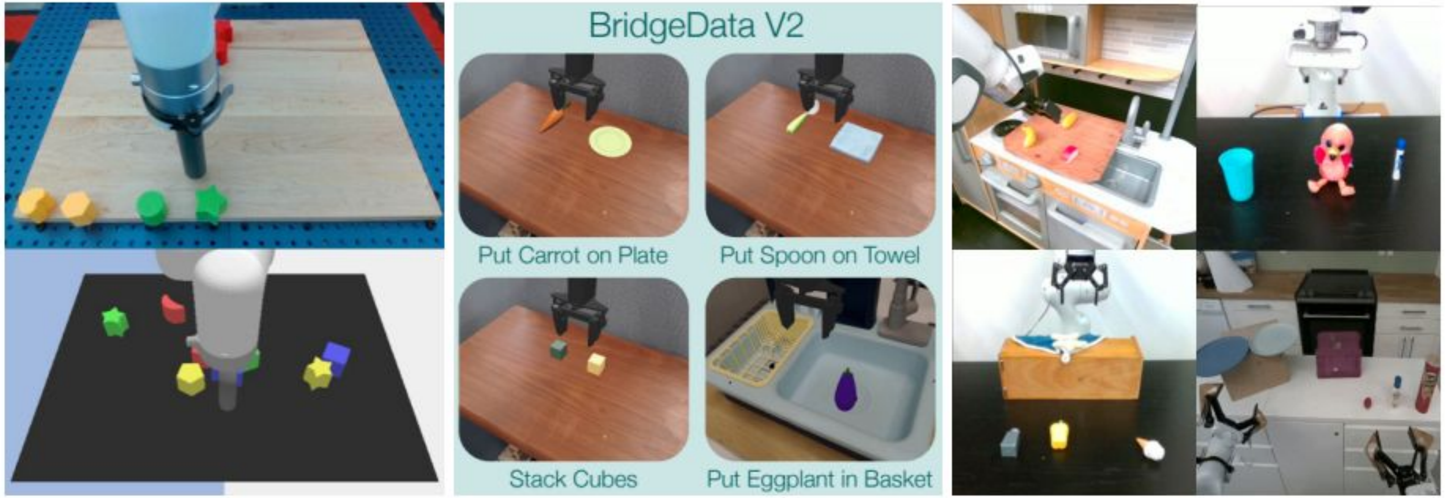
Experiments: Setups

440K Real world trajectories

4 diff eval tasks

4 diff real world tasks

181K Simulation trajectories



(a) LANGUAGE TABLE (b) SIMPLER (c) REAL

Figure 3: **Experimental Setups**. (a) shows an example from the 440k real-world trajectories (top) and the 181k simulation trajectories (bottom) from the Language Table Benchmark. (b) shows the 4 different evaluation tasks we use with the SIMPLER environment. (c) shows the four different tasks that we perform in the real-world.

Pretraining & Finetuning Datasets

Table 1: **Pretraining and fine-tuning dataset for each environment.** Cross-Env denotes cross-environment, Cross-Emb denotes cross-embodiment, and Multi-Emb denotes multi-embodiment. For fine-tuning, MT denotes multi-task training and MI denotes tasks with diverse multi-instructions. Category denotes the main capability we are trying to quantify. Illustration of each environment is shown in Figure 3.

Environment	**Category**	**Pretraining**		**Fine-tuning**	
		Dataset	# Trajs	Dataset	# Trajs
LangTable	In-Domain	Sim (All 5 tasks)	181k	5 Tasks (MT, MI)	1k
	Cross-Task	Sim (All 5 tasks)	181k	1 Task (MI)	7k
	Cross-Env	Real (All 5 tasks)	442k	5 tasks (MT, MI)	1k
SIMPLER	In-Domain	Bridgev2	60k	4 Tasks (MT)	100
	Cross-Emb	Something v2	220k	4 Tasks (MT)	100
Real-World	Cross-Emb	Bridgev2	60k	3 tasks (MI)	450
	Multi-Emb	Open-X	970k	3 tasks (MI)	450
	Cross-Emb	Open-X	970k	1 task (MI, Bi-manual)	150
	Cross-Emb	Something v2	220k	3 tasks (MI)	450

Results

Table 2: **Language Table Results.** Average Success Rate (%) across the three different pretrain-finetune combinations from the Language Table benchmark as described in Table 1. We also note the # of trajectories used for fine-tuning next to each category. We report the performance for individual tasks in Appendix E.1.

	In-domain (1k)		Cross-task (7k)		Cross-env (1k)	
	Seen	Unseen	Seen	Unseen	Seen	Unseen
SCRATCH	$15.6_{\pm 9.2}$	$15.2_{\pm 8.3}$	$27.2_{\pm 13.6}$	$22.4_{\pm 11.0}$	$15.6_{\pm 9.2}$	$15.2_{\pm 8.3}$
UNIPI	$22.0_{\pm 12.5}$	$13.2_{\pm 7.7}$	$20.8_{\pm 12.0}$	$16.0_{\pm 9.1}$	$13.6_{\pm 8.6}$	$12.0_{\pm 7.5}$
VPT	$44.0_{\pm 7.5}$	$32.8_{\pm 4.6}$	$72.0_{\pm 6.8}$	$\mathbf{60.8}_{\pm 6.6}$	$18.0_{\pm 7.7}$	$18.4_{\pm 9.7}$
LAPA	$\mathbf{62.0}_{\pm 8.7}$	$\mathbf{49.6}_{\pm 9.5}$	$\mathbf{73.2}_{\pm 6.8}$	$54.8_{\pm 9.1}$	$\mathbf{33.6}_{\pm 12.7}$	$\mathbf{29.6}_{\pm 12.0}$
ACTIONVLA	$77.0_{\pm 3.5}$	$58.8_{\pm 6.6}$	$77.0_{\pm 3.5}$	$58.8_{\pm 6.6}$	$64.8_{\pm 5.2}$	$54.0_{\pm 7.0}$

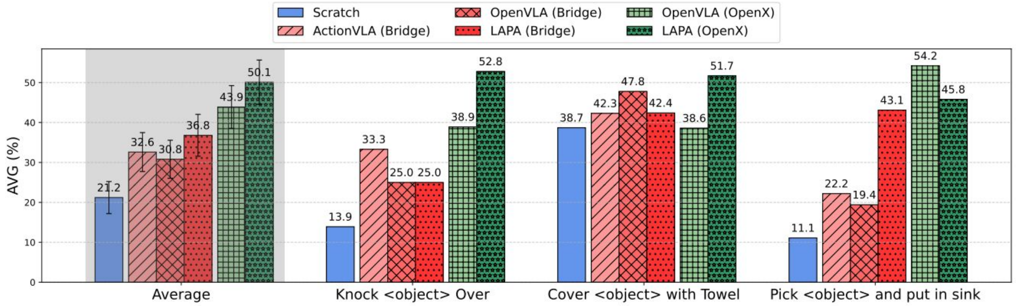


Figure 5: **Real-world Tabletop Manipulation Results.** We evaluate on a total of 54 rollouts for each model encompassing unseen object combinations, unseen objects and unseen instructions. Average success rate (%) are shown. We provide detailed results depedning on the generalization type in Table 12 and individual results in Appendix E.3.

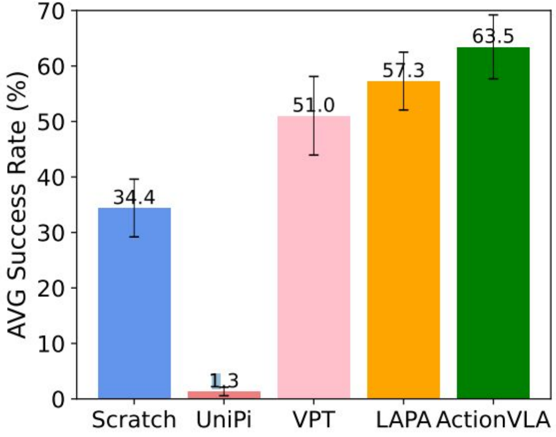
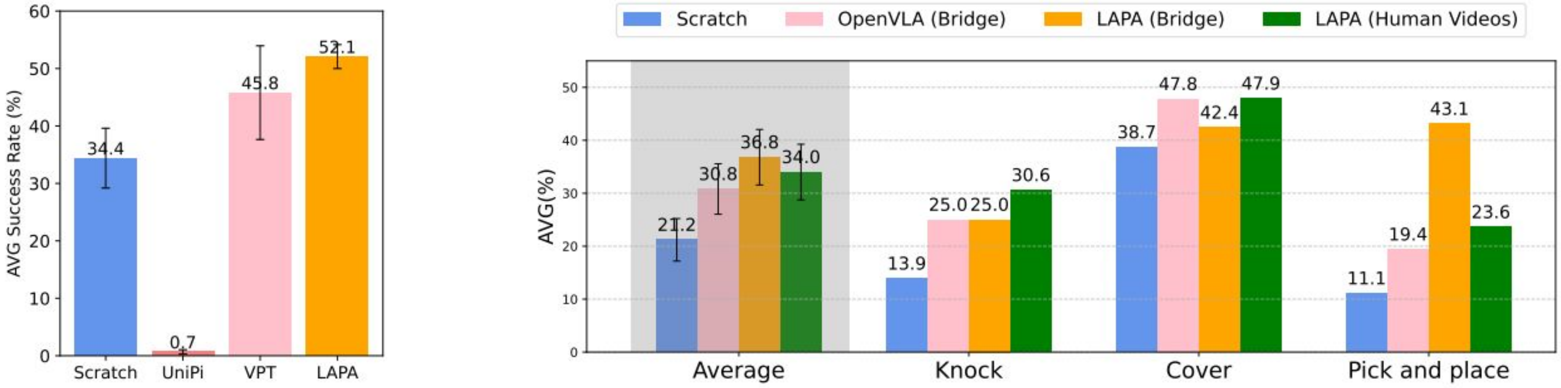


Figure 4: **SIMPLER Results.** Avg. success rate (%) is shown across 4 tasks. Detailed results are in Appendix E.2.

Results



(a) SIMPLER Results

(b) Real-world Tabletop Manipulation Robot Results

Figure 6: **Pretraining from Human Video Results.** Average success rate (%) of LAPA and baselines pre-trained on human manipulation videos where the embodiment and environment gap is extreme. We evaluate on both simulation (left) and real-world robot setup (right).

Latent Action Analysis

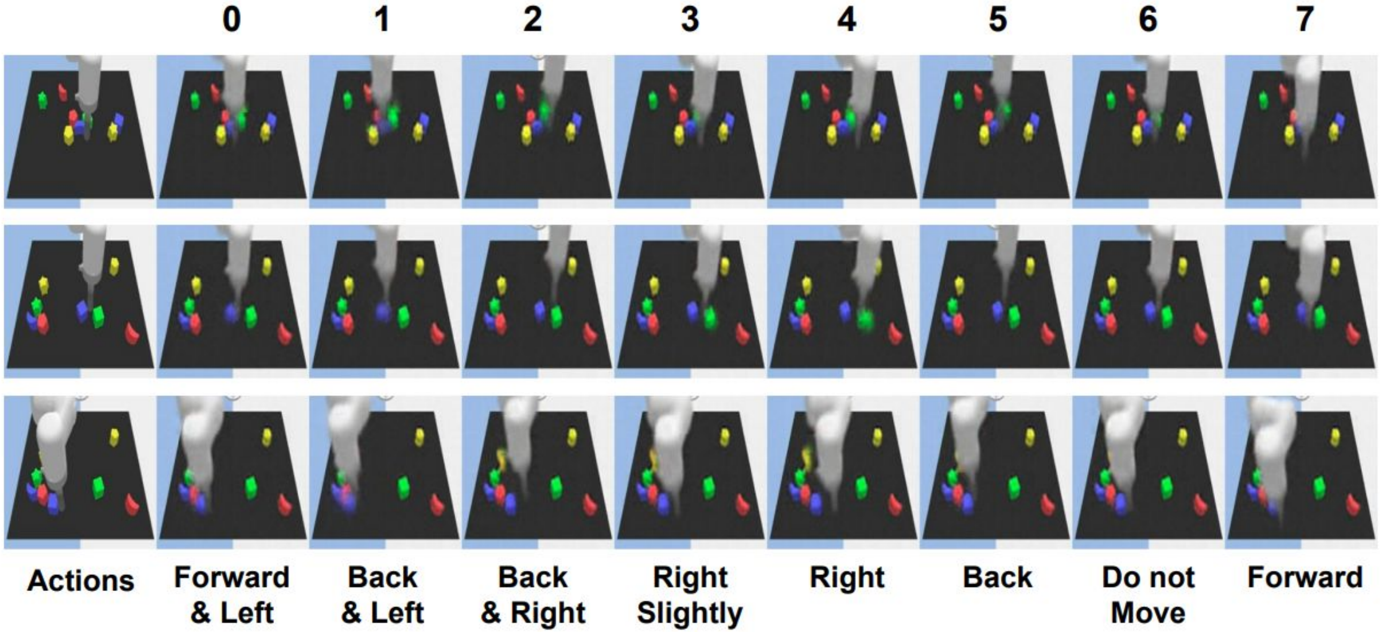


Figure 8: **Latent Action Analysis in Language Table.** We condition the current observation x_1 and quantized latent action to the decoder of the latent action quantization model. We observe that each latent action can be mapped into a semantic action. For example, latent action 0 corresponds to moving a bit left and forward.

Latent Action Analysis

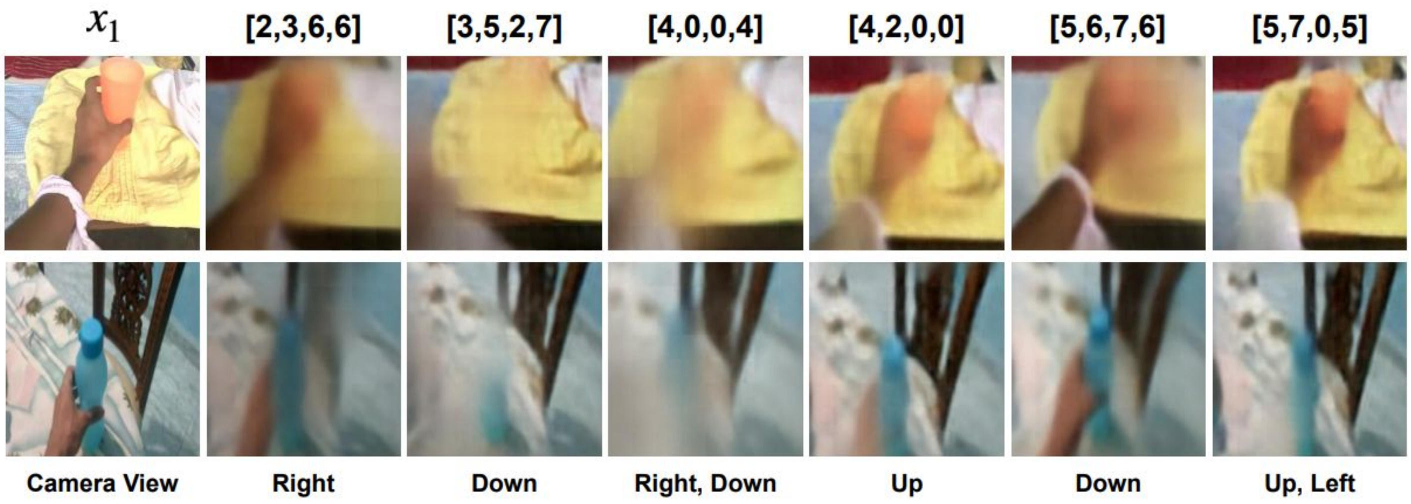


Figure 10: **Latent Action Analysis in Human Manipulation Videos.** We condition the current observation x_1 and quantized latent action to the decoder of the latent action quantization model. We observe that each latent action can be mapped into a semantic action including camera movements. For example, latent action [3,5,2,7] corresponds to moving the camera a bit down while [4,2,0,0] corresponds to moving the camera slightly up.

Something-Something V2 dataset:
https://www.qualcomm.com/developer/software/something-something-v-2-dataset

Latent Action Analysis

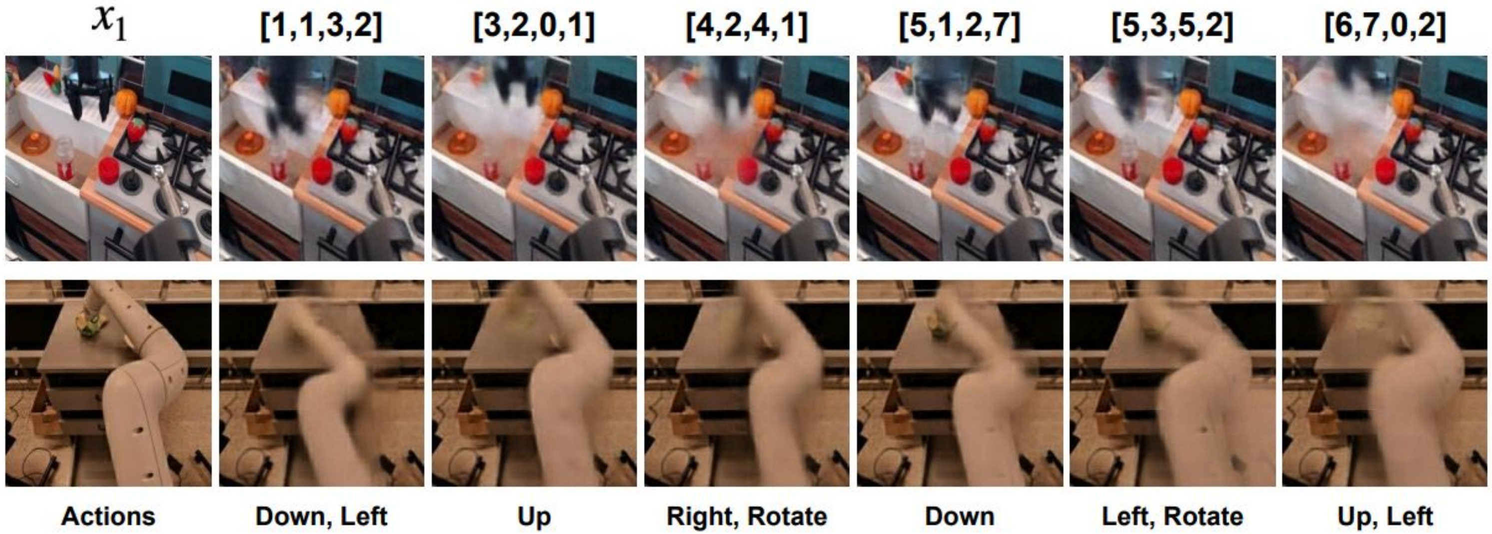


Figure 11: **Latent Action Analysis in Multi-Embodiment Setting.** We condition the current observation x_1 and quantized latent action to the decoder of the latent action quantization model. We observe that each latent action can be mapped into a similar semantic action even though the embodiments are different. For example, latent action [1,1,3,2] corresponds to going down and left while [3,2,0,1] corresponds to going up a little bit.

Latent Action Analysis

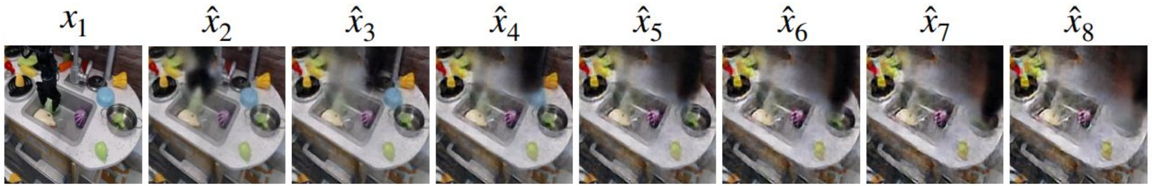
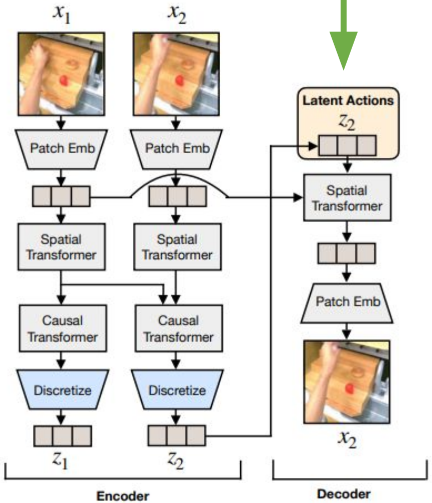
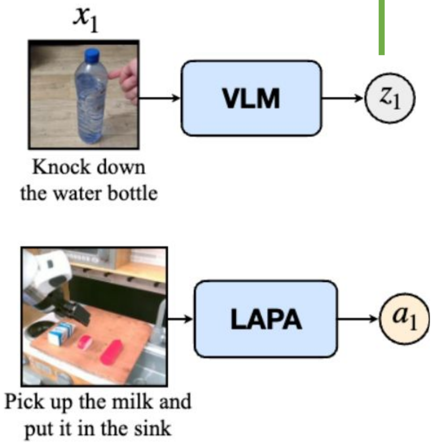


Figure 12: **Closed loop rollout of LAPA.** LAPA is conditioned on current image x_1 and language instruction of 'take the broccoli out of the pot'. We generate rollout images by conditioning the decoder of Latent Action Quantization Model with latent actions generated by LAPA.

Surprisingly can act as a potential world model

GT LAPA





Limitations

LAPA underperforms compared to action pretraining when it comes to fine-grained motion generation tasks like grasping. Increasing the latent action generation space could help address this issue.

Latency challenges during real-time inference. Adopting a hierarchical architecture, where a smaller head predicts actions at a higher frequency, could potentially reduce latency and improve fine-grained motion generation.

The application of LAPA beyond manipulation videos, such as those from self-driving cars, navigation, or landscape scenes need to be explored.

Takeaways

- Same as Mimic-Play
  - Learning latent plans from human play data significantly improves performance.
  - Latent plan pre-training benefits multi-task learning.
- LAPA
  - A scalable pretraining method for building VLAs using actionless videos.
  - A state-of-the-art VLA model that surpasses current models trained on 970K action-labeled trajectories.
  - LAPA can be applied purely on human manipulation videos, where explicit action information is absent, and the embodiment gap is substantial.

Questions?