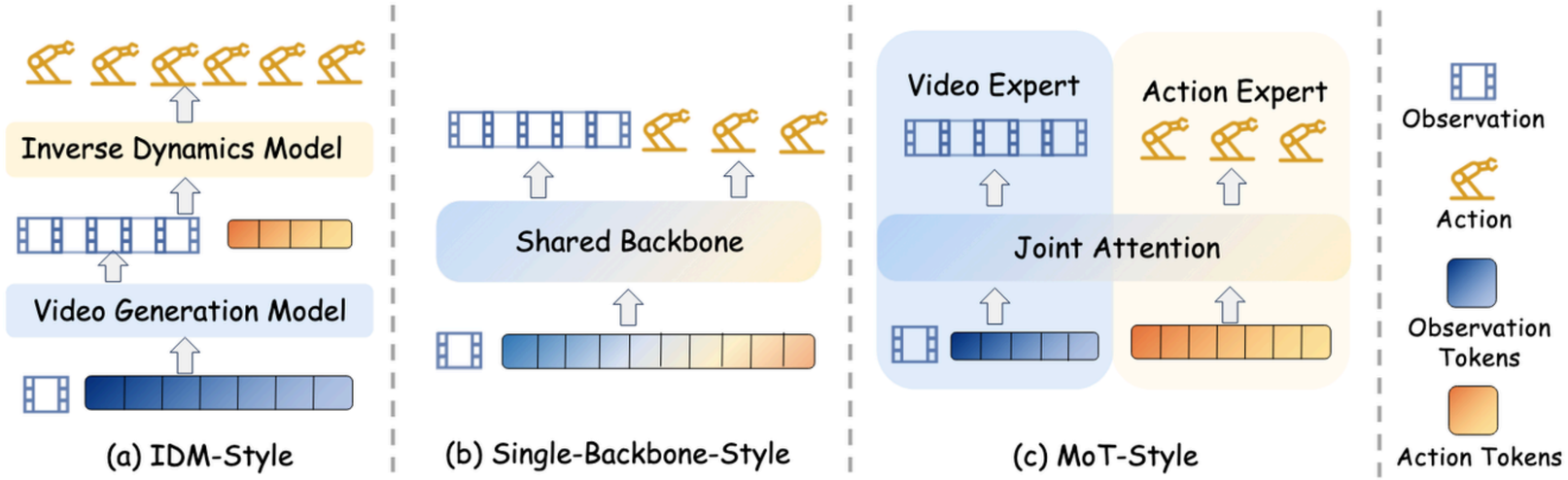
GR-1: LARGE-SCALE VIDEO GENERATIVE PRE-TRAINING FOR VISUAL ROBOT MANIPULATION

World Model Seminar

Presented by Mostafa Kotb

http://www.informatik.uni-hamburg.de/WTM/

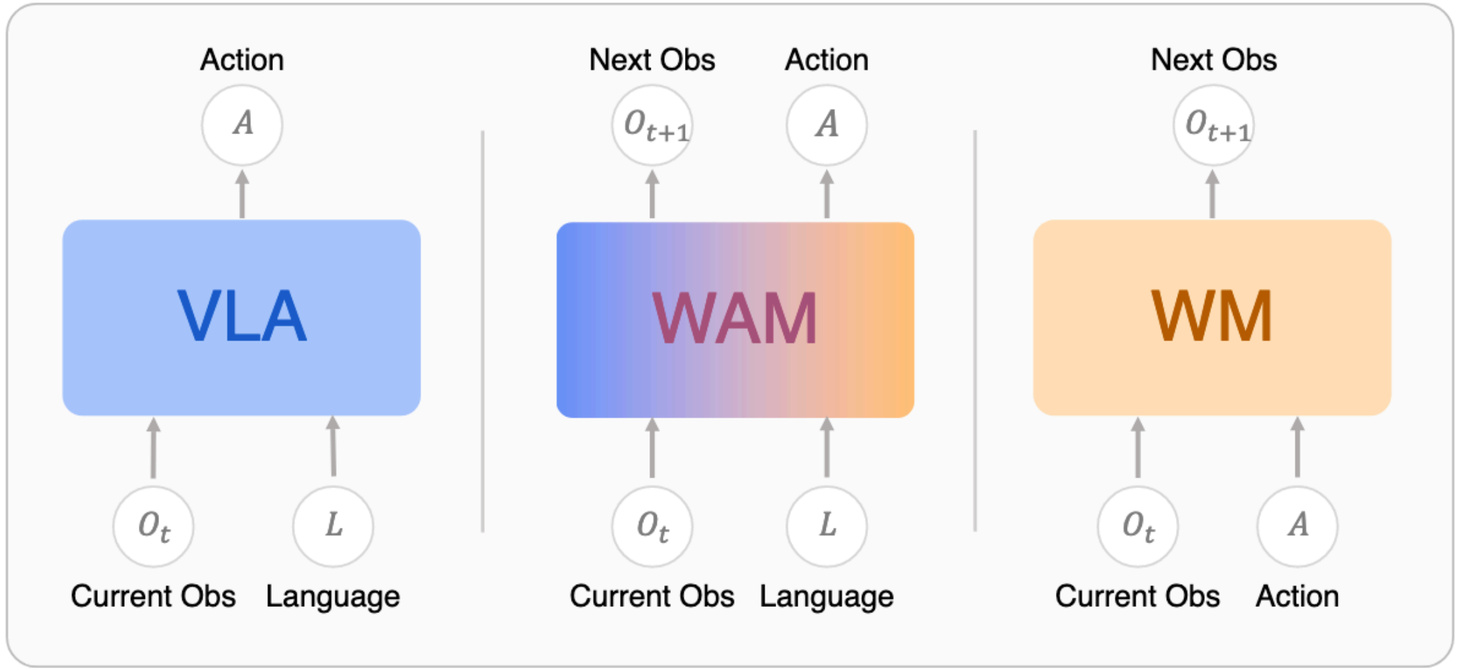
Recap



- The Single-Backbone-Style architecture is referred to **World Action Models (WAMs).**

KNOWLEDGE TECHNOLOGY

WMs,WAMs,VLAs



$$\mathcal{L}_{\text{VLA}} = \mathbb{E}_{(o,l,a)\sim\mathcal{D}}\left[-\log p(a \mid o, l)\right]$$

$$\mathcal{L}_{\text{WM}} = \mathbb{E}_{(o,a,o')\sim\mathcal{D}}\left[-\log p(o' \mid o, a)\right]$$

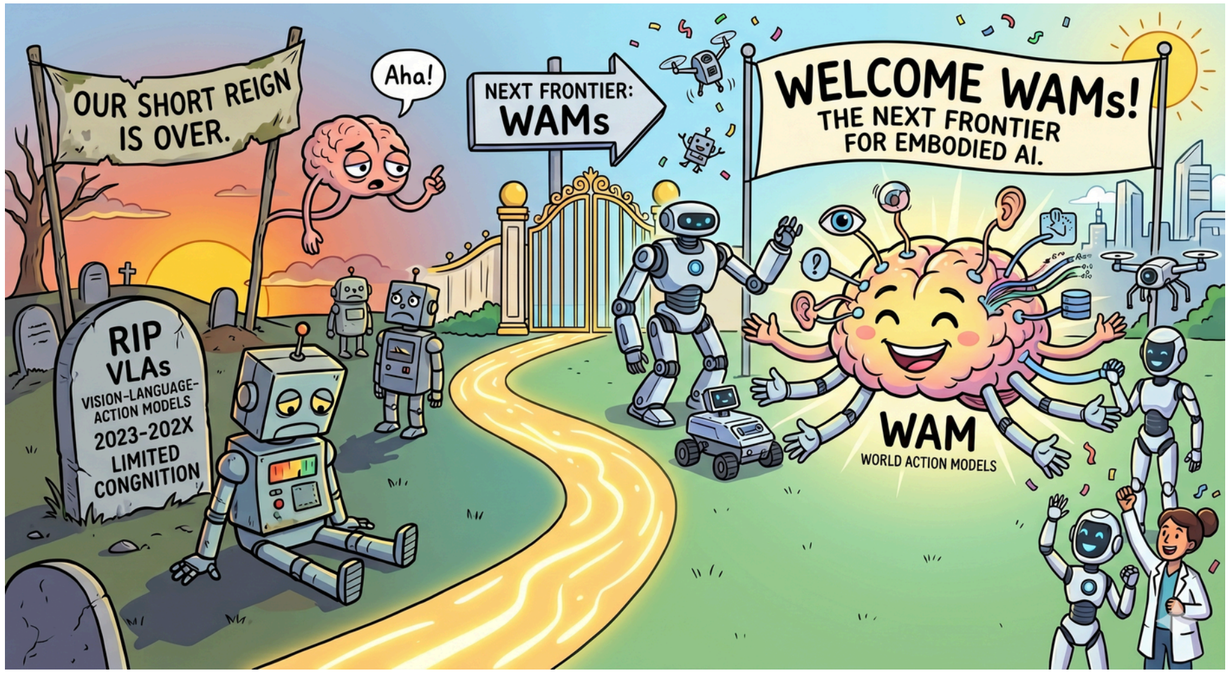
$$\mathcal{L}_{\text{WAM}} = \mathbb{E}_{(o,l,o',a)\sim\mathcal{D}}\left[-\log p(o', a \mid o, l)\right]$$

KNOWLEDGE TECHNOLOGY

WAMs are the next frontier for Embodied AI

Limitations of VLAs:

- *Data Scarcity:* teleoperated robotic data is expensive to collect and difficult to scale.
- *Purely Reactive:* VLAs policies are reactive which remain limited in complex physical tasks.



These limitations stem from the lack of *explicit predictive structure* for anticipating how the world may evolve under the agent's behavior **[1]**.

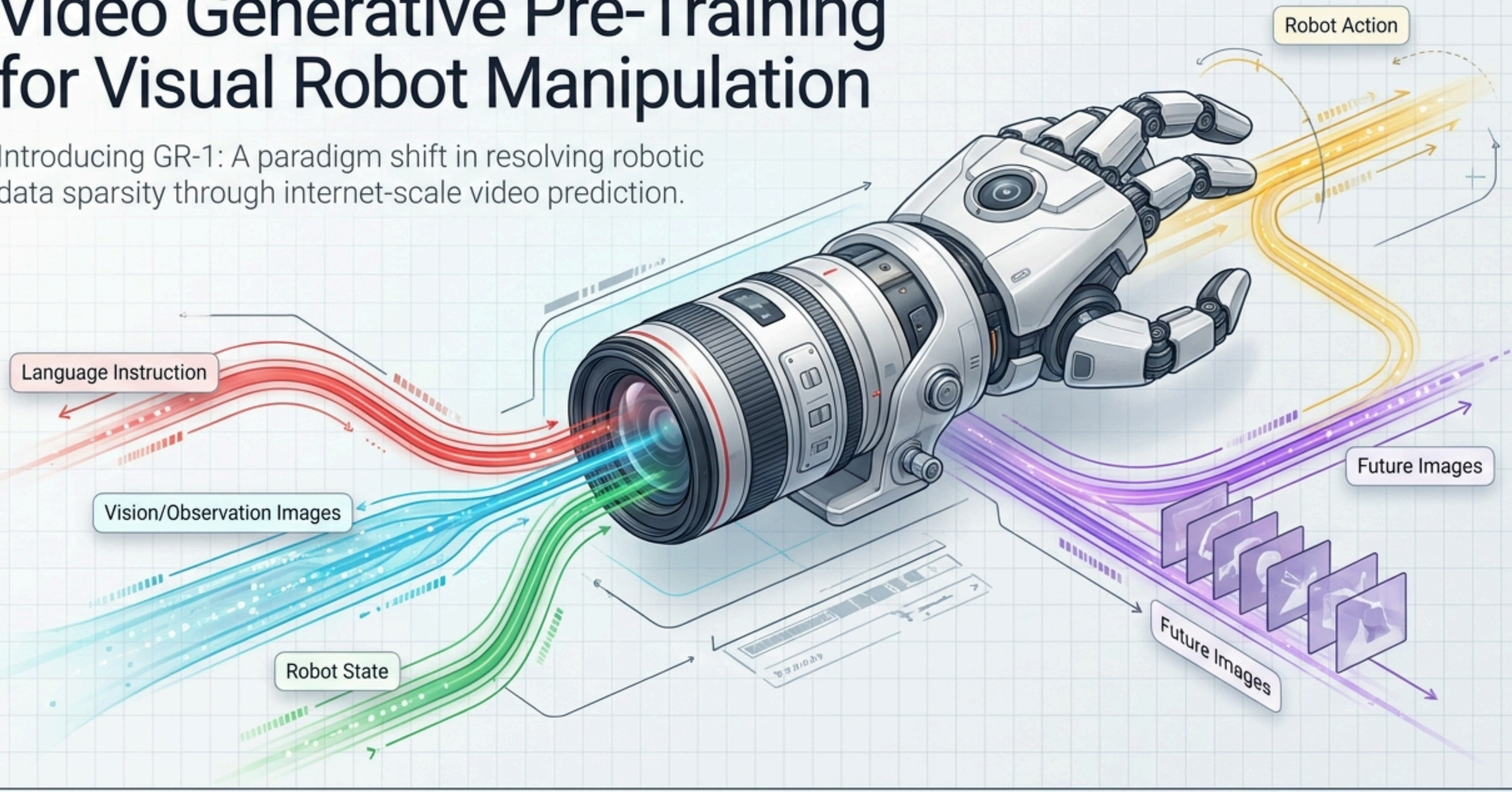
The World Model part in **WAMs** capture environmental dynamics and enable *reasoning about future states before acting*

[1] Ye, Seonghyeon, et al. "World action models are zero-shot policies." arXiv preprint arXiv:2602.15922 (2026).

KNOWLEDGE TECHNOLOGY

Video Generative Pre-Training for Visual Robot Manipulation

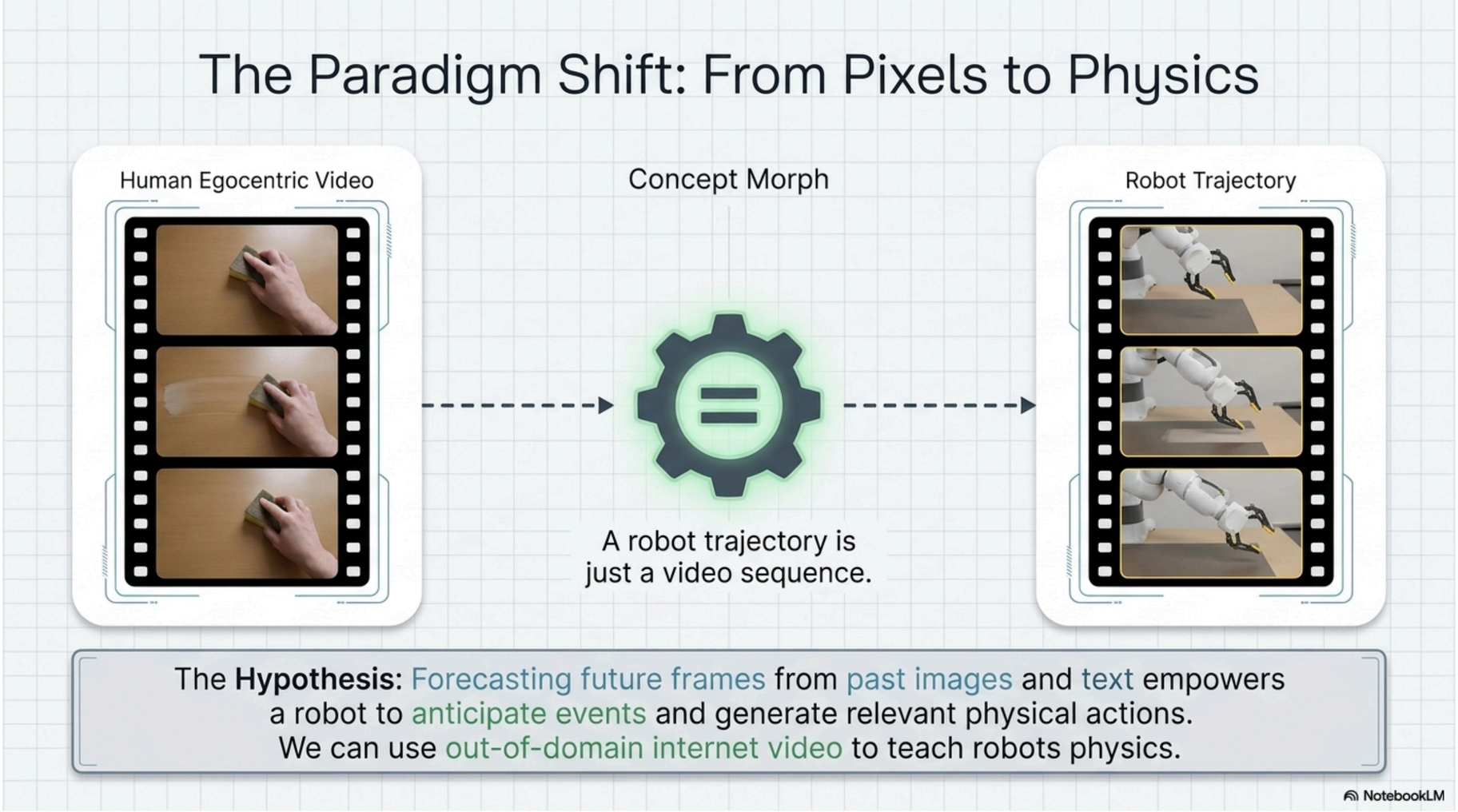
Introducing GR-1: A paradigm shift in resolving robotic data sparsity through internet-scale video prediction.



GR-1

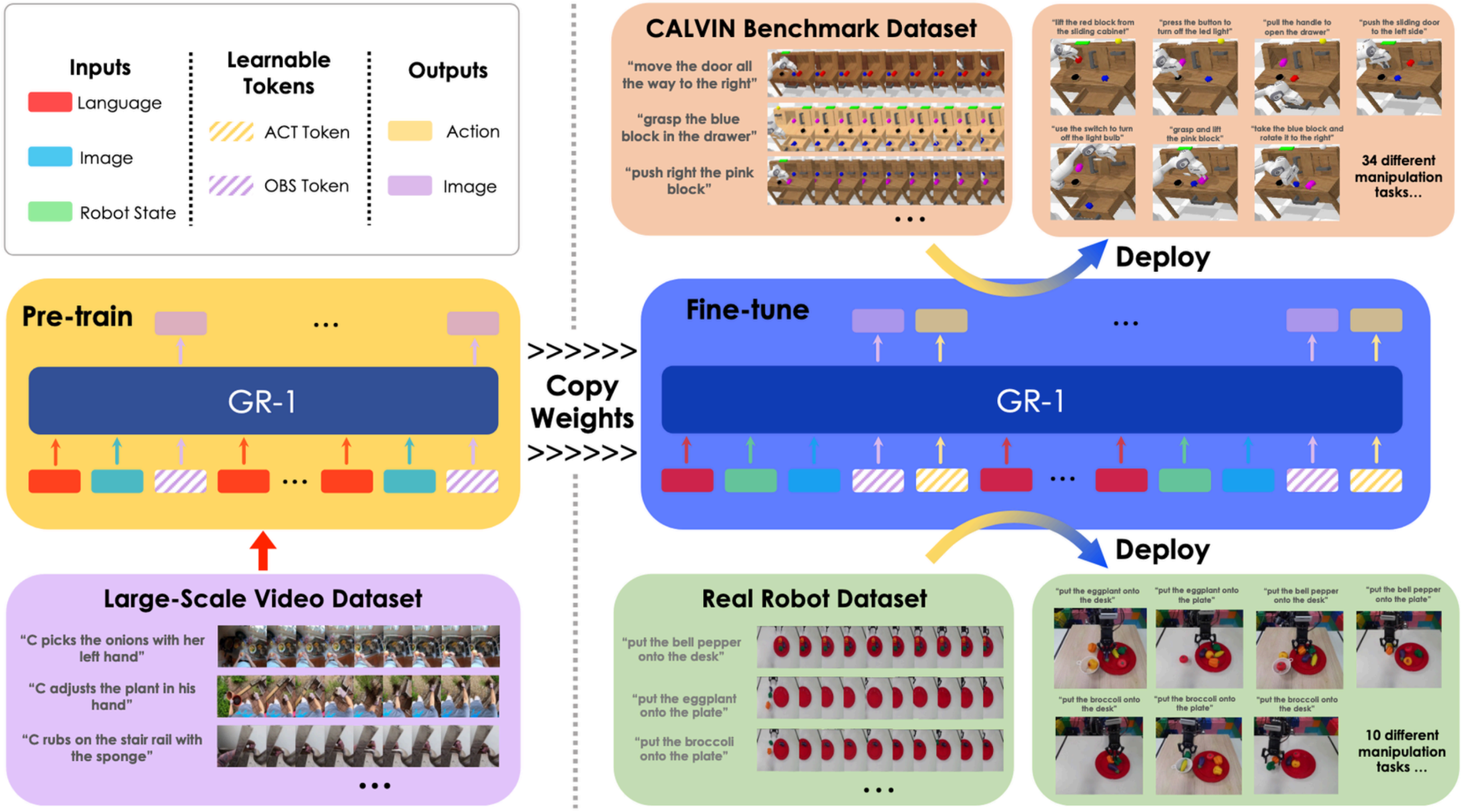
The idea:

- Leverage the large-scale video data for *video generative pre-training*.
- The pre-training phase teaches *physics* to the robot.



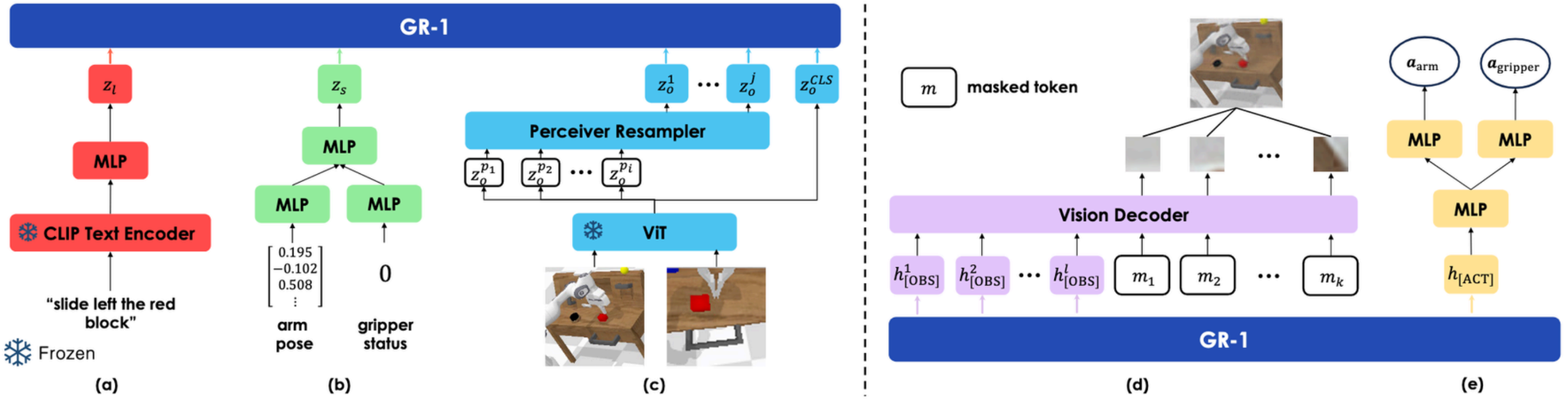
KNOWLEDGE TECHNOLOGY

GR-1 Training Overview



KNOWLEDGE TECHNOLOGY

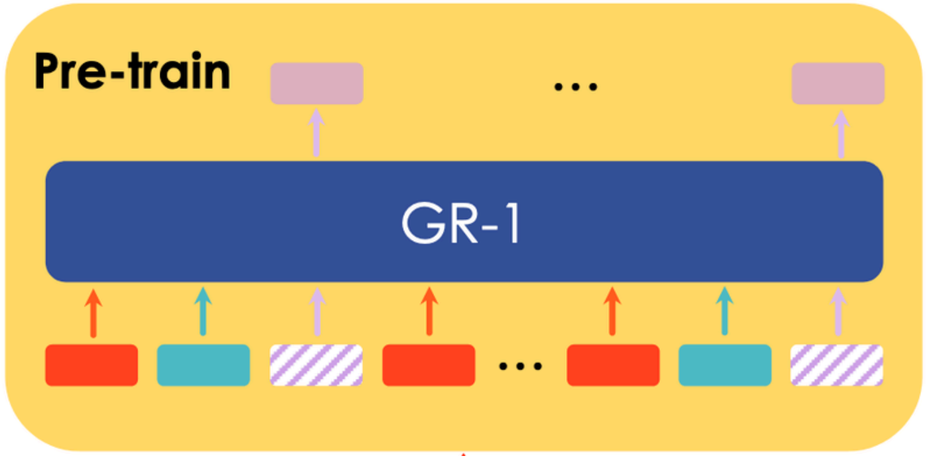
GR-1 Architecture



KNOWLEDGE TECHNOLOGY

Video Generative Pre-training

- **Dataset:** **Ego4D** [1] which contains more than 3,500 hours of language annotated human tasks.
- **Task:** $\pi(l, \mathbf{o}_{t-h:t}) \rightarrow \mathbf{o}_{t+\Delta t}$
- **Input Tokens:** $(l, \mathbf{o}_{t-h}, \texttt{[OBS]}, l, \mathbf{o}_{t-h+1}, \texttt{[OBS]}, ..., l, \mathbf{o}_{t}, \texttt{[OBS]})$
- **Loss:** Image reconstruction **MSE**

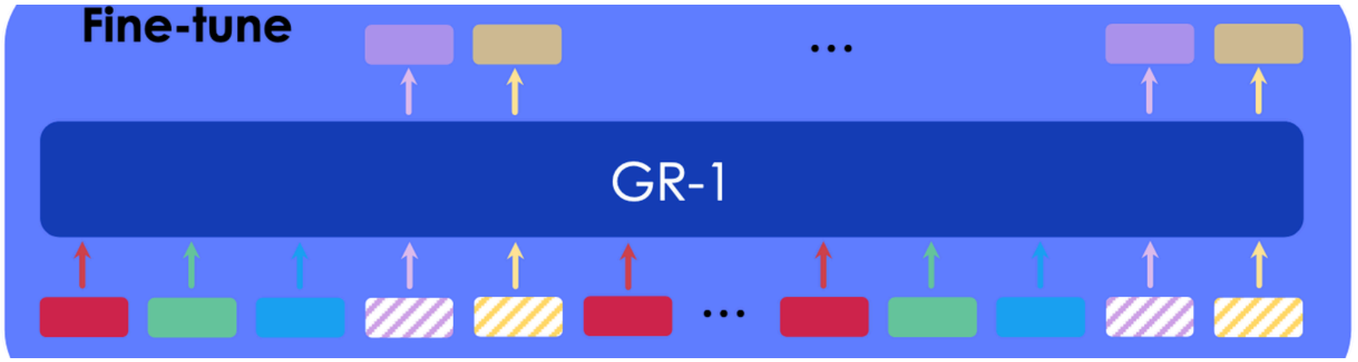


[1] Grauman, Kristen, et al. "Ego4d: Around the world in 3,000 hours of egocentric video." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

KNOWLEDGE TECHNOLOGY

Robot Data Finetuning

- **Dataset:** teleoperated real robot data or simulated such as **CALVIN benchmark** [1]
- **Task:** $\pi(l, \mathbf{o}_{t-h:t}, \mathbf{s}_{t-h:t}) \rightarrow \mathbf{o}_{t+\Delta t}, \mathbf{a}_t$
- **Input Tokens:** $(l, \mathbf{s}_{t-h}, \mathbf{o}_{t-h}, \texttt{[OBS]}, \texttt{[ACT]}, l, \mathbf{s}_{t-h+1}, ..., l, \mathbf{s}_t, \mathbf{o}_t, \texttt{[OBS]}, \texttt{[ACT]})$
- **Loss:** **Smooth-L1** (arm) + **Binary Cross Entropy** (gripper) + **MSE** (future frames)

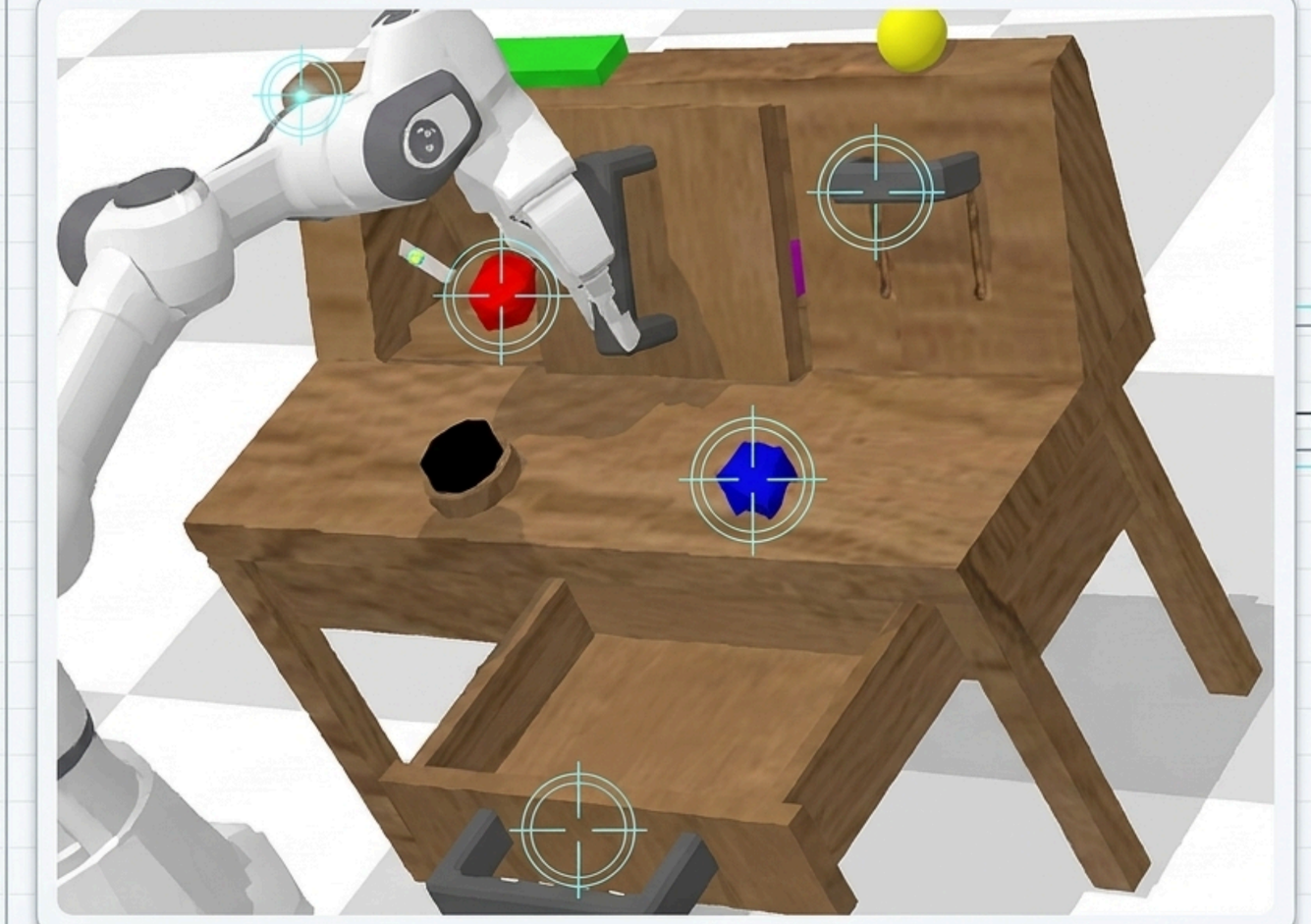


[1] Mees, Oier, et al. "Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks." IEEE Robotics and Automation Letters 7.3 (2022): 7327-7334.

KNOWLEDGE TECHNOLOGY

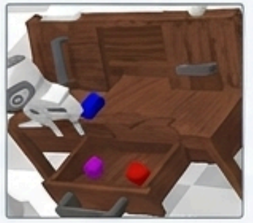
CALVIN Benchmark Experiment

Proving Ground: The CALVIN Benchmark



Env A

Env B



Env C

Env D

Benchmark Specifications:

- 34 distinct manipulation tasks.
- Unconstrained, open-ended language instructions.
- Primary Testing Metric: Sequential execution of 5 complex tasks in a row without failure.

NotebookLM

KNOWLEDGE TECHNOLOGY

CALVIN Benchmark Results

Table 1: CALVIN Benchmark Results.

Method	Experiment	Tasks completed in a row					
		1	2	3	4	5	Avg. Len.
MCIL	ABCD→D	0.373	0.027	0.002	0.000	0.000	0.40
RT-1	ABCD→D	0.844	0.617	0.438	0.323	0.227	2.45
HULC	ABCD→D	0.889	0.733	0.587	0.475	0.383	3.06
MT-R3M	ABCD→D	0.752	0.527	0.375	0.258	0.163	2.08
GR-1 (Ours)	ABCD→D	**0.949**	**0.896**	**0.844**	**0.789**	**0.731**	**4.21**
MCIL	ABC→D	0.304	0.013	0.002	0.000	0.000	0.31
RT-1	ABC→D	0.533	0.222	0.094	0.038	0.013	0.90
HULC	ABC→D	0.418	0.165	0.057	0.019	0.011	0.67
MT-R3M	ABC→D	0.529	0.234	0.105	0.043	0.018	0.93
GR-1 (Ours)	ABC→D	**0.854**	**0.712**	**0.596**	**0.497**	**0.401**	**3.06**
RT-1	10% data	0.249	0.069	0.015	0.006	0.000	0.34
HULC	10% data	0.668	0.295	0.103	0.032	0.013	1.11
MT-R3M	10% data	0.408	0.146	0.043	0.014	0.002	0.61
GR-1 (Ours)	10% data	**0.778**	**0.533**	**0.332**	**0.218**	**0.139**	**2.00**
RT-1	unseen lang	0.494	0.222	0.086	0.036	0.017	0.86
HULC	unseen lang	0.715	0.470	0.308	0.199	0.130	1.82
MT-R3M	unseen lang	0.512	0.249	0.106	0.040	0.017	0.92
GR-1 (Ours)	unseen lang	**0.764**	**0.555**	**0.381**	**0.270**	**0.196**	**2.17**

CALVIN Benchmark Results

Table 6: Examples of Unseen Language Instructions Generated by GPT-4 (OpenAI, 2023) for the Zero-Shot Unseen Language Generalization Experiment in CALVIN.

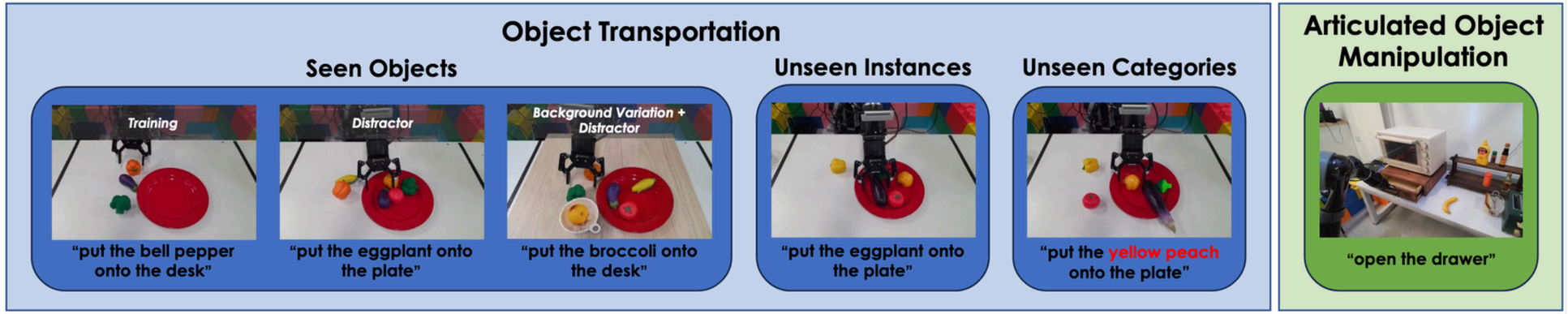
Original	Generated
“use the switch to turn off the light bulb”	“use the switch to stop the light source”
“slide the block that it falls into the drawer”	“Move the block ensuring it goes into the drawer”
“pull the handle to open the drawer”	“Acquire a grip on the handle to slide the drawer out”
“lift the pink block from the sliding cabinet”	“Hoist up the pink block kept in the sliding cabinet”
“store the grasped block in the sliding cabinet”	“Hide the item gripped in sliding stash”
“take the red block and rotate it to the right”	“Twist the red object to the right”
“press the button to turn on the led light”	“press the switch to turn on the glowing LED”
“grasp and lift the blue block”	“grasp firmly and raise the blue block”

CALVIN Benchmark Ablations

Table 4: Ablation Studies.

Pre-Training	Video Prediction	Data	Tasks completed in a row					
			1	2	3	4	5	Avg. Len.
✗	✗	ABCD→D	0.889	0.775	0.661	0.549	0.459	3.33
✗	✓	ABCD→D	0.918	0.833	0.761	0.685	0.619	3.82
✓	✓	ABCD→D	**0.949**	**0.896**	**0.844**	**0.789**	**0.731**	**4.21**
✗	✗	ABC→D	0.823	0.609	0.425	0.318	0.225	2.40
✗	✓	ABC→D	0.815	0.651	0.498	0.392	0.297	2.65
✓	✓	ABC→D	**0.854**	**0.712**	**0.596**	**0.497**	**0.401**	**3.06**
✗	✗	10% data	0.526	0.288	0.138	0.061	0.022	1.04
✗	✓	10% data	0.698	0.415	0.223	0.133	0.052	1.52
✓	✓	10% data	**0.778**	**0.533**	**0.332**	**0.218**	**0.139**	**2.00**

Real Robot Experiments



- **Object Transportation:**
  Collected **1775** demonstrations of transporting one of the tree objects from the plate to the desk and vice versa.

- **Articulated Object Manipulation:**
  Collected **2856** demonstrations of opening and closing the drawer.

KNOWLEDGE TECHNOLOGY

Real Robot Results

Table 2: Real Robot Experiment Results.

Method	Object Transportation			Articulated Object Manipulation
	Seen Objects	Unseen Instances	Unseen Categories	
RT-1	0.27	0.13	0.00	0.35
MT-R3M	0.15	0.13	0.10	0.30
GR-1 (Ours)	**0.79**	**0.73**	**0.30**	**0.75**

```
# Action query token
self.action_queries = nn.Embedding(1, hidden_size) # arm + gripper

# Observation query token
self.obs_queries = nn.Embedding(self.n_patch_latents + 1, self.hidden_size)
self.obs_hand_queries = nn.Embedding(self.n_patch_latents + 1, self.hidden_size)
```