Dynin-Robotics

Omnimodal Unified Diffusion Vision-Language-Action Model

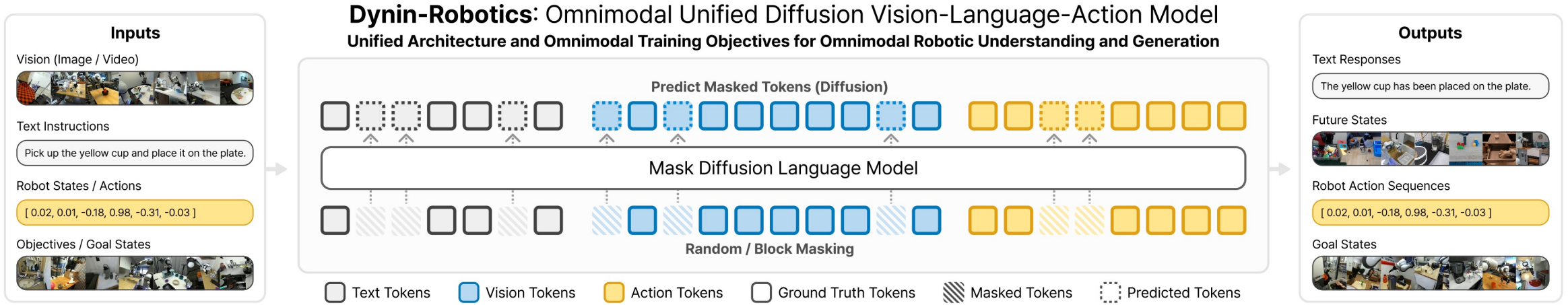
Hoeun Lee[1], Jaeik Kim[2], Jusang Oh[2], Jinhyeok Kim[1], Geon Choi[1], Hyeonggeun Kim[2], Jaeyoung Do[1,2]

AIDAS Laboratory, [1]ECE & [2]IPAI, Seoul National University

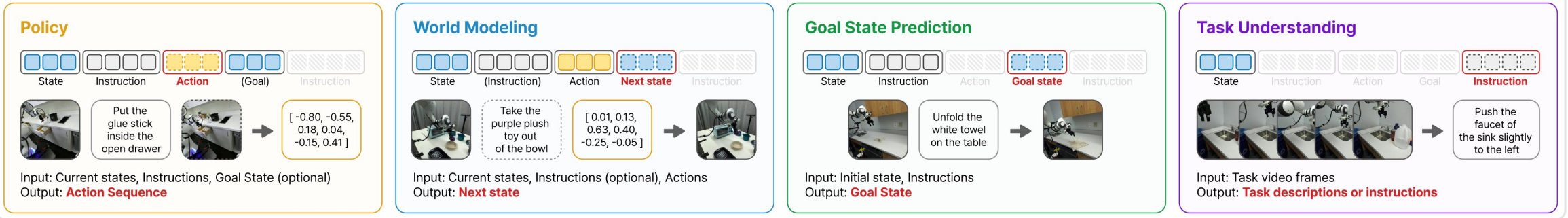
NeurIPS 2026

AIDAS Lab

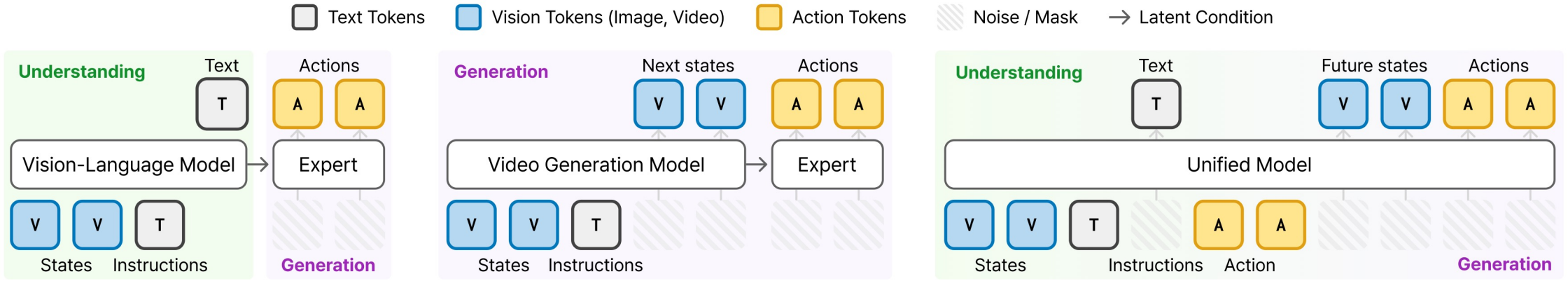
Overview



Unified Objectives Training and Inference



Why Unified Models?



(a) Vision-Language Model **(b) Video Generation Model** **(c) Unified Model (Ours)**

- VLM-based policies: Capture semantic understanding, while relying on a separate action expert for robot control.
- Video-based policies: Predict future visual dynamics, while using a separate action expert to generate robot actions.
- Unified models: Jointly learn semantic understanding, visual prediction, and action generation within a single shared backbone.

Motivation

Table 10: Instruction sensitivity on VLABench. Success rate and progress score are measured with the original instruction and a length-matched random string under the same visual observation and rollout settings. Parentheses show the signed change (random minus original).

		Success Rate ↑		Progress Score ↑	
Model	Task	Original	Random	Original	Random
Mimic-Video [20]	INSERTFLOWER	0.18	0.28 (+0.10)	0.58	0.63 (+0.05)
	SELECTFRUIT	0.08	0.10 (+0.02)	0.53	0.53 (0.00)
$\pi_{0.5}$ [4]	INSERTFLOWER	0.32	0.16 (-0.16)	0.65	0.57 (-0.08)
	SELECTFRUIT	0.64	0.34 (-0.30)	0.82	0.53 (-0.29)

- Random instructions reduce π0.5 success, while Mimic-Video retains success.
- This two-task diagnostic motivates learning language and visual dynamics jointly.

→ **Language sensitivity motivates joint modeling of meaning and dynamics**

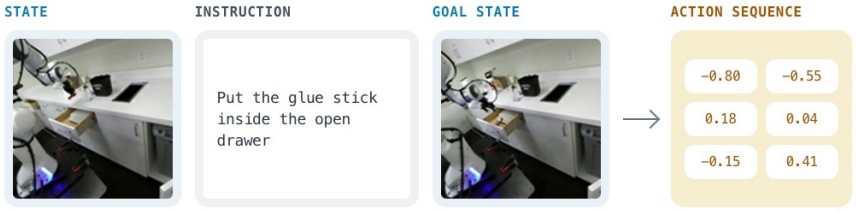
Four Capabilities

Policy

Generate robot action sequences from the current visual state and task instruction.

INPUT: Current states | Instructions | Goal state | Sensor

OUTPUT: Action sequence

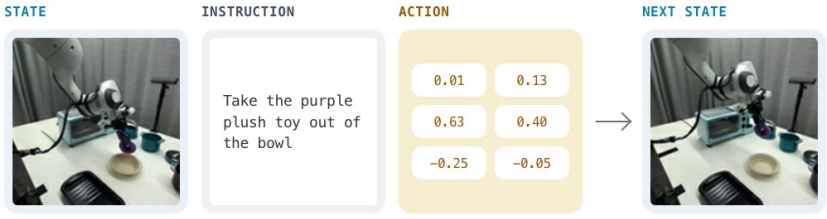


World Modeling

Predict future visual states from observed frames and robot actions.

INPUT: Current states | Instructions | Actions

OUTPUT: Next states



- Policy: current observation/state + instruction → robot action chunk
- World model: current observation/state + robot action → next visual observation/state

Four Capabilities

Goal-State Prediction

Generate a goal state from an initial observation and language instruction.

INPUT Initial state Instructions

OUTPUT Goal state

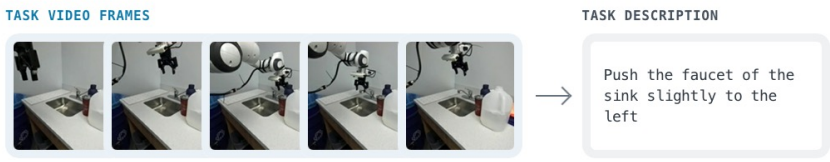


Task Understanding

Decode an observed task trajectory into a natural-language task description.

INPUT Task video frames

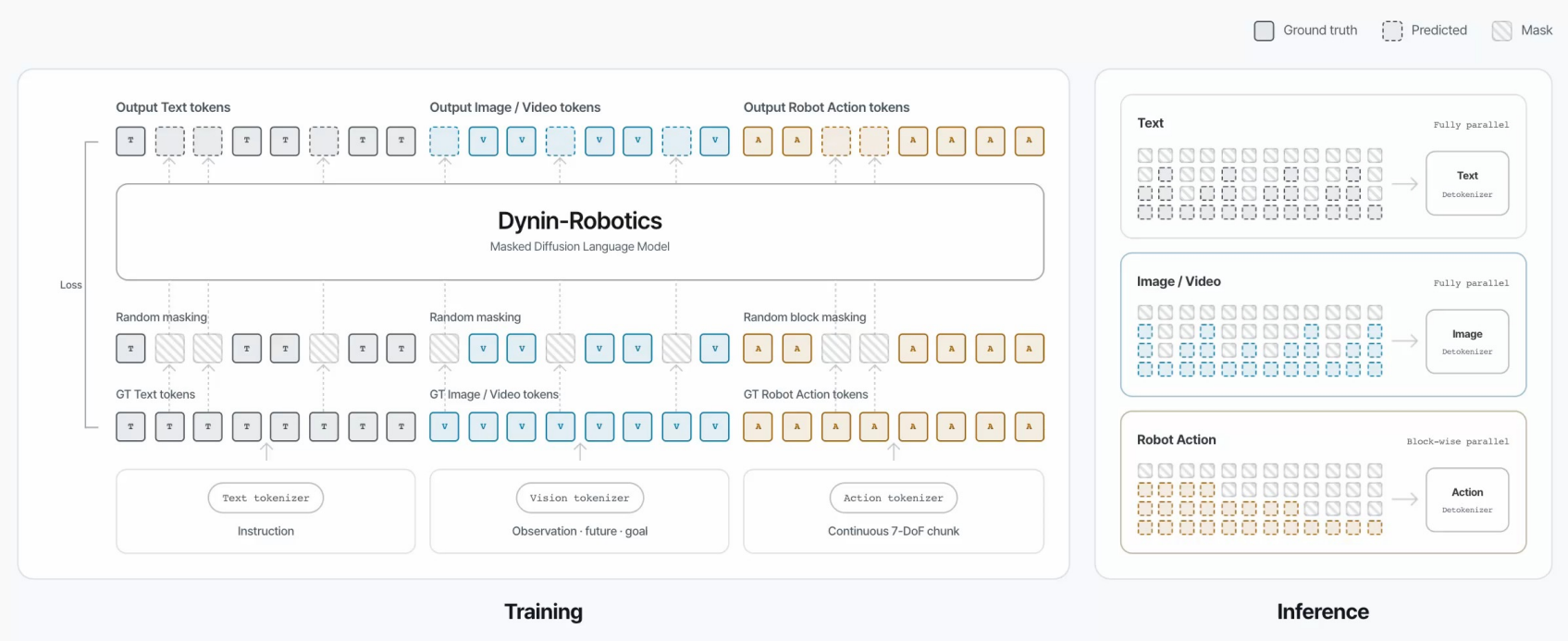
OUTPUT Task description



- Goal prediction: initial observation + instruction → goal state observation
- Task understanding: trajectory video frames → natural-language description/instruction

Methodology: A Shared Diffusion Backbone

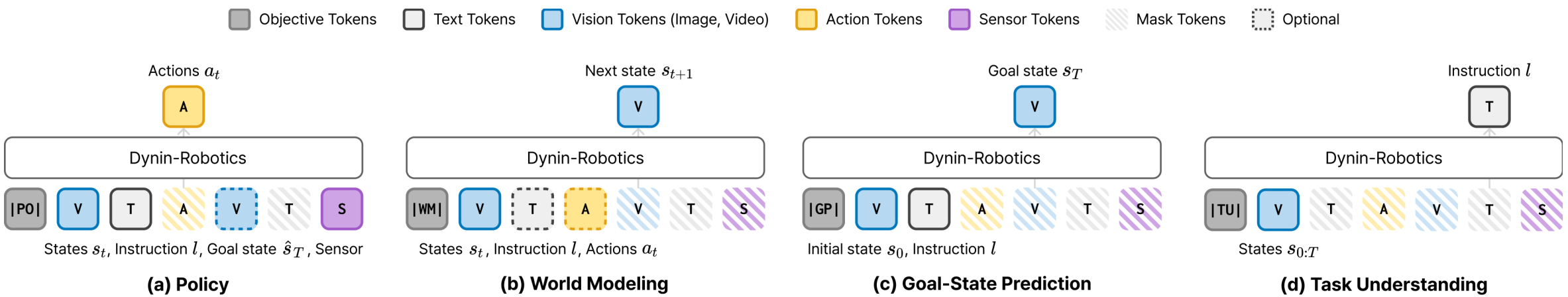
- Text, image, action tokens share a bidirectional diffusion backbone and a categorical prediction head.



- A common token space:
  - Discrete text, image and action tokens
  - One bidirectional diffusion backbone and prediction head
  - Mask target spans; keep conditioning context visible

Omnimodal Unified Objective Training

- Four objectives share one backbone and trajectory representation.
- Each task selects different conditioning tokens and masked targets.



Training Objective: Masked Target Reconstruction

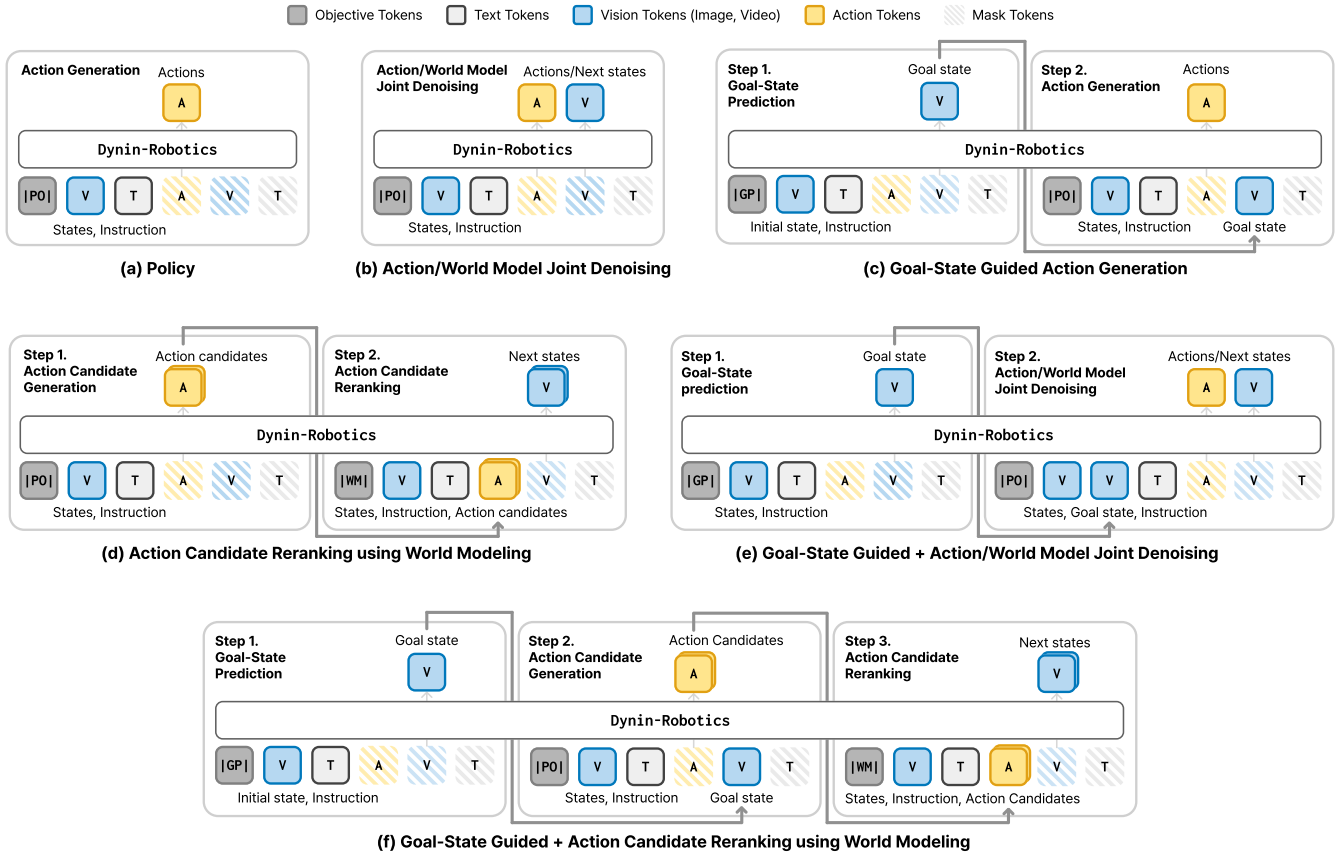
- Compute loss only on masked target tokens.
- Normalize by mask rate and target length; mix the four objectives.

$$\mathcal{L}_{\mathrm{MDM}}(o) = \mathbb{E}\left[-\frac{1}{N_o} \sum_{b=1}^{N_o} \frac{1}{\rho_b L_{o,b}} \times \sum_{i \in \mathcal{M}_{o,b}} \log p_\theta^{(v_o)} \left(x_{b,i}^{(o)} \mid \widetilde{\mathbf{x}}_b^{(o)}, o \right) \right], \tag{9}$$

$$\mathcal{L}_{\mathrm{mix}}^{(n)} = \frac{1}{|\mathcal{A}_n|} \sum_{o \in \mathcal{A}_n} \lambda_o \mathcal{L}_{\mathrm{MDM}}^{(n)}(o). \tag{11}$$

Pre-trained on 1.33M OXE trajectories → adapt to each evaluation domain

Composable Unified Inference



- One checkpoint, six inference modes
  - Predict a visual goal, then generate actions
  - Jointly denoise actions and next-state toke
  - Rerank candidates with the world model

Qualitative Examples

- World Modeling: predict the visual consequences of an action
  - Five observed frames and a robot action condition the visual prediction.
  - Predicted next states expose how an action changes the scene.

Current states, actions

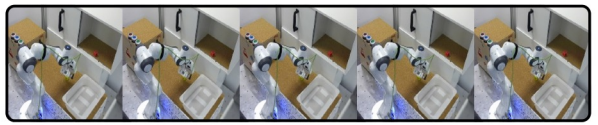


[0.16, 0.20, -0.11, -0.09, 0.22, 0.15, 0.00]

Next-state prediction

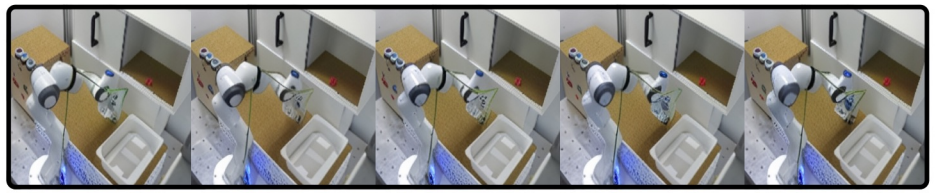


Current states, actions

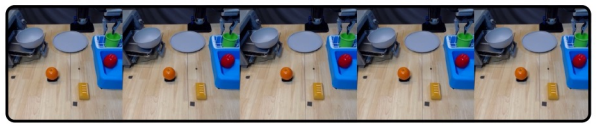


[-0.13, 0.02, -0.49, 0.03, 0.36, -0.04, 0.00]

Next-state prediction



Current states, actions



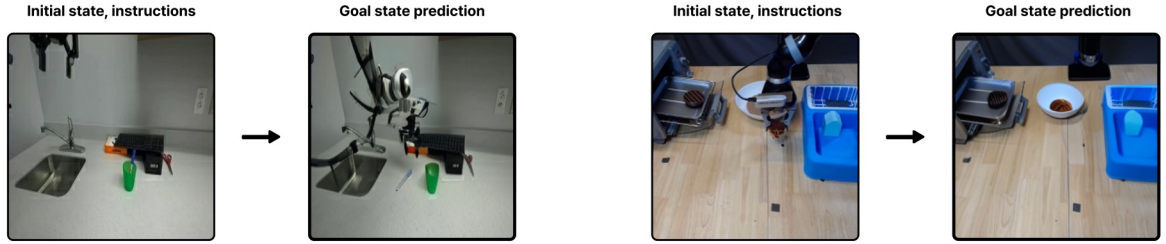
[-0.90, -0.71, 0.03, 0.00, 0.00, 0.00, 0.00]

Next-state prediction



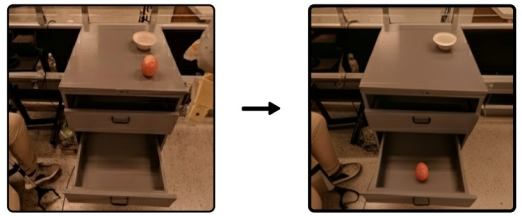
Qualitative Examples

- Goal-State Prediction: imagine the goal from an instruction
  - Predict the goal state from an initial observation and instruction.
  - Use the predicted visual goal as additional context for action generation.

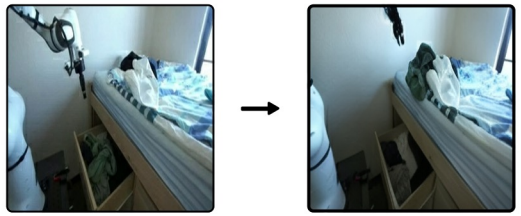


Take the pen out of the cup and place it on the counter

Place the steak meat in the white plate



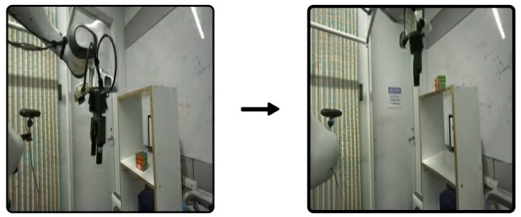
Place apple into middle drawer



Remove the green cloth from the drawer and place it on the bed



Place the t-shirts on the brown box

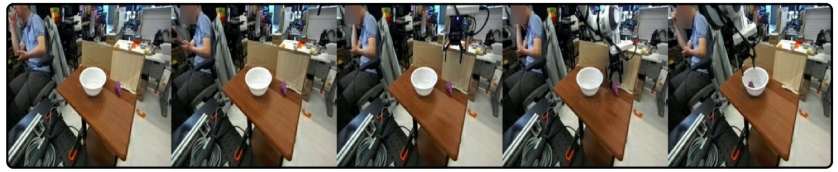


Put the rubik's cube on the top of the shelf

Qualitative Examples

- Task Understanding: infer the task from a visual trajectory
  - Decode the task instruction from an observed visual trajectory.
  - Recognize actions such as placing an object or folding a cloth.

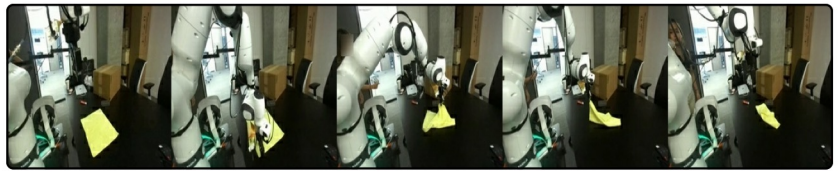
Task video frames



Task instruction prediction

Put the purple object in the bowl

Task video frames



Task instruction prediction

Fold the cloth on the table

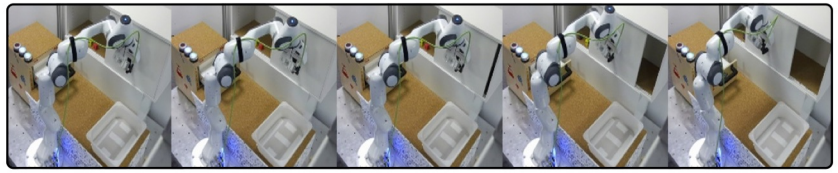
Task video frames



Task instruction prediction

Pick pepsi can from middle drawer and place on counter

Task video frames



Task instruction prediction

Grasp the door handle, slide the door to the left

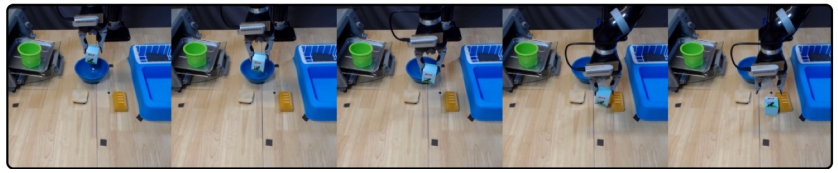
Task video frames



Task instruction prediction

Pick up the yellow cup

Task video frames



Task instruction prediction

Place the milk dairy on the table

Main Results

- LIBERO Benchmark
  - Goal-guided policy inference on the LIBERO benchmark.
  - 98.1% average success; 95.8% on long-horizon tasks

Method	Spatial	Object	Goal	Long	Average
$\pi_{0.5}$ [4]	98.8	98.2	98.0	92.4	96.9
GR00T-N1.6 [5]	97.7	98.5	97.5	94.4	97.0
OpenVLA-OFT [29]	97.6	98.4	97.9	94.5	97.1
X-VLA [6]	98.2	98.6	97.8	97.6	98.1
ABot-M0 [30]	98.8	99.8	99.0	96.6	98.6
Cosmos Policy [19]	98.1	100.0	98.2	97.6	98.5
LingBot-VA [21]	98.5	99.6	97.2	98.5	98.5
WorldVLA [31]	87.6	96.2	83.4	60.0	81.8
MMaDA-VLA [23]	98.8	99.8	98.0	95.2	98.0
Dynin-Robotics	98.9	99.8	97.8	95.8	98.1

Main Results

- LIBERO-Plus Benchmark
  - Evaluate robustness across seven perturbation categories.
  - 73.0% average success without additional LIBERO-Plus adaptation

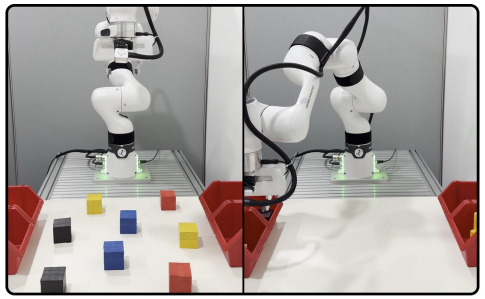
Method	Camera	Robot	Language	Light	Background	Noise	Layout	Average
OpenVLA-OFT	56.4	31.9	79.5	88.7	93.3	75.8	74.2	71.4
RIPT-VLA [33]	55.2	31.2	77.6	88.4	91.6	73.5	74.2	70.2
ABot-M0	60.4	67.9	86.4	96.2	91.6	86.4	82.6	81.6
WorldVLA	0.1	27.9	41.6	43.7	17.1	10.9	38.0	25.6
Dynin-Robotics	59.8	48.2	85.0	83.5	84.6	78.2	71.8	73.0

Main Results

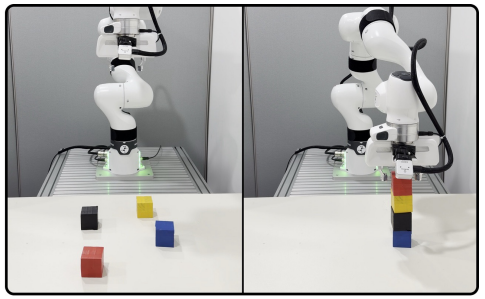
- Real-World Manipulation
  - Tasks: Pick and place the instructed fruit, sort cubes into matching color boxes, and stack four cubes.
  - Color-order stacking: Follow an instructed color sequence; Dynin-Robotics achieves the highest success rate (68.5%).



(a) Fruit PnP: Fruits pick and place tasks



(b) Cube Sort: Sort the cubes and place them in the boxes according to their colors



(c) Cube Stack: Stack four colored cubes in order (Color-Ordered Stack) or in any order (Cube Stack)

Method	Fruit PnP	Cube Sort	Cube Stack	Color-order	Average
$\pi_{0.5}$	96.5	83.5	63.5	64.0	76.9
GR00T-N1.6	91.0	78.0	59.5	63.5	73.0
Cosmos Policy	98.0	87.0	67.5	42.0	73.6
Mimic-Video [20]	96.5	84.5	64.0	39.5	71.1
Dynin-Robotics	97.5	85.5	62.0	68.5	78.4

Ablation Study

- Unified Objective Training Ablation
  - Additional objectives raise OOD success from 33.9% to 47.3%.
  - The decoder is fixed; objective exposure ratios vary.

Initialization	OXE Stage 1	ID (%)	OOD (%)	Gap (pp)
Dynin-Omni	No	3.26	0.70	2.56
Robot-pretrained	Yes	49.61	47.28	2.33
Stage-2 objectives	**Exposure ratio**	**ID (%)**	**OOD (%)**	**Gap (pp)**
Policy only	1.0	47.15	33.88	13.27
+ World Modeling	0.8 : 0.2	46.01	39.54	6.47
+ Task Understanding	0.75 : 0.2 : 0.05	49.38	45.29	4.09
+ Goal-State Prediction	0.65 : 0.2 : 0.05 : 0.10	49.61	47.28	2.33

Ablation Study

- Composable Inference Ablation
  - One checkpoint supports six inference variants.
  - Goal-guided joint denoising improves OOD success by 5.9 percentage points.

Inference variant	ID (%)	OOD (%)	Gap (pp)	TPS
(a) Action-only	45.8	41.4	4.4	9.238
(b) Joint action and next-state	46.4	40.5	5.9	8.805
(c) Goal-guided Policy	45.6	38.2	7.4	9.208
(d) World-model filtering	46.3	42.0	4.3	4.904
(e) Goal-guided joint denoising	49.6	47.3	2.3	8.709
(f) Goal-guided filtering	48.9	47.7	1.2	4.828

- Goal-guided joint denoising: 47.3% OOD success at 8.709 TPS

Acceleration: Parallel Action Decodiing

- Parallel decoding reduces model evaluations from 35 to 1.73 on average.
- Measure action-query latency; goal generation and robot I/O are excluded.

Decoder	Evaluations/chunk	Query latency (ms)	Speedup	Success (%)
Base	35.00	3,795.7	1.00×	97.8 ± 0.4
dInfer-BL7	5.14	383.6	9.89×	97.7 ± 0.2
dInfer-BL35	1.73	130.2	29.15×	97.2 ± 0.4

  - 29.15× action-query speedup
  - 130.2ms / chunk, and 97.2 ± 0.4% success

Thank you!

Hoeun Lee
hoeunlee@snu.ac.kr

AIDAS Lab