第八章 大模型具身智能

修贤超

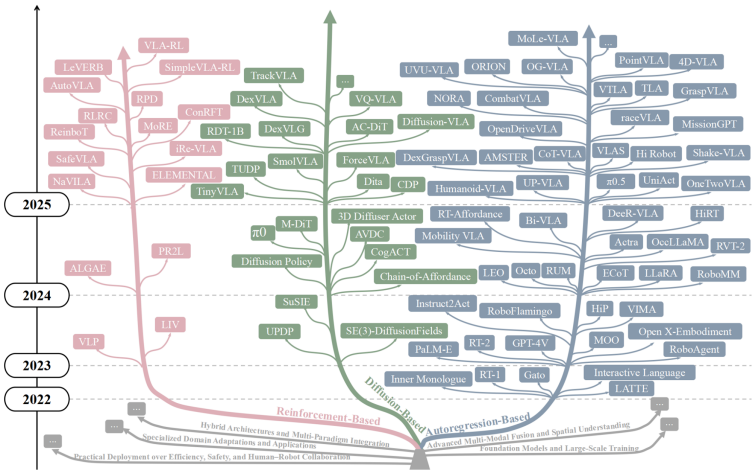
https://xianchaoxiu.github.io

致谢：本教案由陈龙协助准备

目录

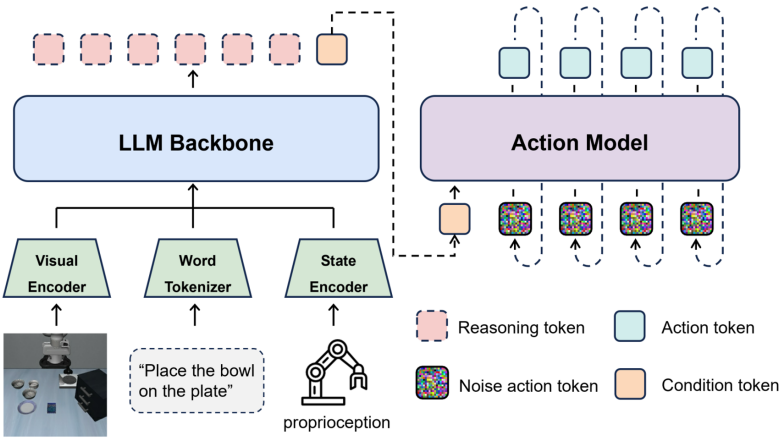
大模型具身智能的发展

- Pure Vision Language Action (VLA) Models: A Comprehensive Survey, 2025



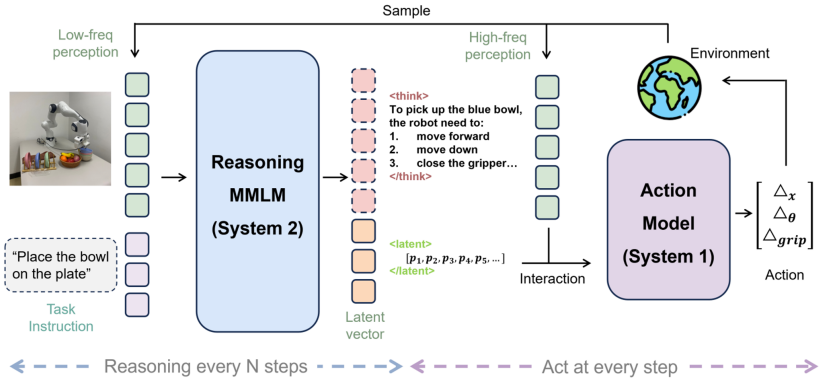
VLA 模型框架

- Efficient Vision-Language-Action Models for Embodied Manipulation: A Systematic Survey, 2025



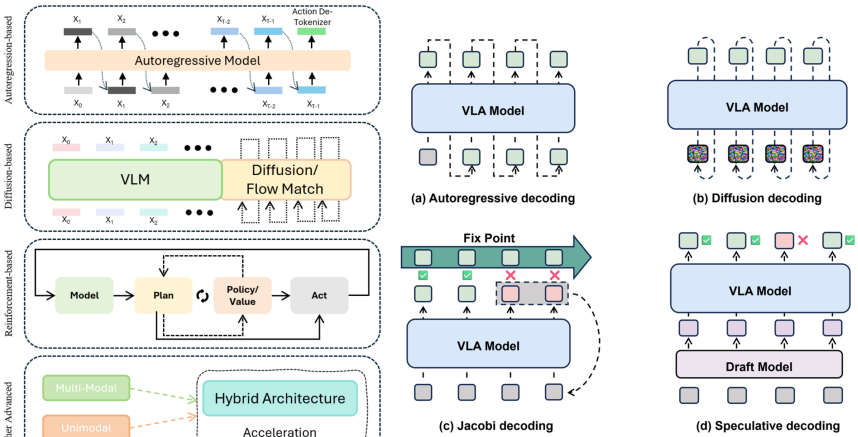
多层 VLA 架构

- Efficient Vision-Language-Action Models for Embodied Manipulation: A Systematic Survey, 2025



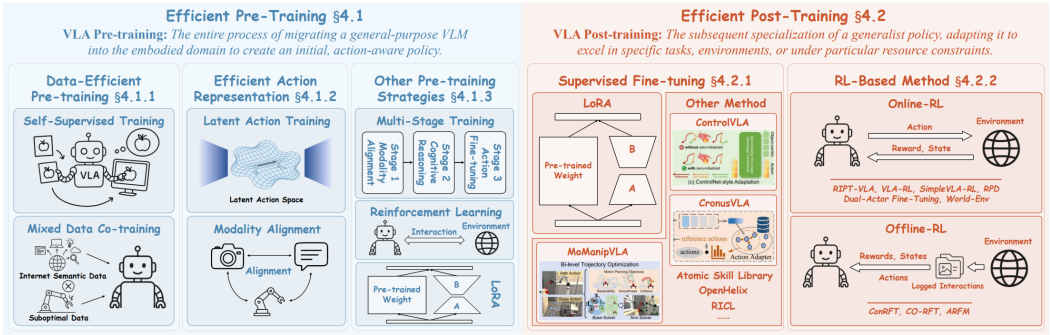
VLA 模型框架

- Pure Vision Language Action (VLA) Models: A Comprehensive Survey, 2025



VLA 预训练和后训练

- A Survey on Efffcient Vision-Language-Action Models, 2024



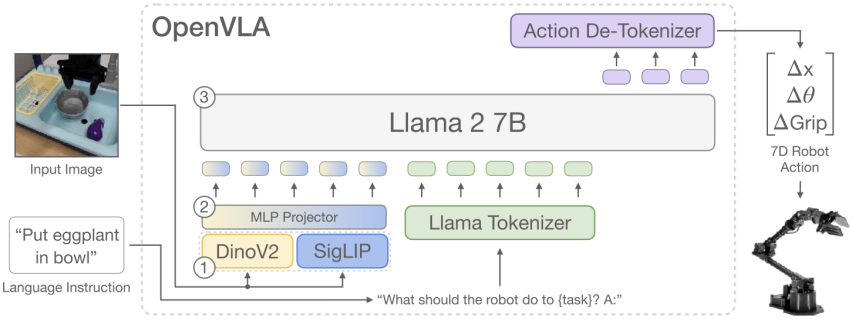
VLA 数据集

- A Survey on Efffcient Vision-Language-Action Models, 2024

目录

OpenVLA

- OpenVLA: An Open-Source Vision-Language-Action Model, 2024

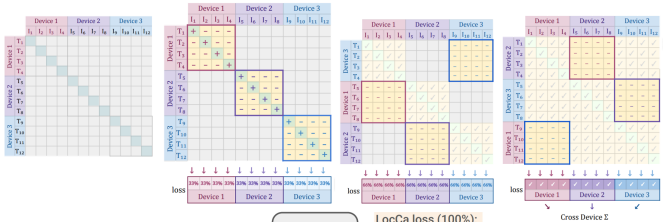


OpenVLA

- DINOv2: Learning Robust Visual Features without Supervision, 2023

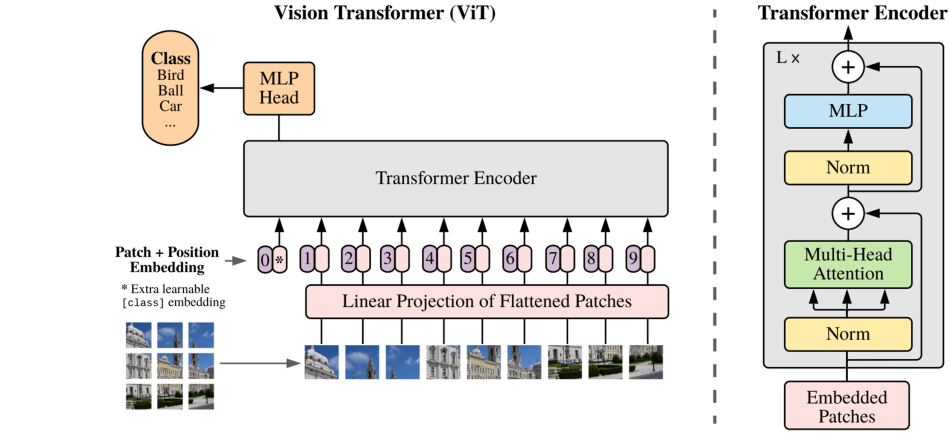
OpenVLA

- Sigmoid Loss for Language Image Pre-Training, 2023

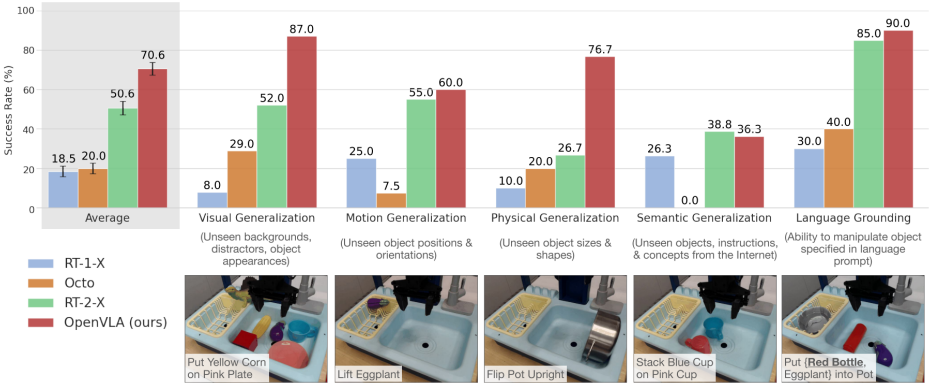


OpenVLA

- An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, 2021

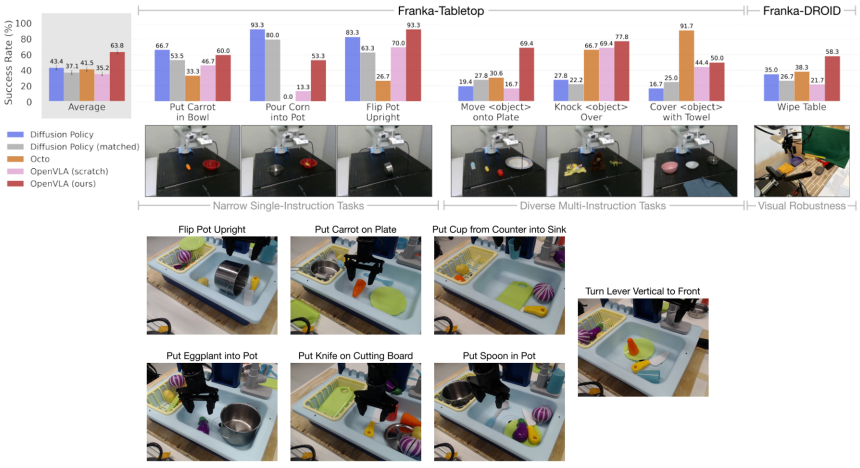


- 仿真实验


Success Rate (%)
100
80
60
40
20
0
18.5
20.0
50.6
70.6
Average
8.0
29.0
52.0
87.0
Visual Generalization
(Unseen backgrounds, distractors, object appearances)
25.0
7.5
55.0
60.0
Motion Generalization
(Unseen object positions & orientations)
10.0
20.0
26.7
76.7
Physical Generalization
(Unseen object sizes & shapes)
26.3
0.0
38.8
36.3
Semantic Generalization
(Unseen objects, instructions, & concepts from the Internet)
30.0
40.0
85.0
90.0
Language Grounding
(Ability to manipulate object specified in language prompt)
RT-1-X
Octo
RT-2-X
OpenVLA (ours)
Put Yellow Corn on Pink Plate
Lift Eggplant
Flip Pot Upright
Stack Blue Cup on Pink Cup
Put {Red Bottle, Eggplant} into Pot

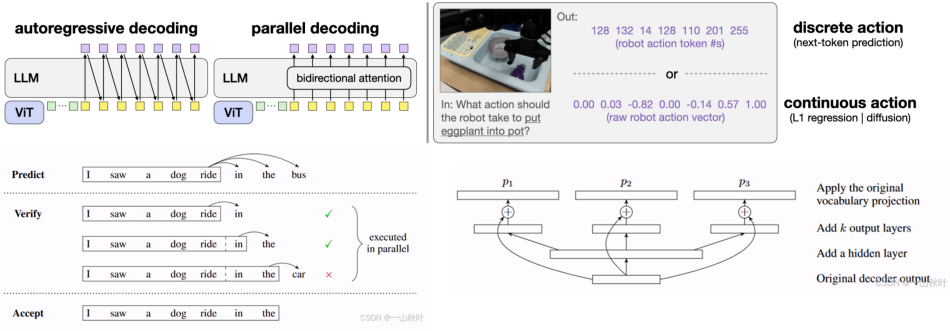
OpenVLA

- https://openvla.github.io



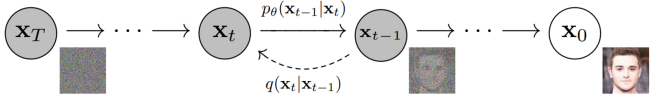
OpenVLA-OFT

- Fine-Tuning Vision-Language-Action Models: Optimizing Speed and Success, 2025



OPENVLA-OFT

- Denoising Diffusion Probabilistic Models, 2020



仿真实验结果

Policy inputs: third-person image, language instruction
(Modified training dataset; unsuccessful demonstrations filtered out)

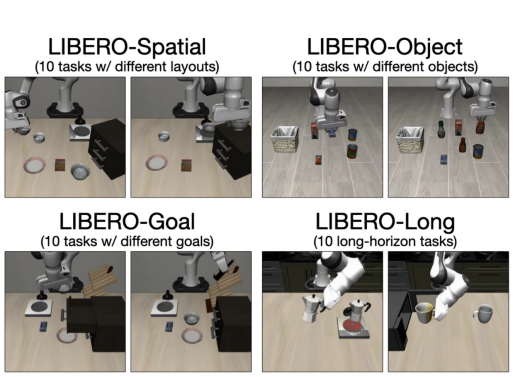
	Spatial SR (%)	Object SR (%)	Goal SR (%)	Long SR (%)	Average SR (%)
Diffusion Policy (scratch) [5]	78.3	92.5	68.3	50.5	72.4
Octo (fine-tuned) [49]	78.9	85.7	84.6	51.1	75.1
DiT Policy (fine-tuned) [13]	84.2	96.3	85.4	63.8	82.4
OpenVLA (fine-tuned) [23]	84.7	88.4	79.2	53.7	76.5
OpenVLA (fine-tuned) + PD&AC	91.3	92.7	90.5	86.5	90.2
OpenVLA (fine-tuned) + PD&AC, Cont-Diffusion	**96.9**	98.1	95.5	**91.1**	**95.4**
OpenVLA-OFT (OpenVLA (fine-tuned) + PD&AC, Cont-L1) (ours)	96.2	**98.3**	**96.2**	90.7	95.3

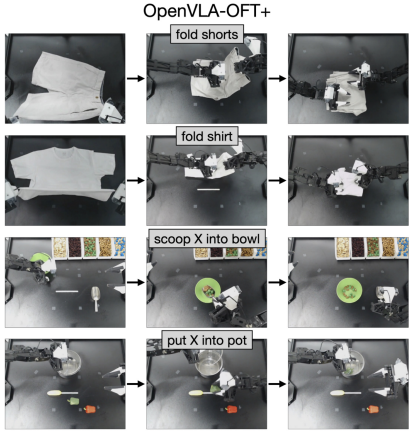
Policy inputs: third-person image, wrist camera image, robot proprioceptive state (optional), language instruction
(Modified training dataset; unsuccessful demonstrations filtered out)

	Spatial SR (%)	Object SR (%)	Goal SR (%)	Long SR (%)	Average SR (%)
π_0 + FAST (fine-tuned) [38]	96.4	96.8	88.6	60.2	85.5
π_0 (fine-tuned) [3]	96.8	**98.8**	95.8	85.2	94.2
OpenVLA-OFT (OpenVLA (fine-tuned) + PD&AC, Cont-L1) (ours)	**97.6**	98.4	**97.9**	**94.5**	**97.1**

OpenVLA-OFT

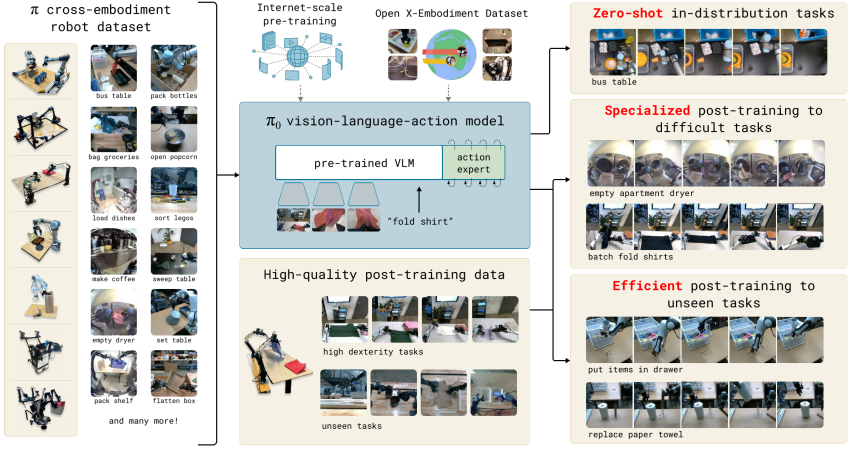
- https://openvla-oft.github.io



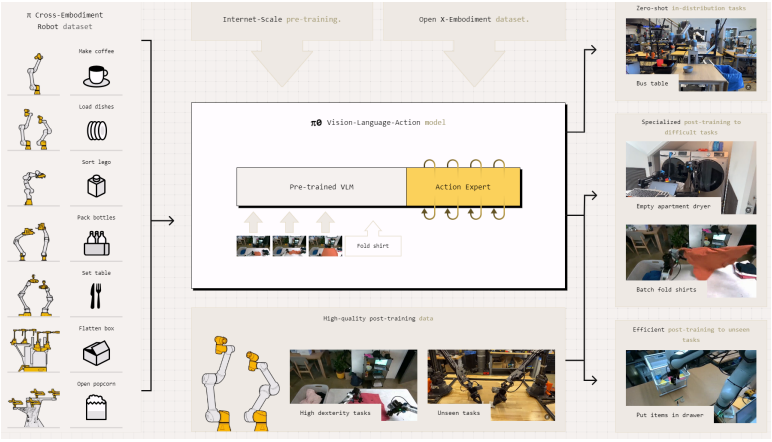


目录

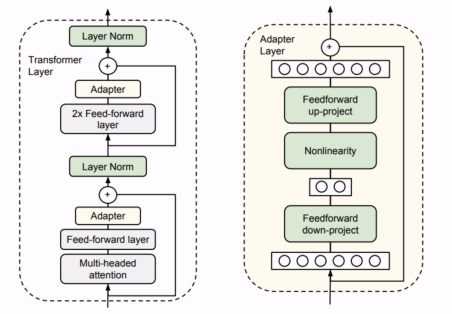
- $\pi 0$: A Vision-Language-Action Flow Model for General Robot Control, 2024

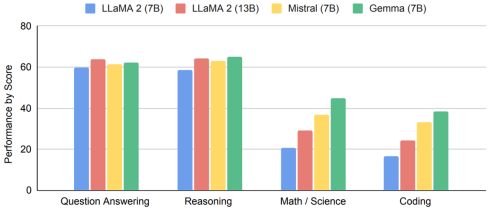


- $\pi 0$: A Vision-Language-Action Flow Model for General Robot Control, 2024

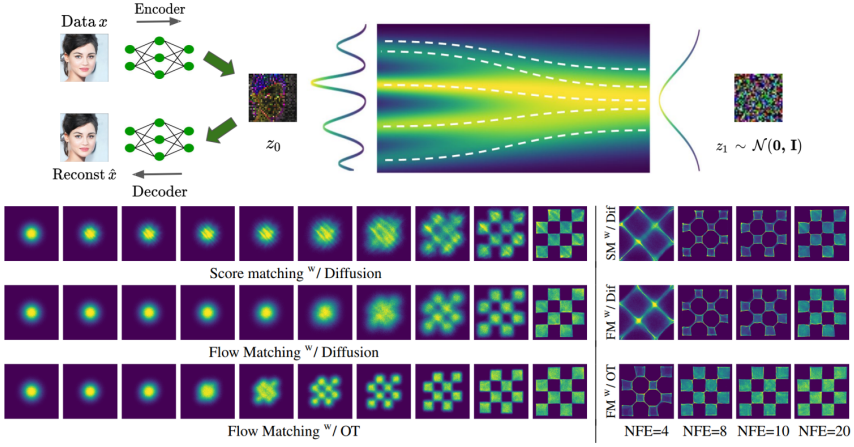


- Gemma: Open Models Based on Gemini Research and Technology, 2024



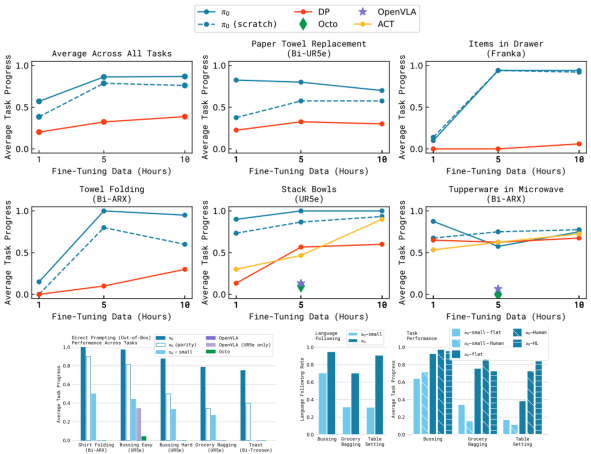


- Flow Matching for Generative Modeling, 2022

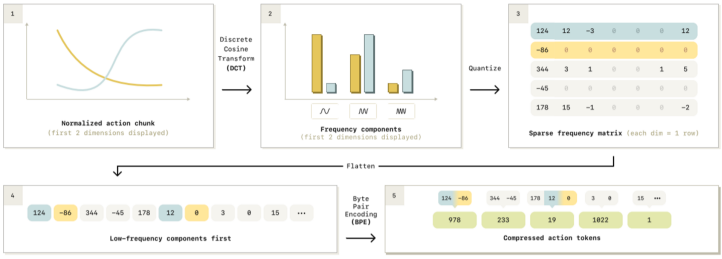


PI0

- https://www.physicalintelligence.company/blog/pi0



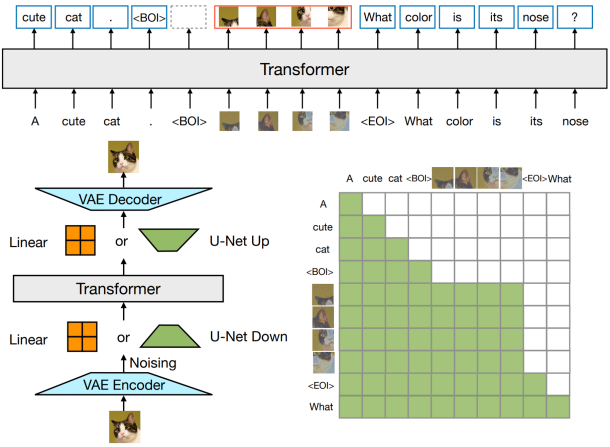
- FAST: Efficient Action Tokenization for Vision-Language-Action Models, 2025



Number	Token	Frequency
1	</w>	23
2	o	14
3	l	14
4	d	10
5	e	16
6	r	3
7	f	9
8	i	9
9	n	9
10	s	13
11	t	13
12	w	4

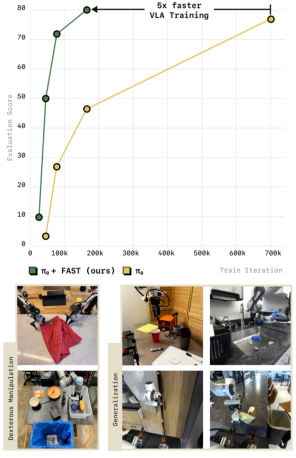
Number	Token	Frequency
1	</w>	23
2	o	14
3	l	14
4	d	10
5	e	16 - 13 = 3
6	r	3
7	f	9
8	i	9
9	n	9
10	s	13 - 13 = 0
11	t	13 - 13 = 0
12	w	4
13	es	9 + 4 = 13 - 13 = 0
14	est	13 - 13 = 0

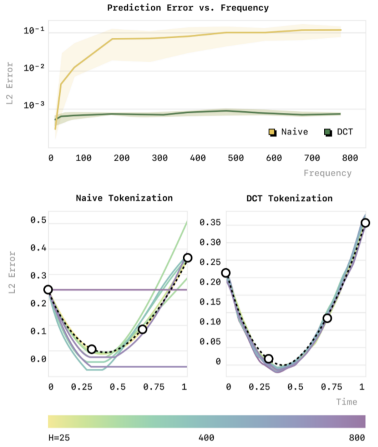
- Transfusion: Predict the Next Token and Diffuse Images with One Multi-Modal Model, 2024



PI0FAST

- https://www.pi.website/research/fast

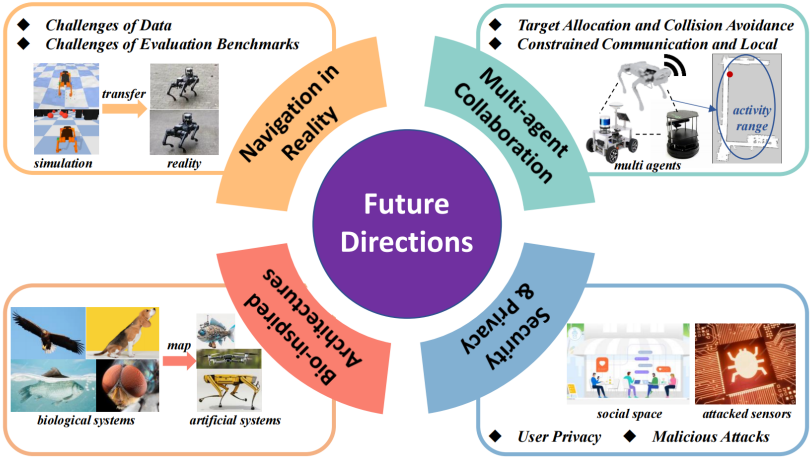




目录

小结

- Embodied Navigation, 2025



小结

- OpenVLA: An Open-Source Vision-Language-Action Model, 2024
- $\pi 0$: A Vision-Language-Action Flow Model for General Robot Control, 2024
- A survey on Vision-Language-Action Models for Embodied AI, 2024
- Pure Vision Language Action (VLA) Models: A Comprehensive Survey, 2025
- Efficient Vision-Language-Action Models for Embodied Manipulation: A Systematic Survey, 2025
- Survey of Vision-Language-Action Models for Embodied Manipulation, 2025
- Vision-Language-Action Models for Robotics: A Review Towards Real-World Applications, 2025

Q&A

Thank you!

感谢您的聆听和反馈