西北工業大學
NORTHWESTERN POLYTECHNICAL UNIVERSITY

Understanding

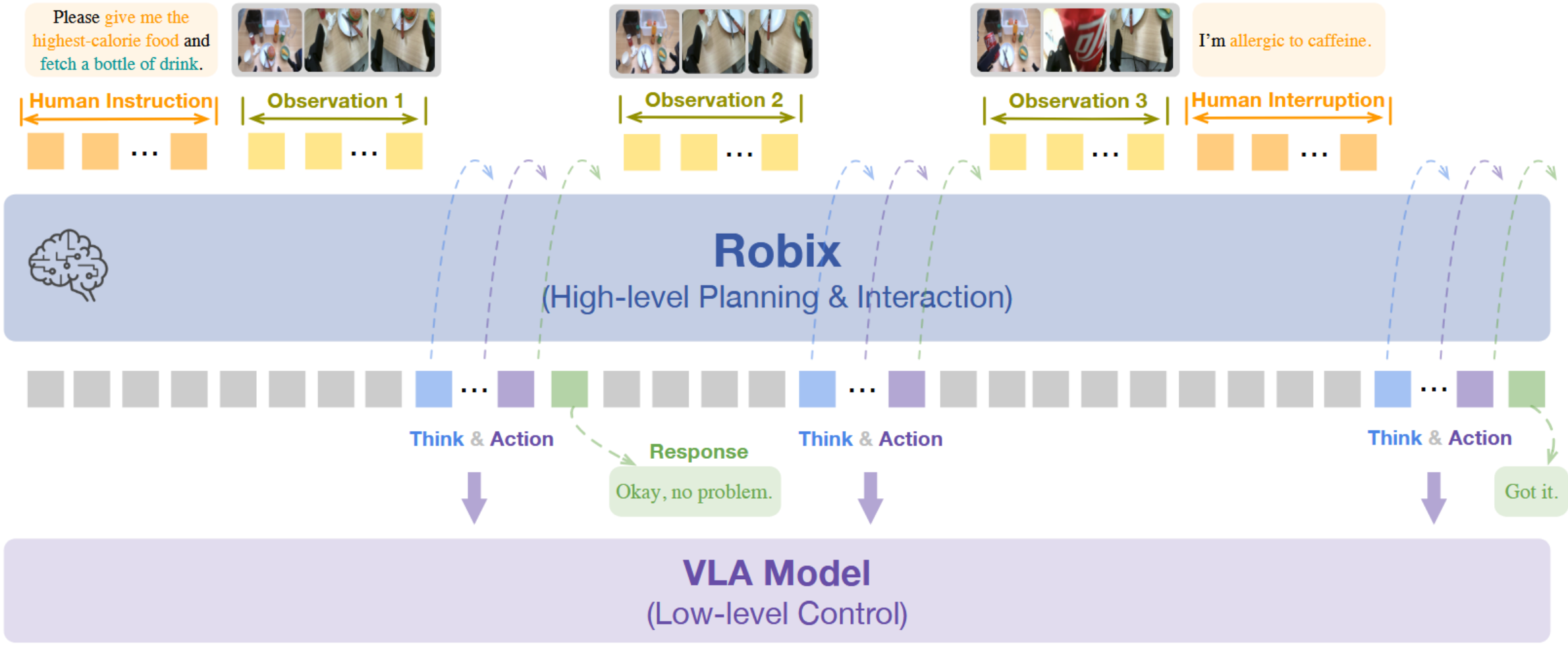
Prediction

Execution

Zhengwei Yang

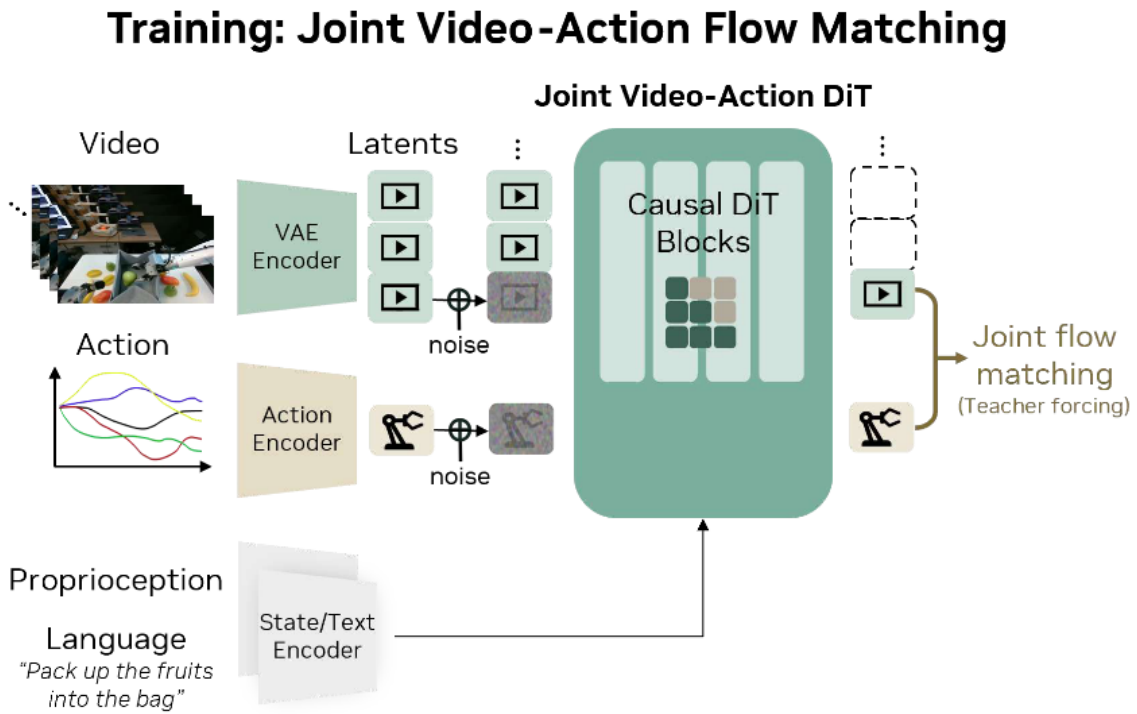
2026.05.31

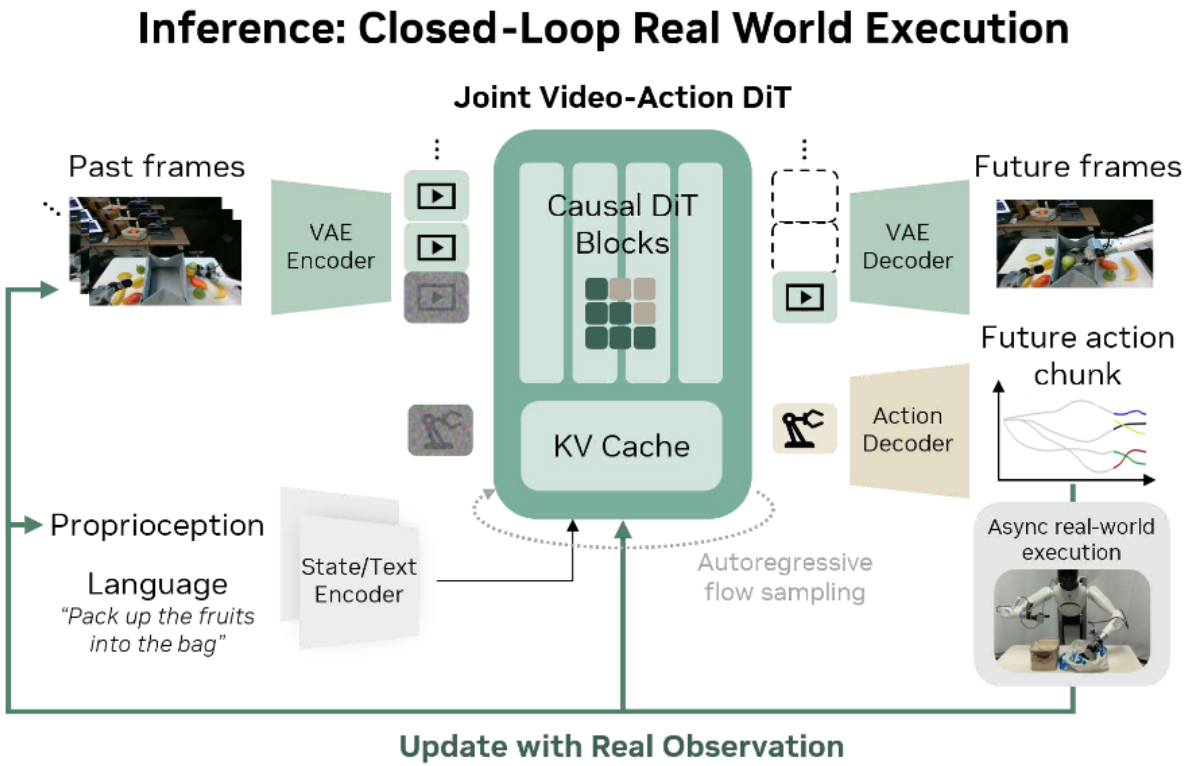
Robix: A Unified Model for Robot Interaction, Reasoning and Planning



- CoT Reasoning
- Three-stage training strategy: Continued pretraining → SFT → RL

World Action Models are Zero-shot Policies

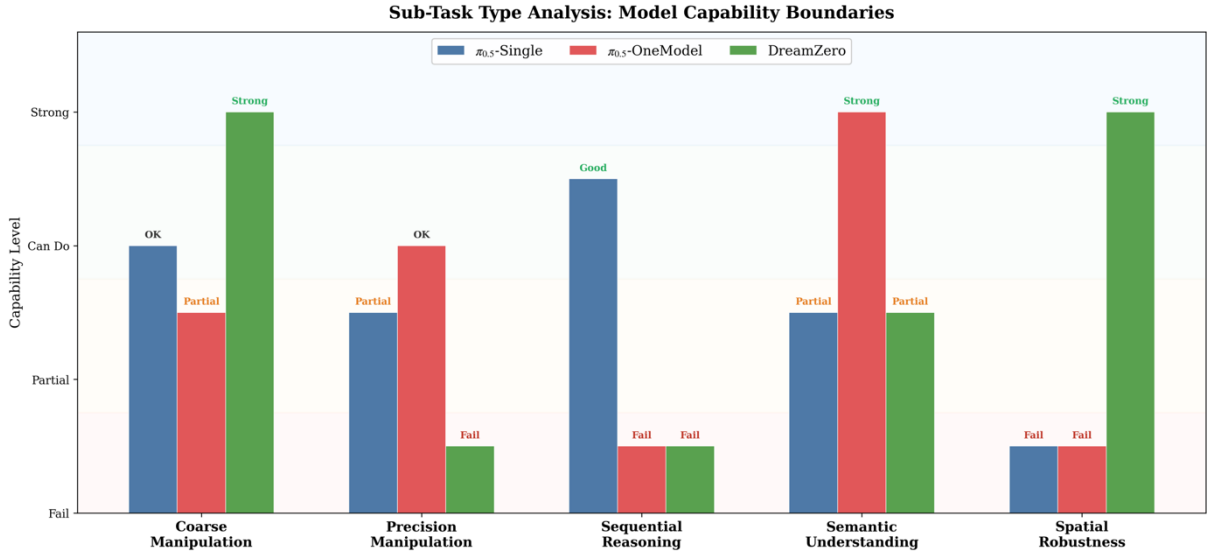




$$\underbrace{\pi_0(\mathbf{o}_{l:l+H}, \mathbf{a}_{l:l+H} \mid \mathbf{o}_{0:l}, \mathbf{c}, \mathbf{q}_l)}_{\textsc{DreamZero}} = \underbrace{\pi_0(\mathbf{o}_{l:l+H} \mid \mathbf{o}_{0:l}, \mathbf{c}, \mathbf{q}_l)}_{\text{video prediction}} \underbrace{\pi_0(\mathbf{a}_{l:l+H} \mid \mathbf{o}_{0:l+H}, \mathbf{q}_l)}_{\text{IDM}}$$

Motivation

Category	Task	$\pi_{0.5}$-Single Score	$\pi_{0.5}$-Single SR	$\pi_{0.5}$-OneModel Score	$\pi_{0.5}$-OneModel SR	DreamZero Score	DreamZero SR
Execution	arrange_cup	18	0%	14	0%	**25**	0%
	put_spoon	52	20%	43	10%	**54**	20%
	put_glasses	80	60%	**87**	70%	37	0%
	put_ring	41	20%	**60**	20%	27	0%
	put_items_drawer	**34**	0%	7	0%	7	0%
	pick_items_basket	46.8	20%	37	10%	**97.8**	90%
	pour_water	8	0%	**19**	0%	12	0%
	insert_wireline	**24**	0%	21	0%	**24**	0%
	put_stationery	21.5	0%	16	0%	**26**	0%
	put_blocks	**48**	30%	42	30%	27	0%
Semantic	sort_headphone	35	20%	**73**	70%	66	60%
	classify_items	59	20%	**63**	40%	36	0%
	press_button	**48**	20%	13	0%	3	0%
	pair_up_items	21	0%	**51.5**	30%	13.5	0%
	pick_fruits	90	80%	**94**	80%	45	0%
Mobile Manipulation (5 tasks): evaluation in progress, to be reported in subsequent update.							
Tabletop Total (/ 1,500)		626.3		**640.5**		500.3	



CVPR 2026 ManipArena Challenge Results

Pros (+)
Cons (-)

Linguistic Planning → LLM-based

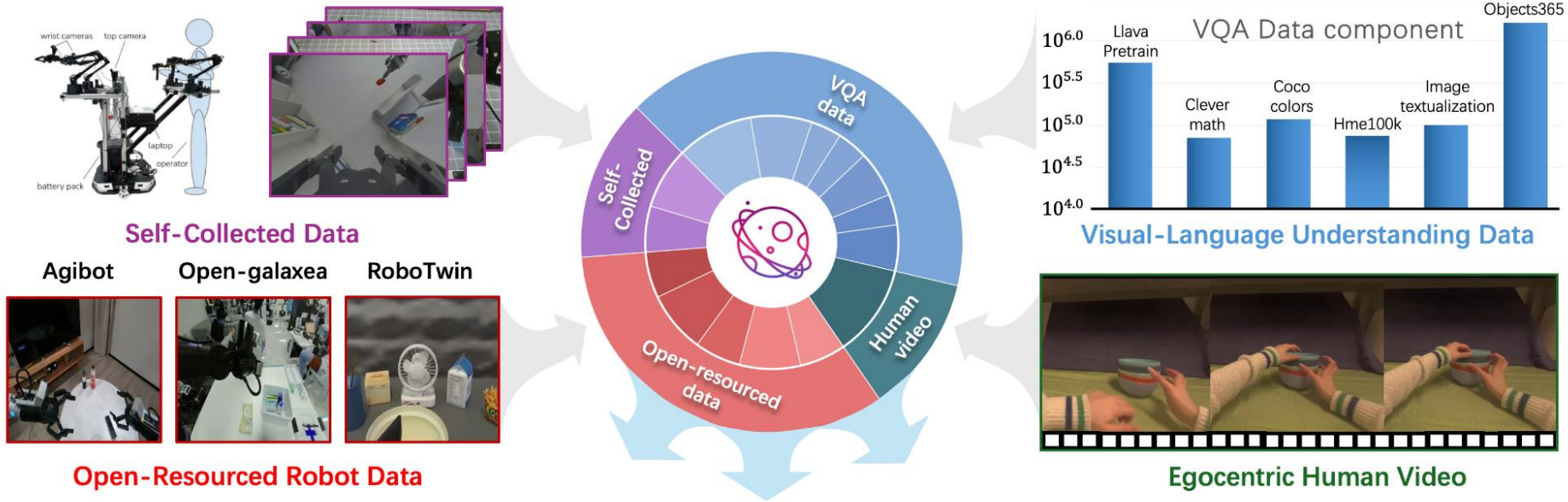
- Strong logical reasoning
- Rich common-sense knowledge
- Strong cross-task transfer capability
- Minimal computational overhead and low latency
- Lack of physical and spatial-geometric awareness
- Poor spatial generalization ability

Visual Forecasting → Video Prediction Policy

- Strong physical consistency and spatial robustness
- Excel at coarse/basic tasks and manipulation
- Latent ability of causal reasoning
- Action-state alignment
- Failed to complete fine-grained tasks
- High latency and huge computational overhead

ByteDance Seed

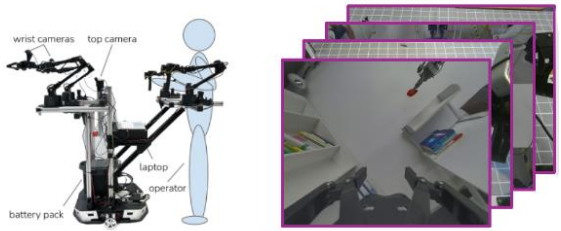
BagelVLA: Enhancing Long-Horizon Manipulation via Interleaved Vision-Language-Action Generation



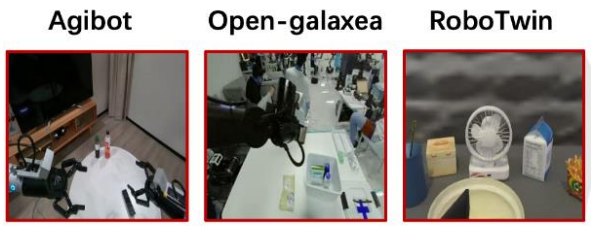
Overview of BagelVLA

Dataset

Large-scale Robotic Datasets

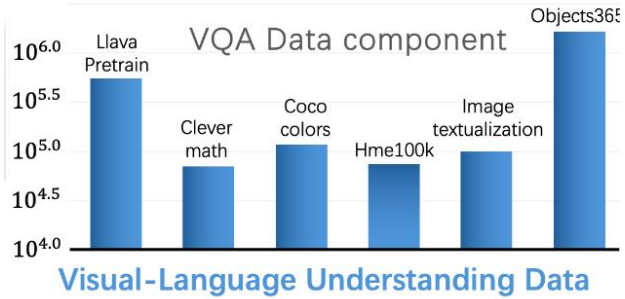


Self-Collected Data



Open-Resourced Robot Data

General Multimodal Data



Visual-Language Understanding Data



Egocentric Human Video

Conventional VLA

Action chunks

$$\mathcal{D} = \{L_i, \tau_i\}_{i=1}^{N} \quad \tau_i = \{(v_1, l_1, a_1), \dots, (v_T, l_T, a_T)\} \Rightarrow p_\theta(a_t | v_t, L)$$

Global instruction

Observation

Stage-specific descriptions

BagelVLA

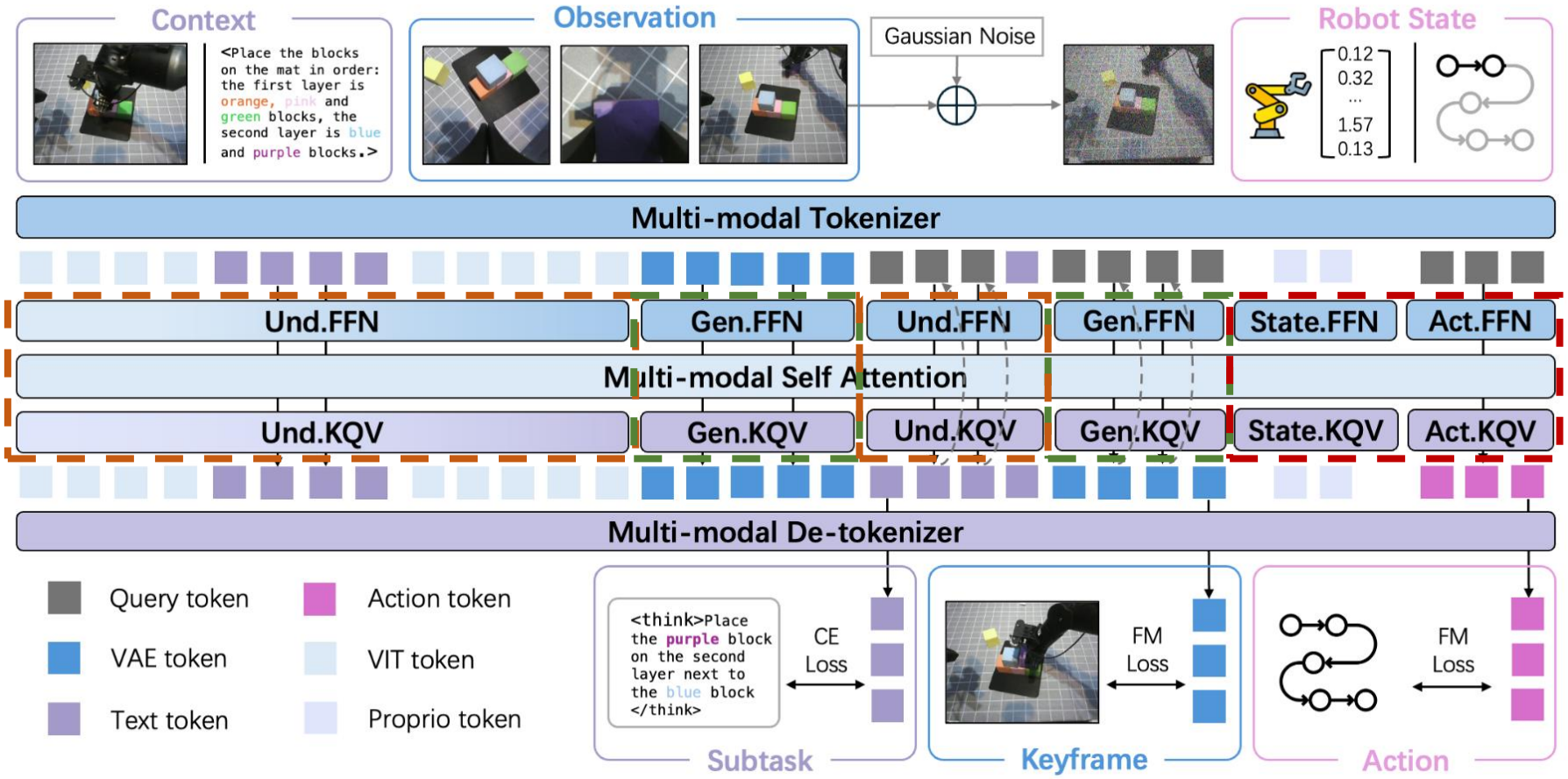
$$p_\theta(a_t, v_{t+k}, l_t | v_t, L) = p_\theta(l_t | v_t, L) \cdot p_\theta\,(v_{t+k} | v_t, L, l_t) \cdot p_\theta\,(a_t | v_t, L, l_t, v_{t+k})$$

Linguistic Planning

Visual Forecasting

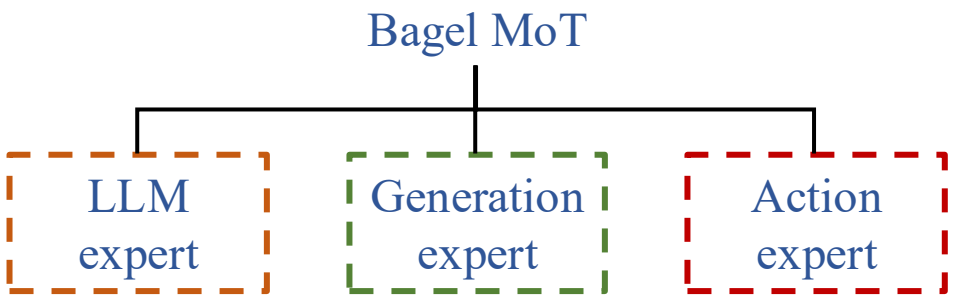
Action Generation

Architecture



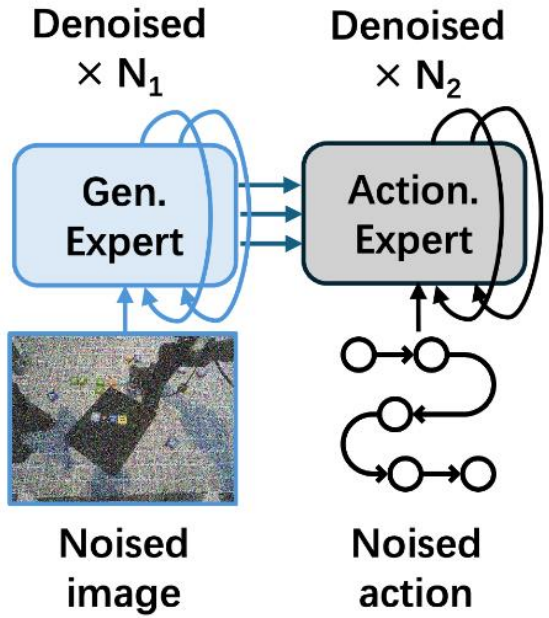
Modules	Understanding Expert	Generation Expert	Action Expert
Size	7B	7B	2B
Input Modality	Image/Text	Image	Proprio/Action
Output Modality	Text	Image	Action
Encoder	ViT+MLP	VAE+MLP	MLP
Image Resolution	256x256	256x256 (VAE)	-
Hidden size	3584	3584	3584
Intermediate size	18944	18944	3584
Layers	28	28	28
Loss Type	CE	MSE(FM)	MSE(FM)
FM Timestep Distribution	-	LogitNormal(0, 1)	Beta(1.5,1)

* Designed for higher execution frequency



Dual Flow-Matching

Pure noise image?
Noised image?



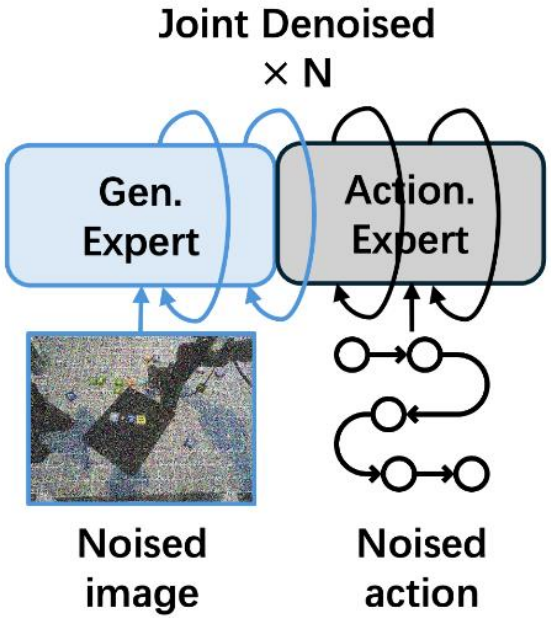
a. Complete Denoise

$$v_{t+k}^{\tau=0} \rightarrow v_{t+k}^{\tau=1} = v_{t+k}$$

$$a_t \mid v_{t+k}$$

~ DreamZero
WM + IDM

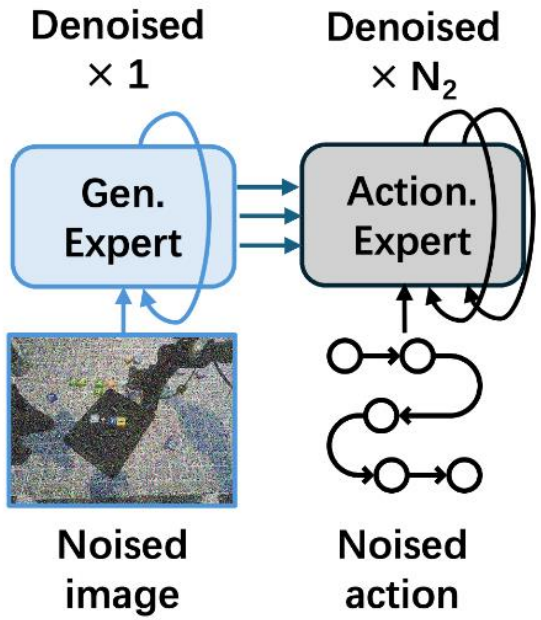
- Based on "a realistic future"
- Inference latency: N1 + N2 steps
- Accumulation of visual errors



b. Joint Denoise

$$a_t^{\tau} \mid v_{t+k}^{\tau}$$

- Faster inference: N steps
- Spurious correlation



c. Single-Step Denoise

$$a_t^{\tau} \mid v_{t+k}^{\tau=0} \quad \times \quad v_{t+k}^{\tau=0} \sim \mathcal{N}(0, I)$$

$$\text{RFG} \quad \checkmark \quad v_{t+k}^{\tau=0} \sim \mathcal{N}(v_t, I)$$

- Fastest inference
- focus on modeling robot manipulation changes

Training

Pretraining - Finetuning
Linguistic Planning and Learning Visual Dynamics

- Understanding Expert
- Generation Expert
- Action Expert

Acquire capabilities in
textual planning and keyframe prediction

Finetuning - Learning Action Planning

- Understanding Expert
- Generation Expert
- Action Expert

Perform robustly in specific scenarios

Trick: Randomly replace the current frame with a preceding image

Inference

48Hz

Asynchronous Execution
Reduce the updating frequency of the KV contexts of
understanding and generation experts

72Hz

At each denoising step, only a single expert is activated

Single-step denoise scheme

Results

Average Completion Length Success Rate

Model	Calvin	Robotwin	
	ABC-D	Clean	Randomized
π_0	3.648	46.42	16.34
RDT	-	34.50	13.72
UP-VLA	4.078	52.92	15.16
VPP	4.329	-	-
w/o Textual-planning	-	54.00	19.20
w/o Keyframe-forecasting	3.345	56.72	15.92
BagelVLA	**4.405**	**75.26**	**20.87**

Simulation

Basic Task

Model	Pick&Place Seen	Pick&Place Unseen	Water Flower	Stack Cubes	Put Flowers in Vase	Stack Bowls	Pour Fries	Sweep Rubbish	Press Button	Drawer Close	Success Average
π_0	**95**	55	50	65	40	70	35	55	**90**	95	65.0
VPP	85	45	**60**	50	**50**	55	30	45	75	100	59.5
BagelVLA	**95**	**85**	**60**	**80**	35	**90**	**45**	**80**	**90**	95	**75.5**

Long-Horizon Task

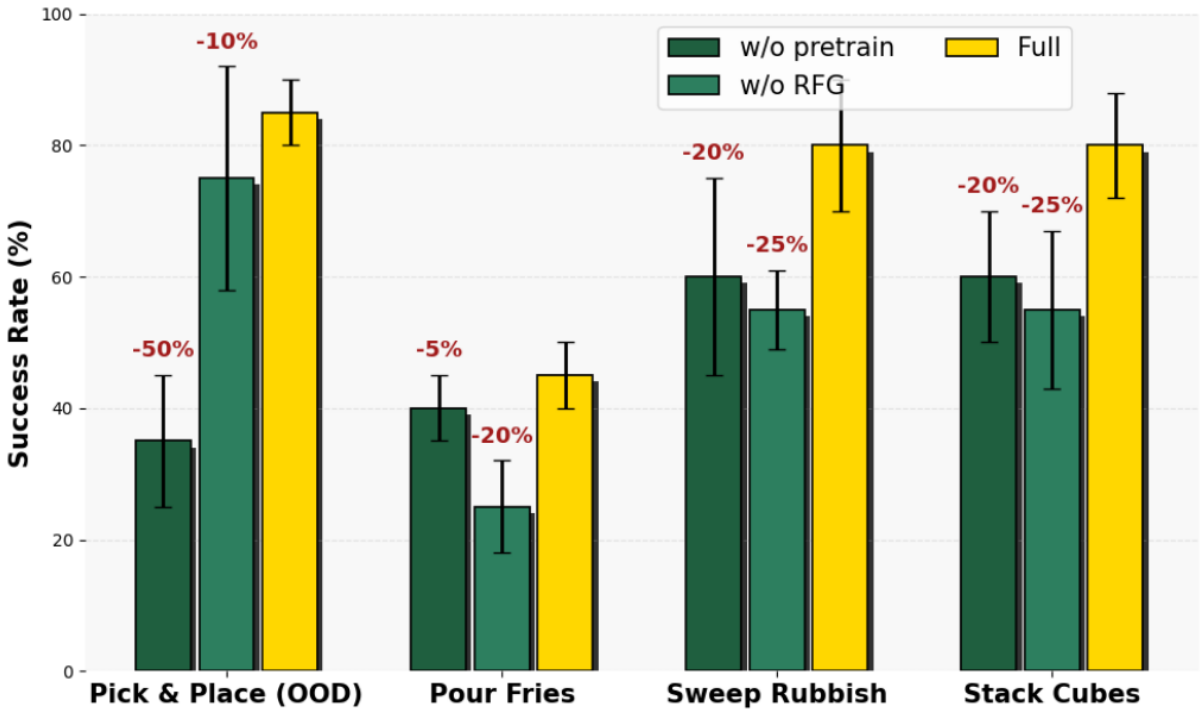
Tasks	Stack Cubes in Requested Order					Calculate and Place Symbol Blocks				
Difficulty	Easy	Middle	Hard	Success Rate↑	Planning Accuracy↑	Easy	Middle	Hard	Success Rate↑	Planning Accuracy↑
π_0	75	35	10	40.0	55	70	25	0	31.7	40
VPP	60	15	0	25.0	45	60	10	0	23.3	30
w/o Keyframe-forecasting	90	45	25	53.3	80	70	50	30	50.0	75
w/o Textual-planning	75	40	15	43.3	70	65	25	10	33.3	50
BagelVLA	**95**	**65**	**60**	**73.3**	**95**	**80**	**65**	**45**	**63.3**	**85**

Real-world

Ablation Study

Dual Flow-Matching Schemes	Latency↓	ABC-D↑
Complete Denoise	6.04s	2.480
Joint Denoise	2.90s	2.038
Single-step Denoise (Eq. 2)	**1.23s**	3.345
RFG (Eq. 3)	**1.23s**	**3.600**

?



Reflections
&
Ideas

RFG for Action Generation + Physical Constraints

Attention-Guided Sampling in Gaussian Latent Manifold

Thanks for listening!

Zhengwei Yang

2026.05.31