Action Generation with World Models

Kyungmin Lee

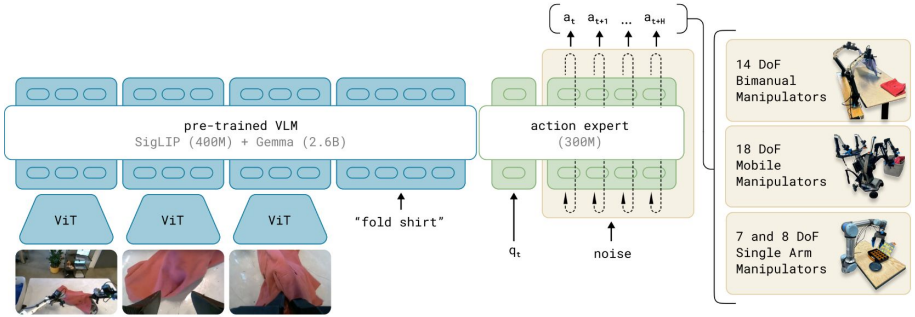
Ph.D. student @ KAIST

Outline of today's talk

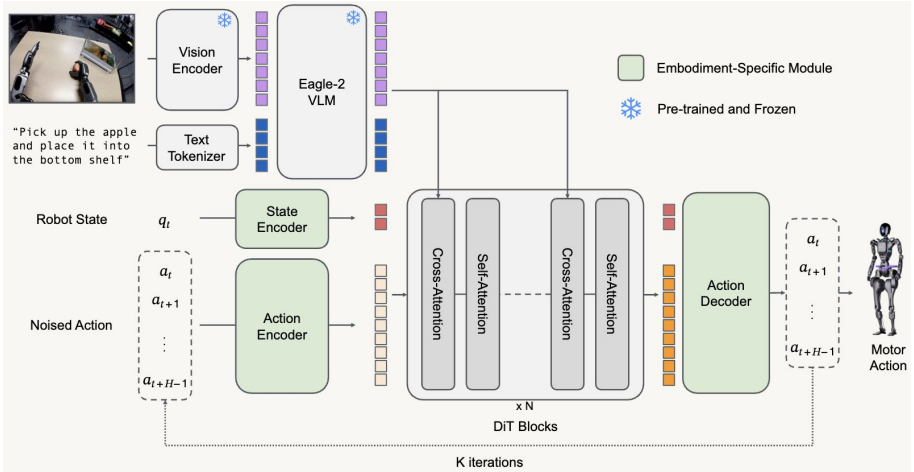
1. Background: VLA and World Models
2. (Main) DUST: Dual Stream Diffusion Model for World Model + VLA
3. (Brief) DreamZero: World Action Models are zero-shot policies
4. Future Direction

Vision-Language-Action (VLA) Model

- Vision-Language model (VLM) + action expert (e.g., flow policy)
- Trained with massive expert trajectories (behavior cloning)
- Good at semantic and object generalization (good at understanding)
- Bad at generalization to skills or environment



PI-0.5 (Physical Intelligence)



GR00T-N1.5 (NVIDIA GEAR)

World Models

- Definition: *A model that predicts next state given past observation and action*
- Recently, video world models have gained attention as
  - Video diffusion models (e.g., Veo, Sora) are able to generate high fidelity real-world
  - Post-trained as action-conditioned video generator (e.g., Genie 3)
  - World model allows to simulate environment to learn policy (e.g., Dreamer v4)

Genie 3

Dreamer 4

World Models for Action Generation

- Consider world model as interim planner
  - Predict the future state through world model
  - Then generate action through inverse dynamics model or optimization (e.g., DINO-WM)
- However, this requires training separate world model and prediction model or require costly optimization at inference time

World Action Models (WAM)

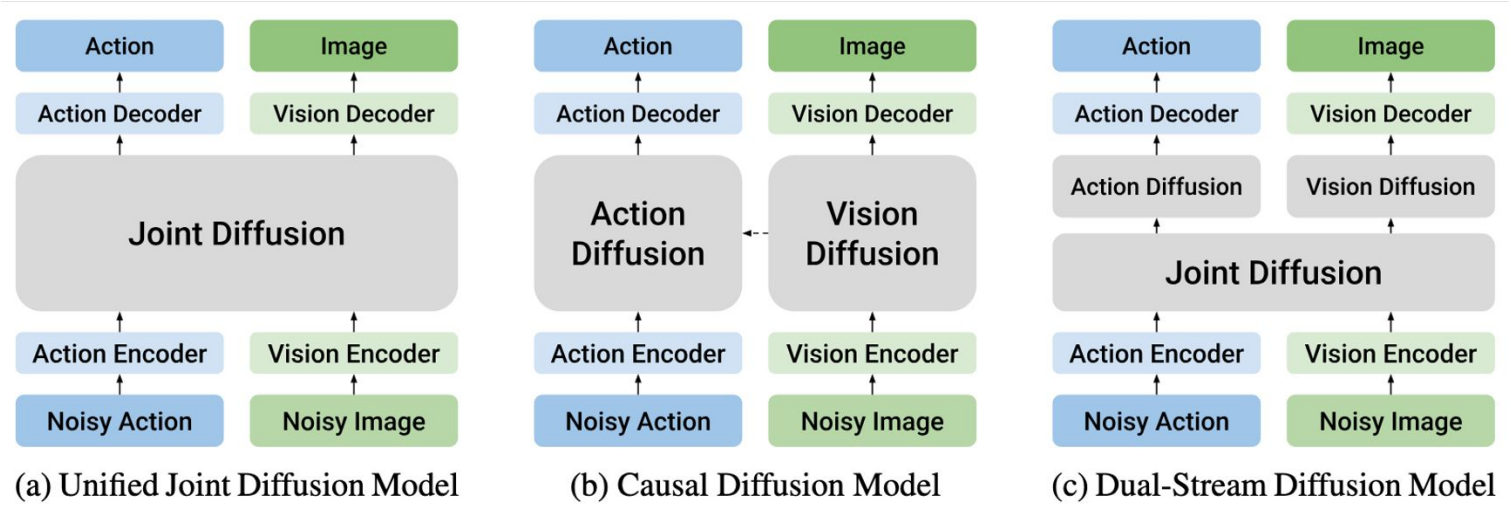
- World Action Model: Joint training of world model and action generation model
  - Joint action - future state prediction within single model
  - A model that learns both transition dynamics and action representation

- Two research papers on WAM
  - Augmenting VLAs with world model through latent world model
  - Video diffusion models as backbone for world action model

DUST: DUal STream Diffusion Model for WAM

Joint Action-Video models

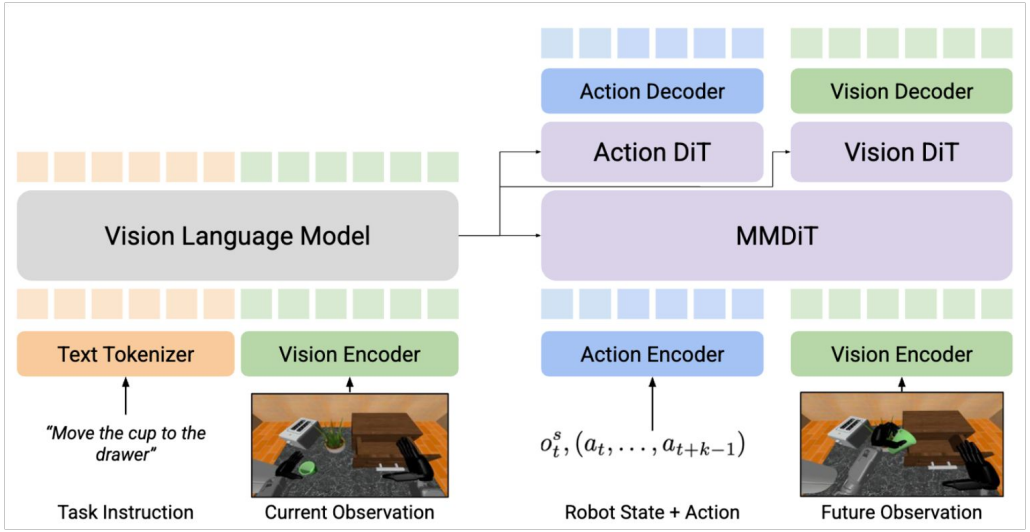
- Unified Joint diffusion model: requires unified latent space, ignores causal relationship
- Causal model: does not allow bidirectional information transfer
- DUST: allows to learn bidirectional transfer, decoupled modality learning



DUST: DUal STream Diffusion Model for WAM

Three technical contributions of DUST

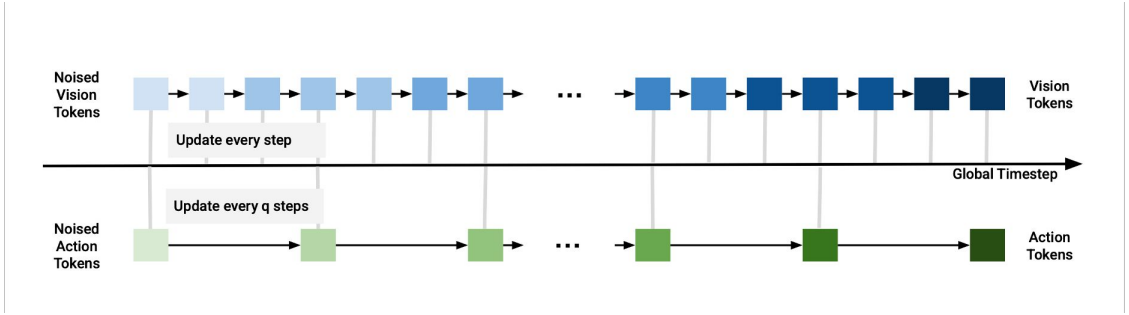
- Multi-modal Diffusion Transformer (MMDiT) architecture for action expert
  - Dual stream expert + Single stream expert for action and world model latents
  - Use SigLIP-2 feature as world model representation



DUST: DUal STream Diffusion Model for WAM

Three technical contributions of DUST

- Decoupling the modality learning
  - Perform diffusion forcing style training for action generation and world modeling
  - Learn diverse distributions of action and latents - allow to generate action with noisy latent
- Asynchronous sampling
  - Action generation requires fewer steps than visual generation (e.g., 4 vs 25)
  - Update vision tokens more frequently than action tokens



DUST: DUal STream Diffusion Model for WAM

- DUST significantly improves performance than VLAs (GR00T, pi-0)
  - On RoboCasa, GR-1, and real-world robotics manipulation task
- Outperforms FLARE (implicit world modeling through alignment loss)

Table 1. **Evaluation on RoboCasa.** Success rates (%) on RoboCasa for 8 pick-and-place (PnP), 6 contraption open/close (OP/CL), and 10 other miscellaneous tasks. 100, 300, and 1,000 demos per task are used for training. †: reproduced results.

	100 Demos				300 Demos				1,000 Demos			
Method	PnP	OP/CL	Other	Avg.	PnP	OP/CL	Other	Avg.	PnP	OP/CL	Other	Avg.
PAD	11.3	42.9	34.6	26.7	14.3	53.8	42.4	33.2	-	-	-	-
VPP	16.0	62.9	46.2	37.1	21.3	72.5	53.2	43.7	-	-	-	-
π_0-FAST	0.7	13.7	22.6	13.1	1.5	48.7	31.0	25.6	12.0	35.7	36.6	28.2
π_0	16.8	67.7	49.2	43.0	15.0	73.3	49.4	43.9	13.0	69.7	58.0	45.9
GR00T-N1.5	21.5	60.3	46.8	41.7	27.2	66.0	46.6	45.0	32.3	75.7	50.8	50.8
+ FLARE†	23.0	64.8	49.8	44.6	38.0	76.7	56.2	55.3	45.9	83.7	68.2	64.6
+ DUST	**29.5**	**76.0**	**51.0**	**50.1**	**42.3**	**80.7**	**58.1**	**58.5**	**48.3**	**86.3**	**68.6**	**66.3**

Table 2. **Evaluation on GR-1.** Success rates (%) on GR-1 for 18 pick-and-place (PnP) and 6 articulated (Art.) tasks. 300 and 1,000 demos per task used. †: reproduced results.

	300 Demos			1,000 Demos		
Method	PnP	Art.	Avg.	PnP	Art.	Avg.
PAD	11.2	15.0	12.2	-	-	-
VPP	20.8	18.7	20.2	-	-	-
π_0-FAST	16.8	29.4	20.0	19.2	30.8	22.1
π_0	19.4	32.3	22.7	20.8	34.3	24.2
GR00T-N1.5	17.6	28.3	20.3	30.7	31.0	30.8
+ FLARE†	34.0	33.0	33.7	39.3	32.4	36.3
+ DUST	**35.8**	**36.7**	**36.0**	**42.2**	**41.3**	**42.0**

Table 3. **Evaluation on real-world tasks.** Success rates (%) of 4 pick-and-place (PnP) tasks, 1 insertion task, and 2 tool-using tasks for real-world Franka Research 3 robot experiments. †: reproduced results.

	Pick-and-Place				Insert	Tool-Use		
Method	PnP-1	PnP-2	PnP-3	PnP-4	Cord	Eraser	Brush	Avg.
π_0	50.0	64.6	45.8	33.3	8.3	37.5	41.7	40.2
GR00T-N1.5	58.3	75.0	50.0	35.4	12.5	50.0	44.4	46.5
+ FLARE†	62.5	72.9	50.0	37.5	20.8	54.2	48.6	49.5
+ DUST	**83.3**	**79.2**	**62.5**	**45.8**	**29.2**	**56.3**	**65.3**	**59.9**

DUST: DUal STream Diffusion Model for WAM

- DUST favors more when jointly trained with human data
  - Best gain when using Human (EgoDex)- Robot (RoboCasa, GR-1) data mixture
- Video pre-training with DUST is effective
  - Using Bridge v2 video pretraining improves performance a lot

Table 4. **Evaluation of joint training.** Success rates (%) on RoboCasa when joint training with or without GR-1 and EgoDex data mixture. For RoboCasa and GR-1, 300 demonstrations for each task is used, while we utilize a Fourier-hand action retargeted subset of EgoDex with 46k episodes. †: reproduced results.

Method	Data Mixture	PnP	OP/CL	Other	Avg.
GR00T-N1.5	✗	27.2	66.0	46.6	45.0
	✓	30.2	69.6	46.6	46.9
+ FLARE†	✗	38.0	76.7	56.2	55.3
	✓	41.1	76.9	59.2	57.6
+ DUST	✗	42.3	80.7	58.1	58.5
	✓	**50.2**	**82.1**	**65.2**	**64.4**

Table 5. **Evaluation of pretraining setup.** Success rates (%) on RoboCasa benchmark comparing performance with or without video data pretraining. BridgeV2 data with only video input is used for the pretraining stage, while 100 RoboCasa demos per task are used for the finetuning stage. †: reproduced results.

Method	Video Pretrain	PnP	OP/CL	Other	Avg.
GR00T-N1.5	✗	21.5	60.3	46.8	41.7
+ FLARE†	✗	23.0	64.8	49.8	44.6
	✓	33.4	77.7	58.8	55.1
+ DUST	✗	29.5	76.0	51.0	50.1
	✓	**42.3**	**80.7**	**58.1**	**58.5**

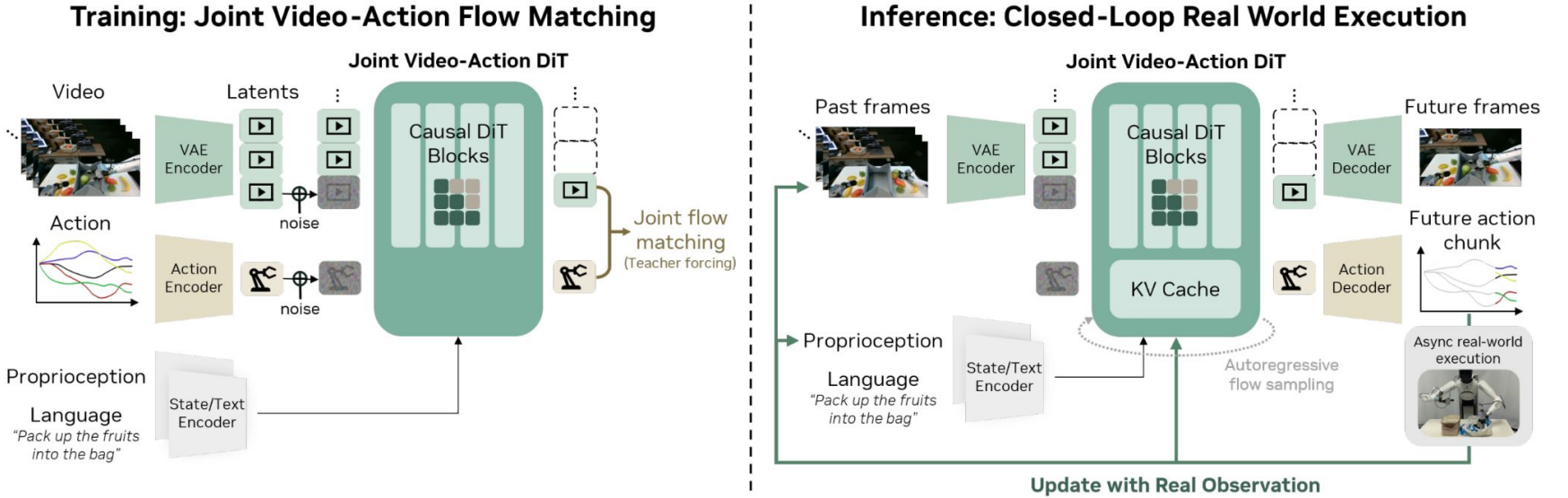
DUST: DUal STream Diffusion Model for WAM

Summary

- We present technical methods to incorporate world modeling to VLA
  - Multi-modal action expert
  - Decoupled diffusion training for joint action-video generation
- We show that augmenting world model to VLAs enhances action generation performance
  - Simulated benchmarks e.g., RoboCasa, GR-1
  - Real-world tasks
  - Video pretraining and human data co-training

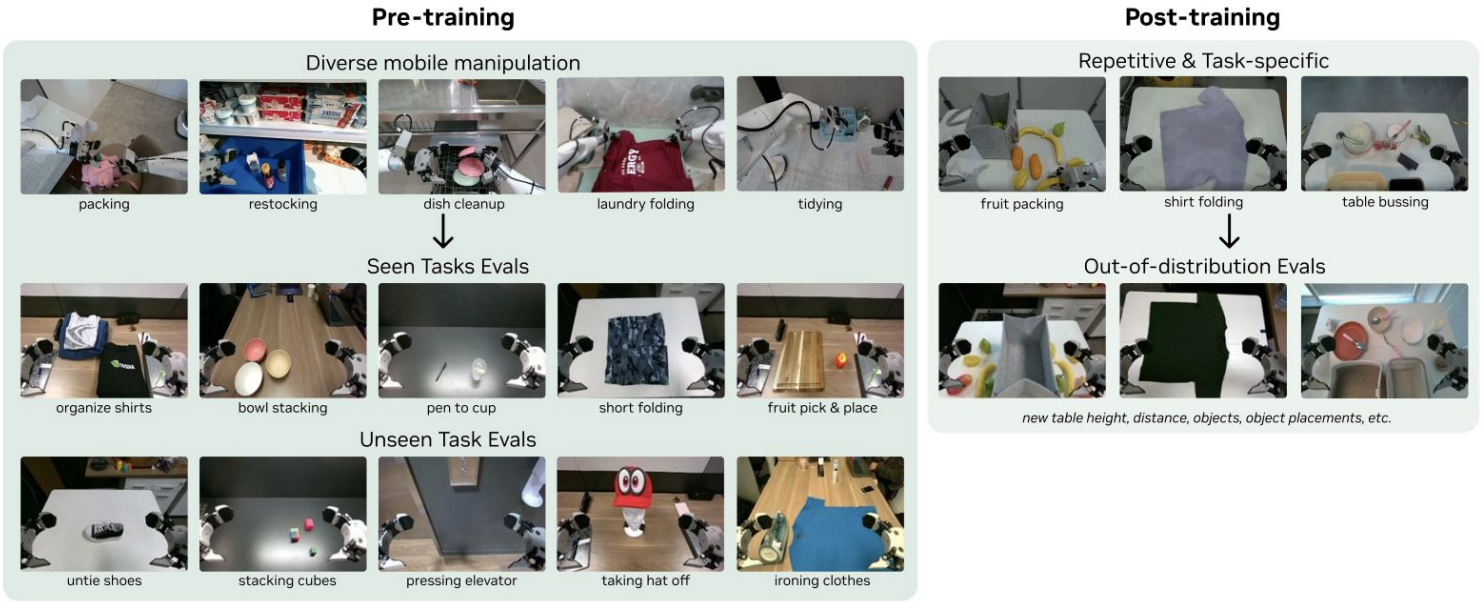
DreamZero: WAM is zero-shot policy

- Video diffusion model (Wan 2.1 14B) as backbone for joint video-action generator
- During training and inference, DreamZero predict action chunks and videos simultaneously
  - To incorporate previous observations, we use autoregressive teacher-forcing
  - KV cache the past observations during inference



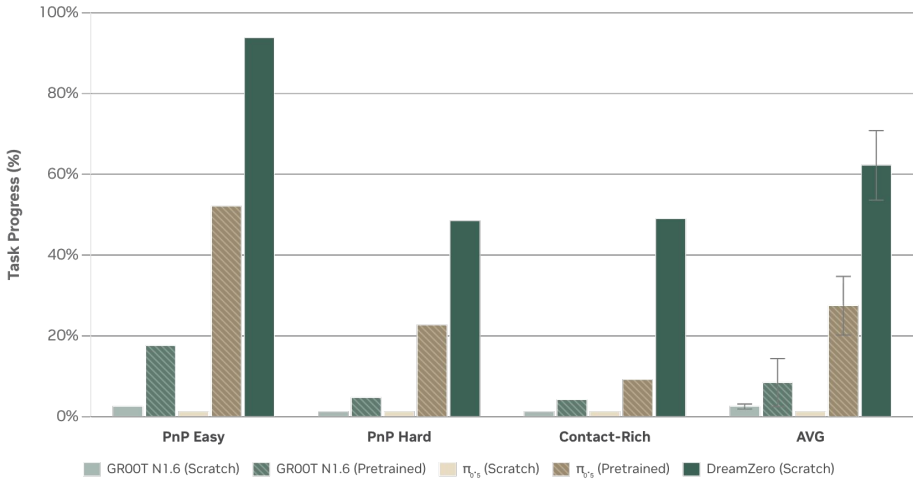
DreamZero: WAM is zero-shot policy

- Pre-training with diverse, non repetitive dataset (Agibot)
- Post-training with small repetitive dataset (Agibot, Droid)

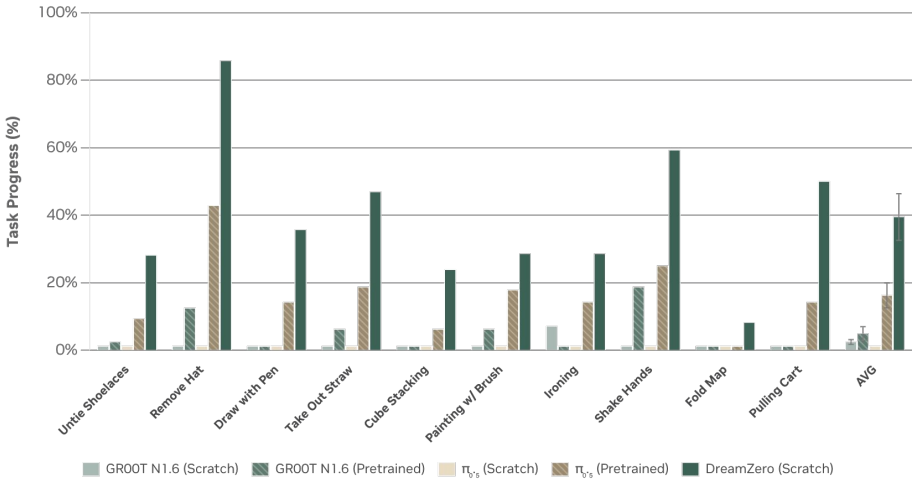


DreamZero: WAM is zero-shot policy

- DreamZero demonstrates superior zero-shot generalization
- Pretrained VLAs fail to generalize, even fail when trained from scratch

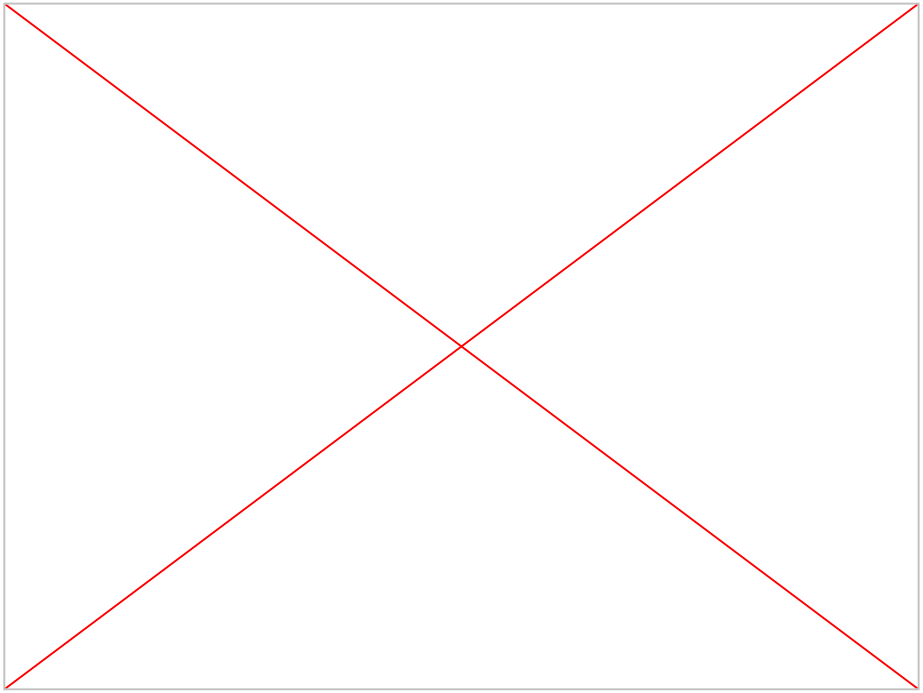


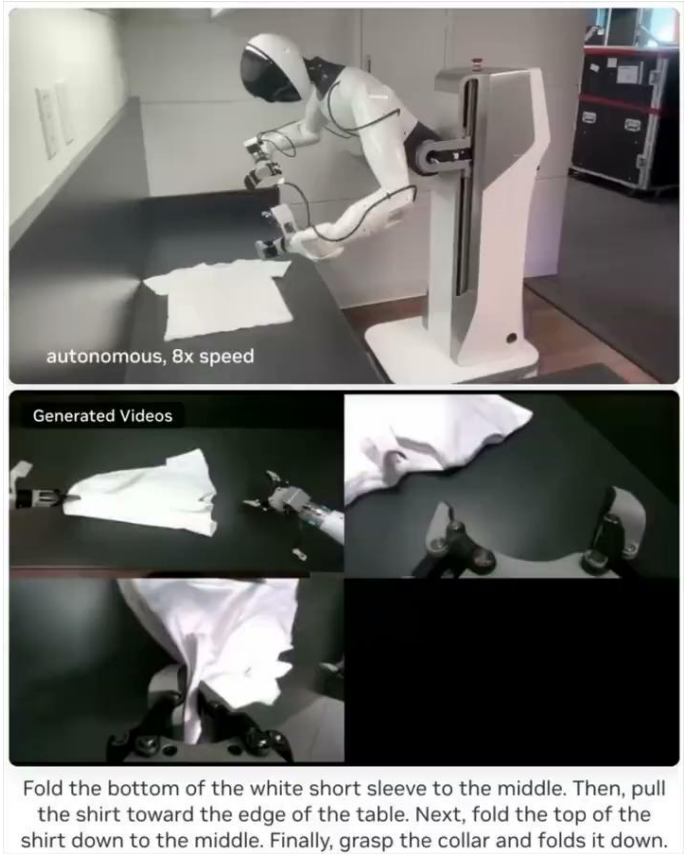
Seen tasks



Unseen tasks

DreamZero: WAM is zero-shot policy





https://dreamzero0.github.io/

DreamZero: WAM is zero-shot policy

Summary

- Video world models have great prior on motion understanding
- DreamZero adds action generation module through joint flow matching
- DreamZero trained on diverse non-repetitive dataset demonstrates zero-shot generalization
    - Show better performance than VLAs on seen and unseen tasks
    - Also, better on post-training and cross-embodiment

Future Directions

- Better video generation quality ⇏ Action generation quality
  - How jointly learning action and video generation enhances performance?
- Heavy inference cost
  - Do we really need to fully sample videos / latents?
- What is good representation for world modeling?
  - Self-supervised visual representation (e.g., DINO, SigLIP) vs pixels
- Can we do RL using world model environment and policy?
  - Dreamer v4 partially shown its promise on Minecraft
- DreamZero cuts rollout to 3-4 chunks
  - Can we extend to extremely long-horizon?