RDT-1B: a Diffusion Foundation Model for Bimanual Manipulation

Songming Liu*, Lingxuan Wu*, Bangguo Li, Hengkai Tan,
Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, Jun Zhu

Background

Problem Formulation

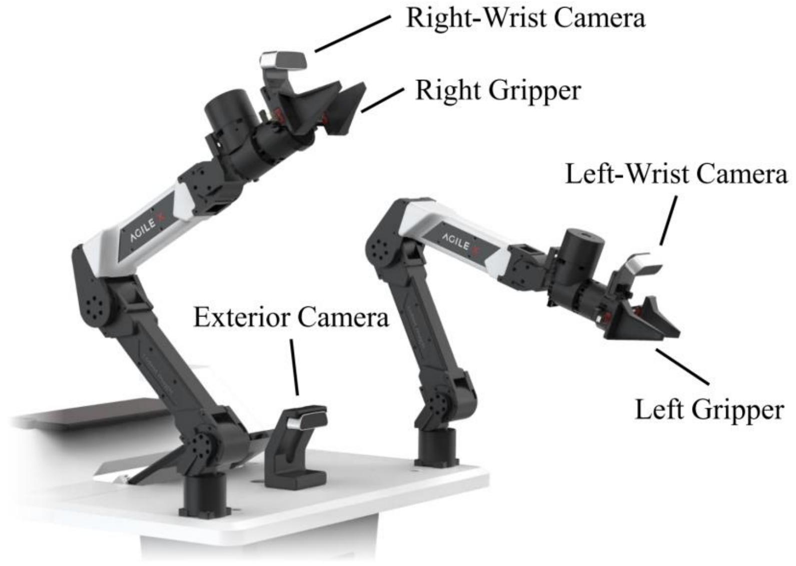
- Bimanual Manipulation
  - Essential for real-world tasks: necessary/faster

UMI [1]

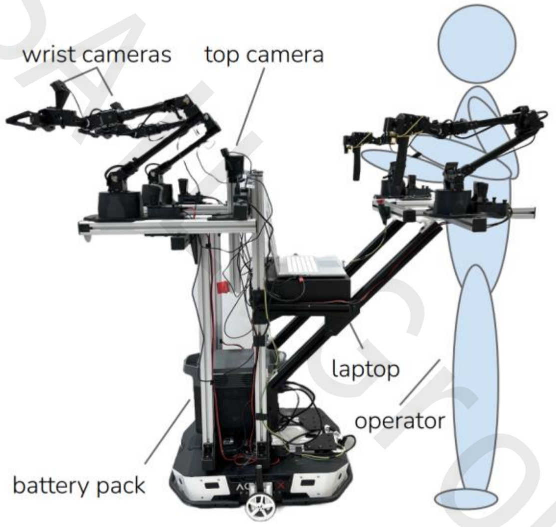
Mobile ALOHA [2]

Problem Formulation

- Hardware
  - ALOHA dual-arm robot by agilex.ai
  - Different from the original ALOHA; wheeled locomotion is not used



Ours



Mobile ALOHA [2]

Problem Formulation

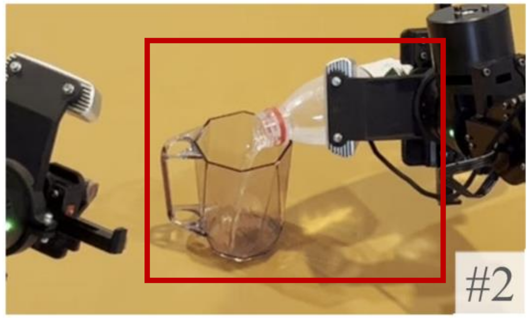
- Input
  - Language Instruction: ℓ
  - Observation $\boldsymbol{o}_t$:
    - RGB Images: $\boldsymbol{X}_{t-1}, \boldsymbol{X}_t$
    - Proprioception: $\boldsymbol{z}_t$
    - Control Frequency: c *(Why? We will discuss it later…)*
- Output
  - Action: $\boldsymbol{a}_t$,

  a subset of desired $\boldsymbol{z}_{t+1}$

Problem Formulation

- A task consists of:
  - Skill: verbs, “**wipe**” or **“open”**
  - Object: nouns, “**bottle**” or **“door”**
  - Scene: task environment, **some room**
  - Modality: how skill is performed, adverbials, “**pick the bottle <u>with the left hand</u>**”
- What is a useful policy?
  - When deployment, it can **generalize** to **unseen** objects, scenes, modalities, and even skills

Object: bottle

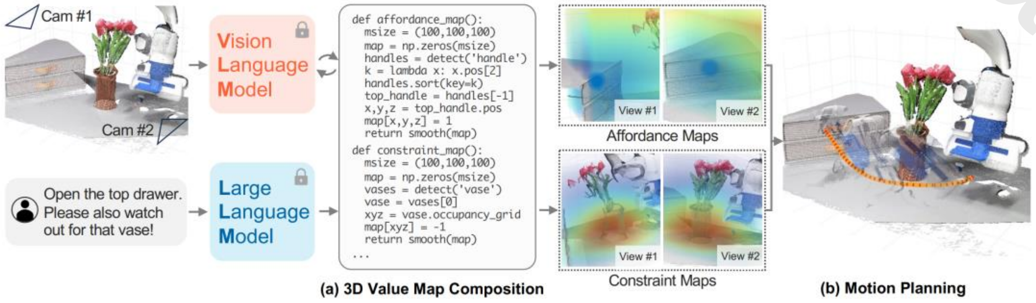


Skill: pour

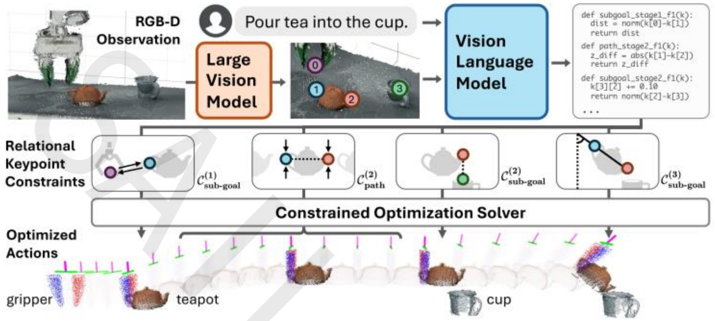
Scene: a meeting room

Previous Methods

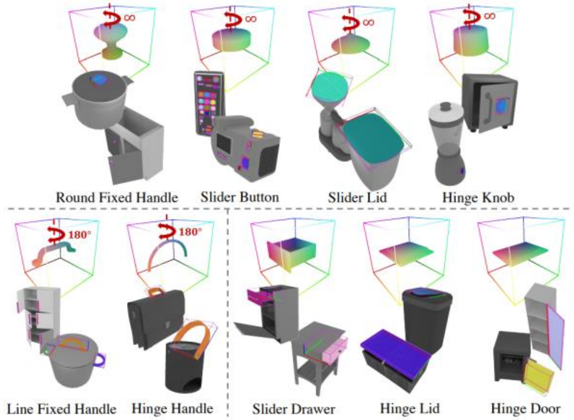
- This is challenging for rule-based methods
  - Depth failure, detection failure, **limited application range**, …



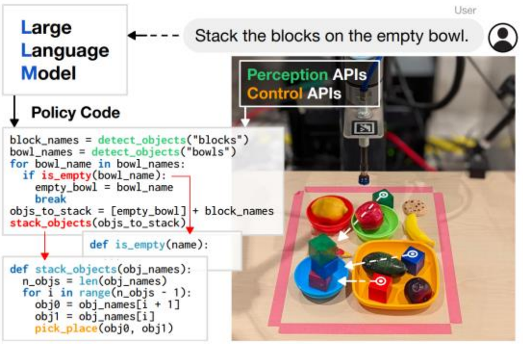
VoxPoser [3]



ReKep [4]

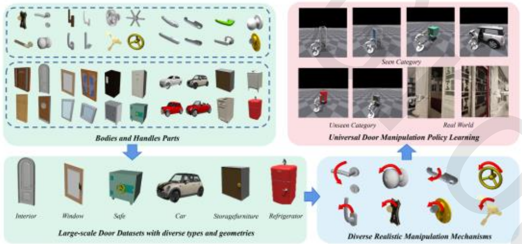


GAPartNet [7]




```
# if you see an orange, move backwards.
if detect_object("orange"):
    robot.set_velocity(x=-0.1, y=0, z=0)
# move rightwards until you see the apple.
while not detect_object("apple"):
    robot.set_velocity(x=0, y=0.1, z=0)
```

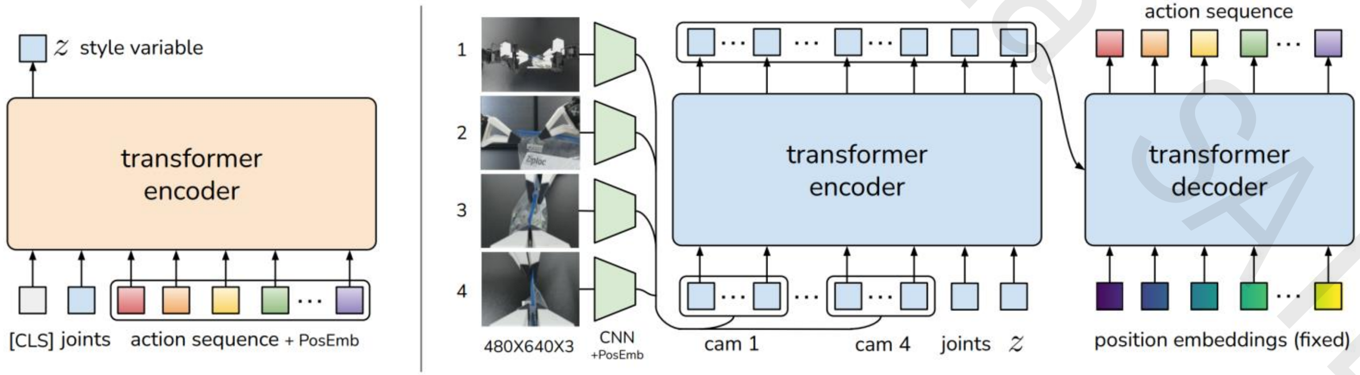
Code as Policies [5]



UniDoorManip [6]

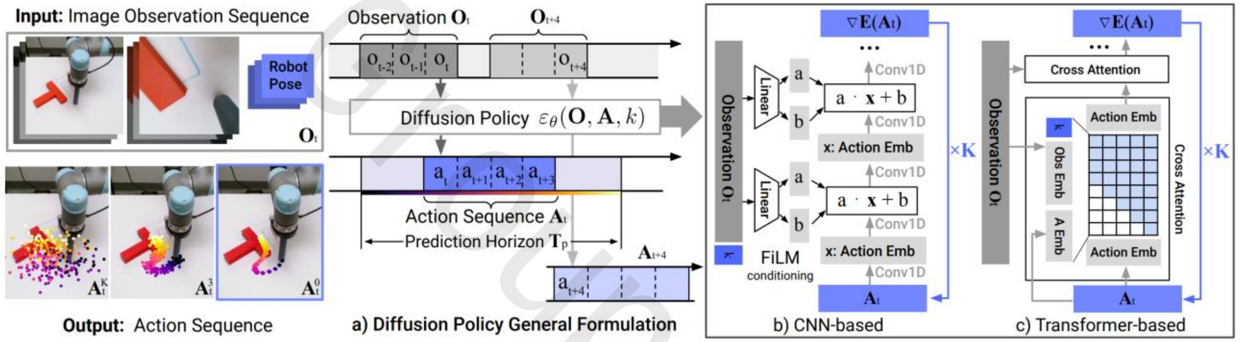
Previous Methods

- This is challenging for small end-to-end methods
  - No bias but limited generalizability



ACT [8]

Diffusion Policy [9]

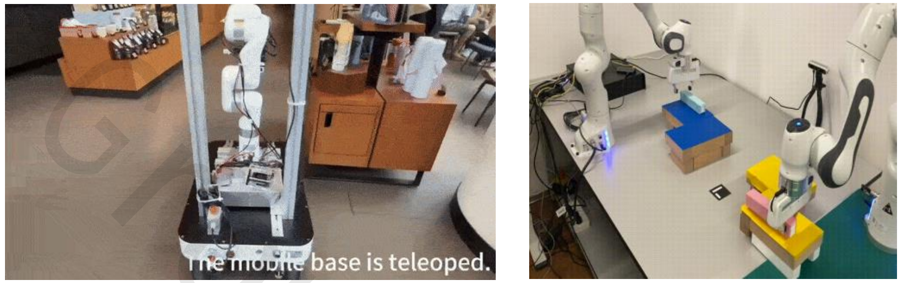


Way Out: Foundation Model

- Imitation Learning -> Foundation Model
  - Given that bimanual data are scarce,
  - Pre-Training: large-scale multi-robot human demonstrations
  - Fine-Tuning: small dataset of the dual-arm robot
- What about RL?
  - Sim2Real gap, reward design, not friendly to large models,...


ManiSkill [10]

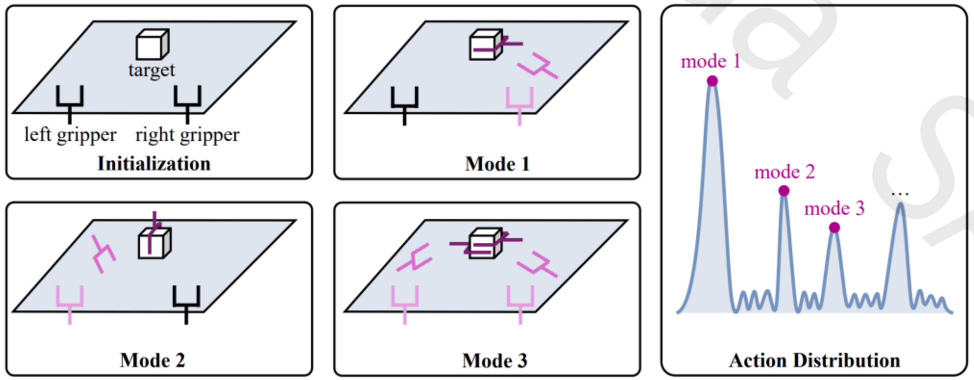

Isaac Lab [11]

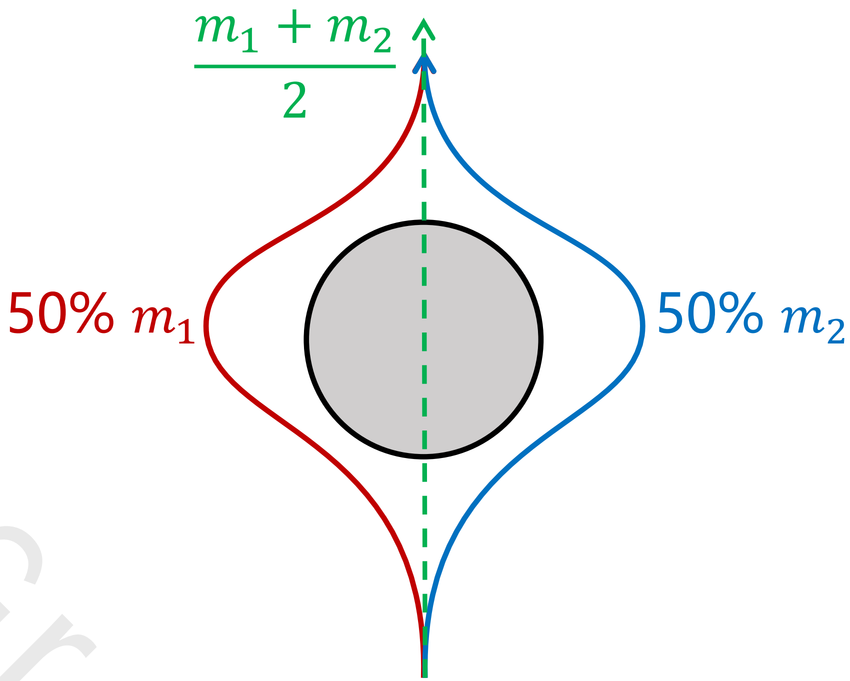


Huazhe's [12-13]

Challenges

- Challenge 1: How to design a powerful architecture?
  - Expressiveness -> multi-modality





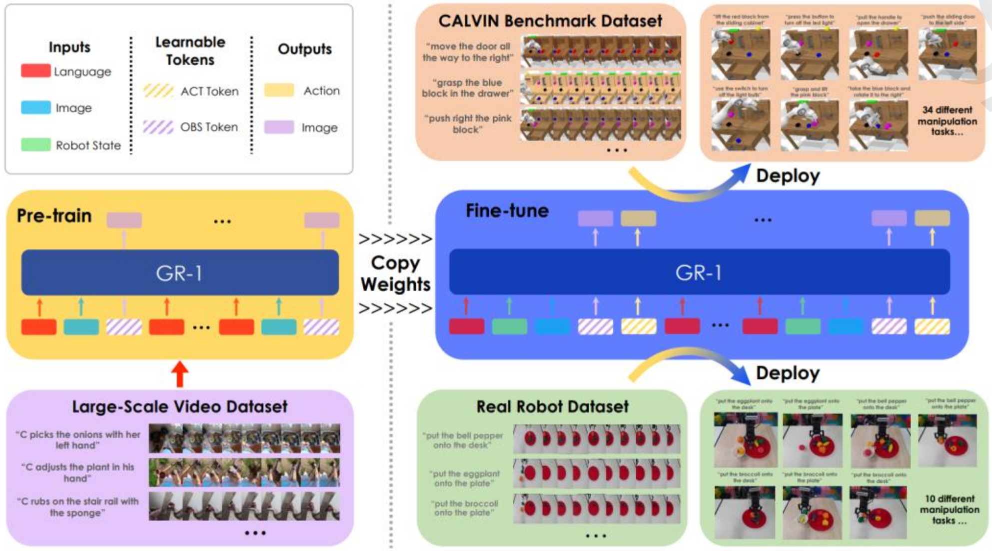
  - Scalability
    - Heterogenous inputs from various modalities
    - Training stability
    - Unique Challenges for Robotic Data *(We will discuss it later...)*

Challenges

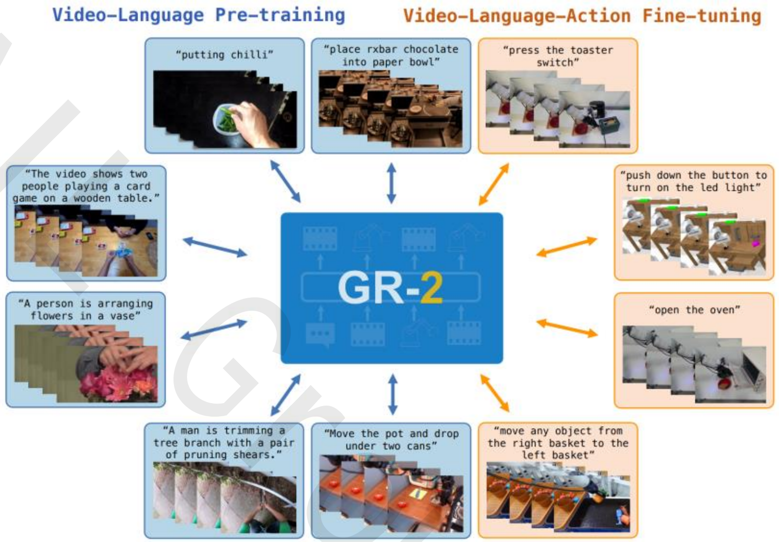
- Challenge 2: How to train on heterogeneous data?
  - Different physical structure -> different action space -> different format
  - Different control frequency *(this is why we feed it into the model)*
  - Different number of sensors, sensor types

Previous Foundation Models

- Transformer + MSE, $(\ell, \boldsymbol{o}_t) \mapsto \boldsymbol{a}_t$
  - Multi-modality -> one $(\ell, \boldsymbol{o}_t)$, many possible $\boldsymbol{a}_t$ -> learn an arithmetic average, which may be infeasible



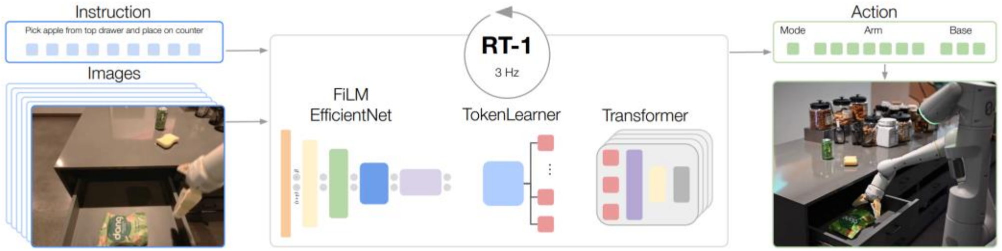
GR-1 [14]

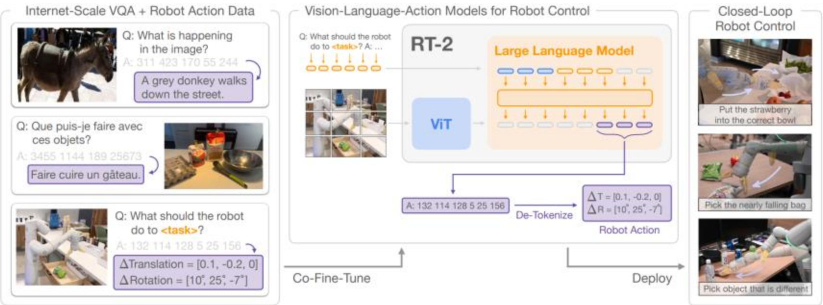


GR-2 [15]

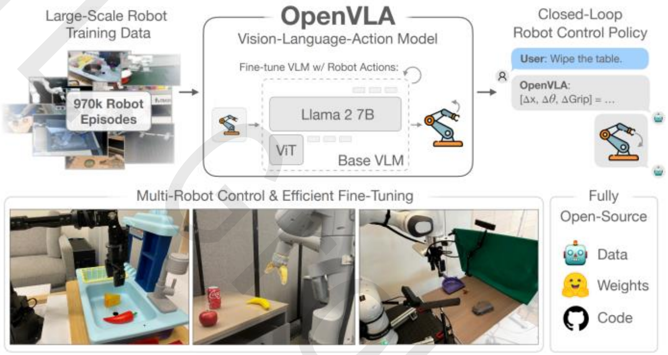
Previous Foundation Models

- Transformer + Discretized Token
  - Quantization errors
  - Classification loss -> lose information of the number
    - Cost(12, 13) == Cost(12, 120)
  - Uncoordinated behaviors [16] -> not a joint distribution





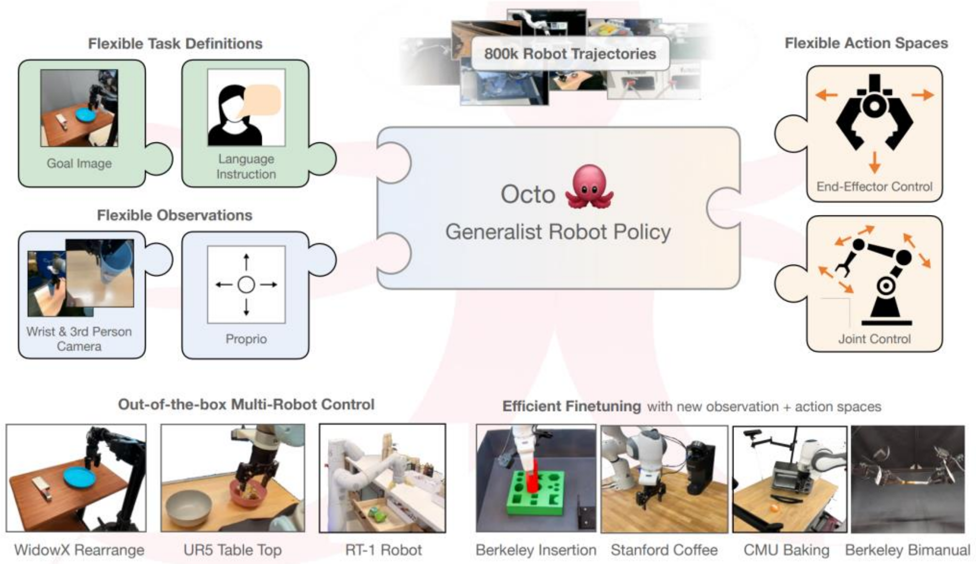
RT-1,2 [17-18]



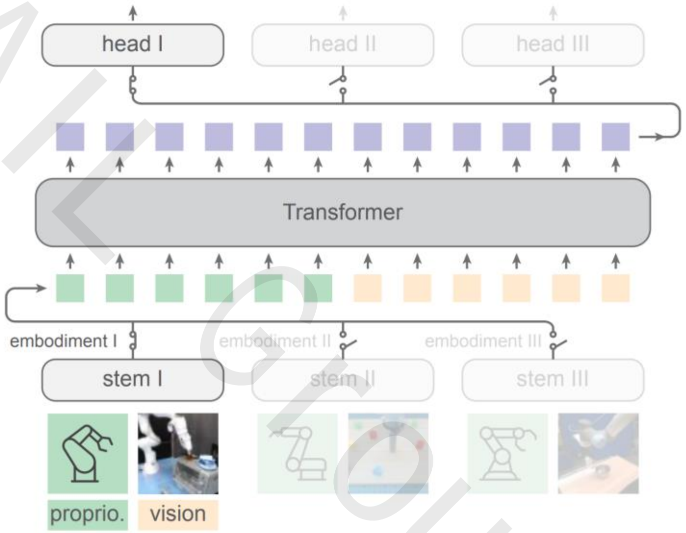
OpenVLA [19]

Previous Foundation Models

- Transformer + Diffusion Head
  - Empirically, we found that it is not as powerful as pure diffusion
  - We speculate that it may be due to the limited expressiveness



Octo [20]



HPT (concurrent work) [21]

Method

Diffusion Modeling

- An ideal choice -> model $p(\boldsymbol{a}_t|\ell,\boldsymbol{o}_t)$
  - Popular choice in history
  - Pros: expressiveness, sampling quality
  - Cons: slow sampling speed (for images/videos)
    - Actions are of much lower dimension; this drawback is minor!
- What is different for action data?
  - Image/video: high-dimensional, temporal and spatial continuity
  - Action: low-dimensional, but:
    - **Nonlinear dynamics**
    - **High-frequency changes**: stemming from collision,…
    - **Extreme values**: unreliable sensors,…

Overall Framework

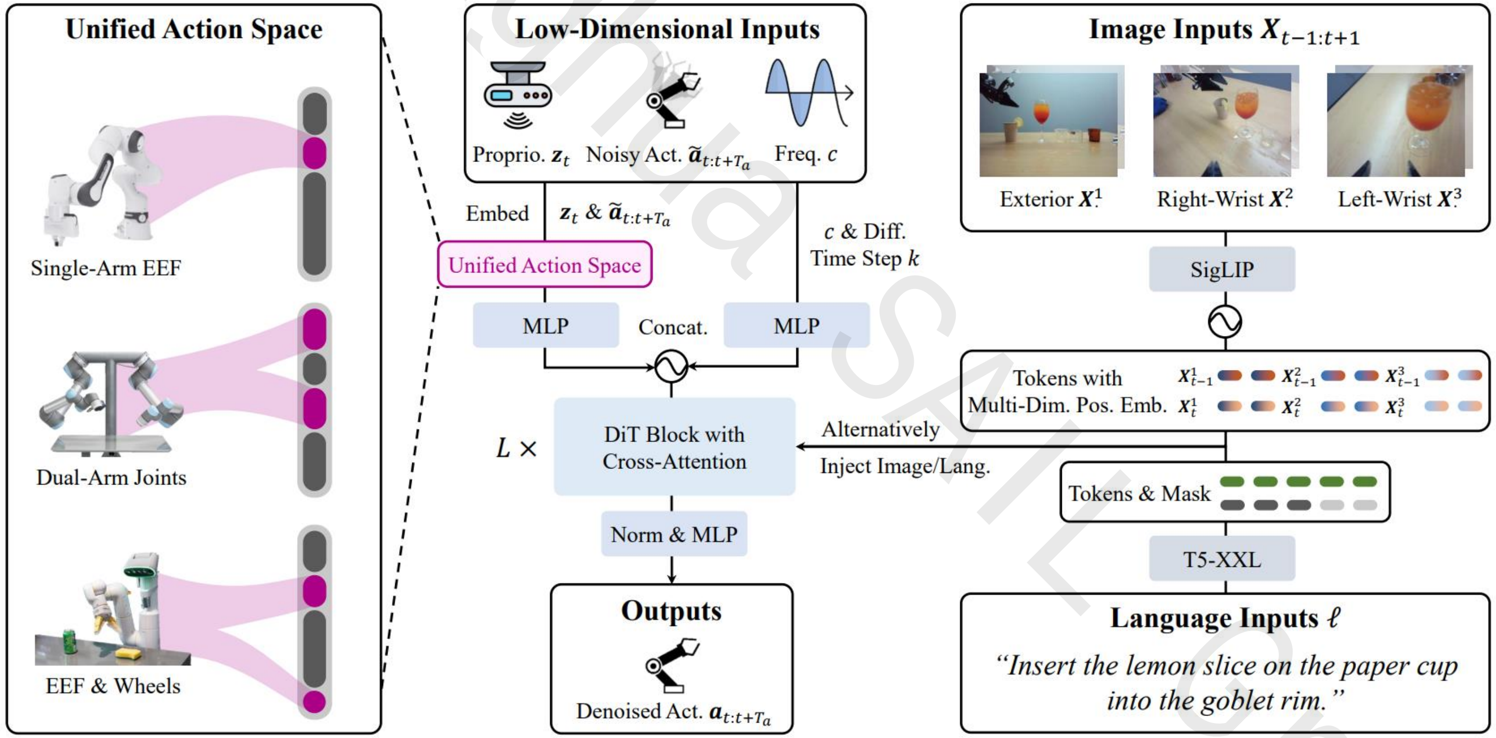


Figure 3: **RDT framework.** Heterogeneous action spaces of various robots are embedded into a unified action space for multi-robot training. **Inputs:** proprioception z_t, noisy action chunk $\tilde{a}_{t:t+T_a}$, control frequency c, and diffusion time step k, acting as denoising inputs; image inputs ($T_{\text{img}} = 2$ and $X_\cdot = \{X^1_\cdot, X^2_\cdot, X^3_\cdot\}$ denotes a set of images from exterior, right-wrist, and left wrist cameras) and language inputs, acting as conditions. **Outputs:** denoised action chunk $a_{t:t+T_a}$.

Encoding of Multi-Modal Inputs

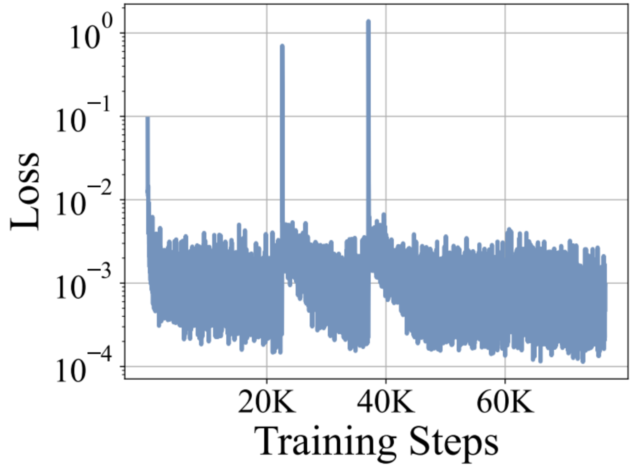
- Unify the format, encode into a single latent space
  - Low-dimensional:
    - MLP with Fourier Features -> high-frequency changes
  - Image inputs:
    - SigLIP -> extract spatial and semantic information
  - Language inputs:
    - T5-XXL -> overcome complexity and ambiguity
- Information Imbalance
  - Info(**exterior** camera) >> Info(**wrist** camera)
  - Info(**image**) >> Info(**language**)
  - Random masking -> avoid learning a **shortcut**

Exterior

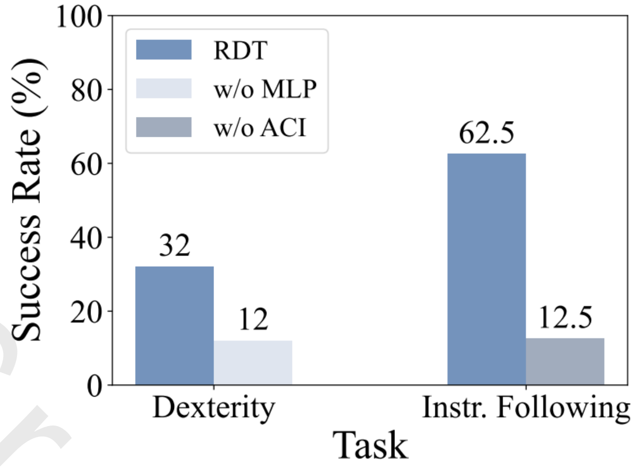
Wrist

Network Structure

- Transformer backbone -> scalability
- Key modifications
  - **QKNorm & RMSNorm**
    - Avoid numerical overflow caused by extreme values
    - Avoid token shift & attention shift caused by LayerNorm [22]
    - W/o this -> **unstable** training
  - **MLP Decoder**
    - Final linear layer -> MLP layer
    - Nonlinear approximation ability⬆️
    - W/o this -> fail to perform **dexterous** tasks



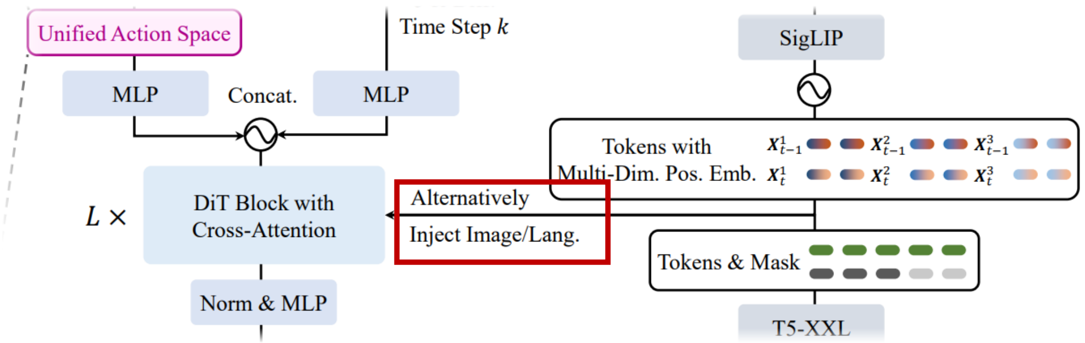
(a) Loss w/o QKN & RMSN

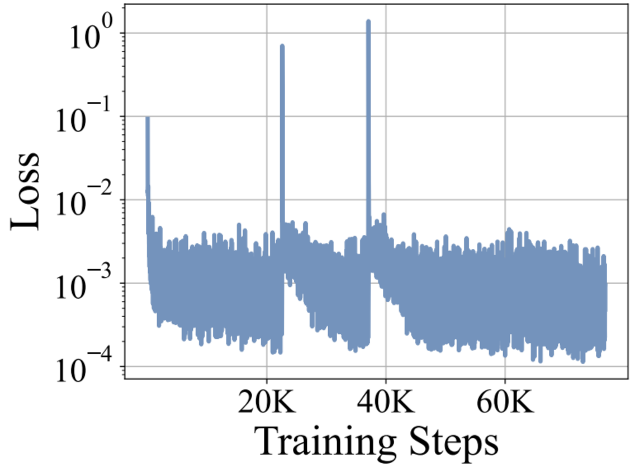


(b) Task w/o MLP or ACI

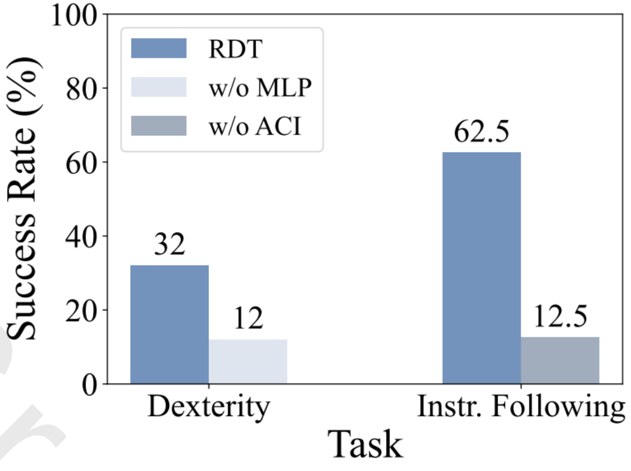
Network Structure

- Key modifications
  - **Alternating Condition Injection (ACI)**
    - #Tokens(**image**) >> #Tokens(**language**)
    - Decouple injection of language and images
    - Alternating injection in successive layers
    - W/o this -> **discard** language inputs
    -> instruction following ⬇️





(a) Loss w/o QKN & RMSN



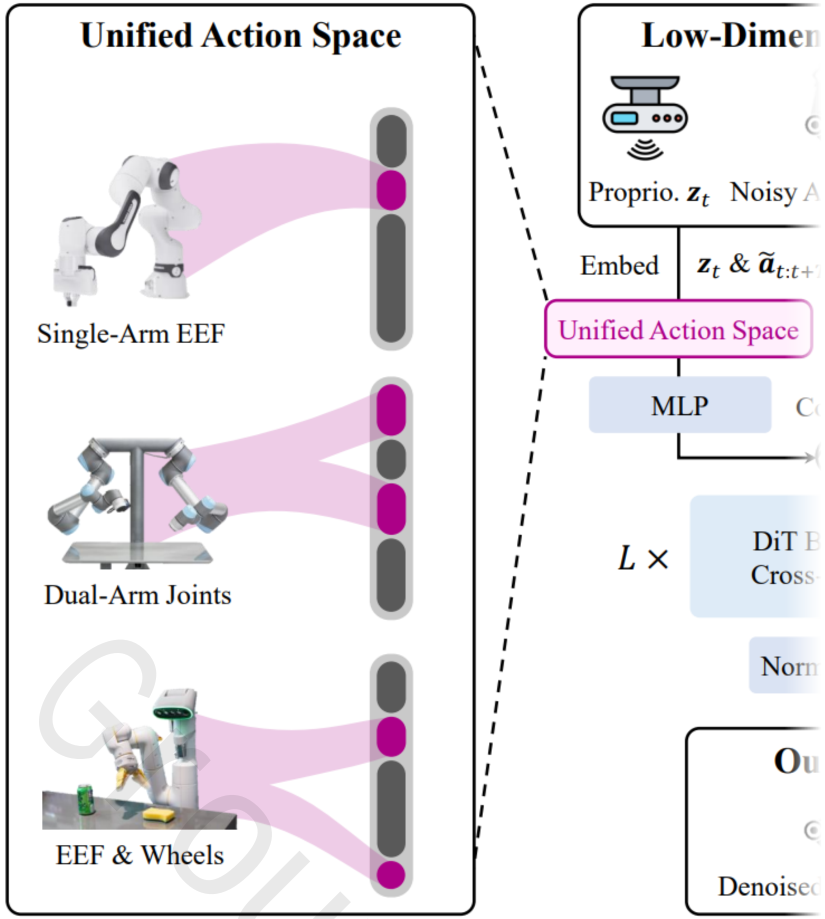
(b) Task w/o MLP or ACI

Training on Multi-Robot Data

- Previous approaches
  - Remove robots with incompatible action spaces
    - Lose valuable data
  - Train different encoders for different robots (Octo/HPT way)
    - Encoder parameters are not shared across robots
    - For a specific robot, less data on representation learning
      - Even worse for Robot learning! (robot data is expensive)
    - Unable to learn physical laws shared across robots

Training on Multi-Robot Data

- Our approach
  - Aggregate **all physical quantities** for manipulators to form a unified space
    - EEF, velocity, joint, wheeled locomotion,…
    - Not too many, only 128 dimensions
  - Each dimension has its physical meaning
  - Can learn shared physical laws across various robotic datasets
  - No normalization
    - "1" in position mean +1m for any robots, aligning the physics standard



Unified Space vs. Separate Encoders

$$\begin{array}{l} \text{EEF Pos.} \\ \\ \\ \text{EEF Ang.} \\ \\ \\ \text{Joints} \end{array} \begin{bmatrix} \boxed{\begin{matrix} 1.1 \\ -0.5 \\ 3 \end{matrix}} \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \xleftarrow{\text{Fill}} \begin{bmatrix} 1.1 \\ -0.5 \\ 3 \end{bmatrix}$$

What does it mean? Oh, it is the EEF coordinate of (+1.1m, -0.5m, 3m)

Pad

Raw Action

This format is shared across robots! Can take advantage of data from other robots...

Our Approach

$$\boxed{\begin{bmatrix} -0.02 \\ 1.01 \\ 0.76 \\ 0.67 \\ -0.64 \\ -0.77 \\ -0.39 \\ 0.21 \\ -1.21 \\ -0.33 \\ 0.51 \\ 0.46 \end{bmatrix}} \xleftarrow{\text{Encoding}} \text{Encoder} \leftarrow \begin{bmatrix} 1.1 \\ -0.5 \\ 3 \end{bmatrix}$$

What does it mean? It takes time and data to learn...

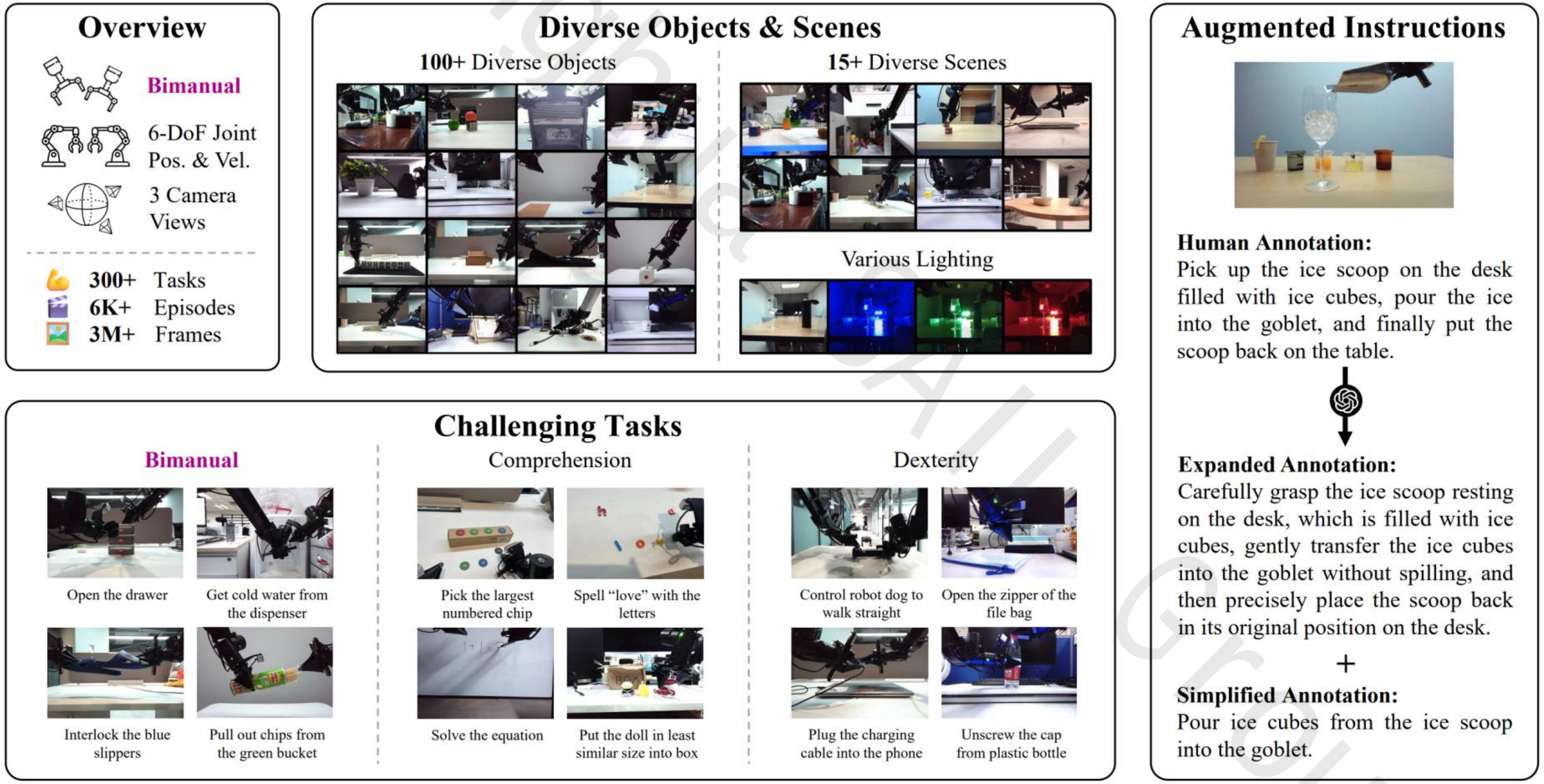
Raw Action

This encoder is not shared! Unfriendly for robots with little data...

Separate Encoders

Pre-Training and Fine-Tuning

-
-

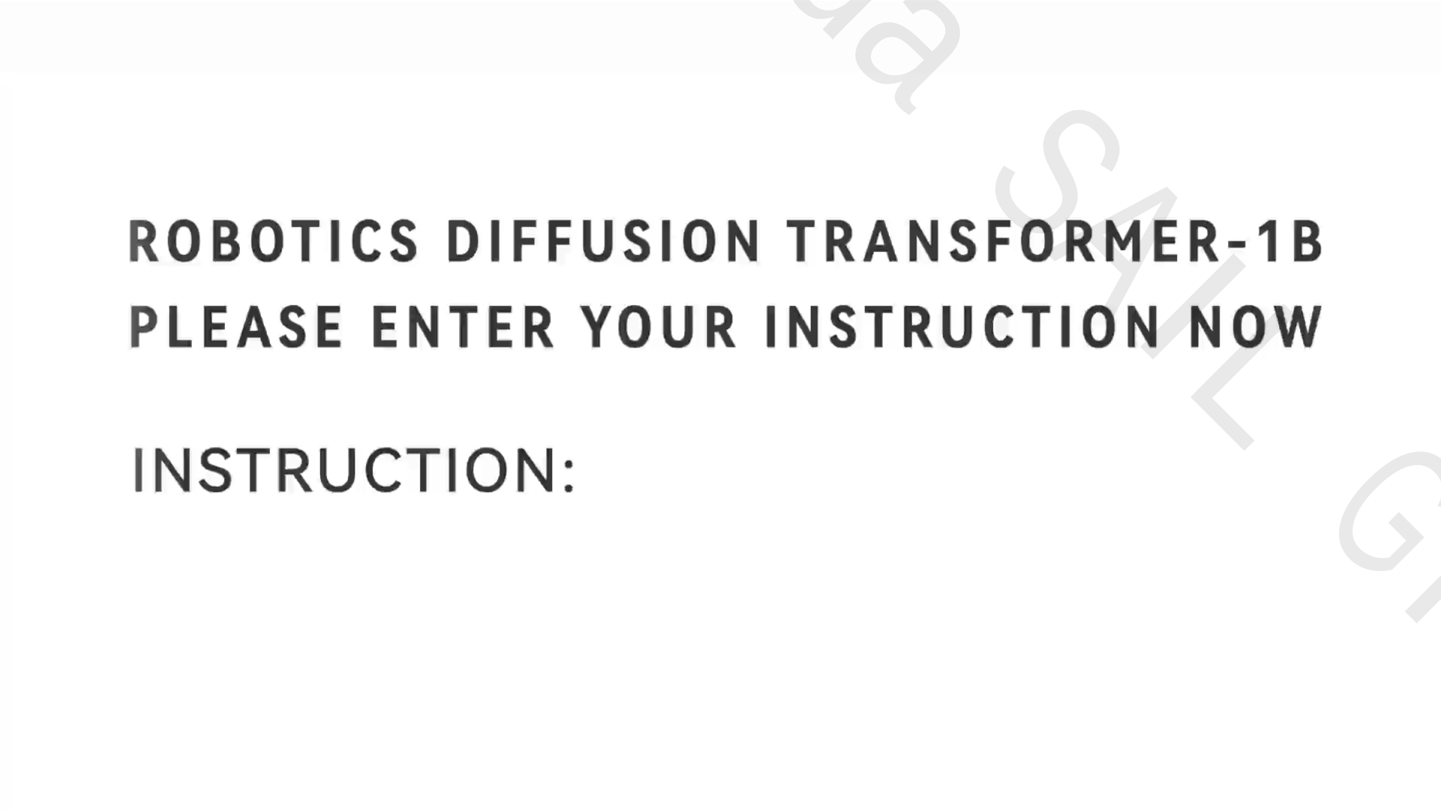


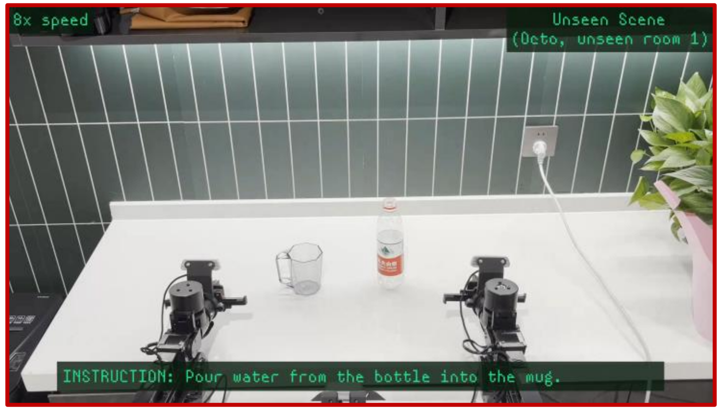
itions

Experiments

Experiments

- Q1: Can RDT **zero-shot** generalize to **unseen objects and scenes**?





Octo

Experiments

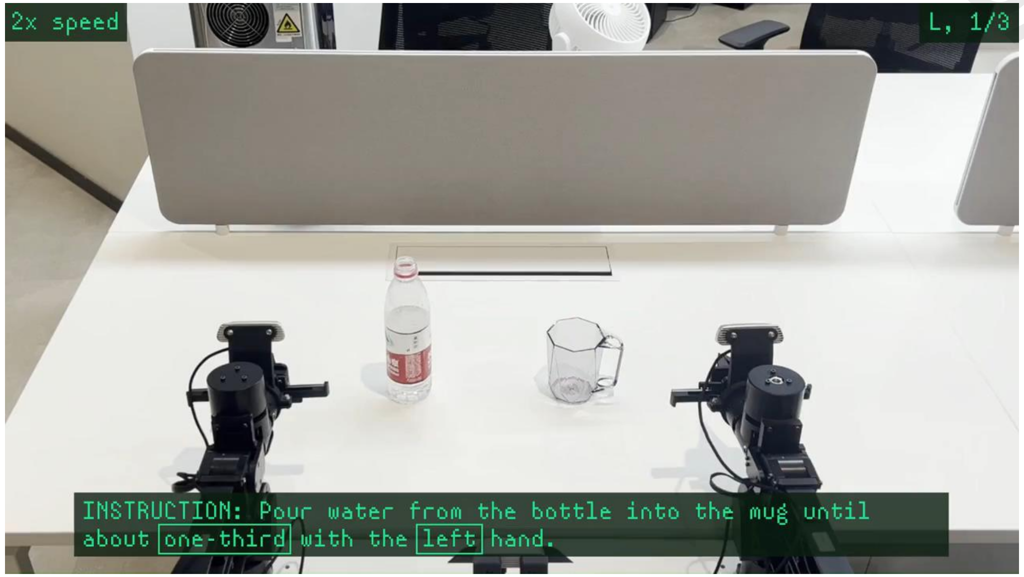
- Q1: Can RDT **zero-shot** generalize to **unseen objects and scenes**?

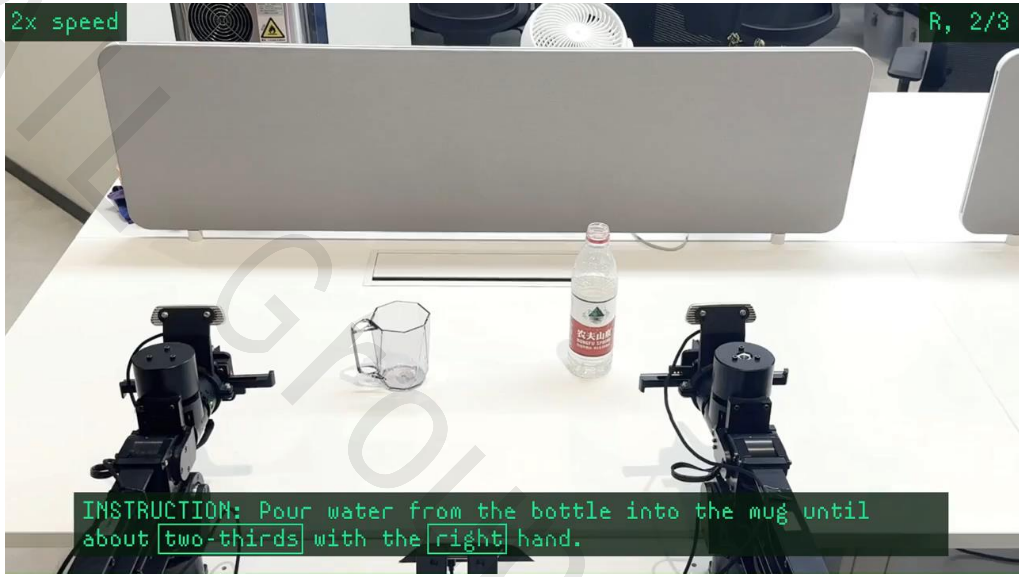
Wash Cup: seen cup 1 | unseen cup 1 | unseen cup 2 (**Unseen Object**)																		
	Pick Up Cup			Turn On Faucet			Get Water			Pour Out Water			Place Back Cup			Total		
ACT	50	12.5	37.5	0	0	0	0	0	0	0	0	0	37.5	0	0	0	0	0
OpenVLA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Octo	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RDT (scratch)	37.5	12.5	0	0	12.5	12.5	0	0	0	37.5	12.5	0	25	0	0	0	0	0
RDT (**ours**)	87.5	87.5	50	62.5	75	50	50	75	50	87.5	75	50	87.5	62.5	50	**50**	**75**	**50**

Pour Water: unseen room 1 | unseen room 2 | unseen room 3 (**Unseen Scene**)												
	Pick Up Bottle			Pour Water			Place Back Bottle			Total		
ACT	25	87.5	25	0	50	12.5	0	37.5	12.5	0	37.5	12.5
OpenVLA	0	0	0	0	0	0	0	0	0	0	0	0
Octo	50	0	12.5	12.5	0	0	12.5	0	0	12.5	0	0
RDT (scratch)	62.5	100	62.5	25	87.5	37.5	25	75	25	25	75	25
RDT (**ours**)	62.5	100	62.5	62.5	100	62.5	62.5	100	62.5	**62.5**	**100**	**62.5**

Experiments

- Q2: How effective is RDT's **zero-shot** instruction-following capability for **unseen modalities**?
  - "1/3" and "2/3" are unseen during training
  - Ground the **language concepts** to the **height** in the physical world





Experiments

- Q2: How effective is RDT's **zero-shot** instruction-following capability for **unseen modalities**?
  - Resulting water levels over **8 trials**

1/3

2/3

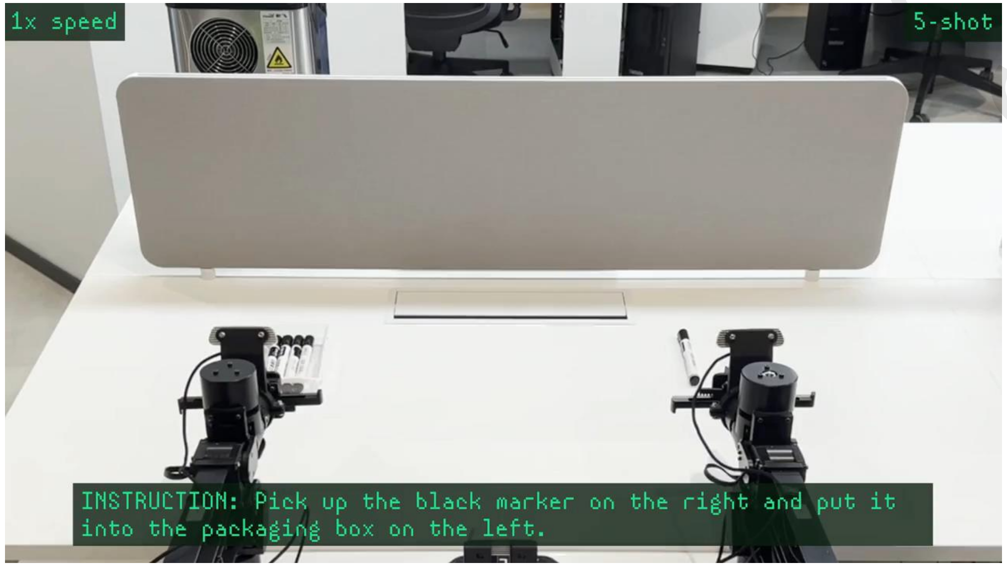
Experiments

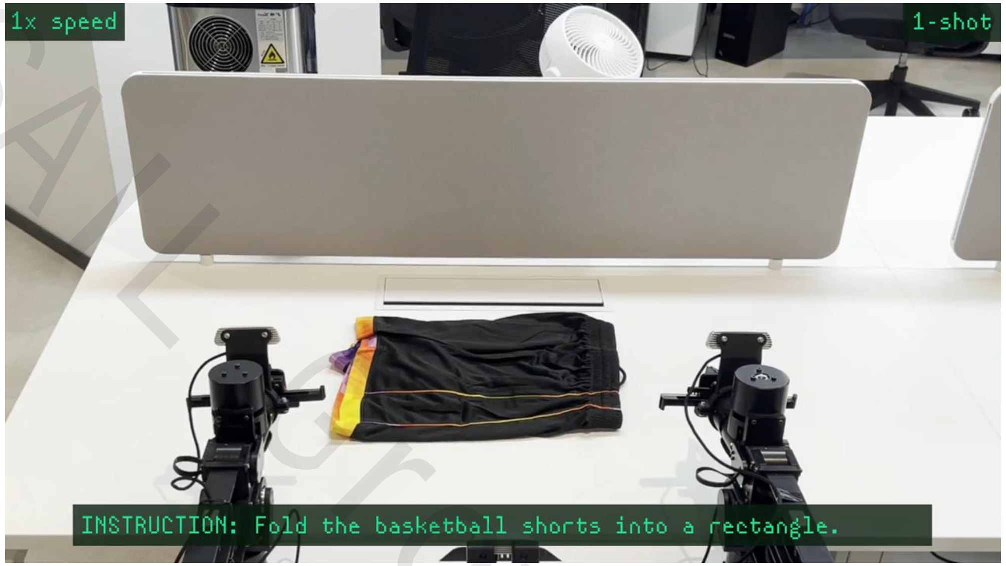
- Q2: How effective is RDT's **zero-shot** instruction-following capability for **unseen modalities**?

Pour Water-L-1/3 | Pour Water-R-2/3 (**Instruction Following**)												
	Pick Up Bottle		Pour Water		Place Back Bottle		Total		Correct Hand		Correct Amount	
OpenVLA	50	0	0	0	0	0	0	0	50	0	0	0
Octo	0	0	0	0	0	0	0	0	0	0	0	0
RDT (scratch)	100	75	75	25	62.5	25	62.5	25	100	75	62.5	12.5
RDT (**ours**)	100	87.5	100	87.5	100	87.5	**100**	**87.5**	**100**	**87.5**	**100**	**75**

Experiments

- Q3: Can RDT facilitate **few-shot learning** for previously **unseen skills**?





Experiments

- Q3: Can RDT facilitate **few-shot learning** for previously **unseen skills**?
  - The baselines…

ACT

OpenVLA

RDT (scratch)

Experiments

- Q3: Can RDT facilitate **few-shot learning** for previously **unseen skills**?

	Handover (**5-Shot**)					Fold Shorts (**1-Shot**)
	Pick Up Pen	Switch Hand	Drop Pen	Fall into Box	Total	Total
ACT	44	0	0	0	0	0
OpenVLA	0	0	0	0	0	0
Octo	12	0	0	0	0	4
RDT (scratch)	88	32	24	16	16	40
RDT (**ours**)	100	56	56	40	**40**	**68**

Experiments

- Q4: Is RDT capable of completing tasks that require **delicate operations**?

ROBOTICS DIFFUSION TRANSFORMER-1B
PLEASE ENTER YOUR INSTRUCTION NOW

INSTRUCTION:

Experiments

- Q4: Is RDT capable of completing tasks that require **delicate operations**?
  - Joystick is only 2cm high
  - Slight push angle -> robot dog deviation

	Robot Dog (**Dexterity**)			
	Grab Remote	Push Joystick	Total	Walk Straight
ACT	88	32	32	32
OpenVLA	84	0	0	0
Octo	100	4	4	0
RDT (scratch)	100	64	64	32
RDT (**ours**)	100	76	**76**	**48**

Experiments

- Q5: Are **large model sizes**, **extensive data**, and **diffusion modeling** helpful for RDT’ s performance?

Table 2: **Ablation study results.** Here are the success rates (%) of the original RDT and its three variants in tasks of *Wash Cup* (unseen cup 2, total success rate), *Pour Water* (unseen room 3, total success rate), and *Pour Water-L-1/3* (correct amount sub-task). All the models except *RDT (scratch)* are pre-trained before fine-tuning.

VARIANT NAME	UNSEEN OBJECT	UNSEEN SCENE	INSTRUCTION FOLLOWING
RDT (regress)	12.5	50	12.5
RDT (small)	37.5	**62.5**	25
RDT (scratch)	0	25	62.5
RDT **(ours)**	**50**	**62.5**	**100**

Pre-Training is crucial for generalizability!

RDT (scratch) performs poorly on unseen objects/scenes...

Practical Tips

部署到我的机器人上需要微调吗

- 如果机器人包含在预训练数据中
  - 如Franka, WidowX, UR5等,
  - 可以直接部署试试效果
- 一般情况下，都建议大家微调
- 收集多大微调数据集?
  - 几十条到几百条，多任务，标注好语言
- 训练多长时间?
  - 等Sampling Error收敛即可，一般不会过拟合
- 数据格式有要求?
  - 腕部相机、外部相机均可，注意放入图片的顺序有要求

部署到我的机器人上需要微调吗

- 如果我的格式比较特殊
  - 三个机械臂
  - 两个外部相机
  - 灵巧手、人形、机械腿
- 也可以训练
  - 需要重新设计动作空间，重新训练encoder和decoder
  - 训练的时候可以给主干网络设一个小的学习率，encoder、decoder设一个大一点的
    - 至少一开始不要给主干网络太大的学习率
  - 性能会有损失吗？会有一点，但是下限是separate encoder的方法，如Octo和HPT

我的显卡能训起来RDT吗

- 如果 > =24GB
  - 直接训即可
  - 可以考虑加gradient checkpointing拉大bs
  - xformers也可以考虑，可以省显存
- 如果12-24GB
  - 可以，试试8bit量化+8bit adam （参考stable diffusion）
- 如果6-12GB
  - 也可以，试试4bit量化
- 如果6GB以下
  - 买个大点的显卡吧，1070 8G目前价格不到1K

我的显卡能训起来RDT吗

- 也可以考虑
  - LoRA、adaptor等高效微调方法
- 如果大家实现了可以给仓库交PR，
- 我们一起努力让更多人用上具身大模型

References

[1] Chi, C., Xu, Z., Pan, C., Cousineau, E., Burchfiel, B., Feng, S., ... & Song, S. (2024). Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots. arXiv preprint arXiv:2402.10329.

[2] Fu, Z., Zhao, T. Z., & Finn, C. (2024). Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation. arXiv preprint arXiv:2401.02117.

[3] Huang, W., Wang, C., Zhang, R., Li, Y., Wu, J., & Fei-Fei, L. (2023). Voxposer: Composable 3d value maps for robotic manipulation with language models. arXiv preprint arXiv:2307.05973.

[4] Huang, W., Wang, C., Li, Y., Zhang, R., & Fei-Fei, L. (2024). Rekep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation. arXiv preprint arXiv:2409.01652.

[5] Liang, J., Huang, W., Xia, F., Xu, P., Hausman, K., Ichter, B., ... & Zeng, A. (2023, May). Code as policies: Language model programs for embodied control. In 2023 IEEE International Conference on Robotics and Automation (ICRA) (pp. 9493-9500). IEEE.

[6] Li, Y., Zhang, X., Wu, R., Zhang, Z., Geng, Y., Dong, H., & He, Z. (2024). Unidoormanip: Learning universal door manipulation policy over large-scale and diverse door manipulation environments. arXiv preprint arXiv:2403.02604.

[7] Geng, H., Xu, H., Zhao, C., Xu, C., Yi, L., Huang, S., & Wang, H. (2023). Gapartnet: Cross-category domain-generalizable object perception and manipulation via generalizable and actionable parts. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 7081-7091).

[8] Zhao, T. Z., Kumar, V., Levine, S., & Finn, C. (2023). Learning fine-grained bimanual manipulation with low-cost hardware. arXiv preprint arXiv:2304.13705.

[9] Chi, C., Xu, Z., Feng, S., Cousineau, E., Du, Y., Burchfiel, B., ... & Song, S. (2023). Diffusion policy: Visuomotor policy learning via action diffusion. The International Journal of Robotics Research, 02783649241273668.

References

[10] Mu, T., Ling, Z., Xiang, F., Yang, D., Li, X., Tao, S., ... & Su, H. (2021). Maniskill: Generalizable manipulation skill benchmark with large-scale demonstrations. arXiv preprint arXiv:2107.14483.

[11] https://docs.omniverse.nvidia.com/isaacsim/latest/isaac_lab_tutorials/index.html

[12] Yuan, Z., Wei, T., Cheng, S., Zhang, G., Chen, Y., & Xu, H. (2024). Learning to manipulate anywhere: A visual generalizable framework for reinforcement learning. arXiv preprint arXiv:2407.15815.

[13] Li, Y., Pan, C., Xu, H., Wang, X., & Wu, Y. (2023, May). Efficient bimanual handover and rearrangement via symmetry-aware actor-critic learning. In 2023 IEEE International Conference on Robotics and Automation (ICRA) (pp. 3867-3874). IEEE.

[14] Wu, H., Jing, Y., Cheang, C., Chen, G., Xu, J., Li, X., ... & Kong, T. (2023). Unleashing large-scale video generative pre-training for visual robot manipulation. arXiv preprint arXiv:2312.13139.

[15] Cheang, C. L., Chen, G., Jing, Y., Kong, T., Li, H., Li, Y., ... & Zhu, M. (2024). GR-2: A Generative Video-Language-Action Model with Web-Scale Knowledge for Robot Manipulation. arXiv preprint arXiv:2410.06158.

[16] Pearce, T., Rashid, T., Kanervisto, A., Bignell, D., Sun, M., Georgescu, R., ... & Devlin, S. (2023). Imitating human behaviour with diffusion models. arXiv preprint arXiv:2301.10677.

[17] Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., ... & Zitkovich, B. (2022). Rt-1: Robotics transformer for real-world control at scale. arXiv preprint arXiv:2212.06817.

[18] Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Chen, X., Choromanski, K., ... & Zitkovich, B. (2023). Rt-2: Vision-language-action models transfer web knowledge to robotic control. arXiv preprint arXiv:2307.15818.

[19] Kim, M. J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., ... & Finn, C. (2024). OpenVLA: An Open-Source Vision-Language-Action Model. arXiv preprint arXiv:2406.09246.

References

[20] Team, O. M., Ghosh, D., Walke, H., Pertsch, K., Black, K., Mees, O., ... & Levine, S. (2024). Octo: An open-source generalist robot policy. arXiv preprint arXiv:2405.12213.

[21] Wang, L., Chen, X., Zhao, J., & He, K. (2024). Scaling proprioceptive-visual learning with heterogeneous pre-trained transformers. arXiv preprint arXiv:2409.20537.

[22] Huang, N., Kümmerle, C., & Zhang, X. (2024). UnitNorm: Rethinking Normalization for Transformers in Time Series. arXiv preprint arXiv:2405.15903.

Thank You!

Page: https://rdt-robotics.github.io/rdt-robotics/

Paper: https://arxiv.org/pdf/2410.07864

Code: https://github.com/thu-ml/RoboticsDiffusionTransformer

Model: https://huggingface.co/robotics-diffusion-transformer/rdt-1b

Discord: https://discord.gg/vsZS3zmf9A