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

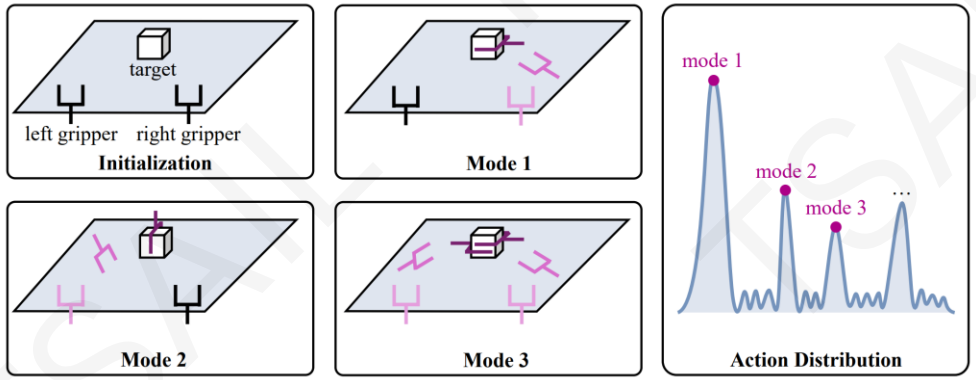
Songming Liu*, Lingxuan Wu*, Bangguo Li, Hengkai Tan,
Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, Jun Zhu

Problem Formulation

- Human Data –Imitation Learning–> Foundation Model
  - Generalizability
- Inputs – Outputs
  - Language Instruction x RGB observation -> Actions
- Bimanual manipulation
  - More efficient/necessary
- Training paradigm
  - Given that the data of a specific robot are scarce,
  - Pre-Training: large-scale multi-robot human demonstrations
  - Fine-Tuning: small dataset of the target dual-arm robot

Diffusion Modeling

- Multi-modality of human data



- To model $p(\boldsymbol{a}_t|\ell, \boldsymbol{o}_t)$
  - An ideal choice: diffusion models
  - Pros: expressiveness -> bimanual distribution, sampling quality
  - Cons: slow sampling speed (for images/videos)
    - Actions are of much lower dimension; this drawback is minor!

Encoding of Multi-Modal Inputs

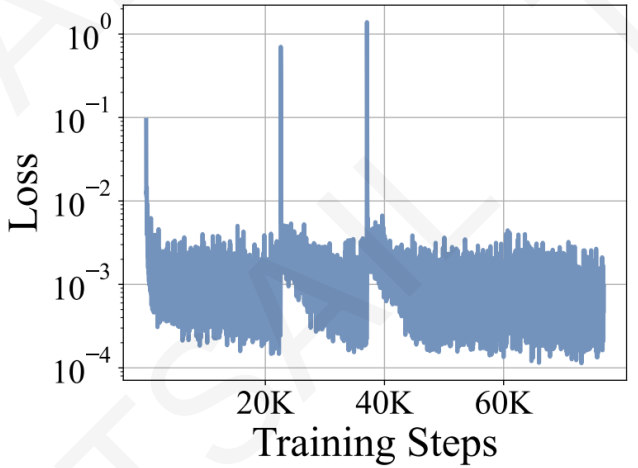
- Encode inputs into a single latent space
  - Low-dimensional:
    - MLP with Fourier Features -> high-frequency changes
  - Image inputs:
    - SigLIP -> extract spatial and semantic information
  - Language inputs:
    - T5-XXL -> overcome complexity and ambiguity
- Information Imbalance
  - Info(**image**)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

Exterior

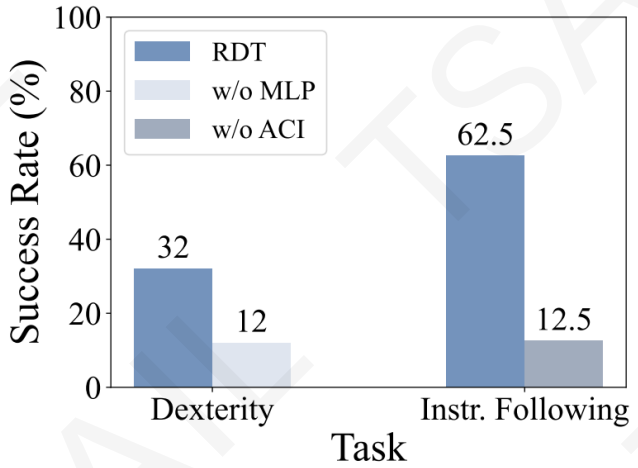
Wrist

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

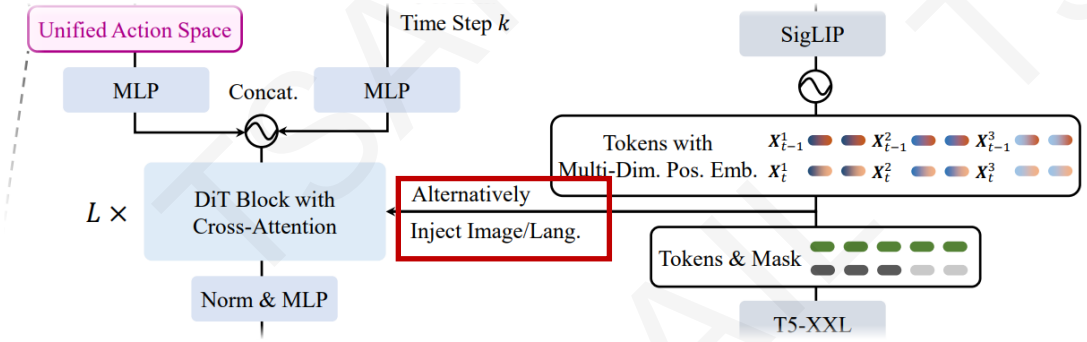
Network Structure

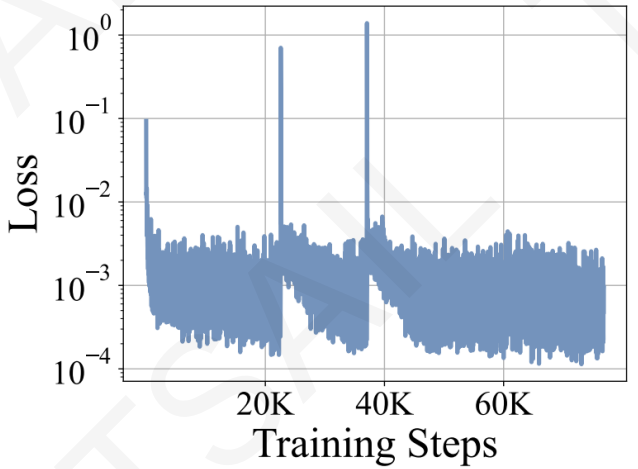
- Key modifications
  - **Alternating Condition Injection (ACI)**
    - #Tokens(**image**) >> #Tokens(**language**)

$$\text{Importance(lang)} = \frac{\sum\exp(\text{lang})}{\sum\exp(\text{lang}) + \sum\exp(\text{image})} \downarrow$$

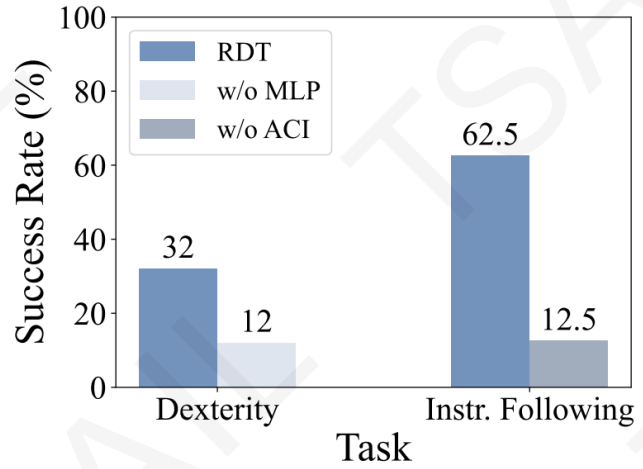
    - Decouple injection of language and images
    - W/o this -> **discard** language inputs

-> instruction following ⬇️





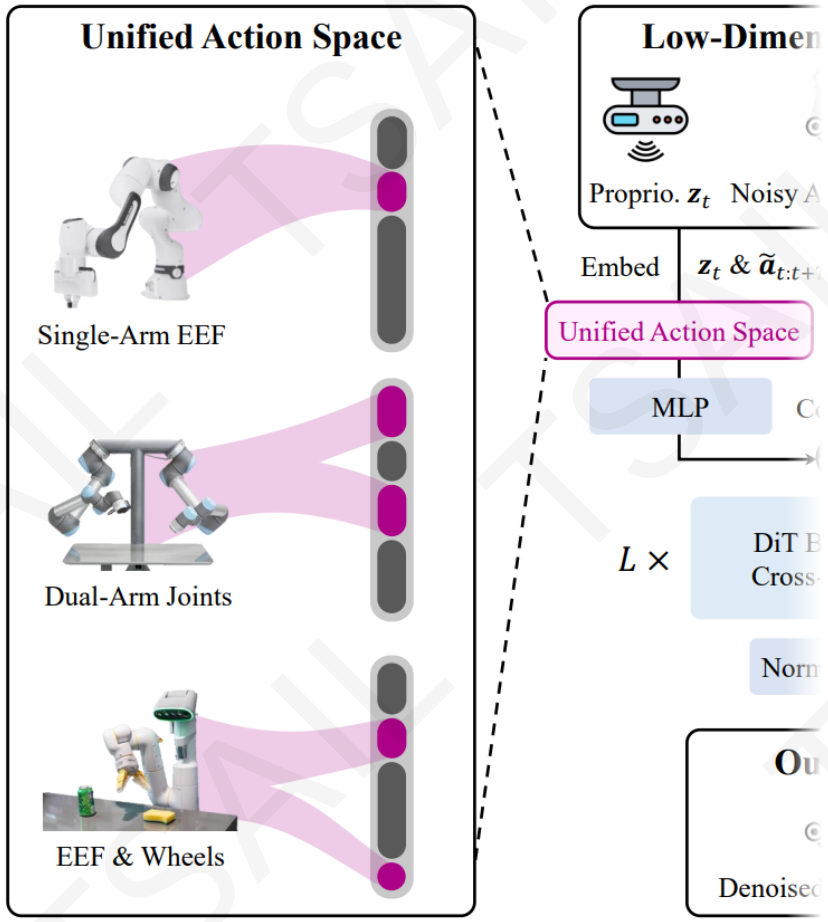
(a) Loss w/o QKN & RMSN



(b) Task w/o MLP or ACI

Training on Multi-Robot Data

- Our approach
  - Aggregate **all physical quantities** for manipulators to form a unified space
    - EEF, velocity, joint, wheeled locomotion,…
    - 128-dimensional unified action space
  - Each dimension has its **physical meaning**
  - Can learn **shared** physical laws across various robotic datasets
  - No normalization
    - "0.1" in position mean +0.1m for any robots, **aligning the physics standard**



Pre-Training and Fine-Tuning

- Pre-Training
  - Thanks to the unified space,
  - **46** datasets of various robots, **1M+** demos, **21TB**
- Fine-Tuning
  - To enhance bimanual capability
  - We collect a high-quality dataset,
    - **Quantity**: 6K+ demos, one of the largest
    - **Comprehensiveness**: 300+ challenging tasks
    - **Diversity**: 100+ objects, 15+ rooms with different lighting conditions

Experiments

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

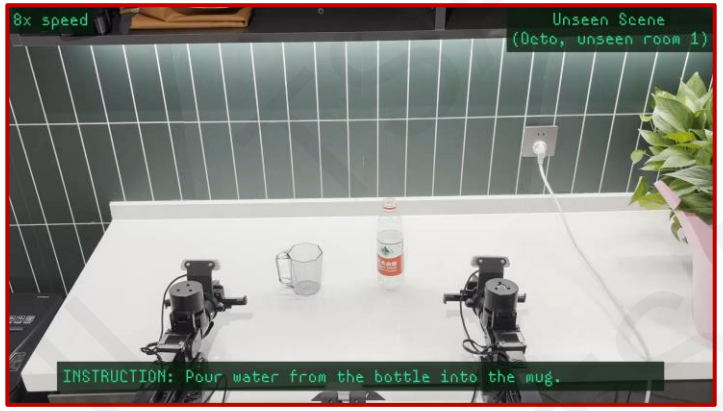
All with 3x speed

Experiments

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

ROBOTICS DIFFUSION TRANSFORMER-1B
PLEASE ENTER YOUR INSTRUCTION NOW

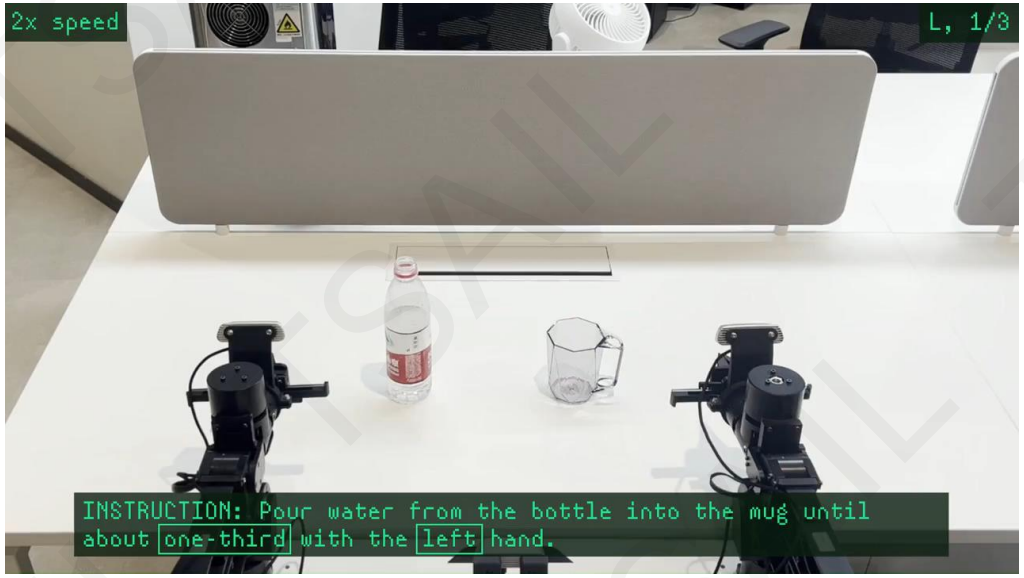
INSTRUCTION:

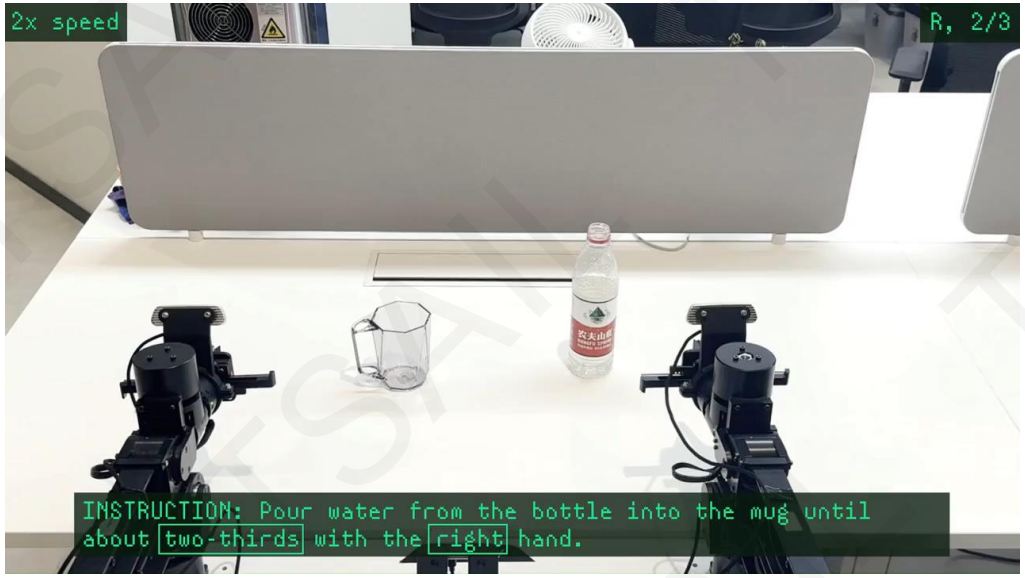


Octo (❌)

Experiments

- Q2: How effective is RDT's **zero-shot** instruction-following capability for **unseen modalities**?
  - "one-third" and "two-thirds" are **unseen** during training
    - The model has only seen "little", "half", and "full" in the instruction
  - Ground the **language concepts** to the **height** in the physical world





Experiments

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

5-demo imitation learning (4x speed)

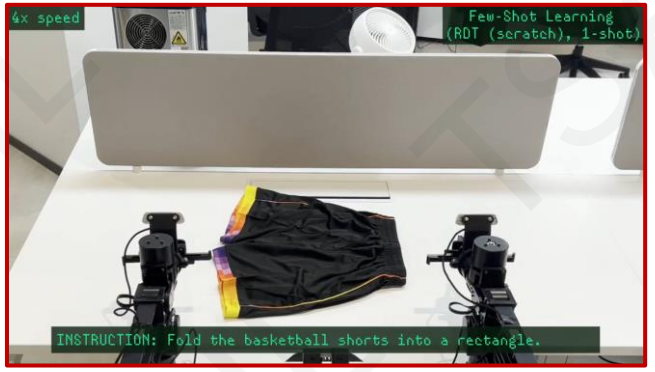
1-demo imitation learning (2x speed)

Experiments

- Q3: Can RDT facilitate **few-shot learning** for previously **unseen skills**?
  - It is really challenging for baselines…

ACT (❌)

OpenVLA (❌)



RDT (scratch) (❌)

Experiments

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

ROBOTICS DIFFUSION TRANSFORMER-1B
PLEASE ENTER YOUR INSTRUCTION NOW

INSTRUCTION:

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.

[23] 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.

Thank You!

Website: 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