Deep Demonstration Tracing: Learning Generalizable Imitator Policy for Runtime Imitation from a Single Demonstration

Pre: Xiong-Hui Chen
LAMDA Group, Nanjing University
Superviosr: Yang Yu

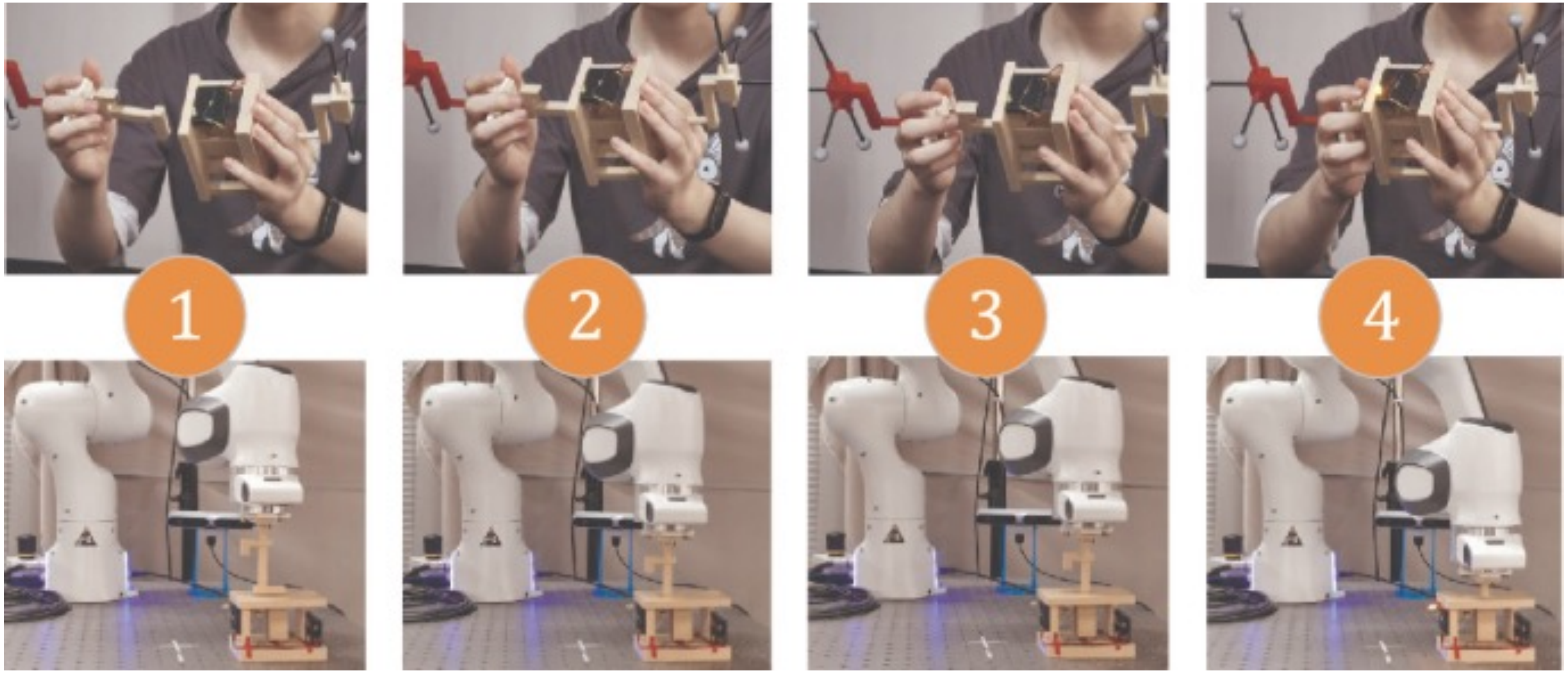
南京大學
NANJING UNIVERSITY

LAMDA
Learning And Mining from DatA

Deep Demonstration Tracing: Generalizable Imitator Policy Learning

1. **Background:**
2. Methodology
3. Experiment
4. Take-home Messages

Vision of Runtime One-Shot Imitation Learning / Learning from a Single Demonstration



Runtime imitator policy: $\Pi(a|s,\tau)$, where $\tau \in \mathrm{T}$ is a unseen human demonstration.

τ like a "prompt" for the imitator policy, guiding the agent achive the tasks as expected

A popular Paradigm: transformer with behavior cloning

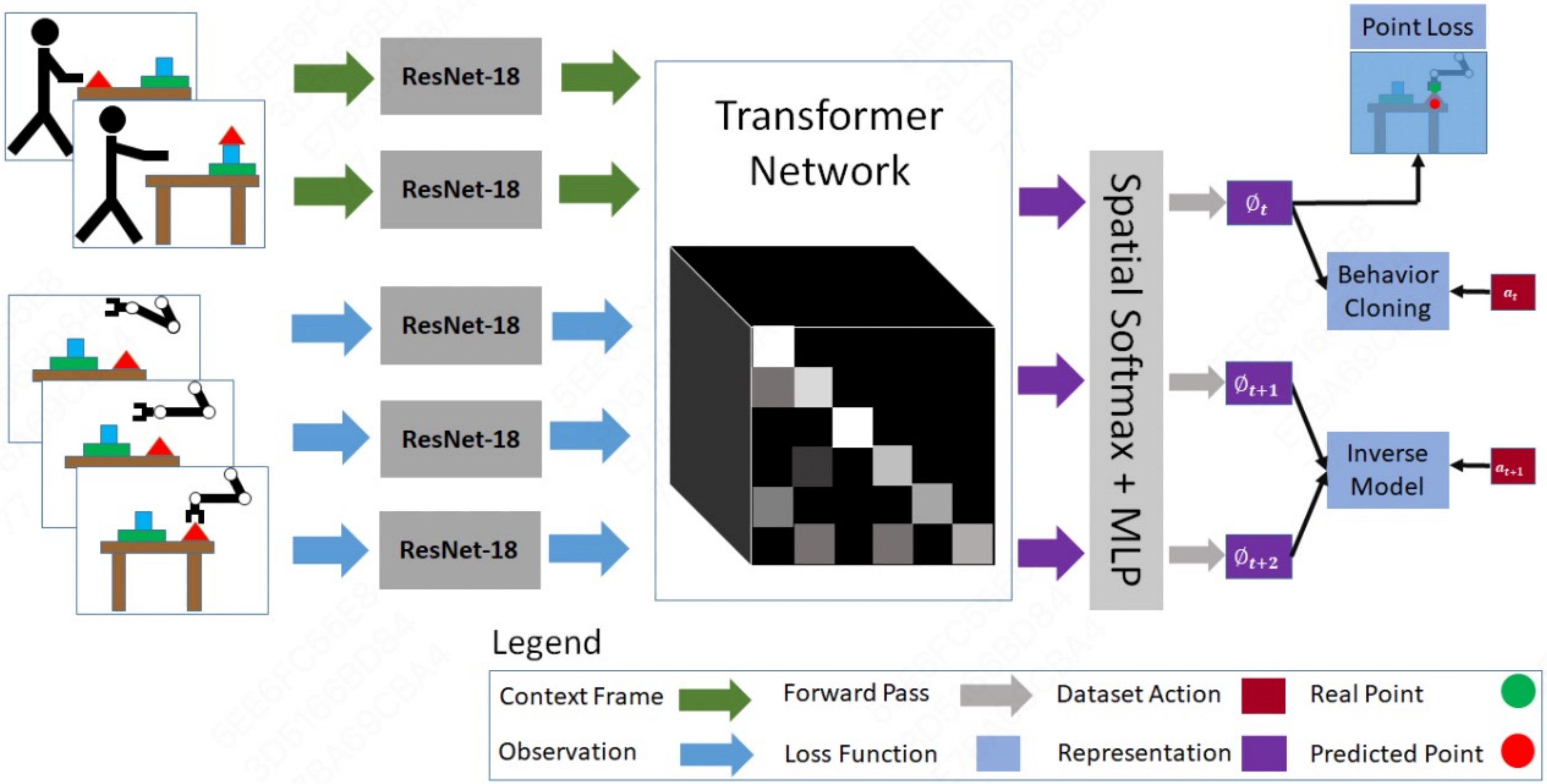
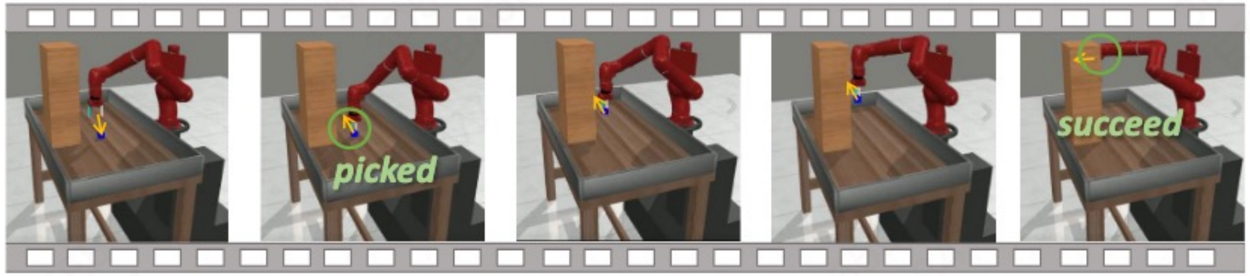
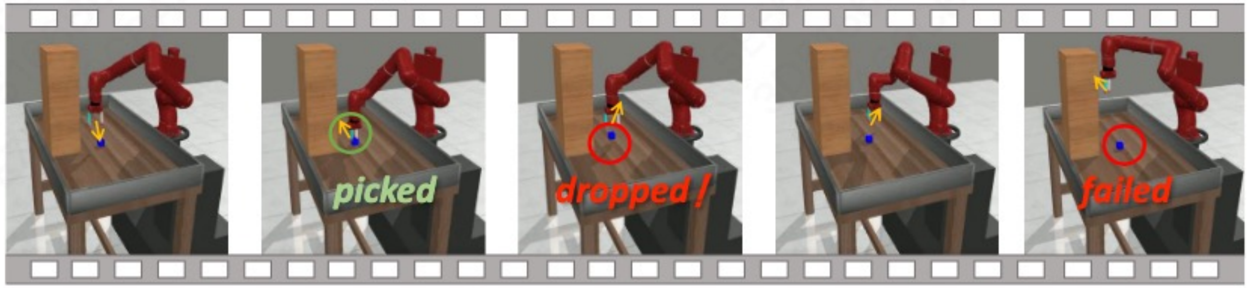


Figure 2: Our method uses a Transformer neural network to create task-specific representations, given context and observation features computed with ResNet-18 (w/ added positional encoding). The attention network is trained end-to-end with a behavior cloning loss, an inverse modelling loss, and an optional point loss supervising the robot's future pixel location in the image.

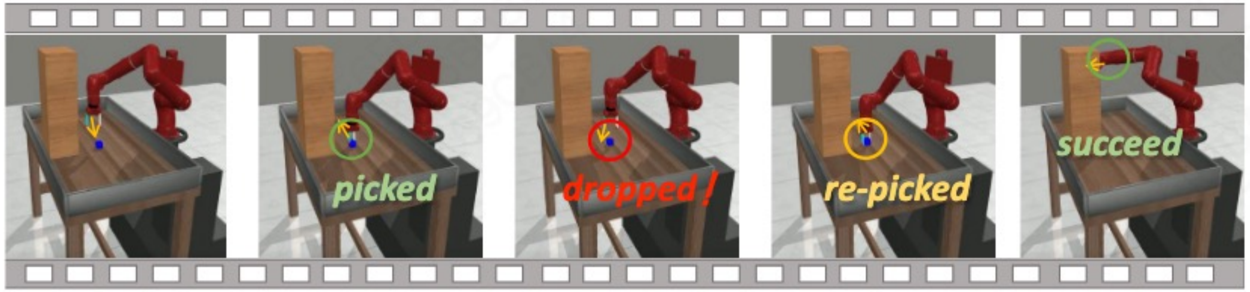
Generlization Challenge of Runtime One-Shot Imitation Learning (OSIL)



(a) Provided demonstration.



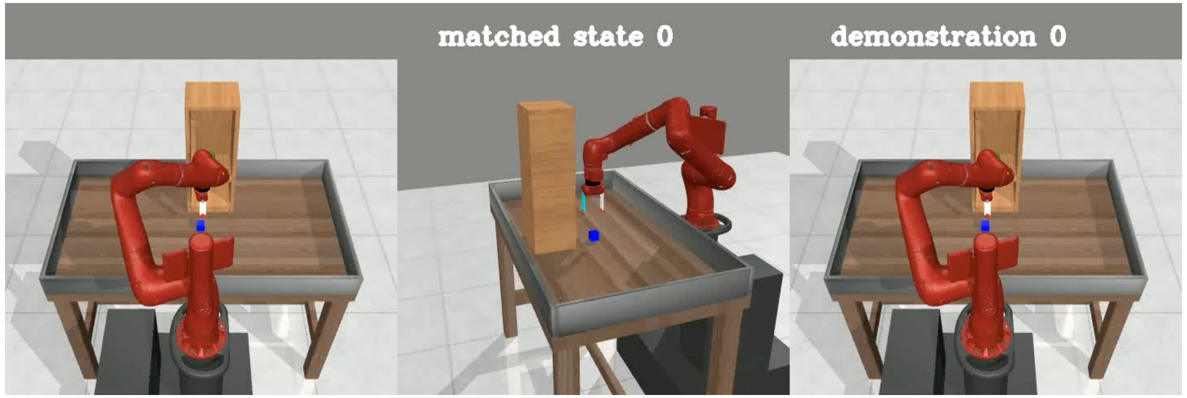
(b) Policy trained by a traditional OSIL method.



(c) Policy trained by DDT.

Gap: Poor Generlization ability in unseen situations.

- unseen demonstrations (transformer)
- emergency eventa unseen when providing the demonstration (bc)



Deep Demonstration Tracing: Generalizable Imitator Policy Learning

1. Background

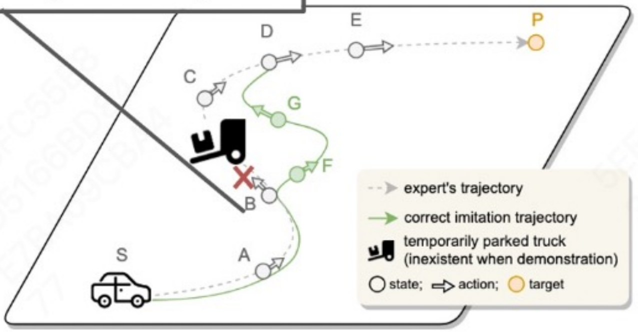
2. **Methodology**
   1. Demonstration transformer
   2. OSIL via meta-RL

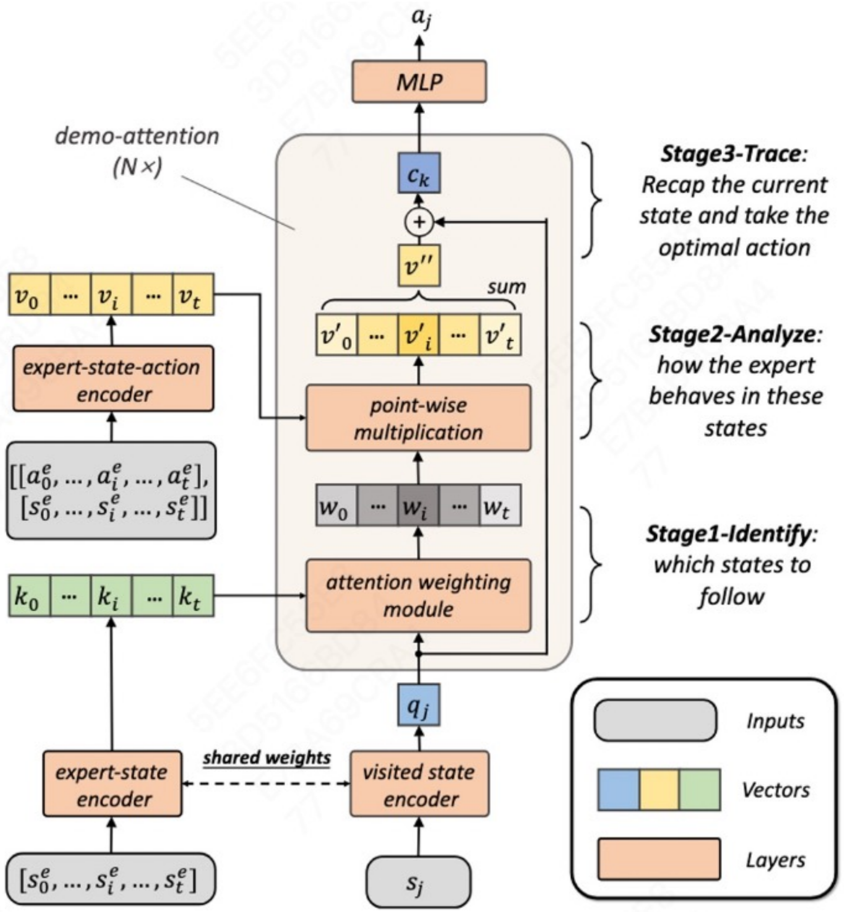
3. Experiment

4. Take-home Messages

Inject the inductive bias of "how human make decisions in runtime OSIL" into the imitator policy network.

- **Stage 1**: Identify relevant states within the trajectory based on the current state. For example, for the state at point B, the related states can be B, C, and D.
- **Stage 2**: Analyze the expert's behavior patterns associated with these states. For example, a human would see that the expert drives forward from B, navigating a turn, to reach D and E.
- **Stage 3**: Trace the expert's demonstrations based on the relationship between the current state and the expert's behavior patterns in the demonstrations. For example, from point S to A, since the agent's state is close to the expert's, it tends to repeat the expert's actions; while in point B, since the observation is different from the demonstrations, the policy should use its common sense to avoid obstacles and traceback to the successor states (like the sequence B-F-G).





Insight:

introduce current inductive bias has proven that is powerful in improving the generalization ability of a neurl network.

-> introduce inductive bias of the 3-stage demonstration tracing principle for imitator policy learning.

南京大學 NANJING UNIVERSITY

LAMDA Learning And Mining from DatA

Solve runtime one-shot imitation learning by context-based meta-RL, instead of supervised learning

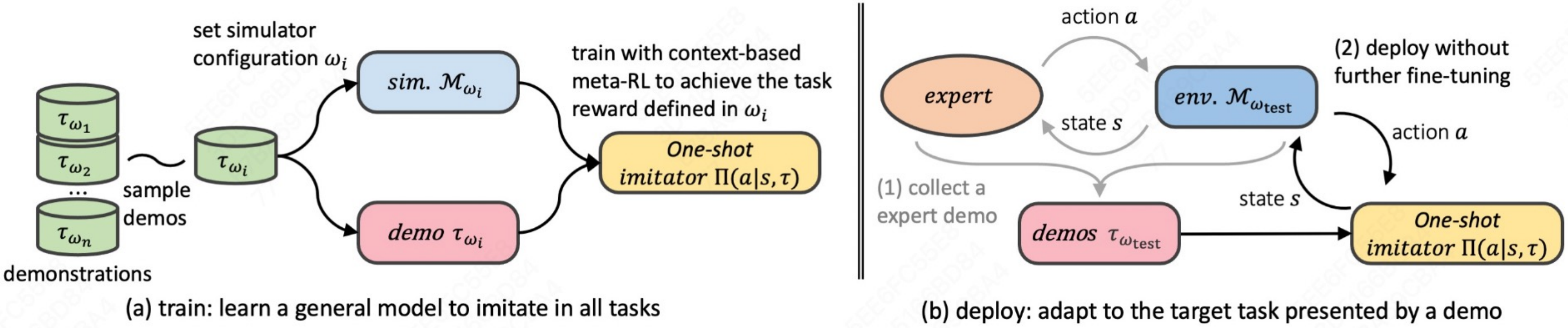


Illustration of the Training and Deploying Workflow for a Runtime One-shot imitator policy via context-based meta-RL.

- The unforeseen changes will randomly apprear in the simulators ($\mathcal{M}$).
- With meta-RL, the imitator policy will try to achieve *all of* the targets the same to the demonstration guided by 0-1 task rewards.
- In the process, the imitator policy will suffer from the unforseen changes and *have to handle them before achieve the targets.*

Deep Demonstration Tracing: Generalizable Imitator Policy Learning

1. Background
2. Methodology
3. **Experiment**
4. Take-home Messages

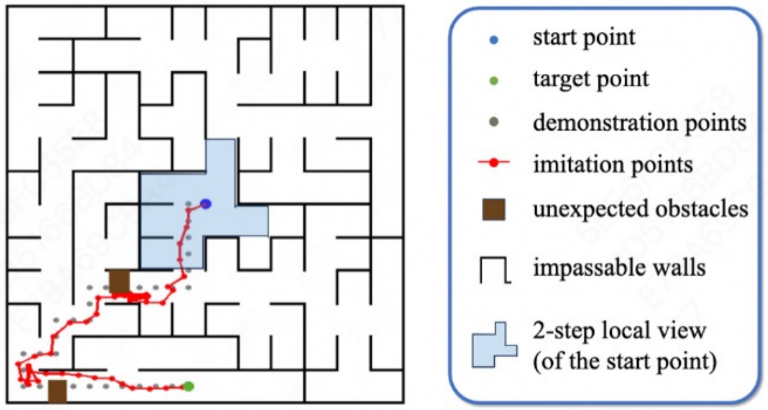
NANJING UNIVERSITY

Learning And Mining from DatA

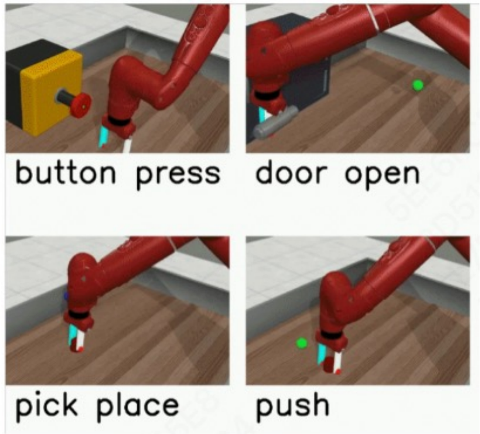
Research questions

1. **RQ1**: The one-shot imitation ability of DDT in unseen situations, including **unseen demonstrations, unseen environments, and unforeseen changes** after demonstration collection.
2. **RQ2**: Does demonstration transformer really imitating via tracing the demonstration?
3. **RQ3**: Can DDT have potential of performance improvement when scaling up the size of parameters and demonstration data, inspired by the **"Scaling Law"** in large language models.

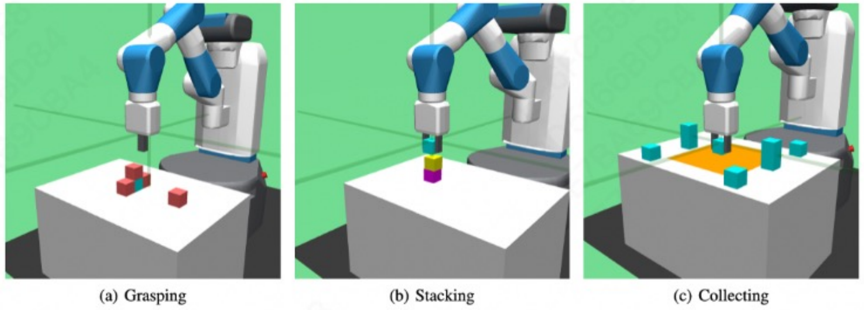
Experiment: Valet Parking Assist in Maze



(A) Valet Parking Assist in Maze (VPAM)



(B) Meta-World



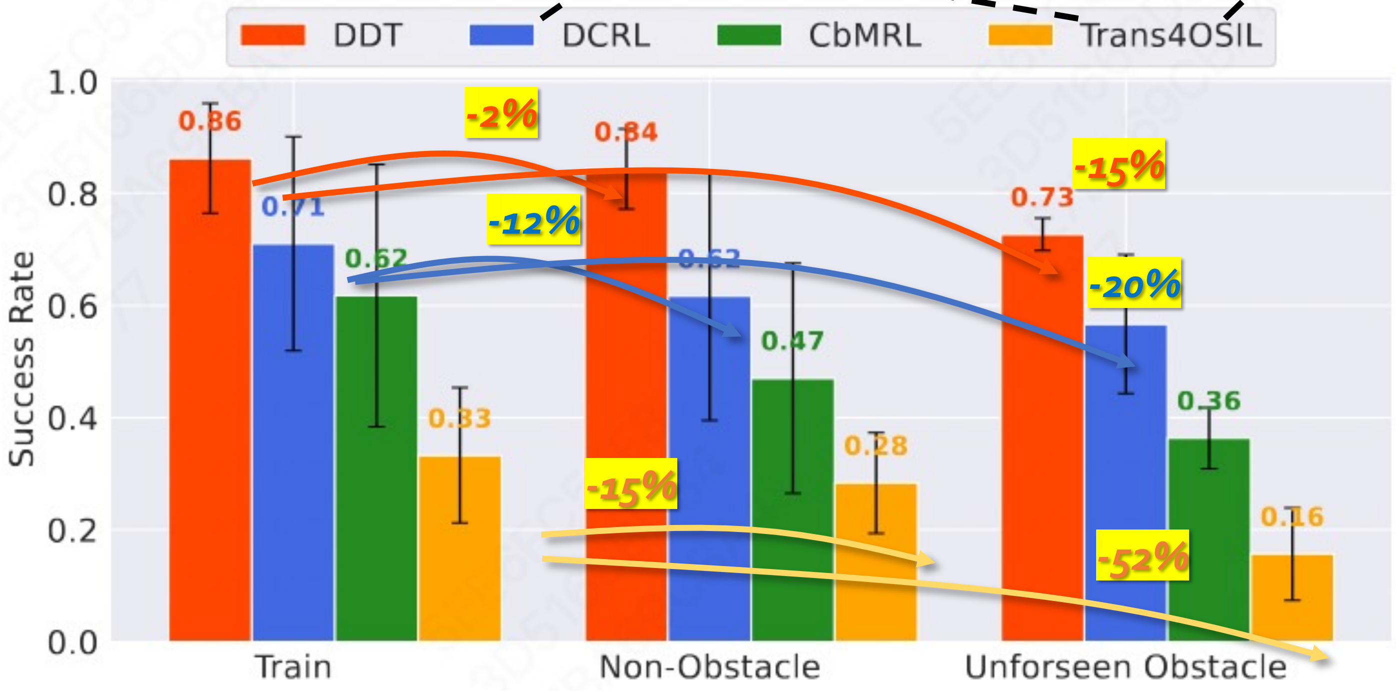
(C) Complex Planning Tasks of Robot Manipulation

Illustration of Major Experiments in this paper. (A) Illustration of the VPAM, which is a new benchmark for OSIL with unforseen changes. The imitation points are provided by our DDT method. (B) Illustration of tasks in Meta-Wolrd. (C) Various Complex tasks of robot manipulation in clutter environments. (a): Grasp the blocked target object (cyan). (b): Stack the objects. (c): Collect the objects scattered over the desk together to the specified area (yellow).

RQ1: One-Shot Imitation Ability in Unseen Situations

Standard Transformer architecture

Behavior cloning



Test :

Test :

Group results averged by 8 settings with 3 seeds (VPAM env).

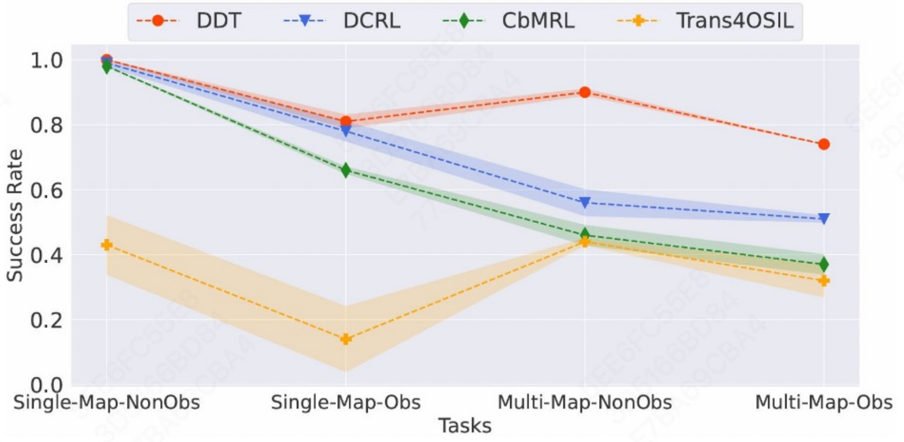


Illustration of the imitation policies' training performance among different settings. The colored areas denote the standard error among the three seeds. DDT displayed a **stable and better performance even in the training tasks**. We attribute this to the integration of the demonstration transformer architecture. This architecture conferred an additional training efficiency boost by implicitly introducing prior knowledge of how OSIL was achieved, facilitating easier adaptation across various tasks and settings with different complexities.

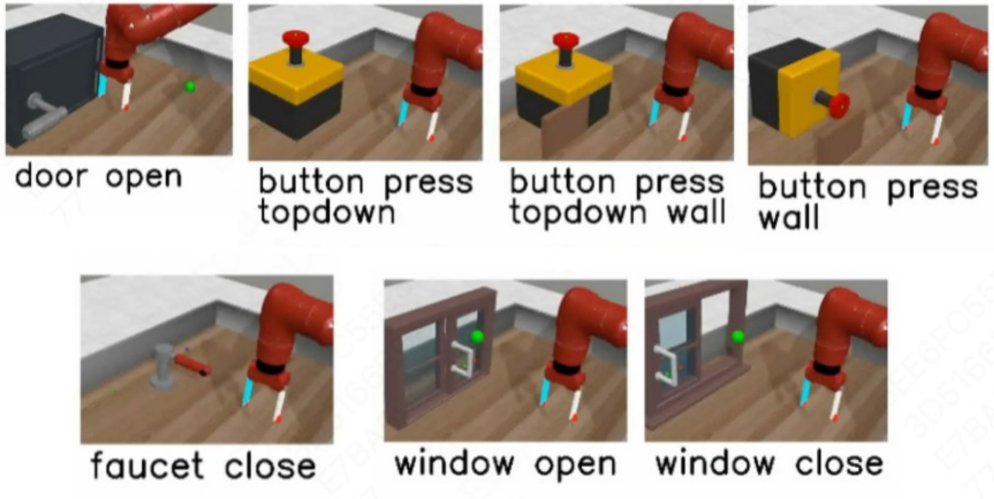
better generlization ability in unseen situations

consistent training performance among different tasks

南京大學 NANJING UNIVERSITY

LAMDA Learning And Mining from DatA

RQ1: One-Shot Imitation Ability in Unseen Situations



Training tasks (reach 100% success rate)

Runtime Imitation

Table 4: Performance on unseen heterogeneous demonstrations.

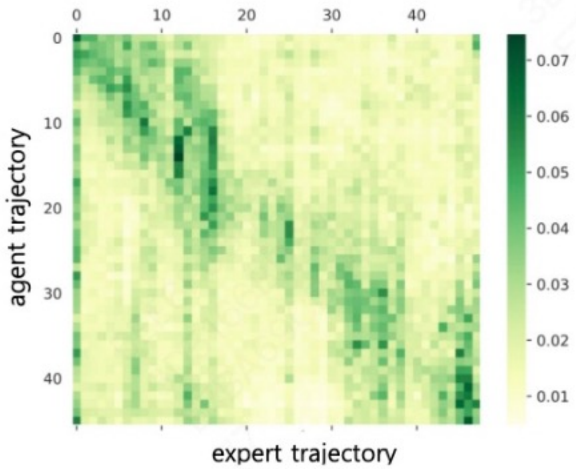
Environment	Button Press	Door Close	Reach
Performance	0.78	1.00	0.75

Test tasks

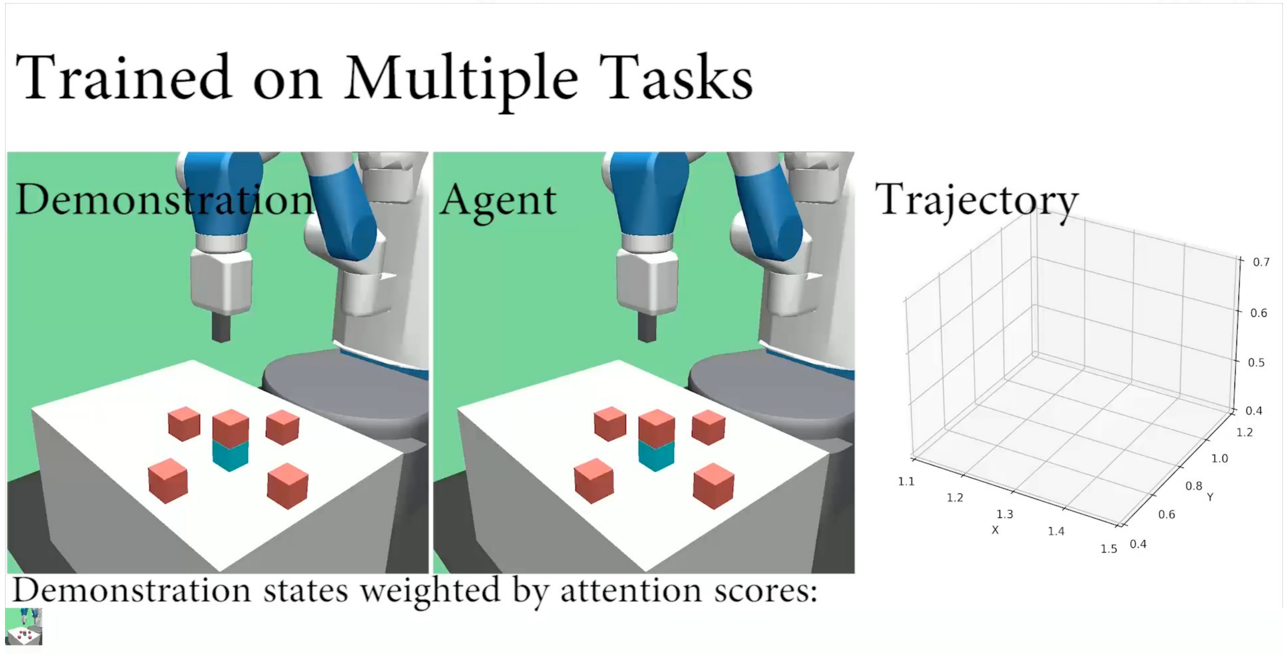
We test and record the generalization performance on three types of unseen heterogeneous demonstrations with all positions of goals without fine-tuning.

RQ2: Demonstration-Attention Mechanism for Demonstration Tracing in DDT

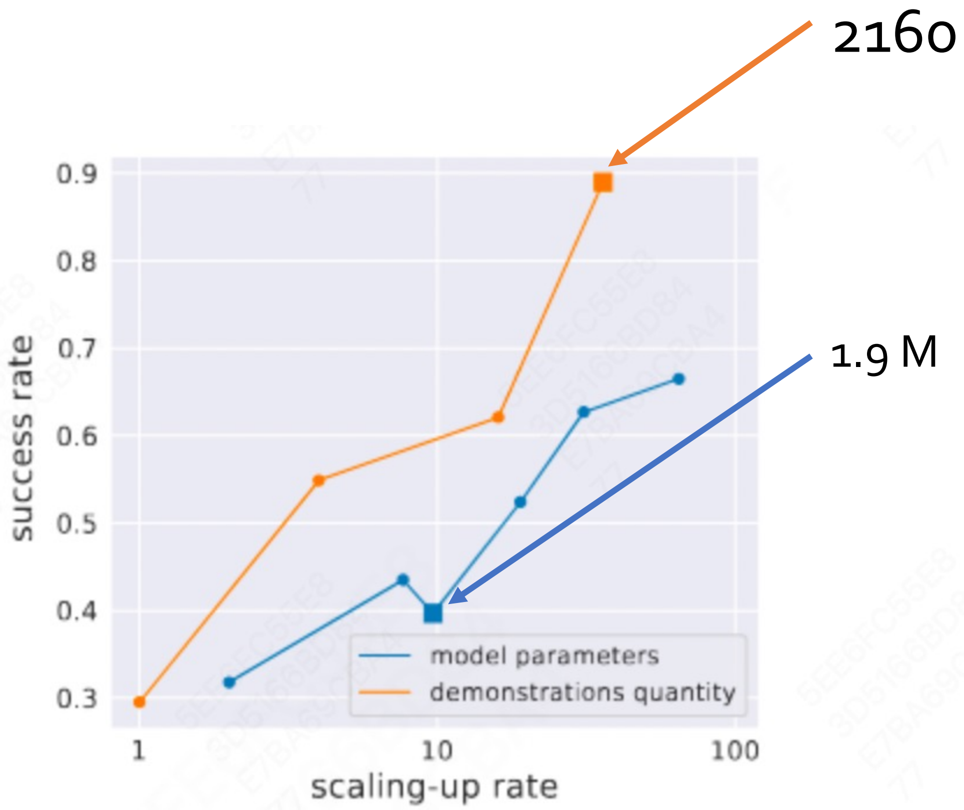
(a) Trajectory of DDT.



(b) Attention score.



RQ3: Similar Scaling Law of DDT when Scaling Up in the OSIL Setting

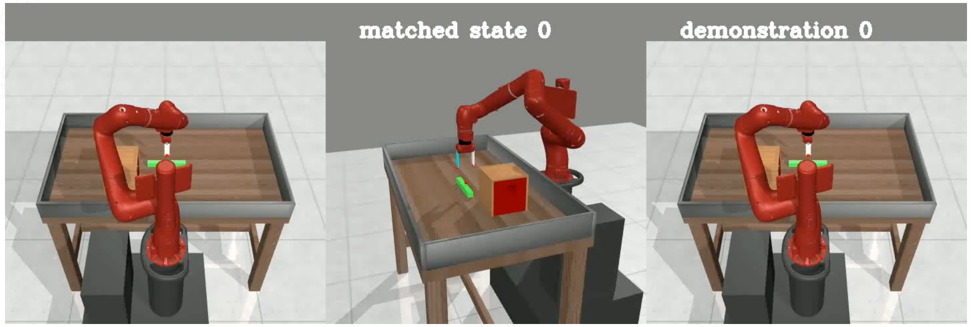


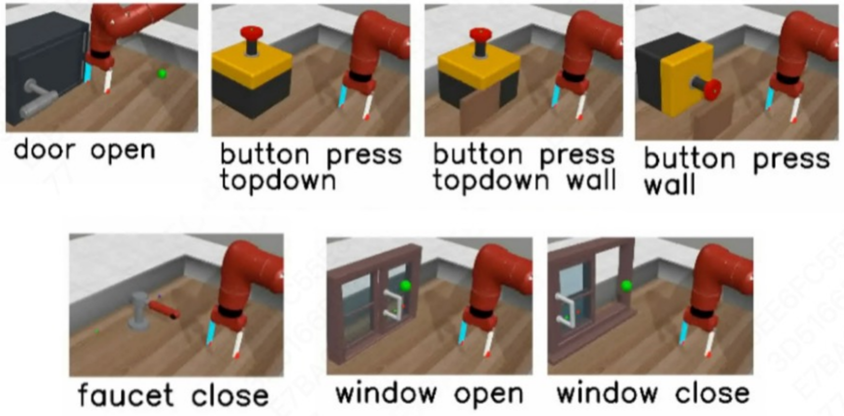
Asymptotic performance of DDT under varying demonstration quantities and model parameters, with each unit on the x-axis representing 60 demonstrations or 0.6 million parameters. The x-axis is on a logarithmic scale. **Square markers** depict the performance of the default DDT parameters.

RQ4: Apply DDT in Other Challenging Tasks

Results on MetaWorld under Disturbance. The video at the start of this project page are rendered from the results in this experiment.

Task		Shelf Place		Peg Insert Side		Pick Place Hole		Sweep	
Demonstration		seen	unseen	seen	unseen	seen	unseen	seen	unseen
No Disturbance	DDT	**1.00**	**0.94**	**1.00**	**0.62**	**1.00**	**0.84**	**1.00**	**1.00**
	DCRL	0.97	0.76	1.00	0.28	0.00	0.00	0.95	0.44
	Trans4OSIL	0.02	0.04	0.04	0.08	0.00	0.00	0.50	0.32
	CbMRL	0.78	0.44	0.84	0.16	0.16	0.00	1.00	0.94
With Disturbance	DDT	**0.79**	**0.44**	**0.92**	**0.70**	**1.00**	**0.90**	**0.61**	**0.40**
	DCRL	0.75	0.34	0.07	0.02	0.00	0.00	0.10	0.10
	Trans4OSIL	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	CbMRL	0.18	0.14	0.02	0.14	0.00	0.00	0.12	0.10





Runtime Imitation

Table 4: Performance on unseen heterogeneous demonstrations.

Environment	Button Press	Door Close	Reach
Performance	0.78	1.00	0.75

Training tasks (reach 100% success rate)

Test tasks

We test and record the generalization performance on three types of unseen heterogeneous demonstrations with all positions of goals without fine-tuning.

Deep Demonstration Tracing: Generalizable Imitator Policy Learning

1. Background
2. Solution
3. Methodology
4. **Take-home Messages**

Take-home Messages

Considering generlized but more specific agents:

1. Transformer might be far from the optimal solution.
    - We still have large room by desigining correct inductive bias.
2. Next token prediction might be far from the optimal solution.
    - Interactive training is a potential way to get a more generalizable result.

>> Thanks