MimicPlay: Long-Horizon Imitation Learning by Watching Human Play

CoRL 2023 ✅

Oral presentation ⭐

Finalist - Best Systems Paper Award ⭐⭐

Finalist - Best Paper/Best Student Paper Awards ⭐⭐⭐

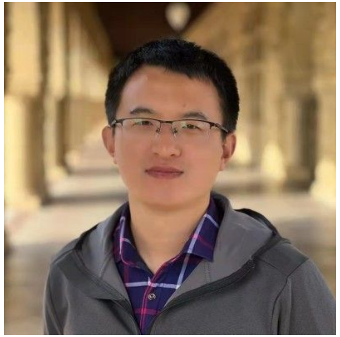
Project: mimic-play.github.io

Jishnu P
Reading Group | IRVL
12/1/23

Authors

Chen Wang
Stanford

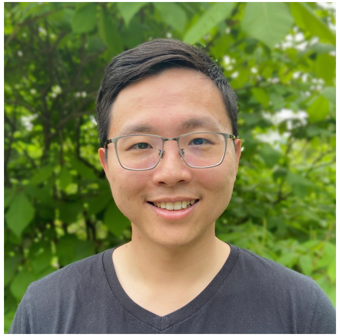
Linxi "Jim" Fan
NVIDIA



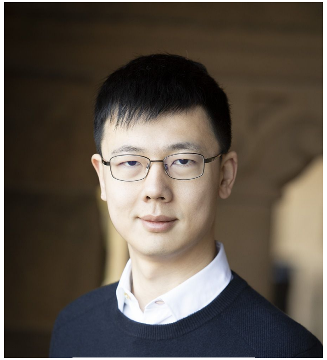
Jiankai Sun
Stanford

Ruohan Zhang
Stanford

Li Fei-Fei
Stanford



Danfei Xu
NVIDIA, Georgia



Yuke Zhu[†]
NVIDIA, UT

Anima Anandkumar[†]
NVIDIA, Caltech

[†]Equal Advising

Coming back to the Title

Long-Horizon Imitation Learning by Watching Human Play

Hierarchical Learning Framework

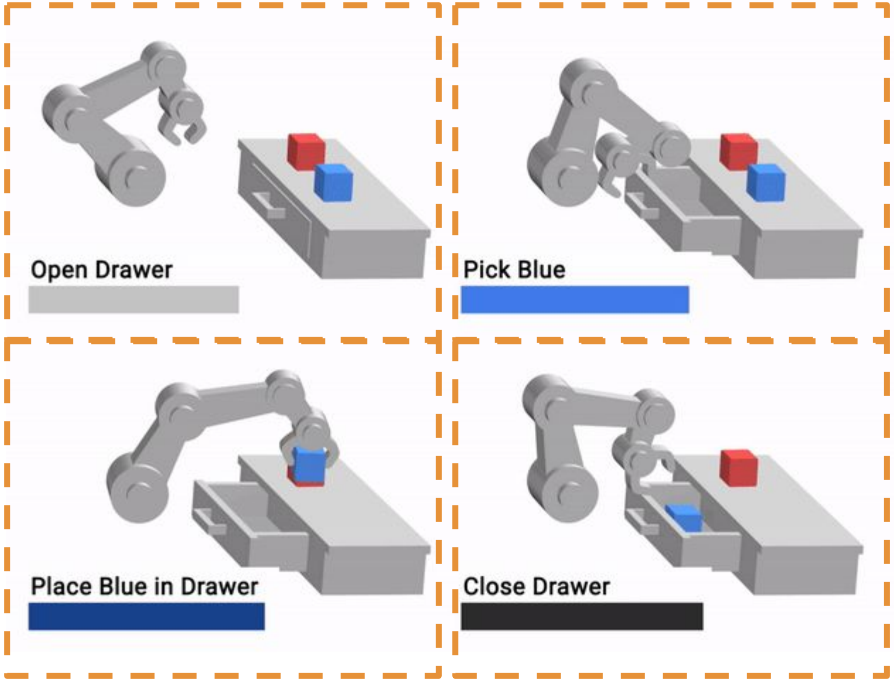
1. High level planner
2. Low level visuomotor controllers

Dataset

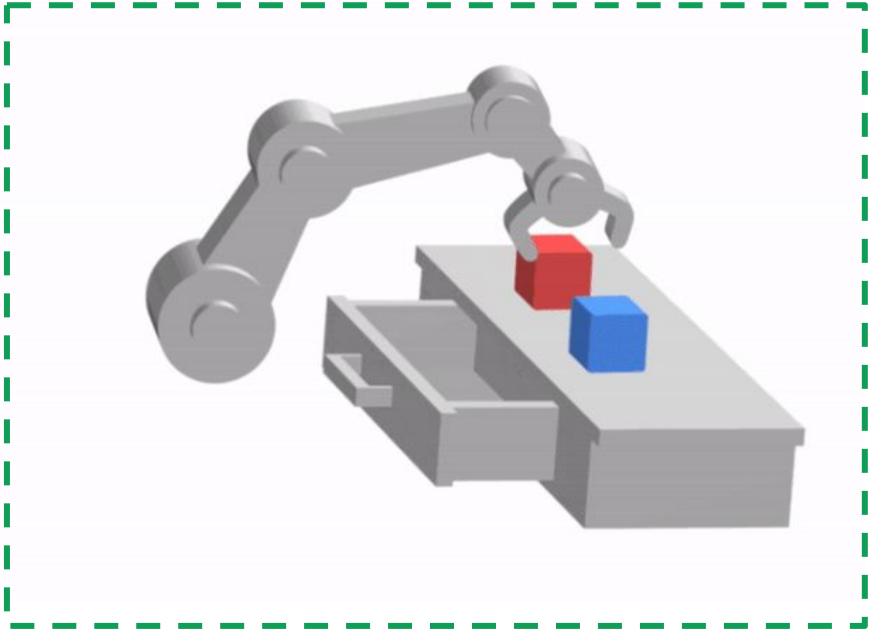
Play data

MimicPlay

Long vs Short horizon tasks



"Move blue to the drawer"



Credits: https://blog.research.google/2022/04/extracting-skill-centric-state.html

Motivation

- **Imitation learning** has shown **promising results** in performing **general purpose manipulation tasks**.
- **Issue:** **It is confined to short-horizon primitives**
  - **Opening a door**
  - **Picking a specific object**
- **Why?**
  - **For long horizon tasks, demonstrations for complex real world are cost and labor intensive.**
- **Existing literature to the rescue**

Connect two directions in literature

Hierarchical Imitation Learning

(decouples the end-to-end deep imitation learning)

1. High level planner: intent dist.
2. Low level visuomotor controllers (goal directed control)

Dataset

Play data

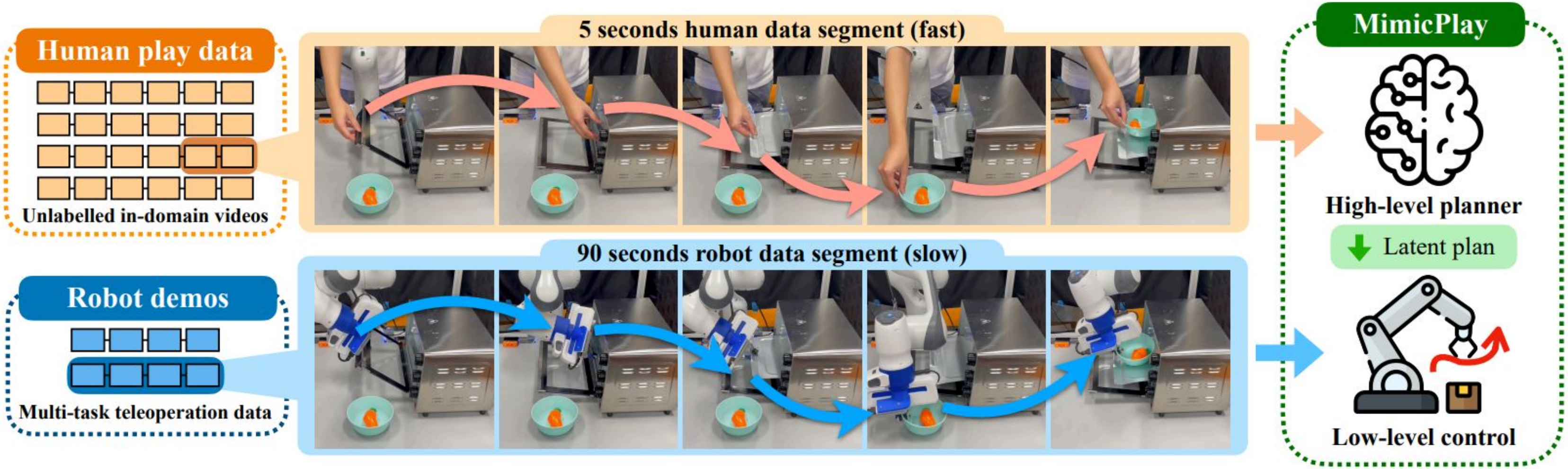
(human teleoperated data interacting with their environment without specific task goals or guidance.)

Issue: However to obtain such data is time costly w.r.t. to the model requirements. C-BeT: 4.5 hours, TACO-RL: 5 hours.

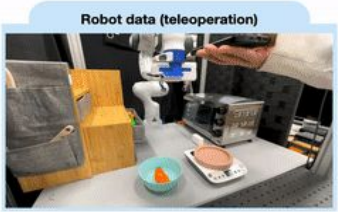
Method + new Human Play Dataset

MimicPlay

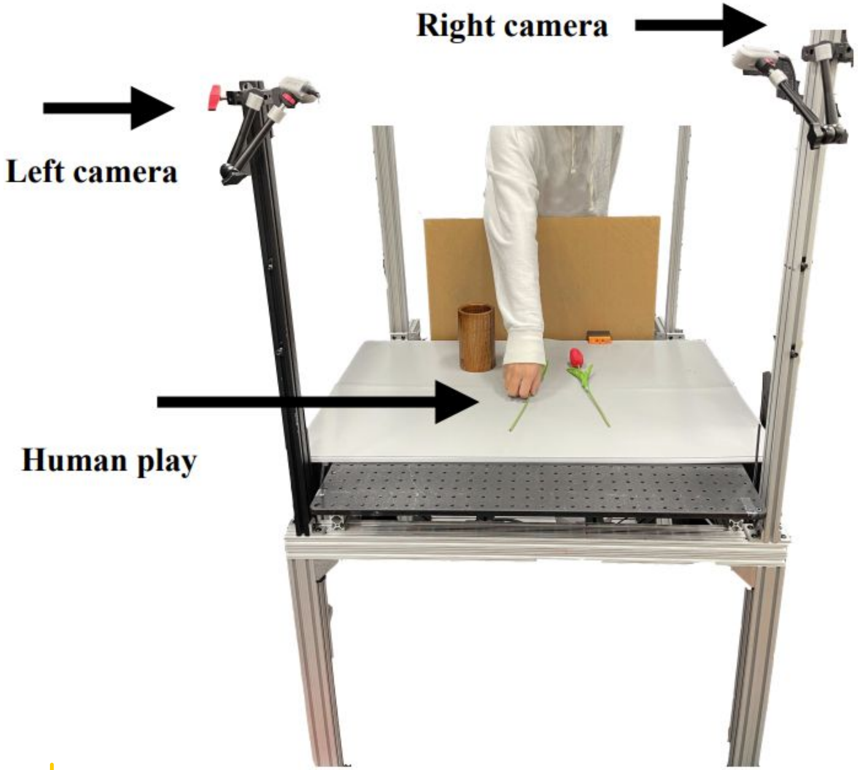
Motivation


Human play data
Unlabelled in-domain videos
5 seconds human data segment (fast)
Robot demos
Multi-task teleoperation data
90 seconds robot data segment (slow)
MimicPlay
High-level planner
Latent plan
Low-level control


Human play data

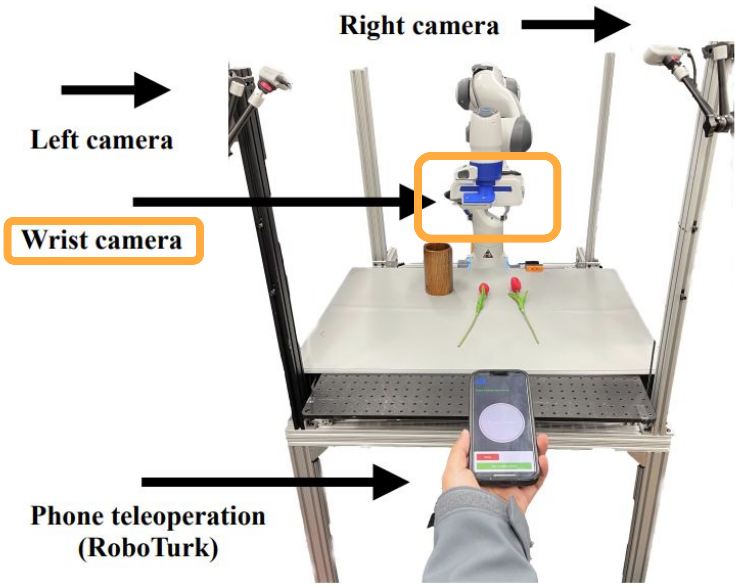

Robot data (teleoperation)

Human Play Dataset



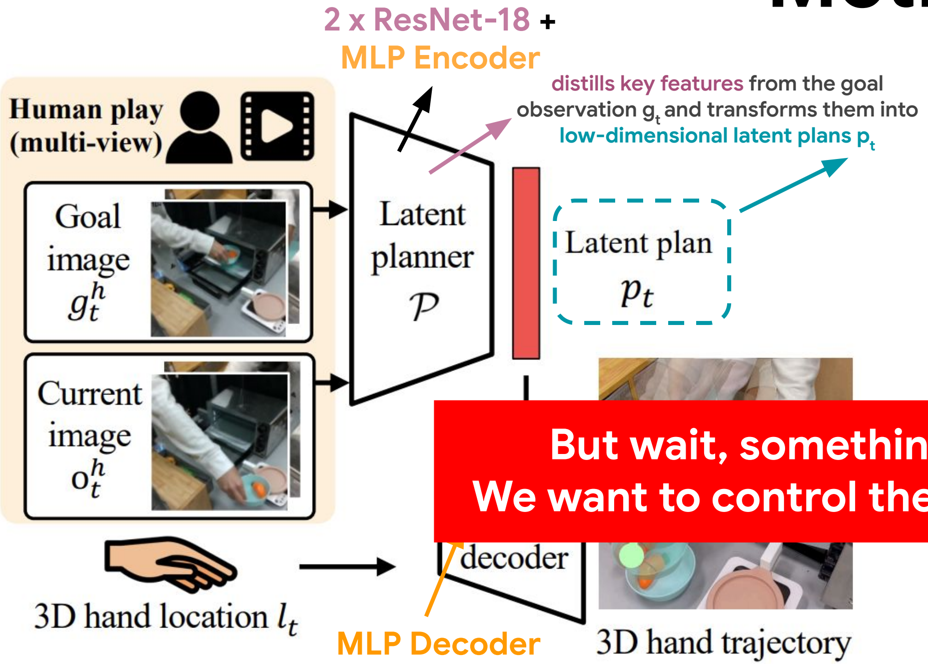
- A human operator directly interacts with the scene with one of the hand and perform interesting behaviors based on curiosity without a specific task goal.
  - Why? - Trajectories contain rich information regarding the individual’s underlying intentions -> better planning
- The left and right cameras record the video at the speed of 100 fps.
  - Common human video datasets comprise single-view observations, providing only 2D hand trajectories. Such trajectories present ambiguities along the depth axis and suffer from occlusions.
  - Two calibrated camera setup to track 3D hand trajectories from human play data.
  - Off-the-shelf hand detector to identify hand locations, reconstructing a 3D hand trajectory based on the calibrated camera parameters
- The human demonstrator completes the assigned sub-goals one by one and finally solves the whole task. 🎯
- For each scene, 10 minutes of human play data is collected. (36K frames)
- The entire trajectory τ is recorded at the speed of 60 fps and is used without cutting or labeling

Robot Demo (Teleoperation)



- A human demonstrator uses a phone teleoperation system (RoboTurk) to control the 6 DoF robot end-effector.
- The gripper of the robot is controlled by pressing a button on the phone interface.
- For each training task, 20 demonstrations are collected.
- The left, right, and end effector wrist cameras record the video at the speed of 20 fps, which is aligned with the control speed of the robot arm (20Hz).
- Each sequence of robot demonstration has a pre-defined task goal 🎯

Method



Once the latent plan is obtained, latent plan p_t + hand location l_t is fed to an MLP-based decoder network -> generates the prediction of the 3D hand trajectory.

However, simple regression of the trajectory cannot fully cover the rich multimodal distribution of human motions. Even for the same human operator, one task goal can be achieved with different strategies.

To address this issue, an MLP-based Gaussian Mixture Model (GMM) is used to model the trajectory distribution from the latent

But wait, something is missing!!!
We want to control the robot, don't we?

Parameters of the GMM

$\sum_{z} \ldots z)p(z|\theta)$ $\quad \theta = \{\mu_k, \sigma_k, \eta_k\}_{k=1}^{K}$

$p(\tau|\theta, z_k)$ is a Gaussian distribution $\mathcal{N}(\tau|\mu_k, \sigma_k)$

Training Stage-1: Learning latent plans

How: Use cheap and sufficiently large human play data to train a goal-conditioned trajectory generation model to build a latent plan space that contains high-level guidance for diverse task goals

❌ paired human-robot video data requirement

$$\mathcal{L}_{\mathrm{KL}} = D_{\mathrm{KL}}(\mathcal{Q}^r || \mathcal{Q}^h)$$

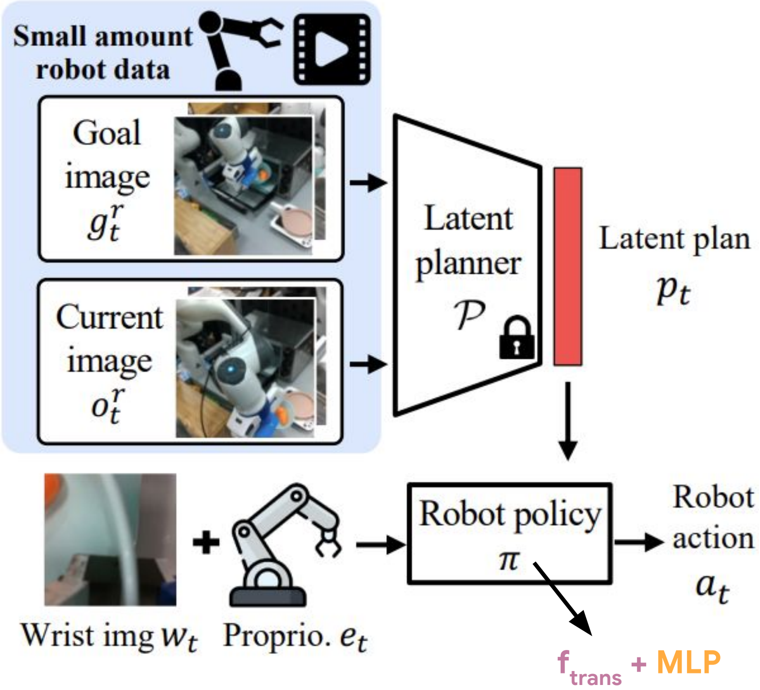
Final learning objective of GMM model is to minimize the negative log-likelihood of the detected 3D human hand trajectory τ. 100k iterations.

$$K = 5$$

$$\mathcal{L}_{\mathrm{GMM}}(\theta) = -\mathbb{E}_{\tau} \log\left(\sum_{k=1}^{K} \eta_k \mathcal{N}(\tau|\mu_k, \sigma_k)\right), \text{where } 0 \leq \eta_k \leq 1, \sum_{k=1}^{K} \eta_k = 1$$

$$\mathcal{L} = \mathcal{L}_{\mathrm{GMM}} + \lambda \cdot \mathcal{L}_{\mathrm{KL}}$$

Method



Training Stage-2

Use a small amount of teleoperation data to train a low-level robot controller conditioned on the latent plans generated by the pre-trained (frozen) planner.

A transformer based policy π is learned

The latent representation for robot's wrist camera observation w_t and proprioception data e_t alongwith latent plan p_t are combined to create a one-step token embedding: $s_t=[w_t,e_t,p_t]$

The sequence of these embeddings over T time steps, s[t:t+T] , is processed through a transformer architecture f_{trans}

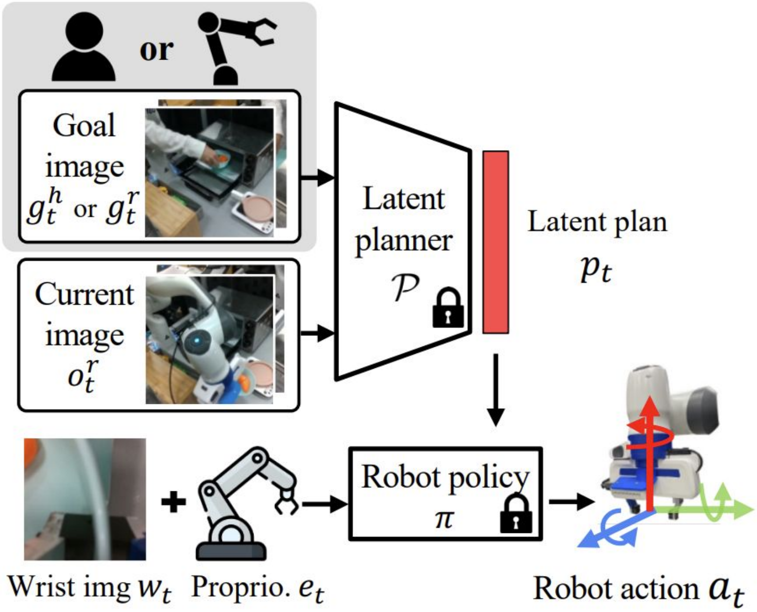
The transformer-based policy, known for its efficacy in managing long-horizon action generation, produces an embedding of action prediction x_t in an autoregressive manner

The final robot control commands at are computed by processing the action feature x_t through *a two-layer fully connected network*

To address the multimodal distribution of robot actions, an MLP-based Gaussian Mixture Model (GMM) is used to for action generation.

100K iterations.

Method



Testing stage

Given a single long-horizon task video prompt (either human motion video or robot teleoperation video), MimicPlay generates latent plans and guides the low-level controller to accomplish the task.

Input: A one-shot video V (either human video V^h or robot video V^r) as a goal specification prompt

V sent to the pre-trained latent planner to generate robot-executable latent plans p_t. (How?)

The one-shot video V is first converted into a sequence of image frames.

At each time step, the high-level planner P takes one image from the sequence as a goal-image input g_t and generates a latent plan p_t to guide the generation of low-level robot action a_t.

After executing a_t, the next image frame in the sequence is used as a new goal image.

During the training, the goal image g^r_t ($g^r_t \in V^r$) is specified as the frame H steps after the current time step in the demonstration.

H is a uniformly sampled integer number within the range of [200,600] (10-30 seconds), which can act as a data augmentation process.

Experiments: Environments & Tasks

3 individual tasks including cooking food with oven

7 individual tasks including tidying up the desk

6 environments with 14 tasks featuring tasks such as contact rich tool use, articulated-object handling, and deformable object manipulation

(c) Flower

(d) Whiteboard

(e) Sandwich

(f) Cloth

Flower insertion into a vase

Erasing curve lines.

Ingredient selection for cheeseburger or sandwich.

Folding a towel twice

Results (Success Rate)

	Subgoal (first subgoal)								Long horizon ($\geq$ 3 subgoals)							
	20 demos				40 demos				20 demos				40 demos			
	Task-1	Task-2	Task-3	ALL	Task-1	Task-2	Task-3	ALL	Task-1	Task-2	Task-3	ALL	Task-1	Task-2	Task-3	ALL
GC-BC (BC-RNN) [20]	0.1	0.0	0.1	0.07	0.1	0.2	0.2	0.17	0.0	0.0	0.0	0.00	0.0	0.0	0.1	0.03
GC-BC (BC-trans) [52]	0.2	0.0	0.0	0.07	0.3	0.7	0.6	0.53	0.0	0.0	0.0	0.00	0.0	0.0	0.1	0.03
C-BeT [6]	0.5	0.6	0.0	0.37	0.4	**1.0**	0.0	0.47	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.00
LMP [5]	0.3	0.1	0.2	0.20	0.6	0.3	0.2	0.37	0.1	0.0	0.1	0.07	0.3	0.1	0.0	0.13
R3M-BC [40]	0.9	0.0	0.0	0.30	0.5	0.4	0.0	0.30	0.0	0.0	0.0	0.00	0.5	0.0	0.0	0.17
Ours (0% human)	**1.0**	0.5	0.3	0.60	**1.0**	0.5	0.5	0.67	0.3	0.1	0.3	0.23	0.4	0.3	0.5	0.40
Ours	**1.0**	**0.8**	**0.7**	**0.83**	**1.0**	0.9	**0.8**	**0.90**	**0.7**	**0.3**	**0.4**	**0.47**	**0.7**	**0.6**	**0.8**	**0.70**

Table 1: Quantitative evaluation results in the Kitchen environment.

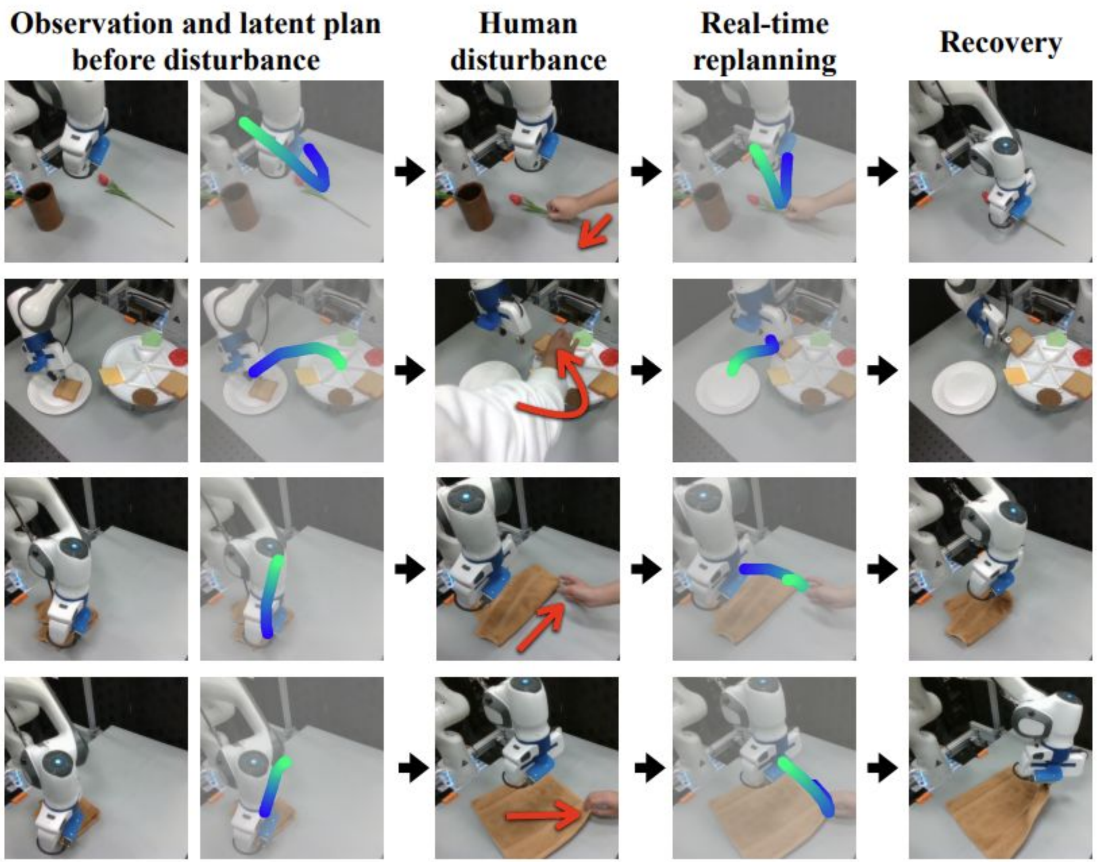
	Trained tasks					Unseen tasks			
	Task-1	Task-2	Task-3	Task-4	ALL	Easy	Medium	Hard	ALL
GC-BC (BC-trans) [52]	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.00
LMP [5]	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.00
Ours (0% human)	0.2	0.3	0.1	0.2	0.20	0.2	0.1	0.0	0.10
Ours (50% human)	0.3	0.4	0.1	0.4	0.30	0.4	0.3	0.1	0.27
Ours (w/o KL)	0.3	**0.7**	0.3	0.2	0.38	0.4	0.2	0.0	0.20
Ours (w/o GMM)	0.4	0.2	0.2	0.3	0.28	0.2	0.0	0.0	0.07
Ours	**0.6**	**0.7**	**0.4**	**0.5**	**0.55**	**0.7**	**0.5**	**0.2**	**0.47**

Table 2: Ablation evaluation results in the Study Desk environment (20 demos).

	Spatial generalization		Extreme long horizon	Deformable	
	Flower	Whiteboard	Sandwich	Cloth	ALL
LMP-single	0.1	0.0	0.1	0.3	0.13
LMP [5]	0.0	0.0	0.0	0.2	0.05
R3M-single	0.2	0.1	0.3	0.4	0.25
R3M [40]	0.1	0.1	0.2	0.2	0.15
Ours-single	**0.5**	**0.5**	0.6	0.7	**0.58**
Ours	0.4	0.2	**0.8**	**0.8**	0.55

Table 3: Quantitative evaluation results of multi-task learning.

Results Cont.



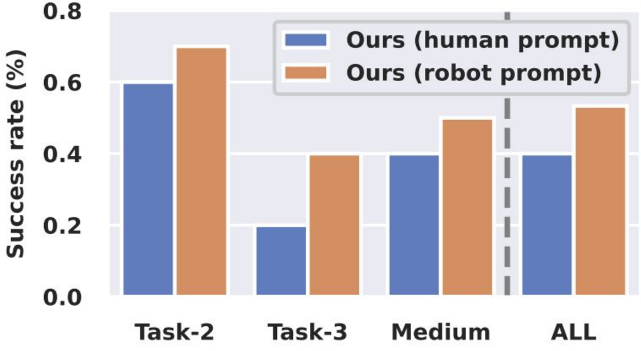
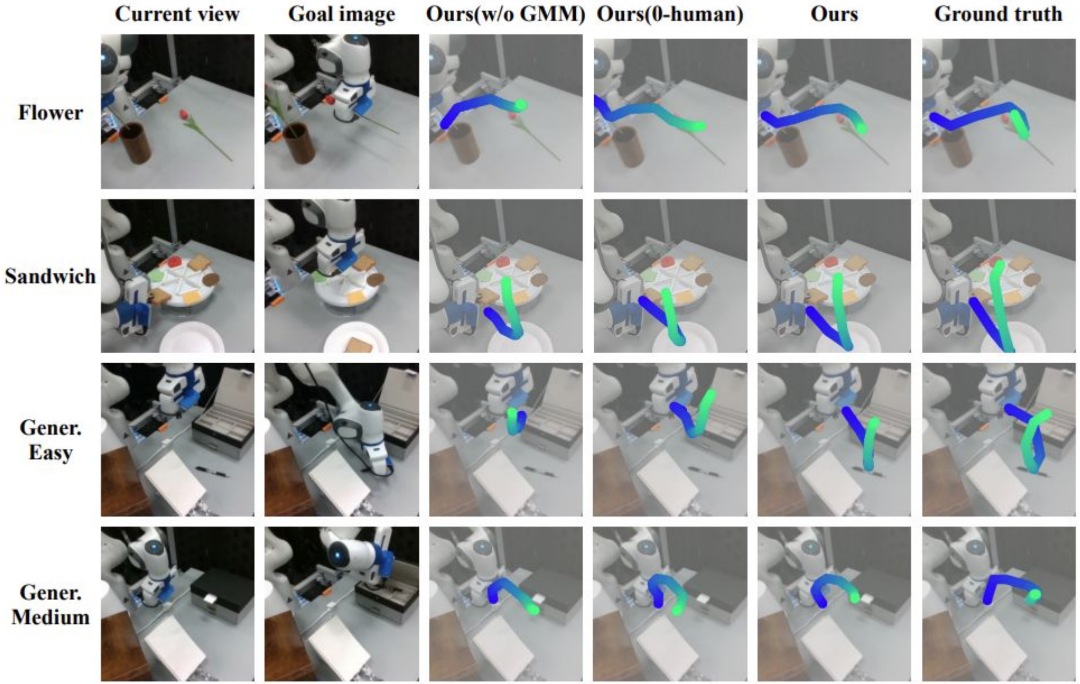
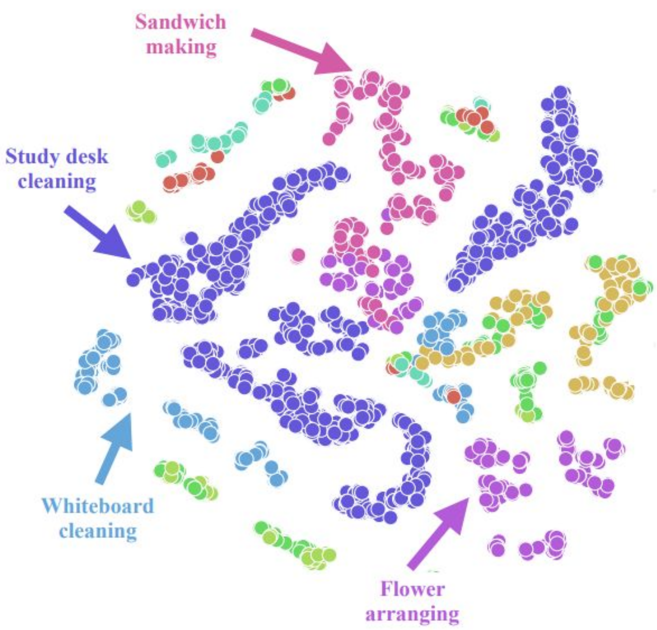


Figure 4: Evaluation of multi-task policy prompted with robot/human videos in the Study Desk environment.

Results Cont.

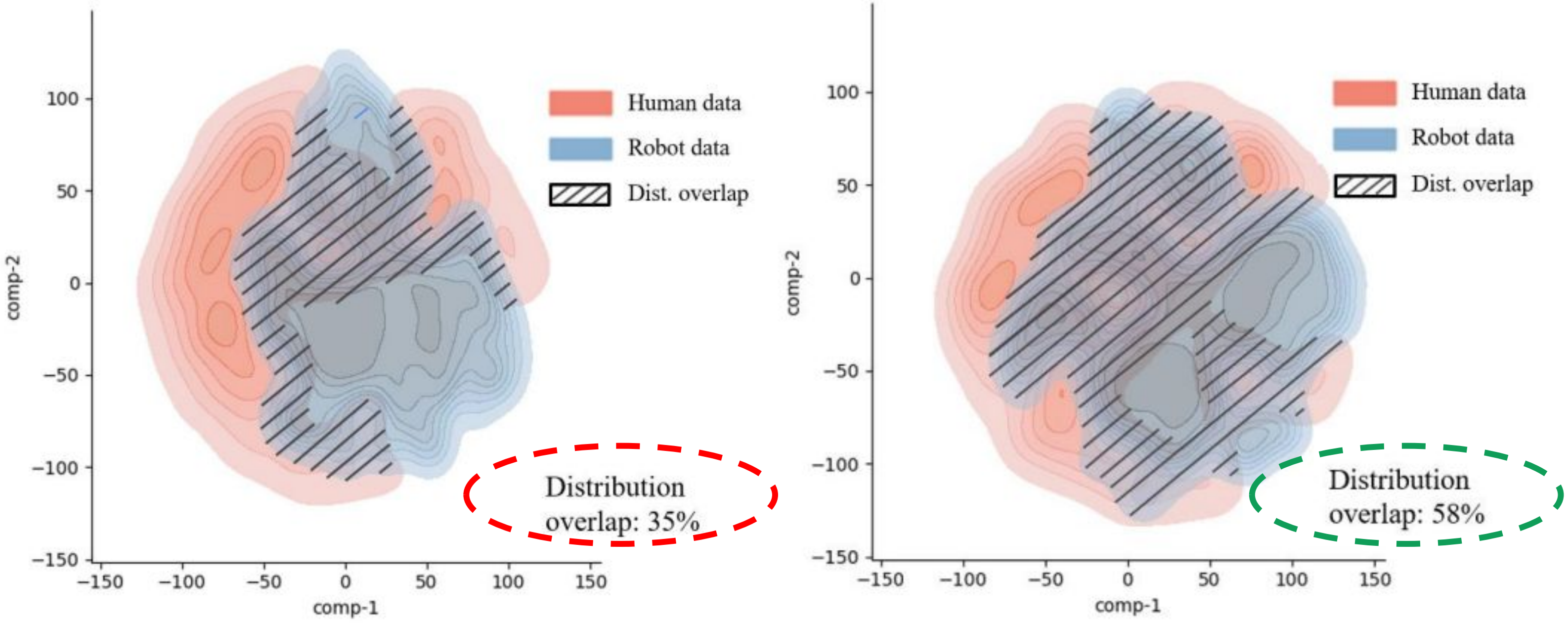


(a) Trajectory prediction results decoded from the latent plans



(b) t-SNE visualization of the latent plans

Results Cont.



(a) Distribution overlap of Ours (w/o KL) **(b) Distribution overlap of Ours**

t-SNE visualization of the generated feature embeddings by taking human data and robot data as inputs.
The slashes refer to the overlap region of two data distributions.

Limitations

The current high-level latent plan is learned from scene-specific human play data. The scalability of MimicPlay can greatly benefit from training on Internet-scale data.

The current tasks are limited to table-top settings. However, humans are mobile and their navigation behaviors contain rich high-level planning information. The current work can be extended to more challenging mobile manipulation tasks.

There is plenty of room to improve on the cross-embodiment representation learning. Potential future directions include temporal contrastive learning and cycle consistency learning from videos.

Takeaways

Learning latent plans from human play data significantly improves performance.

Hierarchical policy is important for learning long-horizon tasks.

Latent plan pre-training benefits multi-task learning.

GMM is crucial for learning latent plans from human play data.

KL loss helps minimize the visual gap between human and robot data.

Questions?