2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

MIMICDROID: In-Context Learning for Humanoid Robot Manipulation from Human Play Videos

Rutav Shah[1] Shuijing Liu*,[1] Qi Wang*,[1] Zhenyu Jiang*,[1] Sateesh Kumar[1] Mingyo Seo[1]
Roberto Martín-Martín[1,2] Yuke Zhu[1,3]
[1]The University of Texas at Austin [2]Amazon Consumer Robotics [3]NVIDIA

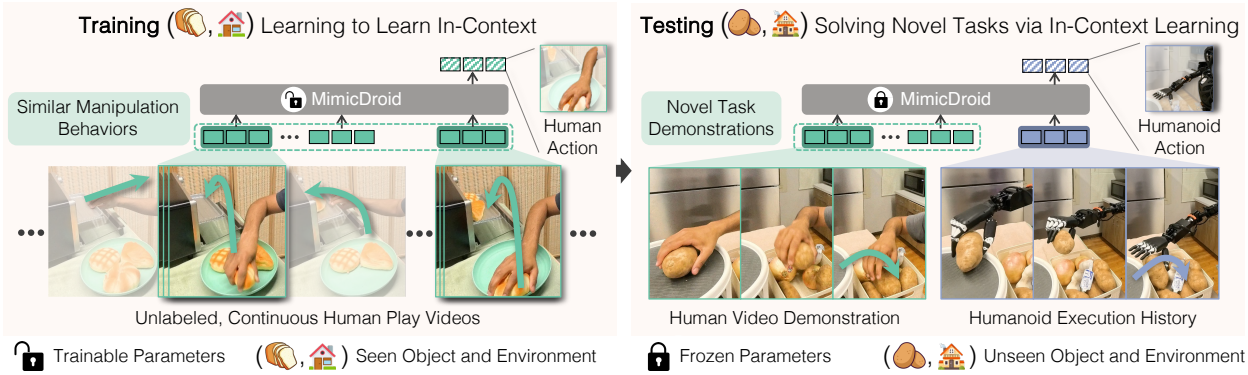

Fig. 1: **Overview.** MIMICDROID enables few-shot learning for humanoid manipulation by training solely on human play videos—a scalable and diverse data source. At test time, it observes human videos of novel tasks and uses in-context learning to perform the same tasks under different object placements.

***Abstract*— We aim to enable humanoid robots to efficiently solve new manipulation tasks from a few video examples. In-context learning (ICL) is a promising framework for achieving this goal due to its test-time data efficiency and rapid adaptability. However, current ICL methods rely on labor-intensive teleoperated data for training, which restricts scalability. We propose using human play videos—continuous, unlabeled videos of people interacting freely with their environment—as a scalable and diverse training data source. We introduce MIMICDROID, which enables humanoids to perform ICL using human play videos as the only training data. MIMICDROID extracts trajectory pairs with similar manipulation behaviors and trains the policy to predict the actions of one trajectory conditioned on the other. Through this process, the model acquired ICL capabilities for adapting to novel objects and environments at test time. To bridge the embodiment gap, MIMICDROID first retargets human wrist poses estimated from RGB videos to the humanoid, leveraging kinematic similarity. It also applies random patch masking during training to reduce overfitting to human-specific cues and improve robustness to visual differences. To evaluate few-shot learning for humanoids, we introduce an open-source simulation benchmark with increasing levels of generalization difficulty. MIMICDROID outperformed state-of-the-art methods and achieved a nearly twofold higher success rate in the real world. Additional materials can be found on: `ut-austin-rpl.github.io/MimicDroid`**

I. INTRODUCTION

Humanoid robots are well-suited for diverse household manipulation tasks due to their human-like morphology. Yet homes exhibit large variability: objects, layouts, and tasks change across time and households. To effectively cope with such variability, humanoids must move beyond pre-defined sets of behaviors and adapt rapidly to novel situations. For instance, when a new appliance is installed, the robot should be able to acquire the necessary skills to manipulate it from a handful of demonstrations, a problem setting known as few-shot learning [1–3].

In-context learning (ICL) has shown promise for few-shot learning, offering data-efficient and rapid adaptation at test time [4–10]. By simply conditioning on a few human demonstrations, ICL can predict robot actions to achieve novel tasks at test time without expensive retraining (Fig. 1, Right). However, effective ICL for few-shot learning relies on large and diverse training data [11–13]. In manipulation, prior ICL methods rely on teleoperated robot demonstrations as training data [9, 14], which are expensive and time-consuming to scale [15–17]. This limitation motivates us to explore scalable training data sources and methods that can leverage such data to enable ICL for humanoid manipulation.

A promising alternative training data source is human play videos—continuous recordings of people interacting with their environments, typically spanning 10–20 minutes of interaction and driven by their curiosity. Human play videos capture task-agnostic, unscripted, and exploratory interactions with environments [18, 19]. Compared to teleoperated demonstrations, they are approximately $18\times$ faster to collect [18] and inherently diverse, covering a broad range of tasks, object configurations, and manipulation behaviors [19]. Leveraging these advantages, we explore scalable and diverse human play videos to train ICL policies that perform few-shot learning for humanoid manipulation.

To realize the potential of human play videos for ICL, two key challenges must be addressed. First, the model should learn tasks from in-context examples, an ability that meta-training can instill through *learning to learn* in-context [4, 20]. However, this requires constructing meta-

training samples from raw human play videos in a scalable, self-supervised way. Second, the kinematic and visual gap between human and robot embodiments presents challenges for applying ICL at test time. The kinematic gap must be bridged so actions learned from human videos can be executed on the robot without losing task intent. The visual gap must be addressed to avoid overfitting to human appearances, which can hinder its ability to apply ICL on robots.

To address these challenges, we develop MIMICDROID (a **Mimic**king an**Droid**), a novel method to perform few-shot humanoid manipulation via ICL using only RGB human play videos for training. At test time, MIMICDROID is provided with 1–3 videos of a human performing a task, potentially involving novel objects and environments, and applies ICL to mimic and perform the task (Fig. 1). Effective ICL relies on exploiting recurring patterns in observation-action relationships, i.e., how visual observations correspond to the actions that follow, enabling the model to predict actions in new scenes. To instill this capability, we generate meta-training samples by pairing trajectory segments with similar patterns, treating one as the target and the others as proxies for test-time demonstrations. This construction encourages the policy to exploit similarities in observation-action relationships across trajectories to predict actions. Human play videos are well-suited for this construction, as they naturally exhibit recurring patterns of similar manipulation behaviors, like "moving *bread* from plate to oven" and later "moving *bagel* from plate to oven," which MIMICDROID retrieves to construct meta-training samples in a self-supervised manner. To bridge the kinematic gap, MIMICDROID retargets predicted human wrist poses to humanoid wrist poses at test time. By operating in task space (Cartesian wrist pose) and exploiting kinematic similarities between embodiments, it preserves the underlying task intent [21–23]. To mitigate the visual gap, it applies random patch masking during training [24, 25], reducing overfitting to human-specific appearances.

We evaluate MIMICDROID in both simulation and real-world settings. We build a new simulation benchmark for evaluating few-shot learning in humanoid manipulation spanning three generalization levels (Sec. IV). By exploiting observation-action relationships in human videos through ICL, MIMICDROID outperforms task-conditioned baselines [26, 27], achieving a twofold improvement in real-world success rate. Compared to parameter-efficient fine-tuning [28], ICL adapts instantaneously and achieves a 26% higher success rate at test time. We show that MIMICDROID scales effectively with training data, yielding a gain of 20% when increasing training human play videos from 64k to 320k frames. Our main contributions are as follows.

1) MIMICDROID enables few-shot learning for humanoid manipulation via ICL using only human play videos.
2) We introduce a new simulation benchmark [29, 30] with 8 hours of play data to evaluate few-shot learning for humanoids.
3) MIMICDROID outperforms prior works and demonstrates scalability with data, while our analysis highlights current limitations and future directions.

II. RELATED WORK

Few-Shot Learning in Manipulation. Few-shot learning is the ability to learn new tasks from just a few demonstrations, enabling robots to adapt to novel tasks and scenes. Meta-learning [31–34] provides one approach to such adaptation, where the model acquires a learning strategy that allows it to quickly learn new skills at test time with only a few examples. Early approaches in robot manipulation instantiated this idea through gradient-based meta-learning methods [1, 2]. While data-efficient, these methods lack on-the-fly adaptation. Recent progress in ICL, driven by long-context transformers [35] and meta-training on diverse data, allows models to adapt on-the-fly by matching patterns in the input-output examples [4, 5]. Meta-training for in-context learning (Meta-ICL) amplifies this ability by training on structured context-target pairs [20]. In robotics, Meta-ICL has shown preliminary promise in learning visuomotor policies with robot teleoperation data or simulated data [9, 10, 14]. However, since the effectiveness of ICL heavily depends on large and diverse training datasets [11–13], these methods are limited by the high cost of data collection [9] and the narrow diversity of simulated tasks [10].

Learning from Human Videos. Human videos offer a more scalable and diverse data source for robot learning, compared to expensive and time-consuming robot teleoperation data. Prior work has explored various strategies to extract knowledge from human videos, such as learning visual representations [36–38] and deriving reward functions [39–42]. However, these approaches often require additional robot data in addition to human videos to accomplish the downstream task [43]. Another line of work focuses on extracting motion priors directly from human videos [18, 22, 44], but often depends on specialized hardware such as VR or hand-tracking devices, which add overhead and limit scalability. To overcome this, recent works leverage advances in hand pose estimation models [45, 46] to extract action information directly from RGB videos [21, 47, 48]. Among these works, some approaches [47, 48] require human demonstrators to mimic the morphology and joint constraints of specific robot manipulators with unnatural motions. Humanoid robots reduce the necessity for such unnatural motions by leveraging their kinematic similarity to humans, which enables a natural mapping from human motions to robot actions [21, 22].

In summary, to solve few-shot learning for humanoid manipulation, MIMICDROID builds on two pillars: (1) ICL to allow the humanoid to adapt to novel objects and environments on the fly; (2) human play videos, consisting of raw RGB frames, as a scalable and diverse training data source to build ICL foundations by leveraging the success of hand pose estimation and the kinematic similarities between humans and humanoids.

III. METHODOLOGY

In this section, we present an overview (Sec. III-A), components involved in constructing the training data (Sec. III-B), and overcoming the embodiment gap between human

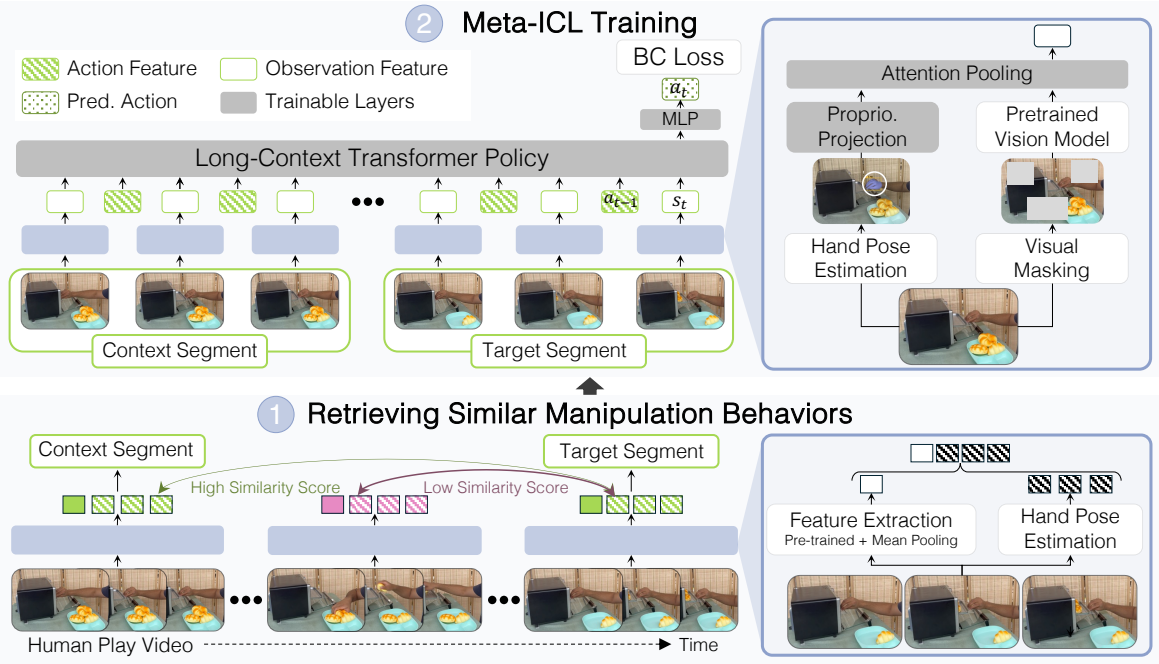

Fig. 2: **Method Overview.** MIMICDROID performs meta-training for in-context learning (Meta-ICL) by constructing context-target pairs from human play videos. For a target segment, we retrieve the top-k most similar trajectory segments (bottom-left) based on observation-action similarity (bottom-right) to serve as context. These context-target pairs are used to teach the policy in-context learning (top-left). To overcome the human-robot visual gap and avoid overfitting to human-specific visual cues, we apply visual masking to input images (top-right), improving transferability.

and robot (Sec. III-C), and finally, meta-training objective (Sec. III-D) to enable test-time ICL (Fig. 2).

A. Overview

Problem Setup. Robots deployed in real-world environments must handle diverse objects, layouts, and environments, making it infeasible to predefine all possible task variations. To enable generalization in such settings, we consider a test-time scenario where the robot is given a small set of human demonstration trajectories for a novel task $\mathcal{T} \sim p(\mathcal{T}_{\text{test}})$: $\mathcal{D}_{\text{test}} = \{\tau_i^{\text{demo}}\}_{i=1}^{k}$, where each $\tau_i^{\text{demo}} = \{s_t\}_{t=1}^{T_i}$ is a sequence of RGB frames of a human performing $\mathcal{T}$. The goal is to learn a visuomotor policy π_θ that can leverage these demonstrations to perform the task.

Training Data. During training, MIMICDROID is provided with a dataset $\mathcal{D}_{\text{train}} = \{\tau_i\}_{i=1}^{N}$ consisting of human play trajectories. Each trajectory τ_i is a sequence of RGB frames from a single play session, where T_i denotes the trajectory length, $\tau_i = \{s_t\}_{t=1}^{T_i}$, typically corresponding to 10–20 minutes of interaction. They span a diverse set of tasks, forming an implicit task distribution $\mathcal{T}_{\text{train}}$.

Policy Learning Using $\mathcal{D}_{\text{train}}$, we aim to train a visuomotor policy π_θ that can perform in-context learning (ICL). We cast this training into the Meta-ICL framework [20], where a task $\mathcal{T}$ is sampled from $p(\mathcal{T}_{\text{train}})$, $\sigma_{\mathcal{T}}^{\text{ctx}}$ denotes a set of *context* trajectories, and $\sigma_{\mathcal{T}}^{\text{tgt}}$ is the *target* trajectory. The context trajectories provide examples of how the task is performed, while the target trajectory supervises the policy's action predictions conditioned on the context. We use behavior cloning to supervise the policy by minimizing the loss $\mathcal{L}_{\mathcal{T}}$ between the predicted and ground-truth actions on $\sigma_{\mathcal{T}}^{\text{tgt}}$. Through this process, π_θ learns to exploit recurring observation-action patterns in context trajectories to predict actions for the target trajectory. Formally, the training objective is

$$\mathbb{E}_{\mathcal{T}\sim p(\mathcal{T}_{\text{train}})}\, \mathbb{E}_{\sigma_{\mathcal{T}}^{\text{ctx}},\, \sigma_{\mathcal{T}}^{\text{tgt}}\sim\mathcal{T}} \left[\mathcal{L}_{\mathcal{T}}\left(\pi_\theta(\sigma_{\mathcal{T}}^{\text{tgt}} \mid \sigma_{\mathcal{T}}^{\text{ctx}})\right)\right], \qquad (1)$$

Unlike traditional meta-learning, which depends on an explicitly defined discrete set of tasks in $p(\mathcal{T}_{\text{train}})$, our setting removes this assumption and instead relies on an implicit task distribution induced by natural human interactions during play. Moreover, play videos lack task labels to sample context-target trajectories for a given task $\sigma_{\mathcal{T}}^{\text{tgt}}, \sigma_{\mathcal{T}}^{\text{ctx}} \sim \mathcal{T}$, and explicit action information to supervise the policy $\mathcal{L}_{\mathcal{T}}$.

Evaluation Protocol. Given $\mathcal{D}_{\text{test}}$, and conditioned on past observations $s_{1:t}$ and past actions $a_{1:t-1}$, the trained policy π_θ predicts the next action a_t to successfully perform the task. The policy uses these demonstrations for in-context learning (ICL), adapting on the fly *without additional parameter updates*. For systematic evaluation, we assess generalization across three test-time task distributions with increasing difficulty introduced via novel objects and environments.

Challenges. To enable effective test-time ICL capabilities, the main challenges include: (a) *Constructing training samples.* Human play videos provide only RGB frames $\tau = \{s_t\}_{t=1}^{T}$, whereas training requires state-action pairs $\tau = \{(s_t, a_t)\}_{t=1}^{T}$ for supervision, along with proprioceptive signals for richer input. Thus, the missing actions and proprioception should be estimated. A subsequent challenge lies

in constructing training samples consisting of context-target pairs from long, continuous human play videos for Meta-ICL (See Eq. 1) (Sec. III-B); (b) *Overcoming the embodiment gap.* Kinematic and visual differences between humans and the humanoid make it challenging to transfer learned ICL capabilities at test time (Sec. III-C).

B. Constructing Training Samples

Extracting low-level action and proprioceptive information. To fill in the missing action information, we use the future human hand pose k timesteps ahead as the intended action for the current timestep. MIMICDROID estimates the hand poses from human RGB frames using an off-the-shelf hand pose estimation model f [45], leveraging recent advances in vision-based hand tracking. This avoids overhead costs introduced by using specialized hardware to predict hand pose [18, 22, 44]. Specifically, given an RGB frame, the model f predicts the wrist pose and a grasp signal derived from finger joint angles. The predicted hand pose $h_{t+k} = f(s_{t+k})$ is treated as the action label a_t. In addition, the hand pose at each time step h_t is included in the observation, serving as the proprioception, to provide richer input for learning. This transforms a raw trajectory from the play data τ into a processed trajectory with inferred low-level actions and proprioceptives, $\tau' = \{s'_t, a_t\}_{t=1}^{T}$, where $s'_t = \{s_t, h_t\}$.

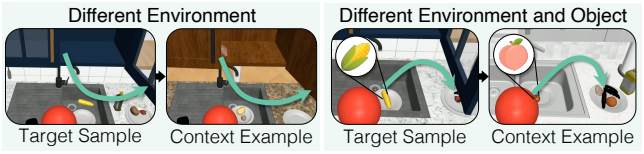

Fig. 3: Examples of target and retrieved context examples.

Constructing context-target training pairs. At test time, the policy must perform novel tasks, *e.g.*, using a new appliance by mimicking human demonstrations. This requires leveraging observation-action relationships and their recurring patterns across demonstrations to predict actions. To enable this, we construct Meta-ICL training samples by pairing trajectory segments with similar patterns, designating one as the target and the others as context (Eq. 1). Human play videos are well-suited for constructing Meta-ICL training data because they naturally exhibit such patterns, *e.g.*, moving an item from one receptacle to another and later moving a different item to another receptacle, each following similar observation-action relationships (Fig. 3).

To retrieve such similar patterns from play videos, specifically, we randomly sample a target trajectory segment $\sigma^{\text{tgt}} \subset \tau'$, and then identify the top k most similar segments $\{\sigma_i^{ctx}\}_{i=1}^{k}$ from the dataset to serve as context, where k is a hyperparameter. The similarity between two trajectory segments σ_x and σ_y is computed as the cosine similarity between their feature embeddings, $d(\sigma_x, \sigma_y) = \cos(\phi(\sigma_x), \phi(\sigma_y))$, where $\phi(\cdot)$ denotes the feature embedding function. To make segment features $\phi(\sigma)$ robust to visual noise (*e.g.*, background clutter) and capture the observation-action distribution, we first extract frame-wise visual features using the pretrained vision model g [49]. We then apply temporal mean pooling over the features, and concatenate the result with the sequence of actions to form the final segment features:

$$\phi(\sigma) = \Big[\frac{1}{T}\sum_{t=1}^{T} g(s_t),\ a_1,\ \ldots,\ a_T\Big] \tag{2}$$

In summary, MIMICDROID leverages the inherent repetitive manipulation behaviors naturally present in human play data to generate training samples for Meta-ICL (Fig. 2, bottom).

C. Overcoming the Embodiment Gap

To transfer policies from human play videos to humanoid robots, MIMICDROID must overcome embodiment gaps between the two. To bridge the kinematic gap, it retargets predicted human wrist poses to humanoid wrist poses and applies inverse kinematics to compute joint angles. By operating in task space and exploiting kinematic similarities, it preserves task intent while avoiding the need for demonstrators to mimic robot morphology [21, 22]. Moreover, differences in body appearance and occlusion patterns between humans and humanoids introduce a visual gap that can hinder ICL at test time. To reduce overfitting to the human-specific visual cues, MIMICDROID incorporates visual masking during training. Specifically, during training, random patches between 1 and n are masked in the input images. This operation is applied with a probability p, encouraging the model to rely less on superficial human-specific cues and instead learn representations that generalize across embodiments (Fig. 2, top-right). As a result, MIMICDROID can learn a policy using *only* human videos without *any* teleoperated demonstrations and deploy it on the robot.

D. Meta-training for In-Context Learning

To train the policy to perform ICL, we optimize the meta-learning objective in Eq. 1 using behavior cloning. Instead of assuming an explicitly defined discrete set of tasks in $p(\mathcal{T}_{\text{train}})$, we rely on an implicit task distribution induced by continuous human play. This is advantageous because human play naturally spans diverse manipulation behaviors, providing a richer and more scalable source of task variation than a manually defined set. We approximate the implicit distribution by uniformly sampling target trajectories from play videos, with each trajectory serving as a task instance. This exposes the policy to a wide variety of behaviors, while context trajectories for training are retrieved as described in Sec. III-B. This meta-training process exposes the policy to many such context-target pairs, enabling it to learn a general strategy for adapting to new tasks at test time.

Formally, each training instance consists of k context trajectory segments $\{\sigma_i^{ctx}\}_{i=1}^{k}$ and a target segment σ^{tgt} where the target segment is $\sigma^{\text{tgt}} = \{(s'_t, a_t)\}_{t=1}^{T}$ with s'_t denotes the RGB images with proprioceptive information, and a_t is the target action extracted from future hand poses (Sec. III-B). To model the multimodal nature of human play data, we adopt action chunking, where the policy predicts a sequence of l future actions at each time step instead of

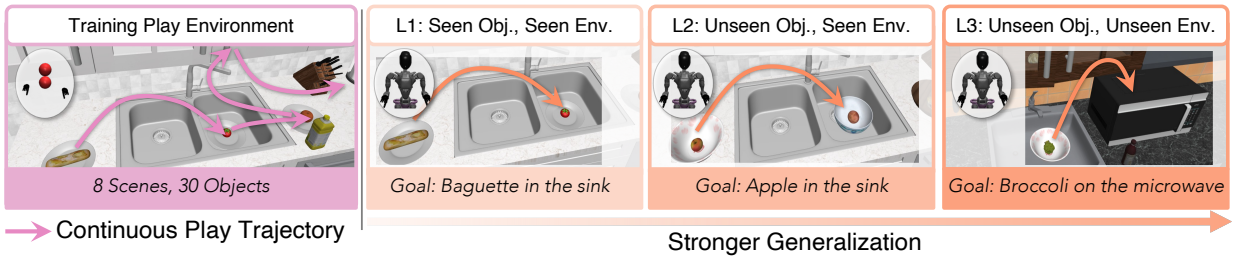

Fig. 4: **Overview of our simulation benchmark.** We introduce a simulation benchmark to evaluate few-shot learning for humanoid manipulation. It contains 8 hours of play data collected using free-floating hands (Left) across 30 objects and 8 kitchen environments. Evaluation is structured into three levels: L1, L2, and L3 with increasing difficulty via novel objects and environments, enabling systematic assessment of generalization to humanoids (target) (Right). The bottom-right of each image indicates the source and target embodiments. Language descriptions are included for clarity.

only the immediate next action [50]. The policy π is trained to imitate the extracted actions from the target segment by minimizing the L1 loss between predicted and ground-truth action chunks (defined for one sample):

$$\mathcal{L}_{\text{BC}} = \sum_{t=1}^{T-1} \left\| \pi\left(a^{\text{pred}}_{t:t+l} \;\middle|\; s'_{1:t}, a_{1:t-1},\ \{\sigma_{c,i}\}_{i=1}^{k} \right) - a_{t:t+l} \right\|_1 \tag{3}$$

Here, $a_{t:t+l}$ denotes the ground-truth l-step action sequence from time t, and $\pi(.)$ denotes the outputs corresponding l-step action prediction conditioned on the past trajectory and the k context examples.

IV. Experiment Setup

Simulation Benchmark. Until now, there has been no standardized benchmark to systematically evaluate few-shot learning in humanoid manipulation policies. To fill this gap and facilitate future research, we introduce a novel simulation benchmark with a wide range of objects, tasks, and environments, building on RoboCasa [29]. The benchmark includes humanoid manipulation tasks, such as pick-and-place tasks involving various objects and receptacles, as well as the manipulation of articulated objects like faucets and cabinets. We systematically categorize the tasks into progressively harder generalization levels (Fig. 4):

L1 (Seen Objects, Seen Environment): The robot must perform manipulation tasks with objects it encountered during training in the environments. This level assesses the model's ability to generalize to new object positions.

L2 (Unseen Objects, Seen Environment): The task requires the robot to apply the learned manipulation skills to novel objects not present in the training set, while the kitchen environments remain the same. It evaluates the ability to adapt to novel objects using only a few demonstrations, *e.g.*, learning the grasping strategy for the new object.

L3 (Unseen Objects, Unseen Environment): This most challenging scenario requires the robot to perform manipulation tasks in entirely new kitchen environments with novel furniture layouts, backgrounds, and novel objects. This level thoroughly tests the robot's ability to generalize from a few demonstrations to completely novel scenarios, often requiring a completely novel motion sequence to solve the task, *e.g.*, using a sink with a different faucet mechanism.

Embodiments. We consider two embodiments in simulation: a free-floating 6-DoF hand (Abstract) and a humanoid robot (GR1), both within the RoboCasa framework [29, 30]. This setup allows us to collect training data using the free-floating hand, while still evaluating the learned policy on the humanoid (GR1) platform, thereby simulating embodiment challenges [51]. It also enables a systematic analysis of the embodiment gap by comparing performance across the two embodiments. In the real world, we collect human play videos and evaluate on the GR1 humanoid.

Data-Collection Setup. We collect human play data by allowing an operator to interact with the scene freely. The operator performs meaningful tasks without specific goals, driven by curiosity. For example, in a kitchen scenario, the operator might open the oven, put bread on the oven tray, and later move it to different receptacles. This free-form approach captures a richer diversity of interactions and object configurations than typical teleoperated demonstrations, since the operator is not constrained by specific task goals or task-specific resets. Each play session lasts 20 minutes in simulation and 10 minutes in the real world. Interaction is done using a spacemouse in simulation and the operator's hand in the real world. We record each session as RGB videos in the real world, with randomized kitchen and object layouts. We collect 8 hours of simulated data (320k timesteps) and 45 minutes of real-world data (80k frames).

Implementation Details. During training, for each randomly sampled target trajectory segment, we find the top $k = 10$ similar trajectories to serve as the context. The features for the trajectory are extracted using $f =$ WiLoR [45] for hand pose estimation and $g =$ DinoV2 [49] for visual features (Sec. III-B). In each iteration, we randomly mask patches of an image with probability $p = 0.8$ and uniformly sample patches between 1–16 (Sec. III-C). The model is trained to predict $l = 32$ actions for each timestep (Sec. III-D).

Baselines. We compare MimicDroid to the following baselines to evaluate its ability to perform ICL from human play videos. (1) *Task-conditioned methods.* We use Vid2Robot [26], which conditions on human videos, and H2R [27], which conditions on final goal images. Both lack action information in the context, a key component for modeling observation-action relationships for ICL. This

TABLE I: Success rates in Abstract and GR1 embodiments in the simulation benchmark.

Method	Abstract			GR1		
	L1	L2	L3	L1	L2	L3
H2R [27]	0.03	0.05	0.03	0.03	0.00	0.00
Vid2Robot [26]	0.44	0.40	0.12	0.41	0.23	0.11
PEFT [28]	0.47	0.35	0.00	0.29	0.21	0.01
MIMICDROID w/o Visual Masking	0.59	**0.51**	0.22	0.37	0.35	0.09
MIMICDROID	**0.73**	0.39	**0.27**	**0.59**	**0.44**	**0.26**

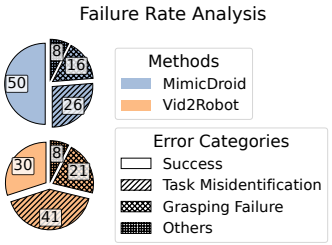

Fig. 5: MIMICDROID reduces task misidentification and grasping errors using ICL compared to the video-conditioned baseline.

comparison isolates the effect of ICL, as these baselines cannot learn in-context. (2) *Fine-tuning methods.* We compare MIMICDROID with the use of parameter-efficient fine-tuning (PEFT) to adapt at the test time for few-shot learning [28]. This comparison highlights the benefit of the instant, gradient-free adaptation of MIMICDROID via ICL over fine-tuning. All baselines are trained on the same human play videos and use identical augmentations and training pipelines to avoid confounding factors and ensure fairness.

V. RESULTS

In our evaluation, we aim to answer the following research questions and analyze MIMICDROID's failure cases.

How well does MIMICDROID achieve generalization through ICL to downstream tasks? We compare MIMICDROID to image goal-conditioned H2R [27] and video-conditioned Vid2Robot [26] (Tab. I), two task-conditioned baselines that receive task specification to recognize the intended behavior, but lack the observation-action pairs needed to perform ICL. In contrast, MIMICDROID using ICL achieves performance gains of $+14\%$ and $+18\%$ in abstract and humanoid embodiments, respectively. Furthermore, we compare with PEFT [28] and find MIMICDROID not only learns instantly compared to test-time finetuning, but also achieves higher success rates ($+29\%$, $+26\%$). Finetuning fails to adapt at the most challenging generalization level (L3), likely due to larger distribution shifts causing forgetting [52, 53], whereas ICL preserves pretrained knowledge due to gradient-free adaptation. We also observe that test-time finetuning leads to overfitting on the abstract embodiment, evidenced by a larger performance drop from abstract to humanoid in PEFT (-10%) compared to ICL (-3%).

Real-world results. Furthermore, we evaluate MIMICDROID on the GR1 humanoid in the real world, using a policy trained solely on human play videos. MIMICDROID achieves a success rate of **0.53** in L1, **0.23** in L2, **0.08** in L3, nearly twofold higher than Vid2Robot, which obtains 0.28, 0.08, 0.00, respectively. These results highlight MIMICDROID's ability to adapt to novel objects and scenes using few demonstrations via ICL (Fig. 9).

How effective is MIMICDROID at bridging the visual gap between embodiments? We ablate visual masking to assess its role in bridging the visual gap between embodiments. In the last two rows of Table I, removing masking (MIMICDROID w/o Visual Masking) causes a sharp drop in transfer performance to the humanoid (-17%) compared to only -3% with MIMICDROID, averaged over all three levels. Although it performs better in source L2, this is likely due to overfitting to the abstract embodiment. Moreover, we evaluate an alternative strategy, EgoMimic [44], which masks the hands with a black patch and red line. In the real world, MIMICDROID's random patch masking during training (0.53%) achieves comparable performance to EgoMimic (0.58%) while eliminating the need for external modules such as SAM [54] for segmentation. These results highlight the effectiveness of random patch masking for robustness to the visual gap between embodiments.

How well MIMICDROID's performance scales with the number of in-context examples? We evaluate how MIMICDROID's performance varies with the number of context demonstrations provided at test time. In Fig. 6, performance improves as the number of context examples increases from 1 to 3. However, the gains drop with 4 to 6 examples due to the limited context length supported by the transformer policy during training. Naively extending the context length is expensive as both memory and compute cost scale linearly with sequence length; future work can explore efficient ways to handle more in-context examples.

How does the quality of training data used in the Meta-ICL affect ICL performance? To evaluate the impact of data quality on Meta-ICL, we vary the number of similar context segments retrieved during training, k. As shown in Fig. 7, performance drops for large k due to the inclusion of dissimilar segments, while small k limits training diversity. We find that $k = 50$ strikes a balance between maintaining training diversity and avoiding excessive noise from dissimilar segments. These results highlight the importance of selecting context-target pairs with meaningful observation-action similarity for effective Meta-ICL. Future work can explore more robust data curation strategies for Meta-ICL.

How does scaling dataset size impact MIMICDROID's ability to perform ICL? We observe consistent performance improvements across all generalization levels (L1-L3) as the amount of training data increases, demonstrating the scalability of learning from RGB play videos (Fig. 8). Specifically, L1 improves from 35% at 128k frames to 59% at 320k, and L2 rises from 21% to 45%, resulting in a $+24\%$ absolute gain for both. L3 also improves, from 16% to 27% ($+11\%$), though the gains are less pronounced compared to L1 and L2. These results affirm the benefits of scaling

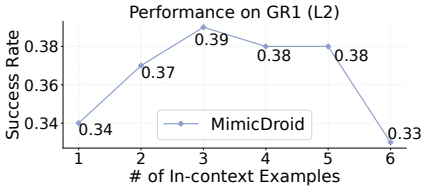

Fig. 6: Performance rises with more in-context examples but plateaus beyond 3 due to training-time context length.

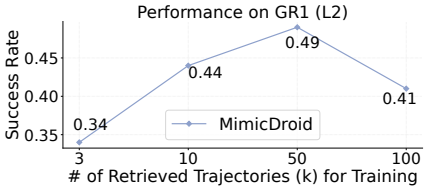

Fig. 7: Performance benefits from more retrieved context segments (k) per target, but high values introduce noise.

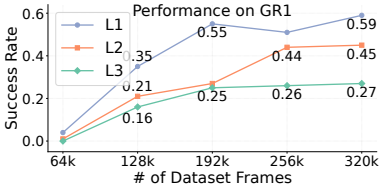

Fig. 8: Performance scales consistently with training data across from L1-L3, showing promise for learning from play.

Fig. 9: **Examples of real-world evaluations (L1-L3).** MIMICDROID generalizes to both seen (*e.g.*, chips, bread, and apples) and unseen objects (*e.g.*, potatoes, garlic, and cloth), as well as novel environments. Evaluation tasks include pick-and-place to different receptacles and articulated object manipulation (*e.g.*, opening or closing an oven tray).

training data, while also motivating a more systematic study of the factors that influence ICL performance on harder generalization tasks like L3.

Failure Analysis. Failures in downstream tasks arise from task misidentification (26%), missed grasps (16%), and other errors (8%) like incomplete cabinet closure, missed placement (Fig. 5). Compared to Vid2Robot, MIMICDROID notably reduces both misidentification (-15%) and grasping errors (-5%) using ICL.

Despite these gains, MIMICDROID still struggles with few-shot performance in L3 (Tab. I). It struggles to learn tasks requiring novel robot motions, potentially due to higher learning complexity. For instance, faucet activation in seen environments involves left-to-right motions, whereas L3 environments introduce sinks that require novel, bottom-to-top motions. Lastly, in real-world settings, MIMICDROID overfits to motions observed during training with specific hand sizes, leading to collisions in cluttered environments. We hypothesize this can be mitigated by using data from diverse operators, which we leave for future work.

VI. Limitations and Future Work

We aim to improve MIMICDROID by addressing several limitations. First, it relies on human play videos, which provide high-quality videos for learning to learn in-context. A natural extension is to augment with web-scale human videos (*e.g.*, YouTube) to expand object and environment diversity. Second, MIMICDROID extracts actions via an off-the-shelf hand pose predictor. While these models handle partial occlusions with hand priors, they fail when hands vanish (*e.g.*, reaching into cupboards or behind furniture). Combining hand and full-body motion estimation may help. Finally, MIMICDROID treats demonstrations as low-level state-action sequences, learning how but not why motions matter. Thus, it cannot generalize across semantically equivalent tasks. Meta-training with language-trajectory pairs, akin to Flamingo [5], may enable this capability.

VII. Conclusion

We introduce MIMICDROID, an in-context learning (ICL) method for few-shot learning in humanoid manipulation. When deployed, MIMICDROID infers humanoid action from a few human videos, possibly involving novel objects and environments via ICL. MIMICDROID acquires this capability by learning from continuous human play videos, leveraging a scalable and diverse data source. It achieves this by leveraging similar manipulation behaviors in human play data as a self-supervised signal to meta-train the policy for ICL. To bridge the visual and kinematic embodiment gaps between humans and humanoids, MIMICDROID uses random patch masking to reduce overfitting to human appearances and retargets human hand poses to humanoid wrist poses to preserve task intent. To support systematic evaluation, we introduce a simulation benchmark to assess few-shot learning for humanoid manipulation. Our results, in simulation and the real world, highlight the promise of leveraging ICL from human play videos, with a notable twofold improvement in performance. In conclusion, this work introduces a method for learning to learn in-context from a diverse, scalable human play videos, laying the groundwork for future research towards adaptive robot assistants for everyday environments.

Acknowledgments

We thank Ben Abbatematteo, Haonan Chen, and Bowen Jiang for providing valuable feedback on the manuscript. This work was partially supported by the National Science Foundation (FRR-2145283, EFRI-2318065), the Office of Naval Research (N00014-24-1-2550), the DARPA TIAMAT program (HR0011-24-9-0428), and the Army Research Lab (W911NF-25-1-0065). This work was further supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean Government (MSIT) (No. RS-2024-00457882, National AI Research Lab Project).

REFERENCES

[1] Finn *et al.*, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International conference on machine learning*, PMLR, 2017.

[2] Yu *et al.*, "One-shot imitation from observing humans via domain-adaptive meta-learning," *arXiv:1802.01557*, 2018.

[3] James *et al.*, "Task-embedded control networks for few-shot imitation learning," in *Conference on robot learning*, PMLR, 2018.

[4] Brown *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, 2020.

[5] Alayrac *et al.*, "Flamingo: A visual language model for few-shot learning," *Advances in neural information processing systems*, 2022.

[6] Laskin *et al.*, "In-context reinforcement learning with algorithm distillation," *arXiv:2210.14215*, 2022.

[7] Sridhar *et al.*, "Regent: A retrieval-augmented generalist agent that can act in-context in new environments," *arXiv:2412.04759*, 2024.

[8] Di Palo *et al.*, "Keypoint action tokens enable in-context imitation learning in robotics," *arXiv:2403.19578*, 2024.

[9] Fu *et al.*, "In-context imitation learning via next-token prediction," *arXiv:2408.15980*, 2024.

[10] Vosylius *et al.*, "Instant policy: In-context imitation learning via graph diffusion," *arXiv:2411.12633*, 2024.

[11] Raparthy *et al.*, "Generalization to new sequential decision making tasks with in-context learning," *arXiv:2312.03801*, 2023.

[12] Wang *et al.*, "Benchmarking general-purpose in-context learning," *arXiv:2405.17234*, 2024.

[13] Kirsch *et al.*, "General-purpose in-context learning by meta-learning transformers," *arXiv:2212.04458*, 2022.

[14] Sridhar *et al.*, "Ricl: Adding in-context adaptability to pre-trained vision-language-action models," *arXiv:2508.02062*, 2025.

[15] Mandlekar *et al.*, "Roboturk: A crowdsourcing platform for robotic skill learning through imitation," in *Conference on Robot Learning*, PMLR, 2018.

[16] Collaboration, *Open X-Embodiment: Robotic learning datasets and RT-X models*, 2023.

[17] al, "Droid: A large-scale in-the-wild robot manipulation dataset," 2024.

[18] Wang *et al.*, "Mimicplay: Long-horizon imitation learning by watching human play," *arXiv:2302.12422*, 2023.

[19] Lynch *et al.*, "Learning latent plans from play," in *Conference on robot learning*, PMLR, 2020.

[20] Min *et al.*, "Metaicl: Learning to learn in context," *arXiv:2110.15943*, 2021.

[21] Li *et al.*, "Okami: Teaching humanoid robots manipulation skills through single video imitation," in *8th Annual Conference on Robot Learning*, 2024.

[22] Qiu *et al.*, "Humanoid policy˜ human policy," *arXiv:2503.13441*, 2025.

[23] Mistry *et al.*, "Representation and control of the task space in humans and humanoid robots," *Humanoid Robotics and Neuroscience: Science, Engineering and Society*, 2015.

[24] He *et al.*, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022.

[25] Fu *et al.*, "Rethinking patch dependence for masked autoencoders," *arXiv:2401.14391*, 2024.

[26] Jain *et al.*, "Vid2robot: End-to-end video-conditioned policy learning with cross-attention transformers," *arXiv:2403.12943*, 2024.

[27] Bharadhwaj *et al.*, "Zero-shot robot manipulation from passive human videos," *arXiv:2302.02011*, 2023.

[28] Hong *et al.*, "Hand me the data: Fast robot adaptation via hand path retrieval," *arXiv:2505.20455*, 2025.

[29] Nasiriany *et al.*, "Robocasa: Large-scale simulation of everyday tasks for generalist robots," in *Robotics: Science and Systems*, 2024.

[30] Zhu *et al.*, "Robosuite: A modular simulation framework and benchmark for robot learning," in *arXiv:2009.12293*, 2020.

[31] Schmidhuber, "Evolutionary principles in self-referential learning," *Diploma thesis, Institut f. Informatik, Tech. Univ. Munich*, 1987.

[32] Naik *et al.*, "Meta-neural networks that learn by learning," in *[Proceedings 1992] IJCNN International Joint Conference on Neural Networks*, IEEE, 1992.

[33] Santoro *et al.*, "Meta-learning with memory-augmented neural networks," in *International conference on machine learning*, PMLR, 2016.

[34] Hochreiter *et al.*, "Learning to learn using gradient descent," in *International conference on artificial neural networks*, Springer, 2001.

[35] Vaswani *et al.*, "Attention is all you need," *Advances in neural information processing systems*, 2017.

[36] Shah *et al.*, "Rrl: Resnet as representation for reinforcement learning," *arXiv:2107.03380*, 2021.

[37] Nair *et al.*, "R3m: A universal visual representation for robot manipulation," *arXiv:2203.12601*, 2022.

[38] Majumdar *et al.*, "Where are we in the search for an artificial visual cortex for embodied intelligence?" *Advances in Neural Information Processing Systems*, 2023.

[39] Shao *et al.*, "Concept2robot: Learning manipulation concepts from instructions and human demonstrations," *The International Journal of Robotics Research*, 2021.

[40] Ma *et al.*, "Liv: Language-image representations and rewards for robotic control," in *International Conference on Machine Learning*, PMLR, 2023.

[41] Ma *et al.*, "Vip: Towards universal visual reward and representation via value-implicit pre-training," *arXiv:2210.00030*, 2022.

[42] Guzey *et al.*, "Bridging the human to robot dexterity gap through object-oriented rewards," *arXiv:2410.23289*, 2024.

[43] Bahety *et al.*, "Screwmimic: Bimanual imitation from human videos with screw space projection," *arXiv:2405.03666*, 2024.

[44] Kareer *et al.*, "Egomimic: Scaling imitation learning via egocentric video," *arXiv:2410.24221*, 2024.

[45] Potamias *et al.*, *Wilor: End-to-end 3d hand localization and reconstruction in-the-wild*, 2024.

[46] Pavlakos *et al.*, "Reconstructing hands in 3D with transformers," in *CVPR*, 2024.

[47] Lepert *et al.*, "Phantom: Training robots without robots using only human videos," *arXiv:2503.00779*, 2025.

[48] Zhu *et al.*, "Vision-based manipulation from single human video with open-world object graphs," *arXiv:2405.20321*, 2024.

[49] Oquab *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv:2304.07193*, 2023.

[50] Zhao *et al.*, "Learning fine-grained bimanual manipulation with low-cost hardware," *arXiv:2304.13705*, 2023.

[51] Seo *et al.*, "Legato: Cross-embodiment imitation using a grasping tool," *IEEE Robotics and Automation Letters*, 2025.

[52] Yang *et al.*, "Self-distillation bridges distribution gap in language model fine-tuning," *arXiv:2402.13669*, 2024.

[53] Zhao *et al.*, "Does continual learning equally forget all parameters?" In *International Conference on Machine Learning*, PMLR, 2023.

[54] Kirillov *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023.