Ag2Manip: Learning Novel Manipulation Skills with Agent-Agnostic Visual and Action Representations

Puhao Li[1,2,⋆], Tengyu Liu[1,⋆], Yuyang Li[1,2,3], Muzhi Han[1,4], Haoran Geng[1,5], Shu Wang[1,4], Yixin Zhu[3], Song-Chun Zhu[1,2,3], Siyuan Huang[1,†]

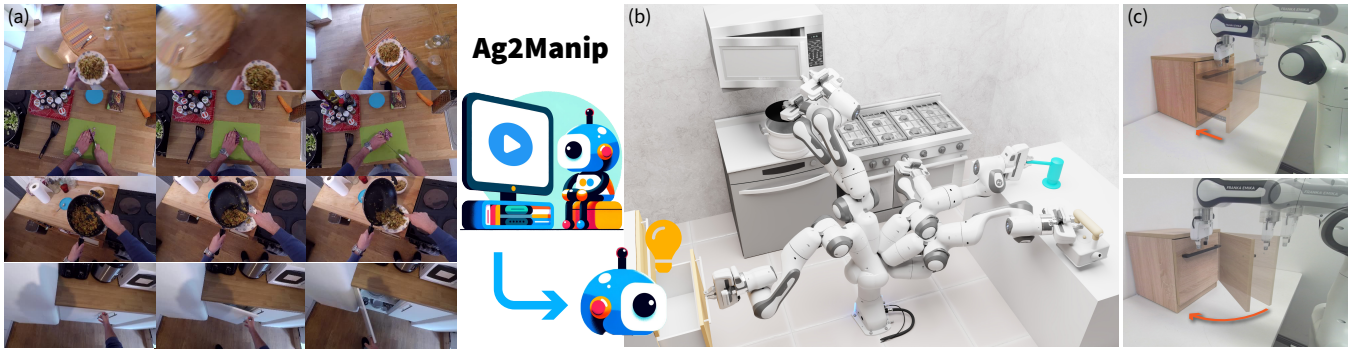

Fig. 1: **Ag2Manip enables various manipulation tasks in scenarios where domain-specific demonstrations are unavailable.** With agent-agnostic visual and action representations, Ag2Manip: (a) learns from human manipulation videos; (b) acquires diverse manipulation skills autonomously in simulation; and (c) supports robust imitation learning of manipulation skills in the real world.

***Abstract*— Enhancing the ability of robotic systems to autonomously acquire novel manipulation skills is vital for applications ranging from assembly lines to service robots. Existing methods (*e.g.*, VIP, R3M) rely on learning a generalized representation for manipulation tasks but overlook (i) the domain gap between distinct embodiments and (ii) the sparseness of successful task trajectories within the embodiment-specific action space, leading to misaligned and ambiguous task representations with inferior learning efficiency. Our work addresses the above challenges by introducing Ag2Manip (Agent-Agnostic representations for Manipulation) for learning novel manipulation skills. Our approach encompasses two principal innovations: (i) a novel agent-agnostic visual representation trained on human manipulation videos with embodiments masked to ensure generalizability, and (ii) an agent-agnostic action representation that abstracts the robot's kinematic chain into an agent proxy with a universally applicable action space to focus on the core interaction between the end-effector and the object. Through our experiments, Ag2Manip demonstrates remarkable improvements across a diverse array of manipulation tasks without necessitating domain-specific demonstrations, substantiating a significant *325%* improvement in average success rate across 24 tasks from FrankaKitchen, ManiSkill, and PartManip. Further ablation studies underscore the critical role of both representations in achieving such improvements.**

I. Introduction

The ability to learn and master new manipulation skills without expert demonstrations (see Fig. 1) is crucial for robots while adapting to evolving tasks and environments. Despite significant advancements in learning manipulation skills [1–7], the autonomous acquisition of these skills without expert demonstrations and task-specific reward remains a challenge. To tackle this, previous works [8–10] explore leveraging extensive pre-training to facilitate manipulation learning. Among them, [8, 9] develop general visual representations from human-centric video datasets such as Epic-Kitchen [11] and Ego4D [12], which captures the essence of tasks and the temporal relationship between visual frames, and then generates rewards that guide robots towards achieving desired goals. Other approaches [10] employ Large Language Models (LLMs) to directly create reward functions for acquiring new manipulation skills. However, these methods usually fall short when facing complex tasks, underscoring three primary challenges in novel skill learning.

First, the visual representations trained on human-centric demonstrations [8, 9] face difficulty reconciling the diverse appearance and kinematics discrepancies between humans and robots. The discrepancy in appearance introduces biases when these models are applied to robots, compromising the models' ability to accurately interpret tasks and the temporal relations between video frames. Additionally, kinematic differences result in divergent task execution strategies between humans and robots; robots may opt for trajectories distinct from those observed in human demonstrations to complete tasks like picking up a cup. This divergence can lead the model to mistakenly view a robot's efficient path as incorrect, a misjudgment rooted in its human-centric training data.

Second, due to the consistent presence of human hands in the training data, these human-centric visual models tend to prioritize the appearance of human hands, focusing on their position and movement rather than the actual task objective. For instance, in cup manipulation, the model might emphasize the upward movement of hands, disregarding

⋆ Puhao Li and Tengyu Liu contributed equally to this paper.
† Corresponding email: syhuang@bigai.ai.
[1] National Key Laboratory of General Artificial Intelligence, Beijing Institute for General Artificial Intelligence (BIGAI). [2] Department of Automation, Tsinghua University. [3] Institute for Artificial Intelligence, Peking University. [4] University of California, Los Angeles. [5] School of Electronics Engineering and Computer Science, Peking University.

whether the cup has been successfully picked up.

Lastly, the precision required in robot manipulation further exaggerates the aforementioned challenges. Minor deviations in trajectories can lead to significant performance drops. While expert-designed rewards offer detailed guidance, those generated from visual or language models tend to be broad and high-level, leading to inaccuracies. This issue becomes particularly pronounced in tasks requiring intricate interaction with the environment, such as opening a door, where precise actions like grasping the handle are crucial.

We introduce **Ag2Manip**: Agent-Agnostic representations for Manipulation to tackle the above challenges. As shown in Fig. 2, Ag2Manip consists of two key designs of generalizable visual and action representations.

To tackle the challenges rooted in human-centric training data, we propose to learn an agent-agnostic visual representation. Inspired by Bahl *et al*. [2], we segment and mask both the humans and the robots from the video frames and then inpaint the videos with the masks. By training on the frames with the agent removed following R3M [8], the learned agent-agnostic visual representation eliminates the domain gap between humans and robots, enabling robust applications to robot-centric scenarios. The rewards derived from this agent-agnostic representation focus on the procedures of the tasks rather than on the human hand, providing more robust and clearer guidance to the goal of manipulation learning.

To tackle the inaccuracies from the visual guidance, we introduce an agent-agnostic action representation. This method abstracts a robot's actions into those of a proxy agent, designed with an action space universally applicable across various embodiments. This representation simplifies manipulation learning into reinforcement learning with two phases: *exploration* and *interaction*. In the *exploration* phase, we focus on learning the trajectory of the proxy's position, which mirrors the end-effector's movements, facilitating environment exploration. Once the proxy is close enough to a tractable area of an object, we switch to the *interaction* phase. Here, the learning objective shifts to understanding the proxy's exerted forces, simulating the physical interactions between the end-effector and objects. This abstraction of the manipulation process into *exploration* and *interaction* phases sidesteps the complexities tied to the robot arm's movements and object handling. By adopting this agent-agnostic action space, we offer a streamlined framework for learning manipulation tasks, enabling the policy to focus on the crucial aspects of tasks while alleviating the impact of lacking detailed, granular guidance. We additionally propose a carefully shaped reward function for reinforcement learning in both stages to encourage interaction and a retargeting approach to recover the robot's arm motions.

We demonstrate the efficacy and robustness of Ag2Manip through goal-conditioned novel skill learning without expert demonstrations and task-specific rewards. Extensive experiments are conducted on diverse tasks in simulation across FrankaKitchen [13], ManiSkill [14] and PartManip [6]. Experiments reveal that Ag2Manip attains an outstanding **78.7%** success rate across all tasks, markedly surpassing baseline methods, which achieve only an 18.5% success rate. By adopting agent-agnostic visual and action representations, Ag2Manip marks a significant stride in the field of manipulation learning without the need for domain-specific demonstrations. This progress equips robots with the capability to adeptly handle novel tasks across diverse scenarios. We further validate our approach through real-world experiments, where the robot acquires diverse manipulation skills by learning from demonstrations. Our model achieves a significantly better success rate compared with others.

In summary, our work presents three key contributions to learning novel manipulation skills without expert demonstrations: (i) an agent-agnostic visual representation that bridges the embodiment gap, enabling more accurate interpretation of visual data for robotic systems; (ii) an agent-agnostic action representation that streamlines the learning of manipulation tasks by abstracting complex robot actions into simpler and universal proxy-agent actions; this representation is further boosted by a carefully shaped reward function that encourages interactions with the environment; and (iii) a significant advancement in the performance of robot novel skill learning, demonstrated across a range of demanding tasks, showcasing the practical efficacy of our approach in improving robotic adaptability and skill acquisition without the need for direct human oversight.

II. Related Works

A. Learning Robotic Manipulation

Learning robotic manipulation relies not only on basic motor skills like grasping [15–17] and manipulation [6, 18–20], but also on obtaining advanced cognitive abilities for discerning task specifics, such as the location, method, and rationale for different tasks [21–23]. Leveraging parallel simulation environments [24, 25], Goal-Conditioned Reinforcement Learning (GCRL) comes in handy in learning skills but often requires manually crafted reward functions for each task [15, 16, 19], even with the help of LLMs and human feedback [10]. One promising solution is learning manipulation skills from demonstrations, circumventing the extensive exploration and simplifying scalability [23]. Robot action trajectories can be obtained via teleoperation [26, 27], Augmented Reality (AR) systems [28], and teach pendant programming [1, 3, 4]. Learning from human videos presents a cost-effective yet challenging method to translate interactions into motor controls [27, 29]. The balance between the cost of data collection and the quality of demonstrations remains a significant obstacle to the direct acquisition of novel skills from such demonstrations.

Drawing inspiration from recent studies [8, 9], our work introduces a generalized visual representation to facilitate novel manipulation skill learning across various tasks, utilizing the rich resource of human demonstrations. This approach aims to overcome the limitations of direct skill learning from videos, offering a scalable and efficient pathway for robots to acquire new abilities.

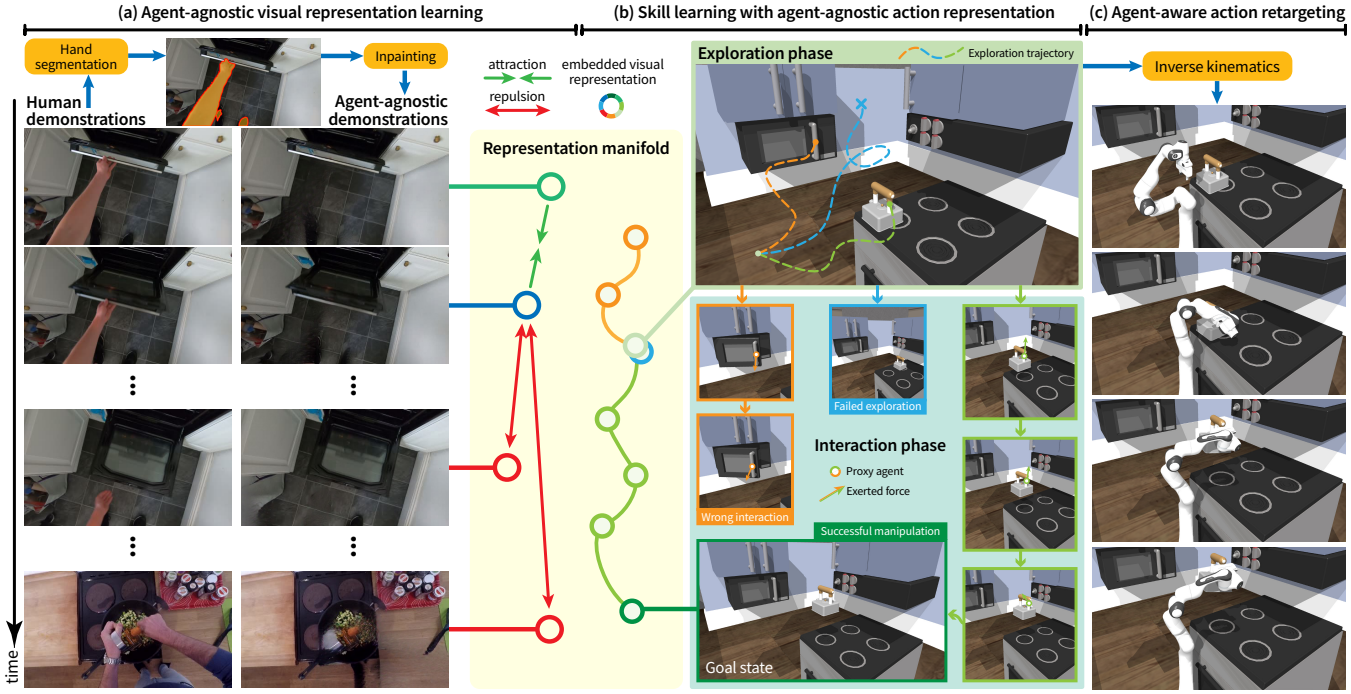

Fig. 2: **Pipeline of Ag2Manip.** Our method consists of three parts: (left) learning an agent-agnostic visual representation; (middle) learning abstracted skills with an agent-agnostic action representation; and (right) retargeting the abstracted skills to a robot.

B. Reward Generation for Skill Learning

Skill learning with model-free Reinforcement Learning (RL) is labor intensive for its requirement of an expert-designed task- and embodiment-specific reward. Tackling this problem requires an autonomously generated reward function for decision-making in each task.

Foundation models like LLMs are capable of directly generating reward functions given the task description [10, 30–32]. However, without environmental awareness, their efficiency is diminished by the necessity of expert feedback [10]. Moreover, this approach often depends on the environmental states, typically inaccessible in the real world.

Perceptual rewards offer a promising alternative for skill learning. By watching videos of humans performing each task [11], robots learn an implicit embedding that reflects how events progress in video and use it as a flexible form of reward [8, 33, 34]. Extending beyond these, researchers propose to learn the temporal dynamics not in task-specific clips but from videos across different tasks to form a task-agnostic visual representation exhibiting greater generalizability [9].

Building upon these advancements, our work isolates agent-aware information from the visual reward to further enhance its robustness and generalizability.

C. Agent-Agnostic Representation

The principle of abstracting actions, objects, and tasks into *agent-agnostic* representations decouples them from the constraints of specific robotic articulations or sensor configurations. This methodology enhances adaptability and transferability across robotic platforms and even human contexts by disentangling low-level perception and control from the modeling, focusing on high-level action abstractions.

A manipulation task can be abstracted to the desired state changes of the world over time [35], which minimizes the direct involvement of an agent. To further describe the interaction between the agent and the object while remaining agent-agnostic, interaction regions (commonly identified as affordances) and trajectories [2, 17, 36–40] define how tasks are performed without relying on the robot's specific motor controls. To represent the interaction region, a simple yet effective way is to use contact points to specify the required contacts between the finger and object [38, 40–42], which is suitable for simple end-effectors like parallel grippers or suckers. In contact-rich interactions, contact maps become necessary to capture the detailed contact or the precise distances from the object surface to each finger [17, 43].

III. Method

We investigate the problem of robotic manipulation learning where expert demonstrations are unavailable. Specifically, given a robot and an image of a desired goal state, we are tasked to learn the robot's motion to accomplish the goal. We devise **Ag2Manip**: Agent-Agnostic representations for Manipulation. Figure 2 illustrates our method's framework. Our approach features two innovative concepts: an agent-agnostic visual representation (Sec. III-A) that bridges the domain gap between humans and robots and an agent-agnostic action representation (Sec. III-B) that abstracts robot actions into a universal proxy agent's action. Leveraging these representations, we employ reinforcement learning to derive a manipulation policy within this abstracted action space, guided by a unique reward function rooted in our agent-agnostic visual framework (Sec. III-C). We then retarget the proxy agent's trajectory to a robot trajectory by

Inverse Kinematics (IK) (Sec. III-D).

A. Agent-Agnostic Visual Representation

Building on visual representations pre-trained on human demonstrations [8, 9], our objective is to develop an agent-agnostic visual representation that bridges the domain gap between human and robot manipulations. This approach aims to enhance the applicability and effectiveness of these representations in robotic contexts, enabling a more flexible acquisition of manipulation skills.

Data Pre-processing: We consider a set of human demonstration video data $\mathcal{D}=\{v^c:=(o_1^c,o_2^c,...,o_{n_c}^c)\}_{c=1}^N$, where $o_f^c\in\mathbb{R}^{H\times W\times 3}$ is the f-th raw frame in the c-th video clip v^c that describes how a human completes a manipulation task. Inspired by Bahl et al. [2], we initiate this process by segmenting the human body from each frame using the ODISE algorithm [44]. Following segmentation, we employ a video inpainting model, E^{2}FGVI [45], to fill in the areas previously occupied by the human. This approach not only removes the human from the video but also ensures a smooth temporal coherence between frames, resulting in a manipulation dataset $\mathcal{D}^a$ that is effectively agent-agnostic.

Time-Contrastive Pre-training: Given the agent-agnostic demonstration dataset $\mathcal{D}^a$, we aim to learn an encoder $\mathcal{F}_\phi\colon\mathbb{R}^{H\times W\times 3}\to\mathbb{R}^K$ that maps a visual observation into a latent embedding, where K denotes the embedding dimension. Following Nair *et al*. [8], we minimize the time-contrastive loss [46] $\mathcal{L}_{\text{tcn}}$ and the regularization penalty $\mathcal{L}_{\text{reg}}$:

$$\mathcal{L}=\lambda_1\mathbb{E}_{o_i^c,o_j^c,o_k^c,o_l^{\neq c}\sim\mathcal{D}^a}\mathcal{L}_{\text{tcn}}+\lambda_2\mathbb{E}_{o\sim\mathcal{D}^a}\mathcal{L}_{\text{reg}},\tag{1}$$

where $(o_i^c,o_j^c,o_k^c)\sim v^c$ indicates a set of temporally ordered 3-frame samples, and each sample in a set is drawn from the same video clip v^c to ensure task proximity. $o_l^{\neq c}$ is a negative sample from a disparate video clip.

The time-contrastive loss is designed to guide the representation so that frames temporally closer to each other are mapped closer in the embedding space, compared to frames that are temporally distant or from disparate video clips:

$$\mathcal{L}_{\text{tcn}}=-\log\frac{e^{\mathcal{S}\left(z_i^c,z_j^c\right)}}{e^{\mathcal{S}\left(z_i^c,z_j^c\right)}+e^{\mathcal{S}\left(z_i^c,z_k^c\right)}+e^{\mathcal{S}\left(z_i^c,z_l^{\neq c}\right)}},\tag{2}$$

where $\mathcal{S}(\cdot,\cdot)$ represents the similarity metric between two embeddings, $z_i^c=\mathcal{F}_\phi(o_i^c)$ denotes the embedding of o_i^c extracted from the encoder $\mathcal{F}_\phi$. The regularization loss encourages a more compact embedding space:

$$\mathcal{L}_{\text{reg}}=\|\mathcal{F}_\phi(o)\|_1+\|\mathcal{F}_\phi(o)\|_2.\tag{3}$$

B. Agent-Agnostic Action Representation

We abstract robotic manipulation learning with an agent-agnostic action representation, where the motion and force of a universal free-floating proxy agent summarize a robot's action. We split the learning into two phases: *exploration* and *interaction*. The *exploration* phase explores the proxy's position, and the *interaction* phase investigates the proxy's force exerted on the environment. For each task represented by a goal image, we learn an RL policy across both phases to minimize the embedding distance between the agent-agnostic visual representations of the current and the goal image.

The *Exploration* Phase: In the *exploration* phase, we abstract the robot into a universal proxy agent, an agent-agnostic sphere to represent its end-effector, and abstract the robot's action into a sequence of positions of the sphere.

We then control the proxy using a proportional-derivative (PD) controller [47]. Note that the proxy has a collision volume with a radius of r_{e}, representing its body. The *exploration* phase halts when the proxy enters a precomputed interactable region in the environment, where the *interaction* phase begins. Since we only consider robots with two-finger grippers in this work, we consider interactable regions to be regions where parallel grasps are highly likely to succeed. Specifically, given the point cloud scan of the environment, we extract the interactive regions as regions with a radius of r_{int} near possible grasping poses detected by GraspNet [48].

While our implementation employs parallel grasp detection for efficiency, it's important to note that general-purpose methods, such as GenDexGrasp [17], can also provide effective interactable regions for various dexterous manipulators.

The *Interaction* Phase: Once the proxy agent enters an interactable region, we consider a grasp available and the object attached to the agent. We then enter the *interaction* phase, where we focus on the motion of the object. We abstract the robot's action space into the force exerted on the environment by the proxy.

C. Reinforcement Learning and Reward Shaping

Given a goal image $g\in\mathbb{R}^{H\times W\times 3}$, our task is to accomplish the task it represents. We use a model-free and GCRL framework to learn the agent-agnostic action policy $\pi=\{\pi_{\text{exp}},\pi_{\text{int}}\}$, with π_{exp} and π_{int} denoting the proxy agent's policies for the *exploration* and *interaction* phases, respectively. The policy π takes the robot states r_t and the environment's states s_t at frame t as its observation and produces the action $a_t=(a_p^t,a_f^t)$, where $a_p^t\in\mathbb{R}^3$ indicates the proxy's desired position in *exploration* and $a_f^t\in\mathbb{R}^3$ indicates the proxy's intended force in *interaction*. A PD controller then guides the proxy to achieve the target action.

To reach the goal depicted by g, we focus on maximizing the similarity $\mathcal{S}(z_t,z_g)$ between the embeddings for current and goal images o_t and g. Recognizing that directly employing $\mathcal{S}$ as a reward function could inappropriately penalize trajectories close but not identical to optimal, we introduce an importance-weighted reward function to promote explorations leading to states that improve upon the initial state:

$$\mathcal{R}(o_t,g;\phi)=\exp\left(\left(1+\alpha\cdot\mathbf{1}_{\mathcal{S}(z_t,z_g)-\beta>0}\right)\frac{\mathcal{S}(z_t,z_g)-\beta}{\beta}\right)-1,\tag{4}$$

where $\beta=\mathcal{S}(z_0,z_g)$ is the similarity between the embeddings of the initial and goal images, and $\alpha>0$ is a tunable hyperparameter. With the indicator function, the proposed reward function emphasizes states closer to the goal image than the initial state and reduces penalties for those that diverge from the goal. This approach encourages more explorations from the proxy and is especially crucial during the early

phase of learning with random policy behaviors.

For policy optimization, we utilize Proximal Policy Optimization (PPO) [49], chosen for its training stability and efficiency in convergence. Through PPO, we aim to maximize the expected cumulative reward $\mathbb{E}\left[\sum_{t=0}^{T-1}\gamma^t\mathcal{R}(o_t, g; \phi)\right]$, thereby effectively guiding the policy π towards the goal.

D. Robot-Specific Action Retargeting

We implement a straightforward retargeting policy to adapt the proxy's trajectory generated by π to real robots, converting proxy actions into robot-specific movements.

In the *exploration* phase, we directly map the proxy agent's positions to the robot's end-effector positions, effectively translating the proxy's trajectory into the end-effector's motion. As we transition from *exploration* to *interaction*, we align the end-effector's 6D pose with the nearest grasping pose identified by GraspNet, a feasible step given the transition occurs only when a viable grasp pose is within reach. For the *interaction* phase, the end-effector's 6D trajectory is derived from the moving object's trajectory, ensuring the robot's actions are synchronized with the target object's dynamics. We solve the robot arm's trajectory using IK, thereby ensuring the practical execution of the task in alignment with the proxy's movements.

E. Implementation Details

In Sec. III-A, we choose Epic-Kitchen [11] as the human demonstration dataset. Echoing the choices of R3M [8] and VIP [9], we use a standard ResNet50 [50] as the architecture of the visual encoder $\mathcal{F}_\phi$. We use the negative L2 distance to measure similarity $\mathcal{S}(\cdot,\cdot)$. The weights for our learning objective are set to $\lambda_1 = \lambda_2 = 1.0$. The visual encoder is optimized using an Adam optimizer, with a learning rate set at 10^{-4}, and the process is run for 24 hours on a single NVIDIA A100 GPU. In Sec. III-B, we define the proxy's collision radius r_e as 2 centimeters and its interactive region radius r_int as 10 centimeters. In Sec. III-C, we empirically set $\alpha = 3.0$ for our reward function across all tasks.

IV. Experiments

To thoroughly evaluate the capabilities of our proposed method, we implemented extensive tests across various tasks and environments. We observe that our method notably enhanced the success rates, from a baseline of 18.5% to an impressive 78.7% across tasks from three distinct environments. We further evaluate the quality of our visual representation and record a substantial boost in imitation learning success rate from 50% to 77.5%. These remarkable advancements underline not only the efficacy of our approach but also its potential to significantly impact practical applications.

A. Experimental Setup

Evaluation Environments: To evaluate our method's effectiveness across a broad spectrum of manipulation tasks, we selected 24 tasks from three distinct simulation environments: FrankaKitchen [13], ManiSkill [14], and PartManip [6], encompassing a diverse array of actions (e.g., opening, pulling, moving) and interacting with various objects (e.g., cabinets, microwaves, kettles). These tasks were executed using a 9-DOF Franka Emika robotic arm and gripper, representing a standard robotic manipulation setup.

We conducted all experiments within the NVIDIA IsaacGym, a GPU-accelerated simulator that facilitates rapid learning through reinforcement learning techniques. The robot was initialized in a default state for consistency across tasks, with each task defined by a goal state illustrated through an image rendered from one of three preset camera angles (front, left, right). A task is considered successfully completed when the relevant object or part achieves the goal state within a margin of error.

For a robust evaluation, each of the 24 tasks was tested under nine different configurations with varying camera perspectives and initialization seeds (3 cameras × 3 seeds) to ensure comprehensive assessment under varying conditions.

Baselines and Ablations: We compare our method against two important baselines, R3M [8] and VIP [9], which both used agent-aware visual representations for manipulation skill learning through time-contrastive learning objectives. Additionally, we benchmark Eureka, a recent method notable for its autonomous skill acquisition capabilities, utilizing LLMs to generate reward functions automatically, showcasing its relevance as a strong competitor.

For consistency and fairness in evaluation, all methods except Eureka utilize a ResNet50 visual backbone and are trained on the Epic-Kitchen dataset. We disabled the human feedback for Eureka to exclude task-specific expert knowledge. This ensures that comparisons emphasize the intrinsic capabilities of each method's learning strategy.

In our ablation study, we remove components from our method to investigate their effectiveness independently. "Ours w/o Act.Repr." learns directly in the robot's original action space while using the agent-agnostic visual representation. "Ours w/o Rew.Shp." replaces the specialized reward function with a basic similarity measure. We don't test the effect of removing only the visual representation because calculating agent-aware visuals is not feasible without agent-aware actions. We also omit the option of removing both representations as they directly resemble the R3M baseline. Further, to probe the performance drop during retargeting, we evaluate the performance of the proxy agent without retargeting its actions to a robot, denoted as "Ours (Proxy)."

Due to space constraints, we consolidate the quantitative results of our experiments in Tab. I, providing a comprehensive overview of our findings.

B. Comparative Study Results

Tab. I shows the average task success rates within and across all 3 environments. Our method, Ag2Manip, achieves an impressive overall task success rate of **78.7%**, significantly outperforming baseline methods (11.1%, 12.0%, and 18.5%). We further report task-specific success details and reveal the notable distinctions of each method's capabilities. We observe that the baseline methods consistently fail to perform tasks involving fine-grained robot-object interactions.

TABLE I: **Main comparison and ablation study**. All tasks have been tested on 3 seeds $\times$ 3 cameras $=$ **9 runs** and the numbers $\mathbf{0-9}$ represent the number of successful runs. We use the characters **a - x** to specify different tasks. Tasks from FrankaKitchen [13]: **a:** open hinge-cabinet **b:** open microwave **c:** open slide-cabinet **d:** close hinge-cabinet **e:** close microwave **f:** close slide-cabinet **g:** move kettle **h:** pick up kettle **i:** turn on switch **j:** turn off switch. Tasks from ManiSkill2 [14]: **k:** open door **l:** close door **m:** pick up cube **n:** stack cube **o:** pick up clutterycb **p:** insert peg **q:** turn left faucet **r:** turn right faucet. Tasks from PartManip [6]: **s:** turn down dishwasher **t:** pull drawer **u:** turn up dishwasher **v:** push drawer **w:** press button **x:** lift lid.

Method	FrankaKitchen											ManiSkill									PartManip							Overall
	a	b	c	d	e	f	g	h	i	j	Avg.	k	l	m	n	o	p	q	r	Avg.	s	t	u	v	w	x	Avg.	
R3M [8]	0	0	0	3	2	0	1	0	0	0	6.7%	0	6	0	0	0	0	0	0	8.3%	0	0	3	9	0	0	22.2%	11.1%
VIP [9]	0	0	0	2	6	0	3	0	0	0	12.2%	0	6	0	0	0	0	0	0	8.3%	0	0	0	9	0	0	16.7%	12.0%
Eureka [10]	0	0	0	7	3	2	3	0	0	0	16.7%	0	9	0	0	0	0	0	1	13.9%	0	0	3	6	0	0	20.0%	18.5%
Ours w/o Act.Repr.	4	1	8	9	9	9	9	1	7	2	65.6%	0	9	0	0	0	0	1	8	25.0%	0	0	8	9	0	0	31.5%	43.5%
Ours w/o Rew.Shp.	8	7	7	9	9	9	7	9	1	0	73.3%	9	9	8	0	3	1	4	5	54.2%	9	6	8	9	0	9	75.9%	67.6%
Ours	7	8	8	8	8	9	8	6	9	9	**88.9%**	7	9	6	0	7	2	8	8	**65.3%**	9	7	9	9	0	9	**79.6%**	**78.7%**
Ours (Proxy)	8	9	9	8	9	9	9	9	9	9	97.8%	7	9	5	5	7	3	8	9	73.6%	9	9	9	9	0	8	81.5%	85.7%

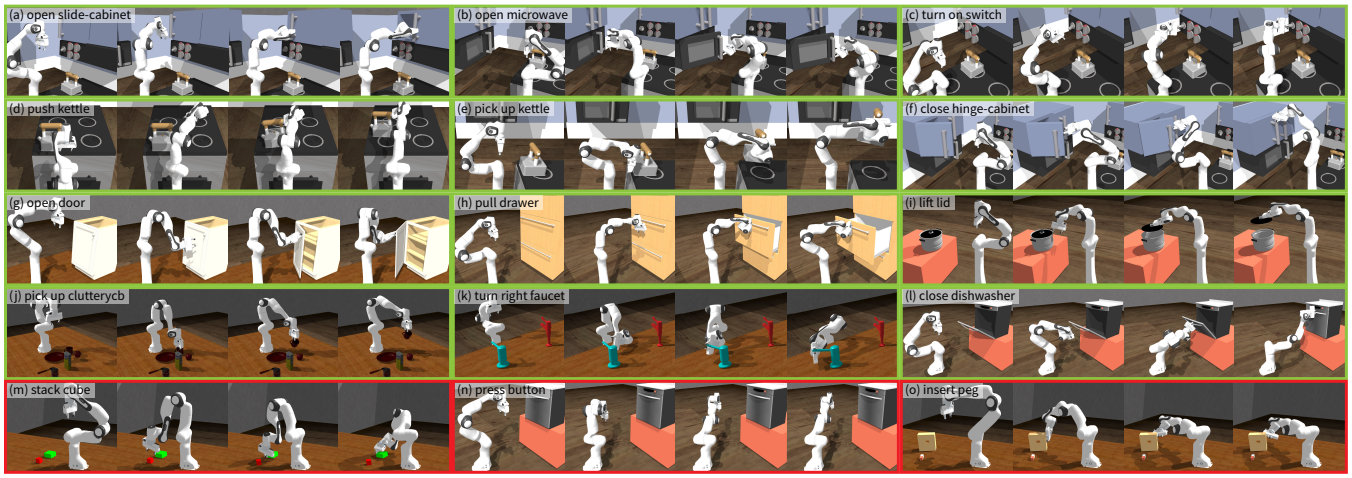

Fig. 3: **Qualitative results of our simulated experiments.** Top four rows are successful executions, and bottom row shows failures.

For instance, tasks such as opening a door or picking up a kettle, which necessitate preliminary attachments of the interested parts, often fail the baselines. Interestingly, we find similar failures occurring with Eureka. We attribute this to the lack of expert-in-the-loop interaction, which limits Eureka's ability to provide high-quality reward terms. In contrast, Ag2Manip effortlessly learns these skills benefiting from the agent-agnostic visual and action representations.

Despite these successful trials, Ag2Manip consistently encounters difficulties with three tasks: stacking cubes, inserting pegs, and pressing buttons. These failures stem from various challenges: collisions involving the robot arm during the retargeting process for stacking cubes, complex object-object interactions that fall outside the training distribution for inserting pegs, and insufficient visual guidance due to sudden and minimal appearance change for pressing buttons. Addressing these issues could involve the adoption of more advanced planning algorithms, training the visual representation on a broader range of human demonstration videos, and incorporating extra guiding elements like the end-effector's intended trajectory for improved task execution.

We also visualize some manipulation trajectories learned by Ag2Manip in Fig. 3. We showcase Ag2Manip's robust performance in manipulating rigid and articulated objects in Fig. 3 (a-l), and failure cases in Fig. 3 (m-o).

C. Ablation Study Results

Replacing the specialized reward function with a simple similarity metric (Ours w/o Rew.Shp.) resulted in a noticeable 11.1% decrease in the overall task success rate. This underscores the crucial role of our reward shaping in handling complex tasks, particularly those requiring fine maneuvers, such as turning and picking. Removing agent-agnostic action representation (Ours w/o Act.Repr.) resulted in a more pronounced decrease in success by 35.2%. This highlights how essential the agent-agnostic action representation is for Ag2Manip's effectiveness, especially in tasks requiring a firm grip, like pulling and opening. We also observe that this ablation shows a 32.4% improvement over the R3M baseline, indicating the improvement brought by our agent-agnostic visual representation.

We further assess the performance of the agent-agnostic proxy agent prior to its retargeting onto the robot (labeled as "Ours (Proxy)"). This result shows that the retargeting process caused a 7.0% drop in success rate, underscoring areas for potential improvement in our method.

TABLE II: **Task progress consistency of visual representation.**

Method	FrankaKitchen	ManiSkill	PartManip	Overall
ResNet50 [50]	$0.535^{\pm.169}$	$0.407^{\pm.182}$	$0.202^{\pm.197}$	$0.418^{\pm.199}$
CLIP [52]	$0.627^{\pm.086}$	$0.381^{\pm.139}$	$0.347^{\pm.151}$	$0.490^{\pm.134}$
R3M [8]	$0.498^{\pm.190}$	$0.393^{\pm.191}$	$0.525^{\pm.123}$	$0.474^{\pm.177}$
VIP [9]	$0.496^{\pm.246}$	$0.251^{\pm.178}$	$0.386^{\pm.121}$	$0.401^{\pm.208}$
Ag2Manip	$\mathbf{0.828}^{\pm.082}$	$\mathbf{0.696}^{\pm.182}$	$\mathbf{0.618}^{\pm.227}$	$\mathbf{0.740}^{\pm.153}$

TABLE III: **Real-world experiment results.**

Method	PushDrawer	CloseCabinet	PickBag	MoveBasket
ResNet50 [50]	$^1/_{10}$	$^5/_{10}$	$^1/_{10}$	$^1/_{10}$
CLIP [52]	$^2/_{10}$	$^3/_{10}$	$^0/_{10}$	$^0/_{10}$
R3M [8]	$^4/_{10}$	$^5/_{10}$	$^4/_{10}$	$^3/_{10}$
VIP [9]	$^6/_{10}$	$^6/_{10}$	$^2/_{10}$	$^6/_{10}$
Ag2Manip	$\mathbf{^7/_{10}}$	$\mathbf{^8/_{10}}$	$\mathbf{^8/_{10}}$	$\mathbf{^8/_{10}}$

D. Additional Evaluations for the Visual Representation

We further evaluate the quality of our agent-agnostic visual representation in two additional experiments.

Task Progress Consistency: To assess if our visual representation consistently reflects the progress of a manipulation task, we analyzed expert trajectories using the Spearman Rank Correlation [51]. This method compares the temporal order of video frames against how similar each frame is to the task's goal state. Effectively, we're checking if earlier frames are generally less similar to the goal than later frames, as expected in a successful task progression.

We compare our approach against well-established baselines, including ResNet50 [50] pre-trained for ImageNet classification, CLIP [52, 53], R3M [8], and VIP [9], all of which have found applications in robotic control. These baselines provide comparisons among visual representations trained for diverse purposes, from fundamental image recognition to specialized robotics applications. We evaluate the baselines on a total of 72 expert trajectories, with three trajectories for each of the 24 tasks used in the above experiments.

Our findings, detailed in Tab. II, show that our agent-agnostic visual representation aligns more consistently with the expected progression of manipulation tasks over time compared to these baselines. This suggests our method offers clearer, more reliable guidance for learning manipulation tasks, enhancing the robot's ability to understand and complete tasks based on visual input.

Efficiency on Real-World Imitation: In our final experiment, we assess our visual representation's effectiveness in real-world few-shot imitation learning. As shown in Fig. 4, we set up a Franka Emika FR3 robot and a Kinect Azure camera, and evaluate four manipulation tasks: PushDrawer, CloseDoor, PickBag, and MoveBasket. We collect 20 demonstrations per task for imitation learning.

We adopt the advantage-weighted regression [54] approach that focuses on transitions that significantly progress the task. This method calculates weights by comparing the current and next observations to the goal, encouraging the robot to prioritize actions that move toward task completion.

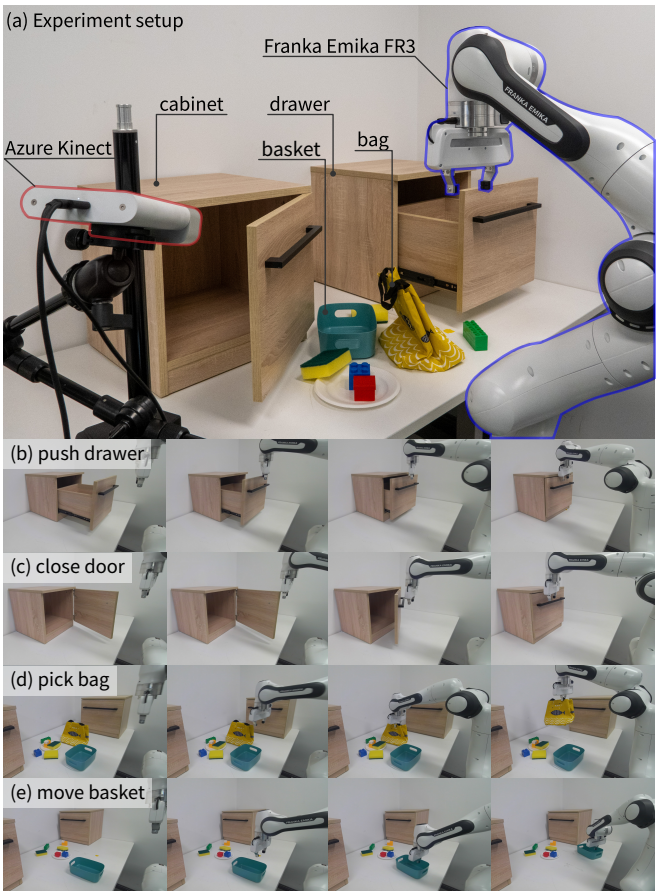

Fig. 4: **Real-world experiment setup.**

Our setup and results are detailed in Tab. III and Fig. 4. We found that our agent-agnostic visual representation more accurately assigns regression weights, leading to better performance over baselines such as ResNet50 and CLIP, which lack task-specific pre-training, and R3M and VIP, which perform relatively well except in certain tasks. Our method effectively bridges the domain gap between the training data and inference observations, capturing essential action trajectories for task completion in few-shot learning scenarios.

V. Conclusion

We present Ag2Manip as a pioneering approach for robots to acquire diverse manipulation skills autonomously, without relying on expert demonstrations. This method integrates innovative agent-agnostic visual and action representations, effectively overcoming the domain gap between different embodiments and tackling the high precision demands of robotic manipulations. Through both simulated and real-world assessments, Ag2Manip demonstrates notable enhancements in the learning of robotic manipulation skills. It facilitates the independent learning of new manipulation abilities in robots, marking a substantial stride towards creating versatile embodied agents capable of adapting to novel challenges.

REFERENCES

[1] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, *et al.*, "Rt-1: Robotics transformer for real-world control at scale," *arXiv preprint arXiv:2212.06817*, 2022.

[2] S. Bahl, A. Gupta, and D. Pathak, "Human-to-robot imitation in the wild," in *RSS*, 2022.

[3] B. Zitkovich, T. Yu, S. Xu, P. Xu, T. Xiao, F. Xia, J. Wu, P. Wohlhart, S. Welker, A. Wahid, *et al.*, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," in *CoRL*, 2023.

[4] A. Padalkar, A. Pooley, A. Jain, A. Bewley, A. Herzog, A. Irpan, A. Khazatsky, A. Rai, A. Singh, A. Brohan, *et al.*, "Open x-embodiment: Robotic learning datasets and rt-x models," *arXiv preprint arXiv:2310.08864*, 2023.

[5] C. Wang, L. Fan, J. Sun, R. Zhang, L. Fei-Fei, D. Xu, Y. Zhu, and A. Anandkumar, "Mimicplay: Long-horizon imitation learning by watching human play," *arXiv preprint arXiv:2302.12422*, 2023.

[6] H. Geng, Z. Li, Y. Geng, J. Chen, H. Dong, and H. Wang, "Partmanip: Learning cross-category generalizable part manipulation policy from point cloud observations," in *CVPR*, 2023.

[7] M. Yang, Y. Du, K. Ghasemipour, J. Tompson, D. Schuurmans, and P. Abbeel, "Learning interactive real-world simulators," *arXiv preprint arXiv:2310.06114*, 2023.

[8] S. Nair, A. Rajeswaran, V. Kumar, C. Finn, and A. Gupta, "R3m: A universal visual representation for robot manipulation," in *CoRL*, 2023.

[9] Y. J. Ma, S. Sodhani, D. Jayaraman, O. Bastani, V. Kumar, and A. Zhang, "Vip: Towards universal visual reward and representation via value-implicit pre-training," in *ICLR*, 2023.

[10] Y. J. Ma, W. Liang, G. Wang, D.-A. Huang, O. Bastani, D. Jayaraman, Y. Zhu, L. Fan, and A. Anandkumar, "Eureka: Human-level reward design via coding large language models," *arXiv preprint arXiv:2310.12931*, 2023.

[11] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, *et al.*, "The epic-kitchens dataset: Collection, challenges and baselines," *TPAMI*, vol. 43, no. 11, pp. 4125–4141, 2020.

[12] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu, *et al.*, "Ego4d: Around the world in 3,000 hours of egocentric video," in *CVPR*, 2022.

[13] A. Gupta, V. Kumar, C. Lynch, S. Levine, and K. Hausman, "Relay policy learning: Solving long-horizon tasks via imitation and reinforcement learning," in *CoRL*, 2019.

[14] J. Gu, F. Xiang, X. Li, Z. Ling, X. Liu, T. Mu, Y. Tang, S. Tao, X. Wei, Y. Yao, *et al.*, "Maniskill2: A unified benchmark for generalizable manipulation skills," in *ICLR*, 2023.

[15] Y. Xu, W. Wan, J. Zhang, H. Liu, Z. Shan, H. Shen, R. Wang, H. Geng, Y. Weng, J. Chen, *et al.*, "Unidexgrasp: Universal robotic dexterous grasping via learning diverse proposal generation and goal-conditioned policy," in *CVPR*, 2023.

[16] Y. Li, B. Liu, Y. Geng, P. Li, Y. Yang, Y. Zhu, T. Liu, and S. Huang, "Grasp multiple objects with one hand," *RA-L*, 2024.

[17] P. Li, T. Liu, Y. Li, Y. Geng, Y. Zhu, Y. Yang, and S. Huang, "Gendexgrasp: Generalizable dexterous grasping," in *ICRA*, 2023.

[18] T. Chen, M. Tippur, S. Wu, V. Kumar, E. Adelson, and P. Agrawal, "Visual dexterity: In-hand reorientation of novel and complex object shapes," *Science Robotics*, vol. 8, no. 84, p. eadc9244, 2023.

[19] Y. Chen, Y. Geng, F. Zhong, J. Ji, J. Jiang, Z. Lu, H. Dong, and Y. Yang, "Bi-dexhands: Towards human-level bimanual dexterous manipulation," *TPAMI*, 2023.

[20] Z. Zhao, Y. Li, W. Li, Z. Qi, L. Ruan, Y. Zhu, and K. Althoefer, "Tac-man: Tactile-informed prior-free manipulation of articulated objects," *arXiv preprint arXiv:2403.01694*, 2024.

[21] A. Billard and D. Kragic, "Trends and challenges in robot manipulation," *Science*, vol. 364, no. 6446, p. eaat8414, 2019.

[22] Y. Zhu, T. Gao, L. Fan, S. Huang, M. Edmonds, H. Liu, F. Gao, C. Zhang, S. Qi, Y. N. Wu, *et al.*, "Dark, beyond deep: A paradigm shift to cognitive ai with humanlike common sense," *Engineering*, vol. 6, no. 3, pp. 310–345, 2020.

[23] O. Kroemer, S. Niekum, and G. Konidaris, "A review of robot learning for manipulation: Challenges, representations, and algorithms," *JMLR*, vol. 22, no. 1, pp. 1395–1476, 2021.

[24] F. Xiang, Y. Qin, K. Mo, Y. Xia, H. Zhu, F. Liu, M. Liu, H. Jiang, Y. Yuan, H. Wang, *et al.*, "Sapien: A simulated part-based interactive environment," in *CVPR*, 2020.

[25] V. Makoviychuk, L. Wawrzyniak, Y. Guo, M. Lu, K. Storey, M. Macklin, D. Hoeller, N. Rudin, A. Allshire, A. Handa, *et al.*, "Isaac gym: High performance gpu-based physics simulation for robot learning," *arXiv preprint arXiv:2108.10470*, 2021.

[26] Y. Qin, W. Yang, B. Huang, K. Van Wyk, H. Su, X. Wang, Y.-W. Chao, and D. Fox, "Anyteleop: A general vision-based dexterous robot arm-hand teleoperation system," in *RSS*, 2023.

[27] Y. Qin, H. Su, and X. Wang, "From one hand to multiple hands: Imitation learning for dexterous manipulation from single-camera teleoperation," *RA-L*, vol. 7, no. 4, pp. 10873–10881, 2022.

[28] J. Duan, Y. R. Wang, M. Shridhar, D. Fox, and R. Krishna, "Ar2-d2: Training a robot without a robot," in *CoRL*, 2023.

[29] Y. Qin, Y.-H. Wu, S. Liu, H. Jiang, R. Yang, Y. Fu, and X. Wang, "Dexmv: Imitation learning for dexterous manipulation from human videos," in *ECCV*, 2022.

[30] L. Fan, G. Wang, Y. Jiang, A. Mandlekar, Y. Yang, H. Zhu, A. Tang, D.-A. Huang, Y. Zhu, and A. Anandkumar, "Minedojo: Building open-ended embodied agents with internet-scale knowledge," in *NeurIPS*, 2022.

[31] M. Kwon, S. M. Xie, K. Bullard, and D. Sadigh, "Reward design with language models," *arXiv preprint arXiv:2303.00001*, 2023.

[32] Y. Du, O. Watkins, Z. Wang, C. Colas, T. Darrell, P. Abbeel, A. Gupta, and J. Andreas, "Guiding pretraining in reinforcement learning with large language models," in *ICML*, 2023.

[33] P. Sermanet, K. Xu, and S. Levine, "Unsupervised perceptual rewards for imitation learning," in *RSS*, 2017.

[34] K. Schmeckpeper, O. Rybkin, K. Daniilidis, S. Levine, and C. Finn, "Reinforcement learning with videos: Combining offline observations with interaction," *arXiv preprint arXiv:2011.06507*, 2020.

[35] M. Chang, A. Prakash, and S. Gupta, "Look ma, no hands! agent-environment factorization of egocentric videos," in *NeurIPS*, 2024.

[36] R. Wu, Y. Zhao, K. Mo, Z. Guo, Y. Wang, T. Wu, Q. Fan, X. Chen, L. Guibas, and H. Dong, "Vat-mart: Learning visual action trajectory proposals for manipulating 3d articulated objects," in *ICLR*, 2022.

[37] Z. Xu, Z. He, and S. Song, "Universal manipulation policy network for articulated objects," *RA-L*, vol. 7, no. 2, pp. 2447–2454, 2022.

[38] Z. Jiang, C.-C. Hsu, and Y. Zhu, "Ditto: Building digital twins of articulated objects from interaction," in *CVPR*, 2022.

[39] H. Bharadhwaj, A. Gupta, S. Tulsiani, and V. Kumar, "Zero-shot robot manipulation from passive human videos," *arXiv preprint arXiv:2302.02011*, 2023.

[40] H. Zhang, B. Eisner, and D. Held, "Flowbot++: Learning generalized articulated objects manipulation via articulation projection," *arXiv preprint arXiv:2306.12893*, 2023.

[41] L. Shao, F. Ferreira, M. Jorda, V. Nambiar, J. Luo, E. Solowjow, J. A. Ojea, O. Khatib, and J. Bohg, "Unigrasp: Learning a unified model to grasp with multifingered robotic hands," *RA-L*, vol. 5, no. 2, pp. 2286–2293, 2020.

[42] T. Liu, Z. Liu, Z. Jiao, Y. Zhu, and S.-C. Zhu, "Synthesizing diverse and physically stable grasps with arbitrary hand structures using differentiable force closure estimator," *RA-L*, vol. 7, no. 1, pp. 470–477, 2021.

[43] S. Brahmbhatt, A. Handa, J. Hays, and D. Fox, "Contactgrasp: Functional multi-finger grasp synthesis from contact," in *IROS*, 2019.

[44] J. Xu, S. Liu, A. Vahdat, W. Byeon, X. Wang, and S. De Mello, "Open-vocabulary panoptic segmentation with text-to-image diffusion models," in *CVPR*, 2023.

[45] Z. Li, C.-Z. Lu, J. Qin, C.-L. Guo, and M.-M. Cheng, "Towards an end-to-end framework for flow-guided video inpainting," in *CVPR*, 2022.

[46] P. Sermanet, C. Lynch, Y. Chebotar, J. Hsu, E. Jang, S. Schaal, S. Levine, and G. Brain, "Time-contrastive networks: Self-supervised learning from video," in *ICRA*, 2018.

[47] J. Tan, K. Liu, and G. Turk, "Stable proportional-derivative controllers," *IEEE Computer Graphics and Applications*, vol. 31, no. 4, pp. 34–44, 2011.

[48] H.-S. Fang, C. Wang, M. Gou, and C. Lu, "Graspnet-1billion: A large-scale benchmark for general object grasping," in *CVPR*, 2020.

[49] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[50] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[51] C. Spearman, "The proof and measurement of association between two things," *The American Journal of Psychology*, vol. 100, no. 3/4, pp. 441–471, 1987.

[52] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, 2021.
[53] R. Shah and V. Kumar, "Rrl: Resnet as representation for reinforcement learning," in *ICML*, 2021.
[54] X. B. Peng, A. Kumar, G. Zhang, and S. Levine, "Advantage-weighted regression: Simple and scalable off-policy reinforcement learning," *arXiv preprint arXiv:1910.00177*, 2019.