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[4], Haoran Geng[1,5], Shu Wang[4], Yixin Zhu[3], Song-Chun Zhu[1,2,3], and Siyuan Huang[1,†]

xiaoyao-li.github.io/research/ag2manip

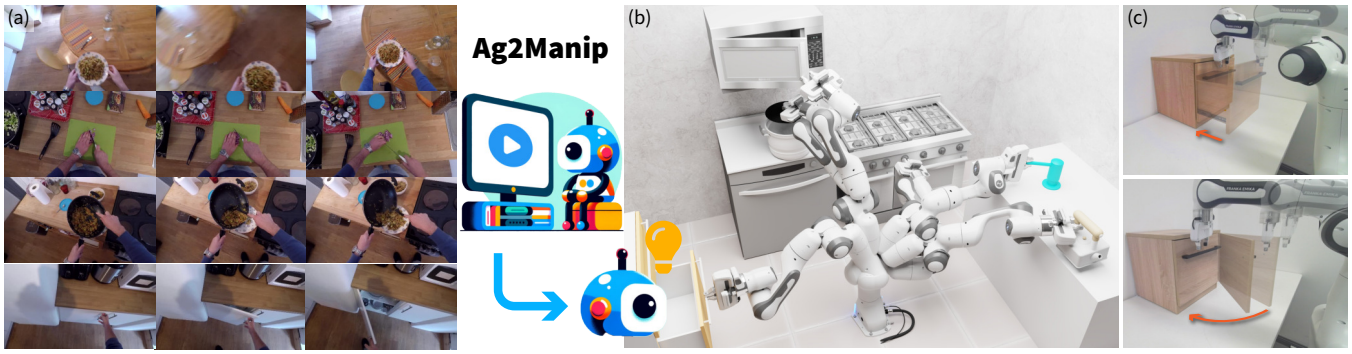

Fig. 1: **Ag2Manip enables various manipulation tasks in scenarios where domain-specific demonstrations are unavailable.** Leveraging agent-agnostic visual and action representations, Ag2Manip (a) learns from human manipulation videos, removing the reliance on domain-specific examples; (b) autonomously acquires diverse manipulation skills in simulation; and (c) facilitates robust imitation learning of manipulation skills in the real world, demonstrating the practical applicability and generalizability of our approach.

***Abstract*— Autonomous robotic systems capable of learning novel manipulation tasks are poised to transform industries from manufacturing to service automation. However, modern methods (*e.g.*, VIP and R3M) still face significant hurdles, notably the domain gap among robotic embodiments and the sparsity of successful task executions within specific action spaces, resulting in misaligned and ambiguous task representations. We introduce Ag2Manip (Agent-Agnostic representations for Manipulation), a framework aimed at surmounting these challenges through two key innovations: a novel agent-agnostic *visual representation* derived from human manipulation videos, with the specifics of embodiments obscured to enhance generalizability; and an agent-agnostic *action representation* abstracting a robot's kinematics to a universal agent proxy, emphasizing crucial interactions between end-effector and object. Ag2Manip's empirical validation across simulated benchmarks like FrankaKitchen, ManiSkill, and PartManip shows a 325% increase in performance, achieved without domain-specific demonstrations. Ablation studies underline the essential contributions of the visual and action representations to this success. Extending our evaluations to the real world, Ag2Manip significantly improves imitation learning success rates from 50% to 77.5%, demonstrating its effectiveness and generalizability across both simulated and physical environments.**

I. Introduction

Robotic systems' capability to autonomously learn and execute novel manipulation skills, without reliance on expert demonstrations, is pivotal as they adapt to changing tasks and environments. Although there have been considerable strides in the domain of learning manipulation skills [1–7], the challenge of autonomously acquiring these skills, devoid of expert guidance and task-specific rewards, persists. In addressing this issue, prior research [8–10] has investigated the use of extensive pre-training to enhance manipulation learning. Notably, recent studies [8, 9] have focused on developing comprehensive visual representations from human-centric video datasets [11, 12]. These datasets are instrumental in capturing the quintessence of tasks and the temporal dynamics between visual frames, subsequently facilitating the generation of rewards that orient robots towards fulfilling specified objectives. Alternatively, other methodologies [10] incorporate Large Language Models (LLMs) to directly craft reward functions that aid in the mastery of new manipulation skills. Despite these advancements, existing strategies often falter when confronted with intricate tasks, highlighting three principal challenges in the realm of novel skill acquisition.

First, **visual representations** derived from human-centric demonstrations [8, 9] encounter challenges in bridging the gap between the varied appearances and kinematic discrepancies of humans and robots. The appearance discrepancy introduces biases when applied to robots, undermining the models' capacity to accurately decode tasks and their temporal sequences. **Kinematic differences**, on the other hand, lead to divergent execution strategies; robots might follow trajectories that differ markedly from those in human demonstrations to accomplish tasks like picking up a cup. This variance can cause the model to erroneously classify a robot's optimal path as incorrect due to its reliance on human-centric training data.

⋆ 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.

This research is in part supported by the National Science and Technology Major Project (2022ZD0114900) and the Beijing Nova Program.

Second, the omnipresence of **human hands** in the training data **biases** these models towards prioritizing hand appearance, focusing on their position and movement over the actual task objective. For example, in tasks involving cup manipulation, the model may highlight the upward movement of the hands rather than ensuring the cup has been successfully grasped.

Last, the **demand for precision** in robotic manipulation exacerbates these challenges. Minor trajectory deviations can result in significant performance degradation. While expert-designed rewards provide detailed guidance, those derived from visual or linguistic models are often too broad and high-level, leading to inaccuracies. This issue is particularly acute in tasks requiring precise interaction with the environment, such as opening a door, where exact actions like grasping the handle are essential.

We introduce **Ag2Manip**: Agent-Agnostic representations for Manipulation to address the challenges outlined above. As depicted in Fig. 2, Ag2Manip features two primary components of generalizable visual and action representations.

To counteract the biases stemming from human-centric training data, we devise an **agent-agnostic visual representation**. Drawing inspiration from Bahl *et al*. [2], we isolate and obscure both humans and robots within video frames, subsequently inpainting the videos. Training on these agent-obscured frames, in the vein of R3M [8], our visual representation transcends the domain gap between humans and robots, fostering robust adaptation to robot-centric tasks. This agent-agnostic visual model prioritizes task processes over human-specific cues, thus providing clearer, more task-focused guidance for manipulation learning.

To mitigate inaccuracies stemming from visual guidance, we propose an **agent-agnostic action representation**. This framework abstracts robot actions into a universal proxy agent, equipped with a universally applicable action space. This representation divides manipulation learning into two phases: *exploration* and *interaction*. In *exploration*, the focus is on learning the proxy's trajectory, akin to the end-effector's movements, to enhance environment exploration. Transitioning to *interaction* when the proxy nears an object's actionable zone shifts the focus to understanding the proxy's exerted forces, simulating end-effector and object interactions. This bifurcation simplifies the learning process, reducing the complexities associated with direct robot and object manipulation. By employing this agent-agnostic action space, our method streamlines task learning, concentrating on pivotal task elements and diminishing the repercussions of sparse guidance. We further complement these representations with a well-structured reward function for each learning stage, fostering interaction and facilitating the translation of learned skills to actual robot arm movements.

Ag2Manip's effectiveness is showcased through goal-conditioned novel skill learning without expert demonstrations or task-specific rewards, across a variety of simulated tasks in FrankaKitchen [13], ManiSkill [14], and PartManip [6]. Our method achieves an impressive **78.7%** success rate, significantly outperforming baseline methods with an 18.5% success rate. By leveraging agent-agnostic visual and action representations, Ag2Manip significantly advances manipulation learning, equipping robots to adeptly navigate novel tasks in varied environments. Further validation in real-world experiments demonstrates the model's superior skill acquisition capabilities.

In summary, our work introduces three pivotal contributions to the field of learning novel manipulation skills **without expert input**: (i) an agent-agnostic visual representation that effectively narrows the embodiment gap, enhancing robotic systems' visual data interpretation; (ii) an agent-agnostic action representation that simplifies complex robot actions into more generalizable proxy-agent actions, augmented by a targeted reward function to encourage environmental interaction; and (iii) substantial progress in robot novel skill learning performance, validated across challenging tasks and affirming our approach's practical benefits in boosting robotic adaptability and autonomy.

II. Related Works

A. Learning Robotic Manipulation

The domain of robotic manipulation encompasses both foundational motor skills such as grasping [15–17] and manipulation [6, 18–21], as well as advanced cognitive capabilities for understanding task specifics, including the location, method, and reasoning behind various tasks [22–24]. The advent of parallel simulation environments [25, 26] has facilitated the learning of such skills, though this often necessitates manually tailored reward functions for each task [15, 16, 19], despite assistance from LLMs and human feedback [10]. A viable alternative involves learning from demonstrations, which bypasses the need for extensive exploration and eases scalability challenges [24]. Robot action trajectories can be captured through teleoperation [27, 28], Augmented Reality (AR) systems [29], and teach pendant programming [1, 3, 4]. However, learning from human videos is a cost-effective yet formidable approach to translating observed interactions into motor controls [28, 30], where balancing data collection costs against the quality of demonstrations poses a substantial hurdle in the direct acquisition of new skills from such sources. Inspired by recent advancements [8, 9], our study introduces generalizable **visual and action representations** for the learning of novel manipulation skills across varied tasks, leveraging the wealth of human demonstrations. This strategy seeks to address the challenges inherent in learning directly from videos, presenting a scalable and efficient solution for robotic systems to assimilate new capabilities.

B. Reward Generation for Skill Learning

Model-free Reinforcement Learning (RL) for skill learning is notably resource-intensive, primarily due to the necessity for expert-crafted, task- and embodiment-specific rewards. Addressing this issue involves devising an autonomously generated reward function for decision-making pertinent to each task. Foundation models, such as LLMs, have shown

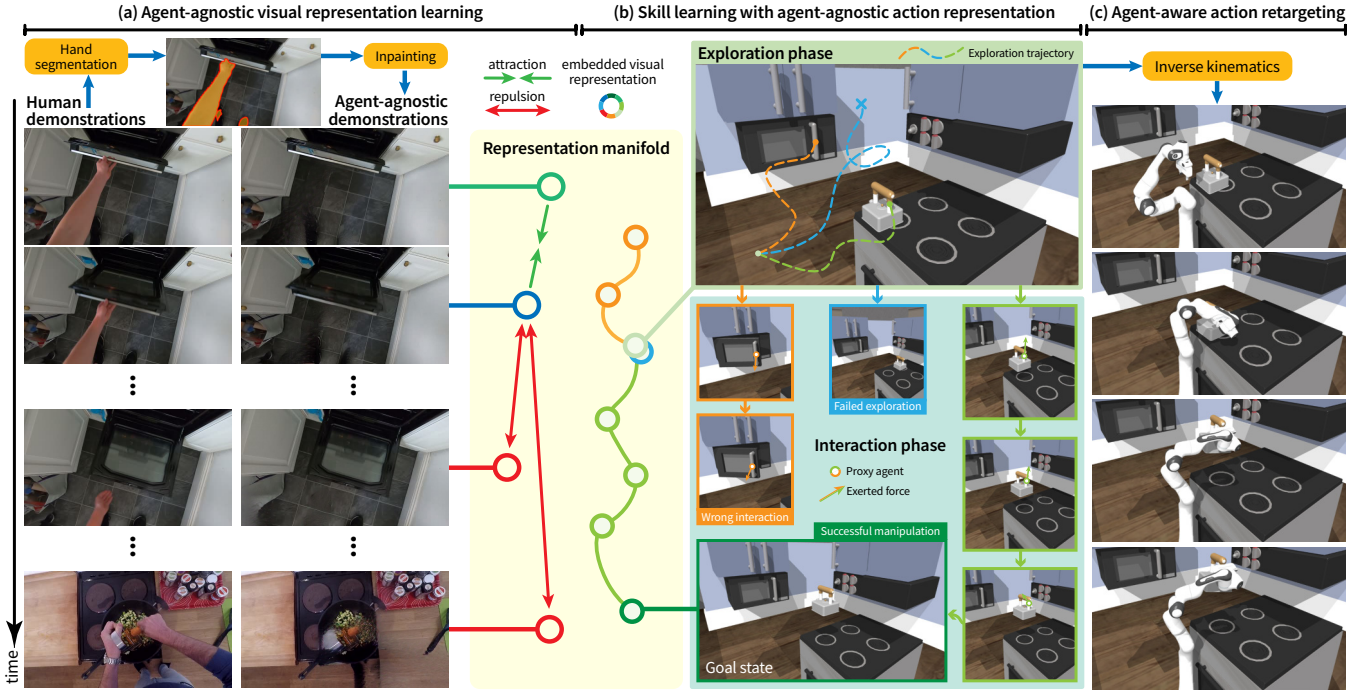

Fig. 2: **Framework of Ag2Manip.** Our approach is structured into three primary components: (a) learning an agent-agnostic visual representation, (b) learning abstracted skills via an agent-agnostic action representation, and (c) retargeting the abstracted skills to a robot.

potential in directly creating reward functions from task descriptions [10, 31–33]. However, their effectiveness is somewhat limited without environmental context, often requiring expert feedback to bridge this gap [10]. Additionally, this method's dependency on environmental states, which are usually not readily available in real-world settings, poses a significant challenge. An alternative, perceptual rewards, emerge as a promising avenue for skill learning. By observing human-executed task videos [11], robots can derive an implicit embedding that captures the sequential unfolding of events, serving as a versatile reward mechanism [8, 34, 35]. Advancing this concept, some researchers suggest learning temporal dynamics not from task-specific footage but from a diverse array of tasks, aiming to establish a task-agnostic visual representation with enhanced generalizability [9]. Our research builds on these innovations, stripping agent-specific information from the visual reward to further boost its robustness and applicability across various contexts.

C. Agent-Agnostic Representation

The concept of crafting *agent-agnostic* representations for actions, objects, and tasks serves to abstract them away from the specificities of robotic articulations or sensory setups. This approach significantly boosts adaptability and transferability across different robotic systems and even into human contexts, by separating low-level perceptual and control details in favor of focusing on high-level action abstractions. Such a framework allows for a manipulation task to be conceptualized as the desired alterations in the world state over time, minimizing the agent's direct engagement [36]. To aptly capture the nuances of agent-object interactions while maintaining agent-agnosticism, the concepts of interaction regions (often correlated with affordances) and trajectories come into play [2, 17, 37–41]. These elements illustrate task execution modalities independent of a robot's specific motor capabilities. For representing interaction zones, a straightforward yet efficacious method involves utilizing contact points to delineate essential contacts between a manipulator (*e.g.*, a finger) and an object [39, 41–43], catering well to simplistic end-effectors like parallel grippers or suction cups. In scenarios characterized by contact-rich interactions, the adoption of contact maps is indispensable for detailing the extensive contact dynamics or for accurately charting the proximity of each finger to the object surface [17, 44].

III. Method

This work delves into robotic manipulation learning in scenarios devoid of expert demonstrations. Our objective is to learn robot motions to accomplish a specified goal, given only the robot and an image of the desired end state. To this end, we introduce **Ag2Manip**: Agent-Agnostic representations for Manipulation, whose framework is depicted in Figure 2. Our methodology is underpinned by two core innovations: an agent-agnostic visual representation (Sec. III-A) that mitigates the domain disparity between humans and robots, and an agent-agnostic action representation (Sec. III-B) that distills robot actions to those of a universal proxy agent. These foundations enable us to harness RL to formulate a manipulation policy within this generalized action space, informed by a novel reward function emerging from our agent-agnostic visual paradigm (Sec. III-C). Subsequently, the trajectory devised for the proxy agent is adapted to the robot through Inverse Kinematics (IK) (Sec. III-D), ensuring the practical applicability of the learned maneuvers.

A. Agent-Agnostic Visual Representation

Our work seeks to develop an agent-agnostic visual representation that transcends the domain gap between human and robot manipulations, building on pre-trained visual representations on human demonstrations [8, 9]. This approach aims to augment the versatility and effectiveness of these representations within robotic contexts, facilitating a more adaptable skill acquisition process.

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 [45]. Following segmentation, we employ a video inpainting model, E^{2}FGVI [46], 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 [47] $\mathcal{L}_{\text{tcn}}$ 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

Our approach encapsulates robotic manipulation learning via an agent-agnostic action representation. This involves abstracting a robot's movements and interactions into those of a universal, free-floating proxy agent, encompassing both motion and exerted forces. The learning process is bifurcated into two stages: *exploration*, concentrating on the proxy's positional adjustments, and *interaction*, focusing on the forces applied by the proxy on the environment. An RL policy is developed to minimize the embedding distance in the agent-agnostic visual representation space between the current state and a goal state depicted by an image.

The *exploration* phase: The robot is abstracted as a universal proxy agent, represented by an agent-agnostic sphere to mimic the end-effector's actions, translating the robot's actions into a sequence of positions for this sphere. Control over the proxy is established through a proportional-derivative (PD) controller [48], with the proxy embodying a collision volume of radius r_{e} to denote its physical presence. This phase concludes upon the proxy's arrival at a precalculated interactable region within the environment, marking the commencement of the *interaction* phase. For the scope of this work which focuses on robots equipped with two-finger grippers, interactable regions are identified as zones where parallel grips are deemed feasible. Utilizing point cloud scans of the environment, these regions are determined based on proximity to potential gripping points identified by GraspNet [49]. Although parallel grip detection is utilized for its efficiency in our setup, general-purpose methods like GenDexGrasp [17] could also delineate interactable regions suitable for a range of dexterous manipulations.

The *interaction* phase: With the proxy's entry into an interactable region, indicating a viable grasp and subsequent object attachment, the focus shifts to the *interaction* phase. This stage is dedicated to the manipulation of the object, abstracting the robot's actions into the forces the proxy exerts upon the environment.

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 Goal-Conditioned Reinforcement Learning (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 $\beta = \mathcal{S}(z_0, z_g)$ denotes the similarity between the embeddings of the start and goal images, and $\alpha > 0$ is an tunable hyperparameter. This reward function, with its indicator function, prioritizes states closer to the goal relative to the starting point and lessens the penalty for deviations, thus promoting exploration beneficial in the policy's early phase of learning with random policy behaviors.

For policy optimization, we utilize Proximal Policy Optimization (PPO) [50], 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

To facilitate the transition of the proxy's trajectory, as determined by π, into actionable movements for real robots, we employ a retargeting policy that translates proxy actions into robot-specific actions. During the *exploration* phase, the positions of the proxy agent are directly mapped to the robot's end-effector positions, thereby converting the proxy's navigational path into corresponding end-effector motions. As the process shifts from *exploration* to *interaction*, the end-effector's 6D pose is adjusted to align with the nearest viable grasp pose as identified by GraspNet, an approach that is feasible because this transition is predicated on the proximity of an achievable grasp. In the *interaction* phase, the movement of the object dictates the end-effector's 6D trajectory, ensuring the robot's actions remain in harmony with the object's dynamics. The trajectory for the robot arm is calculated using IK, aligning the practical task execution with the proxy's intended actions.

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 [51] 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 optimization of the visual encoder is carried out using an Adam optimizer with a learning rate of 10^{-4}, over a duration of 24 hours on a single NVIDIA A100 GPU. In Sec. III-B, the collision and interactive region radii are defined as 2 centimeters (r_{e}) and 10 centimeters (r_{int}). For the reward shaping in Sec. III-C, $\alpha=3.0$ is empirically determined as the hyperparameter of the reward function across all tasks.

IV. Simulations and Experiments

Our comprehensive evaluation of the proposed Ag2Manip showcases significant enhancements in terms of task success rates, achieving a leap from a baseline success rate of 18.5% to an impressive 78.7% across tasks sourced from three different environments. Furthermore, our visual representation contributes to a marked increase in the success rate of imitation learning, which increases from 50% to 77.5%. These advancements highlight the Ag2Manip's effectiveness and its considerable promise for real-world applications.

A. Simulation Setup

Environments: To assess the broad applicability of the proposed Ag2Manip across various manipulation tasks, we select 24 distinct tasks from three varied simulation environments. FrankaKitchen [13], ManiSkill [14], and PartManip [6]. These tasks span a wide range of actions, including opening, pulling, and moving, and involve interactions with a variety of objects like cabinets, microwaves, and kettles, executed using a 9-DOF Franka Emika robotic arm and gripper. This setup typifies a standard in robotic manipulation.

Experiments are conducted within the NVIDIA IsaacGym, leveraging its GPU acceleration for efficient RL-based learning. The robot initiates each task from a standardized default position, with task objectives defined by goal states represented by images rendered from one of three predetermined camera perspectives (front, left, right). Success in a task is determined by the object or component reaching its goal state within a predefined error margin. To ensure a thorough evaluation, each of the 24 tasks undergoes testing in 9 varied setups combining different camera angles and initialization seeds (3 cameras × 3 seeds), providing a comprehensive overview of performance across multiple conditions.

Baselines: Our approach is compared against two baselines, R3M [8] and VIP [9], which utilize agent-aware visual representations and time-contrastive learning objectives for learning manipulation skills. Eureka, a novel method distinguished for its ability to autonomously generate reward functions via LLMs, also stands as a significant benchmark and highlights its strengths in skill learning.

TABLE I: **Comparisons and ablation studies**. Each task was evaluated over 3 seeds × 3 cameras = **9 runs**, with the numbers **0 – 9** indicating the count of successful attempts. The characters **`a` – `x`** denote specific tasks. Tasks from FrankaKitchen [13] include: **`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, and **`j:`** turn off switch. Tasks from ManiSkill2 [14] include: **`k:`** open door, **`l:`** close door, **`m:`** pick up cube, **`n:`** stack cube, **`o:`** pick up clutterycb, **`p:`** insert peg, **`q:`** turn left faucet, and **`r:`** turn right faucet. Tasks from PartManip [6] include: **`s:`** turn down dishwasher, **`t:`** pull drawer, **`u:`** turn up dishwasher, **`v:`** push drawer, **`w:`** press button, and **`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%

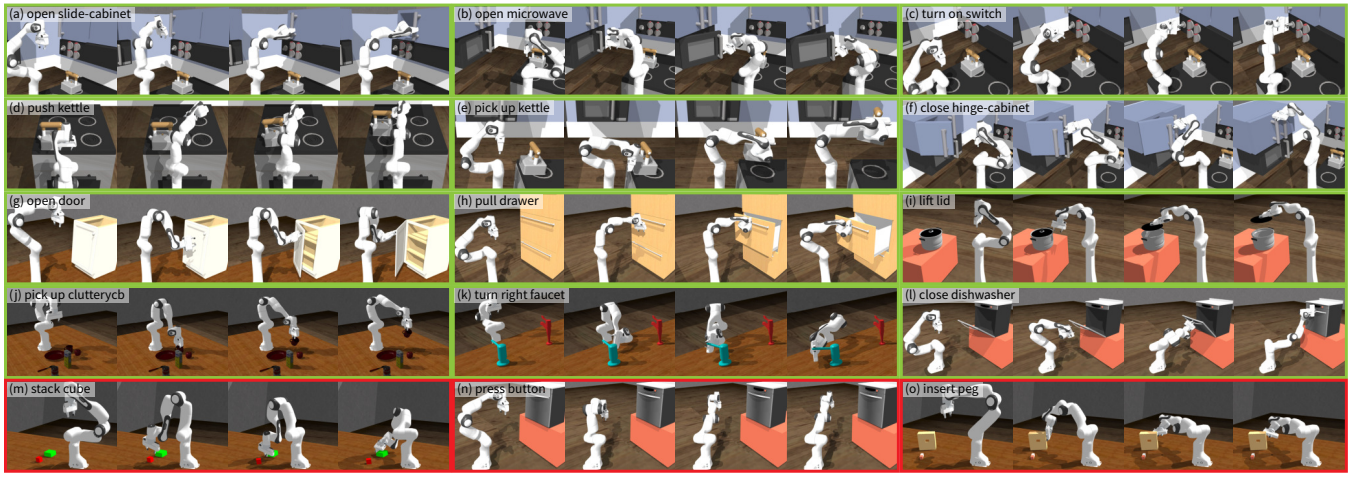

Fig. 3: **Qualitative results in simulation.** The top four rows are successful executions, whereas the bottom row shows failures.

For equitable comparison, all methods, barring Eureka, are built upon a ResNet50 architecture and trained using the Epic-Kitchen dataset. To eliminate the influence of task-specific expert insights, Eureka's human feedback feature was deactivated, ensuring that the evaluation focuses solely on each method's intrinsic learning capabilities.

Ablations: Our ablation study delineates the impact of distinct components by excluding them from our method. *Ours w/o Act.Repr.* investigates learning directly within the robot's native action space while retaining the agent-agnostic visual representation. Conversely, *Ours w/o Rew.Shp.* employs a straightforward similarity metric instead of our tailored reward function. The removal of solely the visual representation was not considered, given the impracticality of computing agent-aware visuals without corresponding actions. Similarly, excluding both representations would essentially replicate the R3M baseline. Additionally, *Ours (Proxy)* examines the efficacy of the proxy agent's performance devoid of action retargeting to a robot, thereby assessing the impact of retargeting on performance.

B. Results: Simulation

The summarized results in Tab. I detail the average task success rates within each of the three environments and cumulatively. Ag2Manip emerges as a standout, securing an overall task success rate of **78.7%**, markedly surpassing the baseline methods which recorded success rates of 11.1%, 12.0%, and 18.5%. Further dissecting the success rates per task reveals the distinct competencies of each method. Notably, baseline approaches underperform in tasks demanding precise robot-object interactions, such as door opening or kettle lifting, which require initial attachment actions that often elude the baselines. Eureka exhibits similar shortcomings, which we ascribe to the absence of expert-in-the-loop feedback, consequently affecting its ability to generate refined reward signals. In contrast, Ag2Manip adeptly acquires these challenging skills, benefitting from its foundational agent-agnostic visual and action representations.

Nonetheless, Ag2Manip does encounter consistent challenges with specific tasks: cube stacking, peg insertion, and button pressing. These difficulties arise from a range of issues including collision occurrences with the robot arm in cube stacking, complex object interactions beyond the training set's scope for peg insertion, and the lack of substantial visual cues for button pressing due to minor appearance changes. Potential resolutions could entail integrating more sophisticated planning methods, broadening the scope of human demonstration videos for training the visual representation, and incorporating more guiding elements like the anticipated trajectory of the end-effector to refine task performance.

Additionally, Fig. 3 illustrates some of the manipulation trajectories learned by Ag2Manip, demonstrating its efficacy in handling both rigid and articulated objects across Fig. 3 (a-l), and delineating instances of failure in Fig. 3 (m-o).

C. Results: Ablation

Substituting our meticulously crafted reward function with a basic similarity metric (*Ours w/o Rew.Shp.*) led to an 11.1% reduction in overall task success rates. This significant decline accentuates the pivotal role our reward shaping plays in facilitating the completion of intricate tasks, particularly those necessitating precise movements like turning and lifting. The omission of the agent-agnostic action representation (*Ours w/o Act.Repr.*) had an even more marked effect, with a 35.2% drop in success, underscoring its critical contribution to Ag2Manip's performance in tasks that demand accurate control, such as pulling and opening. Notably, even with this reduction, this configuration still outperforms the R3M baseline by 32.4%, highlighting the value added by our agent-agnostic visual representation.

Examining the performance of our agent-agnostic proxy agent before retargeting its actions to a robot (*Ours (Proxy)*) revealed that the retargeting step accounts for a 7.0% decrease in success rates. This finding points to the retargeting phase as a promising avenue for further enhancing Ag2Manip's effectiveness.

D. Visual Representation: Task Progress Consistency

To verify the consistency of our visual representation in mirroring the progression within a manipulation task, we

employed the Spearman Rank Correlation [52] to analyze expert trajectories. This approach compares the temporal sequence of video frames with their respective similarities to the task's goal state, aiming to ascertain whether initial frames generally exhibit lesser similarity to the goal than subsequent frames, indicative of coherent task advancement.

The proposed Ag2Manip is benchmarked against several established baselines, such as a ResNet50 [51] model pre-trained on ImageNet for general image classification, CLIP [53, 54], R3M [8], and VIP [9]. These models span a range of applications, from basic image recognition to robotic control tasks, offering a broad spectrum for comparative analysis. The evaluation encompassed 72 expert trajectories—three per task—for the 24 tasks delineated in prior experiments.

According to the results tabulated in Tab. II, our agent-agnostic visual representation demonstrates a higher consistency with the logical task progression over time, surpassing the baseline models. This implies that our approach provides more accurate and dependable cues for task learning, thereby improving the robot's comprehension and execution of tasks through visual guidance.

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

Method	FrankaKitchen	ManiSkill	PartManip	Overall
ResNet50 [51]	$0.535^{\pm.169}$	$0.407^{\pm.182}$	$0.202^{\pm.197}$	$0.418^{\pm.199}$
CLIP [53]	$0.627^{\pm.086}$	$0.381^{\pm.139}$	$0.347^{\pm.151}$	$0.490^{\pm.134}$
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	$\mathbf{0.828^{\pm.082}}$	$\mathbf{0.696^{\pm.182}}$	$\mathbf{0.618^{\pm.227}}$	$\mathbf{0.740^{\pm.153}}$

E. Visual Representation: Experiments on Imitation

This experiment aims to evaluate the effectiveness of our visual representation in real-world few-shot imitation learning scenarios. Utilizing a Franka Emika FR3 robot and a Kinect Azure camera, as depicted in Fig. 4, we explore four manipulation tasks: PushDrawer, CloseDoor, PickBag, and MoveBasket. For each task, we gather 20 demonstrations to facilitate the imitation learning process.

We implement advantage-weighted regression [55] for this experiment, a strategy that accentuates transitions contributing significantly to task progression. This approach assigns weights by assessing the similarity between consecutive observations and the task's goal state, thereby incentivizing actions that evidently advance toward task completion.

The specifics of our experimental setup and the results are shown in Tab. III and Fig. 4. Our findings indicate that the agent-agnostic visual representation notably outperforms the baselines, including ResNet50 and CLIP, which do not undergo task-specific pre-training, as well as R3M and VIP, which exhibit commendable performance barring certain exceptions. Our approach demonstrates a superior capability in narrowing the domain gap that often exists between training datasets and real-world observations, capturing the critical action trajectories necessary for successful task execution within a few-shot learning framework.

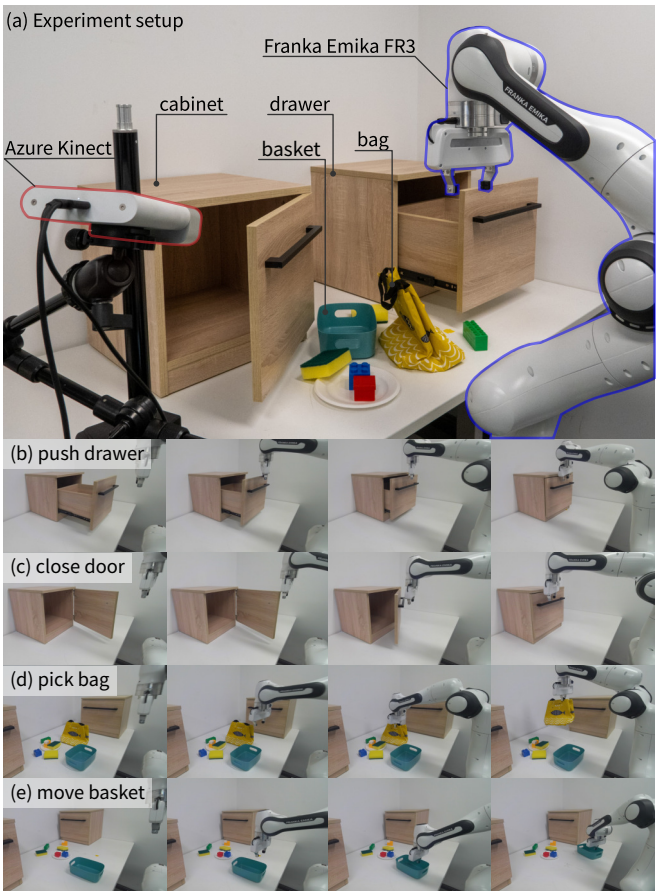

Fig. 4: **Experimental setup.**

TABLE III: **Experimental results.**

Method	PushDrawer	CloseCabinet	PickBag	MoveBasket
ResNet50 [51]	$^1/_{10}$	$^5/_{10}$	$^1/_{10}$	$^1/_{10}$
CLIP [53]	$^2/_{10}$	$^3/_{10}$	$^0/_{10}$	$^0/_{10}$
R3M [8]	$^4/_{10}$	$^5/_{10}$	$^4/_{10}$	$^3/_{10}$
VIP [9]	$^6/_{10}$	$^6/_{10}$	$^2/_{10}$	$^6/_{10}$
Ag2Manip	$\mathbf{^7/_{10}}$	$\mathbf{^8/_{10}}$	$\mathbf{^8/_{10}}$	$\mathbf{^8/_{10}}$

V. Conclusion

In this work, we introduced Ag2Manip, a novel framework enabling robots to acquire a wide array of manipulation skills without the necessity of expert demonstrations. Our method is grounded in the development of novel agent-agnostic visual and action representations, designed to bridge the domain disparities between various robot embodiments and address the intricate precision requirements inherent in robotic manipulations. Evaluated through extensive simulations and real-world experiments, Ag2Manip has proven to significantly improve the process of learning robotic manipulation skills, underscoring its effectiveness in facilitating autonomous skill acquisition in robots. This achievement represents a significant leap towards the realization of versatile embodied agents equipped to navigate and adapt to new challenges seamlessly.

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] H. Geng, H. Xu, C. Zhao, C. Xu, L. Yi, S. Huang, and H. Wang, "Gapartnet: Cross-category domain-generalizable object perception and manipulation via generalizable and actionable parts," *arXiv preprint arXiv:2211.05272*, 2022.

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

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

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

[30] 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.

[31] 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.

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

[33] 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.

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

[35] 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.

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

[37] 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.

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

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

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

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

[42] 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.

[43] 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.

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

[45] 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.

[46] 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.

[47] 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.

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

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

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

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

[52] 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.

[53] 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.

[54] R. Shah and V. Kumar, "Rrl: Resnet as representation for reinforcement learning," in *ICML*, 2021.

[55] 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.