

Bi-KVIL: Keypoints-based Visual Imitation Learning of Bimanual Manipulation Tasks

Jianfeng Gao, Xiaoshu Jin, Franziska Krebs, Noémie Jaquier, and Tamim Asfour

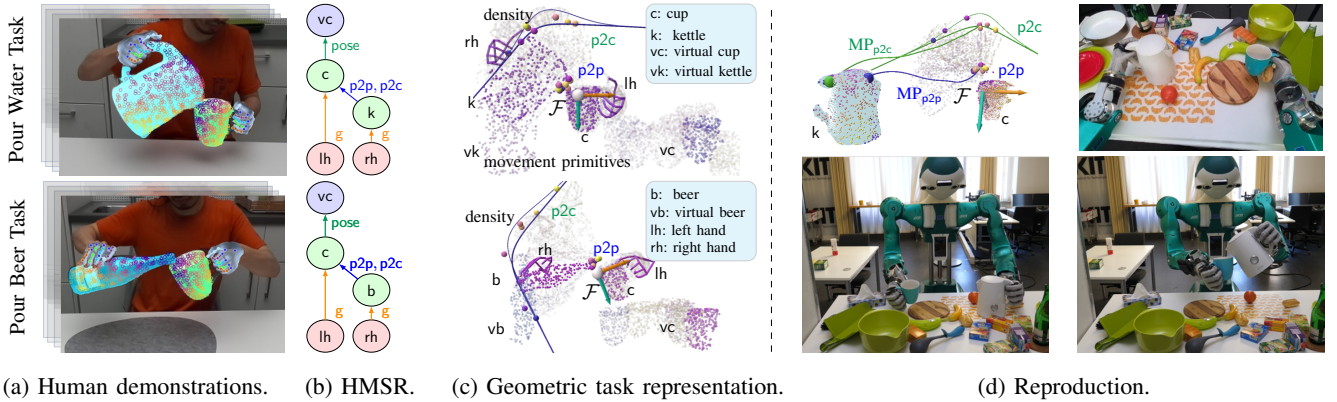


Fig. 1: Overview of Bi-KVIL. (a) Human demonstration videos of the pouring tasks are collected with different styles and pose variations of categorical objects. (b) For each task, we abstract the hand/object relationships into a symbolic *Hybrid Master-Slave Relationship* (HMSR) with (c) sub-symbolic geometric constraints for each master-slave pair to model motion styles. (d) The learned tasks are then reproduced with category-level generalization in cluttered scenes by ARMAR-6.

Abstract—Visual imitation learning has achieved impressive progress in learning unimanual manipulation tasks from a small set of visual observations, thanks to the latest advances in computer vision. However, learning bimanual coordination strategies and complex object relations from bimanual visual demonstrations, as well as generalizing them to categorical objects in novel cluttered scenes remain unsolved challenges. In this paper, we extend our previous work on keypoints-based visual imitation learning (K-VIL) [1] to bimanual manipulation tasks. The proposed Bi-KVIL jointly extracts so-called *Hybrid Master-Slave Relationships* (HMSR) among objects and hands, bimanual coordination strategies, and sub-symbolic task representations. Our bimanual task representation is object-centric, embodiment-independent, and viewpoint-invariant, thus generalizing well to categorical objects in novel scenes. We evaluate our approach in various real-world applications, showcasing its ability to learn fine-grained bimanual manipulation tasks from a small number of human demonstration videos. Videos and source code are available at <https://sites.google.com/view/bi-kvil>.

I. INTRODUCTION

Bimanual manipulation is key to human everyday activities while being significantly more complex than the simple sum of two unimanual tasks. Similarly to the unimanual case, bimanual manipulation tasks are usually characterized by invariant task features over several demonstrations [2],

[3]. For instance, pouring tasks are characterized via the alignment of the mouth of the container to the rim of the cup (see Fig. 1a). In [1], we introduced the Keypoint-based Visual Imitation Learning (K-VIL) framework that leverages this principle to automatically extract sparse, object-centric, and embodiment-independent task representations from few human demonstration videos. However, K-VIL is limited to unimanual tasks with a single master-slave pair and a static master object. Instead, bimanual tasks often involve more than two objects, as well as more complex master-slave relationships since each master object may itself be a motion-salient slave object paired to another master object.

In this paper, we build on our previous work [1] and propose Bi-KVIL, an approach for learning bimanual task representations that capture all relevant temporal and spatial constraints between the hands/objects (see Figs. 1b and 1c). To this end, it is important to understand their roles and relationships in the demonstrated bimanual tasks. Early works focused on hand/arm relationships but overlooked the role of objects. For example, *dominant* and *non-dominant* hands (or arms) are used in [4], [5] to describe their roles in asymmetrical bimanual tasks. Specifically, the non-dominant hand often stabilizes the object and sets a frame of reference defining the motion of the dominant hand. In robotics, the leader-follower [6], [7] and master-slave [1], [8] relationships are also widely used to design control policies for the slave/follower arm within a local frame defined on the master/leader arm. In this paper, we adopt the master-slave relationship (MSR) naming convention following [1], [8],

This work has been supported by the Carl Zeiss Foundation through the JuBot project.

The authors are with the Institute for Anthropomatics and Robotics, Karlsruhe Institute of Technology, Karlsruhe, Germany. E-mails: {jianfeng.gao, asfour}@kit.edu

Accepted in IEEE International Conference on Robotics and Automation (ICRA) 2024.

and extend it from arm coordination to object relationships, while considering the human hands as a special type of object. As a result, we unify the representation of the *roles*, *relationships*, and *task constraints* for both objects and hands. The bimanual manipulation categories [9] of the demonstrated tasks are then derived from the extracted MSR (see Section IV-C). Overall, Bi-KVIL unifies the learning of object-centric uni- and bimanual manipulation tasks, and captures fine-grained manipulation styles. To the best of our knowledge, this work is the first to simultaneously extract bimanual coordination strategies and generalizable geometric task constraints from few (~ 5 -10) visual demonstrations.

The contributions of this paper are twofold: (i) We propose Bi-KVIL for learning bimanual manipulation tasks from a small number of visual demonstrations. Bi-KVIL automatically extracts a *Hybrid Master-Slave Relationship* (HMSR), the corresponding bimanual coordination strategy, and sub-symbolic task representations (see Section IV). These representations include keypoints-based geometric constraints on principal manifolds, their associated local frames, and movement primitives (see Section III); (ii) We present the bimanual keypoint-based admittance controller (Bi-KAC) extended from KAC [1] to handle a set of prioritized geometric constraints for bimanual tasks (see Section V). It allows the reproduction of bimanual tasks corresponding to the bimanual manipulation taxonomy introduced in [9].

II. RELATED WORK

Learning fine-grained bimanual tasks from visual observation of human demonstrations is a long-standing goal in robotics. It combines challenges in computer vision, bimanual coordination, and control. Most previous works focus on one or a few aspects of the problem.

A. Visual Imitation Learning

VIL has made impressive progress thanks to the advances in deep-learning-based computer vision algorithms. Perception pipelines [10], [11] are used to obtain poses of hands and objects from visual demonstrations, which are then used to train reinforcement learning (RL) algorithms for motion policies. Despite their performance, generalization capabilities are not guaranteed in semantic manipulation [12] when objects have large shape variations. To improve category-level generalization, visual object descriptors based on image features [13]–[19] were proposed to find dense correspondences between categorical objects, thus facilitating category-level adaptation of downstream object-centric manipulation skills. Similarly, SE(3)-equivariant object shape features [20]–[22] and space coverage features [23], [24] were proposed to cope with partially-observed object point-clouds. However, VIL based on such features requires manually-annotated keypoints for training [25], [26] and inference [21]. To address this issue, we adopted image features from [13] and proposed a *Principal Constraint Estimation* (PCE) algorithm to automatically extract keypoints-based geometric constraints from demonstrations [1]. This approach outperforms data-driven methods [27] in terms of the number of

demonstrations and category-level generalization. However, considering bimanual coordination, it is crucial to apply any of these approaches to bimanual tasks.

B. Imitation Learning of Bimanual Manipulation

Many works on bimanual manipulation focus on designing controllers coping with known coordination categories [28]–[35] rather than learning the coordination strategies from demonstrations. Such strategies can be either implicitly encoded in the motions or explicitly represented as constraints.

1) *Implicit coordination*: Trajectory-based bimanual imitation learning focuses on learning the spatio-temporal correlations of bilateral motions with different variations of movement primitives [36]–[39] or Transformer-based models [7]. This implicit encoding of coordination strategies overlooks the roles of objects in the task, thus limiting generalization abilities compared to object-centric VIL approaches. Coordination strategies are also implicitly encoded in bimanual deep imitation learning [40]–[46], which additionally requires many demonstrations that are not always available in the real world. Despite the success of their reactive controllers within the trained scenes, these approaches lack generalization abilities as they do not explicitly encode coordination strategies and constraints. In contrast, Bi-KVIL only requires 5-10 demonstrations and improves generalization by explicitly extracting coordination strategies and task constraints.

2) *Explicit coordination*: Abstracting a representation of a coordinated behavior often involves analyzing the contact and grasp state, the role of the objects/hands, as well as the spatio-temporal and force constraints. A rule-based classification was proposed in [9] to determine the category of bimanual actions defined by the bimanual manipulation taxonomy. Other works mainly focused on object-action relation [47] or on learning a specific coordination strategy [8]. In this paper, we focus on spatio-temporal constraints, as force data required by [8] is not available in demonstration videos. Specifically, we unify the bimanual coordination categories of [9] in our MSR representation and controller. Moreover, Bi-KVIL relaxes the need for predefined frames per object as in [8], and combines automatic extraction of MSR, bimanual coordination, and object-centric task representations, thus enabling generalizable fine-grained skills.

III. BACKGROUND

Here, we briefly review the K-VIL framework [1], from which Bi-KVIL is derived. Given a set of human demonstration videos (see Fig. 1a) with different categorical objects (e.g., different cups in Fig. 2a), K-VIL detects and tracks dense correspondence points, i.e., candidate points, on the visible surface of the objects using Dense Object Net (DON) [13]. The motion-salient object is considered a slave object. Candidate local frames are determined by matching the neighboring points between the canonical shape and actual point cloud of the master object (see Fig. 2a). We then align all demonstrations to each candidate local frame on the master object. This allows the geometric constraints, i.e., spatial invariances, to become salient (see Fig. 2b) and to

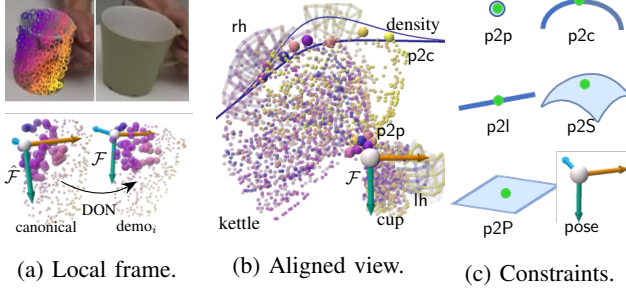


Fig. 2: Illustration of K-VIL on a cup-kettle master-slave pair.

be computed using the Principal Manifold Estimation (PME) algorithm [48]. In this way, the extracted local frame \mathcal{F} , geometric constraints (p2p and p2c in Fig. 2b), and associated via-point movement primitives (VMPs) [49] define the motion of the slave object with respect to the master object. As shown in Fig. 2c, K-VIL considers, in priority order, point-to-point (p2p), point-to-line (p2l), point-to-plane (p2P), point-to-curve (p2c), and point-to-surface (p2S) geometric constraints. Each keypoint is then driven by a spring-damper system following the VMPs and constraints via a keypoint-based admittance controller (KAC). The composed force finally drives the robot hand to reproduce the task. Note that K-VIL focuses on a single master-slave pair, which corresponds to each master-slave pair in the Bi-KVIL HMSR graph (see Fig. 1b and Fig. 3).

IV. BIMANUAL KEYPOINT-BASED VISUAL IMITATION LEARNING (BI-KVIL)

This section presents the Bi-KVIL approach. We first introduce the perception pipeline for preprocessing demonstration videos in Section IV-A. We then detail the proposed master-slave relationship and its extraction in Section IV-B.

A. Preprocessing

(Bi-)K-VIL rely on robust estimation and tracking of dense candidate points of the objects. To this end, we compose various off-the-shell computer vision algorithms into a reliable perception pipeline. We support videos taken by stereo RGB or monocular RGB-D cameras from different viewpoints. We prefer the stereo approach for translucent or thin objects, where the depth is estimated using UniMatch [50].

1) *Candidate points on objects*: Similarly to K-VIL, we train DONs per object category in a self-supervised and task-agnostic manner. To improve the data quality, we replace the traditional 3D reconstruction in DON with a modified Instant-NGP following [51], [52]. We select the first image frame of a random demonstration to create the canonical space of all objects (see Fig. 2a). We use DON to find the initial dense correspondence points for each object *only* on the first image frame of any other demonstrations, and the deep optical flow algorithm RAFT [53] to track the motion of these points in image coordinates. To further restrict the results of DON and RAFT within the region of the objects, we employ the Segment [54] and Track [55] Anything models in combination to the object detection

model Grounding DINO [56]. Finally, we map the motion of the candidate points from image coordinates to 3D using triangular geometry, remove outliers, and smooth the motions with Savitzky-Golay filters.

2) *Human pose estimation*: In addition to candidate points on objects, Bi-KVIL requires keypoints of the human hands and the handedness, i. e., the left/right label of each hand. In natural visual demonstrations, human hands are often heavily (self-)occluded in several image frames, where methods like MediaPipe [57] fail. We found that RTMPose [58] robustly estimates the whole-body human pose in 2D including the handedness, which allows us to map different sub-tasks to the robot's hands. The image patches containing the detected hands and their handedness are used to obtain 3D hand poses using MeshGraphormer [59]. We re-base the hand mesh to the most probable visible keypoint of the hand using object-hand mask overlay and pre-defined priority. We empirically observed that our framework outperforms other models, e. g., MediaPipe 3D, OSX [60] when under heavy (self-)occlusion.

It is important to note that we assume the human demonstrations to be temporally segmented, so we focus on extracting the HMSR and K-VIL's task representation for each motion segment. We do not include evaluations of different computer algorithms, since this is not the focus of our paper.

B. Extraction of Master-Slave Relationships (MSRs)

Given the 3D trajectories of dense points on the objects and hands aligned in the camera frame, we first analyze the object/hands relationships. We propose five types of MSRs that often appear in unimanual and bimanual manipulation tasks, namely, *single*, *multiple*, *multi-level*, *hierarchical*, and *hybrid* MSR, which resemble the definition and graph representation of the inheritance in C++ programming language (see Fig. 3). The single MSR corresponds to the unimanual K-VIL case [1], where only two objects interact within a single master-slave pair. Within multiple MSR, the motion of a slave object is defined in local frames of multiple master objects. In multi-level MSR, a slave object can be the master of another slave object. A hierarchical MSR is a tree-like structure where each master may have multiple slave objects. Finally, the hybrid MSR (HMSR) combines multiple and hierarchical MSRs, i. e., it is a directed acyclic graph (DAG) in which a slave object may have master objects at different levels (e. g., s_1 in Fig. 3e). The HMSR differs from the inheritance in C++ in that a slave object does not inherit geometric constraints from its master. Instead, constraints are explicitly defined between each object pair (see Fig. 3e). In the following, we use HMSR as a general framework that encompasses all other MSRs. To extract HMSR, we first build a rough DAG using motion-saliency, grasping, and pose invariance detection in Sections IV-B.1 to IV-B.3. Since each valid master-slave pair in this graph must have at least one of the constraints in Fig. 2c, we use K-VIL in Section IV-B.4 to truncate master-slave pairs without constraints and finally obtain a compact graph.

1) *Absolute motion saliency detection*: Since any motion must be represented in a local frame, the top-level master

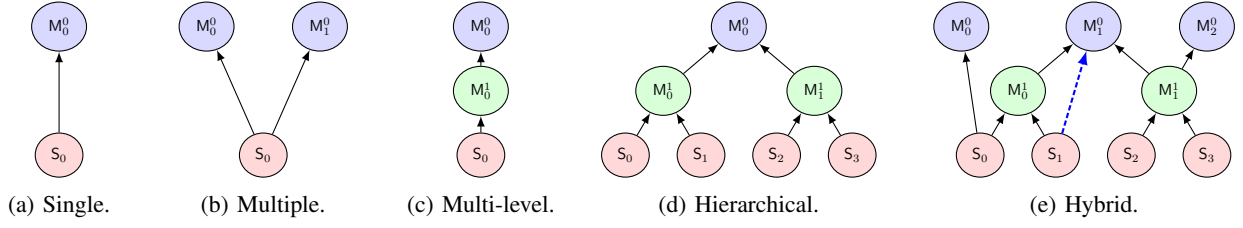


Fig. 3: MSR diagrams. M_j^i represents the j -th master object at level i . For level $i > 0$, a master object is itself a slave object paired to another master object at level $i - 1$. The slave objects are located at the lowest level.

objects in HMSR must be static. We use motion saliency to determine *static* (if any) and *moving* objects based on the average point velocities in the global camera frame. In the absence of static objects, we define local frames on the *initial state* of the moving objects, in which the constraints and its motion following that state are modeled (see Section IV-B.4). We define the initial state, i.e., the observed point cloud at the first timestep, of each moving object as a special object, called *virtual object*, in HMSR (see, e.g., Fig. 1b).

2) *Grasp detection based on relative motion saliency*: Before determining bimanual coordination strategies, we need to estimate the grasping relationships between hands and objects. Similarly to K-VIL, the human hand is considered a special type of object with 21 keypoints (see MANO [61] model). We detect contacts between two objects based on the spatial distances between all candidate point pairs. Additionally, we compute the average change rate of the absolute distance of the $Q = 50$ neighboring points on the object around the hand relative to the hand's local frame. If it drops below a certain threshold for a hand-object pair in contact, a *firm grasp* is detected. In object-centric representations, the *grasps* are modeled and adapted with respect to the object being grasped. Therefore, we set the hand as a slave object paired to the grasped master object (see Fig. 1b).

3) *Pose Invariance Detection*: Our approach relies on the estimated motion of the objects' candidate points without any prior semantic knowledge about the objects or their roles in the task. Therefore, the motion of object A relative to B can equivalently be represented as the motion of B relative to A. This results in a potential bi-directional MSR, leading to improper reproductions. For example, the master cup may move with respect to the slave kettle, resulting in an invalid pouring action. To address this issue, Krebs et. al. [9] chose the master object as the less mobile one using absolute motion saliency detection. However, this is not necessarily correct. For example, the cup is usually considered a master object in the pouring task even if it moves more than the slave kettle (see Fig. 1c). To address this problem, we propose an *invariance criterion*. Given a moving object pair (O_A, O_B) that has a potential bi-directional MSR, we first compute the translational and orientational spatial invariance of both relative to all static objects $\{O_s\}$. The key idea is that if we observe the most salient spatial invariance from object O_l with respect to O_s , $l \in \{A, B\}$, we consider O_l the master, which itself is paired to the master O_s . Specifically,

the master object is obtained by

$$l = \arg \min_l \{r_{s,l}^p, r_{s,l}^o\}_{s \in \mathcal{O}_s, l \in \{A, B\}}, \quad (1)$$

where the ratios $r_{s,l}^p, r_{s,l}^o$ are the normalized translational and orientational spatial variability of the salient object O_l relative to the static object O_s , respectively. This ensures that the HMSR is a DAG. The HMSR may still contain redundant relations, which we then truncate.

4) *Truncation*: For each potential master-slave pair in the HMSR graph, we employ K-VIL as described in Section III and remove the pairs that do not show any constraint between the master and slave objects. This results in a compact HMSR graph associated with sub-symbolic geometric constraints for each master-slave pair (see Section III and Figs. 1b and 1c). In Section VI, we show that the HMSR graph becomes more compact and converges as the number of demonstrations increases. Our insight is that, with scarce demonstrations, any valid salient geometric constraint should be considered as knowledge about the task is limited, while unnecessary constraints can be truncated when statistical evidence becomes available in new demonstrations. When a moving master object O_m is not constrained by any static object after truncation, it is allowed to move freely in space following a task-space VMP. Its pose corresponds to the pose of the local frame that defines the top-priority constraints for its slave object. The frame of reference for the VMP is located on the virtual object, defined by the initial state of object O_m . We model a distribution of the end pose of O_m in the demonstrations, from which we sample a target pose to adapt the VMP for execution. This new type of constraint, called pose constraint, is added to Fig. 2c.

C. Bimanual Coordination Strategies

Given the grasp relationship and HMSR extracted in Section IV-B, we derive bimanual coordination strategies.

1) *Uncoordinated unimanual*: A single grasp relationship is detected between a hand and an object. This corresponds to the K-VIL, i.e., single MSR case (see also Fig. 3a).

2) *Uncoordinated bimanual*: Each hand grasps a different slave object, and these two slave objects have different master objects (see Table II). In this case, the two hands perform different tasks without coordination.

3) *Loosely-coupled coordination*: Interaction forces between two hand groups, i.e., the union of the hand and a grasped object, is key to distinguishing loosely-coupled and tightly-coupled asymmetric coordination strategies [9]. Since estimating interaction forces from visual demonstrations is

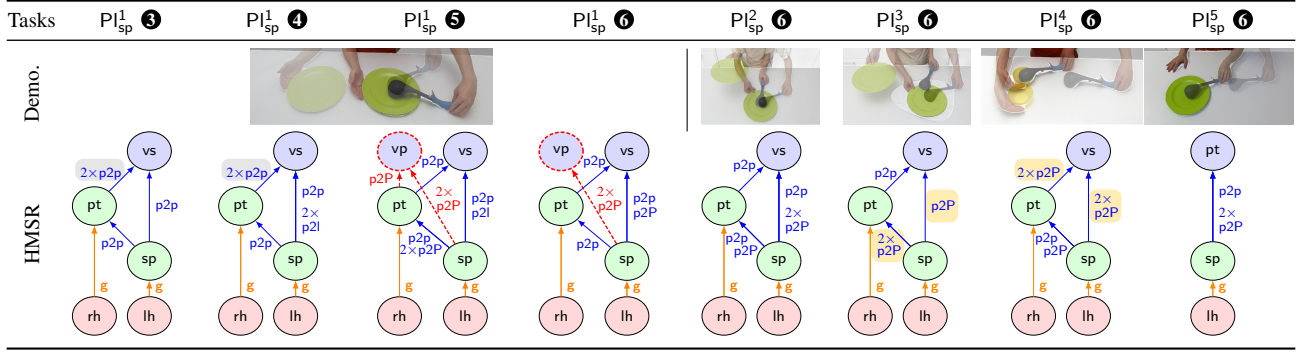


TABLE I: Loosely-coupled task: place spoon (PI_{sp}) on a plate with different styles. Objects include a spoon (sp), a plate (pt), and two hands (lh, rh). We prefix the letter v to the corresponding virtual object, e.g. vs stands for the virtual spoon.

not trivial, we do not distinguish between these two strategies and group them into a single one. That is, if constraints exist between the objects grasped individually by each hand, or if both grasped slave objects share at least one master object, the two hand groups are loosely-coupled. In the former case, one hand is constrained by another, whereas in the latter, the two hands move toward the same master object.

4) *Tightly-coupled symmetric coordination*: For a noticeable time window, (i) both hands grasp the same object, and (ii) the distance change rate (see Section IV-B.2) between two hands drops below a certain threshold.

V. BIMANUAL KEYPOINT-BASED ADMITTANCE CONTROLLER (BI-KAC)

Given the subsymbolic task representations in the HMSR graph including the keypoints, their associated local frames, geometric constraints and MPs, we derive a compliant and torque-controlled bimanual keypoint-based controller extended from KAC [1]. We control the Tool-Center-Point (TCP) of each robot hand with an impedance controller. The TCP task space target is derived using KAC for each arm. Specifically, the spring-damper systems of all constraints defined for an object in a hand contribute to the forces driving this hand. The coordination is achieved via the HMSR. In other words, Bi-KAC is a naive extension of KAC, which handles bimanual coordination via the HMSR representation. For example, in Fig. 1d, the left hand grasps the kettle following p2p and p2c constraints. The corresponding VMPs and constraints are dynamically updated by the moving master cup grasped by the right hand, which itself is controlled by a task-space VMP towards a pose constraint defined on the static virtual cup.

VI. EVALUATION

We evaluate our approach in eight real-world tasks, namely, pour water (PO_w), pour beer (PO_b), place spoon (PI_{sp}), place serving tray (PI_{st}), place spoon and plate ($PI_{sp,pt}$), place cutboard and pan ($PI_{cb,pa}$), place spoon and banana ($PI_{sp,ba}$), and clean table (C_{ta}). Given a few demonstration videos of each task recorded with Azure Kinect or Stereolab ZED camera, we run our perception pipeline to obtain the 3D point trajectories of objects and hands, extract a HMSR and a coordination strategy, and reproduce the tasks

with Bi-KAC in novel scenes. We evaluate Bi-KVIL's ability to 1) extract a consistent HMSR from different styles of task demonstrations, 2) capture these fine-grained styles in its sub-symbolic task representation, and 3) reproduce the learned tasks with categorical generalization.

A. Task Extraction

For each task, we provide different styles and numbers ① of demonstrations, resulting in a total of 14 evaluations. Specifically, in the PI_{sp} task, the motion styles are: (PI_{sp}^1) the plate moves to the spoon and the spoon is lifted up and placed at the center of the plate, (PI_{sp}^2) as PI_{sp}^1 but plates are taken from various positions above the table, (PI_{sp}^3) similar to PI_{sp}^1 , but the spoon is placed at an arbitrary position on the plate, (PI_{sp}^4) the plate moves to an arbitrary position with a spoon at the center, and (PI_{sp}^5) unimanual placement. Results are displayed in Table I). With ③ and ④ demonstrations in PI_{sp}^1 with small pose variations of plates with respect to the virtual plate, Bi-KVIL extracts more p2p (■) constraints than for the other tasks. With additional demonstrations in PI_{sp}^1 ⑤/⑥, p2P constraints for the spoon are extracted with respect to multiple master objects, i.e., the plate, virtual spoon, and virtual plate. Since the plate always remains on the table surface, it is reasonable that multiple p2P constraints exist. When the plate starts from above the table in PI_{sp}^2 ⑥, Bi-KVIL learns to eliminate the redundant master-slave pairs related to the virtual plate and the associated p2P constraints (---), resulting in a more compact HMSR graph. Compared to PI_{sp}^2 ⑥, the spoons are placed at arbitrary positions on the plate in PI_{sp}^3 ⑥, so that the p2p constraints between the spoon and its two masters (■) are truncated and an additional p2P constraint is created. Similarly, in PI_{sp}^4 ⑥, the plate is constrained only by the table surface and the spoon follows its motion, so that p2p (■) are replaced by p2P constraints. Except for the redundant master-slave pair of PI_{sp}^1 ⑤/⑥ and the unimanual case PI_{sp}^5 ⑥, the HMSR graph is structured identically across task styles, but differs in sub-symbolic constraints as different motion styles are captured. Moreover, redundant relations and constraints are eliminated by providing more demonstrations with variations. For all bimanual PI_{sp} tasks, Bi-KVIL extracts a loosely-coupled bimanual coordination strategy with the right hand group being non-dominant since the plate is a master of the spoon.

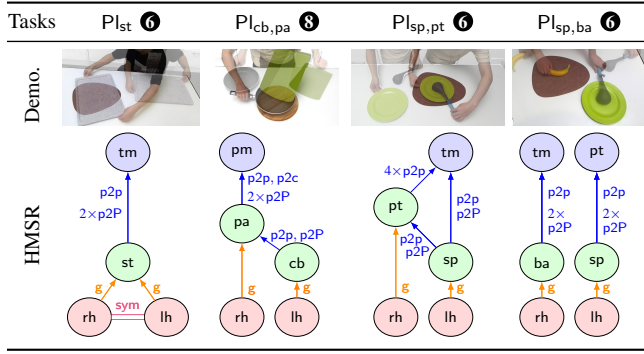


TABLE II: HMSR for PI_{st}, PI_{cb,pa}, PI_{sp,pt}, and PI_{sp,ba} tasks. Legend as Table I and tm, sp, pm, cb, pt, ba stand for tablemat, spoon, potmat, cutboard, plate, and banana.

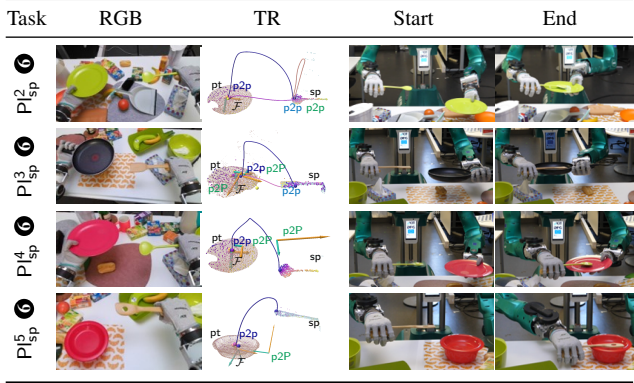


TABLE III: Reproduction of the PI_{sp} tasks with different styles corresponding to PI_{sp}²⁻⁵ ⑥. Images in each row correspond to RGB perception from the robot’s viewpoint, the task representation (TR), and the start and end of execution.

As shown in Fig. 1, different pouring styles also share the same HMSR structure at a symbolic level and differ only in the sub-symbolic definition of the p2p, p2c, and pose constraints. Note that the pose constraints, required to tilt the cup in the Po_b task, are correctly modeled and reproduced by Bi-KAC (see Figs. 1c and 1d). In Fig. 1c, the cup travels longer and faster on average than the kettle or beer bottle. In contrast to the rule-based algorithm in [9], our pose invariance criteria identify the cup as master as it displays less orientation variation.

The virtual object is not required in the presence of a static real object from which the spatial invariances of the moving objects are more salient. When the demonstrations contain sufficient pose or shape variations, Bi-KVIL truncates the virtual objects in the PI_{st}, PI_{cb,pa}, PI_{sp,pt}, and PI_{sp,ba} tasks, and a static real object serves as the top-level master object (see Table II). In PI_{st} ⑥, symmetric coordination is extracted along with sub-symbolic constraints p2p and p2P defining the target pose of the serving tray right above the center of the tablemat. Bi-KVIL also deals with tasks involving more than two objects, e. g., PI_{cb,pa} ⑧, PI_{sp,pt} ⑥ and PI_{sp,ba} ⑥, where a loosely-coupled coordination is extracted for the former two and uncoordinated bimanual coordination for the latter. The master-slave pairs between the hand groups are

truncated as K-VIL finds no salient geometric constraint.

B. Task Reproduction

We evaluate Bi-KAC qualitatively for each task in Section VI-A and refer the reader to [1] for quantitative evaluations, as its behavior for each arm inherits KAC. Here, we select one example per style of the PI_{sp} task to illustrate the behavior of Bi-KAC in reproducing the learned task with the ARMAR-6 humanoid robot [62]. As shown in Table III, the plate is driven to the initial position of the spoon with the spoon head right above the center of the plate in PI_{sp}² ⑥. This is due to the p2p constraints between the plate and the virtual spoon and between the spoon and the plate. In contrast, the plate in PI_{sp}⁴ ⑥ moves on plane constraints, and the spoon is placed anywhere on the pan in PI_{sp}³ ⑥ as p2p constraints were eliminated (). Notice that all tasks were also reproduced using out-of-distribution objects such as spoons of various shapes, plates of different sizes and colors, and cooking pans instead of plates in PI_{sp} in Table III. We refer the interested reader to our website <https://sites.google.com/view/bi-kvil> for results of other tasks with different styles, numbers of demonstrations, and out-of-distribution objects.

VII. CONCLUSION

In this paper, we proposed Bi-KVIL, a novel keypoints-based approach for visual imitation learning of bimanual manipulation tasks. Bi-KVIL simultaneously extracts hybrid master-slave relationships (HMSR) and bimanual coordination strategies at the symbolic level, as well as the task representations capturing the fine-grained motion styles at the sub-symbolic level. The proposed HMSR covers the bimanual manipulation taxonomy [9] and enables unified keypoints-based bimanual controllers for both uni- and bimanual tasks. By explicitly modeling the master-slave relationships and geometric constraints in an object-centric manner, our representation is embodiment-independent and viewpoint invariant (see [1, Section VII]), and generalizes well to categorical objects. Bi-KVIL allows us to learn bimanual task representations while requiring less than 10 human demonstration videos from RGB-D cameras without additional devices. In comparison, other bimanual imitation learning approaches demand a large number of demonstrations, e. g., 20 to 50 in [40], [41], 2500 to 4700 in [44], and 256 to 4000 in [45], [46]. Some approaches additionally require teleoperation data [40], [41] or human pose recorded using motion capture system [7]. Finally, we establish a perception pipeline leveraging advanced computer vision algorithms to provide high-quality datasets for VIL.

Although our perception pipeline includes hand shape completion [59], it does not handle object (self-)occlusion, leading to failures if keypoints are occluded. Failures may also occur due to inaccurate correspondence detection in specific objects poses, which lead to imprecise local frames on master objects and inaccurate target positions, e. g., the spout of the kettle being outside the cup rim. Moreover, Bi-KAC, as a naive extension of KAC, relies entirely on the HMSR for coordination and disregards dual-arm synchronization [32],

[63], [64]. Therefore, it may drop the object in bimanual transport tasks. In future work, we plan to address these limitations and to investigate a comprehensive evaluation benchmark for bimanual imitation learning tasks.

REFERENCES

- [1] J. Gao, Z. Tao, N. Jaquier, and T. Asfour, “K-VIL: Keypoints-Based Visual Imitation Learning,” *IEEE Transactions on Robotics*, vol. 39, no. 5, pp. 3888–3908, 2023.
- [2] M. Muhlig, M. Gienger, J. J. Steil, and C. Goerick, “Automatic selection of task spaces for imitation learning,” in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2009, pp. 4996–5002.
- [3] M. Muhlig, M. Gienger, S. Hellbach, J. J. Steil, and C. Goerick, “Task-level imitation learning using variance-based movement optimization,” in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2009, pp. 1177–1184.
- [4] Y. Guiard, “Asymmetric Division of Labor in Human Skilled Bimanual Action: The Kinematic Chain as a Model,” *Journal of motor behavior*, vol. 19, no. 4, pp. 486–517, 1987.
- [5] M. Kimmerle, C. L. Ferre, K. A. Kotwica, and G. F. Michel, “Development of role-differentiated bimanual manipulation during the infant’s first year,” *Developmental Psychobiology: The Journal of the International Society for Developmental Psychobiology*, vol. 52, no. 2, pp. 168–180, 2010.
- [6] Y. Zhou, M. Do, and T. Asfour, “Coordinate change dynamic movement primitives - a leader-follower approach,” in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2016, p. 5481–5488.
- [7] J. Liu, Y. Chen, Z. Dong, S. Wang, S. Calinon, M. Li, and F. Chen, “Robot Cooking with Stir-fry: Bimanual Non-prehensile Manipulation of Semi-fluid Objects,” *IEEE Robotics and Automation Letters*, vol. 7, pp. 5159–5166, 2022.
- [8] L. P. Ureche and A. Billard, “Constraints extraction from asymmetrical bimanual tasks and their use in coordinated behavior,” *Robotics and Autonomous Systems*, vol. 103, pp. 222–235, 2018.
- [9] F. Krebs and T. Asfour, “A Bimanual Manipulation Taxonomy,” *IEEE Robotics and Automation Letters*, vol. 7, pp. 11 031–11 038, 2022.
- [10] 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 *Euro. Conf. on Computer Vision (ECCV)*, 2022, pp. 570–587.
- [11] A. Patel, A. Wang, I. Radosavovic, and J. Malik, “Learning to Imitate Object Interactions from Internet Videos,” *arXiv:2211.13225*, 2022.
- [12] P. Sundaresan, J. Grannen, B. Thananjeyan, A. Balakrishna, M. Laskey, K. Stone, J. E. Gonzalez, and K. Goldberg, “Learning Rope Manipulation Policies Using Dense Object Descriptors Trained on Synthetic Depth Data,” in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2020, pp. 9411–9418.
- [13] P. Florence, L. Manuelli, and R. Tedrake, “Dense Object Nets: Learning dense visual object descriptors by and for robotic manipulation,” in *Conference on Robot Learning (CoRL)*, 2018, pp. 373–385.
- [14] U. Deekshith, N. Gajjar, M. Schwarz, and S. Behnke, “Visual Descriptor Learning from Monocular Video,” in *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, Valletta, Malta, 2020, pp. 444–451.
- [15] P. Florence, L. Manuelli, and R. Tedrake, “Self-supervised correspondence in visuomotor policy learning,” *IEEE Robotics and Automation Letters*, vol. 5, pp. 492–499, 2020.
- [16] D. Hadjivelichkov, S. Zwane, L. Agapito, M. P. Deisenroth, and D. Kanoulas, “One-Shot Transfer of Affordance Regions? AffCorrs!” in *Conference on Robot Learning (CoRL)*, 2022, pp. 550–560.
- [17] S. Amir, Y. Gandelsman, S. Bagon, and T. Dekel, “Deep ViT Features as Dense Visual Descriptors,” in *ECCVW What is Motion For?*, 2022.
- [18] Y. Liu, Z. Shen, Z. Lin, S. Peng, H. Bao, and X. Zhou, “GIFT: Learning Transformation-Invariant Dense Visual Descriptors via Group CNNs,” in *Neural Information Processing Systems (NeurIPS)*, 2019, pp. 6990–7001.
- [19] L. Yen-Chen, P. Florence, J. T. Barron, T.-Y. Lin, A. Rodriguez, and P. Isola, “NeRF-Supervision: Learning dense object descriptors from neural radiance fields,” in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2022, pp. 6496–6503.
- [20] A. Simeonov, Y. Du, A. Tagliasacchi, J. B. Tenenbaum, A. Rodriguez, P. Agrawal, and V. Sitzmann, “Neural Descriptor Fields: SE(3)-equivariant object representations for manipulation,” in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2022, pp. 6394–6400.
- [21] A. Simeonov, Y. Du, L. Yen-Chen, A. Rodriguez, L. P. Kaelbling, T. Lozano-Perez, and P. Agrawal, “SE(3)-Equivariant Relational Rearrangement with Neural Descriptor Fields,” in *Conference on Robot Learning (CoRL)*, 2022, pp. 835–846.
- [22] E. Chun, Y. Du, A. Simeonov, T. Lozano-Perez, and L. Kaelbling, “Local Neural Descriptor Fields: Locally Conditioned Object Representations for Manipulation,” in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2023, pp. 1830–1836.
- [23] X. Zhao, R. Hu, P. Guerrero, N. Mitra, and T. Komura, “Relationship templates for creating scene variations,” *ACM Transactions on Graphics*, vol. 35, pp. 1–13, 2016.
- [24] Z. Huang, J. Xu, S. Dai, K. Xu, H. Zhang, H. Huang, and R. Hu, “NIFT: Neural Interaction Field and Template for Object Manipulation,” in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2023, pp. 1875–1881.
- [25] P. Sundaresan, S. Belkhale, D. Sadigh, and J. Bohg, “KITE: Keypoint-Conditioned Policies for Semantic Manipulation,” *arXiv:2306.16605*, 2023.
- [26] W. Gao and R. Tedrake, “kPAM 2.0: Feedback control for category-level robotic manipulation,” *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 2962–2969, 2021.
- [27] J. Jin, L. Petrich, M. Dehghan, and M. Jagersand, “A Geometric Perspective on Visual Imitation Learning,” in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2020, pp. 5194–5200.
- [28] A. Ajoudani, N. G. Tsagarakis, J. Lee, M. Gabiccini, and A. Bicchi, “Natural redundancy resolution in dual-arm manipulation using configuration dependent stiffness (CDS) control,” in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2014, pp. 1480–1486.
- [29] S. Savic, M. Rakovic, B. Borovac, and M. Nikolic, “Hybrid motion control of humanoid robot for leader-follower cooperative tasks,” *Thermal Science*, vol. 20, pp. 549–561, 2016.
- [30] D. Almeida and Y. Karayiannidis, “A Lyapunov-Based Approach to Exploit Asymmetries in Robotic Dual-Arm Task Resolution,” in *2019 IEEE 58th Conference on Decision and Control (CDC)*, 2019, pp. 4252–4258.
- [31] S. S. Mirrazavi Salehian, N. Figueroa, and A. Billard, “A unified framework for coordinated multi-arm motion planning,” *The International Journal of Robotics Research*, vol. 37, pp. 1205–1232, 2018.
- [32] J. Gao, Y. Zhou, and T. Asfour, “Projected Force-Admittance Control for Compliant Bimanual Tasks,” in *IEEE/RAS Intl. Conf. on Humanoid Robots (Humanoids)*, 2018, pp. 607–613.
- [33] H. A. Park and C. S. G. Lee, “Extended Cooperative Task Space for manipulation tasks of humanoid robots,” in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2015, pp. 6088–6093.
- [34] J. Lee and P. H. Chang, “Redundancy resolution for dual-arm robots inspired by human asymmetric bimanual action: Formulation and experiments,” in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2015, pp. 6058–6065.
- [35] F. Amadio, A. Colome, and C. Torras, “Exploiting Symmetries in Reinforcement Learning of Bimanual Robotic Tasks,” *IEEE Robotics and Automation Letters*, vol. 4, pp. 1838–1845, 2019.
- [36] È. Pairet, P. Ardón, M. Mistry, and Y. Petillot, “Learning and Composing Primitive Skills for Dual-arm Manipulation,” in *Towards Autonomous Robotic Systems - 20th Annual Conference (TAROS)*, vol. 11649, 2019, pp. 65–77.
- [37] G. Franzese, L. d. S. Rosa, T. Verbarg, L. Peternel, and J. Kober, “Interactive Imitation Learning of Bimanual Movement Primitives,” *IEEE/ASME Transactions on Mechatronics*, pp. 1–13, 2023.
- [38] Z. Dong, Z. Li, Y. Yan, S. Calinon, and F. Chen, “Passive Bimanual Skills Learning From Demonstration With Motion Graph Attention Networks,” *IEEE Robotics and Automation Letters*, vol. 7, pp. 4917–4923, 2022.
- [39] M. Knaust and D. Koert, “Guided Robot Skill Learning: A User-Study on Learning Probabilistic Movement Primitives with Non-Experts,” in *IEEE/RAS Intl. Conf. on Humanoid Robots (Humanoids)*, 2021, pp. 514–521.
- [40] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, “Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware,” in *Robotics: Science and Systems (R:SS)*, 2023.
- [41] Z. Fu, T. Z. Zhao, and C. Finn, “Mobile ALOHA: learning biman-

- ual mobile manipulation with low-cost whole-body teleoperation,” *arXiv:2401.02117*, 2024.
- [42] Y. Chen, T. Wu, S. Wang, X. Feng, J. Jiang, S. M. McAleer, H. Dong, Z. Lu, S.-C. Zhu, and Y. Yang, “Towards Human-Level Bimanual Dexterous Manipulation with Reinforcement Learning,” in *Neural Information Processing Systems (NeurIPS)*, vol. 35, 2022, pp. 5150–5163.
- [43] S. Kataoka, S. K. S. Ghasemipour, D. Freeman, and I. Mordatch, “Bi-Manual Manipulation and Attachment via Sim-to-Real Reinforcement Learning,” *arXiv:2203.08277*, 2022.
- [44] F. Xie and A. Chowdhury, “Deep Imitation Learning for Bimanual Robotic Manipulation,” in *Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 2327–2337.
- [45] H. Kim, Y. Ohmura, and Y. Kuniyoshi, “Robot peels banana with goal-conditioned dual-action deep imitation learning,” *arXiv:2203.09749*, 2022.
- [46] —, “Transformer-based deep imitation learning for dual-arm robot manipulation,” in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2021, pp. 8965–8972.
- [47] C. R. G. Dreher, M. Wächter, and T. Asfour, “Learning object-action relations from bimanual human demonstration using graph networks,” *IEEE Robotics and Automation Letters (RA-L)*, vol. 5, no. 1, pp. 187–194, 2020.
- [48] K. Meng and A. Eloyan, “Principal manifold estimation via model complexity selection,” *Journal of the Royal Statistical Society. Series B, Statistical methodology*, vol. 83, no. 2, pp. 369–394, 2021.
- [49] Y. Zhou, J. Gao, and T. Asfour, “Learning via-point movement primitives with inter- and extrapolation capabilities,” in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2019, pp. 4301–4308.
- [50] H. Xu, J. Zhang, J. Cai, H. Rezatofighi, F. Yu, D. Tao, and A. Geiger, “Unifying Flow, Stereo and Depth Estimation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–18, 2023.
- [51] T. Müller, A. Evans, C. Schied, and A. Keller, “Instant Neural Graphics Primitives with a Multiresolution Hash Encoding,” *ACM Trans. Graph.*, vol. 41, pp. 102:1–102:15, 2022.
- [52] J. Ichnowski, J. Kerr, Y. Avigal, and K. Goldberg, “Dex-NeRF: Using a Neural Radiance Field to Grasp Transparent Objects,” in *Conference on Robot Learning (CoRL)*, vol. 164, 2021, pp. 526–536.
- [53] Z. Teed and J. Deng, “RAFT: Recurrent All-Pairs Field Transforms for Optical Flow,” in *Euro. Conf. on Computer Vision (ECCV)*, 2020, pp. 402–419.
- [54] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, “Segment Anything,” in *Intl. Conf. on Computer Vision (ICCV)*, 2023, pp. 3992–4003.
- [55] Y. Cheng, L. Li, Y. Xu, X. Li, Z. Yang, W. Wang, and Y. Yang, “Segment and Track Anything,” *arXiv:2305.06558*, 2023.
- [56] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu, and L. Zhang, “Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection,” *arXiv:2303.05499*, 2023.
- [57] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee *et al.*, “MediaPipe: A framework for building perception pipelines,” *arXiv:1906.08172*, 2019.
- [58] T. Jiang, P. Lu, L. Zhang, N. Ma, R. Han, C. Lyu, Y. Li, and K. Chen, “RTMPose: Real-Time Multi-Person Pose Estimation based on MMPose,” *arXiv:2303.07399*, 2023.
- [59] K. Lin, L. Wang, and Z. Liu, “Mesh Graphormer,” in *Intl. Conf. on Computer Vision (ICCV)*, 2021, pp. 12 919–12 928.
- [60] J. Lin, A. Zeng, H. Wang, L. Zhang, and Y. Li, “One-Stage 3D Whole-Body Mesh Recovery With Component Aware Transformer,” in *Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 21 159–21 168.
- [61] J. Romero, D. Tzionas, and M. J. Black, “Embodied hands: Modeling and capturing hands and bodies together,” *ACM Transactions on Graphics*, vol. 36, pp. 1–17, 2017.
- [62] T. Asfour, M. Wächter, L. Kaul, S. Rader, P. Weiner, S. Ottenhaus, R. Grimm, Y. Zhou, M. Grotz, and F. Paus, “ARMAR-6: A high-performance humanoid for human-robot collaboration in real world scenarios,” *IEEE Robotics and Automation Magazine*, vol. 26, no. 4, pp. 108–121, 2019.
- [63] H.-C. Lin, J. Smith, K. K. Babarahmati, N. Dehio, and M. Mistry, “A projected inverse dynamics approach for multi-arm cartesian impedance control,” in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2018, pp. 5421–5428.
- [64] E. Shahriari, S. A. B. Birjandi, and S. Haddadin, “Passivity-based adaptive force-impedance control for modular multi-manual object manipulation,” *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 2194–2201, 2022.