Robotics: Science and Systems 2023
Daegu, Republic of Korea, July 10-July 14, 2023

SAR: Generalization of Physiological Dexterity via Synergistic Action Representation

Cameron Berg, Vittorio Caggiano*, Vikash Kumar*
Meta AI
* Equal advising

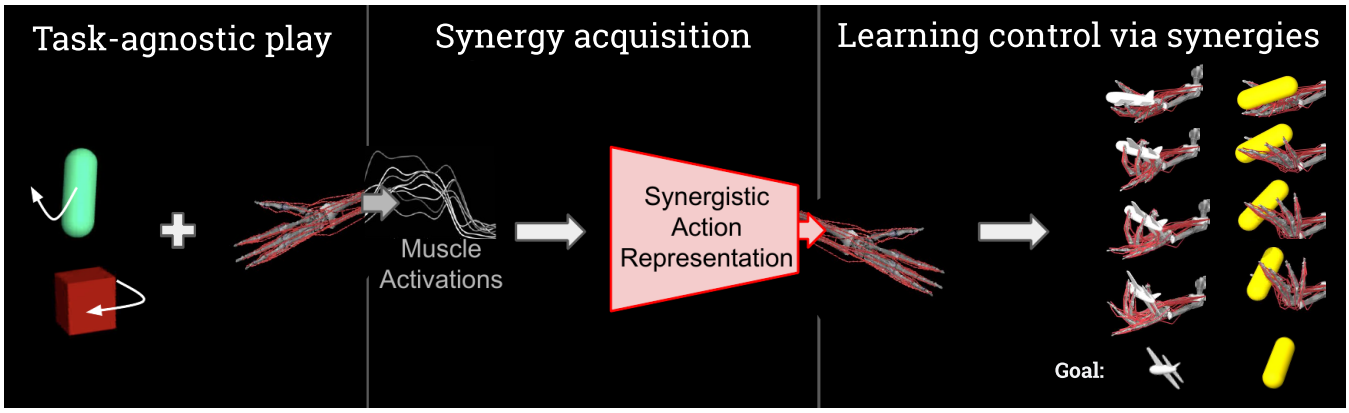

Fig. 1: **SAR pipeline.** A physiologically accurate hand is trained in simulation on a simple manipulation task (left), muscle activations are gathered from this policy (center), and these activations are used to create a synergistic action representation (SAR) that enables robust skill transfer for learning significantly more complex manipulations (right).

***Abstract*—Learning effective continuous control policies in high-dimensional systems, including musculoskeletal agents, remains a significant challenge. Over the course of biological evolution, organisms have developed robust mechanisms for overcoming this complexity to learn highly sophisticated strategies for motor control. What accounts for this robust behavioral flexibility? Modular control via muscle synergies, i.e. coordinated muscle co-contractions, is considered to be one putative mechanism that enables organisms to learn muscle control in a simplified and generalizable action space. Drawing inspiration from this evolved motor control strategy, we use a physiologically accurate hand model to investigate whether leveraging a *Synergistic Action Representation* (*SAR*) acquired from simpler manipulation tasks improves learning and generalization on more complex tasks. We find that *SAR*-exploiting policies trained on a complex, 100-object randomized reorientation task significantly outperformed ($> 70\%$ success) baseline approaches ($< 20\%$ success). Notably, *SAR*-exploiting policies were also found to zero-shot generalize to thousands of unseen objects with out-of-domain size variations, while policies that did not adopt *SAR* failed to generalize. *SAR* also enabled significantly improved transfer learning on real-world objects. Finally, using a robotic manipulation task set and a full-body humanoid locomotion task, we establish the generality of SAR on broader high-dimensional control problems, achieving SOTA performance with an order of magnitude improved sample efficiency. To the best of our knowledge, this investigation is the first of its kind to present an end-to-end pipeline for discovering synergies and using this representation to learn high-dimensional continuous control across a wide diversity of tasks.**

Project website: https://sites.google.com/view/sar-rl

I. Introduction

An important precondition for generalist embodied agents is their ability to exhibit diverse behaviors in response to dynamic environmental conditions, akin to humans and animals. From the practical perspective of building such agents, supporting behavioral diversity requires the integration of mechanisms that can capably handle complex high-dimensional action spaces required to host the behavioral diversity. For example, humans are capable of an extremely large set of feasible motor repertoires, but this behavioral diversity necessitates continuous control of approximately 600 muscles in the human body [21].

Learning control policies for high-dimensional continuous action spaces are significantly more challenging [42] than learning in reduced action spaces that utilize idealized direct control, such as the Atari environments commonly used for reinforcement learning (RL) experiments [28, 4]. However, even the joint-based and end-effector continuous control dynamics commonly utilized in robotics exhibit significant simplifications as compared to the structural and functional physiology of actuated biological organisms [46]. As a case study, consider the human hand, where joints are not controlled directly, but rather, through the proxy of pull-only muscle forces [50]. Further, human physiology is multiarticular, meaning that there exist many-to-one and one-to-many relationships between muscles and joints (e.g., the flexor digitorum profundus muscle controls over 3 joints for each digit of the hand)

[27]. Critically, the human hand is also overactuated, meaning that there are more muscle forces than degrees of freedom (i.e., approximately 39 muscles collectively control 23 joints in the human hand) [49].

Despite the significant challenges inherent in learning musculoskeletal control, humans and animals are able to rapidly learn motor behaviors that are orders of magnitude more complex than the current state of the art in embodied agents[2]. A question naturally emerges from this fact: what strategies have humans and animals evolved that enable them to learn musculoskeletal control so robustly—and can these strategies be utilized to train policies that exhibit similarly robust behaviors in high-dimensional continuous action spaces.

Modular control is one mechanism known to be fundamental to human and animal behavioral flexibility, where motor output is computed as a product of a reduced number of modular spatiotemporal muscle coactivation patterns rather than as separate activations for each individual muscle [10, 8]. These muscle co-contraction modules are referred to as *muscle synergies*. There exists strong evidence not only that motor behavior is produced by coordinated muscle activity that can be represented with high fidelity in terms of a smaller number of synergistic activation patterns [40], but also that the spinal cord explicitly encodes muscle activity at the level of synergies rather than as individual muscle activations [7]. This built-in action representation might be thought of as an embodied form of transfer learning, as it enables efficient adaptations of new motor repertoires that leverage similar—but non-identical—muscle activation patterns. It is suggested that these motor modules are present even in newborns [19], that they can be finetuned and recombined to yield more complex behaviors over the lifetime [14], and that they facilitate skill transfer, wherein similar activation representations can be reused for similar tasks (e.g., writing with a pen vs. writing with a pencil would not require learning two unique motor repertoires from scratch) [6].

In this work, we explore whether the explicit discovery and utilization of muscle synergies improves the performance and generalization power of high-dimensional continuous control policies, with a specific focus on physiological manipulation. We train a physiologically accurate musculoskeletal model of the human hand, MyoHand [11], on a dexterous multi-object reorientation task suite. To the best of our knowledge, this is the first study that demonstrates synergy-exploiting control policies that are able to successfully learn dexterous manipulation tasks unsolved by baseline learning approaches. In particular, our core contributions are as follows:

- We present a simple paradigm to recover a synergistic action representation using task-agnostic play behaviors.
- We demonstrate that training with a *Synergistic Action Representation* enables strong performance ($> 70\%$ success) in a complex multi-object randomized manipulation task that is otherwise unsolved by baselines that lack access to this representation ($< 20\%$ success).
- We further show that policies trained with a *Synergistic Action Representation* zero-shot generalize to thousands of unseen in-domain and out-of-domain objects.
- We demonstrate that a *Synergistic Action Representation* acquired from a simple task enables dexterity transfer to new tasks, accelerating behavior acquisition by $2x$.
- Finally, we demonstrate that *Synergistic Action Representation*s extend beyond physiological control using robotic manipulation and a full-body humanoid locomotion tasks, achieving SOTA performance with an order of magnitude improved sample efficiency on the latter.

II. Related Works

We first provide a background on what is currently understood about muscle synergies in humans and summarize previous work in leveraging synergy representations for learning musculoskeletal control policies.

A. Motor neuroscience of muscle synergies

Over five decades, the motor neuroscience literature on muscle synergies [22] has elucidated their statistical representation [57, 40], control dynamics [45, 25], lifespan finetuning and recombination [19], and cross-species similarity [60, 19]. Muscle synergies can be defined as motor activation modules that impose a dimensionality-reducing spatiotemporal structure to muscle activations by linearly combining (temporal) simplified activation patterns with a set of (spatial) variable activation weights for discrete muscles [19]. It is widely considered that muscle synergies are an evolved mechanism for vertebrate motor control with a direct neurophysiological implementation [51, 24, 32]. Accordingly, muscle synergies are found to be robust to perturbations within a specific motor repertoire; for instance, as few as four muscle synergies in the lower limbs appear to control locomotion in humans and are invariant to movement speed, grade, direction, and skill [19, 45].

B. Synergy representations for motor control

The neuroscience literature on muscle synergies is reminiscent of dynamic movement primitives (DMPs), a popular concept in robotic control where actuation dynamics are decomposed into a smaller set of stable dynamical systems [48]. While these two concepts are similar in their simplification of complex continuous control dynamics into more tractable modular representations, the methods typically used to parameterize and stabilize DMPs can often preclude their generalization power to novel tasks [44]. By contrast, one core property of muscle synergies is their putative role as an action representation template for learning related tasks [14].

A small number of investigations have previously assessed the feasibility of utilizing muscle synergy representations for effectively learning to control musculoskeletal or robotic systems (for a review, see [33]), albeit in fairly constrained task settings. To the best of our knowledge, all previous studies have explored whether synergy representations enable reaching behaviors in a musculoskeletal arm with 7 degrees of freedom (DoF) [30, 1, 18, 44, 12] or 8-DoF thumb/index model [43]. These investigations have found that synergies can enable faster learning and partial generalization. One study also found

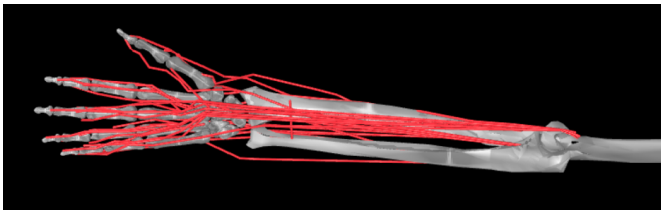

Fig. 3: **Musculoskeletal model of the Hand: *MyoHand***. Musculoskeletal model of the human hand comprising 23 joints, 29 bones, and 39 muscle-tendon units [11]. Muscles are indicated in red.

that synergies improved task performance [18]. While these previous works demonstrate the general promise of leveraging muscle synergies for musculoskeletal control, the nature of these task demonstrations is arguably too low-dimensional both structurally (i.e., a 7-DoF arm) and functionally (i.e., reaching to a small number of fixed points) to yield strong conclusions about the full scope of using synergies for control.

In order to more comprehensively understand whether muscle synergies can enhance the training performance and generalization power of musculoskeletal control policies, we believe that synergy action representations must be tested in significantly more complex musculoskeletal task settings. For this reason, we investigate significantly beyond the current scope of the muscle synergy control literature by exploring (a) how synergistic action representations can be computationally extracted from a musculoskeletal hand model, and (b) how these representations can be leveraged to learn control policies capable of manipulating thousands of complex objects. Finally, we investigate (c) the extent to which this biologically-inspired representation can more generally enable high-dimensional continuous control in task formulations beyond physiological manipulation.

III. Physiological Dexterity

In this work, we first seek to use learned action representations to acquire generalizable manipulation behavior with diverse objects on a physiologically accurate musculoskeletal hand. We begin by introducing the musculoskeletal hand, and then formulate the problem of learning in-hand dexterous behavior before detailing our method in Sec IV.

A. Physiologically accurate musculoskeletal hand model

We investigated various options of musculoskeletal hand [35, 31] before ultimately selecting the *MyoHand* model in favor of its computational efficiency, physiological accuracy, and support for contact dynamics, which is essential for the stated goal of yielding in-hand dexterous manipulation [11]. *MyoHand* is a physiologically realistic model (see Fig. 3) of an adult right hand-wrist comprising 23 joints, 29 bones, and 39 muscle-tendon units [11], implemented in the MuJoCo physics simulator [56]. This hand model exhibits important mechanistic features that differentiate biological actuation from the joint-based control approach traditionally utilized in robotics, including overactuation, multi-articulation, and pull-only actuation with third order muscle dynamics [50]. Important to the current investigation, previous work has also demonstrated that *MyoHand* can learn dexterous manipulation behaviors with single objects (e.g., twirling a pen to a desired orientation) [11]. Here, we aim to go beyond isolated task-specific single-object solutions, by using the *MyoHand* model to learn in-hand manipulation behaviors that generalize from prior experiences to handle new, complex tasks.

Before outlining our methods, we next turn to formalizing our problem formulation.

B. Problem formulation

We pose the problem of learning in-hand behavior with musculoskeletal hand as a Markov Decision Process (MDP) [52], which can be defined as $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \rho, \gamma)$, where $\mathcal{S} \subseteq \mathbb{R}^n$ and $\mathcal{A} \subseteq \mathbb{R}^m$ each represent continuous state and action spaces. $\mathcal{T}$ represents the unknown distribution that describes state transition dynamics such that $s' \sim \mathcal{T}(\cdot \mid s, a)$. The reward function is denoted as $\mathcal{R} : \mathcal{S} \rightarrow [0, R_{\max}]$, $\gamma \in [0, 1)$ indicates the reward discount factor, and ρ denotes the initial distribution of states. The goal in performing RL is to search for policy parameters θ that map from states to a probability distribution over actions $\pi_\theta : \mathcal{S} \rightarrow P(\mathcal{A})$ in order to maximize long-run discounted returns [52]. We optimize our policy using Soft Actor-Critic (SAC) [23], an off-policy RL algorithm used for continuous control which optimizes for both long-run discounted returns, $\sum_t \gamma^t R_t$, and policy entropy, $H(\pi)$, such that $\pi^*_\theta(a \mid s) = \text{argmax}_\theta[J(\pi, \mathcal{M})]$, where

$$J = \max_\theta \mathbb{E}\left[\sum_t \gamma^t \left(R\left(s_t, a_t, s_{t+1}\right) + \alpha H\left(\pi\left(\cdot \mid s_t\right)\right)\right)\right].$$

We later demonstrate (Fig. 6) that the direct optimization of the above objective in an end-to-end learning paradigm on MyoHand does not lead to generalizable in-hand behaviors. We attribute this failure to the challenges in navigating an extremely high-dimensional search space that hosts all internal degrees of freedom of the hand and the object under manipulation in addition to third-order actuation dynamics, intermittent contact dynamics as the policy maneuvers the object, and the high likelihood of catastrophic failures from dropping the object due to uncoordinated movements during early exploration.

In contrast to end-to-end learning, biological organisms faced with similar search spaces are easily able to synthesize sophisticated movements that are robust to this complexity. In the next section, we outline our method that takes inspiration from mechanisms that putatively contribute to this evolved behavioral sophistication.

IV. *SAR*: Synergistic Action Representation

Exploration challenges in dexterous manipulation are well recognized [5], even in the field of robotics [5, 42, 13] and computer vision [63, 36]. Multiple techniques leveraging

expert demonstrations [42], curriculum learning [37, 13], and even human-designed mappings to embed representations into the learning process [29] have been proposed in order to aid and accelerate the acquisition of dexterous behaviors. In contrast to these methods, which lean heavily on experts, we investigate if there exist learning paradigms that do not require such interventions at any point. Our work is heavily inspired by the increasingly robust evidence that the biological motor system enables dexterity by leveraging synergistic control of the actuators [10, 6, 15, 38]. We proceed to ask:

1) How can synergistic representations be automatically acquired?
2) How can synergistic control can be embedded in behavior acquisition paradigms?

A. Synergy Acquisition

Researchers in the fields of biomechanics and neuroscience have long pursued the search for synergies both at the joint-postural and muscle-activation levels. Due to the relative ease of recording joint posture over muscle activations, synergies at the level of joint postures [55, 47, 17] have been more extensively used to synthesize behaviours in robotic control [20, 33]. Nevertheless, although muscle synergies provide direct insight into biological motor control and learning [16, 15, 12, 14], they can only be practically extracted from a limited number of muscles in human hands. This represents the biggest limitation to directly using muscle synergies to synthesize behaviors [53].

In spite of this practical challenge, how can we get access to synergies for the *MyoHand* at the actuation level? We consider the underlying properties of synergies during manipulation: synergies act like lower-dimensional submanifolds in the high-dimensional space that are consistently visited during the course of the manipulation. In order to capture these pathways, we subject *MyoHand* to a task-agnostic play period, during which relatively simple behaviors are learned. To ensure visitation and coverage in the synergistic submanifold, diversity in behaviors was encouraged by randomly varying the objects during play (details in Sec V-C).

We record muscle activation time series data from the resulting play behaviors, $M^{act} \in \mathcal{R}^{a \times t}$. We posit that owing to the diversity of the behaviors captured in the dataset, its visitation density has a high degree of overlap with the synergistic submanifolds. While there are multiple strategies to recover submanifolds from datasets, we resort back to the techniques common in studies investigating synergies from experimental datasets [40]. Applying both PCA and ICA to the data matrix (taken together, ICAPCA) is accepted as a robust method for capturing muscle synergies [57]. In this work, we take the additional step of normalizing this output in the range [-1, 1] and refer to the entire process as *normalized ICAPCA*.

$$SAR := |T_{ICA}(T_{PCA}(M^{act}))|$$

The resulting factorization is what we refer to as the *Synergistic Action Representation* (*SAR*). Practically, it can also be viewed as a way to map any point in lower-dimensional synergy space to full muscle activation space $a_{39}^{syn} = SAR(a_N)$. While $N = 39$ (the number of muscles in *MyoHand*) constitutes a complete reconstruction of the original signal, lower N will allow for capturing the most informative muscle co-activation patterns over time. In this investigation, we define *SAR* based on $N = 20$ (which captured over 80% of the variance in M^{act}, see Appendix: Fig A.15).

B. Behavior Acquisition with SAR

While *SAR* encodes preferred muscle synergies, we still require a method for training manipulation behaviors. The question thus becomes: how can use embed *SAR* into a behavioral learning paradigm? An intuitive strategy is to learn a policy directly in the space of synergies and project it to full muscle activation space via *SAR*: $a_{39}^{syn} = SAR(\pi(a_N|\emptyset_t))$. While the reduced dimensionality helps with the exploration challenges, it also restricts the policy from expressing behaviors significantly outside of the synergistic manifolds. Accordingly, we develop a policy architecture that preserves the benefits of directed search from the synergistic pathways but is still able to learn task-specific behavioral finetuning in the original nonsynergistic manifold.

In order to balance contributions from task-agnostic and task-specific muscle activations, we introduce *SAR* into our behavioral acquisition pipeline as an alternative weighted action representation pathway for our policy to utilize (Fig. 4). At each timestep, the policy is trained to output a $39 + N$-dimensional action vector continuous in the range $[-1, 1]$, where N is the number of synergistic modules in SAR. The first N components of the action vector are input into SAR, which transforms this representation back into synergy-exploiting activations in the original muscle activation space (a_t^{SAR}). The subsequent 39 components of the action vector (a_t^{39}) are additively combined with the synergy-exploiting activations, returning a final action (a_t^*) that steps the environment forward. This simple elementwise vector addition is weighted by φ, a parameter that calibrates the relative influence of the synergistic and nonsynergistic muscle activation vectors such that $a_t^* = \varphi(a_t^{SAR}) + (1 - \varphi)(a_t^{39})$ (Fig. 4).

Contrary to conventional wisdom that suggests a reduction the dimensionality of the action space will best aid exploration challenges, the proposed policy architecture somewhat counterintuitively increases the dimensionality of the search space. Aligned with [59, 62, 61], our policy architecture can be viewed as providing the learning agent the option to mix behavior from task-agnostic synergistic representational pathways (encoded by the first N components of a_t) while simultaneously allowing for the task-specific modulations (encoded by the subsequent 39 components of a_t). As is later demonstrated in our experiments (Sec. VI-B), in accordance with this intuition, our agents learn to leverage the synergistic pathways when helpful to accelerate search and proceed to rely on task-specific pathways for task customization, exploiting the benefits of both learning pathways (see Fig. 4).

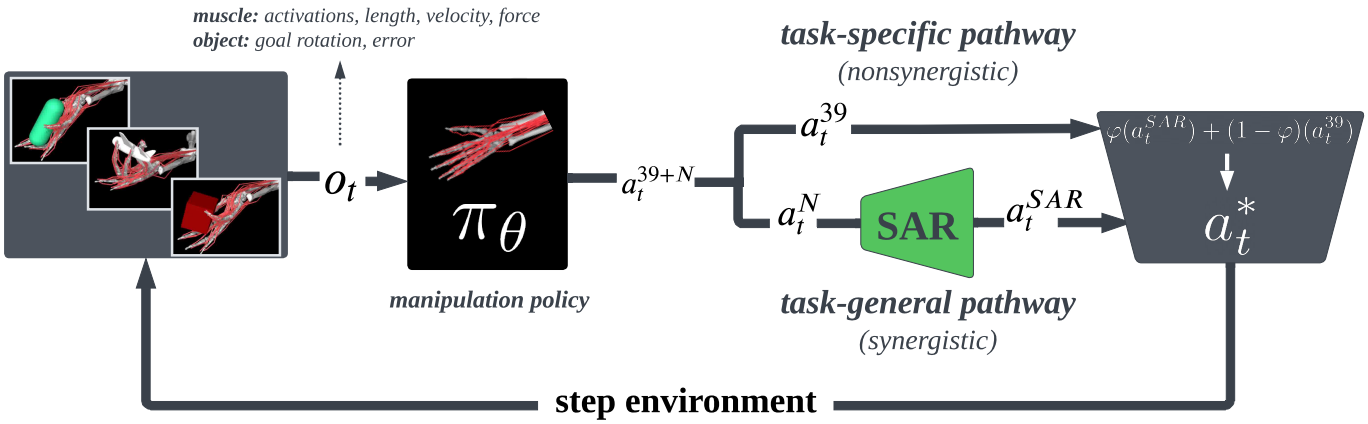

Fig. 4: **Policy architecture.** The policy takes as input o_t and outputs a $39 + N$-dimensional action vector, a_t^{39+N}. The first N synergistic actions are passed through $SAR(a_t^N) = a_t^{SAR}$ to recover the synergistic muscle activations. The subsequent 39 task-specific activations are mixed with the task-general activations using a linear blend φ to recover the final action a_t^* that steps the environment forward.

V. Experimental Design

We structure our main experiment to assess the effectiveness of *SAR* and our proposed policy architecture (Fig. 4) in yielding generalizable in-hand dexterity. Our goals include:

- Leveraging task-agnostic play behaviors in to acquire synergistic representations.
- Accelerating behavior acquisition using our proposed *SAR*-based policy architecture in comparison to traditional pipelines that do not exploit synergies.
- Employing *SAR* to enable zero-shot generalizations of in-hand manipulation both with in-domain and out-of-domain unseen objects.

Before presenting our results in Sec. VI, next we outline details of (a) the in-hand reorientation task suite in Sec. V-A), (b) *SAR* representation acquisition in Sec. V-C, and (c) our *SAR* agent and choice of baselines in Sec. V-E.

A. Task specification

We task our agent with controlling *MyoHand* to reorient an object to a particular goal orientation. Both the specific object and the desired reorientation angle are randomly initialized at the beginning of each training episode. This task requires an agent capable of learning a highly flexible motor control policy that is robust to high variance both in object identity and desired reorientation.

Action space. The action vector a_t consisted of continuous values $[-1, 1]$ for contracting each of the 39 muscles in the musculoskeletal MyoHand model. Given that muscle activations cannot be negative [40], MyoSuite transforms this action vector into the range $[0, 1]$ using the sigmoid transform, $\frac{1}{1+e^{-5*(a_t-0.5)}}$.

State space The state vector $s_t = \left\{\phi_t^{act}, \phi_t^{len}, \dot{\phi}_t, \phi_t^F, \psi_t, \dot{\psi}_t, \psi_t^*, \psi_t^e\right\}$ consisted of continuous values for muscle activations ϕ^{act}, muscle length ϕ^{len},

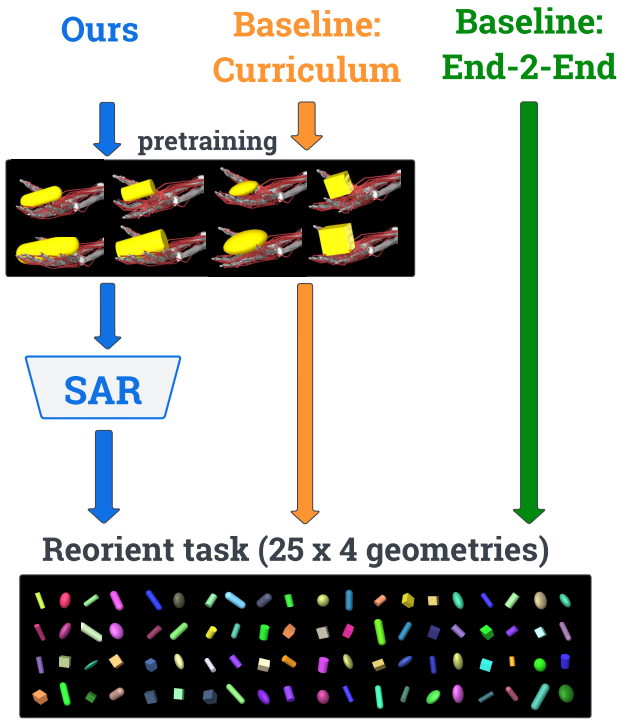

Fig. 5: **Training paradigms for three policy classes.** Ours (left) leverages synergies computed using pre-trained *SAR*. This representation is used to learn a policy that solves reorientations of 100 different objects. Curriculum (center), leverages the pre-training and extends training to re-orient the larger set of objects, without using the *SAR*. End-2-End (right) simply trains directly on the whole set of 100 objects. All policies use identical hyperparameters, including total number of training samples.

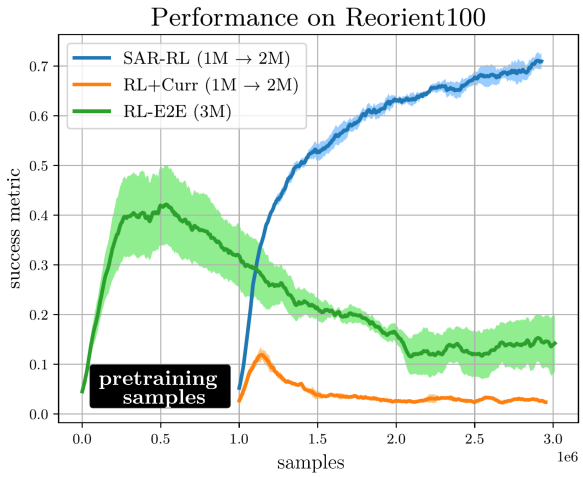

Fig. 6: **Manipulation of 100 objects with *SAR***. Performance learning a multi-task policy to reorient 100 different objects by means of 3 different methods. SAR-RL, depicted in blue, uses the acquired synergistic action representation from a simpler task (1M steps) to facilitate and accelerate learning (2M steps). RL+Curr, depicted in orange, represents the use of *curriculum learning* for 1M steps on a reduced number of objects, followed by finetuning this policy on the full set of 100 objects (2M steps). RL+E2E, depicted in green, represents the use of an *End-2-End* policy that attempts to solve the 100 objects for the whole duration of the learning budget considered (i.e., 3M steps). SAR-RL and RL+Curr are depicted to begin at 1M timesteps in order to account for the 1M steps of pretraining embedded in both methods. The shaded area represents the variability of the policies over 3 different seeds.

muscle velocity $\dot{\phi}$, and muscle force ϕ^F, in addition to object rotation ψ, object velocity $\dot{\psi}$, object goal rotation ψ^*, and object rotational error ψ^e. Note that muscle activations, force, length, and velocity were included in s_t in lieu of joint positions because proprioceptive signals are instantiated in the human nervous system [58], while there is no corresponding evidence that suggests precise joint positions are encoded neurophysiologically.

Reward function We employ the following reward function for learning the multi-object reorientation tasks:

$$R(x_t, \hat{x}_t) := -\lambda_1 \|x_t^{(p)} - \hat{x}_t^{(p)}\|_2 + \lambda_2 |\angle x_t^{(o)} - \hat{x}_t^{(o)}| \} \\ \text{-}\lambda_3 1\{dropped\} - \lambda_4 \|\overline{m}_t\|_2 + \lambda_5 1\{bonus1\} + \lambda_6 1\{bonus2\}$$

where $\angle$ is the quaternion angle between the two current and goal orientations, $x_t^{(p)}$ is the object goal position, $x_t^{(o)}$ is the object goal orientation, $1\{dropped\}$ penalizes object dropping, $1\{bonus1\}$ and $1\{bonus2\}$ incentivize simultaneous rotational and positional alignment above a threshold, and $\bar{m}_t$ encodes overall muscle effort.

Success. We considered a trial successful when object rotational error $\psi^e = |\cos(\psi, \psi^*)|$ was $\leq 5\%$. All trials were 50 timesteps.

B. Reorient dataset and Real-world objects

The objects utilized for the tasks were sampled from two datasets. The *Reorient* dataset consisted of four core geometries in the MuJoCo physics simulator [56]: ellipsoid, cuboid, cylinder, and capsule (i.e., spherocylinder). Those objects could be parameterized to be scaled across the X, Y, and Z axes. In different phases of the study, the 4 different objects were sampled in different ranges of the parameter space. Geometrical dimensions and target poses were determined given both the physiological constraints of the hand and the parameters used in related works involving manipulation [11].

Additionally, a set of real-world objects was obtained by using 8 objects from the ContactDB dataset [9].

C. SAR representation acquisition

Our first step for the definition of the *Synergistic Action Representation* is to train one policy to solve a series of simpler tasks. We sample the 4 objects in the *Reorient* dataset at 2 different sizes (Table A.2). The agent is exposed randomly to all eight of the objects to learn to reorient them. After training the agent for 1M iterations, we can then extract muscle activations which will be transformed into a *SAR* representation that will be used as pretraining (see IV-A).

D. SAR learning and testing generalization

Once *SAR* is computed, we leverage it to train our agent to solve a larger variety of tasks. We use the *SAR*-based policy (see Fig. 4) to solve 25 different geometries (Table A.2) of the 4 objects in the *Reorient* dataset. The set of 100 objects was randomly presented and the policy was trained for 2M iterations.

After training, the policy was evaluated in 3 ways (see Fig. 7). First, in an in-domain test, the policy was tested in a zero-shot inference on 1000 different objects (250 geometries for 4 objects (Table A.2). Second, in an out-of-domain test, an additional 1000 objects (250 geometries for 4 objects) were obtained from a range of parameters outside the one used for learning (Table A.2). Finally, the generalization was tested on a *RealWorldObjs* set (see Sec. V-B; Fig. 7).

E. Baselines

The performance of the *SAR*-based policy was compared against two baselines.

1) Curriculum learning: RL+Curr

First, we compared *SAR* to standard curriculum learning, here named **RL+Curr**, by directly finetuning (i.e., resuming training for) the policy trained on the simpler eight-object task for an additional 2M steps on Reorient100. This baseline (see 'Baseline: Curriculum,' Fig. 5) controls for whether utilizing *any* form of pretraining on a simpler dexterous reorientation environment modulates performance on the more complex target task or whether the synergy action representation *per se* leads to enhanced policies.

2) End-to-end RL: RL-E2E

In addition, we trained a policy end-to-end (**End2End**, or **RL-E2E** for short, see 'Baseline: End-2-End,' Fig. 5) on the full set of 100 objects for 3M timesteps without pretraining or *SAR*. Note that the total training samples and algorithm hyperparameters are identical across all three conditions.

VI. Experimental results

In this section, we present results to show how synergistic action representations uniquely enable learning a single policy that can solve a large set of in-hand reorientations (Sec. VI-A). The same policy is then able to generalize to reorient objects with different shapes that were not part of the training set (Sec. VI-B). Also, the same synergistic representation can be used to learn in-hand reorientation on real objects (Sec. VI-D). Finally, we present a series of ablation studies (Sec. B) investigating our design choices.

A. SAR allows multi-task learning

We started with the hypothesis that learning by means of a *Synergistic Action Representation* (*SAR*) enables learning a single policy capable of solving a large variety of manipulation tasks. We used *SAR* (extracted as described in Section V-C) to train a policy to reorient 100 different objects from the *Reorient* dataset (see Sec. V-B). Figure 6 shows that SAR-RL is able to solve faster and achieve a significantly higher success rate ($> 70\%$) than other methods. In reality, neither learning to solve directly all those tasks (RL-E2E, Fig. 6) nor pre-training (RL+Curr, Fig. 6 orange line) appears to help the generalization to a larger set of objects as indicated by a success rate $< 20\%$ in Figure 6. This indicates that the success at reorienting the object was due to the adoption of *SAR*.

B. SAR zero-shot generalization

Next, we test the hypothesis that the *SAR*-exploiting policy also enables superior zero-shot generalization on 3 different sets of objects (see Fig. 7 - left). First, from the *Reorient* dataset we generated a new set of 1000 in-domain objects within the same range used for the training (*Reorient-ID*). Second, from the *Reorient* dataset we generated a new set of 1000 objects by sampling object dimensions from uniform distributions both above and below those used for the training environment (see Table A.2). This procedure yielded objects that are both larger and smaller than—but never the same size as—those generated in the training environment (*Reorient-OOD*). Third, we tested generalization on 8 different real-world objects (*RealWorldObjs*). Figure 7 shows that the *SAR*-based policy allows generalizing to reorient both in-domain and out-of-domain objects and real-world objects. Indeed, when compared to either end-to-end or curriculum learning, the generalization is $> 3x$ greater for parametric objects and $> 2x$ greater for the real-world objects.

C. SAR stabilizes over learning

Given that *SAR* enables strong generalization, the question that arises is if all synergies were contributing equally to this behavior. Figure 8 shows the normalized mean contributions of each synergy in *SAR* throughout the training at different times. Throughout learning, the first 8 synergies—those that originally contributed to explaining most of the variance of the data (60%, see Fig. A.15)—have a greater contribution to the final solutions as indicated by their larger weight (Fig. 8).

D. SAR transfer learning on Real World Objects

SAR-RL was able to generalize to both parameterized shapes and to real-world objects. Nevertheless, performance on real-world objects seems to be inferior to parameterized objects (see Fig. 7, likely due to the enhanced complexity of their contact dynamics. Accordingly, we investigate if it is possible to use the *SAR* to capture via few-shot learning the details of those real-world objects to reorient them. In this experiment, we leveraged only the first eight synergies as we observed to be the ones contributing most to the generalization.

We find that *SAR* significantly accelerates learning reorientation policies as compared to training without *SAR*, approximately doubling the speed of dexterity acquisition in complex real-world object reorientation tasks (Fig. 9).

VII. Extending *SAR* beyond physiological manipulation

Thus far we demonstrated that SAR enables robust skill transfer for physiological manipulation tasks. To explore the generalizability of the SAR framework to other high-dimensional continuous control settings, we perform two additional sets of experiments on standard continuous control benchmarks. These trials successfully extend SAR to (1) the 20-DoF robotic Shadow Hand in simulation [54], and (2) the Humanoid-v2 gym environment [39], a 17-DoF bipedal walking task.

A. SAR enables robotic skill transfer in Shadow Hand

First, we demonstrate SAR's applicability to robotics problems, demonstrating that SAR enables approximately 2x faster learning of randomized object reorientation on Shadow Hand compared against end-to-end RL (see Fig. 10.

1) Dexterity transfer between geometries

Our first Shadow Hand result demonstrates the generalizability of SAR using synergies computed from policies trained on the official gymnasium-robotics 'HandManipulateEggRotate-v1' environment. It is important to note that the control actions in this environment are absolute angular positions of the actuated joints rather than continuous muscle activations (as in the MyoHand). This requires that the synergistic action representation extracted from this environment would comprise coordinated joint movements rather than muscle co-contractions.

In this experiment, we first train an end-to-end policy to manipulate an egg to a randomized target orientation. We then compute and extract SAR by rolling out this policy. Next, we

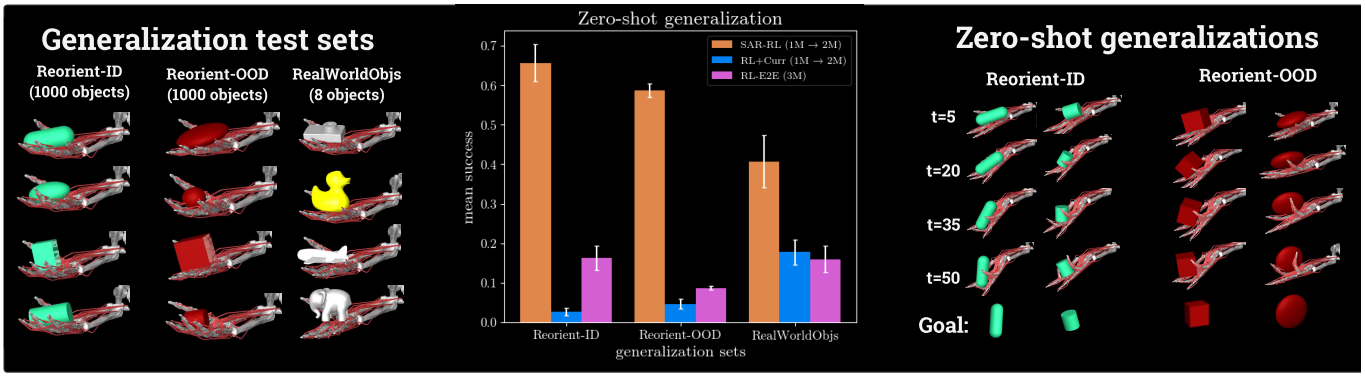

Fig. 7: Left: Test sets used for zero-shot generalization evaluations. Center: Zero-shot testing on 1000 objects obtained within the same parameter set (in-domain, ID) or outside the set of parameters (out-of-domain, OOD) used for generating the objects for training the *SAR* based policy. In addition, a set of eight real-world objects were used. Color coded indicates the *SAR*, curriculum and end-2-end policies described in Figure 6. Right: Behavioral examples of SAR-RL's zero-shot generalization to in-domain and out-of-domain parametric objects.

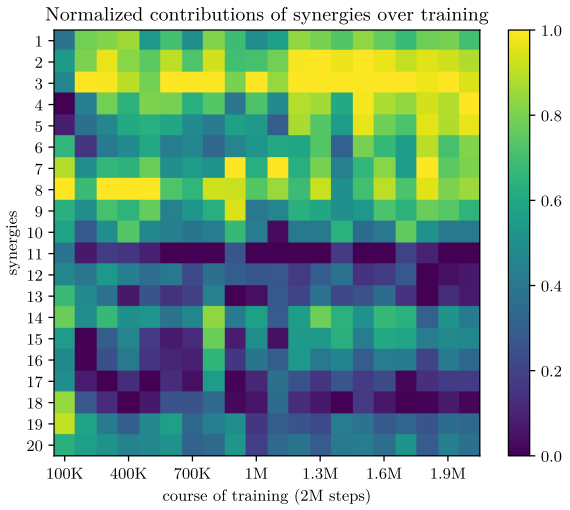

Fig. 8: **Relative contribution of synergies over 100 object training**. Representation of the normalized mean contributions of each synergy over 2M training steps on the 100-object reorientation environment.

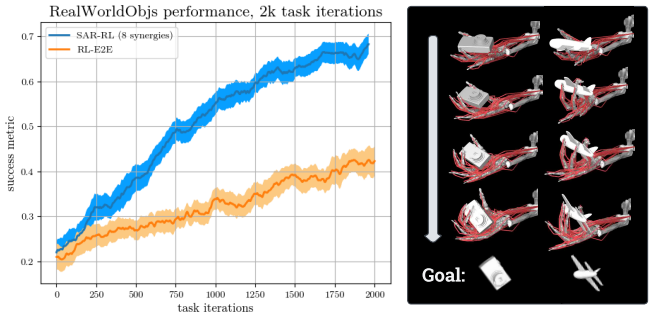

Fig. 9: **Few-shot learning with *SAR* on RealWorldObjs**. Average policy success on the *RealWorldObjs* with SAR-RL (blue) as compared to training without—i.e., RL-E2E (orange).

use SAR (see Fig. 4) to train reorientation policies on two unseen geometries: capsules and cylinders. In comparison to end-to-end RL on cylinder and capsule reorientation, we find that SAR-RL achieves approximately 2x better performance in the same number of samples (see Fig. 10).

2) Replication of Reorient100 experiment

Our second Shadow Hand result reimplements the Reorient100 experiment in the robotic task setting. As in the MyoHand experiment (see Fig. 5), we use an eight-object reorientation task (2 x 4 geometric shapes) to train a policy, compute SAR from rollouts of this trained policy, and use the acquired representation to facilitate learning of the significantly more challenging 100-object (25 x 4 geometric shapes) reorientation task (Fig. 10). We find that, compared to end-to-end training, SAR-RL achieves approximately 2x better performance using the same number of samples on the 100-object task.

B. SAR efficiently yields robust humanoid locomotion

We proceed to extend SAR beyond physiological or robotic hand manipulation, towards full-body motor control on the Humanoid-v2 gym environment. We find that by simply training on this environment for a small number of steps, computing and extracting SAR from this early policy, and using this representation to parameterize the training of a new policy on the same environment, we are able to reach SOTA performance in only 1.25M total training steps (Fig. 11), one to two orders of magnitude more efficiently than competing approaches documented in the literature (see Table I). Of note, we also find that training with SAR yields a significantly more natural gait compared to baseline approaches (i.e., consistent cadence and step length; presence of stance and swing phases; arm swinging and upright posture).

VIII. Discussion

Leveraging SAR computed from simpler tasks as a weighted action representation for learning more complex tasks is

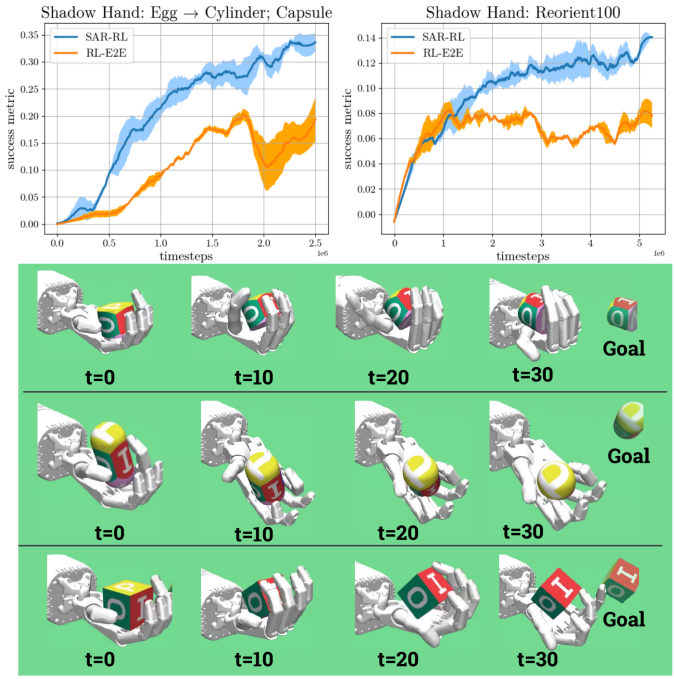

Fig. 10: ***SAR*-enabled dexterity on Shadow Hand**. Top left: Using synergies computed from an egg manipulation policy, SAR-RL achieves approximately 2x better performance on randomized capsule and cylinder reorientation compared to end-to-end RL. Top right: SAR-RL achieves approximately 2x better performance on the Reorient100 task compared to end-to-end RL with synergies computed from an 8-object reorientation policy. Bottom: example manipulations of the policy trained with SAR-RL on Reorient100.

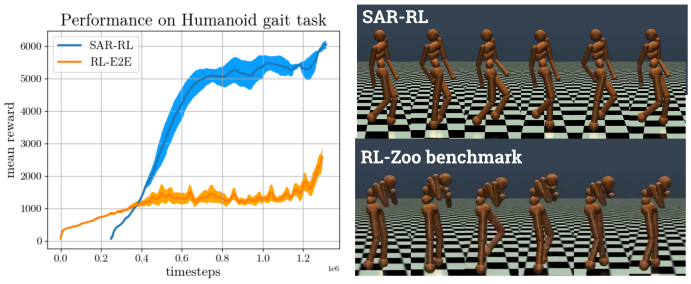

Fig. 11: ***SAR*-enabled locomotion on Humanoid**. Left: SAR-RL yields 3x better performance on Humanoid-v2 than RL-E2E in the same number of total timesteps. Right: behavioral comparison of the locomotion behavior yielded by SAR-RL in comparison to the top RL-Zoo benchmark.

Experiment	Algorithm	Returns at 1.25M samples
SAR (ours)	SAR+SAC	**6104**
Wang & Ni, 2020	meta-SAC	5610
Peng et al, 2010	SAC	4100
Wu et al, 2022	SAC	4600
Toklu et al, 2020	PGPE	715
RL-Zoo baseline	SAC	1552
AI2 baseline	PPO	1120

TABLE I: Comparing SOTA performance on Humanoid-v2 at 1.25M samples.

demonstrated to be a uniquely effective means for enabling strong performance (a) on a large set parameterized objects, (b) on real world objects, and (c) on high-dimensional control problems beyond the scope in-hand physiological manipulation. Alternative solutions that did not leverage synergies did not yield comparably effective or generalizable policies. What accounts for this striking difference in performance?

Framing muscle synergies as an embodied form of transfer learning proves useful for understanding how they enable strong performance on thousands of training and test objects. It is well-known that in humans, skill transfer learning occurs across the entire lifespan and is an important contributing factor to the uniquely sophisticated behavior of humans [3]. Such learning prevents human learners from being required to 'reinvent the wheel' with each new behavioral repertoire by leveraging already-existing stable action representations accrued during lifelong learning. Analogously, in our experiments, normalized contributions of the synergies stabilize over training to utilize a core subset of coordinated muscle co-contraction patterns (Fig. 6). This constitutes strong evidence that a physiologically coherent action representation is transferred from one task to the next.

Framing muscle synergies as an embodied form of transfer learning also suggests why the RL+Curr baseline was unsuccessful in our main experiment. Namely, resuming of training using a task-agnostic play policy on a significantly more complex environment was unsuccessful, while 'concretizing' the dexterity learned in Reorient8 by using SAR stabilizes the representation, which allows for learning a policy that robustly integrates knowledge from across tasks.

Additionally, the successful extension of SAR to robotic and humanoid control demonstrates that the utility of the representation is not constrained to physiological in-hand manipulation and can more generally enable robust high-dimensional continuous control. Taken as a whole, these investigations demonstrate that leveraging synergies—a core mechanism of human motor organization and adaptation—enables a degree of dexterity and agility that was otherwise not reached using baseline learning methods. Accordingly, we propose that learning high-dimensional control using *Synergistic Action Representations* is a promising mechanism for instantiating skill transfer learning and ultimately bootstrapping towards a generalist embodied agent.

IX. Limitations and future research

There are a number of limitations and opportunities for future research from the present study. While the *MyoHand* is physiologically accurate from a structural perspective, there are still fundamental differences between the state and action spaces employed in human motor learning and that of the musculoskeletal model utilized in this research. The most obvious difference is that the *MyoHand* model lacks two sensory modalities: touch and sight. Previous work suggests that both of these sensory pathways in the nervous system help facilitate human-level dexterous manipulation [26] and that a loss of touch in particular can severely compromise dexterity in humans [34]. These facts reveal both a strength and a limitation of this investigation: it is limited insofar as the manipulation policies yielded from the demonstrated synergistic learning paradigm still should not be expected to precisely mirror the behavioral style of human learners given dissimilar sensory data. However, this perceptual deficit also reveals a strength of the synergistic learning paradigm: namely, that sight and touch *were not required* to yield successful dexterous reorientation policies for these complex tasks.

This investigation naturally enables a number of opportunities for future research. First, the method for computing and instantiating muscle synergies developed in this investigation is entirely task-agnostic, meaning that our framework could be conceivably leveraged for any musculoskeletal task set. It is also worth noting that the strategy of using normalized ICAPCA to compute muscle synergies is potentially viable for yielding useful, dimensionality-reduced representations that can be conceivably leveraged for any RL skill transfer problem. In terms of object-level future opportunities, more formally studying the extent to which muscle synergies can be utilized as an embodied form of transfer learning remains highly underexplored and is a promising area of future research.

References

[1] Mazen Al Borno, Jennifer L. Hicks, and Scott L. Delp. The effects of motor modularity on performance, learning and generalizability in upper-extremity reaching: a computational analysis. *Journal of The Royal Society Interface*, 17(167):20200011, June 2020. doi: 10.1098/rsif.2020.0011. URL https://royalsocietypublishing.org/doi/10.1098/rsif.2020.0011. Publisher: Royal Society.

[2] Russell P. Balda, Irene M. Pepperberg, and A. C. Kamil. *Animal Cognition in Nature: The Convergence of Psychology and Biology in Laboratory and Field*. Academic Press, September 1998. ISBN 978-0-08-052723-9. Google-Books-ID: 504iRS01AK0C.

[3] Danielle S. Bassett and Michael S. Gazzaniga. Understanding complexity in the human brain. *Trends in Cognitive Sciences*, 15(5):200–209, May 2011. ISSN 1364-6613. doi: 10.1016/j.tics.2011.03.006. URL https://www.sciencedirect.com/science/article/pii/S1364661311000416.

[4] Marc G. Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The Arcade Learning Environment: An Evaluation Platform for General Agents. *Journal of Artificial Intelligence Research*, 47:253–279, June 2013. ISSN 1076-9757. doi: 10.1613/jair.3912. URL http://arxiv.org/abs/1207.4708. arXiv:1207.4708 [cs].

[5] Antonio Bicchi and Vijay Kumar. Robotic grasping and contact: A review. In *Proceedings 2000 ICRA. Millennium conference. IEEE international conference on robotics and automation. Symposia proceedings (Cat. No. 00CH37065)*, volume 1, pages 348–353. IEEE, 2000.

[6] E. Bizzi, V. C. K. Cheung, A. d'Avella, P. Saltiel, and M. Tresch. Combining modules for movement. *Brain Research Reviews*, 57(1):125–133, January 2008. ISSN 0165-0173. doi: 10.1016/j.brainresrev.2007.08.004.

[7] Emilio Bizzi and Vincent CK Cheung. The neural origin of muscle synergies. *Frontiers in Computational Neuroscience*, 7, 2013. ISSN 1662-5188. URL https://www.frontiersin.org/articles/10.3389/fncom.2013.00051.

[8] Emilio Bizzi, Ferdinando A. Mussa-Ivaldi, and Simon Giszter. Computations underlying the execution of movement: A biological perspective. 253(5017):287–291. doi: 10.1126/science.1857964. URL https://www.science.org/doi/abs/10.1126/science.1857964. Publisher: American Association for the Advancement of Science.

[9] Samarth Brahmbhatt, Cusuh Ham, Charles C. Kemp, and James Hays. ContactDB: Analyzing and Predicting Grasp Contact via Thermal Imaging, April 2019. URL http://arxiv.org/abs/1904.06830. arXiv:1904.06830 [cs].

[10] Vittorio Caggiano, Vincent C. K. Cheung, and Emilio Bizzi. An Optogenetic Demonstration of Motor Modularity in the Mammalian Spinal Cord. *Scientific Reports*, 6(1):35185, October 2016. ISSN 2045-2322. doi: 10.1038/srep35185. URL https://www.nature.com/articles/srep35185. Number: 1 Publisher: Nature Publishing Group.

[11] Vittorio Caggiano, Huawei Wang, Guillaume Durandau, Massimo Sartori, and Vikash Kumar. MyoSuite – A contact-rich simulation suite for musculoskeletal motor control, May 2022. URL http://arxiv.org/abs/2205.13600. arXiv:2205.13600 [cs].

[12] Jiahao Chen and Hong Qiao. Muscle-synergies-based neuromuscular control for motion learning and generalization of a musculoskeletal system. 51(6):3993–4006. ISSN 2168-2232. doi: 10.1109/TSMC.2020.2966818. Conference Name: IEEE Transactions on Systems, Man, and Cybernetics: Systems.

[13] Tao Chen, Jie Xu, and Pulkit Agrawal. A System for General In-Hand Object Re-Orientation. November 2021. URL https://openreview.net/forum?id=7uSBJDoP7tY.

[14] Vincent C. K. Cheung, Ben M. F. Cheung, Janet H. Zhang, Zoe Y. S. Chan, Sophia C. W. Ha, Chao-Ying Chen, and Roy T. H. Cheung. Plasticity of muscle synergies through fractionation and merging during development and training of human runners. *Nature Communications*, 11(1):4356, August 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-18210-4. URL https://www.nature.com/articles/s41467-020-18210-4. Number: 1 Publisher: Nature Publishing Group.

[15] Andrea d'Avella and Emilio Bizzi. Shared and specific muscle synergies in natural motor behaviors. 102(8):3076–3081. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0500199102. URL https://pnas.org/doi/full/10.1073/pnas.0500199102.

[16] Andrea d'Avella, Philippe Saltiel, and Emilio Bizzi. Combinations of muscle synergies in the construction of a natural motor behavior. 6(3):300–308.

[17] Cosimo Della Santina, Matteo Bianchi, Giuseppe Averta, Simone Ciotti, Visar Arapi, Simone Fani, Edoardo Battaglia, Manuel Giuseppe Catalano, Marco Santello, and Antonio Bicchi. Postural hand synergies during environmental constraint exploitation. 11. ISSN 1662-5218. URL https://www.frontiersin.org/articles/10.3389/fnbot.2017.00041.

[18] A. Diamond and O. E. Holland. Reaching control of a full-torso, modelled musculoskeletal robot using muscle synergies emergent under reinforcement learning. *Bioinspiration & Biomimetics*, 9(1):016015, March 2014. ISSN 1748-3190. doi: 10.1088/1748-3182/9/1/016015.

[19] Nadia Dominici, Yuri P. Ivanenko, Germana Cappellini, Andrea d'Avella, Vito Mondì, Marika Cicchese, Adele Fabiano, Tiziana Silei, Ambrogio Di Paolo, Carlo Giannini, Richard E. Poppele, and Francesco Lacquaniti. Locomotor primitives in newborn babies and their development. *Science (New York, N.Y.)*, 334(6058):997–999, November 2011. ISSN 1095-9203. doi: 10.1126/science.1210617.

[20] Fanny Ficuciello, Damiano Zaccara, and Bruno Siciliano. Synergy-based policy improvement with path integrals for anthropomorphic hands. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1940–1945. doi: 10.1109/IROS.2016.7759306. ISSN: 2153-0866.

[21] Henry Gray. *Anatomy of the Human Body*. Lea & Febiger, 1924. Google-Books-ID: RcdqAAAAMAAJ.

[22] S. Grillner. Neurobiological bases of rhythmic motor acts in vertebrates. *Science (New York, N.Y.)*, 228(4696):143–149, April 1985. ISSN 0036-8075. doi: 10.1126/science.3975635.

[23] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learn-

ing with a Stochastic Actor. In *Proceedings of the 35th International Conference on Machine Learning*, pages 1861–1870. PMLR, July 2018. URL https://proceedings.mlr.press/v80/haarnoja18b.html. ISSN: 2640-3498.

[24] Corey B. Hart and Simon F. Giszter. A Neural Basis for Motor Primitives in the Spinal Cord. *Journal of Neuroscience*, 30(4):1322–1336, January 2010. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.5894-08.2010. URL https://www.jneurosci.org/content/30/4/1322. Publisher: Society for Neuroscience Section: Articles.

[25] Y. P. Ivanenko, R. E. Poppele, and F. Lacquaniti. Five basic muscle activation patterns account for muscle activity during human locomotion. *The Journal of Physiology*, 556(1):267–282, 2004. ISSN 1469-7793. doi: 10.1113/jphysiol.2003.057174. URL https://onlinelibrary.wiley.com/doi/abs/10.1113/jphysiol.2003.057174. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1113/jphysiol.2003.057174.

[26] Roland S. Johansson. 19 - Sensory Control of Dexterous Manipulation in Humans. In Alan M. Wing, Patrick Haggard, and J. Randall Flanagan, editors, *Hand and Brain*, pages 381–414. Academic Press, San Diego, January 1996. ISBN 978-0-12-759440-8. doi: 10.1016/B978-012759440-8/50025-6. URL https://www.sciencedirect.com/science/article/pii/B9780127594408500256.

[27] S. L. Kilbreath, R. B. Gorman, J. Raymond, and S. C. Gandevia. Distribution of the forces produced by motor unit activity in the human flexor digitorum profundus. *The Journal of Physiology*, 543(1):289–296, 2002. ISSN 1469-7793. doi: 10.1113/jphysiol.2002.023861. URL https://onlinelibrary.wiley.com/doi/abs/10.1113/jphysiol.2002.023861. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1113/jphysiol.2002.023861.

[28] Petar Kormushev, Sylvain Calinon, and Darwin G. Caldwell. Reinforcement Learning in Robotics: Applications and Real-World Challenges. *Robotics*, 2(3):122–148, September 2013. ISSN 2218-6581. doi: 10.3390/robotics2030122. URL https://www.mdpi.com/2218-6581/2/3/122. Number: 3 Publisher: Multidisciplinary Digital Publishing Institute.

[29] Vikash Kumar, Yuval Tassa, Tom Erez, and Emanuel Todorov. Real-time behaviour synthesis for dynamic hand-manipulation. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6808–6815, May 2014. doi: 10.1109/ICRA.2014.6907864. ISSN: 1050-4729.

[30] Kyo Kutsuzawa and Mitsuhiro Hayashibe. Motor synergy generalization framework for new targets in multi-planar and multi-directional reaching task. *Royal Society Open Science*, 9(5):211721, May 2022. doi: 10.1098/rsos.211721. URL https://royalsocietypublishing.org/doi/full/10.1098/rsos.211721. Publisher: Royal Society.

[31] Jong Hwa Lee, Deanna S. Asakawa, Jack T. Dennerlein, and Devin L. Jindrich. Finger muscle attachments for an opensim upper-extremity model. *PLOS ONE*, 10(4):1–28, 04 2015. doi: 10.1371/journal.pone.0121712. URL https://doi.org/10.1371/journal.pone.0121712.

[32] Ariel J. Levine, Christopher A. Hinckley, Kathryn L. Hilde, Shawn P. Driscoll, Tiffany H. Poon, Jessica M. Montgomery, and Samuel L. Pfaff. Identification of a cellular node for motor control pathways. *Nature Neuroscience*, 17(4):586–593, April 2014. ISSN 1546-1726. doi: 10.1038/nn.3675.

[33] Yinlin Li, Peng Wang, Rui Li, Mo Tao, Zhiyong Liu, and Hong Qiao. A survey of multifingered robotic manipulation: Biological results, structural evolvements, and learning methods. 16. ISSN 1662-5218. URL https://www.frontiersin.org/articles/10.3389/fnbot.2022.843267.

[34] Patrick Luukinen, Olli V. Leppänen, and Jarkko Jokihaara. The effect of digital sensory loss on hand dexterity. *Journal of Hand Surgery (European Volume)*, 46(3):253–259, March 2021. ISSN 1753-1934. doi: 10.1177/1753193420936598. URL https://doi.org/10.1177/1753193420936598. Publisher: SAGE Publications Ltd STM.

[35] Daniel C. McFarland, Benjamin I. Binder-Markey, Jennifer A. Nichols, Sarah J. Wohlman, Marije de Bruin, and Wendy M. Murray. A musculoskeletal model of the hand and wrist capable of simulating functional tasks. *IEEE Transactions on Biomedical Engineering*, pages 1–12, 2022. doi: 10.1109/TBME.2022.3217722.

[36] Igor Mordatch, Zoran Popović, and Emanuel Todorov. Contact-invariant optimization for hand manipulation. In *Proceedings of the ACM SIGGRAPH/Eurographics symposium on computer animation*, pages 137–144, 2012.

[37] OpenAI, Ilge Akkaya, Marcin Andrychowicz, Maciek Chociej, Mateusz Litwin, Bob McGrew, Arthur Petron, Alex Paino, Matthias Plappert, Glenn Powell, Raphael Ribas, Jonas Schneider, Nikolas Tezak, Jerry Tworek, Peter Welinder, Lilian Weng, Qiming Yuan, Wojciech Zaremba, and Lei Zhang. Solving Rubik's Cube with a Robot Hand, October 2019. URL http://arxiv.org/abs/1910.07113. arXiv:1910.07113 [cs, stat].

[38] Simon A. Overduin, Andrea d'Avella, Jose M. Carmena, and Emilio Bizzi. Microstimulation activates a handful of muscle synergies. 76(6):1071–1077. ISSN 1097-4199. doi: 10.1016/j.neuron.2012.10.018.

[39] Matthias Plappert, Marcin Andrychowicz, Alex Ray, Bob McGrew, Bowen Baker, Glenn Powell, Jonas Schneider, Josh Tobin, Maciek Chociej, Peter Welinder, Vikash Kumar, and Wojciech Zaremba. Multi-goal reinforcement learning: Challenging robotics environments and request for research, 2018.

[40] Mohammad Fazle Rabbi, Claudio Pizzolato, David G. Lloyd, Chris P. Carty, Daniel Devaprakash, and Laura E. Diamond. Non-negative matrix factorisation is the most appropriate method for extraction of muscle synergies in walking and running. *Scientific Reports*, 10(1):8266, May 2020. ISSN 2045-2322. doi: 10.1038/s41598-020-65257-w. URL https://www.nature.com/articles/s41598-020-65257-w. Number: 1 Publisher: Nature Publishing Group.

[41] Antonin Raffin, Jens Kober, and Freek Stulp. Smooth Exploration for Robotic Reinforcement Learning, June 2021. URL http://arxiv.org/abs/2005.05719. arXiv:2005.05719 [cs, stat].

[42] Aravind Rajeswaran, Vikash Kumar, Abhishek Gupta, Giulia Vezzani, John Schulman, Emanuel Todorov, and Sergey Levine. Learning Complex Dexterous Manipulation with Deep Reinforcement Learning and Demonstrations, June 2018. URL http://arxiv.org/abs/1709.10087. arXiv:1709.10087 [cs].

[43] Eric Rombokas, Mark Malhotra, Evangelos A. Theodorou, Emo Todorov, and Yoky Matsuoka. Reinforcement learning and synergistic control of the ACT hand. 18(2):569–577. ISSN 1083-4435, 1941-014X. doi: 10.1109/TMECH.2012.2219880. URL http://ieeexplore.ieee.org/document/6341113/.

[44] Elmar Rückert and Andrea d'Avella. Learned parametrized dynamic movement primitives with shared synergies for controlling robotic and musculoskeletal systems. *Frontiers in Computational Neuroscience*, 7:138, October 2013. ISSN 1662-5188. doi: 10.3389/fncom.2013.00138. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3797962/.

[45] Akira Saito, Aya Tomita, Ryosuke Ando, Kohei Watanabe, and Hiroshi Akima. Muscle synergies are consistent across level and uphill treadmill running. *Scientific Reports*, 8(1):5979, April 2018. ISSN 2045-2322. doi: 10.1038/s41598-018-24332-z. URL https://www.nature.com/articles/s41598-018-24332-z. Number: 1 Publisher: Nature Publishing Group.

[46] Gionata Salvietti. Replicating Human Hand Synergies Onto Robotic Hands: A Review on Software and Hardware Strategies. *Frontiers in Neurorobotics*, 12, 2018. ISSN 1662-5218. URL https://www.frontiersin.org/articles/10.3389/fnbot.2018.00027.

[47] Marco Santello, Martha Flanders, and John F. Soechting. Postural hand synergies for tool use. 18(23):10105–10115. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.18-23-10105.1998. URL https://www.jneurosci.org/content/18/23/10105. Publisher: Society for Neuroscience Section: ARTICLE.

[48] Matteo Saveriano, Fares J. Abu-Dakka, Aljaz Kramberger, and Luka Peternel. Dynamic Movement Primitives in Robotics: A Tutorial Survey, February 2021. URL http://arxiv.org/abs/2102.03861. arXiv:2102.03861 [cs].

[49] Pierre Schumacher, Daniel Häufle, Dieter Büchler, Syn Schmitt, and Georg Martius. DEP-RL: Embodied Exploration for Reinforcement Learning in Overactuated and Musculoskeletal Systems, May 2022. URL http://arxiv.org/abs/2206.00484. arXiv:2206.00484 [cs].

[50] Anton R. Sobinov and Sliman J. Bensmaia. The neural mechanisms of manual dexterity. *Nature Reviews Neuroscience*, 22(12):741–757, December 2021. ISSN 1471-0048. doi: 10.1038/s41583-021-00528-7. URL https://www.nature.com/articles/s41583-021-00528-7. Number: 12 Publisher: Nature Publishing Group.

[51] Yunqing Song, Masaya Hirashima, and Tomohiko Takei. Neural Network Models for Spinal Implementation of Muscle Synergies. *Frontiers in Systems Neuroscience*, 16, 2022. ISSN 1662-5137. URL https://www.frontiersin.org/articles/10.3389/fnsys.2022.800628.

[52] Richard S. Sutton and Andrew G. Barto. *Reinforcement learning: An introduction, 2nd ed*. Reinforcement learning: An introduction, 2nd ed. The MIT Press, Cambridge, MA, US, 2018. ISBN 978-0-262-03924-6. Pages: xxii, 526.

[53] Juri Taborri, Valentina Agostini, Panagiotis K. Artemiadis, Marco Ghislieri, Daniel A. Jacobs, Jinsook Roh, and Stefano Rossi. Feasibility of muscle synergy outcomes in clinics, robotics, and sports: A systematic review. 2018:1–19. ISSN 1176-2322, 1754-2103. doi: 10.1155/2018/3934698. URL https://www.hindawi.com/journals/abb/2018/3934698/.

[54] Yuval Tassa, Tom Erez, and Emanuel Todorov. Synthesis and stabilization of complex behaviors through online trajectory optimization. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4906–4913, October 2012. doi: 10.1109/IROS.2012.6386025. ISSN: 2153-0866.

[55] Emanuel Todorov and Zoubin Ghahramani. Analysis of the synergies underlying complex hand manipulation. In *The 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, volume 2, pages 4637–4640. IEEE, 2004.

[56] Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033, October 2012. doi: 10.1109/IROS.2012.6386109. ISSN: 2153-0866.

[57] Matthew C. Tresch, Vincent C. K. Cheung, and Andrea d'Avella. Matrix factorization algorithms for the identification of muscle synergies: evaluation on simulated and experimental data sets. *Journal of Neurophysiology*, 95(4):2199–2212, April 2006. ISSN 0022-3077. doi: 10.1152/jn.00222.2005.

[58] John C. Tuthill and Eiman Azim. Proprioception. *Current Biology*, 28(5):R194–R203, March 2018. ISSN 0960-9822. doi: 10.1016/j.cub.2018.01.064. URL https://www.sciencedirect.com/science/article/pii/S0960982218300976.

[59] Aleš Ude, Andrej Gams, Tamim Asfour, and Jun Morimoto. Task-Specific Generalization of Discrete and Periodic Dynamic Movement Primitives. *IEEE Transactions on Robotics*, 26(5):800–815, October 2010. ISSN 1941-0468. doi: 10.1109/TRO.2010.2065430. Conference Name: IEEE Transactions on Robotics.

[60] Peter C. Wainwright. The evolution of feeding motor patterns in vertebrates. *Current Opinion in Neurobiology*, 12:691–695, 2002. ISSN 1873-6882. doi: 10.1016/S0959-4388(02)00383-5. Place: Netherlands Publisher: Elsevier Science.

[61] T. Y. Wang, T. Bhatt, F. Yang, and Y. C. Pai. Generalization of motor adaptation to repeated-slip perturbation across tasks. *Neuroscience*, 180:85–95, April 2011. ISSN 0306-4522. doi: 10.1016/j.neuroscience.2011.02.039. URL https://www.sciencedirect.com/science/article/pii/S0306452211002090.

[62] Chenguang Yang, Chao Zeng, Cheng Fang, Wei He, and Zhijun Li. A DMPs-Based Framework for Robot Learning and Generalization of Humanlike Variable Impedance Skills. *IEEE/ASME Transactions on Mechatronics*, 23(3):1193–1203, June 2018. ISSN 1941-014X. doi: 10.1109/TMECH.2018.2817589. Conference Name: IEEE/ASME Transactions on Mechatronics.

[63] He Zhang, Yuting Ye, Takaaki Shiratori, and Taku Komura. Manipnet: neural manipulation synthesis with a hand-object spatial representation. *ACM Transactions on Graphics (ToG)*, 40(4):1–14, 2021.

APPENDIX

A. Real world objects set used in this study

We begin by presenting the complete set of real world object we used in our experiments in Sec.VI-B.

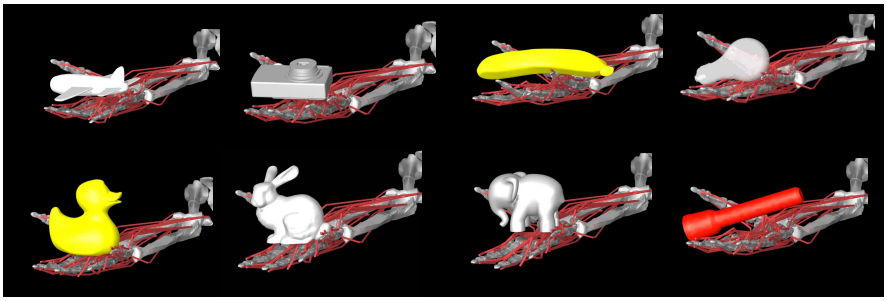

Fig. A.12: Eight objects used for RealWorldObjs data set (see Sec. VI-B).

B. Ablations

We present a series of ablation experiments to understand how our policy and representation design choices in the Reorient100 experiment impacted results.

1) Relative contribution of synergies to the solution

First, we investigate the choice and sensitivity of our blend weight (φ, see Sec. IV-B) between the synergistic and non-synergistic dimensions. We analyzed the contribution of *SAR* in terms of success rate in Figure A.13. While an action contribution of *SAR* between 0.6 and 0.8 does not change greatly the results, the solution greatly benefits from the use of *SAR*. Indeed, choosing a very weak contribution or no contribution of the synergies negatively impacts the overall performance, leading to success rate < 0.25. This result indicates that SAR is in fact utilized by our trained policy.

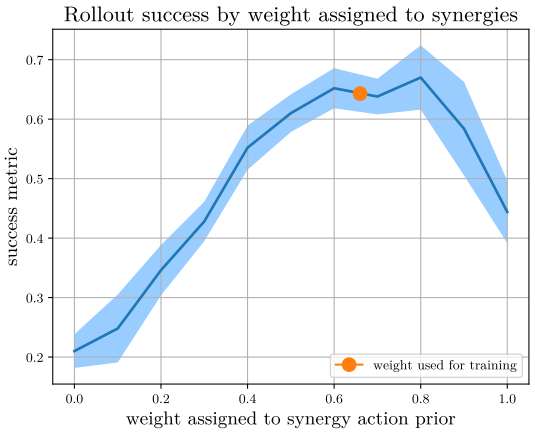

Fig. A.13: **Post hoc modulation of weight assigned to *SAR*** Analysis of the relationship of φ to the success rate. The value of φ used for the main experiment is depicted in orange.

2) Relative contribution of most vs least meaningful synergies

We have demonstrated (Fig. 8) that during training, the earlier synergies that explain most of the variance of the original signal have an outsized contribution to the final policy. Nevertheless, a question arises whether any choice of synergies produce the same results? Here, we tested the contribution of the synergies that contribute most versus the ones that contributed the least to the original tasks on which the *SAR* was built (see Appendix Fig. A.15). Figure A.14 shows that when the least informative synergies for the pre-training task were chosen to compute *SAR*, their ability to facilitate the learning of a larger set of reorientation was heavily compromised, reaching about half of the performance when more significant synergies were utilized for *SAR*. Overall, this result is consonant with the observation of Figure 8 where the use of synergies that capture a disproportionate amount of the information from the

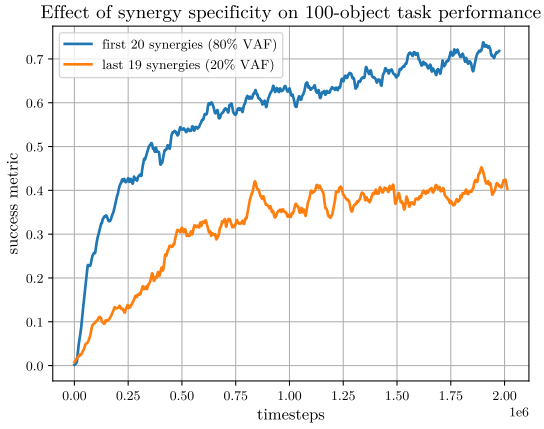

Fig. A.14: **Comparison SAR performance using more vs. less informative synergies**. Utilizing synergies that explain more variance in a muscle activation dataset yields more robust policies. It should be noted that both of these policies significantly outperform the two baselines tested in this investigation (see Fig. 6).

signal—i.e. first synergies have higher VAF (see Fig. A.15)—also contribute a disproportionate amount to the solution.

C. Dimensionality (N) of the Synergistic Action Representation (SAR)

For the purposes of this investigation, SAR is constructed from the first 20 synergies computed from the task-agnostic play period described in Sec. IV-A. This quantity of muscle synergies was selected because it balances (a) explaining a large proportion of the variance ($> 80\%$) in the initial activation data while simultaneously (b) reducing the dimensionality of the original data by a factor of 2 (see Fig. A.15). We also ablate in Fig. A.14 that training with synergies that explain more variance in the the muscle activation data leads to improved performance as compared to training with later synergies.

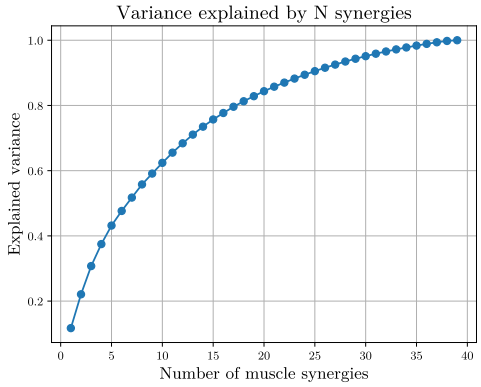

Fig. A.15: Proportion of variance in muscle activation data explained from task-agnostic play policy as the number of computed muscle synergies increases.

D. Synergistic Action Representation vs random action representation

Additionally, in order to further demonstrate the utility of the latent representation captured by SAR, we compare performance between training with SAR and training with randomized PCA and ICA component matrices that share the same dimensions as SAR (see Fig. A.16). This analysis demonstrates that randomized representations are ineffective for learning the multi-object reorientation task, which further indicates that SAR *per se* has a representational utility.

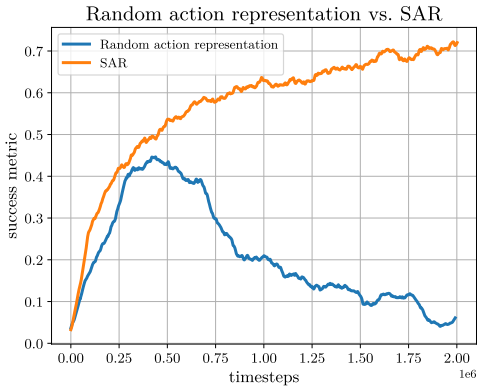

Fig. A.16: Training performance when SAR is computed as described in Sec. IV-A as compared to initializing a random representation.

E. Effect of blend weight (φ) on effectiveness of SAR

An important consideration for our investigation was the optimal blend weight, φ^*, for mixing synergistic and nonsynergistic activations for each final action, a_t^* (see Fig. 4). In Fig. A.17 we present an ablation study involving various choices of blend weight. We observe that an approximate ratio of $2/3$ SAR activations, $1/3$ task-specific activations (i.e., $\varphi \approx .66$) facilitated best performance. More specifically, we find that as as φ is increased from this value and SAR is more heavily weighted in a_t^*, performance decrements slightly (0.7 success rate $\rightarrow$ 0.5-0.6 success rate). As φ is decreased from this value and SAR is less heavily weighted in a_t^*, performance decrements significantly (0.7

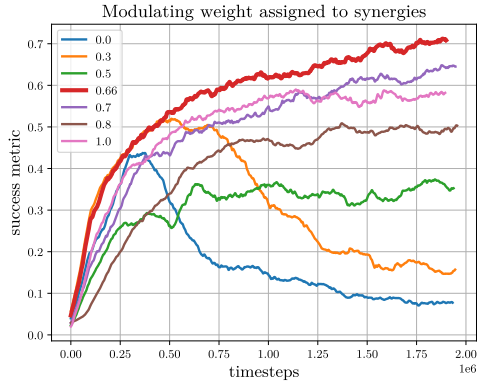

Fig. A.17: Training performance at different values of φ. In bolded red is $\varphi^* = .66$, the value used for the main experiment. We find that as φ is increased from this value (i.e., SAR is given more weight in a_t^*), performance decrements slightly (0.7 $\rightarrow$ 0.5-0.6 success rate). As φ is decreased from this value (i.e., SAR is given less weight in a_t^*), performance decrements significantly (0.7 $\rightarrow$ 0.1-0.4 success rate).

success rate $\rightarrow$ 0.1-0.4 success rate) (see Fig. A.17).

This result can be thought as *proactively* searching over values of φ—in other words, different values can be specified for φ prior to training, and variation in performance can be compared. In contrast, we can also consider *retroactively* searching over values of φ—in other words, SAR-RL trained at some value (.66 in this case) of φ, but the φ passed into the trained policy can be modulated during test to assess the impact of SAR on a_t^* on the policy. This latter approach is precisely what was done for SAR-RL in Fig. A.13.

Thus, comparing proactive and retroactive modulation of φ yields a similar result: training performance is best at $\varphi \approx .66$, decreases slightly as $\varphi > .66$, and decreases significantly as $\varphi < .66$. This result indicates that the combination of synergistic and nonsynergistic activations leads to best performance (see Sec. IV-A), but that more heavily weighting SAR within this mix is required for achieving this level of performance. It should be emphasized that at the extremes, $\varphi = 0$—i.e., training with no SAR—leads to poor performance (success rate < 0.2), while $\varphi = 1$—i.e., training only with SAR—leads to significantly improved but still suboptimal performance (success rate ≈ 0.6) (see Fig. A.17).

F. Parameter and hyperparameter selection

We present parameters utilized for the training and testing regimes of our main experiments (see Secs. VI-B; V-C). We selected parameters for the X, Y, and Z axes of the parametric objects used in Secs. VI-B and V-C by sampling over ranges for each geometry that were conducive to in-hand manipulation (i.e., excluding shapes that were too large to fit in-hand or too small to properly manipulate). (see Table A.2; Fig. 5)

Additionally, we utilize hyperparameters for SAC (Table A.3) inpsired by previous successful utilizations of this learning algorithm for continuous control in robotic RL [41]. We also include a linearly-scheduled learning rate to prevent unstable learning as the end of the training regime is approached.

TABLE A.2: Parameter distributions used for generating pretraining, in-domain, and out-of-domain parametric objects (see Sec. VI-B). Note that for out-of-domain samples, either a larger or smaller uniform distribution was first randomly selected from which to sample parameters for each specified axis.

Geometry	Pretraining	In-domain	Out-of-domain
Ellipsoid	obj1=[0.011,0.025,0.025], obj2=[0.019,0.040,0.040]	$X \sim \mathcal{U}_{[0.008,0.02]}$, $Y, Z \sim \mathcal{U}_{[0.020,0.045]}$	$X \sim \mathcal{U}_{[0.008,0.02]}$, $Y, Z \sim \mathcal{U}_{[0.015,0.020]} \vee \mathcal{U}_{[0.045,0.050]}$
Box	obj1=[0.017,0.017,0.017], obj2=[0.023,0.023,0.023]	$X, Y, Z \sim \mathcal{U}_{[0.015,0.025]}$	$X, Y, Z \sim \mathcal{U}_{[0.025,0.030]} \vee \mathcal{U}_{[0.010,0.015]}$
Capsule	obj1=[0.013,0.025,0.025], obj2=[0.019,0.040,0.040]	$X \sim \mathcal{U}_{[0.010,0.022]}$, $Y, Z \sim \mathcal{U}_{[0.020,0.045]}$	$X \sim \mathcal{U}_{[0.010,0.022]}$, $Y, Z \sim \mathcal{U}_{[0.015,0.020]} \vee \mathcal{U}_{[0.045,0.050]}$
Cylinder	obj1=[0.013,0.025,0.025], obj2=[0.019,0.040,0.040]	$X \sim \mathcal{U}_{[0.010,0.022]}$, $Y, Z \sim \mathcal{U}_{[0.020,0.045]}$	$X \sim \mathcal{U}_{[0.010,0.022]}$, $Y, Z \sim \mathcal{U}_{[0.015,0.020]} \vee \mathcal{U}_{[0.045,0.050]}$

TABLE A.3: Hyperparameters used for training with SAC across all experiments. The stable-baselines3 implementation of SAC was utilized for this investigation (see Sec. III).

Soft Actor-Critic (SAC) hyperparameters	
Policy type	MlpPolicy
Actor architecture	$[400, 300]$
Critic architecture	$[400, 300]$
Learning rate	$linearschedule(.001)$
Start learning	$t = 3000$
Batch size	256
τ	.02
γ	.98