MIMo: A Multi-Modal Infant Model for Studying Cognitive Development in Humans and AIs

Dominik Mattern
Dept. of Computer Science and Mathematics
Goethe-University Frankfurt
Frankfurt am Main, Germany
d.mattern@stud.uni-frankfurt.de

Francisco M. López
Frankfurt Institute for Advanced Studies
Frankfurt am Main, Germany
lopez@fias.uni-frankfurt.de

Markus R. Ernst
Frankfurt Institute for Advanced Studies
Frankfurt am Main, Germany
mernst@fias.uni-frankfurt.de

Arthur Aubret
University Clermont Auvergne
CNRS, Pascal institute
Clermont-Ferrand, France
arthur.aubret@uca.fr

Jochen Triesch
Frankfurt Institute for Advanced Studies
Frankfurt am Main, Germany
triesch@fias.uni-frankfurt.de

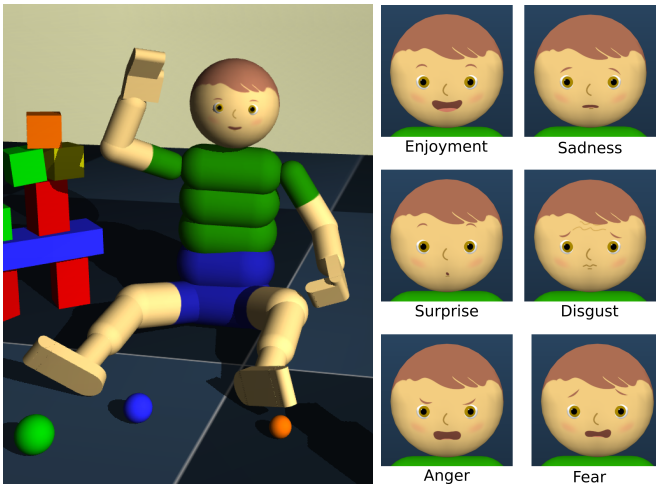
***Abstract*—A central challenge in the early cognitive development of humans is making sense of the rich multimodal experiences originating from interactions with the physical world. AIs that learn in an autonomous and open-ended fashion based on multimodal sensory input face a similar challenge. To study such development and learning *in silico*, we have created MIMo, a multimodal infant model. MIMo's body is modeled after an 18-month-old child and features binocular vision, a vestibular system, proprioception, and touch perception through a full-body virtual skin. MIMo is an open-source research platform based on the MuJoCo physics engine for constructing computational models of human cognitive development as well as studying open-ended autonomous learning in AI. We describe the design and interfaces of MIMo and provide examples illustrating its use.**

***Index Terms*—cognitive development, developmental AI, infant model, multimodal perception, physics simulation**

I. Introduction

What does it mean to *understand* cognitive development? While there are many possible answers to this question, a good measure of our understanding of cognitive development is our ability to *rebuild* it. Rebuilding cognitive development means to construct a computational model of the developing brain controlling the developing body. This entails learning to interact with the environment in a way that matches that of humans and explains the mental representations forming during the process and the underlying brain mechanisms. The idea of building computers that can learn more autonomously like children has also been a major motivation of AI research, that can be traced back all the way to Turing [1]. However such "Developmental Robotics" or "Developmental AI" efforts have become more common only in the last 20 years, for reviews see [2]–[7].

Human cognitive development critically depends on the embodied interaction with a physical environment, and this may be just as important for human-like AIs with a commonsense understanding of the physical world. Thus, recreating the physical interaction with the environment is a central aspect of rebuilding cognitive development. There are two options for this. The first option is using humanoid robots as is common in Developmental Robotics research. The main advantage of this approach is its realism. However, working with humanoid robots tends to be expensive, time-consuming, and suffers from the brittleness of today's humanoid hardware. Importantly, all these factors impede the reproducibility of research. Furthermore, the sensing abilities of today's robots are usually not comparable to that of humans. This is particularly problematic for the sense of touch. The human body is covered by a flexible skin containing different kinds of mechanoreceptors, thermoreceptors, and nociceptors (pain receptors), which allow us to sense touch, pressure, vibration,



(a) Full body view (b) Facial expressions

Fig. 1. MIMo, the multimodal infant model.

This research was supported by "The Adaptive Mind" and "The Third Wave of Artificial Intelligence" funded by the Excellence Program of the Hessian Ministry of Higher Education, Science, Research and Art. JT was supported by the Johanna Quandt foundation.

temperature, and pain. Reproducing such a human-like skin in humanoid robots is still out of reach.

The second option is modeling the physical interaction *in silico*. This requires physics simulators or game engines that approximate the physics of such interactions. Many such simulators are available today [8]–[10], for review see [11]. Disadvantages of this approach are inevitable approximations leading to inaccuracies of such simulations and the high computational costs, especially when non-rigid body parts and objects are considered. However, this approach avoids all the problems of working with humanoid robot hardware mentioned above and ensures perfect reproducibility. Furthermore, with powerful compute infrastructures simulations can run faster than real-time, facilitating the simulation of developmental processes unfolding over months and years.

To support such research, we have developed MIMo, a Multi-Modal Infant Model (Fig. 1). MIMo is a research platform for developing computational models of human cognitive development and building developmental AIs. Its body is modeled after an average 18-month-old child. MIMo has 38 degrees of freedom of the body, 6 degrees of freedom of the eyes, and allows the animation of different facial expressions (for studies on social development). MIMo's interaction with the physical environment is simulated using the MuJoCo physics engine [12], which is particularly strong at simulating contact physics including friction. In the design of MIMo we have aimed for a balance between realism and computational efficiency. To accelerate the simulation of physics and touch sensation, MIMo's body is composed of simple rigid shape primitives such as a sphere for the head and capsules for most other body parts. In this first release of MIMo, it is equipped with four sensory modalities: binocular vision, proprioception, full-body touch sensation, and a vestibular system. In the following we describe the design of MIMo and illustrate how it may be used in developmental science and AI research. Concretely, we consider three scenarios where MIMo learns to 1) reach for an object, 2) stand up, 3) touch different locations on his body.

The remainder of this article is organized as follows. In Section II we briefly review related efforts to create platforms for simulating cognitive development. In Section III we present the physical design of MIMo and in Section IV we describe the sensory modalities. Section V summarizes MIMo's application programming interface (API) and Section VI presents some experiments to measure performance and illustrate potential uses of MIMo. Finally, Section VII discusses our approach and points out directions for future work. Our code is available at https://github.com/trieschlab/MIMo.

II. Related Work

We focus on two major classes of software platforms for simulating cognitive development during embodied interactions with the environment. The first kind is designed to simulate a particular physical robot used in Developmental Robotics research and intended to complement the work with that physical robot. Examples are the iCub simulator [13] and the simulator for the NICO robot [14] that has been implemented using the V-Rep robotics simulation environment [9]. Such platforms typically aim to faithfully reproduce the design and behavior of the physical robot. Doing so helps to reduce the notorious gap between simulation and real world. However, such platforms also inherit any shortcomings of the robot design relative to the human body and human sensing capabilities. For example, if the robot possesses only poor touch sensation, its simulated counterpart will suffer from the same limitation.

The second kind of platform emulates human body and sensing abilities directly and thus is not restricted by limitations of current robotics technology in general or that of specific robots in particular. An early example is the seminal work by Kuniyoshi and Sangawa [15]. More frequently, simulation models of specific aspects of sensorimotor development have been proposed. These typically encompass only a small subset of degrees of freedom and sensory modalities. An early example is work on the development of grasping by Oztop and colleagues [16]. A more recent example is the OpenEyeSim simulator, which has been designed to support modeling the development of active binocular vision [17] and is built using the OpenSim software for simulating neuromusculoskeletal systems [18]. While widely used in reinforcement learning for locomotion tasks [19], standard humanoids introduced within the MuJoCo [20] or Bullet [10] platforms only incorporate very limited haptic modality and do not model a child-like appearance.

III. Physical Design

MIMo's overall body dimensions and proportions were adapted from anthropometric measurements of 16–19 month old infants [21], treating the unit of distance in MuJoCo as one meter. To keep computational costs down he is composed of simple geometric primitives.

All joints are modeled as a series of one-axis hinges and split into flexion/extension, abduction/adduction or internal/external rotation as appropriate for the joint. For the shoulders we merged the commonly used abduction and flexion axes, instead using horizontal flexion, abduction and internal rotation. We found this necessary to keep the range of motion realistic as MuJoCo does not allow for sufficiently complex joint limits without modeling the actual muscle and bone structure. Keeping abduction and flexion as two separate axes would allow MIMo to "double-dip" on both axes depending on the position of the other two.

Muscles are modeled as a combination of a motor and a weak spring directly applying torque at the joint. The motor acts as the voluntary muscle force while the spring will weakly try to return the joint to its neutral position. Finally the motion of each joint is weakly damped, loosely modeling the decreasing force of muscles with increasing velocity. This model is an intentional trade-off between accuracy and complexity. We plan to provide an additional actuation model using MuJoCo's built-in muscle model in the future.

Since there does not appear to exist a single source providing a full set of range of motion or strength measurements for our target age of about 18 months, we inferred data from a large number of sources [22]–[33]. For range of motion we took values from the youngest age reported in the various studies and assumed no change. For muscle strengths we took data from [22] as a baseline. These authors consider children aged 3–9 and we used values from the lower end of this range for all joints reported. Values for other joints were taken from studies on adults or older children and then scaled down for MIMo. We did this by assuming that the relative strengths of joints stay constant, using the strengths of the knee or elbow from [22] as reference values. Where required we converted forces to torques using the appropriate lever arms from MIMo based on the methodologies used in the respective source.

Table I shows the range of motion and strengths for all joints. We treat extension, adduction and internal rotation as positive and flexion, abduction and external rotation as negative. Citations show the source of the data, while entries without marked sources indicate best guesses based on the other values.

By default, MIMo has a neutral facial expression. For studies of social learning in multi-agent setups we created six additional facial expressions. These correspond to the six basic emotions proposed by [34] (enjoyment, sadness, surprise, disgust, anger, and fear) and enable MIMo to convey an internal emotional state. Expressions are implemented as textures of the head sphere (see Fig. 1).

TABLE I
JOINT RANGE OF MOTION AND STRENGTH FOR MIMO.

Joint	ROM [°]	Voluntary Torque [Nm]
Neck flexion/ext.	-70 [22] to 80 [22]	-1.17 [26]* to 2.10 [26]*
Neck lateral flex.	-70 [27] to 70 [27]	-1.17 to 1.17
Neck rotation	-111 [27] to 111 [27]	-1.17 to 1.17
Trunk flexion/ext.	-61 [25] to 34 [25]	-8.13 [25]* to 10.58 [25]*
Trunk lateral flex.	-41 [25] to 41 [25]	-7.25 [25]* to 7.25 [25]*
Trunk rotation	-36 [25] to 36 [25]	-3.63 [25]* to 3.63 [25]*
Shoulder horizontal	-118 [29] to 28 [29]	-1.8 [31]* to 1.8 [31]*
Shoulder flexion/ext.	-183 [28] to 84 [28]	-2.75 [30]* to 4 [30]*
Shoulder rotation	-99 [22] to 67 [22]	-1.6 [30]* to 2.5 [30]*
Elbow flexion/ext.	-146 [22] to 5 [28]	-3.6 [22] to 3.0 [22]
Wrist palmar/dorsi	-92 [28] to 86 [28]	-1.24 [35] to 0.7 [23]*
Wrist ulnar/radial	-53 [32] to 48 [32]	-0.83 [35] to 0.95 [35]
Wrist rotation	-90 [28] to 90 [28]	-0.7 to 0.7
Mitt fingers flex/ext.	-160 to 8	-0.69 [22] to 0.23
Hip flexion/ext.	-133 [22] to 20 [24]	-8 [23]* to 8* [23]*
Hip ab-/adduction	-51 [24] to 17 [24]	-6.24 [23]* to 6.24 [22]
Hip rotation	-32 [22] to 41 [22]	-2.66 [22] to 3.54 [22]
Knee flexion/ext.	-145 [22] to 4 [22]	-6.5 [22] to 10 [22]
Ankle plantar/dorsi	-63 [22] to 32 [22]	-3.78 [22] to 1.89 [22]
Ankle e-/inversion	-33 [33] to 31 [33]	-1.06 [33]* to 1.16 [33]*
Ankle rotation	-20 to 30	-1.2 to 1.2

* Reported value scaled to be proportional to knee or elbow reference.

IV. MULTIMODAL SENSING

A. Binocular Vision

Vision is arguably the most informative and reliable sensory modality for humans, yet it is the latest one to be developed. Newborns have poor visual systems with low sensitivities to depth, movement, and color. Concurrently with the development of the visual neural pathway, infants learn to perform coordinated movements with both eyes. Binocular vision is typically fully functional at 5 months of age but continues to improve throughout childhood (see [36] for a review).

MIMo is equipped with two cameras, situated on the surfaces of the two independently-controlled eyeballs. Each eye can rotate on three orthogonal axes resulting in horizontal and vertical movements, often called pan and tilt, as well as torsional movement. The range of motion is $\pm 45°$ horizontally, -47° to 33° vertically [37] and $\pm 8°$ torsionally [38]. The cameras render two RGB images with a 60º field of view, equivalent to the central vision of humans [39]. Beyond this range, visual acuity and color perception drop significantly.

B. Proprioception

Proprioception in mammals is driven by two main groups of sensory organs [40]. Muscle spindles run in parallel to the muscles and contain multiple types of nerve structures. These measure the length and the rate of change of the length of the muscle. The spindles for all the muscles acting on a given joint thus effectively measure the position and velocity of the joint. Golgi tendon organs are embedded at the connection points between muscles and their tendons and measure the effective load on the muscle. In addition to these main groups there are mechanoreceptors in the joints. These support joint position sensing but they respond strongest when a limb is reaching the limit of its range of motion, effectively acting as limit sensors.

MIMo does not simulate this diverse set of receptors explicitly, but tries to capture the essence of the complex proprioception system. Specifically, position and velocity is measured for each joint degree of freedom. In addition, torque sensors for each joint measure the applied torque, i.e., the sum of applied motor torque, external torques and apparent torques due to inertia. Finally there are limit sensors that activate when a joint moves to within a certain threshold of its range of motion limit, with the activation increasing linearly as the joint approaches the limit.

C. Vestibular System

The vestibular system is composed of the semicircular canals, the urticule, and the saccule, all located in the inner ears of humans and other mammals. This sensory modality provides information about the linear and rotational acceleration of the head, in particular indicating the direction of gravity to maintain balance. An impaired vestibular function during infancy results in deficits or delays in motor development [41].

MIMo's vestibular system consists of a single three-axis accelerometer and a single three-axis gyroscope located at the

center of his head. This is a simpler setting than the bilateral vestibular system of humans, without any considerable loss in functionality.

D. Touch Perception

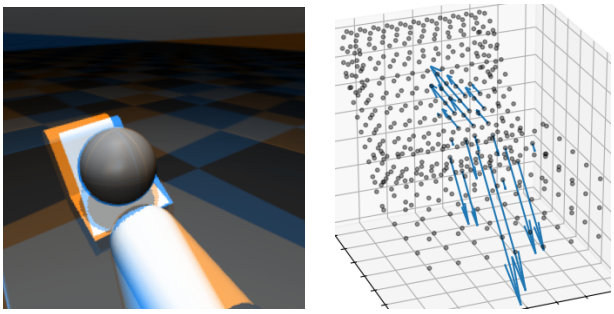
Human touch sensation is produced by a variety of receptors responding towards specific aspects of touch, such as the *Slowly Adapting type 1* or SA1 type, which responds primarily to direct pressure and coarse texture or the *Rapidly Adapting* or RA type for slip and fine texture [42]. Due to the computation overhead that comes from a full emulation of this process, we simplify our model significantly. We ignore signal travel times and condense the various types of receptors into a single generic "touch" sensor type. This type can directly provide normal and tangential friction forces at its location. The sensors are spread over the whole body with variable resolution according to the two-point discrimination distance based on data by [43]. The simple nature of the geometries of the different body parts is mirrored in the sensor distributions, the sensor density varies across body parts, but not on the same body part. Thus, for example, the front of the leg has the same sensor density as the back.

When an object touches the skin, the associated forces, provided by MuJoCo, are distributed to the nearby sensors according to a simple surface response function only dependent on the distance between the contact and the sensor. The sensor density for each body part is configurable during initialization.

Significant limits are imposed by the contact simulation of MuJoCo. All contacts are treated as point contacts between rigid bodies. For example, a box sitting on a flat surface will produce four contact points, one at each corner. Soft-body physics can only be simulated as a connected mesh of rigid particles, and is very expensive computationally, which represents a significant hurdle for training reinforcement learning systems. The surface response function can weakly simulate soft-body sensing without actually performing soft-body physics. A further limitation is that our touch sensors currently do not measure contact slip. We plan to address this in a future release.

V. Application Programming Interface

Our code is written in Python and built as an OpenAI gym [44] environment to allow easy integration into existing experimental setups and take advantage of the large amount of documentation and third-party libraries that already exist, such as the stable baselines library [45]. This environment is intended as an abstract base class that will be subclassed and adapted by other environments for specific experiments. These subclasses would handle reward structure, sensor limits or any additional constraints. Underdeveloped or limited sensors can be implemented through their configurations, but for the most part the user would implement any perceptual constraints. The environment is set up to facilitate this in a straightforward way. The configuration of the sensory modalities, such as the density of the touch sensors or the field of view and resolution of the visual system can be adjusted or disabled easily during initialization without modifying the underlying MuJoCo XMLs.



(a) Binocular vision (b) Touch perception

Fig. 2. MIMo's multimodal perception while holding a ball. (a) Anaglyph of left and right eye views. (b) The ball touches the hand in two locations, leading to activations of touch sensors (points) in the vicinity of the two contact points proportional to contact force and distance from the contact point (arrows). Note that the touch resolution was slightly lowered for readability.

The action and observation spaces are generated automatically based on the configuration of the MuJoCo XMLs and the sensor modules. Disabling touch perception also removes the associated entry from the observation space. All of the sensory modalities are programmed as separate modules and can be readily attached to any MuJoCo-based gym environment.

VI. Experiments

A. Benchmarking MIMo

In a first set of experiments we benchmark the simulation speed of MIMo. In particular, we are interested in assessing under what conditions we can achieve faster than real-time simulations.

We benchmark on a dummy environment, consisting of MIMo and two objects. MIMo takes random actions continuously. Each benchmark runs for 60 episodes, each lasting 6000 environment steps. The configuration of the MuJoCo wrappers is such that there is a certain number of physics steps for each environment step, with the duration of each physics step determined in the model XML. Increasing the number of physics steps thus reduces the number of environment steps within a given time. For this benchmark each physics step lasts 5ms with two physics steps per environment step. The total simulation time of each run is therefore 1 hour with 1-minute episodes and 100 steps per simulation second. We measure the real-time spent in each run, as well as in each of the different components of the system: MuJoCo and the sensory modalities. We test performance with multiple configurations for the different sensory modalities, focusing on the vision and touch modules since they are most sensitive to their configuration and consume the bulk of the processing time. The test system is equipped with an AMD FX-8350, 16GB RAM and a GTX 1070. The execution times are measured using Python's cProfile library. The results can be seen in Fig. 3.

The default configuration of MIMo (vision resolution of 256x256 pixels) performs significantly faster than real-time at 1.36 simulation seconds for each real second. In addition to adjusting the configuration of the modalities or increasing the duration of each physics step, the performance of the simulation could also be improved by increasing the number of physics steps for each environment step, since that reduces the amount of time spent in the sensory modalities.

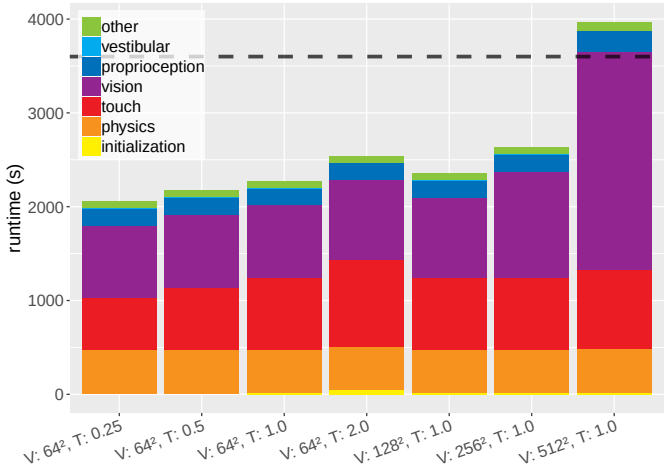


Fig. 3. Results of the performance benchmarks. Each bar represents one run consisting of 60 episodes of 1-minute length each. The labels indicate the pixel resolution for the visual system (V) and a scalar multiplier for the sensor density for the touch system (T) used. The 4 leftmost bars correspond to configurations with increasing touch sensor density and constant visual resolution, the 3 rightmost bars to increasing visual resolution and constant touch sensor density. All configurations with visual resolutions lower than 512x512 pixels run significantly faster than real-time (dashed horizontal line).

B. Illustrations of learning

In the following, we illustrate learning from multimodal input using three examples: reaching for objects, standing up, and self-body knowledge. For the sake of simplicity, MIMo is trained with extrinsic rewards using two different state-of-the-art deep reinforcement learning algorithms, Proximal Policy Optimization (PPO) [46] and Soft Actor-Critic (SAC) [19], with their default hyper-parameters from the Stable-Baselines3 library. Performance is compared for PPO and SAC with 10 different seeds (Fig. 4). We do not claim that such extrinsically motivated learning is how human infants learn these skills. We merely use these examples to showcase MIMo learning from multimodal input.

a) Reaching for Objects: Reaching is a complex behavior that emerges in the first 6 months of age. Since it requires hand-eye coordination, infants must combine vision, proprioception, and touch to produce the desired motion [47]. They initially perform a visual search for an object of their interest in order to determine the position of a desired target. A motor command is generated for the arm and hand muscles to produce the reaching movement, with immediate haptic feedback about its success.

In our illustration, MIMo learns to reach for a ball. He is standing in front of the target, which changes position randomly in each episode, always within reach of MIMo's right hand. He can only move his right shoulder, elbow, and hand joints. His head and eyes are set to look directly at the ball, i.e., the initial visual search and object fixation is assumed. The observation space only includes the proprioception sensory modality. MIMo can use the joint angles of his head and eyes to determine the position of the target. The reward function

$$r = \begin{cases} 100 & \text{if target reached,} \\ -\|\mathbf{p}_{\text{fingers}} - \mathbf{p}_{\text{target}}\| & \text{otherwise} \end{cases} \quad (1)$$

is the negative distance between the positions of the fingers and the target, with a sparse positive reward when contact is detected. Each episode lasts 1000 timesteps or until MIMo successfully touches the ball.

b) Standing up: Infants learn to stand up by themselves at around 10 months and start walking shortly after. This is a gradual process that includes previous stages such as crawling, maintaining balance, and gradually standing up with the help of adults [48]. One particular stage is marked by the emergence of pulling-to-stand, when infants who are unable to stand without support grasp the edge of a solid surface and pull themselves upwards, thus combining the strengths of their arms and legs [49]. This behavior appears as early as 7 months of age and is a necessary milestone during independent locomotion development.

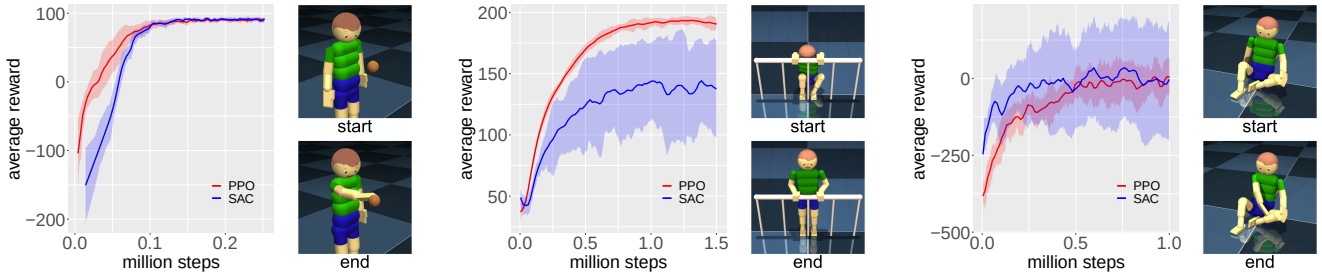
To reproduce the pulling-to-stand behavior, we design an environment where MIMo is placed sitting inside a crib. His feet are fixed to the ground and his hands are fixed to the crib's rail guard, at a height of 45 cm. He can move the joints in his arms, torso, and legs, with the aim of standing up. The observation space includes the proprioception and vestibular sensory modalities. The latter can be particularly useful by providing information about vertical acceleration. The extrinsic reward is given by

$$r = z_{\text{head}} - 0.01 \sum_{j \in \text{joints}} c_j^2 \quad (2)$$

where z_{head} is the head's height measured from an initial height of 20 cm, c_j is the control force of joint j, and the sum is taken over all active joints. This reward function favors standing positions while penalizing states that require excessive force. The parameters are set to balance the two components. All episodes last 500 timesteps.

c) Self-body Knowledge: Infants learn not only about the world but also about themselves and their own bodies. In fact, this begins as a tactile exploratory behavior before birth and continues over the first few months of life [50]. Infants develop a self-body knowledge that allows them to map the multimodal sensory inputs to the different parts of their bodies.

MIMo can learn this self-body knowledge by using his touch perception. We design an environment where MIMo is sitting with his legs crossed, such that his right arm can reach all of his body parts. In each episode he is given a target body part sampled uniformly at random from the geometric



(a) Reaching for objects (b) Standing up (c) Self-body knowledge

Fig. 4. Comparison of learning curves for PPO (red) and SAC (blue) in the three illustrative environments, with 10 different seeds each. Similar performance is achieved for both algorithms in the reaching and body knowledge environments. In the standing up environment PPO outperforms SAC, likely because the latter favors more stochastic exploration, causing MIMo to fall due to the instability of the standing straight posture. Snapshots show typical postures of MIMo at the start (top) and end (bottom) of an episode. Videos are available at https://tinyurl.com/MIMo-playlist

primitives that make up his body. By only moving his right arm, he is trained to activate the touch sensors on the target body part. The observation space includes proprioception and touch, as well as the target as a vector with one-hot encoding. The reward function

$$r = \begin{cases} 500 & \text{if target touched,} \\ -\|\mathbf{p}_{\text{touched}} - \mathbf{p}_{\text{target}}\| & \text{if other part touched,} \\ -1 & \text{otherwise} \end{cases} \quad (3)$$

is positive only if the target body part is touched. Otherwise, it is either the negative distance to the target body part if another touch signal is activated or a fixed negative value if there is no touch signal. Each episode lasts 500 time steps or until MIMo successfully touches the target.

VII. Discussion

We have presented MIMo, a multimodal infant model. MIMo is an open-source software platform for studying the principles of cognitive development in humans and AIs. A main strength of MIMo is the combination of state-of-the-art physics simulation based on MuJoCo (https://mujoco.org) with an efficient simulation of a full-body touch-sensitive skin. We believe that these ingredients are essential for advancing computational models of, e.g., the development of an infant's self-model or their object manipulation skills.

Overall, the design of MIMo reflects a number of trade-offs between realism and computational efficiency. While these permit faster than real-time simulations on standard hardware, they have also resulted in a number of limitations. First, while resembling the dimensions of an 18-month-old infant, the body of MIMo had to be simplified to shape primitives undergoing rigid body dynamics. Future versions could consider more realistic body shapes and soft body dynamics. This is particularly relevant for MIMo's hands, which currently have a simplistic morphology. Future work will add more detailed multi-fingered hands. Second, actuation of joints is so far limited to either position or torque control. Future versions could consider joint actuation by more human-like muscle tendon systems. Third, MIMo currently features simple implementations of only four sensory modalities (binocular vision, proprioception, touch, and a vestibular system). More sophisticated versions of, e.g., MIMo's touch perception could be developed. Furthermore, future work could incorporate nociception (pain perception), audition, or olfaction.

Despite these limitations, we hope that MIMo will facilitate a transition towards models of development that can learn large numbers of skills in an open-ended fashion based on rich multimodal input. At the very least, it should make such efforts easier and more reproducible. Furthermore, it will enable a cumulative approach to such research, where models of the development of higher-level cognitive functions are built on top of previously published models of the development of their precursor skills. After all, human development is often a cumulative process, where new representations, skills, and competences are built on top of existing ones.

Acknowledgment

We thank Layla Selestrini for her contribution to the design of MIMo's facial expressions.

References

[1] A. M. Turing, "Computing machinery and intelligence," *Mind*, vol. 59, no. 236, pp. 433–460, 1950.

[2] M. Asada, K. F. MacDorman, H. Ishiguro, and Y. Kuniyoshi, "Cognitive developmental robotics as a new paradigm for the design of humanoid robots," *Robotics and Autonomous Systems*, vol. 37, no. 2, pp. 185–193, 2001, humanoid Robots.

[3] M. Lungarella, G. Metta, R. Pfeifer, and G. Sandini, "Developmental robotics: a survey," *Connection Science*, vol. 15, no. 4, pp. 151–190, 2003.

[4] J. Schmidhuber, "Developmental robotics, optimal artificial curiosity, creativity, music, and the fine arts," *Connection Science*, vol. 18, no. 2, pp. 173–187, 2006.

[5] M. Asada, K. Hosoda, Y. Kuniyoshi, H. Ishiguro, T. Inui, Y. Yoshikawa, M. Ogino *et al.*, "Cognitive developmental robotics: A survey," *IEEE Transactions on Autonomous Mental Development*, vol. 1, no. 1, pp. 12–34, 2009.

[6] A. Cangelosi and M. Schlesinger, "From babies to robots: The contribution of developmental robotics to developmental psychology," *Child Development Perspectives*, vol. 12, no. 3, pp. 183–188, 2018.

[7] K. Doya and T. Taniguchi, "Toward evolutionary and developmental intelligence," *Current Opinion in Behavioral Sciences*, vol. 29, pp. 91–96, 2019, artificial Intelligence.

[8] C. Gan, J. Schwartz, S. Alter, D. Mrowca, M. Schrimpf, J. Traer, J. D. Freitas *et al.*, "ThreeDWorld: A platform for interactive multi-modal physical simulation," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. [Online]. Available: https://openreview.net/forum?id=db1InWAwW2T

[9] E. Rohmer, S. P. N. Singh, and M. Freese, "V-rep: A versatile and scalable robot simulation framework," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2013, pp. 1321–1326.

[10] E. Coumans and Y. Bai, "Pybullet, a python module for physics simulation for games, robotics and machine learning," 2016.

[11] J. Collins, S. Chand, A. Vanderkop, and D. Howard, "A review of physics simulators for robotic applications," *IEEE Access*, vol. 9, pp. 51 416–51 431, 2021.

[12] E. Todorov, T. Erez, and Y. Tassa, "Mujoco: A physics engine for model-based control," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012, pp. 5026–5033.

[13] V. Tikhanoff, A. Cangelosi, P. Fitzpatrick, G. Metta, L. Natale, and F. Nori, "An open-source simulator for cognitive robotics research: The prototype of the iCub humanoid robot simulator," in *Proceedings of the 8th Workshop on Performance Metrics for Intelligent Systems*, ser. PerMIS '08. New York, NY, USA: Association for Computing Machinery, 2008, p. 57–61.

[14] M. Kerzel, E. Strahl, S. Magg, N. Navarro-Guerrero, S. Heinrich, and S. Wermter, "Nico — neuro-inspired companion: A developmental humanoid robot platform for multimodal interaction," in *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2017, pp. 113–120.

[15] Y. Kuniyoshi and S. Sangawa, "Early motor development from partially ordered neural-body dynamics: experiments with a cortico-spinal-musculo-skeletal model," *Biological Cybernetics*, vol. 95, no. 6, pp. 589–605, 2006.

[16] E. Oztop, N. S. Bradley, and M. A. Arbib, "Infant grasp learning: a computational model," *Experimental brain research*, vol. 158, no. 4, pp. 480–503, 2004.

[17] A. Priamikov, M. Fronius, B. Shi, and J. Triesch, "Openeyesim: a biomechanical model for simulation of closed-loop visual perception," *Journal of Vision*, vol. 16, no. 15, pp. 25–25, 12 2016.

[18] S. L. Delp, F. C. Anderson, A. S. Arnold, P. Loan, A. Habib, C. T. John, E. Guendelman *et al.*, "Opensim: Open-source software to create and analyze dynamic simulations of movement," *IEEE Transactions on Biomedical Engineering*, vol. 54, no. 11, pp. 1940–1950, 2007.

[19] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. PMLR, 10–15 Jul 2018, pp. 1861–1870.

[20] Y. Tassa, T. Erez, and E. Todorov, "Synthesis and stabilization of complex behaviors through online trajectory optimization," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012, pp. 4906–4913.

[21] S. Ressler, "Anthrokids-anthropometric data of children," *National Institute of Standards and Technology*, 1977. [Online]. Available: https://math.nist.gov/~SRessler/anthrokids/

[22] M. J. McKay, J. N. Baldwin, P. Ferreira, M. Simic, N. Vanicek, J. Burns, . N. P. Consortium *et al.*, "Normative reference values for strength and flexibility of 1,000 children and adults," *Neurology*, vol. 88, no. 1, pp. 36–43, 2017.

[23] M. N. Eek, A.-K. Kroksmark, and E. Beckung, "Isometric muscle torque in children 5 to 15 years of age: Normative data," *Archives of Physical Medicine and Rehabilitation*, vol. 87, no. 8, pp. 1091–1099, 2006.

[24] W. N. Sankar, C. T. Laird, and K. D. Baldwin, "Hip range of motion in children: what is the norm?" *Journal of Pediatric Orthopaedics*, vol. 32, no. 4, pp. 399–405, 2012.

[25] T. Gomez, G. Beach, C. Cooke, W. Hrudey, and P. Goyert, "Normative database for trunk range of motion, strength, velocity, and endurance with the isostation b-200 lumbar dynamometer," *Spine*, vol. 16, no. 1, p. 15—21, January 1991.

[26] A. Jordan, J. Mehlsen, P. M. Bülow, K. Østergaard, and B. Danneskiold-Samsøe, "Maximal isometric strength of the cervical musculature in 100 healthy volunteers," *Spine*, vol. 24, no. 13, p. 1343, 1999.

[27] A. M. Öhman and E. R. Beckung, "Reference values for range of motion and muscle function of the neck in infants," *Pediatric Physical Therapy*, vol. 20, no. 1, pp. 53–58, 2008.

[28] H. Watanabe, K. Ogata, T. Amano, and T. Okabe, "The range of joint motions of the extremities in healthy japanese people–the difference according to the age (author's transl)," *Nihon Seikeigeka Gakkai Zasshi*, vol. 53, no. 3, pp. 275–261, 1979.

[29] I. Günal, N. Köse, O. Erdogan, E. Göktürk, and S. Seber, "Normal range of motion of the joints of the upper extremity in male subjects, with special reference to side," *Journal of Bone & Joint Surgery*, vol. 78, no. 9, p. 1401, 1996.

[30] R. E. Hughes, M. E. Johnson, S. W. O'Driscoll, and K.-N. An, "Age-related changes in normal isometric shoulder strength," *The American Journal of Sports Medicine*, vol. 27, no. 5, pp. 651–657, 1999.

[31] M. Katoh, "Test-retest reliability of isometric shoulder muscle strength measurement with a handheld dynamometer and belt," *Journal of Physical Therapy Science*, vol. 27, no. 6, pp. 1719–1722, 2015.

[32] S. N. Da Paz, A. Stalder, S. Berger, and K. Ziebarth, "Range of motion of the upper extremity in a healthy pediatric population: introduction to normative data," *European Journal of Pediatric Surgery*, vol. 26, no. 05, pp. 454–461, 2016.

[33] S.-K. Bok, T. H. Lee, and S. S. Lee, "The effects of changes of ankle strength and range of motion according to aging on balance," *Annals of Rehabilitation Medicine*, vol. 37, no. 1, pp. 10–16, 2013.

[34] P. Ekman, "An argument for basic emotions," *Cognition and Emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[35] J. M. Vanswearingen, "Measuring wrist muscle strength," *Journal of Orthopaedic & Sports Physical Therapy*, vol. 4, no. 4, pp. 217–228, 1983.

[36] S. P. Johnson, "Chapter 14 - development of the visual system," in *Neural Circuit Development and Function in the Brain*, J. L. Rubenstein and P. Rakic, Eds. Oxford: Academic Press, 2013, pp. 249–269.

[37] W. J. Lee, J. H. Kim, Y. U. Shin, S. Hwang, and H. W. Lim, "Differences in eye movement range based on age and gaze direction," *Eye*, vol. 33, no. 7, pp. 1145–1151, 2019.

[38] A. L. Rosenbaum and A. P. Santiago, *Clinical strabismus management: principles and surgical techniques*. W.B. Saunders, 1999.

[39] H. Strasburger, I. Rentschler, and M. Jüttner, "Peripheral vision and pattern recognition: A review," *Journal of Vision*, vol. 11, no. 5, pp. 13–13, 12 2011.

[40] J. C. Tuthill and E. Azim, "Proprioception," *Current Biology*, vol. 28, no. 5, pp. R194–R203, 2018.

[41] S. Wiener-Vacher, D. Hamilton, and S. Wiener, "Vestibular activity and cognitive development in children: perspectives," *Frontiers in Integrative Neuroscience*, vol. 7, 2013.

[42] K. O. Johnson and S. S. Hsiao, "Neural mechanisms of tactual form and texture perception," *Annual Review of Neuroscience*, vol. 15, no. 1, pp. 227–250, 1992.

[43] F. Mancini, A. Bauleo, J. Cole, F. Lui, C. A. Porro, P. Haggard, and G. D. Iannetti, "Whole-body mapping of spatial acuity for pain and touch," *Annals of Neurology*, vol. 75, no. 6, pp. 917–924, 2014.

[44] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, "OpenAI Gym," *arXiv preprint arXiv:1606.01540*, 2016.

[45] A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus, and N. Dormann, "Stable-baselines3: Reliable reinforcement learning implementations," *Journal of Machine Learning Research*, vol. 22, no. 268, pp. 1–8, 2021.

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

[47] D. Corbetta, R. F. Wiener, S. L. Thurman, and E. G. McMahon, "The embodied origins of infant reaching: Implications for the emergence of eye-hand coordination," *Kinesiology Review*, vol. 7, no. 1, pp. 10–17, 2018.

[48] H. Sigmundsson, H. W. Lorås, and M. Haga, "Exploring task-specific independent standing in 3- to 5-month-old infants," *Frontiers in Psychology*, vol. 8, 2017.

[49] O. Atun-Einy, S. E. Berger, and A. Scher, "Pulling to stand: Common trajectories and individual differences in development," *Developmental Psychobiology*, vol. 54, no. 2, pp. 187–198, 2012.

[50] L. Jacquey, J. Fagard, K. O'Regan, and R. Esseily, "Development of body know-how during the baby's first year of life," *Enfance*, vol. 2, no. 2, pp. 175–192, 2020.