VECA: A New Benchmark and Toolkit for General Cognitive Development

Kwanyoung Park*, **Hyunseok Oh***, **Youngki Lee**

Department of Computer Science and Engineering, Seoul National University, South Korea
william202@snu.ac.kr, ohsai@snu.ac.kr, youngkilee@snu.ac.kr

Abstract

The developmental approach, simulating a cognitive development of a human, arises as a way to nurture a human-level commonsense and overcome the limitations of data-driven approaches. However, neither a virtual environment nor an evaluation platform exists for the overall development of core cognitive skills. We present the **VECA**(**V**irtual **E**nvironment for **C**ognitive **A**ssessment), which consists of two main components: (i) a first benchmark to assess the overall cognitive development of an AI agent, and (ii) a novel toolkit to generate diverse and distinct cognitive tasks. VECA benchmark virtually implements the cognitive scale of *Bayley Scales of Infant and Toddler Development-IV*(Bayley-4), the gold-standard developmental assessment for human infants and toddlers. Our VECA toolkit provides a human toddler-like embodied agent with various human-like perceptual features crucial to human cognitive development, e.g., binocular vision, 3D-spatial audio, and tactile receptors. We compare several modern RL algorithms on our VECA benchmark and seek their limitations in modeling human-like cognitive development. We further analyze the validity of the VECA benchmark, as well as the effect of human-like sensory characteristics on cognitive skills.

Introduction

Building a cognitive intelligent agent with human-like commonsense is a milestone of artificial intelligence (Zhu et al. 2020). Human cognition is an interpretable and sample-efficient general intelligence, encompassing diverse abilities like information processing, intuitive psychology, and goal setting (Lake et al. 2017; Sloman 1999). These core cognitive skills naturally construct in an early stage of human development, often with limited experiences. Hence one of the emerging paths towards such mental capabilities is to mimic such cognitive development of a human, i.e., a developmental approach. (Doya and Taniguchi 2019; Silver et al. 2021) The goal of the developmental approach is to simulate a human's neuro-cognitive developmental process and enable continual life-long learning with active interactions.

However, the developmental approach currently lacks both the assessment for cognitive development and the general environment to simulate the development. Several benchmarks only target specific cognitive skills of an AI agent, such as intuitive physics (Bakhtin et al. 2019) or intuitive psychology (Shu et al. 2021). None of them covers the overall development of distinct cognitive skills like object-relatedness, memory, and sensorimotor development. Embodied agent simulators are the closest thing to mimic human biological features, since they give realistic egocen-

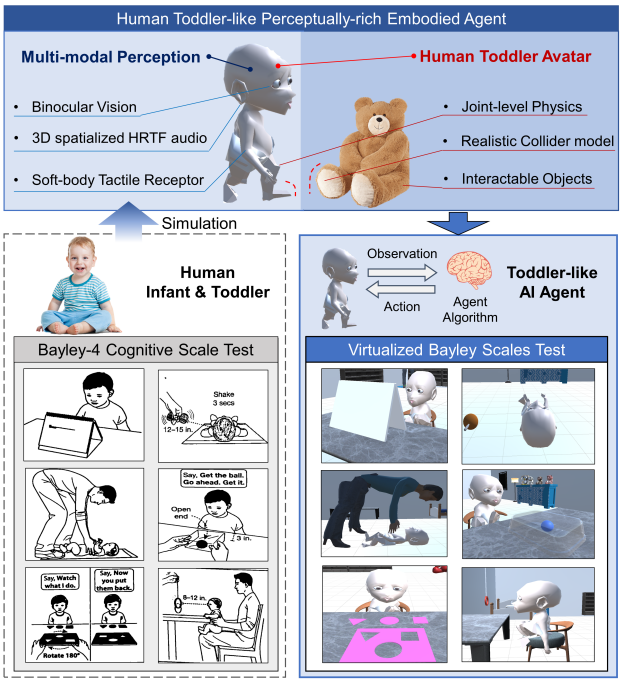


Figure 1: We present the VECA benchmark, a novel benchmark to assess the overall cognitive development of an AI agent. Our VECA benchmark virtually implements the *Bayley Scales of Infant and Toddler Development-IV*, a gold-standard developmental test for early human infancy. VECA benchmark is built upon our VECA toolkit, that can easily develop diverse domains of cognitive tasks with controllable difficulties. To bridge the perception and interaction gap of human toddler and AI agent, VECA toolkit provides a human toddler-like multisensory embodied agent.

*These authors contributed equally.

tric perceptions or active interactions (Xia et al. 2018; Kolve et al. 2017; Wu et al. 2018; Chen et al. 2020). Unfortunately, they neither evaluate the general cognitive development nor they focus on human-like sensory characteristics crucial to cognitive development.

In this light, we propose the VECA (stands for **V**irtual **E**nvironment for **C**ognitive **A**ssessment), which consists of two major components: (i) VECA benchmark, the first benchmark to assess the general cognitive development of virtual AI agents, and (ii) VECA toolkit, a novel task-generating toolkit that can easily create diverse tasks evaluating distinct cognitive capabilities, e.g., object understanding, multimodal learning, or sensorimotor development. For our VECA benchmark, we virtually implement the cognitive scale of *Bayley Scales of Infant and Toddler Development-IV*(Bayley-4), the gold-standard developmental assessment for a human (Bayley 1999). Our VECA benchmark consists of 81 tasks evaluating various cognitive functions, as demonstrated in Figure 2. Under a standardized scenario, the tester observes behavioral responses of an agent to determine the mastery, emergence, or inability of a certain cognitive skill. Since Bayley-4 is initially designed for a human infant and toddler, a virtual agent of VECA should embody human-like multimodal perceptions and interaction capabilities.

VECA toolkit provides a series of essential features to simulate human cognitive development and virtualize the Bayley-4 test: 1) human-like multimodal perceptions, 2) comprehensive interaction capability with the environment, 3) extensibility for developing various custom-built tasks. to model human sensory characteristics influential to cognitive development, e.g., binocular vision, spatialized audio with HRTF, and human-like tactile receptors. These features are critical in human's learning; humans learn by collecting **rich multimodal perception** (e.g., vision, audio, tactile) from their surroundings (Landau, Smith, and Jones 1998; Tacca 2011) and **actively interacting** (Franchak, van der Zalm, and Adolph 2010; Vogt et al. 2018) with objects. We plan to open-source both our VECA benchmark and toolkit on a public repository.

We assess several representative RL algorithms with our VECA benchmark, including policy gradient methods (Espeholt et al. 2018; Schulman et al. 2017; Haarnoja et al. 2018) and curiosity-driven learning (Burda et al. 2019), and find that there is still a long way to go to reach the human-level cognitive capabilities. Experimental results show that goal-driven learning (IMPALA, the policy gradient method) initially outperforms the unsupervised exploration without explicit reward (curiosity-driven learning), but it prematurely converges and marginally improves from a random policy. Furthermore, we demonstrate the validity of our VECA benchmark by measuring the solvability and complexity of its tasks. Our results show that all the tasks are solvable, and their difficulties are well-distributed. We also observe that the mastery of cognitive skill is much more difficult to acquire than the emergence of the skill. Ablation study reveals that our VECA's human-like sensory features meaningfully affect the development of cognitive skills.

In summary, our key contributions are as follows:

- We develop the VECA benchmark, a first benchmark to assess the overall development of core cognitive skills of an AI agent. Our benchmark virtually implements the Bayley-4, a standard developmental delay assessment tool of a human.
- We introduce a novel VECA toolkit that can easily generate various tasks measuring diverse cognitive skills. Our VECA toolkit supports human toddler-like agents with rich human-like perception and interaction capabilities.
- Our work is first to provide diverse human biomemetic sensory features on multimodal sensation, e.g., binocular vision, HRTF-based spatialized audio, and human-like tactile receptor.
- Using our VECA benchmark, We study the limitation of several modern RL algorithms in simulating human-like cognitive development. Moreover, we analyze the validity of VECA benchmark and VECA toolkit's human-like sensory characteristics.

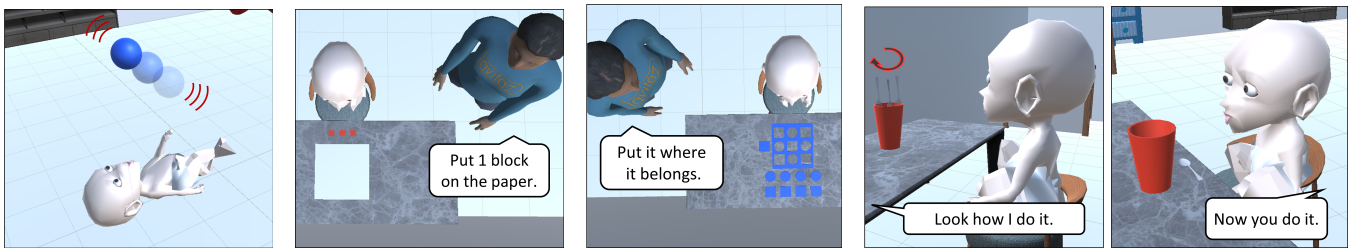
Background & Related Works

In terms of human-like embodiment or cognitive development assessment, our work is related to prior works on (1) cognitive tests for AI agents, (2) cognitive developmental robotics, and (3) embodied agent simulator.

Cognitive Test for AI Agents To verify the human-like cognitive capability of AI, prior works introduced cognitive tests for AI agents. For instance, *TOMNet* (Rabinowitz et al. 2018) applies the Sally-Anne test to an AI agent, a psychological test measuring a person's socio-cognitive ability of false belief (Wimmer and Perner 1983). Thorough benchmarks exist for specific cognitive skills, e.g., physical reasoning and intuitive physics (Bakhtin et al. 2019), productive and systematic generalization under uncertainty (Vedantam et al. 2021), and intuitive psychology (Shu et al. 2021). In contrast, our VECA benchmark assesses diverse core cognitive functions which emerge in the early stage of human development. To narrow the gap between the human test and its virtual counterpart, we use a toddler avatar embodied with a number of human-like features, unlike puzzle-solving (Bakhtin et al. 2019) or dataset-based (Shu et al. 2021) benchmarks.

Cognitive Developmental Robotics Cognitive Developmental Robotics *physically embodies* an agent to a human baby-like robot to study how human's higher cognitive functions emerge through real-world interaction (Asada et al. 2009). Humanoid robot platforms in these works faithfully mimic the human toddler's body (Metta et al. 2010), e.g., joint flexibility, soft skin, and human-like appearance. However, robot platforms are cost-inefficient to train and test AI algorithms; It is more cost-effective, scalable, and safe using virtual environments and agents (Zhao, Queralta, and Westerlund 2020). Furthermore, it is challenging to enable the standard and repeatable testing procedures of Bayley-4 assessment with existing robotics platforms.

Realistic Simulators for Embodied AI. Embodied AI researchers hypothesize that intelligence emerges from interacting with its surroundings, just like humans (Smith and



(a) # 2 Looks at Object.

(b) # 61 Understands Concept of One.

(c) # 47 Blue Board 3 Pieces.

(d) # 29 Stirs Spoon.

Figure 2: Four example VECA benchmark tasks. VECA encompasses core cognitive functions which develop in early infancy: For instance, (2a) visual attention and object recognition (Reynolds 2015), (2b) context understanding, cardinality, and counting (Sarnecka and Wright 2013), (2c) problem-solving and concept of shape (Clements et al. 1999), (2d) memory (Rovee-Collier and Hayne 1987) and cognitive imitation (Subiaul et al. 2004).

Gasser 2005). To verify it and faithfully simulate the real-world entity, virtual agents embody egocentric sensory inputs as well as interaction capability with the environment. Prior environments for embodied AI agents focus on photo-realistic indoor simulation (Gibson 1988), vision-language tasks (Wu et al. 2018), robot simulation with realistic robot sensors and dynamics (Koenig and Howard 2004), audio-visual multimodal learning (Chen et al. 2020), and predefined object-specific interaction (Kolve et al. 2017). By contrast, our VECA provides unique features to model cognitive development, which none of the existing environments offers. In particular, VECA incorporates a suite of human-like features such as tactile sensing, soft skin, HRTF spatialized audio, and baby-like morphology, which are all essential to implementing the *Cognitive Scale* of Bayley-4.

VECA Environment & Toolkit

Bayley-4 is a valid and reliable assessment tool for human development (Bayley 1999), but our focus is an AI agent, not a human. The question naturally arises: "Is the Bayley-4 test also valid for virtual AI agents?". Bayley-4 consists of tasks that are most meaningful with a human-like embodiment; For instance, task (2a) needs binocular human eyes and a toddler's posture. General cognitive development is closely related to biological factors (Ranjitkar et al. 2019), which virtual AI agents do not acquire or develop. To substantiate the virtualization of Bayley-4, the test subject agent should embody human toddler-like biological features in both sensory input and action capability.

VECA toolkit introduces a *human-like multisensory embodied agent* and an *immersive virtual environment*, as shown in the overview Figure 1. VECA agent receives highly multimodal sensations simulating human characteristics and does joint-level or animation-based actions. The immersive VECA environment enables physical and animated interactions with surrounding objects.

Human-like Multisensory Embodied Agent

A VECA agent embodies three major components: rich human-like multimodal(multisensory) perception, human toddler avatar, and joint-level physics. Inspired by researches in human multisensory learning (Stein 2012; Moustafa 1999), we provide four important sensory modalities: vision, audio, tactile, and proprioception. Studies show that multi-modal sensory experiences, mainly vision, audio, tactile, and proprioception, facilitates early development (Murphy 1997) and learning (Chandrasekaran 2017). We further augment them with human biological characteristics crucial for cognitive development, which is detailed in Figure 3.

Human-like Vision. In imitation of human binocular vision, the VECA agent receives binocular vision input through two eye pupils of the toddler avatar. Binocular vision plays a key role in a part of core cognitive abilities and their development, e.g., depth and space perception (van Hof, van der Kamp, and Savelsbergh 2006). Another important trait of an infant's vision system is biological development, in which visual acuity (Dobson and Teller 1978) and color sensitivity (Adams, Maurer, and Cashin 1990) steadily grow in the first few months (Valenti 2006). We model such deficient color vision and sharpness with multiple visual filters varying, e.g., focal length, grayscale, and blur. We allow the parametrized manipulation of these features to simulate a particular developmental stage of the vision system.

Human-like Audio. We introduce HRTF (head-related transfer function) spatialization filter to facilitate blind-spot recognition and audio source localization of a VECA agent. Diffraction and reflection properties of human body structures like the head or torso greatly affect auditory processing (Bögelein et al. 2018) and early auditory development (Tollin 2009). The phase and impulse difference of audio signal between two ears makes it possible (Potisk 2015). HRTF models such physical interaction between human anatomy and the sound source with a transfer function of azimuth, elevation, and frequency. HRTF datasets collect audio data from human subject's ears varying source locations to model the input-output transfer curve. We use the KEMAR dataset (Gardner 1994), which uses a dummy head instead of a real person. Since the KEMAR dataset only supports discrete spherical coordinates, we apply bilinear interpolation as in (Sousa and Queiroz 2010) to enable arbitrary

coordinates. Post-processed audio data $y_{i=L,R}$ our VECA agent perceives on its left(L) and right(R) ear is,

$$y_i(t, d_i, \theta_i, \varphi_i) = \underbrace{\min\{1/d_i^2, d_{TH}\}}_{\text{Inverse Square Law}} \underbrace{F(\hat{\mathbf{H}}(f(x(t)), \theta_i, \varphi_i))}_{\text{HRTF-based Magnitude Scaling}}$$

$$\hat{\mathbf{H}}(x, \theta, \varphi) = \begin{pmatrix} \lceil\theta\rceil - \theta \\ \theta - \lfloor\theta\rfloor \end{pmatrix}^T \mathbf{H}(x, \theta, \varphi) \begin{pmatrix} \lceil\varphi\rceil - \varphi \\ \varphi - \lfloor\varphi\rfloor \end{pmatrix}$$

$$\mathbf{H}(x, \theta, \varphi) = \begin{pmatrix} H(x, \lceil\theta\rceil, \lceil\varphi\rceil) & H(x, \lceil\theta\rceil, \lfloor\varphi\rfloor) \\ H(x, \lfloor\theta\rfloor, \lceil\varphi\rceil) & H(x, \lfloor\theta\rfloor, \lfloor\varphi\rfloor) \end{pmatrix}$$

where $x(t)$ the source audio data, $i = \{L, R\}$ indicating the left(L) or right(R) ear, d_i the distance of the sound source, θ_i [°] the azimuth, φ_i [°] the polar angle, d_{TH} a minimum audible distance, H the KEMAR HRTF function, f the discrete fourier transform, and F the inverse discrete fourier transform. The function $\hat{\mathbf{H}}$ is thus the composition of bilinear interpolation and the discrete HRTF function H.

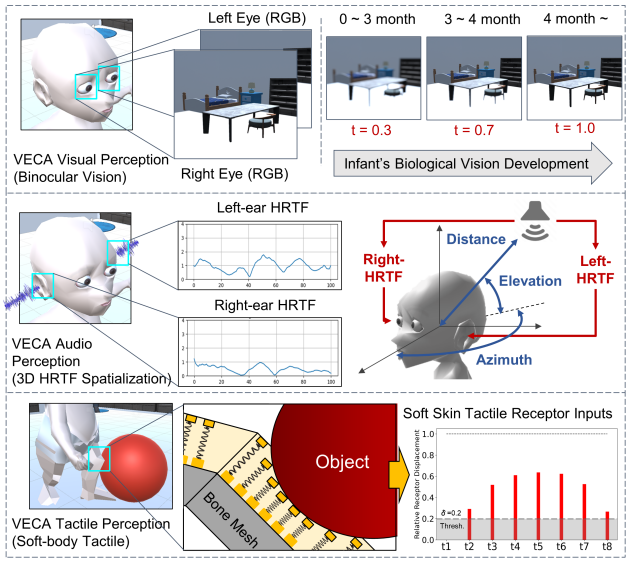


Figure 3: Human-like multimodal perceptions of VECA toolkit's embodied agent. Unlike existing embodied AI environments, we provide the human toddler agents with various human-like features, e.g., binocular vision, soft-skin based tactile sensation, and HRTF-based spatialized audio.

Human-like Tactile. We simulate the human tactile sensation by modeling a soft and flexible skin covering a toddler avatar's rigid-body bones. Tactile perception plays a significant role in early cognitive development; toddlers unconsciously learn cognitive skills through tactile interaction like mouthing or grabbing. (Gibson 1988; Piaget and Cook 1952). Tactile perception is critical in sensorimotor development (Dusing 2016), which consists a large portion of the Bayley-4 test. Prior work naively simulates such tactile sensation by measuring collision force on a single contact point (Juliani et al. 2018). In contrast, we mimic four essential features of a biological tactile receptor: soft skin elasticity (Wang et al. 2021), multiple contact points, sensory threshold (Lawless and Heymann 1999), and sensory habituation (Song, Banks, and Bewick 2015). First, a tactile sensor converts the elastic deformation of the local skin area above it into a sensor signal. Inspired from Hooke's law for elastic body and prior works (El Bab et al. 2008; Ren et al. 2018), the initial sensor value is proportional to the displacement of the skin area. Second, we place multiple tactile sensors on each triangle face of the agent's mesh, which can simultaneously activate. Third, we cut off the sensory value with an absolute threshold. Finally, we model the sensory habituation with exponential decay, following the study of (Thompson and Spencer 1966). The tactile sensory input $T_i(s, t)$ of i-th sensor is thus formulated as follows:

$$T_i(s, t) = \sigma_\delta(\underbrace{\min\{1, s/s_{max}\}}_{\text{Soft-skin displacement}} \underbrace{e^{-\lambda t}}_{\text{Habituation}})$$

$$\text{where} \quad \sigma_\delta(x) = \begin{cases} 0, & \text{if } x < \delta \\ x, & \text{if } x \geq \delta \end{cases} \quad \text{(Cutoff function)}$$

s is the displacement of skin area around i-th sensor, t the time (number of frames) passed from an initial stimulus, s_{max} the maximum displacement of skin area, λ the decay rate, and δ the absolute threshold of sensory value.

Proprioception. To supply the kinesthetic senses like limb position and movement, our VECA provides the raw vector quantity of bones and joints, e.g., bone orientation, current angle, and angular velocity of the joints. These proprioceptive senses play a critical role in human motor and sensorimotor development. However, their biological receptors are noisy and difficult to simulate; these proprioreceptors rely heavily on the human musculoskeletal anatomy since they are pressure sensors within muscles and joints.

Human Toddler Avatar. We model a humanoid agent with a human toddler-like appearance and joint-level motion capability, as studies show that baby-like morphology affects the human cognitive development (Dusing 2016) and AI agent's learning (Bambach et al. 2018). The physical avatar has 47 bones with 82 degrees of freedom and human joint-like angular constraints. Skin mesh overlays the bones, and the mesh adaptively changes with the bone orientation to model the soft skin. Accurate mesh collider and bone modeling enable the agent to interact physically with the objects.

We also support an animation-based avatar and interaction to trade-off task complexity with the physical plausibility. For complex tasks, controlling an agent entirely with the joint-level actions and physical interactions may be difficult. Stable primitive toddler-like actions (e.g., walk, rotate, crawl) and object interactions (e.g., grab, open, step on) are animated, similar to (Kolve et al. 2017).

Implementation Detail We use *Unity3D* game engine as an environment simulator to support a 3D realistic scene rendering and physics engine. VECA environment is a 3D Euclidean space with Newtonian physics and a downward gravitational force. VECA provides a series of features to facilitate AI algorithm development: First, VECA supports parallel execution of environment along with batched sampling from a multi-task or multi-agent environment. Second, to enable training on a remote server with abundant

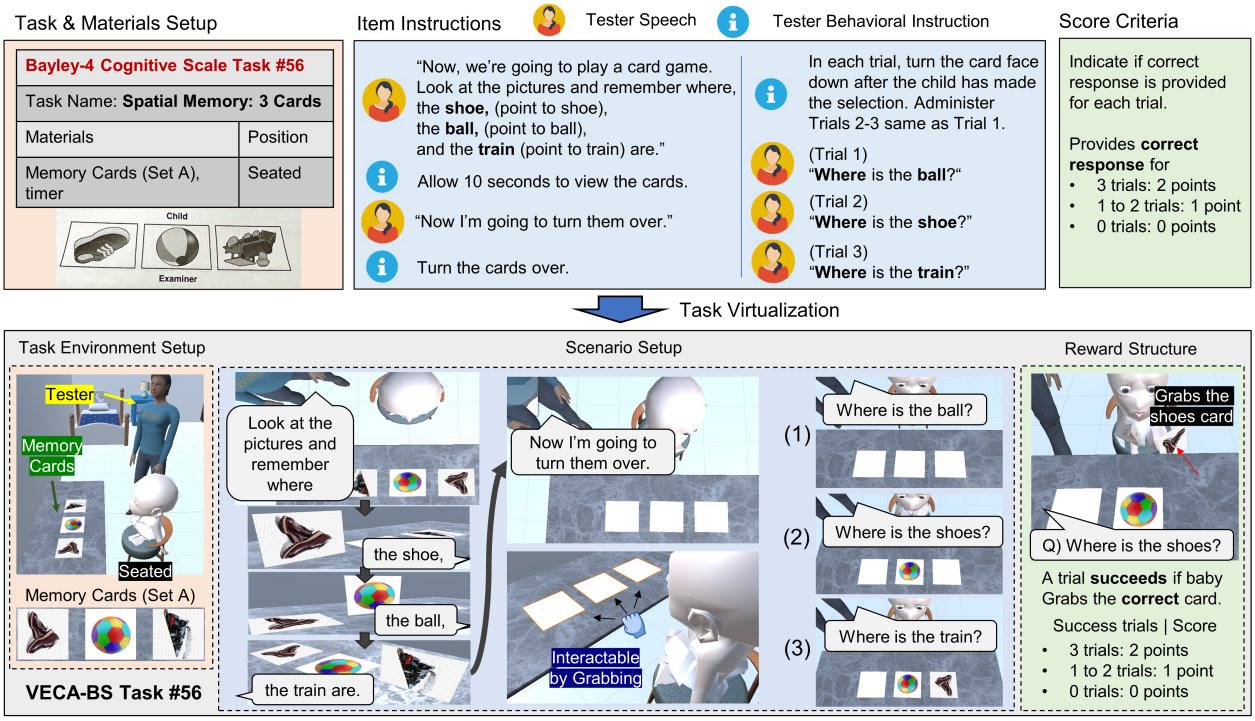


Figure 4: Example case of virtualizing a Bayley-4 cognitive scale task into the VECA environment. This case virtually implements cognitive scale task #56, which assesses the spatial memory capacity.

resources, VECA can communicate information through the socket network interface. Third, VECA includes an easy-to-use python API resembling the OpenAI Gym interface that can easily integrate with existing AI algorithms.

VECA task-generation toolkit is an extensive set of APIs for creating cognitive tasks in a VECA environment. Using the toolkit, we additionally generated a number of embodied AI tasks for various cognitive skills: joint-level control, understanding the context of objects, multimodal learning, and multi-agent RL. Further details of the VECA environment, toolkit, usage of the toolkit, and generated task set are thoroughly described in the appendix (Park, Oh, and Lee 2021).

Virtualized Bayley-Scale Assesment

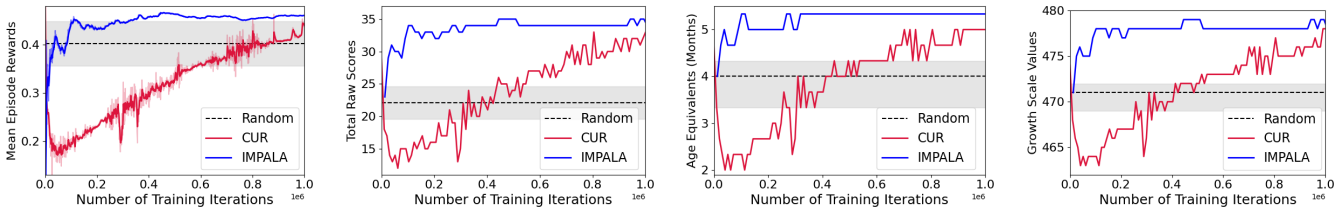
Bayley Scales Assessment The Bayley Scales of Infant and Toddler Development (Bayley Scales Test) (Bayley 1999) is a gold standard (Carey et al. 2009) of development assessment tool for a child aged 1 to 42 months. Its main goal is to monitor the child's developmental progress longitudinally or to identify a human child with developmental delay. A structured series of developmental tasks constitutes the Bayley Scales Test. The score is given per task when the test subject makes a correct behavioral response in the task scenario (Albers and Grieve 2007). The test's standardized administration and scoring procedures should not be violated to precisely compare the child's performance. To claim the validity and clinical utility of the test and its metrics, large-scale standardization research is conducted on 1,700 typically developing children.

The latest edition of Bayley Scales Test (Bayley-4) (Aylward 2017) uses the five-scale framework: Cognitive, Language, Motor, Social-Emotional, and Adaptive Behavior. We use *Cognitive* scale in this work, since it focuses on the development of cognitive processing aspects.

Cognitive Scale The Cognitive Scale of Bayley-4 consists of 81 task items that measure diverse cognitive processing aspects in early development, e.g., exploration and manipulation, object relatedness, concept formation, memory, sensorimotor development (Bayley 1999). We list several notable cognitive processes the scale examines.

- Information-processing tasks including novelty preference, habituation, and anticipation of patterns, which correlate with later cognitive functioning (Rose et al. 2012).
- Problem-solving, a higher-order information processing that involves thinking or reasoning, memory, and synthesis of information (Greiff et al. 2015).
- Play activities facilitating cognitive growth and symbol understanding (Frost, Wortham, and Reifel 2000).
- One-to-one correspondence, counting, and cardinality (Geary et al. 2018).

Language development, one of the fundamental cognitive abilities, is mainly measured on a separate *Language Scale*. Note that the goal of the scale is not to quantify the *entire* cognition or intelligence of a human subject; rather, it checks whether core cognitive skills expected in a certain age have actually emerged.



(a) Mean Episode Rewards (b) Total Raw Scores. (c) Age Equivalents. (d) Growth Scale Values.

Figure 5: VECA benchmark baselines results on a training steps of policy learners (IMPALA, CUR) (x-axis). *Bayley-4* metrics (5b), (5c), (5d) of policy learners are measured per 2×10^4 steps. We report human baseline results here for clarity: total raw scores 130, age equivalent 33 months, GSV 529, and mean reward 1.7105.

Virtualizing the Bayley-4 Cognitive Scale

For a benchmark of general cognitive development, we virtually implement the cognitive scale of Bayley-4 with our VECA toolkit. We are permitted to adapt the Bayley-4 in a virtual environment for research purposes. Figure 4 describes how we port a Bayley-4 task to the VECA environment. A Bayley-4 task contains three main components: task & materials setup, item instructions, and scoring. We first model the task environment and materials in a VECA environment; for example, we arrange prop materials or prepare a caregiver's audio clips. Next, we develop the item instructions as a scenario and produce it with materials and a caregiver avatar. Finally, we design a reward structure that returns a score depending on the agent's behavior. Score 2 means a consistent proficiency of skill, whereas score 1 implies that skill is inconsistent but emerging. Score 0 represents the absence of skill. Note that the *correctness* of behavioral response is algorithmically determined, unlike the real-world Bayley-4 test in which a human tester subjectively determines its correctness. For instance, the task *Looks at Object* in Figure 2a defines the "looking" as the cosine similarity of head and object direction > 0.95.

Metrics Bayley-4 provides several standardized metrics for its sub-scales (Aylward 2017) that our VECA benchmark can leverage. The total raw score is simply a sum of each task score, which converts to three standardized metrics: scaled scores, age equivalent, and growth scale values. Scaled Scores uses the biological age of the participant to normalize the raw score. Age Equivalent shows the developmental age of a normative human child equivalent to the raw score. Growth Scale Values (GSV) are used to track the child's growth over time, and it has a mean and std value of 500 and 25. A score conversion table exists that maps the total raw score to other metrics. We only use three metrics from Bayley-4, except the scaled scores, since it is difficult to decide the VECA agent's exact biological age.

Experiments

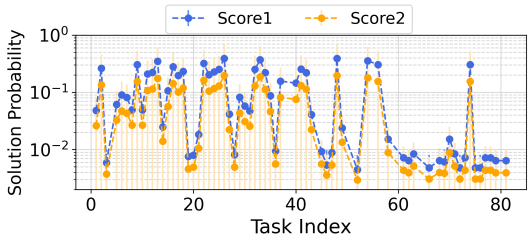
We evaluate four aspects of our VECA benchmark toolkit. First, we share the benchmark results of four representative baselines and show that using standardized metrics of Bayley-4, the cognitive capability of AI can be directly compared to the human. Second, we validate our virtualized Bayley-4 tasks by analyzing the solvability and complexity of each task, following the protocol of (Bakhtin et al. 2019). Third, we show that VECA's human-like sensory characteristics meaningfully affect the development of cognitive skills by training certain tasks where these features are crucial. Finally, we demonstrate that our VECA toolkit can create diverse cognitive tasks with different difficulties by varying the difficulty setup of our toolkit-generated tasks. We list detailed experimental setups in the appendix (Park, Oh, and Lee 2021).

Baselines. Four distinct types of baseline methods are used for our experiments: (i) Policy Gradient, (ii) Curiosity-driven Learning, (iii) Random Agent, and (iv) Human Baseline. We use three modern policy gradient algorithms for our evaluation: IMPALA (Espeholt et al. 2018) as a baseline of our benchmark, and PPO (Schulman et al. 2017) and SAC (Haarnoja et al. 2018) to assess the VECA toolkit itself. These methods represent goal-driven learning with explicit rewards. We first implemented parallelized PPO and SAC that can train with our VECA environment. We also revise the IMPALA implementation of (Küttler et al. 2019) to support our VECA with diverse settings. For curiosity-driven Learning (**CUR**), we used the dynamics-based curiosity model of intrinsic reward (Pathak et al. 2017) revised for our benchmark. CUR learns without any supervision; the only reward signal is the intrinsic prediction error of observation input. It represents unsupervised exploration, which aligns with how humans learn in the early stage of development (Gibson 1988). Random agent samples actions from the uniform distribution in 6-dimensional action space. No training is thus held for this agent. Finally, to measure the human baseline performance, three human adult participants play our VECA benchmark and report their scores.

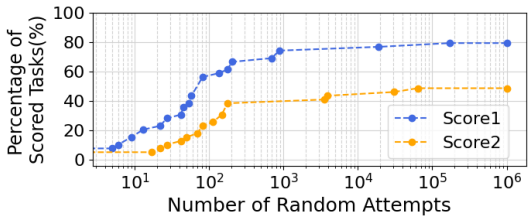
VECA Benchmark Baselines and Metrics. We compare four baselines on our VECA benchmark. We train IMPALA and CUR(policy learners) with the entire VECA tasks, which are uniformly sampled at random per episode. Both the CUR and IMPALA agents are trained for 1M steps. In addition to Bayley-4 metrics, we measure the mean episode reward over the previous 100 episodes.

Figure 5 shows that the policy learners train to surpass the random agent over time, but it is way below the human baseline, leaving plenty of room for reaching human-level

cognition. Policy learners achieve an age equivalent of 5 months and GSV value 477, which is a lower 17.88% of normative human baby data. It shows that these developmental psychology-based metrics give a more intuitive understanding of the developmental status, unlike RL metrics. Note that the IMPALA achieves early improvement with faster convergence, but the CUR constantly improves towards a higher score. Such finding is in contrast with how humans learn; unsupervised exploration dominates in the early development, but it proceeds to goal-driven learning in the later stage.



(a) Probability (y-axis) of random agent scoring for each VECA benchmark task(x-axis).



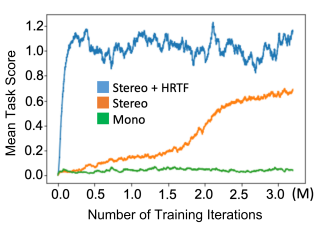
(b) Percentage (y-axis) of VECA tasks scored by random agent as a function of number of attempts (x-axis).

Figure 6: Analysis of complexity and solvability of VECA benchmark. All the tasks are solvable, but it gets more difficult for a task with higher index. Score 2 (Mastery) is much harder to achieve than Score 1 (Emerging).

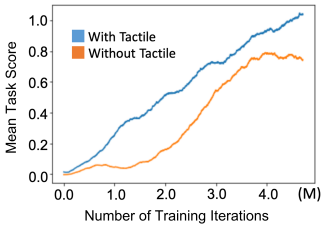
Validity Analysis. We evaluate the validity our VECA benchmark in two folds: (i) solvability and the complexity of each task, (ii) the complexity of the overall VECA benchmark. We use a random agent to analyze the task complexity, e.g., a solution probability of random agent or the number of random trials to solve a portion of tasks.

Figure (6a) shows that all the VECA benchmark tasks are solvable with well-distributed task complexity. Tasks with a higher template index tend to be more difficult, which aligns with how the original Bayley-4 tasks are organized. A clear percentage gap of score-1 curve and score-2 curve in Figure (6b) implies that score-2 of VECA benchmark, a mastery of a certain cognitive skill, is much more difficult to achieve than the emergence of the skill. Furthermore, the random agent fails to solve 100% of tasks even after a large number of trials, suggesting that completely solving our benchmark is a highly challenging goal.

Effect of Human-like Perception. We now study the effect of human-like sensory features on cognitive tasks. We



(a) Auditory Features, *Kicks Squeaky Ball* task.



(b) Tactile Sensation, *Grabs Block* task.

Figure 7: Learning curve of mean task score varying the audio (7a) and tactile (7b) setup. Human-like sensory features noticeably affect the learning of the relevant cognitive task.

use the *Kicks Squeaky Ball* task for auditory perception, which should 3D-localize the transient and dynamic sound source. For tactile perception, we use the *Grabs Block* task, a sensorimotor task that should physically grab a large block with both hands and lift it. PPO is used for the agent training.

Noticeable performance gap of learning curves in Figure 7 shows that VECA's human-like sensory features are crucial in the development of cognitive skills. The agent with HRTF-spatialized audio attains much higher performance faster than the agents without them. Like a human, audio spatialization due to physical body structure appears to be crucial in the sound source localization of a VECA agent. The same applies to the VECA's tactile sensation; tactile sensation seems to plays an essential role in the sensorimotor development of a VECA agent.

Usefulness of VECA Toolkit. To show that our VECA toolkit covers a diverse domain of tasks with variable difficulty, we pick several representative toolkit-generated tasks assessing distinct cognitive skills, and train VECA agents on them with PPO and SAC. Diverse tasks and their domains that we evaluate include *GrabObject* for joint-level control, *ObjectNav* for navigation and visual recognition, *KickTheBall* for multimodal learning, and *MultiAgentNav* for multi-agent RL. Results show that these tasks are trainable, and the user can easily diversify the task difficulty. Due to space, we illustrate the result plot in the appendix (Park, Oh, and Lee 2021).

Conclusion & Future Works

We introduced a new VECA toolkit that can generate diverse and distinct cognitive tasks for a human-like agent. Using our toolkit, we developed a novel VECA benchmark that measures the overall cognitive development of an AI agent for the first time. Our evaluation with the VECA benchmark revealed that current RL algorithms need a significant improvement to acquire general cognitive skills like a human.

In future works, we plan to extend our benchmark to the different *Language and Motor Scales* of Bayley-4, and test cognitive models beyond RL agents, e.g., NLP models, planning, and cognitive architectures. To evaluate the fidelity of our VECA toddler agent to the real human toddler, we also plan to motion-capture a human toddler's movement and compare it with the motion dynamics of our VECA agent.

References

Adams, R. J.; Maurer, D.; and Cashin, H. A. 1990. The influence of stimulus size on newborns' discrimination of chromatic from achromatic stimuli. *Vision research*, 30(12): 2023–2030.

Albers, C. A.; and Grieve, A. J. 2007. Test review: Bayley, N.(2006). Bayley scales of infant and toddler development–third edition. San Antonio, TX: Harcourt assessment. *Journal of Psychoeducational Assessment*, 25(2): 180–190.

Asada, M.; Hosoda, K.; Kuniyoshi, Y.; Ishiguro, H.; Inui, T.; Yoshikawa, Y.; Ogino, M.; and Yoshida, C. 2009. Cognitive Developmental Robotics: A Survey. *IEEE Transactions on Autonomous Mental Development*, 1: 12–34.

Aylward, G. 2017. Bayley Scales of Infant and Toddler Development.

Bakhtin, A.; Maaten, L. V. D.; Johnson, J.; Gustafson, L.; and Girshick, R. B. 2019. PHYRE: A New Benchmark for Physical Reasoning. In *NeurIPS*.

Bambach, S.; Crandall, D. J.; Smith, L. B.; and Yu, C. 2018. Toddler-Inspired Visual Object Learning. In *NeurIPS*.

Bayley, N. 1999. Bayley Scales of Infant Development.

Bögelein, S.; Brinkmann, F.; Ackermann, D.; and Weinzierl, S. 2018. Localization cues of a spherical head model.

Burda, Y.; Edwards, H.; Pathak, D.; Storkey, A.; Darrell, T.; and Efros, A. A. 2019. Large-Scale Study of Curiosity-Driven Learning. *ArXiv*, abs/1808.04355.

Carey, W. B.; Crocker, A. C.; Elias, E. R.; Feldman, H. M.; and Coleman, W. L. 2009. *Developmental-Behavioral Pediatrics E-Book*. Elsevier Health Sciences.

Chandrasekaran, C. 2017. Computational principles and models of multisensory integration. *Current Opinion in Neurobiology*, 43: 25–34.

Chen, C.; Jain, U.; Schissler, C.; Gari, S. V. A.; Al-Halah, Z.; Ithapu, V. K.; Robinson, P.; and Grauman, K. 2020. Soundspaces: Audio-visual navigation in 3d environments. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, 17–36. Springer.

Clements, D.; Swaminathan, S.; Hannibal, M. A.; and Sarama, J. 1999. Young Children's Concepts of Shape. *Journal for Research in Mathematics Education*, 30: 192–212.

Dobson, V.; and Teller, D. Y. 1978. Visual acuity in human infants: a review and comparison of behavioral and electrophysiological studies. *Vision research*, 18(11): 1469–1483.

Doya, K.; and Taniguchi, T. 2019. Toward evolutionary and developmental intelligence. *Current Opinion in Behavioral Sciences*, 29: 91–96.

Dusing, S. 2016. Postural variability and sensorimotor development in infancy. *Developmental Medicine & Child Neurology*, 58.

El Bab, A. M. F.; Tamura, T.; Sugano, K.; Tsuchiya, T.; Tabata, O.; Eltaib, M. E.; and Sallam, M. M. 2008. Design and simulation of a tactile sensor for soft-tissue compliance detection. *IEEJ Transactions on Sensors and Micromachines*, 128(5): 186–192.

Espeholt, L.; Soyer, H.; Munos, R.; Simonyan, K.; Mnih, V.; Ward, T.; Doron, Y.; Firoiu, V.; Harley, T.; Dunning, I.; Legg, S.; and Kavukcuoglu, K. 2018. IMPALA: Scalable Distributed Deep-RL with Importance Weighted Actor-Learner Architectures. *ArXiv*, abs/1802.01561.

Franchak, J. M.; van der Zalm, D. J.; and Adolph, K. E. 2010. Learning by doing: Action performance facilitates affordance perception. *Vision research*, 50(24): 2758–2765.

Frost, J.; Wortham, S.; and Reifel, R. S. 2000. Play and Child Development.

Gardner, B. 1994. HRTF Measurements of a KEMAR Dummy-Head Microphone.

Geary, D. C.; et al. 2018. Growth of symbolic number knowledge accelerates after children understand cardinality. *Cognition*, 177: 69–78.

Gibson, E. J. 1988. Exploratory behavior in the development of perceiving, acting, and the acquiring of knowledge. *Annual review of psychology*, 39(1): 1–42.

Greiff, S.; Wüstenberg, S.; Goetz, T.; Vainikainen, M.-P.; Hautamäki, J.; and Bornstein, M. H. 2015. A longitudinal study of higher-order thinking skills: working memory and fluid reasoning in childhood enhance complex problem solving in adolescence. *Frontiers in psychology*, 6: 1060.

Haarnoja, T.; Zhou, A.; Abbeel, P.; and Levine, S. 2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *arXiv preprint arXiv:1801.01290*.

Juliani, A.; Berges, V.-P.; Vckay, E.; Gao, Y.; Henry, H.; Mattar, M.; and Lange, D. 2018. Unity: A general platform for intelligent agents. *arXiv preprint arXiv:1809.02627*.

Koenig, N.; and Howard, A. 2004. Design and use paradigms for gazebo, an open-source multi-robot simulator. In *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)(IEEE Cat. No. 04CH37566)*, volume 3, 2149–2154. IEEE.

Kolve, E.; Mottaghi, R.; Han, W.; VanderBilt, E.; Weihs, L.; Herrasti, A.; Gordon, D.; Zhu, Y.; Gupta, A.; and Farhadi, A. 2017. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*.

Küttler, H.; Nardelli, N.; Lavril, T.; Selvatici, M.; Sivakumar, V.; Rocktäschel, T.; and Grefenstette, E. 2019. TorchBeast: A PyTorch Platform for Distributed RL. *arXiv preprint arXiv:1910.03552*.

Lake, B. M.; Ullman, T. D.; Tenenbaum, J. B.; and Gershman, S. J. 2017. Building machines that learn and think like people. *Behavioral and brain sciences*, 40.

Landau, B.; Smith, L.; and Jones, S. 1998. Object perception and object naming in early development. *Trends in cognitive sciences*, 2(1): 19–24.

Lawless, H. T.; and Heymann, H. 1999. Measurement of sensory thresholds. In *Sensory evaluation of food*, 173–207. Springer.

Metta, G.; Natale, L.; Nori, F.; Sandini, G.; Vernon, D.; Fadiga, L.; Von Hofsten, C.; Rosander, K.; Lopes, M.; Santos-Victor, J.; et al. 2010. The iCub humanoid robot:

An open-systems platform for research in cognitive development. *Neural Networks*, 23(8-9): 1125–1134.

Moustafa, B. M. 1999. Multisensory Approaches and Learning Styles Theory in the Elementary School: Summary of Reference Papers.

Murphy, N. 1997. A Multisensory vs. Conventional Approach to Teaching Spelling.

Park, K.; Oh, H.; and Lee, Y. 2021. Technical Appendix. https://github.com/GGOSinon/VECA/blob/master/docs/Supplementary_Materials.pdf. Accessed: 2021-12-16.

Pathak, D.; Agrawal, P.; Efros, A. A.; and Darrell, T. 2017. Curiosity-Driven Exploration by Self-Supervised Prediction. *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 488–489.

Piaget, J.; and Cook, M. 1952. *The origins of intelligence in children*, volume 8. International Universities Press NY.

Potisk, T. 2015. Head-related transfer function.

Rabinowitz, N. C.; Perbet, F.; Song, H. F.; Zhang, C.; Eslami, S.; and Botvinick, M. 2018. Machine Theory of Mind. *ArXiv*, abs/1802.07740.

Ranjitkar, S.; Hysing, M.; Kvestad, I.; Shrestha, M.; Ulak, M.; Shilpakar, J. S.; Sintakala, R.; Chandyo, R.; Shrestha, L. S.; and Strand, T. 2019. Determinants of Cognitive Development in the Early Life of Children in Bhaktapur, Nepal. *Frontiers in Psychology*, 10.

Ren, Z.; Nie, J.; Shao, J.; Lai, Q.; Wang, L.; Chen, J.; Chen, X.; and Wang, Z. L. 2018. Fully elastic and metal-free tactile sensors for detecting both normal and tangential forces based on triboelectric nanogenerators. *Advanced Functional Materials*, 28(31): 1802989.

Reynolds, G. D. 2015. Infant visual attention and object recognition. *Behavioural Brain Research*, 285: 34–43.

Rose, S. A.; Feldman, J. F.; Jankowski, J. J.; and Van Rossem, R. 2012. Information processing from infancy to 11 years: Continuities and prediction of IQ. *Intelligence*, 40(5): 445–457.

Rovee-Collier, C.; and Hayne, H. 1987. Reactivation of infant memory: implications for cognitive development. *Advances in child development and behavior*, 20: 185–238.

Sarnecka, B.; and Wright, C. 2013. The Idea of an Exact Number: Children's Understanding of Cardinality and Equinumerosity. *Cognitive science*, 37 8: 1493–506.

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

Shu, T.; Bhandwaldar, A.; Gan, C.; Smith, K. A.; Liu, S.; Gutfreund, D.; Spelke, E.; Tenenbaum, J.; and Ullman, T. D. 2021. AGENT: A Benchmark for Core Psychological Reasoning. In *ICML*.

Silver, D.; Singh, S.; Precup, D.; and Sutton, R. 2021. Reward is enough. *Artificial Intelligence*, 299: 103535.

Sloman, A. 1999. What sort of architecture is required for a human-like agent? In *Foundations of rational agency*, 35–52. Springer.

Smith, L. B.; and Gasser, M. 2005. The Development of Embodied Cognition: Six Lessons from Babies. *Artificial Life*, 11: 13–29.

Song, Z.; Banks, R. W.; and Bewick, G. 2015. Modelling the mechanoreceptor's dynamic behaviour. *BMC Neuroscience*, 16: P16 – P16.

Sousa, G. H.; and Queiroz, M. 2010. Two approaches for HRTF interpolation.

Stein, B. 2012. The new handbook of multisensory processes.

Subiaul, F.; Cantlon, J.; Holloway, R.; and Terrace, H. 2004. Cognitive Imitation in Rhesus Macaques. *Science*, 305: 407 – 410.

Tacca, M. C. 2011. Commonalities between perception and cognition. *Frontiers in psychology*, 2: 358.

Thompson, R. F.; and Spencer, W. 1966. Habituation: a model phenomenon for the study of neuronal substrates of behavior. *Psychological review*, 73 1: 16–43.

Tollin, D. J. 2009. Development of sound localization. *Oxford handbook of developmental behavioral neuroscience*.

Valenti, C. 2006. Infant Vision Guidance: Fundamental Vision Development in Infancy.

van Hof, P.; van der Kamp, J.; and Savelsbergh, G. 2006. Three- to eight-month-old infants' catching under monocular and binocular vision. *Human movement science*, 25 1: 18–36.

Vedantam, R.; Szlam, A. D.; Nickel, M.; Morcos, A. S.; and Lake, B. 2021. CURI: A Benchmark for Productive Concept Learning Under Uncertainty. In *ICML*.

Vogt, F.; Hauser, B.; Stebler, R.; Rechsteiner, K.; and Urech, C. 2018. Learning through play–pedagogy and learning outcomes in early childhood mathematics. *European Early Childhood Education Research Journal*, 26(4): 589–603.

Wang, Y.; Huang, W.; Fang, B.; Sun, F.; and Li, C. 2021. Elastic Tactile Simulation Towards Tactile-Visual Perception. *ArXiv*, abs/2108.05013.

Wimmer, H.; and Perner, J. 1983. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13: 103–128.

Wu, Y.; Wu, Y.; Gkioxari, G.; and Tian, Y. 2018. Building generalizable agents with a realistic and rich 3D environment. *arXiv preprint arXiv:1801.02209*.

Xia, F.; R. Zamir, A.; He, Z.-Y.; Sax, A.; Malik, J.; and Savarese, S. 2018. Gibson env: real-world perception for embodied agents. In *Computer Vision and Pattern Recognition (CVPR), 2018 IEEE Conference on*. IEEE.

Zhao, W.; Queralta, J. P.; and Westerlund, T. 2020. Sim-to-Real Transfer in Deep Reinforcement Learning for Robotics: a Survey. *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, 737–744.

Zhu, Y.; Gao, T.; Fan, L.; Huang, S.; Edmonds, M.; Liu, H.; Gao, F.; Zhang, C.; Qi, S.; Wu, Y.; Tenenbaum, J.; and Zhu, S.-C. 2020. Dark, Beyond Deep: A Paradigm Shift to Cognitive AI with Humanlike Common Sense. *ArXiv*, abs/2004.09044.