Toddler-Guidance Learning: Impacts of Critical Period on Multimodal AI Agents

Junseok Park
Seoul National University
Seoul, Republic of Korea

Kwanyoung Park
Seoul National University
Seoul, Republic of Korea

Hyunseok Oh
Seoul National University
Seoul, Republic of Korea

Ganghun Lee
Seoul National University
Seoul, Republic of Korea

Minsu Lee
Seoul National University, AIIS
Seoul, Republic of Korea

Youngki Lee
Seoul National University
Seoul, Republic of Korea

Byoung-Tak Zhang
Seoul National University, AIIS
Seoul, Republic of Korea

ABSTRACT

Critical periods are phases during which a toddler's brain develops in spurts. To promote children's cognitive development, proper guidance is critical in this stage. However, it is not clear whether such a critical period also exists for the training of AI agents. Similar to human toddlers, well-timed guidance and multimodal interactions might significantly enhance the training efficiency of AI agents as well. To validate this hypothesis, we adapt this notion of critical periods to learning in AI agents and investigate the critical period in the virtual environment for AI agents. We formalize the critical period and Toddler-guidance learning in the reinforcement learning (RL) framework. Then, we built up a toddler-like environment with VECA toolkit to mimic human toddlers' learning characteristics. We study three discrete levels of mutual interaction: weak-mentor guidance (sparse reward), moderate mentor guidance (helper-reward), and mentor demonstration (behavioral cloning). We also introduce the EAVE dataset consisting of 30,000 real-world images to fully reflect the toddler's viewpoint. We evaluate the impact of critical periods on AI agents from two perspectives: how and when they are guided best in both uni- and multimodal learning. Our experimental results show that both uni- and multimodal agents with moderate mentor guidance and critical period on 1 million and 2 million training steps show a noticeable improvement. We validate these results with transfer learning on the EAVE dataset and find the performance advancement on the same critical period and the guidance.

CCS CONCEPTS

• **Computing methodologies → Generative and developmental approaches**; **Sequential decision making**; **Transfer learning**; **Neural networks**.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
ICMI '21, October 18–22, 2021, Montréal, QC, Canada
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8481-0/21/10…$15.00
https://doi.org/10.1145/3462244.3479932

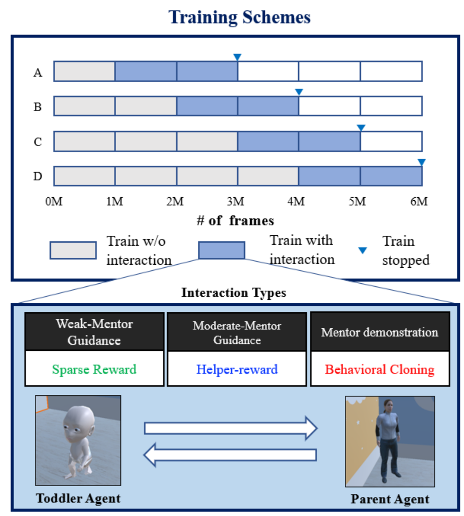


Figure 1: Overall approach to testing the critical period. We first trained the agent for 1/2/3/4 million frames only with sparse rewards. We then continued training the agent with guidance (helper-reward, behavior cloning) or without guidance (sparse reward) for an additional 2 million frames.

KEYWORDS

Reinforcement Learning; Toddler Object Learning; Virtual Environment for Cognitive Agents; Guidance

ACM Reference Format:
Junseok Park, Kwanyoung Park, Hyunseok Oh, Ganghun Lee, Minsu Lee, Youngki Lee, and Byoung-Tak Zhang. 2021. Toddler-Guidance Learning: Impacts of Critical Period on Multimodal AI Agents. In *Proceedings of the 2021 International Conference on Multimodal Interaction (ICMI '21), October 18–22, 2021, Montréal, QC, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3462244.3479932

1 INTRODUCTION

Mentor guidance is critical in the early stages of cognitive development. Learning in the child stage significantly influences learning-to-learn capabilities like goal-setting and self-control [56]. For this reason, parent-child interaction is crucial in the overall cognitive development process, including linguistics and object understanding [28]. Adequate supervision in the early stages of learning is also important for machine learning (ML), as the sample efficiency of supervised learning is in general much higher than that of unsupervised approaches [26]. Prior works have investigated diverse forms of guided learning on ML agents such as distillation [2], imitation learning, weakly-supervised RL [23]. However, there has been limited cross-domain research to find an optimal guidance strategy for a given task.

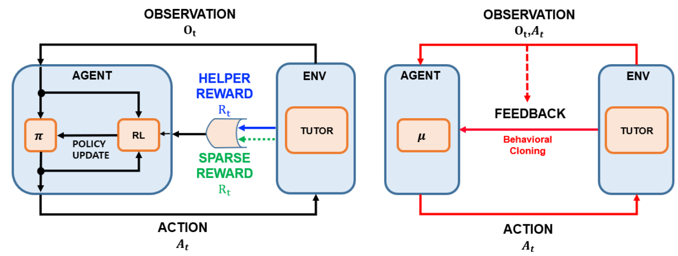
In this work, we are inspired by the concept of *critical periods* in the infant-toddler learning, a specific period in which the acquisition of language and visual/auditory processing capabilities are highly accelerated [10, 16, 21, 22, 34, 40, 41], and we explore the question, "does such a critical period also exist for machine learning (ML) agents?". In particular, we aim to identify the optimal guidance during training on a certain task, and study the correspondence between the optimal duration of guidance in ML algorithms and the critical period in human learning. This study is significant in that the optimal guidance strategy for a task can significantly enhance the training efficiency of the ML model.

To this end, we define and empirically study the optimal initial time of mentor guidance as the counterpart of the critical period in AI agents. To do this we formulate a reward structure that is variable through the training iterations, in terms of the Partially Observable Markov Decision Process(POMDP) in RL framework. We then formalize mentor guidance as a policy-invariant mutable reward structure on a given task. Finally, we formally define the optimal initial time of mentor guidance, and study it as the AI agent's equivalent to the *critical period* seen in toddlers.

Our key finding from experiments is that reinforcement learning agents indeed have a critical period much like toddlers in the moderate guidance learning setting. To demonstrate this, we have implemented guidance inside the VECA [29] virtual environment in a multimodal agent RL framework to model mentor guidance. First, we have modeled learning-through-play of toddlers with an object navigation RL task using VECA. Second, we introduce the EAVE dataset, a real-world image and audio dataset accurately reflecting the toddler's point of view. For our experiments, we have imposed three levels of guidance as seen in Fig 2: weak-mentor guidance (sparse reward), moderate mentor guidance (helper-reward), and mentor demonstration (behavioral cloning) and we adjust these guidances and their duration to find the best-performing setting. Our studies on unimodal and multi-modal RL agents show that the critical period has appeared only when the toddler agent has moderate guidance while learning. Moreover, we validate our results using the EAVE dataset via transfer learning. Results confirm that the models trained on their appropriate critical period (1M&2M iterations) noticeably outperform the models trained after (3M&4M iterations) their critical period.

To summarize, this paper's contributions are threefold:

- We use a policy-invariant mutable reward structure to formulate the human toddler's critical period in an ML model, especially the multimodal humanoid RL agent we developed.
- We construct a multimodal Navigation RL task and a real-world EAVE dataset that can effectively demonstrate the critical period in a multimodal RL agent.



(a) Reward bonus (b) Behavioral cloning

Figure 2: Three types of guidance. (a) Weak-Mentor Guidance (green) and Moderate-Mentor Guidance (blue) use a reward bonus method based on RL. (b) Mentor demonstration (red) use behavioral cloning (Imitation Learning)

- We empirically study the critical period in a unimodal and multimodal RL task with four different periods: 1/2/3/4 million training iterations. Agent trained in a critical period outperforms others significantly with proper guidance. Furthermore, the performance advantage transfers well to the real-world downstream task.

2 RELATED WORKS

Curriculum Learning [4, 9, 31, 37, 44, 50] is an easy-to-hard training strategy in ML that gradually increases the complexity of the data samples used during the training process [43]. Humans learn the easy and preliminary concepts sooner and the difficult and advanced concepts later. Moreover, humans learn much better when the examples are not randomly presented but are organized in a meaningful order [13, 43]. Curriculum learning applies a similar strategy for ML model training, and achieves two major advantages: faster convergence and better performance. There is a number of domains including face recognition [24], object segmentation [55] and reinforcement learning [46] in which curriculum learning has been successfully applied. However, using curriculum learning does not always benefit performance. It requires careful design of the difficulty hierarchy before training. For instance, the diversity of the data samples may be restricted if the task difficulty measure prefers choosing easy examples only from a small subset of classes [42, 43]. An approach diametrical to easy-to-hard curriculum learning, that is emphasizing harder examples as in hard example mining [18, 39] or anti-curriculum learning [7, 32], can achieve improved results in some cases.

ML researchers have recently taken interest in ideas from child learning further technical advances [3, 36, 49, 51, 52]. Learning properly in a child stage is crucial for learning-to-learn capabilities like goal-setting and self-control [48, 56]. Life experiences and prior knowledge learned in one's childhood are known to influence adult learning as well. The latest works in deep learning like visual object learning [3] are inspired by the way children learn. Achille et al. [1] is the first work to show that deep artificial neural networks also exhibit *critical periods* like humans, a decisive time window for the post-natal developmental stage. In deep neural network training, a rapid growth in information is followed by a reduction of information from an analysis with Fisher Information. Unfortunately, their analysis is limited to supervised learning in convolution neural networks (CNNs), and not executed on either RL or multimodal setup. To the best of our knowledge, our work is the first attempt to apply the toddler's developmental characteristic of humans to effective RL training. We aim to understand how the functionality and learning-to-learn capabilities are nurtured through appropriate supervision.

Several works in developmental psychology note that more guidance does not mean better human development. A general understanding of objects develops in an early stage of toddler learning with weak supervision as well as through interactions like mouthing, chewing, and rotating [14, 33]. Furthermore, teaching formal subjects too early for a child is counterproductive in the long run [45]. Inspired by human toddler learning, we conjecture that there is also a form of optimal supervision and an optimal time period for RL training of intelligent machines. We empirically search for a type of guidance that accelerates the training, but does not change the optimal policy of the environment.

3 METHOD

Reinforcement Learning. Reinforcement learning (RL) is a framework for experience-driven autonomous learning [46] in which an agent interacts with the environment to learn optimal behaviour, improving over time through trial and error [2]. Upon observing the consequent state, an agent learns to alter its policy in response to received rewards. The environment in RL is typically assumed to be a Markov Decision Process (MDP). MDP is memoryless; the conditional probability of the next state and reward are stochastically determined only from the current state and action, and do not depend on the past [38].

3.1 Guidance in a POMDP

In social psychology, a *social behavior* or *social action* is a behavior that affects the other person and provokes a response [17, 47]. A general social interaction sequence is defined as a sequence of a person acting toward another (under the expectation), and the other acting towards the person in response (under the interpretation) [11]. Guided learning, or *mentor guidance*, can be defined as a process in which learners achieve their learning goal with the supervision of a more experienced mentor [5].

Inspired by these psychological discoveries, we formalize the notion of *Guidance* in the context of Partially Observable Markov Decision Process (POMDP). We find it analogous of a person as an agent, and an effect of the behavior as a change in the observation.

Partially-observable Markov decision processes (POMDP) extend RL frameworks so that an agent observes only part of the world state [12]. A POMDP is a 7-tuple $\langle S, A, T, R, \Omega, O, \gamma \rangle$ consisting of: S, a set of possible states; A, a set of actions; T, a set of transition probabilities; R, a reward function $R(s, a)$ for a state s and action a; Ω, a set of possible observations, O, a set of observation probabilities $o(s, a) = Pr(o|a, s)$ for given state s and action a; and γ a discount factor. We extend this to include a no-op action (do nothing) in the action space as $A' = A \bigcup \{a_{nop}\}$.

3.1.1 *Mutable Reward Structure.*

Mutable reward structure is a reward function that is variable under number of iterations. Since MDP is memoryless, the state should incorporate the number of training iteration as a reference. Suppose the state consists of (s, t) where $t \in \mathbb{N}$ is the training iteration. The reward structure $\tilde{R}$ is mutable if such state $s \in S$, action $a \in A$ exists.

$$\tilde{R}((s, t), a) \neq \tilde{R}((s, t + 1), a) \tag{1}$$

The mutable reward structure is dependent on the past, so the properties under the MDP do not hold anymore. However, the policy is invariant [27] if two MDPs become graph-isomorphic after a positive linear transformation $f : \mathbb{R} \to \mathbb{R}$ is applied between corresponding weights. Policy invariance also holds when the reward structure is a potential-based shaping function. Reward structure morphing to a policy-invariant MDP preserves the optimal policy, and the value of the current policy. We further define such reward function as a policy-invariant mutable reward structure.

3.1.2 *Mentor Guidance.*

Mentor guidance can be interpreted as a policy-invariant mutable reward structure that efficiently trains the learner (mentee). Parent-child interaction, one of the major mentor guidance in human, is a special form of interaction which establishes a strong bond between the parent and the child and greatly facilitates the cognitive development of a child [28]. The parent agent observes the child's state and has interactions with the child to optimize the child's ability on various tasks. It shows that continuous care and guidance of mentor can hugely facillitate the mentee's performance. Motivated by it, we investigate a specific case of policy-invariant mutable reward structure, e.g. helper reward, aiming to optimize the learner's performance. We also study the behavior cloning, a mutable reward structure that the policy invariance is not known. The guidance G_T on a task T is defined as a reward structure $G_T = \tilde{R}_T(s, a)$ where $s \in S$, $a \in A$ with agents performance metric as J_T.

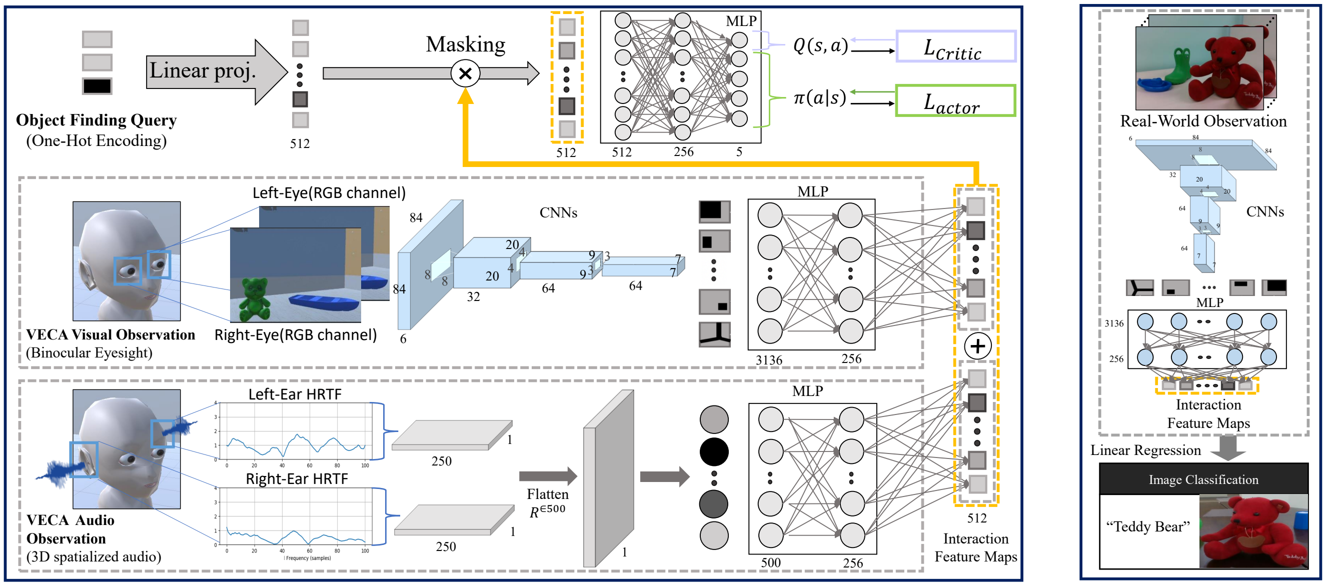
3.2 Critical Period

Critical periods are maturational stages in the lifespan of humans that act as decisive period of learning in a particular learning domain as human grow up [16]. The critical period for the development of the human visual system is considered to be between three and eight months [34]. Other critical periods have also been identified for the development of hearing and the vestibular system [6, 21]. Besides, critical periods for language acquisition indicate that the first few years of life constitute the time during which language develops readily, and after which language acquisition is much more difficult. [40] To sum up, critical period stands for the effective initiation time for supervising the cognitive development after birth. We can see the correspondence between the critical period of a toddler and the morphing of a reward structure. In this work, we seek the critical period of an ML agent. We empirically search whether the initial morphing period of best guidance is finite, and can be reached in a reasonable number of training iterations.

3.3 Optimal Guidance as a Critical Period

3.3.1 *Initial Time of Guidance.*

For human, the parent's guidance initiates from the beginning of the child's life. For the machine intelligence however, the guidance should persist only for a certain period for a learner to solve the task and learn on their own. We model this characteristic to the RL setting. Given the guidance, the reward morphing starts from an initial time t_G and continues only for a certain amount of time. The guidance resets from this point on.



(a) Training Process **(b) Transfer Learning**

Figure 3: Architecture of the baby agent. (a) Its visual observation is encoded as a feature map from the CNN encoder, and the audio observation is encoded with an MLP encoder. These feature maps are concatenated and linearly combined to make an *interaction feature map*. It is masked with a linearly projected object embedding and fed to an MLP, which outputs the Q-function $Q(s,a)$ and movement policy $\pi(a|s)$. (b) We validate helper models in another relevant task via transfer learning

We denote the *initial morphing time* of the guidance as t_G. So for a series of learning episodes $\{E_i^G\}$ with guidance G and their duration $\{t_i\}$, the time interval of guidance lies in $[t_G, t_G + \sum_i t_i]$. This raises the question: What is the optimal initial period of guidance for the learner's performance?

3.3.2 Optimal Guidance. We can define an optimal guidance policy on a set of guidance reward structures $\{G_j\}$ as the one which results in the best performance with their optimal initial morphing time. Suppose the $J_T(\pi_\theta)$ is the performance of policy parameter θ on task T, and $\Pr(\theta|G, t_G)$ is a trained policy distribution under the guidance $R = G \in \{G_j\}$ and its duration t_G. The optimal guidance can be defined as the guidance and its initial morphing period tuple (G^*, t_G^*) providing the best performance on a specific task T.

$$(G^*, t_G^*) = \arg\max_{G, t_G} \left(\max_{\theta \sim Pr(\theta|G, t_G)} \mathbb{E}_{\pi_\theta} \left(J_T(\pi_\theta) \right) \right) \tag{2}$$

3.4 Transfer Learning

Transfer learning aims to help improve learning of some task in the target domain using knowledge learned from the source domain [53]. Transfer learning is used for few-shot learning on new datasets or in previously unseen domains both to reduce training effort and to transfer some of the knowledge and inductive biases acquired during training in the source domain. One recent work modeled learning-through-play of a toddler by training a Deep Neural Network (DNN) model in a RL framework and transfer this model to a different domain to evaluate it in terms of visual object understanding [30]. They aim to acquire a general understanding of objects by exploring the environment and actively interacting with the objects, as a toddler does.

Hyperparameter	Candidate values	Optimal value
α (entropy coefficient)	{0.003, 0.01, 0.03}	0.01
Learning rate	{0.0001, 0.0003, 0.001}	0.0003
γ (discount factor)	{0.95, 0.99}	0.99

Table 1: Tuned hyperparameters used in SAC.

4 EXPERIMENTS AND RESULTS

4.1 Implementation Details

We designed the architecture of the agent as seen in Fig. 3 (a) to learn and store transferable knowledge. We used Soft Actor-Critic (SAC) [15] to train the agent. In particular, we used 8 workers and updated the parameter per 256 steps with batch size 512 using the replay buffer size of 20,000. Other important hyperparameters of SAC are tuned as shown in Table. 1.

Visual observations of the agent are encoded using a Convolutional Neural Network (CNN) encoder and an Multi-Layer Perceptron (MLP), resulting in *interaction feature maps*. These features are masked with a linearly embedded representation of the object and determine the action of the agent. Since the agent's movement only depends on masked features, the agent must learn to represent abstract features of its observation corresponding to the target object. Then the visual feature extractor of the agent is transferred to the EAVE dataset and learned with the Adam [19] gradient descent optimizer. We trained the network on a server with Ubuntu 18.04 LTS, Xeon Gold 5128 Scalable CPU and six RTX3090 GPUs.

4.2 Interactive Virtual Environment

We implemented an interactive virtual environment supporting toddler-like visual observations and active physical interaction with objects, to train and evaluate guidance on different levels

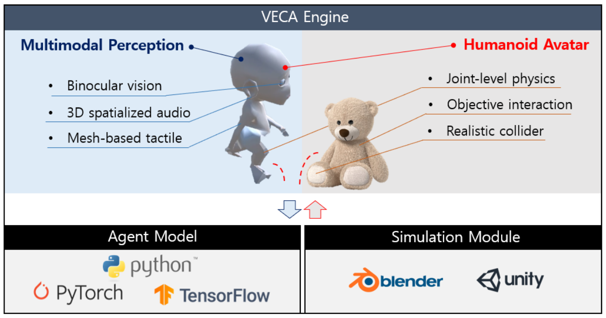
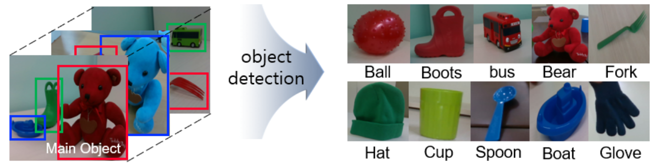


Figure 4: VECA engine characteristics. Unlike other existing RL environments, we trained the toddler AI agents in a setting endowed with various human features and characteristics (e.g. binocular vision).

and for different durations. We used VECA toolkit [29], a virtual environment generation toolkit for human-like agents, to build the toddler-guidance learning environment. In particular, we leveraged a toddler AI agent to be a humanoid avatar including several human characteristics: 1) binocular vision, 2) 3D spatialized audio, 3) mesh-based tactiles, 4) joint-level physics, 5) objective interaction, and a 6) realistic collider as seen in Fig. 4. The concept of the critical period was founded on human learning. Therefore, we believe that a humanoid agent is able to more precisely learn from observations than most of the existing RL environment agents, which don't reflect the human features for testing the critical period [8, 20, 35, 54]. The agent's goal is to learn the ability to visually locate distant objects, which are the same 10 objects as in the EAVE real-world dataset seen in Fig. 5 (b), while freely exploring and observing nearby objects as seen in Fig 6 (b) and (c). The environment's reward structure is modified to implement the different levels and durations of guidance. We aim to observe the influence of guidance on establishing a profound visual understanding of an object during the RL stage.

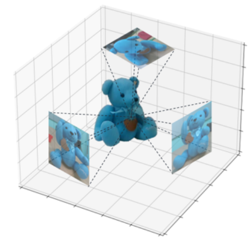
4.3 EAVE Dataset Collection

We have collected an Egocentric Audio-Visual Exploration Dataset for Unsupervised Alignment Learning (EAVE dataset) in order to approximate everyday images observed by toddlers. Most existing visual image datasets [25] do not take into account how humans actually perceive the input, meaning their "first-person" perspective, thus they cannot fully capture the features of real human-visual input. In particular, the way toddlers perceive the world has two main unique aspects: 1) in the input observed by toddlers objects occupy a relatively larger size compared to an adult's view of the scene (so the image has a much higher resolution); 2) the world is observed by toddlers from a larger variety of different angles than by adults [3]. To complement the aforementioned characteristics of the toddler's point of view, we used 10 toys in red, green and blue (for a total of 30 objects) as can be seen in Fig. 5 (a) and collected 30,000 images using a depth camera. Because toddlers do not typically see only a single object in real-world environments, each image includes three objects much like the "real-world observations" from Fig. 5 (a). We are attempting to collect the image perspectives by assuming the baby's point of view from a variety of angles Fig. 5 (c).



(a) Real-World dataset(EAVE) objects

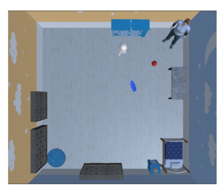
(b) Virtual Environment (VECA) objects



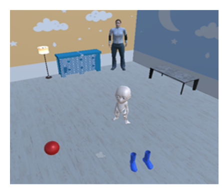
(c) Example of datasets(Teddy bear) captured from various angles

Figure 5: (a) Example of visual images with bounding boxes and the 10 objects used in the EAVE dataset, (b) The same 10 objects in the Virtual Environment (VECA), (c) The EAVE visual dataset includes egocentric images from the toddler's POV at various angles.

Moreover, each image has one main object which has the baby's attention. For each image, we label the bounding box information of the three objects and the main object such that it is possible to crop each object individually as a single-object image. Note that the EAVE dataset is only used for transfer learning experiments in the Study 3.



(a)



(b)

(c)

Figure 6: Various viewpoints of our 3D environment. (a) Developed 3D-toddler playroom Environment top view, (b) Third-person point of view, (c) First-person point of view of a toddler agent while learning the objects.

4.4 Study 1: Examining the impact of the level of interaction in unimodal learning

The purpose of study1 is to compare the performance of the agent in a visual (unimodal) environment trained with various start times (for a fixed duration of guidance) and mutual interaction in the form of guidance, and find if there exists a critical period for each mutual interaction. To observe the effect of mutual interaction, we designed the procedure of the experiment as shown in the left side of Fig. 3. We first trained the agent for 1/2/3/4 million frames only with sparse rewards. For each pre-trained agent, we then continued training the agent with guidance (helper-reward, behavior cloning) or without guidance for 2 million frames as seen in Fig 1.

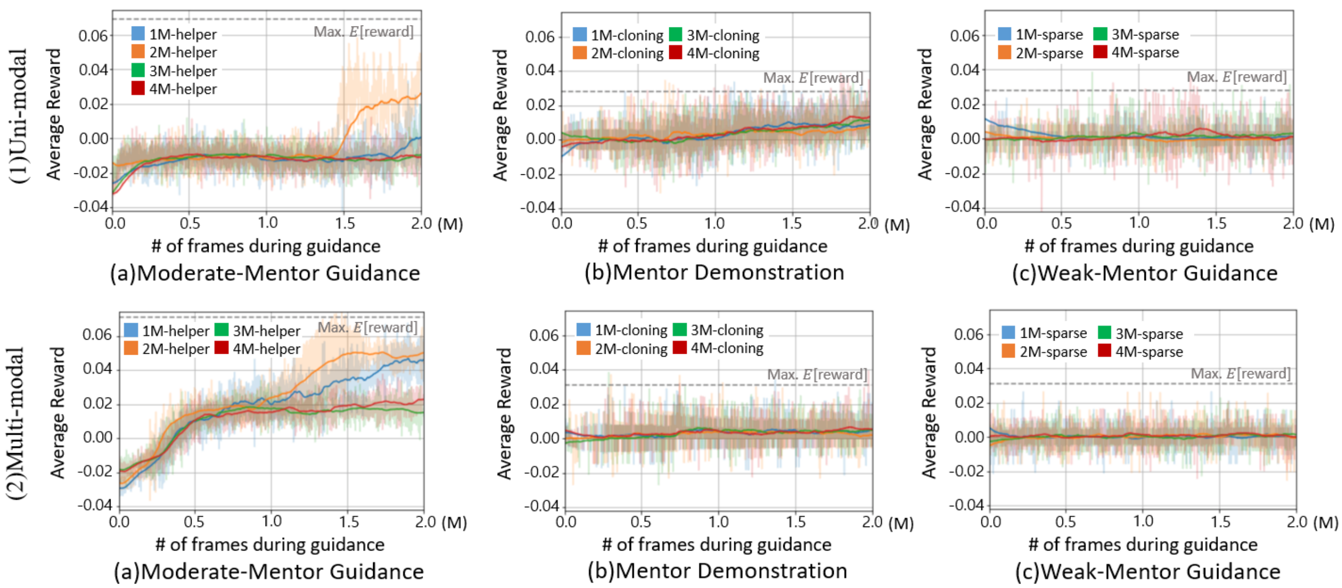


Figure 7: Performance of the agents during the guidance. Vertical axis indicates the average reward while the gray dotted line indicates average reward of the ideal policy (0.069/0.028 and 0.077/0.031 for unimodal and multimodal environment with/without helper reward). Horizontal axis indicates the number of frames that the guidance is provided. Agents guided with helper-reward show a remarkable difference in the performance. Agents guided with behavior-cloning improves, but the difference is negligible. Without guidance, agents fail to improve. For helper-reward in multi-modal environment, we used two seeds to ensure that the result is not by the randomness of training. For other experiments, we used single seed.

4.4.1 Task Design. We designed an object-finding task to train the agent in the virtual environment. We used the VECA [29] toolkit to implement and built up an interactive 3D environment that includes ten objects (of the same categories as in the EAVE dataset) and a baby agent. For each episode, two objects are colored randomly (in red, green, blue) and randomly placed with the agent in the square room with length of 18 units, with the distance of at least 4 units for each other as seen in Fig 6 (a). The agent receives the reward +1 for reaching the target object, and receives -1 for reaching the wrong object. We say that the agent reached the object if the agent is looking towards the object and the distance is less than 2.5 units. Note that the distance is necessary because the object, which is on the floor, will disappear from the eyesight if it gets closer since the agent can't move its head vertically. The maximum length of each episode is 256. The agent receives RGB 84x84 binocular vision and receives the information about the target object as a 10-dimensional one-hot vector. The agent can walk around the environment with maximum speed of 0.33 units per frame, which is represented by two continuous action variables a_f (within [0, 1], forward walking speed) and a_r (within [-1, 1], direction of walking).

Please note that this task is not easy to solve with SAC without guidance. First, random (maximum-entropy) actions in this environment are likely to lead the agent to head into the wall, which fills most of the replay buffer with unhelpful data. Moreover, the agent has to learn color- and perspective-independent features of the object since the object is placed and colored randomly. Also, the presence of a wrong object requires the agent not only to navigate to the target object, but also to avoid the wrong object.

This reward function (3) encourages the agent to look towards the object, visually recognize it, and move towards the object.

4.4.2 Design of guidance. For moderate mentor guidance, we gave the agent helper reward formulated as follows:

$$r' = \begin{cases} -0.03a_f & \text{If target is out of eyesight} \\ 0.05a_r + 0.03a_f & \text{If target is left of eyesight} \\ -0.05a_r + 0.03a_f & \text{If target is right of eyesight} \end{cases} \tag{3}$$

For mentor demonstration, we set the mentor's policy according to its visual information. In particular, if the target object is out of eyesight then the agent turns right to find the object. If the target object is in its eyesight then the agent walks forward and turns slightly left/right if the object is on the right/left side of eyesight.

4.4.3 Results. For guidance from helper-rewards, we found that there exists a noticeable difference in performance depending on the time from which the guidance is provided as shown in Fig. 7(1)(a). For convenience, we denote nM-helper as the model that is trained with helper reward guidance starting from n million frames. In detail, the 2M-helper had much higher average reward (0.03) than the other agents (0.00/-0.01/-0.01 for 1M/3M/4M). This shows that a critical regime indeed exists for guidance from helper-rewards in the object-finding task. Note that for guidance from helper-rewards, helper reward is included when calculating the reward plotted in Fig. 7(1)(a) so the reward is negative on the early stage. Fig. 8(1)(a) compares the trajectories of the agents, visually showing that the policy of the 2M-helper performs better than the other agents.

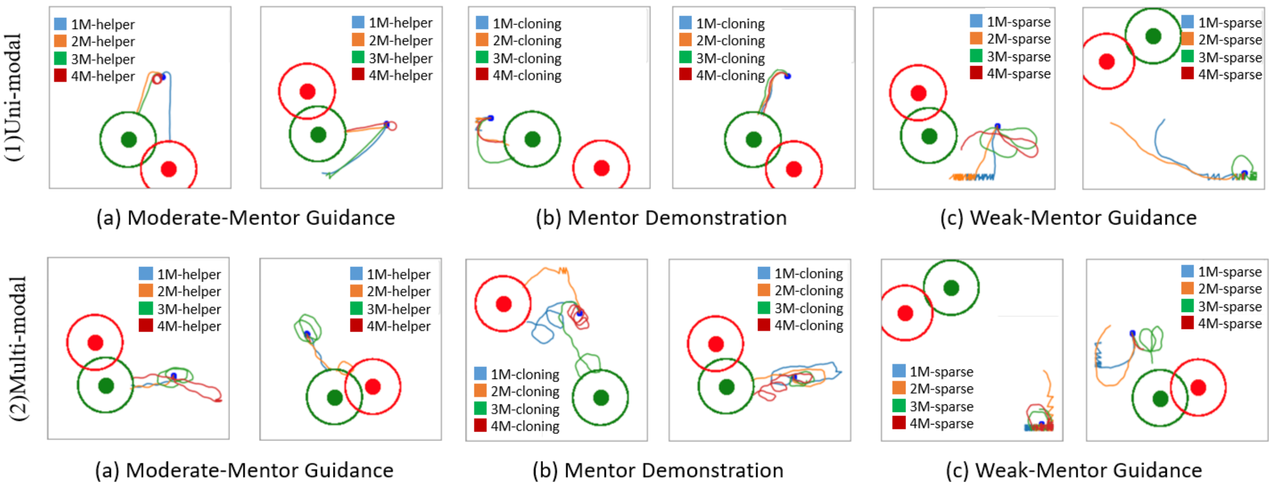


Figure 8: Visualization of the trajectory for each agent's policy in unimodal and multimodal environments. Blue dots represent starting points of the agents. Green/red dots and circles represent the target/wrong object and the boundary for reaching it respectively. Note that the agent has to get inside the boundary and also face towards the object to clear the task.

However, we found that there was no remarkable difference between the performance for behavior-cloning. As shown in Fig. 7(1)(b), the average rewards are similar (around 0.01) for all agents. These agents learned how to navigate to the target object, but could not learn how to avoid the wrong object as shown in Fig. 8(1)(b).

Without guidance, the agent was unable to learn to solve the task properly as shown in Fig. 7(1)(c) and Fig. 8(1)(c). Compared to the results with guidance, these experiments show that guidance does aid in the agents' learning process.

4.5 Study 2: Examining the impact of the level of interaction in multimodal learning

The purpose of study2 is to examine the effect of interaction in a multimodal environment and observe how the modality of the agent affected its performance and the critical period. To do so, we modified the object-finding task in Study 1 to an audio-visual task. Specifically, each object has its unique sound and makes this sound during the whole episode. Since the audio observation of the agent is 3D-spatialized, the agent can recognize the direction of each sound source. Thus, we expect that this task would be easier than Study 1, but still challenging since the agent has to learn the relation between the object and the sound. Other detailed setups about the learning process and guidance are identical to Study 1.

4.5.1 Results. Similar to Study 1, we find that there exists a noticeable difference in performance for guidance from helper-rewards, as shown in Fig. 7(2)(a). This shows that a critical regime also exists for guidance from helper-rewards in the multimodal object-finding task. However, the critical regime differed from Study 1: the 1M- and 2M-helper had much higher average reward (0.046 and 0.050) than 3M-(0.02), 4M-(0.02) agents. Fig. 8(2)(a) compares the trajectories of the agents, visually showing that the policy of the 1M- and 2M-helper performs better than the other agents. We observed that the agent majorly uses the audio to get the direction of the object, so it turns right/left even if the object is slightly left/right, which results in wiggling behavior.

The agents guided with behavior-cloning had only a negligible difference in performance as shown in Fig. 7(2)(b). These agents showed behavior similar to the agents in Study 1: they learned to navigate to the target object but didn't learn to avoid the wrong object as shown in Fig. 8(2)(b). However, the average reward was lower than in the unimodal environment. We think that because the target policy is only based on visual observation, so the audio observation worked as noise during behavior cloning.

Without guidance, the agent couldn't solve the task properly as shown in Fig. 7(2)(c) and Fig. 8(2)(c). These results show that the multi-modality of the agent didn't give rise to a new critical period for guidance methods without critical regime, but affected the time of critical period of the guidance with a critical regime.

4.6 Study 3: Transfer to a real-world dataset

In study1, we trained with three levels of guidance learning (Sparse-reward, helper-reward, and behavioral cloning) Fig. 2. This leads to three distinct sets of CNN and MLP weights from the trained models, and we can evaluate each of these sets of weights in another relevant task via transfer learning as seen in Fig. 3 (b).

We previously observed that the 2M- and 1M-helper models showed a noticeable difference in performance compared to the models with no critical period (3M-, 4M-helper) on both uni- and multimodal AI agents. If such a disparity in performance is valid, we believe that the models pre-trained with helper-reward guidance should show comparable results on real-world data input which better reflects the characteristics of toddlers' observations.

To verify the validity of our claims, we evaluated the transfer performance of image encoders (consisting of the CNN and the MLP) from 1M-, 2M-, 3M-, and 4M-helper models. Specifically, we fixed the parameters of the image encoders after pre-training and connected additional linear layers for image classification. Next, we used the EAVE dataset images, which was collected in the real world in accordance with the toddler's viewpoint, as input (size 84x84) with a total of 6 channels, consisting of the toddler agent's

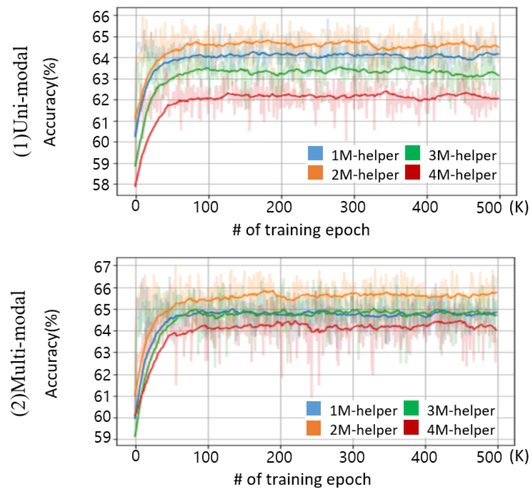


Figure 9: Transfer learning performance with the moderate guidance learning (helper-reward) using the EAVE dataset on unimodal (1) and multimodal (2) agents.

binocular vision with RGB channels each to the right and left side. We trained the model for 500 000 iterations with batch size 32.

As shown in Fig. 9, we found that the transferred 2M- (Uni:64.6/ Multi:65.8%) and 1M-helper (Uni:64/Multi:64.9%) models with a critical period achieved a better performance in terms of accuracy than the 3M- (Uni:63.3/multi:64.9%) and 4M-helper (Uni:62.1/multi:64.2%) models without a critical period on the EAVE dataset. The multimodal agent's set of CNN and MLP weights has shown a much higher average accuracy than the unimodal agent. Hence, we confirm that multimodal learning rather than unimodal can be helpful in efficiently learning the representation of objects during learning. Furthermore, similar to study1 (see Fig. 7 (1),(2)-(a)). we see a distinct difference in moderate guidance learning performance with a helper-reward and can accordingly rank the models in the same order where the models (2M,1M) with a critical period significantly outperform the ones without. Through this, we validate that the learning performance of toddler AI agents indeed depends on the critical period for the setting with moderate guidance learning.

5 DISCUSSIONS

Our paper has made three key contributions as a first step toward discovering the optimal guidance corresponding to critical periods in AI agents. First, to prove that RL-based humanoid AI agents also have critical periods like a toddler has, we found for the first time that AI agents also show specific critical periods (2M, 1M-helper) within the only moderate guidance learning setting (see Fig. 7 (1),(2)-(a)). We assume that 3M/4M frames are past the critical period, so the policy overfits more to the sparse reward, unable to reach the performance of the 2M/1M model. Behavioral Cloning merely imitates the mentor's trajectory, unable to generalize beyond the mentor-given policy. However, the helper reward is the optimal guidance among them.

Second, we collected the EAVE dataset, which contains images that imitate a toddler's point of view and its special characteristics and which, along with the developed virtual environment with a humanoid agent, allowed us to verify the qualitative notion of critical periods in a computational and cognitive way. Furthermore, we specifically designed the agent as a humanoid agent, mimicking the toddler's viewpoint's characteristics in VECA, so that a humanoid agent can closely learn and more efficiently observe the features of objects in the same way a toddler does. We believe that those human-like learning features enable us to test and validate the critical learning period more reliably.

Lastly, Our findings are especially significant for both Human-AI interaction domain and Human-in-the-loop RL, possibly used to determine the best time for human intervention in the training of AI agents. We developed and designed both uni- and multimodal agents, and confirm that the multimodal agent achieves a considerable improvement over every result of the unimodal agent, inspired by actual human cognitive development. The multimodal agent has a higher average reward and accuracy on object-finding task (Fig. 7 (2)-(a)) and transfer learning on the EAVE dataset (Fig. 9 (b)). We validate that multimodal learning at 2M&1M accompanying the critical period is more effective for learning useful representation compared to the unimodal approach.

In the future, we aim to analyze the critical period based on a more toddler-like RL algorithm deeper in the multi-task&meta-RL domain or develop an algorithm that automatically finds the critical learning period of the agent with additional verification studies. In this work, however, we employed a conventional RL algorithm with a light deep learning structure as a first step towards this goal. Also, testing the generalizability of the critical period concept requires further experiments. Since the concept is from humans, we speculate that physically grounded humanoid agents will more likely show critical periods here. We plan to study this in various environments including non-humanoid agent in the future.

6 CONCLUSION

Toddlers usually have a distinctive increase in performance during the critical period [34]. At this time, the increase in learning performance depends on the timing of mutual interaction. To investigate a similar effect in learning in AI agents, we studied three levels of mutual interaction learning (forms of guidance) in the virtual environment based on our formulation of toddler guidance learning terms in the RL framework. We validated our observations of the model's critical period on real-world observations (EAVE dataset). Corresponding to the critical period in toddlers, we observed a similar increase in performance using optimal guidance during the training of an AI agent. Specifically, we found that only moderate mutual interaction with a multimodal agent leads to a distinct increase in performance within the critical period and further affects the period during which the AI agent achieves remarkable performance compared to the unimodal agent.

ACKNOWLEDGMENTS

This work was partly supported by the IITP (2015-0-00310-SW.StarLab/16%, 2017-0-01772-VTT/16%, 2018-0-00622-RMI/16%, 2019-0-01371-BabyMind/20%) grants and the NRF of Korea (2021R1A2C10-10970/16%) grant funded by the Korean government, and the CARAI (UD190031RD/16%) grant funded by the DAPA and ADD.

REFERENCES

[1] Alessandro Achille, Matteo Rovere, and Stefano Soatto. 2018. Critical learning periods in deep networks. In *International Conference on Learning Representations*.

[2] Kai Arulkumaran, Marc Peter Deisenroth, Miles Brundage, and Anil Anthony Bharath. 2017. A brief survey of deep reinforcement learning. *arXiv preprint arXiv:1708.05866* (2017).

[3] Sven Bambach, David Crandall, Linda Smith, and Chen Yu. 2018. Toddler-inspired visual object learning. In *Advances in neural information processing systems*. 1201–1210.

[4] Yoshua Bengio, J. Louradour, Ronan Collobert, and J. Weston. 2009. Curriculum learning. In *ICML '09*.

[5] Stephen Billett. 2000. Guided learning at work. *Journal of Workplace learning* (2000).

[6] Michael S Brainard and Eric I Knudsen. 1998. Sensitive periods for visual calibration of the auditory space map in the barn owl optic tectum. *Journal of Neuroscience* 18, 10 (1998), 3929–3942.

[7] Stefan Braun, Daniel Neil, and Shih-Chii Liu. 2017. A curriculum learning method for improved noise robustness in automatic speech recognition. *2017 25th European Signal Processing Conference (EUSIPCO)* (2017), 548–552.

[8] Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. 2020. Soundspaces: Audio-visual navigation in 3d environments. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer.

[9] Xinlei Chen and Abhinav Gupta. 2015. Webly supervised learning of convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*. 1431–1439.

[10] Barry R Chiswick and Paul W Miller. 2008. A test of the critical period hypothesis for language learning. *Journal of multilingual and multicultural development* 29, 1 (2008), 16–29.

[11] John M Darley and Russell H Fazio. 1980. Expectancy confirmation processes arising in the social interaction sequence. *American Psychologist* 35, 10 (1980), 867.

[12] Finale Doshi-Velez. 2009. The infinite partially observable Markov decision process. *Advances in neural information processing systems* 22 (2009), 477–485.

[13] J. Elman. 1993. Learning and development in neural networks: the importance of starting small. *Cognition* 48 (1993), 71–99.

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

[15] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. 2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*. PMLR, 1861–1870.

[16] Takao K Hensch. 2004. Critical period regulation. *Annu. Rev. Neurosci.* 27 (2004), 549–579.

[17] George C Homans. 1974. Social behavior: Its elementary forms. (1974).

[18] Andrew Jesson, N. Guizard, Sina Hamidi Ghalehjegh, D. Goblot, F. Soudan, and Nicolas Chapados. 2017. CASED: Curriculum Adaptive Sampling for Extreme Data Imbalance. *ArXiv* abs/1807.10819 (2017).

[19] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[20] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. 2017. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474* (2017).

[21] A Kral. 2013. Auditory critical periods: a review from system's perspective. *Neuroscience* 247 (2013), 117–133.

[22] Stephen D Krashen. 1973. Lateralization, language learning, and the critical period: Some new evidence. *Language learning* 23, 1 (1973), 63–74.

[23] Lisa Lee, Ben Eysenbach, Russ R Salakhutdinov, Shixiang Shane Gu, and Chelsea Finn. 2020. Weakly-Supervised Reinforcement Learning for Controllable Behavior. *Advances in Neural Information Processing Systems* 33 (2020).

[24] L. Lin, Keze Wang, Deyu Meng, W. Zuo, and Lei Zhang. 2018. Active Self-Paced Learning for Cost-Effective and Progressive Face Identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40 (2018), 7–19.

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

[26] Tyler Tian Lu. 2009. *Fundamental limitations of semi-supervised learning*. Master's thesis. University of Waterloo.

[27] A. Ng, D. Harada, and Stuart J. Russell. 1999. Policy Invariance Under Reward Transformations: Theory and Application to Reward Shaping. In *ICML*.

[28] Sheryl L Olson, John E Bates, and Kathryn Bayles. 1990. Early antecedents of childhood impulsivity: The role of parent-child interaction, cognitive competence, and temperament. *Journal of abnormal child psychology* 18, 3 (1990), 317–334.

[29] Kwanyoung Park, Jeong Heo, and Youngki Lee. 2020. VECA: A VR Toolkit for Training and Testing Cognitive Agents. https://github.com/GGOSinon/VECA.

[30] Kwanyoung Park, Junseok Park, Hyunseok Oh, Byoung-Tak Zhang, and Youngki Lee. 2021. Learning task-agnostic representation via toddler-inspired learning. arXiv:2101.11221 [cs.AI]

[31] Anastasia Pentina, Viktoriia Sharmanska, and Christoph H Lampert. 2015. Curriculum learning of multiple tasks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5492–5500.

[32] Te Pi, Xi Li, Zhongfei Zhang, Deyu Meng, Fei Wu, Jun Xiao, and Yueting Zhuang. 2016. Self-Paced Boost Learning for Classification. In *IJCAI*.

[33] Jean Piaget and Margaret Cook. 1952. *The origins of intelligence in children*. Vol. 8. International Universities Press New York.

[34] Ann L Robson. 2002. Critical/sensitive periods. *Child Develop-ment* (2002), 101–103.

[35] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. 2019. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9339–9347.

[36] Roger C Schank. 1972. Conceptual dependency: A theory of natural language understanding. *Cognitive psychology* 3, 4 (1972), 552–631.

[37] Yangyang Shi, Martha Larson, and Catholijn M Jonker. 2015. Recurrent neural network language model adaptation with curriculum learning. *Computer Speech & Language* 33, 1 (2015), 136–154.

[38] Yoav Shoham, Rob Powers, and Trond Grenager. 2003. Multi-agent reinforcement learning: a critical survey. (2003).

[39] Abhinav Shrivastava, A. Gupta, and Ross B. Girshick. 2016. Training Region-Based Object Detectors with Online Hard Example Mining. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), 761–769.

[40] David Michael Singleton and Zsolt Lengyel. 1995. *The age factor in second language acquisition: A critical look at the critical period hypothesis*. Multilingual Matters.

[41] Catherine E Snow and Marian Hoefnagel-Höhle. 1978. The critical period for language acquisition: Evidence from second language learning. *Child development* (1978), 1114–1128.

[42] Petru Soviany, Claudiu Ardei, Radu Tudor Ionescu, and M. Leordeanu. 2020. Image Difficulty Curriculum for Generative Adversarial Networks (CuGAN). *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)* (2020), 3452–3461.

[43] Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and N. Sebe. 2021. Curriculum Learning: A Survey. *ArXiv* abs/2101.10382 (2021).

[44] Valentin I Spitkovsky, Hiyan Alshawi, and Daniel Jurafsky. 2009. Baby Steps: How "Less is More" in unsupervised dependency parsing. (2009).

[45] Sebastian P Suggate, Elizabeth A Schaughency, and Elaine Reese. 2013. Children learning to read later catch up to children reading earlier. *Early Childhood Research Quarterly* 28, 1 (2013), 33–48.

[46] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction*. MIT press.

[47] Piotr Sztompka. 2002. Socjologia. *Analiza społeczeństwa, Znak, Kraków* (2002), 324.

[48] Sebastian Thrun and Lorien Pratt. 2012. *Learning to learn*. Springer Science & Business Media.

[49] Satoshi Tsutsui, Arjun Chandrasekaran, Md Alimoor Reza, David Crandall, and Chen Yu. 2020. A Computational Model of Early Word Learning from the Infant's Point of View. *arXiv preprint arXiv:2006.02802* (2020).

[50] Radu Tudor Ionescu, Bogdan Alexe, Marius Leordeanu, Marius Popescu, Dim P Papadopoulos, and Vittorio Ferrari. 2016. How hard can it be? Estimating the difficulty of visual search in an image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2157–2166.

[51] Matteo Turchetta, Andrey Kolobov, Shital Shah, Andreas Krause, and Alekh Agarwal. 2020. Safe reinforcement learning via curriculum induction. *arXiv preprint arXiv:2006.12136* (2020).

[52] Alan M Turing. 2009. Computing machinery and intelligence. In *Parsing the Turing Test*. Springer, 23–65.

[53] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. 2016. A survey of transfer learning. *Journal of Big data* 3, 1 (2016), 9.

[54] Fei Xia, Amir R Zamir, Zhiyang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. 2018. Gibson env: Real-world perception for embodied agents. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 9068–9079.

[55] Y. Zhang, P. David, and Boqing Gong. 2017. Curriculum Domain Adaptation for Semantic Segmentation of Urban Scenes. *2017 IEEE International Conference on Computer Vision (ICCV)* (2017), 2039–2049.

[56] Jennifer N Zosh, Emily J Hopkins, Hanne Jensen, Claire Liu, Dave Neale, Kathy Hirsh-Pasek, S Lynneth Solis, and David Whitebread. 2017. *Learning through play: a review of the evidence*.