Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Semantic Visual Representations Emerge from Embodied Self-Supervised Learning

Arthur Aubret

aubret@fias.uni-frankfurt.de

Frankfurt Institute for Advanced Studies

Article

Keywords:

Posted Date: July 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10140180/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

Semantic Visual Representations Emerge from Embodied Self-Supervised Learning

Anonymous

June 24, 2026

Abstract

The principles underpinning the development of human visual perception remain poorly understood. Current computational models of human visual learning still fail at capturing humans' rich semantic recognition abilities. Here, we introduce Embodied Self-Supervised Learning (E-SSL), a class of computational models that harness the rich statistical structure of concurrent visual and motor signals during embodied interaction with the environment for visual representation learning. E-SSL relies on two prominent unsupervised learning principles: temporal consistency and sensorimotor contingencies. When E-SSL is applied to large-scale recordings of human egocentric videos, eye movements and body movements during natural behavior, it produces high-level semantic visual representations in the absence of additional language input or explicit supervision. The learned representations substantially outperform those of previous models of human vision on a range of downstream tasks. We further find that representations learned via E-SSL become more aligned with those of visual cortex of pre-verbal children than those of an established self-supervised model of biological vision. Finally, E-SSL predicts that the onset of locomotion substantially enhances visual perception and suggests that the dichotomy between the ventral and dorsal visual streams may have its origin in the two complementary learning principles of E-SSL. Together, these results shed light on the learning principles and experiences that jointly drive the development of human vision.

1 Introduction

Within a few years, humans acquire powerful visual abilities with minimal supervision, such as semantic object recognition. The principles that drive this development remain poorly understood, however. Bio-inspired models of learning fall short of reproducing human visual performance [1–5], whereas high-performing artificial vision systems depend on training data and learning mechanisms that drastically differ from those of humans [6–8]. For instance, common self-supervised learning (SSL) methods depend on artificial image transformations unlikely to be implemented by the brain, such as aggressive cropping or masking, to generate learning signals [7]. It is currently unclear how this gap between human-level performance and human-like learning principles can be bridged [5].

An important difference between learning in high-performing artificial vision models and humans is that humans actively explore their environment through eye and body movements. This allows them to exploit the statistical structure of concurrent sensory and motor signals during their interactions with the world when acquiring visual representations. This observation has laid the groundwork for two prominent theories of visual learning. The theory of temporal consistency (TC) posits that mammalian vision systems learn similar representations for visual inputs that occur close in time [9–11]. In natural settings, TC has been argued to support the formation of semantic object representations. For example, rotating an object can allow the visual system to associate different views of the same object [12–14]. A complementary account is provided by the theory of sensorimotor contingencies (SMC), which proposes that visual learning arises from understanding the relationship between actions and the visual consequences they produce [15, 16]. For instance, rotating a mug may reveal its handle, making the handle's presence and shape informative about the applied rotation. Learning this SMC facilitates the recognition of mugs, as the presence of a handle is a defining feature of the concept of "mug". Although both TC [17, 18] and SMC [19–21] are supported by neurobiology, existing computational investigations are based on simplified datasets and additional artificial learning signals [12, 13, 22–24]. Thus, it is presently unclear if these principles suffice to drive the development of human-like visual representations.

In this work, we test to what extent temporal consistency and sensorimotor contingencies are sufficient to support the emergence of human-like visual representations from human experience. For this purpose, we introduce Embodied SSL (E-SSL), a powerful SSL framework designed to model human visual learning through TC and SMC. E-SSL receives gaze-centered visual input (cf. Figure 1A) and learns from the temporal structure of visual observations and associated movements (cf. Figure 1B).

Concretely, we simulate the daily experience of humans with Nymeria [25], a dataset comprising 300 hours of naturalistic videos recorded with head-cameras. Nymeria also contains gaze locations from eye tracking as well as the trajectory of the participants' head, from which we can derive participants' eye and head movements. We evaluate the resulting visual representations across a diverse set of visual tasks mastered by humans such as object categorization or relative object pose estimation (cf. Figure 1C-D-E).

Our results demonstrate that utilizing the principles of TC and SMC during learning from human experience yields powerful semantic representations, substantially narrowing the performance gap between existing bio-inspired learning models [2, 3, 26, 27] and humans. The learned representations support a wide range of visual tasks and exhibit properties similar to those observed in pre-verbal children. Moreover, our experiments predict that sensorimotor experiences following the onset of locomotion improve children's ability to perceive the semantic and geometric structure of their environment. Finally, we find that the principles of temporal consistency and sensorimotor contingencies naturally give rise to a functional organization reminiscent of the ventral and dorsal visual pathways. Together, these findings suggest that temporal consistency and sensorimotor contingencies constitute fundamental principles underlying human visual development.

2 Results

2.1 E-SSL with human experience acquires powerful semantic visual representations

We first want to assess whether the principles of TC and SMC, as embodied by E-SSL, together with human experience are sufficient for the emergence of object recognition abilities. To compare E-SSL with other models, we use the ImageNet-1k and Cifar10, two classical benchmarks for evaluating category recognition in computer vision and computational neuroscience. We follow the evaluation protocol illustrated in Figure 1C-D. In Figure 2A, we observe that the complete version of E-SSL can achieve 71.04% of Top-5 accuracy on ImageNet-1k, outperforming Prednet, an established model of biological learning [3], by up to 58.73%. When controlling the training data (here Kitti [35]), a drastically ablated version of E-SSL trained solely with TC surpasses Prednet by 1.72%. In Cifar10, E-SSL yields a similar increase in object recognition performance and surpasses state-of-the-art bio-inspired models (LPL [2], SH [26] and SCFF [27]), despite the fact that these models were evaluated on in-distribution images (Cifar10 pre-training), vs. out-of-distribution images for E-SSL (Nymeria pre-training).

E-SSL does not process any semantic information during training, making it comparable to pre-verbal infants. Since infants' performance is unavailable and difficult to obtain with classical vision benchmarks, we compare E-SSL to an upper-bound performance represented by human adults who have extended language-enhanced visual experience. In Figure 2A, E-SSL is only $\sim$ 20 and $\sim$ 10 percentage points lower than current estimates of adult performance on ImageNet-1k [30] and Cifar10 [31], respectively. In general, our results demonstrate that semantic visual representations emerge from TC and SMC learning with human experience, despite using no form of semantic information during pre-training.

2.2 E-SSL reproduces visual abilities similar to those of pre-verbal children

Next, we test whether E-SSL develops visual properties similar to those of children with limited language experience. We consider complete E-SSL (TC+SMC), E-SSL without sensorimotor contingencies (TC) and E-SSL without temporal consistency (SMC). We first verify whether the model can form infancy-level attribute prototypes (cf. Figure 3A). In Figure 3B, we observe that E-SSL forms reliable color, texture and size prototypes. In contrast, considered baselines heterogeneously fail in at least one of the three tasks: a random baseline with the same architectural inductive biases ("Random") does not build strong size or texture prototypes, and an established model of the visual cortex (Prednet) fails to build color or size prototypes. Removing sensorimotor learning (TC) does not significiantly impair the quality of the prototypes, whereas removing temporal consistency (SMC) impairs color recognition. Other E-SSL models that implement the same embodied learning principles with different loss functions and architectures show similar results (cf. Appendix C). These results suggest that TC is sufficient to support the emergence of visual properties similar to those observed in infancy

If E-SSL models infancy-level properties, it may also develop visual properties achieved by young toddlers. As presented in Figure 3A, we investigate whether the model can recognize shape caricatures of object categories, a skill that begins to emerge at around 18 months of age in human development [38]. In Figure 3B, E-SSL correctly matches a prototype to the corresponding natural image $\sim$ 65% of the time. In contrast, "Random" and "Prednet" achieve chance-level performance. Furthermore, ablating sensorimotor learning (TC) yields a small decrease in caricature recognition, though consistent across E-SSL models (cf. section C). This shows that TC and SMC promote the extraction of features that

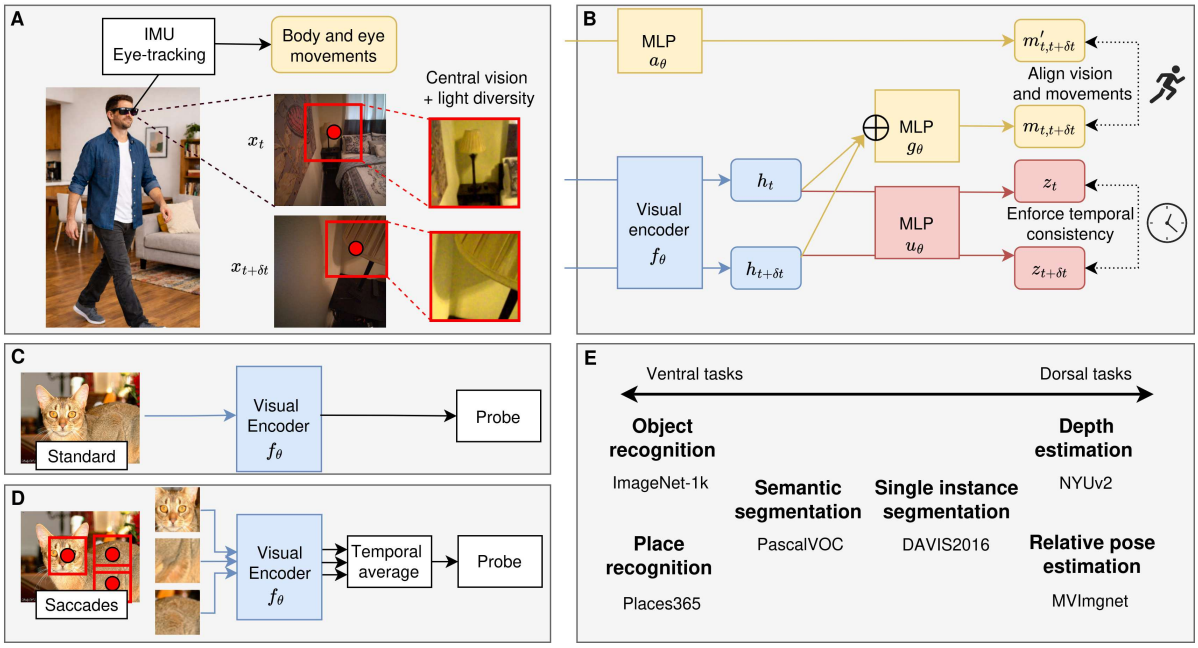


Figure 1: Overview of E-SSL and the evaluation protocol. **A)** We randomly sample pairs of temporally adjacent images x_t and $x_{t+\delta t}$ separated by $\delta t \sim \mathcal{U}(-\Delta T, \Delta T)$ seconds from Nymeria, a dataset of human egocentric data captured with ARIA glasses. For each pair, we extract a 224×224 crop centered on the participant's gaze location. To simulate diverse lighting conditions that are consistent across time (bulb, sunset etc.), we apply the same color transformation to both images. We additionally extract the head and eye movements executed between the two time points, derived from IMU and eye-tracking measurements. **B)** The generic computational model of learning consists of two loss functions based on SSL. The first promotes similar embeddings z_t and $z_{t+\delta t}$ for temporally adjacent visual inputs. The second one aligns a parameterized movement $m'_{t,t+\delta t}$ with an embedding of the visual transition $m_{t,t+\Delta t}$ obtained by concatenating the visual representations before and after the movement. We implement E-SSL with MoCoV3[28], SimCLR[7] or VICReg[29], which we refer to as E-MoCoV3, E-SimCLR and E-VICReg. For E-MoCoV3, the model also incorporates momentum neural networks and predictors [28]. Our main experiments are all done with E-MoCoV3, but all experimental conclusions are replicated with the two other E-SSL models Appendix C. **C)** For object recognition, we assess the quality of the visual encoder by training a probe to solve downstream visual recognition tasks, following standard conventions. **D)** During classification, humans may have performed different eye movements, image comparisons [30, 31] etc. To account for this, we emulate multiple eye fixations by sampling N image portions at a fixed scale ratio (0.15), encode them with the visual model, and average their representations, following [32]. For each image, we sample one training fixation and $N = 16$ evaluation fixations. **E)** Extended set of visual tasks and their respective dataset, arranged based on whether they are related to the ventral or dorsal stream in the visual cortex [33, 34].

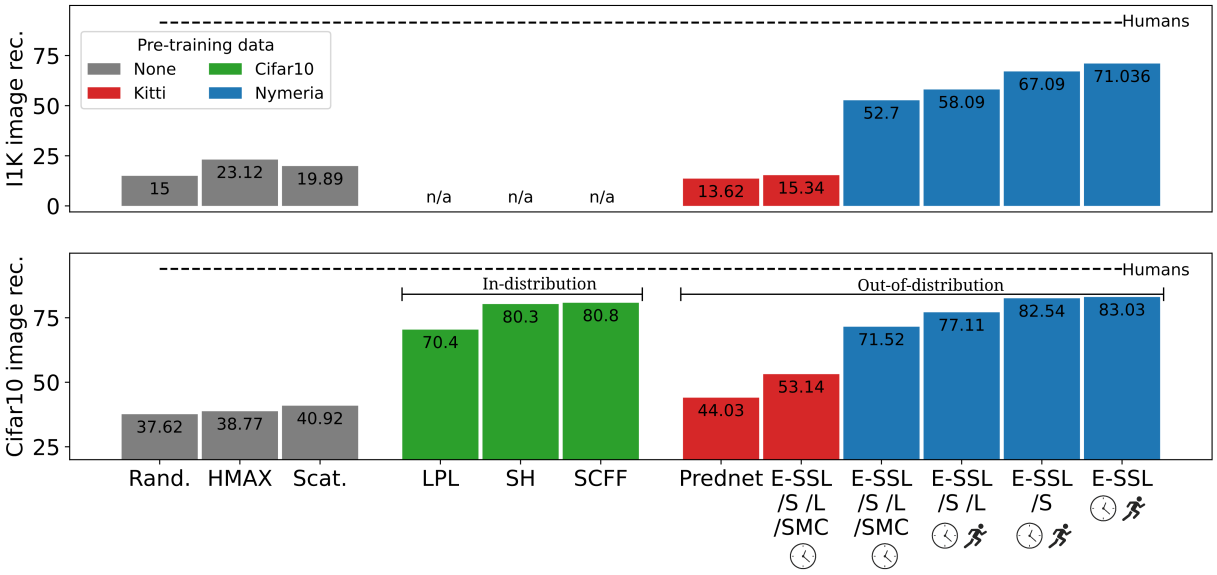


Figure 2: Imagenet-1k Top-5 accuracy and Cifar10 Top-1 accuracy for bio-inspired models of object recognition, E-SSL and humans. As baseline models, we consider untrained models of the visual cortex (HMAX [36], Scattering [37]), recent state-of-the-art bio-inspired models (LPL [2], SH [26] and SCFF [27]) trained on static data, and Prednet [3], an established model of visual learning. We also consider E-SSL without saccades (/S), without light augmentations (/L) and without learning sensorimotor contingencies (/SMC). We are not aware of other bio-inspired unsupervised models that report Imagenet-1k accuracy. We train only E-SSL /S /L /SMC with Kitti to align training assumptions to that of Prednet.

support generalization over categories of shapes, similar to young toddlers.

Recognizing shape caricatures requires complex shape processing. Toddlers progressively and preferentially recognize objects using shapes versus textures [39, 40]. Here, we hypothesize that texture statistics and shape features are stable and unstable over time, respectively. By definition, TC prioritizes the extraction of features that are stable over time, whereas SMC favors features related to (fast) movements. We therefore ask whether TC and SMC yield an opposite specialization towards texture and shape encoding. To this end, we use a developmentally-inspired cue-conflict evaluation protocol [41, 42] illustrated in Figure 3A. In Figure 3B, the results show that TC favors texture encoding, while adding SMC boosts shape encoding. The shape bias of untrained networks has already been reported before [42] and tends to decrease during training in most models. Overall, this indicates that SMC can favor the emergence of a visual shape bias.

Since E-SSL captures similar semantic visual properties to that of 1-year-old children, it may capture the visual geometry of semantic representations in the infant brain. To test this, we compare the structure of model representations to fMRI data recorded for 24 images in the ventral visual cortex (VVC), a region associated with semantic processing [48]. Data were recorded for 2-month-old infants, 9-month-old infants, and adults. In Figure 3C, the "Prednet" model fails to reproduce the representational structure observed in infants, whereas E-SSL captures meaningful patterns, notably for animate objects (top-left square block). This is quantitatively confirmed in Figure 3D, where E-SSL has higher correlations with VVC representations in 9-month-old infants ($\rho = 0.289$) compared to the Random model and Prednet. Notably, E-SSL exhibits a particularly strong alignment with the brain representations of 9-month-old infants, compared to adults and 2-month-old infants. Taken together, these results show that TC and SMC yield visual representations that capture important aspects of vision in pre-verbal children.

2.3 Visuomotor alignment from self-induced locomotion promotes visual learning

Since E-SSL captures important aspects of children's visual development, we wondered how specific visuomotor experiences may shape learning. In particular, we hypothesize that the onset of self-induced locomotion provides to the human brain an unprecedented opportunity to relate efference copies of motor commands and their visual effects. To investigate this hypothesis, we first simulate different embodied conditions, ranging from an "embodied" agent that leverages aligned visuomotor signals to an externally manipulated "puppet" ("Puppet" and "Puppet swap") in which motor signals are uncorrelated to the visual effects. We see in Figure 4A that the "embodied" agents consistently outperform "Puppet" and "Puppet swap" on all visual tasks. Interestingly, removing eye movement signals ("Puppet w/o

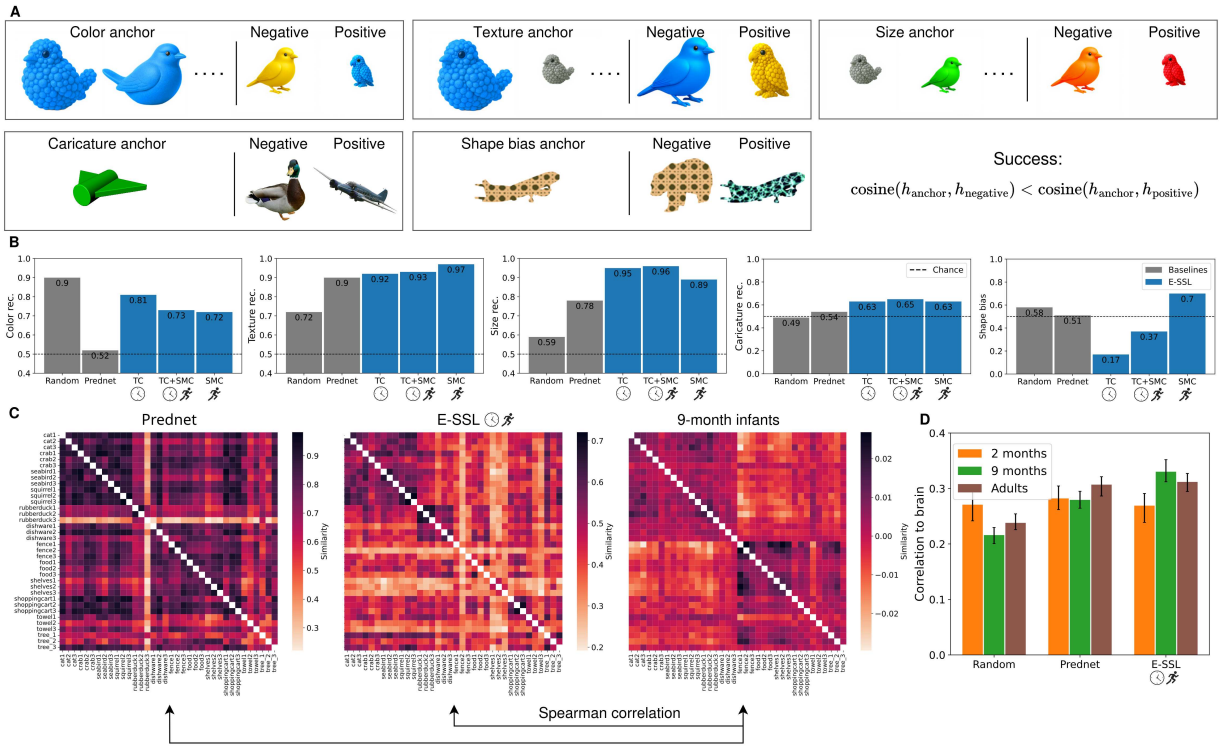


Figure 3: **A)** Developmentally-relevant evaluation protocol for few-shot recognition. For all given triplets, We compare the representation of an anchor to that of two controlled stimuli using the cosine similarity. We test color [43], texture [44], and size recognition [45] by emulating preferential-looking paradigms [43], following [46]. We compute attribute prototypes (anchors) by averaging the representations of images that share a given attribute (*e.g* "blue") within a semantic category. Using cosine similarity, the representation of the prototype anchor is then compared to the representation of a held-out image that shares the same attribute or an image with a different value of that attribute (*e.g.* "yellow"). Both images belong to the same category as the prototype anchor. To emulate shape caricature recognition [38], we leverage a previously constructed set of simplified shapes [47] following [42]. Here, the anchor is an image representation of a single shape caricature. To quantify the shape bias of a representation, we assess whether an image for which a given shape and texture (anchor) is closer to an image with the same shape/different texture versus to an image with a different shape/same texture, in the representation space [41, 42]. **B)** We compute the success rate of E-SSL and baseline models across triplets, for each few-shot recognition test. **C)** Representation Similarity Matrix (RSM) of "No Pretraining" model, E-SSL, and fMRI data from the ventral visual cortex of 9-month-old infants (ROIs FG1/VO1, FG3/PHC, FG2 and FG4). To compute similarity scores, we compute spearman scores between pairs of image representations. **D)** Representational Similiarity Analysis (RSA) between models' representations and fMRI recordings from 2-month-old infants, 9-month-old infants, and adults [48]. We compute the Spearman correlation score between RSM upper triangles (diagonal excluded) and normalize the score by the split-halves noise ceiling (0.90, 0.87, 0.76 for adults, 9-month-old and 2-month-old infants, respectively). Error bars represent 95% confidence intervals, calculated by bootstrap resampling (50 bootstraps) across subject/run pairs, with group average RSM. We retain only the highest score across layers of the ResNet50 (Layer 1-4 and Average Pooling), which is obtained by the last layer. We refer to [48] for more details about human fMRI data.

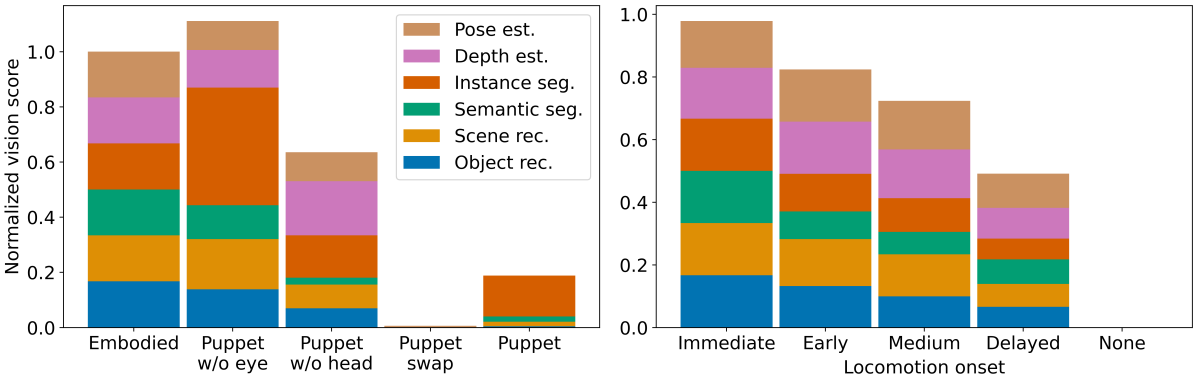


Figure 4: **A)** Normalized performance of E-SSL when removing eye or head movements from parameterized movements, or randomly swapping parameterized movements m across the dataset. For each of the six downstream tasks, we normalize the performance such that $\frac{1}{6}$ is the best performance of E-SSL and 0 is the worst performance for a given SSL method. We set eye and head signals to zero for puppets w/o eye and w/o head, respectively. For Puppet "swap", we switch the motor signals associated with a pair of visuals inputs with that of a randomly sampled pair in the dataset. "Puppet" is E-SSL without SMC. **B)** Normalized performance similar to A), while varying the epoch at which the model has access to the head movement associated to a visual transition. Prior to that, head movements are generated with Gaussian noise. "Immediate", "Early", "Medium", "Delayed", "None" refer to epochs 0, 5, 10, 15, never (over a total of 20 training epochs), respectively.

eye") does not significantly impede the performance. This may stem from the difficulty of inferring eye movement signals from visual transitions derived from extended body movements. In contrast, discarding head movement signals ("Puppet w/o head") hinders visual learning, even though the models still outperform the full "Puppet" baseline (no motor signals). This indicates that the alignment between the motor signals of the head and the visual experience plays an important role in visual learning.

The onset of self-induced locomotion is an important driver of alignment between head movement signals and visual transitions. This raises the question of whether delayed visuomotor alignment, induced by delayed locomotion development, may impede visual development. To answer this question, we varied the time point at which head movement signals become aligned with visual transitions, during training. As observed in Figure 4A, the earlier the alignment, the better the final performance of the model on all visual tasks, on average. We also notice that delayed alignment fills an important part of the performance gap relative to "Immediate" locomotion, though not entirely. Together, these results show that early access to visuomotor regularities induced by self-locomotion boosts visual representation learning.

2.4 E-SSL induces ventral and dorsal stream-like specializations

We found that the TC and SMC objectives preferentially promote the representation of, respectively, texture and shape information: hence, they should also support different classes of visual tasks. In particular, the human visual cortex progressively separates visual processing into ventral and dorsal streams [33, 49]. Functionally, ventral areas support semantic perception, while the dorsal stream processes visual information required for embodied actions such as reaching [49]. Therefore, we explore whether the learning principles of TC and SMC promote features related to visual or dorsal areas, respectively. To do so, we vary the importance of SMC during training with E-SSL. In Figure 5A, we first observe that considering SMC with medium importance ($\alpha_{\text{SMC}} = 0.1$) consistently improves over TC alone ($\alpha_{\text{SMC}} = 0$) across both kind of visual tasks. This indicates that learning contingencies between egocentric movements and visual changes supports the development of general-purpose visual representations. However, the sweet spot of α_{SMC} varies across tasks. Ventral-related tasks perform best with intermediate values ($\alpha_{\text{SMC}} \in [0.1, 0.3]$), whereas tasks associated with the dorsal pathway [33] benefit from stronger SMC learning ($\alpha_{\text{SMC}} = 1$). This pattern is consistent with all three E-SSL methods (cf. Appendix C). This demonstrates that TC and SMC preferentially promote neural representations aligned with human ventral and dorsal visual functions, respectively.

Our experiments show that TC and SMC induce a functional specialization towards ventral and dorsal streams. Since dorsal processing is tuned to faster temporal dynamics than the ventral pathway [50–52], varying our hyper-parameter ΔT, which governs the timescale for learning TC and SMC, may analogously trade off performance between dorsal- and ventral-related tasks. As seen in Figure 4B,

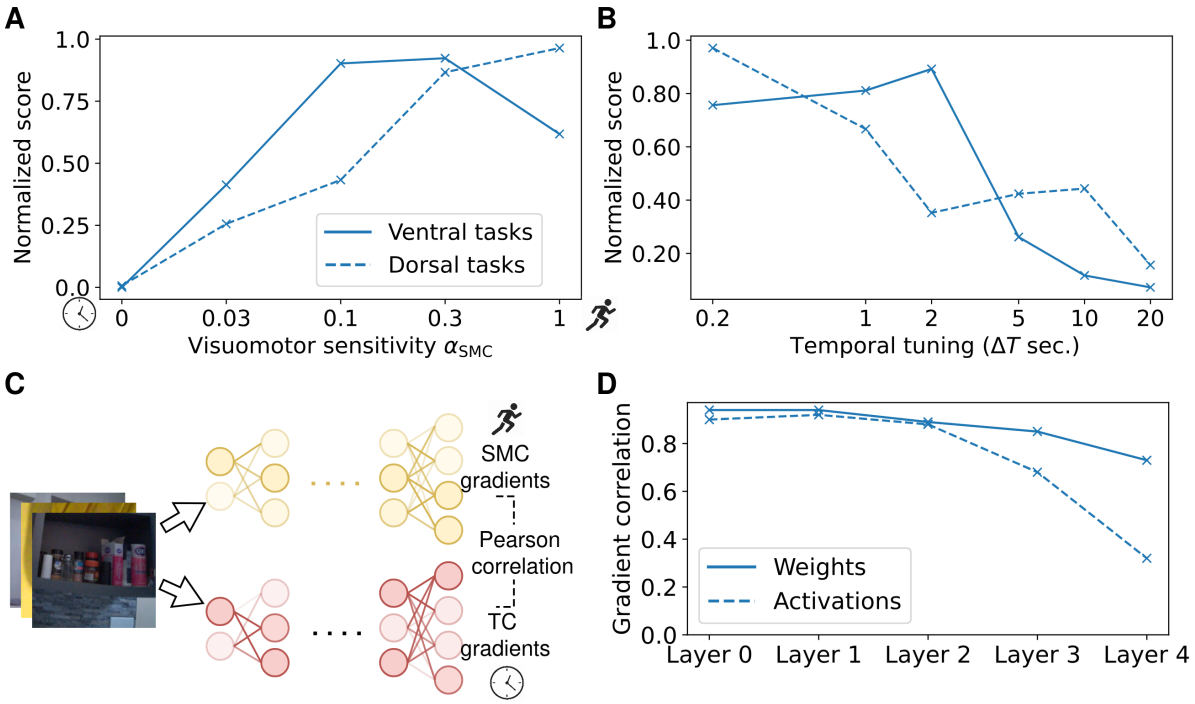


Figure 5: Downstream performance of E-SSL on **A)** ventral-related and dorsal-related visual tasks. We vary the importance of SMC learning by varying the hyper-parameter α_{SMC} on a logarithmic scale. $\alpha_{\text{SMC}} = 0$ denotes the absence of SMC while $\alpha_{\text{SMC}} = 1$ means equal contribution from TC and SMC. We normalize all scores such that 0 and 1 denote the worst and best performance, respectively. For ventral-related tasks, we average the normalized performance over ImageNet-1k, Places365 and PascalVOC. For Dorsal-related tasks, we average the normalized performance over Davis, NYUv2 and MVImgNet. **B** Similar to A, we plot results of E-SSL while varying the hyper-parameter ΔT on a logarithmic scale, which defines the temporal range for aligning frames and movements. **C)** To compare the pathways recruited by TC and SMC, we compute the correlation between TC and SMC gradients. Using 1,280 randomly sampled training pairs of images, we group the gradients for each ResNet50 layer and compute the Pearson correlation score between the TC gradient groups and their analog SMC gradient groups. We replicate the experiment with gradients of neurons (after ReLU activation). **D)** Gradient correlation between TC and SMC.

E-SSL with short timescales ($\Delta T = 0.2\,\mathrm{s}$) achieves strong performance for dorsal-related tasks. In contrast, ventral-related tasks (such as object and scene recognition) benefit from intermediate movements ($\Delta T = 2\,\mathrm{s}$). These results suggest that learning TC and SMC over short timescales supports dorsal-related tasks, while focusing on intermediate timescales promotes semantic perception.

Next, we investigate whether this functional matching originates from the learned vision models mirroring the neuronal organization of the visual cortex into ventral and dorsal streams. To characterize the pathway recruited by TC or SMC, we use a gradient-based approach to determine which neuron and weights are involved in optimizing TC or SMC, as illustrated in Figure 5C. In Figure 5D, the neurons and weights associated with each optimization objective progressively diverge in deeper layers of the network. Notably, this pattern mirrors the increasing anatomical separation between ventral and dorsal pathways in the visual cortex. This demonstrates that the learning principles of TC and SMC naturally give rise to an organizational structure resembling the ventral and dorsal streams.

3 Conclusion

How do humans learn to visually perceive their environment? A key challenge for addressing this long-standing question lies in establishing links between neural mechanisms, human experience and emerging visual abilities. In this work, we introduced E-SSL, a self-supervised learning framework that exclusively models the well-supported biological learning principles of temporal consistency (TC) and sensorimotor contingency (SMC). Three instances of E-SSL were trained on human egocentric data that encompass first-person naturalistic video recordings as well as eye and head movement signals. After training, the models were probed and tested on a diverse range of visual tasks. This framework provides a computational approach to studying how biologically grounded learning principles may support the emergence of human visual perception.

Our results demonstrate that TC and SMC yield powerful semantic visual representations. E-SSL substantially narrows the gap between established neuroscience models and human perception, supporting semantic visual perception, and capturing key visual properties emerging in young toddlers. Moreover, TC and SMC naturally give rise to a neural and functional organization reminiscent of the ventral and dorsal pathways of the visual cortex. Our analysis indicates that the visuomotor alignment induced by self-locomotion plays a substantial role in this development. These findings support the view that TC and SMC constitute fundamental principles underlying human visual learning.

These findings have broader implications for developmental research. E-SSL may enable researchers to rigorously generate testable predictions about how variations in human experience shape visual development [53–55]. Many computational models used in developmental psychology rely on learning principles that differ from those of humans, neglecting the temporal or sensorimotor structure of human visual experience. This limits their ability to capture developmental trajectories. For instance, the learning mechanisms by which self-locomotion promotes distance and spatial cognition [56, 57] remain unclear. By systematically controlling visuomotor alignment during training, our study suggests that developmental improvements following locomotion onset may arise from learning contingencies between internal efference copies of motor signals and their visual effects. Moreover, our results predict that self-induced locomotion may facilitate not only spatial abilities but also semantic recognition, extending beyond the restricted set of tasks studied so far [58].

E-SSL could also open avenues for elucidating the neuroanatomical development of the visual system. In particular, our experiments contribute to the long-standing debate about the extent to which the functional structure of the cortex stems from innate factors versus experience [59, 60]. When training a classical ResNet50 from scratch, we observed that neurons and weights in later layers tended to specialize for either the TC or SMC learning objective. This result suggests that functional specialization resembling dorsal and ventral pathways may emerge from general unsupervised learning dynamics rather than being rigidly specified by innate architectural constraints [61].

Furthermore, E-SSL offers a bio-inspired alternative to standard approaches of machine learning. Current machine learning models struggle to fully capture the representational structure of human perception [62–64]. In addition, they need far more diverse data than humans and remain strongly misaligned in object shape and relational processing [42, 65–67]. Recent research suggests that the latter may stem from the absence of dorsal-related visual learning in these models [68, 69]. Our experiments provide mixed evidence for this hypothesis: we found that SMC does promote shape-based object encoding, but only to some extent this supports object recognition. Nonetheless, we focused on human head and eye movements. Future work should incorporate richer sensorimotor signals, such as hand-manipulation behaviors and action selection, which have been argued to play an important role for human learning [70].

Despite the progress achieved by E-SSL in modeling human visual learning, a substantial gap remains between E-SSL and humans in object recognition performance. This limitation is also reflected in the visual abilities acquired by the model: E-SSL only begins to recognize abstract shapes composed of a few geometric parts, and the best-performing models (for object recognition) remain strongly biased toward encoding textures rather than object shape. In contrast, 3-year-old toddlers can readily

recognize abstract shapes and tend to generalize object categories based on shape rather than texture [38–40].

Several additional factors may contribute to narrowing the gap between E-SSL and toddlers. Although our work focuses on unsupervised learning, language and social interactions play an important role in refining visual representations [71–73]. In addition, recent research suggests that children's early visual experience, compared to the adult experience used in the present work, may be particularly suited to visual learning [14, 53, 55, 74–76]. In the present study, we restricted our experiments to adult data as there is no open dataset of children's egocentric videos that comprise motor signals, which are indispensable to our aim of testing the importance of SMC in visual learning. Another possibility is that a more mechanistic model of brain computations may be required to close the remaining gap, though the current panorama of vision models rather reflects a trade-off between accurately modeling brain learning mechanisms and modeling human object recognition [2, 5, 26]. In this context, E-SSL provides a simple yet powerful framework for investigating the principles that underpin humans' impressive visual development.

4 Method

4.1 Training data

Nymeria[25] is by far the largest dataset of human movements with 300 hours of daily activities distributed into 1,200 sequences. The participants were given high-level instructions such as "grab food in the cafeteria and eat on the patio", ensuring that they perform an activity. All participants wore ARIA glasses, recording raw egocentric view and movements. After preprocessing, the dataset comprises the gaze location of participants in their egocentric visual field and the closed-loop trajectory of participants' head. We extract crops of images centered on gaze locations at 5 fps, yielding a dataset of 6.1M images. When the boundaries of the crop area go beyond the field of view of the camera, we minimally shift the crop location to fit the field of view.

4.2 Embodied Self-Supervised Learning

Construction of image pairs. At each training iteration, E-SSL samples close-in-time images x_t and $x_{t+\delta t}$ where $\delta t \sim \mathcal{U}(-\Delta T, \Delta T)$ seconds. Following [23], E-SSL attenuates peripheral vision by applying a 224×224 crop centered on gaze locations c_t and $c_{t+\delta t}$. Humans redundantly interacts with the same environment exposed to different lighting conditions, such as sunset, artificial light, darkness etc. This lighting diversity is not reflected in the raw data. Therefore, we augment pairs of training images with time-consistent color jittering ($p = 0.8$) and grayscaling ($p = 0.2$), with the same parameters as in [7].

Derivation of eye and head movements. Alongside image sampling, we derive eye and body movements. For eye movements, we compute a two-dimensional translation in the image place, $c_{t,t+\delta t} = c_{t+\delta t} - c_t$ from the gaze locations c_t and $c_{t+\delta t}$. For body movements, we compute the relative rotation R_{rel} and translation p_{rel} of the camera as

$$R_{t,t+\delta t} = R_t^\top R_{t+\delta t},$$
$$p_{t,t+\delta t} = R_t^\top (p_{t+\delta t} - p_t),$$

where R_t is the camera-to-world rotation matrix and p_t represents the camera position in world coordinates at time t. $R_{t,t+\delta t}$ is converted into a quaternion $q_{t,t+\delta t}$. Finally, the full movement $m_{t,t+\delta t}$ is represented as a 9-dimensional vector $[c_{t+\delta t}, q_{t,t+\delta t}, p_{t,t+\delta t}]$.

Self-supervised learning models. At each training step, a mini batch of images and movements feed our neural networks following Figure 1B. This yields a minibatch of image embeddings $\mathcal{Z}$, vision-derived movements $\mathcal{M}$ and movements $\mathcal{M}'$. For generality, we instantiate E-SSL with three SSL approaches[7, 28, 29], namely E-MoCoV3 , E-SimCLR, E-VICReg. This choice aims to cover the diversity of SSL methods: VICReg uses a different loss function and similarity metric than SimCLR and MoCoV3. MoCoV3 stands out by incorporating predictors and momentum networks during training. We essentially apply their loss functions between two close-in-time visual representations (TC) as well as between the concatenate of visual representations and the movement they induce (SMC) [77].

E-SimCLR minimizes

$$\mathcal{L}_{\text{E-SimCLR}} = \frac{1}{|\mathcal{Z}|} \sum_{z_t \in \mathcal{Z}} -\log \frac{\exp\left(\text{sim}(z_t, z_{t+\delta t})/\tau\right)}{\sum_{z_k \in \mathcal{Z}} \mathbf{1}_{[k \neq t]} \exp\left(\text{sim}(z_t, z_k)/\tau\right)} \quad \textbf{TC}$$

$$+\alpha_{\text{SMC}} \frac{1}{|\mathcal{M} \cup \mathcal{M}'|} \sum_{m_{t,t+\delta T} \in \mathcal{M}} -\log \frac{\exp\left(\text{sim}(m_{t,t+\delta T}, m'_{t,t+\delta T})/\tau\right)}{\sum_{m_k \in \mathcal{M} \cup \mathcal{M}'} \mathbf{1}_{[k \neq \{t, t+\delta T\}]} \exp\left(\text{sim}(m_{t,t+\delta T}, m_k)/\tau\right)}, \quad \textbf{SMC}$$

where τ is the temperature hyper-parameter and sim is the cosine similarity.

E-MoCoV3 minimizes

$$\mathcal{L}_{\text{E-MoCoV3}} = \frac{1}{|\mathcal{Z}|} \sum_{z_t \in \mathcal{Z}} -\log \frac{\exp\left(\text{sim}(p_\theta(z_t), z^d_{t+\delta t})/\tau\right)}{\sum_{z_k \in \mathcal{Z}} \mathbf{1}_{[k \neq t]} \exp\left(\text{sim}(p_\theta(z_t), z^d_k)/\tau\right)} \quad \textbf{TC}$$

$$+ \alpha_{\text{SMC}} \frac{1}{|\mathcal{M} \cup \mathcal{M}'|} \sum_{m_{t,t+\delta T} \in \mathcal{M}} -\log \frac{\exp\left(\text{sim}(q_\theta(m_{t,t+\delta T}), m'^d_{t,t+\delta T})/\tau\right)}{\sum_{m_k \in \mathcal{M} \cup \mathcal{M}'} \mathbf{1}_{[k \neq \{t, t+\delta T\}]} \exp\left(\text{sim}(q_\theta(m_{t,t+\delta T}), m^d_k)/\tau\right)}, \quad \textbf{SMC}$$

where sim is the cosine similarity. p_θ and q_θ are predictor networks and the superscript d represents the embedding of a frozen neural architecture that mirrors the original one. The weights of this neural network are updated at each iteration according to an exponential moving average over the original network $\mu = 0.996\mu + (1 - 0.996)\theta$.

E-VICReg minimizes

$$\mathcal{L}_{\text{E-VICReg}} = \lambda \frac{1}{|\mathcal{Z}|} \sum_{z_t \in \mathcal{Z}} \|z_t - z_{t+\delta t}\|^2 + \mu \sum_{j=1}^{d} \max(0, 1 - \sqrt{\text{Var}(z^j) + \epsilon}) + \nu \frac{1}{d} \sum_{j \neq k} [C(\mathcal{Z})]_{j,k}^2 \quad \textbf{TC}$$

$$+ \alpha_{\text{SMC}} \Big[\lambda \frac{1}{|\mathcal{M} \cup \mathcal{M}'|} \sum_{m_{t,t+\delta t} \in \mathcal{M}} \|m_{t,t+\delta t} - m'_{t,t+\delta t}\|^2 + \mu \sum_{j=1}^{d} \max(0, 1 - \sqrt{\text{Var}(m^j) + \epsilon})$$

$$+ \nu \frac{1}{d} \sum_{j \neq k} [C(\mathcal{M} \cup \mathcal{M}')]_{j,k}^2 \Big] \quad \textbf{SMC}$$

where C is the covariance matrice between embeddings within a minibatch and $d = 2048$ the dimension size of embeddings. For all methods, we optimize $\mathcal{L}$ and its symmetric loss function, which can be derived by swapping $m_{t,t+\delta t}$ with $m'_{t,t+\delta t}$ and z_t with $z_{t,t+\delta t}$ in the loss functions $\mathcal{L}$.

4.3 Hyper-parameters

The code is implemented using the solo-learn code base [78]. As visual encoder, we use a ResNet50, a well-established convolutional neural network that has been used to model human vision [8]. Important hyper-parameters were selected based on ImageNet-100 k-nearest-neighbors accuracy [79]. For E-SimCLR, we set $\tau = 0.1$, $\alpha_{\text{SMC}} = 0.1$, and $\Delta T = 10\,\text{s}$ and all projection heads a_θ, g_θ and u_θ have 2 hidden layers. For E-MoCoV3, we use $\tau = 0.2$, $\alpha_{\text{SMC}} = 0.1$, and $\Delta T = 2\,\text{s}$. Hidden layers of all projection and prediction networks have 2 hidden layers, except for the predictor p_θ for which we use 1 hidden layer. Regarding E-VICReg, we set $\lambda = 10$, $\mu = 10$, $\nu = 1$, $\alpha_{\text{SMC}} = 0.1$, $\Delta T = 5\,\text{s}$ and use 3 hidden layers in the projection heads.

Nymeria We train the models for 20 epochs. After hyper-parameter search, we configure the initial learning rate to 1.6 for all methods, scaled by $\frac{256}{\text{Batch size}}$. Following standard practices [7], we apply a linear warmup schedule from 0.00003 to 1.6 during the first training epoch. Then, we decrease the learning rate following a cosine decay scheduler. We use a Lars optimizer and a batch size of 512 distributed into 8 GPUs.

To adapt to Cifar10 resolution, we pre-train a different model. We employ the modified version of Resnet50 (No max pooling, and first kernel size of 3×3 without strides). We also resize the pre-training images to 32×32. That way, we observed a small improvement in object recognition accuracy.

Kitti We use the preprocessed dataset provided along Prednet [3]. Similar to Prednet, we resize the training images to 128×160 and pre-train E-SSL (Kitti) for 200 epochs.

4.4 Evaluation protocol

When possible, we followed task-specific established evaluation protocols from the SSL literature.

Object and Scene perception. Following standard linear evaluation protocols [7], we freeze the visual encoder and train a linear classifier on top of the learned representations. Object and scene perception are evaluated on ImageNet-1k [30] and Places365 [80], respectively. We also consider Cifar10 in some experiment for comparison purpose. To ensure that performance differences do not stem from suboptimal learning-rate choices, we simultaneously train classifiers with base SGD learning rates in $0.05, 0.1, 0.2, 0.4, 0.8, 1.6$ [23]. We report the maximum accuracy across epochs and learning rates.

To maintain consistency between training and testing for both Prednet and E-SSL (Kitti), we always resize the images to 128×128 during evaluation, versus 224×224 for other models (32×32 for Cifar10). Due to computational constraints (for Prednet), we train the linear probe of Kitti models for 20 epochs.

Semantic segmentation. We evaluate semantic segmentation on PascalVOC 2012 using the BYOL protocol [81]. We modify the fourth ResNet50 block to use dilation 2 and stride 1. The resulting feature maps feed a randomly initialized convolutional head composed of two 3×3 layers with padding 6 and dilation 6, each followed by Batch Normalization and ReLU activations. A final 1×1 convolution predicts, after up-sampling, per-feature class labels. The model is trained with cross-entropy for 100 epochs and we report the best validation mean intersection-over-union (mIoU).

Single instance segmentation. We evaluate single-instance segmentation on DAVIS 2016 [82] using the same architecture as for semantic segmentation. The visual encoder is frozen and only the convolutional head is trained for 100 epochs with a base learning rate of 3. Performance is measured using the Jaccard index $\mathcal{J}$ (IoU). Due to high variance, results are averaged over five random seeds.

Depth estimation. We evaluate depth estimation on NYUv2 [83], mainly following the protocol used in BYOL [81]. Frozen Layer 4 features feed four fast up-projection blocks with filter sizes 512, 256, 128, and 64 [84]. A final 3×3 convolution predicts depth values that are then upsampled to the original resolution. Training runs for 40 epochs using an L1 loss with a base learning rate of 0.1. We report the best test relative mean squared error across epochs.

Pose estimation. We evaluate pose estimation on MVImgNet [85], which provides camera poses alongside object-centric video clips. At each iteration, two images are sampled from a clip. From their poses we compute the ground-truth relative rotation quaternion q_{gt} and translation vector p_{gt}. Because clips have different world scales, translations are normalized to represent only the direction of motion. Each image is processed independently by the frozen visual encoder. The resulting representations are concatenated and passed through a four-layer fully connected network with output dimensions $512, 256, 128$, and 7. Hidden layers are followed by Batch Normalization and ReLU activations. The prediction head is trained for 20 epochs using the losses

$$\mathcal{L}_{\text{rot}} = 1 - \left| q_{\text{pred}}^{\top} q_{\text{gt}} \right|, \qquad \mathcal{L}_{\text{trans}} = \text{SmoothL1}(p_{\text{pred}} - p_{\text{gt}}). \tag{1}$$

We use the 10% train split defined in [77] and report the loss function over the validation split.

References

[1] Lipshutz D, Windolf C, Golkar S, Chklovskii D. A biologically plausible neural network for slow feature analysis. Advances in neural information processing systems. 2020;33:14986-96.

[2] Halvagal MS, Zenke F. The combination of Hebbian and predictive plasticity learns invariant object representations in deep sensory networks. Nature neuroscience. 2023;26(11):1906-15.

[3] Lotter W, Kreiman G, Cox D. Deep Predictive Coding Networks for Video Prediction and Unsupervised Learning. In: International Conference on Learning Representations; 2017. .

[4] Maye A, Engel AK. A discrete computational model of sensorimotor contingencies for object perception and control of behavior. In: 2011 IEEE International Conference on Robotics and Automation. IEEE; 2011. p. 3810-5.

[5] Richards BA, Lillicrap TP, Beaudoin P, Bengio Y, Bogacz R, Christensen A, et al. A deep learning framework for neuroscience. Nature neuroscience. 2019;22(11):1761-70.

[6] Oquab M, Darcet T, Moutakanni T, Vo H, Szafraniec M, Khalidov V, et al. DINOv2: Learning Robust Visual Features without Supervision. Transactions on Machine Learning Research Journal. 2024.

[7] Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. In: International conference on machine learning. PMLR; 2020. p. 1597-607.

[8] Zhuang C, Yan S, Nayebi A, Schrimpf M, Frank MC, DiCarlo JJ, et al. Unsupervised neural network models of the ventral visual stream. Proceedings of the National Academy of Sciences. 2021;118(3):e2014196118.

[9] Wiskott L, Sejnowski TJ. Slow feature analysis: Unsupervised learning of invariances. Neural computation. 2002;14(4):715-70.

[10] Miyashita Y. Neuronal correlate of visual associative long-term memory in the primate temporal cortex. Nature. 1988;335(6193):817-20.

[11] Li N, DiCarlo JJ. Unsupervised natural experience rapidly alters invariant object representation in visual cortex. science. 2008;321(5895):1502-7.

[12] Franzius M, Wilbert N, Wiskott L. Invariant object recognition and pose estimation with slow feature analysis. Neural computation. 2011;23(9):2289-323.

[13] Aubret A, Ernst M, Teulière C, Triesch J. Time to augment self-supervised visual representation learning. arXiv preprint arXiv:220713492. 2022.

[14] Yu Z, Aubret A, Raabe MC, Yang J, Yu C, Triesch J. Active Gaze Behavior Boosts Self-Supervised Object Learning. arXiv preprint arXiv:241101969. 2024.

[15] O'regan JK, Noë A. A sensorimotor account of vision and visual consciousness. Behavioral and brain sciences. 2001;24(5):939-73.

[16] Di Paolo EA, Barandiaran XE, Beaton M, Buhrmann T. Learning to perceive in the sensorimotor approach: Piaget's theory of equilibration interpreted dynamically. Frontiers in human neuroscience. 2014;8:551.

[17] Miyashita Y, et al. Inferior temporal cortex: where visual perception meets memory. Annual review of neuroscience. 1993;16(1):245-63.

[18] Schapiro AC, Kustner LV, Turk-Browne NB. Shaping of object representations in the human medial temporal lobe based on temporal regularities. Current biology. 2012;22(17):1622-7.

[19] Bohg J, Hausman K, Sankaran B, Brock O, Kragic D, Schaal S, et al. Interactive perception: Leveraging action in perception and perception in action. IEEE Transactions on Robotics. 2017;33(6):1273-91.

[20] Kawato M. Internal models for motor control and trajectory planning. Current opinion in neurobiology. 1999;9(6):718-27.

[21] Rao RP. A sensory-motor theory of the neocortex based on active predictive coding. bioRxiv. 2022:2022-12.

[22] Pandey L, Wood S, Wood J. Are vision transformers more data hungry than newborn visual systems? Advances in Neural Information Processing Systems. 2024;36.

[23] Schaumlöffel T, Aubret A, Roig G, Triesch J. Temporal Slowness in Central Vision Drives Semantic Object Learning. In: The Fourteenth International Conference on Learning Representations; 2026. Available from: https://openreview.net/forum?id=tEezDE0vWt.

[24] Scherr F, Guo Q, Moraitis T. Self-supervised learning through efference copies. Advances in Neural Information Processing Systems. 2022;35:4543-57.

[25] Ma L, Ye Y, Hong F, Guzov V, Jiang Y, Postyeni R, et al. Nymeria: A Massive Collection of Multimodal Egocentric Daily Motion in the Wild. arXiv preprint arXiv:240609905. 2024.

[26] Journé A, Rodriguez HG, Guo Q, Moraitis T. Hebbian Deep Learning Without Feedback. In: The Eleventh International Conference on Learning Representations;. .

[27] Chen X, Liu D, Laydevant J, Grollier J. Self-contrastive forward-forward algorithm. Nature Communications. 2025;16(1):5978.

[28] Chen X, Xie S, He K. An empirical study of training self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision; 2021. p. 9640-9.

[29] Bardes A, Ponce J, Lecun Y. VICReg: Variance-Invariance-Covariance Regularization For Self-Supervised Learning. In: ICLR 2022-International Conference on Learning Representations; 2022. .

[30] Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. Imagenet large scale visual recognition challenge. International journal of computer vision. 2015;115:211-52.

[31] Ho-Phuoc T. CIFAR10 to compare visual recognition performance between deep neural networks and humans. arXiv preprint arXiv:181107270. 2018.

[32] Chen Y, Bardes A, Li Z, LeCun Y. Bag of Image Patch Embedding Behind the Success of Self-Supervised Learning. Trans Mach Learn Res. 2023;2023. Available from: https://openreview.net/forum?id=r06xREo3QG.

[33] Erlikhman G, Caplovitz GP, Gurariy G, Medina J, Snow JC. Towards a unified perspective of object shape and motion processing in human dorsal cortex. Consciousness and Cognition. 2018;64:106-20.

[34] Dwivedi K, Bonner MF, Cichy RM, Roig G. Unveiling functions of the visual cortex using task-specific deep neural networks. PLoS computational biology. 2021;17(8):e1009267.

[35] Geiger A, Lenz P, Stiller C, Urtasun R. Vision meets robotics: The kitti dataset. The international journal of robotics research. 2013;32(11):1231-7.

[36] Riesenhuber M, Poggio T. Hierarchical models of object recognition in cortex. Nature neuroscience. 1999;2(11):1019-25.

[37] Mallat S. Group invariant scattering. Communications on Pure and Applied Mathematics. 2012;65(10):1331-98.

[38] Smith LB. Learning to recognize objects. Psychological Science. 2003;14(3):244-50.

[39] Smith LB, Jones SS, Landau B, Gershkoff-Stowe L, Samuelson L. Object name learning provides on-the-job training for attention. Psychological science. 2002;13(1):13-9.

[40] Diesendruck G, Bloom P. How specific is the shape bias? Child development. 2003;74(1):168-78.

[41] Tartaglini AR, Vong WK, Lake B. A Developmentally-Inspired Examination of Shape versus Texture Bias in Machines. In: Proceedings of the Annual Meeting of the Cognitive Science Society. vol. 44; 2022. .

[42] Aubret A, Triesch J. Do vision models perceive objects like toddlers? In: The Fourth Blogpost Track at ICLR 2025; 2025. .

[43] Bornstein MH, Kessen W, Weiskopf S. Color vision and hue categorization in young human infants. Journal of Experimental Psychology: Human Perception and Performance. 1976;2(1):115.

[44] Atkinson J, Braddick O. Visual segmentation of oriented textures by infants. Behavioural Brain Research. 1992;49(1):123-31.

[45] Sensoy Ö, Culham JC, Schwarzer G. Do infants show knowledge of the familiar size of everyday objects? Journal of experimental child psychology. 2020;195:104848.

[46] Batsell P, Satoshi T, Wen B. Learning Through Little Eyes: Attribute Discrimination Beyond Objects. arXiv preprint arXiv:251218951. 2025.

[47] Sheybani S, Smith L, Tiganj Z, Maini S, Dendukuri A. ModelVsBaby: A developmentally motivated benchmark of out-of-distribution object recognition. Preprint at https://osf io/preprints/psyarxiv/83gae_v1. 2024.

[48] O'Doherty C, Dineen ÁT, Truzzi A, King G, Zaadnoordijk L, Harrison K, et al. Infants have rich visual categories in ventrotemporal cortex at 2 months of age. Nature Neuroscience. 2026:1-10.

[49] Milner AD, Goodale MA. Two visual systems re-viewed. Neuropsychologia. 2008;46(3):774-85.

[50] Livingstone M, Hubel D. Segregation of form, color, movement, and depth: anatomy, physiology, and perception. Science. 1988;240(4853):740-9.

[51] Kristensen S, Garcea FE, Mahon BZ, Almeida J. Temporal frequency tuning reveals interactions between the dorsal and ventral visual streams. Journal of cognitive neuroscience. 2016;28(9):1295-302.

[52] Ayzenberg V, Simmons C, Behrmann M. Temporal asymmetries and interactions between dorsal and ventral visual pathways during object recognition. Cerebral Cortex Communications. 2023;4(1):tgad003.

[53] Bambach S, Crandall D, Smith L, Yu C. Toddler-inspired visual object learning. Advances in neural information processing systems. 2018;31.

[54] Sheybani S, Hansaria H, Wood J, Smith L, Tiganj Z. Curriculum learning with infant egocentric videos. Advances in Neural Information Processing Systems. 2024;36.

[55] Lu Z, Thorat S, Cichy RM, Kietzmann TC. Adopting a human developmental visual diet yields robust, shape-based AI vision. arXiv preprint arXiv:250703168. 2025.

[56] Campos JJ, Anderson DI, Barbu-Roth MA, Hubbard EM, Hertenstein MJ, Witherington D. Travel broadens the mind. Infancy. 2000;1(2):149-219.

[57] Anderson DI, Campos JJ, Witherington DC, Dahl A, Rivera M, He M, et al. The role of locomotion in psychological development. Frontiers in psychology. 2013;4:440.

[58] Schwarzer G, Gehb G, Kelch A, Gerhard-Samunda T, Jovanovic B. Locomotion training contributes to 6-month-old infants' mental rotation ability. Human Movement Science. 2022;85:102979.

[59] Johnson MH. Functional brain development in humans. Nature Reviews Neuroscience. 2001;2(7):475-83.

[60] Li J, Osher DE, Hansen HA, Saygin ZM. Innate connectivity patterns drive the development of the visual word form area. Scientific reports. 2020;10(1):18039.

[61] Bakhtiari S, Mineault P, Lillicrap T, Pack C, Richards B. The functional specialization of visual cortex emerges from training parallel pathways with self-supervised predictive learning. Advances in neural information processing systems. 2021;34:25164-78.

[62] Muttenthaler L, Dippel J, Linhardt L, Vandermeulen RA, Kornblith S. Human alignment of neural network representations. In: The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net; 2023. Available from: `https://openreview.net/forum?id=ReDQ1OUQR0X`.

[63] Mahner FP, Muttenthaler L, Güçlü U, Hebart MN. Dimensions underlying the representational alignment of deep neural networks with humans. Nature Machine Intelligence. 2025;7(6):848-59.

[64] Xu Y, Vaziri-Pashkam M. Limits to visual representational correspondence between convolutional neural networks and the human brain. Nature communications. 2021;12(1):2065.

[65] Doshi FR, Fel T, Konkle T, Alvarez GA. Visual Anagrams Reveal Hidden Differences in Holistic Shape Processing Across Vision Models. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems;. .

[66] Geirhos R, Jacobsen JH, Michaelis C, Zemel R, Brendel W, Bethge M, et al. Shortcut learning in deep neural networks. Nature Machine Intelligence. 2020;2(11):665-73.

[67] Bowers JS, Malhotra G, Dujmović M, Montero ML, Tsvetkov C, Biscione V, et al. Deep problems with neural network models of human vision. Behavioral and Brain Sciences. 2023;46:e385.

[68] Ayzenberg V, Behrmann M. Does the brain's ventral visual pathway compute object shape? Trends in Cognitive Sciences. 2022;26(12):1119-32.

[69] Ayzenberg V, Behrmann M. The dorsal visual pathway represents object-centered spatial relations for object recognition. Journal of Neuroscience. 2022;42(23):4693-710.

[70] Gibson JJ. The ecological approach to visual perception: classic edition. Psychology press; 2014.

[71] Waxman SR. Links between object categorization and naming: Origins and emergence in human infants. Early category and concept development: Making sense of the blooming, buzzing confusion. 2003:213-41.

[72] LaTourrette A, Chan DM, Waxman SR. A principled link between object naming and representation is available to infants by seven months of age. Scientific reports. 2023;13(1):14328.

[73] Schaumlöffel T, Aubret A, Roig G, Triesch J. Caregiver Talk Shapes Toddler Vision: A Computational Study of Dyadic Play. In: 2023 IEEE International Conference on Development and Learning (ICDL). IEEE; 2023. p. 67-72.

[74] Sheybani S, Hansaria H, Wood J, Smith L, Tiganj Z. Curriculum learning with infant egocentric videos. Advances in Neural Information Processing Systems. 2023;36:54199-212.

[75] Long B, Xiang V, Stojanov S, Sparks RZ, Yin Z, Keene GE, et al. The BabyView dataset: High-resolution egocentric videos of infants' and young children's everyday experiences. arXiv preprint arXiv:240610447. 2024.

[76] Orhan AE, Lake BM. Learning high-level visual representations from a child's perspective without strong inductive biases. Nature Machine Intelligence. 2024;6(3):271-83.

[77] Aubret A, Teulière C, Triesch J. Self-supervised visual learning from interactions with objects. In: European Conference on Computer Vision. Springer; 2024. p. 54-71.

[78] da Costa VGT, Fini E, Nabi M, Sebe N, Ricci E. solo-learn: A Library of Self-supervised Methods for Visual Representation Learning. Journal of Machine Learning Research. 2022;23(56):1-6. Available from: `http://jmlr.org/papers/v23/21-1155.html`.

[79] Tian Y, Krishnan D, Isola P. Contrastive multiview coding. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16. Springer; 2020. p. 776-94.

[80] Zhou B, Lapedriza A, Khosla A, Oliva A, Torralba A. Places: A 10 million Image Database for Scene Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2017.

[81] Grill JB, Strub F, Altché F, Tallec C, Richemond P, Buchatskaya E, et al. Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems. 2020;33:21271-84.

[82] Perazzi F, Pont-Tuset J, McWilliams B, Van Gool L, Gross M, Sorkine-Hornung A. A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation. In: Computer Vision and Pattern Recognition; 2016. .

[83] Nathan Silberman PK Derek Hoiem, Fergus R. Indoor Segmentation and Support Inference from RGBD Images. In: ECCV; 2012. .

[84] Laina I, Rupprecht C, Belagiannis V, Tombari F, Navab N. Deeper depth prediction with fully convolutional residual networks. In: 2016 Fourth international conference on 3D vision (3DV). IEEE; 2016. p. 239-48.

[85] Yu X, Xu M, Zhang Y, Liu H, Ye C, Wu Y, et al. MVImgNet: A Large-scale Dataset of Multi-view Images. In: CVPR; 2023. .

[86] Schaumlöffel T, Aubret A, Roig G, Triesch J. Human Gaze Boosts Object-Centered Representation Learning. arXiv preprint arXiv:250102966. 2025.

[87] Everingham M, Eslami SMA, Van Gool L, Williams CKI, Winn J, Zisserman A. The Pascal Visual Object Classes Challenge: A Retrospective. International Journal of Computer Vision. 2015 Jan;111(1):98-136.

[88] Yang Z, Wang Q, Bertinetto L, Hu W, Bai S, Torr PH. Anchor diffusion for unsupervised video object segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision; 2019. p. 931-40.

[89] Bendidi I, Bardes A, Cohen E, Lamiable A, Bollot G, Genovesio A. No free lunch in self supervised representation learning. arXiv preprint arXiv:230411718. 2023.

Table 1: Comparison of E-SSL models with SSL models that use bio-implausible learning signals.

	Pre-training data	Vision encoder	Top-1 I1K acc.
Bio-inspired SSL			
Prednet	1 car ego-hours	Specific	4.59%
E-VICReg (TC+SMC)	300 adult ego-hours	ResNet50	38.87%
E-SimCLR (TC+SMC)	300 adult ego-hours	ResNet50	41.31%
E-MoCoV3 (TC+SMC)	300 adult ego-hours	ResNet50	44.23 %
SSL with bio-implausible learning signals			
MoCoV3	300 adult ego-hours	ResNet50	46.30 %
MoCoV3 (TC) [86]	3,600 adult ego-hours	ResNet50	49.58%
MoCoV3	100k curated images (I1K 10%)	ResNet50	50.9%
DinoV2 [75]	868 baby ego-hours	ViT-B/14	53.28%
DinoV2 [75]	3,600 adult ego-hours	ViT-B/14	54.39%
MoCoV3	1M curated I1K images (I1K 100%)	ResNet50	75.3%

Table 2: Raw models performance as well as estimated lower-bound and upper-bound for each visual task.

Task	E-MoCoV3	No pretraining	Minimum	Maximum
ImageNet-1k ↑	**44.23**	6.05	0.1 (Chance)	74.3 [81]
Places365 ↑	**41.59**	9.25	0.27 (Chance)	53.2[81]
PascalVOC (segm.) ↑	**34.69**	1.5	0	76.3[81]
Davis2016 ↑	**36.05**	12.73	0	81.7 [88]
NYUv2 ↓	**0.756**	1.128	1.41 (First epoch)	0.573 [84]
MVImgNet ↓	**0.137**	0.334	0.37 (First epoch)	0

A E-SSL approaches the performance of SSL based on bio-implausible learning signals

Here, we compare our Embodied SSL framework (E-MoCoV3, E-VICReg, E-SimCLR), which rest on biologically inspired learning principles, to standard SSL models (MoCoV3, DinoV2) that on non-biological inductive biases, such as invariance to color jitter and horizontal flips. We use the conventional SSL evaluation protocol illustrated in Figure 1C. As shown in Table 1, E-MoCoV3 (44.23%) nearly matches the performance of MoCoV3 (46.30%) when pre-trained on 300 hours of egocentric data. Furthermore, E-SSL performs competitively compared to training regimes that leverage substantially larger video corpora, infant-view videos, or curated subsets of data. Overall, this demonstrates that the biological principles TC and SMC achieve performance comparable to the bio-implausible learning signals employed by artificial SSL models.

B E-SSL yields versatile visual representations

In daily life, humans must visually estimate distances and egomotion, categorize scenes, locate objects, or infer motion. To assess whether the visual representations learned by E-SSL support solving such diverse visual tasks, we probe the pre-trained vision encoder on a range of datasets. As observed in Figure 2B, E-SSL yields drastic performance gains compared to "No pre-training" across all considered tasks. We also visualize downstream depth and object location predictions of E-SSL in Figure 2C-D-E. The models evaluated without pre-training fails to capture scene geometry or localize salient objects reliably. In contrast, E-SSL exhibits a coarse but consistent understanding of scene geometry and salient object location. This indicates that the learned representations capture generic and diverse visual features transferable to many downstream tasks.

The performance reported in Figure 6 is calculated as $\frac{x-\text{minimum}}{\text{maximum}-\text{minimum}}$ where x is the performance of a given model. "minimum" and "maximum" are defined according to Table 2.

C Results are consistent across E-SSL models

In this section, we complement the experiments presented in the main paper. In Figure 7, we consider a model that leverages bio-implausible learning signals (MoCoV3) as well as various E-SSL models (E-SimCLR and E-VICReg). We observe that the performance of E-MoCoV3 is consistent with that of other E-SSL models, across various tasks, loss functions (TC, TC+SMC) and metrics for computing

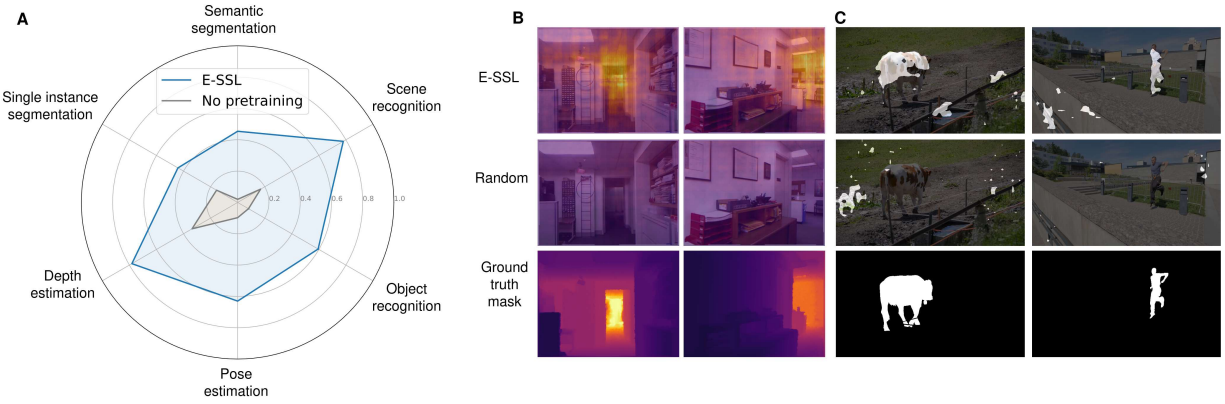


Figure 6: **A)** Normalized performance of E-MoCoV3 and "No pretraining" for diverse visual tasks: scene categorization with Places365 [80], semantic segmentation with PascalVOC [87], depth estimation with NYUv2 [83], single instance segmentation with Davis2016 [82], and estimation of egomotion with MVImgNet [85]. We restrict our comparison to a Random neural network (Resnet50) as the evaluation protocol can not be easily adapted to prior bio-inspired models. For each dataset, we perform a min-max normalization based on the theoretical optimal performance or state-of-the-art results from computer vision methods. **B-C-D)** Visualization of the predictions of E-MoCoV3 and "No pretraining" for downstream depth estimation (B), single instance segmentation (C).

RDMs. MoCoV3 is reasonably well-aligned with infants' representations, except that this model fails to build color prototypes, which is likely due to the model being trained to learn invariance to color transformations [89].

In Figure 8, we present the raw performance of three E-SSL models on our set of visual tasks for various hyper-parameters. Overall, we see that the results obtained in section 2.4 are consistent across E-SSL models. Figure 9 further demonstrates that the findings made in section 2.3 generalize to diverse E-SSL models.

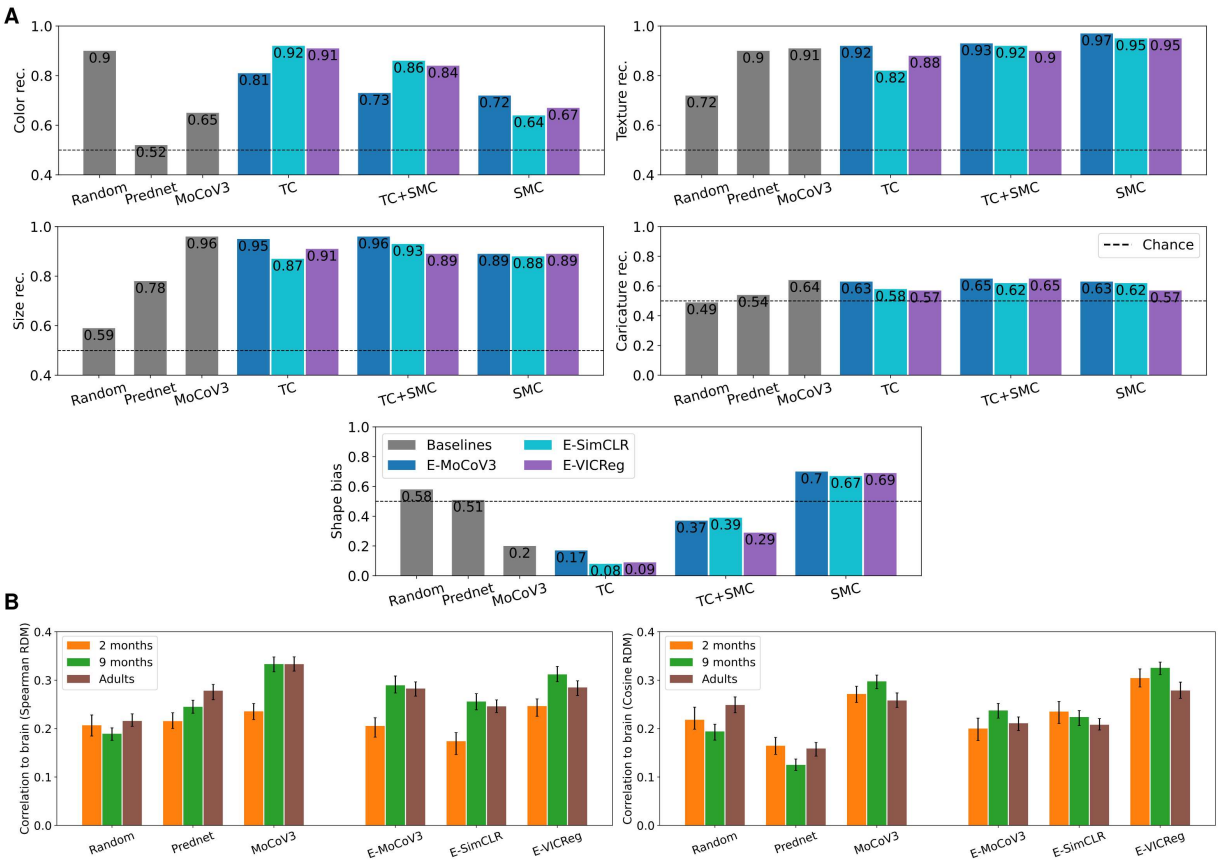


Figure 7: Developmental alignment for additional SSL and E-SSL models. **A)** We follow the evaluation protocol presented in Figure 3A-B. **B)**Spearman correlation scores when models' RDMs are computed with Spearman score (left) or cosine similarity (right). To present raw results, we do not normalize the scores by the noise ceiling. See Figure 3D for more details.

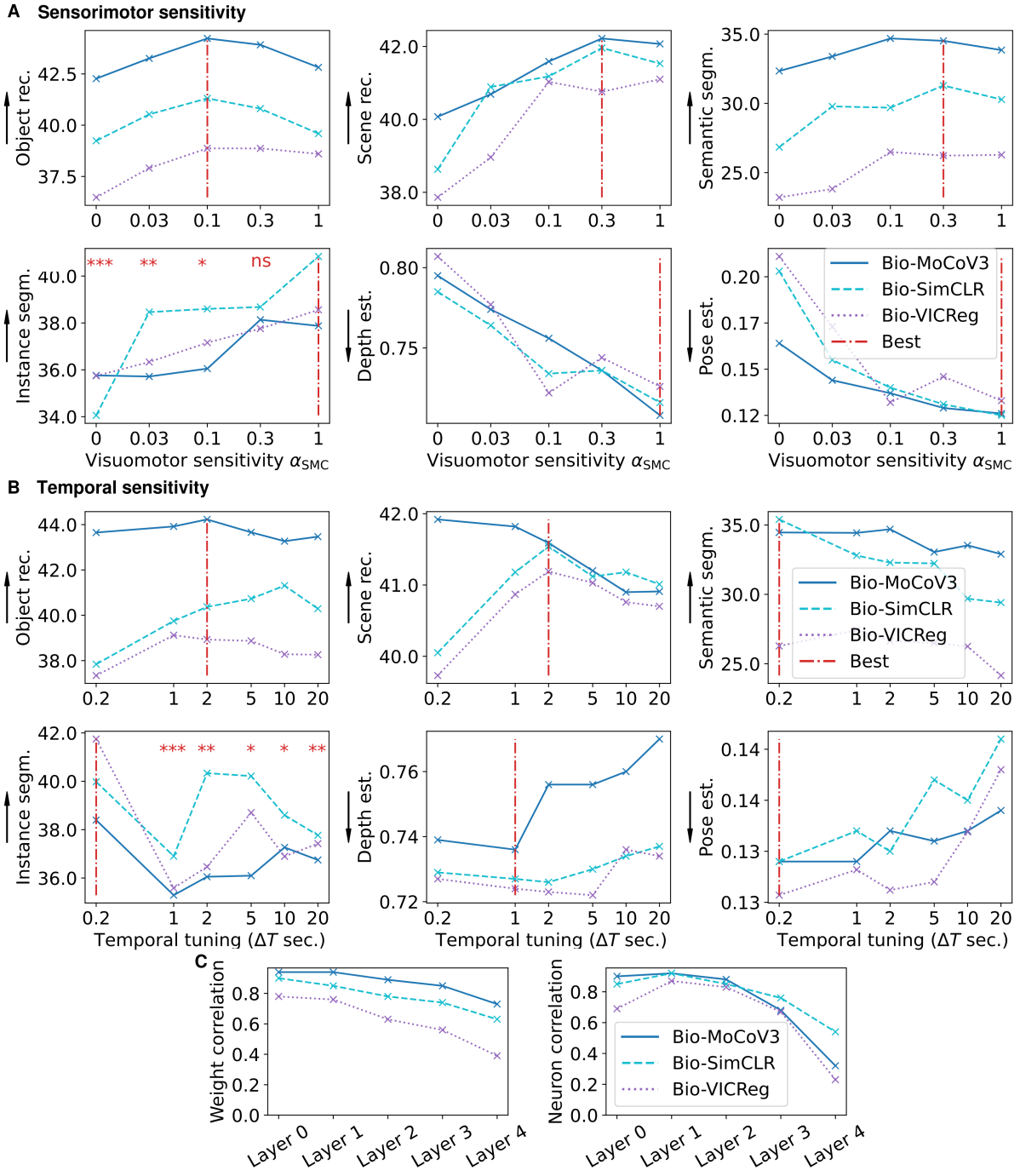


Figure 8: Detailed results of Figure 5 for each E-SSL model and downstream task. Because of high variance (omitted for clarity), Davis2016 results are averaged over 5 random seeds and we compute an unpaired t-test across E-SSL methods between $\text{IoU}(\alpha_{\text{SMC}} = 1) - \text{IoU}(\alpha_{\text{SMC}} = v)$ and zeros, for each $v \in [0, 0.03, 0.1, 0.03]$. $***$, $**$ and $*$ indicate a p-value $p < 0.001$, $p < 0.01$ and $p < 0.05$, respectively. "Best" denotes the best performance averaged across E-SSL methods.

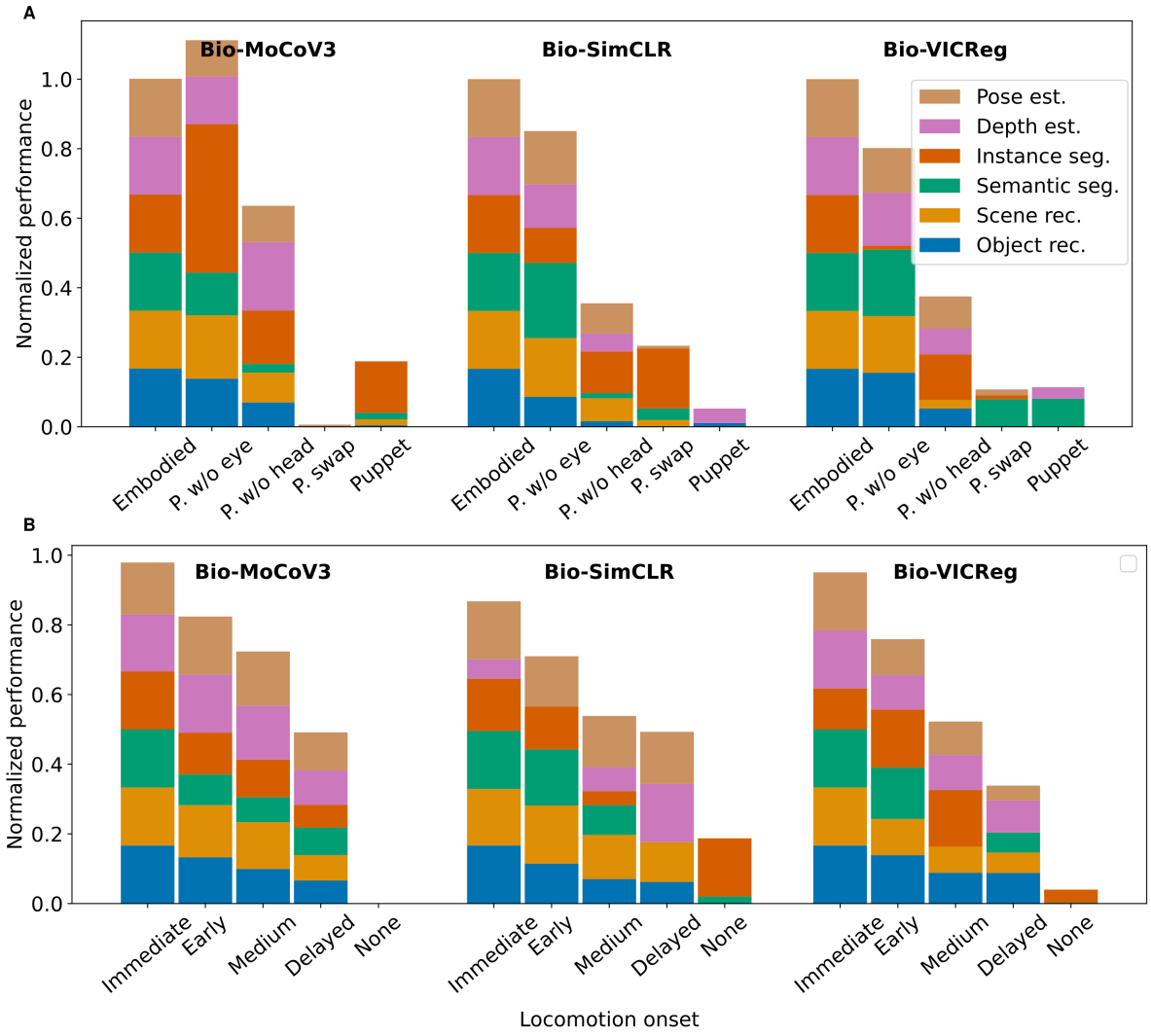


Figure 9: Complementary results of Figure 4.