

Learning Like Infants: Input Constraints in Predictive World Models

Somesh Mohapatra^{1*}

¹Independent.

*Address correspondence to: pikulsomesh@gmail.com

Abstract

Infants learn from input that is degraded, ordered in time and delivered by senses that mature at different rates, whereas machine-learning systems are usually trained on clean, shuffled data. We ask how such constraints change learning in predictive world models, networks that learn by forecasting their own sensory input. We prove that, for a linear read-out under gradient flow, input contrast sets the speed of learning (scaling it by κ delays learning by exactly κ^{-2}) while blur lowers the reachable ceiling; that widening a network delays learning; and that a squared-error forecaster is more surprised by an object’s reappearance than by its disappearance. We test these results in a controlled multisensory simulation with one world model per sense and 5 seeds per training regime. A newborn-quality front end slows vision 39-fold in theory, and the trained network lags the best score its input allows (0.02 against 0.60); ablations show that contrast, not blur, produces the lag. A gentle start helps hearing and costs vision a little, growth halves compute but slows learning, and fusing senses tracks a hidden object better than any one sense. In this small synthetic setting, the quality and order of input change learning in predictable ways, which suggests, as a hypothesis for future tests at scale, that they could be a cheap lever for training world models when data are limited.

1 Introduction

Most machine-learning systems learn from data that are clean, shuffled and presented at full quality from the first step. Biological learners do not. A human infant receives a single continuing stream of experience through sensory organs that are immature at birth and that mature on different schedules: newborn visual acuity is roughly one thirtieth of the adult’s and contrast sensitivity is a small fraction of it [1, 2], touch and hearing function earlier than vision, and the order in which senses come online may itself shape how they are combined [3]. Infants nevertheless learn quickly from a few thousand waking hours of such experience [4], and networks trained on one child’s head-camera video now learn useful visual representations and word meanings [5–7]. This raises a question that is about learning rather than about infants: when the input to a learner is staged, degraded and ordered, and when its capacity may grow, what happens to the dynamics of learning, and which of these constraints help?

The evidence from machine learning is mixed. Starting with blurred or colourless images can make recognition more robust [8, 9], slowly changing video helps self-supervised learning [10], and networks show critical periods of their own, including for multisensory integration [11, 12]. Yet the benefit of “starting small” [13] was challenged by later simulations [14], ordered training helps little on standard benchmarks

and mostly when training time is short or labels are noisy [15], and stage-like learning can emerge from the dynamics of learning alone [16]. What is missing is an account of *how* an input constraint changes learning, specific enough to predict which constraint will matter and by how much.

Predictive world models make this question pressing. A world model learns a compact internal model of its environment and uses it to forecast what follows from what it senses and does [17]. LeCun argues that such models, learned by prediction in an abstract representation space rather than in pixels, are a missing ingredient of autonomous intelligence [18], and joint-embedding predictive architectures (JEPA) put the idea into practice for images and video [19–21]. Li describes world models as generative, multimodal and interactive systems that predict the next state of a scene given an action, and names spatial intelligence as the next frontier [22]. These systems are trained, like language models, on very large corpora, yet training datasets are projected to grow to roughly the size of the whole stock of public human-written text between 2026 and 2032 [23], and multisensory, embodied data for world models are scarcer still. How data are staged, filtered and ordered, which an infant’s development does by design, may therefore matter as much as how much of it there is.

We provide an account of how input constraints change learning, for predictive world models trained to forecast their own sensory input from its recent past. We formalise a staged-input problem, prove four propositions about it (an information bound on what can be learned from filtered input, an exact learning-time law for a linear read-out, the delay that follows growth with function-preserving initialisation, and the conditional-mean account of why squared-error forecasters are not surprised by disappearance) and test each in a controlled simulation with one world model per sense. Whether the effects hold beyond this small simulation, with larger networks, richer scenes and real video, is not tested here; if they do, the quality and order of input could be a cheap lever for training world models when data are limited, and the Discussion says what our evidence supports and what it does not.

Companion paper. The same simulation and age-indexed front ends underlie a companion manuscript on sensory perception across infant ages, which collects published milestones for each sense, shows the front-end tables and the application built on them, and uses three of our modelling studies for what they say about perception [24]. The two papers share data and say so where they do. Shared with the companion paper: the front-end parameter tables (Table 3 here; Fig. 2 there), the information-ceiling and own-age probe data for vision and hearing (Fig. 3 and Extended Data Fig. 3 here; Fig. 3 there, plotted differently), and the one-sentence finding that the ceiling-based half-age measure does not recover the infant order. New in this paper and absent there: the formal problem, the four propositions and their proofs, the spectral analysis (Fig. 2), the six training regimes with statistics and the front-end ablations, the growth analysis, the fusion model, the violation-of-expectation dissociation and the follow-up experiments of Extended Data Fig. S4. The companion experiments (critical period, sense-onset order, perceptual narrowing, training beyond 12 months, a fit of the hearing front end to infant thresholds and the training of a latent-prediction variant of the vision model) are reported only in the companion paper; the latent variant is the same model in both, and its surprise at the disappearance events is reported here.

2 Theory

The problem. A scene state s_t evolves under the agent’s action u_t . Each sense s delivers an adult-quality observation o_t^s , a deterministic function of the scene (Methods), and a property $Y_t = g(s_t)$, such as the position or identity of a toy, is the quantity we ask a representation to retain. At age a the sense delivers $\tilde{o}_t^s = \Phi_s(o_t^s; a)$, where the front end Φ_s is a fixed, non-learned filter whose parameters (acuity, contrast,

colour, frequency resolution, noise) are interpolated from published infant values for vision and from coarse tables for the other senses (Table 3, Methods). The age a indexes input quality, never network structure: nothing in the networks is switched on or off by age.

Each sense has a world model, a network with an encoder e , a dynamics model over the last k latents and the action, and a decoder d (Methods). Its training objective at age a is a reconstruction error plus a one-step prediction error on $\tilde{o}$, so the network predicts its own filtered input; a fusion model then combines the frozen latents of the four sense models. We judge a model’s internal state with a probe, a linear read-out fitted to it and scored by R^2 , the fraction of a quantity’s variance the probe recovers (0 is no better than the mean, 1 is perfect); a probe measures what is stored, not what the network does with it. The information ceiling is the best score reachable from the same filtered input, estimated with a separate network trained directly on the task.

Training proceeds in four stages and the model is read out at the end of each, which we call checkpoints 0, 3, 6 and 12 months (names for training stages, not real time). Six regimes (Table 1, Fig. 1) cross the input schedule (adult quality throughout, immature input improving with age, quality drawn at random per batch, or immature input in reverse order) with capacity (fixed or grown by function-preserving widening). All regimes see the same data, the same number of updates and the same hyperparameters, so any difference is due to the schedule or to growth.

Table 1. Training regimes. All see the same episodes for the same number of updates.

Regime	Network	Input quality across the four stages
Immature input (primary)	fixed (12-month size)	$0 \rightarrow 3 \rightarrow 6 \rightarrow 12$ months
Immature input + growth	grows $\times 1 \rightarrow \times 4$	$0 \rightarrow 3 \rightarrow 6 \rightarrow 12$
Growth only	grows $\times 1 \rightarrow \times 4$	adult throughout
Adult input	fixed	adult throughout
Shuffled input quality	fixed	each batch from a random stage
Reversed immaturity	fixed	$12 \rightarrow 6 \rightarrow 3 \rightarrow 0$

We state four results about this setting. Each proof is short and given in full; Fig. 2 and the Results test each result against the simulation.

Proposition 1 (information bound). Let Y be a scene property, O the adult-quality observation and $\tilde{O}_a = \Phi(O; a)$ the filtered observation, with Φ a stochastic degradation of O , so that $Y \rightarrow O \rightarrow \tilde{O}_a$ is a Markov chain. Then for every representation $Z = e(\tilde{O}_a)$,

$$I(Z; Y) \leq I(\tilde{O}_a; Y) \leq I(O; Y), \quad (1)$$

and the best squared-error read-out has ceiling $R_{\star}^2(a) = 1 - \mathbb{E}[\text{Var}(Y \mid \tilde{O}_a)] / \text{Var}(Y)$, and it does not increase from age a to age a' whenever $\tilde{O}_{a'}$ is a further degradation of $\tilde{O}_a$ (nested filters, so that $Y \rightarrow \tilde{O}_a \rightarrow \tilde{O}_{a'}$ is a Markov chain).

Proof. The inequalities are the data-processing inequality applied twice. For the ceiling, if $\tilde{O}_{a'}$ is a further degradation of $\tilde{O}_a$ then $\mathbb{E}[Y \mid \tilde{O}_{a'}] = \mathbb{E}[\mathbb{E}[Y \mid \tilde{O}_a] \mid \tilde{O}_{a'}]$, and by Jensen’s inequality the variance of this conditional mean cannot exceed that of $\mathbb{E}[Y \mid \tilde{O}_a]$. Since $\text{Var}(Y) = \mathbb{E} \text{Var}(Y \mid \cdot) + \text{Var}(\mathbb{E}[Y \mid \cdot])$, the residual variance cannot decrease. ■

The proposition separates what the input contains from what learning extracts. The supervised ceiling we report estimates $R_{\star}^2(a)$ from below; the gap between a self-supervised model and that ceiling is a property

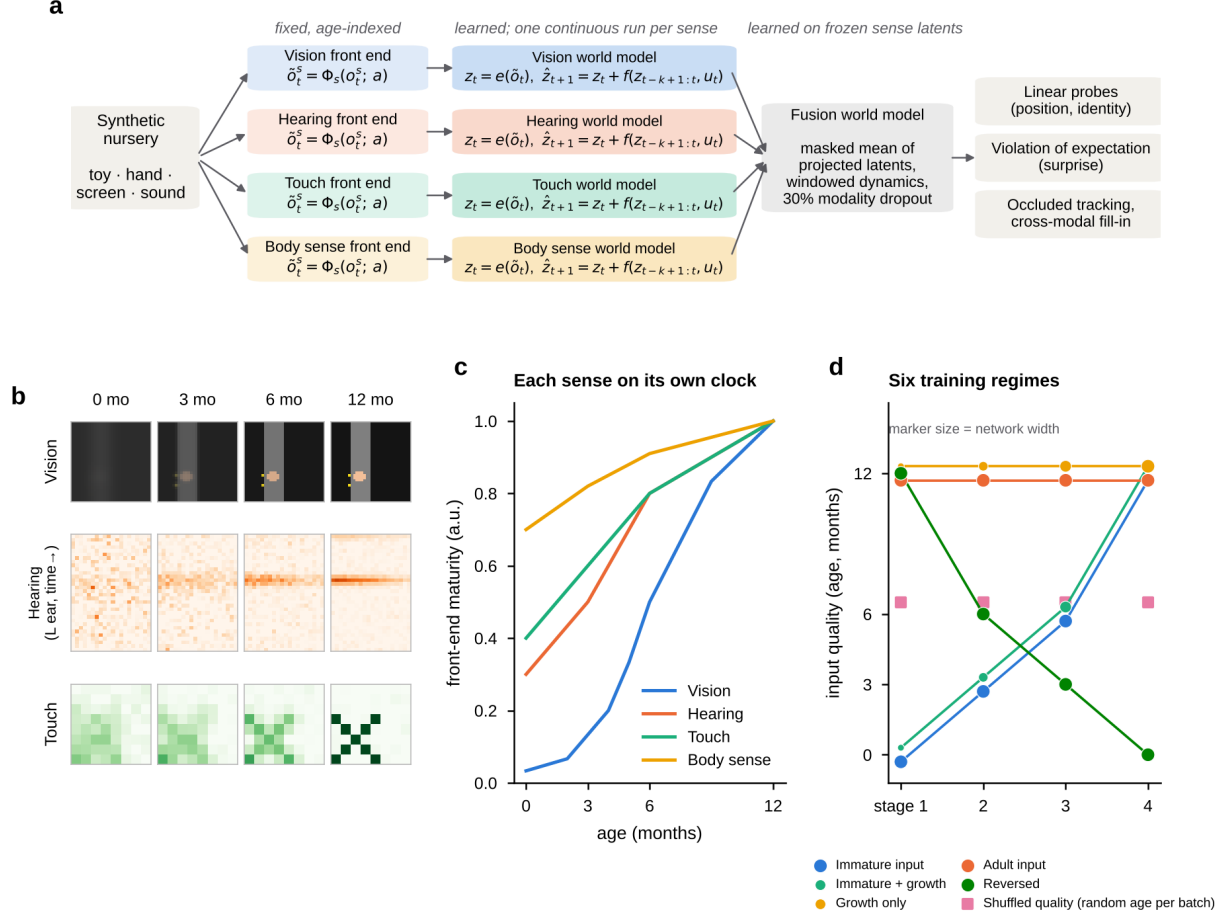


Figure 1. Sense-specific developmental world models. **a**, Each sense passes a synthetic nursery (a toy that is seen, heard and felt, the infant’s babbling hand, and an occluding screen) through a fixed filter $\Phi_s(\cdot; a)$ indexed by age, and learns its own world model: an encoder e , a dynamics model f over the last k latents and the infant’s motor command u_t , and a decoder. A fusion world model combines the frozen sense latents. Read-outs are linear probes, violation-of-expectation surprise and occluded tracking. **b**, The same moment delivered to vision, hearing (left ear, 32 frequency bins over the episode) and touch (an 8×8 patch of skin) at 0, 3, 6 and 12 months. **c**, Front-end maturity per sense (normalised parameters of Table 3): touch and proprioception start close to mature, hearing partly, vision far behind. **d**, The six training regimes: input quality over the four training stages; marker size is network width.

of learning.

Proposition 2 (learning time under gradient flow). Let $\tilde{x} = \Phi(o; a) - \mathbb{E}\Phi(o; a)$ be the centred filtered input, y a standardised target, $\Sigma = \mathbb{E}\tilde{x}\tilde{x}^\top = \sum_i \lambda_i v_i v_i^\top$ its covariance and $c_i = v_i^\top \mathbb{E}[\tilde{x}y]$. A linear read-out $W(0) = 0$ trained by gradient flow $\dot{W} = -\nabla_W \mathbb{E}\|y - W^\top \tilde{x}\|^2/2$ has explained variance

$$R^2(t) = \sum_{i:\lambda_i>0} \frac{c_i^2}{\lambda_i} (1 - e^{-2\lambda_i t}), \quad R^2(\infty) = \sum_{i:\lambda_i>0} \frac{c_i^2}{\lambda_i}. \quad (2)$$

Consequently: (a) if the front end scales input fluctuations by κ , then $\lambda_i \rightarrow \kappa^2 \lambda_i$ and $c_i \rightarrow \kappa c_i$, so $R_\kappa^2(t) = R_1^2(\kappa^2 t)$: the same ceiling is reached κ^{-2} times later. (b) If the front end is a Gaussian blur with standard deviation σ and the input statistics are approximately shift-invariant, the variance of a mode at angular spatial frequency ω is multiplied by $e^{-\sigma^2 \omega^2}$, so its learning time is multiplied by $e^{\sigma^2 \omega^2}$: coarse modes are barely affected, fine modes are learned exponentially later, and modes whose variance falls below a noise floor or the resolution of a finite training budget are not learned at all.

Proof. In the eigenbasis of Σ the modes decouple: $\dot{w}_i = c_i - \lambda_i w_i$, so $w_i(t) = (c_i/\lambda_i)(1 - e^{-\lambda_i t})$. The variance explained by mode i is $2w_i c_i - \lambda_i w_i^2 = (c_i^2/\lambda_i)(1 - e^{-2\lambda_i t})$, and the modes add. For (a), scaling $\tilde{x}$ by κ scales Σ by κ^2 and the cross-covariance by κ , so c_i^2/λ_i is unchanged and $\lambda_i t$ becomes $\kappa^2 \lambda_i t$. For (b), the Fourier transform of a Gaussian of width σ is $e^{-\sigma^2 \omega^2/2}$, so power at frequency ω is multiplied by $e^{-\sigma^2 \omega^2}$. ■

Two points limit how far this carries. The statement is for a linear read-out and plain gradient descent; deep linear networks show a similar inverse dependence of learning time on the strength of an input-output mode, with sigmoidal rather than exponential timing [16], and nonlinear networks can only be expected to follow it loosely. And adaptive optimisers such as Adam divide each step by a running estimate of gradient magnitude, which weakens the dependence on the global scale κ (for a linear read-out, from κ^{-2} towards roughly κ^{-1}); whether contrast or blur dominates in a trained network is therefore an empirical question, which the front-end ablations below answer.

Proposition 3 (growth delay). Widen a hidden layer from n to n' units by $W'_1 = \begin{bmatrix} W_1 \\ R \end{bmatrix}$ and $W'_2 = [W_2 \ 0]$, with R arbitrary. Then the network's function is unchanged, $W'_2 \phi(W'_1 x) = W_2 \phi(W_1 x)$, the gradient on the new incoming weights is zero,

$$\frac{\partial \mathcal{L}}{\partial R} = (\mathbf{0}^\top \delta) \odot \phi'(Rx) x^\top = 0, \quad (3)$$

while the gradient on the new outgoing weights A (initially 0) is $\partial \mathcal{L}/\partial A = \mathbb{E}[\delta \phi(Rx)^\top]$, which is nonzero only if the new units' activations are correlated with the residual error δ (for $\phi(0) = 0$ this requires $R \neq 0$). New capacity therefore starts to learn only through its outgoing weights: once $A \neq 0$ the new units act as fixed random features, and their incoming weights R adapt at a rate proportional to $\|A\|$.

Proof. The first claim follows because the new units' contribution is multiplied by zero. The gradient with respect to R passes through the new outgoing weights 0, which gives the displayed expression. The gradient with respect to the new outgoing weights is $\mathbb{E}[\delta \phi(Rx)^\top]$, which is nonzero unless $\phi(Rx)$ is uncorrelated with the residual. ■

The origin $(r, a) = (0, 0)$ of an incoming and an outgoing weight is a saddle of the loss. With $\dot{a} = \rho r$ and $\dot{r} = \rho a$, where ρ measures the correlation of the new unit with the residual, $a(t) = a_0 \cosh \rho t + r_0 \sinh \rho t$, so a reaches a given size a^* after $t \approx \rho^{-1} \log(2a^*/|a_0 + r_0|)$: with $a_0 = 0$ the delay grows with $\log(1/|r_0|)$ and

shrinks with the step size and with ρ , which itself falls as the existing units absorb the residual. Growth therefore buys a cheaper early network (fewer parameter-updates) at the price of a plateau after each widening.

Proposition 4 (surprise of a squared-error forecaster). The predictor $\hat{o}_{t+1}$ minimising $\mathbb{E}\|\hat{o}_{t+1} - o_{t+1}\|^2$ given the past is the conditional mean $\mathbb{E}[o_{t+1} \mid o_{\leq t}]$. Consider a pixel behind a screen, which shows o^{toy} if the toy reappears at the next frame, with probability p , and o^{screen} otherwise, and let $\Delta = o^{\text{toy}} - o^{\text{screen}} \neq 0$. In a violation-of-expectation test, the surprise at a possible event is compared with the surprise at an impossible one. The optimal forecast is $(1 - p)o^{\text{screen}} + po^{\text{toy}}$ and the surprise (squared error) is

$$S_{\text{reappear}} = (1 - p)^2 \Delta^2, \quad S_{\text{vanish}} = p^2 \Delta^2. \quad (4)$$

The forecaster is more surprised by reappearance than by non-reappearance exactly when $p < \frac{1}{2}$.

Proof. The conditional mean minimises mean squared error (the decomposition $\mathbb{E}\|\hat{o} - o\|^2 = \mathbb{E}\|\hat{o} - \mathbb{E}[o \mid \cdot]\|^2 + \mathbb{E} \text{tr Cov}(o \mid \cdot)$). Substituting the two outcomes gives the displayed errors, with Δ^2 read as $\|\Delta\|^2$ for vector observations. ■

When the time of reappearance is uncertain, $p < \frac{1}{2}$ for most frames of an occlusion, so a negative violation-of-expectation index, averaged over frames, is the correct behaviour of a well-trained pixel forecaster and is not a failure of training or of memory. (The argument is exact for the unweighted loss. The vision networks are trained with extra weight $w_r \geq 1$ on moving pixels, which replaces the threshold $\frac{1}{2}$ by $1/(1 + w_r)$.) A positive index requires either a forecaster whose probability of reappearance exceeds one half, or an additional signal (such as a sound) that makes the impossible event informative.

Fig. 2 shows the spectral quantities of Proposition 2 measured on the simulation’s own frames.

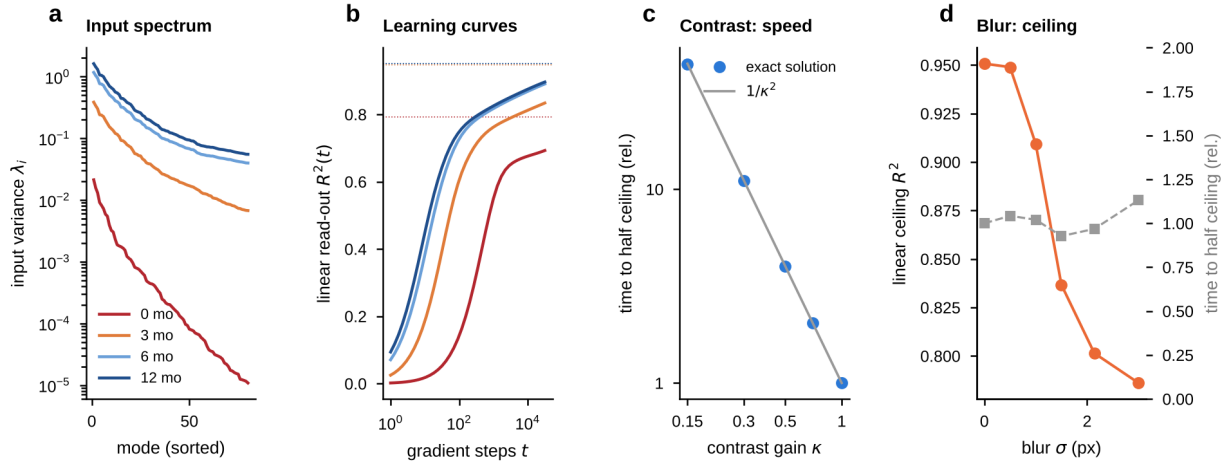


Figure 2. Spectral account of learning time. For the toy-position read-out from filtered frames. **a**, Eigenvalue spectrum of the input covariance at the four ages: immature front ends shrink it by orders of magnitude. **b**, Exact gradient-flow solution $R^2(t)$ of Proposition 2 for each age with the step size fixed by the adult input; dotted lines, the linear ceilings. **c**, Time to half the ceiling when only contrast is scaled by κ : the exact solution follows κ^{-2} . **d**, Blur alone: the ceiling falls with the blur width while the time to half ceiling barely moves.

3 Results

A controlled simulation isolates each constraint. The simulated world renders one toy of four types, a babbling hand and, in half the scenes, a vertical screen into four streams: 32×32 colour vision, binaural hearing (2 ears $\times$ 32 frequency bins, panned by position and passing through the screen), an 8×8 tactile patch on the hand, and four-dimensional proprioception (Fig. 1a; Methods). Each toy type is specified redundantly across senses, by shape and colour, pitch and felt shape [25]. Each sense is filtered by a fixed front end whose parameters follow the age a (Fig. 1b, c; Table 3); for vision the parameters are published infant acuity and contrast sensitivity [1, 2], and for hearing, touch and proprioception they are coarse placeholder tables that respect the known order of maturation. Behind each front end is a world model of the same family (Methods). Each sense is trained in one continuous run of four stages and read out at the end of each. The six regimes (Fig. 1d; Table 1) cross the two levers, input schedule and network growth, with two order controls.

Vision develops last and learns slowly from newborn-quality input. With immature input, the vision model cannot localise the toy at the first checkpoint (position R^2 0.02 ± 0.00) and is still poor at the second (0.31 ± 0.15), then rises to 0.85 ± 0.04 (Fig. 3a). Trained on adult-quality input the same network localises the toy from the first checkpoint (0.69 ± 0.04) and ends at 0.89 ± 0.03 . Hearing, touch and the body senses are close to their final level much earlier (Fig. 3c, e), so vision is the last sense to develop, which is put there by the front-end tables. Measured by the age at which each sense’s information ceiling reaches half its 12-month value (touch 0, hearing 4, vision 0 months; Kendall $\tau = 0.00$ against the order touch, hearing, vision; Fig. 3f) the infant order is not recovered, because the vision ceiling is already high at birth even though the learned model is not.

The early failure is a property of learning, not of the input. Proposition 1 asks whether this early failure is the input or the learner. Probing on each checkpoint’s own input (Fig. 3b) gives 0.00 ± 0.01 at 0 months and 0.85 ± 0.04 at 12, so it is not a distribution shift from training on blurred and testing on sharp images. The newborn input is not empty: a supervised network trained directly on it reaches $R^2 = 0.60$ (Fig. 3a, dashed), and the exact linear solution of Proposition 2 reaches 0.79. The gap between 0.60 and 0.02 is therefore a property of learning. Proposition 2 also quantifies it: with the step size set by adult input, the newborn front end shrinks the total input variance from 32.20 to 0.12 and lengthens the time to half ceiling 39-fold (Fig. 2a, b). The vision model approaches its ceiling only as the front end sharpens (Fig. 3d).

In the linear model, contrast sets the speed of learning and blur sets the ceiling. Scaling contrast alone by κ delays the linear solution by κ^{-2} to within 1% over a 44-fold range, and leaves its ceiling unchanged (Fig. 2c). Blur alone leaves the time to half ceiling almost unchanged but lowers the ceiling from 0.95 without blur to 0.80 at the newborn width of 2.14 px (Fig. 2d). The linear read-out is not the network, and the networks are trained with Adam, which removes the global scale effect; ablations of the front end test which component matters in the trained vision model.

In the trained network, contrast, not blur, produces the early lag. When only blur follows age, the vision model already localises the toy at the first checkpoint (0.34 ± 0.05 , against 0.69 ± 0.04 for adult input and 0.02 ± 0.00 for the full schedule), reaches 0.79 ± 0.05 at the second and ends at 0.88 ± 0.02 (Fig. 4a). When only contrast follows age, the model reproduces the full lag (0.04 ± 0.02 at the first checkpoint, 0.51 ± 0.04 at the second) and identity accuracy follows the same pattern (0.37 ± 0.03 against 0.60 ± 0.06 for blur only at the second checkpoint; Fig. 4b). Contrast therefore accounts for nearly all of the early delay, as Proposition 2 predicts, even though the networks are trained with Adam; blur alone costs little speed at

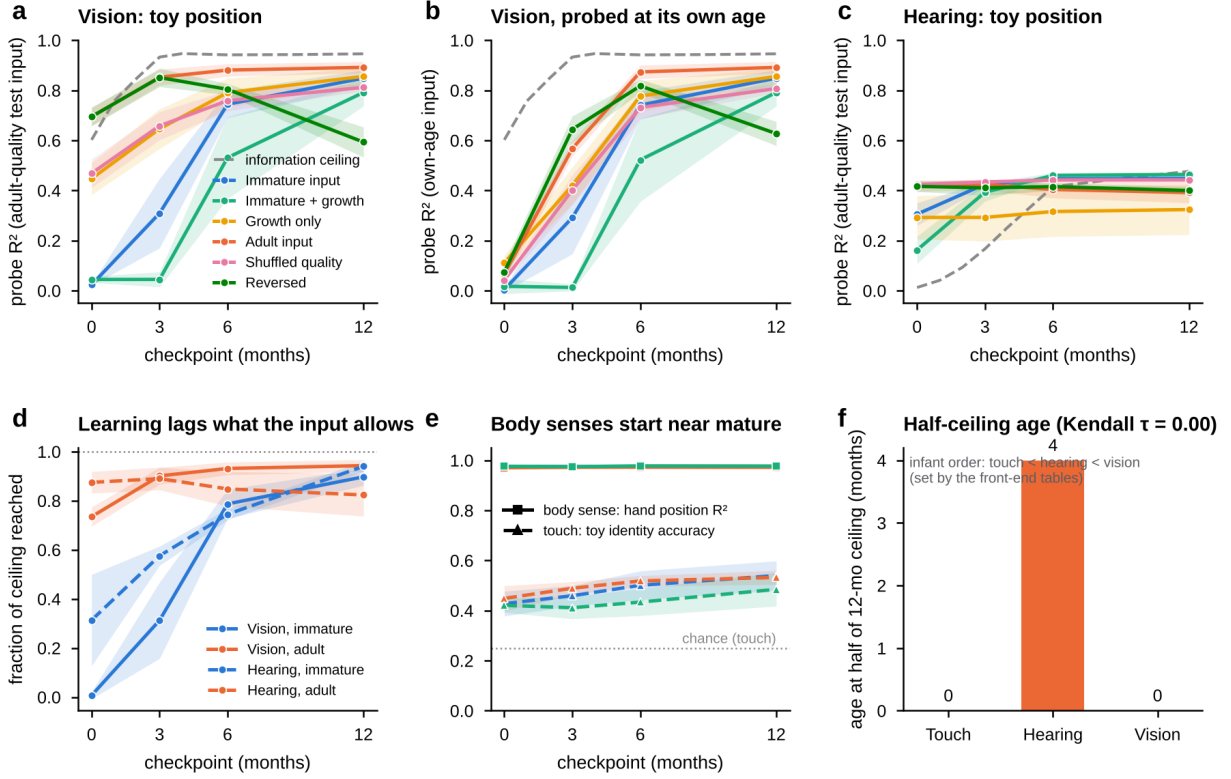


Figure 3. Development of each sense. The ceiling and own-age data for vision and hearing are also plotted against age in the companion paper [24]. **a**, Toy position decoded from the vision model’s latent at each checkpoint (probe fitted and tested on adult-quality input), for the six regimes; dashed grey line, information ceiling (a supervised network trained directly on that age’s filtered input). **b**, The same with the probe fitted and tested on the input quality of the checkpoint’s own age. **c**, Hearing. **d**, Fraction of the age-matched ceiling reached by the immature-input regime (own-age probe) and by the adult-input regime, for vision (solid) and hearing (dashed). **e**, Proprioception (hand position R^2 , squares) and touch (toy identity accuracy, triangles; chance 0.25) are close to their final level from the first checkpoint. **f**, Age at which each sense’s information ceiling first reaches half its 12-month value; this measure does not recover the infant order (Kendall τ against touch < hearing < vision in the panel title), because the vision ceiling is already high at birth. Lines, mean over seeds; bands, ± 1 s.d.; $n = 5$ seeds per regime.

this image scale. One caveat applies: colour saturation was held at the adult value in both ablations (5 seeds each, like every other regime), so its contribution is untested.

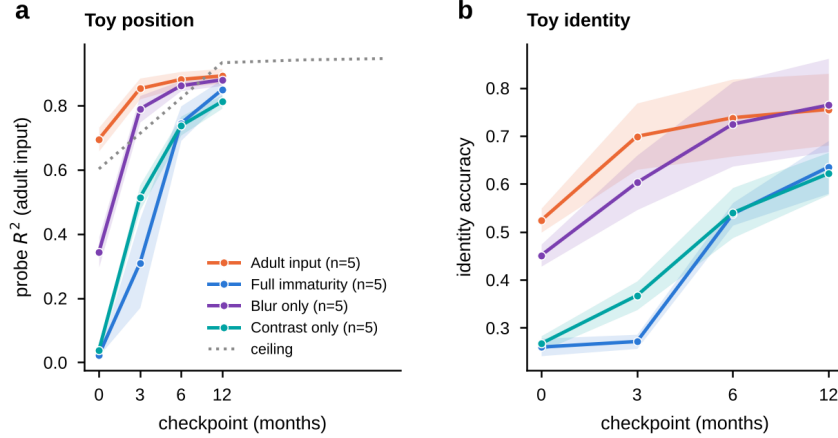


Figure 4. Front-end ablations in the trained vision model. Vision probe score (a, toy position; b, toy identity; adult-quality test input) at the four checkpoints for adult input, the full immature schedule, and schedules in which only blur or only contrast follows age (colour is adult throughout in both). Lines, mean over seeds; bands, ± 1 s.d.; $n = 5$ seeds per regime.

Growth delays learning and halves the parameter-updates. Adding growth to immature input does not change the shape of the curve; it delays it (Fig. 3a). The grown vision model is still at R^2 0.04 ± 0.03 at the 3-month checkpoint and ends at 0.79 ± 0.06 , against 0.85 ± 0.04 for the fixed-size network; identity accuracy is 0.49 ± 0.06 against 0.64 ± 0.06 (Hedges’ $g = -2.2$, $p = 0.005$). Growth without immature input learns quickly and ends near the adult-input model (0.86 ± 0.04). At matched updates, growing networks use 0.48 of the fixed network’s parameter-updates (Fig. 5d). This is the plateau Proposition 3 predicts: new units begin learning only through their outgoing weights, so each widening is followed by a stall, and growth trades accuracy for compute.

Early immaturity costs vision a little. On clean input the adult-input vision model is best. Immature input costs $0.89 - 0.85$ in R^2 ($g = -1.2$, $p = 0.06$). Tested on newborn-quality input, every model fails (Fig. 5a), with regimes that saw immature input failing less on average (-0.22 ± 0.14 against -0.40 ± 0.22 for adult input), but the difference is within seed variability ($p = 0.16$); we do not claim a robustness benefit in vision.

A gentle start helps hearing. Models that heard immature sound early localise better on clean input at 12 months (0.45 ± 0.01 immature, 0.46 ± 0.01 immature with growth) than models that only heard adult-quality sound (0.39 ± 0.05 adult input, 0.32 ± 0.11 growth only; Fig. 5b; $g = 1.5$, $p = 0.05$). Shuffled input quality (0.44 ± 0.01) is close to the immature schedule, so the benefit comes from having heard degraded input at all, a form of augmentation (the mixture objective of training at several qualities), more than from its order. Reversing the schedule is the clearest negative control: the vision model that ends on newborn-quality input falls from 0.85 ± 0.04 to 0.59 ± 0.07 , so what the model has learned does not survive a late degradation.

Fusing senses recovers what each sense alone loses. During a long occlusion, the hidden toy’s horizontal position can be decoded from the vision model’s memory (R^2 0.60 ± 0.08 at 12 months), from hearing, which passes through the screen (0.42 ± 0.09), and best from the fused world model (0.75 ± 0.06 ; Fig. 6a; paired difference fused minus vision positive in every seed). The advantage appears at 6 months and

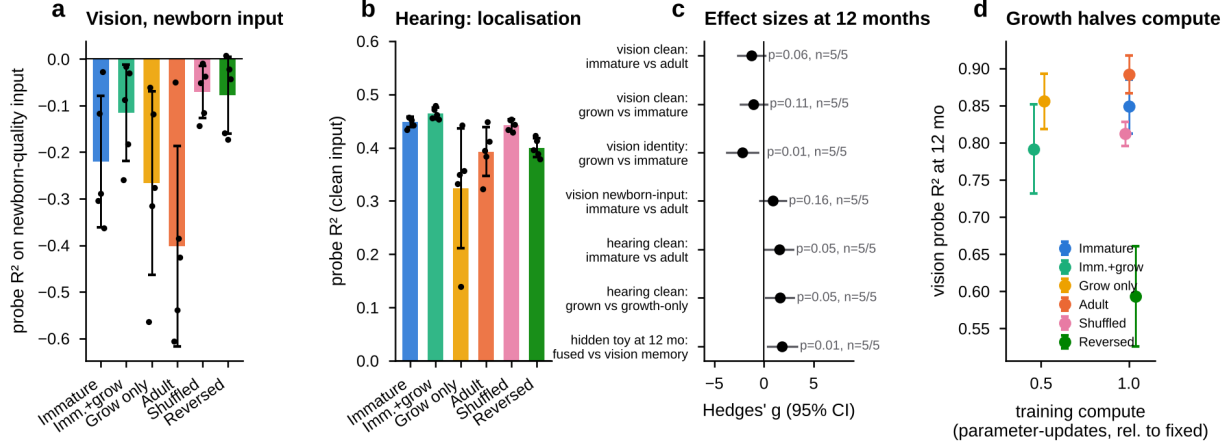


Figure 5. Costs and benefits at the 12-month checkpoint. **a**, Vision probe R^2 when tested on newborn-quality input (bars, mean; dots, seeds; whiskers, s.d.). All values are negative (worse than predicting the mean), so differences are in how badly models fail. **b**, Hearing localisation on clean input. **c**, Effect sizes (Hedges' g with approximate 95% CI) and Welch-test p values for the planned comparisons; n per group shown. **d**, Vision accuracy against training compute (parameter-updates relative to a fixed-size network).

nothing is decodable at 0 to 3 months, when neither vision nor binaural hearing can localise. For a visible toy, the fused state is better than vision alone (0.87 ± 0.03 against 0.79 ± 0.06) and, with vision switched off, still recovers position from hearing, touch and the hand (0.53 ± 0.01 ; Fig. 6b).

No unimodal vision model is surprised by an object's disappearance. No unimodal vision model, in any regime, has a positive violation-of-expectation index (Fig. 6c): all are more surprised when the toy reappears than when it fails to, like a predictor that copies the last frame (Fig. 6e). This holds although the hidden toy's position is decodable from the same model's memory. It is the behaviour Proposition 4 predicts for a squared-error forecaster when the reappearance time is uncertain.

The fused model is surprised only when a sound gives the event away. The fused model does show a positive index when the toy vanishes from every sense (0.32 ± 0.11 at 12 months), but its channels show where the surprise comes from: hearing 0.75 ± 0.05 , vision -0.04 ± 0.05 (Fig. 6d). The model notices that the sound stopped. When the toy vanishes from sight but keeps sounding, the overall index is -0.02 ± 0.01 and the vision channel -0.09 ± 0.07 ; when a silent toy vanishes, they are -0.04 ± 0.02 and -0.23 ± 0.11 . Without a sound to give the event away, the fused model does not expect the toy back either: a read-out can show that information about a hidden object is present while the model's predictions do not use it.

The trained encoder responds first to the coarse, high-variance input modes and widens as the front end sharpens. To ask whether the spectral account carries over to trained networks, we measured the Jacobian of each vision encoder on adult-quality frames and expressed it in the eigenbasis of the adult input covariance (Extended Data Fig. S4c, d; 5 seeds). At the first checkpoint the immature-input encoder puts 0.52 ± 0.09 of its Jacobian energy in the ten input modes that carry 0.69 of the input variance, against 0.09 ± 0.01 for the encoder trained on adult input throughout, and its effective rank is 4.3 ± 0.6 against 12.9 ± 0.8 . The share falls and the rank rises as the front end sharpens (0.06 ± 0.02 and 13.5 ± 1.2 at 12 months), but the rank stays below that of the adult-input encoder (17.9 ± 1.0). Early learning from immature input is therefore concentrated on the modes that dominate the input, as Proposition 2 implies for a linear read-out, and the fine modes are acquired later. This is a measurement on the encoder, not on

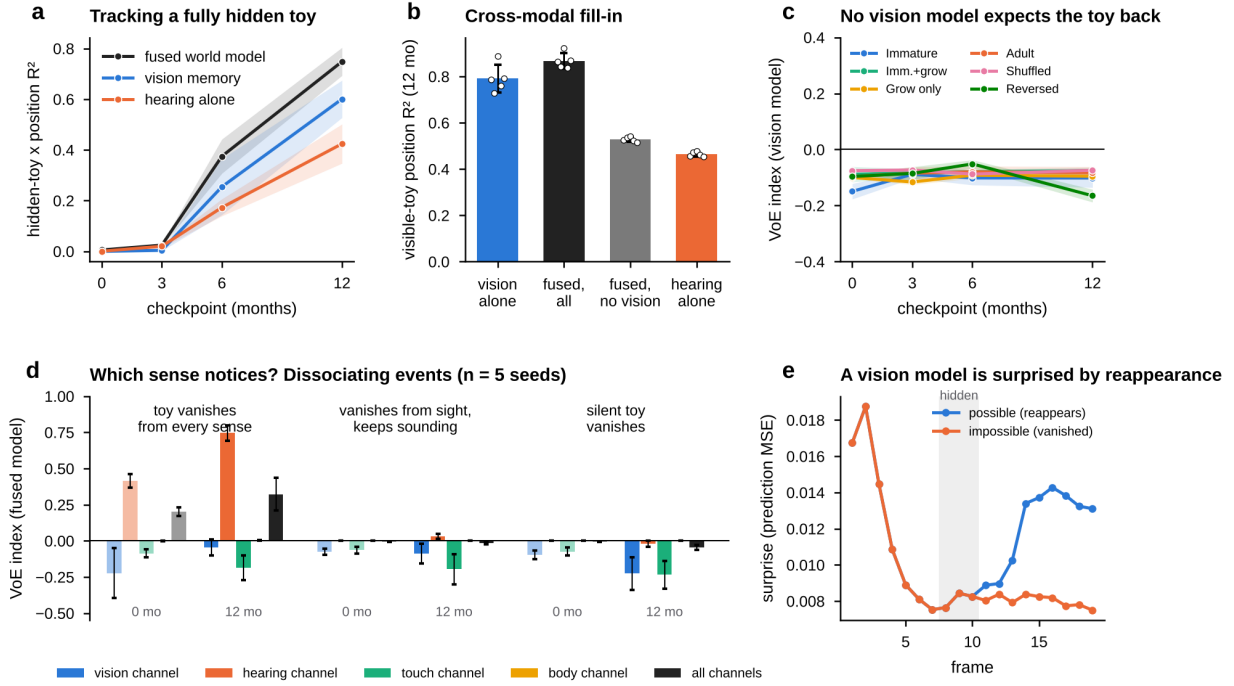


Figure 6. Fusion and expectation. **a**, Horizontal position of a fully hidden toy decoded from the vision model’s memory window, from hearing alone and from the fused world model. **b**, Position of a visible toy at 12 months from vision alone, the fused state, the fused state with vision switched off, and hearing alone. **c**, Violation-of-expectation (VoE) index of the unimodal vision models; positive means more surprise when the toy fails to reappear. **d**, VoE index of the fused model broken down by channel, for three events: the toy vanishes from every sense; it vanishes from sight and touch but keeps sounding; a silent toy vanishes. Faded bars, 0 months; solid, 12 months. **e**, Frame-by-frame surprise of one vision model (immature input, 12 months) on a possible and an impossible event; grey, toy hidden. Bands and whiskers, ± 1 s.d. over seeds.

the map to the toy’s position, and it supports the spectral account without proving it for deep networks.

Matching experience to age leaves the 12-month result unchanged. In the staged schedule every checkpoint receives the same number of updates, so the labels 0, 3, 6 and 12 months name stages, not amounts of experience. In a single run in which the front-end age rises with the number of updates, so that experience is proportional to age, vision reaches position R^2 0.86 ± 0.01 at 12 months, against 0.85 ± 0.04 for the staged schedule, and hearing reaches 0.45 ± 0.01 in both (Extended Data Fig. S4a, b). The intermediate checkpoints differ, as they should: the matched vision model has had half as many updates at 3 months as at 6, and reads out 0.03 ± 0.02 at 3 months and 0.67 ± 0.05 at 6 months, against 0.29 ± 0.16 and 0.74 ± 0.06 in the staged schedule. The 12-month results of the main study do not depend on the choice of staging.

Predicting an uncertainty, or predicting in latent space, does not make the model surprised by disappearance. Proposition 4 attributes the missing surprise to a squared-error forecaster that predicts a conditional mean. We therefore tested two forecasters that predict something else. A vision model whose forecast comes with a per-pixel variance, trained by the Gaussian likelihood, tracks the toy as well as the main model (position R^2 0.85 ± 0.04 at 12 months) and still has a negative violation-of-expectation index in all five seeds (-0.35 ± 0.23 by negative log-likelihood; -0.10 ± 0.04 by squared error of the mean, as in the main model). A latent-space forecaster with a slowly updated target encoder and variance–covariance regularisation has a negative index in all five seeds as well (-0.22 ± 0.14), but it does not learn to represent the toy (position R^2 0.02 ± 0.04), so this second test is uninformative about expectation (Extended Data Fig. S4e, f). Predicting an uncertainty did not change the sign of the index, so it is not only a consequence of the squared-error loss; whether a latent objective that does track the toy changes it remains open.

4 Discussion

What the theory explains. Table 2 sets each proposition against what we observed. The information bound and the learning-time law together account for the main shape of development in the vision model: a front end that shrinks input variance and removes fine detail makes the early read-out both lower and slower, and the self-supervised model reaches only a small fraction of the ceiling until the front end sharpens. The growth delay accounts for why growing the network did not help, and the conditional-mean result accounts for the sign of the surprise index.

Table 2. Predictions of the theory and what was observed.

Proposition	Prediction	Observation
1. Information bound	Learned score cannot exceed the ceiling at that input quality	Holds at every age; early gap is large for vision
2. Learning time	Contrast scales speed by κ^{-2} ; blur limits the ceiling	Exact in the linear model; in the trained network, contrast-only reproduces the lag and blur-only barely slows learning
3. Growth delay	A plateau after each widening; fewer parameter-updates	Grown model lags; uses about half the parameter-updates
4. Conditional mean	Negative surprise index when reappearance is uncertain	Negative for every unimodal vision model; positive only when a sound gives the event away

What it does not explain. The propositions are exact for a linear read-out under gradient flow, whereas the networks we train are nonlinear and use Adam, so the theory and the experiments are linked by analogy and by ablation, not by derivation. The trained vision model lags its ceiling far more than the linear model lags its own, which suggests that most of the delay lies in learning features, not in fitting a read-out. Proposition 2 does not apply as stated to adaptive optimisers, yet the ablations show that contrast, the component it singles out, still accounts for the early lag in the trained network. Extending the theory to deep linear networks, building on the analysis of their learning dynamics [16], would close part of this gap and is future work. The Jacobian spectra of the trained encoders show the same ordering, coarse modes first (Extended Data Fig. S4c, d), but we have not derived the delay for deep networks. Nor does the theory explain the benefit of early immaturity in hearing: the shuffled-order control suggests augmentation, but we have not derived it.

Implications for staged training. Three design rules follow, each stated for the setting studied here. Reduce contrast or gain only if the cost in speed is acceptable, because it delays learning by the square of the scale with plain gradient descent and gives no lasting information benefit; use blur when the aim is to withhold fine detail early, and expect a lower ceiling in a finite budget; and do not use growth as a substitute for a schedule, because new units contribute nothing until their outgoing weights move. When surprise is the read-out, a squared-error pixel forecaster is a poor instrument in this setting. A forecaster that predicts a per-pixel variance and a latent-space forecaster were not surprised by the toy’s disappearance either (Extended Data Fig. S4e, f), but the latent model did not learn to track the toy at this scale [24], so objectives that do, such as masked prediction over many patches and more data [21], or training on longer horizons, remain to be tested against the same dissociating events.

Limitations. Everything is synthetic and small (32×32 scenes with one toy, at most 0.6 million parameters per sense) and is evaluated mainly with linear probes, which measure decodability, not behaviour. We did not test whether these schedules help or hurt larger architectures or standard benchmarks, so the findings describe this setting and are not evidence about training at scale. Maturity is defined by the task: touch reaches its 12-month level at once because tracking a toy needs no fine tactile discrimination, and a harder tactile task would give a different order. The theory is exact for a linear read-out under gradient flow and is suggestive for trained networks. Hearing, touch and proprioception tables are placeholders, and the labels 0, 3, 6 and 12 months name training stages and carry no claim about hours of experience. With 5 seeds per regime, several comparisons have large effect sizes but are not significant (Fig. 5c); we report them as estimates. The Jacobian spectra describe the encoder and not the map to the read-out, and the experience-matched schedule has no 0-month checkpoint (its first is at 1 month).

Future work. The immediate extensions are late fusion with sense-onset order as a factor; comparing the biological order of the senses (touch, then hearing, then vision) with artificial and arbitrary schedules for multisensory integration; latent-space objectives, with an inductive bias or masking strategy that lets them survive newborn-level blur, tested against the dissociating events; and extending the theory to deep linear networks and measuring the delay from the Jacobian spectrum. The decisive tests are at scale: whether staged input improves sample efficiency or robustness on standard benchmarks and with modern architectures, in a 3D embodied environment and on real infant egocentric video. The data path runs from this synthetic simulation to an embodied simulator for touch and proprioception [26] and to head-camera corpora for vision and hearing [27, 28].

5 Materials and Methods

5.1 Experimental design

The study compares six training regimes (Table 1), plus two front-end ablations for vision, for four sense-specific world models at four checkpoints, with six analyses specified before the final runs: a spectral analysis of the filtered input (below); probes at adult-quality, own-age and newborn-quality input; a supervised information ceiling; violation-of-expectation surprise with dissociating events; occluded tracking and fill-in with a fusion model; and planned statistical comparisons (Table 4). Each regime is trained with the same data, the same number of updates and shared hyperparameters.

5.2 Environment

Episodes last $T = 20$ frames. A toy of type $c \in \{1, \dots, 4\}$ moves with constant velocity and bounces in a 32×32 scene; the hand’s position follows a smoothed random motor command $u_t \in \mathbb{R}^2$ and stays in the lower half. Half the episodes contain a vertical screen. Vision is an RGB image on a dark background; the toy is drawn with a type-specific shape and colour, the screen in grey and the hand in skin tone. Hearing is a binaural spectrum, $o_t^{\text{aud}} \in \mathbb{R}^{2 \times 32}$, a Gaussian bump at a type-specific pitch whose loudness falls with distance and whose left/right balance follows horizontal position; sound passes through the screen. Touch is an 8×8 image of the toy’s shape when the hand overlaps the toy, otherwise zero. Proprioception is the hand’s normalised position and velocity. Ground truth includes toy position, type, visibility and contact. Training sets have 3,000 episodes per seed, test sets 400.

Probe events. *Possible*: the toy crosses a central screen horizontally and reappears. *Impossible*: the toy vanishes while hidden, from every sense. *Vanish-visual*: it vanishes from vision and touch but keeps sounding as if still moving. *Silent*: possible and impossible events with no sound at all (hearing receives noise only). *Track*: like possible, with a wider screen (14 px) and varied speed, so the toy is hidden for about seven frames. Paired events share random draws so that they differ only in the manipulation.

5.3 Front ends

All front ends are fixed functions $\tilde{o} = \Phi_s(o; a)$ whose parameters are linearly interpolated in age from Table 3.

Vision. Write $\text{cpd}(a)$ for acuity (cycles per degree), $\text{CS}(a)$ for contrast sensitivity and $\gamma(a)$ for colour saturation at age a ; then The first equation sets how much to blur the image, with more blur at lower acuity. The equations then apply that blur, pull the contrast towards the image average, and fade the colour towards grey, each by an amount that depends on age.

$$\sigma(a) = \frac{1}{2} \cdot \frac{30}{\text{cpd}(a)} \cdot \frac{32}{224}, \quad x = G_{\sigma(a)} * o, \quad x \leftarrow \bar{x} + \kappa(a)(x - \bar{x}), \quad \kappa(a) = \text{clip}\left(\frac{1}{2} \log_{10} \text{CS}(a), 0.15, 1\right), \quad (5)$$

$$\tilde{o} = \text{clip}_{[0,1]}(\ell + \gamma(a)(x - \ell)), \quad \ell = 0.299x_R + 0.587x_G + 0.114x_B, \quad (6)$$

where G_σ is a separable Gaussian with replicate padding and $\bar{x}$ the image mean. The log compression of contrast replaces a linear gain $\text{CS}/100$, which would leave a newborn 2% of contrast; contrast sensitivity is a threshold, not a gain.

Hearing. Per ear, $x = g_{\beta(a)} * o$ along frequency, then a high-frequency roll-off $x_f \leftarrow x_f \text{sigmoid}((h(a) \cdot 32 - f)/1.5)$, then reduced interaural difference $x_{L,R} \leftarrow m + \rho(a)(x_{L,R} - m)$ with $m = \frac{1}{2}(x_L + x_R)$, then additive Gaussian noise $\eta(a)$.

Touch and proprioception. Receptive-field blur $G_{\tau(a)}$ on the 8×8 patch plus noise; additive noise on proprioception.

Table 3. Developmental parameters at the four checkpoints (intermediate months are interpolated). The same tables are shown graphically in the companion paper [24].

Sense	0 months	3 months	6 months	12 months
Vision: acuity (cpd) / contrast sensitivity / colour	1 / 2 / 0.1	4 / 10 / 0.7	15 / 50 / 1.0	30 / 100 / 1.0
Hearing: frequency blur (bins) / high-frequency pass / interaural difference / noise	2.5 / 60% / 0.3 / 0.08	1.8 / 75% / 0.5 / 0.06	1.0 / 90% / 0.8 / 0.04	0.5 / 100% / 1.0 / 0.02
Touch: receptive-field blur (taxels) / noise	1.2 / 0.06	0.9 / 0.04	0.6 / 0.03	0.3 / 0.02
Proprioception: noise	0.12	0.08	0.05	0.02
Capacity (growing regimes): width multiplier / memory window	$\times 1$ / 1	$\times 2$ / 2	$\times 3$ / 4	$\times 4$ / 6

5.4 Spectral analysis and front-end ablations

Spectral analysis. For each age we generated 800 episodes without occluders, kept the 6,000 frames in which the toy is visible, applied the vision front end, centred the pixels and computed the covariance Σ ($3,072 \times 3,072$) and its cross-covariance with the standardised toy position. The closed-form solution of Proposition 2 gives $R^2(t)$, the linear ceiling $\sum_i c_i^2 / \lambda_i$ (modes with $\lambda_i < 10^{-9} \lambda_{\max}$ are dropped, which acts as a numerical noise floor) and the time to half the ceiling, found by bisection. The step size is fixed at $1/\lambda_{\max}$ of the adult input so that every age is stable under the same step. We verified the closed form against explicit gradient descent at age 0 (agreement to 10^{-3} in R^2). Controlled sweeps scale contrast or blur alone at adult quality otherwise. Code: `scripts/analyze_spectral.py`; output: `results/spectral.json`.

Front-end ablations. Two further vision regimes follow the immature schedule for one component of the front end only: *blur only* (acuity follows age, contrast and colour adult) and *contrast only* (contrast sensitivity follows age, acuity and colour adult). All other settings are those of the immature-input regime, with the same seeds.

5.5 Sense world models

For sense s with latent size D_s (32 for vision, 16 for hearing and touch, 8 for proprioception), the encoder e turns what the sense receives into a short summary z_t . The next summary is the current one plus a correction computed from the last k summaries and the baby’s own movement u_t , and the decoder d turns a summary back into sensory input:

$$z_t = e_{\theta}(\tilde{o}_t), \quad \hat{z}_{t+1} = z_t + f_{\theta}([z_t, z_{t-1}, \dots, z_{t-k+1}], u_t), \quad \hat{o} = d_{\theta}(z), \quad (7)$$

where the window is padded with the first frame. The network is penalised for failing to reproduce what it senses now and for failing to predict what it will sense next, with extra weight on pixels that change, much as babies’ eyes are drawn to motion, and a small last term keeps the predicted summary close to the real next one. On a sub-sequence of 10 frames the loss is

$$\mathcal{L} = \frac{\sum_{t,q} m_{t,q} (d(z_t)_q - \tilde{o}_{t,q})^2}{\sum_{t,q} m_{t,q}} + \frac{\sum_{t,q} m_{t+1,q} (d(\hat{z}_{t+1})_q - \tilde{o}_{t+1,q})^2}{\sum_{t,q} m_{t+1,q}} + \frac{0.1}{T-1} \sum_t \|\hat{z}_{t+1} - \text{sg}(z_{t+1})\|^2, \quad (8)$$

where q indexes pixels or channels, sg is the stop-gradient and $m_{t,q} = 1$ except for vision, where per-pixel weights $m_{t,q} = 1 + 20 \max_c |\tilde{o}_t - \tilde{o}_{t-1}|$ up-weight what moves, a stand-in for the orienting to motion without which the small toy is ignored. The vision encoder is a three-layer CNN with two coordinate channels and the decoder a spatial-broadcast decoder [29]; other senses use two-layer MLPs. Widths scale with a multiplier in $\{1, 2, 3, 4\}$.

Growth. Growing a network widens every hidden layer as in Proposition 3 and lengthens the window by adding input weights of zero for the new (older) memory slots, so the model’s function is unchanged (checked for every layer by a numerical test). Vision models grow from 43,507 to 586,499 parameters and hearing models from 13,792 to 149,088.

Training. Adam, learning rate 2×10^{-3} , batch 32, 2,000 updates per stage for vision and 1,500 for other senses, four stages in one continuous run; weights carry over between stages and a fresh Adam optimiser is started at each stage, so all regimes see the same optimiser schedule. Hyperparameters were set once on the immature-input regime and shared by all regimes. Seeds: 5 for each regime (fusion uses the grown regime’s checkpoints).

5.6 Fusion world model

The fusion model takes frozen sense latents z_t^s and a mask $\mu_t^s \in \{0, 1\}$ marking which senses are present. It averages the summaries of the senses that are present, after mapping each to a common size and tagging it with its sense, and predicts forward from the last six steps:

$$h_t = \text{MLP}\left(\frac{\sum_s \mu_t^s (P_s z_t^s + \epsilon_s)}{\max(1, \sum_s \mu_t^s)}\right), \quad \hat{h}_{t+1} = h_t + g([h_t, \dots, h_{t-5}], u_t), \quad (9)$$

with projections P_s to 64 dimensions and learned modality embeddings ϵ_s . Read-out heads r_s reconstruct every sense latent from h_t and from $\hat{h}_{t+1}$, so the model is trained to recover each sense’s own summary from the combined one, both now and one step ahead:

$$\mathcal{L}_{\text{fus}} = \frac{1}{|S|} \sum_s \left(\|r_s(h_t) - z_t^s\|^2 + \|r_s(\hat{h}_{t+1}) - z_{t+1}^s\|^2 \right), \quad (10)$$

with each sense dropped independently with probability 0.3 during training. Latents are standardised per sense.

5.7 Follow-up experiments

All follow-up experiments use the main study’s full-size networks, data generator, optimiser and seeds (0–4), and report the own-age probe.

Experience-matched schedule. The staged schedule gives every checkpoint the same number of updates (2,000 for vision, 1,500 for hearing). In the matched schedule one network is trained in a single run of four times that many updates in which the front-end age rises linearly with the update count from 0 to 12 months; checkpoints are taken after 1/12, 3/12, 6/12 and 12/12 of the updates, so that experience is proportional to age.

Uncertainty-aware forecaster. The main-study vision model gains a second decoder with the same architecture whose sigmoid output is scaled to a log-variance $\log \sigma_q^2$ between -7 and 1 for every pixel q of the forecast. It is trained on the Gaussian negative log-likelihood $\frac{1}{2}(\log \sigma_q^2 + (\tilde{o}_{t+1,q} - \hat{o}_{t+1,q})^2 / \sigma_q^2)$ with the same motion weights, on detached inputs (the predicted latent and the predicted mean), so the forecaster itself receives exactly the main-study gradients and the head learns how uncertain that forecaster is. Surprise is the unweighted mean of the negative log-likelihood over pixels; the squared error of the mean is also reported.

Latent forecaster. The decoder is removed and the dynamics module predicts the next frame’s code from a target encoder whose weights are an exponential moving average (rate 0.99) of the encoder’s, with no gradient through the target, plus variance and covariance penalties on the codes (weights 25, 25 and 1). Surprise is the mean squared error between the predicted code and the target encoder’s code of the next frame. This is the model trained in the companion paper, and the same runs are used.

Violation-of-expectation index. For these forecasters the index uses the same events and frames as the main study, $(S_{\text{imp}} - S_{\text{pos}}) / (|S_{\text{imp}}| + |S_{\text{pos}}|)$; the absolute values leave it unchanged when the surprises are positive and keep it bounded when a negative log-likelihood is negative.

Jacobian spectrum. For each vision checkpoint of the immature-input and adult-input regimes (seeds 0–4; seed 0 from the stored checkpoints, seeds 1–4 retrained with identical settings and reproducing the main-study read-outs exactly) we took 400 adult-quality frames in which the toy is visible, computed the Jacobian of the 32-dimensional latent with respect to the 3,072 input values for each frame by automatic differentiation, and expressed it in the principal directions of the adult frames. The energy of mode k is $\sum_n \|J_n v_k\|^2$; the share in the top 10 and 100 modes and the effective rank $(\sum_i \sigma_i)^2 / \sum_i \sigma_i^2$ of the frame-averaged Jacobian are reported.

5.8 Read-outs

Linear probes. A probe is a straight-line fit from the network’s internal state to the quantity of interest, with a small penalty λ against over-fitting. For features X (a latent, or a memory window of latents) and targets Y we use ridge regression with centred targets and an unpenalised intercept, whose closed-form solution is

$$\hat{B} = \arg \min_B \|(Y - \bar{Y}) - X_c B\|^2 + \lambda \|B\|^2 = (X_c^\top X_c + \lambda I)^{-1} X_c^\top (Y - \bar{Y}), \quad (11)$$

where X_c are the standardised features. The probe is fitted on training episodes and evaluated on test episodes, $R^2 = 1 - \sum \|y - \hat{y}\|^2 / \sum \|y - \bar{y}\|^2$. Identity uses one-hot targets and the arg-max (chance 0.25). Touch is probed on contact frames, proprioception on hand position. *Adult-input* probes use adult-quality input; *own-age* probes use the input quality of the checkpoint’s age; *newborn-input* probes are fitted on adult-quality input and tested on newborn-quality input.

Information ceiling. For every sense and each age $a \in \{0, 1, 2, 3, 4, 6, 9, 12\}$ we train a separate supervised network, using the encoder design of the world models, for 1,500 updates on single filtered frames to regress the toy’s position and classify its identity, and score it on held-out frames.

Violation of expectation. We define surprise as the squared error of the one-step forecast, $S_t = \|d(\hat{z}_t) - \tilde{o}_t\|^2$ (unimodal) or $\|r_s(\hat{h}_t) - z_t^s\|^2$ per channel (fused, standardised latents), averaged over the hidden stretch and the two frames after expected re-emergence. With S_{pos} and S_{imp} the mean surprise on paired possible and impossible events, the index VoE balances surprise at the impossible event against surprise at the possible one: it runs from -1 , when only the possible event surprises, through 0 , when both do equally, to $+1$, when only the impossible one does. The effect size d_z is the average difference divided by its spread across episodes:

$$\text{VoE} = \frac{S_{\text{imp}} - S_{\text{pos}}}{S_{\text{imp}} + S_{\text{pos}}}, \quad d_z = \frac{\overline{S_{\text{imp}} - S_{\text{pos}}}}{\text{sd}(S_{\text{imp}} - S_{\text{pos}})}. \quad (12)$$

Occluded tracking and fill-in. The hidden toy’s horizontal position is probed on fully hidden frames of the *track* events from the vision model’s memory window ($k = 6$ latents), from the hearing latent, and from the fused state, fitting on half of the episodes and testing on the other half. Fill-in probes toy position from the fused state with all senses, without vision, and from single senses.

Statistics. Bands and whiskers are ± 1 s.d. over seeds. Planned comparisons at the 12-month checkpoint use Welch’s t-test and Hedges’ $g = \frac{\bar{a} - \bar{b}}{s_p} \left(1 - \frac{3}{4(n_a + n_b) - 9}\right)$ with approximate 95% interval $g \pm 1.96 \sqrt{\frac{n_a + n_b}{n_a n_b} + \frac{g^2}{2(n_a + n_b)}}$; where s_p is the pooled standard deviation of the two groups; per-cell 95% bootstrap intervals are reported with the results. No correction for multiple comparisons is applied; we list all planned comparisons.

Table 4. Planned comparisons.

Comparison (12-month checkpoint)	Hedges’ g	p (Welch)	n
Vision clean, immature vs adult input	−1.2	0.06	5/5
Vision clean, grown vs immature	−1.0	0.11	5/5
Vision identity, grown vs immature	−2.2	0.005	5/5
Vision on newborn input, immature vs adult	0.9	0.16	5/5
Hearing clean, immature vs adult	1.5	0.05	5/5
Hearing clean, grown vs growth only	1.6	0.05	5/5
Hidden toy, fused vs vision memory	1.8	0.01	5/5

Compute. Training compute is approximated by $\sum_{\text{stages}} (\text{parameters}) \times (\text{updates})$; relative to a fixed 12-month-size network, grown vision models use 0.48 and grown hearing models 0.49.

Milestones. For each sense, the half-ceiling age is the first age at which the ceiling reaches half of its 12-month value (position R^2 for vision and hearing; identity accuracy above chance for touch). Kendall’s τ compares the resulting order with the order of functional maturity usually described for the three senses, touch before hearing before vision [3].

Human and animal studies. The study uses only synthetic data; no human or animal participants were involved.

During the preparation of this work, the author used an AI assistant (Claude, Anthropic) to help write and run analysis code, draw figures and draft text. The author reviewed and edited the content and takes full responsibility for the content of the publication.

All code, configurations, results and trained checkpoints are at <https://github.com/pikulsomesh/developmental-world-models> (MIT licence). The data are synthetic and are regenerated by the code from the random seeds listed in the repository. An interactive companion is at <https://pikulsomesh.github.io/developmental-world-models/>.

S. Mohapatra conceived the study, directed the work and approved the final text.

This work received no specific funding and was carried out independently.

The author declares that there is no conflict of interest regarding the publication of this article.

References

1. Banks MS and Bennett PJ. Optical and photoreceptor immaturities limit the spatial and chromatic vision of human neonates. *Journal of the Optical Society of America A* 1988;5:2059–79.
2. Norcia AM and Tyler CW. Spatial frequency sweep VEP: Visual acuity during the first year of life. *Vision Research* 1985;25:1399–408.
3. Turkewitz G and Kenny PA. Limitations on input as a basis for neural organization and perceptual development: A preliminary theoretical statement. *Developmental Psychobiology* 1982;15:357–68.
4. Smith LB, Jayaraman S, Clerkin E, and Yu C. The Developing Infant Creates a Curriculum for Statistical Learning. *Trends in Cognitive Sciences* 2018;22:325–36.
5. Orhan AE, Gupta V, and Lake BM. Self-supervised learning through the eyes of a child. In: *Advances in Neural Information Processing Systems*. Vol. 33. 2020:9960–71.
6. Orhan AE and Lake BM. Learning high-level visual representations from a child’s perspective without strong inductive biases. *Nature Machine Intelligence* 2024;6:271–83.
7. Vong WK, Wang W, Orhan AE, and Lake BM. Grounded language acquisition through the eyes and ears of a single child. *Science* 2024;383:504–11.
8. Vogelsang L, Gilad-Gutnick S, Ehrenberg E, et al. Potential downside of high initial visual acuity. *Proceedings of the National Academy of Sciences* 2018;115:11333–8.
9. Vogelsang M, Vogelsang L, Gupta P, et al. Impact of early visual experience on later usage of color cues. *Science* 2024;384:907–12.
10. Sheybani S, Hansaria H, Wood JN, Smith LB, and Tiganj Z. Curriculum Learning With Infant Ego-centric Videos. In: *Advances in Neural Information Processing Systems*. Vol. 36. 2023.
11. Achille A, Rovere M, and Soatto S. Critical Learning Periods in Deep Neural Networks. arXiv preprint arXiv:1711.08856 (v3, February 2019; first posted 2017). A conference version was reported at ICLR 2019 but could not be confirmed against a primary record. 2019. arXiv: 1711.08856.
12. Kleinman M, Achille A, and Soatto S. Critical Learning Periods for Multisensory Integration in Deep Networks. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023:24296–305. DOI: 10.1109/CVPR52729.2023.02327.
13. Elman JL. Learning and development in neural networks: The importance of starting small. *Cognition* 1993;48:71–99.

14. Rohde DLT and Plaut DC. Language acquisition in the absence of explicit negative evidence: How important is starting small? *Cognition* 1999;72:67–109.
15. Wu X, Dyer E, and Neyshabur B. When Do Curricula Work? In: *International Conference on Learning Representations*. 2021.
16. Saxe AM, McClelland JL, and Ganguli S. A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences* 2019;116:11537–46.
17. Ha D and Schmidhuber J. World Models. 2018. DOI: 10.5281/zenodo.1207631. arXiv: 1803.10122.
18. LeCun Y. A Path Towards Autonomous Machine Intelligence. OpenReview position paper. 2022. URL: <https://openreview.net/forum?id=BZ5a1r-kVsf>.
19. Assran M, Duval Q, Misra I, et al. Self-Supervised Learning From Images With a Joint-Embedding Predictive Architecture. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023:15619–29.
20. Bardes A, Garrido Q, Ponce J, et al. Revisiting Feature Prediction for Learning Visual Representations from Video. *Transactions on Machine Learning Research* 2024.
21. Garrido Q, Ballas N, Assran M, et al. Intuitive physics understanding emerges from self-supervised pretraining on natural videos. arXiv preprint arXiv:2502.11831 2025.
22. Li FF. From Words to Worlds: Spatial Intelligence is AI’s Next Frontier. Substack essay, 10 November 2025. 2025. URL: <https://drfeifei.substack.com/p/from-words-to-worlds-spatial-intelligence>.
23. Villalobos P, Ho A, Sevilla J, Besiroglu T, Heim L, and Hobbhahn M. Position: Will We Run Out of Data? Limits of LLM Scaling Based on Human-Generated Data. In: *Proceedings of the 41st International Conference on Machine Learning*. Vol. 235. *Proceedings of Machine Learning Research*. 2024:49523–44.
24. Mohapatra S. Simulating Infant Senses with Age-Tuned Networks. Companion paper; manuscript and source code on GitHub. 2026. URL: <https://github.com/pikulso mesh/developmental-world-models/tree/main/papers/baby-senses>.
25. Bahrick LE and Lickliter R. Intersensory redundancy guides attentional selectivity and perceptual learning in infancy. *Developmental Psychology* 2000;36:190–201.
26. Mattern D, Schumacher P, López FM, et al. MIMo: A Multimodal Infant Model for Studying Cognitive Development. *IEEE Transactions on Cognitive and Developmental Systems* 2024;16:1291–301.
27. Sullivan J, Mei M, Perfors A, Wojcik E, and Frank MC. SAYCam: A Large, Longitudinal Audiovisual Dataset Recorded From the Infant’s Perspective. *Open Mind* 2021;5:20–9.
28. Long B, Xiang V, Stojanov S, et al. The BabyView dataset: High-resolution egocentric videos of infants’ and young children’s everyday experiences. arXiv preprint arXiv:2406.10447 2024.
29. Watters N, Matthey L, Burgess CP, and Lerchner A. Spatial Broadcast Decoder: A Simple Architecture for Learning Disentangled Representations in VAEs. arXiv preprint arXiv:1901.07017 2019.

Extended Data (Supplementary Figures)

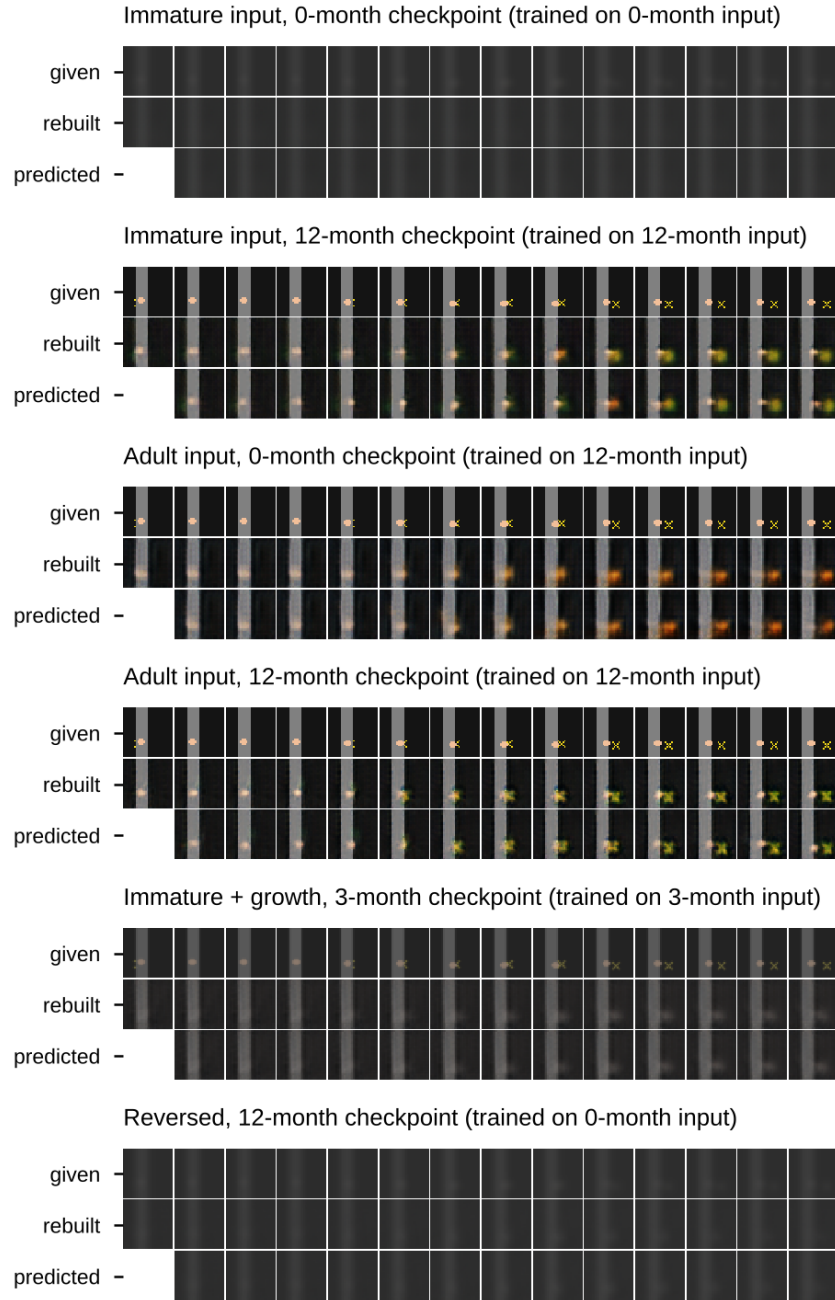


Figure S1. What the vision models see, rebuild and predict. For selected regimes and checkpoints: the input given (at the age that stage was trained on), the reconstruction from the latent, and the one-step prediction, for the first 14 frames of one episode.

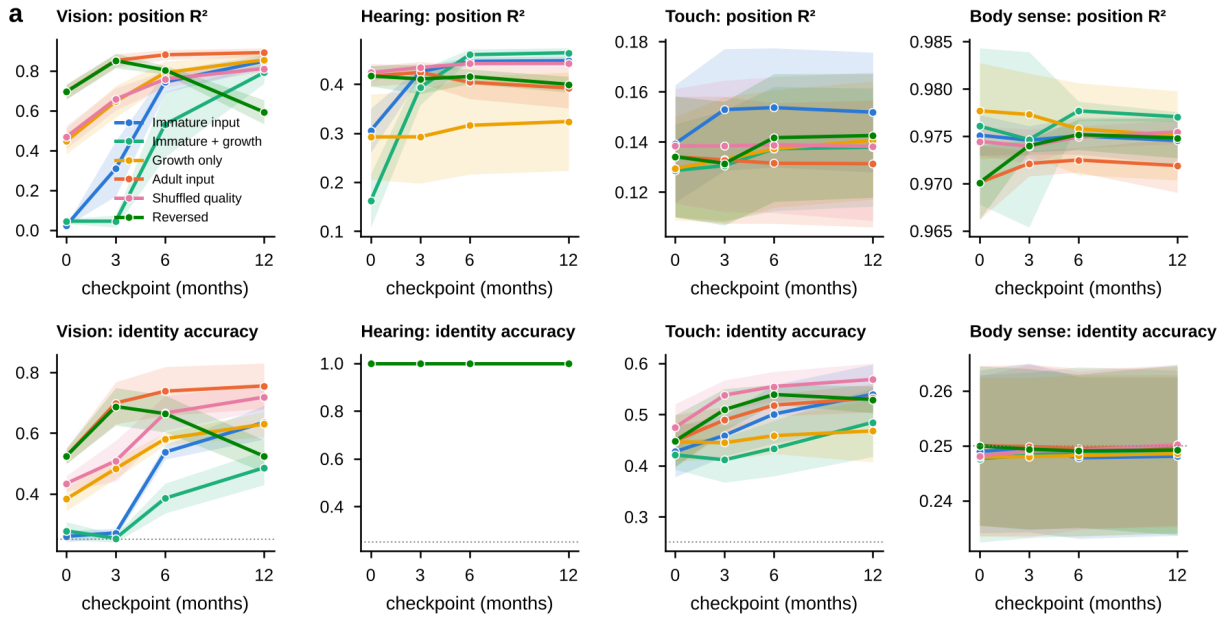


Figure S2. All senses, all regimes. Position R^2 (top) and identity accuracy (bottom; dotted, chance) per checkpoint.

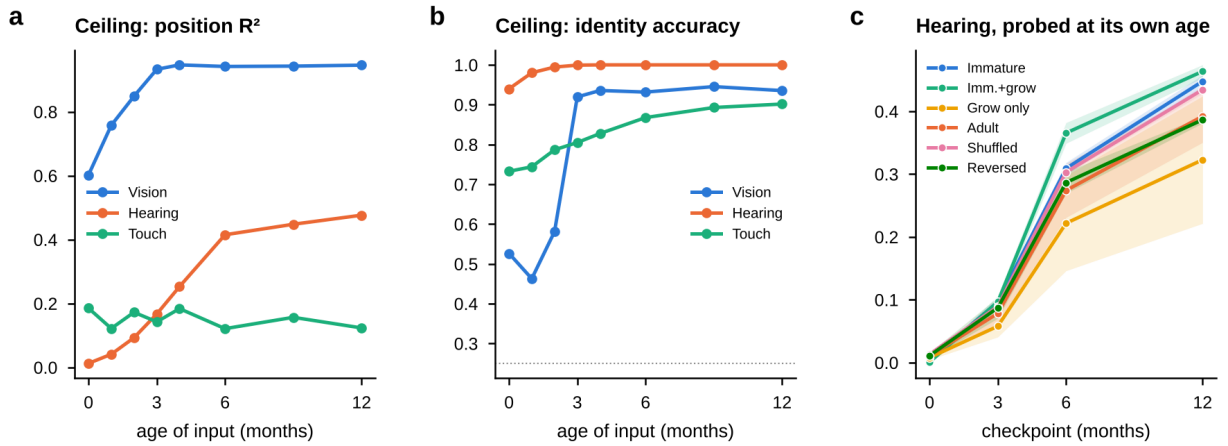


Figure S3. Information ceilings and own-age probes. The same ceiling data are plotted by age in Fig. 3 of the companion paper [24]. a, b, Supervised ceilings per sense and age. c, Hearing probed on own-age input.

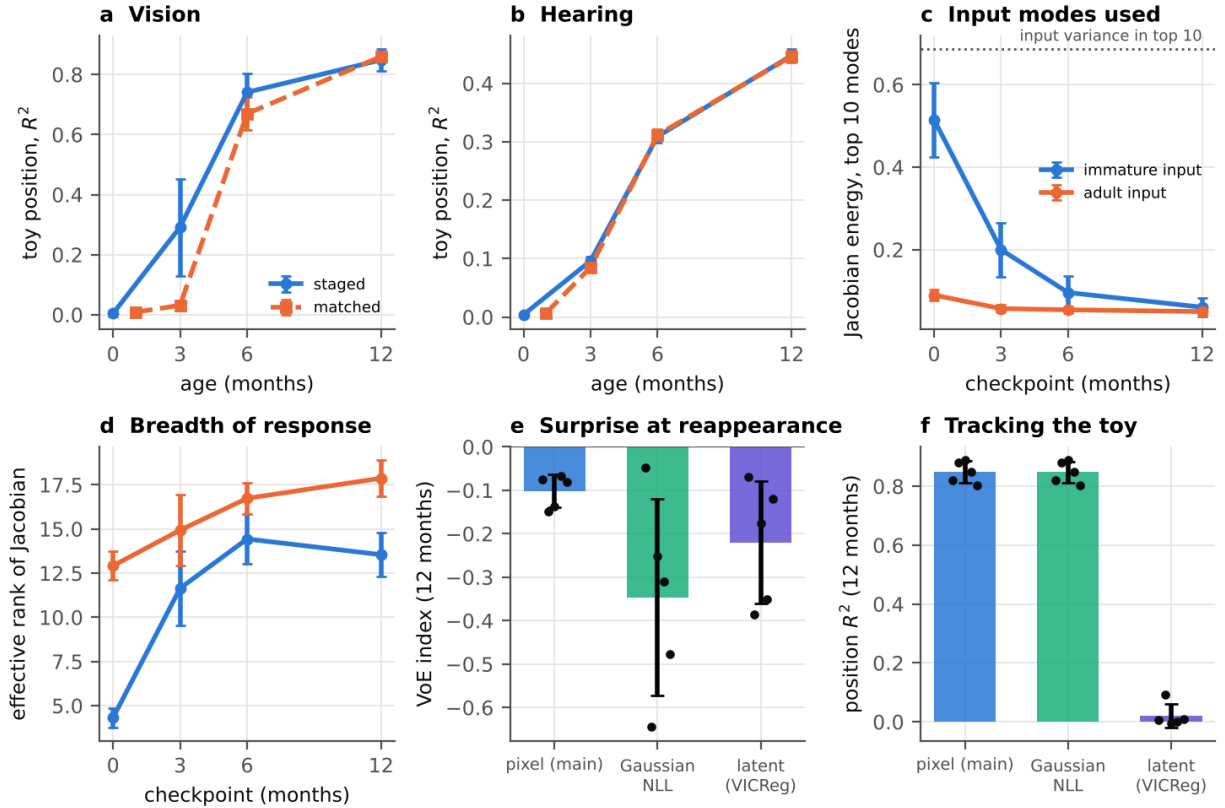


Figure S4. Follow-up experiments. **a, b**, Toy position read-out for vision (**a**) and hearing (**b**) in the staged schedule (equal updates per checkpoint) and in a single run in which experience is proportional to age (matched; first checkpoint at 1 month). **c, d**, Jacobian of the vision encoder on adult-quality frames, in the principal directions of the adult input: share of its energy in the ten leading modes (**c**; dotted line, share of the input variance in those modes) and effective rank (**d**), for the immature-input and adult-input regimes. **e**, Violation-of-expectation index at 12 months for the main pixel forecaster, a forecaster with a predicted variance (negative log-likelihood) and a latent-space forecaster. **f**, Position R^2 of the same three models. Lines and bars, mean over 5 seeds; whiskers, ± 1 s.d.; dots, seeds.