arXiv:2604.10333v2 [cs.AI] 9 Sep 2026

Zero-shot World Models Are Developmentally Efficient Learners

Khai Loong Aw, **Klemen Kotar**, **Wanhee Lee**, **Seungwoo Kim**, **Khaled Jedoui**, **Rahul Venkatesh**, **Lilian Naing Chen**, **Michael C. Frank**, **Daniel L.K. Yamins**

Stanford University
Corresponding author: khaiaw@stanford.edu

Abstract. Young children demonstrate early abilities to understand their physical world, estimating depth, motion, object coherence, interactions, and many other aspects of physical scene understanding. Children are both data-efficient and flexible cognitive systems, creating competence despite extremely limited training data, while generalizing to myriad untrained tasks – a major challenge even for today's best AI systems. Here we introduce a novel computational hypothesis for these abilities, the Zero-shot World Model (ZWM). ZWM is based on three principles: a sparse temporally-factored predictor that decouples appearance from dynamics; zero-shot estimation through approximate causal inference; and composition of inferences to build more complex abilities. We show that ZWM can be learned from the first-person experience of a single child, rapidly generating competence across multiple physical understanding benchmarks. It also shows progressive, staged emergence of capacities during learning and builds brain-like internal representations. Our work presents a blueprint for efficient and flexible learning from human-scale data, advancing both a computational account of children's early physical understanding and a path toward data-efficient AI systems.

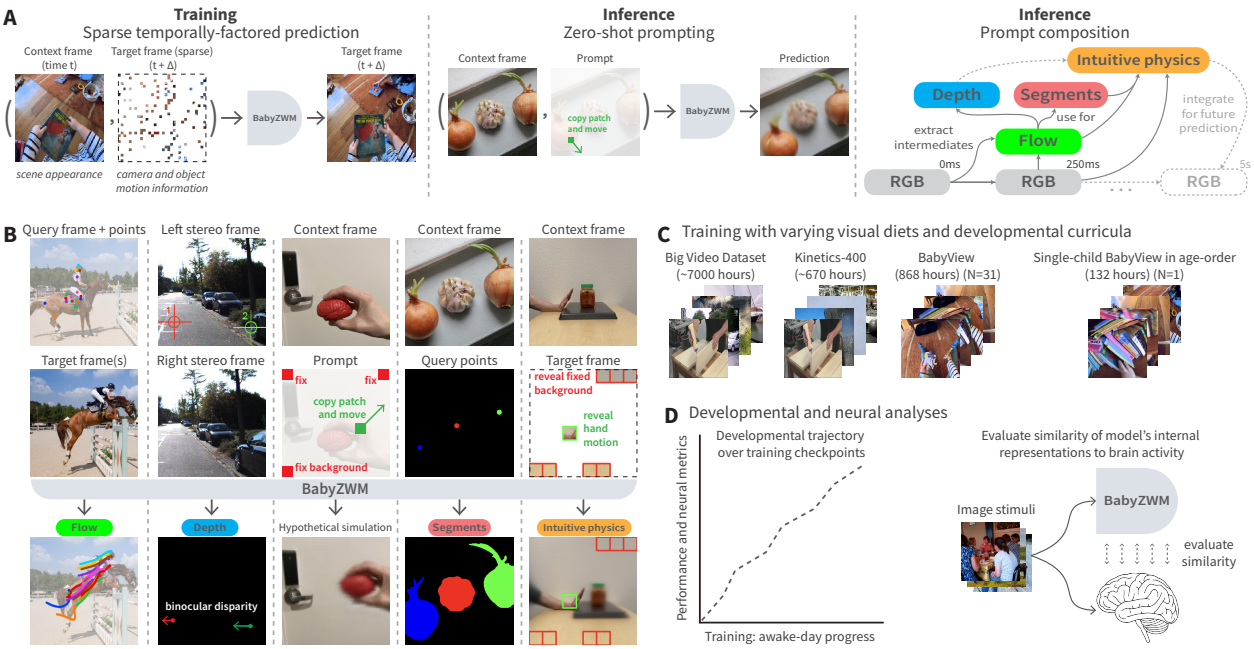


Figure 1: Overview. (**A**) The Zero-shot World Model (ZWM) framework has three design principles: temporally-factored prediction to flexibly separate appearance from dynamics; zero-shot extraction of visual-cognitive structures from the predictor through approximate causal inference; and composing extractors together to achieve increasingly complex inference abilities. (**B**) After self-supervised pretraining, ZWM can perform diverse visual-cognitive tasks zero-shot, i.e., without any additional training or examples. (**C**) We train ZWM with varying visual diets and single-child developmental curricula. (**D**) We evaluate BabyZWM's task performance across training checkpoints (developmental trajectory) and the similarity of its internal representations with brain responses.

By early childhood, humans demonstrate rich visual-cognitive abilities, exhibiting strong capacities on a diverse set of physical understanding tasks [1, 2, 3, 4, 5, 6]. The mechanisms behind these abilities are remarkably powerful, in two related but distinct senses. First, they are *data-efficient*, in that children demonstrate these capacities despite highly limited "training data" afforded by the first-person experience of a single individual. Second, they are *flexible*, supporting performance of new task abilities from an existing general-purpose representation without task-specific examples ("zero-shot"), e.g., tracking motion, estimating depth, and intuitive physics. Motivated by the sophistication and early emergence of infants' object and physical knowledge, some researchers have argued that babies have innate biases for visual cognition [3, 4, 5, 6]. However, "innateness" can mean different things, e.g., the *learning machinery* (objectives, architectures, programs) or *content* (representational primitives and concepts). Here we ask: what ingredients are required to achieve data-efficient and flexible (zero-shot) visual cognition from early experience?

We approach this question using computational models of visual learning and development. Inspired by neurophysiological observations [7, 8], deep neural networks (DNNs) have emerged as the most task-performant models for visual tasks as well as the most accurate models of neural responses across the visual cortex [9, 10, 11, 12, 13] and for human-like error patterns [14]. Initially, DNNs required supervision on large labeled datasets [15, 16] and did not transfer broadly to downstream tasks. These issues motivated a shift to self-supervised models, which learn representations by grouping similar or temporally-proximate images [17, 18, 19, 20, 21, 22, 23, 24]. Such self-supervised models also turn out to provide a description of neural and cognitive patterns in the primate visual system, rivaling or exceeding the accuracy of the earlier supervised systems [25, 26].

However, from a developmental point of view, modern self-supervised learning methods are a "glass half full". When trained on the visual data diet of real infants and children, they achieve substantial improvements compared to earlier methods such as predictive coding [25, 27, 28], but still are far from matching human abilities and perform much worse than similar methods trained on curated, highly non-natural image databases such as ImageNet [27, 25, 29, 30, 31]. It has remained very difficult for even the best visual learning algorithms from the AI literature to efficiently extract powerful representations from natural data, perhaps due to camera motion/blur, occlusions, and the relatively low diversity of environments children encounter [32, 33, 34, 35, 36]. This *ecological data learning gap* is also strikingly present for natural language processing, where large language models (LLMs) need large-scale, highly-curated data to achieve linguistic competency [37, 38, 39].

Equally important, while current self-supervised visual models may learn useful representations, they cannot perform tasks directly in the flexible zero-shot manner of humans. Instead, for each downstream task, a separate readout must be trained using labeled data, making the overall pipeline for any given task (e.g., image segmentation) ecologically implausible. By contrast, in the language domain, modern LLMs can perform diverse tasks flexibly and zero-shot – although this behavior only emerges with extremely large amounts of data. The dual challenges of data efficiency and flexibility reflect our incomplete understanding of what inductive structure to build into our models to match human cognition.

Here, we report substantial progress on the algorithmic foundations of data efficiency and flexibility, by building the Zero-shot World Model (ZWM), a self-supervised neural network that flexibly performs a broad suite of visual-cognitive tasks zero-shot, i.e., without task-specific training/examples (Figure 1B). ZWM is based on three key principles (Figure 1A). First, ZWM's underlying learned component is a *sparse temporally-factored* predictor model, a neural network that learns to make predictions given sparse and variable amounts of information. Second, after training, the model can be manipulated to yield a universal zero-shot prompting interface by comparing predictions under

ground-truth inputs to predictions under minimally modified inputs – a form of approximate causal inference. Third, ZWM composes simple prompts into more complex queries – for example, simulating hypothetical motions of objects and then computing the resulting optical flow to segment those objects – thus, building a computational graph of visual representations that progressively extracts and integrates increasingly complex structure. Taken together, these components constitute a kind of data-driven world model that is able to forecast the effect of proxy actions on the visual scene.

To test the developmental hypothesis of the ZWM framework and probe its data efficiency under naturalistic conditions, we leverage the BabyView dataset [31], a set of egocentric video recordings from young children aged $\sim$5 months to 5 years. Trained solely on 868 hours of BabyView videos ($\sim$3 months of waking experience), the BabyZWM model performs competitively with supervised, state-of-the-art models on challenging real-world datasets across a wide variety of visual cognitive tasks despite receiving no task-specific supervision or labeled probes. Beyond strong behavioral performance, BabyZWM exhibits a developmental trajectory with progressive emergence of visual-cognitive capacities, and internal representations that align with biological neural responses. Taken together, our results suggest one potential answer to the long-standing question of how general-purpose, zero-shot visual cognition can emerge from limited experience – and, simultaneously, a new approach to building data-efficient AI systems that learn flexibly from limited, uncurated data.

The ZWM framework

The Zero-shot World Model (ZWM) concept operationalizes our approach to data-efficient, zero-shot visual world modeling in three design principles (Figure 1A): temporally-factored prediction to flexibly separate appearance from dynamics; zero-shot extraction of visual-cognitive structures from the predictor through approximate causal inference; and composing extractors together to achieve increasingly complex inference abilities. Taken together, the three components of ZWM form a kind of data-driven "world model" – so-called because the system can be used to forecast the effect of proxy actions (the probes) on a scene.

Sparse temporally-factored prediction. The core learned component of the ZWM is a *sparse temporally-factored* masked multi-frame visual predictor, which we will denote by Ψ. Though the concept can be applied in longer-range many-frame videos (or even non-visual data domains), it is easiest to understand in the two-frame setting we use here (Figure 1A). Given two RGB video frames f_1 and f_2 separated by a short time gap, let Ψ_Θ be a parameterized function that seeks to predict f_2 from f_1 together with a small fraction of pixel patches from f_2. The inputs to Ψ_Θ are thus of the form $(f_1, f_2^{\text{masked}})$, where f_2^{masked} is a subset of the patches of f_2 after a mask has been applied. Ψ_Θ then outputs an estimate of the whole of f_2, denoted $\widehat{f_2}$. During training on ground truth frame pairs, the parameters Θ^* are optimized to minimize the average L2-loss of the prediction across the training dataset $\mathcal{D}$:

$$\Psi_{\Theta^*} : \left(f_1, f_2^{\text{masked}}\right) \longmapsto \widehat{f_2}; \qquad \Theta^* := \arg\min_\Theta \left\langle \|f_2 - \widehat{f_2}\|^2 \right\rangle_{(f_1,f_2)\in\mathcal{D}}. \tag{1}$$

Two critical aspects of the training of Ψ are that (i) the masks applied are very sparse – e.g. no more than 10% of the patches of f_2 are revealed; and (ii) training is performed with randomly-chosen masks, requiring no situation-specific knowledge of the semantic contents of the frames. Ψ is a type of *masked autoencoder* [40, 41], in which the training mask is temporally biased to reveal all of one frame (f_1) and very little of the other (f_2). The constraints on this prediction problem are

very generic – merely that masked training is performed, and that the masks are temporally biased. However, it turns out that this apparently weak generic constraint forces Ψ to learn a very structured representation of the visual scene. Specifically, to successfully reconstruct f_2, Ψ implicitly must: infer object *appearance* from the dense patches of f_1, as the small number of patches revealed from f_2 are insufficient to do so, while inferring object and camera *motion transformations* from the sparse revealed patches in f_2^{masked}. Ψ thus implicitly *factorizes* appearance and motion, compressing low-dimensional motion data into a compact, but naturally interpretable, set of patches.

Zero-shot extraction via approximate causal inference. A key insight of ZWM is that the highly compressed but interpretable motion patches that the training process creates can be manipulated with "zero-shot prompts" to extract key visual quantities by making the trained predictor's implicit knowledge explicitly available (Figure 1A). The core mechanism of this process is to: (i) formulate a *minimal perturbation* of a ground-truth input; and (ii) *compare* the predictor Ψ's output in both the original ground-truth case and the minimally perturbed case, and (iii) *aggregate* the difference to hone in on the quantity of interest. Formally, this process can be represented as:

$$x_\delta := \textbf{perturb}(x); \qquad \delta\Psi := \textbf{compare}(\Psi(x), \Psi(x_\delta)); \qquad \text{output} := \textbf{aggregate}(\delta\Psi). \tag{2}$$

For example, to segment an object, the **perturb** function can simply induce hypothetical motion by translating one small patch on the object, causing the predictor Ψ to propagate hypothetical motion to the rest of the object, but not other components in the scene; the **compare** function computes the optical flow between the perturbed and unperturbed cases; and **aggregate** thresholds the flow to determine which pixels belong to the object. We show that a wide variety of visual concepts can be extracted in a zero-shot manner from Ψ by choosing different but extremely simple **perturb**, **compare**, and **aggregate** functions.

ZWM's approach to zero-shot extraction is a form of *approximate causal inference*. As discussed in the causality literature [42, 43], the process of causal inference asks how an outcome of a dynamic process changes when a minimal change is made to its antecedents. Analogously, the Ψ function acts as a learned structural equation for the world's dynamics, whose temporally-factored nature permits the construction of minimal perturbations that expose some aspect of the causal structure of the world. For example, ZWM's object segmentation procedure uses a motion perturbation to expose the underlying causal structure of the world — groups of pixels move together due to the latent *cause* of belonging to the same physical object.

Compositional prompting. ZWM composes simple prompts to construct more complex queries, progressively extracting and integrating increasingly abstract visual structures, e.g., motion and objects, rather than RGB pixels (Figure 1A). ZWM (i) estimates optical flow from RGB; (ii) computes optical flow on binocular views for relative depth; (iii) simulates hypothetical motions and computes optical flow to segment objects; and (iv) uses flow and segments for intuitive physics. Consequently, this composition builds a computational graph of visual intermediates.

Model implementation. We implement the predictor Ψ as a neural network, and perform learning via stochastic gradient descent on its parameters. The base network architecture is a Vision Transformer (ViT) backbone [44], with versions at two sizes (170 million and 1 billion parameters). We compare models trained on a variety of datasets, as described below. In each case, training datapoints consist of RGB frame pairs taken from a real-world video distribution, with an inter-frame gap sampled uniformly between 150ms and 450ms. The images are input into the model as square

256x256 pixel arrays, and then patchified into 8x8-pixel patches. During training, masks are chosen randomly on each example, with 10% of patches in the second frame revealed.

Results

ZWM performs diverse visual-cognitive tasks zero-shot

How well can ZWM flexibly perform a broad suite of visual-cognitive tasks zero-shot? We evaluate a spectrum of visual-cognitive tasks that humans perform, from lower- to higher-level, including optical flow, relative depth estimation, object segmentation, and intuitive physical reasoning. To test robustness across visual diets, we train ZWM on **BabyView** (N=34, 868 hours, 2025.1 release) [31] (BabyZWM), **Kinetics-400** [45] ($\sim$670 hours) (smaller than BabyView but far more diverse Internet videos), and a **Big Video Dataset** (BVD) [46] ($\sim$7000 hours; computer vision datasets and Internet videos; an approximate upper bound on what can be achieved with high visual diversity and scale).

We evaluate ZWM against a range of alternative hypotheses, including both representation-based models and task-specific systems. Representation-based systems learn general-purpose visual features from pretraining that are typically transferred to downstream tasks via finetuning or lightweight task-specific heads/objectives. Unlike ZWM, these models are not natively zero-shot, a key limitation for modeling human vision, so we design simple zero-shot probes to evaluate these models. We sweep each baseline's evaluation hyperparameters (e.g., feature layer and thresholds), reporting its best configuration (Supplementary Tables S7, S9, S17). As a **"standard" supervised static image model,** we evaluate ResNet-50 trained on ImageNet-1k with category-label supervision. For a **task-generic self-supervised static image model,** we evaluate DINOv3 [47] (off-the-shelf and trained on BabyView), which learns strong single-image representations by training the model to produce consistent features across different views of the same image. As an example of a **task-generic self-supervised video model**, we evaluate V-JEPA2 [24] (off-the-shelf and trained on BabyView), a model that learns by predicting masked regions of a video in feature space rather than in raw pixels.

Poor results across visual-cognitive tasks for these representation-based comparison models do not imply they do not develop useful and powerful visual-cognitive representations, but rather that these models provide no means of accessing these representations zero-shot. We therefore also benchmark ZWM against state-of-the-art task-specific baselines, including models trained directly for individual benchmarks (e.g., supervised networks optimized for flow, depth, or segmentation). These baselines can be viewed as concrete instantiations of the alternative hypothesis that human-like competence is achieved via separate, specialized systems rather than a unified world model. Due to the paucity of benchmarks for direct model-to-human comparisons, and because humans would be expected to perform near ceiling on these everyday visual tasks, we instead treat supervised state-of-the-art systems as a strong proxy baseline. Outperforming them therefore provides a strong test of BabyZWM's zero-shot data efficiency.

We evaluate on established benchmarks designed to be challenging (real-world motion, occlusions, and lighting changes). For fair comparisons, we provide all models with the same inputs. We describe detailed methods for each task in the Supplementary Materials.

Optical flow. To rigorously evaluate algorithm performance on optical flow, we use several recent computer vision benchmarks: TAP-Vid-DAVIS [48], consisting of challenging real-world videos

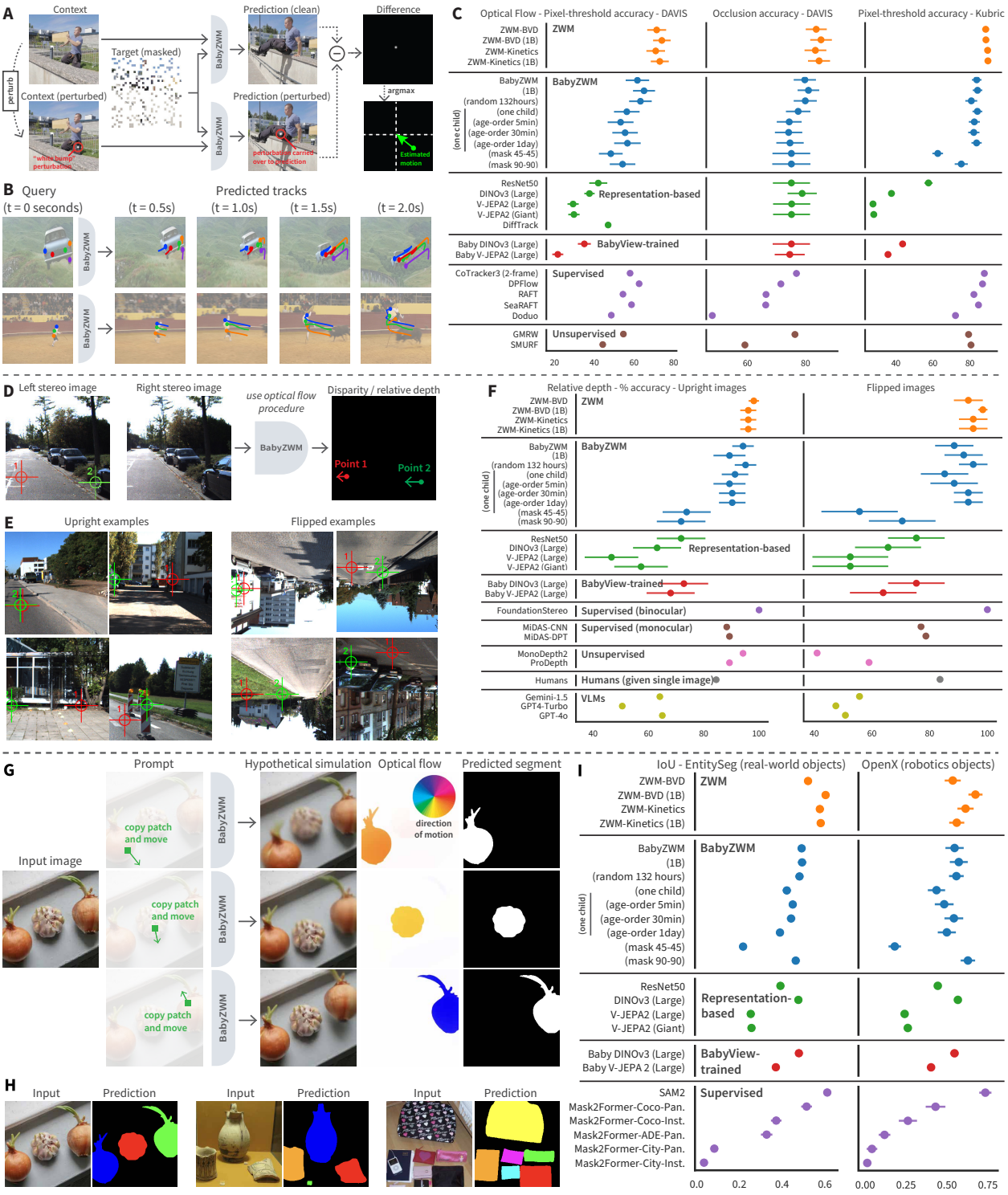


Figure 2: BabyZWM estimates optical flow, relative depth, and object segments zero-shot. (**A**) Optical flow method. (**B**) Flow predictions as tracks. (**C**) BabyZWM matches state-of-the-art optical flow. (**D**) Relative depth method (built on optical flow). (**E**) Benchmark examples (upright and flipped). (**F**) BabyZWM beats supervised monocular depth models. (**G**) Object segmentation and hypothetical motion procedure. (**H**) Segmentation predictions. (**I**) BabyZWM matches supervised segmenters, except SAM2. Error bars indicate bootstrap 95% intervals throughout the paper. Representation baselines are each read at their best configuration over a layer $\times$ threshold sweep (Supplementary Table S17).

with human-annotated "ground-truth"; and TAP-Vid-Kubric [49], consisting of synthetic, simulator-generated videos where ground-truth flows are known by construction. To enable controlled comparisons, we pass in two frames (query, target) to all models. We measure pixel-threshold accuracy (percentage of predictions within a pixel-radius of ground truth) and occlusion/out-of-frame detection accuracy. ZWM achieves state-of-the-art results (Figure 2C); BabyZWM is competitive with label-supervised CoTracker3 [50], DPFlow [51], and SeaRAFT [52] on TAP-Vid-DAVIS and matches supervised baselines at detecting occlusions. On TAP-Vid-Kubric, BabyZWM is strong but slightly below supervised models (which use synthetic training). BabyZWM outperforms the DINOv3 and V-JEPA2 models.

Relative depth estimation. We evaluate depth perception on UniQA-3D [53]: point pairs that require judging which is further. Depth is extracted zero-shot from ZWM by computing optical flow between stereo images (Figure 2D). Both ZWM and BabyZWM exceed 90% accuracy (Figure 2F). They surpass large vision–language models (Gemini-1.5, GPT-4-Turbo, GPT-4o) [54, 55, 56], are comparable to supervised (MiDaS-CNN [57]) and self-supervised (MonoDepth2 [58]) monocular estimators, and trail only a supervised binocular model [59].

Object discovery. We evaluate object segmentation on SpelkeBench [60], a class-agnostic benchmark defining objects as distinct, bounded entities. On this benchmark, BabyZWM rivals supervised Mask2Former variants [61] trained on large-scale COCO [62], though it performs slightly below SAM2 [63] which leverages large-scale human annotations (Figure 2I).

Intuitive physical understanding. We develop a novel short-timescale physical reasoning benchmark (Figure 3A) to evaluate models, featuring interactions between a hand and 1–2 objects that test 5 categories of reasoning: object cohesion, support relations with motion of either the top or bottom object, force transfer, and force separation. A prediction is counted as correct if it is closer to the ground-truth than the context frame, using mean squared error and LPIPS perceptual similarity [64] (Methods). ZWM, BabyZWM and V-JEPA2 approach 100% performance on almost all categories, but not Baby V-JEPA2 (Figure 3C). We apply model interpretability techniques to BabyZWM, revealing several attention heads that consistently follow the hand (causal agent) when predicting the motion of the object of interest (Figure 3B; Supplementary Figure S1). Randomly selected qualitative predictions for every category are shown in Supplementary Figure S2. The benchmark does not test object permanence, which requires multi-frame memory, nor mass sensitivity, since a hand-driven push moves objects regardless of mass. A preliminary qualitative test of the latter is reported in the Supplement (Figure S3).

ZWM achieves data efficiency and continual learning

A correct theory of visual-cognitive learning must be able to achieve effective learning on the real datastreams that a human experiences.

BabyZWM retains most of its performance compared to the same architecture trained on much more diverse datasets, such as Kinetics-400 and BVD (Figures 2, 3), emphasizing the data-efficiency of the ZWM architecture.

To perform an even more stringent test, we next trained ZWM on **Single-Child BabyView**, a subset of BabyView consisting of 132 hours of recordings from a single individual (age 9–30 months). The Single-Child dataset represents a stricter test for model learning, because it requires algorithms to

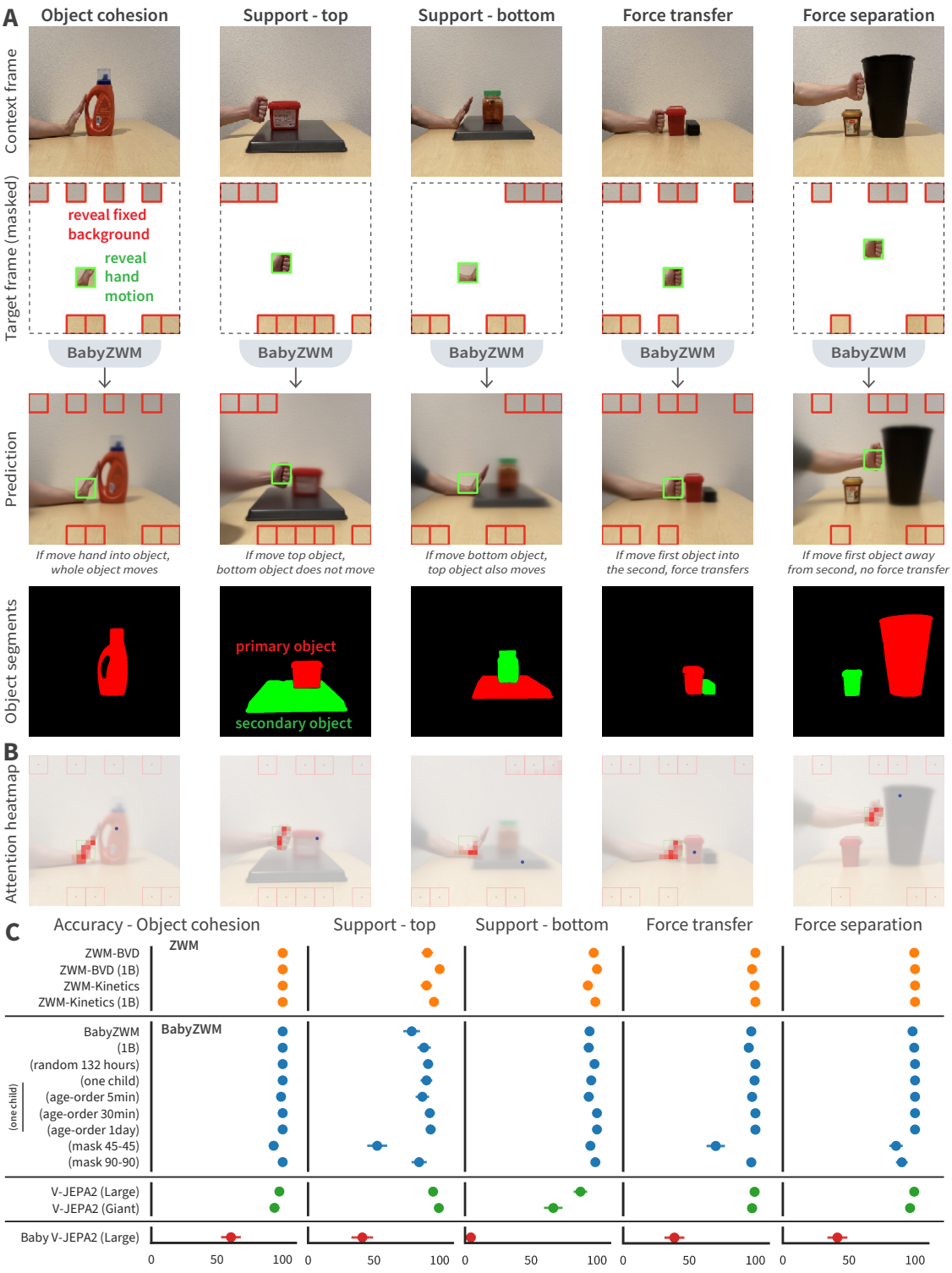


Figure 3: BabyZWM exhibits object knowledge and intuitive physics. (**A**) Short-timescale benchmark testing cohesion, support (move top/bottom object), force transfer, and force separation; given a context frame and a few target patches, predict remaining patches. We reveal the hand's target location, and fixed border patches to signal a static camera. (**B**) Interpretability methods reveal attention heads that track the hand (causal agent) when predicting the target object (marked with a blue point). (**C**) ZWM, BabyZWM, and V-JEPA2 near 100% on most categories; Baby V-JEPA2 does not.

be able to learn generalizable capacities from the highly restricted visual diversity of one child's experience. (We additionally train on a random 132-hour subset, allowing us to disentangle the contributions of diversity versus total exposure.) Single-Child BabyZWM performs similarly to BabyZWM across most tasks (Figures 2, 3).

Additionally, we trained a version of Single-Child BabyZWM on a single pass through a version of the data in which the video clips were ordered by the child's age. This is an important test of developmental robustness and continual/life-long learning, a setting also studied outside of developmental contexts [65]. We create a set of curricula by shuffling within various temporal windows (5 minutes, 30 minutes, and 1 day), loosely approximating different degrees of experience consolidation (e.g., within-episode mixing vs. sleep-like reordering). The age-ordered Single-Child BabyZWM models performed similarly to Single-Child BabyZWM across all tasks (Figures 2, 3).

Finally, "Standard" BabyZWM uses asymmetric masking (fully visible f_1, 90% masked f_2), explicitly prioritizing the learning of motion dynamics. Because this temporally-factored mask structure contributes a conceptually core component of the ZWM concept, we explore simpler alternatives. We evaluate symmetric masking variants of BabyZWM (mask 45%-45% and mask 90%-90%), which perform substantially worse (Figures 2, 4), showing that emphasizing motion information is useful for data efficiency and zero-shot abstraction.

Representation and readout are both important

A natural question is whether ZWM's zero-shot performance is due to a better learned representation or a better readout mechanism. To answer this question, we crossed the two training objectives (temporally-factored asymmetric masking vs. symmetric masking) with two readout mechanisms (our approximate causal inference vs. the label-free feature nearest-neighbor probe used for the representation baselines). This yields four settings, evaluated on optical flow, relative depth, and object segmentation (Figure 4; Supplementary Table S18). Notably, asymmetric masked autoencoding read out by feature-matching is the setting closest to prior work, in particular SiamMAE [66].

Neither factor alone accounts for ZWM's performance: the full combination is the best of the four settings on every task, and degrading either factor costs performance on every task. Read with the feature nearest-neighbor probe, the BabyZWM representation performs at roughly the level of state-of-the-art self-supervised baselines – above DINOv3 and V-JEPA2 on flow and depth, but below DINOv3 on segmentation. Yet, when read with our approximate causal-inference mechanism, BabyZWM alone becomes competitive with state-of-the-art *task-specialized supervised* solutions (Figure 2).

We investigated the specific mechanism underlying this result (Figure 4 B-C). We compared a symmetric-trained model given the same total amount of information as the asymmetric ZWM objective, but in which the information was distributed differently (that is, symmetrically) during training. This is closer to the "usual" VideoMAE masking objective [21]. The representation thus trained is similar in nearest-neighbor geometry, but has drastically worse causal-readout performance. This is because, since the symmetric objective does not concentrate dynamics information in sparse tokens, minimal interventions are less effective, causal inference is heavily confounded, and the readout quantities are poor.

The temporally-factored objective therefore builds the correspondence structure that the causal-

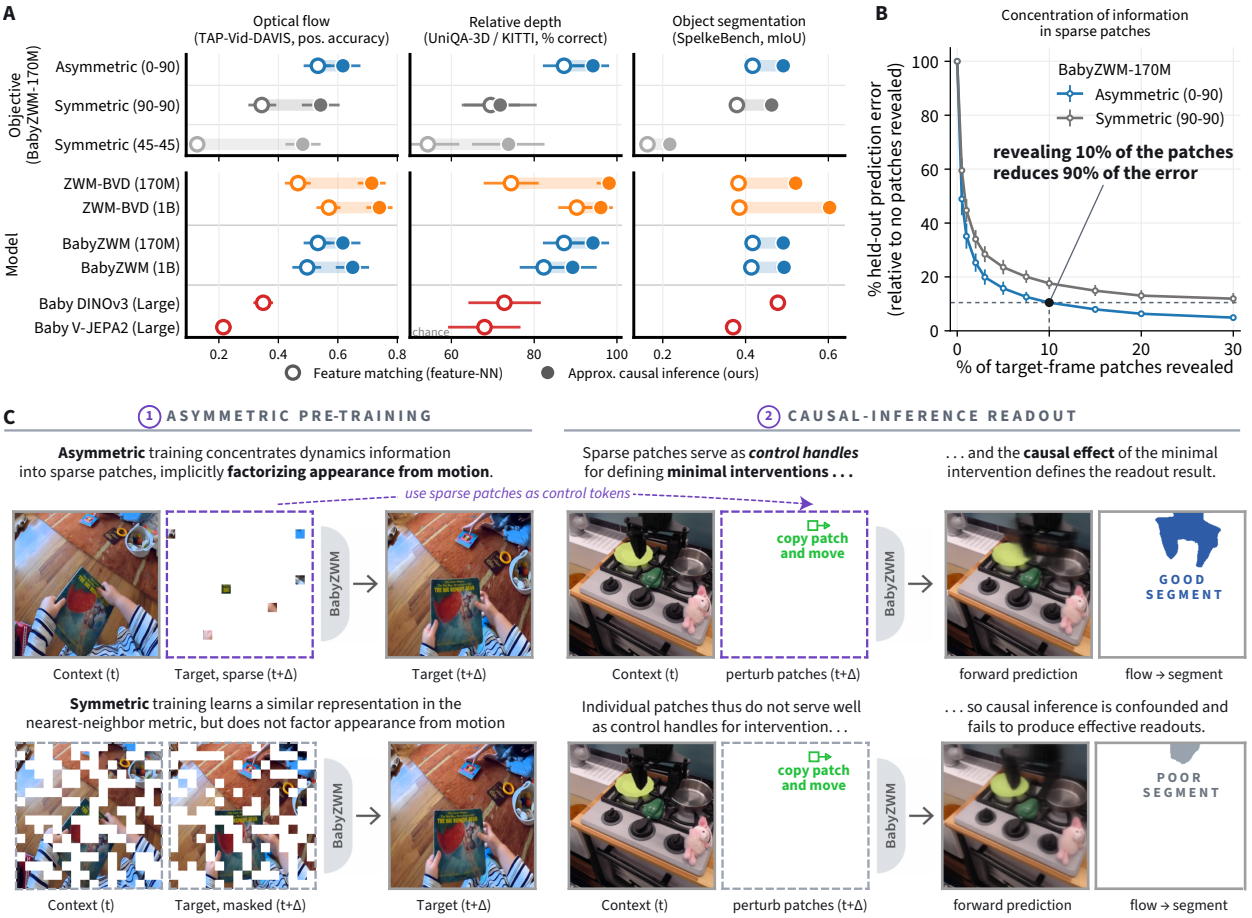


Figure 4: Representation and readout are both important to ZWM's success. (**A**) Objective × readout for BabyZWM-170M (top), and readouts across models (bottom). Both representation-learning choices and read-out method matter; degrading either costs performance on every task. (**B**) With the context frame fully visible, we reveal different proportions of target-frame patches and score prediction error on 20% of patches never revealed – across 100 Kinetics-400 videos (Supplementary Table S20). Revealing 10% of patches reduces 90% of the error – sparse patches carry most of the target frame's predictive information (how camera and objects move). Asymmetric training concentrates this information more than the symmetric ablation. (**C**) (*Top row*) Asymmetric (ZWM) versus symmetric training objective. ZWM's temporal-factoring asymmetric objective concentrates dynamics information into sparse patches, which can then be used as control tokens to create minimal interventions for powering causal-inference readout. (*Bottom row*) Symmetric control training objective with the same amount of information does *not* produce usable control handles for minimal interventions, confounding causal inference.

inference readout exploits, and the readout is what converts that structure into task performance. Under this view the readout is part of the model in a deep way – a point that matters especially for a model of child development, where no task labels are available to train a readout.

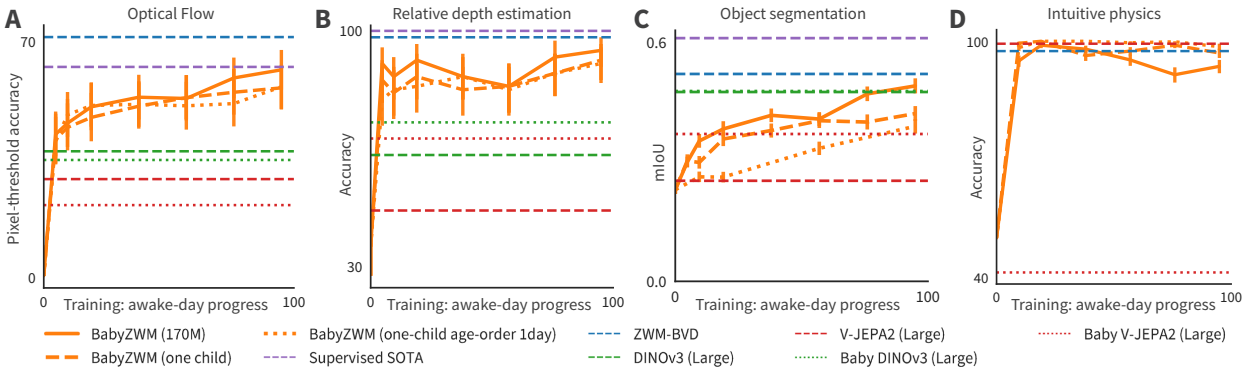


Figure 5: BabyZWM develops zero-shot capacities across training checkpoints. We plot the developmental trajectories of BabyZWM, Single-Child BabyZWM, and Single-Child BabyZWM (age-order, shuffle within each day) to observe how fast different visual-cognitive capacities emerge. We evaluate these models across a full training run, which corresponds to roughly 95 days of waking experience assuming ∼10 awake hours/day [67]. We also compare these to ZWM trained on BVD, supervised state-of-the-art baselines, and other alternative hypotheses. We plot these developmental trajectories for (**A**) optical flow, (**B**) relative depth estimation, (**C**) object segmentation, and (**D**) intuitive physical reasoning.

Progressive, staged emergence of visual-cognitive capacities

Having evaluated the BabyZWM models, we next ask what we can learn from looking at their developmental trajectories and when different visual-cognitive capacities emerge.

BabyZWM's optical flow and object segmentation capabilities increase across the entire training period, whereas depth and intuitive physical capacities rise steeply and peak early (Figure 5). We measure each capacity's emergence over its *learnable gain* – the rise from a floor (chance for the two-alternative depth and physics benchmarks, the value at initialization otherwise) to its maximum across checkpoints. We report the percentage of training needed to reach 80% of that gain (also 50%; Supplementary Table S23). Relative depth emerges earliest (4.6% of training, to 80% gain), with intuitive physics next (8.7%), optical flow later (17.4%), and object segmentation latest (69.8%).

These comparisons should be interpreted cautiously: they partly reflect benchmark-specific design choices – especially differences in task difficulty, metrics, and ceiling effects – rather than a clean ordering of underlying capability development. What the model does supply is quantitative, testable predictions of the rate at which each capacity emerges. These emergence dynamics can be compared against the order and timing of capacity development in children – object tracking [68, 69], early stereopsis [70, 71, 72, 73], object segmentation and intuitive physics [74, 75, 76, 77]. However, adjudicating that comparison requires systematic, matched developmental benchmarks for humans and models, which is an important future direction.

ZWM representations align with neural responses

Having characterized ZWM's behavioral trajectories, we next ask whether ZWM also develops brain-like internal representations. The human visual system is organized hierarchically, transforming retinal inputs into increasingly complex representations [79, 80, 81] that develop over childhood [82, 83, 84, 85, 86]. We evaluate the similarity of our models' internal representations (across

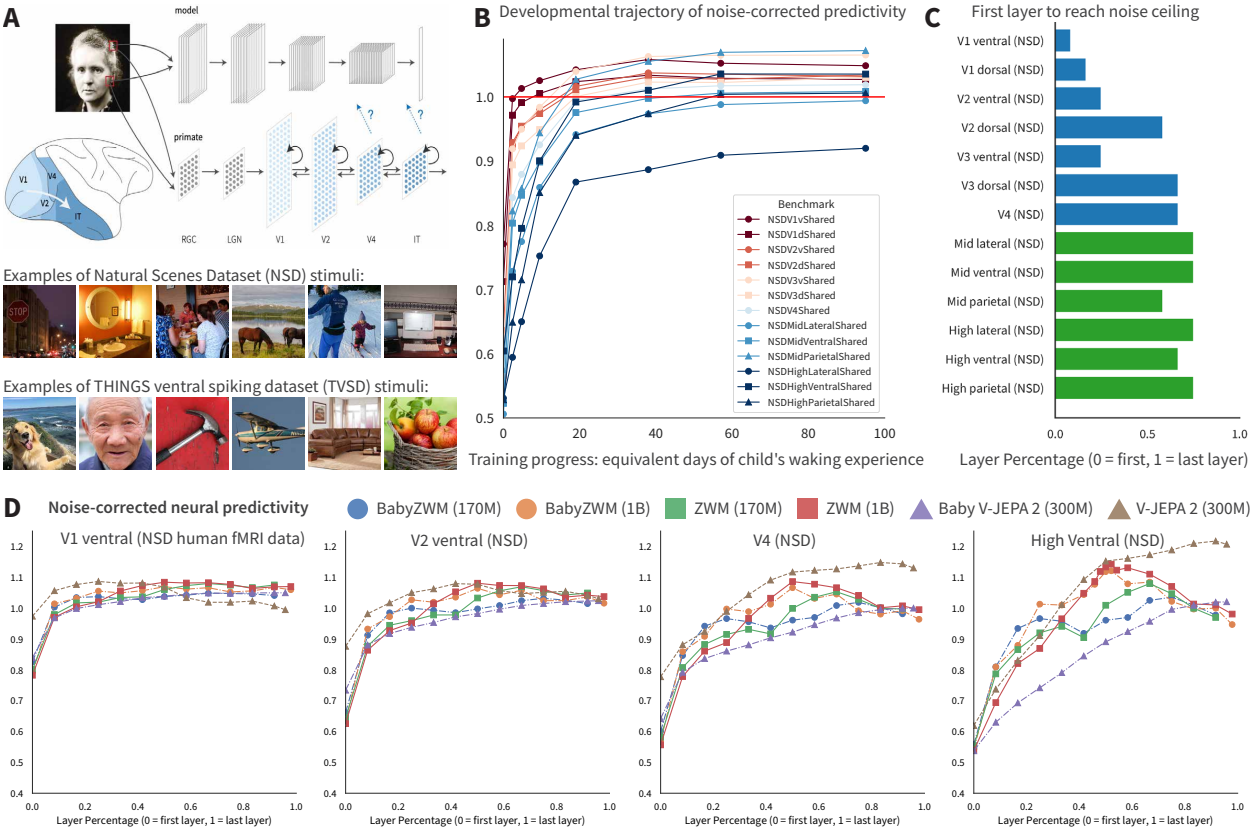


Figure 6: BabyZWM successfully develops internal representations that align with neural responses from human fMRI and macaque electrophysiology datasets. (**A**) Neural predictivity schematic [78], with example images from NSD and TVSD. (**B**) Developmental trajectory: BabyZWM's neural predictivity for early visual areas increases quickly in training, while it takes longer for the later areas, exhibiting an "early-first" developmental trajectory. We observe this both in the steeper slope for neural predictivity of V1 than higher regions, as well as neural predictivity reaching V1's noise ceiling at an earlier checkpoint. (**C**) For various visual areas in the brain, we plot the first layer of BabyZWM that reaches the noise ceiling. For earlier cortical regions, earlier model layers reach the noise ceiling, whereas later cortical regions align with deeper layers. This exhibits neuroanatomical consistency with several accounts of hierarchical visual organization. (**D**) Detailed plots for noise-corrected neural predictivity for the ventral stream for NSD human fMRI.

various training checkpoints) with brain responses by computing neural predictivity [9, 10, 11]: fit a cross-validated linear probe from model features to neural responses, then report noise-corrected correlations (Figure 6A). We evaluate two complementary benchmarks: the Natural Scenes Dataset (NSD) [87] for human fMRI and the THINGS Ventral Stream Spiking Dataset (TVSD) [88] for macaque electrophysiology. fMRI captures large-scale representational geometry; electrophysiology reveals fine-grained single-neuron tuning/timing.

Across NSD and TVSD, the BabyZWM model exhibits neural alignment consistent with hierarchical visual development. Neural predictivity for early visual cortex approaches its noise ceiling at relatively early training checkpoints, whereas higher regions improve more gradually, an "early-first" developmental trajectory (Figure 6B). Layer–area correspondence is hierarchically aligned: for earlier cortical regions, earlier model layers reach the noise ceiling, whereas later cortical regions align with deeper layers (Figure 6C). This pattern is consistent with several accounts of hierarchical visual organization [79, 89, 80, 81] and prior modeling findings [9, 10], supporting an explicit "mechanistic mapping" between model layers and cortical regions [90, 91, 92]. A single, self-supervised world model thus captures representational structure shared across species and measurement scales, and BabyZWM recapitulates brain-like signatures of both developmental dynamics and hierarchical organization.

The ZWM objective is also robust to a developmental data diet. We quantified the change in peak neural predictivity when an architecture is trained on BabyView instead of its standard training set (Figure 6D; Supplementary Table S24). Averaged over higher-order visual areas (NSD V4, mid- and high-ventral; macaque V4 and IT), peak predictivity drops by 10.4% for V-JEPA2 but only 2.1% (170M) and 1.5% (1B) for ZWM – a five-fold smaller "baby-data" penalty. Furthermore, the V-JEPA2 penalty grows monotonically along the hierarchy (-3.4% V1, -16.2% high ventral) while ZWM's stays small and roughly flat across areas.

Expanded layer-by-region profiles for both benchmarks, baseline comparisons for the layer-area correspondence and the developmental trajectory, and per-area converged predictivity for every model, are given in the Supplement (Supplementary Figures S4, S5, S6; Supplementary Tables S24, S25). An exciting future direction is to extract features while ZWM runs its causal-inference readouts, rather than from static images, which may yield stronger/distinct neural predictions — though this likely needs neural datasets with dynamic stimuli that engage these inferences.

Discussion

Modern visual learning algorithms are highly data inefficient when compared to humans, experiencing substantial performance gaps when trained on the real datastreams experienced by human children [27, 25, 29, 30, 31]. Here, we describe a novel approach to visual learning, Zero-shot World Modeling, that is able to bridge this gap. To acquire diverse visual-cognitive capacities without labels, ZWM represents a shift from the dominant paradigm of representation learning with task-specific readouts to unified, zero-shot world models. In representation learning, each downstream task needs its own labeled readout [17, 18, 19, 20, 21, 22, 23, 24], limiting the range of feasible tasks and encouraging overfitting to sparse labels. In contrast, ZWM achieves zero-shot, out-of-distribution generalization to challenging real-world scenes, synthetic simulations, and flipped images. Moreover, ZWM gains competence even when trained with limited data from one individual child, presented in an online single-epoch fashion.

Beyond its implications for cognitive science, ZWM's zero-shot capability addresses a pressing

challenge in AI. Current self-supervised visual models, despite learning rich representations, remain locked behind task-specific labeled readouts – an expensive and brittle dependency that limits practical deployment. ZWM eliminates this bottleneck: a single learned predictor yields optical flow, depth, segmentation, and physical reasoning zero-shot, through a universal interface. This mirrors the paradigm shift in NLP when LLMs replaced task-specific fine-tuned models – but ZWM achieves this in vision with orders of magnitude less data. That this is achievable from just hundreds of hours of a single child's naturalistic, uncurated video – rather than millions of hours of curated internet data – suggests that the right inductive structure can dramatically reduce the data requirements for broad visual competence. This has direct relevance for domains such as robotics, medical imaging, and embodied AI, where large-scale labeled data is unavailable.

ZWM is a natural hybrid between two polar concepts of the role of intermediate structure in cognition and learning. The first is a "pure learning" alternative, embodied by Richard Sutton's Bitter Lesson [93] – that complex hand-built inductive biases are unnecessary in formulating effective learning machines. The second is the idea, emanating from computational cognitive science, that human learning is best understood as embodying strong priors about the world, citing the sophistication and early emergence of infants' object and physical knowledge [3, 4, 5, 6] and poverty-of-stimulus claims that children's input is too noisy to support learning [94, 95]. The ZWM principles draw on both of these ideas, illustrating how explicit structure can be created within a minimally-biased learned network. Scale suffices when data is abundant, but when it is finite, as in developmental experience, built-in structure is the right design. Such structure can itself be found by search and learning (possibly instantiated by evolution), making it a product, rather than a violation, of the Bitter Lesson.

The fact that ZWM can implement this hybrid, and the observation that doing so leads to substantial gains in learning efficiency, has implications for the long-standing debate between developmental nativism and empiricism. Specifically, ZWM instantiates a form of "architectural nativism" [96] that includes readout programs: a small, fixed scaffold is innate – architecture, learning algorithm, and task-specific readout programs – while the representational content and network parameters are learned from experience. The readouts are "zero-shot": they specify *how* to perform a task without supplying labels/examples. Any readout—including the feature-matching probes applied to the representation baselines—encodes an instruction of this kind. Importantly, our results provide proof-of-concept validation that this mechanism supports acquisition of visual-cognitive capacities and object- and physics-like representations from naturalistic visual experience, challenging strong nativist accounts that posit extensive innate biases for representational content and concepts.

Under this interpretation, zero-shot readouts may correspond to evolutionarily-specified, hard-wired neural circuits that map learned dynamics to visual-cognitive percepts. This scaffold is compact enough to pass through the genomic bottleneck (compact programs, not stored knowledge) [97, 98]. Future work can explore if they might alternatively be learned during development as flexible adapters, or constructed online as query-like cognitive inference routines over the learned predictor.

ZWM achieves zero-shot visual cognition by being a world model, which forecasts the consequences of actions. This concept has a long tradition within model-based reinforcement learning [99, 100] and model-predictive control [101, 102, 103, 104, 105, 106, 107, 108, 109]. It might at first seem odd that we discuss ZWM as a world model—after all, the inputs to Ψ are just data, so where are the actions whose consequences are to be forecasted? ZWM is a "data-driven world model", in which expensive-to-obtain true action data is proxied by cheap data (e.g. pixel-patch) operations that approximate simple actions – the "tracers" and "motions" used to create hypotheticals for computing flow, object segments, etc. ZWM formally treats data patches in the same way that true

action data would be, and reaps the reward of doing so, because training on raw data enables the underlying model to learn enough about the way the world works that it can competently perform hypotheticals. Future work could seek to learn an interactive *policy* for choosing such "actions", setting up comparisons to observed child hand and head motions captured in the BabyView dataset.

Our present work has a number of important limitations. First, by focusing on physically-grounded quantities that are learned by very young infants, ZWM leaves unaddressed how semantic concepts – e.g. named linguistic categories of objects, relationships, activities – arise developmentally. We hope that future work will integrate the world model learned by ZWM with the rich linguistic/auditory data experienced by children. Second, a core empirical limitation of the present work is the paucity of detailed developmental behavioral and neural comparisons. Such datasets are very challenging to produce and will require concerted collaborative efforts. Finally, as a deterministic regression model, ZWM's Ψ predictor is subject to *mode collapse*, leading to blurry predictions in situations in which there is underlying uncertainty about how the future will resolve. This design limits our ability to study longer-horizon prediction and control; extending to multi-frame training, richer temporal memory, and long-horizon tasks (e.g., object permanence) is an important next step [46].

An intriguing line for future work is *integrating* ZWM's zero-shot task extractions into its underlying predictor Ψ, so Ψ can be conditioned on, and make predictions of, these intermediate quantities. Recent work in world modeling has suggested a possible mechanism [46, 110, 111], creating a bootstrapping cycle in which every additional intermediate could contribute a learnable target for enriching the predictor, in turn enabling increasingly efficient learning and the identification of more sophisticated intermediates. This might pave the way for even more flexible, data-efficient learning.

Acknowledgments

We are very grateful to Cameron Ellis, Hyowon Gweon, James (Jay) McClelland, Cliona O'Doherty, and Alison Gopnik for helpful feedback on our manuscript.

Funding: This work was supported by the following awards. D.L.K.Y.: Simons Foundation grant 543061, National Science Foundation CAREER grant 1844724, National Science Foundation Grant NCS-FR 2123963, Office of Naval Research grant S5122, ONR MURI 00010802, ONR MURI S5847, and ONR MURI 1141386 - 493027. We also thank the Stanford HAI, Stanford Data Sciences, Stanford Marlowe team, and the Google TPU Research Cloud team for computing support.

Author contributions: K.L.A., M.C.F., and D.L.K.Y. designed research, analyzed data, and wrote the paper; K.L.A., K.K., and W.L. implemented and trained models; S.K. implemented optical flow algorithms and analyses; K.J. implemented neural predictivity algorithms and analyses; L.N.C. and R.V. implemented object segmentation algorithms and analyses.

Competing interests: There are no competing interests to declare.

Data and materials availability: Code for reproducible model training and evaluation is available at `https://github.com/awwkl/ZWM`. Trained model checkpoints and our intuitive-physics benchmark are released at `https://huggingface.co/awwkl`. BabyView is available to researchers through Databrary at `https://www.databrary.org/volume/1882`.

References

[1] Philip J. Kellman and Elizabeth S. Spelke. Perception of partly occluded objects in infancy. *Cognitive Psychology*, 15(4):483–524, October 1983. ISSN 00100285. doi: 10.1016/0010-0285(83)90017-8. URL https://linkinghub.elsevier.com/retrieve/pii/0010028583900178.

[2] Renée Baillargeon, Elizabeth S. Spelke, and Stanley Wasserman. Object permanence in five-month-old infants. *Cognition*, 20(3):191–208, January 1985. ISSN 00100277. doi: 10.1016/0010-0277(85)90008-3. URL https://linkinghub.elsevier.com/retrieve/pii/0010027785900083.

[3] Elizabeth S. Spelke, Karen Breinlinger, Janet Macomber, and Kristen Jacobson. Origins of knowledge. *Psychological Review*, 99(4):605–632, 1992. ISSN 1939-1471, 0033-295X. doi: 10.1037/0033-295X.99.4.605. URL https://doi.apa.org/doi/10.1037/0033-295X.99.4.605.

[4] Elizabeth S. Spelke. Core knowledge. *American Psychologist*, 55(11):1233–1243, November 2000. ISSN 1935-990X, 0003-066X. doi: 10.1037/0003-066X.55.11.1233. URL https://doi.apa.org/doi/10.1037/0003-066X.55.11.1233.

[5] Susan Carey. *The origin of concepts*. Oxford series in cognitive development. Oxford University Press, Oxford ; New York, 2009. ISBN 978-0-19-536763-8.

[6] Elizabeth S. Spelke. *What Babies Know: Core Knowledge and Composition Volume 1*. Oxford University PressNew York, 1 edition, November 2022. ISBN 978-0-19-061824-7 978-0-19-061825-4. doi: 10.1093/oso/9780190618247.001.0001. URL https://academic.oup.com/book/43912.

[7] Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4):193–202, April 1980. ISSN 0340-1200, 1432-0770. doi: 10.1007/BF00344251. URL http://link.springer.com/10.1007/BF00344251.

[8] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, 1(4):541–551, December 1989. ISSN 0899-7667, 1530-888X. doi: 10.1162/neco.1989.1.4.541. URL https://direct.mit.edu/neco/article/1/4/541-551/5515.

[9] Daniel L. K. Yamins, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, June 2014. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1403112111. URL https://pnas.org/doi/full/10.1073/pnas.1403112111.

[10] U. Guclu and M. A. J. Van Gerven. Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. *Journal of Neuroscience*, 35(27):10005–10014, July 2015. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.5023-14.2015. URL https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.5023-14.2015.

[11] Daniel L K Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3):356–365, March 2016. ISSN 1097-6256, 1546-1726. doi: 10.1038/nn.4244. URL https://www.nature.com/articles/nn.4244.

[12] Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLoS Computational Biology*, 10 (11):e1003915, November 2014. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1003915. URL https://dx.plos.org/10.1371/journal.pcbi.1003915.

[13] Santiago A. Cadena, George H. Denfield, Edgar Y. Walker, Leon A. Gatys, Andreas S. Tolias, Matthias Bethge, and Alexander S. Ecker. Deep convolutional models improve predictions of macaque V1 responses to natural images. *PLOS Computational Biology*, 15(4):e1006897, April 2019. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1006897. URL https://dx.plos.org/10.1371/journal.pcbi.1006897.

[14] Rishi Rajalingham, Elias B. Issa, Pouya Bashivan, Kohitij Kar, Kailyn Schmidt, and James J. DiCarlo. Large-Scale, High-Resolution Comparison of the Core Visual Object Recognition Behavior of Humans, Monkeys, and State-of-the-Art Deep Artificial Neural Networks. *The Journal of Neuroscience*, 38(33):7255–7269, August 2018. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.0388-18.2018. URL https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.0388-18.2018.

[15] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, May 2017. ISSN 0001-0782, 1557-7317. doi: 10.1145/3065386. URL https://dl.acm.org/doi/10.1145/3065386.

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, Miami, FL, June 2009. IEEE. ISBN 978-1-4244-3992-8. doi: 10.1109/CVPR.2009.5206848. URL https://ieeexplore.ieee.org/document/5206848/.

[17] Zhirong Wu, Yuanjun Xiong, Stella Yu, and Dahua Lin. Unsupervised Feature Learning via Non-Parametric Instance-level Discrimination, May 2018. URL http://arxiv.org/abs/1805.01978. arXiv:1805.01978 [cs].

[18] Chengxu Zhuang, Alex Zhai, and Daniel Yamins. Local Aggregation for Unsupervised Learning of Visual Embeddings. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6001–6011, Seoul, Korea (South), October 2019. IEEE. ISBN 978-1-7281-4803-8. doi: 10.1109/ICCV.2019.00610. URL https://ieeexplore.ieee.org/document/9011034/.

[19] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations, June 2020. URL http://arxiv.org/abs/2002.05709. arXiv:2002.05709 [cs, stat].

[20] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised Learning, September 2020. URL http://arxiv.org/abs/2006.07733. arXiv:2006.07733 [cs].

[21] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training, October 2022. URL http://arxiv.org/abs/2203.12602. arXiv:2203.12602 [cs].

[22] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski,

and Armand Joulin. Emerging Properties in Self-Supervised Vision Transformers, May 2021. URL http://arxiv.org/abs/2104.14294. arXiv:2104.14294 [cs].

[23] Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting Feature Prediction for Learning Visual Representations from Video, February 2024. URL http://arxiv.org/abs/2404.08471. arXiv:2404.08471 [cs].

[24] Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba, Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, et al. V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning, June 2025. URL http://arxiv.org/abs/2506.09985. arXiv:2506.09985 [cs].

[25] Chengxu Zhuang, Siming Yan, Aran Nayebi, Martin Schrimpf, Michael C. Frank, James J. DiCarlo, and Daniel L. K. Yamins. Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, 118(3):e2014196118, January 2021. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2014196118.

[26] Talia Konkle and George A. Alvarez. A self-supervised domain-general learning framework for human ventral stream representation. *Nature Communications*, 13(1):491, January 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-28091-4. URL https://www.nature.com/articles/s41467-022-28091-4.

[27] A. Emin Orhan, Vaibhav V. Gupta, and Brenden M. Lake. Self-supervised learning through the eyes of a child, December 2020. URL http://arxiv.org/abs/2007.16189. arXiv:2007.16189 [cs].

[28] William Lotter, Gabriel Kreiman, and David Cox. A neural network trained for prediction mimics diverse features of biological neurons and perception. *Nature Machine Intelligence*, 2(4):210–219, April 2020. ISSN 2522-5839. doi: 10.1038/s42256-020-0170-9. URL https://www.nature.com/articles/s42256-020-0170-9.

[29] Saber Sheybani, Himanshu Hansaria, Justin N Wood, Linda B Smith, and Zoran Tiganj. Curriculum Learning with Infant Egocentric Videos. 2023.

[30] A. Emin Orhan, Wentao Wang, Alex N. Wang, Mengye Ren, and Brenden M. Lake. Self-supervised learning of video representations from a child's perspective, July 2024. URL http://arxiv.org/abs/2402.00300. arXiv:2402.00300 [cs, q-bio].

[31] Bria Long, Violet Xiang, Stefan Stojanov, Robert Z. Sparks, Zi Yin, Grace E. Keene, Alvin W. M. Tan, Steven Y. Feng, Chengxu Zhuang, Virginia A. Marchman, et al. The BabyView dataset: High-resolution egocentric videos of infants' and young children's everyday experiences, June 2024. URL http://arxiv.org/abs/2406.10447. arXiv:2406.10447 [cs].

[32] Elizabeth M. Clerkin, Elizabeth Hart, James M. Rehg, Chen Yu, and Linda B. Smith. Real-world visual statistics and infants' first-learned object names. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1711):20160055, January 2017. ISSN 0962-8436, 1471-2970. doi: 10.1098/rstb.2016.0055. URL https://royalsocietypublishing.org/doi/10.1098/rstb.2016.0055.

[33] Elizabeth M. Clerkin and Linda B. Smith. Real-world statistics at two timescales and a mechanism for infant learning of object names. *Proceedings of the National Academy of*

Sciences, 119(18):e2123239119, May 2022. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2123239119. URL https://pnas.org/doi/full/10.1073/pnas.2123239119.

[34] Alvin Wei Ming Tan, Jane Yang, Tarun Sepuri, Khai Loong Aw, Robert Z. Sparks, Zi Yin, Virginia A. Marchman, Michael C. Frank, and Bria Long. Assessing the alignment between infants' visual and linguistic experience using multimodal language models, November 2025. URL http://arxiv.org/abs/2511.18824. arXiv:2511.18824 [cs].

[35] Tarun Sepuri, Khai Loong Aw, Alvin Wei Ming Tan, Robert Zane Sparks, Virginia A. Marchman, Michael C. Frank, and Bria Long. Characterizing young children's everyday activities using video question-answering models, October 2025. URL https://osf.io/gndy9_v1.

[36] Jane Yang, Tarun Sepuri, Alvin Wei Ming Tan, Michael C. Frank, and Bria Long. Quantifying infants' everyday experiences with objects in a large corpus of egocentric videos, June 2025. URL https://osf.io/jqmf3_v1.

[37] Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, et al. Findings of the BabyLM Challenge: Sample-Efficient Pretraining on Developmentally Plausible Corpora. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 1–6, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.conll-babylm.1. URL https://aclanthology.org/2023.conll-babylm.1.

[38] Alex Warstadt, Leshem Choshen, Aaron Mueller, Adina Williams, Ethan Wilcox, and Chengxu Zhuang. Call for Papers – The BabyLM Challenge: Sample-efficient pretraining on a developmentally plausible corpus, January 2023. URL http://arxiv.org/abs/2301.11796. arXiv:2301.11796 [cs].

[39] Michael C. Frank. Bridging the data gap between children and large language models. *Trends in Cognitive Sciences*, 27(11):990–992, November 2023. ISSN 13646613. doi: 10.1016/j.tics.2023.08.007. URL https://linkinghub.elsevier.com/retrieve/pii/S1364661323002036.

[40] Daniel M. Bear, Kevin Feigelis, Honglin Chen, Wanhee Lee, Rahul Venkatesh, Klemen Kotar, Alex Durango, and Daniel L. K. Yamins. Unifying (Machine) Vision via Counterfactual World Modeling, June 2023. URL http://arxiv.org/abs/2306.01828. arXiv:2306.01828 [cs].

[41] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners, December 2021. URL http://arxiv.org/abs/2111.06377. arXiv:2111.06377 [cs].

[42] Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, USA, 2nd edition, 2009. ISBN 052189560X.

[43] Tobias Gerstenberg. Counterfactual simulation in causal cognition, January 2024. URL https://osf.io/72scr.

[44] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, June 2021. URL http://arxiv.org/abs/2010.11929. arXiv:2010.11929 [cs].

[45] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The Kinetics Human

Action Video Dataset, May 2017. URL http://arxiv.org/abs/1705.06950. arXiv:1705.06950 [cs].

[46] Klemen Kotar, Wanhee Lee, Rahul Venkatesh, Honglin Chen, Daniel Bear, Jared Watrous, Simon Kim, Khai Loong Aw, Lilian Naing Chen, Stefan Stojanov, et al. World Modeling with Probabilistic Structure Integration, September 2025. URL http://arxiv.org/abs/2509.09737. arXiv:2509.09737 [cs].

[47] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. DINOv3, August 2025. URL http://arxiv.org/abs/2508.10104. arXiv:2508.10104 [cs].

[48] Carl Doersch, Ankush Gupta, Larisa Markeeva, Adrià Recasens, Lucas Smaira, Yusuf Aytar, João Carreira, Andrew Zisserman, and Yi Yang. TAP-Vid: A Benchmark for Tracking Any Point in a Video, March 2023. URL http://arxiv.org/abs/2211.03726. arXiv:2211.03726 [cs].

[49] Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J. Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, et al. Kubric: A scalable dataset generator, March 2022. URL http://arxiv.org/abs/2203.03570. arXiv:2203.03570 [cs].

[50] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. CoTracker: It is Better to Track Together, October 2024. URL http://arxiv.org/abs/2307.07635. arXiv:2307.07635 [cs].

[51] Henrique Morimitsu, Xiaobin Zhu, Roberto M. Cesar Jr, Xiangyang Ji, and Xu-Cheng Yin. DPFlow: Adaptive Optical Flow Estimation with a Dual-Pyramid Framework, March 2025. URL http://arxiv.org/abs/2503.14880. arXiv:2503.14880 [cs].

[52] Yihan Wang, Lahav Lipson, and Jia Deng. SEA-RAFT: Simple, Efficient, Accurate RAFT for Optical Flow, May 2024. URL http://arxiv.org/abs/2405.14793. arXiv:2405.14793 [cs].

[53] Yiming Zuo, Karhan Kayan, Maggie Wang, Kevin Jeon, Jia Deng, and Thomas L. Griffiths. Towards Foundation Models for 3D Vision: How Close Are We?, October 2024. URL http://arxiv.org/abs/2410.10799. arXiv:2410.10799 [cs].

[54] Gemini, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, December 2024. URL http://arxiv.org/abs/2403.05530. arXiv:2403.05530 [cs].

[55] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, et al. GPT-4 Technical Report, March 2024. URL http://arxiv.org/abs/2303.08774. arXiv:2303.08774 [cs].

[56] OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, A. J. Ostrow, Akila Welihinda, Alan Hayes, et al. GPT-4o System Card, October 2024. URL http://arxiv.org/abs/2410.21276. arXiv:2410.21276 [cs].

[57] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-shot Cross-dataset Transfer, August 2020. URL http://arxiv.org/abs/1907.01341. arXiv:1907.01341 [cs].

[58] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel Brostow. Digging Into Self-Supervised Monocular Depth Estimation, August 2019. URL http://arxiv.org/abs/1806.01260. arXiv:1806.01260 [cs].

[59] Bowen Wen, Matthew Trepte, Joseph Aribido, Jan Kautz, Orazio Gallo, and Stan Birchfield. FoundationStereo: Zero-Shot Stereo Matching, April 2025. URL http://arxiv.org/abs/2501.09898. arXiv:2501.09898 [cs].

[60] Rahul Venkatesh, Klemen Kotar, Lilian Naing Chen, Seungwoo Kim, Luca Thomas Wheeler, Jared Watrous, Ashley Xu, Gia Ancone, Wanhee Lee, Honglin Chen, et al. Discovering and using Spelke segments, July 2025. URL http://arxiv.org/abs/2507.16038. arXiv:2507.16038 [cs].

[61] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention Mask Transformer for Universal Image Segmentation, June 2022. URL http://arxiv.org/abs/2112.01527. arXiv:2112.01527 [cs].

[62] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common Objects in Context, February 2015. URL http://arxiv.org/abs/1405.0312. arXiv:1405.0312 [cs].

[63] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. SAM 2: Segment Anything in Images and Videos, October 2024. URL http://arxiv.org/abs/2408.00714. arXiv:2408.00714 [cs].

[64] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric, April 2018. URL http://arxiv.org/abs/1801.03924. arXiv:1801.03924 [cs].

[65] João Carreira, Michael King, Viorica Pătrăucean, Dilara Gokay, Cătălin Ionescu, Yi Yang, Daniel Zoran, Joseph Heyward, Carl Doersch, Yusuf Aytar, et al. Learning from One Continuous Video Stream, March 2024. URL http://arxiv.org/abs/2312.00598. arXiv:2312.00598 [cs.CV].

[66] Agrim Gupta, Jiajun Wu, Jia Deng, and Li Fei-Fei. Siamese Masked Autoencoders, May 2023. URL http://arxiv.org/abs/2305.14344. arXiv:2305.14344 [cs.CV].

[67] Ivo Iglowstein, Oskar G. Jenni, Luciano Molinari, and Remo H. Largo. Sleep Duration From Infancy to Adolescence: Reference Values and Generational Trends. *Pediatrics*, 111(2):302–307, February 2003. ISSN 0031-4005, 1098-4275. doi: $10.1542/\mathrm{peds}.111.2.302$. URL https://publications.aap.org/pediatrics/article/111/2/302/66745/Sleep-Duration-From-Infancy-to-Adolescence.

[68] Lana M. Trick, Fern Jaspers-Fayer, and Naina Sethi. Multiple-object tracking in children: The "Catch the Spies" task. *Cognitive Development*, 20(3):373–387, July 2005. ISSN 08852014. doi: $10.1016/\mathrm{j.cogdev}.2005.05.009$. URL https://linkinghub.elsevier.com/retrieve/pii/S0885201405000249.

[69] Tashauna L. Blankenship, Roger W. Strong, and Melissa M. Kibbe. Development of multiple object tracking via multifocal attention. *Developmental Psychology*, 56(9):1684–

1695, September 2020. ISSN 1939-0599, 0012-1649. doi: 10.1037/dev0001064. URL https://doi.apa.org/doi/10.1037/dev0001064.

[70] R Held, E Birch, and J Gwiazda. Stereoacuity of human infants. *Proceedings of the National Academy of Sciences*, 77(9):5572–5574, September 1980. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.77.9.5572. URL https://pnas.org/doi/full/10.1073/pnas.77.9.5572.

[71] Robert Fox, Richard N. Aslin, Sandra L. Shea, and Susan T. Dumais. Stereopsis in Human Infants. *Science*, 207(4428):323–324, January 1980. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.7350666. URL https://www.science.org/doi/10.1126/science.7350666.

[72] Eileen E. Birch, Jane Gwiazda, and Richard Held. Stereoacuity development for crossed and uncrossed disparities in human infants. *Vision Research*, 22(5):507–513, January 1982. ISSN 00426989. doi: 10.1016/0042-6989(82)90108-0. URL https://linkinghub.elsevier.com/retrieve/pii/0042698982901080.

[73] Anthony M. Norcia, Milena Kaestner, Yulan D. Chen, and Caroline S. Clement. Late Development of Sensory Thresholds for Horizontal Relative Disparity in Human Visual Cortex in the Face of Precocial Development of Thresholds for Absolute Disparity. *The Journal of Neuroscience*, 45(7):e0216242024, February 2025. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.0216-24.2024. URL https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.0216-24.2024.

[74] Scott P. Johnson. How Infants Learn About the Visual World. *Cognitive Science*, 34(7):1158–1184, September 2010. ISSN 0364-0213, 1551-6709. doi: 10.1111/j.1551-6709.2010.01127.x. URL https://onlinelibrary.wiley.com/doi/10.1111/j.1551-6709.2010.01127.x.

[75] Renée Baillargeon, Maayan Stavans, Di Wu, Yael Gertner, Peipei Setoh, Audrey K. Kittredge, and Amélie Bernard. Object Individuation and Physical Reasoning in Infancy: An Integrative Account. *Language Learning and Development*, 8(1):4–46, January 2012. ISSN 1547-5441, 1547-3341. doi: 10.1080/15475441.2012.630610. URL http://www.tandfonline.com/doi/abs/10.1080/15475441.2012.630610.

[76] Renée Baillargeon, Amy Needham, and Julie Devos. The development of young infants' intuitions about support. *Early Development and Parenting*, 1(2):69–78, January 1992. ISSN 1057-3593, 1099-0917. doi: 10.1002/edp.2430010203. URL https://onlinelibrary.wiley.com/doi/10.1002/edp.2430010203.

[77] Susan J Hespos and Renée Baillargeon. Reasoning about containment events in very young infants. *Cognition*, 78(3):207–245, March 2001. ISSN 00100277. doi: 10.1016/S0010-0277(00)00118-9. URL https://linkinghub.elsevier.com/retrieve/pii/S0010027700001189.

[78] Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, et al. Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like? preprint, Neuroscience, September 2018. URL http://biorxiv.org/lookup/doi/10.1101/407007.

[79] D. J. Felleman and D. C. Van Essen. Distributed Hierarchical Processing in the Primate Cerebral Cortex. *Cerebral Cortex*, 1(1):1–47, January 1991. ISSN 1047-3211, 1460-2199. doi: 10.1093/cercor/1.1.1. URL https://academic.oup.com/cercor/article-lookup/doi/10.1093/cercor/1.1.1.

[80] James J. DiCarlo, Davide Zoccolan, and Nicole C. Rust. How Does the Brain Solve

Visual Object Recognition? *Neuron*, 73(3):415–434, February 2012. ISSN 08966273. doi: 10.1016/j.neuron.2012.01.010. URL https://linkinghub.elsevier.com/retrieve/pii/S089662731200092X.

[81] Kalanit Grill-Spector and Kevin S. Weiner. The functional architecture of the ventral temporal cortex and its role in categorization. *Nature Reviews Neuroscience*, 15(8):536–548, August 2014. ISSN 1471-003X, 1471-0048. doi: 10.1038/nrn3747. URL https://www.nature.com/articles/nrn3747.

[82] Nitin Gogtay, Jay N. Giedd, Leslie Lusk, Kiralee M. Hayashi, Deanna Greenstein, A. Catherine Vaituzis, Tom F. Nugent, David H. Herman, Liv S. Clasen, Arthur W. Toga, et al. Dynamic mapping of human cortical development during childhood through early adulthood. *Proceedings of the National Academy of Sciences*, 101(21):8174–8179, May 2004. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0402680101. URL https://pnas.org/doi/full/10.1073/pnas.0402680101.

[83] Golijeh Golarai, Dara G Ghahremani, S Whitfield-Gabrieli, Allan Reiss, Jennifer L Eberhardt, John D E Gabrieli, and Kalanit Grill-Spector. Differential development of high-level visual cortex correlates with category-specific recognition memory. *Nature Neuroscience*, 10(4):512–522, April 2007. ISSN 1097-6256, 1546-1726. doi: 10.1038/nn1865. URL https://www.nature.com/articles/nn1865.

[84] Kalanit Grill-Spector, Golijeh Golarai, and John Gabrieli. Developmental neuroimaging of the human ventral visual cortex. *Trends in Cognitive Sciences*, 12(4):152–162, April 2008. ISSN 13646613. doi: 10.1016/j.tics.2008.01.009. URL https://linkinghub.elsevier.com/retrieve/pii/S1364661308000570.

[85] Oliver Braddick and Janette Atkinson. Development of human visual function. *Vision Research*, 51(13):1588–1609, July 2011. ISSN 00426989. doi: 10.1016/j.visres.2011.02.018. URL https://linkinghub.elsevier.com/retrieve/pii/S004269891100068X.

[86] Lynne Kiorpes. Visual development in primates: Neural mechanisms and critical periods. *Developmental Neurobiology*, 75(10):1080–1090, October 2015. ISSN 1932-8451, 1932-846X. doi: 10.1002/dneu.22276. URL https://onlinelibrary.wiley.com/doi/10.1002/dneu.22276.

[87] Emily J. Allen, Ghislain St-Yves, Yihan Wu, Jesse L. Breedlove, Jacob S. Prince, Logan T. Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, et al. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 25(1):116–126, January 2022. ISSN 1097-6256, 1546-1726. doi: 10.1038/s41593-021-00962-x. URL https://www.nature.com/articles/s41593-021-00962-x.

[88] Paolo Papale, Feng Wang, Matthew W. Self, and Pieter R. Roelfsema. An extensive dataset of spiking activity to reveal the syntax of the ventral stream. *Neuron*, 113(4):539–553.e5, February 2025. ISSN 08966273. doi: 10.1016/j.neuron.2024.12.003. URL https://linkinghub.elsevier.com/retrieve/pii/S089662732400881X.

[89] Melvyn A. Goodale and A.David Milner. Separate visual pathways for perception and action. *Trends in Neurosciences*, 15(1):20–25, January 1992. ISSN 01662236. doi: 10.1016/0166-2236(92)90344-8. URL https://linkinghub.elsevier.com/retrieve/pii/0166223692903448.

[90] David Michael Kaplan and Carl F. Craver. The Explanatory Force of Dynamical and Mathematical Models in Neuroscience: A Mechanistic Perspective. *Philosophy of Science*, 78(4):601–627, October 2011. ISSN 0031-8248, 1539-767X. doi: 10.1086/661755. URL https://www.cambridge.org/core/product/identifier/S0031824800016093/type/journal_article.

[91] Rosa Cao and Daniel Yamins. Explanatory models in neuroscience: Part 1 – taking mechanistic abstraction seriously, April 2021. URL http://arxiv.org/abs/2104.01490. arXiv:2104.01490 [cs, q-bio].

[92] Michael C. Frank. Cognitive modeling using artificial intelligence, March 2025. URL https://osf.io/wv7mg_v1.

[93] Richard Sutton. The bitter lesson. Essay, 2019.

[94] Noam Chomsky. Rules and representations. *Behavioral and Brain Sciences*, 3(1):1–15, March 1980. ISSN 0140-525X, 1469-1825. doi: 10.1017/S0140525X00001515. URL https://www.cambridge.org/core/product/identifier/S0140525X00001515/type/journal_article.

[95] Howard Lasnik and Jeffrey L. Lidz. The Argument from the Poverty of the Stimulus. In Ian Roberts, editor, *The Oxford Handbook of Universal Grammar*, pages 220–248. Oxford University Press, December 2016. ISBN 978-0-19-957377-6. doi: 10.1093/oxfordhb/9780199573776.013.10. URL https://academic.oup.com/edited-volume/27996/chapter/211719982.

[96] Jeffrey Elman, Elizabeth Bates, Mark H. Johnson, Annette Karmiloff-Smith, Domenico Parisi, and Kim Plunkett. *Rethinking Innateness: A Connectionist Perspective on Development*. The MIT Press, October 1996. ISBN 978-0-262-27229-2. doi: 10.7551/mitpress/5929.001.0001. URL https://direct.mit.edu/books/book/3912/Rethinking-InnatenessA-Connectionist-Perspective.

[97] Anthony M. Zador. A critique of pure learning and what artificial neural networks can learn from animal brains. *Nature Communications*, 10(1):3770, August 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-11786-6. URL https://www.nature.com/articles/s41467-019-11786-6.

[98] Sergey Shuvaev, Divyansha Lachi, Alexei Koulakov, and Anthony Zador. Encoding innate ability through a genomic bottleneck. *Proceedings of the National Academy of Sciences*, 121(38):e2409160121, September 2024. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2409160121. URL https://pnas.org/doi/10.1073/pnas.2409160121.

[99] Richard S. Sutton. Integrated Architectures for Learning, Planning, and Reacting Based on Approximating Dynamic Programming. In *Machine Learning Proceedings 1990*, pages 216–224. Elsevier, 1990. ISBN 978-1-55860-141-3. doi: 10.1016/B978-1-55860-141-3.50030-4. URL https://linkinghub.elsevier.com/retrieve/pii/B9781558601413500304.

[100] Marc Peter Deisenroth and Carl Edward Rasmussen. Pilco: A model-based and data-efficient approach to policy search. In *International Conference on Machine Learning*, 2011. URL https://api.semanticscholar.org/CorpusID:14273320.

[101] Carlos E. García, David M. Prett, and Manfred Morari. Model predictive control: Theory and practice—A survey. *Automatica*, 25(3):335–348, 1989. ISSN 0005-1098. doi: https://doi.org/10.1016/0005-1098(89)90002-2. URL https://www.sciencedirect.com/science/article/pii/0005109889900022.

[102] David Ha and Jürgen Schmidhuber. World Models. March 2018. doi: 10.5281/zenodo.1207631. URL http://arxiv.org/abs/1803.10122. arXiv:1803.10122 [cs].

[103] Danijar Hafner, Timothy P. Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *ArXiv*, abs/1912.01603, 2019. URL https://api.semanticscholar.org/CorpusID:208547755.

[104] Danijar Hafner, Timothy P. Lillicrap, Ian S. Fischer, Ruben Villegas, David R Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. *ArXiv*, abs/1811.04551, 2018. URL https://api.semanticscholar.org/CorpusID:53280207.

[105] Danijar Hafner, Kuang-Huei Lee, Ian S. Fischer, and P. Abbeel. Deep hierarchical planning from pixels. *ArXiv*, abs/2206.04114, 2022. URL https://api.semanticscholar.org/CorpusID:249538516.

[106] Lukasz Kaiser, Mohammad Babaeizadeh, Piotr Milos, Blazej Osinski, Roy H. Campbell, K. Czechowski, D. Erhan, Chelsea Finn, Piotr Kozakowski, Sergey Levine, et al. Model-based reinforcement learning for atari. *ArXiv*, abs/1903.00374, 2019. URL https://api.semanticscholar.org/CorpusID:67856232.

[107] Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, L. Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, et al. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588:604 – 609, 2019. URL https://api.semanticscholar.org/CorpusID:208158225.

[108] Philipp Wu, Alejandro Escontrela, Danijar Hafner, Ken Goldberg, and P. Abbeel. Daydreamer: World models for physical robot learning. In *Conference on Robot Learning*, 2022. URL https://api.semanticscholar.org/CorpusID:250088882.

[109] Klemen Kotar, Aaron Walsman, and Roozbeh Mottaghi. Entl: Embodied navigation trajectory learner. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10863–10872, October 2023.

[110] Wanhee Lee, Klemen Kotar, Rahul Mysore Venkatesh, Jared Watrous, Honglin Chen, Khai Loong Aw, and Daniel L. K. Yamins. 3D Scene Understanding Through Local Random Access Sequence Modeling, April 2025. URL http://arxiv.org/abs/2504.03875. arXiv:2504.03875 [cs].

[111] Seungwoo Kim, Khai Loong Aw, Klemen Kotar, Cristobal Eyzaguirre, Wanhee Lee, Yunong Liu, Jared Watrous, Stefan Stojanov, Juan Carlos Niebles, Jiajun Wu, et al. Taming generative video models for zero-shot optical flow extraction, November 2025. URL http://arxiv.org/abs/2507.09082. arXiv:2507.09082 [cs.CV].

[112] Zhenyu Jiang, Hanwen Jiang, and Yuke Zhu. Doduo: Learning Dense Visual Correspondence from Unsupervised Semantic-Aware Flow, September 2023. URL http://arxiv.org/abs/2309.15110. arXiv:2309.15110 [cs.CV].

[113] Rhodri Guerrier, Adam W. Harley, and Dima Damen. PointSt3R: Point Tracking through 3D Grounded Correspondence, October 2025. URL http://arxiv.org/abs/2510.26443. arXiv:2510.26443 [cs.CV].

[114] James J. Gibson. Optical motions and transformations as stimuli for visual perception. *Psychological Review*, 64(5):288–295, 1957. ISSN 1939-1471, 0033-295X. doi: 10.1037/h0044277. URL https://doi.apa.org/doi/10.1037/h0044277.

[115] Lu Qi, Jason Kuen, Weidong Guo, Tiancheng Shen, Jiuxiang Gu, Jiaya Jia, Zhe Lin, and Ming-Hsuan Yang. High-Quality Entity Segmentation, April 2023. URL http://arxiv.org/abs/2211.05776. arXiv:2211.05776 [cs].

[116] Open X.-Embodiment Collaboration, Abby O'Neill, Abdul Rehman, Abhinav Gupta, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, et al. Open X-Embodiment: Robotic Learning Datasets and RT-X Models, May 2025. URL http://arxiv.org/abs/2310.08864. arXiv:2310.08864 [cs].

[117] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition, December 2015. URL http://arxiv.org/abs/1512.03385. arXiv:1512.03385 [cs].

[118] Sungmin Woo, Wonjoon Lee, Woo Jin Kim, Dogyoon Lee, and Sangyoun Lee. ProDepth: Boosting Self-Supervised Multi-Frame Monocular Depth with Probabilistic Fusion, July 2024. URL http://arxiv.org/abs/2407.09303. arXiv:2407.09303 [cs].

Supplementary Materials for Zero-shot World Models Are Developmentally Efficient Learners

Khai Loong Aw*, Klemen Kotar, Wanhee Lee, Seungwoo Kim, Khaled Jedoui, Rahul Venkatesh, Lilian Naing Chen, Michael C. Frank, Daniel L.K. Yamins
*Corresponding author. Email: khaiaw@stanford.edu

Methods

Model architecture

The ZWM predictor Ψ is implemented as a Vision Transformer (ViT) [44]. Input frames are resized to 256×256 pixels and divided into non-overlapping 8×8-pixel patches, yielding $32 \times 32 = 1024$ patch tokens per frame. We evaluate two model sizes (Table S1):

- **ZWM-170M**: 24 transformer layers, 12 attention heads, embedding dimension 768, totaling $\sim$170 million parameters.
- **ZWM-1B**: 48 transformer layers, 16 attention heads, embedding dimension 1280, totaling $\sim$1 billion parameters.

Two-frame input tokenization. Given a frame pair (f_1, f_2), the first frame f_1 is fully patchified into 1024 tokens, each a flattened $8 \times 8 \times 3 = 192$-dimensional vector. The second frame f_2 is masked: only 10% of its patches (approximately 102 tokens) are revealed, with the remaining patches replaced by a shared learnable mask token. Both sets of tokens receive positional embeddings (learned, not sinusoidal) before being concatenated and fed into the transformer.

Masking strategy. During training, the mask for f_2 is sampled uniformly at random on each example, with exactly 10% of patches revealed. This ensures the model encounters diverse masking patterns and cannot rely on fixed spatial positions. The asymmetric structure—fully visible f_1, 90% masked f_2—is a core design choice that encourages temporal factorization of appearance and motion.

Symmetric masking ablation. To test whether this asymmetry is necessary, we trained BabyZWM variants with symmetric masking policies:

- **Symmetric 45%-45%**: Both frames are masked at 45%, so each frame reveals 55% of patches.
- **Symmetric 90%-90%**: Both frames are masked at 90%, so each frame reveals only 10% of patches.

Both symmetric variants perform substantially worse across all zero-shot visual-cognitive tasks (Figures 2 and 4), demonstrating that the temporally-biased mask structure—rather than masking per se—is critical for learning representations that support flexible zero-shot extraction.

Output and loss. The model outputs a prediction $\widehat{f_2}$ for the full second frame, including all masked positions. The training objective is the mean squared error (MSE) between the predicted and ground-truth pixel values of f_2, computed over masked patches:

$$\mathcal{L} = \left\langle \| f_2 - \widehat{f_2} \|^2 \right\rangle_{(f_1, f_2) \in \mathcal{D}}. \tag{S1}$$

Table S1: Model architecture configurations. Architectural hyperparameters for the two ZWM model sizes evaluated in this work.

	ZWM-170M	ZWM-1B
Transformer layers	24	48
Attention heads	12	16
Embedding dimension	768	1280
Patch size	8×8	8×8
Input resolution	256×256	256×256
Tokens per frame	1024	1024
Total parameters	$\sim$170M	$\sim$1B

Training procedure

Each ZWM model is trained for 200,000 steps with a batch size of 512 (Table S2). As the videos are stored at 30 frames per second, this corresponds to $\sim$950 video hours, or roughly 95 days of waking experience assuming $\sim$10 awake hours per day for young children [67].

Training datapoints consist of RGB frame pairs sampled from real-world video, with the inter-frame temporal gap randomly and uniformly chosen in the range 150–450ms (corresponding to 5–14 frames at 30 fps).

Optimization. We use AdamW with a peak learning rate of 3e-4, weight decay of 1e-1, and $(\beta_1, \beta_2) = (0.9, 0.95)$. The learning rate follows a cosine decay schedule with 2000 warmup steps. We use gradient clipping with a maximum norm of 1.0.

Data augmentation. No data augmentation (e.g., random crops, color jitter, horizontal flips) is applied during training. The model is trained directly on raw RGB frame pairs.

Compute. All models are trained using PyTorch with Distributed Data Parallel (DDP) and mixed-precision (bfloat16). The ZWM-170M model is trained on 4 nodes of 8 NVIDIA H100 GPUs each (32 GPUs total) for approximately 11 hours ($\sim$352 GPU-hours). The ZWM-1B model is trained on 8 nodes of 8 H100 GPUs each (64 GPUs total) for approximately 24 hours ($\sim$1,536 GPU-hours).

Training datasets

We train ZWM on a spectrum of visual diets to test data efficiency and robustness (Table S3):

BabyView. The BabyView dataset [31] consists of 868 hours, mostly longitudinal, egocentric video recordings from $N = 34$ children aged $\sim$5 months to 3 years, and $\sim$100 hours from 3–5-year-olds recorded in a preschool setting. Videos are recorded using head-mounted cameras worn by the children during natural daily activities. We refer to the ZWM model trained on the full BabyView dataset as "BabyZWM." The raw BabyView videos, recorded by families in their homes, are typically several minutes in duration. We preprocess all videos by splitting them into 10-second clips, stored at 30 fps with the shorter spatial dimension resized to 256 pixels (the longer side scaled proportionally to preserve aspect ratio). During training, we randomly sample a 256×256 crop from

Table S2: Training hyperparameters. Training configuration shared across all ZWM models unless otherwise noted.

Hyperparameter	Value
Optimizer	AdamW
AdamW (β_1, β_2)	(0.9, 0.95)
Peak learning rate	3e-4
Weight decay	1e-1
LR schedule	Cosine decay
Warmup steps	2000
Batch size	512
Total training steps	200,000
Inter-frame gap	150–450 ms
Patches revealed during training (f_1)	100%
Patches revealed during training (f_2)	10%

each frame pair, with the same crop applied to both f_1 and f_2. No additional augmentations are applied.

Single-Child BabyView. To test learning from even more restricted experience, we construct a subset of BabyView consisting of 132 hours of recordings from a single individual (child S00320001, aged 9-30 months). This represents the most stringent test of data efficiency, requiring the model to learn generalizable capacities from the highly restricted visual diversity of one child's experience.

Random 132-hour subset. To disentangle the contributions of visual diversity from total exposure, we also train on a random 132-hour subset of BabyView, sampled uniformly across all 34 children. This subset matches the Single-Child dataset in total duration but contains substantially more environmental diversity.

Age-ordered curricula. To test continual learning and robustness to catastrophic forgetting, we train Single-Child BabyZWM in an online, single-epoch fashion on the age-ordered video stream. We create curricula by shuffling within temporal windows of varying durations:

- **5-minute shuffle**: 10-second clips are randomly shuffled within contiguous 5-minute windows, preserving the coarse temporal order while introducing local mixing (analogous to within-episode consolidation).
- **30-minute shuffle**: 10-second clips are shuffled within 30-minute windows.
- **1-day shuffle**: 10-second clips are shuffled within full recording days, loosely approximating overnight sleep-like reordering.

Kinetics-400. Kinetics-400 [45] consists of $\sim$670 hours of 10-second video clips from YouTube, spanning 400 human action categories. This dataset is smaller than BabyView but contains substantially more environmental and semantic diversity due to its Internet-sourced content.

Table S3: Training dataset statistics. Summary of the video datasets used to train ZWM variants.

Dataset	Hours	Source	ZWM variant
BabyView	868	34 children, egocentric	BabyZWM
Single-Child BabyView	132	1 child, egocentric	Single-Child BabyZWM
Random 132h subset	132	34 children, egocentric	—
Kinetics-400	∼670	Internet (YouTube)	ZWM-Kinetics
Big Video Dataset (BVD)	∼7,000	CV datasets + Internet	ZWM-BVD

Table S4: Summary of zero-shot prompts for visual-cognitive tasks. Each prompt extracts a visual quantity by perturbing the predictor's input, comparing the perturbed output to the unperturbed prediction, and aggregating the difference. Later prompts compose earlier ones.

Task	**Perturb**	**Compare**	**Aggregate**	**Composes**
Optical flow	Add Gaussian tracer to query point in f_1	RGB difference between perturbed and clean predictions of f_2	Argmax of difference gives flow vector	— (primitive)
Hyp. motion	Displace object patch to new location in f_2^{masked}	Predict remaining masked regions	Full hypothetical scene	— (primitive)
Relative depth	Apply optical flow tracer to query point in binocular f_L	Optical flow from f_L to f_R via perturbed vs. clean predictions	Rank points by disparity magnitude	Optical flow
Object segment.	Displace object patch via hyp. motion	Optical flow between original and hypothetical scene	Threshold flow; aggregate over directions	Hyp. motion, optical flow
Intuitive physics	Reveal hand's ground-truth location in f_2^{masked}; fix background	MSE and LPIPS between prediction and target f_2	Closer to target or context; compose flow + segments for "what moved"	Hyp. motion, optical flow, object segment.

Big Video Dataset (BVD). BVD [46] consists of ∼7,000 hours of video drawn from a combination of computer vision datasets and Internet videos. This serves as an approximate upper bound on performance achievable with high visual diversity and scale.

Zero-shot prompt design

Here, we describe how diverse visual-cognitive quantities are extracted using ZWM's zero-shot prompts, which act as approximate *causal inferences* by comparing hypothetical or counterfactual predictions against the ground-truth. Each prompt follows a common structure: (i) construct a minimal **perturbation** that intervenes on a latent *cause* governing some visual quantity; (ii) **compare** the predictor's output under the perturbation against the unperturbed ground-truth prediction; and (iii) **aggregate** the difference to extract the quantity of interest. Simple prompts compose to extract increasingly complex visual structures, building a computational graph of visual intermediates. Table S4 summarizes the structure of each prompt.

Optical flow (Figure 2A). *Latent cause*: Between two frames (f_1, f_2), a point at position x_q in f_1 *causes* a corresponding point in f_2, due to the underlying causal structure of motion.

1. **Perturb**: Duplicate the initial frame f_1 and add a white-dot tracer to form $\tilde{f}_1$, using a Gaussian centered at the query location x_q with amplitude 255 on each RGB channel and standard deviation $\sigma = 3.0$ pixels.
2. **Compare**: Run the model twice with the same masked second frame f_2^{masked}: once with the clean frame $(f_1, f_2^{\text{masked}}) \to \hat{f}_2$ and once with the perturbed frame $(\tilde{f}_1, f_2^{\text{masked}}) \to \tilde{f}_2^{\text{pred}}$. Compute the RGB difference $\Delta = \tilde{f}_2^{\text{pred}} - \hat{f}_2$.
3. **Aggregate**: Take the argmax of $|\Delta|$ to identify where the perturbation was carried to. The flow vector at x_q is argmax$(|\Delta|) - x_q$.

Optical flow is the most primitive prompt and does not compose from other prompts; all subsequent prompts build on it. The masked patches in f_2^{masked} are randomly selected and differ across evaluations, but are held fixed between the perturbed and unperturbed forward passes to ensure the flow signal can be localized.

Relative depth (Figure 2D). *Latent cause*: The depth of a point is the latent cause governing its displacement under binocular separation; farther points exhibit smaller binocular disparity.

1. **Perturb**: Given a binocular image pair (f_L, f_R) from stereo cameras, apply the optical flow tracer (as above) to a query point x_q in f_L. Note that binocular image pairs are provided by the evaluation dataset; this is ecologically plausible, as humans possess binocular vision.
2. **Compare**: Compute the optical flow from f_L to f_R at x_q by comparing the perturbed and unperturbed predictions, composing the optical flow prompt described above.
3. **Aggregate**: The magnitude of the resulting flow vector gives the binocular disparity at x_q, which is inversely related to depth. To compare relative depth of multiple points, compose multiple optical flow prompts and rank by disparity magnitude.

Ecological validity and the rigidity question. The relative-depth readout uses a binocular (stereo) image pair, which humans with two eyes have access to, so a binocular input is ecologically natural. Stereopsis is widespread among animals with overlapping visual fields, and human infants are sensitive to binocular disparity within the first few months [70]. This readout also does not assume that scenes are rigid. The model uses binocular disparity at each timepoint independently, rather than motion parallax over time, so there is no scene-rigidity-over-time assumption. The two stereo views are captured at the same instant—simulating two eyes—so the disparity between them arises purely from the binocular baseline (parallax) and encodes depth by correspondence regardless of whether objects in the scene are moving. Both static and moving objects have a well-defined stereo disparity.

Hypothetical motion (Figure 2G). Before describing the remaining prompts, we introduce a key primitive: hypothetical motion. ZWM simulates "what if this object moved?" by selecting one or more patches on an object and displacing them to a new location in f_2^{masked}, then predicting the remaining masked regions. The predictor propagates this local displacement to the rest of the object, producing a full hypothetical scene. While displacing a single patch often suffices, displacing multiple patches from the same object generally produces more coherent hypothetical scenes. This primitive is not evaluated directly but serves as a building block for object segmentation and intuitive physics below.

Object segmentation (Figure 2G). *Latent cause*: Groups of pixels move together due to the latent cause of belonging to the same physical object—a learned form of "common fate" [1].

1. **Perturb**: Select a patch on a candidate object and displace it using the hypothetical motion primitive (above), producing a hypothetical scene $\tilde{f}$ in which the object has moved.
2. **Compare**: Compose the optical flow prompt to compute flow between the original image f and the hypothetical prediction $\tilde{f}$. Pixels belonging to the perturbed object will exhibit coherent flow; other pixels will not.
3. **Aggregate**: Threshold the flow magnitude to produce a binary mask. Repeat over 8 displacement directions, with displacement magnitudes between 25 and 35 pixels, then aggregate the resulting masks to obtain the full object segment.

Signalling a fixed camera. During training, the model learns to distinguish camera motion from object motion: when the camera moves, the background and the objects in the scene move together in a coherent manner, whereas when an individual object moves, it moves differently from the background and from other objects. At inference, the benchmark indicates to the model that the camera is not moving by holding the image corners at their original, unmoved appearance. The procedure therefore reveals only (i) the displaced object patch and (ii) these corner blocks. The corners are fixed locations rather than chosen per object; a corner is removed only if it lies close to the displaced source or destination block. This anchoring matters because, without it, an imposed perturbation could itself be read as camera motion, with the whole scene predicted to shift together and the object-specific signal lost. Empirically the anchoring works: in the model's predictions the surrounding background does not change, confirming that the model interprets the grounding patches as signalling a fixed camera, so the resulting segment reflects the imposed object motion rather than camera motion.

Signalling a fixed camera also operationalizes how a human observer might perform this segmentation: faced with a scene, the observer runs the same hypothetical simulation—"what if I nudged this part?"—from a fixed vantage point. Holding the camera fixed is a natural condition for that mental simulation, and it applies equally to static scenes and to scenes containing moving objects. Predicting a hypothetical conditioned on a static camera relies on a signal that biological organisms already possess, since organisms can tell whether their head is moving using efference copy and vestibular input.

Intuitive physics (Figure 3A). *Latent cause*: Physical interactions transmit forces between objects—e.g., pushing one object into another *causes* the second to move, exposing the underlying causal structure of contact dynamics.

1. **Perturb**: Reveal a 32×32-pixel green intervention patch in f_2^{masked} at the hand's ground-truth location in the target frame, providing information about where the hand has moved. The hand location is annotated by human labelers, with care taken to ensure the patch does not reveal the object's position. This acts as a perturbation relative to the unperturbed case (where the hand's motion is masked and unknown), prompting the model to predict the physical consequences of the hand's action on the rest of the scene. Additional 32×32-pixel red background patches are revealed to fix illumination and camera pose, isolating the causal effect of the hand's motion on objects. These border patches are revealed at their true, unchanged appearance to signal a static camera. The model operates directly on the revealed pixels, so it has access to their fine detail; empirically its predictions keep the table and surrounding background fixed, confirming that it reads the grounding patches as a stationary camera.
2. **Compare**: Compare the model's prediction (given the revealed hand motion) against the

ground-truth target frame f_2, using both MSE and LPIPS perceptual similarity [64].

3. **Aggregate**: Determine whether the prediction is closer to the ground-truth target f_2 or the context frame f_1. Additionally, compose optical flow and object segmentation prompts on the predicted scene to evaluate *what* moved and *how*—e.g., whether force transferred to a second object.

Evaluation benchmarks

Optical flow: TAP-Vid benchmarks. We evaluate optical flow on two benchmarks from the TAP-Vid suite:

- **TAP-Vid-DAVIS** [48]: Real-world videos with human-annotated ground-truth point correspondences, featuring challenging scenarios including fast motion, occlusions, and appearance changes.
- **TAP-Vid-Kubric** [49]: Synthetic, simulator-generated videos where ground-truth flows are known by construction, providing a complementary evaluation without annotation noise.

Query protocol and temporal-gap distribution. We use TAP-Vid's standard "first" query mode, in which every visible point is matched from its first visible occurrence to every subsequent target frame. To enable a controlled comparison across model classes, every model receives exactly two frames—the query frame and the target frame—and nothing else; this is the native input format for ZWM and for the two-frame optical-flow models. The temporal gap between those two frames is therefore not a parameter we choose, but the distribution induced by the benchmark's own query mode. Across the 38,881 query–target pairs of TAP-Vid-DAVIS (650 query points over 30 videos), the gap ranges from 0 to 103 frames, with a median of 31 and a mean of 34. The distribution spans the full range rather than concentrating at a single stride (Table S5), so the benchmark tests correspondence across both short gaps and large displacements.

Table S5: Temporal-gap distribution of the TAP-Vid-DAVIS "first" query protocol. Share of the 38,881 query–target pairs falling in each gap range. The distribution is induced by the benchmark's query mode rather than chosen: every visible point is matched from its first visible occurrence to every subsequent frame.

Temporal gap (frames)	Share of pairs
0–10	18.2%
11–30	31.3%
31–60	34.4%
$>$60	16.1%
Median 31; mean 34; range 0–103	38,881 pairs

Optical flow versus tracking. These two literatures use different input regimes. Optical-flow methods estimate motion from a pair of frames, whereas point-tracking methods consume a whole video and chain predictions across it. Although we evaluate in a two-frame optical-flow setting, we use the TAP-Vid tracking benchmark because it contains real-world, challenging dynamics, whereas optical-flow benchmarks are mostly synthetically generated (e.g., Sintel, FlyingThings). Passing every model the same two frames therefore lets us compare across both model classes on the

same input, and this pairwise implementation of the TAP-Vid "first" protocol has been used in prior work [48, 112, 113].

Why two frames. Two-frame optical flow is a classic and important scientific concept [114], and historically it played a key role in thinking about visual development. Two-frame evaluation is the fundamental unit of motion understanding — tracking itself is built from estimating how a point moves or becomes occluded across two frames — and it isolates the underlying motion representation by removing algorithm-specific heuristics for chaining frame-pair predictions into tracks. It is the standard regime of the optical-flow community, whereas multi-frame evaluation is standard in the tracking community; the two simply reflect different goals.

Tracking algorithms and temporal-smoothness priors. Tracking algorithms depend on temporal-smoothness priors, so the tracking evaluation — although useful for computer-vision applications — is less indicative of a model's capacity to understand the core motion-dynamics computation we are concerned with. These models build in many assumptions about smoothness, which is orthogonal to the vision-science question. Precisely because the multi-frame numbers are naturally higher, they allow an important dependence on smoothness priors that can blunt the assessment of how good a model's understanding of basic dynamics is. The two-frame CoTracker3 results are nonetheless reasonable: its performance is degraded reasonably rather than fully, still matching the supervised flow models DPFlow and SeaRAFT, which makes sense because all three take in two frames; and multi-frame CoTracker3, trained on multiple labelled datasets, achieves similar performance to ZWM-1B, which reaches 74.0% while seeing only two frames. We read this not as a deficiency of the two-frame protocol but as evidence that much of a tracker's advantage comes from aggregating intermediate frames rather than from stronger core two-frame dynamics understanding.

Downstream relevance. We are primarily interested in optical flow as a scientific concept and an instantiation of a concrete visual ability, rather than as a computer-vision application. Nevertheless, two-frame correspondence is the basic primitive behind frame interpolation, video compression, and motion-based segmentation — settings where motion must be estimated from a pair of frames rather than from a long continuous clip. We view two-frame flow and multi-frame tracking as complementary rather than opposed: algorithms that are good at two-frame flow can be made into tracking algorithms, and historically advances in optical flow have driven better multi-frame trackers [112].

All evaluations are conducted at 256×256 resolution. For each algorithm, we report two standard TAP-Vid metrics [48]:

- **Position accuracy ($< \delta^x_{\text{avg}}$)**: For visible points, the fraction of predicted correspondences falling within a pixel-distance threshold of the ground-truth position, averaged over five thresholds (1, 2, 4, 8, and 16 pixels).
- **Occlusion accuracy (OA)**: Binary classification accuracy for predicting whether each query point is occluded or out of frame on each time step.

Relative depth: UniQA-3D. We evaluate relative depth estimation on UniQA-3D [53], which presents pairs of points and requires judging which is farther from the camera. The upright data originally contains 500 samples, but after filtering to ensure the query points fall unambiguously within the center crop of the image and that each image contains a stereo pair from the original KITTI dataset, the upright set is filtered to 103 examples and the flipped set to 61 examples.

Object segmentation: SpelkeBench. We evaluate class-agnostic object segmentation on SpelkeBench [60], which defines objects as distinct, bounded physical entities. The benchmark draws images from two sources: 496 images from EntitySeg (real-world scenes) [115] and 51 images from Open X-Embodiment (real-world robot interactions) [116], totaling 547 images. We measure performance using intersection-over-union (IoU) [62].

Intuitive physics benchmark. We develop a novel short-timescale physical reasoning benchmark to evaluate models on intuitive physics (Figure 3A). The benchmark features tabletop interactions between a hand and 1–2 objects, testing five categories of physical reasoning:

1. **Object cohesion**: When one part of an object is moved, the entire object moves together.
2. **Support (top object moves)**: When a supporting surface is removed, the supported object falls.
3. **Support (bottom object moves)**: When the bottom object in a stack is moved, the top object moves with it.
4. **Force transfer**: Pushing one object into another causes the second object to move.
5. **Force separation**: Moving one object does not affect a spatially separated object.

The benchmark contains 20 image pairs per category (100 image pairs total). Each image pair is evaluated under 8 different random mask configurations for the revealed patches in f_2, yielding $5 \times 20 \times 8 = 800$ total evaluations per model.

We first considered classic object-permanence tasks (e.g., violation-of-expectation under occlusion). But these primarily probe working memory, rather than reasoning about physical dynamics. This leads to unnecessary confounds; if a model fails, it is not obvious whether it is due to limited working memory. Existing object-permanence benchmarks were set up this way because they were inspired by tests on human infants, whereas our intuitive-physics benchmark is designed for models and deliberately short-timescale (two frames), isolating physical reasoning — cohesion, support, force transfer and force separation — from the memory demands that object permanence introduces. Tackling object permanence calls for longer-horizon prediction and temporal memory, which we flag in the Discussion as important future work.

Each example consists of a context frame and a target frame. Accuracy is defined as the proportion of examples for which the model's prediction is closer to the ground-truth target than to the context frame, evaluated using both MSE and LPIPS [64].

Accuracy, stated precisely. For each example, the model's prediction is counted as correct if it is closer—that is, at a smaller distance—to the ground-truth target frame than to the context frame. This is computed under two distance measures, mean squared error (MSE) and LPIPS perceptual similarity [64], run separately, yielding two accuracy numbers per category. MSE and LPIPS capture different things—raw pixel error versus learned perceptual similarity—yet accuracy is very similar under both, which adds robustness, since the result does not hinge on a particular choice of distance measure.

A graded alternative to the binary metric. Because this binary accuracy saturates near ceiling and gives only a coarse pass/fail picture, we additionally report a finer, graded measure of how faithfully each prediction matches the real outcome. It is a continuous closeness score—the fraction of the

context-to-target gap that the prediction closes,

$$\text{graded} = \frac{d(\text{pred}, \text{context})}{d(\text{pred}, \text{context}) + d(\text{pred}, \text{target})} \in [0, 1], \tag{S2}$$

where a value of 0 means the prediction is identical to the context (no change predicted) and a value of 1 means it is identical to the ground-truth outcome. We compute d under both the MSE and the optical-flow distances, and within each object region (primary, secondary, and overall), turning the binary accuracy into a continuous fidelity scale. These graded scores rank the models exactly as the binary accuracy does, confirming the original result, but resolve differences that the binary number hides (Tables S21 and S22).

Scoring a latent-prediction model. V-JEPA2 does not generate pixels, so we do not force it into a pixel metric. It is evaluated in its native feature space — we compare whether its predicted features are closer to the target frame's features than to the context frame's — which is the appropriate way to score a latent-prediction model, and prior work has used V-JEPA2 to predict and compare in embedding space, including the original V-JEPA2 paper [24]. Given that some models predict pixels whereas others predict latent representations, the fairest approach is to evaluate each model in its native prediction space. The prediction spaces are nonetheless different and not directly comparable, so the feature-based and pixel-based columns of Table S22 should not be read against one another.

Developmental trajectories. To analyze developmental curves, we evaluate BabyZWM at various training checkpoints (0, 5k, 10k, 20k, 40k, 80k, 120k, 160k, 200k). Each ZWM model is trained for 200,000 steps with a batch size of 512. As the videos are stored at 30 frames per second, this corresponds to $\sim$950 video hours, or roughly 95 days of waking experience assuming $\sim$10 awake hours per day for young children [67]. The x-axis represents training steps.

Neural predictivity. We evaluate the alignment between model representations and biological neural responses using two complementary benchmarks:

- **Natural Scenes Dataset (NSD)** [87]: Human fMRI responses to natural images, capturing large-scale representational geometry.
- **THINGS Ventral Stream Spiking Dataset (TVSD)** [88]: Macaque single-neuron electrophysiology, revealing fine-grained neural tuning and timing.

For each model and brain region, we:

1. Extract features from every other layer of the model for each stimulus image.
2. Fit a cross-validated ridge regression from model features to neural responses.
3. Report noise-corrected Pearson correlations as the measure of neural predictivity.

For both NSD and TVSD, we use 10-fold cross-validation to split the data into training and test sets. Ridge regression regularization is performed using `RidgeCV`, which evaluates 21 regularization strengths (α). Importantly, α is selected independently for each target (i.e., per voxel for fMRI, per neuron for electrophysiology), allowing the regularization to adapt to the noise characteristics of each recording site.

Noise ceilings are estimated differently for each dataset. For NSD, we use the reliability estimation method described by Allen et al. [87]. For TVSD, noise ceilings are computed via split-half

correlations. For NSD, neural predictivity is evaluated for V1, V2, V3, V4, and the mid- and high-ventral visual regions. For TVSD, we evaluate using V1, V4, and inferior temporal (IT) cortex.

Baselines

We compare ZWM against both representation-based models and task-specific systems.

Representation-based models. Unlike ZWM, representation-based models are not natively zero-shot and typically require labeled supervision (fine-tuning or linear probes) for each downstream task. To enable fair comparison, we design simple zero-shot probes for these models:

- **ResNet50 (ImageNet-supervised)**: A standard ResNet50 [117] pretrained on ImageNet-1K with category-label supervision.
- **Baby DINOv3**: DINOv3 [47] (ViT-Large) learns single-image representations by training the model to produce consistent features across different augmented views of the same image. We train DINOv3 on BabyView (868 hours).
- **Baby V-JEPA2**: V-JEPA2 is a self-supervised video model that learns by predicting masked regions of a video in feature space rather than in raw pixels. We train a 300-million parameter V-JEPA2 model on BabyView (868 hours) using the official implementation and default hyperparameters. We verified successful training via frozen linear probes on held-out subsets, yielding 54.2% top-1 accuracy on Kinetics-400 (400-way classification) and 53.45% on ImageNet-1K (1000-way classification) [16].

In addition to these BabyView-trained variants, we evaluate the publicly released, standard-data-trained DINOv3 (ViT-Large) and V-JEPA2 (Large and Giant) models as non-developmental reference points; these are the "standard" / non-"Baby" rows in the supplementary comparison tables.

Zero-shot probe designs for representation-based baselines. Since ResNet50, DINOv3 and V-JEPA2 are representation-based models that do not natively support zero-shot visual-cognitive extraction, we design simple probe procedures to enable fair comparison.

Optical flow. For ResNet50, DINOv3 and V-JEPA2, we pass both frames through the model and extract patch-level feature representations. To estimate the flow at a query point in the first frame, we compute the cosine similarity between the query point's patch feature in the first frame and all patch features in the second frame. The target location is taken as the patch in the second frame with the highest cosine similarity, and the flow vector is the displacement between the query and target positions. This cosine-similarity nearest-neighbor matching follows the protocol of the DINOv3 paper [47] and is a standard label-free readout in prior self-supervised work [22, 66]; we additionally apply multi-iteration zooming to raise the baselines' effective feature resolution.

Relative depth. We apply the same cosine-similarity correspondence matching procedure described above for optical flow to binocular image pairs, and infer relative depth from the magnitude of the resulting disparity (optical flow) vector, following the same logic as the ZWM depth prompt.

Object segmentation. We compute pairwise cosine similarity between all patch features within each image. Object segments are obtained by thresholding the cosine similarity to a seed patch, grouping patches with high feature similarity as belonging to the same object.

Intuitive physics. ResNet50 and DINOv3 are single-image models and cannot be meaningfully evaluated on our temporal intuitive physics benchmark, so they are excluded from this comparison. For V-JEPA2, we provide the same revealed patches (hand location and background grounding patches) as input and use V-JEPA2 to predict representations for the masked patches. We then compute the cosine similarity between the predicted representations and the representations of the same regions extracted from (i) the initial context frame f_1 and (ii) the ground-truth target frame f_2. If the predicted representations are more similar to f_2 than to f_1, the model is scored as correct.

Task-specific baselines. For each visual-cognitive task, we compare against state-of-the-art task-specific models:

- **Optical flow**: CoTracker3 [50], DPFlow [51], and SeaRAFT [52]. All are supervised models trained with ground-truth flow annotations.
- **Relative depth**: MiDaS-CNN (supervised monocular) [57], MonoDepth2 (self-supervised monocular) [58], ProDepth (self-supervised monocular) [118], and FoundationStereo (supervised binocular) [59]. We also compare against large vision-language models: Gemini-1.5 [54], GPT-4-Turbo [55], and GPT-4o [56].
- **Object segmentation**: Mask2Former [61] (trained on COCO [62]) and SAM2 [63] (trained with large-scale human annotations).
- **Intuitive physics**: No established baselines exist for our novel benchmark; we compare against V-JEPA2 and Baby V-JEPA2.

Supplementary Text

Relation to prior work on frame-pair masked prediction

We separate ZWM's core idea into two components: (i) the representation-learning objective (temporally-factored masked prediction), and (ii) how that representation is used to read out visual structure. On (i) we overlap with prior masked auto-encoding work, e.g., SiamMAE [66]. Our key contribution is (ii), a zero-shot causal-inference mechanism that reads flow, depth, segmentation, and intuitive physics out of the frozen predictor.

ZWM's zero-shot extraction mechanism is different from prior work; it uses approximate causal inference that can be further composed toward increasingly complex abilities. We perturb the input, run the predictor, and compare the perturbed prediction against the unperturbed one, exposing the causal structure of the world (for instance, pixels move together because they belong to the same physical object). This perturb–predict–compare is generic and unifies the extraction of a broad, composable set of quantities — optical flow, relative depth, object segmentation. Feature-propagation approaches such as SiamMAE [66] show neither depth nor intuitive-physics tests, nor a compositional interface.

Furthermore, at its time of release, SiamMAE [66] was the strongest self-supervised correspondence learner on propagation-style benchmarks, but well below task-specialized supervised state-of-the-art. On optical flow specifically, SiamMAE did not evaluate head-to-head against flow methods. By contrast, ZWM is competitive with supervised, task-specialized state-of-the-art across optical flow, relative depth, and segmentation.

Empirically, ZWM's zero-shot causal-inference readout is an important source of performance. Read with feature-propagation, the standard BabyZWM-170M model scores 0.526 on optical flow, 87.2% on relative depth, and 0.417 mIoU on segmentation. Read with approximate causal inference, it achieves 0.616, 94.2%, and 0.496 (Table S18; main-text Figure 4). Hence, applying a SiamMAE-style feature-propagation readout to ZWM recovers substantially less of the representation than our mechanism does, and the matched-probe comparisons in the following subsections show the same pattern across every model we evaluate. The full crossing of training objective with readout mechanism, including the symmetric-masking variants, is reported below under "Ablation: training objective $\times$ readout mechanism".

What we mean by "zero-shot"

By "zero-shot" we mean the model performs a task with no task-specific training and examples.

We wish to highlight an important distinction between *examples* and *instructions*. In the LLM literature, models are commonly provided with task instructions (what the model is supposed to do) and examples of the task being completed. "Zero-shot" refers to performance when these task-specific examples are omitted. This aligns with the machine learning literature, which contrasts "few-shot" against "zero-shot" learning. In few-shot learning, several examples are provided to the model, either in context or for updating the weights through learning; in image recognition, for example, few-shot learning refers to providing a few exemplars of a novel, unseen category to a model, and using it to perform image recognition for the category. "Zero-shot" is the complement of the term, meaning that no examples are provided. ZWM's readout programs serve as *instructions* — without examples — just like prompts for LLMs.

Our usage of the term also aligns with a developmental framing. A child — whom the model is meant to model — comes to estimate motion, depth, object segments, and physical outcomes without ever being trained on those tasks or shown labeled examples of them. ZWM mirrors this.

Some task-specific "domain knowledge" or "instructions" must nonetheless be supplied to any model. This can come in the form of supervised labels, or alternatively as instructions that communicate a procedure for solving a task, exactly what our zero-shot probes are. Otherwise there is no way for a system to know what task to perform and how. Consequently, the nearest-neighbor feature-matching probe we apply to the representation baselines is also a form of task-specific instruction. For flow extraction using feature-NN matching, the specific set of instructions are: pass both frames into the model, build representations across the model, pick a specific layer (the specific layer is an important, task-specific choice), take the query patch's representation, use cosine similarity to find the target patch with the closest representation. All these steps are important innate elements provided to the model, and they are task-specific; the steps are different for performing object segmentation, which instead involves single images and thresholding cosine similarity. In this sense, feature-NN matching encodes domain knowledge, and is similarly "zero-shot" as our approximate causal-inference mechanism.

Although we provide innate task-specific structure, our framework of approximate causal inference is generic and unifies several tasks, each using the same high-level recipe. First, construct a minimal perturbation that intervenes on a latent cause governing some visual quantity. Second, compare the prediction under the perturbed input against the prediction under the original input. Third, aggregate the difference to extract the quantity of interest. Flow, depth, segments, and intuitive physics differ only in which latent cause is intervened upon; the overall recipe — perturb, predict, compare/aggregate — is unchanged. The developmental plausibility of this built-in structure, and the sense in which the readout programs are hypothesized innate priors rather than learned task heads, are discussed in the main text.

Furthermore, our usage of the term "zero-shot" is stricter and more principled than common usage in AI in several respects. Recent computer-vision work uses "zero-shot" to mean out-of-distribution generalization: a model is trained on a task with labels, then transferred "zero-shot" to the same task on unseen benchmarks [63]. Our usage is stricter, since we never train on the task and use no task labels at any point. In the LLM literature, "zero-shot" has come to mean flexibly performing a new task from a prompt; but because such models are trained on enormous corpora, many "zero-shot" tasks are basically interpolations of the training distribution, and what counts as genuinely "zero-shot" is ambiguous. By that measure our setting is again more demanding: BabyZWM learns from a few hundred hours of children's video and is tested for capabilities it was never trained on.

Optical flow: Matched feature-NN probing across models

We apply the same matched feature-NN probe to every model, sweeping the transformer layer from which the frozen features are read (Table S6) and reporting each model at its best-performing layer (Table S7).

Table S6: Matched feature-NN flow probe: per-layer sweep, all ZWM models. The same nearest-neighbor cosine patch-feature probe used for the representation baselines, applied to each model's own frozen features (in-context two-frame encoding, parameter-free LayerNorm), on TAP-Vid-DAVIS two-frame "first" (position accuracy, higher is better). 170M sweeps layers $\{0, 4, 8, 12, 16, 20, -1\}$; 1B sweeps $\{0, 8, 16, 24, 32, 40, -1\}$; — = layer not in this model's sweep grid.

Layer	ZWM-170M (BigVideo)	ZWM-1B (BigVideo)	BabyZWM-170M (BabyView)	BabyZWM-1B (BabyView)
L0	0.191	0.142	0.222	0.159
L4	0.220	—	0.317	—
L8	0.240	0.179	0.410	0.315
L12	0.325	—	0.489	—
L16	0.423	0.303	0.526	0.414
L20	0.455	—	0.526	—
L24	—	0.440	—	0.469
L32	—	0.514	—	0.491
L40	—	0.565	—	0.484
-1	0.312	0.453	0.449	0.276

Table S7: Optical flow, all methods (TAP-Vid-DAVIS "first" position accuracy, Pct). Same 38,881-point set; res = input resolution (patch grid). feature-NN = best over the layer sweep (best layer in parens).

Model	probe	res	Pct
ZWM: native zero-shot inference			
ZWM-170M (BigVideo)	**native** (perturb-compare)	256	0.713
ZWM-1B (BigVideo)	**native** (perturb-compare)	256	**0.740**
BabyZWM-170M (BabyView)	**native** (perturb-compare)	256	0.616
BabyZWM-1B (BabyView)	**native** (perturb-compare)	256	0.650
feature-NN probe			
ZWM-170M (BigVideo)	feature-NN (L20)	256 (32×32)	0.455
ZWM-1B (BigVideo)	feature-NN (L40)	256 (32×32)	**0.565**
BabyZWM-170M (BabyView)	feature-NN (L16)	256 (32×32)	0.526
BabyZWM-1B (BabyView)	feature-NN (L32)	256 (32×32)	0.491
ResNet50	feature-NN (final)	256	0.419
DINOv3 (Large)	feature-NN (final)	224 (14×14)	0.375
Baby DINOv3 (Large)	feature-NN (final)	224 (14×14)	0.349
V-JEPA2 (Large)	feature-NN (final)	256 (16×16)	0.293
Baby V-JEPA2 (Large)	feature-NN (final)	256 (16×16)	0.215
multi-frame tracker (full video, reference)			
CoTracker3 (multi-frame)	supervised (full-video track)	full video	0.745
supervised flow / 2-frame (reference)			
DPFlow	supervised	2-frame	**0.625**
SeaRAFT	supervised	2-frame	0.587
CoTracker3 (2-frame)	supervised	2-frame	0.580
RAFT	supervised	2-frame	0.544
Doduo	supervised	2-frame	0.485
self-supervised (reference)			
GMRW	self-supervised	2-frame	0.546
DiffTrack	self-supervised	2-frame	0.469
SMURF	self-supervised	2-frame	0.442

Relative depth: Matched feature-NN probing across models

The same probe is applied to the binocular stereo pair, again sweeping the feature layer (Table S8) and reporting each model at its best-performing layer (Table S9).

Table S8: Matched feature-NN relative-depth probe: per-layer sweep, all ZWM models (% accuracy). Cosine patch-feature NN probe applied to the binocular stereo pair (disparity correspondence) on UniQA-3D / KITTI (164 pairs); chance $=$ 50%. 170M sweeps $\{0, 4, 8, 12, 16, 20, -1\}$; 1B sweeps $\{0, 8, 16, 24, 32, 40, -1\}$; — $=$ layer not in this model's sweep grid.

Layer	ZWM-170M (BigVideo)	ZWM-1B (BigVideo)	BabyZWM-170M (BabyView)	BabyZWM-1B (BabyView)
L0	53.7	48.8	57.3	51.2
L4	54.9	—	72.0	—
L8	57.9	54.3	81.7	63.4
L12	67.1	—	84.8	—
L16	68.3	64.0	86.0	76.8
L20	74.4	—	87.2	—
L24	—	69.5	—	80.5
L32	—	79.3	—	82.3
L40	—	90.2	—	82.3
-1	74.4	80.5	82.3	68.9

Table S9: Relative depth, all methods (% accuracy; chance = 50%). ZWM-native and reference rows are canonical values; feature-NN-on-ZWM rows are the matched runs (best completed layer in parens).

Model	probe	acc
ZWM: native zero-shot inference		
ZWM-170M (BigVideo)	**native** (perturb-compare)	**98.1**
ZWM-1B (BigVideo)	**native** (perturb-compare)	96.1
BabyZWM-170M (BabyView)	**native** (perturb-compare)	94.2
BabyZWM-1B (BabyView)	**native** (perturb-compare)	89.3
feature-NN probe: ZWM (ours, new)		
ZWM-170M (BigVideo)	feature-NN (L20)	74.4
ZWM-1B (BigVideo)	feature-NN (L40)	**90.2**
BabyZWM-170M (BabyView)	feature-NN (L20)	87.2
BabyZWM-1B (BabyView)	feature-NN (L40)	82.3
feature-NN probe: representation baselines		
ResNet50	feature-NN	71.8
DINOv3 (Large)	feature-NN	63.1
Baby DINOv3 (Large)	feature-NN	72.8
V-JEPA2 (Large)	feature-NN	46.6
V-JEPA2 (Giant)	feature-NN	57.3
Baby V-JEPA2 (Large)	feature-NN	68.0
supervised (reference)		
FoundationStereo (binocular)	supervised	**100.0**
MiDaS-DPT (monocular)	supervised	89.3
MiDaS-CNN (monocular)	supervised	88.3
self-supervised (reference)		
MonoDepth2	self-supervised	94.2
ProDepth	self-supervised	89.3
humans / vision-language (reference)		
Humans (single image)	—	84.5
GPT-4o	VLM	65.0
Gemini-1.5	VLM	64.1
GPT-4-Turbo	VLM	50.5

Object segmentation: Matched feature-NN probing across models

For segmentation the probe has two free choices, the feature layer and the bipartition threshold τ, so we sweep both (Tables S10 and S11) and report each model at its best-performing configuration (Table S17).

Table S10: Matched feature-NN segmentation probe: τ sweep, ZWM-1B (BigVideo). Cosine patch-feature probe on SpelkeBench (547 images): cosine-affinity from each GT seed patch, threshold at bipartition τ, connected component containing the seed, upsample. Metric: IoU. Rows $=$ feature layer; columns $=$ bipartition τ; **bold** $=$ best per row.

Layer	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
L0	.096	.096	.096	.096	.097	.101	.135	**.188**	.180
L8	.096	.096	.096	.098	.109	.149	.264	**.302**	.148
L16	.096	.103	.126	.187	.314	**.385**	.293	.134	.045
L24	.100	.120	.181	.301	**.385**	.346	.208	.090	.037
L32	.102	.123	.184	.296	**.374**	.328	.184	.075	.033
L40	.110	.145	.222	.303	**.327**	.267	.155	.068	.032
L47	.096	.097	.101	.109	.121	.153	.194	**.200**	.089
-1	.099	.106	.119	.134	.165	.207	**.240**	.216	.104

Table S11: Segmentation feature-NN: layer $\times$ model, best over τ. Each cell $=$ best IoU over $\tau \in \{0.1, \ldots, 0.9\}$; — $=$ layer not in this model's sweep grid (170M sweeps $\{0, 4, 8, 12, 16, 20, 23, -1\}$; 1B sweeps $\{0, 8, 16, 24, 32, 40, 47, -1\}$). **bold** $=$ best layer per model.

Layer	ZWM-170M (BigVideo)	ZWM-1B (BigVideo)	BabyZWM-170M (BabyView)	BabyZWM-1B (BabyView)
L0	.258	.188	.295	.197
L4	.268	—	.405	—
L8	.295	.302	.403	.348
L12	**.383**	—	.365	—
L16	.377	**.385**	**.417**	.398
L20	.360	—	.369	—
L23	.214	—	.238	—
L24	—	.385	—	**.414**
L32	—	.374	—	.384
L40	—	.327	—	.344
L47	—	.200	—	.220
-1	.264	.240	.251	.224

Object segmentation: Matched DINOv3 re-run (SpelkeBench)

Table S12: DINOv3 (Large) SpelkeBench IoU (layer $\times$ τ). DINOv3 run through the identical probe as ZWM (same cosine-NN rule, parameter-free LayerNorm, its own τ and layer sweep, 512px $\rightarrow 32 \times 32$ grid, no CRF). **bold** = best per row; `final` = post-LayerNorm.

Layer	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
L0	.096	.097	.101	.111	.132	.166	.216	**.245**	.220
L4	.096	.101	.121	.164	.240	.331	**.335**	.215	.084
L8	.099	.106	.119	.144	.198	.313	**.421**	.247	.064
L12	.105	.144	.251	.418	**.437**	.307	.171	.075	.034
L16	.138	.287	.430	**.435**	.327	.208	.113	.055	.032
L20	.099	.154	.277	.377	**.396**	.333	.219	.113	.047
L23	.096	.096	.096	.096	.096	.096	.096	.149	**.547**
final	.161	.294	.439	.538	**.568**	.538	.462	.336	.157

Table S13: Baby DINOv3 (Large) SpelkeBench IoU (layer $\times$ τ). Same matched probe as Table S12. **bold** = best per row; `final` = post-LayerNorm.

Layer	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
L0	.212	.242	**.248**	.242	.223	.198	.171	.142	.103
L4	.179	.266	**.285**	.253	.201	.148	.103	.069	.045
L8	.148	.272	**.345**	.293	.205	.129	.077	.048	.034
L12	.142	.331	**.440**	.376	.263	.166	.094	.052	.034
L16	.148	.337	.482	**.510**	.446	.335	.217	.116	.048
L20	.118	.235	.394	.495	**.519**	.476	.374	.228	.091
L23	.109	.173	.291	.418	.497	**.507**	.437	.289	.114
final	.154	.264	.387	.482	**.518**	.498	.411	.263	.104

For completeness, the same layer $\times$ τ sweep was run for the remaining representation baselines (Tables S14–S16), so that each of these models is reported at its own best configuration rather than at a setting inherited from another model.

Table S14: V-JEPA2 (Large) SpelkeBench IoU (layer $\times$ τ). Same matched probe as Table S12, at 256px. **Bold** = best per row; `final` = post-LayerNorm. Cells at $\approx$0.096 are the whole-image floor, where the affinity threshold groups the entire image into one segment.

Layer	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
L0	.096	.096	.097	.098	.102	.129	.184	**.216**	.190
L4	.096	.096	.096	.096	.096	.096	.096	.096	**.129**
L8	.096	.096	.096	.096	.096	.096	.096	.096	**.110**
L12	.096	.096	.096	.096	.096	.096	.096	.096	**.178**
L16	.096	.096	.096	.096	.096	.096	.096	.107	**.232**
L20	.096	.096	.096	.096	.096	.103	.203	**.221**	.101
L23	.099	.128	.209	**.252**	.208	.145	.107	.092	.091
final	.099	.128	.209	**.252**	.208	.145	.107	.092	.091

Table S15: Baby V-JEPA2 (Large) SpelkeBench IoU (layer $\times$ τ). Same matched probe, at 256px. **Bold** $=$ best per row.

Layer	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
L0	.181	.220	.254	**.274**	.268	.250	.224	.194	.157
L4	.098	.102	.110	.122	.147	.195	.261	**.268**	.175
L8	.096	.097	.100	.107	.122	.152	.217	**.298**	.199
L12	.096	.096	.097	.100	.113	.142	.216	**.318**	.191
L16	.096	.096	.097	.103	.128	.187	.313	**.323**	.150
L20	.096	.096	.100	.121	.179	.297	**.372**	.268	.125
L23	.096	.096	.103	.128	.198	.319	**.373**	.255	.120
final	.096	.096	.103	.128	.198	.319	**.373**	.255	.120

Table S16: ResNet50 (ImageNet-supervised) SpelkeBench IoU (stage $\times$ τ). Same matched probe, at 256px; rows are residual stages. **Bold** $=$ best per row.

Stage	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
S0	.096	.096	.096	.098	.103	.114	.136	.179	**.243**
S1	.096	.096	.096	.096	.097	.108	.148	**.177**	.070
S2	.096	.096	.099	.121	.214	**.248**	.149	.065	.034
S3	.096	.112	.235	.388	**.397**	.324	.225	.141	.097
S4	.234	.300	.331	**.338**	.319	.284	.245	.216	.204

Table S17: SpelkeBench segmentation, all methods (overall mIoU, 547 imgs). res = input resolution (patch grid for the feature-NN probes); feature-NN = best over the layer $\times$ τ sweep. ZWM is evaluated at 256px throughout, so the 256 rows are the like-for-like comparison; the 512 rows give those baselines four times ZWM's input resolution. †V-JEPA2 at 512 is additionally out of distribution for its processor, which normally fixes input to 256px. ‡V-JEPA2 (Giant) is run at its native 384px (24×24), the same checkpoint used for the other evals.

Model	probe	res	mIoU
ZWM: native zero-shot inference			
ZWM-170M (BigVideo)	**native** (perturb-compare)	256	0.522
ZWM-1B (BigVideo)	**native** (perturb-compare)	256	**0.610**
BabyZWM-170M (BabyView)	**native** (perturb-compare)	256	0.496
BabyZWM-1B (BabyView)	**native** (perturb-compare)	256	0.500
feature-NN probe: 256px input			
ZWM-170M (BigVideo)	feature-NN	256 (32×32)	0.383
ZWM-1B (BigVideo)	feature-NN	256 (32×32)	0.385
BabyZWM-170M (BabyView)	feature-NN	256 (32×32)	0.417
BabyZWM-1B (BabyView)	feature-NN	256 (32×32)	0.414
DINOv3 (Large)	feature-NN	256 (16×16)	0.484
Baby DINOv3 (Large)	feature-NN	256 (16×16)	0.484
V-JEPA2 (Large)	feature-NN	256 (16×16)	0.252
Baby V-JEPA2 (Large)	feature-NN	256 (16×16)	0.373
V-JEPA2 (Giant)	feature-NN	384 (24×24)‡	0.257
ResNet50	feature-NN	256 (16×16)	0.396
feature-NN probe: 512px input			
DINOv3 (Large)	feature-NN	512 (32×32)	0.568
Baby DINOv3 (Large)	feature-NN	512 (32×32)	0.519
V-JEPA2 (Large)	feature-NN	512 (32×32)†	0.305
Baby V-JEPA2 (Large)	feature-NN	512 (32×32)†	0.383
supervised (reference)			
SAM2	supervised	256	**0.622**
Mask2Former-COCO-Panoptic	supervised	256	0.506
Mask2Former-COCO-Instance	supervised	256	0.362
Mask2Former-ADE-Panoptic	supervised	256	0.308
Mask2Former-Cityscapes-Pan.	supervised	256	0.078
Mask2Former-Cityscapes-Inst.	supervised	256	0.032

Ablation: training objective × readout mechanism

To separate the contributions of the training objective and the readout mechanism, we cross the two dimensions while holding everything else fixed (same BabyZWM-170M architecture, same BabyView data, same capacity). "Standard" BabyZWM uses an asymmetric mask (f_1 fully visible, f_2 90%-masked), which prioritizes learning motion dynamics; we compare it against symmetric variants (45%–45% and 90%–90%) that resemble the masking used by popular masked video models such as VideoMAE [21]. We read each checkpoint two ways: with the matched, label-free feature-NN probe (which measures how decodable the frozen representation is) and with our native perturb-predict-compare mechanism (main-text Figure 4). Crossing the objective with the readout gives four primary settings (plus the 45%–45% variant, which is strictly worse on every task). Neither factor alone accounts for ZWM's performance: the full combination (asymmetric objective, causal-inference readout) is the best setting on every task, and degrading either factor costs performance on every task.

Table S18: Training objective × readout ablation (BabyZWM-170M). Same architecture, BabyView data, and capacity; only the masking objective differs. Each checkpoint is read with the matched, label-free feature-NN probe (best over the layer sweep, best layer in parentheses) and with our native perturb-predict-compare mechanism. Optical flow: TAP-Vid-DAVIS position accuracy; relative depth: % of pairs ordered correctly (chance = 50%); segmentation: SpelkeBench overall mIoU (547 images). Under the fixed probe, both symmetric variants show substantially degraded representations across all three tasks — the representation itself is worse, not merely less amenable to the mechanism.

	optical flow		relative depth		segmentation	
objective	feature-NN	native	feature-NN	native	feature-NN	native
asymmetric 90% (standard)	0.526 (L16)	0.616	87.2 (L20)	94.2	0.417 (L16)	0.496
symmetric 90%–90%	0.344 (L16)	0.542	69.5 (L16)	71.8	0.379 (L12)	0.479
symmetric 45%–45%	0.127 (L8)	0.482	54.3 (L12)	73.8	0.163 (L0)	0.213

Table S19: Feature-NN relative-depth per-layer sweep for the objective ablation (% accuracy, chance = 50%). The standard asymmetric objective climbs to 87.2% at intermediate/deep layers, whereas both symmetric variants stay far lower at every layer — the 45%–45% variant never exceeds ∼54% (near chance). **Bold** = best layer per model.

Layer	standard	symmetric 45%–45%	symmetric 90%–90%
L0	57.3	53.1	53.7
L4	72.0	48.1	54.9
L8	81.7	52.4	61.5
L12	84.8	**54.3**	64.6
L16	86.0	53.0	**69.5**
L20	**87.2**	53.0	59.8
final	82.3	48.2	60.4

Both readouts degrade together when the objective is weakened: the feature-NN probe does *not* stay flat while only the native mechanism drops. Because the architecture and capacity are

held fixed and only the objective changes, this isolates the temporally-factored, motion-prioritizing objective as what *produces* the correspondence structure that both a fixed feature-matching probe and our zero-shot mechanism rely on. Mechanistically, predicting a heavily-masked second frame from a fully-visible first frame forces the model to learn cross-frame correspondence, which symmetric masking never requires; accordingly the most correspondence-dependent readouts (optical flow and binocular stereo depth) collapse hardest, with the 45%–45% variant falling to near chance on depth at every layer (Table S19).

Concentration of information in sparse patches

Main-text Figure 4B asks how much of the target frame the predictor needs before it can reconstruct the rest. We sample 100 clips at random (fixed seed) from the Kinetics-400 validation set, skipping any clip shorter than 20 frames. From each we draw one frame pair: a uniformly random start index and a second frame 5–15 frames later, comparable to the inter-frame gap used in training. Frames are resized and centre-cropped to 256×256 and patchified into 8×8 patches, giving $32 \times 32 = 1024$ patches per frame.

The context frame f_1 is fully visible throughout; only the proportion of f_2 revealed varies, over $\{0, 0.5, 1, 2, 3, 5, 7.5, 10, 15, 20, 30\}\%$ of the 1024 patches. Two details make the measurement non-trivial. First, before revealing anything we reserve a random 20% of f_2's patches (204 patches) as an evaluation set that is *never* revealed at any proportion, and we score mean squared error *only* on those patches; revealed patches are drawn exclusively from the remaining 820 patches. Without this reservation the error would fall simply because more of the frame is handed to the model verbatim rather than predicted, and the curve would measure the protocol rather than the representation. Second, the revealed sets are nested — each larger set contains the smaller — so adding a point only ever adds information.

We normalize each clip's curve to its own error with no patches revealed and report percentages; the mean is taken across clips and the 95% confidence interval comes from 10,000 bootstrap resamples of the 100 clips, computed after normalization.

Table S20: Held-out prediction error versus the proportion of target-frame patches revealed. BabyZWM-170M under the two training objectives, as % of each clip's error with no patches revealed (mean over 100 Kinetics-400 clips, 95% bootstrap CI in brackets). Error is scored only on a reserved 20% of patches never revealed at any proportion.

% revealed	temporally factored (0–90)	symmetric (90–90)
0	100.0	100.0
0.5	48.9 [42.9, 55.0]	59.5 [54.9, 64.0]
1	35.1 [30.5, 39.8]	44.7 [40.4, 48.9]
2	25.3 [21.9, 28.7]	34.0 [30.5, 37.4]
3	19.9 [17.1, 22.8]	28.4 [25.5, 31.4]
5	15.8 [13.4, 18.2]	23.6 [21.0, 26.3]
7.5	12.6 [10.8, 14.5]	20.1 [17.8, 22.5]
10	**10.5 [9.0, 12.0]**	**17.6 [15.5, 19.9]**
15	8.0 [6.8, 9.2]	14.9 [12.9, 17.0]
20	6.3 [5.4, 7.4]	13.1 [11.2, 15.1]
30	4.9 [4.1, 5.8]	11.9 [10.1, 14.0]

Revealing 1% of patches (10 of 1024) already reduces prediction error by 65%, and 10% reduces it by 90%: the information needed to predict the frame is concentrated in a small subset of patches. ZWM's temporally-factored objective concentrates it more than the symmetric ablation, with lower error at every proportion.

Intuitive physics: Attention head analysis

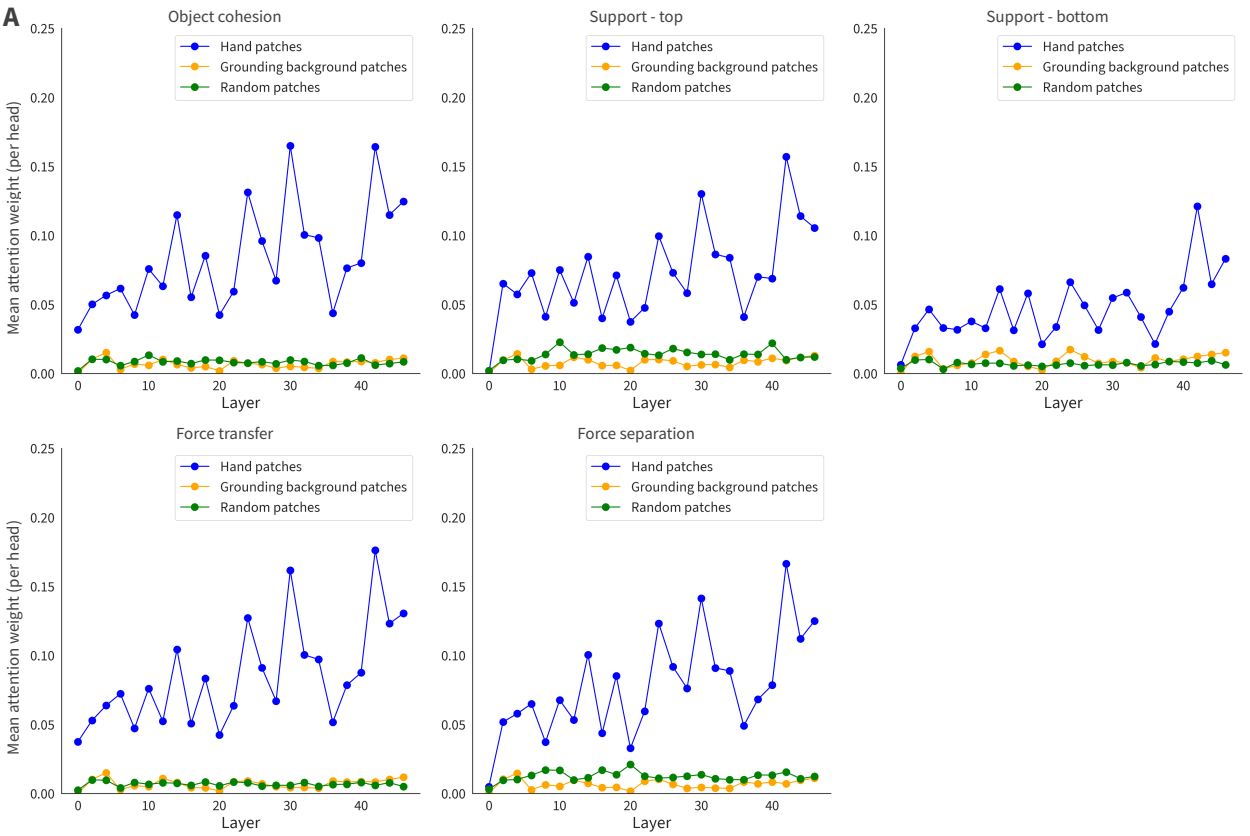


Figure S1: Attention head analysis for intuitive physics. Layer-wise average attention weights from the moved object's query patch to hand patches, background grounding patches, and random patches, shown for each intuitive physics category. In deeper layers, attention is disproportionately allocated to the hand—the causal agent of object motion—relative to background and random patches.

To understand how BabyZWM implements intuitive physical reasoning internally, we analyze the attention patterns of individual transformer heads during the intuitive physics evaluation.

Methodology. For each intuitive physics example, we extract the full attention weight tensor across all layers and heads during the factual prediction forward pass. The model receives two frames as input: the context frame f_1 (1024 tokens) and the partially unmasked target frame f_2 (1024 tokens), yielding 2048 total tokens. We select a *query patch* located on the moved object in f_2 (identified from human annotations of the object's position in the target frame) and examine which tokens this query patch attends to across all layers.

We partition the key tokens into three groups:

- **Hand patches**: The 16 patches (forming a 4×4 patch region, i.e., 32×32 pixels) centered on the hand's revealed location in f_2—the causal agent responsible for the object's motion.
- **Background grounding patches**: The patches revealed at fixed background locations (top and bottom image borders) that anchor camera pose and illumination, providing non-causal

contextual information.

- **Random patches**: An equal number of patches sampled randomly from f_2, excluding hand and background patches, serving as a baseline.

Quantitative analysis. For each layer, we compute the average attention weight (averaged across all heads) from the query patch to each of the three patch groups. To ensure comparability across groups of different sizes, we normalize the background attention by the ratio of background patches to hand patches. We average these layer-wise attention profiles across all examples within each intuitive physics category (e.g., object cohesion, support, force transfer) and across 8 random seeds.

Results. The layer-wise attention profiles reveal that in deeper transformer layers, attention from the moved object's query patch tends to be disproportionately allocated to the hand patches relative to both background grounding patches and random patches. This pattern is broadly consistent across intuitive physics categories, suggesting that the model may have learned to preferentially attend to the hand—the causal agent of object motion—when predicting physical outcomes. While these attention patterns are consistent with a "causal attention" interpretation, we note that attention weights are an indirect measure of the model's internal computations, and further mechanistic work would be needed to confirm whether these heads play a causal role in the model's physical predictions. Nonetheless, the emergence of hand-directed attention in deeper layers is suggestive of a hierarchical computation in which later layers begin to integrate information about agent–object relationships relevant to physical prediction. Layer-wise attention profiles for each intuitive physics category are shown in Figure S1.

Intuitive physics: Flow-based motion-fidelity metric

Table S21: Flow-based motion-fidelity metric for intuitive physics (BabyZWM-1B), per category. "Primary" denotes the manipulated-object region. Acc = % of examples within 50% normalized flow error; EPE in pixels (lower is better); graded = fraction of the context→target gap the prediction closes (0–1, higher is better). For force-separation, the secondary object has EPE 1.1 px (correctly near-zero motion), so read EPE rather than % accuracy there.

Category	Prim. acc	Prim. EPE (px)	Prim. graded	Overall acc
Cohesion	0.98	6.9	0.93	0.98
Support (top)	0.84	22.6	0.79	0.81
Support (bottom)	0.83	21.3	0.74	0.83
Force transfer	0.93	14.4	0.89	0.86
Force separation	0.99	9.2	0.90	0.99
Overall	**0.91**	**14.9**	**0.85**	**0.90**

Table S22: Graded closeness score: model comparison across all categories. Graded = $d(\text{pred}, f_1)/[d(\text{pred}, f_1) + d(\text{pred}, f_2)] \in [0, 1]$ (1 = matches the correct outcome f_2; 0 = predicts no change, i.e. stays at the context f_1). MSE and flow graded require a pixel prediction and exist only for the generative ZWM models; feature graded (cosine distance) is the only version available for V-JEPA2, which outputs no pixels. Accuracy (all categories combined) is whole-image accuracy—whether the prediction is closer to the correct outcome (f_2) than to the context (f_1)—and is reported under each distance that applies to a given model class: pixel MSE and LPIPS for the generative ZWM models, and feature cosine for V-JEPA2, which outputs no pixels (the same whole-image metric used for V-JEPA2 in the main-text object-reasoning figure). MSE and LPIPS capture different things, raw pixel error versus learned perceptual similarity, yet agree to within 0.04 on every ZWM model, so the result does not hinge on the choice of distance. Accuracy is near-ceiling for the stronger models; the graded columns, on the primary (manipulated-object) region, provide the finer comparison. V-JEPA2's feature graded sits near 0.5 because cosine distances between frames are compressed, so it is not on the same scale as the ZWM pixel and flow graded.

	accuracy			graded		
Model	MSE	LPIPS	feature	MSE	flow	feature
ZWM-170M (BigVideo)	0.98	1.00	—	0.87	0.90	—
ZWM-1B (BigVideo)	1.00	1.00	—	0.90	0.93	—
BabyZWM-170M	0.94	0.98	—	0.77	0.79	—
BabyZWM-1B	0.95	0.99	—	0.84	0.85	—
V-JEPA2 (Large)	—	—	0.96	—	—	0.57
V-JEPA2 (Giant)	—	—	0.91	—	—	0.62
Baby V-JEPA2 (Large)	—	—	0.37	—	—	0.43

Intuitive physics: Qualitative examples

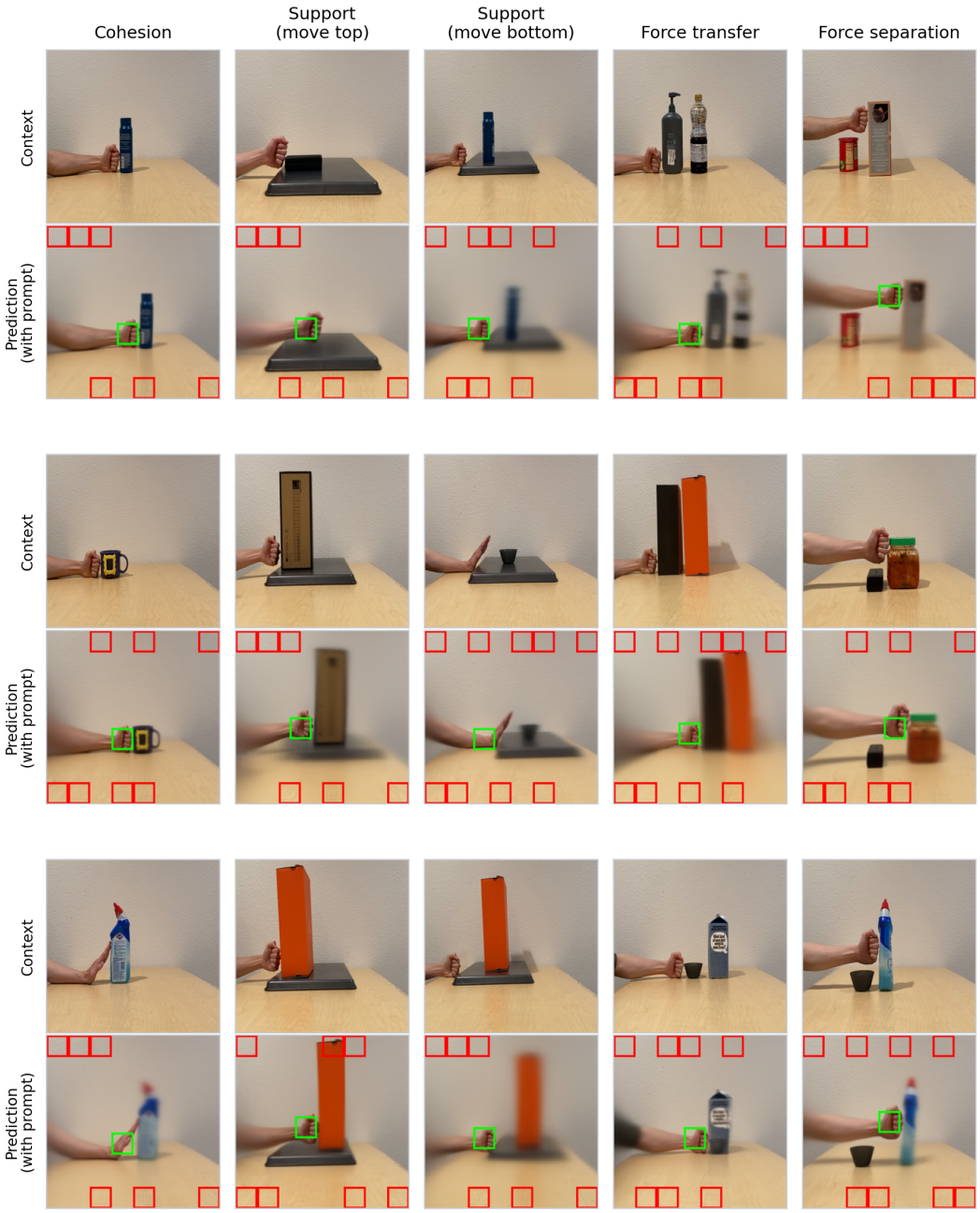


Figure S2: Qualitative intuitive-physics predictions (BabyZWM-1B). Three randomly selected examples per category; each pair shows the context frame (top) above the model's prediction for the masked target frame (bottom). Green marks the revealed patch giving the hand's target location, red the fixed border patches signalling a static camera — these are the only target-frame patches the model receives.

Support predictions are visibly less sharp than cohesion and force separation (Figure S2), consistent with the graded motion-fidelity scores in Table S21.

Intuitive physics: Mass sensitivity

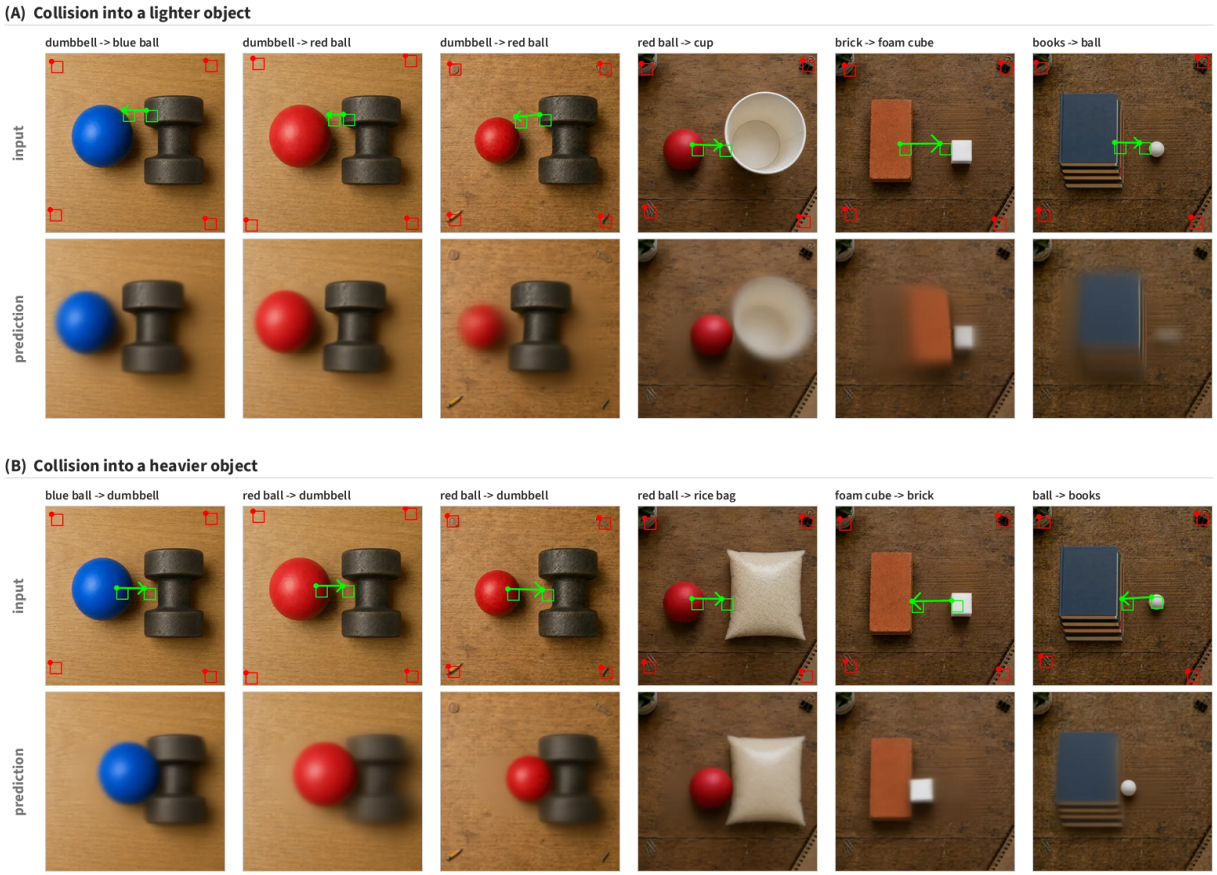


Figure S3: Preliminary qualitative test of mass sensitivity (BabyZWM-1B). Each column is one scene: the annotated input above, the model's prediction below. Green marks the prompted motion (the start and end position of the moved patch); red marks the background patches that hold the scene fixed. No human hand appears in any scene. **(A)** The heavier object is moved into the lighter one, which the model predicts to be additionally displaced by the contact. **(B)** The lighter object is moved into the heavier one, which the model predicts to stay in place while the lighter object comes to rest against it.

Our main intuitive-physics benchmark is not designed to test mass sensitivity. There, a human hand drives the motion, and a hand-pushed light object is still physically expected to cause a heavier one to move—so the two directions carry the same expected outcome and no asymmetry is available to measure. We therefore ran a separate preliminary qualitative test in which no hand appears, so that the only thing that can displace the struck object is the object that strikes it.

Stimuli. We use tabletop scenes containing a heavier and a lighter object, whose masses are evident from familiar appearance. Heavy objects include a dumbbell, a brick, a stack of books, and a bag of rice; light objects include a rubber ball, a foam cube, and an empty paper cup. We hold the scene fixed using background patches, as in the main intuitive-physics evaluation, and vary whether we move the lighter object into the heavier object, or vice versa. Notably, we do not use a human hand to drive the motion, for the reason above. In most scenes we hold the scene itself

fixed—the same two objects, positions, and contact geometry—and reverse only which of the two is moved; in one further contrast we instead hold the moved object fixed (the same ball, the same motion) and vary only the object it is moved into, from a light paper cup to a heavy bag of rice. This gives 12 scenes in total, all run through BabyZWM-1B.

Results. The predictions are mass-sensitive (Figure S3). When moving a lighter object into a heavier one, the model predicts the lighter object to travel and come to rest against the heavier one, which stays in place. Whereas when moving the heavier object into the lighter one, the model predicts the lighter object to be additionally displaced by the contact. However, these results are preliminary and qualitative: the sample is small and run at a single seed, and we do not report a quantitative mass-sensitivity metric or vary the mass ratio parametrically.

An in-principle argument. Mass is not a special or separate property of the world requiring dedicated machinery. It is simply another regularity in how objects behave on contact, alongside support, cohesion, and force transfer — all of which BabyZWM does acquire from BabyView. A child's everyday visual experience contains many interactions between objects of differing apparent mass, so a model that learns those other regularities from naturalistic video faces no principled barrier to also learning that heavier-looking objects resist being displaced.

Why a rigorous test needs a richer interface. Testing this rigorously calls for a richer interface than our current ZWM model. These tests carry inherent ambiguity: how fast or forcefully the moving object was travelling (our prompt specifies its displacement, not its speed), and how heavy either object is, are not observable from the input — and therefore neither is how likely the struck object is to move, nor how far. ZWM makes a single deterministic prediction, so it cannot express such a graded expectation; and because it is trained with a regression objective, its prediction under ambiguity averages over the possible outcomes, which shows up as blurry predictions. In follow-up work we are therefore building a variant of ZWM that is distributional; its readout exposes the predicted distribution over how likely the struck object is to move, and how far. It is also multi-frame and has time embeddings, allowing us to specify the speed of object motion. We plan to use it to test mass sensitivity and other physical concepts more rigorously and quantitatively.

Developmental trajectory: emergence timing

Table S23: Developmental emergence timing (BabyZWM-170M). Emergence is measured over each capability's learnable gain (from a baseline to its max across checkpoints). % training = (training step / 200,000) $\times$ 100, and the equivalent days of waking experience = % training $\times$ 0.95 (a full run $\approx$ 95 days); lower = emerges earlier. 95% CIs in brackets (bootstrap).

Capability	% training $\rightarrow$ 50% gain	% training $\rightarrow$ 80% gain	Final-third trend
Relative depth	3.3 [3.0, 3.8]	4.6 [4.2, 15.1]	fluctuates near ceiling
Optical flow	3.6 [3.1, 4.5]	17.4 [5.3, 63.6]	rising slowly
Intuitive physics	5.4 [5.3, 5.7]	8.7 [8.4, 9.0]	peaked early
Object segmentation	12.6 [9.2, 20.0]	69.8 (lower bd)	**rising** (+40%/100k, CI 28–52)

Neural alignment: predictivity results

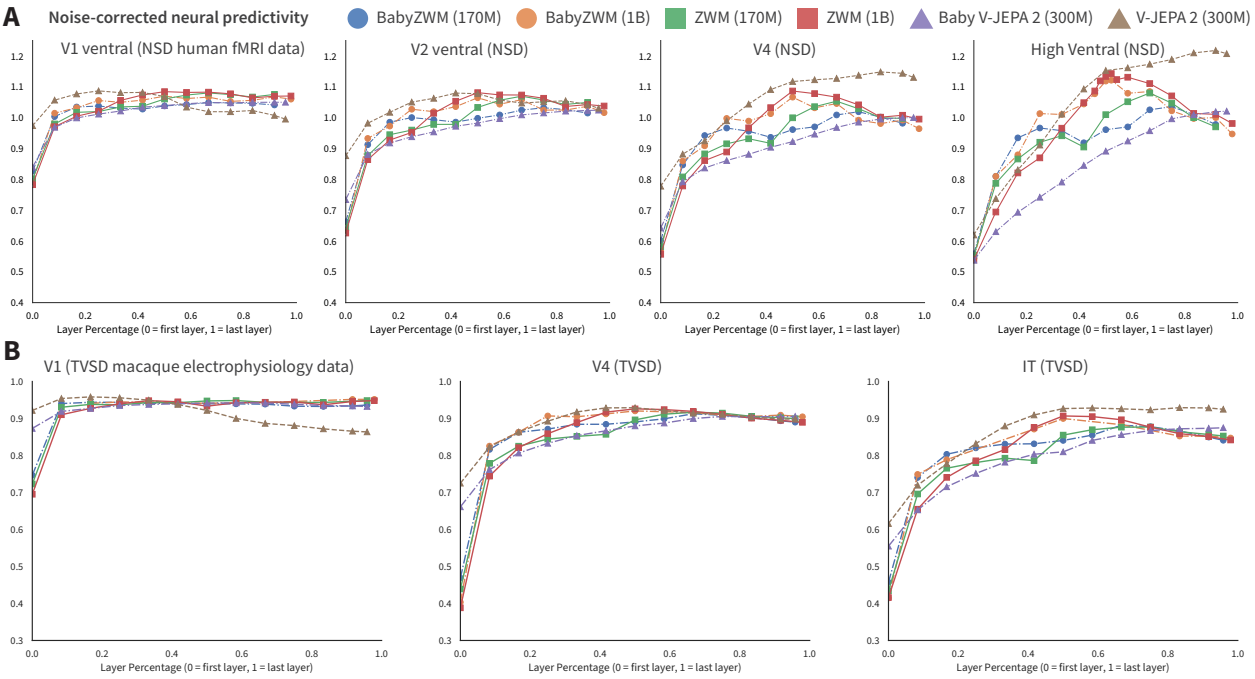


Figure S4: Neural predictivity across NSD and TVSD. (**A**) Cross-validated noise-corrected correlations between model features and human fMRI responses (NSD). (**B**) Cross-validated noise-corrected correlations between model features and macaque single-neuron electrophysiology responses (TVSD). Both benchmarks show hierarchically organized layer-area correspondences for BabyZWM and baseline models.

In the main text, we reported that BabyZWM's internal representations align with hierarchical visual cortex organization using the Natural Scenes Dataset (NSD) [87]. Here we present expanded neural predictivity results across both NSD (human fMRI) and the THINGS Ventral Stream Spiking Dataset (TVSD; macaque single-neuron electrophysiology) [88], providing converging evidence across species and measurement modalities.

Across both benchmarks, BabyZWM exhibits hierarchically organized layer-area correspondence: earlier model layers best predict earlier cortical regions, while deeper layers align with higher visual areas. In NSD, this pattern is evident across all evaluated ROIs; TVSD corroborates these findings at single-neuron resolution, confirming that the representational hierarchy is not an artifact of the fMRI measurement scale. BabyZWM achieves predictivity comparable to ZWM variants trained on substantially larger and more diverse datasets, whereas Baby V-JEPA2 shows lower neural alignment than its larger-data counterpart. Full layer-by-region predictivity profiles for both benchmarks are shown in Figure S4.

To read this layer-area correspondence relative to alternatives, Figure S5 repeats the main-text Figure 6C analysis for every model: per NSD visual area, the fractional encoder depth of the earliest layer to reach the noise ceiling (else the best layer). Two patterns hold. First, hierarchical layer-area correspondence is present in every model — earlier layers reach the ceiling for earlier visual areas and progressively deeper layers for higher areas — so BabyZWM's hierarchy, learned from a single child's video, is neither an artifact nor unique to it, and is comparable to models trained on orders of

magnitude more data, consistent with established accounts of hierarchical visual representation. Second, the developmental diet affects the two objectives differently (each column pairs a standard model, top, with its BabyView-trained counterpart, bottom): ZWM's correspondence is essentially unchanged under the baby diet, whereas V-JEPA2's shifts markedly deeper, reaching the ceiling only in its late layers even for early areas. This complements the baby-data penalty (Table S24) — feature-prediction degrades under a child's data budget while the ZWM objective remains robust — here visible as *where in depth* alignment is achieved rather than its magnitude. Because the readout marks the first layer to reach the noise ceiling (else the best layer), it is a coarse, ordinal summary; we therefore read the rising trend and the standard-vs-baby shift, not exact bar heights.

An objective-based account of the layer-wise profiles. As a provisional hypothesis, we offer an objective-based account of the layer-wise shapes in main-text Figure 6D: the profiles separate cleanly by training target, with ZWM peaking at intermediate depth and then declining toward its final layers (most so for higher areas), whereas V-JEPA2 rises monotonically to a late peak. We interpret this as a consequence of what each model predicts. ZWM's terminal layers are optimized to predict pixels, an appearance-level target not aligned with high-level cortex, so its deepest representations specialize away from IT- and high-ventral-like features; V-JEPA2 predicts high-level latent features and so keeps its deepest layers aligned with higher areas.

Robustness to a developmental data diet. To read BabyZWM's alignment relative to alternatives, we quantify the *baby-data penalty*: the change in peak (best-layer) noise-corrected predictivity when a fixed architecture is retrained on BabyView rather than its standard training set, which isolates the effect of the data regime from the architecture and objective (Table S24). Averaged over higher-order visual areas, V-JEPA2 loses 10.4% of its peak predictivity, whereas ZWM loses only 2.1% (170M) and 1.5% (1B) — a roughly five-fold smaller penalty. Moreover, the V-JEPA2 penalty grows monotonically along the visual hierarchy (from -3.4% at V1 ventral to -16.2% at high ventral), most damaging exactly the higher-area representations of greatest interest, while ZWM's penalty stays small and roughly flat across areas and species. This pattern replicates across both NSD and TVSD, indicating that the ZWM objective is more robust to a developmental data diet than feature-prediction.

Table S24: Baby-data penalty: robustness of neural predictivity to a developmental data diet. Percentage change in peak (best-layer) noise-corrected predictivity when a fixed architecture is retrained on BabyView rather than its standard training set (negative = drop under the developmental diet). Comparisons are architecture-matched: V-JEPA2 vs. Baby V-JEPA2, and ZWM vs. BabyZWM at 170M and 1B. The final row averages over higher-order visual areas (NSD V4, mid-ventral, and high-ventral; macaque V4 and IT).

Region	V-JEPA2 (300M)	ZWM (170M)	ZWM (1B)
V1 ventral (NSD)	-3.4	-2.9	-1.3
V2 ventral (NSD)	-5.1	-3.5	-1.5
V3 ventral (NSD)	-8.3	-3.3	-1.5
V4 (NSD)	-12.9	-3.2	-1.8
Mid ventral (NSD)	-14.6	-3.3	-2.3
High ventral (NSD)	-16.2	-4.1	-1.9
V1 (TVSD)	-1.7	-0.5	$+0.4$
V4 (TVSD)	-2.5	-0.4	-0.6
IT (TVSD)	-5.8	$+0.6$	-0.8
Higher-area mean	$\mathbf{-10.4}$	$\mathbf{-2.1}$	$\mathbf{-1.5}$

Developmental trajectory relative to converged baselines. The main-text developmental trajectory (Figure 6B) plots predictivity against training progress for BabyZWM alone, because its x-axis — equivalent days of a child's waking experience — is calibrated to BabyZWM's developmental diet and is not shared by models trained on different data and schedules; a baseline therefore cannot be placed on it as a trajectory. To still read BabyZWM's alignment relative to alternatives, Figure S6 keeps BabyZWM's trajectory and adds each baseline's converged (best-layer) predictivity as endpoint markers in a horizontally extended margin, on the same predictivity axis. The comparison reinforces the baby-data-penalty result: under a matched developmental diet, BabyZWM aligns better with cortex than Baby V-JEPA2 at every area, and although the off-the-shelf V-JEPA2 is the single strongest higher-area predictor, that advantage is bought entirely by its large-scale, non-developmental training data.

Table S25: Converged neural predictivity across models and cortical areas. Peak (best-layer) noise-corrected predictivity at convergence, for each model and cortical area; NSD values are averaged over the ROIs within each area group, TVSD values are single ROIs. The final column averages over higher-order visual areas (NSD V4, mid-ventral, high-ventral; macaque V4 and IT), matching the baby-data-penalty definition (Table S24). These are the converged endpoints plotted in Figure S6. Under a matched developmental diet, BabyZWM aligns better than Baby V-JEPA2 at every area, and the gap widens up the hierarchy. Off-the-shelf V-JEPA2 is the strongest higher-area predictor, but retraining the same architecture on the developmental BabyView diet (Baby V-JEPA2) collapses it to the weakest model in higher areas. Values above 1.0 exceed the noise ceiling and should be read cautiously, as the noise-corrected ratio there is increasingly dominated by ceiling-estimation noise.

	NSD (human fMRI)						TVSD (macaque)			Higher-area
Model	V1	V2	V3	V4	Mid	High	V1	V4	IT	mean
BabyZWM (170M)	1.04	1.03	1.05	1.02	1.02	0.99	0.94	0.91	0.88	0.97
BabyZWM (1B)	1.06	1.06	1.08	1.07	1.09	1.08	0.95	0.92	0.90	1.01
ZWM (170M)	1.07	1.07	1.09	1.05	1.06	1.03	0.95	0.92	0.88	0.99
ZWM (1B)	1.07	1.08	1.10	1.09	1.11	1.11	0.95	0.93	0.91	1.03
Baby V-JEPA2 (300M)	1.04	1.02	1.03	1.00	1.00	0.97	0.94	0.91	0.87	0.96
V-JEPA2 (300M)	1.07	1.08	1.12	1.15	1.18	1.19	0.96	0.93	0.93	1.07

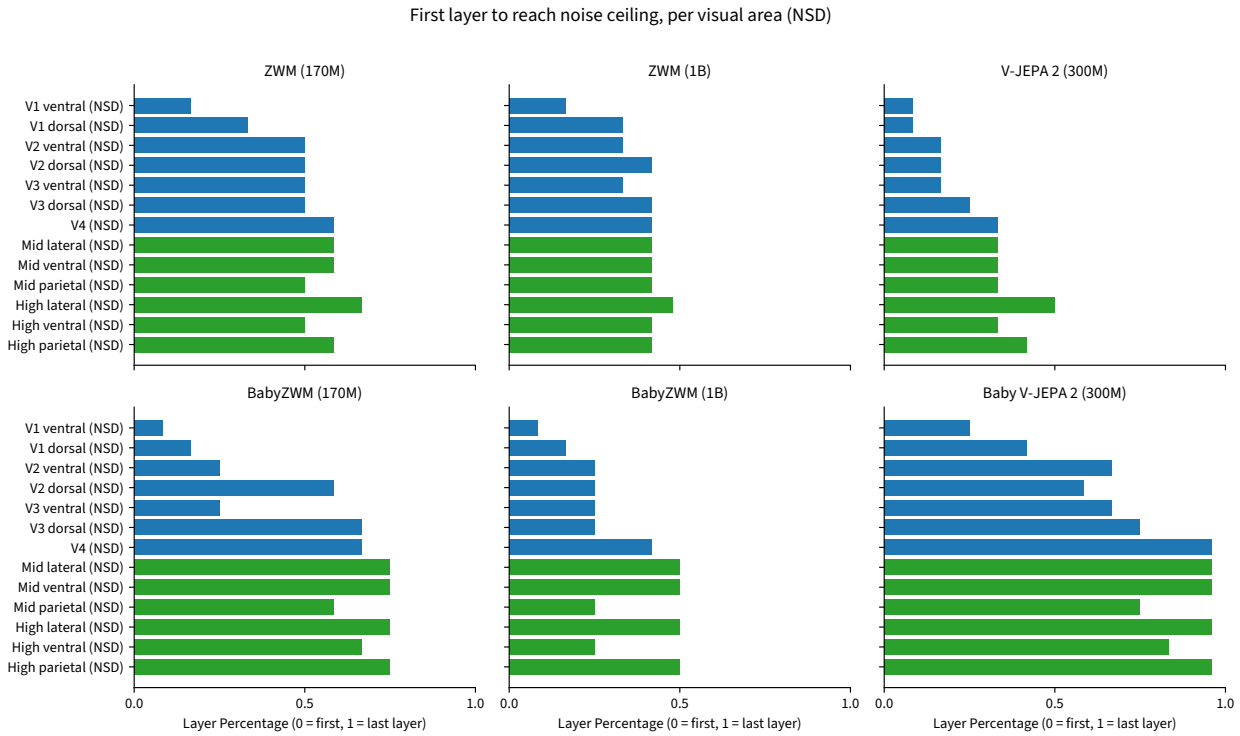


Figure S5: Neural predictivity: layer-area correspondence across models (NSD). One panel per model, arranged so each column pairs an architecture (170M, 1B, V-JEPA2) with its BabyView-trained counterpart (top: standard data; bottom: developmental diet). For each NSD visual area (rows, early areas in blue at top to mid/high areas in green at bottom), the bar shows the fractional encoder depth ($0 =$ first layer, $1 =$ last layer) of the earliest layer to reach the noise ceiling (else the best layer); a profile whose bars lengthen from early (blue) to higher (green) areas indicates hierarchical layer-area correspondence, with earlier model layers best predicting earlier cortical regions and deeper layers higher regions. Every model shows this hierarchy, so BabyZWM's correspondence — from a single child's video — matches the much-larger-data baselines. Across the developmental diet (each column, top vs. bottom), ZWM's correspondence is largely unchanged whereas V-JEPA2's shifts to its late layers, complementing the baby-data penalty (Table S24). The readout is a coarse, ordinal summary, so the rising trend and the standard-vs-baby shift should be read rather than exact bar heights.

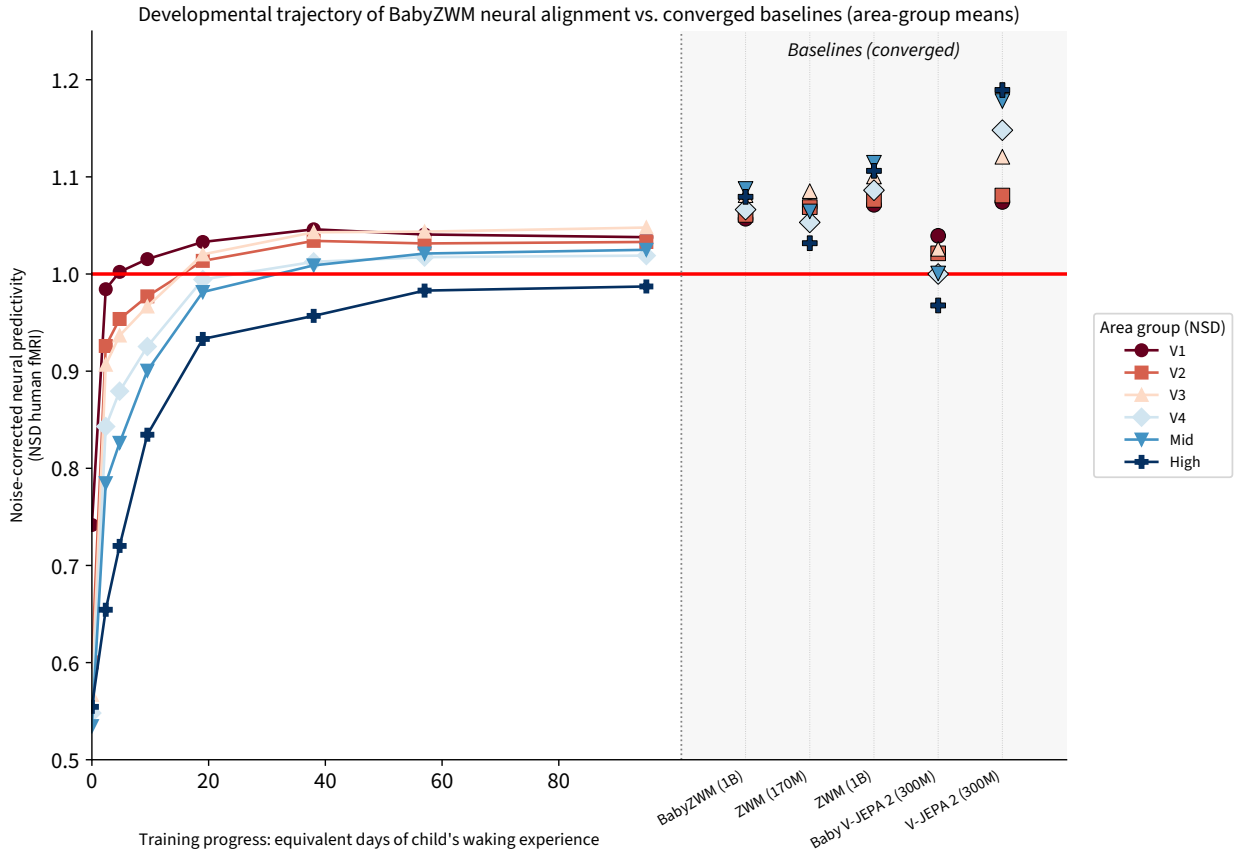


Figure S6: Neural predictivity: developmental trajectory of BabyZWM relative to converged baselines (NSD). BabyZWM-170M's noise-corrected predictivity is plotted against training progress (equivalent days of a child's waking experience), summarized as the mean over NSD ROIs within each cortical area group (V1–High); the bold red line marks the noise ceiling (predictivity $= 1$). To the right of the divider, each baseline's converged (best-layer) predictivity is shown as endpoint markers on the same axis; the baselines have no comparable training-progress axis (it is calibrated to BabyZWM's developmental diet), so they are shown as converged endpoints rather than trajectories. Two comparisons matter. (i) Under a matched developmental diet, BabyZWM aligns better than Baby V-JEPA2 at every area, and the gap widens up the hierarchy (higher-area mean predictivity 0.97 vs. 0.96 at 170M, and 1.01 vs. 0.96 at 1B). (ii) The off-the-shelf V-JEPA2 is the single strongest higher-area predictor, but this advantage is bought entirely by large-scale, non-developmental training data: retraining the same architecture on BabyView (Baby V-JEPA2) collapses it to the weakest model in higher areas, below BabyZWM. Predictivity above the noise ceiling falls in a regime where the noise-corrected ratio is dominated by ceiling-estimation noise and should be read cautiously. Together these support reading BabyZWM's brain alignment as strong relative to developmentally-matched alternatives.