# Learning to segment self-generated from externally caused optic flow through sensorimotor mismatch circuits

Matthias Brucklacher[1,*], Giovanni Pezzulo[2], Francesco Mannella[2], Gaspare Galati[3], Cyriel M. A. Pennartz[1]

[1]Cognitive and Systems Neuroscience, University of Amsterdam, Amsterdam, 1098XH Amsterdam, Netherlands

[2]Institute of Cognitive Sciences and Technologies, National Research Council, 00185 Rome, Italy

[3]Brain Imaging Laboratory, Department of Psychology, Sapienza University, 00185 Rome, Italy

*Corresponding author. Email address: m.m.brucklacher@uva.nl

Preprint posted on BioRxiv

**Classification:** Biological Sciences/Neuroscience

**Keywords:** Generative model, Object segmentation, Predictive coding, Optic Flow, Sensorimotor

**Abstract**

Efficient sensory detection requires the capacity to ignore task-irrelevant information, for example when optic flow patterns created by egomotion need to be disentangled from object perception. Distinguishing self- from externally caused changes in visual input is thus an important problem that the visual system needs to solve. Predictive coding with sensorimotor mismatch detection is an attractive starting point to investigate this question computationally. Although experimental evidence for sensorimotor mismatch signals in early visual areas exists, it is not understood how it is functionally integrated into cortical networks that perform input segmentation and categorization. Our model advanced a novel, biologically plausible solution to this question, which extends predictive coding models with the ability to distinguish self-generated from externally caused optic flow. We first show that a simple three neuron microcircuit produces experience-dependent sensorimotor mismatch responses, in agreement with calcium imaging data from mice. This microcircuit is then integrated into a predictive coding neural network with two generative streams. The first stream is motor-to-visual and consists of many microcircuits in parallel. This stream learns to spatially predict optic flow resulting from self-motion and mirrors connections from motor cortex to V1. The second stream is visual-to-visual. Bidirectionally connecting Middle Temporal cortex to V1, it assigns a crucial role to the abundant feedback connections between these areas: the maintenance of a generative model of externally caused optic flow. In the model, area MT learns to segment moving objects from the background, and facilitates object categorization. Our model extends the framework of Hebbian predictive coding to sensorimotor settings, in which the agent is not a passive observer of external inputs, but actively moves - and learns to predict the consequences of its own movements.

## Significance statement

This research addresses a fundamental challenge in sensory perception: how the brain distinguishes between self-generated and externally caused visual motion. Using a computational model inspired by predictive coding and sensorimotor mismatch detection, the study proposes a biologically plausible solution. The model incorporates a neural microcircuit that generates sensorimotor mismatch responses, aligning with experimental data from mice. This microcircuit is integrated into a neural network with two streams: one predicting self-motion-induced optic flow and another maintaining a generative model for externally caused optic flow. The research advances our understanding of how the brain segments visual input into object and background, shedding light on the neural mechanisms underlying perception and categorization.

# 1 Introduction

Efficient sensory detection requires the capacity to ignore task-irrelevant information. In visual object perception for example, optic flow patterns generated by external objects are more informative than those resulting from self-motion (Gibson, 1950). Consequently, discerning changes in visual input caused by self-movement versus external factors is a crucial challenge for the visual system (see Fig. 1). In computer vision, optic flow finds frequent application, particularly in figure-ground or object segmentation paradigms (see Anthwal and Ganotra, 2019). Indeed, motion has long been recognized to be a powerful Gestalt cue for distinguishing visual objects from the background (Wertheimer, 1923).

In the biological context, optic flow has been suggested to be segmented into self- and externally generated components by sensorimotor mismatch - an error signal arising from the disparity between expectation and bottom-up sensory input (Keller & Mrsic-Flogel, 2018). This conjecture gains support from accumulating evidence for the presence of sensorimotor mismatch (or error) signals in early auditory (Audette & Schneider, 2023) and visual areas (Attinger et al., 2017; Zmarz and Keller, 2016, see Fig. 1A), forming parts of motor correlates found in sensory cortex (Kaplan & Zimmer, 2020; Lohuis et al., 2022).

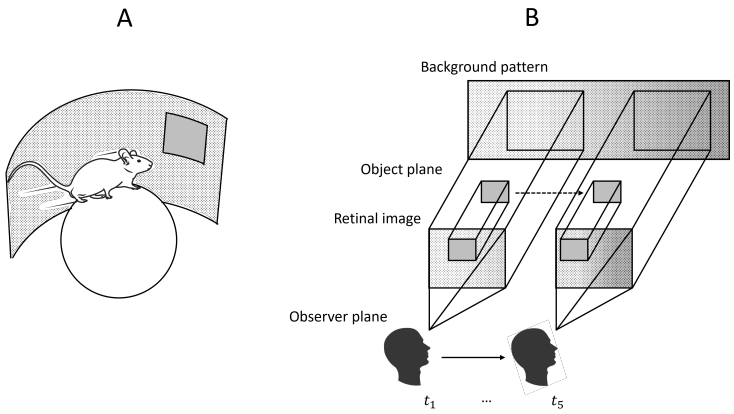


Figure 1: **Two causes of sensed optic flow: self-induced and externally generated.** (A) Experimental setup for mismatch detection as used in Zmarz and Keller, 2016. The VR environment allows control over the visual input of the mouse as it is moving on a spheric treadmill. (B) Configuration of our simulations, mirroring the image construction process in A. The apparent movement of the background pattern is anticipated based on the motor state of the model. Deviations in optic flow arise when an object moves independently between the background and the observer. Time points $t_1$ and $t_5$ depict distinct moments in the simulation.

A recent study on saccadic eye movements by Miura and Scanziani, 2022 demonstrated that direction-selective V1 neurons differentiate between apparent stimulus movement being self- or externally caused. Computationally, these findings suggest the presence of a motor-to-sensory forward model conveying corollary discharges from motor areas to early sensory areas and transforming them into sensory coordinates (see Frith et al., 2000 for an overview). This endows the brain with neural circuits for mismatch (or prediction error)-based segmentation. While models demonstrating sensorimotor mismatch responses exist (Hertäg & Sprekeler, 2020; Mikulasch et al., 2022), it remains unclear if and how these can functionally contribute to segmentation operations. Moreover, given that a significant portion of visual input is generated by external motion, inferring the causes of optic flow becomes a two-fold problem. This raises the question how predictions from the motor-to-sensory forward model and predictions about external objects integrate mechanistically to distinguish between self-caused and externally generated optic flow and interpret them.

Here, we show that the framework of predictive coding elegantly and parsimoniously allows integration of these two generative processes. Predictive coding describes perception as a hierarchical generative process in which higher cortical areas improve their top-down prediction of activity in lower areas (Friston, 2005; Lee & Mumford, 2003; Rao & Ballard, 1999). It forms an important computational building block for understanding how perception – and in particular a motion-corrected, stable world representation - is constructed through learning and inference (R. A. Andersen et al., 1985; Crapse & Sommer, 2008; Creutzig & Sprekeler, 2008; Pennartz, 2015; Whishaw & Brooks, 1999). In pure form, both inference (updating neural activities) and learning of synaptic weights are driven by prediction errors. In contrast to most predictive coding models derived from Rao and Ballard, 1999, however, our model does not predict luminosity, but optic flow. First, a core microcircuit is constructed, with its biological plausibility strengthened by replication of experimentally observed sensorimotor mismatch (Attinger et al., 2017). Second, the model is scaled up and extended by circuits for the representation of externally generated optic flow, leading to segmentation of self- vs. externally generated inputs (Fig. 1B). Lastly, we demonstrate that the higher visual areas of the model, analogous to brain areas MT/MST, become tuned to optic flow patterns caused by external objects, facilitating classification of the perceived object. The novel model proposed here thus extends biologically plausible predictive coding, by functionally integrating motor-to-visual feedback signals in multi-area generative inference of optic flow.

# 2 Methods

## 2.1 A microcircuit for sensorimotor mismatch detection

We developed a microcircuit for sensorimotor mismatch calculation that is shown in Fig. 2A. In inference, optic flow elicits neural activity $y_{vis} \in [0, 1]$ of a direction selective cell in V1 (as originally observed by Hubel and Wiesel, 1959). Simultaneously, the motor area (roughly corresponding to motor cortex or closely connected areas (Guitchounts et al., 2020; Leinweber et al., 2017), here represented by only one unit) codes for the current motor state via neural activity $y_{mot} \in [0, 1]$ of a subpopulation. This motor state is transformed to sensory coordinates via a forward model (Frith et al., 2000).

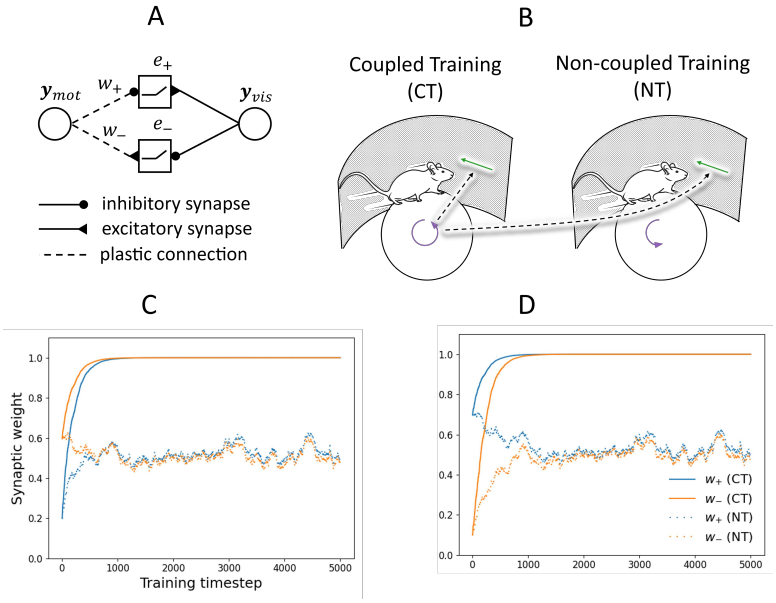


Figure 2: **Proposed microcircuit for mismatch calculation and weight evolution during training.** (A) Proposed microcircuit with error neurons depicted as squares and representation neurons as circles. $w_-$ is the weight of the forward model's excitatory connection to the negative prediction error neuron, $w_+$ the weight of the inhibitory connection to the positive prediction error neuron (see Equation 1. The predicted optic flow is given as the outcome of the forward model, i.e. the synaptic currents arriving at the error neurons from the left. (B) Illustration of coupled and non-coupled training: in both cases, visual inputs are the same, given by the motor movements in the CT condition. In NT, however, motor states and visual inputs are completely uncorrelated. (C) Illustration of robustness to initial conditions proven in section 7.1: Shown is the synaptic weight in coupled (CT) and non-coupled (NT) training (see main text) over training time. (D) Same as (C) for a different weight initialization scheme.

Inversely connected via excitatory and inhibitory synapses, positive and negative error neurons then compare the sensory input to the predictions from the forward model. We chose to model positive and negative error neurons separately, as they are considered to offer a better fit to experimental data (Keller & Mrsic-Flogel, 2018). Their firing rate is given as

$$e_{\pm}(t) = \phi(\pm g(t) \cdot (y_{vis}(t) - w_{\pm}(t-1) \cdot y_{mot}(t))) \tag{1}$$

(note the inverse wiring of positive and negative error neurons illustrated in Figure 2A), with synaptic weights $w_{\pm}$ of the forward model (determined by learning at the previous time step described below or in the initial condition), gating function $g(t)$ that keeps the firing rate at baseline if the model is not sending a movement signal ($y_{mot} = 0$). This gating is putatively mediated by signalling from higher-order thalamus or neuromodulation (Keller & Mrsic-Flogel, 2018):

$$g(t) = \begin{cases} 0 & \text{if } y_{mot} = 0 \\ 1 & \text{otherwise} \end{cases} \tag{2}$$

and activation function

$$\phi(x) = \max(0, x + e_0) \tag{3}$$

with baseline firing rate $e_0 = 0.5$ and the total synaptic input $x$. Supportive evidence for bipartite error computation comes from biologically detailed models that show how these circuits can develop in cortical tissue (Hertäg & Sprekeler, 2020; Mirasso et al., 2023). Learning is then mediated by a Hebbian rule with a switch between long-term potentiation and long-term depression mediated by NMDA receptor activation of the postsynapse (Lüscher & Malenka, 2012; Malenka, 1994):

$$\Delta w_{\pm}(t) = \pm \epsilon_l \cdot y_{mot}(t) \cdot (e_{\pm}(t) - e_0) \tag{4}$$

with learning rate $\epsilon_l = 0.01$. In supplementary section 7.1, a proof of convergence of this learning rule is provided.

For comparison with Attinger et al., 2017, where two cohorts of mice were raised differently, under controlled regimes of sensorimotor experience, we trained two copies of the microcircuit with input contingencies based on their paradigm. Importantly, both visual input and motor state are one-dimensional at this point. In coupled training (CT), the motor state of the animal, sampled from a Bernoulli distribution, is reflected in the optic flow: $y_{vis} = y_{mot} \sim \mathcal{B}(1, 0.5)$. In the non-coupled training paradigm (NT), motor states are the same as in the CT condition, but optic flow is independently sampled from an identical distribution. In both conditions, convergence of the forward model weights was robust to different weight initialization and depended strongly on the training paradigm (CT or NT; Fig. 2B-C).

## 2.2 Integration with hierarchical predictive coding

### 2.2.1 Inference in the two-stream architecture

The presence of mismatch computation in primary visual cortex fits well with theories of generative perception in which the brain interprets visual inputs by identifying the causes that are most likely to have produced them (Dayan et al., 1995; Friston, 2005; Gregory, 1980; Lee & Mumford, 2003; Pennartz, 2015; Rao & Ballard, 1999; von Helmholtz, 1860). We thus integrated the microcircuit developed above into a well-known framework of modeling generative perception, hierarchical predictive coding. Due to its modular structure (Fig. 3A), the model relates more easily to primates than to rodents, although also mouse visual cortex is functionally modularized to some degree (Marshel et al., 2011; Wang et al., 2011).

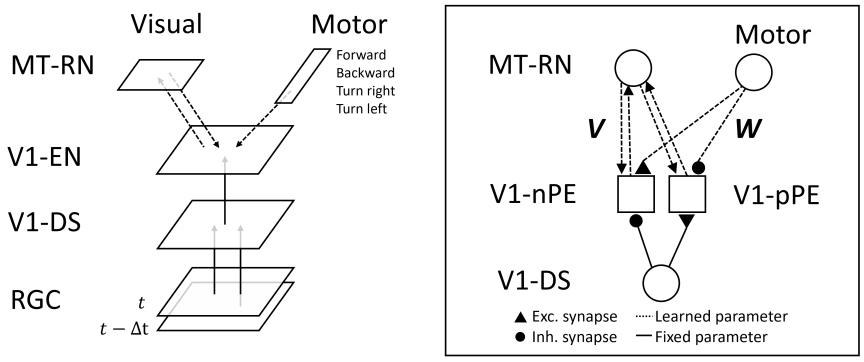


Figure 3: **Integration of motor-to-sensory predictions with predictions from higher visual areas.** Left: Error neurons in visual area V1 receive predictions about optic flow from two streams. After training, the motor-to-visual stream sends predictions about expected optic flow as a result of egomotion. The remaining prediction errors in V1 are fed forward to higher visual area MT. Right: Synaptic wiring of subpopulations with error neurons depicted as squares and representation neurons as circles. Synaptic connections subject to the Hebbian learning rule of Equation 4 are shown as dashed lines: $\boldsymbol{V}$ for the visual-to-visual stream and $\boldsymbol{W}$ for the motor-to-visual stream. Connections between MT and V1 are not *a priori* constrained to be excitatory or inhibitory and are thus denoted as arrows.

Visual inputs to the model are multidimensional (we used a retinotopic field of 40 x 40 or 80 x 80 units depending on the dataset used), covering the visual field. As in most predictive coding models, updating of neural activities (inference) and updating of synaptic strengths (learning) are split into separate phases. During inference, the model receives grayscale video as input. From this, optic flow is extracted through the Farneback algorithm (Farnebäck, 2003), mimicking the functioning of direction selective cells in V1. As a result, four populations of cells encode orthogonal stimulus movement directions (left/right/up/down) leading to

a V1 population of 40 x 40 x 4 = 6,400 neurons (for the modified FashionMNIST dataset) and 80 x 80 x 4 = 25,600 neurons for the Animal dataset respectively). An equal number of positive and negative error neurons then compares velocity of sensory, bottom-up optic flow in the respective direction to top-down predictions, fitting well with reports of retinotopic mismatch calculation (Zmarz & Keller, 2016). These are mediated via two pathways shown in Fig. 3 that jointly attempt to explain the sensed optic flow:

1. The motor-to-sensory forward model consisting of many microcircuits, as developed in section 2.1. The parallel microcircuits form fully connected layers of synaptic connections from a small number $d_{mov}$ ('dimensionality of the movement space') of neurons or subpopulations in the motor area to the nPEs and pPEs in V1. Each of the motor neurons codes for a motor-program (moving forward, turning sideways etc.). We begin with a single motor-program (moving forward) and then later extend to $d_{mov} > 1$ (see Fig. 10). In this pathway, initial weights are drawn from a Gaussian distribution $\mathcal{N}(0.6, 0.6)$.

2. The visual-to-visual stream originating from model area MT (middle temporal) and projecting to lower visual areas here summarized as V1. The name of this model area was based on the properties shared with the eponymous cortical area in monkeys: direction-selectivity (differently weighted inputs from pPE and nPE; (Albright, 1984; Maunsell & Van Essen, 1983b)), large receptive fields, strong feedback connections to V1 (Clavagnier et al., 2004), and its functional role in perceiving structure from motion (R. Andersen et al., 1996; Born & Bradley, 2005; Buračas & Albright, 1997; Duncan et al., 2000; Grunewald et al., 2002; Handa et al., 2008).

The updating of neural activity in model area MT is based on the inference mechanism of Rao and Ballard, 1999: MT representation neurons receive inputs from V1 error neurons through convolution kernels of size 4 x 4 x 4 (stride 1), mimicking receptive fields, and inhibit the error neurons in return (Fişek et al., 2023). We used separate sets of weights connecting to pPEs and nPEs, as well as for the four directions each. Depending on the resolution of the dataset, area MT thus contains 77 x 77 x 4 = 23,716 or 37 x 37 x 4 = 4,356 neurons. Initial weights of the visual-to-visual were sampled from $\mathcal{U}(-\sqrt{k}, \sqrt{k})$ where

$$k = \frac{1}{C_{\text{in}} * \prod_{i=0}^{1} \text{kernel\_size}[i]} \tag{5}$$

with $C_{in}$ being the number of feature maps in the higher area and the index $i$ enumerating the two spatial dimensions of the quadratic kernel. The equation governing updating of an MT neuron's state variable (akin to the membrane potential) in area MT is:

$$\boldsymbol{x}_{MT}(t) = \boldsymbol{x}_{MT}(t-1) + \epsilon_{inf}(\boldsymbol{V}_{+}^{\intercal}(\boldsymbol{e}_{+}(t-1) - e_0) + \boldsymbol{V}_{-}^{\intercal}(\boldsymbol{e}_{-}(t-1) - e_0)) \tag{6}$$

with inference rate $\epsilon_{inf} = 0.05$, visual weights $\boldsymbol{V}_{\pm}$ from MT to V1-EN (cf. Figure 3), and firing rate $\boldsymbol{e}$ of positive and negative error neurons in V1. $\boldsymbol{e_0} = 0.5$ again denotes the baseline firing rate of V1-PE neurons. A ReLU activation function transfers $\boldsymbol{x}_{MT}$ into the output firing rate

$$\boldsymbol{y}_{MT} = \psi(\boldsymbol{x}_{MT}) \tag{7}$$

Error neurons in turn have output firing rates determined by (cf. Eq. 1)

$$\boldsymbol{e}_{\pm} = \phi(\boldsymbol{y}_{vis} - \boldsymbol{W}_{\pm}\boldsymbol{y}_{mot} - \boldsymbol{V}_{\pm}\boldsymbol{y}_{MT}) \tag{8}$$

where $\boldsymbol{y}_{vis}$ is the magnitude of optic flow in the given direction and $\boldsymbol{W}_{\pm}$ denotes the weights of the motor-to-sensory forward model with motor state $\boldsymbol{y}_{mot}$. For the representational readout experiments, we also experimented with the addition of another area on top of MT that is referred to as a model of the medial superior temporal area, MST, based on its hierarchical superordination relative to MT (Born & Bradley, 2005; Felleman & Van Essen, 1991) (for details see section 7.6). After inferring neural activity for multiple time steps (15 proved to be sufficient), the Hebbian learning rule from Eq. 4 was used to update synaptic weights.

### 2.2.2 Training the retinotopic model

Training was conducted in the setting of Fig. 1B, which is comparable to the behavioral paradigm of Attinger et al., 2017; Zmarz and Keller, 2016 (Fig. 1A), but with interpretable objects instead of abstract brick patterns. This paradigm allowed us to study the processing of background and foreground in a controlled manner. However, it also required video inputs with an independently moving object and information about the locomotion state of the observer. As we could not find a dataset of suitable complexity, we constructed three novel datasets in which movement of the background image is controlled by the model's motor state:

a) Ten sequences with different animal images moving in front of a background with a grass texture of 80 x 80 pixels (for details see section 7.2). The one-dimensional self-movement signal drives movement of the background patterns across the retina. Sequences consist of blocks with uniform movement speed reflecting locomotion interleaved with blocks of no movement, with a total sequence length of 50 frames. Concerning object motion, several settings are implemented, among others a stable perception paradigm (such as during parallel movement of object and observer, or during gaze following (Zuberbühler, 2008)) and movement of the object image across the retina, independent of self-motion. The dataset also contains a setting with pure background under observer movement and no objects present.

b) Sequences of retinocentric FashionMNIST objects in 40 x 40 resolution with otherwise the same properties as in a), although the FashionMNIST objects are only used in the stable perception paradigm where objects remain retinocentric. Since our modification of this datasets contains a

significantly larger number of samples (we use 5000) and is organized in ten object classes, it is suitable to test how useful the neural representations are in a given model area for downstream classification. Across the ten object classes, we use 400 samples for training and 100 different samples per class for testing. Here, only the stable perception paradigm is implemented (objects appear static on the retina).

c) Six different optic flow patterns including non-homogeneous expanding and contracting flow fields without object present. Each flow pattern is associated with one of six dimensions of the motor state encoded in one-hot manner (turning leftwards, moving forward, etc.).

Unless noted differently, the two streams (MT to V1 and motor cortex to V1) were trained in the following manner: During pretraining, only the motor-to-visual stream was activated and trained on visual inputs coupled to the model's motor state. In this phase, no external objects were presented. Then, in the training phase, objects were introduced (with the observer moving along them unless where noted differently) and the visual stream was activated. Training proceeded until prediction errors were sufficiently low. To separately analyze learning in the two streams, the weights in the motor-to-visual stream were frozen during training, but see section 7.4 for experiments with joint training.

## 2.3 Segmentation experiments

### 2.3.1 Obtaining a segmentation mask from the model

Segmentation masks highlighting external objects were obtained from MT→V1-EN feedback, based on reports of figure-ground segmentation in V1 through feedback from higher visual areas (Self et al., 2013). Excitatory top-down signals from area MT to V1-EN were obtained by matrix multiplication of the top-down synaptic weights $\boldsymbol{V}_{\pm}$ from Equation 6 and MT activity $y_{MT}$). The summed up top-down signals were then thresholded at each point in retinotopic space to yield a segmentation mask: Where signals are larger than

$$C_{thresh} = c_{thresh} \cdot \max\left(\boldsymbol{V}_{+}\boldsymbol{y}_{MT} + \boldsymbol{V}_{-}\boldsymbol{y}_{MT}\right) \tag{9}$$

an external object is assumed, with a heuristically chosen scaling constant $c_{thresh}$. Physiologically, this thresholding may be achieved via a neuron with appropriately set firing threshold via baseline inputs.

### 2.3.2 Quantifying segmentation accuracy and comparison to baseline

We quantified segmentation accuracy with an Intersection over Union (IoU) measure. IoU is given as the ratio of overlap between predicted and true object area relative to the union and thus ranges between zero and one. Measured at an arbitrarily selected timepoint in the sequence (we chose nine frames into it unless where noted differently), the only important constraint here was object movement relative to the background.

As a strong baseline, we computed a binary segmentation mask $\boldsymbol{S}_{baseline}$ by thresholding optic flow signals $\boldsymbol{y}_{vis}$:

$$\boldsymbol{S}_{baseline} = \Theta(\pm(\boldsymbol{y}_{vis} - \boldsymbol{\theta})) \tag{10}$$

with the Heaviside function $\Theta$: $\mathbb{R} \to \{0, 1\}$ and threshold $\boldsymbol{\theta}$. The optimal baseline model was then identified by maximizing its IoU score through a grid search of $\boldsymbol{\theta}$.

### 2.4 Linear readout to analyze representational content

Simulating a class-dependent downstream task such as class-specific approach-or-flight behavior, we investigated coding of population activity for object category. Neuronal representations in model areas MT, MST, V1 and retina were inferred for training (400 stimuli from each of the ten classes) and test images (100 stimuli per class) from the modified FashionMNIST dataset. Then, a linear classifier (one per each area) was trained to map the activity patterns to class labels using a cross-entropy loss.

## 3 Results

### 3.1 Reproduction of sensorimotor mismatch effects

Before analyzing how well the complete hierarchical network performs computationally, we first examined the biological feasibility of the underlying microcircuit (section 2.1). This involved investigating whether and under what conditions sensorimotor mismatch responses, as discussed in the Introduction, actually appear. Recall that networks were trained in coupled and non-coupled conditions akin to the mice in Attinger et al., 2017. After training, we exposed the networks to the same testing conditions. As in the experimental study, we termed them *mismatch* (continuously moving model, sudden halt of optic flow) and *playback halt* (passive observation of constant optic flow that is suddenly halted). Across both conditions, the error neurons in the model reproduced observed firing rate patterns from mouse V1. As shown in Fig. 4A, a strong increase in neural activity was observed during mismatch after CT training, but not after NT training. No increase to playback halt is observed in either condition, in line with the recordings from Attinger et al. shown for comparison in 4B. The model also reproduced recovery of mismatch responses observed by Attinger et al: Subsequently training the model originally trained in the NT paradigm by exposing it to the CT conditions installed an increase in neural activity under the mismatch condition 4C and 4D). One source of difference between the simulations and experimental results is the simple nature of the neuron model. Since no receptor dynamics are modeled and the experimental data is a low-pass filtered $Ca^{2+}$-response, the rise of activity is faster in the model. Without an explicit model of somatic $Ca^{2+}$, the activity decay is instantaneous in the model, whereas the recordings in Figure 4B, D show a slow signal decay characteristic for calcium imaging.

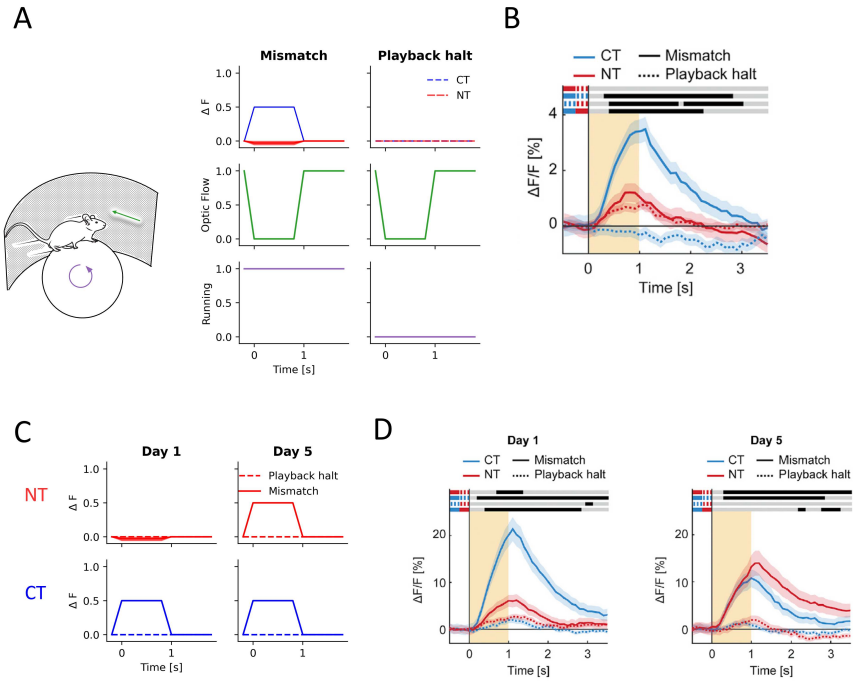


Figure 4: **The proposed microcircuit reproduces experimentally observed mismatch responses.** (A) Left: VR setup with optic flow stimuli (green) independently controllable of motor state (purple). Right: Response of model V1 error neurons in the model to mismatch and playback halt conditions defined by sudden halt of optic flow from $t = 0s$ to $t = 1s$ while running (mismatch) and standing still (playback halt). After experiencing coupled training (CT), i.e. contingent sensorimotor experience, a mismatch response was observed. This was not the case after non-coupled training (NT) with visual feedback uncorrelated to locomotion. Error bands (partially invisible) indicate standard deviation across five random seeds influencing the sampling of the training sequences. (B) Calcium imaging responses to the two conditions in mice V1. Delta F/F (in somatic $Ca^{2+}$) on the vertical axis is a reflection of neural firing rate. The orange area indicates duration of visual halt, shading indicates SEM. (C) Responses of model error neurons after training in CT and NT conditions (left, "day 1") and after the model has been subsequently trained in the coupled paradigm (right, "day 5"). Note the appearance of the mismatch response in the red continuous curve. (D) Observed neural responses in mice initially trained in the CT and NT conditions, before and after subsequent coupled training. Panels B and D were reproduced from Attinger et al., 2017, published in Cell, Copyright Elsevier.

### 3.2 Area MT contributes to reducing prediction errors

After positive evaluation of the microcircuit, we tested its capacity to reduce prediction errors when integrated into a hierarchical predictive coding model of visual processing (see section 2.2). In the pretraining phase illustrated in the left panel of Fig. 5A, we restricted predictions about optic flow to top-down feedback from the motor area to V1-EN. As shown in Fig. 5B, coupled training minimized prediction errors in this phase much more efficiently than non-coupled training. Here, we used the Animal dataset described in section 2.2.2. In the non-coupled paradigm, mean squared error (MSE) as measured by input to error neurons amounted to 3.19 ± 0.00 (a.u.) after this phase, compared to 0.74 ± 0.00 in the coupled condition. This correlates well with the experimentally observed necessity of coupled training to truthfully detect sensorimotor mismatch (cf. Fig. 4B). While in the Animal dataset used here the lateral movement of the background image created homogeneous optic flow patterns, we also tested learning of multiple non-homogeneous flow patterns. As shown in Suppl. Fig. 11, the parallel microcircuits from motor neurons to V1-EN (Suppl. Fig. 10) were able to accurately learn the corresponding forward models.

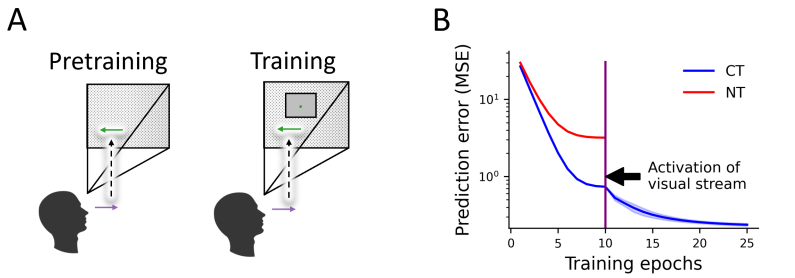


Figure 5: **Learning to predict external and internal causes of optic flow.** (A) Pretraining setup: the model is trained to predict the consequences of its own movements in an empty environment. Training setup: objects are introduced, that cast a static image onto the retina due to the observer moving along them, while the background patterns move across the retina. As in previous figures, the colored vectors refer to absolute/world-centric movement (purple) and pattern movement relative to the observer's retina (green). (B) Pretraining (epoch 1-10) reduces prediction errors most effectively if conducted in the coupled paradigm (CT) as opposed to the non-coupled paradigm (NT) both introduced in Fig. 2B and in the main text. Activation and training of the second visual stream from MT to V1 after epoch ten further reduces prediction error. Error bands (too small to be visible in epoch 1-10) indicate one standard deviation computed across four runs with randomly initialized weights.

After training the motor-to-visual stream, we then tested in the training phase whether addition of the second visual stream (MT to V1) provided a reduction in prediction errors quantified by activity in V1 prediction error neurons. To do so, the visual-to-visual stream of the model was activated and the weights of the pretrained motor-to-visual stream were frozen, putatively corresponding to a halt of critical period plasticity. In Fig. 5B, this moment is indicated by the vertical line. Training then proceeded until convergence of the error signal which took approximately 15 epochs (each epoch consisting of a single iteration through the ten sequences of the dataset). This further reduced the MSE from $0.74 \pm 0.00$ to $0.24 \pm 0.01$. Qualitatively, the effect of feedback from area MT to V1-EN can be understood by the following scenario: With the motor-to-visual stream trained on background-only inputs (i.e. the pretraining paradigm described in 2.2.2), and before activation of the second visual stream, an external object moving independently from the background elicited an increase in nPE activity (Fig. 6A, second column). After activating and training the visual stream, predictions from model area MT successfully minimized prediction errors (Fig. 6A, third column), i.e. optic flow including the moving animal was better predicted. This can be explained by representation of the object-generated optic flow patterns in model area MT that is then fed back as top-down prediction to V1 where the full bottom-up optic flow is represented, cancelling the object-generated prediction errors there (cf. Eq 8). While separate training of both streams proved to be the most effective strategy, we demonstrated that in principle also both streams can be trained simultaneously (Supplementary Material 7.4).

### 3.3 Visual predictions segment external causes

If, as hypothesized above, model area MT indeed represents external causes of optic flow that cannot be explained by self-motion, it should be possible to use its activity to segment moving objects from the background. We obtained segmentation masks from the model as described in section 2.3. The model accurately captured the outline of external objects as shown in Fig. 6B.

We evaluated segmentation performance of the model on objects moving relative to both the observer and background. To do so, the Animal dataset was modified: In addition to background motion correlating with the movement state of the agent, the external object now moved at an independent speed in the opposite direction (Fig. 6C). We found that the model was capable of shifting the segmentation mask across retinal space as indicated by the high IoU scores (introduced in section 2.3) in Fig. 6D. Here we selected appropriate timepoints for evaluating readout accuracy to ensure that the object was still in the model's field of view: For speeds [0, 1, 2, 4] these were [9, 9, 6, 4] frames into the sequence. Across relative movement speeds, model performance was on par with the baseline derived by optimal segmentation of the instantaneous optic flow (described in more detail in section 2.3), and both mildly decreased as expected with faster movement speed.

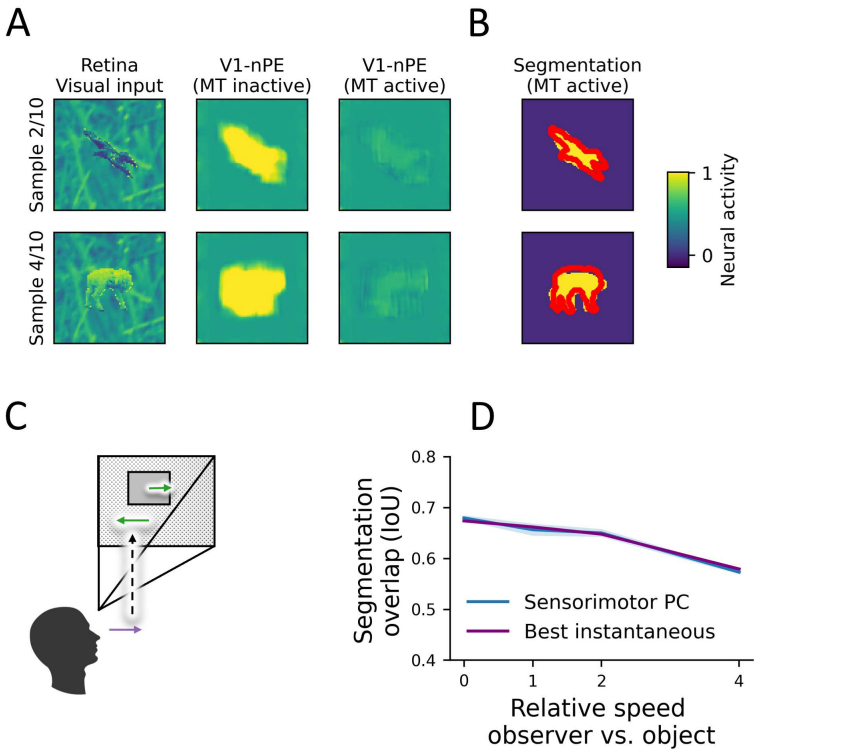


Figure 6: **Segmentation of external causes through visual feedback** (A) Same setting as in Figure 5. First column: Retinal inputs showing an eagle and a lamb against a similarly colored background. Second column: A model with the motor-to-visual stream trained in the coupled paradigm correctly minimizes prediction errors in the background, but the external object moving slower than expected elicits activity in nPEs. pPEs are not shown for simplicity but show the inverse pattern. Third column: after activating and training feedback from MT to V1, prediction errors are further minimized. (B) Outlining the fully trained model's guess about the spatial extent of the external object, the segmentation mask obtained from top-down signals from model area MT to V1 is shown in yellow over the ground truth outlined in red. (C) Non-zero relative speed of the object relative to the observer. In contrast to the stable perception paradigm, the object image moves across the retina (upper green arrow). (D) Segmentation performance under varying relative speeds between object and observer during observer movement. Prior to evaluation, the two-stream model was trained in the paradigm of C on the Animal dataset. The error bands indicate one standard deviation, calculated across four randomly initialized runs. The baseline in purple is the optimum at the time step of evaluation (see section 2.3).

### 3.4 Higher visual areas encode object identity

For the visual system to guide behavior, it does not suffice to know where an externally moving object is located (shown by the correct placement of the segmentation masks); also its identity needs to be inferred. Although classically, such core object recognition is attributed to inferotemporal cortex in non-human primates, it would also be expected (to a lesser degree) in area MT based on the neuronal responses of macaque MT to structure-from-motion displays (R. Andersen et al., 1996; Grunewald et al., 2002). Furthermore, evaluating the information content of model area MT may allow to better understand its computational role.

To read out representational content from the model, we used the modified FashionMNIST dataset described in section 2.2.2. First, both generative streams were trained according to the procedure described in section 3.2. As before, the binary locomotion state was coupled to movement of the background pattern across the retina, while the object's image remained stable in the center. In this phase, use of 100 stimuli (10 per class) proved sufficient. Then linear decoding was used to estimate readout accuracy of object class as described in section 2.4. The resulting classification accuracy of 81.6% on unseen (test) images in area MT and 82.3% in MST shows class-specificity (Tab. 1) and decent generalization to unseen data. Succeeding rejection of the null hypothesis ('all populations have the same readout accuracy') using a Welch's ANOVA, we performed a Games-Howell post-hoc test to determine the significance of the pairwise differences between the populations (for a detailed description of the statistical methods see section 7.7).

| Area | Acc. train | Acc. test |
|---|---|---|
| MT | 94.0 ± 0.2 | 81.6 ± 0.7 |
| MST | 93.3 ± 0.7 | 82.3 ± 0.5 |
| V1-DS | 88.8 ± 0.3 | 79.6 ± 0.4 |
| Full vis. | 92.7 ± 0.2 | 76.8 ± 0.1 |
| Chance | 10.00 | 10.00 |

Table 1: **Model area MT allows readout of object class.** Percentage of correctly classified stimuli evaluated on the train/test split of the ten classes of the modified FashionMNIST dataset as described in the main text. Rows denote the cell population: MT, MST, V1-Direction Selective (DS) cells, full retinal input as a baseline, and chance level. The standard deviation is calculated across four randomly initialized runs.

Showing an increase in decodability compared to lower areas, readout accuracy on test data was slightly but significantly higher in area MT (81.6%) than in V1-DS cells (79.6%) with $p < 0.03$, and also significantly higher than the baseline obtained from directly reading out the raw retinal activity patterns (76.8%, $p <$ 2.5e-3). The difference between areas MST and MT was not significant.

Lastly, in all areas, the drop between readout accuracy on training and test data suggested overfitting of the readout classifier. We conclude that learning optic flow patterns drives learning of behaviorally useful and generalized representations and emphasize that more class-specific representations are likely to be acquired via additional learning mechanisms in the ventral stream. Due to the invariance of optic-flow to changes in luminance, the approach employed here can be expected to be more robust to changes in lighting condition than approaches that rely on texture, i.e. patterns in static luminosity.

### 3.5 Unexpected events elicit elevated neuronal responses

To further examine the general conjecture that model area MT codes for optic flow caused by externally moving objects, we recorded the neural dynamics of the network in two conditions. In a shape-from-motion paradigm illustrated in Fig. 7A, the external object from the Animal dataset began to move in front of the background, followed by movement of the observer so that the object image appeared statically on the retina. In the model trained in the coupled-training paradigm, this movement onset elicited an activity increase in MT neurons with receptive fields on the object, but not for those with receptive fields on the background (Fig. 7C, bottom row). Such coding for world-centric movement was indeed found in area MT (Erickson and Thier, 1991, see also section 4.2 of the Discussion). In V1-PE neurons with receptive fields on the object, but not outside, a strong transient response was observed, reminiscent of increased activity to unexpected events in oddball paradigm experiments (Squires et al., 1975). Due to the suppressive effect of area MT on V1-PEs (see 5), the rapid decay in activity was most likely caused by top-down feedback. We note that the "stable perception" paradigm in which the observer maintains the object image statically on the retina was chosen for consistency with the previously described simulations and was not necessary to elicit the described neural responses (object movement sufficed).

To confirm that the response in area MT was truly linked to object presence, we recorded the model's response to unexpected object disappearance (Fig. 7B). This setup is reminiscent of temporally unexpected stimuli in the oddball paradigm (Squires et al., 1975). The network with two streams trained (CT) on the Animal dataset was first moving to maintain an image from the Animal dataset statically on the retina. After a fixed amount of time steps, the image was suddenly removed (illustrated in Fig. 7B). Following object removal, a non-instantaneous decrease of MT activity was observed for neurons with receptive fields on the object. Furthermore, we again observed a spike of activity in V1-PE neurons with receptive fields on the object (Fig. 7D), fitting well with the response increase in oddball paradigm experiments.

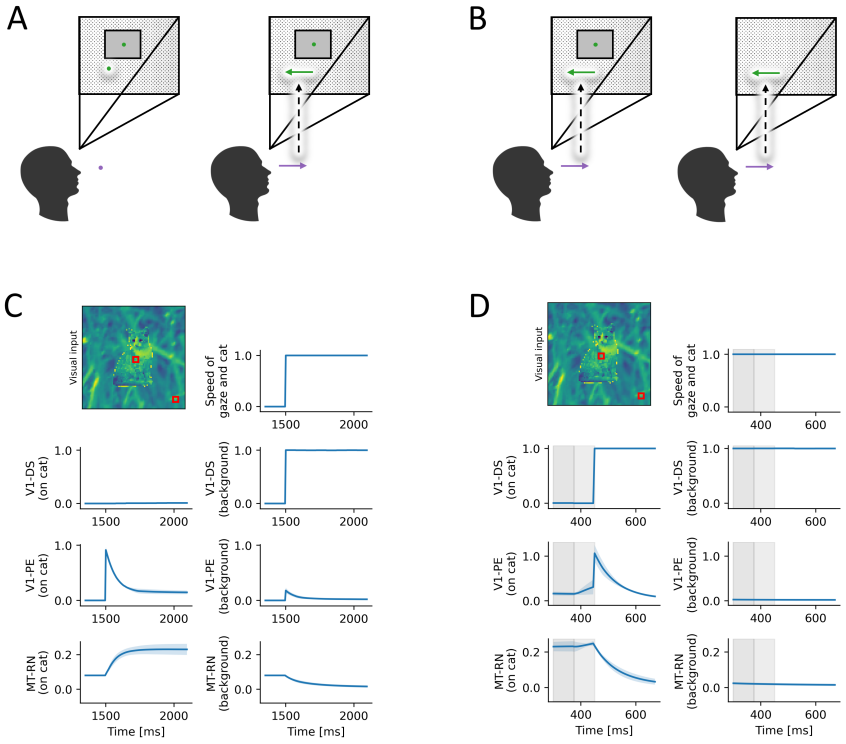


Figure 7: **Neuronal dynamics during movement onset and object removal.** (A) Illustration of movement onset. Green arrows indicate movement of patterns across the retina, the purple arrow movement of the observer. Left: the starting point is static, symbolized by points in green (relative to the retina) and purple (world-centric). Right: both the object and the observer begin to move, so that the object appears static on the retina. (B) Object removal: after the observer moves synchronously with it (left), the object suddenly disappears (right). (C) Neuronal activity during movement onset as illustrated in A. Responses were recorded across two sites indicated by red squares. The light blue band (invisible) indicates one standard deviation calculated across four randomly initialized runs. (D) Neuronal activity during the object removal outlined in B. At the end of the area shaded in grey, the cat image was suddenly removed, where the lightly shaded area indicates a transition time of one frame during which optic flow estimation is disturbed until a reliable estimate could again be made. Representations in area MT take a number of time steps to decay (bottom row). This difference between prediction and presence results in a 'spike' of error signals in V1 (third row).

In this setting, responses can thus be explained by model area MT representing the distortions in optic flow caused by the external object. Sudden removal then creates a divergence between the still present prediction of optic flow-distortion by a moving object and the sensed optic flow that is now completely explained by egomotion. Consequently, MT reduces its activity, reflecting the novel absence of external optic-flow distortions.

# 4 Discussion

## 4.1 Summary of results

We developed a scalable microcircuit with an analytically tractable learning rule to learn the visual consequences of egomotion (Fig. 2). The microcircuit replicated sensorimotor mismatch signals observed in V1 of mice under various conditions and in an experience-dependent manner (Fig. 4). Subsequently, we integrated the microcircuit into a novel model designed for inferring the causes of optic flow (Fig. 3). To our knowledge, this joint generative model is the first to functionally incorporate motor-to-sensory microcircuits into a biological model of visual inference, conforming to hierarchical predictive coding. This extended model successfully learned to predict self-generated optic flow patterns across the visual field (Fig. 5B), and was capable of learning multiple non-homogeneous optic flow patterns (Supplementary Fig. 11). The benefit of the added visual-to-visual stream was firstly demonstrated by its capacity to reduce prediction errors in V1 more effectively than a pure motor-to-visual model when external objects were introduced (Fig. 5A). We then showed that this efficiency in error-minimization was due to top-down signalling of a segmentation mask by model area MT, highlighting the external object. Lastly, we confirmed that model area MT not only learned the 2D shape of the object, but that population activity of MT neurons also contains enough information for accurate readout of object class.

## 4.2 Real motion coding in the middle temporal cortex

The functional role assigned to model area MT in terms of segmentation of object from background aligns well with the interpretation of Born and Bradley, 2005 who after a review of mostly passive recordings arrives at the conclusion that "[a]ll together the evidence rather strongly suggests that MT neurons are critically involved in segmenting an image into separately moving parts". Our model postulates that this is achieved through neurons tuned to world-centric as opposed to retinocentric movement. Such *real motion cells* were indeed found in multiple brain areas (Galletti & Fattori, 2003; Nau et al., 2018; Sasaki et al., 2020) including MT (Erickson & Thier, 1991). Coding for real motion in human MT is supported by more recent experiments by Pitzalis et al., 2020 and, with independent object and self-motion condition, Sulpizio et al. (manuscript in preparation), although in both studies subjects do not actively move, but

infer their movement state from background motion. While the flow-parsing theory of Warren and Rushton, 2009 explains such real motion coding in MT with a subtraction operation *within* MT (bottom-up received optic flow minus self-generated optic flow), our model allocates only the outcome to MT and the subtraction operation to error neurons in V1. At least in mice, reports of sensorimotor mismatch in V1 indeed speak for the latter. In primates, little is known about sensorimotor mismatch computation. Translating the VR paradigm of Attinger et al., 2017; Zmarz and Keller, 2016 to primates or constructing a VR mismatch paradigm for humans in an fMRI scanner akin to (Di Marco et al., 2021) could help to elucidate where sensorimotor mismatch is computed. Lastly, our model also provides a novel perspective on the large number of feedback connections from MT to V1 (Maunsell & Van Essen, 1983a) as part of a larger generative model employed in the interpretation of sensed optic flow.

### 4.3 The neural circuitry of predictive coding

While the predictive coding framework offers a compelling account of inference and learning in perception, its empirical predictions are still under scrutiny (Green et al., 2023; Leinweber et al., 2017; Pennartz et al., 2019; Walsh et al., 2020). With regard to sensorimotor mismatch, the question is to what extent neuronal mismatch signals can be explained purely by locomotion-induced gain as argued by Muzzu and Saleem, 2021, 2023. A recent case in favor of true generative motor-to-visual feedback was made by Vasilevskaya et al., 2023. Together with the alignment with experimental results, use of a biologically plausible learning rule and its strong computational benefits (object representation learning without external supervision, optic flow-based segmentation), the predictive coding account appears compelling.

Another point of debate is the necessity of dedicated error neurons. Although recent evidence suggests prediction error-coding by SST interneurons in the posterior parietal cortex of mice (Green et al., 2023), dedicated error neurons are not required per se for prediction-based inference and learning. Indeed, Mikulasch et al., 2023 developed a model of hierarchical (unimodal) predictive coding with error computation in basal dendrites (see also Urbanczik and Senn, 2014). Due to the algorithmic similarity of how bottom-up sensory inputs are compared to top-down predictions in their model compared to ours (since both are derived from Rao and Ballard, 1999), many aspects proposed here, such as the learning rule, functional modularization and its mapping to brain areas are translatable to an implementation with dendritic error computation.
It should be mentioned that sensorimotor mismatch is not a purely cortical phenomenon and has also been observed in the cerebellum (Hull, 2020, see also Lisberger, 1988). There, however, mismatch responses are mostly linked to learning of sensorimotor transformations and motor control (Albus, 1971; Hull, 2020; Ito, 1970; Lisberger, 1988; Marr, 1969; Stone & Lisberger, 1986) without object segmentation, whereas the circuits modeled here underlie a perceptual function: the parsing of visual inputs in a dynamic world. It thus appears that

the brain developed the same principle of comparing the predicted outcome of movement to the observed state at least twice for different purposes.

## 4.4 Related work

Based on psychophysical evidence of object movement perception during self-motion, Royden and Holloway, 2014 constructed a model to identify the outlines of moving objects. This was achieved through comparison of bottom-up optic flow patterns to a memory bank of optic flow templates. In contrast to our model, however, no neural mechanisms for the learning of the templates, nor for the wiring of the comparison operation were implemented.

Within the framework of predictive coding, most models typically operate purely in the visual domain, whether static (Rao & Ballard, 1999), or dynamic (Brucklacher et al., 2023; Lotter et al., 2020). Extensions of predictive coding that take into account the multimodality of sensory processing in the brain were proposed by Keller and Mrsic-Flogel, 2018; Pennartz, 2015, 2022 in theory and implemented e.g. by Pearson et al., 2021. Another type of multimodal generative model is learning of motor-to-visual forward models based on corollary discharges, wherein motor commands are taken to constitute a non-sensory modality. On a neural level, such models have been proposed by Hertäg and Sprekeler, 2020; Mikulasch et al., 2022, in contrast to our model without the ability to account for externally-caused optic flow. Interestingly, Hertäg and Sprekeler, 2020 demonstrated that signed prediction error responses emerge under visual feedback contingent with motor state in a biologically detailed model, and investigate the role of the involved interneurons. The microcircuit developed in section 2.1 can thus be seen as a useful abstraction over the interneuron-level circuits developed by Hertäg et al., lending it more easily to functionally powerful computations and the abovementioned integration with a purely visual stream. In contrast to Mikulasch et al., 2022 and Hertäg and Sprekeler, 2020, the used datasets go beyond one-dimensional visual inputs and the model is capable of solving visual tasks such as figure-ground segmentation and, to a limited degree, classification. To our knowledge, the model put forward here is the first to integrate motor-to-sensory forward models and sensory-sensory predictive coding within a functional model of generative visual perception.

## 4.5 Limitations in performance

Segmentation performance (Fig. 6B and D) is naturally dependent on the accuracy of the underlying optic flow-extraction through V1 direction-selective cells. Here, accuracy can be assumed to correlate positively with the resolution of the input signal. Based on the relatively low resolution used due to limitations in computational resources (40 x 40, 80 x 80 pixels per frame), we expect significant improvements when using input sequences of higher resolution (and thus wider networks). Another factor that would significantly improve segmentation is the use of depth information through stereovision or motion parallax. Since

area MT displays strong tuning to binocular disparity and thus depth (Born & Bradley, 2005; Cumming & DeAngelis, 2001), stereoscopic integration would fit well with the functional neuroanatomy. Thus, an interesting extension of the current model would be a joint generative model of optic flow and depth for 3D segmentation and recognition from binocular inputs. Depth is also an informative cue for segmentation of partially overlapping objects moving at the same speed - whereas the current model depends on distinct speeds. Lastly, it should be stressed that self-generated optic flow is by no means irrelevant (Gibson, 1950), but provides an important reafference signal to inform motor execution. Interestingly, as discussed in section 4.3, such closed-loop control also relies on sensorimotor prediction errors (Albus, 1971).

### 4.6 Conclusion

In summary, this study presents a novel computational model that seamlessly integrates motor-to-sensory microcircuits into a hierarchical predictive coding framework for visual perception. Notably, the model's use of generative feedback for external object segmentation provides a novel angle to the ongoing discourse regarding the functional importance of top-down connections to early sensory areas in neural networks.

## 5 Acknowledgments

We thank Sander Bohte, Valentina Sulpizio and Thede Witschel for helpful discussions.
This project has received funding from the European Union's Horizon 2020 Framework Programme for Research and Innovation under the Specific Grant Agreement No. 945539 (Human Brain Project SGA3 to G.P., F.M. and C.P.).
This project has received funding from the European Research Council under the Grant Agreement No. 820213 (ThinkAhead), the PNRR MUR projects PE0000013-FAIR, and IR0000011–EBRAINS-Italy to G.P.
A research stay connected to this project was funded for M.B. through the NENS Exchange Grant program of the Federation of European Neuroscience Societies (FENS).

## 6 Data availability

The Python code to reproduce the results of this paper can be found at: https://github.com/matthias-brucklacher/LearningMotorFeedback

# References

Albright, T. D. (1984). Direction and orientation selectivity of neurons in visual area mt of the macaque. *Journal of neurophysiology*, *52*(6), 1106–1130.

Albus, J. S. (1971). A theory of cerebellar function. *Mathematical biosciences*, *10*(1-2), 25–61.

Andersen, R., Bradley, D., & Shenoy, K. (1996). Neural mechanisms for heading and structure-from-motion perception. *Cold Spring Harbor symposia on quantitative biology*, *61*, 15–25.

Andersen, R. A., Essick, G. K., & Siegel, R. M. (1985). Encoding of spatial location by posterior parietal neurons. *Science*, *230*(4724), 456–458.

Anthwal, S., & Ganotra, D. (2019). An overview of optical flow-based approaches for motion segmentation. *The Imaging Science Journal*, *67*(5), 284–294.

Attinger, A., Wang, B., & Keller, G. B. (2017). Visuomotor coupling shapes the functional development of mouse visual cortex. *Cell*, *169*(7), 1291–1302.

Audette, N. J., & Schneider, D. M. (2023). Stimulus-specific prediction error neurons in mouse auditory cortex. *Journal of Neuroscience*. https://doi.org/10.1523/JNEUROSCI.0512-23.2023

Baldwin, M. K., Kaskan, P. M., Zhang, B., Chino, Y. M., & Kaas, J. H. (2012). Cortical and subcortical connections of v1 and v2 in early postnatal macaque monkeys. *Journal of Comparative Neurology*, *520*(3), 544–569.

Born, R. T., & Bradley, D. C. (2005). Structure and function of visual area mt. *Annu. Rev. Neurosci.*, *28*, 157–189.

Brucklacher, M., Bohte, S. M., Mejias, J. F., & Pennartz, C. M. (2023). Local minimization of prediction errors drives learning of invariant object representations in a generative network model of visual perception. *Frontiers in Computational Neuroscience*, *17*.

Buračas, G., & Albright, T. (1997). Contribution of area mt to perception of three-dimensional shape: A computational study. *Ophthalmic Literature*, *1*(50), 46.

Clavagnier, S., Falchier, A., & Kennedy, H. (2004). Long-distance feedback projections to area v1: Implications for multisensory integration, spatial awareness, and visual consciousness. *Cognitive, Affective, & Behavioral Neuroscience*, *4*, 117–126.

Crapse, T. B., & Sommer, M. A. (2008). Corollary discharge across the animal kingdom. *Nature Reviews Neuroscience*, *9*(8), 587–600.

Creutzig, F., & Sprekeler, H. (2008). Predictive coding and the slowness principle: An information-theoretic approach. *Neural Computation*, *20*(4), 1026–1041.

Cumming, B. G., & DeAngelis, G. C. (2001). The physiology of stereopsis. *Annual review of neuroscience*, *24*(1), 203–238.

Dayan, P., Hinton, G. E., Neal, R. M., & Zemel, R. S. (1995). The helmholtz machine. *Neural computation*, *7*(5), 889–904.

Di Marco, S., Sulpizio, V., Bellagamba, M., Fattori, P., Galati, G., Galletti, C., Lappe, M., Maltempo, T., & Pitzalis, S. (2021). Multisensory inte-

gration in cortical regions responding to locomotion-related visual and somatomotor signals. *Neuroimage*, *244*, 118581.

Dora, S., Bohte, S. M., & Pennartz, C. M. (2021). Deep gated hebbian predictive coding accounts for emergence of complex neural response properties along the visual cortical hierarchy. *Frontiers in Computational Neuroscience*, *15*, 666131.

Duncan, R. O., Albright, T. D., & Stoner, G. R. (2000). Occlusion and the interpretation of visual motion: Perceptual and neuronal effects of context. *Journal of Neuroscience*, *20*(15), 5885–5897.

Erickson, R., & Thier, P. (1991). A neuronal correlate of spatial stability during periods of self-induced visual motion. *Experimental brain research*, *86*, 608–616.

Farnebäck, G. (2003). Two-frame motion estimation based on polynomial expansion. *Image Analysis: 13th Scandinavian Conference, SCIA 2003 Halmstad, Sweden, June 29–July 2, 2003 Proceedings 13*, 363–370.

Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex (New York, NY: 1991)*, *1*(1), 1–47.

Fişek, M., Herrmann, D., Egea-Weiss, A., Cloves, M., Bauer, L., Lee, T.-Y., Russell, L. E., & Häusser, M. (2023). Cortico-cortical feedback engages active dendrites in visual cortex. *Nature*, 1–8.

Friston, K. (2005). A theory of cortical responses. *Philosophical transactions of the Royal Society B: Biological sciences*, *360*(1456), 815–836.

Frith, C. D., Blakemore, S.-J., & Wolpert, D. M. (2000). Abnormalities in the awareness and control of action. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *355*(1404), 1771–1788.

Galletti, C., & Fattori, P. (2003). Neuronal mechanisms for detection of motion in the field of view. *Neuropsychologia*, *41*(13), 1717–1727.

Gibson, J. J. (1950). The perception of the visual world.

Green, J., Bruno, C. A., Traunmüller, L., Ding, J., Hrvatin, S., Wilson, D. E., Khodadad, T., Samuels, J., Greenberg, M. E., & Harvey, C. D. (2023). A cell-type-specific error-correction signal in the posterior parietal cortex. *Nature*, 1–8.

Gregory, R. L. (1980). Perceptions as hypotheses. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, *290*(1038), 181–197.

Grunewald, A., Bradley, D. C., & Andersen, R. A. (2002). Neural correlates of structure-from-motion perception in macaque v1 and mt. *Journal of Neuroscience*, *22*(14), 6195–6207.

Guitchounts, G., Masis, J., Wolff, S. B., & Cox, D. (2020). Encoding of 3d head orienting movements in the primary visual cortex. *Neuron*, *108*(3), 512–525.

Handa, T., Katai, S., Kuno, R., Unno, S., Inoue, M., & Mikami, A. (2008). Differential activity to shapes under shape-from-motion condition in macaque middle temporal area. *Neuroscience*, *156*(4), 1118–1135.

Hertäg, L., & Sprekeler, H. (2020). Learning prediction error neurons in a canonical interneuron circuit. *Elife*, *9*, e57541.

Hubel, D. H., & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *The Journal of physiology*, *148*(3), 574.

Hull, C. (2020). Prediction signals in the cerebellum: Beyond supervised motor learning. *elife*, *9*, e54073.

Ito, M. (1970). Neurophysiological aspects of the cerebellar motor control system. *International journal of neurology*, *7*, 162–176.

Kaplan, H. S., & Zimmer, M. (2020). Brain-wide representations of ongoing behavior: A universal principle? *Current opinion in neurobiology*, *64*, 60–69.

Keller, G. B., & Mrsic-Flogel, T. D. (2018). Predictive processing: A canonical cortical computation. *Neuron*, *100*(2), 424–435.

Lee, T. S., & Mumford, D. (2003). Hierarchical bayesian inference in the visual cortex. *JOSA A*, *20*(7), 1434–1448.

Leinweber, M., Ward, D. R., Sobczak, J. M., Attinger, A., & Keller, G. B. (2017). A sensorimotor circuit in mouse cortex for visual flow predictions. *Neuron*, *95*(6), 1420–1432.

Lisberger, S. G. (1988). The neural basis for learning of simple motor skills. *Science*, *242*(4879), 728–735.

Lohuis, M. N. O., Marchesi, P., Olcese, U., & Pennartz, C. (2022). Triple dissociation of visual, auditory and motor processing in primary visual cortex. *bioRxiv*, 2022–06.

Lotter, W., Kreiman, G., & Cox, D. (2020). A neural network trained for prediction mimics diverse features of biological neurons and perception. *Nature machine intelligence*, *2*(4), 210–219.

Lüscher, C., & Malenka, R. C. (2012). Nmda receptor-dependent long-term potentiation and long-term depression (ltp/ltd). *Cold Spring Harbor perspectives in biology*, *4*(6), a005710.

Malenka, R. C. (1994). Synaptic plasticity in the hippocampus: Ltp and ltd. *Cell*, *78*(4), 535–538.

Marr, D. (1969). A theory of cerebellar cortex. *The Journal of Physiology*, *202*(2), 437–470.

Marshel, J. H., Garrett, M. E., Nauhaus, I., & Callaway, E. M. (2011). Functional specialization of seven mouse visual cortical areas. *Neuron*, *72*(6), 1040–1054.

Maunsell, J. H., & Van Essen, D. C. (1983a). The connections of the middle temporal visual area (mt) and their relationship to a cortical hierarchy in the macaque monkey. *Journal of Neuroscience*, *3*(12), 2563–2586.

Maunsell, J. H., & Van Essen, D. C. (1983b). Functional properties of neurons in middle temporal visual area of the macaque monkey. i. selectivity for stimulus direction, speed, and orientation. *Journal of neurophysiology*, *49*(5), 1127–1147.

Mikulasch, F. A., Rudelt, L., & Priesemann, V. (2022). Visuomotor mismatch responses as a hallmark of explaining away in causal inference. *Neural computation*, *35*(1), 27–37.

Mikulasch, F. A., Rudelt, L., Wibral, M., & Priesemann, V. (2023). Where is the error? hierarchical predictive coding through dendritic error computation. *Trends in Neurosciences*, *46*(1), 45–59.

Mirasso, C. R., Fraile, J. G., Scherr, F., Ramasco, J. J., Arkhipov, A., & Maass, W. (2023). Competition between bottom-up visual input and internal inhibition generates error neurons in a model of the mouse primary visual cortex. *bioRxiv*, 2023–01.

Miura, S. K., & Scanziani, M. (2022). Distinguishing externally from saccade-induced motion in visual cortex. *Nature*, *610*(7930), 135–142.

Muzzu, T., & Saleem, A. B. (2021). Feature selectivity can explain mismatch signals in mouse visual cortex. *Cell reports*, *37*(1), 109772.

Muzzu, T., & Saleem, A. B. (2023). Redefining sensorimotor mismatch selectivity in the visual cortex. *Cell Reports*, *42*(3).

Nau, M., Schindler, A., & Bartels, A. (2018). Real-motion signals in human early visual cortex. *NeuroImage*, *175*, 379–387.

Pearson, M. J., Dora, S., Struckmeier, O., Knowles, T. C., Mitchinson, B., Tiwari, K., Kyrki, V., Bohte, S., & Pennartz, C. (2021). Multimodal representation learning for place recognition using deep hebbian predictive coding. *Frontiers in Robotics and AI*, *8*, 732023.

Pennartz, C. M. (2015). *The brain's representational power: On consciousness and the integration of modalities*. MIT Press.

Pennartz, C. M. (2022). What is neurorepresentationalism? from neural activity and predictive processing to multi-level representations and consciousness. *Behavioural Brain Research*, *432*, 113969.

Pennartz, C. M., Dora, S., Muckli, L., & Lorteije, J. A. (2019). Towards a unified view on pathways and functions of neural recurrent processing. *Trends in neurosciences*, *42*(9), 589–603.

Pitzalis, S., Serra, C., Sulpizio, V., Committeri, G., de Pasquale, F., Fattori, P., Galletti, C., Sepe, R., & Galati, G. (2020). Neural bases of self-and object-motion in a naturalistic vision. *Human brain mapping*, *41*(4), 1084–1111.

Price, D. J., Kennedy, H., Dehay, C., Zhou, L., Mercier, M., Jossin, Y., Goffinet, A. M., Tissir, F., Blakey, D., & Molnár, Z. (2006). The development of cortical connections. *European Journal of Neuroscience*, *23*(4), 910–920.

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, *2*(1), 79–87.

Royden, C. S., & Holloway, M. A. (2014). Detecting moving objects in an optic flow field using direction-and speed-tuned operators. *Vision research*, *98*, 14–25.

Salvatori, T., Pinchetti, L., Millidge, B., Song, Y., Bao, T., Bogacz, R., & Lukasiewicz, T. (2022). Learning on arbitrary graph topologies via predictive coding. *Advances in neural information processing systems*, *35*, 38232–38244.

Sasaki, R., Anzai, A., Angelaki, D. E., & DeAngelis, G. C. (2020). Flexible coding of object motion in multiple reference frames by parietal cortex neurons. *Nature neuroscience*, *23*(8), 1004–1015.

Self, M. W., van Kerkoerle, T., Super, H., & Roelfsema, P. R. (2013). Distinct roles of the cortical layers of area v1 in figure-ground segregation. *Current biology*, *23*(21), 2121–2129.

Squires, N. K., Squires, K. C., & Hillyard, S. A. (1975). Two varieties of long-latency positive waves evoked by unpredictable auditory stimuli in man. *Electroencephalography and clinical neurophysiology*, *38*(4), 387–401.

Stone, L., & Lisberger, S. (1986). Detection of tracking errors by visual climbing fiber inputs to monkey cerebellar flocculus during pursuit eye movements. *Neuroscience letters*, *72*(2), 163–168.

St-Yves, G., Allen, E. J., Wu, Y., Kay, K., & Naselaris, T. (2023). Brain-optimized deep neural network models of human visual areas learn non-hierarchical representations. *Nature communications*, *14*(1), 3329.

Suzuki, M., Pennartz, C. M., & Aru, J. (2023). How deep is the brain? the shallow brain hypothesis. *Nature Reviews Neuroscience*, 1–14.

Urbanczik, R., & Senn, W. (2014). Learning by the dendritic prediction of somatic spiking. *Neuron*, *81*(3), 521–528.

Vasilevskaya, A., Widmer, F. C., Keller, G. B., & Jordan, R. (2023). Locomotion-induced gain of visual responses cannot explain visuomotor mismatch responses in layer 2/3 of primary visual cortex. *Cell Reports*, *42*(3).

von Helmholtz, H. (1860). In southall (ed.), handbuch der physiologischen optik, vol. 3.

Walsh, K. S., McGovern, D. P., Clark, A., & O'Connell, R. G. (2020). Evaluating the neurophysiological evidence for predictive processing as a model of perception. *Annals of the new York Academy of Sciences*, *1464*(1), 242–268.

Wang, Q., Gao, E., & Burkhalter, A. (2011). Gateways of ventral and dorsal streams in mouse visual cortex. *Journal of Neuroscience*, *31*(5), 1905–1918.

Warren, P. A., & Rushton, S. K. (2009). Perception of scene-relative object movement: Optic flow parsing and the contribution of monocular depth cues. *Vision research*, *49*(11), 1406–1419.

Wertheimer, M. (1923). Untersuchungen zur lehre von der gestalt (condensed translation from 1950). *A Source Book of Gestalt Psychology*.

Whishaw, I. Q., & Brooks, B. L. (1999). Calibrating space: Exploration is important for allothetic and idiothetic navigation. *Hippocampus*, *9*(6), 659–667.

Zmarz, P., & Keller, G. B. (2016). Mismatch receptive fields in mouse visual cortex. *Neuron*, *92*(4), 766–772.

Zuberbühler, K. (2008). Gaze following. *Current Biology*, *18*(11), R453–R455.

# 7 Supplementary material

## 7.1 Loss function and convergence of learning rule

To derive a principled loss function corresponding to the learning rule of Eq. 4, we first consider $w_-$ and then use the symmetry of the circuit to generalize to $w_+$. Plugging the firing rate of the nPE neuron from Eq. 1 into the weight update yields

$$\Delta w_- = \epsilon_l \cdot y_{mot} \cdot (\max(0, -(y_{vis} - w_- \cdot y_{mot}) + e_0) - e_0) \tag{11}$$

Here we assume the model to be moving ($g = 1$), otherwise no learning would take place as can be seen from Eq. 4. In the coupled training paradigm, it holds true that

$$\frac{y_{vis}}{y_{mot}} = \alpha \tag{12}$$

with $\alpha \in \mathbb{R}$. The resulting weight update is then the rectified linear function shown in blue in Fig. 8 with zero value derived from Eq. 11 at

$$w_-^* = \frac{y_{vis}}{y_{mot}} = \alpha \tag{13}$$

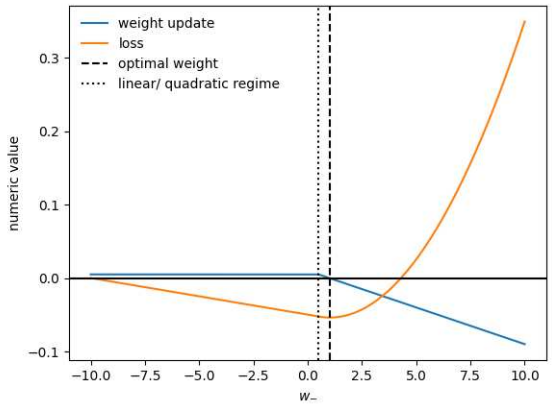


Figure 8: **Loss function underlying the learning rule.** As can be seen from the global minimum of the loss function (Eq. 15) depicted in orange, the weight updates (blue) are guaranteed to drive the weight towards the optimum value $w_-^*$ (dashed vertical line).

At this value, the net input to the error neuron is zero. Integrating weight updates yields the (negative) loss function on which learning performs gradient descent, shown in orange in Fig. 8. Based on the constant and negatively sloped linear parts of its negative derivative shown in blue, it can be seen that its global minimum lies at the optimal weight $w_-^*$ (denoted by the dashed line in Fig. 8) with strictly monotonic increase in both directions. This guarantees convergence of the learning rule to $w_-^* \pm \epsilon_l$. The transition from a linear to a quadratic regime (dotted line in Fig. 8) occurs at the firing threshold of the error neuron, where the net input to the neuron becomes zero

$$w_-^{thresh} = \frac{y_{vis} - e_0}{y_{mot}} \tag{14}$$

Integration of the weight update in each regime and division by the learning rate then yields the loss

$$L = \begin{cases} -y_{mot} \cdot e_0 \cdot w_-, & \text{if } w_- < w_-^{thresh} \\ y_{mot} \cdot \left(y_{vis} \cdot w_- - 1/2 y_{mot} w_-^2\right), & \text{otherwise} \end{cases} \tag{15}$$

with integration constants appropriately chosen (integration from $w_- = 0$). Analogous derivation for the pPE neurons yields the optimal weight

$$w_+^* = \frac{y_{vis}}{y_{mot}} = \alpha = w_-^* \tag{16}$$

## 7.2 Construction of the Animal dataset

To construct the Animal dataset, ten images of animals where obtained from pixhttps://pixabay.com under the Pixabay content license allowing free use and modification of the images for commercial and non-commercial use https://pixabay.com/service/terms/. As an illustration, one image can be found under https://pixabay.com/de/photos/pferd-grau-konik-weide-l%C3%A4uft-2388274/, and links to all images are listed in data/original/image_credits.txt in the GitHub repository). The animal images were cropped out of their original background, grayscaled, and resized so that the longer edge was 50 pixels long. Then, the images were pasted in front of the grass background texture (80 x 80 pixels) obtained from https://www.patternpictures.com/full-frame-green-grass-texture/ under the Pattern Picture license allowing unrestricted use (https://www.patternpictures.com/license/) allowing lateral shifting according to movement of the observer and the object (as described in section 2.2.2).

## 7.3 Hyperparameters of the two-stream model

The used hyperparameters are listed in Tab. 2.

| Symbol | Parameter | Value |
|---|---|---|
| $c_{thresh}$ | Segmentation threshold | 0.71 |
| $e_0$ | Baseline activity V1-EN | 0.5 a.u. |
| $\epsilon_{inf}$ | Inference rate MT & MST | 0.05 |
| $\epsilon_l$ | Learning rate pretraining | 200 |
| $\epsilon_l$ | Learning rate training | 0.02 |
| - | MT feature maps | 4 |
| - | MST feature maps | 4 |
| kernel_size | Kernel size MT-V1 | [4, 4] |
| - | Kernel size MST-MT | [4, 4] |
| - | Convolutional stride MT-V1 | 1 |
| - | Convolutional stride MST-MT | 1 |

Table 2: **Hyperparameters of the two-stream model.**

## 7.4 Simultaneous learning of visual and visuomotor stream

In the main results, motor-to-visual and visual-to-visual streams were trained independently. Upon activation and training of the visual stream, the connections in the motor-to-visual stream were frozen. In the development of cortico-cortical brain connectivity (Price et al., 2006), not much is known about the developmental order of feedback connections from MT to V1 relative to connections from motor areas to V1. Hints come from the work of Baldwin et al., 2012 that suggests the presence of feedback connections from macaque MT to V1 about two weeks after birth. To investigate sensitivity of the model to separate training phases, we studied simultaneous training of all plastic connections. To investigate this, we combined both pretraining (background only) and training (Animals) data into an interleaved dataset. Here, one in ten sequences contained an animal, the other nine were empty. Inference was conducted as described in section 2.2 and plasticity was enabled in both streams. In general we found this training paradigm to be quite unstable compared to the separate training phases. To achieve segmentation performance, several modifications proved helpful:

1. A high learning rate in the motor-to-visual stream compared to the visual-to-visual stream. This enforced learning of a correct motor-to-visual stream on a fast timescale and thus partially decoupled the two learning problems.

2. A large ratio (>10:1) of empty sequences over sequences with objects in the training data to ground the motor-to-visual stream.

3. Adapting the segmentation threshold. We found predictions of the visual stream to be about 30% weaker in this joint training paradigm compared

to the separate training paradigm. Lowering the threshold (of summed up predictions) for assigning a point in space to an external object prevented the segmented areas from becoming too small.

With these tweaks, the jointly trained model reached an IoU of 0.56 compared to 0.68 of the separately trained model. The worse performance illustrated in Fig. 9 was likely due to noisier predictions resulting from interactions between the two streams via V1 error neurons.

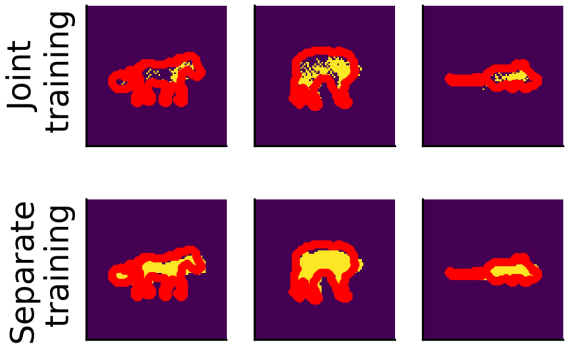


Figure 9: **Joint training leads to noisier segmentation performance compared to separate training of both streams.** Segmentation mask from MT→V1-feedback (yellow) over the ground truth of the object outline (red) on the Animal dataset (cf. Fig. 6). Top row: model trained with plasticity enabled in both streams, bottom row: separate pretraining and training phases as described in section 2.2.2

.

## 7.5 Learning distinct forward models from multiple actions

Distinct movements create differing optic flow patterns, such as when walking forward as opposed to turning around. The model's capacity to learn such mappings was demonstrated by training it on the multidimensional dataset of artificially generated optic flows described in section 2.2.2. Figure 10 illustrates the wiring of the model. After training the motor-to-visual stream for five epochs, the model's prediction from the motor stream closely matched the original optic flow patterns, as shown in Fig. 11.

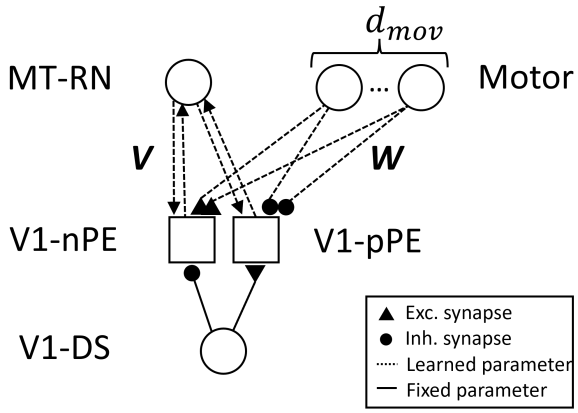


Figure 10: **Forward models from multidimensional action states.** In an extension of the model shown in Figure 3, multiple motor neurons encode a $d_{mov}$-dimensional action space. For abbreviations see Fig. 3.

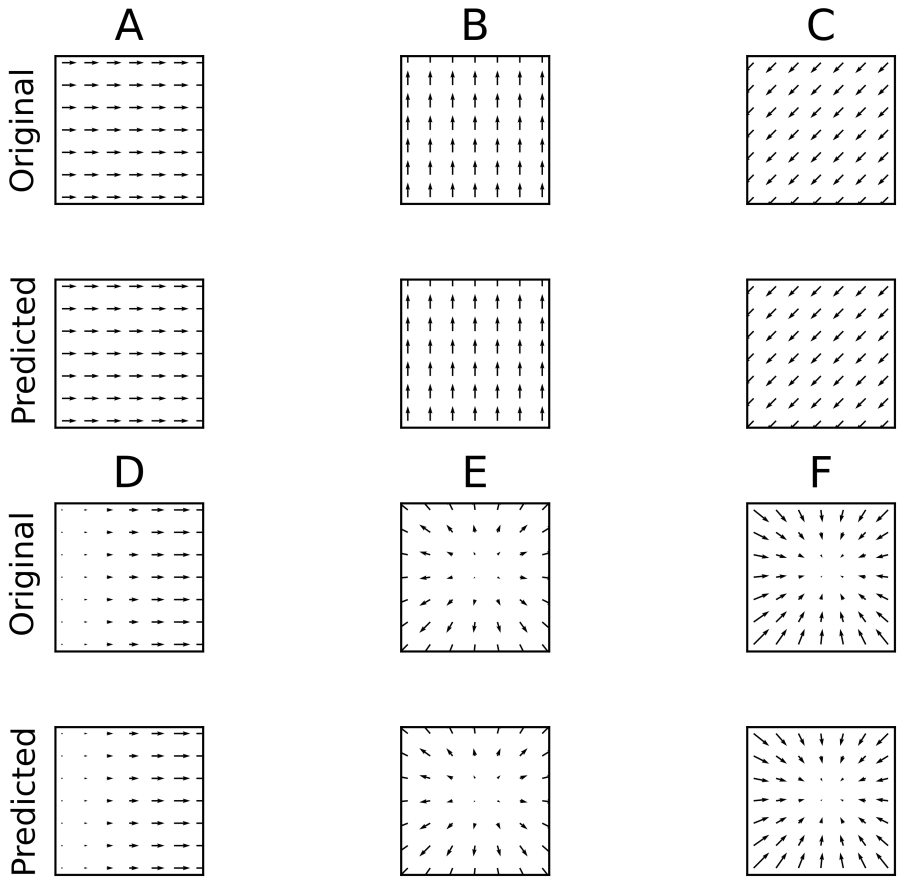


Figure 11: **Learning outcomes of multiple movements in one model.** Optic flow input patterns of the six movement dimensions are shown above predictions from motor stream to V1 after training. "Original" optic flow is obtained from V1-DS neurons and predicted optic flow from mtr-V1 signals (averaged across V1-nPE and V1-pPE at each point in retinotopic space). The vectors (plotted at every fourth pixel column/row) are computed by subtracting opposing signal components: the $x$-component of the vector is given by rightward minus leftward (predicted) optic flow, the $y$-component by upwards minus downwards (predicted) optic flow. The actions corresponding to the chosen optic flow patterns are (A) turning leftwards, (B) looking downwards, (C) looking up and to the right, (D) following a leftward curve while walking, (E) moving forward and (D) moving backwards.

### 7.6 Addition of a second purely visual area

Vision is generally understood as a hierarchical cortical process (Lee & Mumford, 2003)), although see Suzuki et al., 2023 and St-Yves et al., 2023. Also most neural networks models of generative perception are strictly hierarchically organized (Dayan et al., 1995; Dora et al., 2021; Rao & Ballard, 1999), although see Salvatori et al., 2022. To extend our model in depth, we added an area on top of MT (Fig. 12). Neuroanatomically, this area roughly maps onto the medial superior temporal cortex (MST) because it is further removed from thalamic sensory input in the visual hierarchy than MT (Born & Bradley, 2005; Felleman & Van Essen, 1991). Interconnected with MT neurons via linear error neurons (MT-EN, for simplicity not split into positive and negative counterparts), area MST learns a generative model of activity patterns in area MT. Connections are convolutional with stride 1, kernel size 4 and mapping to four feature maps in area MST. With the input resolution of the modified FashionMNIST dataset described in section 2.2.2, the number of neurons in MST is thus 34 x 34 x 4 = 4,624.

As in area MT (see section 2.2), inference and learning in area MST are driven by error signals from the area below. Since we were interested in the representations formed in MST, we did not implement an influence of MT-EN on area MT-RN as common in other predictive coding implementations and Eq. 6 remains unchanged. Thus, the remaining network function is not affected by the addition of area MST. Training and readout evaluation then progressed as described in sections 2.2.2 (training) and 3.4.

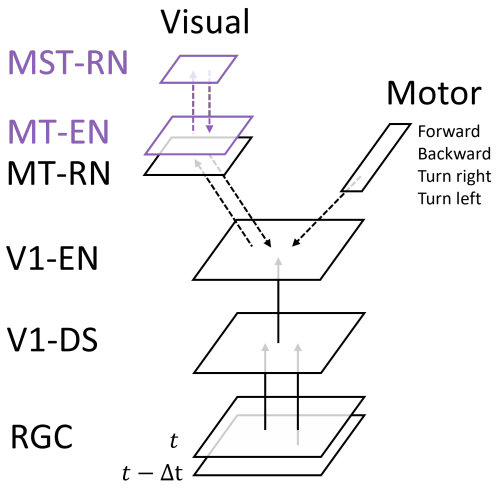


Figure 12: **Extension of the visual-to-visual stream.** Shown in purple is the added model area MST and its connection to area MT via an added population of error neurons (MT-EN). Connections subject to synaptic plasticity are depicted as dashed lines. Compare to Fig. 3.

## 7.7 Statistical methods

To analyze whether significant differences between readout accuracy (on the test dataset) across network areas were present, we first conducted a Welch ANOVA using the pingouin package in Python (https://pingouinstats.org/build/html/index.html). The null hypothesis (i.e., no significant difference in readout accuracy across representation neurons in the three network areas) was rejected with $p <$ 1e-5 (Tab. 7.7).

| | Source | ddof1 | ddof2 | F | p-unc | np2 |
|---|---|---|---|---|---|---|
| 0 | Population | 3 | 5.698583 | 142.910062 | 0.000009 | 0.953495 |

Table 3: **Full report on the outcome of the Welch's ANOVA for the readout of object class.** ddof1: degrees of freedom (numerator), ddof2: degrees of freedom (denominator), p-unc: uncorrected p-values, np2: Partial eta-square effect sizes. This figure supports main text section 3.4.

Subsequently, multiple pairwise comparisons were conducted using the *pingouin* implementation of the Games-Howell post-hoc test, which led to the corrected *p*-values reported in section 3.4 of the main text. All comparisons are shown in Tab. 7.7.

Games-Howell post-hoc

| | A | B | mean(A) | mean(B) | diff | se | T | df | pval | hedges |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Full vis. | MST | 0.76775 | 0.82275 | -0.05500 | 0.003075 | -17.883614 | 3.497725 | 0.000481 | -10.996195 |
| 1 | Full vis. | MT | 0.76775 | 0.81550 | -0.04775 | 0.003955 | -12.071843 | 3.292599 | 0.002467 | -7.422680 |
| 2 | Full vis. | V1-DS | 0.76775 | 0.79625 | -0.02850 | 0.002407 | -11.842487 | 3.846641 | 0.001247 | -7.281654 |
| 3 | MST | MT | 0.82275 | 0.81550 | 0.00725 | 0.004863 | 1.490940 | 5.615486 | 0.499418 | 0.916743 |
| 4 | MST | V1-DS | 0.82275 | 0.79625 | 0.02650 | 0.003714 | 7.135718 | 5.603904 | 0.002063 | 4.387578 |
| 5 | MT | V1-DS | 0.81550 | 0.79625 | 0.01925 | 0.004470 | 4.306674 | 4.825987 | 0.029702 | 2.648068 |

Table 4: **Full report of the post-hoc pairwise test (Games-Howell).** Columns A and B refer to the neuronal populations tested in a pairwise manner. Columns: mean: average readout accuracy on the test data across the randomly initialized runs, diff: difference between mean values, se: standard error, df: adjusted degrees of freedom, pval: Games-Howell corrected p-values, hedges: Hedges effect size.