Bayesian causal inference unifies perceptual and neuronal processing of center-surround motion in area MT.

Gabor Lengyel[1,3†], Sabyasachi Shivkumar[1,2,†], Gregory C. DeAngelis[1] & Ralf M. Haefner[1]

October 23, 2025

Abstract

Center-surround processing is a ubiquitous computation in the brain, operating across levels from early sensory encoding to abstract conceptual representation and shaping both behavior and neural responses. However, current models of these neural interactions, typically based on divisive normalization or descriptive frameworks, fail to account for the wide range of center-surround effects. Focusing on the domain of motion processing, we propose that both neural and perceptual phenomena related to center-surround motion emerge from a single normative principle: Bayesian causal inference over the latent structure of motion. We derived neural predictions from a causal inference model of perception that dynamically infers the most probable reference frame for moving stimuli. Specifically, we generated predictions for both the mean response and variability of single neurons across the full 4D stimulus space of center and surround motion (i.e., directions and speeds). These predictions exhibit complex interactions that are unexplainable by classic DN models but qualitatively match the wide range of neural data from the primate middle temporal area (MT). Our model resolves two long-standing puzzles: it explains when and why surround stimuli suppress versus facilitate neural responses, and it shows why some MT neurons are found to encode motion in retinal coordinates while others encode motion relative to an inferred reference frame. By unifying disparate neural observations with perception through this model, our work suggests new, theory-driven experiments to probe the circuit-level implementation of these computations.

Keywords: Bayesian causal inference; Center-surround processing; Motion perception; Neural coding in MT; Motion integration; Motion segmentation

[1] Department of Brain and Cognitive Science, University of Rochester, NY, US
[1] Center for Visual Sciences, University of Rochester, NY, US
[2] Zuckerman Institute, Columbia University, New York, NY, US
[3] Institute of Science and Technology Austria, Klosterneuburg, AT
[†] Equal contribution

Correspondence should be addressed to: lengyel.gaabor@gmail.com

1 Introduction

Center-surround (CS) processing is a canonical computation in the brain, shaping vision from its earliest stages (Allman et al., 1985b; Angelucci et al., 2017). Starting in the retina to enhance contrast sensitivity and facilitate the detection of boundaries (Baden et al., 2020; Euler et al., 2014; Gaynes et al., 2022; Kuffler, 1953; Turner et al., 2018), this principle of surround-dependent contextual modulation is repeated throughout the visual hierarchy. In the ventral stream, it refines feature selectivity in V1 (Angelucci et al., 2017; Nurminen & Angelucci, 2014) and shapes object representations in the inferotemporal cortex (Angelucci et al., 2017; Carandini & Heeger, 2012; Zoccolan et al., 2005). This principle is also fundamental to the dorsal stream's processing of motion (Born, 2000; Born & Bradley, 2005; Britten & Heuer, 1999; DeAngelis & Uka, 2003; Pack et al., 2004; Rust et al., 2006). In the middle temporal area (MT), for instance, a moving surround can powerfully suppress a neuron's response or, under different conditions, facilitate it (Allman et al., 1985a; Born, 2000; Born & Bradley, 2005; Born & Tootell, 1992; Bradley & Andersen, 1998; DeAngelis & Uka, 2003; Huang et al., 2007; Inaba et al., 2011, 2007; Newsome et al., 1988; Pack et al., 2004, 2005; Raiguel et al., 1995; Tanaka et al., 1986; Tzvetanov & Womelsdorf, 2008; Xiao et al., 1997; Zemel et al., 1998). Perceptually, this same stimulus can cause a central motion to be integrated with the surround into a unified whole object motion, or to be segmented from the surround in stark contrast (Bill et al., 2022, 2020; Braddick, 1993; Dakin & Mareschal, 2000; Gershman et al., 2016; Penaloza et al., 2024; Shivkumar et al., 2025; Tadin et al., 2003, 2019; Zarei Eskikand et al., 2020).

Despite its ubiquity, a unifying computational model explaining these diverse CS interactions on both at the perceptual and neural levels has remained elusive. At the neural level, we have a rich, process-level description of what happens, but no normative theory for why it happens and what computational principles can produce such profoundly different and context-dependent outcomes. For instance, decades of neurophysiological research have meticulously catalogued the variety of CS interactions in MT, from the direction-selective antagonistic suppression first described by Allman et al. (1985a) and Tanaka et al. (1986), to the reinforcing facilitation and more complex patterns later categorized by Born (2000), Born and Bradley (2005), and Born and Tootell (1992). However, existing models of the V1-MT pathway, often based on linear-nonlinear integration or divisive normalization, can account for a subset of suppressive phenomena but struggle to explain the full range of effects (Adelson & Bergen, 1985; Albright, 1984; Barth & Watson, 2000; Movshon et al., 1985; Qian et al., 1994a; Rust et al., 2006; Simoncelli & Heeger, 1998; Wilson et al., 1992; Wilson & Kim, 1994; Zarei Eskikand et al., 2020). Conversely, influential models of perception successfully describe phenomena like motion integration and segmentation but have not been linked to the rich diversity of single-neuron CS response properties (Bill et al., 2022, 2020; Braddick, 1993; Dakin & Mareschal, 2000; Gershman et al., 2016; Penaloza et al., 2024; Shivkumar et al., 2025; Tadin et al., 2003, 2019; Zarei Eskikand et al., 2020).

Recent normative models of motion perception suggest that causal inference over reference frames provides a unifying framework for both motion integration and segmentation (Bill et al., 2022, 2020; Gershman et al., 2016; Penaloza et al., 2024; Shivkumar et al., 2025). Causal inference is a Bayesian computational motif for inferring when to combine information across cues (Körding et al., 2007) and it has recently been proposed to be a universal computation across the cortex (Shams & Beierholm, 2022). This line of work proposed that the brain infers the causal structure underlying the elements moving in a visual scene, and uses that structure to determine the reference frames within which to represent the motion of each element. These models successfully explained the observed behavior in psychophysical experiments using CS motion stimuli, effectively bridging the gap between normative computations and behavior (Bill et al., 2022, 2020; Gershman et al., 2016; Penaloza et al., 2024; Shivkumar et al., 2025).

In this study, we built on one of these recently proposed models, which accurately captures the perceptual effects of CS stimuli (Fig. 1a,d) using a biologically plausible architecture (Fig. 1b) (Shivkumar et al., 2025). From this model, we derived neural predictions for mean responses (tuning curves, Figs. 4 and 5, and Figs. S2 to S4) and neural variability (Fano factor, Fig. 5, and Figs. S2 and S3) across the entire 4-dimensional (motion direction and speed in center and surround) stimulus space. The neural predictions of Bayesian causal inference captured a wide range of neural CS interaction effects (Figs. 5 to 7, and Figs. S2 to S5). The predicted CS interactions reflected both the representation of motion relative to a reference frame, inferred by causal inference, and the representation of motion in retinal coordinates. Finally, we show that these predictions are compatible with classic neurophysiological findings (Figs. 6 and 7, and Fig. S5), providing a normative explanation why the surround motion facilitates or suppresses neural responses, and when these effects are direction dependent. Our results suggest future experiments – both to provide more comprehensive tests of the normative theory and to link it to circuit-level mechanistic

models.

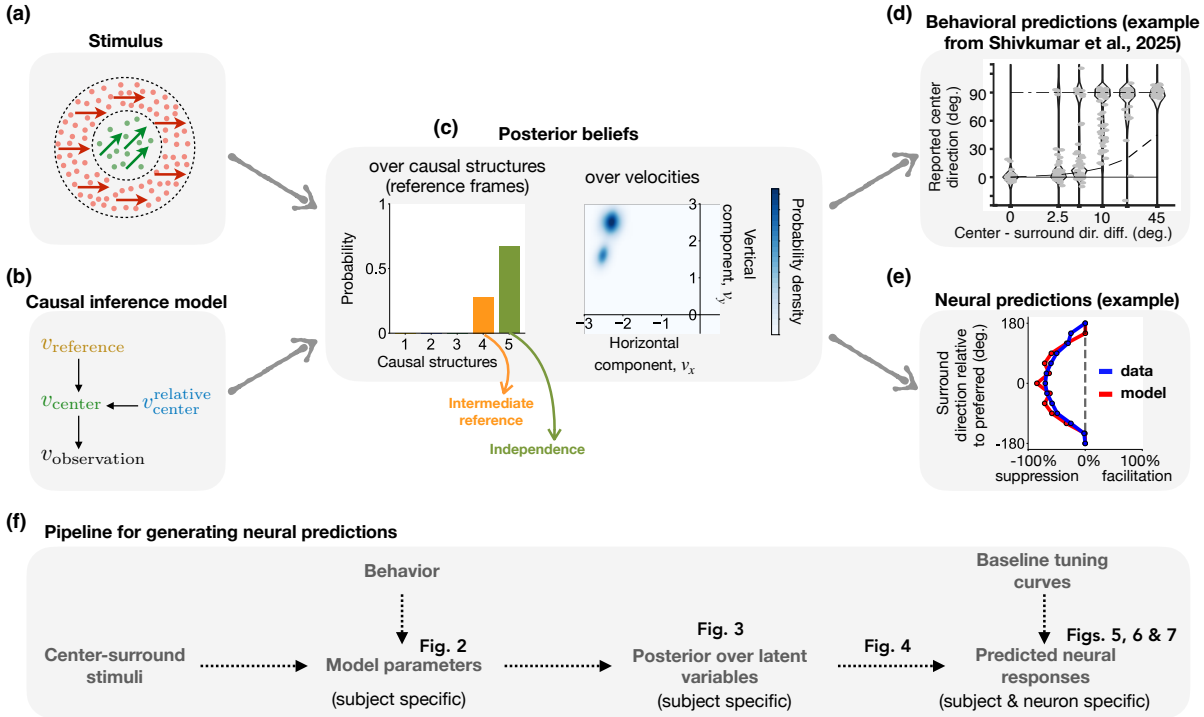

Figure 1: **A normative, Bayesian framework for linking center-surround processing in perception and neural activity.** This figure provides a conceptual overview of the study's approach, from the experimental paradigm and Bayesian model to its connection with behavioral and neural data. **(a) Human experimental paradigm.** Observers view a central moving dot patch (green) within a moving surround (red) and report the center's perceived direction. The retinal velocities of the center and surround are depicted by green and red arrows, respectively. **(b) Generative model for the observed stimulus velocity.** The model assumes the true retinal center velocity, v_{center} , is equal to the vector sum of the reference frame velocity, $v_{\text{reference}}$, and the center velocity relative to that reference frame, $v^{\text{relative}}_{\text{center}}$. This true center velocity then produces a noisy sensory measurement, $v_{\text{observation}}$. **(c) Posterior beliefs.** The brain infers posterior beliefs over latent variables, such as causal structure and velocities. The example shows posteriors where the observer infers two possible causal structures (on the left): Intermediate reference (perceiving motion relative to a reference frame intermediate between the center and surround motions) and Independence (perceiving the two patches as moving independently). The posterior over $v^{\text{relative}}_{\text{center}}$ (on the right) reflects this, showing two modes that correspond to these two structures. **(d) Connecting the model to behavior.** The model is constrained by fitting it to human psychophysical data. The panel shows reports (gray dots) and the model's posterior predictions (black violins) for an example observers from Shivkumar et al. (2025). The plots display the center's perceived direction (y-axis) as a function of the directional difference between the center and surround (x-axis). **(e) Connecting the model to neural activity.** The model's posterior beliefs, constrained by behavior, are then used to predict the modulation (suppression/facilitation) of single MT neurons in response to complex surround motion. **(f) Overall research pipeline.** In summary, the framework uses a pipeline where the model is first fitted to behavioral responses (Fig. 2). These subject-specific parameters are then used to compute posterior beliefs for any center-surround stimulus (Fig. 3). Finally, these posteriors are combined with baseline neural tuning curves (measured with a stationary surround) to generate testable predictions (Fig. 4) for how single neurons will respond to complex moving surrounds (Fig. 5).

2 Results

Our paper is structured as follows. First, we introduce the Bayesian causal inference model we adapt, along with the CS motion experiment and the behavioral results used to fit the model parameters. Next, we examine the variety and complexity of posterior beliefs inferred by the model in CS motion tasks. We then describe how neural predictions – including tuning curves and their variability – are derived from the model. Following this, we characterize the CS interactions present in these predictions. Finally, we demonstrate that Bayesian causal inference qualitatively accounts for a wide range of neural CS interactions observed in MT.

2.1 Task, model, and prior empirical results

We used the model of Shivkumar et al. (2025) (Fig. 2c-g) with parameters inferred from behavioral data of five human observers in their experiment. Participants reported their perceived motion direction of

a central target patch composed of multiple moving dots (Fig. 2a, green dots), while the target was surrounded by a ring of moving dots (Fig. 2a, red dots). The experiment was designed to measure how participants' motion direction estimates are influenced by the directions and speeds of both the center and surround. This stimulus is commonly used to investigate CS processing in motion perception and in neurophysiology (e.g., Adelson and Bergen, 1985; Born, 2000; Tadin et al., 2008, 2019; Tanaka et al., 1986).

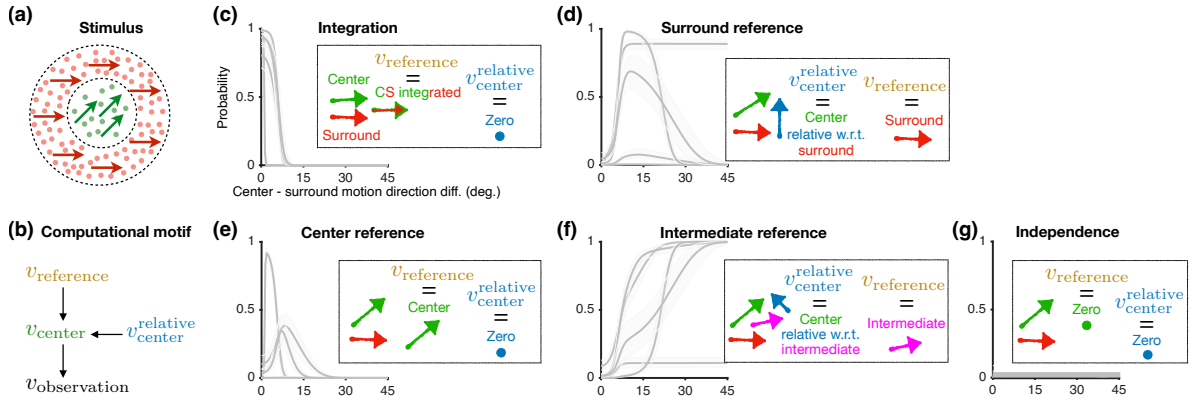

Figure 2: **A hierarchical Bayesian model for inferring the causal structure of visual motion. (a)** Same as Fig. 1a. **(b)** Same as Fig. 1b. **(c-g)** Posterior probability assigned to competing causal motion structures. The model considers 12 hypotheses for the relationship between center and surround motion. Here, we show 5 of the 12 structures depicted schematically (arrows: motion; dots: stationary) in the inset boxes. These include four structures assuming a common cause (c-f) and one assuming independent causes (g). The plots show the median posterior probability across five observers (solid grey lines) with 95% credible intervals (shading) as a function of the directional difference between the center and surround stimuli. Observers consistently assigned negligible probability to the independent-causes structure (g). Adapted from Shivkumar et al. (2025).

The hierarchical Bayesian causal inference model from Shivkumar et al. (2025) proposes that the brain jointly infers the appropriate reference frames ($v_{\text{reference}}$ in Fig. 2) for each stimulus and its motion in that reference frame ($v^{\text{relative}}_{\text{center}}$ in Fig. 2). Applied to this CS stimulus (Fig. 2a), the model assumes that observers infer the probability of 12 possible causal structures. Notably, in Shivkumar et al. (2025), four of these had significant probability mass. While we use these four for illustration, our model's predictions incorporate all 12 structures. In addition to these four, we also include a fifth causal structure in which the center and surround move independently. In this "independence" structure, motion is perceived relative to a world reference frame (Fig. 2g and Fig. S1k). This can be considered as motion segmentation with respect to a world reference[1]. This structure had negligible mass in Shivkumar et al. (2025), presumably due to the patches' proximity, shared temporal dynamics, and small motion differences. However, the wider stimulus range (than those used in the original experiment) we explore here makes the independence structure as a plausible alternative.

The first of the four main causal structures is classical motion "integration" (Braddick, 1993; Tadin & Lappin, 2005; Tadin et al., 2003) (see Fig. 2c; generative model in Fig. S1c). Observers infer this structure as the most probable when the differences between center and surround patches' motions are minimal, leading to a "cue-combined" perception in which both motions are integrated (Fig. 2c, curves). In contrast, motion segmentation (Braddick, 1993; Tadin et al., 2019), the second causal structure, arises when the motion differences are significant. We term this second structure "surround reference" (Fig. 2d and Fig. S1e) because observers segment the center's motion from the surround perceiving it relative to the surround's motion. Interestingly, this structure was dominant for only one of five observers, playing a less significant role for the other participants (Fig. 2d).

The third and fourth causal structures account for the behavioral variability beyond canonical integration and segmentation. In the third structure, observers perceive the surround motion relative to the center group, using the center as a reference frame (Fig. 2e and Fig. S1g). Since observers in this task reports the motion direction of the center patch, this "center reference" structure generates responses that are unaffected by the surround patch, much like the independence structure (Fig. 2g, Fig. S1k). The fourth structure captures when observers adopt an "intermediate reference" frame, between the center and surround patches' motions, perceiving motions of both patches relative to this intermediate

[1] In an experiment where participants' heads and eyes are fixed, retinal, ego-centered, and world coordinates are indistinguishable – and our predictions do not depend on that distinction.

frame. Most observers put significant probability mass to this structure across many different CS stimuli configuration (Fig. 2f and Fig. S1i).

These latter two structures can be easily miscategorized as classical integration or segmentation. For instance, the center reference structure can be confused with motion integration when center and surround motions are very similar, because the theoretically integrated percept would lie very close to the center's actual motion. Similarly, the intermediate reference structure can be mistaken for motion segmentation if one assumes that the subtraction of the surround's velocity from the center's is incomplete (Shivkumar et al., 2025). The Bayesian causal inference model resolves these ambiguities, differentiating between all these structures and explaining diverse behavioral phenomena through a single computational principle.

2.2 Causal inference produces complex posterior beliefs

In order to generate neural predictions for arbitrary CS stimulus, we independently varied CS velocities (consisting of direction and speed each) and computed the posteriors over all latent variables in the Bayesian causal inference model using parameters fitted to the five observers in Shivkumar et al. (2025). The posteriors can be computed using any parameter values hypothesized by the researcher. However, the parameters fitted to real behavioral responses will reveal posterior beliefs that are backed up by empirical data. The model (Fig. 2b) contains three latent variables that are involved in representing the center patch's motion: $v_{\text{reference}}$, v_{center} , and $v_{\text{center}}^{\text{relative}}$. $v_{\text{reference}}$ denotes the motion of the reference frame in retinal coordinates. v_{center} represents the motion in retinal coordinates while $v_{\text{center}}^{\text{relative}}$ represents the motion relative to the reference frame. As a general principle, the posterior over v_{center} is only minimally affected by surround motion, typically forming a simple, Gaussian-like distribution centered on the physical velocity of the center patch, reflecting the uncertainty due to observation noise (Fig. 3g, bottom). In contrast, the posteriors for $v_{\text{reference}}$ and $v_{\text{center}}^{\text{relative}}$ are strongly modulated by the surround, often displaying complex, multi-modal forms characteristic of causal inference (Fig. 3c-f). In the main text, we illustrate our findings using a representative observer (observer #2 in Exp. 1 in Shivkumar et al. (2025); Figs. 3 to 5), with the results for all participants available in Figs. S1 and S3.

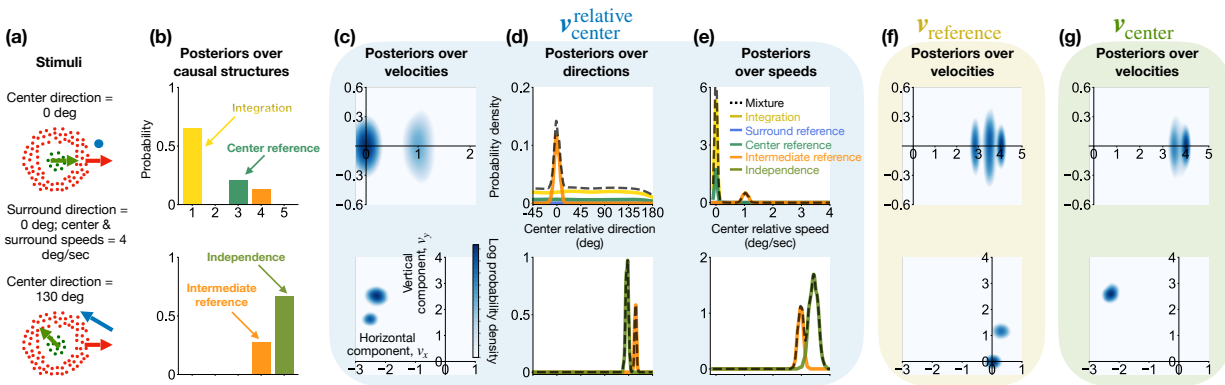

Figure 3: **Causal inference generates complex, multi-modal posterior beliefs about latent motion variables.** Posteriors were computed for observer #2 from Shivkumar et al. (2025) in response to two distinct center-surround stimuli (top vs. bottom row). **a:** The CS stimuli. The surround motion was fixed at 0° (red arrow), while the center motion differed between the two examples (green arrow). The relative velocity vector (center - surround velocity) is shown in blue. The blue dot in the top row represents zero velocity. **b:** Posterior probabilities over the five causal motion structures (see Fig. 2c-g). Note how the most probable structure changes depending on the stimulus. **c-e:** Posterior beliefs about relative motion ($v_{\text{center}}^{\text{relative}}$). These panels show the full 2D posterior over velocity, and the marginal posteriors for direction and speed (dashed line: posterior beliefs; colored lines: individual mixture components corresponding to the motion structures). The multi-modal and complex shapes reflect the observer's uncertainty over the causal structure. **f and g:** Posterior beliefs about the reference frame ($v_{\text{reference}}$) and retinal motion (v_{center}). Marginal posteriors for direction and speed are not shown for these variables. To aid visualization, all 2D velocity posteriors (c, f, g) are displayed using log probability density, which makes secondary modes visible. In contrast, the 1D marginal posteriors (d, e) are plotted using linear probability density.

To illustrate the complexity arising from the causal inference process, we first examine the model's posterior beliefs for a nominally simple stimulus where the center and surround move identically (Fig. 3, top row). Contrary to expectation for such a simple stimulus, most observers did not infer a single causal structure. Instead, they assigned significant probability mass to a mixture of competing structures. For instance, observer #2 in Shivkumar et al. (2025) assigned high probability to three structures: motion integration, the center patch acting as a reference frame, and an intermediate reference frame. This

mixture of inferred structures directly shapes the posterior distributions. The posterior over the reference frame ($v_{\text{reference}}$; Fig. 3f) consequently shows three distinct peaks corresponding to these probable structures. The dominant peak aligns with motion integration (around 0° direction and 4 deg/s speed), while secondary modes correspond to the center-reference (middle peak at 3.7 deg/s) and intermediate-reference (left peak at 2.8 deg/s) interpretations. Notably, the peak for the center-reference mode reflects a slower motion than the true stimulus speed; this discrepancy is a direct consequence of the slow-speed prior inherent in the model (Shivkumar et al., 2025).

Correspondingly, the posterior over relative velocity ($v_{\text{center}}^{\text{relative}}$; Fig. 3c-e) is not a simple peak at zero. While its main component is concentrated at zero velocity (corresponding to both integration and center-reference structures), it exhibits one additional mode that reflects the observer's belief in the intermediate-reference structure. This additional mode is visible as a smaller peak in the full posterior view (top, blue contour around $v_x = 1$ deg/s in Fig. 3c). This single mode manifests in the marginal distributions as both a peak at 0° for direction (in addition to the uniform distribution representing the stationary component; Fig. 3d) and a secondary peak near 1 deg/s for speed (Fig. 3e).

The posterior over retinal motion (v_{center} ; Fig. 3g), while typically unimodal, also displays a subtle bimodality in this case. This corresponds to the center-reference structure, which, as noted, was inferred to be slightly slower than the true stimulus speed. This could indicate a small perceptual bias for this observer even under this simple CS stimulus. Together, this example demonstrates that due to uncertainty over the underlying causal structure, the model's posterior beliefs are highly complex even when the stimulus itself seems unambiguous to the experimenter.

A second example highlights how the model arbitrates between competing causal structures when the directional difference between the center and surround motion is large (Fig. 3, bottom row). The same example observer infers two probable structures for this CS stimulus (Fig. 3b): (1) that the motions are independent, or (2) that they share a common intermediate reference frame. Each of these interpretations corresponds to a distinct mode in the bimodal posteriors for $v_{\text{reference}}$ and $v_{\text{center}}^{\text{relative}}$. The independent-motion structure generates a mode for $v_{\text{reference}}$ at zero velocity (Fig. 3f), with a corresponding mode for $v_{\text{center}}^{\text{relative}}$ at the retinal velocity of the center patch (top blue contour in Fig. 3c). Conversely, the intermediate-reference structure produces a second mode for $v_{\text{reference}}$ at the inferred reference velocity (bottom blue contour around $v_y = 1$ deg/s in Fig. 3f), with $v_{\text{center}}^{\text{relative}}$ peaking at the center's velocity relative to this moving frame (bottom blue contour in Fig. 3c).

2.3 A novel sampling-based method for predicting single neural activity from posterior beliefs

To bridge the gap between our normative, Bayesian causal inference model and neural data, we introduce a novel linking method which posits that single-neuron activity reflects samples from the model's posterior beliefs (i.e., the neural sampling hypothesis; Fiser et al., 2010; Hoyer and Hyvärinen, 2002). Neural sampling enables us to generate neural predictions that reflect the complex, multi-modal nature of the model posteriors. Our method provides a general framework for predicting single neural responses to any "test" stimulus from any Bayesian model Fig. 4. Specifically, it posits that the response to the test stimulus can be derived from the same neuron's tuning curve measured for a simpler "baseline" stimulus set. This baseline tuning curve acts as a "look-up table": each sample from the test posterior is converted into an individual firing rate (using the look-up table), and aggregating all these rates forms a full distribution of predicted responses (more details about the method are provided in the Methods 4.3). This method can be applied in any domain, as long as its two key components are well-defined: (1) the Bayesian model parameters to compute posterior beliefs and (2) the tuning curve parameters to link posterior samples to firing rates.

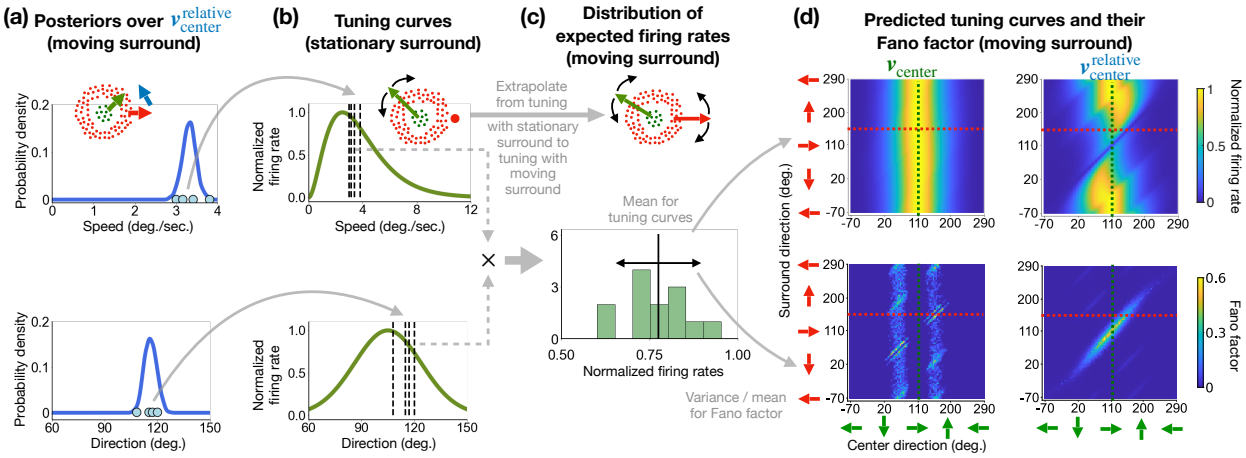

Figure 4: **Linking posterior beliefs to neural activity via the neural sampling hypothesis.** Schematic of the linking model used to generate neural predictions from the Bayesian observer model's posterior distributions. All firing rates are normalized. **(a)** For a given center-surround stimulus (inset), the model's posterior belief over a latent variable (here, $v_{\text{center}}^{\text{relative}}$) is approximated by a set of discrete samples (blue dots), in line with the neural sampling hypothesis (Fiser et al., 2010; Hoyer & Hyvärinen, 2002). **(b)** A baseline tuning curve is measured (here, in response to a stimulus with a stationary surround). This baseline serves as a look-up table to convert each sample from the test posterior (here, in response to a stimulus with a moving surround) into a predicted firing rate. **(c)** The mapped activities for speed and direction are multiplied (symbol $\times$), yielding a full distribution of predicted firing rates. The mean (solid black line) and variance (black arrow) of this distribution form our predictions for the neuron's response to the specific stimulus in (a). **(d)** This entire process is repeated for a wide range of CS stimuli to generate complete 2D tuning curves. Shown are the predicted mean firing rate (top) and Fano factor (bottom) for a neuron assumed to represent either retinal motion (v_{center} , left) or relative motion ($v_{\text{center}}^{\text{relative}}$, middle). Model parameters are from observer #2 shown in Fig. 2.

To apply this method to our CS motion stimuli, we define these two components as follows. For the Bayesian model parameters, we used the unique parameters fitted to each of the five human observers in Shivkumar et al. (2025). For the tuning curve parameters, we used the median of the parameters of empirically measured MT tuning curves from monkeys (DeAngelis & Uka, 2003). While this combines human behavioral models with monkey neurophysiology, it provides a strong proof of principle. Ideally, both behavioral and neural data would be collected from the same subjects.

We then implemented the linking procedure (Fig. 4). First, for a given test stimulus (moving surround), we generated samples from the model's posterior distribution over a latent variable (Fig. 4a). Second, we used the baseline tuning curve (derived from the median parameters of MT neurons' tuning curves measured to a stationary surround in DeAngelis and Uka (2003)) as a look-up table (Fig. 4b) to map each posterior sample to a predicted firing rate yielding a full distribution of predicted firing rates (Fig. 4c). Finally, we calculated the mean and variance of this distribution, which serve as our predictions for the neuron's mean firing rate and response variability (Fig. 4d). In essence, this method leverages a neuron's simple motion tuning curve to predict its complex response to center motion with moving surround. It achieves this by assuming the neuron's firing is determined not by the raw stimulus, but by the brain's internal belief about the probable motion components and their reference frames, as represented by the samples from model's the posteriors (full details of the method are provided in Methods 4.3).

2.4 Predicted 4D tuning curves reveal complex signatures of causal inference

Applying our linking method, we generated neural predictions for neurons representing posterior beliefs over either v_{center} or $v_{\text{center}}^{\text{relative}}$. These variables correspond to classical receptive fields overlapping the center stimulus, matching the empirical studies we aim to explain (Allman et al., 1985a; Born, 2000; Born & Bradley, 2005; Born & Tootell, 1992; Tanaka et al., 1986). In contrast, the spatial extent of a receptive field representing $v_{\text{reference}}$ is more ambiguous, and we therefore do not analyze predictions for $v_{\text{reference}}$ here.

First, we tested the hypothesis that the two main classes of MT neurons (Born & Bradley, 2005) – those insensitive to the surround and those with strong surround modulation – correspond to neurons encoding v_{center} and $v_{\text{center}}^{\text{relative}}$, respectively. We generated four-dimensional (4D) tuning curves (Fig. 5, and Figs. S2 and S3) as a function of center direction (x-axis in each panel), surround direction (y-axis in each panel), center speed (increasing across columns), and surround speed (increasing across rows). Consistent with our hypotheses, the predictions for a neuron encoding retinal-centric motion (v_{center}) show minimal modulation by the surround, appearing as uniform horizontal bands across all conditions (Fig. 4d, top-

left panel, and Fig. S3), and thus align with the surround-insensitive neuron class. In stark contrast, predictions for a neuron encoding relative motion ($v_{\text{center}}^{\text{relative}}$) are substantially modulated by the surround, revealing complex CS interactions that match the surround-dependent neuron class (Fig. 4d, top-right panel and Fig. S2). We will now use the neural predictions of $v_{\text{center}}^{\text{relative}}$ for the representative observer #2 in Shivkumar et al. (2025) to explain the key features of the causal inference model's predictions below (predictions for v_{center} and for the remaining four observers shown in Figs. S2 and S3).

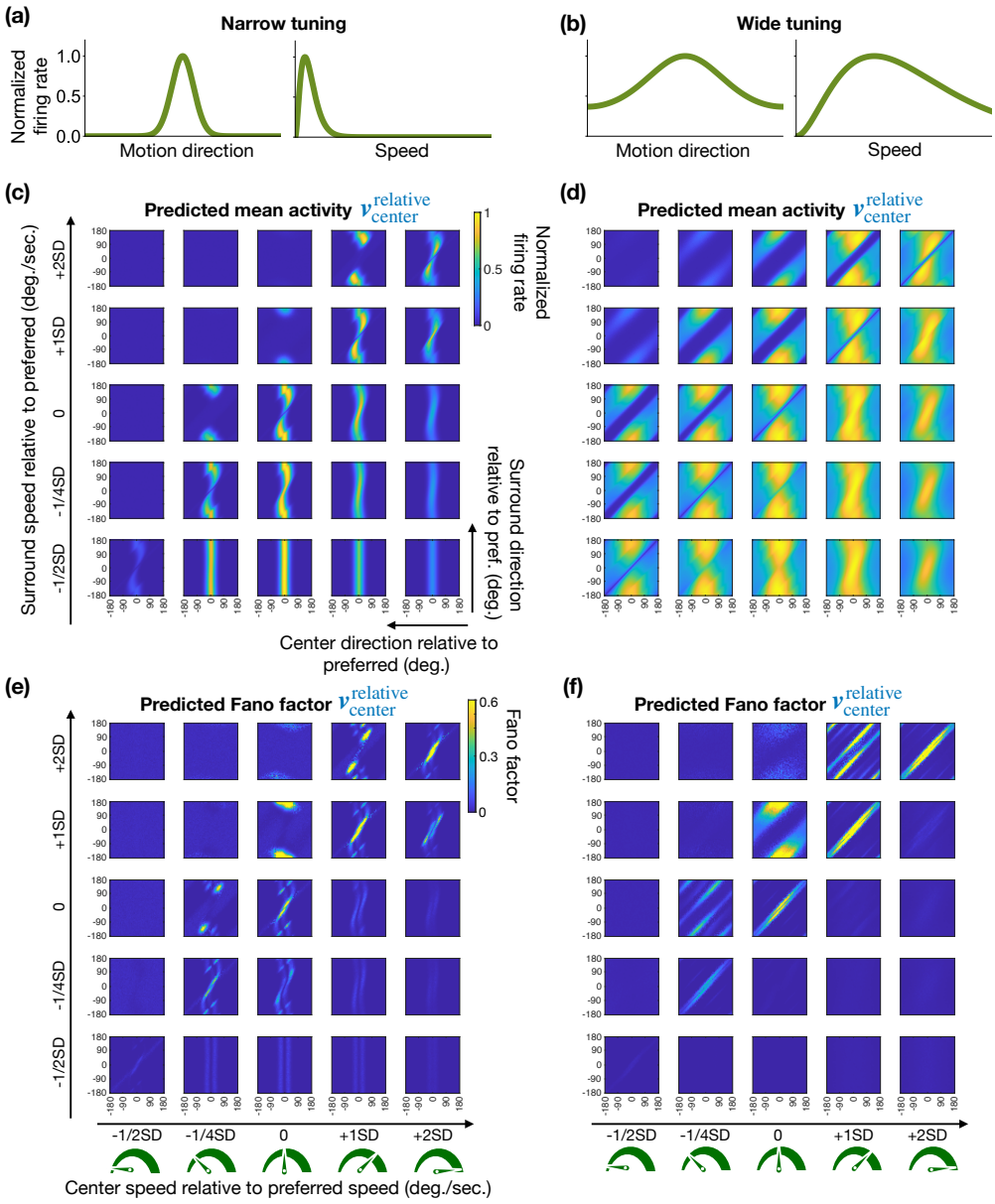

Figure 5: **Predicted 4D tuning curves reveal complex signatures of causal inference.** Predictions are shown for observer #2 in Shivkumar et al. (2025) and two hypothetical neurons with different tuning curve shapes. **(a)-(b)** Schematics of the narrow and wide tuning curves used to generate predictions. **(c)-(d)** Predicted mean firing rates as a function of center direction (x-axis), surround direction (y-axis), center speed (columns), and surround speed (rows). Axes are relative to the neuron's preferred direction and speed (where zero on both axes indicates the preferred direction). The key signature of causal inference is the complex pattern in the tuning curve, which consists of a mixture of different diagonal interactions corresponding to the different reference frames under different motion structures. **(e)-(f)** Predicted Fano factor for the same conditions as (c-d). All firing rates are normalized.

Examining the predictions for $v_{\text{center}}^{\text{relative}}$ in Fig. 5c-d, a key feature emerges: a strong diagonal interaction between center and surround directions. This diagonal pattern is a direct signature of the neuron representing center motion relative to a reference frame influenced by the surround. The more an observer interprets the surround as a reference frame, the stronger this diagonal interaction becomes (Fig. S2). This interaction is strongest under classical motion segmentation, a perceptual outcome where the sur-

round is treated as the sole reference frame for the center's motion (Braddick, 1993; Tadin & Lappin, 2005). Indeed, observer #1 inShivkumar et al. (2025) whose behavior most closely matched this outcome exhibits the sharpest diagonal pattern (see Observer #1 in Fig. S2).

The mechanism for this diagonal is twofold. First, the observer's posterior over $v_{\text{center}}^{\text{relative}}$ peaks at the velocity difference between the center and the inferred reference frame. Second, the predicted firing rate is highest when these posterior peaks align with the neuron's preferred direction and speed. Consequently, a pure surround-relative representation produces the strong diagonal structure seen for Observer 1 (Fig. S2), while a motion representation relative to an intermediate reference frame results in the weaker diagonal interaction seen for observer #2 in Fig. 5.

It is important to note that all predictions in Fig. 5 (and Figs. S2 and S3) are displayed relative to the neuron's preferred speed and direction, making these 4D tuning curves generalizable to any neuron with a similar tuning profile. The specific patterns of CS interaction, however, depend on two key factors. First is the neuron's intrinsic tuning shape, with narrower and broader tuning curves yielding qualitatively different interactions (Fig. 5, left vs. right columns). Second is the observer's perception, as variations in the fitted model parameters lead to substantial differences in the predicted responses across individuals (compare panels in Figs. S2 to S4).

Importantly, causal inference across potential reference frames shapes the diagonal interaction. When an observer infers a mixture of probable motion structures, the model's posterior becomes multimodal (Fig. 3). This mixing of beliefs translates into complex response patterns. For instance, in Fig. 4d (top-right panel), the neural prediction for observer #2 is shaped by a mixture of three stimulus-dependent interpretations: (1) for large CS directional differences, the patches are inferred as independent, producing vertical bands; (2) for moderate differences, an intermediate reference frame is used, creating a curved peak near the diagonal; and (3) for similar directions, integration occurs (i.e., relative motion, $v_{\text{center}}^{\text{relative}}$, is zero), suppressing activity along the diagonal.

A second key characteristic of our Bayesian models is the assumption that people not only form beliefs about motion but also represent the uncertainty in those beliefs. This allows us to generate neural predictions capturing this uncertainty as variability in the firing rates. Thus, instead of computing the mean of the predicted firing rates mapped from posterior samples, we computed the variance of these firing rates across the same four dimensions (CS directions and speeds). Then, we computed the Fano factor and found it also exhibits diagonal structures (Fig. 4d, bottom panels, Fig. 5e & f, and Figs. S2 and S3, bottom rows). This diagonal structure appears because the highest uncertainty in posterior beliefs – and thus the highest predicted firing rate variance – occurs when posteriors are multimodal. This happens when an observer considers multiple causal structures to be simultaneously plausible, for example, when CS directions are similar making both integration and segmentation with respect to an intermediate reference plausible, (see the high Fano factor regions along the diagonal in Fig. 5e & f) or far apart making both independence and segmentation plausible, (see the regions of Fano factor further away from the diagonal in Fig. 5e & f).

It is important to note that this predicted neural variability relates to uncertainty about the latent variables (v_{center} and $v_{\text{center}}^{\text{relative}}$). This does not necessarily imply high behavioral variability, as the final behavioral percept in the model is derived after integrating over all latent uncertainty, often resulting in a unimodal predictive posterior for decision (Shivkumar et al., 2025).

2.5 Comparing causal inference predictions to the empirical literature

The role of area MT in encoding motion remains a topic of debate. While some neurons were found to represent retinal motion and are unaffected by surround motion (Inaba et al., 2011, 2007; Newsome et al., 1988), many others appear to encode motion relative to the surround (Allman et al., 1985a; Born & Tootell, 1992; Huang et al., 2007, 2008; Tanaka et al., 1986; Tzvetanov & Womelsdorf, 2008). Complicating this picture, studies have found a wide variety of additional CS interactions, such as direction-selective suppression, non-selective suppression, and even direction-selective and non-selective facilitation (Allman et al., 1985a; Born, 2000; Born & Bradley, 2005; Tanaka et al., 1986). No single theory yet captures all these diverse CS effects (Born & Bradley, 2005). In this section, we show that our causal inference model predicts this observed diversity. These varied response patterns emerge naturally from the model, depending on whether a given neuron encodes the retinal-centric (v_{center}) or relative-motion ($v_{\text{center}}^{\text{relative}}$) variable, and on the inferred motion causal structure determining the reference frame in which motion is perceived.

We qualitatively compared our model's predictions against published data from three key neurophysiology studies that used conceptually identical CS stimuli: Allman et al. (1985a) (61 single units, 3 owl

monkeys), Tanaka et al. (1986) (105 single units, 4 Japanese monkeys), and Born (2000) (169 single units, 18 owl monkeys). A key challenge in this comparison is that the model's predictions depend on both observer-specific model parameters and neuron-specific tuning curve parameters. Ideally, one would fit the model to an animal's behavior and use tuning curves recorded from the same animal. Since this was not possible, we generated a diverse set of predictions using the five sets of model parameters from the human observers in Shivkumar et al. (2025) and a library of nine hypothetical but plausible tuning curves (Fig. S5a). This library was constructed by combining three distinct tuning widths for both speed and direction (3 speed widths $\times$ 3 direction widths), approximating the 25th, 50th, and 75th percentiles of MT tuning curve parameters measured in DeAngelis and Uka (2003). To evaluate whether these observed MT responses support our model, we compared this large set of predictions against the empirically measured responses. We first categorized the recorded neurons by their functional type of CS interaction. This grouping allowed us to assess whether each distinct neural response pattern could be explained by the causal inference framework. For each empirical CS interaction type, we then identified the best-matching prediction from our library (i.e., the one with the lowest root mean squared error; more details are provided in Methods 4.6).

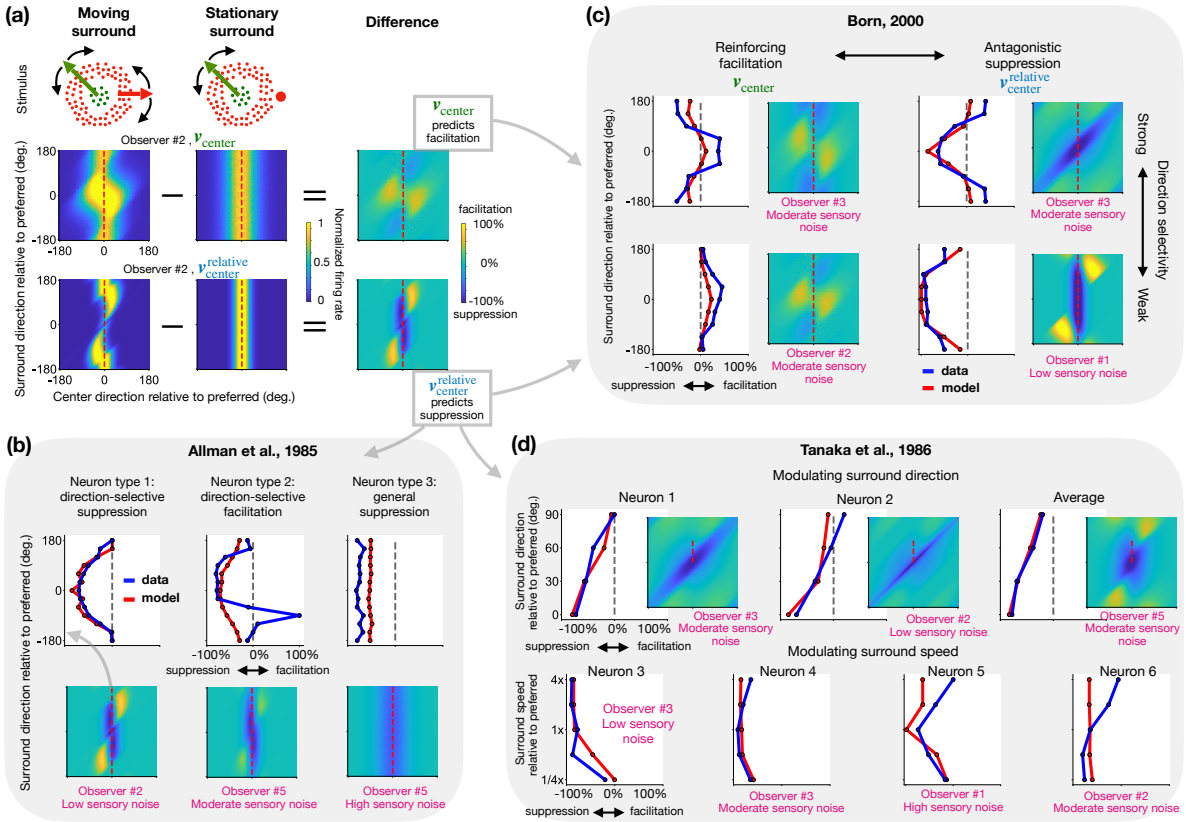

Figure 6: **Model predictions qualitatively capture diverse surround modulation patterns in empirical MT data.** The model's predictions are compared to previously published single-neuron data from monkeys. In all heatmaps and plots, axes are expressed relative to the neuron's preferred direction (labeled as zero). **(a) Method for calculating surround modulation.** Surround modulation is computed by subtracting the predicted response to a stationary surround (middle heatmap) from the response to a moving surround (left heatmap). The top inset illustrates the stationary- and moving-surround stimulus conditions. The model predicts suppression for neurons encoding $v^{\text{relative}}_{\text{center}}$ (e.g., bottom row) and minimal modulation for neurons encoding v_{center} . Facilitation is predicted for v_{center} under moderate to high sensory uncertainty (e.g., top row; see main text). The dashed red lines indicate the cross-section corresponding to the experimental data that the model aims to replicate. **(b) Comparison to three modulation types from Allman et al. (1985a).** We performed qualitative model fitting; see main text and Methods 4.6 for details. Model predictions assuming $v^{\text{relative}}_{\text{center}}$ coding (red lines) are overlaid on average empirical data (blue lines). The plot shows modulation (facilitation/suppression, x-axis) as a function of surround direction (y-axis), while the center patch moves in the neuron's preferred direction. The full 2D prediction is shown in the heatmap, with the dashed red line indicating the plotted cross-section which corresponds to the experimental data we aim to match. Pink text indicates the best-matching observer model and sensory uncertainty level ("low," "moderate," and "high" correspond to the fitted noise value from Shivkumar et al. (2025), 10x, and 100x that value). **(c) Comparison to four functional classes from Born (2000).** The format is the same as in (b). The model accounts for the different neural classes by assuming different latent variable encoding. Predictions based on v_{center} under moderate uncertainty capture the observed facilitation effects (left column). In contrast, predictions based on $v^{\text{relative}}_{\text{center}}$ capture the observed suppression effects (right column). **(d) Comparison to surround modulation from Tanaka et al. (1986).** Following the same format, model predictions for $v^{\text{relative}}_{\text{center}}$ (red lines) are compared to example neurons and population data where surround properties were varied: in the top row, direction was varied while speed was fixed at the neuron's preferred speed; in the bottom row, speed was varied while direction was fixed at the neuron's preferred direction. The heatmap and its red dashed cross-section correspond to the direction-varying data in the top row. A corresponding heatmap for the speed-varying data (bottom row) is not shown, as it would represent a different 3D parameter space that also contains surround speed.

2.5.1 Neurons unaffected by surround motion

A subset of MT neurons shows no measurable modulation by surround motion. This response profile aligns with classical accounts of MT as representing motion in retinal coordinates (Inaba et al., 2011, 2007; Newsome et al., 1988). This pattern is well-captured by the causal inference model when a neuron is assumed to encode the latent variable for retinal-centric motion (v_{center}), which, in most cases, remains unaffected by the surround (Fig. 4d, left column, and Fig. S3). Allman et al. (1985a) found that approximately 8% of MT neurons exhibited this response profile, responding selectively to center motion while remaining unaffected by motion in the surround.

2.5.2 Neurons exhibiting direction-selective antagonistic suppression

The most common class of MT neurons exhibits strong suppression when the center and surround move in the same direction, with suppression decreasing as their directions diverge (Allman et al., 1985a; Born, 2000; Tanaka et al., 1986). The causal inference model robustly replicates this profile when neurons are assumed to encode the relative motion variable ($v_{\text{center}}^{\text{relative}}$). Mechanistically, this suppression arises because the model for relative motion ($v_{\text{center}}^{\text{relative}}$) predicts a diagonal CS interaction for the moving surround condition, where activity is lowest along the diagonal (i.e., for same-direction motion). When the response to a stationary surround is subtracted, this results in strong, direction-selective suppression (Fig. 6a, bottom row). This pattern emerges across all observer parameters and tuning curve shapes (Fig. S4, Fig. S5b), and the model's predictions closely align with the measured neural data from multiple studies (Fig. 6b, left; Fig. 6c, right column; Fig. 5d, top). Also, the closer the inferred reference frame to the surround motion (i.e., surround reference causal motion structure, see Fig. 2e) the strongest the diagonal CS interaction is in the neural predictions.

Although this modulation is consistent with relative motion coding, it is not sufficient proof. The single cross-section of the 2D tuning space measured in most studies (e.g., Fig. 6, red cross-sections) is insufficient to distinguish a true diagonal interaction, which is the key signature of relative motion, from other forms of surround suppression. Verifying such a representation would require testing multiple center and surround motion directions in combination.

2.5.3 Neurons showing non-directional surround suppression

Approximately 30% of MT neurons in Allman et al. (1985a) (and a smaller fraction in other studies) exhibited suppression from the surround regardless of its direction (Born, 2000; Tanaka et al., 1986). Our model captures this response pattern for a neuron encoding $v_{\text{center}}^{\text{relative}}$ when sensory uncertainty of the observer is assumed to be moderate or high (Fig. 6b, right). High uncertainty broadens the posterior distributions over $v_{\text{center}}^{\text{relative}}$, which averages out directional effects and leads to a more uniform, direction-independent suppression, consistent with these neurons' responses. In this context, "low", "moderate", and "high" uncertainty correspond to the sensory noise level fitted to behavioral data in Shivkumar et al. (2025), and to levels 10x and 100x greater than that fitted value, respectively

2.5.4 Neurons with reinforcing surround modulation

Our model also captures the "reinforcing modulation" observed in a subset of MT neurons, where responses increase when the surround moves in the same direction as the center (Born, 2000). This facilitation is predicted for two of our five observer models under specific conditions: that the neuron encodes the retinal-centric variable (v_{center}) and that sensory uncertainty is moderate (10x the empirically fitted value) (Fig. 6c, left column).

2.5.5 Neurons showing surround facilitation orthogonal to their preferred direction

A small fraction of neurons (8%) in Allman et al. (1985a) were facilitated by surround motion orthogonal to their preferred direction (Fig. 6b, middle column). Our model does not reproduce this effect under any parameter settings, suggesting these neurons may perform a different computation than the one described by our framework. Given their small proportion and the lack of replication in subsequent studies (Born, 2000; Tanaka et al., 1986), we do not view this discrepancy as a significant challenge to the model's overall explanatory power.

2.5.6 Neurons showing speed-dependent suppression

Tanaka et al. (1986) reported that surround suppression in MT is modulated by surround speed in addition to direction. Our model successfully accounts for the primary pattern they observed, where higher surround speeds led to stronger suppression (Fig. 6d, neurons 3 and 4). This effect is robustly predicted for a neuron encoding $v_{\text{center}}^{\text{relative}}$, as its activity consistently decreases with increasing surround speed across most parameter settings (Fig. S5c). In contrast, the model does not account for the opposite pattern – stronger suppression at lower surround speeds – which was observed in two other neurons from the same study (Fig. 6d, neurons 5 and 6). However, as this latter effect has not been widely replicated, we do not consider it a major challenge to the model.

2.5.7 Neurons with minimal shifts in tuning under surround manipulation

To directly test whether neurons encode relative motion with respect to the surround (i.e., surround reference causal motion structure), Born (2000) recorded from neurons while jointly varying center and surround direction or speed. If neurons truly represent center motion relative to the surround, we would expect the center tuning to shift accordingly. However, the authors found little to no tuning curve shift, merely an overall response modulation. While this finding seems to challenge pure relative-motion models, the small number of neurons in these experiments (3 for direction, 9 for speed) and sparse stimulus sampling limit strong conclusions.

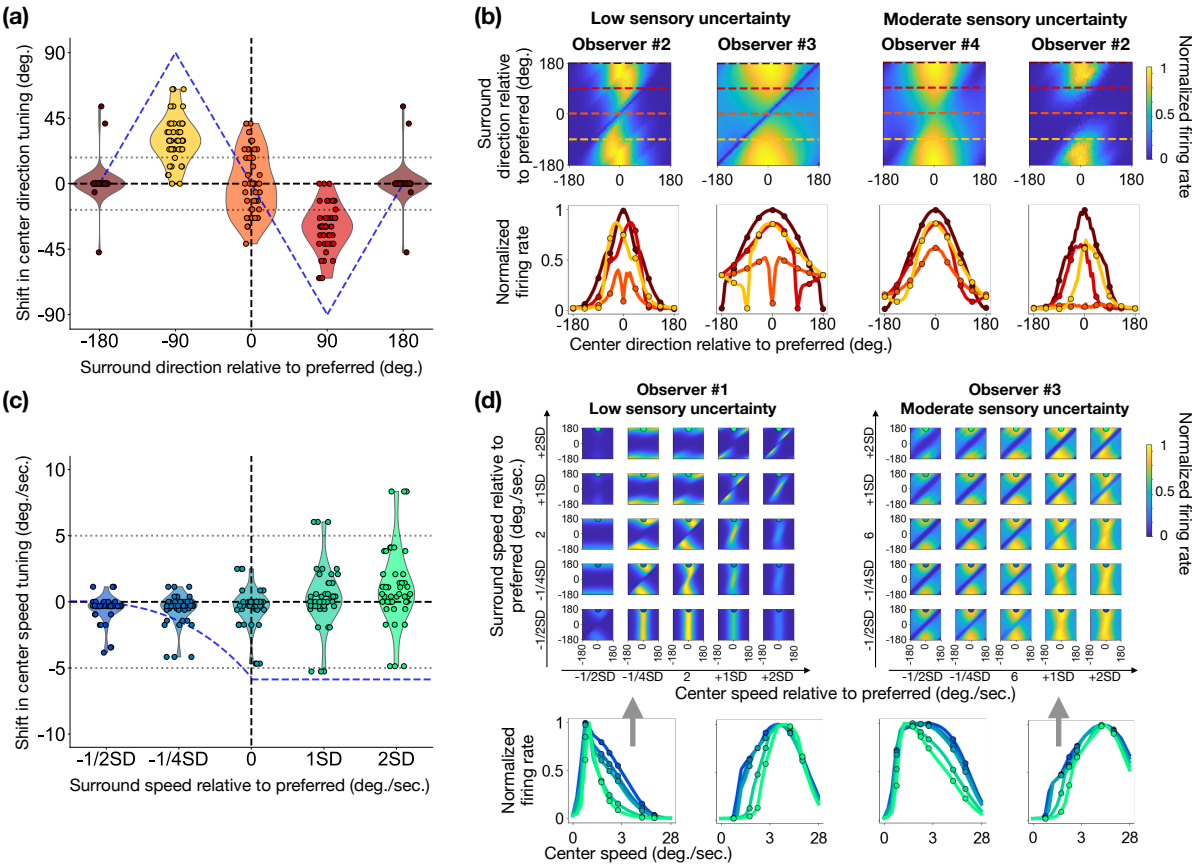

Figure 7: **The model predicts when tuning curve shifts due to surround motion are large and when they are minimal.** In all panels, axes are plotted relative to the neuron's preferred direction and speed (labeled as zero). Specific points on the tuning curves correspond to the stimulus conditions tested in Born (2000). **(a)** Predicted shifts in center direction tuning (y-axis) as a function of four different surround directions (x-axis). Violin plots show the distribution of shifts across 45 combinations of observer parameters and neuron tuning profiles (5 observers $\times$ 9 tuning curves). The predicted shifts for some of the observers falls within the empirically observed range from Born (2000) (dotted horizontal lines) and most shift are below predicted by a pure surround-relative model under the surround reference structure (blue dashed line). **(b)** Example 2D tuning curves (top) and their 1D cross-sections (bottom) for direction. The colored dashed lines in the heatmaps indicate the surround directions for which the center tuning curves below are plotted. Predictions are shown for low (fitted values) and moderate (10x fitted values) levels of sensory uncertainty. The shift of the peaks are smaller for moderate sensory noise, consistent with the points around y=0 in (a) **(c)** Predicted shifts in center speed tuning, analogous to (a). The x-axis represents surround speed scaled by the tuning width (SD). The predicted shifts are smaller than what is predicted by a pure relative-motion model (blue dashed line) and mostly fall within the empirically measured range (dotted horizontal lines). **(d)** Example 4D tuning curves and 1D cross-sections for speed. The top row displays two examples of the predicted 4D tuning curves, showing firing rate as a function of center speed (columns) and surround speed (rows). The bottom row shows four example 1D center speed tuning curves. The first and last of these are cross-sections corresponding to the 4D predictions shown directly above them, while the two middle curves are additional examples.

Our causal inference framework provides specific predictions for when tuning shifts should be large and when they should be minimal or even zero, thereby accounting for the seemingly contradictory findings in the literature. In the model, the magnitude of the shift depends on the inferred reference frame. When the model infers an intermediate reference frame lying between the center and surround, it predicts minimal shifts with direction-independent response modulation observed by Born (2000) (see observers #3 and #4 in Fig. 7b). As shown in Fig. 7a for direction tuning and Fig. 7c for speed tuning,

the model's predictions for many parameter combinations (points) fall within the small range empirically reported (dotted gray lines). Under certain conditions, such as moderate sensory uncertainty (10x the fitted value), the model predicts near-zero shifts (Fig. 7b and d). Interestingly, for center speed tuning, the model can even predict shifts in the opposite direction of a simple surround-relative prediction (first and third panels in Fig. 7d, bottom row), precisely because motion is referenced to the inferred intermediate frame.

Thus, the data from Born (2000) do not rule out relative motion coding in MT. They may challenge the assumption of a fixed, pure surround-based reference frame. However, even this conclusion is limited by the small number of neurons and sparse stimulus sampling used in the experiments. Importantly, our model provides a way to resolve this ambiguity. It generates specific, quantitative predictions for the expected shift based on an observer's behavioral parameters and their neuron's baseline tuning. In fact, for some observers from Shivkumar et al. (2025), the minimal shifts reported by Born (2000) are exactly what our Bayesian causal inference model would predict. This highlights the need for such a principled model to guide new, densely-sampled experiments that can definitively distinguish among competing theories of motion encoding.

3 Discussion

In this study, we demonstrated that a normative Bayesian causal inference model of motion perception can qualitatively account for the diverse center-surround (CS) interaction effects reported in the primate middle temporal area (MT). By deriving comprehensive predictions from a model fitted to human psychophysical data, we showed that the diverse, and often seemingly contradictory, response properties of MT neurons can be understood as principled manifestations of a single underlying computation: inferring the latent causal structure of the visual world. This framework unifies disparate findings in motion perception and neurophysiology, reinterprets long-standing debates about neural coding in MT, and provides a clear, theory-driven path for future experiments.

3.1 A single model for diverse neural effects

The function of CS interactions in area MT has been a puzzle, primarily due to the diversity of the observed phenomena. Our framework reframes this diversity not as a catalogue of many distinct neuron types with fixed properties, but as the output of a single inferential process: Bayesian causal inference.

Classic neurophysiological studies categorized the variety of surround modulations. A small subset of MT neurons (∼8%) shows no surround modulation at all, consistent with the traditional view of MT encoding motion in purely retinal coordinates (Inaba et al., 2011, 2007; Newsome et al., 1988). Our model naturally accounts for these neurons through the retinal-centric latent variable (v_{center}). The most common class of neurons exhibits direction-selective antagonistic suppression, strongest when center and surround move together (Allman et al., 1985a; Born, 2000; Born & Bradley, 2005; Born & Tootell, 1992; Tanaka et al., 1986). Our model robustly reproduces this effect as a signature of relative motion coding ($v_{\text{center}}^{\text{relative}}$) with respect to the inferred reference frame.

Crucially, the model also explains phenomena that have challenged simpler descriptive or mechanistic models. For instance, reinforcing modulation (facilitation), where responses are enhanced by same-direction surrounds, has been difficult to reconcile with a purely suppressive mechanism like classic divisive normalization (Coen-Cagli et al., 2015). Our framework predicts this exact effect for neurons encoding retinal-centric motion (v_{center}) under conditions of moderate to high sensory uncertainty, providing a normative explanation for its emergence. Our model also resolves the key finding from Born (2000) that center tuning curves modulate their gain but do not shift their peak in response to surround motion – a result that poses a direct challenge to any pure surround-relative coding scheme. The causal inference model predicts this when there is larger uncertainty about the causal structures, and hence ambiguity about the correct reference frame. Because the inferred reference is often intermediate – not locked to the surround – the posterior belief about the center's motion is modulated while its peak shifts only minimally, in line with empirical observations.

Our framework not only reinterprets past findings but also generates new, testable predictions. The model makes directly testable predictions for surround modulation. Facilitation, for instance, is predicted for neurons encoding retinal-centric motion (v_{center}) under high sensory uncertainty. In contrast, suppression is the hallmark of neurons encoding relative motion ($v_{\text{center}}^{\text{relative}}$) with respect to a surround-influenced reference frame. Furthermore, the model explains how direction-dependence arises from sensory noise: high uncertainty leads to broad, direction-independent modulation, whereas low uncertainty produces

sharply tuned, direction-specific effects. Furthermore, our linking method generates highly specific predictions for neural CS interactions. It directly links an individual's perception (via behaviorally fitted model parameters) with their neural encoding (via their recorded baseline tuning profile) to produce these specific predictions. By making explicit predictions for tuning curves and firing rate variability across a broad range of CS stimuli, our work moves beyond explaining prior studies – which are insufficient to rule out competing relative motion models – and provides a clear path forward for targeted new experiments.

3.2 Reinterpreting MT: from motion filters to causal inference

Our findings suggest a conceptual shift in our understanding of MT's role. In traditional models of neural activity, CS processing in motion perception arises from a cascade of linear and nonlinear computations that integrate local motion signals into the representation of a more global motion of the visual scene (Adelson & Bergen, 1985; Albright, 1984; Barth & Watson, 2000; Movshon et al., 1985; Rust et al., 2006; Simoncelli & Heeger, 1998; Wilson et al., 1992). In these models, V1 simple cells function as linear space-time filters, detecting only local motion energy within their passband, without encoding the global motion in the scene. The role of V1 complex cells remains debated: some models suggest they compute local averages of V1 activity to represent phase-insensitive motion energy (Adelson & Bergen, 1985; Albright, 1984; Movshon et al., 1985; Rust et al., 2006; Simoncelli & Heeger, 1998), while others propose they already signal global motion direction by selectively responding to endpoints of long contours, thereby avoiding the aperture problem (Pack et al., 2003a; Pack et al., 2003b; Zarei Eskikand et al., 2016). For MT, most models posit that its cells integrate local motion energy signals from V1 and apply nonlinear computations, such as divisive normalization and half squaring, to compute the global motion in the visual scene (Adelson & Bergen, 1985; Albright, 1984; Movshon et al., 1985; Rust et al., 2006; Simoncelli & Heeger, 1998). Alternative theories argue that nonlinear computations, such as texture boundary motion and feature extraction, occur earlier in V1 and V2, where global motion in the visual scene may already be represented. In this framework, MT primarily pools these inputs and reduces noise (Barth & Watson, 2000; Wilson et al., 1992; Zarei Eskikand et al., 2019, 2020, 2016). While these traditional models successfully explain many aspects of neural activity along the V1–MT pathways, these models are fundamentally descriptive; they explain how a computation might be implemented but not why it takes the particular form it does. They struggle to explain the functional purpose behind the diverse CS effects and do not clarify how the brain determines the appropriate reference frame for motion integration or segmentation. Consequently, these models fail to explain CS effects in perception.

In contrast, our model assumes that neural activity represents posterior beliefs about latent variables in a generative model of retinal motion. These latent variables represent not only the velocities of moving objects but also the reference frames within which the velocities are represented, both of which are inferred from sensory input. The resulting relationships between the 4-dimensional stimulus (center and surround motion, each with speed and direction) and neural responses are complex and difficult to capture with a feedforward model. Furthermore, these relationships are dynamically modulated by contextual variables that influence the brain's belief about causal structure, such as stimulus uncertainty (number of elements and/or contrast) and the distance between center and surround. These variables help account for reported differences between studies.

Therefore, our results reframe MT not merely as a sophisticated filter for integrating V1 signals using simple, fixed rules, but as an inferential system that deduces the relationships between moving objects to form coherent perceptual representations. This perspective provides the missing functional role ("why") for the complex repertoire of CS interactions, suggesting they emerge as a solution to a causal inference problem.

3.3 Bridging perceptual phenomena and neural circuits

A central contribution of our work is bridging the gap between perceptual studies of motion and their neural underpinnings. Decades of psychophysical research have established two dominant effects in motion perception: integration and segmentation (Braddick, 1993; Dakin & Mareschal, 2000; Penaloza et al., 2024; Tadin & Lappin, 2005; Tadin et al., 2003; Zarei Eskikand et al., 2024, 2019). Integration is thought to improve the signal-to-noise ratio, especially in low-contrast conditions (Britten & Heuer, 1999; Britten et al., 1993; Tadin et al., 2003), while segmentation enhances the perception of relative motion, aiding in object tracking and parsing scenes relative to a reference frame (Huang et al., 2007; Johansson, 1950; Tadin et al., 2019). MT neurons seem to exhibit shifting responses between integration and segmentation in a dynamic, stimulus-dependent way (Huang et al., 2007, 2008).

Our results provide a unifying explanation for these observations, both at the perceptual and at the neural levels. The causal inference model casts both integration and segmentation as outcomes of probabilistic inference over latent motion structures (Penaloza et al., 2024; Shivkumar et al., 2025; Yang et al., 2021). We show that MT neurons can exhibit response properties consistent with either computation by representing posterior beliefs over latent variables that flexibly change with stimulus context. This offers a principled account of how the same population of neurons can support both perceptual states (integration and segmentation) and why certain CS stimuli elicit ambiguous or mixed neural responses.

At the circuit level, divisive normalization is often proposed as a canonical computation for CS effects (Carandini & Heeger, 2012; Coen-Cagli et al., 2015). While simple divisive normalization models cannot capture the full range of effects predicted by our framework, extended variants that include multiplicative interactions within their normalization pool may provide a mechanistic approximation (Penaloza et al., 2023). Several mechanistic models have been proposed to explain the integration and segmentation properties of MT neurons (Kim & Wilson, 1997; Qian et al., 1994a, 1994b; Wilson & Kim, 1994; Zarei Eskikand et al., 2019, 2020). These models generally fall into three categories: (1) those attributing segmentation to inhibitory interactions between direction-tuned neurons (Kim & Wilson, 1997; Qian et al., 1994a, 1994b; Wilson & Kim, 1994), (2) those proposing distinct populations of neurons specialized for integration versus segmentation (Beck & Neumann, 2011; Zarei Eskikand et al., 2020), and (3) those positing a single, adaptive population that dynamically adjusts its CS interactions based on stimulus context (Zarei Eskikand et al., 2019). While these models can account for many CS interactions, they struggle to explain cases that contradict simple relative motion encoding (e.g., Born (2000) and Tanaka et al. (1986)). It also remains unclear if circuit models based on divisive normalization or stabilized supralinear networks (Rubin et al., 2015) can support the probabilistic computations required by Bayesian causal inference. Future work could explore this connection, using the detailed predictions from our framework as a clear computational target for mechanistic modeling.

3.4 Limitations and Future Directions

This study provides a qualitative, proof-of-principle demonstration that a normative model of perception can account for a wide range of neurophysiological data. Two main limitations, however, point toward crucial avenues for future research.

First, linking Bayesian models to neural activity requires an assumption about the neural code. We used the neural sampling hypothesis (e.g., Buesing et al., 2011; Fiser et al., 2010; Haefner et al., 2016; Hoyer and Hyvärinen, 2002; Orbán et al., 2016), which belongs to a broad class of Linear Distributional Codes (LDCs) that also includes Distributed Distributional Codes (DDCs) (Lange & Haefner, 2022). While our predictions for mean tuning curves are generalizable across LDCs (see Methods 4.3 and Lengyel et al., 2023), our predictions for response variability are specific to sampling-based codes. The broader debate over whether the brain uses LDCs or other encoding schemes, like Probabilistic Population Codes (PPCs), is ongoing (Beck et al., 2012, 2020; Fiser et al., 2010; Haefner et al., 2024; Lange et al., 2022; Ma et al., 2006; Pouget et al., 2013; Tajima et al., 2016; Ujfalussy & Orbán, 2022; Vértes & Sahani, 2018, 2019). A critical next step is to derive predictions for non-LDC codes like PPCs, which would generate directly competing and testable hypotheses.

Second, our comparison to historical data is qualitative. While we show that the patterns of neural activity are consistent with our model, rigorous quantitative model comparison against alternatives (like extended divisive normalization models) is required. This would be best achieved by fitting the model parameters to an animal's behavior while simultaneously recording from MT neurons. Our framework is ideally suited to guide such work, as it can generate predictions for any stimulus, allowing researchers to design optimized experiments that can maximally differentiate between competing models (i.e., controversial stimuli; Golan et al., 2020).

3.5 Conclusion

This study establishes Bayesian causal inference as a unifying normative principle for center-surround interactions in motion. Our framework parsimoniously explains a wide array of disparate findings – from surround suppression and facilitation to minimal tuning curve shifts in response to surround motion – as the principled outputs of a single computation over latent reference frames. By reinterpreting classic neurophysiological data and bridging the long-standing gap between models of perception and neural activity, this work reframes our understanding of MT's role from that of a motion filter to that of an inference engine. This provides a clear foundation for designing new, theory-driven experiments to dissect the circuit-level implementation of these abstract cognitive computations.

4 Methods

4.1 The behavioral task and data source

The Bayesian causal inference model was constrained using behavioral data from five observers in Experiment 1 of Shivkumar et al. (2025). In that study, observers used a dial to report their perceived motion direction of a central patch of green dots surrounded by a peripheral patch of red dots (Fig. 2a). The surrounding dots were either stationary or moved horizontally (0 or 180°) at a speed of 1 deg./sec. Stimuli were displayed at an eccentricity of 5° in the periphery. On each trial, the direction of the central patch was randomly selected from the set 0°, ±2.5°, ±5°, ±10°, ±20°, ±45°. Both the center and surround patches shared a common horizontal velocity (0°or 180°), resulting in the center's velocity relative to the surround being ±90°, depending on the direction of the center patch. After a fixation period of 0.5 seconds, the stimulus appeared and oscillated back and forth for 1.5 cycles. The patch envelopes moved at a constant velocity and reversed direction after 1.5 seconds, following a square wave velocity profile with a period of 3 seconds. This oscillatory motion ensured that the envelopes remained within a fixed area on the screen. For a detailed description of the stimuli, task, and design, see Shivkumar et al. (2025).

When the surround was stationary, reported directions closely followed the true motion of the center patch as displayed on the screen. However, when the surround patch was moving, observers' responses deviated systematically from the true motion of the center patch. For small center directions, responses were biased towards zero degrees (the surround direction), consistent with observers integrating the center and surround velocities. For larger center directions, responses were biased towards 90°, indicating that observers perceived the relative velocity between the center and surround. Additional findings from this experiment, along with further analyses, are detailed in Shivkumar et al. (2025).

4.2 The Bayesian causal inference model

The Bayesian causal inference model described in Shivkumar et al. (2025) was used to derive posterior beliefs for center-surround (CS) motion stimuli. In this generative model, the retinal velocity of each visual element (i.e., the center and surround patches) is the sum of the reference frame's velocity to which the element belongs and its relative velocity with respect to that reference frame. The prior over relative velocity is modeled as a mixture of a delta function centered at zero and a zero-mean Gaussian distribution. This prior naturally enables the model to infer whether elements are stationary relative to their reference frame or moving within their reference frame. A detailed description of the full generative model, inference equations, and model fitting procedures can be found in Shivkumar et al. (2025).

Observers allocated most of the posterior probability mass to the delta component, reflecting the brain's strong expectation that relative motion in the world is exactly zero, rather than merely slow. Moreover, observers attributed significant probability mass to only four out of the 12 possible causal motion structures:

1. Motion integration: Integrating center and surround velocities, leading to the perception of cue-combined motion velocity (Fig. 2c). The posterior probability of this structure was highest when the center and surround moved at the same velocity and decreased as the separation between center and surround velocities increased (Fig. S1c and d).

2. Surround reference: Perceiving center motion in the reference frame defined by the surround (i.e., pure relative motion or motion segmentation, see Fig. 2d). This structure was prominent for intermediate differences between center and surround motion directions for a subset of observers (Fig. S1e and f).

3. Center reference: Perceiving surround motion in the reference frame of the center. In this case, the perception of center motion largely aligned with its retinal motion (Fig. 2e), with only a few observers attributing significant probability mass to this structure (Fig. 2g and h).

4. Intermediate reference: Perceiving both center and surround motion relative to a reference frame intermediate between the two (Fig. 2f). This structure accounted for incomplete subtraction of the surround velocity from the center velocity at larger differences between center and surround motion directions and was prominent among most observers (Fig. 2i and j).

Notably, none of the observers attributed any probability mass to the possibility that the center and surround patches were independent, indicating that they did not perceive the two patches as causally unrelated (Fig. 2g and Fig. S1 k and l).

Interestingly, motion structures with intermediate reference frames have higher posterior probabilities than other structures at large separations. This is due to the Gaussian slow-speed component in the prior, which favors smaller relative velocities under an intermediate reference frame, outweighing the mass concentrated in the delta component of the prior. Additional insights from fitting the model to observers' responses, along with further analyses, can be found in Shivkumar et al. (2025).

We adopted a simplified version of the Bayesian causal inference model in (Shivkumar et al., 2025) which includes 10 parameters: (1) two sensory uncertainty parameters associated with the center and surround patches' observed velocities, (2) one computational noise parameter, (3) three mixture prior parameters corresponding to the reference, center, and surround velocities, (4) three parameters for the prior widths of the zero mean Gaussian slow speed priors over the reference, center, and surround velocities, and (5) one parameter representing the probability of grouping the center, and the surround patches into a single causal structure. Detailed descriptions of these parameters and their fitted values can be found in Shivkumar et al. (2025).

4.3 Generating neural predictions from Bayesian models

We generated neural predictions from the model's posterior beliefs assuming that they are represented in the single neuron activity via Neural Sampling (Fiser et al., 2010; Hoyer & Hyvärinen, 2002).

We used the following algorithm to link a posterior distribution to mean firing rates and variability in the firing rates:

1. Draw samples, S, from the posterior distribution over latent variables, $p^{v^{\text{relative}}_{\text{center}}}$, $p^{v_{\text{center}}}$, and $p^{v_{\text{reference}}}$:

$$S \sim p^{(k)} \tag{1}$$

   where $k = \{v^{\text{relative}}_{\text{center}}, v_{\text{center}}, v_{\text{reference}}\}$, and $S = \{(u, \theta) \mid u \in U, \theta \in \Theta\}$ containing pairs of speeds, u, and directions, θ (U and Θ represent the set of all possible speeds and directions, respectively).

2. Link each sample, containing a combination of a speed and a direction value, in S with a predicted instantaneous firing rate using the measured (or hypothesized) tuning curve of the neuron in response to stimuli with stationary surround:

$$R = f_{\text{speed}}(S_{\text{speed}}) * f_{\text{direction}}(S_{\text{direction}}) \tag{2}$$

   where f_{speed} and $f_{\text{direction}}$ denote the speed and direction tuning functions of the neuron, respectively. $S_{\text{speed}} = \{u \mid (u, \theta) \in S\}$ and $S_{\text{direction}} = \{\theta \mid (u, \theta) \in S\}$ while $*$ denotes element wise multiplication. R represents the distribution over the predicted instantaneous firing rates corresponding to the pairs of speed and direction samples.

3. Take the mean or the variance to compute the predicted firing rate and variability in the firing rate.

$$\bar{R} = \frac{1}{N} \sum_{i}^{N} R_i \tag{3}$$

$$\bar{R} = \frac{1}{N-1} \sum_{i}^{N} \left(R_i - \bar{R}\right)^2 \tag{4}$$

   where N is the number of samples in S.

4. Repeat this process for each CS stimulus.

We used the algorithm above to generate the predicted tuning curves and variability in the firing rates for all plots in this study.

4.4 Predicted tuning curves holds for all Linear Distributional Codes

The method described above can be regarded as a special case of the approach introduced in Lengyel et al. (2023). In that work, Lengyel et al. (2023) developed a method for testing Bayesian models using neural activity, which applies to all Linear Distributional Code (LDC; Lange and Haefner, 2022) based encoding models. This approach also enables researchers to generate predicted tuning curves for single neurons that are consistent across all LDCs (Lange and Haefner, 2022). LDCs encompass a broad class

of encoding models, including schemes within two of the three principal families of hypotheses about neural encoding: Neural Sampling Codes (Fiser et al., 2010; Hoyer & Hyvärinen, 2002) and Distributed Distributional Codes (DDCs; Vértes and Sahani, 2019). Consequently, the neural predictions derived from this method hold for any Neural Sampling or DDC-based encoding used by the brain to represent probability distributions. However, the method does not apply to Probabilistic Population Codes (PPCs; Ma et al., 2006) or other percentile codes (see Lengyel et al., 2023 for details).

In essence, the method identifies predicted, "test" stimuli whose posteriors can be expressed as mixtures of the posteriors of other, "baseline" stimuli, referred to as predictor stimuli. It demonstrates that, under the assumption of any LDC-based encoding, these predicted test stimuli are expected to elicit neural responses that are mixtures of the neural responses to the predictor baseline stimuli. Crucially, the mixture weights used to generate the neural responses correspond to the same mixture weights used to match the posteriors for the predicted test stimuli (see the derivation in Lengyel et al., 2023).

The method we used in this work to generate tuning curves (i.e., mean firing rates) is mathematically equivalent to the approach in Lengyel et al. (2023), under the following assumptions. To predict neural activity in response to a CS motion stimulus with a moving surround (the predicted test stimulus), we first compute the posterior belief for this stimulus using a Bayesian model with fixed or fitted parameters (e.g., $p^{(k)}$ in step 1 above). Next, we draw samples from this posterior (e.g., S in step 1 above) and use these as predictor baseline CS motion stimuli, with a stationary surround. We assume that the posteriors corresponding to these predictor baseline stimuli, i, are delta posteriors:

$$p_i^{\text{predictor}} = \delta(u_i, \theta_i) \tag{5}$$

Then, the posterior for the predicted test stimulus is expressed as a mixture of the predictor baseline stimuli's posteriors:

$$p^{(k)} \approx \sum_i w_i p_i^{\text{predictor}} \tag{6}$$

The optimal match is achieved when the mixture weights, w, are equal to the probability density of the posterior samples drawn for the predicted test stimulus:

$$w_i = p^{(k)}(u_i, \theta_i) \tag{7}$$

Using measured or hypothesized tuning curves for baseline CS motion stimuli with a stationary surround, we look up the firing rate for each predictor baseline stimulus, i:

$$R_i = f_{\text{speed}}(u_i) * f_{\text{direction}}(\theta_i) \tag{8}$$

where, similar to step 2 above, f_{speed}, $f_{\text{direction}}$, u_i, and θ_i denote the speed and direction tuning functions of the neuron, and the speed and direction samples drawn from the posterior, respectively. For a large number of samples drawn from the posterior, the mixture of these neural responses to the predictor baseline stimuli weighted by the mixture weights, w, matches the predicted mean firing rate, $\bar{R}$, computed in step 3 of the algorithm described above:

$$\bar{R} \approx \sum_i w_i R_i \tag{9}$$

Since the weighted sum $\sum_i w_i R_i$ converges to the integral $\int p^{(k)}(u, \theta) \cdot R(u, \theta)\, d(u, \theta)$, which defines the expected firing rate.

Therefore, the predicted tuning curves generated with the algorithm described above hold for any LDC, whether Neural Sampling or DDC-based. However, the method does not apply to other encoding schemes such as PPCs. Additionally, the predicted variability in neural responses applies only to Neural Sampling codes.

4.5 Analyzing the predicted neural activity

The CS stimulus described in 4.1 has four important properties that modulates posterior beliefs in the Bayesian causal inference model: (1) the direction of the center, (2) the direction of the surround, (3) the speed of the center, and (4) the speed of the surround patches. Therefore, we plotted the predicted tuning curves and variability as a function of these four stimulus dimensions (see Fig. 5, and Figs. S2 and S3). We then demonstrate that these predictions are different for different observers via the model parameter values fitted to their behavior (see Figs. S2 and S3). Finally, we demonstrate that these predictions are different for different neurons via their different tuning curves (see Figs. 5 and S5).

In order to investigate the variety of possible neural CS interactions predicted by Bayesian causal inference, we generated neural predictions using a wide range of possible model parameters besides the fitted parameters for the $v_{\text{center}}^{\text{relative}}$ latent variable. In short, we used the five sets of parameters fitted to the five observers in Shivkumar et al. (2025) and sampled the model parameter values around the fitted values to generate 210 different predictions for each of the 9 neuron tuning profiles and for each of the 5-5 CS speeds and 60-60 CS directions. This resulted in 47250 different prediction maps across CS directions (210 (model parameters) x 9 (tuning profiles) x 5 (center speed) x 5 (surround speed) = 47250). The tuning profiles included three levels of spreads for the speed and direction tunings from sharp to wide tuning curves (i.e., 3 speed tunings x 3 direction tunings = 9 tuning profiles, see Fig. S5). The center of the tuning curves always represented the preferred direction or speed of the neuron. Fig. S4 shows the types of CS interactions predicted by the causal inference model assuming $v_{\text{center}}^{\text{relative}}$ is encoded.

The main signature of our causal inference model encoding relative motion is the diagonal interaction between the CS directions. The prediction maps in Fig. S4 are ordered from strongest (upper left corner) to weakest (lower right corner) diagonal interactions. We measured diagonal interaction as how symmetrical the prediction was to the horizontal cross-section line in the middle (i.e. when surround direction equals 0 with varying center directions). The lower the symmetry scores, the stronger the CS interaction (see the scores on top of each prediction map). This diagonal CS interaction reflects the fact that these predictions assume that the neuron represents the center motion relative to the inferred reference frame's motion. The more observers infer the surround patch as a reference frame and thus perceive the center motion relative to the surround, the stronger the diagonal CS interaction will be predicted for mean neural responses. Importantly, while this library of 47250 maps was generated to explore the model's full predictive range (Fig. S4), it was not used for the subsequent comparison with empirical data. For that validation, we only used the parameters fitted to the five observers in Shivkumar et al. (2025). This allowed us to test the specific hypothesis that the same parameters explaining human behavior could also predict the neural responses.

4.6 Qualitative comparison to empirical data

To qualitatively compare our model's predictions to previously published neural data, we followed a three-step procedure. First, we generated a catalogue of potential neural responses. Second, we defined the specific stimulus conditions ("cross-sections") needed to match each historical study. Third, for each study, we selected the best-fitting prediction from our catalogue.

Generation of the neural prediction catalogue: Our predictions needed to account for variability across both observers and neurons. We therefore generated a catalogue spanning two dimensions: (1) Observer Parameters: We used the five unique parameter sets fitted to the human observers in Shivkumar et al. (2025). (2) Neuron Tuning Profiles: We created nine plausible tuning profiles by combining three distinct tuning widths (approximating the 25th, 50th, and 75th percentiles from DeAngelis and Uka (2003)) for both speed and direction (3 speed widths $\times$ 3 direction widths; see Fig. S5a).

For each of the 45 combinations (5 observers $\times$ 9 tuning profiles), we generated predicted mean firing rates (tuning curves) and response variability (Fano factor) across a dense, four-dimensional grid of center and surround speeds and directions.

Stimulus cross-section selection and matching procedure: To compare our predictions with the modulation effects reported in the literature, we first calculated a surround modulation index (facilitation or suppression) by subtracting the predicted response to a stationary surround from the predicted response to a moving surround (Fig. 6a).

We then extracted specific "cross-sections" from our 4D predictions to precisely match the stimulus conditions of each empirical study. For each dataset, we identified the best-fitting prediction from our 45-member catalogue by selecting the observer/tuning-profile combination that yielded the lowest root mean squared error (RMSE). The specific cross-sections were defined as follows:

(1) To match Allman et al. (1985a): We fixed the center motion to the neuron's preferred direction and speed, and varied the surround direction (0°, 30°, 60°, 90°, 120°, 150°, 180°).

(2) To match Born (2000): We used a similar cross-section, varying the surround direction from 0° to 180° in 45° increments.

(3) To match Tanaka et al. (1986): For direction modulation, we varied the surround direction (0°, 30°, 60°, 90°). For speed modulation, we fixed both center and surround to the preferred direction and varied the surround speed (0.25x, 1x, and 4x the preferred speed).

References

Adelson, E. H., & Bergen, J. R. (1985). Spatiotemporal energy models for the perception of motion. *J. Opt. Soc. Am. A*, *2*(2), 284–299. https://doi.org/10.1364/JOSAA.2.000284

Albright, T. D. (1984). Direction and orientation selectivity of neurons in visual area mt of the macaque [PMID: 6520628]. *Journal of Neurophysiology*, *52*(6), 1106–1130. https://doi.org/10.1152/jn.1984.52.6.1106

Allman, J, Miezin, F, & McGuinness, E. (1985a). Direction- and velocity-specific responses from beyond the classical receptive field in the middle temporal visual area (MT). *Perception*, *14*(2), 105–126.

Allman, J., Miezin, F., & McGuinness, E. (1985b). Stimulus specific responses from beyond the classical receptive field: Neurophysiological mechanisms for local-global comparisons in visual neurons. *Annual Review of Neuroscience*, *8*(Volume 8, 1985), 407–430. https://doi.org/https://doi.org/10.1146/annurev.ne.08.030185.002203

Angelucci, A., Bijanzadeh, M., Nurminen, L., Federer, F., Merlin, S., & Bressloff, P. C. (2017). Circuits and mechanisms for surround modulation in visual cortex. *Annual Review of Neuroscience*, *40*(Volume 40, 2017), 425–451. https://doi.org/https://doi.org/10.1146/annurev-neuro-072116-031418

Baden, T., Euler, T., & Berens, P. (2020). Understanding the retinal basis of vision across species. *Nature Reviews Neuroscience*, *21*(1), 5–20.

Barth, E., & Watson, A. B. (2000). A geometric framework for nonlinear visual coding. *Opt. Express*, *7*(4), 155–165. https://doi.org/10.1364/OE.7.000155

Beck, C., & Neumann, H. (2011). Combining feature selection and integration—a neural model for mt motion selectivity. *PLOS ONE*, *6*(7), 1–15. https://doi.org/10.1371/journal.pone.0021254

Beck, J., Pouget, A., & Heller, K. A. Complex inference in neural circuits with probabilistic population codes and topic models (F. Pereira, C. Burges, L. Bottou, & K. Weinberger, Eds.). In: *Advances in neural information processing systems* (F. Pereira, C. Burges, L. Bottou, & K. Weinberger, Eds.). Ed. by Pereira, F., Burges, C., Bottou, L., & Weinberger, K. *25*. Curran Associates, Inc., 2012. https://proceedings.neurips.cc/paper_files/paper/2012/file/b7087c1f4f89e63af8d46f3b20271153-Paper.pdf

Beck, J., Ralf, H., Xaq, P., Cristina, S., & Eszter, V. Competing theories of probabilistic computations in the brain. In: In *Ccn 2020 workshop gac*. 2020. https://openreview.net/forum?id=uZMMO2obl50

Bill, J., Gershman, S. J., & Drugowitsch, J. (2022). Visual motion perception as online hierarchical inference. *Nature Communications*, *13*(1), 7403.

Bill, J., Pailian, H., Gershman, S. J., & Drugowitsch, J. (2020). Hierarchical structure is employed by humans during visual motion perception. *Proceedings of the National Academy of Sciences*, *117*(39), 24581–24589. https://doi.org/10.1073/pnas.2008961117

Born, R. T. (2000). Center-surround interactions in the middle temporal visual area of the owl monkey [PMID: 11068007]. *Journal of Neurophysiology*, *84*(5), 2658–2669. https://doi.org/10.1152/jn.2000.84.5.2658

Born, R. T., & Bradley, D. C. (2005). Structure and function of visual area mt [PMID: 16022593]. *Annual Review of Neuroscience*, *28*(1), 157–189. https://doi.org/10.1146/annurev.neuro.26.041002.131052

Born, R. T., & Tootell, R. B. H. (1992). Segregation of global and local motion processing in primate middle temporal visual area. *Nature*, *357*(6378), 497–499. https://doi.org/10.1038/357497a0

Braddick, O. (1993). Segmentation versus integration in visual motion processing. *Trends in Neurosciences*, *16*(7), 263–268. https://doi.org/https://doi.org/10.1016/0166-2236(93)90179-P

Bradley, D. C., & Andersen, R. A. (1998). Center-surround antagonism based on disparity in primate area MT. *J Neurosci*, *18*(18), 7552–7565.

Britten, K. H., & Heuer, H. W. (1999). Spatial summation in the receptive fields of mt neurons. *Journal of Neuroscience*, *19*(12), 5074–5084. https://doi.org/10.1523/JNEUROSCI.19-12-05074.1999

Britten, K. H., Shadlen, M. N., Newsome, W. T., & Movshon, J. A. (1993). Responses of neurons in macaque mt to stochastic motion signals. *Visual Neuroscience*, *10*(6), 1157–1169. https://doi.org/10.1017/S0952523800010269

Buesing, L., Bill, J., Nessler, B., & Maass, W. (2011). Neural dynamics as sampling: A model for stochastic computation in recurrent networks of spiking neurons. *PLOS Computational Biology*, *7*(11), 1–22. https://doi.org/10.1371/journal.pcbi.1002211

Carandini, M., & Heeger, D. J. (2012). Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, *13*(1), 51–62. https://doi.org/10.1038/nrn3136

Coen-Cagli, R., Kohn, A., & Schwartz, O. (2015). Flexible gating of contextual influences in natural vision. *Nature Neuroscience*, *18*(11), 1648–1655.

Dakin, S. C., & Mareschal, I. (2000). The role of relative motion computation in 'direction repulsion'. *Vision Research*, *40*(7), 833–841. https://doi.org/https://doi.org/10.1016/S0042-6989(99)00226-6

DeAngelis, G. C., & Uka, T. (2003). Coding of horizontal disparity and velocity by mt neurons in the alert macaque [PMID: 12574483]. *Journal of Neurophysiology*, *89*(2), 1094–1111. https://doi.org/10.1152/jn.00717.2002

Euler, T., Haverkamp, S., Schubert, T., & Baden, T. (2014). Retinal bipolar cells: Elementary building blocks of vision. *Nature Reviews Neuroscience*, *15*(8), 507–519.

Fiser, J., Berkes, P., Orbán, G., & Lengyel, M. (2010). Statistically optimal perception and learning: From behavior to neural representations. *Trends Cogn. Sci.*, *14*(3), 119–130.

Gaynes, J. A., Budoff, S. A., Grybko, M. J., Hunt, J. B., & Poleg-Polsky, A. (2022). Classical center-surround receptive fields facilitate novel object detection in retinal bipolar cells. *Nature Communications*, *13*(1), 5575.

Gershman, S. J., Tenenbaum, J. B., & Jäkel, F. (2016). Discovering hierarchical motion structure [Quantitative Approaches in Gestalt Perception]. *Vision Research*, *126*, 232–241. https://doi.org/https://doi.org/10.1016/j.visres.2015.03.004

Golan, T., Raju, P. C., & Kriegeskorte, N. (2020). Controversial stimuli: Pitting neural networks against each other as models of human cognition. *Proceedings of the National Academy of Sciences*, *117*(47), 29330–29337. https://doi.org/10.1073/pnas.1912334117

Haefner, R. M., Beck, J., Savin, C., Salmasi, M., & Pitkow, X. (2024). How does the brain compute with probabilities?

Haefner, R. M., Berkes, P., & Fiser, J. (2016). Perceptual Decision-Making as probabilistic inference by neural sampling. *Neuron*, *90*(3), 649–660.

Hoyer, P. O., & Hyvärinen, A. Interpreting neural response variability as monte carlo sampling of the posterior. In: In *Proceedings of the 15th international conference on neural information processing systems*. NIPS'02. Cambridge, MA, USA: MIT Press, 2002, 293–300.

Huang, X., Albright, T. D., & Stoner, G. R. (2007). Adaptive surround modulation in cortical area mt. *Neuron*, *53*(5), 761–770. https://doi.org/10.1016/j.neuron.2007.01.032

Huang, X., Albright, T. D., & Stoner, G. R. (2008). Stimulus dependency and mechanisms of surround modulation in cortical area mt. *Journal of Neuroscience*, *28*(51), 13889–13906. https://doi.org/10.1523/JNEUROSCI.1946-08.2008

Inaba, N., Miura, K., & Kawano, K. (2011). Direction and speed tuning to visual motion in cortical areas mt and mstd during smooth pursuit eye movements [PMID: 21273314]. *Journal of Neurophysiology*, *105*(4), 1531–1545. https://doi.org/10.1152/jn.00511.2010

Inaba, N., Shinomoto, S., Yamane, S., Takemura, A., & Kawano, K. (2007). Mst neurons code for visual motion in space independent of pursuit eye movements [PMID: 17329625]. *Journal of Neurophysiology*, *97*(5), 3473–3483. https://doi.org/10.1152/jn.01054.2006

Johansson, G. (1950). Configurations in the perception of velocity. *Acta Psychologica*, *7*, 25–79. https://doi.org/https://doi.org/10.1016/0001-6918(50)90003-5

Kim, J., & Wilson, H. R. (1997). Motion integration over space: Interaction of the center and surround motion**this research was first reported at the annual meeting of the association for research in vision and ophthalmology, may 1994 and 1995. *Vision Research*, *37*(8), 991–1005. https://doi.org/https://doi.org/10.1016/S0042-6989(96)00254-4

Kuffler, S. W. (1953). Discharge patterns and functional organization of mammalian retina. *J Neurophysiol*, *16*(1), 37–68.

Körding, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B., & Shams, L. (2007). Causal inference in multisensory perception. *PLOS ONE*, *2*(9), 1–10. https://doi.org/10.1371/journal.pone.0000943

Lange, R. D., & Haefner, R. M. (2022). Task-induced neural covariability as a signature of approximate bayesian learning and inference. *PLOS Computational Biology*, *18*(3), 1–39. https://doi.org/10.1371/journal.pcbi.1009557

Lange, R. D., Shivkumar, S., Chattoraj, A., & Haefner, R. M. (2022). Bayesian encoding and decoding as distinct perspectives on neural coding. *bioRxiv*. https://doi.org/10.1101/2020.10.14.339770

Lengyel, G., Shivkumar, S., & Haefner, R. M. A general method for testing bayesian models using neural data. In: In *Unireps: The first workshop on unifying representations in neural models*. 2023. https://openreview.net/forum?id=oWJP0NhcY7

Ma, W. J., Beck, J. M., Latham, P. E., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience*, *9*(11), 1432–1438. https://doi.org/10.1038/nn1790

Movshon, J, Adelson, E., Gizzi, M., & Newsome, W. (1985). The analysis of moving visual patterns. In C. Chagas, R. Gattass, & C. Gross (Eds.), *Pattern recognition mechanisms* (pp. 117–151). Vatican Press.

Newsome, W. T., Wurtz, R. H., & Komatsu, H. (1988). Relation of cortical areas mt and mst to pursuit eye movements. ii. differentiation of retinal from extraretinal inputs [PMID: 3171644]. *Journal of Neurophysiology*, *60*(2), 604–620. https://doi.org/10.1152/jn.1988.60.2.604

Nurminen, L., & Angelucci, A. (2014). Multiple components of surround modulation in primary visual cortex: Multiple neural circuits with multiple functions? [The Function of Contextual Modulation]. *Vision Research*, *104*, 47–56. https://doi.org/https://doi.org/10.1016/j.visres.2014.08.018

Orbán, G., Berkes, P., Fiser, J., & Lengyel, M. (2016). Neural variability and Sampling-Based probabilistic representations in the visual cortex. *Neuron*, *92*(2), 530–543.

Pack, C. C., Born, R. T., & Livingstone, M. S. (2003a). Two-dimensional substructure of stereo and motion interactions in macaque visual cortex. *Neuron*, *37*(3), 525–535. https://doi.org/https://doi.org/10.1016/S0896-6273(02)01187-X

Pack, C. C., Gartland, A. J., & Born, R. T. (2004). Integration of contour and terminator signals in visual area mt of alert macaque. *Journal of Neuroscience*, *24*(13), 3268–3280. https://doi.org/10.1523/JNEUROSCI.4387-03.2004

Pack, C. C., Hunter, J. N., & Born, R. T. (2005). Contrast dependence of suppressive influences in cortical area mt of alert macaque [PMID: 15483068]. *Journal of Neurophysiology*, *93*(3), 1809–1815. https://doi.org/10.1152/jn.00629.2004

Pack, C. C., Livingstone, M. S., Duffy, K. R., & Born, R. T. (2003b). End-stopping and the aperture problem: Two-dimensional motion signals in macaque v1. *Neuron*, *39*(4), 671–680. https://doi.org/https://doi.org/10.1016/S0896-6273(03)00439-2

Penaloza, B., Shivkumar, S., Lengyel, G., DeAngelis, G. C., & Haefner, R. M. (2024). Causal inference predicts the transition from integration to segmentation in motion perception. *Scientific Reports*, *14*(1), 27704. https://doi.org/10.1038/s41598-024-78820-6

Penaloza, B., Shivkumar, S., Lengyel, G., Xu, X., DeAngelis, G. C., & Haefner, R. Divisive normalization as a mechanism for hierarchical causal inference in motion perception. In: In *Cosyne*. COSYNE. Montréal, Canada, 2023.

Pouget, A., Beck, J. M., Ma, W. J., & Latham, P. E. (2013). Probabilistic brains: Knowns and unknowns. *Nat. Neurosci.*, *16*(9), 1170–1178.

Qian, N, Andersen, R., & Adelson, E. (1994a). Transparent motion perception as detection of unbalanced motion signals. i. psychophysics. *Journal of Neuroscience*, *14*(12), 7357–7366. https://doi.org/10.1523/JNEUROSCI.14-12-07357.1994

Qian, N, Andersen, R., & Adelson, E. (1994b). Transparent motion perception as detection of unbalanced motion signals. iii. modeling. *Journal of Neuroscience*, *14*(12), 7381–7392. https://doi.org/10.1523/JNEUROSCI.14-12-07381.1994

Raiguel, S., Van Hulle, M. M., Xiao, D.-K., Marcar, V. L., & Orban, G. A. (1995). Shape and spatial distribution of receptive fields and antagonistic motion surrounds in the middle temporal area (v5) of the macaque. *European Journal of Neuroscience*, *7*(10), 2064–2082. https://doi.org/https://doi.org/10.1111/j.1460-9568.1995.tb00629.x

Rubin, D., Van Hooser, S., & Miller, K. (2015). The stabilized supralinear network: A unifying circuit motif underlying multi-input integration in sensory cortex. *Neuron*, *85*(2), 402–417. https://doi.org/https://doi.org/10.1016/j.neuron.2014.12.026

Rust, N. C., Mante, V., Simoncelli, E. P., & Movshon, J. A. (2006). How MT cells analyze the motion of visual patterns. *Nature Neuroscience*, *9*(11), 1421–1431.

Shams, L., & Beierholm, U. (2022). Bayesian causal inference: A unifying neuroscience theory. *Neuroscience & Biobehavioral Reviews*, *137*, 104619. https://doi.org/https://doi.org/10.1016/j.neubiorev.2022.104619

Shivkumar, S., DeAngelis, G. C., & Haefner, R. M. (2025). Hierarchical motion perception as causal inference. *Nature Communications*, *16*(1), 3868. https://doi.org/10.1038/s41467-025-58797-0

Simoncelli, E. P., & Heeger, D. J. (1998). A model of neuronal responses in visual area mt. *Vision Research*, *38*(5), 743–761. https://doi.org/https://doi.org/10.1016/S0042-6989(97)00183-1

Tadin, D., & Lappin, J. S. 278279linking psychophysics and physiology of center-surround interactions in visual motion processing. In: In *Seeing spatial form*. Oxford University Press, 2005, Novem-

ber. ISBN: 9780195172881. https://doi.org/10.1093/acprof:oso/9780195172881.003.0014 https://academic.oup.com/book/0/chapter/193335756/chapter-ag-pdf/44625360/book“˙25762“˙section“˙1933357

Tadin, D., Lappin, J. S., Gilroy, L. A., & Blake, R. (2003). Perceptual consequences of centre–surround antagonism in visual motion processing. *Nature*, *424*(6946), 312–315. https://doi.org/10.1038/nature01800

Tadin, D., Paffen, C. L. E., Blake, R., & Lappin, J. S. (2008). Contextual modulations of center-surround interactions in motion revealed with the motion aftereffect. *Journal of Vision*, *8*(7), 9–9. https://doi.org/10.1167/8.7.9

Tadin, D., Park, W. J., Dieter, K. C., Melnick, M. D., Lappin, J. S., & Blake, R. (2019). Spatial suppression promotes rapid figure-ground segmentation of moving objects. *Nature Communications*, *10*(1), 2732. https://doi.org/10.1038/s41467-019-10653-8

Tajima, C. I., Tajima, S., Koida, K., Komatsu, H., Aihara, K., & Suzuki, H. (2016). Population code dynamics in categorical perception. *Scientific Reports*, *6*(1), 22536.

Tanaka, K, Hikosaka, K, Saito, H, Yukie, M, Fukada, Y, & Iwai, E. (1986). Analysis of local and wide-field movements in the superior temporal visual areas of the macaque monkey. *Journal of Neuroscience*, *6*(1), 134–144. https://doi.org/10.1523/JNEUROSCI.06-01-00134.1986

Turner, M. H., Schwartz, G. W., & Rieke, F. (2018). Receptive field center-surround interactions mediate context-dependent spatial contrast encoding in the retina (A. J. King, Ed.). *eLife*, *7*, e38841. https://doi.org/10.7554/eLife.38841

Tzvetanov, T., & Womelsdorf, T. (2008). Predicting human perceptual decisions by decoding neuronal information profiles. *Biological Cybernetics*, *98*(5), 397–411. https://doi.org/10.1007/s00422-008-0226-0

Ujfalussy, B. B., & Orbán, G. (2022). Sampling motion trajectories during hippocampal theta sequences (A. Peyrache & L. L. Colgin, Eds.). *eLife*, *11*, e74058. https://doi.org/10.7554/eLife.74058

Vértes, E., & Sahani, M. Flexible and accurate inference and learning for deep generative models. In: In *Proceedings of the 32nd international conference on neural information processing systems*. NIPS'18. Montréal, Canada: Curran Associates Inc., 2018, 4170–4179.

Vértes, E., & Sahani, M. A neurally plausible model learns successor representations in partially observable environments (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett, Eds.). In: *Advances in neural information processing systems* (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett, Eds.). Ed. by Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., & Garnett, R. *32*. Curran Associates, Inc., 2019. https://proceedings.neurips.cc/paper/2019/file/dea184826614d3f4c608731389ed0c74-Paper.pdf

Wilson, H. R., Ferrera, V. P., & Yo, C. (1992). A psychophysically motivated model for two-dimensional motion perception. *Visual Neuroscience*, *9*(1), 79–97. https://doi.org/10.1017/S0952523800006386

Wilson, H. R., & Kim, J. (1994). A model for motion coherence and transparency. *Visual Neuroscience*, *11*(6), 1205–1220. https://doi.org/10.1017/S0952523800007008

Xiao, D. K., Marcar, V. L., Raiguel, S. E., & Orban, G. A. (1997). Selectivity of macaque MT/V5 neurons for surface orientation in depth specified by motion. *Eur J Neurosci*, *9*(5), 956–964.

Yang, S., Bill, J., Drugowitsch, J., & Gershman, S. J. (2021). Human visual motion perception shows hallmarks of bayesian structural inference. *Scientific Reports*, *11*(1), 3714. https://doi.org/10.1038/s41598-021-82175-7

Zarei Eskikand, P., Grayden, D. B., Kameneva, T., Burkitt, A. N., & Ibbotson, M. R. (2024). Understanding visual processing of motion: Completing the picture using experimentally driven computational models of mt. *Reviews in the Neurosciences*, *35*(3), 243–258. https://doi.org/doi:10.1515/revneuro-2023-0052

Zarei Eskikand, P., Kameneva, T., Burkitt, A. N., Grayden, D. B., & Ibbotson, M. R. (2019). Pattern motion processing by mt neurons. *Frontiers in Neural Circuits*, *13*. https://doi.org/10.3389/fncir.2019.00043

Zarei Eskikand, P., Kameneva, T., Burkitt, A. N., Grayden, D. B., & Ibbotson, M. R. (2020). Adaptive surround modulation of mt neurons: A computational model. *Frontiers in Neural Circuits*, *14*. https://doi.org/10.3389/fncir.2020.529345

Zarei Eskikand, P., Kameneva, T., Ibbotson, M. R., Burkitt, A. N., & Grayden, D. B. (2016). A possible role for end-stopped v1 neurons in the perception of motion: A computational model. *PLOS ONE*, *11*(10), 1–27. https://doi.org/10.1371/journal.pone.0164813

Zemel, R. S., Dayan, P., & Pouget, A. (1998). Probabilistic Interpretation of Population Codes. *Neural Computation*, *10*(2), 403–430. https://doi.org/10.1162/089976698300017818

Zoccolan, D., Cox, D. D., & DiCarlo, J. J. (2005). Multiple object response normalization in monkey inferotemporal cortex. *Journal of Neuroscience*, *25*(36), 8150–8164. https://doi.org/10.1523/JNEUROSCI.2058-05.2005