Hierarchical motion perception as causal inference

Sabyasachi Shivkumar,[1,2*] Gregory C. DeAngelis,[1,3] Ralf M. Haefner[1,3*]

[1]Brain and Cognitive Sciences, University of Rochester, Rochester, NY 14627, USA,
[2]Zuckerman Mind Brain Behavior Institute, Columbia University, NY 10027, USA
[3]Center for Visual Science, University of Rochester, Rochester, NY 14627, USA

One-sentence summary: Recursive generative model motif with most prior mass on stationarity quantitatively explains hierarchical motion perception

Since motion can only be defined relative to a reference frame, which reference frame guides perception? A century of psychophysical studies has produced conflicting evidence: retinotopic, egocentric, world-centric, or even object-centric. We introduce a hierarchical Bayesian model mapping retinal velocities to perceived velocities. Our model mirrors the structure in the world, in which visual elements move within causally connected reference frames. Friction renders velocities in these reference frames mostly stationary, formalized by an additional delta component (at zero) in the prior. Inverting this model automatically segments visual inputs into groups, groups into supergroups, etc. and "perceives" motion in the appropriate reference frame. Critical model predictions are supported by two new experiments, and fitting our model to the data allows us to infer the subjective set of references frames used by individual observers. Our model provides a quantitative normative justification for key Gestalt principles providing inspiration for building better models of visual processing in general.

Introduction

If motion can only be defined relative to a reference frame (*1*), what is the brain's reference frame for the perception of a moving object? A century of psychophysical studies has provided us with evidence that motion is alternatively perceived in a retinotopic reference frame (*2*), in allocentric (world) coordinates (*3*), or coordinate frames defined by other objects in a visual scene (*4–8*). Interestingly, perceived motion can rarely be explained by a single reference frame.

For instance, in the famous Johansson illusion (*4*), while the perceived velocity of the center dot is clearly biased away from the observed retinal velocity, it is not vertical as predicted by a reference frame defined by the flanker dots. Equally, the "flow-parsing" hypothesis (*3*) proposes that the brain subtracts optic flow signals that are compatible with self-motion in order to make us perceive object motion in allocentric coordinates; yet, the empirically observed subtraction is rarely complete (*9*)

Importantly, our perception of motion appears to be closely linked to an observer's perception of the "Gestalt" of a scene, its structure, or configuration (*10*). Recent work has made some progress in mathematically formalizing this elusive concept of a Gestalt: first in information-theoretic terms (*5,11*), and more recently in closely-related Bayesian terms (*12–14*). A Bayesian formulation, compatible with the widely influential idea that the brain combines incoming sensory information with prior expectations to form subjective beliefs about the outside world (*15, 16*), has the advantage that these priors can be justified by the statistics and structure of the outside world. Yet, existing Bayesian accounts of motion perception are either formulated in a purely retinotopic reference frame (*17, 18*) or use priors that are not justified by our knowledge about the world but instead "put in the Gestalt law that we want to get out" (*14*). Furthermore, quantitative empirical tests of these models are based on explicit questions like "Do you perceive structure A or B?" (*7, 19–21*) which are known to be susceptible to decision biases (*22*).

On a mechanistic level, motion signals are processed locally in early visual areas, and the brain needs to combine information across these local motion detectors to form a coherent percept. A long line of research has modeled how these early areas detect local motion (*2, 23*) and how, through a series of linear-nonlinear stages, this local motion can be combined into a more global motion percept (*2, 24*). While integrating these local motion signals allows the brain to solve the 'aperture' problem (*25*), it is not always useful for the brain to integrate information. In fact, local motion differences are a powerful segmentation cue, and several studies (*6, 25*) have shown how our brain contrasts local motions to perceive relative motion. It is however unclear how the brain decides between these two opposing operations, integration and segmentation.

A separate line of research (*26*) in multisensory integration has modeled how the brain solves a similar problem of deciding when to combine information across cues in a Bayesian framework ('causal inference'). Given the general nature of this problem of deciding when to combine information, causal inference has been proposed to be a universal computational motif across the sensory cortex (*27*). In motion perception, causal inference has been used to successfully explain biases in estimating heading direction from both visual and vestibular cues (*28, 29*), but not the perception of moving objects.

Here we present a hierarchical causal inference model that overcomes all of the above challenges by performing joint inference over the hierarchical structure of a scene and the motion of individual visual elements within it. Importantly, the motion priors in our model are justified by motion in the real world, in which most objects are not merely slow, but exactly stationary with respect to their canonical reference frame (*16*). We also present new data from two psychophysical experiments in humans that probe the hierarchical perception of motion and that

provide strong support for the key elements of our model, in particular its hierarchical structure, novel prior, and approximate computations.

Results

Motion is perceived in dynamically inferred reference frames that reflect the causal structure of the world

Much of the world consists of approximately rigid objects who in turn are made up of approximately rigid parts. During translation, all points on a rigid object move in the world at the same speed. Consequently, a common velocity for multiple moving elements in a visual scene acts as a strong cue that the elements belong to a single object. Not surprisingly, when we observe a group of dots moving at the same speed (Figure 1A, top), our brain combines them into one object that it perceives as moving (Figure 1A, bottom), rather than perceiving the individual parts as moving independently (*10, 30*). This common velocity cue also allows the brain to segment the scene into multiple moving objects. For example, when we observe the dots moving as shown in Figure 1B, top, we perceive two partially overlapping objects moving at their respective velocities (Figure 1B, bottom).

Importantly, objects do not simply move independently in the world, but are related to each other through hierarchical whole-part relationships. When a part moves differently from the whole, the whole becomes a natural frame of reference in which to represent that part's motion. For example, the body is the natural frame of reference for the motion of an arm because of the causal whole-part relationship between the two: any change in the motion of the body is directly translated into a change in the motion of the arm.

Our key idea linking retinal observations to percepts in Figure 1A-D is that the brain dynamically constructs reference frames within which most of the visual elements it contains are stationary. We formalize this aspect of the physical world by extending the traditional slow-speed prior (*18*) over moving objects to include a mixture component consisting of a delta at 0. This is shown graphically in Figure 1E, where α denotes the prior probability that an object is stationary. This prior acts on the relative velocities of the visual elements represented in this reference frame. This brings us to the central motif in the generative model we propose is used by the brain to perform inference (Figure 1F). The motif specifies how the velocity of an object is the sum of the velocity of the reference frame and the velocity of the object within that reference frame. This sum is probabilistic, allowing for computational imprecision as quantified by σ^2_Δ. Inference within this generative model motif leads to a decomposition of the observed velocities of visual elements into the velocity of a (shared) reference frame and each element's velocity relative to that reference frame. The degeneracy of this decomposition is broken by the mixture prior over the relative velocities, which leads to an automatic "chunking" of moving elements into groups that are inferred to move together. We hypothesize that the perceived velocity of a visual element moving in a reference frame is its relative velocity to

the reference frame velocity if the relative velocity is non-zero and the reference frame velocity otherwise. This is compactly illustrated in our model by adding a shaded gray box around the candidate variables for perception in the generative model. Under this illustration, the percept corresponds to the candidate velocity that is non-zero and lowest in the model hierarchy (closest to the observations).

Recursively applying this motif leads to our proposed hierarchical causal inference model describing velocity percepts in a scene consisting of dots moving according to hierarchical causal relationships. Our model combines dots into groups, groups into supergroups, and so on (Figure 1G). At the bottom of the model are the actually observed velocities in retinal coordinates, $\vec{o}$, which are linked to the latent variables $\vec{v}$ by a Gaussian likelihood whose width represents the observational noise. The top level of the hierarchy is the velocity corresponding to the stationary objects in the world in the egocentric reference frame. For stationary observers this is zero, but for moving observers it is equal to the optic flow velocity caused by self-motion. This allows for a natural extension of this model to explain deviations in perceived velocities due to self-motion (*3*).

Performing Bayesian inference in this generative model requires computing a posterior belief over all possible structures in which the visual elements could be grouped, and over all the velocities in each of the structures. Before empirically testing the quantitative predictions of this model, we next explain the intuitive impact of its key elements using an increasing number of moving dots, building up to explain the classic Johansson illusion.

Illustrating the causal inference model for dot stimuli consisting of 1-3 dots

We illustrate the key features of the model by applying it to very simple stimuli consisting of two or three dots. The model infers full posteriors over all possible structures and the velocities within each structure. For compactness, we focus on the most probable structures and show the most likely inferred velocity using vectors instead of variables in the generative model, as shown in Figure 2A for a simple stimulus in which a single dot is moving. We explicitly show the variables for the rest of the structures in Figure S1.

When observing two dots that move with the same velocity, there are two primary structures that can explain the observations: one in which both dots are moving independently, each represented in the egocentric coordinate system as an individual dot (Figure 2B, left structure), or a structure in which each dot can be inferred to be stationary with respect to a group (an abstract object), consisting of both dots, with the velocity of the group corresponding to the retinal velocity of the dots (Figure 2B, right structure). The delta component of our mixture prior over the relative velocities ensures that the latter hierarchical structure has the highest likelihood given the data since it has the fewest non-zero relative velocities. Furthermore, since the observed velocities for each dot will slightly deviate from each other due to observation noise, the group velocity combines both observations to obtain a more reliable velocity percept of the group (*31*). This combination of dots into a single group occurs for all dots in the scene that are inferred to move with the same velocity.

If a third dot is added to the scene that moves with a different velocity from the other two dots, the two plausible structures to explain the observations are: (a) an object consisting of the two coherently moving dots plus an independently moving third dot (Figure 2C, left) and (b) a single object consisting of all three dots in which the differently moving dot is represented as moving in the object's reference frame (Figure 2C, right). The slow speed component in our mixture prior favors the latter structure if the differently moving dot has a smaller speed in the object's reference frame than in the stationary egocentric reference frame. For instance, in the Johansson illusion (Figure 2C), the third dot has a small relative velocity with respect to the two coherently moving dots, and its perceived velocity is indeed biased towards its velocity in the reference frame provided by the group made up of the two dots (dark green vertical arrow in Figure 2C, right). However, it is important to recognize that even for as little as three moving elements, there are 16 different structures, for instance one in which dots 2 & 3 move relative to dot 1 rather than the other way around, or where all dots move relative to a reference frame defined by all of the dots together, moving at an intermediate velocity. So it may not be surprising that structure (b) alone cannot explain human observers' percepts which sometimes correspond to the retinal velocity suggesting structure (a), sometimes correspond to the relative velocity (structure (b)), and sometimes lie in between (*32, 33*), suggesting that the brain performs inference over multiple, if not all, of these possible structures.

Furthermore, model predictions will depend on how perception is related to the posterior over structures and velocities. Prior work (*7*) has suggested the mean for the most likely structure, but it could also correspond to the mean across structures as in other work on causal inference (*26*), or posterior sampling (*34, 35*).

New empirical tests of model predictions

In order to quantitatively distinguish between our model and previous proposals, we designed two motion estimation tasks that tested the key elements of our model: (i) the novel mixture prior over relative velocity with a delta at zero, and (ii) the linking hypothesis mapping the posterior distributions over velocities in the model to the distribution over perceived velocities. Importantly, by using a motion estimation task, we test both causal inference over reference frames, and the perception of motion within a reference frame.

Experiment 1: Test using stimuli with two potential levels of hierarchy

In order to test our model, and to constrain its parameters, Experiment 1 was designed to precisely measure human motion perception during the transition from integration to segmentation. Observers used a dial to report their perceived motion direction of a patch of green dots, surrounded by a variable number of patches of red dots (Figure 3A). Surround dots were either stationary, or moving horizontally (direction 0 degrees). The number of surround patches was randomly chosen every trial from $\{1, 2, 3, 5, 10\}$, while the center always consisted of a single patch of dots. Additionally, the retinal center direction (direction on the screen) was randomly

chosen on each trial from the set $\{0, \pm 2.5, \pm 5, \pm 10, \pm 20, \pm 45\}^\circ$. The center and surround had a common horizontal velocity (0°), such that the direction of the center's velocity relative to the surround was $\pm 90^\circ$ depending on the sign of the center direction (more details in Methods). As expected, reported directions lie along the identity line when the surround is stationary (Figure 3B).

When the surround is moving, observer responses systematically deviate from the identity line (Figures 3C-G, S4). Specifically, responses are biased towards zero degrees (surround direction) for small center directions, consistent with the observer integrating the center and surround velocities and reporting the cue-combined velocity. The reported velocities are biased towards 90° for larger center directions, consistent with the observer perceiving the relative velocity between the center and surround.

The responses are in excellent agreement with our causal inference model predictions which are overlaid as violin plots. The absolute goodness of fit was measured by variance explained (VE) to be between 92-96% across observers. Also note the clear evidence of causal inference – uncertainty whether to integrate or segment – in the form of bimodal responses (or increased variability) for intermediate center directions around 5° visible in both empirical responses and model fits. These characteristic deviations between retinal motion and perceptual reports depend on the number of surround patches, also in agreement with the model fits (Figure S3 with an average VE of 94%).

We also found clear quantitative empirical signatures of the shift from integration to segmentation using a model-free analysis. To get an interpretable estimate of the effect of the surround on the center, we mapped responses to a modulation index. This index lies between -1 and $+1$, where -1 corresponds to complete integration (perceiving the surround direction), 0 corresponds to the surround having no effect, and $+1$ corresponds to complete segmentation (perceiving the relative velocity between the center and surround; see Methods for details).

The mean modulation index, averaged across observers (Figure 3H) is negative indicating integration, for 2.5° ($p < 0.001$ for the group, $p < 0.05$ individually for 4 out of 5 observers, based on 5000 bootstrapped samples). For larger separations (greater than 10°), the average and individual mean modulation indices are positive indicating segmentation for larger separations ($p < 0.001$). The standard deviation of modulation index, averaged across observers (Figure 3F) is largest for intermediate separations ($p < 0.01$ for the difference in the standard deviations of the modulation index at 5° and at 2.5° and 45°) indicating a higher variability due to uncertainty over causal structures, in agreement with our model predictions.

Insights from model fitting

We next fit the model to individual responses and obtained posterior distributions over the parameters for each observer (Methods). This allowed us to gain three key insights about the model: (a) whether the mass in the delta component was required to explain the pattern of responses, (b) how observers map the inferred posteriors to responses on each trial, and, (c) how the different causal structures contribute depending on center direction.

Remarkably, for all observers, most of the prior mass was in the delta-component, indicating the strength the brain's expectation that relative motion in the world is exactly zero, rather than merely slow (Figure 4A). We further confirmed the presence of a delta component in the mixture prior by a formal model comparison (Figure 4B, leftmost column) and found strong evidence against a model without as compared to models with mass in the delta component.

Our model comparison (Figure 4B) also revealed that observers' responses are best described as arising from approximate inference (posterior sampling) in the full Bayesian model. We compared this to other previously proposed maps from posteriors to responses: (a) reporting the posterior mean (model averaging) which is the optimal strategy for estimation tasks (*36*), (b) reporting the conditional mean under the most probable structure (model selection) which maximizes consistency (*37*) and minimizes the description length (*38*), and (c) reporting the mean by sampling the structure (structure sampling) which is a more precise approximation than posterior sampling (*35*) (Figure S7). For all observers, the model favored by the evidence had different prior widths and different masses on the delta component for different velocity variables (center, surround, group), performed joint inference over structure and motion, and converted posteriors into responses by sampling.

Our model also allowed us to determine the causal structures underlying each observer's subjective percepts. We found that for all observers, only four out of 16 possible structures were assigned a significant posterior mass (Figure 4C-F for 10 surround patches, and Figures S5, S6 for all other conditions). Under structure 1, the observer integrates center and surround, thus perceiving the cue-combined velocity (Figure 4C). As expected, the posterior probability of this structure was highest when the center and surround moved with the same velocity and decreased with an increase in separation between center and surround velocities (Figure 4G). The bias in the perceived center velocity towards the surround was determined by the weight given to the surround (Figure 4K), which we quantified using the modulation index (as in the previous section). Under structure 2, center motion is perceived in the reference frame defined by the surround – the canonical structure typically assumed for center-surround motion segmentation (Figure 4D,H, L). Unexpectedly, this structure is dominant for only 1 out of 5 observers, plays a transient role for intermediate differences between center and surround motion directions for just 2 observers, and only plays a very minor role for the remaining observers. The same is true for structure 3, under which the surround is perceived in the reference frame of the center, and perception of center motion mostly coincides with retinal motion (Figure 4E,I,M). Finally, structure 4 implies that both center and surround motion are perceived with respect to a reference that moves at a velocity intermediate to both center and surround (Figure 4F,J,N). This is the structure that carries primary responsibility for intermediate percepts (i.e. the apparently incomplete subtraction of the surround from center velocity) at large differences between center and surround motion directions. Surprisingly, none of the observers places any mass on the possibility that center and surround might belong to different causal structures (the 99th percentile of mass on this structure is below 1% for all observers) – an alternative potential explanation for intermediate percepts. However, one prediction of such a structure would have been to find a substantial fraction of responses along the identity line (given that responses are best explained

by posterior matching overall), something that we did not observe (Figure 3C-G). The reason that structure 4 has higher posterior mass than structures 2 and 3 for large separations is the Gaussian slow speed component in the prior: the smaller relative velocities under structure 4 overpower the mass in the delta component of the prior.

As predicted by the model, we also find that both the integration effect (quantified by the model predicted modulation index at 0°) and segmentation effect (quantified by the model predicted modulation index at 45°) become stronger with increase in the number of surround patches (Figure S8B and S8C respectively).

Experiment 2: Test using stimuli with three potential levels of hierarchy

Next, we added a second surround of moving dots to test three key qualitative behavioral predictions of the hierarchical replication of our causal inference motif (possible structures shown in Figure S9). First, motion perception during segmentation is local, i.e. the perceived motion of an element inferred to be moving relative to a group is independent of other motion in the visual scene. Second, if a visual element is integrated with a group, then it inherits the reference frame of that group. Third, the retinal velocity of a group, and not the velocity with which the group is perceived, determines the reference frame velocity for the elements that are part of that group (compare the alternative model in Figure S10 to Figure 1G).

As in Experiment 1, observers reported the direction of the center patchmoving in the presence of a surround but in Experiment 2, the surround consisted of inner and outer rings of dots (Figure 5A). The inner and outer ring velocities were on opposite sides of the center patch velocity with the inner ring moving at $\{0^\circ, -3^\circ, -10^\circ, -30^\circ, -45^\circ\}$ counter-clockwise from the center. The outer ring moved at 60° counter-clockwise from the center when the inner ring moved at 0° and the outer ring's velocity was adjusted to maintain a constant relative velocity between outer and inner rings as the direction of inner ring was varied.

The data clearly confirms our model predictions: when the center and inner ring are integrated to form a group, this group is perceived in the reference frame formed with the outer ring (significant biases towards -90°, Figure 5B magenta box, $p<0.01$ for all observers). However, when the center is segmented from the inner ring (when its velocity is very different than that of the inner ring), then the center is perceived as moving in the reference frame formed with the inner ring, ignoring the outer ring (reports are biased towards 90°, Figure 5B orange box, significant for 7/9 observers with $p<0.01$).

Our data also supports our model's prediction that the retinal velocity of the inferred group determines the group's reference frame velocity (Figures 5C and 5D respectively). Specifically, as the center is perceived relative to the inner ring when they move differently, the outer ring should have no effect on the percept. Consistent with this, we found no significant difference in reported center directions for different outer ring directions, for inner ring directions of -30° and -45° ($p> 0.05$ evaluated through bootstrapping with 1/9 observers showing significant difference for both inner ring directions). For these inner ring directions, the observers predominantly segment the center and inner rings (Figure 5B, orange box). The individual trial

responses are shown in Figure S11.

In addition to the conditions shown in Figure 5B, we also test how: (a) the observers perceive the inner ring in the presence of the center and/or surround, (b) how the center ring percept is biased by just the inner ring moving coherently (similar to Experiment 1) and just the outer ring moving coherently. Responses in these conditions also agree with predictions of the causal inference model (Figure S12C-J).

Discussion

We have presented a new model of complex motion perception, and new data to support this model. The key normative ingredients, reflecting the structure of the physical world, are that the motion of visual elements should be represented in reference frames that they are causally connected to, that due to friction forces, the velocity in such a reference frame is mostly zero, and that the world is compositional and hierarchically organized. We designed two experiments in order to test all key elements of our model, to constrain its parameters, and to generate new insights for motion perception in scenes with a richer motion structure than earlier experiments.

Our work advances the large body of work trying to understand sensory processing in terms of natural input statistics (*39–42*). Whereas prior work started with the input statistics at the sensory periphery, our model explicitly incorporates causal relationships (moving the whole will move the part) and the importance of friction (the relative velocity is zero for most objects that are touching each other) for giving rise to these statistics. Strikingly, fitting our model to perceptual reports revealed that most prior mass is in fact at zero for all of our human observers (Figure 4A).

The richness of our our psychophysical data, together with formal model comparison, also allowed us to answer the important question of how the often-complex posterior distributions of a Bayesian observer are related to our deterministic-appearing percept (*43, 44*). In prior work, inference over motion structures and velocities within each structure are treated as separate problems in which either the most likely structure is inferred (*7*) or the structure is assumed to be known or learned over longer time scales (*20*). Our experiments and analyses provide overwhelming evidence that, at least in the context of motion processing, the brain performs joint inference over both structure and motion, and that perceptual reports are best described by samples from this joint distribution (Figure 4B).

Approximate joint inference in our model can also explain why effect sizes in prior studies usually deviated from previously proposed theories like (Bayesian) Vector Analysis (*4, 33*) or flow-parsing (*3*): a range of different causal explanations are all consistent with the presented impoverished psychophysical stimuli. We confirmed a closely related prediction of our model – the increase in integration and segmentation effect sizes with the number of dots, i.e. the uncertainty in the surround (*45*) of our experiment 1 (Figure S8).

Fitting our model to individual responses revealed unexpected variability in the causal structures inferred by different observers. We had expected that partial segmentation for large dif-

ferences in motion direction between center and surround was best explained by observers' lack of confidence that center and surround were indeed part of the same object. However, our quantitative analysis showed this not to be true: partial segmentation was actually explained by observers perceiving both center and surround to be slowly moving with respect to an abstract reference frame moving with an intermediate velocity. A corollary to this result is that the larger the integration bias (i.e. the more the reference frame velocity aligns with the surround), the larger is the segmentation bias, in agreement with earlier work (*8*). This result also demonstrated the importance of the 'slow-speed' component of the prior on relative motion – in contrast to the otherwise analogous spike and slab prior (*46*) common in machine learning.

Earlier work on the influence of context on perception found this influence to be a function of the percept of the context, not the direct sensory input defining the context – both for temporal context effects (*47*), and during binocular rivalry (*48*). Our data from Experiment 2 differs from these findings, confirming that our percepts reflect inference in a model based on the physical laws of motion and suggesting differences between the causal models underlying the ventral and the dorsal processing streams.

While it has long been recognized that both integration and segmentation are key operations underlying motion perception (*25*), our model shows that there are in fact two qualitatively different kinds of segmentation: (1) do two visual elements belong to the same causal structure, and, if the answer is yes, (2) is a visual element moving with respect to its reference frame? While our model answers both questions using the framework of 'causal inference' from multisensory integration (*26*), which has been proposed as a 'universal computation' for cortex (*27*), only the first question corresponds to a question about causality as statistically defined (*49*).

Our model also lends itself to making predictions for neurophysiological data. We leave for future work the tantalizing possibility that the two kinds of variables in our model (corresponding to the left and right sides of Figure 1G) are represented by the two major classes of neurons who have been reported (*50*) in cortical motion area MT (surround-suppressed and not surround-suppressed).

Bayesian causal inference has recently been proposed as a unifying theory for neuroscience (*27*). Our model extends the simple 'same' or 'different' scenarios in previous causal inference work to hierarchical whole-part relationships reflecting the compositionality of the world (*51*). The computations at the lowest level of our model in fact resembles a recent probabilistic model of neural responses in primary visual cortex (*52*), and the hierarchical architecture of our model directly suggests an equivalent one for the ventral stream, potentially allowing us to understand both dorsal and ventral stream as performing inference in closely related generative models.

Acknowledgements

We would like to thank Jan Drugowitsch, Gabor Lengyel, Boris Penaloza, Johannes Bill, Jean-Paul Noel, Greg Horwitz, Steven Grisafi, and Frank Jäkel for their helpful comments on our manuscript.

Funding

This work was supported by:

National Institutes of Health grant U19NS118246 (to SS, GCD, and RMH)
Division of Information and Intelligent Systems IIS-2143440 (to RMH)
National Science Foundation grant NSF-1449828 (to SS).

Author Contributions

Conceptualization: SS, GCD, RMH
Methodology: SS, GCD, RMH
Investigation: SS
Formal Analysis: SS, RMH
Visualization: SS, GCD, RMH
Funding acquisition: GCD, RMH
Writing – original draft: SS, GCD, RMH
Writing – review & editing: SS, GCD, RMH

Competing Interests

Authors declare that they have no competing interests.

Code and Data availability

Code and data is available at https://osf.io/f9dsg/.

List of Supplementary Materials

Methods
Supplementary Text
Figs. S1 to S12
Tables S1

References

1. C. Hoefer, N. Huggett, J. Read, *The Stanford encyclopedia of philosophy* (2023).
2. E. H. Adelson, J. R. Bergen, *Journal of the Optical Society of America A* **2**, 284 (1985).
3. P. A. Warren, S. K. Rushton, *Current Biology* **19**, 1555 (2009).

4. G. Johansson, *Acta Psychologica* pp. 1–55 (1950).

5. F. Restle, *Psychological Review* **86**, 1 (1979).

6. W. C. Gogel, M. Koslow, *Perception & Psychophysics* **11**, 309 (1972).

7. S. J. Gershman, J. B. Tenenbaum, F. Jäkel, *Vision Research* **126**, 232 (2016).

8. X. Wu, M. Spering, *PLOS ONE* **17**, e0275324 (2022).

9. D. C. Niehorster, L. Li, *i-Perception* **8**, 204166951770820 (2017).

10. M. Wertheimer, *Zeitschrift fur psychologie* **61**, 161 (1912).

11. J. R. Pomerantz, M. Kubovy (1986).

12. V. Froyen, J. Feldman, M. Singh, *Psychological Review* **122**, 575 (2015).

13. J. Feldman, *Psychological review* **116**, 875 (2009). Publisher: American Psychological Association.

14. F. Jäkel, M. Singh, F. A. Wichmann, M. H. Herzog, *Vision Research* **126**, 3 (2016).

15. H. Von Helmholtz, *Handbuch der physiologischen Optik: mit 213 in den Text eingedruckten Holzschnitten und 11 Tafeln*, vol. 9 (1867).

16. D. C. Knill, W. Richards, *Perception as bayesian inference* (1996).

17. Y. Weiss, E. P. Simoncelli, E. H. Adelson, *Nature Neuroscience* **5**, 598 (2002).

18. A. A. Stocker, E. P. Simoncelli, *Nature Neuroscience* **9**, 578 (2006).

19. J. Bill, H. Pailian, S. J. Gershman, J. Drugowitsch, *Proceedings of the National Academy of Sciences* **117**, 24581 (2020). Publisher: National Acad Sciences.

20. J. Bill, S. J. Gershman, J. Drugowitsch, *Nature communications* **13**, 7403 (2022).

21. S. Yang, J. Bill, J. Drugowitsch, S. J. Gershman, *Scientific Reports* **11**, 3714 (2021).

22. J.-P. Noel, S. Shivkumar, K. Dokka, R. M. Haefner, D. E. Angelaki, *Elife* **11**, e71866 (2022).

23. B. Hassenstein, W. Reichardt, *Zeitschrift für Naturforschung B* **11**, 513 (1956).

24. N. C. Rust, V. Mante, E. P. Simoncelli, J. A. Movshon, *Nature Neuroscience* **9**, 1421 (2006).

25. O. Braddick, *Trends in Neurosciences* **16**, 263 (1993).

26. K. P. Körding, *et al.*, *PLoS ONE* **2**, e943 (2007).

27. L. Shams, U. Beierholm, *Neuroscience & Biobehavioral Reviews* **137**, 104619 (2022).

28. L. Acerbi, K. Dokka, D. E. Angelaki, W. J. Ma, *PLOS Computational Biology* **14**, e1006110 (2018).

29. K. Dokka, H. Park, M. Jansen, G. C. DeAngelis, D. E. Angelaki, *Proceedings of the National Academy of Sciences* **116**, 9060 (2019).

30. G. Jansson, G. Johansson, *Perception* **2**, 321 (1973).

31. M. O. Ernst, M. S. Banks, *Nature* **415**, 429 (2002).

32. J. Hochberg, P. Fallon, *Science (New York, N.Y.)* **194**, 1081 (1976).

33. K. H. Shum, G. L. Wolford, *Perception & Psychophysics* **34**, 17 (1983).

34. E. Vul, N. Goodman, T. L. Griffiths, J. B. Tenenbaum, *Cognitive Science* **38**, 599 (2014).

35. D. R. Wozny, U. R. Beierholm, L. Shams, *PLoS Computational Biology* **6**, e1000871 (2010).

36. K. P. Körding, D. M. Wolpert, *Nature* **427**, 244 (2004).

37. A. A. Stocker, E. Simoncelli, *Advances in neural information processing systems* **20** (2007).

38. N. Chater, *Psychological Review* **103**, 566 (1996).

39. F. Attneave, *Psychological review* **61**, 183 (1954).

40. H. B. Barlow, others, *Sensory communication* **1**, 217 (1961).

41. A. Hyvärinen, J. Hurri, P. O. Hoyer, *Natural image statistics: A probabilistic approach to early computational vision.*, vol. 39 (2009).

42. S. Laughlin, *Zeitschrift für Naturforschung C* **36**, 910 (1981).

43. N. Block, *Philosophical Transactions of the Royal Society B: Biological Sciences* **373**, 20170341 (2018).

44. D. Rahnev, N. Block, R. N. Denison, J. Jehee (2021).

45. Y. Zhou, L. Acerbi, W. J. Ma, *PLoS computational biology* **16**, e1006308 (2020).

46. T. J. Mitchell, J. J. Beauchamp, *Journal of the american statistical association* **83**, 1023 (1988).

47. G. M. Cicchini, A. Benedetto, D. C. Burr, *Current Biology* **31**, 1245 (2021).

48. A. Chopin, P. Mamassian, R. Blake, *Vision Research* **63**, 63 (2012).

49. J. Pearl, *Causality* (2009).

50. R. T. Born, D. C. Bradley, *Annual Review of Neuroscience* **28**, 157 (2005).

51. B. M. Lake, R. Salakhutdinov, J. B. Tenenbaum, *Science (New York, N.Y.)* **350**, 1332 (2015).

52. R. Coen-Cagli, A. Kohn, O. Schwartz, *Nature Neuroscience* **18**, 1648 (2015).

53. T. Mori, *Perceptual and Motor Skills* **48**, 587 (1979).

54. L. Acerbi, W. J. Ma, S. Vijayakumar, *Advances in neural information processing systems* **27** (2014).

55. R. Nishihara, I. Murray, R. P. Adams, *The Journal of Machine Learning Research* **15**, 2087 (2014).

56. R. F. Murray, Y. Morgenstern, *Journal of vision* **10**, 15 (2010).

57. R. S. Russell, W. Bernard, *Operations management. Hoboken: Wiley* pp. 497–8 (2006).

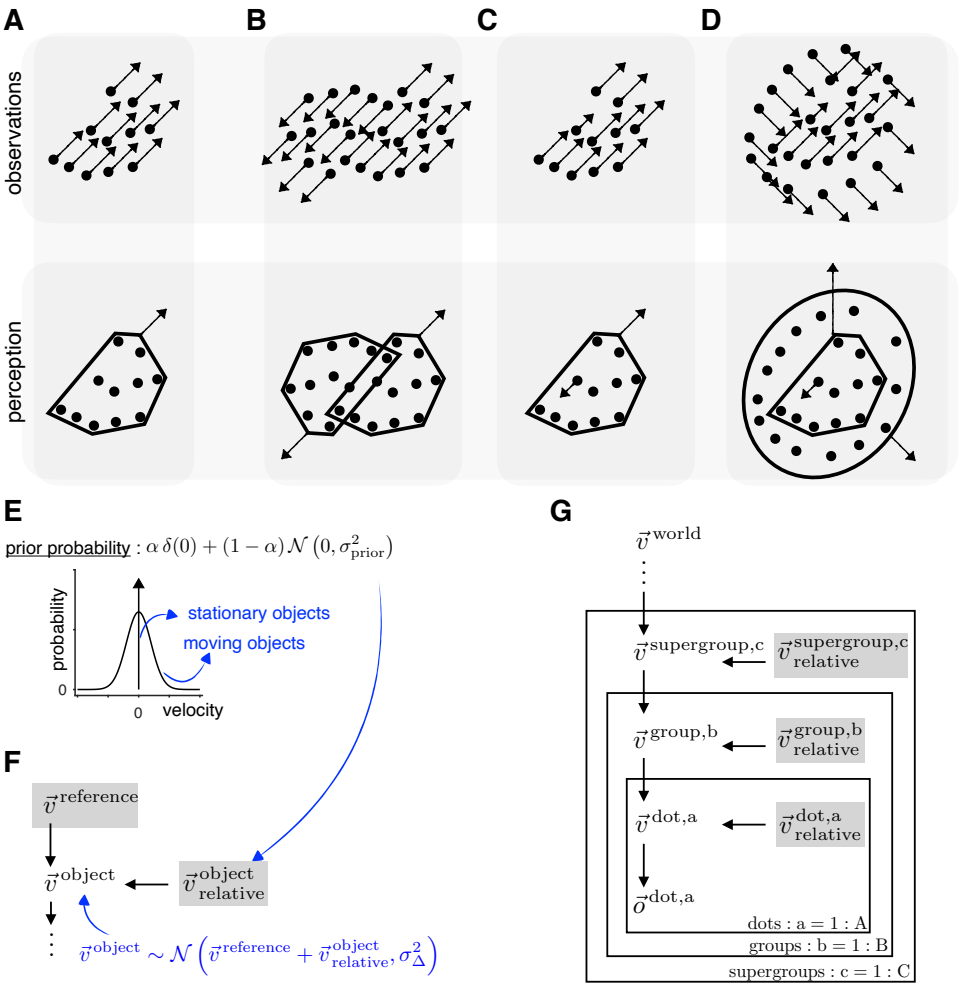


Figure 1: **Causal inference model for hierarchical motion perception.** **(A-D)** Illustration of four different dot patterns that are observed to move as shown in the top row with the arrows indicating retinal velocity vectors. The bottom row shows our predicted motion percept where the brain uses common velocity information to combine dots into objects and group objects into a hierarchical structure. **(E)** Prior that consists of a mixture of a delta distribution at 0 and a Gaussian distribution centered at zero, reflecting the knowledge that elements are either exactly stationary in an appropriate reference frame or are likely to move with a slow speed. **(F)** Generative model motif in which the object's observed retinal velocity is the sum of the reference frame velocity ($\vec{v}^{\,\text{reference}}$) and its velocity with respect to the reference frame ($\vec{v}^{\,\text{object}}_{\text{reference}}$). **(G)** Hierarchical causal inference model obtained by repeatedly applying the motif in F. Inference in this model leads to hierarchical grouping of dots, and representing dot motion in reference frames defined by the groups they belong to. The percept is determined by the non-zero relative variable lowest in the hierarchy.

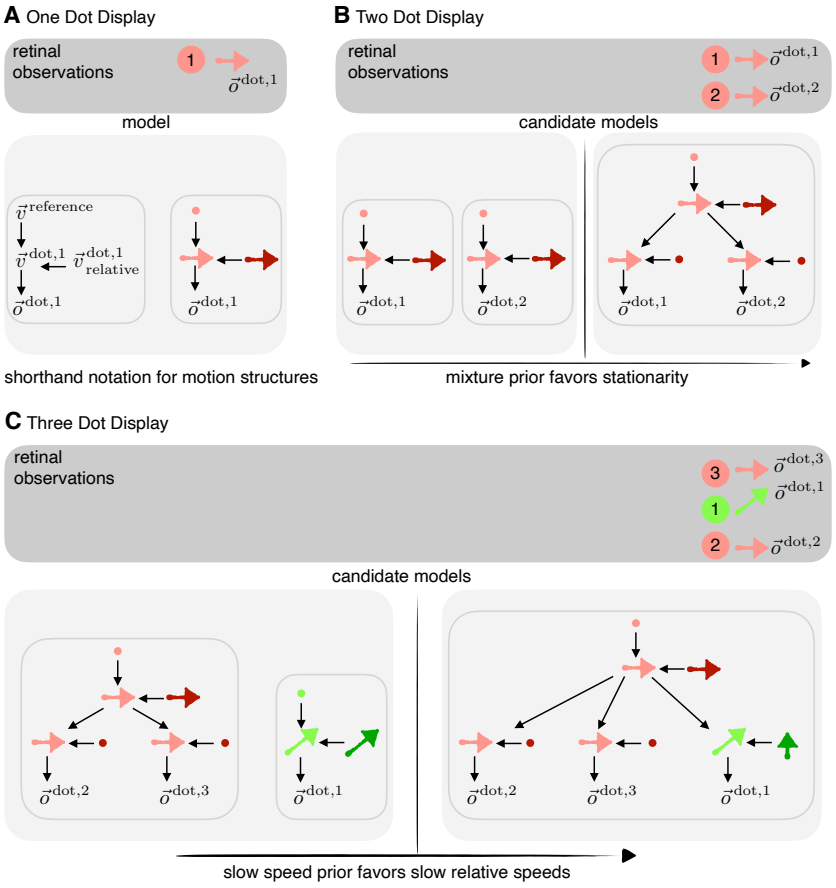


Figure 2: **Model predictions for simple dot stimuli.** **(A)** The motion of a single dot is inferred in the reference frame of a stationary world. In our shorthand notation, velocity variables have been replaced by their most likely value shown as a motion vector. Darker shades are used to indicate relative velocities and filled circles indicate zero velocity. **(B)** Two moving dots are explainable by two possible structures (left and right). If they move coherently, such that both are stationary with respect to a moving group, the delta component in the prior implies that most posterior mass lies on the combined structure (right). As a result we perceive a moving object consisting of both dots. **(C)** If a third dot is added to the display in (B), the observations are explainable by 8 different structures (Methods, Figure S9), two of which are shown here. On the left, the green dot is perceived as independent of, and unaffected by the motion of the red dots. On the right, the green dot is part of the same structure as the red dots, and perceived in a reference frame defined by a group in which two out of the three dots are stationary (favored by the delta component in the prior). The Gaussian component of the prior favors the right structure over the left one since the velocity of the green dot in the reference frame defined by the red dots is smaller than its velocity with respect to the stationary world. This explains the Johansson illusion (*4*).

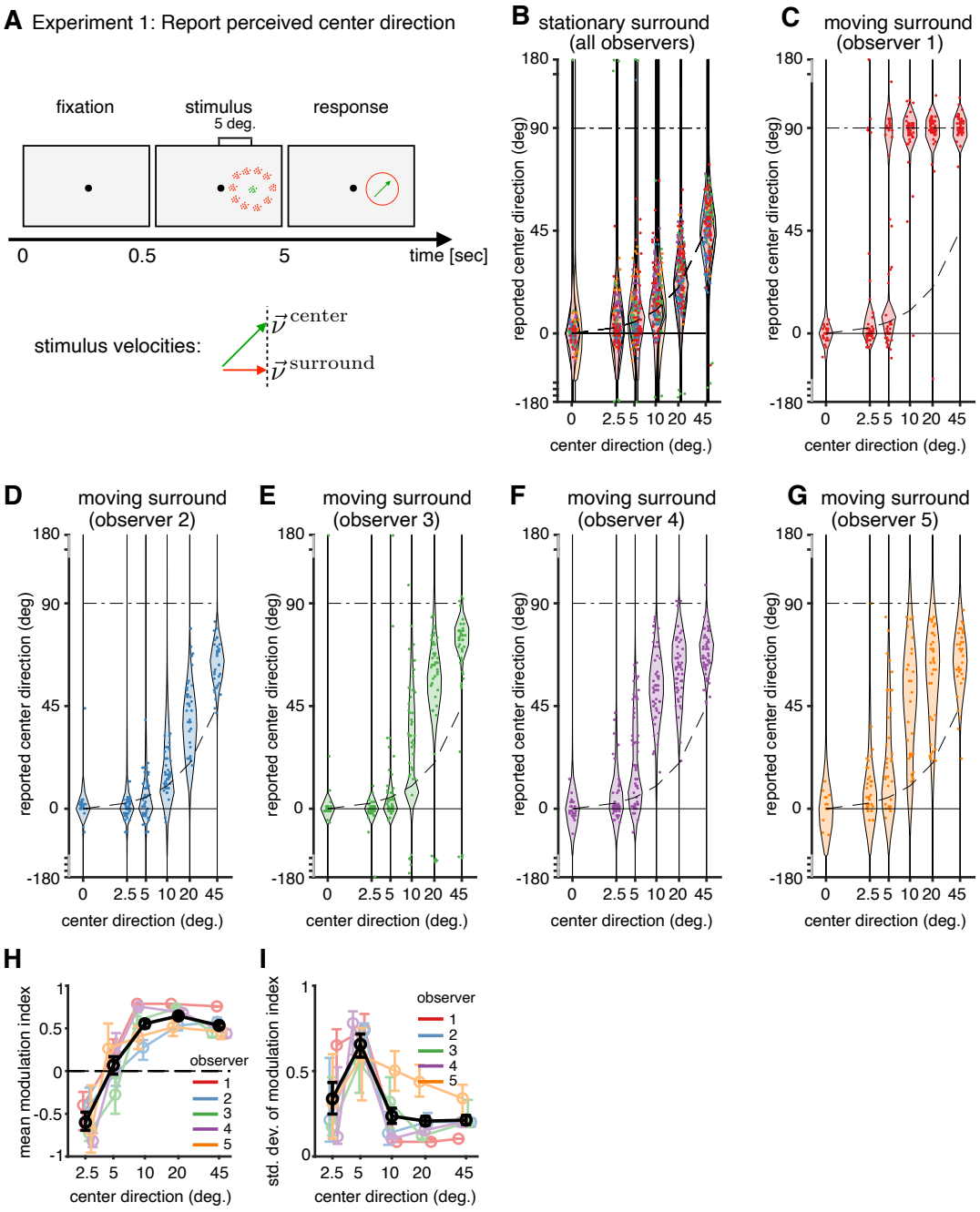

A Experiment 1: Report perceived center direction
fixation
stimulus
5 deg.
response
0
0.5
5
time [sec]
stimulus velocities:
$\vec{v}$center
$\vec{v}$surround
B stationary surround (all observers)
C moving surround (observer 1)
D moving surround (observer 2)
E moving surround (observer 3)
F moving surround (observer 4)
G moving surround (observer 5)
reported center direction (deg)
180
90
45
0
-180
0 2.5 5 10 20 45
center direction (deg.)
H
mean modulation index
1
0.5
0
-0.5
-1
2.5 5 10 20 45
center direction (deg.)
observer
1
2
3
4
5
I
std. dev. of modulation index
1
0.5
0
2.5 5 10 20 45
center direction (deg.)
observer
1
2
3
4
5

Figure 3: **Experiment 1 – design and results.** **(A)** During fixation, two groups of dots (red and green) appear and move back and forth three times for 4.5 seconds before disappearing. During the last phase of the movement, the fixation dot turned green. The observers adjusted a dial to report their perceived direction of the green dots during the last movement phase. The red (surround) dots were either stationary or moved horizontally (0°), while the green (center) dots varied in direction from trial to trial while keeping the horizontal component of their velocity matched with the surround. **(B)** Responses of all five observers overlaid for the condition where the surround is stationary. Each dot represents a single trial. All responses lie around the identity line (warped due to non-linear spacing of the $x-$axis). Responses were flipped for negative center directions to match the positive directions after verifying that the responses were symmetric. **(C-G)** Responses for each observer when the surround is moving. The horizontal lines at 0° and 90° indicate the predicted reports for complete integration (perceiving the surround) and complete segmentation, i.e., perceiving the relative velocity, respectively. **(B-G)** The overlaid violin plots show the model predictions (not data distributions). One model was fit jointly on all data for each observer. **(H,I)** Mean and standard deviation in modulation index (68% confidence intervals) defined such that -1 corresponds to pure integration, $+1$ to pure segmentation, and 0 to retinal motion. Different colors indicate different observers; black line denotes the average across observers.

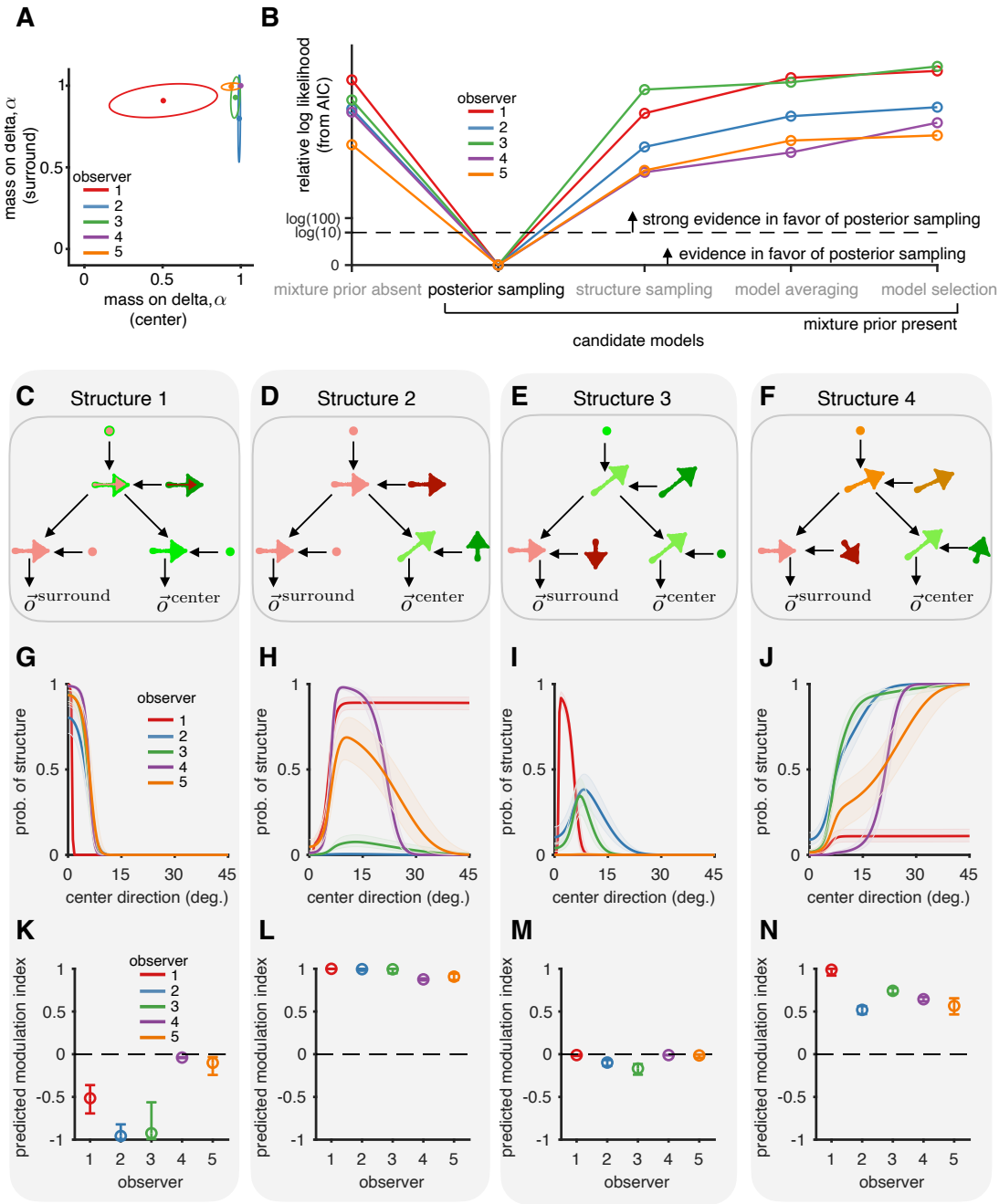

A
B
mass on delta, α (surround)
mass on delta, α (center)
observer
1
2
3
4
5
relative log likelihood (from AIC)
log(100)
log(10)
0
strong evidence in favor of posterior sampling
evidence in favor of posterior sampling
mixture prior absent
posterior sampling
structure sampling
model averaging
model selection
candidate models
mixture prior present
C Structure 1
D Structure 2
E Structure 3
F Structure 4
$\vec{o}^{\text{surround}}$
$\vec{o}^{\text{center}}$
G
H
I
J
prob. of structure
center direction (deg.)
K
L
M
N
predicted modulation index

Figure 4: **Model fitting insights from experiment 1**.**(A)** Fitted mass on the delta component in the prior for center and surround. Each colored line show the mean and 95% CI for each observer. **(B)** Model comparison. For each model, the difference in AIC score to the posterior sampling model is show. We find strong evidence against all alternative models as compared to the posterior sampling model. **(C-F)** The four causal structures that have a non-zero probability in our model fits. (See main text for description.) **(G-J)** The posterior probability assigned to each structure by each of the observers as a function of center direction. All observers integrate the center and surround for center directions close to zero (C+G) and segments the center and surround otherwise. However, they differ to what degree they rely on each of the three different reference frames implied by D, E, and F. **(K-N)** As in Figure 3E, the modulation index predicted under each structure quantifies the influence of the surround on the perceived center direction. The modulation index for a structure is independent of the probability assigned to that structure. Together, they determine the influence of the surround on the center.

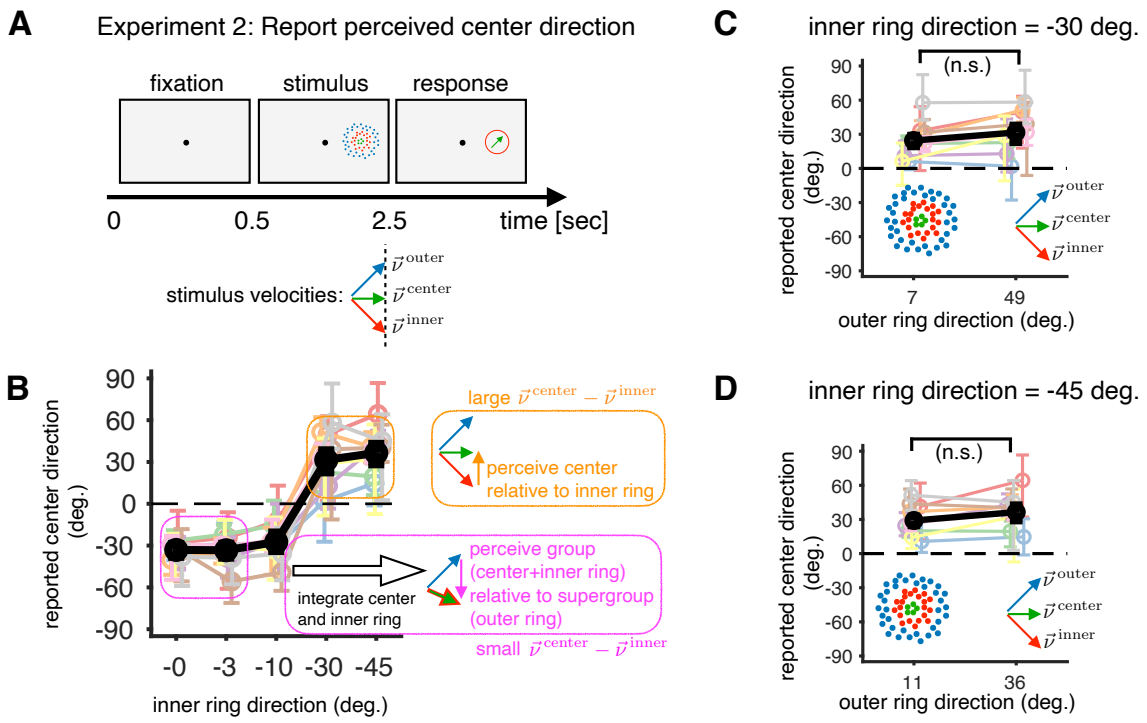


Figure 5: **Design and results of Experiment 2 (three moving groups). (A)** The observer performs an estimation task in which they have to report the direction of green dots (center) using a dial. While center dot directions are randomized from 0 to 360°, results are combined after rotating all velocities such that $\vec{v}^{\text{center}}$ moves horizontally (0°). **(B)** Our model predicts that we will perceive the center dots in the inner ring's reference frame if both move in noticably different directions (segmentation, orange box) and cue combine the center and the inner ring motion if they are sufficiently similar (integration, magenta box). In the latter case, our percept would be the cue-combined center and inner ring's velocity in the outer ring's reference frame. The observer responses support these model predictions. **(C)** The causal inference model predicts that during segmentation, the percept of center motion should only depend on the retinal inner ring motion, not the perceived inner ring motion. The data clearly supports this prediction since the outer ring motion as no influence on the reported center directions (even though it influences the inner ring percepts, see Figure S11). **(D)** Same as (C) but for conditions where the inner ring moves at $-45°$ and the outer ring moves at either $11°$ or $36°$. We consider the two inner velocities that are most different from the center in our experiment to minimize the probability that the observer integrates the center and the inner ring velocities.

Methods

Observers

Five naive observers participated in Experiment 1, and 10 naive observers participated in Experiment 2. Observers provided written informed consent, and were financially compensated for their time. Experiments were approved by the Office for Human Subject Protection (OHSP) at the University of Rochester (IRB number 0003909). 1/10 observers in Experiment 2 was excluded based on their large response variability (standard deviation greater than 30deg) in the control condition and their data was omitted from further analysis.

Experiment 1 details: two moving elements

The stimulus consisted of a 'center patch' of green dots, presented at 5 degree eccentricity. Dots were 0.1 degree in diameter and distributed uniformly within the patch with a density of $6.88\,\mathrm{dots/degree}^2$. The center patch had a radius of 0.68 degrees and was surrounded by a ring of radius 2.72 degrees, consisting of a variable number of patches of red dots (Figure 3A). Each surround patch had a radius of 0.54 degrees and a dot density of $10.91\,\mathrm{dots/degree}^2$. Dot displays were viewed binocularly, and no disparity cues were added, such thatall the dots moved in the plane of the display. The stimuli were presented on a 27-inch monitor with a refresh rate of 60 Hz and a resolution of 1920×1080 at a viewing distance of 105 cm. Eye movements were tracked using an Eyelink 1000 system and trials were discarded in which eyes moved within 1 degree of the center patch.

The stimuli were presented at an eccentricity of 5 degrees in the periphery. The number of dots in a patch was fixed to 10. There was one center patch and the number of surround patches was chosen in every trial from the set [1,2,3,5,10]. The center retinal direction was chosen from the set $[0, \pm 2.5, \pm 5, \pm 10, \pm 20, \pm 45]$ and the surround was either stationary (stationary surround always had 5 patches) or moved at 0 degrees (horizontally rightwards) at a speed of 1 deg/sec.

After a fixation period of 0.5s, the stimulus appeared and moved back and forth for 1.5 cycles. The patch envelopes moved at a constant velocity and reversed their velocity after 1.5 seconds (square wave velocity profile with a time period of 3 seconds). The back-and-forth movement ensured that the envelopes stayed within a fixed area of the screen. In the last half cycle, the fixation dot turned green indicating that the observer had to report the direction during the last half cycle.

After stimulus offset, an arrow appeared at the location of the center patch, and observers used a dial to adjust the arrow direction to match their motion direction percept. The stationary surround condition served as a control. Observers who had a response standard deviation larger than 30 degrees in this condition were removed from subsequent analysis. The fixation period and stimulus had a total duration of 5 seconds following which the observers could make a response at any time to proceed to the next trial. Each observer participated in three sessions to get 22 trials per condition on average for a total of 1446 trials per observer on average.

Experiment 2 details: three moving elements

In experiment 2, the stimulus consisted of a center patch, an inner ring, and an outer ring, all arranged concentrically centered at 5 degrees eccentricity (Figure 5A) separated by 0.5 degrees. On every trial, either the center dots, or the inner ring dots, were colored green, indicating whose direction had to be reported. The other parts of the stimulus contained red and blue dots, respectively, with the color assignment randomly drawn every trial. The number of dots in the center was 10 (density of 3.2 dots/degrees^2 with a center patch radius of 1 degree), the inner ring contained 50 dots (density of 4 dots/degrees^2 with a inner ring width of 1 degree) and the outer ring contained 250 dots(density of 11.4 dots/degrees^2 with a outer ring width of 1 degree). The viewing distance, eye recording details, and the criteria for discarding trials due to fixation breaks and control condition response variability were the same as in experiment 1.

Each trial started with a 0.5 s fixation period during which only the fixation dot was shown on the screen. This was followed by the random-dot stimulus which moved for 2 seconds within an aperture at a constant velocity. Unlike for Experiment 1, the moving dots were presented inside fixed apertures, and the stimulus direction was not reversed. The center moved with a speed of 0.5 degrees/sec and the center direction was randomized across trials, within a range of 360°, to minimize the effect of reporting biases. The inner ring's retinal direction was chosen from the set $[0, -3, -10, -30, -45]$ relative to the center where negative angles indicate clockwise rotations. The outer ring moved in a counter-clockwise direction chosen such that the same relative velocity was maintained between the outer and inner rings across different inner ring directions. Furthermore, the speeds of inner and outer ring were chosen such that the relative velocity between either ring and center was perpendicular to the center direction.

Different conditions were interleaved across trials in which the inner and outer rings could either move randomly or coherently and the observer had to report the direction of the center or inner ring. As in experiment 1, the observer made their responses by adjusting the direction of the arrow that appeared at the location of the center or inner ring after the stimulus offset with its size matched to the size of the corresponding target. Each observer performed three sessions to get 46 trials per condition on average for a total of 2239 trials per observer on average.

Modulation index to summarize observer responses in Experiment 1

We defined the modulation index, w_{MI}, such that the percept of the center patch ($\vec{v}_{\text{percept}}^{\,\text{center}}$) predicted for a given w_{MI} is given by

$$\vec{v}_{\text{percept}}^{\,\text{center}} = \begin{cases} \vec{o}^{\,\text{center}} - w_{\text{MI}}\vec{o}^{\,\text{surround}} & w_{\text{MI}} \geq 0 \\ (1 + w_{\text{MI}})\vec{o}^{\,\text{center}} - w_{\text{MI}}\vec{o}^{\,\text{surround}} & w_{\text{MI}} \leq 0 \end{cases} \tag{1}$$

where $\vec{o}^{\,\text{center}}$ and $\vec{o}^{\,\text{surround}}$ are the observed center and surround velocities from the brain's perspective for a given trial. This definition incorporates partial subtraction of the surround (case $w_{\text{MI}} \geq 0$) for relative velocities when the effect on perception is repulsive (*53*) and cue com-

bination (*31*) when the effect is attractive (case $w_{\mathrm{MI}} \leq 0$). Under the simplifying assumption that $\vec{o}^{\,\mathrm{center}}$ and $\vec{o}^{\,\mathrm{surround}}$ correspond to the experimenter-controlled velocities on the screen (ignoring observation noise), and that the perceptual reports exactly reflect the perceived variable $\vec{v}^{\,\mathrm{center}}_{\mathrm{percept}}$ (ignoring motor noise), one could estimate a per-trial modulation index to obtain a distribution over w_{MI} for each separation of center and surround in order to estimate means and standard deviations which unfortunately would be biased. In order to obtain the unbiased estimates reported in Figures 3H and I, we therefore modeled observation noise, motor noise, and motor bias explicitly which allowed us to infer the distribution over w_{MI} from the distribution over perceptual reports (see supplementary section S1 for details).

Causal inference model for hierarchical motion perception

Generative model in a scene with two moving elements

The observed retinal velocities, $\vec{o}^{\,\mathrm{center}}$ and $\vec{o}^{\,\mathrm{surround}}$, were modeled as the true retinal velocities, $\vec{v}^{\,\mathrm{center}}$ and $\vec{v}^{\,\mathrm{surround}}$ corrupted by additive Gaussian noise with variance $\sigma^2_{\mathrm{center}}$ and $\sigma^2_{\mathrm{surround}}$, respectively ($\mathbb{I}$ refers to a 2×2 identity matrix) :

$$\vec{o}^{\,\mathrm{center}} \sim \mathcal{N}(\vec{v}^{\,\mathrm{center}}, \sigma^2_{\mathrm{center}}\,\mathbb{I}) \quad \text{and} \quad \vec{o}^{\,\mathrm{surround}} \sim \mathcal{N}(\vec{v}^{\,\mathrm{surround}}, \sigma^2_{\mathrm{surround}}\,\mathbb{I}). \tag{2}$$

The velocities were parameterized as two dimensional vectors reflecting the x and y components. In order to model the inference over the different causal structures (here, whether center and surround are part of the same moving group), we follow (*26*) in introducing a binary (logical) variable $S^{\mathrm{center,surround}} \in \{0, 1\}$ (corresponding to the left and right side of Figures 2B and S2B). This allows us to write the conditional probabilities compactly as:

$$\vec{v}^{\,\mathrm{center}} \sim \mathcal{N}(S^{\mathrm{center,surround}}\vec{v}^{\,\mathrm{group}} + \vec{v}^{\,\mathrm{center}}_{\mathrm{relative}} + (1 - S^{\mathrm{center,surround}})\vec{v}^{\,\mathrm{world}}, \sigma^2_{\Delta}\,\mathbb{I}) \quad \text{and} \tag{3}$$

$$\vec{v}^{\,\mathrm{surround}} \sim \mathcal{N}(S^{\mathrm{center,surround}}\vec{v}^{\,\mathrm{group}} + \vec{v}^{\,\mathrm{surround}}_{\mathrm{relative}} + (1 - S^{\mathrm{center,surround}})\vec{v}^{\,\mathrm{world}}, \sigma^2_{\Delta}\,\mathbb{I}). \tag{4}$$

where σ^2_{Δ} models the uncertainty in the velocity composition, e.g. due to computational noise. The prior over $S^{\mathrm{center,surround}}$ is given by $\beta^{\mathrm{center,surround}}$ which represents the prior probability that center and surround belong to a common structure (based on prior experience, or other non-motion cues). In a scene with only two moving elements, the group velocity is inferred in the stationary world reference frame with $\vec{v}^{\,\mathrm{world}} = 0$ (Figure 1G):

$$\vec{v}^{\,\mathrm{group}} \sim \mathcal{N}(\vec{v}^{\,\mathrm{group}}_{\mathrm{relative}} + \vec{v}^{\,\mathrm{world}}, \sigma^2_{\Delta}\,\mathbb{I}). \tag{5}$$

The prior over each of the relative velocities is a mixture prior of a delta function at zero and a normal distribution centered at zero

$$\vec{v}^{\,\mathrm{center/\,surround/\,group}}_{\mathrm{relative}} \sim \alpha\,\delta(0) + (1 - \alpha)\mathcal{N}(0, \sigma^2_{\mathrm{prior}}\,\mathbb{I}) \tag{6}$$

where α represents the expectation that the (relative) motion is exactly zero (Figure 1E,F).

Mapping inferred latent variables to percepts

Our model predicts that if the center is inferred to be part of the same group a the surround, then the perceived center motion corresponds to the center's relative velocity if it is inferred to be nonzero, and to the group velocity if the center's relative velocity is inferred to be zero. If the center is not inferred to be part of the same group as the surround, then its motion relative to stationary world is perceived. By defining $C^{(\cdots)} \in \{0, 1\}$ to denote whether the velocity $\vec{v}^{(\cdots)}_{\text{relative}}$ is zero ($C^{(\cdots)} = 0$) or not ($C^{(\cdots)} = 1$), we can compactly write the percept as:

$$\vec{v}^{\,\text{center}}_{\text{percept}} = C^{\text{center}}\vec{v}^{\,\text{center}}_{\text{relative}} + (1 - C^{\text{center}})S^{\text{center,surround}}C^{\text{group}}\vec{v}^{\,\text{group}}_{\text{relative}}. \tag{7}$$

Therefore, the distribution over the observer's percept is a mixture distribution with the mixture weights corresponding to the posterior probability of the different causal structures (characterized by possible S) and the different nested structures within each causal structure (characterized by possible C):

$$\begin{aligned} p(\vec{v}^{\,\text{center}}_{\text{percept}}|\underline{\vec{o}}) = &\sum_{c_1=\{0,1\}}\sum_{c_2=\{0,1\}}\sum_{c_3=\{0,1\}}\sum_{c_4=\{0,1\}} \\ &p(C^{\text{center}} = c_1, C^{\text{surround}} = c_2, C^{\text{group}} = c_3, S^{\text{center,surround}} = c_4|\underline{\vec{o}}) \\ &p(\vec{v}^{\,\text{center}}_{\text{percept}}|\underline{\vec{o}}, C^{\text{center}} = c_1, C^{\text{surround}} = c_2, C^{\text{group}} = c_3, S^{\text{center,surround}} = c_4). \end{aligned} \tag{8}$$

where for compactness we define $\underline{\vec{o}} \equiv (\vec{o}^{\,\text{center}}, \vec{o}^{\,\text{surround}})$. For any number of moving elements, this posterior has the general form:

$$p(\vec{v}^{\,\text{center}}_{\text{percept}}|\underline{\vec{o}}) = \sum_{i}^{N} w_i(\underline{\vec{o}}_j)\mathcal{N}(\vec{v}^{\,\text{center}}_{\text{percept}}; \vec{\mu}_i(\underline{\vec{o}}_j), \sigma_i^2\,\mathbb{I}) \tag{9}$$

where w_i, μ_i, and σ_i^2 correspond to the probability, mean and variance associated with each causal structure, respectively. For n moving elements, the total number of mixture components in the posterior is $N = \#S(n)\,2^{\#C(n)}$ where $\#S(n)$ is the number of possible causal structures and $\#C(n)$ is the maximum possible number of relative velocities (each of which could be zero or not thereby resulting in $2^{\#C(n)}$ nested structures). $\#S(n)$ satisfies the recurrence relation given by $\#S(n) = 2+\sum_{k=2}^{N-1}\binom{n}{k}\#S(n-k+1)$ as in addition to all elements being independent or part of the same group, grouping k elements results in $n-k+1$ moving groups. $\#C(n) = \frac{n(n+1)}{2}$ as at each hierarchical level i, we can have $n-i$ moving elements where $i \in \{0, N-1\}$. The expressions for the terms in Eq. 9 are derived in supplementary section S2.

The distribution over perceived velocity can be mapped onto the distribution over the perceived center direction, $\theta^{\text{center}}_{\text{percept}}$, by mapping each velocity to its corresponding direction:

$$p(\theta^{\text{center}}_{\text{percept}}|\underline{\vec{o}}) = \int_{\vec{v}^{\,\text{center}}_{\text{percept}}} p(\theta|\vec{v}^{\,\text{center}}_{\text{percept}})p(\vec{v}^{\,\text{center}}_{\text{percept}}|\underline{\vec{o}}). \tag{10}$$

Inserting Eq. (9) into (10), we can approximate the integral to get:

$$p(\theta^{\text{center}}_{\text{percept}}|\underline{\vec{o}}_j) = \sum_{i=1}^{N} w_i(\underline{\vec{o}})\mathcal{N}_{\text{circular}}\{\theta; \hat{\theta}_i(\underline{\vec{o}}), \kappa_i\}. \tag{11}$$

$\mathcal{N}_{\text{circular}}$ represents the von Mises distribution pdf with mean parameter $\hat{\theta}_i(\underline{\vec{o}}) = \arctan(\vec{\mu}_{i,y}, \vec{\mu}_{i,x})$ and concentration parameter $\kappa_i = T_2(\sigma_i^2/||\vec{\mu}_i||_2^2)$ which can be computed numerically. We allow for lapses in responses by adding another component to the distribution over the center direction that is a von Mises distribution characterized by μ^{lapse} and κ^{lapse}.

$$p(\theta^{\text{center}}_{\text{percept}}|\underline{\vec{o}}) = \lambda\mathcal{N}_{\text{circular}}(\theta; \mu^{\text{lapse}}, \kappa^{\text{lapse}}) + (1-\lambda)\sum_{i=1}^{N} w_i(\underline{\vec{o}})\mathcal{N}_{\text{circular}}\{\theta; \hat{\theta}_i(\underline{\vec{o}}), \kappa_i\}. \tag{12}$$

Perceptual estimation

We consider the following four, previously proposed (*26, 35*), ways in which the brain may convert the posterior $p(\theta^{\text{center}}_{\text{percept}}|\underline{\vec{o}})$ into a perceptual point estimate, $\theta^{\text{center}}_{\text{estimate}}$:

Model averaging: $\theta^{\text{center}}_{\text{estimate}}$ is the mean of the joint posterior over all structures: $\theta^{\text{center}}_{\text{estimate}} = \sum_{i=0}^{N} w_i'\mu_i'(\underline{\vec{o}})$.

Model selection: $\theta^{\text{center}}_{\text{estimate}}$ is the posterior mean over direction for the most likely structure: $\theta^{\text{center}}_{\text{estimate}} = \mu_{i^*}'(\underline{\vec{o}})$ where $i^* = \arg\max w_i'$.

Structure sampling: $\theta^{\text{center}}_{\text{estimate}}$ is the posterior mean over direction for a single structure sampled from the posterior over structures: $\theta^{\text{center}}_{\text{estimate}} = \mu_i'(\underline{\vec{o}})$ where probability of $\mu_i'(\underline{\vec{o}}) \propto w_i'(\underline{\vec{o}})$.

Posterior Sampling: $\theta^{\text{center}}_{\text{estimate}}$ is a direction sampled from the joint posterior over all structures and directions: $\theta^{\text{center}}_{\text{estimate}} \sim p(\theta^{\text{center}}_{\text{percept}}|\underline{\vec{o}})$.

Predicted distribution over observer responses

The distribution over observer reports, R, for a set of experimenter-defined directions $\underline{\vec{\nu}} \equiv (\vec{\nu}^{\text{center}}_{\text{retina}}, \vec{\nu}^{\text{surround}}_{\text{retina}})$ can be obtained by marginalizing out all possible sensory observations (*26, 54*) :

$$p(R|\underline{\vec{\nu}}) = \int_{\underline{\vec{o}}} p(R|\underline{\vec{o}})p(\underline{\vec{o}}|\underline{\vec{\nu}}). \tag{13}$$

We model the distribution over observer reports, R, as a von Mises distribution centered on $\theta^{\text{center}}_{\text{estimate}}$ allowing for a reporting bias b and a motor noise κ_m:

$$p(R|\underline{\nu}) = \int_{\underline{\vec{o}}} \mathcal{N}_{\text{circular}}(R; \theta^{\text{center}}_{\text{estimate}} + b, \kappa_m) p(\underline{\vec{o}}|\underline{\vec{\nu}}). \tag{14}$$

Since this integral is intractable to evaluate analytically, we approximate this using Gaussian quadratures evaluated at points $\vec{o_j}$ with weights w_j^{quad}:

$$p(R|\underline{\nu}) = \sum_j w_j^{\text{quad}} \mathcal{N}_{\text{circular}}(R; \theta^{\text{center}}_{\text{estimate}} + b, \kappa_m). \tag{15}$$

The distribution over observer responses for different perceptual estimates described in the previous section are give in supplementary section S3.

Model fitting details

We obtained the maximum a posteriori (MAP; for initializing the sampler) and maximum likelihood estimate (MLE; to compute AIC) for the model parameters under weakly informative priors (details in Table S1) using a quasi-newton Broyden-Fletcher-Goldfarb-Shanno (BFGS) unconstrained optimization procedure (fminunc in MATLAB). We obtained full posteriors over all model parameters using generalized elliptical slice sampling (*55*) which allowed us to get uncertainty estimates for all parameter estimates. We used 144 chains with 25000 samples per chain to estimate the posterior distribution over the parameters (average $\hat{R} \leq 1.1$).

To evaluate the absolute goodness of fit of the model, for each combination of center and surround velocities, we compared the empirical CDF from the reported directions with the corresponding model prediction. We quantified the overall match as variance explained across all conditions.

Supplementary Text

S1 Estimating the distribution over modulation indices

In this section we present the details of the procedure to compute the distribution of modulation indices described in the main text. The predicted distribution over reports across trials for a given combination of veridical center and surround velocities can be expanded for a given modulation index w_{MI} assuming Gaussian observation noise:

$$p(R|\vec{\nu}^{\,\text{center}}, \vec{\nu}^{\,\text{surround}}, w_{\text{MI}}) = \begin{cases} \mathcal{N}(R; b + \vec{\nu}^{\,\text{center}} - w_{\text{MI}}\vec{\nu}^{\,\text{surround}}, \\ \sigma^2_{\text{motor}} + \sigma^2_{\text{center}} + w^2_{\text{MI}}\sigma^2_{\text{surround}}) & w_{\text{MI}} \geq 0 \\ \mathcal{N}(R; b + (1 + w_{\text{MI}})\vec{\nu}^{\,\text{center}} - w_{\text{MI}}\vec{\nu}^{\,\text{surround}}, \\ \sigma^2_{\text{motor}} + (1 + w_{\text{MI}})^2\sigma^2_{\text{center}} + w^2_{\text{MI}}\sigma^2_{\text{surround}}) & w_{\text{MI}} \leq 0 \end{cases} \tag{1}$$

with the corresponding CDF

$$\text{CDF}(R|\vec{\nu}^{\,\text{center}}, \vec{\nu}^{\,\text{surround}}, w_{\text{MI}}) = \begin{cases} \Phi(R; b + \vec{\nu}^{\,\text{center}} - w_{\text{MI}}\vec{\nu}^{\,\text{surround}}, \\ \sigma^2_{\text{motor}} + \sigma^2_{\text{center}} + w^2_{\text{MI}}\sigma^2_{\text{surround}}) & w_{\text{MI}} \geq 0 \\ \Phi(R; b + (1 + w_{\text{MI}})\vec{\nu}^{\,\text{center}} - w_{\text{MI}}\vec{\nu}^{\,\text{surround}}, \\ \sigma^2_{\text{motor}} + (1 + w_{\text{MI}})^2\sigma^2_{\text{center}} + w^2_{\text{MI}}\sigma^2_{\text{surround}}) & w_{\text{MI}} \leq 0 \end{cases} \tag{2}$$

where Φ denotes the normal CDF. The variance over reports in the stationary surround condition, given by $\text{Var}(R_{\text{stationary}}) = \sigma^2_{\text{motor}} + \sigma^2_{\text{center}}$ is an upper bound on the σ^2_{center} (as $\sigma^2_{\text{motor}} \geq 0$). This allowed us to parameterize $\sigma^2_{\text{center}} = \text{Var}(R_{\text{stationary}})c_{\text{center}}$ and $\sigma^2_{\text{surround}} = \text{Var}(R_{\text{stationary}})c_{\text{surround}}$ where $0 \leq c_{\text{center}}, c_{\text{surround}} \leq 1$. In order to estimate c_{center}, c_{surround}, and π^i_{MI}, we minimized the L2 norm between the predicted CDF over the reported directions and the empirically estimated CDF in the moving surround condition. The predicted CDF can be expanded as a mixture of the CDF predicted for each w_{MI} weighted by the corresponding probability:

$$\text{CDF}(R|\vec{\nu}^{\,\text{center}}, \vec{\nu}^{\,\text{surround}}) = \sum_{i=1}^{33} \text{CDF}(R|\vec{\nu}^{\,\text{center}}, \vec{\nu}^{\,\text{surround}}, w^i_{\text{MI}})\pi^i_{\text{MI}}\Delta w_{\text{MI}}. \tag{3}$$

w^i_{MI} are the bin centers separated by Δw_{MI}. We assumed a weakly informative lognormal prior over σ^2_{center} and $\sigma^2_{\text{surround}}$ (with logarithmic mean -3.75 and standard deviation 1.15). We also added two costs on π^i_{MI} as regularization to ensure convergence: (a) L1 regularization penalizing the sum of absolute values of π^i_{MI} that ensured that π^i_{MI} corresponding to non-required regions of w^i_{MI} went to zero, and (b) regularization on the curvature (second derivative) to ensure smooth distributions over w^i_{MI}. The regularization constants λ_1 and λ_2 corresponding to the two terms of the regularization were estimated by cross validation on synthetic ground truth data.

S2 Inference in the hierarchical causal inference model

In this section, we derive the expressions for posteriors over the inferred latents in our causal inference model. Eq. 9 in the main text can be expanded in the context of our model as

$$p(\vec{v}^{\text{center}}_{\text{percept}}|\underline{\vec{o}}) = \sum_{c_1=\{0,1\}} \sum_{c_2=\{0,1\}} \sum_{c_3=\{0,1\}} \sum_{c_4=\{0,1\}} p(C^{\text{center}} = c_1, C^{\text{surround}} = c_2, C^{\text{group}} = c_3, S^{\text{center,surround}} = c_4|\underline{\vec{o}}) \; p(\vec{v}^{\text{center}}_{\text{percept}}|\underline{\vec{o}}, C^{\text{center}} = c_1, C^{\text{surround}} = c_2, C^{\text{group}} = c_3, S^{\text{center,surround}} = c_4). \tag{4}$$

where $w_i(\underline{\vec{o}}_j) = p(C^{\text{center}} = c_1, C^{\text{surround}} = c_2, C^{\text{group}} = c_3, S^{\text{center,surround}} = c_4|\underline{\vec{o}})$, $\mu_i(\underline{\vec{o}}_j)$ and σ_i^2 are the mean and variance of $p(\vec{v}^{\text{center}}_{\text{percept}}|\underline{\vec{o}}, C^{\text{center}} = c_1, C^{\text{surround}} = c_2, C^{\text{group}} = c_3, S^{\text{center,surround}} = c_4)$ respectively.

To get the required expression for Eq. 4, it is sufficient to derive expressions for

$$p(C^{\text{ center}}, S^{\text{center,surround}}, C^{\text{group}}, C^{\text{ surround}}|\vec{o}^{\text{ center}}, \vec{o}^{\text{ surround}}) \tag{5}$$

$$p(\vec{v}^{\text{group}}_{\text{relative}}|\vec{o}^{\text{ center}}, \vec{o}^{\text{ surround}}, C^{\text{ center}}, S^{\text{center,surround}}, C^{\text{group}}, C^{\text{ surround}}) \tag{6}$$

$$p(\vec{v}^{\text{center}}_{\text{relative}}|\vec{o}^{\text{ center}}, \vec{o}^{\text{ surround}}, C^{\text{ center}}, S^{\text{center,surround}}, C^{\text{group}}, C^{\text{ surround}}) \tag{7}$$

as

$$p(\vec{v}^{\text{center}}_{\text{percept}}|\vec{o}^{\text{ center}}, \vec{o}^{\text{ surround}}, C^{\text{ center}} = 0, S^{\text{center,surround}} = 1, C^{\text{group}} = 1, C^{\text{ surround}} \in \{0,1\}) = p(\vec{v}^{\text{group}}_{\text{relative}}|\vec{o}^{\text{ center}}, \vec{o}^{\text{ surround}}, C^{\text{ center}} = 0, S^{\text{center,surround}} = 1, C^{\text{group}} = 1, C^{\text{ surround}} \in \{0,1\}) \tag{8}$$

$$p(\vec{v}^{\text{center}}_{\text{percept}}|\vec{o}^{\text{ center}}, \vec{o}^{\text{ surround}}, C^{\text{ center}} = 1, S^{\text{center,surround}} \in \{0,1\}, C^{\text{group}} \in \{0,1\}, C^{\text{ surround}} \in \{0,1\}) = p(\vec{v}^{\text{center}}_{\text{relative}}|\vec{o}^{\text{ center}}, \vec{o}^{\text{ surround}}, C^{\text{ center}} = 1, S^{\text{center,surround}} \in \{0,1\}, C^{\text{group}} \in \{0,1\}, C^{\text{ surround}} \in \{0,1\}) \tag{9}$$

The joint distribution over the latents and observations for a given structure can be expanded in terms of their definitions (main text)

$$p(\vec{o}^{\,\text{center}}, \vec{o}^{\,\text{surround}}, \vec{v}^{\text{group}}_{\text{relative}}, \vec{v}^{\text{center}}_{\text{relative}}, \vec{v}^{\text{surround}}_{\text{relative}}, \vec{v}^{\text{group}} | C^{\,\text{center}}, S^{\text{center,surround}}, C^{\text{group}}, C^{\,\text{surround}}) =$$
$$\mathcal{N}(\vec{o}^{\,\text{center}} - \vec{v}^{\text{center}}_{\text{relative}}; S^{\text{center,surround}}\vec{v}^{\text{group}}, \sigma^2_{\text{center}} + \sigma^2_\Delta)$$
$$\mathcal{N}(\vec{o}^{\,\text{surround}} - \vec{v}^{\text{surround}}_{\text{relative}}; S^{\text{center,surround}}\vec{v}^{\text{group}}, \sigma^2_{\text{surround}} + \sigma^2_\Delta)$$
$$\mathcal{N}(\vec{v}^{\text{group}}; \vec{v}^{\text{group}}_{\text{relative}}, \sigma^2_\Delta)\mathcal{N}(\vec{v}^{\text{group}}_{\text{relative}}; 0, C^{\text{group}}\sigma^2_{\text{prior}})$$
$$\mathcal{N}(\vec{v}^{\text{center}}_{\text{relative}}; 0, C^{\,\text{center}}\sigma^2_{\text{prior}})\mathcal{N}(\vec{v}^{\text{surround}}_{\text{relative}}; 0, C^{\,\text{surround}}\sigma^2_{\text{prior}}) \tag{10}$$

We have assumed the inferred $\vec{v}^{\,\text{world}} = 0$ since our stimuli in the experiment are local moving patches and are unlikely to introduce non-zero self motion velocities. Marginalizing out the center and surround velocities

$$p(\vec{o}^{\,\text{center}}, \vec{o}^{\,\text{surround}}, \vec{v}^{\text{group}}_{\text{relative}}, \vec{v}^{\text{group}} | C^{\,\text{center}}, S^{\text{center,surround}}, C^{\text{group}}, C^{\,\text{surround}}) =$$
$$\mathcal{N}(\vec{o}^{\,\text{center}}; S^{\text{center,surround}}\vec{v}^{\text{group}}, \sigma^2_{\text{center}} + \sigma^2_\Delta + C^{\,\text{center}}\sigma^2_{\text{prior}})$$
$$\mathcal{N}(\vec{o}^{\,\text{surround}}; S^{\text{center,surround}}\vec{v}^{\text{group}}, \sigma^2_{\text{surround}} + \sigma^2_\Delta + C^{\,\text{surround}}\sigma^2_{\text{prior}})$$
$$\mathcal{N}(\vec{v}^{\text{group}}; \vec{v}^{\text{group}}_{\text{relative}}, \sigma^2_\Delta)\mathcal{N}(\vec{v}^{\text{group}}_{\text{relative}}; 0, C^{\text{group}}\sigma^2_{\text{prior}}) \tag{11}$$

By defining

$$\sigma^2_5 = \sigma^2_{\text{center}} + \sigma^2_{\text{surround}} + 2\sigma^2_\Delta + C^{\,\text{center}}\sigma^2_{\text{prior}} + C^{\,\text{surround}}\sigma^2_{\text{prior}} \tag{12}$$

$$\gamma_5 = \frac{\sigma^2_{\text{surround}} + \sigma^2_\Delta + C^{\,\text{surround}}\sigma^2_{\text{prior}}}{\sigma^2_5} \tag{13}$$

we can apply the formula for product of Gaussian pdf to get

$$p(\vec{o}^{\,\text{center}}, \vec{o}^{\,\text{surround}}, \vec{v}^{\text{group}}_{\text{relative}} | C^{\,\text{center}}, S^{\text{center,surround}}, C^{\text{group}}, C^{\,\text{surround}}) = \mathcal{N}(\vec{o}^{\,\text{center}}; \vec{o}^{\,\text{surround}}, \sigma^2_5)$$
$$\mathcal{N}(\vec{o}^{\,\text{center}}\gamma_5 + \vec{o}^{\,\text{surround}}(1-\gamma_5); S^{\text{center,surround}}\vec{v}^{\text{group}}_{\text{relative}}, \sigma^2_5\gamma_5(1-\gamma_5) + S^{\text{center,surround}}\sigma^2_\Delta)$$
$$\mathcal{N}(\vec{v}^{\text{group}}_{\text{relative}}; 0, C^{\text{group}}\sigma^2_{\text{prior}}) \tag{14}$$

Similarly by defining

$$\sigma^2_6 = \sigma^2_5\gamma_5(1-\gamma_5) + S^{\text{center,surround}}(\sigma^2_\Delta + C^{\text{group}}\sigma^2_{\text{prior}}) \tag{15}$$

$$\gamma_6 = \frac{S^{\text{center,surround}} C^{\text{group}} \sigma^2_{\text{prior}}}{\sigma_6^2} \tag{16}$$

we can apply the product of Gaussian pdf to get

$$\begin{aligned} p(\vec{o}^{\text{ center}}, \vec{o}^{\text{ surround}}, \vec{v}^{\text{group}}_{\text{relative}} | C^{\text{ center}}, S^{\text{center,surround}}, C^{\text{group}}, C^{\text{ surround}}) = \\ \mathcal{N}(\vec{o}^{\text{ center}}; \vec{o}^{\text{ surround}}, \sigma_5^2) \mathcal{N}(\vec{o}^{\text{ center}} \gamma_5 + \vec{o}^{\text{ surround}} (1 - \gamma_5); 0, \sigma_6^2) \\ \mathcal{N}(\vec{v}^{\text{group}}_{\text{relative}}; \vec{o}^{\text{ center}} \gamma_5 \gamma_6 + \vec{o}^{\text{ surround}} (1 - \gamma_5) \gamma_6, C^{\text{group}} \sigma^2_{\text{prior}} (1 - S^{\text{center,surround}} \gamma_6)) \end{aligned} \tag{17}$$

We can substitute the above equation into the likelihood for the posterior over the group relative velocity. By marginalizing the group relative velocity, we can also substitute the above equation to the posterior over the different causal structures

$$\begin{aligned} p(\vec{v}^{\text{group}}_{\text{relative}} | \vec{o}^{\text{ center}}, \vec{o}^{\text{ surround}}, C^{\text{ center}}, S^{\text{center,surround}}, C^{\text{group}}, C^{\text{ surround}}) = \\ \mathcal{N}(\vec{v}^{\text{group}}_{\text{relative}}; \vec{o}^{\text{ center}} \gamma_5 \gamma_6 + \vec{o}^{\text{ surround}} (1 - \gamma_5) \gamma_6, C^{\text{group}} \sigma^2_{\text{prior}} (1 - S^{\text{center,surround}} \gamma_6)) \end{aligned} \tag{18}$$

$$\begin{aligned} p(C^{\text{ center}}, S^{\text{center,surround}}, C^{\text{group}}, C^{\text{ surround}} | \vec{o}^{\text{ center}}, \vec{o}^{\text{ surround}}) = \\ \mathcal{N}(\vec{o}^{\text{ center}}; \vec{o}^{\text{ surround}}, \sigma_5^2) \mathcal{N}(\vec{o}^{\text{ center}} \gamma_5 + \vec{o}^{\text{ surround}} (1 - \gamma_5); 0, \sigma_6^2) \\ (1-\alpha)^{(C^{\text{ center}} + C^{\text{group}} + C^{\text{ surround}})} (\alpha)^{(3 - C^{\text{ center}} - C^{\text{group}} - C^{\text{ surround}})} \beta_{\text{gr}}^{S^{\text{center,surround}}} (1-\beta_{\text{gr}})^{(1 - S^{\text{center,surround}})} \end{aligned} \tag{19}$$

To get the posterior over the center relative velocity, we start with joint distributions over the latents for each causal structures and repeatedly apply the Gaussian pdf products and marginalize out the variables other than $\vec{v}^{\text{center}}_{\text{relative}}$

$$p(\vec{o}^{\,\text{center}}, \vec{o}^{\,\text{surround}}, \vec{v}^{\text{group}}_{\text{relative}}, \vec{v}^{\text{center}}_{\text{relative}}, \vec{v}^{\text{surround}}_{\text{relative}}, \vec{v}^{\text{group}} | C^{\,\text{center}}, S^{\text{center,surround}}, C^{\text{group}}, C^{\,\text{surround}}) =$$
$$\mathcal{N}(\vec{o}^{\,\text{center}} - \vec{v}^{\text{center}}_{\text{relative}}; S^{\text{center,surround}}\vec{v}^{\text{group}}, \sigma^2_{\text{center}} + \sigma^2_{\Delta})$$
$$\mathcal{N}(\vec{o}^{\,\text{surround}} - \vec{v}^{\text{surround}}_{\text{relative}}; S^{\text{center,surround}}\vec{v}^{\text{group}}, \sigma^2_{\text{surround}} + \sigma^2_{\Delta})$$
$$\mathcal{N}(\vec{v}^{\text{group}}; \vec{v}^{\text{group}}_{\text{relative}}, \sigma^2_{\Delta})\mathcal{N}(\vec{v}^{\text{group}}_{\text{relative}}; 0, C^{\text{group}}\sigma^2_{\text{prior}})\mathcal{N}(\vec{v}^{\text{center}}_{\text{relative}}; 0, C^{\,\text{center}}\sigma^2_{\text{prior}})\mathcal{N}(\vec{v}^{\text{surround}}_{\text{relative}}; 0, C^{\,\text{surround}}\sigma^2_{\text{prior}})$$

(20)

By defining

$$\sigma_1^2 = \sigma^2_{\text{surround}} + \sigma^2_{\text{center}} + 2\sigma^2_{\Delta} \tag{21}$$

$$\gamma_1 = \frac{\sigma^2_{\text{surround}} + \sigma^2_{\Delta}}{\sigma_1^2} \tag{22}$$

we can apply the formula for product of Gaussian pdf to get

$$p(\vec{o}^{\,\text{center}}, \vec{o}^{\,\text{surround}}, \vec{v}^{\text{group}}_{\text{relative}}, \vec{v}^{\text{center}}_{\text{relative}}, \vec{v}^{\text{surround}}_{\text{relative}} | C^{\,\text{center}}, S^{\text{center,surround}}, C^{\text{group}}, C^{\,\text{surround}}) =$$
$$\mathcal{N}(\vec{o}^{\,\text{center}} - \vec{v}^{\text{center}}_{\text{relative}}; \vec{o}^{\,\text{surround}} - \vec{v}^{\text{surround}}_{\text{relative}}, \sigma_1^2)$$
$$\mathcal{N}(S^{\text{center,surround}}\vec{v}^{\text{group}}_{\text{relative}}; (\vec{o}^{\,\text{center}} - \vec{v}^{\text{center}}_{\text{relative}})\gamma_1 + (\vec{o}^{\,\text{surround}} - \vec{v}^{\text{surround}}_{\text{relative}})(1-\gamma_1), \sigma_1^2\gamma_1(1-\gamma_1) + S^{\text{center,surround}}\sigma^2_{\Delta})$$
$$\mathcal{N}(\vec{v}^{\text{group}}_{\text{relative}}; 0, C^{\text{group}}\sigma^2_{\text{prior}})\mathcal{N}(\vec{v}^{\text{center}}_{\text{relative}}; 0, C^{\,\text{center}}\sigma^2_{\text{prior}})\mathcal{N}(\vec{v}^{\text{surround}}_{\text{relative}}; 0, C^{\,\text{surround}}\sigma^2_{\text{prior}})$$

(23)

By marginalizing the group relative velocity

$$p(\vec{o}^{\,\text{center}}, \vec{o}^{\,\text{surround}}, \vec{v}^{\text{center}}_{\text{relative}}, \vec{v}^{\text{surround}}_{\text{relative}} | C^{\,\text{center}}, S^{\text{center,surround}}, C^{\text{group}}, C^{\,\text{surround}}) =$$
$$\mathcal{N}(\vec{o}^{\,\text{center}} - \vec{v}^{\text{center}}_{\text{relative}}; \vec{o}^{\,\text{surround}} - \vec{v}^{\text{surround}}_{\text{relative}}, \sigma_1^2)$$
$$\mathcal{N}(0; (\vec{o}^{\,\text{center}} - \vec{v}^{\text{center}}_{\text{relative}})\gamma_1 + (\vec{o}^{\,\text{surround}} - \vec{v}^{\text{surround}}_{\text{relative}})(1-\gamma_1), \sigma_1^2\gamma_1(1-\gamma_1) + S^{\text{center,surround}}(\sigma^2_{\Delta} + C^{\text{group}}\sigma^2_{\text{prior}}))$$
$$\mathcal{N}(\vec{v}^{\text{center}}_{\text{relative}}; 0, C^{\,\text{center}}\sigma^2_{\text{prior}})\mathcal{N}(\vec{v}^{\text{surround}}_{\text{relative}}; 0, C^{\,\text{surround}}\sigma^2_{\text{prior}})$$

(24)

Rearraging the different terms

$$p(\vec{o}^{\,\text{center}}, \vec{o}^{\,\text{surround}}, \vec{v}^{\text{center}}_{\text{relative}}, \vec{v}^{\text{surround}}_{\text{relative}} | C^{\,\text{center}}, S^{\text{center,surround}}, C^{\text{group}}, C^{\,\text{surround}}) =$$
$$\mathcal{N}(\vec{v}^{\text{surround}}_{\text{relative}}; \vec{o}^{\,\text{surround}} - \vec{o}^{\,\text{center}} + \vec{v}^{\text{center}}_{\text{relative}}, \sigma_1^2)$$
$$\mathcal{N}((1-\gamma_1)\vec{v}^{\text{surround}}_{\text{relative}}; (\vec{o}^{\,\text{center}} - \vec{v}^{\text{center}}_{\text{relative}})\gamma_1 + \vec{o}^{\,\text{surround}}(1-\gamma_1), \sigma_1^2\gamma_1(1-\gamma_1) + S^{\text{center,surround}}(\sigma_\Delta^2 + C^{\text{group}}\sigma^2_{\text{prior}}))$$
$$\mathcal{N}(\vec{v}^{\text{center}}_{\text{relative}}; 0, C^{\,\text{center}}\sigma^2_{\text{prior}})\mathcal{N}(\vec{v}^{\text{surround}}_{\text{relative}}; 0, C^{\,\text{surround}}\sigma^2_{\text{prior}}) \tag{25}$$

$$p(\vec{o}^{\,\text{center}}, \vec{o}^{\,\text{surround}}, \vec{v}^{\text{center}}_{\text{relative}}, \vec{v}^{\text{surround}}_{\text{relative}} | C^{\,\text{center}}, S^{\text{center,surround}}, C^{\text{group}}, C^{\,\text{surround}}) =$$
$$\mathcal{N}(\vec{v}^{\text{surround}}_{\text{relative}}; \vec{o}^{\,\text{surround}} - \vec{o}^{\,\text{center}} + \vec{v}^{\text{center}}_{\text{relative}}, \sigma_1^2)\frac{1}{(1-\gamma_1)^2}$$
$$\mathcal{N}(\vec{v}^{\text{surround}}_{\text{relative}}; (\vec{o}^{\,\text{center}} - \vec{v}^{\text{center}}_{\text{relative}})\frac{\gamma_1}{(1-\gamma_1)} + \vec{o}^{\,\text{surround}}, \frac{\sigma_1^2\gamma_1(1-\gamma_1) + S^{\text{center,surround}}(\sigma_\Delta^2 + C^{\text{group}}\sigma^2_{\text{prior}})}{(1-\gamma_1)^2})$$
$$\mathcal{N}(\vec{v}^{\text{center}}_{\text{relative}}; 0, C^{\,\text{center}}\sigma^2_{\text{prior}})\mathcal{N}(\vec{v}^{\text{surround}}_{\text{relative}}; 0, C^{\,\text{surround}}\sigma^2_{\text{prior}}) \tag{26}$$

By defining

$$\sigma_2^2 = \sigma_1^2 + \frac{\sigma_1^2\gamma_1(1-\gamma_1) + S^{\text{center,surround}}(\sigma_\Delta^2 + C^{\text{group}}\sigma^2_{\text{prior}})}{(1-\gamma_1)^2} \tag{27}$$

$$\gamma_2 = 1 - \frac{\sigma_1^2}{\sigma_2^2} \tag{28}$$

we can apply the formula for product of Gaussian pdf and marginalizing the surround relative velocity to get

$$p(\vec{o}^{\,\text{center}}, \vec{o}^{\,\text{surround}}, \vec{v}^{\text{center}}_{\text{relative}} | C^{\,\text{center}}, S^{\text{center,surround}}, C^{\text{group}}, C^{\,\text{surround}}) =$$
$$\frac{1}{(1-\gamma_1)^2}\mathcal{N}((\vec{o}^{\,\text{center}} - \vec{v}^{\text{center}}_{\text{relative}})\frac{\gamma_1}{(1-\gamma_1)} + \vec{o}^{\,\text{surround}}; \vec{o}^{\,\text{surround}} - \vec{o}^{\,\text{center}} + \vec{v}^{\text{center}}_{\text{relative}}, \sigma_2^2)$$
$$\mathcal{N}(0; (\vec{o}^{\,\text{surround}} - \vec{o}^{\,\text{center}} + \vec{v}^{\text{center}}_{\text{relative}})\gamma_2 + ((\vec{o}^{\,\text{center}} - \vec{v}^{\text{center}}_{\text{relative}})\frac{\gamma_1}{(1-\gamma_1)} + \vec{o}^{\,\text{surround}})(1-\gamma_2), \sigma_2^2\gamma_2(1-\gamma_2) + C^{\,\text{surround}}\sigma^2_{\text{pr}}$$
$$\mathcal{N}(\vec{v}^{\text{center}}_{\text{relative}}; 0, C^{\,\text{center}}\sigma^2_{\text{prior}}) \tag{29}$$

By rearranging terms

$$
\begin{aligned}
&p(\vec{o}^{\,\text{center}}, \vec{o}^{\,\text{surround}}, \vec{v}^{\text{center}}_{\text{relative}} | C^{\,\text{center}}, S^{\text{center,surround}}, C^{\text{group}}, C^{\,\text{surround}}) = \\
&(1-\gamma_1)^2 \mathcal{N}(\vec{o}^{\,\text{center}}\gamma_1 - \vec{v}^{\text{center}}_{\text{relative}}\gamma_1 + \vec{o}^{\,\text{surround}}(1-\gamma_1); \vec{o}^{\,\text{surround}}(1-\gamma_1) - \vec{o}^{\,\text{center}}(1-\gamma_1) + \vec{v}^{\text{center}}_{\text{relative}}(1-\gamma_1), \sigma_2^2(1-\gamma_1)^2) \\
&\mathcal{N}(0; (\vec{o}^{\,\text{surround}} - \vec{o}^{\,\text{center}} + \vec{v}^{\text{center}}_{\text{relative}})\gamma_2(1-\gamma_1) + (\vec{o}^{\,\text{center}}\gamma_1 - \vec{v}^{\text{center}}_{\text{relative}}\gamma_1 + \vec{o}^{\,\text{surround}}(1-\gamma_1))(1-\gamma_2), \dots \\
&\sigma_1^2\gamma_2(1-\gamma_1)^2 + C^{\,\text{surround}}\sigma^2_{\text{prior}}(1-\gamma_1)^2) \\
&\mathcal{N}(\vec{v}^{\text{center}}_{\text{relative}}; 0, C^{\,\text{center}}\sigma^2_{\text{prior}})
\end{aligned}
\tag{30}
$$

$$
\begin{aligned}
&p(\vec{o}^{\,\text{center}}, \vec{o}^{\,\text{surround}}, \vec{v}^{\text{center}}_{\text{relative}} | C^{\,\text{center}}, S^{\text{center,surround}}, C^{\text{group}}, C^{\,\text{surround}}) = \\
&(1-\gamma_1)^2 \mathcal{N}(\vec{v}^{\text{center}}_{\text{relative}}; \vec{o}^{\,\text{center}}, \sigma_2^2(1-\gamma_1)^2) \\
&\mathcal{N}((\gamma_1-\gamma_2)\vec{v}^{\text{center}}_{\text{relative}}; (\gamma_1-\gamma_2)\vec{o}^{\,\text{center}} + \vec{o}^{\,\text{surround}}(1-\gamma_1), \sigma_1^2\gamma_2(1-\gamma_1)^2 + C^{\,\text{surround}}\sigma^2_{\text{prior}}(1-\gamma_1)^2) \\
&\mathcal{N}(\vec{v}^{\text{center}}_{\text{relative}}; 0, C^{\,\text{center}}\sigma^2_{\text{prior}})
\end{aligned}
\tag{31}
$$

$$
\begin{aligned}
&p(\vec{o}^{\,\text{center}}, \vec{o}^{\,\text{surround}}, \vec{v}^{\text{center}}_{\text{relative}} | C^{\,\text{center}}, S^{\text{center,surround}}, C^{\text{group}}, C^{\,\text{surround}}) = \\
&(1-\gamma_1)^2 \mathcal{N}((\gamma_1-\gamma_2)\vec{v}^{\text{center}}_{\text{relative}}; (\gamma_1-\gamma_2)\vec{o}^{\,\text{center}} + \vec{o}^{\,\text{surround}}(1-\gamma_1), \sigma_1^2\gamma_2(1-\gamma_1)^2 + C^{\,\text{surround}}\sigma^2_{\text{prior}}(1-\gamma_1)^2) \\
&\mathcal{N}(\vec{v}^{\text{center}}_{\text{relative}}; \vec{o}^{\,\text{center}}, \sigma_2^2(1-\gamma_1)^2)\mathcal{N}(\vec{v}^{\text{center}}_{\text{relative}}; 0, C^{\,\text{center}}\sigma^2_{\text{prior}})
\end{aligned}
\tag{32}
$$

$$
\begin{aligned}
&p(\vec{o}^{\,\text{center}}, \vec{o}^{\,\text{surround}}, \vec{v}^{\text{center}}_{\text{relative}} | C^{\,\text{center}}, S^{\text{center,surround}}, C^{\text{group}}, C^{\,\text{surround}}) = \frac{(1-\gamma_1)^2}{(\gamma_1-\gamma_2)^2} \\
&\mathcal{N}(\vec{v}^{\text{center}}_{\text{relative}}; \vec{o}^{\,\text{center}} + \vec{o}^{\,\text{surround}}\frac{(1-\gamma_1)}{(\gamma_1-\gamma_2)}, \sigma_1^2\gamma_2\frac{(1-\gamma_1)^2}{(\gamma_1-\gamma_2)^2} + C^{\,\text{surround}}\sigma^2_{\text{prior}}\frac{(1-\gamma_1)^2}{(\gamma_1-\gamma_2)^2}) \\
&\mathcal{N}(\vec{v}^{\text{center}}_{\text{relative}}; \vec{o}^{\,\text{center}}, \sigma_2^2(1-\gamma_1)^2)\mathcal{N}(\vec{v}^{\text{center}}_{\text{relative}}; 0, C^{\,\text{center}}\sigma^2_{\text{prior}})
\end{aligned}
\tag{33}
$$

By defining

$$
\sigma_3^2 = (\sigma_1^2\gamma_2 + C^{\,\text{surround}}\sigma^2_{\text{prior}} + \sigma_2^2(\gamma_1-\gamma_2)^2)(1-\gamma_1)^2 \tag{34}
$$

$$\gamma_3 = \frac{\sigma_2^2(1-\gamma_1)^2(\gamma_1-\gamma_2)^2}{\sigma_3^2} \tag{35}$$

we can apply the formula for product of Gaussian pdf

$$\begin{aligned} &p(\vec{o}^{\,\text{center}}, \vec{o}^{\,\text{surround}}, \vec{v}^{\,\text{center}}_{\text{relative}} | C^{\,\text{center}}, S^{\text{center,surround}}, C^{\text{group}}, C^{\,\text{surround}}) = \\ &\frac{(1-\gamma_1)^2}{(\gamma_1-\gamma_2)^2}\mathcal{N}(\vec{o}^{\,\text{center}}; \vec{o}^{\,\text{center}} + \vec{o}^{\,\text{surround}}\frac{(1-\gamma_1)}{(\gamma_1-\gamma_2)}, \frac{\sigma_3^2}{(\gamma_1-\gamma_2)^2}) \\ &\mathcal{N}(\vec{v}^{\,\text{center}}_{\text{relative}}; \vec{o}^{\,\text{center}} + \vec{o}^{\,\text{surround}}\frac{(1-\gamma_1)}{(\gamma_1-\gamma_2)}S^{\text{center,surround}}\gamma_3, \sigma_2^2(1-\gamma_1)^2(1-S^{\text{center,surround}}\gamma_3)) \\ &\mathcal{N}(\vec{v}^{\,\text{center}}_{\text{relative}}; 0, C^{\,\text{center}}\sigma^2_{\text{prior}}) \end{aligned} \tag{36}$$

By rearranging terms

$$\begin{aligned} &p(\vec{o}^{\,\text{center}}, \vec{o}^{\,\text{surround}}, \vec{v}^{\,\text{center}}_{\text{relative}} | C^{\,\text{center}}, S^{\text{center,surround}}, C^{\text{group}}, C^{\,\text{surround}}) = \\ &(1-\gamma_1)^2\mathcal{N}((\gamma_1-\gamma_2)\vec{o}^{\,\text{center}} + \vec{o}^{\,\text{surround}}(1-\gamma_1); (\gamma_1-\gamma_2)\vec{o}^{\,\text{center}}, \sigma_3^2) \\ &\mathcal{N}(\vec{v}^{\,\text{center}}_{\text{relative}}; \vec{o}^{\,\text{center}} + \vec{o}^{\,\text{surround}}\frac{(1-\gamma_1)}{(\gamma_1-\gamma_2)}S^{\text{center,surround}}\gamma_3, \sigma_2^2(1-\gamma_1)^2(1-S^{\text{center,surround}}\gamma_3)) \\ &\mathcal{N}(\vec{v}^{\,\text{center}}_{\text{relative}}; 0, C^{\,\text{center}}\sigma^2_{\text{prior}}) \end{aligned} \tag{37}$$

We can expand and simplify some of the terms in the pdf to define a gain g

$$\sigma_2^2 = \frac{\sigma_1^2}{(1-\gamma_1)} + \frac{S^{\text{center,surround}}(\sigma_\Delta^2 + C^{\text{group}}\sigma^2_{\text{prior}})}{(1-\gamma_1)^2} \tag{38}$$

$$\sigma_2^2 = \frac{\sigma_1^2(1-\gamma_1) + S^{\text{center,surround}}(\sigma_\Delta^2 + C^{\text{group}}\sigma^2_{\text{prior}})}{(1-\gamma_1)^2} \tag{39}$$

$$\sigma_2^2 - \sigma_1^2 = \frac{\sigma_1^2(1-\gamma_1)\gamma_1 + S^{\text{center,surround}}(\sigma_\Delta^2 + C^{\text{group}}\sigma^2_{\text{prior}})}{(1-\gamma_1)^2} \tag{40}$$

$$\gamma_2 = \frac{\sigma_1^2\gamma_1(1-\gamma_1) + S^{\text{center,surround}}(\sigma_\Delta^2 + C^{\text{group}}\sigma^2_{\text{prior}})}{\sigma_1^2(1-\gamma_1) + S^{\text{center,surround}}(\sigma_\Delta^2 + C^{\text{group}}\sigma^2_{\text{prior}})} \tag{41}$$

$$\gamma_1 - \gamma_2 = \gamma_1 - \frac{\sigma_1^2\gamma_1(1-\gamma_1) + S^{\text{center,surround}}(\sigma_\Delta^2 + C^{\text{group}}\sigma^2_{\text{prior}})}{\sigma_1^2(1-\gamma_1) + S^{\text{center,surround}}(\sigma_\Delta^2 + C^{\text{group}}\sigma^2_{\text{prior}})} \tag{42}$$

$$\gamma_1 - \gamma_2 = \frac{S^{\text{center,surround}}\gamma_1(\sigma_\Delta^2 + C^{\text{group}}\sigma_{\text{prior}}^2) - S^{\text{center,surround}}(\sigma_\Delta^2 + C^{\text{group}}\sigma_{\text{prior}}^2)}{\sigma_1^2(1-\gamma_1) + S^{\text{center,surround}}(\sigma_\Delta^2 + C^{\text{group}}\sigma_{\text{prior}}^2)} \tag{43}$$

$$\gamma_1 - \gamma_2 = -\frac{S^{\text{center,surround}}(\sigma_\Delta^2 + C^{\text{group}}\sigma_{\text{prior}}^2)(1-\gamma_1)}{\sigma_1^2(1-\gamma_1) + S^{\text{center,surround}}(\sigma_\Delta^2 + C^{\text{group}}\sigma_{\text{prior}}^2)} \tag{44}$$

$$(\gamma_1 - \gamma_2)^2 = \frac{S^{\text{center,surround}}(\sigma_\Delta^2 + C^{\text{group}}\sigma_{\text{prior}}^2)^2(1-\gamma_1)^2}{[\sigma_1^2(1-\gamma_1) + S^{\text{center,surround}}(\sigma_\Delta^2 + C^{\text{group}}\sigma_{\text{prior}}^2)]^2} \tag{45}$$

$$\sigma_2^2(\gamma_1 - \gamma_2)^2 = \frac{S^{\text{center,surround}}(\sigma_\Delta^2 + C^{\text{group}}\sigma_{\text{prior}}^2)^2}{[\sigma_1^2(1-\gamma_1) + S^{\text{center,surround}}(\sigma_\Delta^2 + C^{\text{group}}\sigma_{\text{prior}}^2)]} \tag{46}$$

$$\frac{1}{(\sigma_1^2\gamma_2 + C^{\text{surround}}\sigma_{\text{prior}}^2 + \sigma_2^2(\gamma_1 - \gamma_2)^2)} = \frac{(1-\gamma_1)^2}{\sigma_3^2} \tag{47}$$

$$\gamma_3 = \frac{\sigma_2^2(1-\gamma_1)^2(\gamma_1 - \gamma_2)^2}{\sigma_3^2} \tag{48}$$

$$\gamma_3 = \frac{S^{\text{center,surround}}(\sigma_\Delta^2 + C^{\text{group}}\sigma_{\text{prior}}^2)^2}{[(\sigma_1^2\gamma_2 + C^{\text{surround}}\sigma_{\text{prior}}^2 + \sigma_2^2(\gamma_1 - \gamma_2)^2)]} \tag{49}$$

$$\gamma_3 = -\frac{S^{\text{center,surround}}(\sigma_\Delta^2 + C^{\text{group}}\sigma_{\text{prior}}^2)g}{[\sigma_\Delta^2 + \sigma_g^2 + S^{\text{center,surround}}(\sigma_\Delta^2 + C^{\text{group}}\sigma_{\text{prior}}^2)]} \tag{50}$$

$$\frac{(1-\gamma_1)}{(\gamma_1 - \gamma_2)}\gamma_3 = -\frac{S^{\text{center,surround}}(\sigma_\Delta^2 + C^{\text{group}}\sigma_{\text{prior}}^2)}{(\sigma_1^2\gamma_2 + C^{\text{surround}}\sigma_{\text{prior}}^2 + \sigma_2^2(\gamma_1 - \gamma_2)^2)} \tag{51}$$

$$g = \frac{(1-\gamma_1)}{(\gamma_1 - \gamma_2)}\gamma_3 \tag{52}$$

$$g = -\frac{S^{\text{center,surround}}(\sigma_\Delta^2 + C^{\text{group}}\sigma_{\text{prior}}^2)}{\sigma_\Delta^2 + \sigma_r^2 + C^{\text{surround}}\sigma_{\text{prior}}^2 + S^{\text{center,surround}}(\sigma_\Delta^2 + C^{\text{group}}\sigma_{\text{prior}}^2)} \tag{53}$$

Substituting in the original expression

$$p(\vec{o}^{\,\text{center}}, \vec{o}^{\,\text{surround}}, \vec{v}_{\text{relative}}^{\text{center}} | C^{\,\text{center}}, S^{\text{center,surround}}, C^{\text{group}}, C^{\,\text{surround}}) =$$
$$(1-\gamma_1)^2\mathcal{N}(\vec{o}^{\,\text{surround}}(1-\gamma_1); 0, \sigma_3^2)\mathcal{N}(\vec{v}_{\text{relative}}^{\text{center}}; \vec{o}^{\,\text{center}} + \vec{o}^{\,\text{surround}}g, \sigma_2^2(1-\gamma_1)^2(1-S^{\text{center,surround}}\gamma_3))$$
$$\mathcal{N}(\vec{v}_{\text{relative}}^{\text{center}}; 0, C^{\,\text{center}}\sigma_{\text{prior}}^2) \tag{54}$$

By defining

$$\sigma_4^2 = \sigma_2^2(1-\gamma_1)^2(1 - S^{\text{center,surround}}\gamma_3) + C^{\text{ center}}\sigma_{\text{prior}}^2 \tag{55}$$

$$\gamma_4 = \frac{C^{\text{ center}}\sigma_{\text{prior}}^2}{\sigma_4^2} \tag{56}$$

We can apply the rule for multiplying Gaussians and substitute the expression in the definition of the posterior of relative center velocity to get

$$\begin{aligned} p(\vec{v}_{\text{relative}}^{\text{center}} | \vec{o}^{\text{ center}}, \vec{o}^{\text{ surround}}, C^{\text{ center}}, S^{\text{center,surround}}, C^{\text{group}}, C^{\text{ surround}}) = \\ \mathcal{N}(\vec{v}_{\text{relative}}^{\text{center}}; (\vec{o}^{\text{ center}} + \vec{o}^{\text{ surround}} g)\gamma_4, \sigma_2^2(1-\gamma_1)^2(1 - S^{\text{center,surround}}\gamma_3)\gamma_4) \end{aligned} \tag{57}$$

which can be simplified as follows

$$\begin{aligned} p(\vec{v}_{\text{relative}}^{\text{center}} | \vec{o}^{\text{ center}}, \vec{o}^{\text{ surround}}, C^{\text{ center}}, S^{\text{center,surround}}, C^{\text{group}}, C^{\text{ surround}}) = \\ \mathcal{N}(\vec{v}_{\text{relative}}^{\text{center}}; (\vec{o}^{\text{ center}} + \vec{o}^{\text{ surround}} g)\gamma_4, [\sigma_4^2 - C^{\text{ center}}\sigma_{\text{prior}}^2]\gamma_4) \end{aligned} \tag{58}$$

$$\begin{aligned} p(\vec{v}_{\text{relative}}^{\text{center}} | \vec{o}^{\text{ center}}, \vec{o}^{\text{ surround}}, C^{\text{ center}}, S^{\text{center,surround}}, C^{\text{group}}, C^{\text{ surround}}) = \\ \mathcal{N}(\vec{v}_{\text{relative}}^{\text{center}}; (\vec{o}^{\text{ center}} + \vec{o}^{\text{ surround}} g)\gamma_4, C^{\text{ center}}\sigma_{\text{prior}}^2(1-\gamma_4)) \end{aligned} \tag{59}$$

S3 Distribution over responses under perceptual estimates

The distribution over responses can be evaluated for the different perceptual estimates by substituting the different estimates detailed in the main text (Section: Perceptual estimation) into Eq. 15.

Posterior sampling

$$p(R|\underline{\nu}) = \sum_j \sum_{i=0}^{N} w_j^{\text{quad}} w_i'(\vec{\underline{o}}) \int_\theta \mathcal{N}_{\text{circular}}(R; \theta + b, \kappa_m) \mathcal{N}_{\text{circular}}[\theta; \mu_i'(\vec{\underline{o}}), \kappa_i'] \tag{60}$$

which can be simplified using the product rule of von-Mises pdf (*56*) to get

$$p(R|\underline{\nu}) = \sum_j \sum_{i=0}^{N} w_j^{\text{quad}} w_i'(\vec{\underline{o}}) \frac{I_0[\sqrt{(\kappa_i')^2 + \kappa_m + 2\kappa_i'\kappa_m \cos(R - b - \mu_i'(\vec{\underline{o}}))}]}{2\pi I_0(\kappa_m) I_0(\kappa_i')} \tag{61}$$

Structure sampling

$$p(R|\underline{\nu}) = \sum_j \sum_{i=0}^{N} w_j^{\text{quad}} w_i'(\vec{\underline{o}}) \mathcal{N}_{\text{circular}}(R; \mu_i'(\vec{\underline{o}}) + b, \kappa_m). \tag{62}$$

Model averaging

$$p(R|\underline{\nu}) = \sum_j w_j^{\text{quad}}(\vec{\underline{o}}) \mathcal{N}_{\text{circular}} \left(R; \sum_{i=0}^{N} w_i' \mu_i'(\vec{\underline{o}}) + b, \kappa_m \right) \tag{63}$$

Model selection

$$p(R|\underline{\nu}) = \sum_j w_j^{\text{quad}}(\vec{\underline{o}}) \mathcal{N}_{\text{circular}} \left(R; \mu_{i^*}'(\vec{\underline{o}}) + b, \kappa_m \right) \tag{64}$$

Supplementary Figures

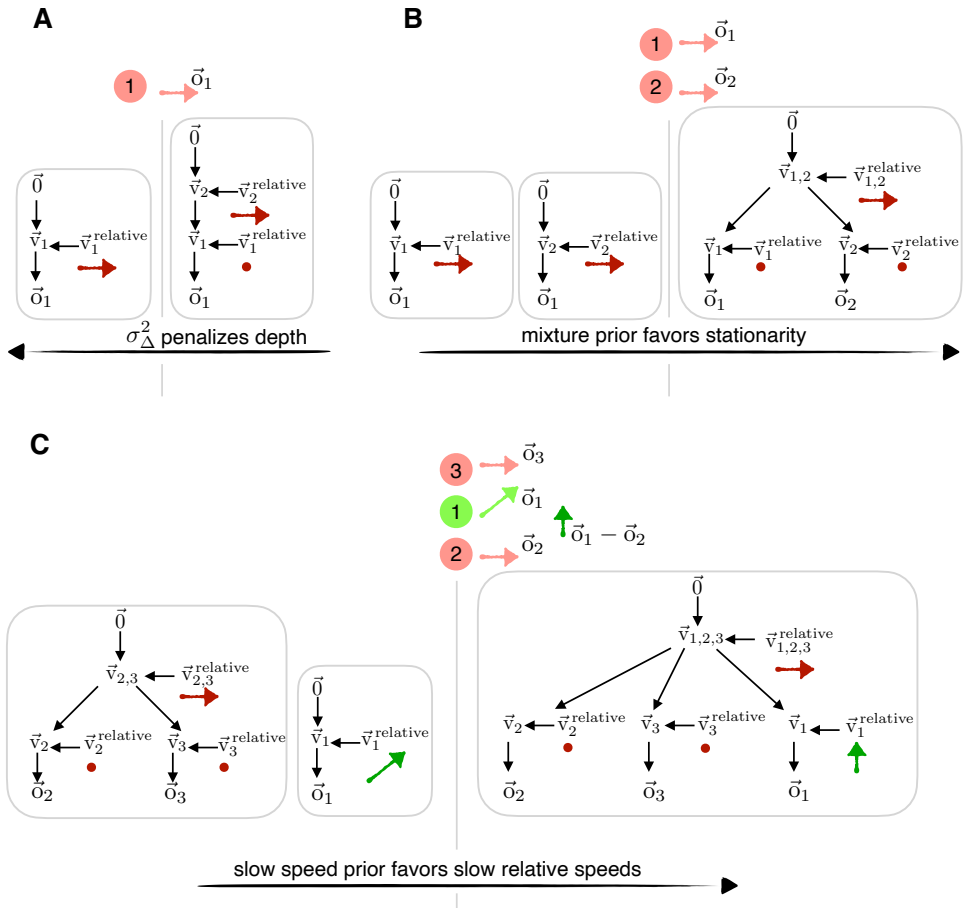


Figure S1: **Generative model in Figure 2 with the inferred velocities indicated.** Dot velocities and the corresponding inferred structures for stimuli with one moving dot (**A**), two moving dots (**B**) or three moving dots (**C**)

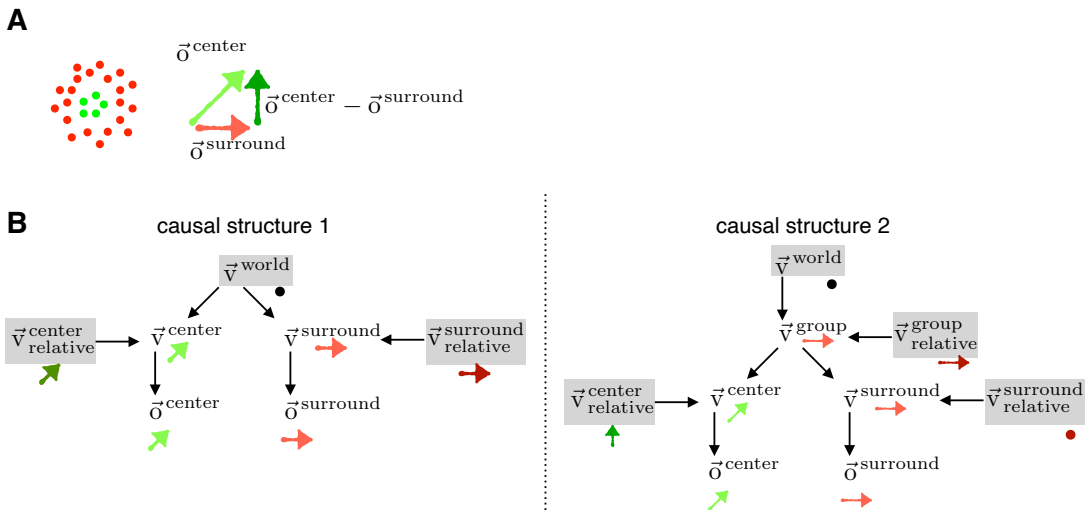


Figure S2: **Scenes with two moving elements. A** Stimulus with two coherently moving groups of dots whose velocities are chosen such that the target (green dots) moves perpendicularly to the surround (red dots). The stimulus is designed to separate three possible percepts of the target velocity as predicted by our model: (a) perceiving the retinal velocity (light green vector) if the target and group are not inferred to be part of the same motion structure (b) perceiving the group velocity (red vector) if the target is inferred to be part of the same motion structure as the surround and the target moves with the surround (c) perceiving the relative velocity (dark green vector) if the target is inferred to be part of the same motion structure as the surround but the target moves differently from the surround **B** The different motion structures that the model can infer with two moving elements. The colored arrows indicate the relative velocities inferred in each motif for the stimulus in **A**. Black dots indicate stationary relative velocities.

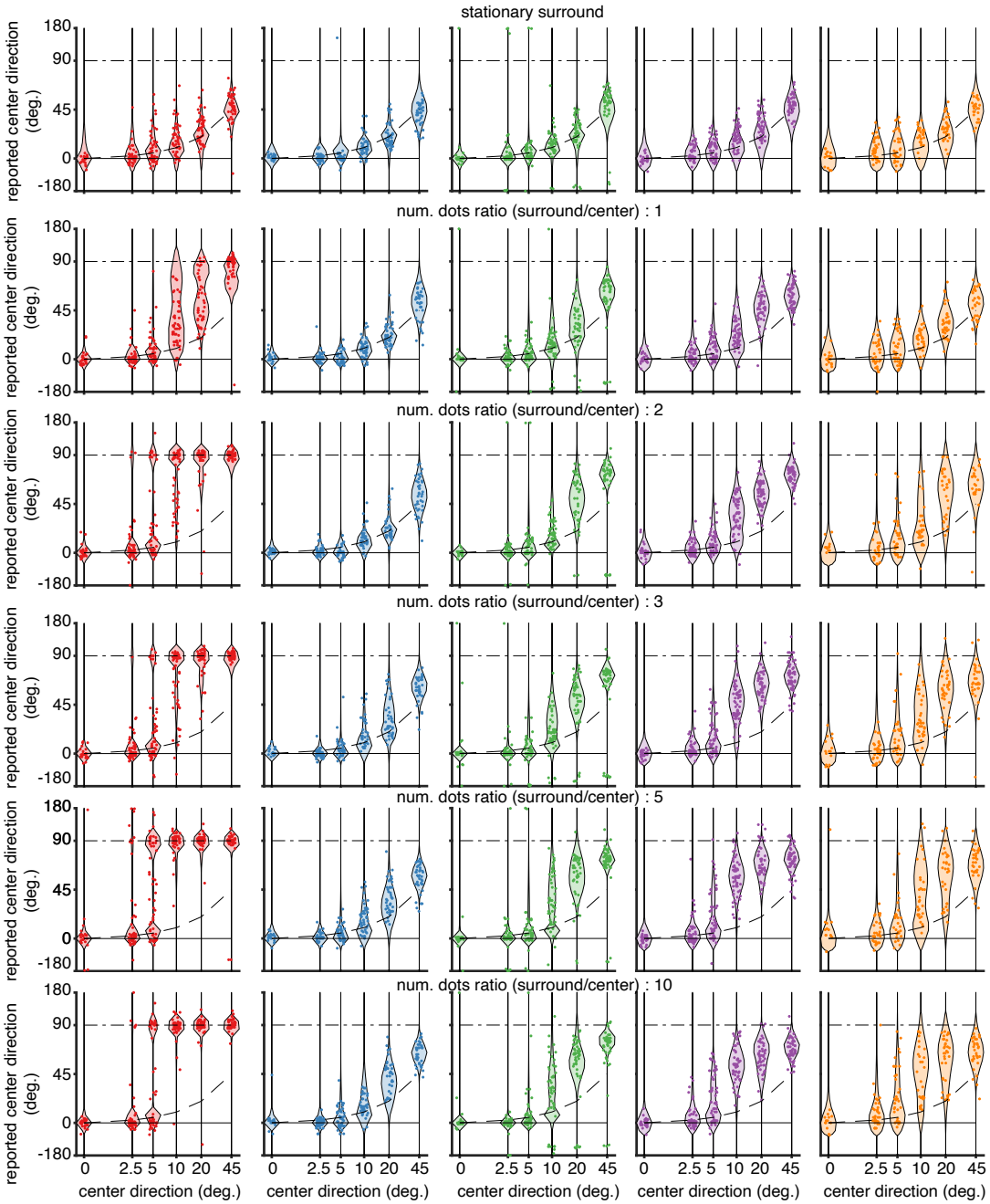


Figure S3: **Individual observer responses in experiment 1.** Response across all observers and different ratios of surround to center dots along with best fit model predictions. Each color indicates a different observer and each row corresponds to a different number of surround patches (indicated by the title).

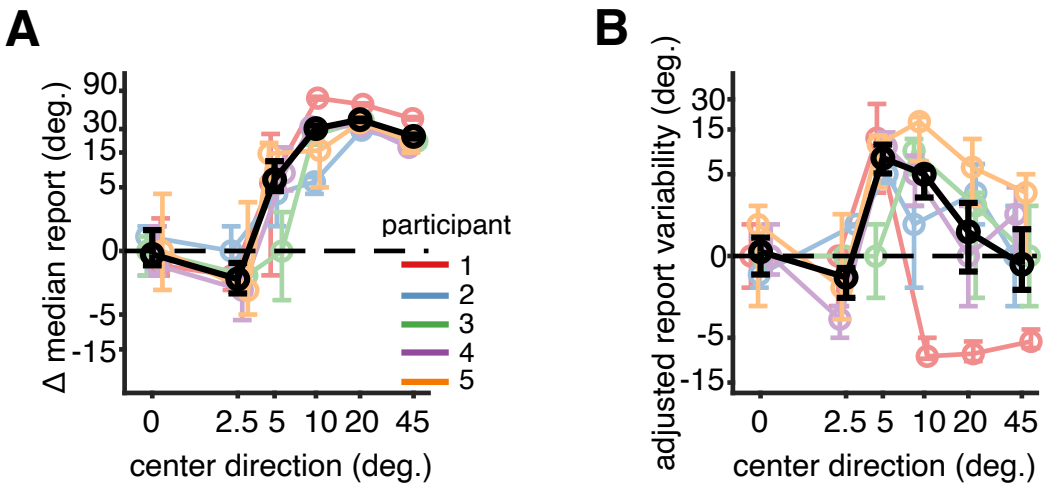


Figure S4: **Summary statistics of observer responses. (A)** Difference in median reports between moving and stationary surround conditions with 68% confidence intervals. Negative values indicate integration and positive values indicate segmentation.**(B)** Difference in report variability between moving and stationary surround with 68% confidence intervals, quantified by the difference in median absolute deviation (MAD). The MAD (*57*) is a robust estimate of variability defined as the median absolute difference between individual trial reports and median reports across trials. Each color corresponds to a different observer and the black line shows the average across observers. The average difference in median reports is consistently negative for $2.5°$ (not significant with $p = 0.07$ across observers, with $p < 0.05$ individually for 3 out of 5 observers) indicating integration, and is significantly positive for larger separations (greater than 10 degrees; $p < 0.001$ across observers, also $p < 0.001$ individually for all observers) indicating segmentation. The small effect sizes for integration are a consequence of the experiment design as the maximum possible difference in median reports for a retinal center direction $2.5°$ is $-2.5°$ which is small compared to the reporting noise. This shortcoming is addressed in Experiment 2 described in the next section. The average MAD at $5°$ (Fig. 3D) is significantly greater than the MAD at $2.5°$ and $45°$ ($p < 0.001$ across observers, with $p < 0.05$ for 3 out of 5 observers individually) indicating greater variability in reports for intermediate separations reflecting the higher uncertainty in causal structures expected from the causal inference model.Significance was estimated by bootstrapping using 10^4 samples.

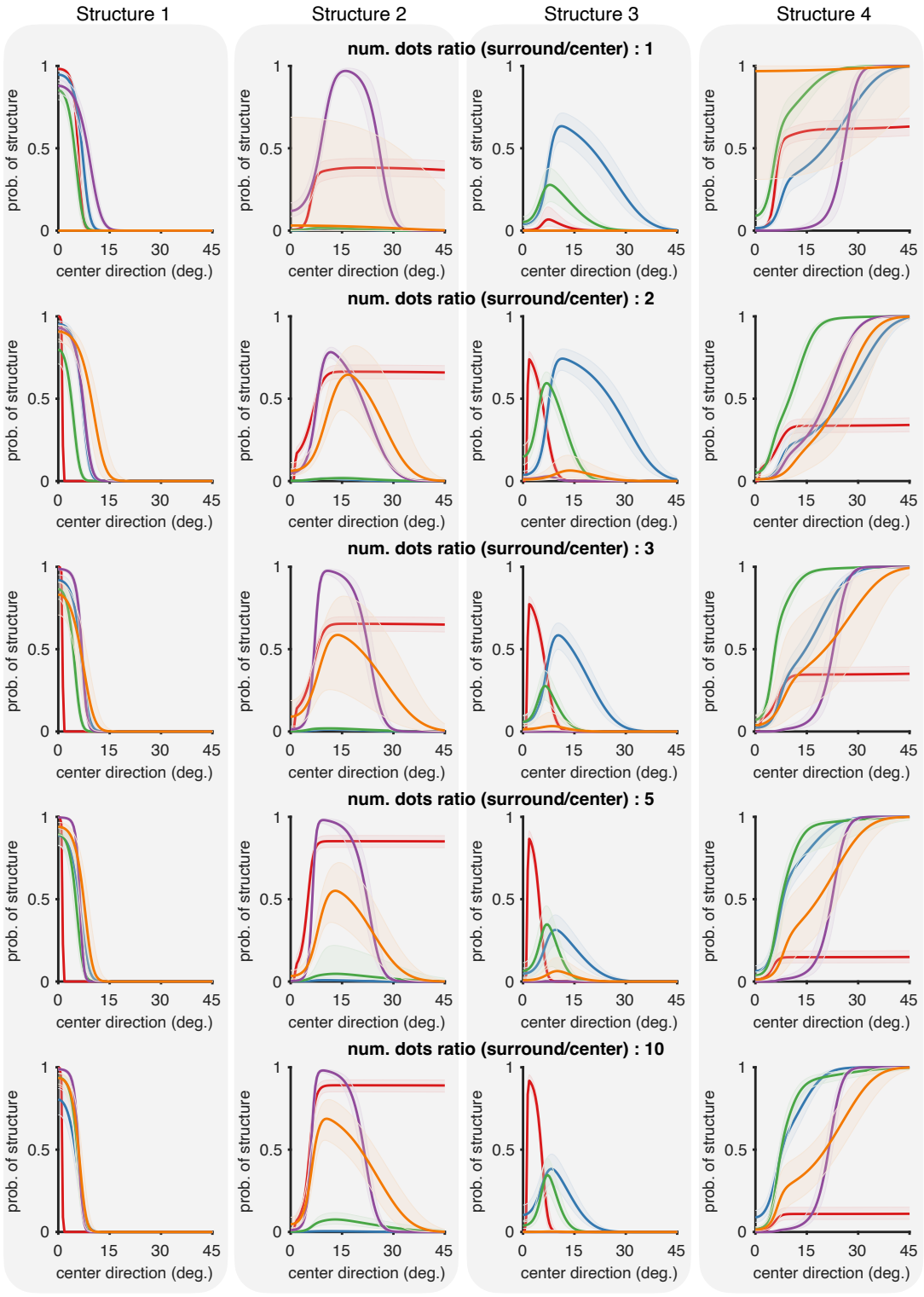

Structure 1
Structure 2
Structure 3
Structure 4
num. dots ratio (surround/center) : 1
num. dots ratio (surround/center) : 2
num. dots ratio (surround/center) : 3
num. dots ratio (surround/center) : 5
num. dots ratio (surround/center) : 10
prob. of structure
center direction (deg.)
0
0.5
1
0
15
30
45

Figure S5: **Posterior probability of structures for varying numbers of dots.** Posterior probability of the four structures in Figure 4A-D similar to those shown in Figure 4E-H for 10 surround patches, but for all numbers of surround patches (1,2,3,5 and 10).

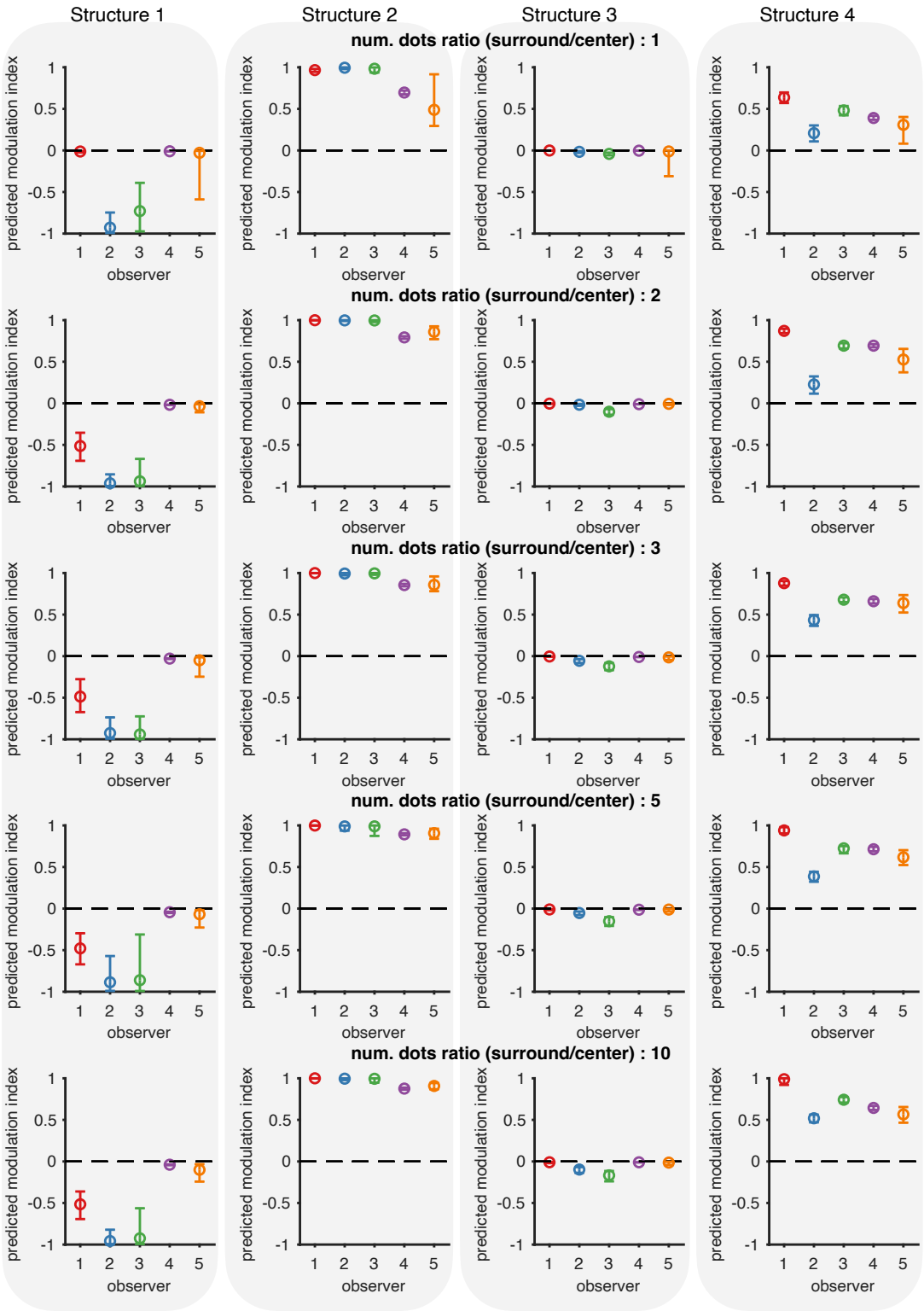
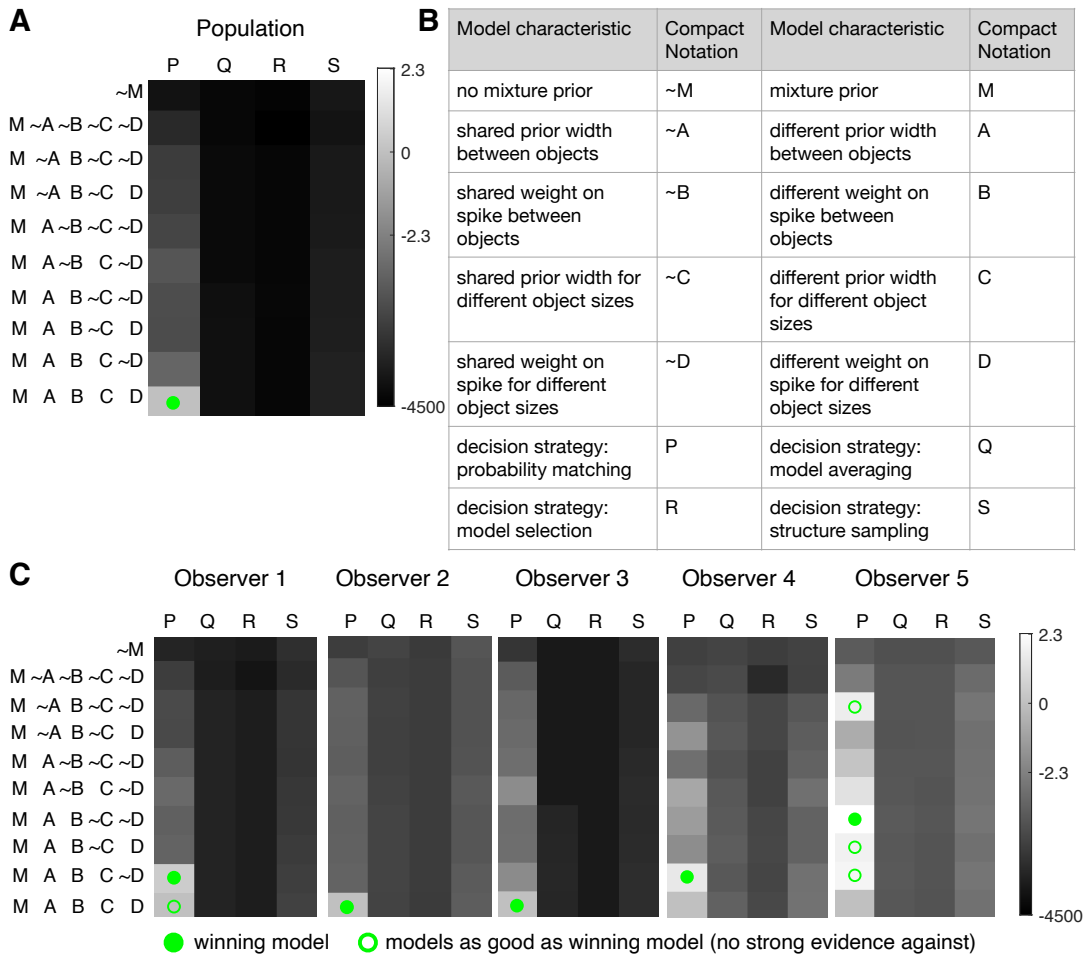

Structure 1
Structure 2
Structure 3
Structure 4
num. dots ratio (surround/center) : 1
num. dots ratio (surround/center) : 2
num. dots ratio (surround/center) : 3
num. dots ratio (surround/center) : 5
num. dots ratio (surround/center) : 10
predicted modulation index
observer

Figure S6: **Modulation indices corresponding to each structure for varying numbers of dots.** Modulation indices predicted under the four structures in Figure 4A-D similar to those shown in Figure 4I-L for 10 surround patches, but for all numbers of surround patches (1,2,3,5 and 10).



Model characteristic	Compact Notation	Model characteristic	Compact Notation
no mixture prior	~M	mixture prior	M
shared prior width between objects	~A	different prior width between objects	A
shared weight on spike between objects	~B	different weight on spike between objects	B
shared prior width for different object sizes	~C	different prior width for different object sizes	C
shared weight on spike for different object sizes	~D	different weight on spike for different object sizes	D
decision strategy: probability matching	P	decision strategy: model averaging	Q
decision strategy: model selection	R	decision strategy: structure sampling	S

Figure S7: **Factorial model comparison.** Factorial model comparison using relative log likelihood (shown in grayscale) computed using AIC as in Figure 4B but across all simplifications of the model depicted in **(B)** shown for each observer in **(C)** and across all observers in **(A)**. The filled green circle indicates the best model through model comparison and the open green circles show models that are not the best but the evidence against them is not strong as compared to the best model. The most simplified model had the same prior parameters (weight on the delta and width of the slow speed prior) for: (a) the different velocities variables (i.e. center, surround, group) in our model, and (b) the inferred velocities for stimuli with different numbers of surround patches. We systematically allowed these parameters to vary to get to the most complex model where all parameters were allowed to vary (but with a weak prior preferring a shared value).

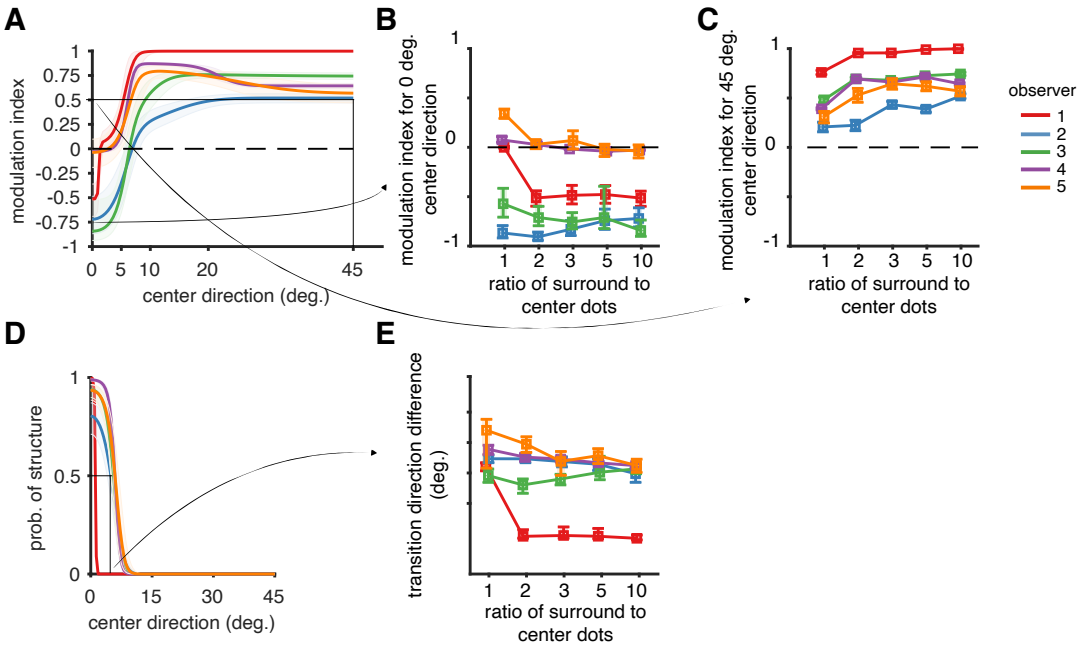
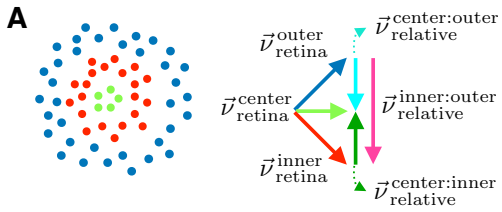
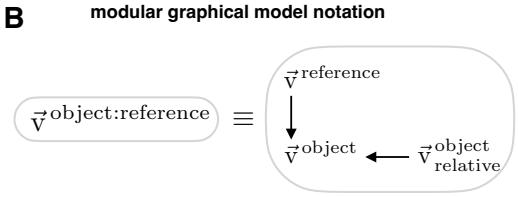
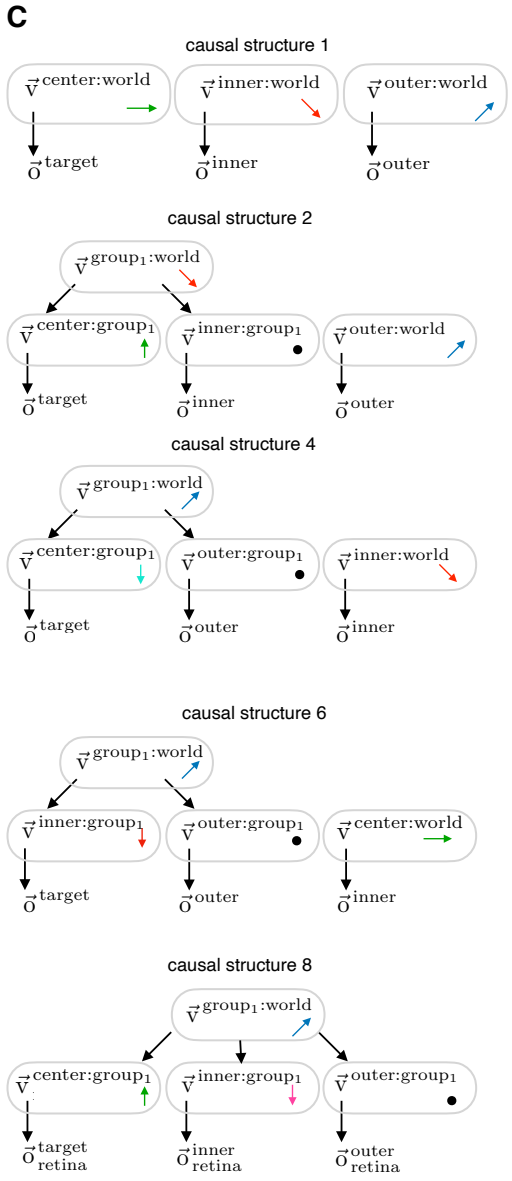

A
B
C
D
E
modulation index
center direction (deg.)
modulation index for 0 deg. center direction
modulation index for 45 deg. center direction
ratio of surround to center dots
observer
1
2
3
4
5
prob. of structure
transition direction difference (deg.)

Figure S8: **Experiment 1: dependency on number of surround dots. (A)** The posterior predictive distribution over modulation indices across observers. **(B)** Predicted modulation index at 0° from (A) plotted as a function of the number of surround patches. 4/5 observers show a decrease in the predicted modulation index at 0 degrees (3 significantly with $p< 0.001$) with increasing number of surround patches indicating a higher strength of integration as the reliability of the surround increases. **(C)** Predicted modulation index at 45° from (A) plotted as a function of the number of surround patches. All observers show a significant increase in modulation index ($p< 0.001$) with increasing number of surround patches indicating a higher strength of segmentation as the reliability of the surround increases. The square markers in (B,C) indicate that these are model-predicted transition directions and not empirically measured data. All the statistical tests were done using samples obtained through posterior sampling of the model parameters. Two statistical tests were done to assess change in modulation indices as a function of the number of surround patches. First, the probability of modulation indices for 10 surround patches was estimated as being greater/lesser than those for 1 surround patch by comparing the posterior samples over the corresponding modulation indices. Second, the posterior slope distribution of the line fit to the modulation indices as a function of the number of dots was used to assess the increasing/decreasing trend by estimating the proportion of samples from the slope distribution that were greater/lesser than zero. **(D)** We quantify the region of integration by defining a transition direction difference which is the difference in direction between center and surround where the probability of integration is 0.5. The probability of integration is plotted in Figure S5 (first column) for different number of surround dots. **(E)** We find that the transition direction difference decreases with increase in number of surround dots in agreement with earlier causal inference studies that the region of integration increases with increase in cue uncertainty. The difference is significant ($p< 0.05$) for 4/5 observers.


A
$\vec{v}^{\text{outer}}_{\text{retina}}$
$\vec{v}^{\text{center:outer}}_{\text{relative}}$
$\vec{v}^{\text{center}}_{\text{retina}}$
$\vec{v}^{\text{inner:outer}}_{\text{relative}}$
$\vec{v}^{\text{inner}}_{\text{retina}}$
$\vec{v}^{\text{center:inner}}_{\text{relative}}$


B
modular graphical model notation
$\vec{V}^{\text{object:reference}}$
≡
$\vec{V}^{\text{reference}}$
$\vec{V}^{\text{object}}$
$\vec{V}^{\text{object}}_{\text{relative}}$


C
causal structure 1
$\vec{V}^{\text{center:world}}$
$\vec{V}^{\text{inner:world}}$
$\vec{V}^{\text{outer:world}}$
$\vec{O}^{\text{target}}$
$\vec{O}^{\text{inner}}$
$\vec{O}^{\text{outer}}$
causal structure 2
$\vec{V}^{\text{group}_1\text{:world}}$
$\vec{V}^{\text{center:group}_1}$
$\vec{V}^{\text{inner:group}_1}$
$\vec{V}^{\text{outer:world}}$
$\vec{O}^{\text{target}}$
$\vec{O}^{\text{inner}}$
$\vec{O}^{\text{outer}}$
causal structure 4
$\vec{V}^{\text{group}_1\text{:world}}$
$\vec{V}^{\text{center:group}_1}$
$\vec{V}^{\text{outer:group}_1}$
$\vec{V}^{\text{inner:world}}$
$\vec{O}^{\text{target}}$
$\vec{O}^{\text{outer}}$
$\vec{O}^{\text{inner}}$
causal structure 6
$\vec{V}^{\text{group}_1\text{:world}}$
$\vec{V}^{\text{inner:group}_1}$
$\vec{V}^{\text{outer:group}_1}$
$\vec{V}^{\text{center:world}}$
$\vec{O}^{\text{target}}$
$\vec{O}^{\text{outer}}$
$\vec{O}^{\text{inner}}$
causal structure 8
$\vec{V}^{\text{group}_1\text{:world}}$
$\vec{V}^{\text{center:group}_1}$
$\vec{V}^{\text{inner:group}_1}$
$\vec{V}^{\text{outer:group}_1}$
$\vec{O}^{\text{target}}_{\text{retina}}$
$\vec{O}^{\text{inner}}_{\text{retina}}$
$\vec{O}^{\text{outer}}_{\text{retina}}$

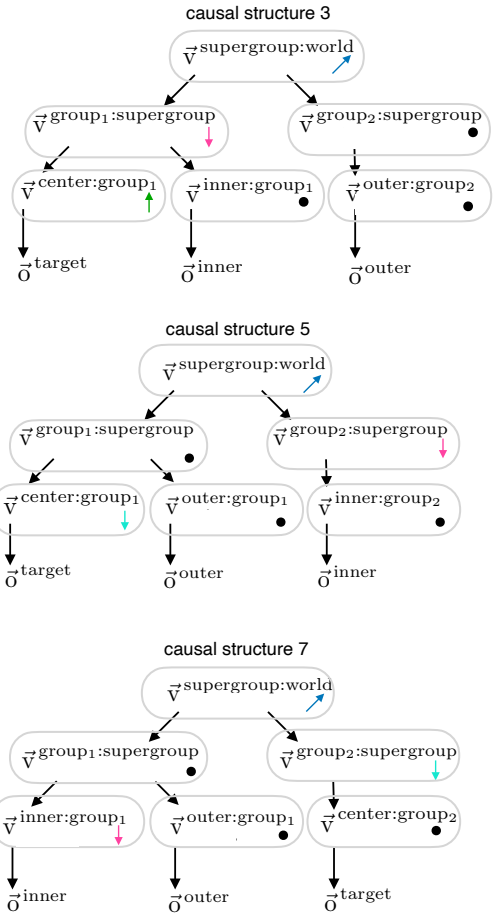

causal structure 3
$\vec{V}^{\text{supergroup:world}}$
$\vec{V}^{\text{group}_1\text{:supergroup}}$
$\vec{V}^{\text{group}_2\text{:supergroup}}$
$\vec{V}^{\text{center:group}_1}$
$\vec{V}^{\text{inner:group}_1}$
$\vec{V}^{\text{outer:group}_2}$
$\vec{O}^{\text{target}}$
$\vec{O}^{\text{inner}}$
$\vec{O}^{\text{outer}}$
causal structure 5
$\vec{V}^{\text{supergroup:world}}$
$\vec{V}^{\text{group}_1\text{:supergroup}}$
$\vec{V}^{\text{group}_2\text{:supergroup}}$
$\vec{V}^{\text{center:group}_1}$
$\vec{V}^{\text{outer:group}_1}$
$\vec{V}^{\text{inner:group}_2}$
$\vec{O}^{\text{target}}$
$\vec{O}^{\text{outer}}$
$\vec{O}^{\text{inner}}$
causal structure 7
$\vec{V}^{\text{supergroup:world}}$
$\vec{V}^{\text{group}_1\text{:supergroup}}$
$\vec{V}^{\text{group}_2\text{:supergroup}}$
$\vec{V}^{\text{inner:group}_1}$
$\vec{V}^{\text{outer:group}_1}$
$\vec{V}^{\text{center:group}_2}$
$\vec{O}^{\text{inner}}$
$\vec{O}^{\text{outer}}$
$\vec{O}^{\text{target}}$

Figure S9: **Scenes with three moving elements.** **(A)** Stimulus with three coherently moving groups of dots whose velocities are such that the center (green dots) move perpendicularly to the surrounding inner and outer rings. The three velocities share a common component such that the relative velocities between any two elements are either 90 degrees or -90 degrees. **(B)** A compact notation for the causal inference motif where an object is inferred in a reference frame by inferring whether its relative velocity to the reference frame is zero or not. **(C)** The different motion structures that the model can infer with three moving elements. Each gray outline box implies two nested tructures corresponding to whether the relative velocity is inferred to be zero or not. The colored arrows indicate the inferred relative velocities for the stimulus in **A**.

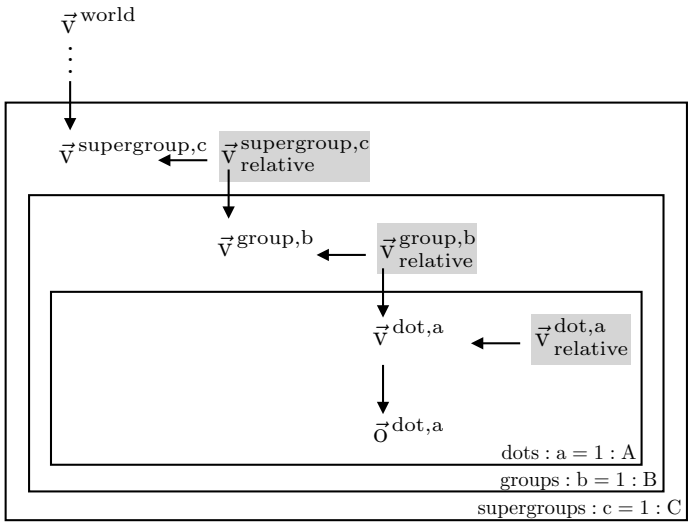


Figure S10: **Alternative (hypothetical) model for hierarchical motion perception.** The motif presented in Figure 1F could be stacked alternatively to Figure 1G such that the perceived velocity (gray shaded box) forms the reference frame for the level below. Such a model would be consistent with previous work in orientation perception (*47, 48*) where the perceived context value influences the perception of target stimuli. Experiment 1 cannot separate between this structure and Figure 1G as both the perceived and retinal group velocities are the same. But results from experiment 2 (with an additional level of hierarchy) flalsify this model and support our causal inference model presented in Figure 1G.

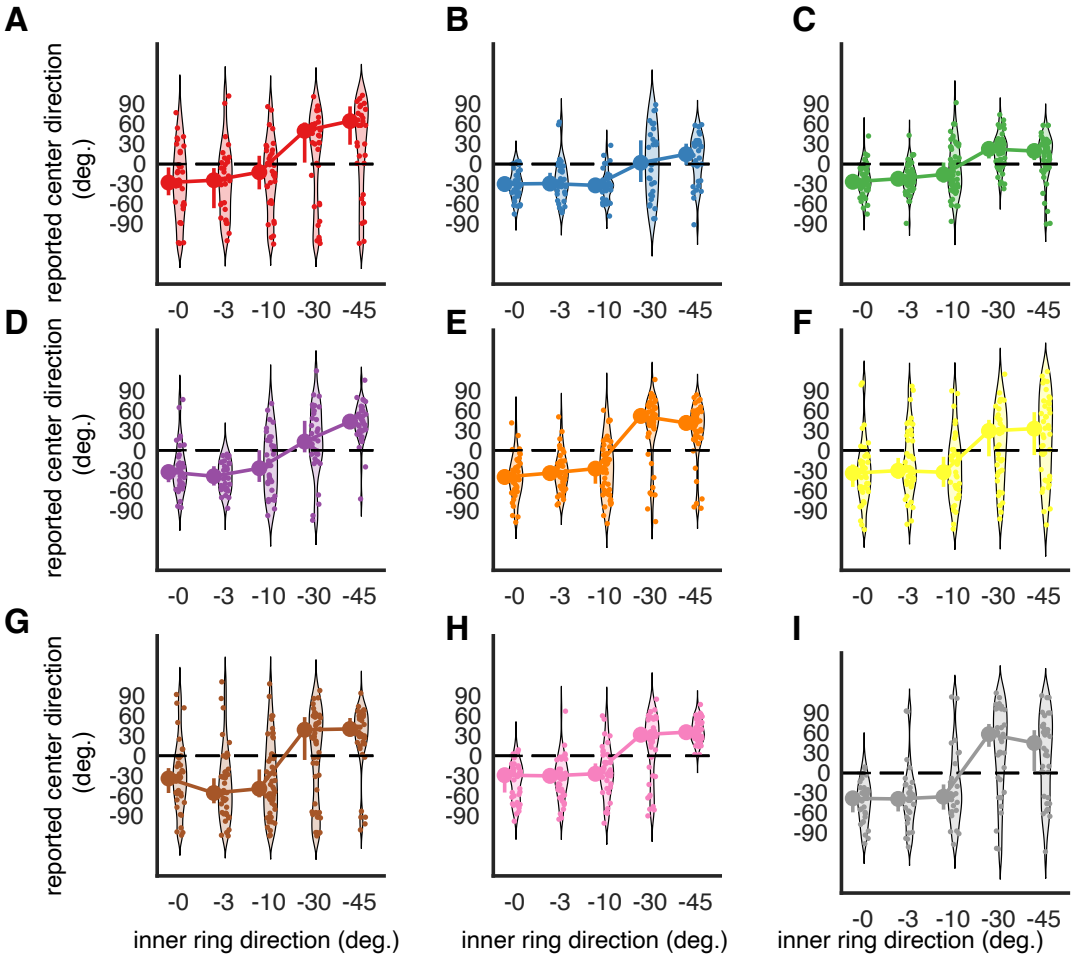


Figure S11: **Individual responses in experiment 2.** Individual responses across trials for observers in experiment 2 for the experimental condition shown in Figure 5A. Each dot corresponds to the response on a single trial and the violin plot is the empirical histogram of responses. The responses are summarized using the circular marginal median along with 95% CI as indicated by the errorbars to the left of each violin plot.

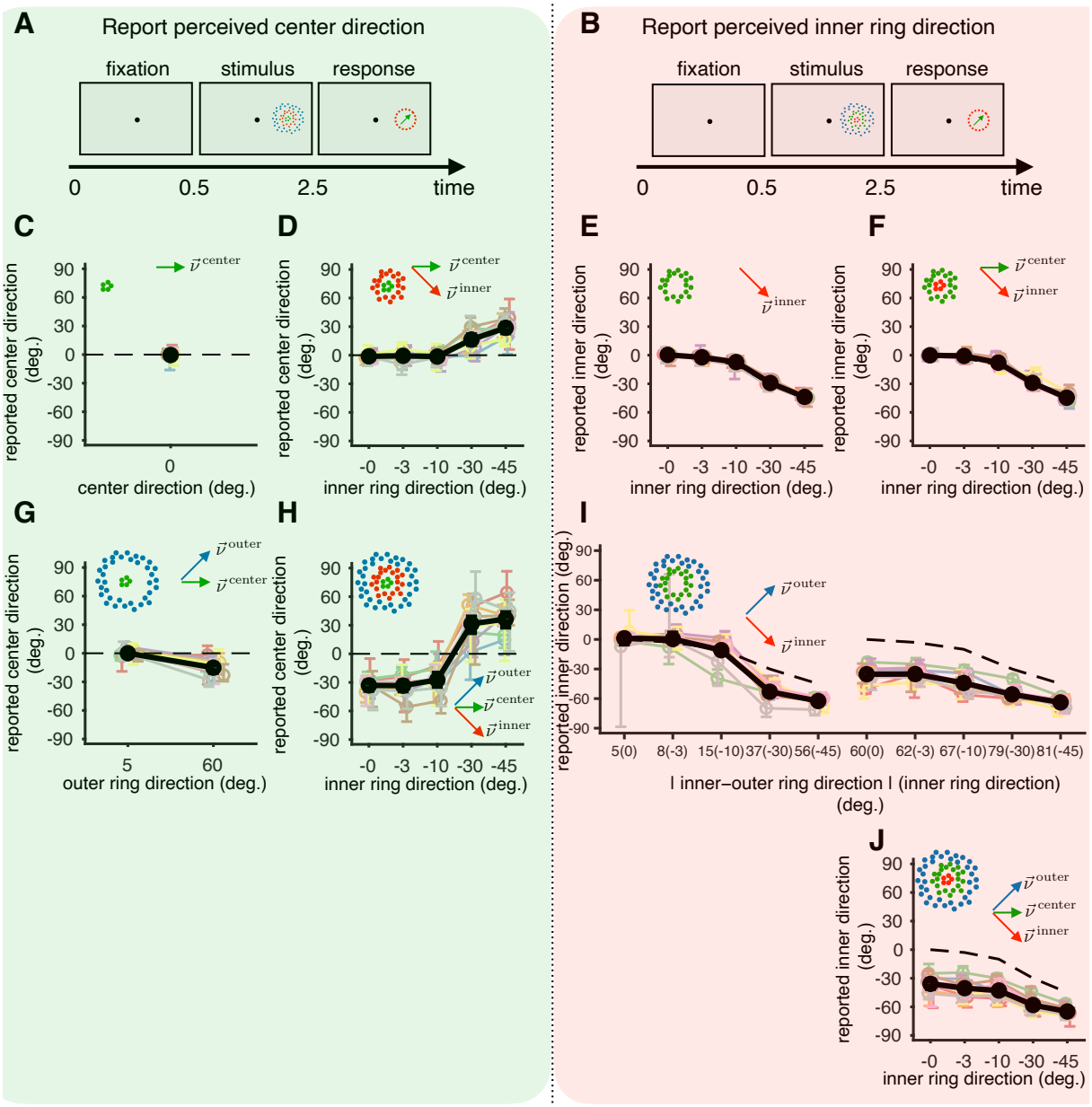

A
Report perceived center direction
fixation
stimulus
response
0
0.5
2.5
time
B
Report perceived inner ring direction
fixation
stimulus
response
0
0.5
2.5
time
C
D
E
F
G
H
I
J
reported center direction (deg.)
reported inner direction (deg.)
center direction (deg.)
inner ring direction (deg.)
outer ring direction (deg.)
| inner−outer ring direction | (inner ring direction) (deg.)
5(0) 8(-3) 15(-10) 37(-30) 56(-45) 60(0) 62(-3) 67(-10) 79(-30) 81(-45)

Figure S12: **Additional empirical results with three moving elements.** **(A)** Stimulus schematic. The observer performs an estimation task in which they have to report the direction of green dots (center) using a dial. **(B)** Stimulus schematic. The observer performs an estimation task in which they have to report the direction of green dots (inner ring) using a dial.**(C,D,G,H)** Observer responses (and 95 % CI) for center percept for different stimulus conditions that define the velocities of the center, inner ring, and outer ring. The group of dots that are displayed in the inset are the coherently moving dots in that condition with velocities indicated by the vectors. The dots not displayed move randomly. The velocities are chosen such that the center always moves at 0 degrees (horizontally) and the two rings move with the same horizontal component but opposite vertical components such that the relative velocities are either 90 degrees (center and inner ring) or -90 degrees (center and outer ring). This arrangement is rotated around the circle across trials to account for any reporting biases. **(E,F,I,J)** Observer responses (and 95 % CI) for inner ring percepts for different stimulus conditions similar to **C,D,G,H**. The observer always reports the direction of the green dots and in these trials the green dots form the inner ring whereas they formed the center in **C,D,G,H**. For (I), the tick-labels indicate the absolute angular difference between the inner and outer ring directions and the brackets indicate the inner ring direction.
In two control conditions (C,E), the observers reported either the center or the inner ring's direction respectively, which was found to be aligned with the presented retinal direction demonstrating no reporting biases. When the center and inner ring moved coherently (D,F), the center direction reports followed the same pattern as predicted in experiment 1 (integration for small inner ring directions and segmentation for larger inner ring directions). The inner ring percepts (F) were largely unaffected by the center suggesting that the inner ring predominantly determines the group velocity. When the center and outer ring moved coherently (G), seven out of nine observers displayed significant negative biases. This outcome supports the notion of perceiving the relative velocity of the center to the outer ring, albeit with a low modulation index suggesting the effect of a surround on the center diminishes with increasing spatial separation between them. Similarly, when the inner and outer rings moved together (I), a weak integration was visible (flattening of the curve beyond the dashed line) for an absolute angular difference is 8 degrees (significant with $p<0.05$ across observers; significant with $p<0.05$ for 5/9 observers). We also see significant biases ($p<0.001$ across observers and also for each observer) consistent with perceiving the inner ring relative to the outer ring for larger separations. Reporting the inner ring directions in the condition in which all the elements move coherently (J), the observer responses are significantly negative across and for individual observers ($p<0.001$) consistent with the model predicts that the observers perceive the inner ring relative to the outer ring. The absolute angular differences between the inner and outer ring lie between 60 to 81 degrees and are similar to the responses when only the inner and outer rings move coherently (for large separations) supporting the model prediction that the percept of the inner ring is the perceived velocity of the center-inner group in the outer ring reference frame and that the group velocity is predominantly determined by the inner ring.

Parameter	Prior	Remarks
σ^2_{prior}	$\text{Lognormal}(-2, 2)$	$\sim 90\%$ mass $\in [0.5°, 75°]$;
σ^2_{Δ}	$\text{Lognormal}(-2, 2)$	see previous
$\sigma^2_{\text{center}}/\sigma^2_{\text{prior}}$	$\text{Beta prime}(1, 1)$	chosen this way such that weight on the observation in the posterior given by $\left(\frac{\sigma^2_{\text{prior}}}{\sigma^2_{\text{center}}+\sigma^2_{\text{prior}}}\right)$ is uniformly distributed
$\sigma^2_{\text{surround}}/\sigma^2_{\text{prior}}$	$\text{Beta prime}(1, 1)$	see previous
α	$\text{Beta}(1.25, 1.25)$	weakly informative prior with a mode at 0.5
$\beta^{\text{center,surround}}$	$\text{Beta}(1.25, 1.25)$	see previous
λ	$\text{Beta}(1, 5)$	prior incorporating knowledge that lapse rates are likely to be small
μ^{lapse}	$\text{Normal}(0, \pi)$	symmetric prior over lapse mean probabilities in radians
κ^{lapse}	$\text{Lognormal}(0, 2.35)$	95% mass $\in [0.01, 100]$
b	$\text{Normal}(0, \pi/2)$	symmetric prior over response biasin radians
κ_{m}	$\text{Lognormal}(0, 2.35)$	95% mass $\in [0.01, 100]$

Table S1: The weakly informative priors used over parameters for MAP and posterior estimation. When different parameters for σ^2_{prior} are estimated for center and surround, and for different number of surround patches, a Normal prior [$\mathcal{N}(0, 4)$] is placed on the difference between the parameters incorporating knowledge that apriori all parameters are likely to be similar for reduced displays. Similarly, a Normal prior [$\mathcal{N}(0, 0.5)$] is placed on the log odds between the estimated α values for center and surround, and for different number of surround patches preferring similar estimates