Flexible coding of object motion in multiple reference frames by parietal cortex neurons

Ryo Sasaki[1]*, Akiyuki Anzai[1], Dora E. Angelaki[2] and Gregory C. DeAngelis[1]

[1] Department of Brain and Cognitive Sciences, Center for Visual Science, University of Rochester, Rochester, New York USA
[2] Center for Neural Science, New York University, New York, New York, USA

*Current institution for R.S.: Department of Neuroscience, Graduate School of Medicine, Kyoto University, Kyoto, Japan

Correspondence should be addressed to R.S. (sasaki.ryo.3r@kyoto-u.ac.jp).

ABSTRACT

Neurons represent spatial information in diverse reference frames, but it remains unclear whether neural reference frames change with task demands and whether these changes can account for behavior. We examined how neurons represent the direction of a moving object during self-motion, while monkeys switched, from trial to trial, between reporting object direction in head- and world-centered reference frames. Self-motion information is needed to compute object motion in world coordinates, but should be ignored when judging object motion in head coordinates. Neural responses in the ventral intraparietal area are modulated by the task reference frame, such that population activity represents object direction in either reference frame. In contrast, responses in the lateral portion of the medial superior temporal area primarily represent object motion in head coordinates. Our findings demonstrate a neural representation of object motion that changes with task requirements.

INTRODUCTION

Sensory signals are encoded in modality-specific reference frames at the sensory periphery, such as an eye-centered reference frame for visual signals and a head-centered reference frame for vestibular signals. In downstream brain areas, signals are often transformed into non-native reference frames, including intermediate or mixed reference frames [1-13], and it is generally presumed that different reference frames are useful for guiding different behaviors.

Natural behavior is flexible, however, and may require the observer to interpret the same sensory signals in different reference frames, depending on task context (e.g., a soccer player might judge motion of the ball relative to their head or relative to the goal). Having neural representations that flexibly adapt to task demands may thereby simplify sensorimotor transformations. While task-switching has been studied extensively with behavior and neuroimaging [14, 15], and single-neuron correlates of switching task sets have been reported [16, 17], little is known about whether the reference frame of neural representations changes dynamically when judgements are made in different reference frames. When the reference frame required for a task changes, do neural representations change accordingly or are neural reference frames fixed?

Perception of object motion during self-motion provides an attractive model system for studying this issue. When an observer is stationary, there is a unique mapping of object motion in the world to image motion on the retina. However, when a translating observer views the same moving object, image motion also depends on self-motion (Fig. 1). To judge object motion relative to their head (e.g., a soccer player who wants to head the ball), one can rely on retinal image motion, which is equivalent to motion in a head-centered reference frame if the eyes do not move. However, to judge object motion relative to the world (e.g., a soccer goalie trying to judge whether a shot will be on goal), one needs to compensate for the visual consequences of self-motion. How neural circuits incorporate information about self-motion to represent object motion in the world is not well understood [18, 19]. Moreover, it is not known whether neural representations are dynamically updated to represent object motion in head or world coordinates based on task instructions.

Numerous psychophysical studies have examined how perception of object motion discounts image motion caused by self-motion, and have identified visual mechanisms that attempt to isolate image motion produced by independent object motion [20-24]. Vestibular signals also aid in dissociating the components of retinal image motion due to object motion and self-motion [24-26]. Thus, we hypothesize that vestibular input may contribute to generating flexible representations of object motion that update with task reference frame.

We trained macaque monkeys to discriminate object direction in either a world-centered or a head-centered reference frame. Monkeys successfully switched between reference frames, randomly across trials, and their performance was enhanced when both visual and vestibular self-motion signals were available. We recorded neural activity from the lateral subdivision of the medial superior temporal (MSTl) area, which has been suggested to play a role in coding object motion [27, 28]. MSTl neurons combine retinal image motion with extra-retinal signals related to eye and head rotation [29], suggesting that MSTl might be a viable candidate for representing direction of object motion in different reference frames. We also recorded neural activity from the ventral intraparietal (VIP) area, which is well known for its roles in representing visual motion, as well as for carrying multisensory representations of visual, vestibular, tactile, and auditory signals in diverse reference frames [2, 4, 11, 30, 31]. Thus, VIP is a good candidate to flexibly represent object motion in a head- or world-centered reference frame.

We find that responses of individual VIP neurons are modulated by the task reference frame, such that their tuning shifts toward world coordinates when the task requires a world-centered reference frame. In contrast, MSTl neurons do not show this effect. At the population level, linear decoding of VIP activity accounts nicely for behavioral effects, whereas decoding of MSTl activity does not. Strikingly, a single set of decoding weights can accurately classify object direction in either head or world coordinates based on VIP activity, but not MSTl activity. The time course of the reference frame transformation in VIP is delayed relative to onset of visual responses. Together, our findings demonstrate that VIP flexibly represents object motion in different reference frames depending on task instructions, with self-motion signals being incorporated into the computation when needed. More generally, our results provide striking evidence that the reference frames of neural representations can be highly dynamic, and that the same neural populations can carry information in different reference frames from moment to moment.

RESULTS

We trained two macaque monkeys to report whether an object moves upward and rightward or upward and leftward during lateral self-motion (Fig. 2a). In the world coordinate task, monkeys judged object motion relative to vertical in a world-centered reference frame; in the head coordinate task, animals reported object motion relative to a head-centered vertical reference. Thus, for some stimulus conditions, the same patterns of random-dot motion on the retina could give rise to opposite perceptual reports in the two different reference frames. In each trial, the color and shape of the fixation point instructed the reference frame to be used (Fig. 2b), and the two reference frame conditions were randomly interleaved. Self-motion information was provided by optic flow of background dots or by a congruent combination of optic flow and physical translation of the animal on a motion platform. A partial cube frame that faded out during object motion helped to instruct the task reference frame (Fig. 2c). If monkeys could switch perfectly between task reference frames, psychometric functions for the two directions of self-motion should overlap in the world coordinate task (Fig. 2d). In contrast, for the head coordinate task, psychometric functions for the two opposite directions of self-motion should be shifted by a predictable amount (ΔPSE = 35.5 deg, see Methods).

Monkeys can switch between world- and head-centered reference frames while judging object motion

To summarize behavior, we computed average psychometric functions across all recording sessions from the two monkeys (Fig. 3a). In the absence of self-motion (Object Only condition), monkeys reported object direction with very little bias (Fig. 3a, black curve). During self-motion, monkeys also performed quite accurately in the head coordinate task, as evidenced by psychometric functions that were shifted by approximately the predicted amount (Fig. 3a, blue and cyan curves). In the world coordinate task, psychometric functions for the two directions of self-motion showed much smaller shifts, indicating that animals largely compensated for the effect of their self-motion on object motion (Fig. 3a, magenta and brown curves). In the Object+Combined condition, this compensation was nearly complete (magenta), whereas compensation was substantially less complete in the Object+Visual condition (brown). This finding demonstrates that vestibular signals enhance the monkeys' ability to judge object direction in a world-centered reference frame. This effect presumably arises because adding vestibular signals provides a more accurate and precise estimate of self-motion velocity, rather than facilitating switching between reference frames per se. However, we cannot differentiate these possibilities.

To quantify these effects, we fit psychometric functions with cumulative Gaussian curves and measured the point of subjective equality (PSE) for each self-motion direction, reference frame, and recording session. Then, we computed the difference in PSE (ΔPSE) between leftward and rightward self-motion directions and compared the results between Object+Visual and Object+Combined conditions (Fig. 3b). For the head coordinate task, mean ΔPSE values for both conditions (Fig. 3b, black triangle) are close to predicted values, although both are 10-15% greater than expected (two-tailed t-test, Object+Visual: t(184)=2.03, p=0.044; Object+Combined: t(184)=6.05, $p=8.0\times10^{-9}$). Moreover, there is no substantive difference between mean ΔPSE values for Object+Visual and Object+Combined conditions in the head coordinate task (two-tailed paired t-test, t(368)=-0.71, p=0.47). This is expected because self-motion information is not needed to perform the head coordinate task.

The pattern of results is strikingly different for the world coordinate task, where ΔPSE values are much smaller. The mean ΔPSE value in the Object+Visual condition is small (9.4 deg) but consistently greater than zero (t(184)=13.9, $p=1.3\times10^{-50}$), whereas the corresponding value (0.75 deg) for the Object+Combined condition is not substantially different from zero (two-tailed t-test, t(184)=1.34, p=0.054). The difference in ΔPSE values between Object+Visual and Object+Combined conditions is quite robust across sessions in the world coordinate task (two-tailed paired t-test, t(368)=11.9, $p=1.7\times10^{-35}$), unlike the head coordinate task. These behavioral results demonstrate that monkeys successfully switch between world and head reference frames from trial to trial, and that vestibular signals facilitate this transformation.

Note that the slope of psychometric functions in Fig. 3a depends on the presence of self-motion. Additional analysis showed that object motion discrimination thresholds are significantly lower in the Object Only condition than in the conditions with self-motion, with only modest differences between Object+Visual and Object+Combined conditions (Extended Data Fig. 1).

Effects of task reference frame on single-unit responses in VIP and MSTl

We next investigated whether the activity of VIP and MSTl neurons was modulated by task reference frame. We recorded from 223 VIP neurons and 177 MSTl neurons that met basic inclusion criteria (see Methods). In general, responses of neurons in both areas are influenced by both object motion and self-motion directions (Fig. 4, filled vs. open symbols). This is not surprising since motion stimuli were transparent such that both object and self-motion vectors impinged upon the receptive fields of the recorded populations (Extended Data Fig. 2). The key question is whether responses depend on the instructed reference frame for the same object and background motions, which can be assessed by comparing responses in the world coordinate task with responses in the head coordinate task. Differences would suggest a representation of object motion that changes with task demands. Indeed, the two example VIP neurons in Fig. 4a show clear response differences between task reference frames in the Object+Combined condition (magenta vs. cyan). In contrast, for the two example MSTl neurons in Fig. 4b, responses are much more similar between the two task reference frames. These examples suggest that VIP responses are more strongly modulated by task reference frame than MSTl responses. Data from the Object+Visual condition for these same example neurons are shown in Fig. 4c,d; data from additional example neurons are shown in Extended Data Fig. 3.

To quantify response modulations related to task reference frame (without assuming a functional form of those modulations), we first computed a modulation index (MI) that captures the net response difference between head and world coordinate tasks (see Methods): larger values of MI indicate greater response differences between task reference frames. We find that MI values are significantly greater for VIP than MSTl neurons (Fig. 5a; Wilcoxon signed rank test, Object+Visual condition: Z=4.10, p=4.1×10^{-5}; Object+Combined: Z=4.46, p=8.2×10^{-6}). This suggests that the representation of object motion in VIP is more dependent on task requirements than that in MSTl.

To characterize the temporal dynamics of response modulations related to task reference frame, we computed MI values within a 300 ms sliding window that was shifted in increments of 50 ms. While MI values are greater in VIP than MSTl at almost all time points (Fig. 5b), they grow substantially over time in both areas. Comparison of the time course of MI (Fig. 5b) with the time course of population responses to the most effective object direction for each neuron (Extended Data Fig. 4a) reveals that response modulations related to task reference frame arise later than stimulus-driven modulations in both MSTl and VIP.

Differences in receptive field sizes and/or locations between brain areas could potentially confound interpretation of the MI data. To evaluate this possibility, we performed an analysis of covariance (ANCOVA) to test whether the difference in MI between MSTl and VIP was robust to including receptive field size and eccentricity as covariates. We found a significant main effect of brain area (F(1, 201)= 5.0, p = 0.026), with no significant dependence on receptive field size (F(3, 201)= 0.7, p = 0.55) or eccentricity (F(3, 201)= 1.85, p = 0.14).

While the MI data of Fig. 5a,b suggest greater task dependency of responses to object motion in VIP, MI is sensitive to any differences in response between tasks, and does not necessarily reflect a shift in the neural representation toward world coordinates in the world task condition. To assess this, we computed a (Pearson) correlation coefficient between tuning in the Object Only condition and tuning in the conditions with self-motion. These correlations were computed among tuning curves expressed in world coordinates (e.g., bottom row of Fig. 4), such that alignment of the tuning curves in world coordinates would yield a large positive correlation coefficient. This analysis was performed by pooling data across the Object+Visual and Object+Combined conditions to gain statistical power, but results were very similar when computed for the Object+Visual and Object+Combined conditions separately (not shown). Across the populations of VIP and MSTd neurons, tuning correlations cover a broad range of values (Fig. 5c), indicating that object tuning is not generally expressed in world coordinates. Our data did not allow us to compute the correlation between tuning curves in head coordinates, since there was insufficient overlap of object directions across self-motion directions when expressed in head coordinates (e.g., top row of Fig. 4).

Critically, our design allowed us to test the hypothesis that correlation among tuning curves in world coordinates becomes stronger in the world coordinate task as compared to the head coordinate task. Indeed, for VIP, we find a robust increase in this correlation for the world coordinate task (Fig. 5c, orange), consistent with a shift in neural tuning toward world coordinates (Wilcoxon signed-rank test, Z=-4.45, $p=8.4\times10^{-6}$). The time course of this correlation reveals that the shift toward world coordinates begins around 1000ms (Fig. 5d), just as stimulus-driven responses in VIP are rising rapidly (Extended Data Fig. 4a). In contrast, we find no significant dependence of tuning correlation on task reference frame for MSTl (Fig. 5c, green, Z=0.175, p=0.86), with only a small shift toward world coordinates occurring late in the trial (Fig. 5d). These data suggest that VIP activity might account for the reference frame shifts seen in behavior, whereas MSTl activity cannot.

Linear decoding of VIP population activity predicts behavioral performance across reference frame conditions

To test whether the observed changes in single-unit responses with task reference frame could account for behavioral performance, we used linear decoders (Fisher Linear Discriminant) to classify object motion as rightward or leftward relative to vertical, based on responses of pseudo-populations of 223 VIP neurons or 177 MSTl neurons (see Methods). We first considered whether neural activity could account for behavioral performance if we trained separate decoders to classify object motion direction in world or head coordinates (Fig. 6a). For the head coordinate task, decoding either MSTl or VIP activity produced a pattern of results (Fig. 6b, blue and cyan curves) similar to the measured behavior (Fig. 3a, blue and cyan curves). This implies that head-centered motion signals can be extracted from VIP and MSTl in the presence of self-motion. Critically, the world coordinate task revealed clear differences in decoding performance between brain areas. For MSTl, decoder performance curves (Fig. 6b right, brown and magenta curves) shifted with self-motion direction to a much greater extent than seen in behavior (Fig. 3a). In contrast, decoding of VIP responses (Fig. 6b left, brown and magenta curves) reveals a pattern of results quite similar to behavior in the world coordinate task, including substantially smaller shifts in the Object+Combined condition than the Object+Visual condition. Because decoder weights were common across Object+Combined and Object+Visual conditions, this improvement in decoder performance with inclusion of vestibular signals is not guaranteed.

To summarize decoder performance, we computed ΔPSE values from the decoder-predicted psychometric functions exactly as done for behavioral data. Results for the VIP decoder lie fairly close to behavioral performance for both world and head coordinate tasks (Fig. 6c, orange vs. black symbols). In contrast, ΔPSE values for the MSTl decoder differ greatly from behavioral metrics in the world coordinate task (Fig. 6c, green vs. black symbols). These findings demonstrate that MSTl responses only partially integrate self-motion signals and cannot be effectively decoded in world coordinates, whereas the representation of object motion in VIP can be decoded linearly to estimate object motion in either reference frame. Very similar results (Extended Data Fig. 5) were obtained using a logistic regression decoder [32], and results were consistent across animals (Extended Data Fig. 6a-d).

In the analysis described above, we computed separate decoders for head and world coordinate tasks, effectively assuming that the brain can apply different read-out weights to the two task contexts. We can further ask whether it is possible to find a single set of decoding weights that allows object direction to be read out in either reference frame. For this purpose, we computed a single decoder that classifies object direction across both task reference frames (Fig. 6d). Strikingly, we found that a single decoder of VIP activity predicts performance in both reference frames that is similar to behavior, whereas a single decoder of MSTl activity fails almost completely for the world coordinate task (Fig. 6e). This finding, which was also quite consistent across animals (Extended Data Fig. 6e-h), suggests that self-motion signals are incorporated into VIP activity when animals are instructed to perform the task in a world-centered reference frame (see Discussion). These results provide strong evidence for a novel role of VIP in constructing a flexible representation of object motion.

A potential confound is that differential responses between the world and head coordinate tasks might be driven by retinal motion of the partial cube frame, which is different between task reference frames. The partial cube frame was generally kept outside of the receptive fields and was faded out during motion of the object and background dots (Fig. 2c) to avoid this confound; nevertheless, it is possible that this could account for some of the task-related response modulations. To assess this possibility, we computed a measure of direction-selective response to the partial cube frame (cube effect index, CEI, see Methods) for both world and head coordinate tasks, focusing on the first 500ms of the trial during which the partial cube is visible and moving but background dots are still largely invisible. For both brain areas, we found very few cells that showed significant directionally-selective responses to the partial cube frame (Extended Data Fig. 7 a,b,d,e), with little difference in median CEI between world and head task conditions (Wilcoxon signed rank test, VIP: Z=1.62, p=0.10; MSTl: Z=-0.83, p=0.41). Nevertheless, we divided the neural populations in half based on the absolute difference in CEI, |ΔCEI|, between world and head tasks (Extended Data Fig. 7c,f), and we performed population decoding for these two subsets of neurons. We found no reliable differences in decoder accuracy between subset of neurons with relatively small and large values of |ΔCEI| (Extended Data Fig. 7g,h for separate decoders, Extended Data Fig. 7i,j for single decoder), suggesting that the partial cube frame was not responsible for differences in VIP responses between world and head task conditions.

Temporal dynamics of reference frame transformations

Since monkeys were trained to switch reference frames based on a visual cue, there might be some delay in gating self-motion signals into the computation of object direction, especially if monkeys rely on judging motion of the partial cube frame, in addition to the color of the fixation point. Thus, we examined the time course of decoder performance using a 300 ms sliding window that was shifted in increments of 50 ms. For this analysis, we used separate decoders for head and world coordinate tasks to give each area the best chance of success. We computed the time course of decoder performance separately for stimulus conditions in which the correct answer is the same for both tasks (“response matched”) and conditions in which the correct answers are different for the two task reference frames (“response conflict”, Fig. 7a). Decoder performance on response conflict trials is of special interest, as it should reveal the clearest differences in neural representations across task reference frames.

Indeed, time courses of decoder accuracy for response conflict trials revealed striking differences between task reference frames and brain areas (Fig. 7b). For VIP on response conflict trials, decoder accuracy rose ~500ms later for the world coordinate task than the head coordinate task (Fig. 7b, left), roughly consistent with the time course of average MI values (Fig. 5b). This late rise in VIP decoder performance for the world task, starting around 1000ms, is too early to be attributed to re-appearance of the partial cube frame, which reaches zero luminance at ~1120ms and does not become clearly visible for another few hundred ms. In contrast, for MSTl, decoder accuracy never reached much above chance for response conflict trials in the world coordinate task, whereas accuracy rose quickly to high levels in the head coordinate task (Fig. 7b, right). Thus, VIP responses undergo a delayed transformation that represents object motion in world coordinates, whereas MSTl responses do not reliably represent motion in world coordinates at any time. Note that decoder performance for VIP in the world task initially dips below chance (0.5) levels, before rising precipitously (Fig. 7b, left). This below-chance performance on response conflict trials is consistent with an early representation of object direction in head coordinates in VIP, which later transitions to world coordinates. Below chance performance for response conflict trials is also seen for MSTl during much of the stimulus period, again reflecting a representation of object direction in head coordinates (Fig. 7b, right).

In response matched conditions, the time course of decoder performance is largely similar for MSTl and VIP in both head and world coordinate tasks (Fig. 7c). The only notable difference is that MSTl decoder accuracy drops off somewhat toward the end of the stimulus period, whereas accuracy of the VIP decoder is more sustained. This difference was not attributable to receptive field coverage (Extended Data Fig. 2) or to the temporal profiles of neural responses (Extended Data Fig. 4). We also examined decoder performance separately for trials that followed a switch between task reference frames, as compared to trials for which the reference frame did not change, and we found no clear differences (data not shown).

Our findings suggest that self-motion signals are incorporated into the computation of object motion in VIP during the world coordinate task. To probe this idea further, we took advantage of the fact that monkeys judged object direction in both reference frames for each unique random-dot stimulus. This allowed us to perform cross-task decoding by training a classifier to perform the world coordinate task using neural responses from the head coordinate task, or vice-versa. For this analysis, we focused on the response conflict conditions, for which differences between areas and task reference frames are most clear.

For VIP, the decoder could perform the head coordinate task using responses from the world coordinate task (Fig. 8b, gray/black), but could not perform the world coordinate task using responses from the head coordinate task (Fig. 8a, gray/black). This suggests that VIP activity contains information about object motion in head coordinates in both task conditions, but only represents object motion in world coordinates when the animal is performing the world coordinate task. By comparison, a decoder of MSTl activity fails to perform the world coordinate task at all times using responses from either task condition (Fig. 8c), whereas it accurately reports object direction in head coordinates when trained on responses from either task condition (Fig. 8d). These observations are consistent with the hypothesis that both areas carry robust information about object direction in a head-centered reference frame under all conditions, whereas self-motion signals are incorporated into the computation of object motion in VIP (but not MSTl) when the task requires a world-centered reference frame.

Dissociation of choice signals from task reference frame signals

Given that monkeys' choices are clearly different between the world and head task conditions (Fig. 3), can the effects of task instruction on VIP responses be simply accounted for by choice-related modulations? This question is especially relevant given that VIP neurons often have strong choice-related activity that is not predictable from their stimulus tuning [33]. We performed analyses, at both the single neuron and population levels, which demonstrate that response modulations related to task instruction are distinct from choice-related activity.

At the single-unit level, we separately quantified choice- and task-related activity (see Methods). For choice-related activity, we computed the familiar choice probability (CP) metric [34], which involves sorting responses into choice groups separately for each distinct stimulus and task condition. Analogously, we quantified task-related modulations by computing a 'task probability (TP)' metric, which involves sorting responses into groups based on task reference frame. Critically, this is done separately for each choice and stimulus condition before z-scoring and pooling, to ensure that TP is not influenced by choice. Many neurons in VIP and MSTl show significant CP and/or TP values (Extended Data Fig. 8a,b); however, we find no correlation between CP and TP across the population (VIP: r=-0.062, p=0.36; MSTl: r = -0.041, p=0.59, Pearson correlation). To probe this dissociation further, we computed TP separately for left and right choices (Extended Data Fig. 8c,d), and we find these values to be strongly correlated (VIP: r=0.76, p=9.9x10^{-42}; MSTl: r = 0.76, p=1.6x10^{-32}, Pearson correlation), indicating that task-related modulations for individual neurons are consistent across left and right choices. Similarly, we find that CP values are correlated across task reference frames (VIP: r=0.50, p=2.3x10^{-15}; MSTl: r = 0.33, p=1.7x10^{-5}, Extended Data Fig. 8e,f, Pearson correlation). Together, these results show that choice and task reference frame have separable effects on responses of individual neurons.

To assess whether performance of our decoders could be confounded with choice-related activity, we devised an approach (see Methods) to largely remove either the choice-related or task-related response modulations for each neuron. Our approach for removing choice-related activity virtually eliminated significant CP values while leaving TP values largely unchanged (compare Extended Data Fig. 9a,b with Extended Data Fig. 8a,b). Similarly, our method for removing task-related modulations largely eliminated significant TP values while leaving CP values largely unchanged (Extended Data Fig. 9c,d vs. Extended Data Fig. 8a,b). Critically, decoder performance on response conflict trials was greatly impaired when task-related modulations were removed (compare Extended Data Fig. 9e to Fig. 7b), whereas there was little effect on decoder performance of removing choice-related modulations in response (Extended Data Fig. 9f). These findings demonstrate that the flexible reference frame exhibited by VIP activity cannot simply be attributed to choice-related activity.

DISCUSSION

By training monkeys to report object motion in either head or world coordinates, we examined whether neurons represent object direction in a fixed reference frame or one that changes with task requirements. Our findings demonstrate that VIP, but not MSTl, contains a flexible representation of object motion that dynamically changes from moment to moment to represent object direction relative to the head or world. The dynamics of these effects suggest that self-motion signals are incorporated into the representation of object motion in VIP when the task requires a world-centered representation.

Our findings provide an important advance in understanding how the brain represents object motion during self-motion, providing the first evidence for a flexible multi-sensory representation that can signal object motion relative to the head or world. The fact that addition of vestibular stimulation facilitates reference frame transitions is consistent with previous psychophysical and neurophysiological studies showing that vestibular input helps to dissociate object motion and self-motion [19, 24, 25, 35, 36]. More generally, our findings provide compelling evidence that the reference frame of neural representations is not static, and can be powerfully modulated by task instructions.

Caveats and limitations

It was difficult for animals to switch reference frames from trial to trial without the partial cube frame, potentially because the screen edges provide a strong head-centered frame in the absence of the partial cube (see Methods). Although the partial cube frame gradually became invisible while the moving object became visible (Fig. 2c), it is possible that animals might have tried to judge the horizontal velocity of object motion relative to specific edges or corners of the partial cube, rather than judging object velocity relative to the world. Two factors argue strongly against this interpretation: 1) a visual strategy of judging object velocity relative to specific features of the partial cube frame would not explain the behavioral and neural effects of physical motion of the platform, which provided vestibular signals. 2) We performed control experiments in which we varied the location of the partial cube frame in depth from trial to trial. If animals were reporting object motion relative to the near or far edges of the cube frame, their ΔPSE values would be expected to depend systematically on cube depth, since the retinal velocity of cube features depends on their distance from the observer. In contrast, we found no significant dependence on depth of the cube (Extended Data Fig. 10), suggesting that the animals did not employ this strategy.

Because we were not able to record from large ensembles of neurons simultaneously, our decoding analyses were based on pseudo-populations of neurons for which the noise correlations were largely unknown (see Methods). Thus, our analyses effectively assumed that neurons had independent noise, which is not accurate [37, 38]. It is well established that correlated noise among neurons can influence the information content (or sensitivity) of a population code [39-41]. Importantly, all of our main conclusions are based upon estimates of biases in decoding performance, not on sensitivity measures. While we cannot rule out the possibility that correlated noise would influence biases in decoder performance, it seems unlikely that the pattern of results would change qualitatively.

Our analyses assume that the monkeys always identified the object as moving relative to the background field of dots, which is a safe assumption for two reasons. First, object motion always contained a substantial vertical component that was not compatible with the horizontal self-motion of the animal. Second, the moving object was visually distinct from background dots (see Methods), such that it was easily segmented from the background. More generally, the brain has to solve a causal inference problem to discern whether the retinal motion of an object is produced by self-motion or also reflects independent movement of the object relative to the scene [42]. The neural basis of this causal inference process will be the topic of future studies.

Relationship to previous studies

A previous study [29] reported that visual tracking neurons in area MSTl represent visual target motion in world coordinates while macaques tracked a target using voluntary eye and head rotations. It was suggested that MSTl neurons represent world-centered target motion by combining retinal motion signals, efference copy signals related to smooth eye movement, and vestibular signals related to head rotation. On the surface, the findings of Ilg et al. [29] appear to conflict with our finding that MSTl does not represent object motion in world coordinates. However, our subjects performed the discrimination task while their eyes and head remained oriented straight ahead. Whereas the tracking task used by Ilg et al. [29] elicited extra-retinal signals related to eye and head rotation, our stimuli involved real or simulated head translations. The findings of Ilg et al. [29] are therefore not incompatible with ours, and collectively they suggest that some MSTl neurons may account for eye and head rotations, but that MSTl does not contain a generalized representation of world-referenced object motion.

Reference frames of different sensory signals in area VIP have been the focus of several previous studies. Facial tactile receptive fields (RFs) are coded in a head-centered reference frame [2], whereas auditory RFs are organized in a continuum between eye- and head-centered coordinates [10, 11]. Visual RFs and heading tuning (optic flow) are represented mainly in an eye-centered reference frame [11, 31, 43], although some studies have described head-centered visual RFs in VIP as well [4]. In contrast, vestibular heading signals in VIP are coded in body- or world-centered reference frames [30, 44]. One recent study showed that the reference frame of vestibular heading tuning in VIP depended on whether gaze was focused on a head- or world-fixed target [44], consistent with the idea that VIP has flexible reference frames. However, our findings show that VIP reference frames can change just by task instructions, and do not require different motor actions [44].

Our findings substantially extend previous work on the context-dependence and dynamics of spatial reference frames. Human neuroimaging studies have reported that visual motion signals can be represented in retinal or head coordinates depending on the spatial allocation of attention [45], although this finding has been refuted by other studies [46]. Human fMRI studies also demonstrated that activity in parietal and premotor cortex reflected different spatial reference frames depending on the sensory modality used to specify target location [47]. Previous studies have also demonstrated that the reference frame of neural activity in monkeys is dynamic, changing over time relative to task events [8, 48]. Our findings extend this work in two important ways. First, we demonstrate that neural activity is modified by task instructions to represent object motion in the reference frame required for each task condition. Second, we directly compare neural and behavioral correlates of dynamically changing task reference frames, allowing for a more direct assessment of whether changes in neural activity with task reference frame can explain behavior. In contrast, previous neurophysiological studies of reference frames have generally just varied the position of an effector without requiring animals to make a perceptual report.

The Duncker illusion [49, 50] describes biases in the perceived trajectory of an object when it moves relative to a moving background. The perceptual biases exhibited by our monkeys cannot simply be accounted for by the Duncker illusion because the image motion of the target object and background dots are identical in the world and head coordinate tasks, yet the perceptual reports are strikingly different (Fig. 3a).

Implications of flexible reference frames

The primary contribution of our study is to demonstrate that neural reference frames can change dramatically based on task instructions. Secondarily, these results have implications for understanding the variability of outcomes across previous studies of neural reference frames. Two examples of such variability, as noted above, include the incidence of head-centered visual receptive fields in VIP [4, 43] and the existence of spatiotopic representations in human visual cortex [45, 46]. Given that the vast majority of neurophysiological studies of reference frames have not used a behavioral task that enforces a specific task reference frame, findings could vary with the intrinsic (and uncontrolled) reference frame that the animal employs, which in turn may depend on the animal's previous experience or training history. Findings could also vary with the stimuli used, which might bias the animal toward adopting a specific task reference frame.

We found that a single set of decoding weights could be used to classify object direction in either head or world coordinates, based on VIP activity. This result could arise because task instructions simply shift the 'population hill' of neural activity along the stimulus axis, similar to changing object direction itself. A pure horizontal shift of the population hill, in which the pattern of population activity is simply translated along the object direction axis, would occur if all tuning curves for object direction simply shifted with self-motion in the world coordinate task. This was clearly not the case based on inspection of tuning curves from individual neurons (Fig. 4 and Extended Data Fig. 3), as well as the broad distribution of tuning correlation values in Fig. 5c. We found that self-motion has diverse effects on object motion tuning in the world coordinate task, including shifts, gain changes, and changes in shape of tuning. Thus, it remains an interesting topic for future studies to determine how a single decoder can estimate object direction in head or world coordinates based on such diverse modulations at the single-unit level.

Acknowledgements: This work was supported by NIH grant EY01618 (to GCD), The Uehara Memorial Foundation (to RS), the Japan Society for the Promotion of Science (to RS), and an NEI CORE grant (EY001319). We thank D. Graf, S. Shimpi, and E. Murphy for excellent technical support, and J. Wen and A. Yung for programming support.

Author contributions: R.S. and G.C.D. conceived and designed research; R.S. performed experiments; R.S. analyzed data; A.A. built recording system; R.S., A.A., D.E.A., and G.C.D. interpreted results of experiments; R.S. prepared figures; R.S. and G.C.D. drafted manuscript; R.S., A.A., D.E.A., and G.C.D. edited and revised manuscript; R.S., A.A., D.E.A., and G.C.D. approved final version of manuscript.

Competing interests: The authors declare no competing interests.

REFERENCES

1. Andersen, R.A., Essick, G.K. & Siegel, R.M. Encoding of spatial location by posterior parietal neurons. *Science* **230**, 456-458 (1985).
2. Avillac, M., Deneve, S., Olivier, E., Pouget, A. & Duhamel, J.R. Reference frames for representing visual and tactile locations in parietal cortex. *Nat Neurosci* **8**, 941-949 (2005).
3. Batista, A.P., Buneo, C.A., Snyder, L.H. & Andersen, R.A. Reach plans in eye-centered coordinates. *Science* **285**, 257-260 (1999).
4. Duhamel, J.R., Bremmer, F., Ben Hamed, S. & Graf, W. Spatial invariance of visual receptive fields in parietal cortex neurons. *Nature* **389**, 845-848 (1997).
5. Fetsch, C.R., Wang, S., Gu, Y., Deangelis, G.C. & Angelaki, D.E. Spatial reference frames of visual, vestibular, and multimodal heading signals in the dorsal subdivision of the medial superior temporal area. *J Neurosci* **27**, 700-712 (2007).
6. Galletti, C., Battaglini, P.P. & Fattori, P. Parietal neurons encoding spatial locations in craniotopic coordinates. *Exp Brain Res* **96**, 221-229 (1993).
7. Jay, M.F. & Sparks, D.L. Auditory receptive fields in primate superior colliculus shift with changes in eye position. *Nature* **309**, 345-347 (1984).
8. Lee, J. & Groh, J.M. Auditory signals evolve from hybrid- to eye-centered coordinates in the primate superior colliculus. *J Neurophysiol* **108**, 227-242 (2012).
9. Mullette-Gillman, O.A., Cohen, Y.E. & Groh, J.M. Eye-centered, head-centered, and complex coding of visual and auditory targets in the intraparietal sulcus. *J Neurophysiol* **94**, 2331-2352 (2005).
10. Mullette-Gillman, O.A., Cohen, Y.E. & Groh, J.M. Motor-related signals in the intraparietal cortex encode locations in a hybrid, rather than eye-centered reference frame. *Cereb Cortex* **19**, 1761-1775 (2009).
11. Schlack, A., Sterbing-D'Angelo, S.J., Hartung, K., Hoffmann, K.P. & Bremmer, F. Multisensory space representations in the macaque ventral intraparietal area. *J Neurosci* **25**, 4616-4625 (2005).
12. Snyder, L.H., Grieve, K.L., Brotchie, P. & Andersen, R.A. Separate body- and world-referenced representations of visual space in parietal cortex. *Nature* **394**, 887-891 (1998).
13. Sajad, A.*, et al.* Visual-Motor Transformations Within Frontal Eye Fields During Head-Unrestrained Gaze Shifts in the Monkey. *Cereb Cortex* **25**, 3932-3952 (2015).
14. Kiesel, A.*, et al.* Control and interference in task switching--a review. *Psychol Bull* **136**, 849-874 (2010).
15. Ruge, H., Jamadar, S., Zimmermann, U. & Karayanidis, F. The many faces of preparatory control in task switching: reviewing a decade of fMRI research. *Hum Brain Mapp* **34**, 12-35 (2013).
16. Stoet, G. & Snyder, L.H. Neural correlates of executive control functions in the monkey. *Trends Cogn Sci* **13**, 228-234 (2009).
17. Stoet, G. & Snyder, L.H. Single neurons in posterior parietal cortex of monkeys encode cognitive set. *Neuron* **42**, 1003-1012 (2004).
18. Kim, H.R., Pitkow, X., Angelaki, D.E. & DeAngelis, G.C. A simple approach to ignoring irrelevant variables by population decoding based on multisensory neurons. *J Neurophysiol* **116**, 1449-1467 (2016).

19. Sasaki, R., Angelaki, D.E. & DeAngelis, G.C. Dissociation of Self-Motion and Object Motion by Linear Population Decoding That Approximates Marginalization. *J Neurosci* **37**, 11204-11219 (2017).

20. Rushton, S.K. & Warren, P.A. Moving observers, relative retinal motion and the detection of object movement. *Curr Biol* **15**, R542-543 (2005).

21. Warren, P.A. & Rushton, S.K. Optic flow processing for the assessment of object movement during ego movement. *Curr Biol* **19**, 1555-1560 (2009).

22. Royden, C.S. & Connors, E.M. The detection of moving objects by moving observers. *Vision Res* **50**, 1014-1024 (2010).

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

24. Fajen, B.R. & Matthis, J.S. Visual and non-visual contributions to the perception of object motion during self-motion. *PLoS One* **8**, e55446 (2013).

25. Dokka, K., MacNeilage, P.R., DeAngelis, G.C. & Angelaki, D.E. Multisensory self-motion compensation during object trajectory judgments. *Cereb Cortex* **25**, 619-630 (2015).

26. MacNeilage, P.R., Zhang, Z., DeAngelis, G.C. & Angelaki, D.E. Vestibular facilitation of optic flow parsing. *PLoS One* **7**, e40264 (2012).

27. Eifuku, S. & Wurtz, R.H. Response to motion in extrastriate area MSTl: center-surround interactions. *J Neurophysiol* **80**, 282-296 (1998).

28. Tanaka, K., Sugita, Y., Moriya, M. & Saito, H. Analysis of object motion in the ventral part of the medial superior temporal area of the macaque visual cortex. *J Neurophysiol* **69**, 128-142 (1993).

29. Ilg, U.J., Schumann, S. & Thier, P. Posterior parietal cortex neurons encode target motion in world-centered coordinates. *Neuron* **43**, 145-151 (2004).

30. Chen, X., DeAngelis, G.C. & Angelaki, D.E. Diverse spatial reference frames of vestibular signals in parietal cortex. *Neuron* **80**, 1310-1321 (2013).

31. Chen, X., DeAngelis, G.C. & Angelaki, D.E. Eye-centered representation of optic flow tuning in the ventral intraparietal area. *J Neurosci* **33**, 18574-18582 (2013).

32. Berens, P., *et al.* A fast and simple population code for orientation in primate V1. *J Neurosci* **32**, 10618-10626 (2012).

33. Zaidel, A., DeAngelis, G.C. & Angelaki, D.E. Decoupled choice-driven and stimulus-related activity in parietal neurons may be misrepresented by choice probabilities. *Nat Commun* **8**, 715 (2017).

34. Britten, K.H., Newsome, W.T., Shadlen, M.N., Celebrini, S. & Movshon, J.A. A relationship between behavioral choice and the visual responses of neurons in macaque MT. *Vis Neurosci* **13**, 87-100 (1996).

35. Dokka, K., DeAngelis, G.C. & Angelaki, D.E. Multisensory Integration of Visual and Vestibular Signals Improves Heading Discrimination in the Presence of a Moving Object. *J Neurosci* **35**, 13599-13607 (2015).

36. Sasaki, R., Angelaki, D.E. & DeAngelis, G.C. Processing of object motion and self-motion in the lateral subdivision of the medial superior temporal area in macaques. *J Neurophysiol* **121**, 1207-1221 (2019).

37. Chen, A., DeAngelis, G.C. & Angelaki, D.E. Functional specializations of the ventral intraparietal area for multisensory heading discrimination. *J Neurosci* **33**, 3567-3581 (2013).

38. Gu, Y., *et al.* Perceptual learning reduces interneuronal correlations in macaque visual cortex. *Neuron* **71**, 750-761 (2011).

39. Kohn, A., Coen-Cagli, R., Kanitscheider, I. & Pouget, A. Correlations and Neuronal Population Information. *Annu Rev Neurosci* **39**, 237-256 (2016).

40. Averbeck, B.B., Latham, P.E. & Pouget, A. Neural correlations, population coding and computation. *Nat Rev Neurosci* **7**, 358-366 (2006).

41. Moreno-Bote, R., *et al.* Information-limiting correlations. *Nat Neurosci* **17**, 1410-1417 (2014).

42. Dokka, K., Park, H., Jansen, M., DeAngelis, G.C. & Angelaki, D.E. Causal inference accounts for heading perception in the presence of object motion. *Proc Natl Acad Sci U S A* **116**, 9060-9065 (2019).

43. Chen, X., DeAngelis, G.C. & Angelaki, D.E. Eye-centered visual receptive fields in the ventral intraparietal area. *J Neurophysiol* **112**, 353-361 (2014).

44. Chen, X., DeAngelis, G.C. & Angelaki, D.E. Flexible egocentric and allocentric representations of heading signals in parietal cortex. *Proc Natl Acad Sci U S A* **115**, E3305-E3312 (2018).

45. Crespi, S., *et al.* Spatiotopic coding of BOLD signal in human visual cortex depends on spatial attention. *PLoS One* **6**, e21661 (2011).

46. Merriam, E.P., Gardner, J.L., Movshon, J.A. & Heeger, D.J. Modulation of visual responses by gaze direction in human visual cortex. *J Neurosci* **33**, 9879-9889 (2013).

47. Bernier, P.M. & Grafton, S.T. Human posterior parietal cortex flexibly determines reference frames for reaching based on sensory context. *Neuron* **68**, 776-788 (2010).

48. Bremner, L.R. & Andersen, R.A. Temporal analysis of reference frames in parietal cortex area 5d during reach planning. *J Neurosci* **34**, 5273-5284 (2014).

49. Duncker, K. Uber induzierte Bewegung. *Psychologische Forschung* **12**, 180-259 (1929).

50. Zivotofsky, A.Z. The Duncker illusion: intersubject variability, brief exposure, and the role of eye movements in its generation. *Invest Ophthalmol Vis Sci* **45**, 2867-2872 (2004).

FIGURE LEGENDS

Figure 1. Schematic illustration of interactions between object motion and self-motion. (a) An object (gray sphere) moves upward in the world while an observer is translated rightward or leftward at two speeds. (b) Resultant image motion vectors. Without self-motion, image motion is upward (white). During self-motion, image motion is biased according to the direction and speed of self-motion. For simplicity, the image view in panel b does not reflect image reversals that would be caused by projection onto the retina.

Figure 2. Behavioral task design and predicted psychometric functions. (a) A sphere of dots moves up-right (+θ) or up-left (-θ) in the world. Rightward or leftward self-motion occurs while the animal views the moving object. (b) In the world coordinate task, a world-fixed partial cube indicates that the monkey should report object motion relative to the world. In the head coordinate task, the partial cube remains fixed relative to the head, and cues a report in head coordinates. Dashed vertical lines indicate fixed world-centered locations as a reference. Background dots were presented at 40% coherence, but background motion here is depicted with 100% coherence for visual clarity. (c) Time course of the luminance of visual stimulus components. The luminance of the object and partial cube were changed dynamically such that the partial cube faded out during the portion of the trial when the object faded in. (d) Hypothetical psychometric functions that plot the proportion of 'rightward' choices as a function of object direction in world coordinates. If the animal compensates fully for self-motion in the world coordinate task, psychometric functions for rightward and leftward self-motion should overlap (magenta). On the other hand, those functions should shift with self-motion by a specific

amount (horizontal bar) in the head coordinate task (cyan). Dashed/solid curves: leftward/rightward self-motion.

Figure 3. Summary of behavioral performance for each task reference frame. (a) Psychometric functions showing the proportion of rightward choices as a function of object direction in world coordinates (positive = rightward). Data are shown for trials in which there is no self-motion (Object Only, black), for trials with self-motion in which the animal performs the head coordinate task (cyan/navy), and for trials in which the animal performs the world-coordinate task (magenta/brown). Darker colors (navy/brown) represent the Object+Visual condition and lighter colors (cyan/magenta) represent the Object+Combined condition. Filled/open symbols: rightward/leftward self-motion. Smooth curves are cumulative Gaussian fits to data pooled across 128 sessions for Monkey N and 57 sessions for monkey K. (b) Summary of behavioral biases, quantified as the difference in point of subjective equality (ΔPSE) between psychometric functions for rightward and leftward self-motion. ΔPSE values are compared between the Object+Combined and Object+Visual conditions for each recording session and each monkey (squares: monkey N; diamonds: monkey K). Data are shown separately for the world coordinate (red) and head coordinate (blue) task conditions. Error bars represent 95% confidence intervals around the mean values (black symbols) across 185 sessions.

Figure 4. Data from example neurons recorded from areas VIP and MSTl. (a) Data from two example VIP neurons (one neuron per column) recorded in the Object+Combined condition. The top and bottom rows plot firing rates as a function of object direction in head and world coordinates, respectively. (b) Data from two example MSTl neurons in the Object+Combined

condition. Note the greater differences in response between the head (cyan) and world (magenta) coordinate task conditions for the example VIP neurons, as compared to the example MSTl neurons. Error bars denote SEM (across n=10 stimulus repetitions). (c, d) Data for the same 4 example neurons from the Object+Visual condition.

Figure 5. Summary of single-unit results for VIP and MSTl. (a) Summary of modulation index (MI) values for populations of neurons recorded from VIP (orange, N=223) and MSTl (green, N=177) in the Object+Visual (top) and Object+Combined (bottom) conditions. MI measures the response difference between a pair of object tuning curves in the head- and world-coordinate tasks. Filled bars represent MI values significantly greater than zero (permutation test, $p < 0.05$). Numbers in the legend indicate the total number of neurons, as well as the number with MI values significantly greater than zero, for each brain area. Arrowheads and numbers indicate the median values for each brain area and self-motion condition. (b) Time course of average MI values for VIP (orange, N=223) and MSTl (green, N=177) neurons in the Object+Visual (lighter hues) and Object+Combined (darker hues) conditions. Error bars represent 95% confidence intervals. Gray curve shows the Gaussian temporal profile of object speed. (c) The correlation between object direction tuning (computed in world coordinates) is compared for the world and head coordinate tasks. Data from VIP and MSTl are shown in orange and green, respectively. Data are included in this panel only for neurons (VIP: N=57; MSTl: N=44) that had significant tuning (ANOVA, $p<0.05$) in the Object Only condition. Star symbols denote the three neurons in Figure 4 that met this criterion. (d) Time course of the difference in tuning correlation between world and head coordinate tasks for the same

populations of VIP (orange, N=223) and MSTl (green, N=177) neurons described in panel c. Error bars represent 95% confidence intervals.

Figure 6. Summary of population decoding results. Panels a-c correspond to results from training separate decoders to perform the world and head coordinate tasks; panels d-f correspond to results from a single decoder trained to perform in both task reference frames. (a) Schematic diagram of separate decoders for the head and world coordinate task conditions. **(**b) Results for separate world/head task decoders, plotted in the same format as the behavioral data of Fig. 3a. Decoding VIP activity produces a pattern of results very similar to behavior, whereas decoding MSTl produces large biases in the world coordinate task. (c) Summary comparison of monkey behavior and performance of the separate decoders. ΔPSE for the Object+Combined condition is plotted against ΔPSE for the Object+Visual condition. Results from the VIP decoder (orange) are largely similar to behavior (black, same data from Fig. 3b), whereas results from the MSTl decoder (green) depart sharply from behavioral performance for the world coordinate task. Error bars on decoder performance values represent 95% confidence intervals obtained by bootstrapping (n=1000 bootstraps, see Methods). Pink and cyan dashed lines: expected ΔPSE for perfect performance in the world and head coordinate tasks, respectively. (d) Schematic diagram for the single decoder. (e) Analogous results to panel b, but from a single decoder trained to classify object direction in both task reference frames. (f) Summary comparison of single decoder results with behavior. Format as in panel c; error bars represent 95% confidence intervals across n=1000 bootstraps.

Figure 7. Time course of decoder performance. (a) Schematic illustration of examples of "response matched" and "response conflict" conditions. In a response matched condition (left), correct answers are the same for both task reference frames; in a response conflict condition (right), correct reports are opposite for the two reference frames. Magenta and cyan vectors indicate object direction in world and head coordinates, respectively. (b) Time course of decoder classification accuracy for populations of VIP (left, n=223) and MSTl (right, n=177) neurons, evaluated in the subset of response conflict conditions. Separate decoders were trained to classify object direction in the world (magenta/brown) and head (cyan/navy) reference frames at each time point. Error bars represent 95% confidence intervals (across n=100 bootstraps). (c) Time course of decoder classification accuracy in the subset of response matched conditions, format as in b. Time courses were obtained by computing each variable within a 300 ms sliding time window that was advanced across the trial epoch in steps of 50 ms.

Figure 8. Time courses of classification accuracy using within-task vs. cross-task decoding. (a, b) Results for decoding VIP activity. In panel a, the decoder is trained to classify object direction in world coordinates using responses from the world task condition (magenta/brown, within-task) or using responses from the head task condition (light/dark gray, cross-task). In panel b, the decoder is trained to classify object direction in head coordinates using responses from the head task condition (cyan/navy, within-task) or the world task condition (light/dark gray, cross-task). (c,d) Analogous results for within-task (colors) and cross-task (gray) decoders of MSTl activity. In all panels, error bars denote 95% confidence intervals (across n=100 bootstraps).

METHODS

General

Two male rhesus monkeys (*Macaca mulatta*) participated in this study. During this study, monkey K ranged in age from 4 to 6 years and ranged in weight from 5.8 to 8.5 kg. Monkey N ranged in age from 5 to 7 years and weight from 7.2 to 9.7 kg. General procedures have been described previously [19, 51]. All experimental procedures conformed to National Institutes of Health guidelines and were approved by the University Committee on Animal Resources at the University of Rochester. Additional information can be found in the Life Sciences Reporting Summary.

Vestibular and visual stimuli

A 6 degree-of-freedom motion platform (MOOG 6DOF2000E; Moog) was used to passively translate animals leftward or rightward along the interaural axis. Visual stimuli were projected onto a tangent screen by a three-chip digital light projector (Mirage S^+3K ; Christie Digital Systems, Cypress, CA). The display screen measured 60 x 60 cm and was mounted ~30 cm in front of the monkey, thus subtending ~90 x 90° of visual angle. Visual stimuli simulated translational self-motion through a three-dimensional field of stars. Each star was a triangle that measured 0.15 cm x 0.15 cm, and the field of stars measured 100 cm wide by 100 cm tall by 40 cm deep, with a star density of 0.01 stars per cm^3. To provide stereoscopic cues, the star field was rendered as a red-green anaglyph and viewed through custom red-green goggles, consisting of Kodak Wratten2 filters (#29 and #61). The entire display was visible through the colored filters.

The optic flow field contained naturalistic cues simulating lateral translation of the observer in the horizontal plane; these included motion parallax, size, and binocular disparity cues. While the monkey was translated leftward or rightward, an object also moved upward in the world with a small leftward or rightward component (Fig. 1). The moving object was a transparent sphere (diameter 10°) composed of random dots, with a density (0.25 dots/cm^3) that was higher than that of the star-field background, such that the object was easily segmented from the background. The moving object's center was located in depth within the plane of the visual display, such that it consisted of dots with a mixture of crossed and uncrossed disparities. At the start of each trial, the object appeared with its center located 5 deg left of fixation and 10 deg below fixation. The object moved in one of 7 directions relative to upward in the world: -21, -14, -7, 0, 7, 14 and 21 deg, where negative angles represent upward/left motion, positive angles represent upward/right motion, and 0 means straight upward (in world coordinates). All self-motion and object motion trajectories were straight translational movements with a duration of 2 sec, and having a Gaussian velocity profile with a SD of 1/3 sec [51]. Peak stimulus velocity occurred at 1120ms after stimulus onset, due to delays and dynamics of the motion platform; visual stimuli were synchronized to platform motion. The total excursion (0.25 m) and peak velocity (0.75 m/s) of object motion were greater than those for self-motion (0.08 m and 0.24 m/s, respectively). Because the head-centered velocity of the object is determined by both self-motion velocity and object velocity relative to the world, the object could move up/left in world coordinates and up/right in head coordinates, or vice-versa. While we shall distinguish between world- and head-centered references frames in this study, we cannot distinguish head-centered and retinal reference frames because the fixation target was always head-fixed.

Two different versions of the task were interleaved that required the animal to report object direction in either head or world coordinates (Fig. 2). In the world coordinate task (Fig. 2b), a partially-visible cube defined a world-fixed reference frame that was updated every video frame. The cube dimensions were 76 cm wide by 76 cm tall by 40 cm deep and the center of the cube was located in depth at the fixation point. Thus, the cube moved relative to the head during self-motion in the world coordinate task (Fig. 2b) but remained head-fixed in the head coordinate task (Fig. 2b).

We attempted to train monkeys to switch between the world and head coordinate tasks based solely on the color of the fixation point. While one animal could partially achieve this, the other animal could not. Both animals were much better able to switch between tasks when the partially-visible cube was presented. We think that the partial cube was particularly important because it was not possible to eliminate luminance boundaries at the edge of the display. Since the display screen translated with the animal, the luminance boundaries always provided a head-centered reference frame, and thus the partially visible cube was important to help define a world reference frame (anecdotally, this was the case for human observers also).

A potential concern about use of the partially visible cube is that animals might learn to report object direction relative to the moving visual elements of the cube. Two measures helped to prevent this possibility. First, the luminance of the moving object was dynamically changed according to the same Gaussian envelope that governed the speed of object and self-motion, such that the moving object was initially invisible, reached maximum brightness in the middle of the presentation when it also reached maximum speed, and then decayed again to become invisible at the end of the trial (Fig. 2b,c). Simultaneously, the luminance of the partially visible cube followed an inverted Gaussian velocity profile, such that the cube was maximally visible at the beginning and end of each trial and disappeared in the middle of the trial (Fig. 2b,c). This allowed the partial cube to define the reference frame while having little overlap with the visibility of the moving object.

Second, to assess whether animals might have still judged object motion relative to the partially-visible cube, we performed behavioral control experiments in which we randomly varied the location of the partial cube in depth from trial to trial. If animals reported object direction relative to the cube, then their performance should depend systematically on the depth of the cube. We found no such dependence (Extended Data Fig. 10), indicating that animals were successfully prevented from adopting this strategy. Thus, we believe that both animals successfully learned to report object direction in world or head coordinates.

Behavioral task

Monkeys were trained to report whether the object moved up-left or up-right in either head- or world-coordinates (Fig. 2). This was a very challenging task for animals to learn, and required 1.5-2 years of training for each animal. We initially trained the animals to perform the head and world tasks in separate blocks of trials. We then gradually reduced the length of these blocks, and then transitioned animals to trial-by-trial interleaving of the two tasks.

In each trial, a fixation point initially appeared. Once the monkey looked at the fixation point, the object appeared and moved upward in the virtual environment, with a small rightward or leftward component. The monkey reported whether the object moved upward/rightward or upward/leftward by making a saccade to one of two targets that appeared (10 degrees to the right and left of the fixation target) after a 500 ms delay period following the end of the visual stimulus. In the world and head coordinate tasks (Fig. 2b), the monkey reported whether object motion moved leftward or rightward relative to vertical in world or head coordinates, respectively. The two versions of the task were cued by the shape and color of the fixation point (Fig. 2b), such that they could be randomly interleaved. Animals were rewarded for reporting the correct direction of object motion in each reference frame condition. When object direction was exactly vertical (in the relevant reference frame), monkeys were rewarded randomly on 50% of trials.

Crucially, for each particular combination of a self-motion direction and an object motion direction in the world, the motion trajectory of the object in head (or retinal) coordinates was identical across the two task conditions. Thus, for the same exact motion of all of the dots on the screen, the animal might be required to make a rightward choice in the world coordinate task and a leftward choice in the head coordinate task, or vice-versa. For other stimulus conditions, the correct choice would be the same in both reference frames. This allowed us to compare decoder performance for subsets of trials in which the correct answers were the same or different for the two tasks.

Three self-motion conditions were interleaved for each task reference frame. 1) In the Object Only condition, there was no self-motion such that world- and head-centered reference frames are aligned. Background dots were stationary on the display in the Object Only condition, since the only source of image motion for the background dots is self-motion. 2) In the Object+Visual condition, the motion platform remained stationary while a background of random dots provided optic flow that simulated leftward or rightward self-motion. Background optic flow had a motion coherence of 40% such that the object was easy to segment from the background. 3) In the Object+Combined condition, self-motion was indicated by both optic flow and physical translation of the motion platform (which provided vestibular cues). Since cue combination is known to enhance heading perception [52, 53], we expected that the monkeys would be best able to compensate for self-motion in this condition.

Behavioral data analysis

Psychometric functions were constructed by plotting the proportion of 'rightward' choices as a function of object direction in world coordinates. Plotted in this fashion, psychometric functions for rightward and leftward self-motion should overlap in the world coordinate task if the animal compensates fully for self-motion (Fig. 2d). If the animal does not account for self-motion and reports object direction in head coordinates, there will be a large horizontal shift between psychometric functions corresponding to leftward and rightward self-motion (Fig. 2d). To quantify these shifts, we fit each psychometric curve with a cumulative Gaussian function and used the mean parameter of the fit to estimate the point of subjective equality (PSE) for each direction of self-motion. The difference in PSE (ΔPSE) between rightward and leftward self-motion directions was then taken as an index of the reference frame used by the monkeys to judge object direction (Fig. 3b). If the monkey correctly estimates object direction in head coordinates, then we expect ΔPSE=35.5 deg (horizontal bar, Figs. 2d, 3a).

Physiological recording procedures

Neural recordings were obtained from the right hemisphere of two monkeys while the animals performed the behavioral task. We recorded 223 VIP neurons (monkey N, N=93; monkey K, N=130) and 177 MSTl neurons (monkey N, N=94; monkey K, N=83), with most neurons recorded in separate sessions. We attempted to record from any VIP and MSTl neuron that could be isolated; there were no selection criteria based on response properties other than receptive field location. Recordings were included if we obtained data for at least 3 repetitions for each stimulus condition. For Monkey N, 68 VIP and 50 MSTl neurons were recorded with single tungsten microelectrodes (FHC, Bowdoinham, ME; 0.5 – 1 MΩ impedance). Single-unit action potentials were sorted on-line using a hardware window discriminator (Bak Electronics). The remaining 25 VIP and 44 MSTl neurons from Monkey N were recorded with linear electrode arrays that were inserted into either VIP or MSTl daily (Plexon U-probes with two rows of 12 channels spaced 100µm vertically and 50 µm horizontally or Plexon V-probes with 24 channels spaced 50 µm vertically). For array recordings, single-unit action potentials were isolated using Plexon Offline Sorter. For Monkey K, all 130 VIP and 83 MSTl neurons were recorded with linear arrays. There were a total of 104 recording sessions for VIP (monkey N, N=69; monkey K, N=35) and 81 recording sessions for MSTl (monkey N, N=59; monkey K, N=22). In experiments using linear arrays, a mean of 3.8 and 3.7 neurons were recorded simultaneously for areas MSTl and VIP, respectively.

Both VIP and MSTl were initially localized via structural MRI scans as described previously for VIP [54] and MST [55]. We were careful to distinguish MSTl from the dorsal subdivision of area MST (MSTd) and MT [36]. To do this, we carefully mapped the portions of area MT that were found beneath MSTd, in the posterior bank of the superior temporal sulcus. We located the foveal representation of area MT, which is generally located at the anterior-lateral extent of MT. We then carefully mapped regions around that area, and MSTl was localized primarily in regions anterior to the foveal representation of MT.

Experimental protocol

We first performed standard tests to map receptive fields and assess response properties qualitatively. These tests, along with mapping recording sites onto structural MRI images [54], allowed us to confidently assign recording sites to MSTl or VIP. Neurons were isolated while presenting a large field of flickering dots that could be varied in position, size, and velocity. For some neurons, we used a reverse-correlation technique to measure the spatial and directional receptive field structure of each neuron [56]. From these maps, we fit the receptive field with a two-dimensional Gaussian, and used the contour of the Gaussian at half-maximal response to define the receptive field contours shown in Extended Data Fig. 2 (17% of VIP neurons and 13% of MSTl neurons). Due to a technical difficulty, reverse correlation maps were not available for a substantial fraction of neurons. In other recordings, receptive fields were mapped by hand, and receptive field location and size was estimated when the map was clearly noted. We also performed a standard measurement of directional tuning within the fronto-parallel plane by presenting 8 directions of motion, 45 deg apart. These preliminary tests typically required 150-200 trials of fixation behavior.

We recorded from all neurons regardless of their direction and speed preferences. To facilitate population decoding, we used exactly the same stimulus set for all recorded neurons. This allowed us to construct pseudo-population responses for decoding, although these pseudo-population responses do not contain accurate correlated noise since the vast majority of neurons were recorded separately. After extensively mapping out the receptive field coverage of neurons in areas VIP and MSTl, we focused our recordings on a set of penetrations for which receptive fields were concentrated on the same region of space for both brain areas (Extended Data Fig. 2). We carefully selected the starting location and trajectory of object motion based on the distributions of receptive fields of MSTl and VIP neurons in our selected penetrations, such that object motion was centered on the receptive field locations for the populations of neurons in both VIP and MSTl (Extended Data Fig. 2). The main experimental protocol involved 7 directions of object motion, 2 reference frame conditions (head or world), 2 self-motion directions, and 10 stimulus repetitions for each of the Object+Visual and Object+Combined conditions (560 trials), as well as 7 directions and 20 stimulus repetitions for the Object Only condition (140 trials), for a total of 700 trials.

Neural data analyses

Neural responses were computed as firing rates over a time window from 500-2500 ms following stimulus onset. Since the stimulus duration was 2000 ms, this window included most of the stimulus period during which neurons were active, as well as the 500ms delay period after stimulus offset. This analysis window was based on inspection of population response profiles (Extended Data Fig. 4a). The initial 500ms of the stimulus period was not included in our main analysis window because there is an early response to luminance onset of dots during this time (see Extended Data Fig. 4a) and because object and background motion is small during the first 500ms (due to the Gaussian velocity profile used). Our main analyses were also conducted as a function of time, using a moving window of 300 ms that was slid across the data in steps of 50 ms. All analyses are performed on all trials, including both correct and incorrect trials, unless indicated otherwise.

Modulation Index for the effect of reference frame on neural responses: To quantify how neural responses are modulated by the task reference frame, we computed a modulation index (MI) as follows:

$$MI = \frac{1}{N}\left(\frac{|\sum_{\theta}(R(\theta)_{W,L}-R(\theta)_{H,L})|+|\sum_{\theta}(R(\theta)_{W,R}-R(\theta)_{H,R})|}{|\sum_{\theta}(R(\theta)_{W,L}+R(\theta)_{H,L})|+|\sum_{\theta}(R(\theta)_{W,R}+R(\theta)_{H,R})|}\right) \qquad (1)$$

In this formulation, R_W and $R_{H,}$ denote the mean responses of a neuron in the world and head-coordinate tasks, respectively, whereas additional subscripts L and R denote leftward and rightward self-motion directions. θ represents object direction, and N denotes the number of object directions. MI ranges from 0 (no difference between responses in the two reference frames) to 1 (if, for example, responses to one reference frame condition are completely suppressed).

In formulating MI, we sought a simple metric to quantify response modulations related to the task reference frame across all object and self-motion directions. If neural responses were identical in the world and head coordinate tasks, MI would be zero; however, in practice, MI values are unlikely to be very close to zero due to response variability. If world and head coordinate tasks produce different average neural responses, then MI values will become substantially greater than zero. Note that MI is not sensitive to the nature of response modulations (e.g., peak shifts vs. gain modulations vs. tuning shape changes). Given that our direction discrimination task covered a relatively narrow range of directions relative to the full tuning curves, it is difficult to examine the exact nature of tuning changes from our data.

Direction discrimination index: To quantify the strength of tuning for object direction, we used a direction discrimination index (DDI) that was defined as follows:

$$DDI = \frac{R_{\max} - R_{\min}}{R_{\max} - R_{\min} + 2\sqrt{SSE/(N-M)}} \qquad (2)$$

where *R_{max} and R_{min}* represent the maximum and minimum responses from the measured direction tuning function, respectively. *SSE* is the sum squared error around the mean responses, *N* is the total number of observations (trials), and *M* is the number of tested object directions (*M* = 7). DDI is a signal-to-noise metric, conceptually similar to d', that is normalized to range from 0 to 1. Neurons with stronger response modulations relative to their variability will take on values closer to 1.

Effect of partial cube frame on responses: To quantify effects of the partial cube frame on neural responses, we computed a cube effect index (CEI). This index measures neural responses over the initial 500ms of each trial, when the partial cube is visible while background dots are largely invisible (Fig. 2c). For each object direction, θ, CEI takes the absolute difference in response to the cube frame between rightward and leftward self-motion directions, and divides by the sum of those responses. The resultant is then averaged across the *N* object directions. For the world coordinate task, the calculation of CEI is as follows:

$$CEI_W = \frac{1}{N}\left(\frac{|\sum_{\theta}(R(\theta)_{W,R}-R(\theta)_{W,L})|}{|\sum_{\theta}(R(\theta)_{W,R}+R(\theta)_{W,L})|}\right)_{t\in[0-500\ ms]} \qquad (3)$$

where R_W denotes the mean responses of a neuron in the world task, and subscripts *L* and *R* denote leftward and rightward self-motion directions. The calculation of CEI for the head coordinate task is identical, with replacing R_W by R_H.

Metrics of choice-related and task-related activity: To quantify choice-related activity in single neurons, we computed the well-established choice probability (CP) metric [34]. For each unique object direction, self-motion direction, self-motion modality (visual, combined), and task reference frame condition (world vs. head), the distribution of responses was z-scored and then divided into two groups based on whether the animal made a leftward or rightward saccade. Z-scored responses were then pooled across unique stimulus/task conditions as long as there were at least 3 choices made in each direction. ROC analysis was then applied to the pooled z-scores for the two choice groups, and CP was defined as the area under the ROC curve. For our purposes, CP was not referenced to each neuron's preferred direction; rather CP > 0.5 corresponds to a preference for rightward choices and CP < 0.5 corresponds to a preference for leftward choices. This avoids potential issues with defining the "preferred" stimulus when choice effects are large [33].

We devised an analogous ROC-based metric to quantify single-unit activity related to task reference frame. This 'task probability' (TP) metric is computed just like CP, but swapping the roles of variables that represent choice (left vs. right) and task (head vs. world). For each distinct combination of object direction, self-motion direction, self-motion modality, and choice, responses were z-scored and sorted into two groups based on task reference frame. If there were at least 3 trials for world and head reference frames, normalized responses from that condition were pooled with other conditions that met the same criteria. ROC analysis was applied to the pooled z-scores that were sorted into world and head task groups. TP > 0.5 corresponds to greater responses in the world coordinate task, and TP < 0.5 corresponds to greater responses in the head coordinate task.

Removal of choice- and task-related response modulations: To test whether choice- or task-related signals make specific contributions to decoder performance, we devised a method to remove either choice- or task-related response modulations from neural activity. First, we identified a set of trials corresponding to each unique combination of object direction, self-motion direction, and self-motion modality. If this set of trials included at least 3 trials each for left and right choices and 3 trials each for world and head task conditions, then we proceeded to remove either the choice- or task-related response component. To remove the choice-related response component, we shifted the mean responses for right and left choices toward each other to equate the mean responses. Comparison of Extended Data Fig. 9a,b to Extended Data Fig. 8a,b indicates that this manipulation eliminated most of the choice-related modulations, while preserving task-related modulations. Similarly, to remove the task-related component, we shifted the mean responses for world and head task conditions to equate the means. Extended Data Fig. 9c,d indicates that this manipulation was successful in eliminating most of the task-related modulations, while preserving choice effects. These manipulations cannot completely remove all choice- or task-related activity because they can only be performed when there are at least a few trials in each choice and task group, and estimates of mean responses based on a few trials are noisy.

Population decoding by a linear classifier: Linear decoding was performed to classify object direction as rightward or leftward of vertical in each reference frame. Pseudo-population responses of 223 neurons for VIP and 177 neurons for MSTl were used for this purpose. We used a linear classifier to categorize object motion as rightward or leftward relative to vertical in either world or head coordinates:

$$f = \sum_{i=1}^{N} w_i \cdot r_i + \mathrm{k} \qquad (4)$$

Here, N is the number of neurons in the pseudo-population for either VIP or MTl, r_i is the response of the i^{th} neuron, w_i is the decoding weight for the i^{th} neuron, and k is constant scalar. The decoder's choice is determined by the sign of the output variable, f. We used a Fisher linear discriminant (FLD) to compute the parameters (w, k) as follows:

$$w = \Sigma^{-1} \cdot (\mu_R - \mu_L) \qquad (5)$$

$$k = \frac{1}{2} \cdot [(\mu_L^T \cdot \Sigma^{-1} \cdot \mu_L) - (\mu_R^T \cdot \Sigma^{-1} \cdot \mu_R)] \qquad (6)$$

where μ_L and μ_R indicate the mean population response vectors for rightward and leftward object directions relative to the world (for the world task decoder) or head (for the head task decoder), and Σ is the response covariance matrix.

Since most of our neurons were not recorded simultaneously, all neurons did not see the same number of repetitions of each unique stimulus. Thus, we constructed a population response matrix in which each neuron had responses corresponding to 10 stimulus repetitions. For neurons recorded for >10 repetitions (199/223 for VIP, 152/177 for MSTl), we randomly removed some repetitions; for neurons recorded for <10 repetitions (23/223 for VIP, 24/177 for MSTl), we filled in data by sampling with replacement. Once this was done, we computed the covariance matrix using the 'cov()' function in Matlab, as though all neurons had been recorded simultaneously. Since most pairs of neurons were not recorded simultaneously (simultaneous pairs: 226/24753 for VIP, 187/15576 for MSTl), the off-diagonal elements of the resulting covariance matrix do not reflect correlated noise for the vast majority of neuron pairs. However, the off-diagonal elements are generally non-zero since they reflect covariance that is driven by stimulus variations, and which is also dependent on the similarity of tuning properties of a pair. Separate covariance matrices were computed for leftward and rightward object direction classes and were averaged to get the covariance matrix used in Eqn. 5, $\Sigma = \frac{1}{2} \cdot (\Sigma_R + \Sigma_L)$. However, results were very similar if a single covariance matrix was used for both object direction classes. We also compared our results to performance of a standard decoder based on logistic regression [32, 57], which was trained on data and does not require explicit computation of a covariance matrix. Cross-validated output of the logistic regression decoder produced nearly identical results (Extended Data Fig. 5).

We took multiple approaches to decoding object direction from VIP and MSTl responses. 1) *Separate decoders for each task reference frame.* In this approach, we trained separate decoders to classify object direction in world or head coordinates for each brain area. This approach assumes that the animal could have learned to read out VIP or MSTl activity in different ways for each task reference frame. For each task condition (head vs. world), FLD parameters (w_{world}, k_{world} *and* w_{head}, k_{head}) were computed separately from neural responses recorded in the corresponding task condition. Otherwise, each decoder was trained to report object direction across all stimulus conditions, including both self-motion directions and both Object+Visual and Object+Combined conditions. For each decoder, we randomly sampled 20 trials (with replacement) from each neuron. 80% of these trials were used for computing the classifier parameters as described above, and the remaining 20% were used for computing classifier performance (fivefold cross-validation approach). This was repeated 1000 times and overall performance was found by averaging the results. For sliding window analyses, this resampling approach was repeated 100 times for each time bin.

2) *Common decoder for both reference frame conditions.* We also investigated whether a single decoder with one set of common weights could correctly classify object direction in both head and world coordinates. This decoder examines the hypothesis that VIP or MSTl responses are modulated by self-motion signals in a task-dependent manner that allows for the same readout weights to be used for computing object direction in either head or world coordinates. For this analysis, FLD parameters were computed from neural responses that were recorded in both the head- and world-coordinate task conditions, as well as across both self-motion directions and both Object+Visual and Object+Combined conditions. All other aspects of the computation (e.g., cross-validation) were as described above for the separate decoders.

3) *Cross-task decoders.* Because neural responses were obtained for identical conditions of object and background-dot motion (in screen coordinates) under both task reference frames, we could test how well a decoder trained to perform the task in a particular reference frame would perform when supplied with neural responses from the other task reference frame condition. Specifically, for the cross-task decoders, we trained a decoder to perform the world coordinate task based on neural responses from the head coordinate task, and we trained a decoder to perform the head coordinate task using responses from the world task. All other aspects of the decoding procedure were as described above. This approach allowed us to test how well the neural representations in VIP or MSTl could generalize across tasks.

Statistics and reproducibility

In cases where the data met assumptions of normality, as assessed by Lilliefors test, parametric statistical tests were used, including t-tests, paired t-tests, and Pearson correlations. When data were not normally distributed, we used non-parametric tests, including the Wilcoxon rank sum test and the Wilcoxon signed rank test (for paired data).

No statistical methods were used to pre-determine sample sizes but our sample sizes are comparable to, if not greater than, those reported in previous publications of a similar nature [37, 52, 53]. Data collection and analysis were not performed blind to the conditions of the experiments. However, all experimental conditions followed a standard protocol for each recording site and were entirely under computer control. Within each recording session, all stimulus conditions were block-randomized, such that the distinct stimuli were presented in a random order for each repetition. No animals were excluded from the analysis. Neurons were selected for analysis only based on their receptive field location (as described above), and if they could be recorded for at least 3 stimulus repetitions in the main experiment.

Code availability. Custom analysis code was written using MATLAB (v. 2018a). Matlab scripts employed are available from the corresponding author upon reasonable request.

Data availability: The data that support the findings of this study are available from the corresponding author upon reasonable request.

51. Gu, Y., Watkins, P.V., Angelaki, D.E. & DeAngelis, G.C. Visual and nonvisual contributions to three-dimensional heading selectivity in the medial superior temporal area. *J Neurosci* **26**, 73-85 (2006).

52. Fetsch, C.R., Pouget, A., DeAngelis, G.C. & Angelaki, D.E. Neural correlates of reliability-based cue weighting during multisensory integration. *Nat Neurosci* **15**, 146-154 (2012).

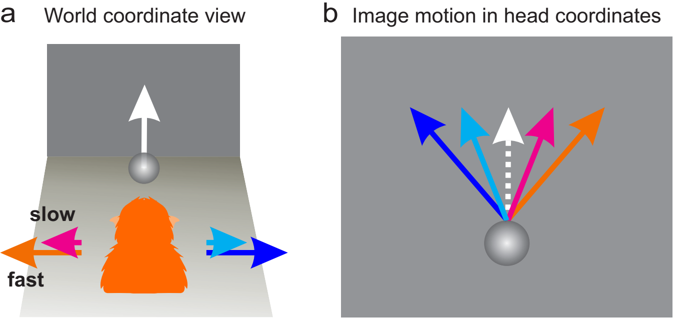
53. Gu, Y., Angelaki, D.E. & DeAngelis, G.C. Neural correlates of multisensory cue integration in macaque MSTd. *Nat Neurosci* **11**, 1201-1210 (2008).

54. Chen, A., DeAngelis, G.C. & Angelaki, D.E. Representation of vestibular and visual cues to self-motion in ventral intraparietal cortex. *J Neurosci* **31**, 12036-12052 (2011).

55. Chen, A., DeAngelis, G.C. & Angelaki, D.E. Macaque parieto-insular vestibular cortex: responses to self-motion and optic flow. *J Neurosci* **30**, 3022-3042 (2010).

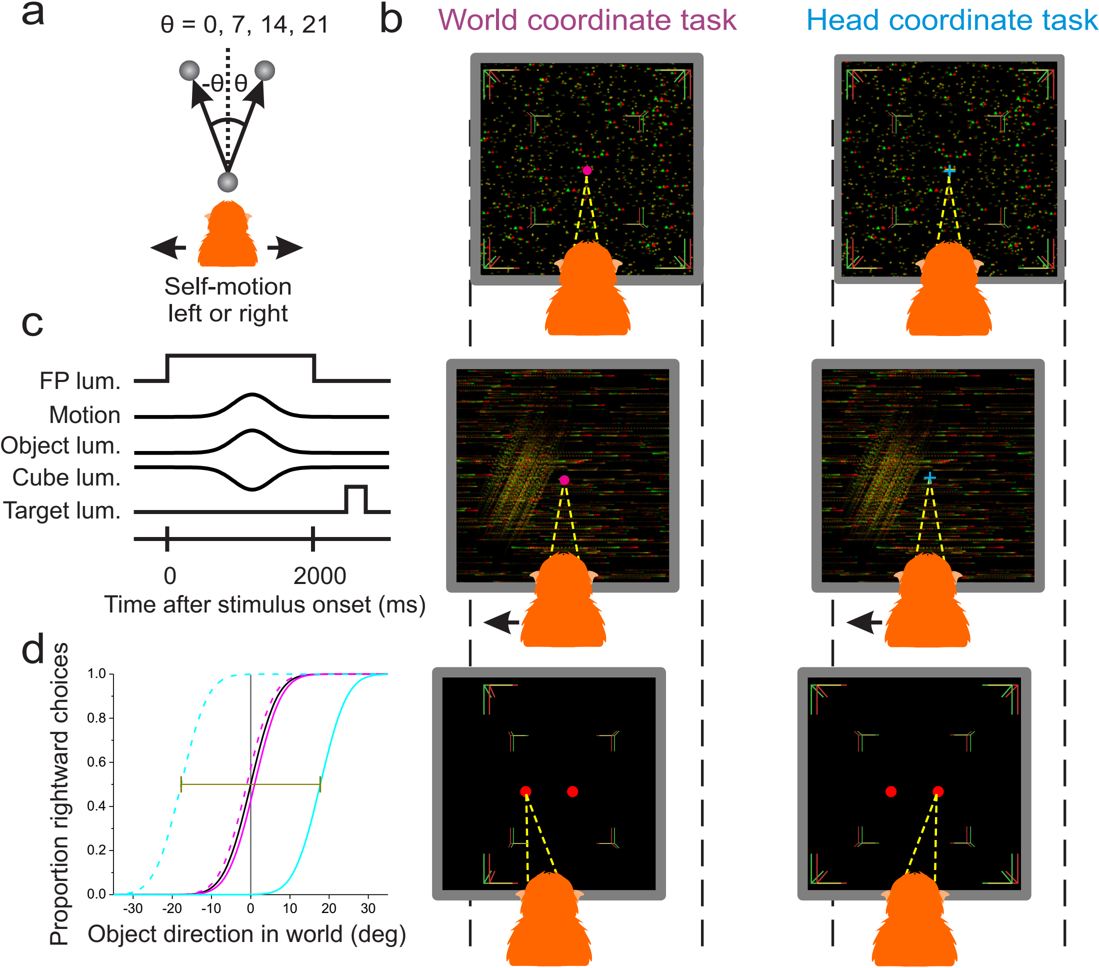
56. Chen, A., Gu, Y., Takahashi, K., Angelaki, D.E. & DeAngelis, G.C. Clustering of self-motion selectivity and visual response properties in macaque area MSTd. *J Neurophysiol* **100**, 2669-2683 (2008).

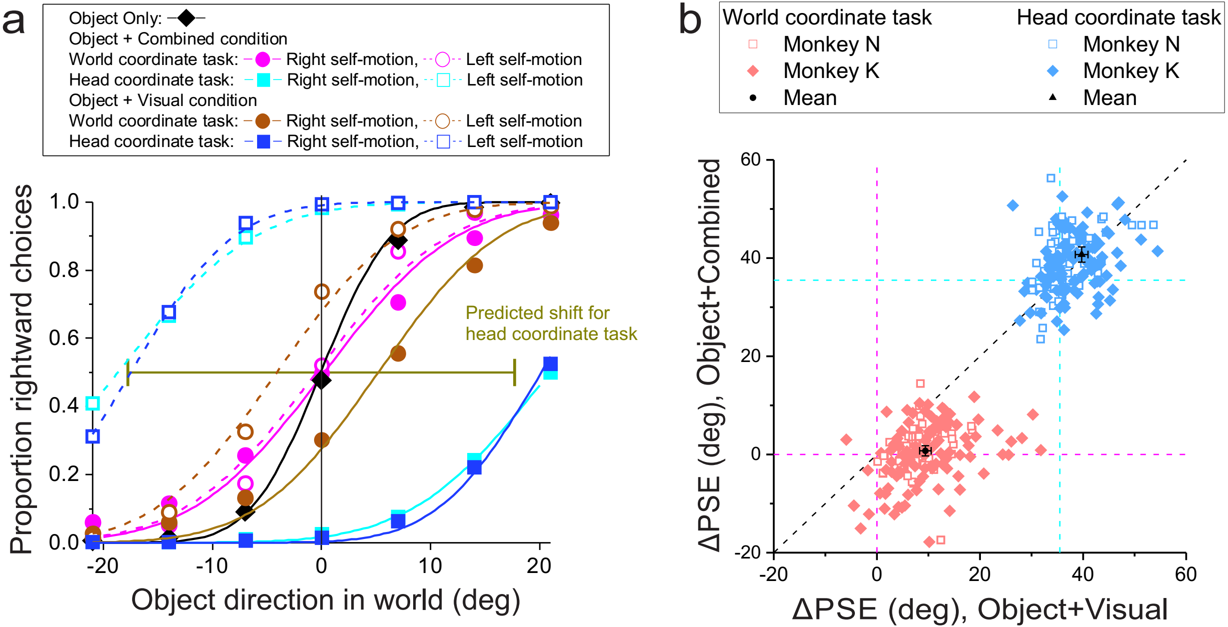
57. Bishop, C.M. *Pattern Recognition and Machine Learning* (Springer, New York, 2006).

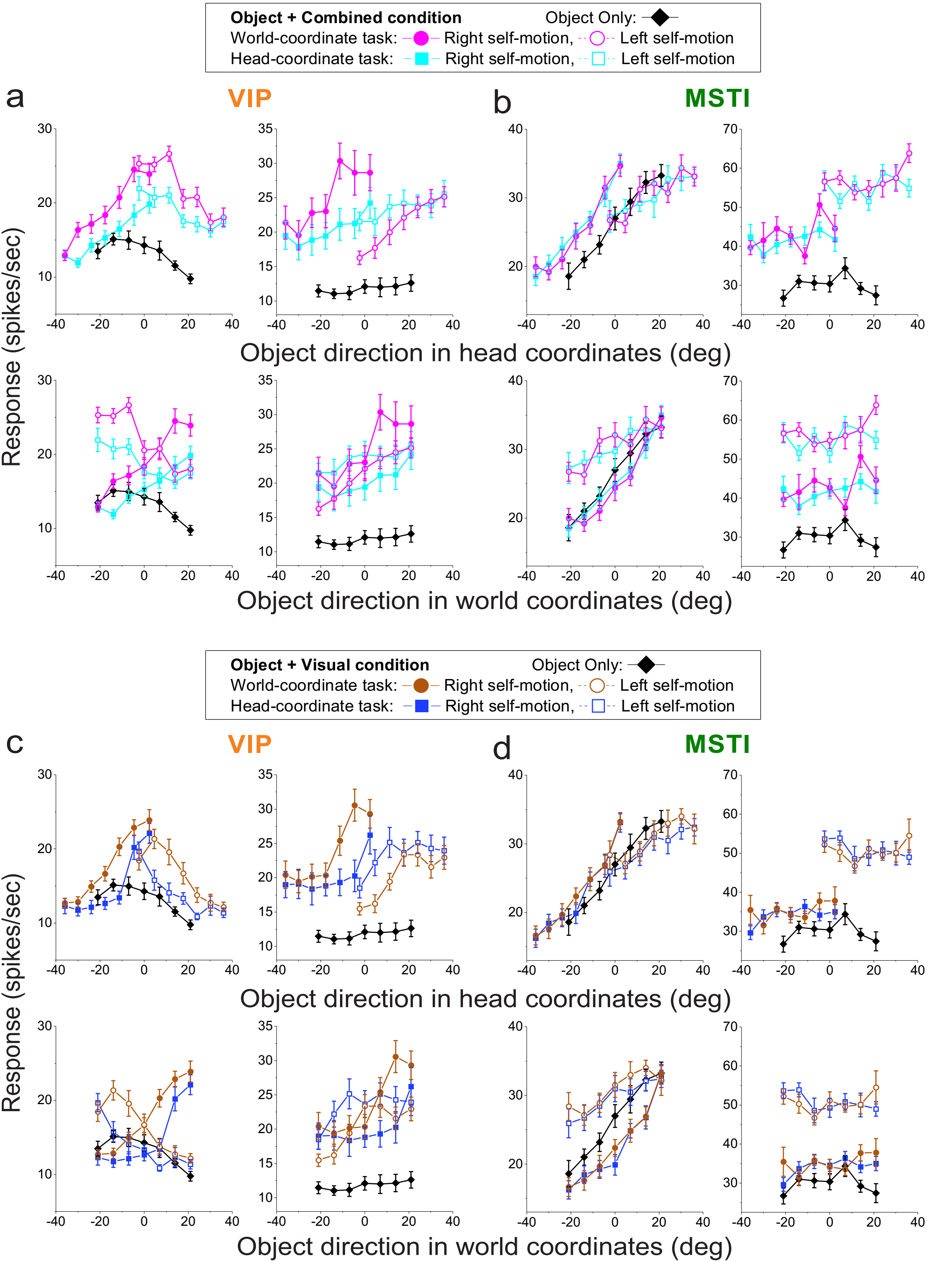


a World coordinate view

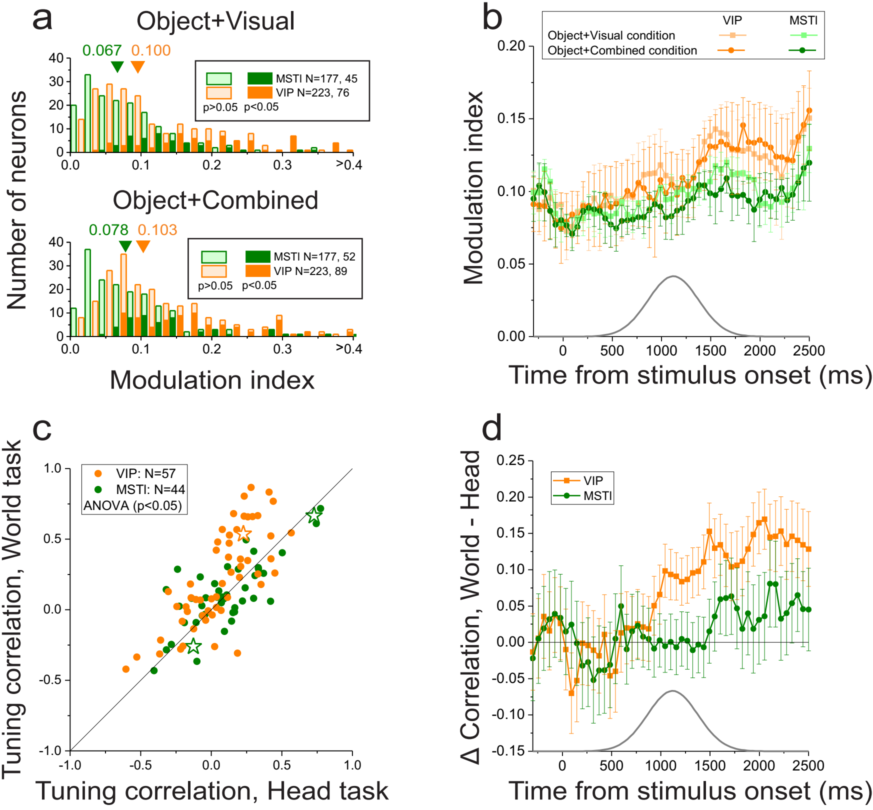
b Image motion in head coordinates

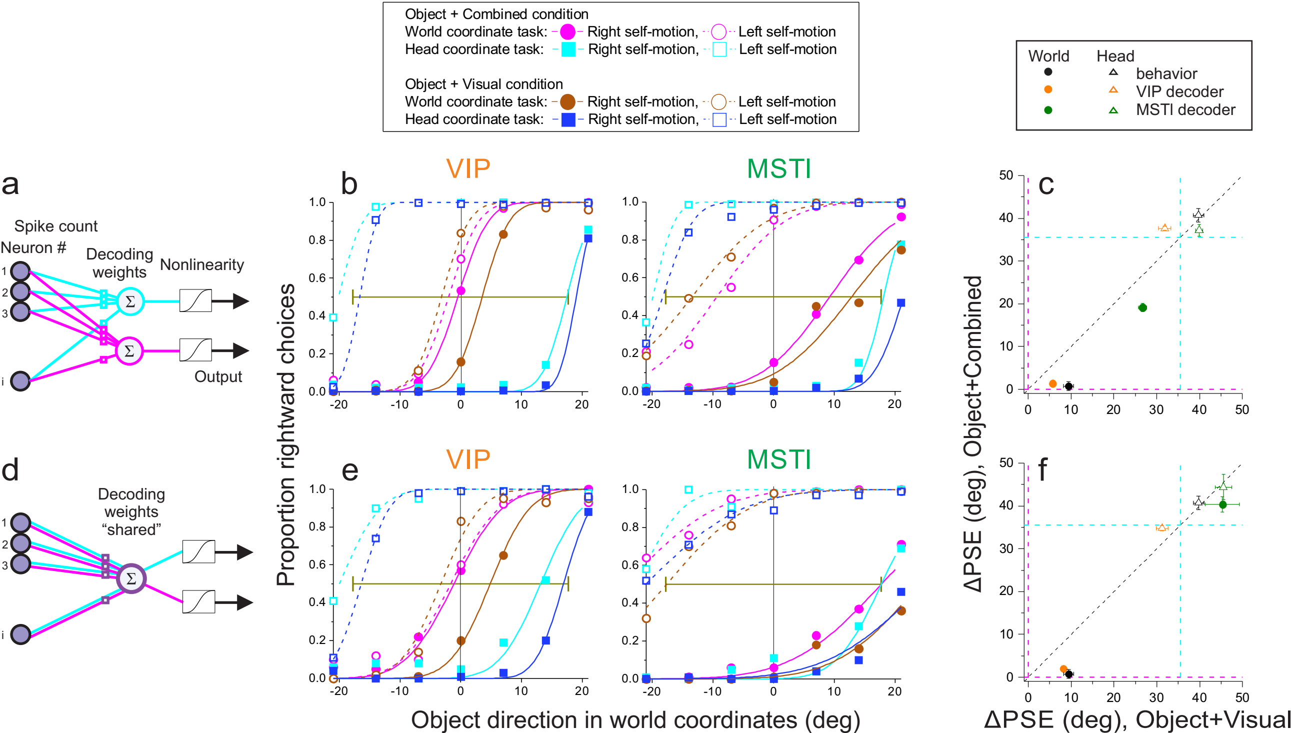

a
θ = 0, 7, 14, 21
-θ
θ
Self-motion
left or right
b
World coordinate task
Head coordinate task
c
FP lum.
Motion
Object lum.
Cube lum.
Target lum.
0
2000
Time after stimulus onset (ms)
d
Proportion rightward choices
1.0
0.8
0.6
0.4
0.2
0.0
-30
-20
-10
0
10
20
30
Object direction in world (deg)

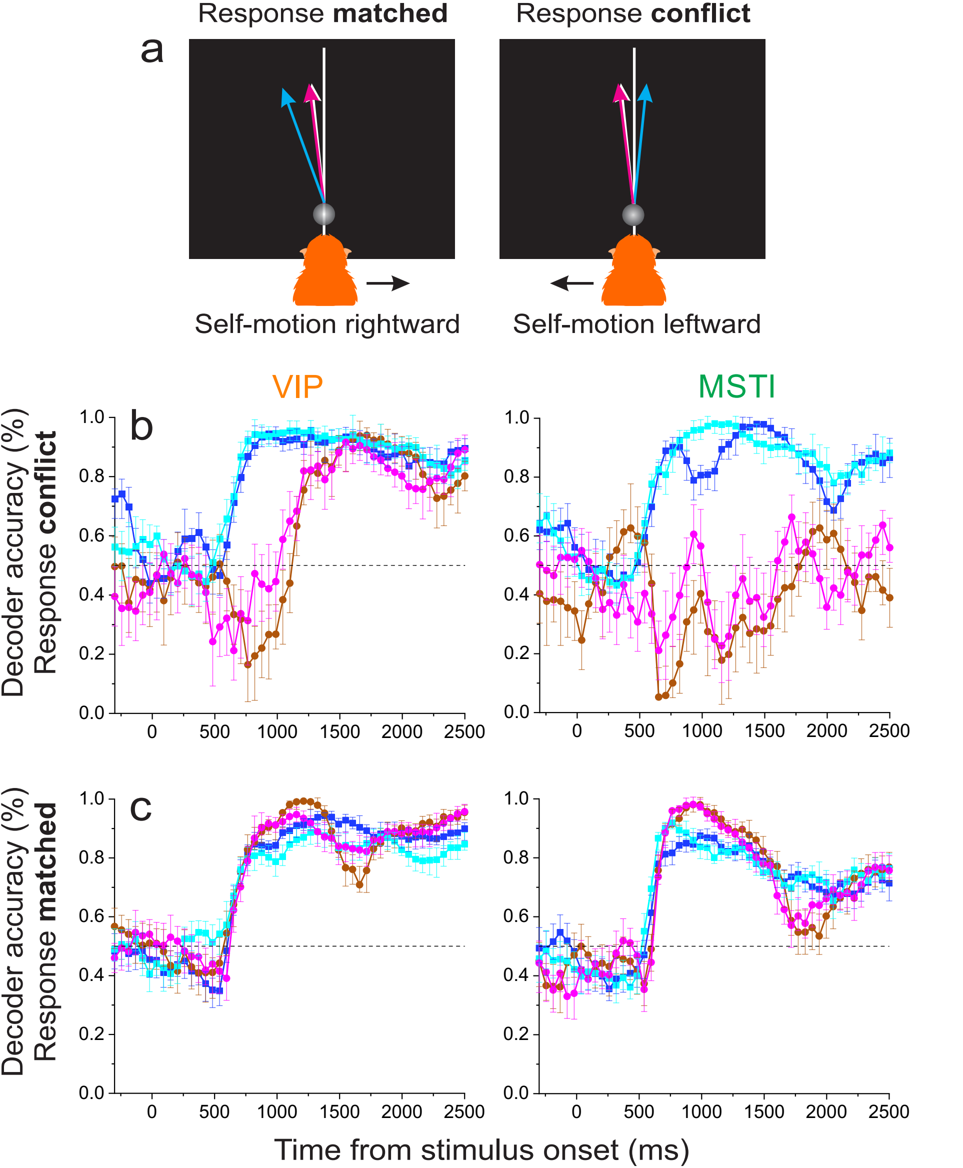

a
Object Only: ◆
Object + Combined condition
World coordinate task: ● Right self-motion, ○ Left self-motion
Head coordinate task: ■ Right self-motion, □ Left self-motion
Object + Visual condition
World coordinate task: ● Right self-motion, ○ Left self-motion
Head coordinate task: ■ Right self-motion, □ Left self-motion
Predicted shift for
head coordinate task
Proportion rightward choices
0.0
0.2
0.4
0.6
0.8
1.0
-20
-10
0
10
20
Object direction in world (deg)
b
World coordinate task
□ Monkey N
◆ Monkey K
● Mean
Head coordinate task
□ Monkey N
◆ Monkey K
▲ Mean
ΔPSE (deg), Object+Combined
ΔPSE (deg), Object+Visual
-20
0
20
40
60

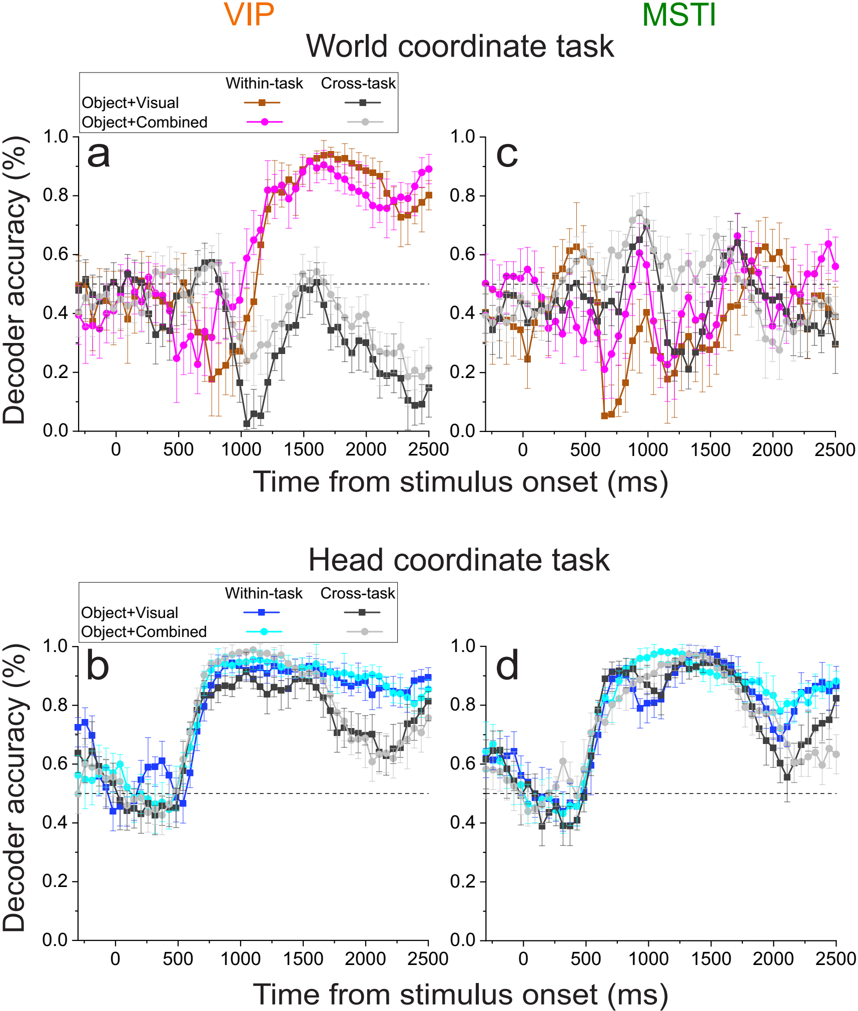


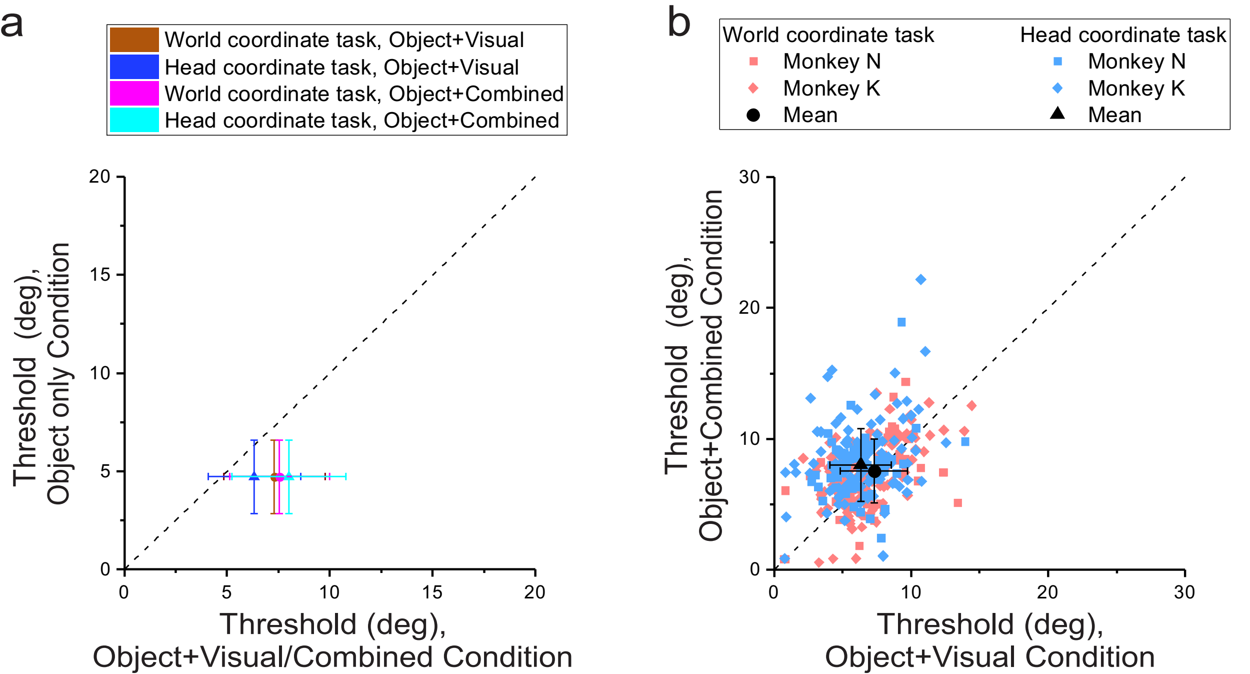
Object + Combined condition
Object Only:
World-coordinate task: Right self-motion, Left self-motion
Head-coordinate task: Right self-motion, Left self-motion
a
VIP
b
MSTl
Response (spikes/sec)
Object direction in head coordinates (deg)
Object direction in world coordinates (deg)
Object + Visual condition
Object Only:
World-coordinate task: Right self-motion, Left self-motion
Head-coordinate task: Right self-motion, Left self-motion
c
VIP
d
MSTl
Response (spikes/sec)
Object direction in head coordinates (deg)
Object direction in world coordinates (deg)
10
20
30
35
15
25
40
50
60
70
-40
-20
0
20
40

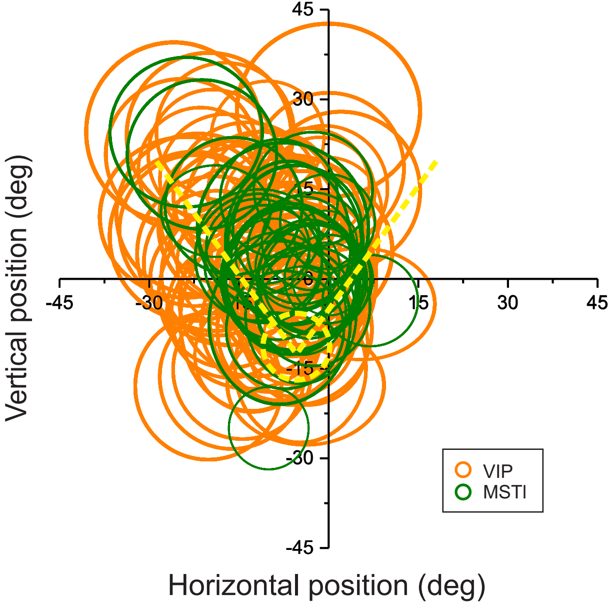

a
Object+Visual
0.067
0.100
MSTI N=177, 45
VIP N=223, 76
p>0.05 p<0.05
Object+Combined
0.078
0.103
MSTI N=177, 52
VIP N=223, 89
p>0.05 p<0.05
Number of neurons
0 10 20 30 40
0.0 0.1 0.2 0.3 >0.4
Modulation index
b
VIP MSTI
Object+Visual condition
Object+Combined condition
Modulation index
0.00 0.05 0.10 0.15 0.20
0 500 1000 1500 2000 2500
Time from stimulus onset (ms)
c
VIP: N=57
MSTI: N=44
ANOVA (p<0.05)
Tuning correlation, World task
Tuning correlation, Head task
-1.0 -0.5 0.0 0.5 1.0
d
VIP
MSTI
Δ Correlation, World - Head
-0.15 -0.10 -0.05 0.00 0.05 0.10 0.15 0.20 0.25
0 500 1000 1500 2000 2500
Time from stimulus onset (ms)

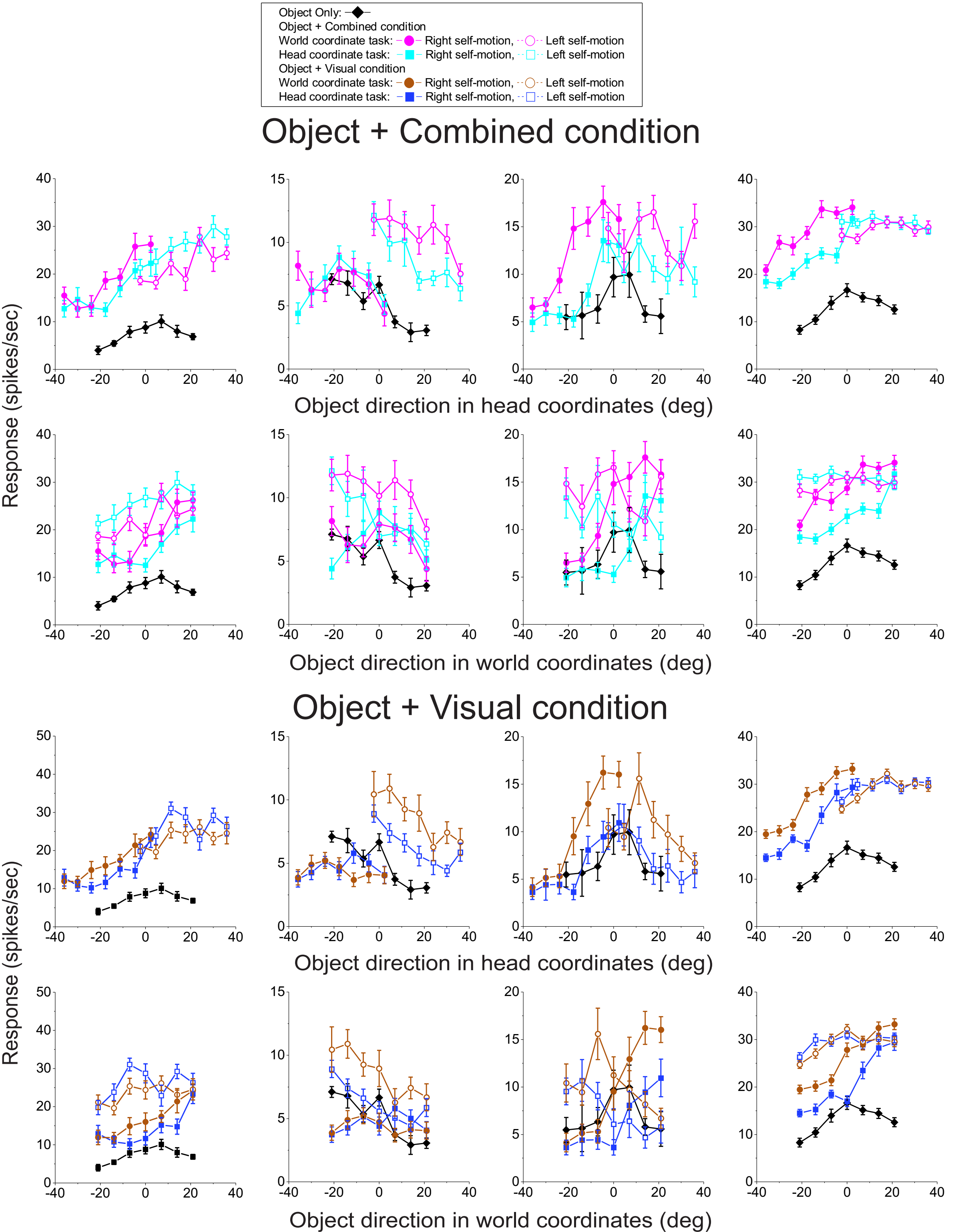

Object + Combined condition
World coordinate task: Right self-motion, Left self-motion
Head coordinate task: Right self-motion, Left self-motion
Object + Visual condition
World coordinate task: Right self-motion, Left self-motion
Head coordinate task: Right self-motion, Left self-motion
World
Head
behavior
VIP decoder
MSTl decoder
a
Spike count
Neuron #
1
2
3
i
Decoding weights
Nonlinearity
Σ
Σ
Output
b
VIP
MSTl
1.0
0.8
0.6
0.4
0.2
0.0
-20
-10
0
10
20
c
50
40
30
20
10
0
0
10
20
30
40
50
d
Decoding weights "shared"
1
2
3
i
Σ
e
VIP
MSTl
f
Proportion rightward choices
Object direction in world coordinates (deg)
ΔPSE (deg), Object+Combined
ΔPSE (deg), Object+Visual

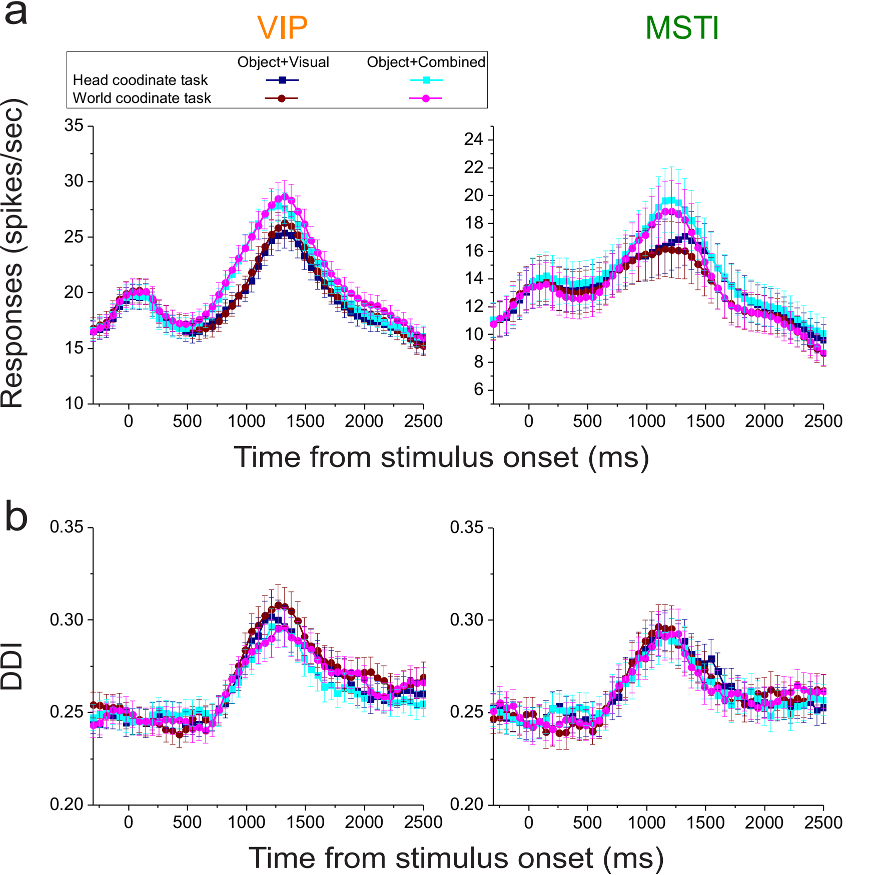

Response matched
Response conflict
a
Self-motion rightward
Self-motion leftward
VIP
MSTl
b
c
Decoder accuracy (%)
Response conflict
Response matched
1.0
0.8
0.6
0.4
0.2
0.0
0
500
1000
1500
2000
2500
Time from stimulus onset (ms)

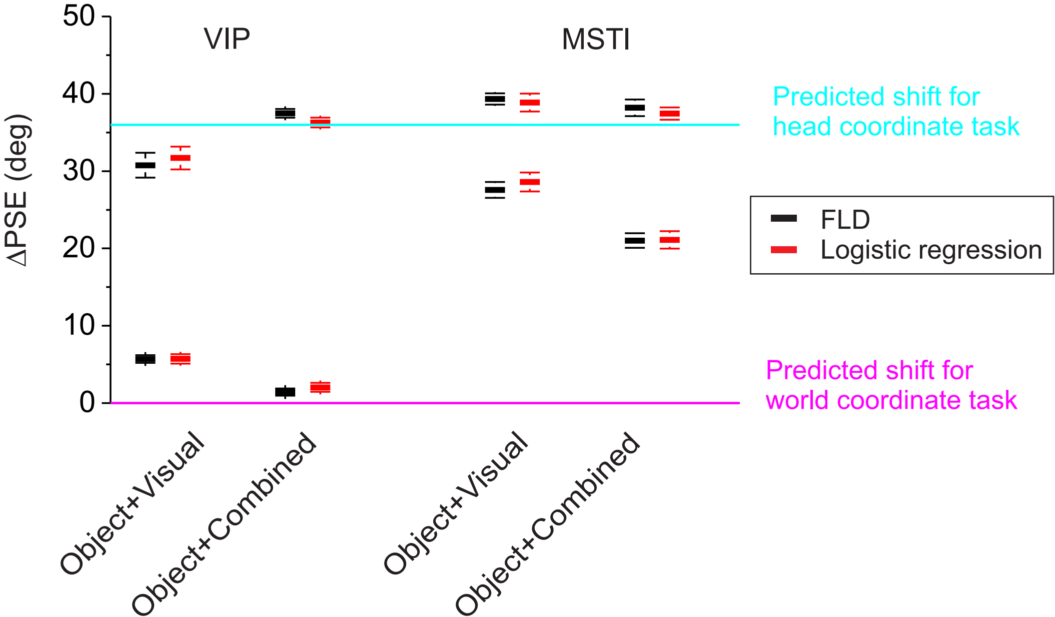

VIP
MSTl
World coordinate task
Within-task
Cross-task
Object+Visual
Object+Combined
a
c
Decoder accuracy (%)
1.0
0.8
0.6
0.4
0.2
0.0
0
500
1000
1500
2000
2500
Time from stimulus onset (ms)
Head coordinate task
Within-task
Cross-task
Object+Visual
Object+Combined
b
d
Decoder accuracy (%)
Time from stimulus onset (ms)

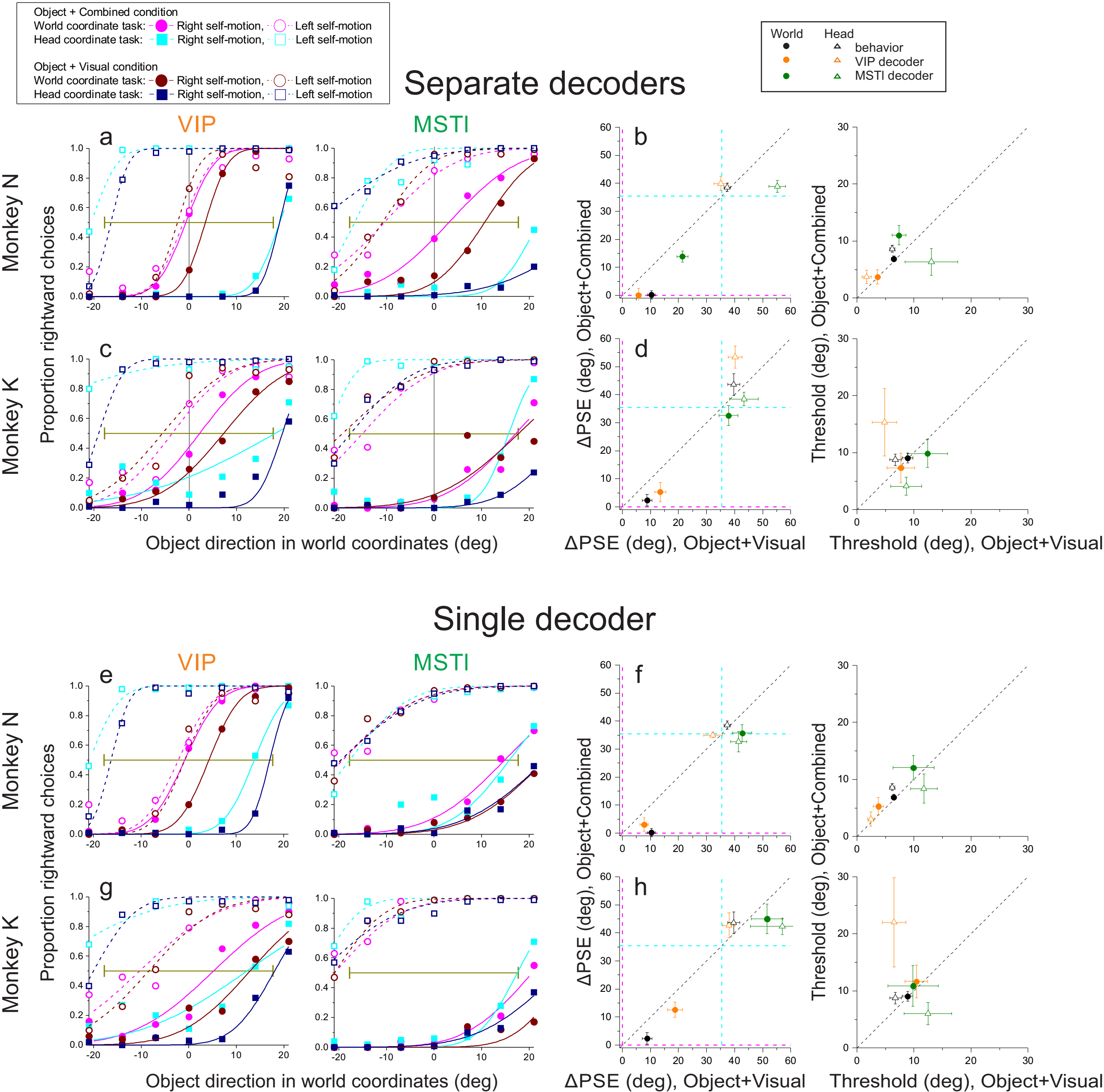

a
World coordinate task, Object+Visual
Head coordinate task, Object+Visual
World coordinate task, Object+Combined
Head coordinate task, Object+Combined
0
5
10
15
20
0
5
10
15
20
Threshold (deg), Object only Condition
Threshold (deg), Object+Visual/Combined Condition
b
World coordinate task
Monkey N
Monkey K
Mean
Head coordinate task
Monkey N
Monkey K
Mean
0
10
20
30
0
10
20
30
Threshold (deg), Object+Combined Condition
Threshold (deg), Object+Visual Condition

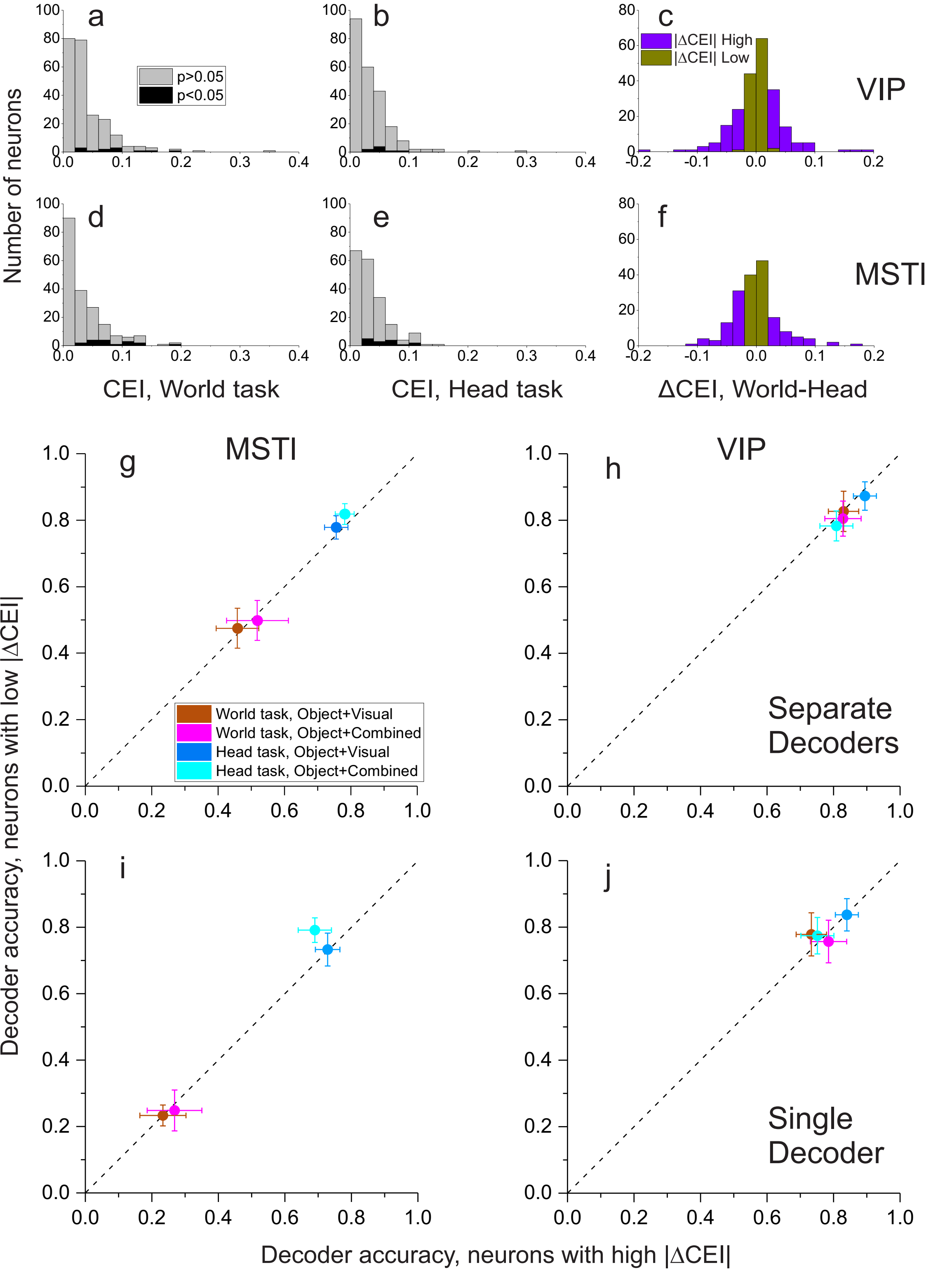

45
30
15
0
-15
-30
-45
-45
-30
-15
15
30
45
Vertical position (deg)
Horizontal position (deg)
VIP
MSTl

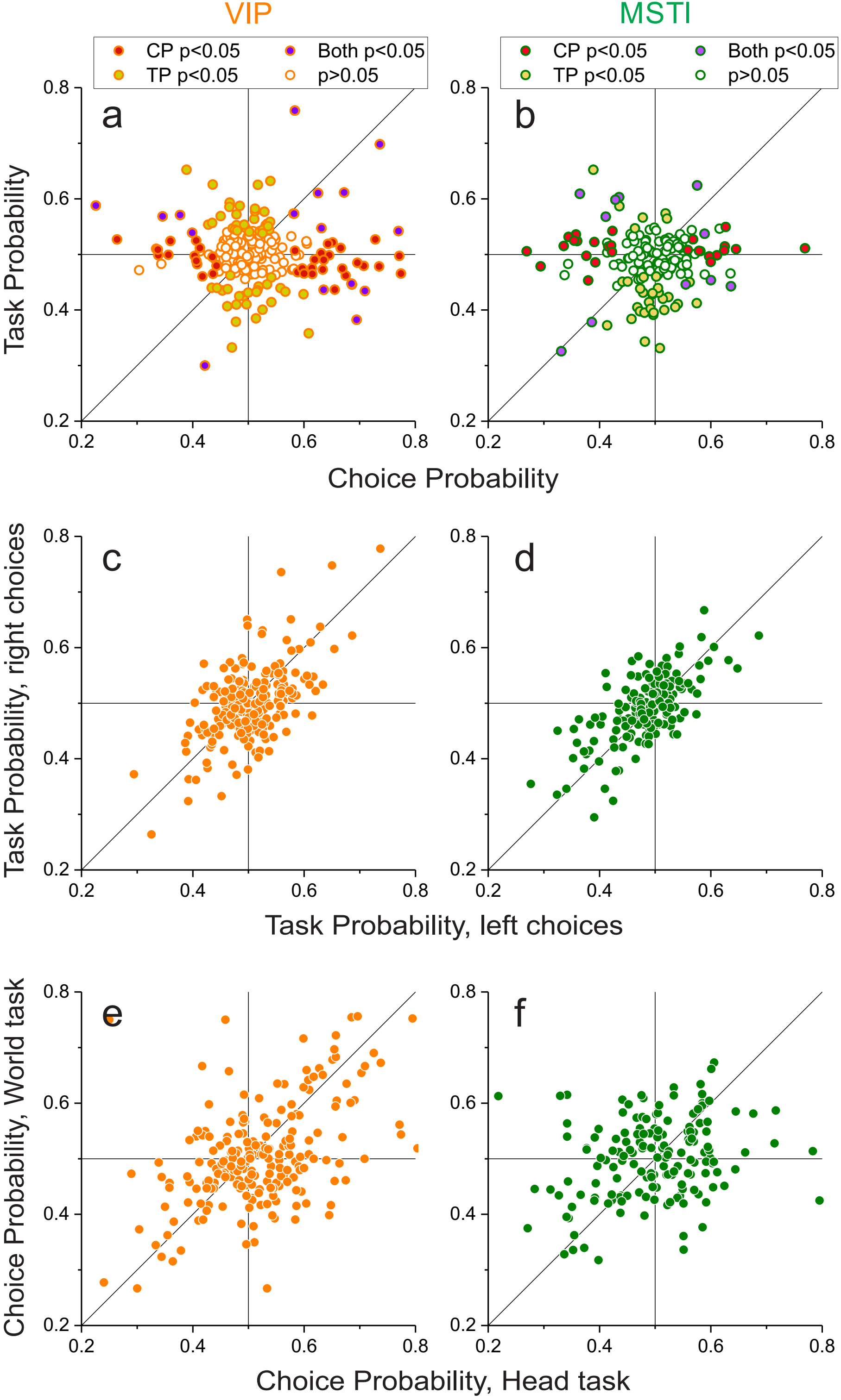

Object Only:
Object + Combined condition
World coordinate task: Right self-motion, Left self-motion
Head coordinate task: Right self-motion, Left self-motion
Object + Visual condition
World coordinate task: Right self-motion, Left self-motion
Head coordinate task: Right self-motion, Left self-motion
Object + Combined condition
Response (spikes/sec)
Object direction in head coordinates (deg)
Object direction in world coordinates (deg)
Object + Visual condition
Response (spikes/sec)
Object direction in head coordinates (deg)
Object direction in world coordinates (deg)
-40
-20
0
20
40
0
5
10
15
20
30
40
50

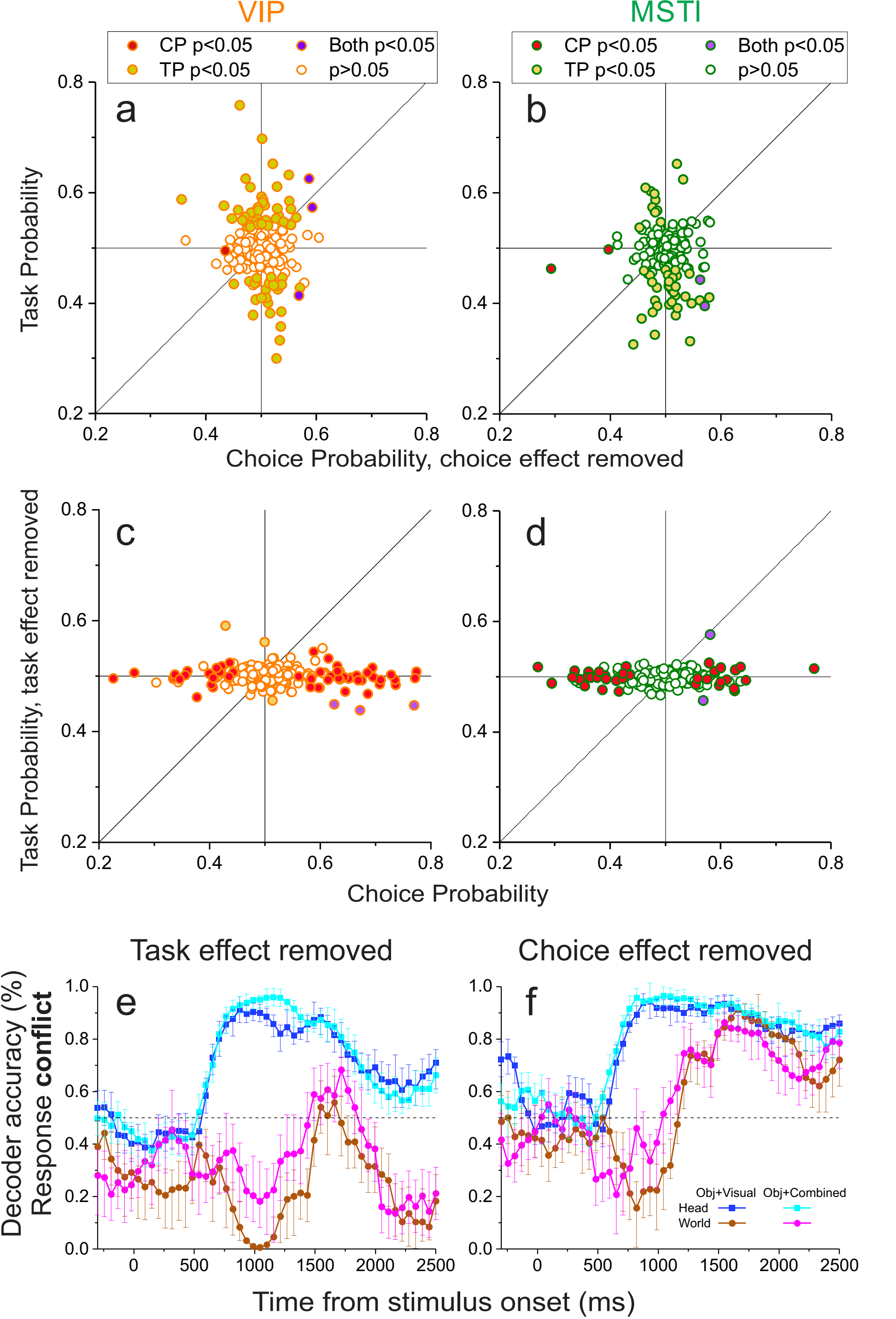

a
VIP
MSTl
Object+Visual
Object+Combined
Head coodinate task
World coodinate task
Responses (spikes/sec)
35
30
25
20
15
10
24
22
20
18
16
14
12
10
8
6
0
500
1000
1500
2000
2500
Time from stimulus onset (ms)
b
DDI
0.35
0.30
0.25
0.20
Time from stimulus onset (ms)

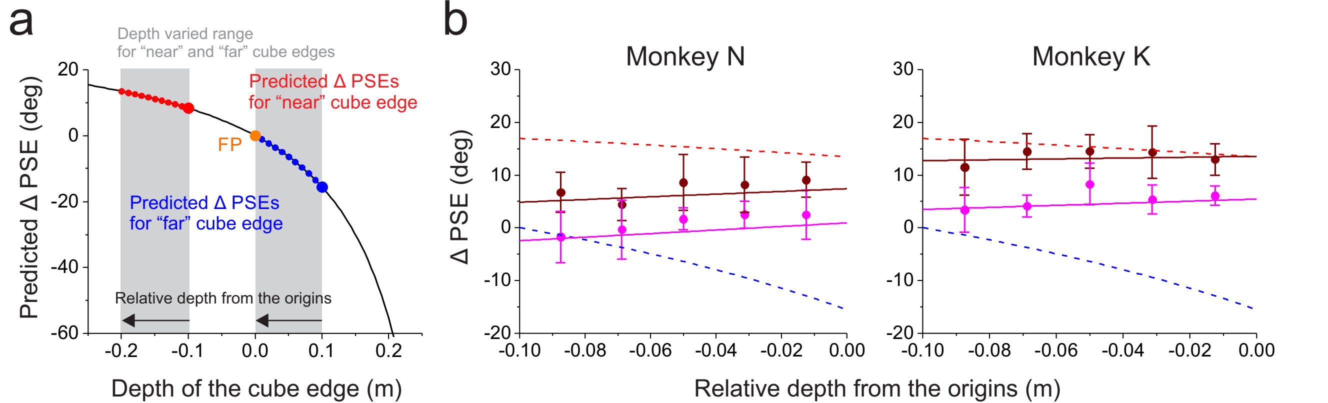

ΔPSE (deg)
50
40
30
20
10
0
VIP
MSTI
Predicted shift for
head coordinate task
Predicted shift for
world coordinate task
FLD
Logistic regression
Object+Visual
Object+Combined
Object+Visual
Object+Combined


Object + Combined condition
World coordinate task: ● Right self-motion, ○ Left self-motion
Head coordinate task: ■ Right self-motion, □ Left self-motion
Object + Visual condition
World coordinate task: ● Right self-motion, ○ Left self-motion
Head coordinate task: ■ Right self-motion, □ Left self-motion
World
Head
behavior
VIP decoder
MSTl decoder
Separate decoders
VIP
MSTl
a
b
c
d
Monkey N
Monkey K
Proportion rightward choices
1.0
0.8
0.6
0.4
0.2
0.0
-20
-10
0
10
20
Object direction in world coordinates (deg)
ΔPSE (deg), Object+Combined
ΔPSE (deg), Object+Visual
0
10
20
30
40
50
60
Threshold (deg), Object+Combined
Threshold (deg), Object+Visual
0
10
20
30
Single decoder
VIP
MSTl
e
f
g
h
Monkey N
Monkey K
Proportion rightward choices
Object direction in world coordinates (deg)
ΔPSE (deg), Object+Combined
ΔPSE (deg), Object+Visual
Threshold (deg), Object+Combined
Threshold (deg), Object+Visual


a
b
c
d
e
f
g
h
i
j
p>0.05
p<0.05
|ΔCEI| High
|ΔCEI| Low
VIP
MSTl
Number of neurons
CEI, World task
CEI, Head task
ΔCEI, World-Head
MSTl
VIP
World task, Object+Visual
World task, Object+Combined
Head task, Object+Visual
Head task, Object+Combined
Separate Decoders
Single Decoder
Decoder accuracy, neurons with low |ΔCEI|
Decoder accuracy, neurons with high |ΔCEI|
0.0
0.1
0.2
0.3
0.4
-0.2
-0.1
0
20
40
60
80
100
0.2
0.4
0.6
0.8
1.0


VIP
MSTl
CP p<0.05
TP p<0.05
Both p<0.05
p>0.05
a
b
c
d
e
f
Task Probability
Choice Probability
Task Probability, right choices
Task Probability, left choices
Choice Probability, World task
Choice Probability, Head task
0.2
0.4
0.6
0.8


VIP
MSTI
CP p<0.05
TP p<0.05
Both p<0.05
p>0.05
a
b
c
d
e
f
Task Probability
Choice Probability, choice effect removed
Task Probability, task effect removed
Choice Probability
0.2
0.4
0.6
0.8
Task effect removed
Choice effect removed
Decoder accuracy (%)
Response **conflict**
Time from stimulus onset (ms)
0.0
0.2
0.4
0.6
0.8
1.0
0
500
1000
1500
2000
2500
Obj+Visual
Obj+Combined
Head
World


a
Depth varied range
for "near" and "far" cube edges
Predicted Δ PSEs
for "near" cube edge
FP
Predicted Δ PSEs
for "far" cube edge
Relative depth from the origins
Predicted Δ PSE (deg)
20
0
-20
-40
-60
-0.2
-0.1
0.0
0.1
0.2
Depth of the cube edge (m)
b
Monkey N
Monkey K
Δ PSE (deg)
30
20
10
0
-10
-20
-0.10
-0.08
-0.06
-0.04
-0.02
0.00
Relative depth from the origins (m)

Inventory of Supporting Information

Manuscript #: NN-A69354B
Corresponding author name(s): Ryo Sasaki

Instructions:

Please complete each of the Inventory Tables below to outline your Extended Data and Supplementary Information items.

There are four sections; *1. Extended Data, 2A. Supplementary Information: Flat Files, 2B. Supplementary Information: Additional Files,* and *3. Source Data.* Each section includes specific instructions. Please complete these tables as fully as possible. We ask that you avoid using spaces in your file names, and instead use underscores, i.e.: *Smith_ED_Fig1.jpg* ***not*** *Smith ED Fig1.jpg*

Please note that titles and descriptive captions will only be lightly edited, so please ensure that you are satisfied with these prior to submission.

If you have any questions about any of the information contained in this inventory, please contact the Editorial Assistant: neurosci@us.nature.com

1. Extended Data

Complete the Inventory below for all Extended Data figures.

- Keep Figure Titles to one sentence only
- File names should include the Figure Number. i.e.: *Smith_ED Fig1.jpg*
- Please be sure to include the file extension in the Filename. Note that Extended Data files must be submitted as .jpg, .tif or .eps files *only*
- All Extended Data figure legends must be provided in the Inventory below and should not exceed 300 words each *(if possible)*
- Please include Extended Data *ONLY* in this table

Figure #	**Figure title** One sentence only	**Filename** This should be the name the file is saved as when it is uploaded to our system. Please include the file extension. i.e.: *Smith_ED_Fig1.jpg*	**Figure Legend** If you are citing a reference for the first time in these legends, please include all new references in the Online Methods References section, and carry on the numbering from the main References section of the paper.
Extended Data Fig. 1	**Summary of psychophysical thresholds (inverse of sensitivity) across task conditions.**	Sasaki_ED_Fig1.eps	**Extended Data Figure 1. Summary of psychophysical thresholds (inverse of sensitivity) across task conditions.** (a) Average threshold for the Object Only condition (no self-motion) is plotted against average thresholds for the Object+Visual and Object+Combined conditions for the world (brown/magenta) and head (blue/cyan) coordinate tasks. Error bars represent 95% confident intervals. Averages taken over n=185 sessions across the two animals. (b) For each session, threshold in the Object+Combined condition is plotted against the corresponding threshold in the Object+Visual condition. Black symbols show mean thresholds

			and error bars represent 95% confidence intervals. Data from 128 sessions for Monkey N and 57 sessions for monkey K.
Extended Data Fig. 2	**Summary of receptive field locations for populations of VIP (orange, N=66) and MSTl (green, N=44) neurons.**	Sasaki_ED_Fig2.eps	**Extended Data Figure 2. Summary of receptive field locations for populations of VIP (orange, N=66) and MSTl (green, N=44) neurons.** Cells are included here if they had significant structure in receptive field maps obtained by reverse correlation (17% of VIP and 13% of MSTl neurons) or if they had clear hand-mapped receptive fields for which good estimates of RF center and size were obtained (13% of VIP neurons and 12% of MSTl neurons). Significant structure in reverse correlation maps was assessed by a two-sided permutation test ($p<0.05$), in which we scrambled the relationship between response amplitude and stimulus location within the RF, as described previously [56]. Ellipses approximate the RF dimensions and were derived either from a two-dimensional Gaussian fit (contour at half-maximal response) to receptive field maps obtained by reverse correlation (VIP: N=38; MSTl: N=23), or from hand mapping (VIP: N=28; MSTl: N=21). Coordinate (0, 0) represents the center of the visual display, where the fixation target was located. Yellow dashed lines represent the starting location of the moving object and the range of directions in head coordinates.
Extended Data Fig. 3	**Data from four additional VIP neurons, illustrating diversity of effects of self-motion on tuning curves.**	Sasaki_ED_Fig3.eps	**Extended Data Figure 3. Data from four additional VIP neurons, illustrating diversity of effects of self-motion on tuning curves.** Top: Object+Combined condition. Bottom: Object+Visual condition. Format as in Fig. 4. Error bars denote SEM (n=10 stimulus repetitions per datum).
Extended Data Fig. 4	**Summary of time courses of average firing rates and directional selectivity.**	Sasaki_ED_Fig4.eps	**Extended Data Figure 4. Summary of time courses of average firing rates and directional selectivity.** (a) Average response across all 223 VIP and 177 MSTl neurons is shown for each stimulus condition for both the head and world coordinate task conditions. For each neuron, responses were taken from the object motion direction that elicited the maximum firing

			rate. Error bars represent SEM. Color coding as in Fig. 7. Results were nearly identical if the responses of neurons were normalized before averaging. (b) Average direction discrimination index (DDI) for populations of VIP (n=223) and MSTl (n=177) neurons (see Methods, Eqn. 2). DDI values were computed separately for leftward and rightward self-motion and then averaged for each neuron. Error bars represent 95% confidence intervals. For this figure, both average responses and DDI values were computed within a 300 ms sliding time window that was advanced across the stimulus epoch in steps of 50 ms.
Extended Data Fig. 5	**Decoder results are robust to the type of classifier used.**	Sasaki_ED_Fig5.eps	**Extended Data Figure 5. Decoder results are robust to the type of classifier used.** Black data points represent results from the FLD classifier used in all main figures. Red data points show results from a logistic regression decoder. For this comparison, the same population responses were used for training and testing each decoder. The results are very robust to the type of decoder used. Error bars represent 95% confidence intervals (across n=1000 bootstraps).
Extended Data Fig. 6	**Comparison of decoder results across animals.**	Sasaki_ED_Fig6.eps	**Extended Data Figure 6. Comparison of decoder results across animals.** (a-d) Results for separate decoders trained to perform the world and head coordinate tasks. Format as in Figure 6. Each row shows results separately for each animal. Pink and cyan dashed lines in panels b and d: expected ΔPSE for perfect performance in the world and head coordinate tasks, respectively. Error bars in panels b and d represent 95% confidence intervals (across n=1000 bootstraps). (e-h) Results for the single decoder, shown separately for each animal. Decoders were trained separately using responses from each animal, yet main results are conserved across subjects. Error bars represent 95% confidence intervals (across n=1000 bootstraps). Format as in panels a-d.
Extended Data Fig. 7	**Effect of partial cube frame on single-unit**	Sasaki_ED_Fig7.eps	**Extended Data Figure 7. Effect of partial cube frame on single-unit responses and population decoding.** (a, d)

	responses and population decoding.		Distributions of the cube effect index (CEI, see Methods) for areas VIP and MSTl, respectively, in the world coordinate task. Black and gray shading denotes neurons with CEI values that are significantly different from zero and non-significant, respectively (two-sided permutation test, p<0.05). (b, e) Distributions of CEI for VIP and MSTl, respectively, in the head coordinate task condition. (c, f) Distributions of the difference in CEI (ΔCEI) between world and head task conditions for VIP and MSTl, respectively. Green and purple shading indicates a median split of the data based on the absolute value, |ΔCEI|. (g, h) Comparison of decoder accuracy (proportion correct) for populations of neurons with above-median |ΔCEI| (abscissa) and below-median |ΔCEI| (ordinate) values, for areas VIP and MSTl, respectively. Error bars represent 95% confidence intervals (across n=1000 bootstraps). Data in these panels come from decoders that were trained separately for the world and head coordinate task conditions. (i, j) Same as panels g and h, except for a single decoder trained to perform the task across both reference frame conditions. Format as in g, h.
Extended Data Fig. 8	**Summary of choice-related and task-related response modulations.**	Sasaki_ED_Fig8.eps	**Extended Data Figure 8. Summary of choice-related and task-related response modulations.** (a) Scatter plot of task probability (TP) and choice probability (CP) values for VIP neurons (N=223). Color of the symbol centers corresponds to significance of TP and CP values as follows: blue center, both TP and CP are significantly different from 0.5 (two-sided permutation test, p<0.05); red center, only CP is significantly different from 0.5; gold center, only TP is significantly different from 0.5; white center, neither TP nor CP is significant. The observation that TP and CP values are largely uncorrelated here is an empirical observation that is not enforced by the analysis. (b) Scatter plot of TP and CP values for MSTl neurons (N=177). Symbol center color conventions as in panel a. (c) Scatter plot of TP values for VIP neurons

			computed separately for right and left choices (N=223). (d) Same as panel c but for MSTl neurons (N=177). (e) Scatter plot comparing CP values from VIP for the world and head coordinate task conditions (N=223). (f) Same as panel e but for MSTl neurons (N=177).
Extended Data Fig. 9	**Effects of selectively removing choice- or task-related response modulations.**	Sasaki_ED_Fig9.eps	**Extended Data Figure 9. Effects of selectively removing choice- or task-related response modulations.** (a) Scatter plot of TP and CP values for VIP (N=223) after selective removal of choice-related response modulations (see Methods for details). Format as in Extended Data Fig. 8a. (b) Same as panel a except for MSTl (N=177). Format as in Extended Data Fig. 8b. (c) Scatter plot of TP and CP values for VIP after selective removal of task-related response modulations. (d) Same as panel c, except for MSTl. (e) Time course of decoder performance based on activity of 223 VIP neurons on response conflict trials, after removal of task-related response modulations. Data are shown for the case of separate decoders for world and head coordinate task conditions. Format as in Fig. 7b. Error bars represent 95% confidence intervals (across n=100 bootstraps). (f) Time course of VIP decoder performance, as in panel e, but after removal of choice-related response modulations.
Extended Data Fig. 10	**Results from behavioral control sessions in which the depth of the partial cube was varied across trials.**	Sasaki_ED_Fig10.eps	**Extended Data Figure 10. Results from behavioral control sessions in which the depth of the partial cube was varied across trials.** (a) Predicted ΔPSE values are shown as a function of the depth of the partial cube. Red and blue data points show predicted ΔPSE values and depths for the near and far edges of the cube. (b) Dashed curves replot the predictions from panel a, where the horizontal axis is now depth relative to the origins for the near (red) and far (blue) cube edges (where the origins are the farthest depths for each edge). Data points represent behavioral ΔPSE values for the two monkeys (n=7 sessions for each animal); magenta and brown data points show results for the Object+Combined and

			Object+Visual conditions. Error bars show 95% confidence intervals, and lines show regression fits. The slopes of the linear fits were not significantly different from zero for either animal or either self-motion condition (two-tailed t-test, p > 0.15 for all four cases).

Delete rows as needed to accommodate the number of figures (10 is the maximum allowed).

2. Supplementary Information:

A. Flat Files

Complete the Inventory below for all additional textual information and any additional Supplementary Figures, which should be supplied in one combined PDF file.

- **Row 1:** A combined, flat PDF containing any Supplementary Methods, Discussion, Equations, Notes, Additional Supplementary Figures, simple tables, and all associated legends. Only one such file is permitted.

- **Row 2:** Nature Research's Reporting Summary; please provide an updated Summary, fully completed, without any mark-ups or comments. **Please note that this is a required document.**

Item	**Present?**	**Filename** This should be the name the file is saved as when it is uploaded to our system, and should include the file extension. The extension must be .pdf	**A brief, numerical description of file contents.** i.e.: *Supplementary Figures 1-4, Supplementary Discussion, and Supplementary Tables 1-4.*
Supplementary Information	No		

Reporting Summary	Yes	Sasaki_nr-reporting-summary_RENEWED

B. Additional Supplementary Files

Complete the Inventory below for all additional Supplementary Files that cannot be submitted as part of the Combined PDF.

- Do not list Supplementary Figures in this table (see section 2A)
- Where possible, include the title and description within the file itself
- Spreadsheet-based tables and data should be combined into a workbook with multiple tabs, not submitted as individual files.
- Please note that the *ONLY* allowable types of additional Supplementary Files are:
  - Supplementary Tables
  - Supplementary Videos
  - Supplementary Audio
  - Supplementary Data
  - NMR Data
  - Cryo-EM Data
  - Computational Data
  - Suppl. Software

Type	**Number** If there are multiple files of the same type this should be the numerical indicator. i.e. "1" for Video 1, "2" for Video 2, etc.	**Filename** This should be the name the file is saved as when it is uploaded to our system, and should include the file extension. i.e.: *Smith_ Supplementary_Video_1.mov*	**Legend or Descriptive Caption** Describe the contents of the file
Choose an item.			
Choose an item.			
Choose an item.			
Choose an item.			

Choose an item.			
Choose an item.			

Add rows as needed to accommodate the number of files.

3. Source Data

Complete the Inventory below for all Source Data files.

- Acceptable types of Source Data are:
  - Statistical Source Data
    - Plain Text (ASCII, TXT) or Excel formats only
    - One file for each relevant Figure, containing all source data
  - Full-length, unprocessed Gels or Blots
    - JPG, TIF, or PDF formats only
    - One file for each relevant Figure, containing all supporting blots and/or gels

Figure	**Filename** **This should be the name the file is saved as when it is uploaded to our system, and should include the file extension. i.e.: *Smith_SourceData_Fig1.xls,* or *Smith_Unmodified_Gels_Fig1.pdf***	**Data description** **i.e.: Unprocessed Western Blots and/or gels, Statistical Source Data, etc.**
Source Data Fig. 1		
Source Data Fig. 2		
Source Data Fig. 3		
Source Data Fig. 4		
Source Data Fig. 5		
Source Data Fig. 6		
Source Data Fig. 7		

Source Data Fig. 8		
Source Data Extended Data Fig. 1		
Source Data Extended Data Fig. 2		
Source Data Extended Data Fig. 3		
Source Data Extended Data Fig. 4		
Source Data Extended Data Fig. 5		
Source Data Extended Data Fig. 6		
Source Data Extended Data Fig. 7		
Source Data Extended Data Fig. 8		
Source Data Extended Data Fig. 9		
Source Data Extended Data Fig. 10		