Low-dimensional encoding of decisions in parietal cortex reflects long-term training history

Authors: Kenneth W. Latimer[1*] & David J. Freedman[1]

[1]Department of Neurobiology, University of Chicago

*Correspondence; E-mail: latimerk@uchicago.edu.

October 7, 2021

Abstract

Neurons in parietal cortex exhibit task-related activity during decision-making tasks. However, it remains unclear how long-term training to perform different tasks over months or even years shapes neural computations and representations. We examine lateral intraparietal area (LIP) responses during a visual motion delayed-match-to-category (DMC) task. We consider two pairs of monkeys with different training histories: one trained only on the DMC task, and another first trained to perform fine motion-direction discrimination. We introduce generalized multilinear models to quantify low-dimensional, task-relevant components in population activity. During the DMC task, we found stronger cosine-like motion-direction tuning in the pretrained monkeys than in the DMC-only monkeys, and that the pretrained monkeys' performance depended more heavily on sample-test stimulus similarity. These results suggest that sensory representations in LIP depend on the sequence of tasks that the animals have learned, underscoring the importance of training history in studies with complex behavioral tasks.

1 Introduction

Activity of single neurons in the macaque lateral intraparietal area (LIP) encodes task-relevant information in a variety of decision-making tasks (Freedman & Ibos, 2018). As a result, LIP has been proposed to support many different neural computations underlying perceptual decision making, including abstract visual categorization (Freedman & Assad, 2016; Huk et al., 2017). Throughout a lifetime, animals learn to make many different kinds of decisions in a variety of tasks and contexts, and different animals collect a unique set of experiences that shape their perceptual and decision-making skills and strategies (Summerfield & De Lange, 2014; Goldstone & Byrge, 2015). In contrast, experiments designed to study neural mechanisms of decision making often focus on neurons recorded during a specific task in isolation. However,

previously learned neural representations and strategies may impact how a cortical region is recruited when learning a new task. To understand the generality and flexibility of neural representations which support decision making, we aim to compare decision-related LIP activity in animals performing the same tasks, but with different long-term training histories.

We examine LIP recordings in four monkeys performing two related tasks in which they were required to determine if sequentially presented motion directions matched according to a learned rule (Fig. 1A). In both tasks, the monkey views two random dot motion stimuli (sample and test) separated by a delay period. To receive a reward, the animal responds by releasing a touch bar if the two stimuli match or by continuing to hold the touch bar on non-match trials. The delayed-match-to-sample (DMS) task is a memory-based, fine-direction discrimination task in which the sample and test motion stimuli match only if they are in the exact same direction. In the delayed-match-to-category (DMC) task, the stimuli match if the directions belong to the same category (red or blue) according to a learned arbitrary category rule. In the DMC task, two matching stimuli may be nearly 180° apart but belong to matching categories, while neighboring directions on different sides of the category boundary do not match. Thus, while the tasks use the same structure and stimuli, they require performing different perceptual and/or cognitive computations.

We consider two pairs of monkeys with two different training histories (Fig. 1B). In one pair of monkeys (B and J; Swaminathan & Freedman, 2012), the monkeys were trained only to perform the DMC task (i.e., without first training the monkeys on fine discrimination), and LIP recordings were made after training was completed (DMC only populations). The second pair of monkeys (pretrained monkeys D and H; Sarma et al., 2016), was first trained extensively on the DMS task, and LIP recordings were obtained after training (DMS population). The monkeys were then retrained on the DMC task, and a set of LIP recordings was made during an intermediate stage of training (when the monkeys' performance stabilized; DMC early populations). After the DMC-early recordings, the monkeys received additional training which overemphasized near-category-boundary sample stimuli (the most difficult conditions where the monkeys' performance was lowest) so that the monkeys' performance increased. After the second training stage was complete, a final set of LIP neurons was recorded during the DMC task (DMC late populations). In this study, the monkeys did not perform both tasks during a single session; they were switched exclusively to the DMC task and retrained over the course of months. In total, we analyzed eight LIP populations from four animals.

Not only do the DMS and DMC tasks share the same structure, timings, and stimuli, many of the sample-test pairs are rewarded for the same responses in both contexts (e.g., sample and test stimuli of the same direction match in both tasks). It is plausible that pretrained monkeys may reuse strategies acquired for performing the discrimination task while learning the DMC task. Similarly, different training histories may lead to different behavioral strategies to perform the DMC task, and different strategies may give rise to different patterns of activity in LIP (Tang et al., 2020). Previous studies have found that stimulus encoding and working-memory dependent sustained-firing activity in the prefrontal cortex during cognitive tasks depends on training (Qi & Constantinidis, 2013; Li et al., 2020). Additionally, the causal contribution of the middle temporal (MT) area of macaque visual cortex, which encodes motion and projects to LIP (Born & Bradley, 2005), depends on training history during motion-direction

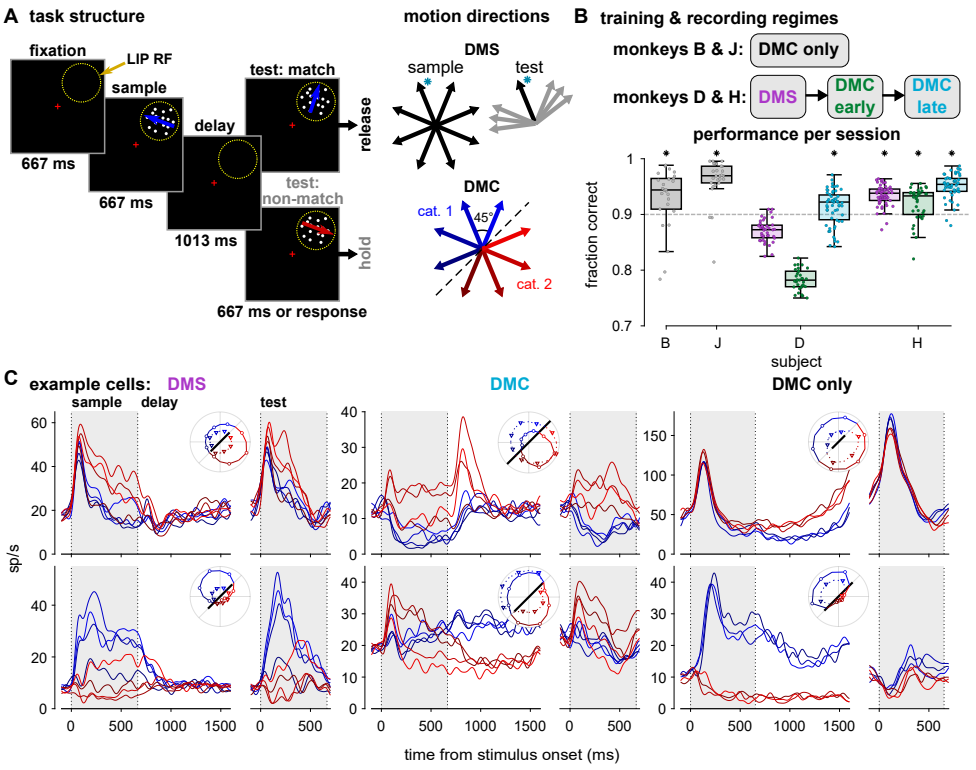

Figure 1: LIP recordings during DMS and DMC tasks. (**A**) In both tasks, the animal fixated and viewed a motion direction stimulus (sample). Following a delay period, a second stimulus (test) was presented. The monkey signaled if the sample and test stimuli matched by releasing a touch bar, otherwise the monkey was required to hold the touch bar. In the DMS task, the sample and test stimuli matched only if the directions were exactly the same. In the DMC task, the stimuli matched if they belonged to the same category: the motion directions were split into two equally sized categories with a 45°-225° boundary (red and blue directions; the boundary was constant for all sessions). The motion stimuli were placed inside the LIP cell's response field (yellow circle) during recording. (**B**) (top) Training and recording regimes for the four monkeys. (bottom) Performance during each recording session (dots) for each animal are summarized by the box plots. Colors correspond to the task and training period (DMS, DMC early or late, and DMC only). Asterisks indicate median per-session performance is greater than 90 % ($p < 0.01$, one-sided sign test, Holm-Bonferroni corrected). All four monkeys learned to perform the DMC task with a median performance of at least 90 % per session. (**C**) Mean firing rates of six single LIP cells recorded during each task. Colors correspond to the stimulus direction and category. Firing rates aligned to the sample stimulus onset are averaged by sample direction (left), and the test stimulus aligned rates are averaged over test direction (right). Although motion categories were not part of the DMS task, the directions are labeled blue or red for consistency. (inset polar plots) The mean firing rate for each sample direction during the sample stimulus presentation (circles and solid lines; 0 ms to 650 ms after motion onset) and delay period (triangles and dotted lines; 800 ms to 1450 ms after motion onset). The solid black line denotes 20 sp/s.

discrimination (Liu & Pack, 2017) and coarse depth discrimination (Chowdhury & DeAngelis, 2008). We therefore hypothesize that differences in training history could result in differences

in LIP population activity during the DMC task which reflect behaviorally relevant aspects of the neural computations underlying categorization (Churchland & Kiani, 2016).

Direction and category selectivity is visible in the average firing rates of single LIP neurons in both pairs of monkeys during the DMC task (Fig. 1C). However, based on single cells alone it is difficult to uncover the computations involved in the DMC task: how sample category is computed and then stored during the delay period, or how the test stimulus is compared to the sample. Instead, we take a dimensionality reduction approach to compare the low-dimensional geometry of population responses to better illuminate how LIP encodes different tasks (Okazawa et al., 2021).

While many methods of dimensionality reduction are available, we sought a compact, low-dimensional description of LIP responses that quantified the population responses as a function of the task variables. Moreover, we wished to perform dimensionality reduction on the trial-by-trial spike train responses (as opposed to trial-averaged spike rates) within each population in order to account for structure in the neural activity beyond mean firing rates (e.g., bursting or oscillations). We therefore introduce the generalized multilinear model (GMLM) as a model-based dimensionality reduction method for population activity during flexible cognitive tasks (Fig. 2A). The GMLM is a tensor-regression extension of the generalized linear model (GLM) which describes a single neuron's spiking response to different task events through a set of linear weights or kernels (Zhou et al., 2013; Park et al., 2014; Robinson et al., 2016; Kossaifi et al., 2020). The GMLM fits the data from all cells in a dataset into one compact representation by taking a low-rank tensor decomposition of linear kernels of the task events — the sample and test stimuli and the touch-bar release — that best describes the shared response dynamics across the entire population as a function of the task variables. This contrasts with the GLM, which fits each cell individually without directly recovering low-dimensional structure. Similarly to exponential family principal components analysis (Mohamed et al., 2008), the GMLM can be applied directly to binned spike count data rather than smoothed spike rates. The GMLM inherits the GLM's flexibility for modeling trials with variable structure. For example, the timing of the end of the trial is controlled by the animal via their releasing the touch-bar. In contrast, dimensionality reduction approaches based on peristimulus time histograms (PSTH) require temporally aligned trials (e.g., principal components analysis-based methods; Kobak et al., 2016; Aoi et al., 2020), thereby limiting those approaches' ability to quantify motion and category tuning or touch-bar response related activity during the test period. Stimulus category and direction are low-dimensional variables, and motion direction tuning in sensory regions such as area MT can be well-captured by simple parametric models (Rust et al., 2006). Therefore, the GMLM is well-suited for modeling how LIP populations represent combinations of these task variables during the DMC task.

By applying the GMLM to the LIP populations, we quantified population-level differences in LIP activity between animals and compared those differences with behavioral performance with respect to the animals' training histories. We found category and direction selectivity in LIP during the DMC task in all subjects. However, we found stronger cosine-like motion direction tuning in LIP during the DMC task in monkeys first trained on the DMS task compared to monkeys trained only on categorization. During the test stimulus presentation when the monkeys had to compare the incoming test stimulus to the remembered sample stimulus,

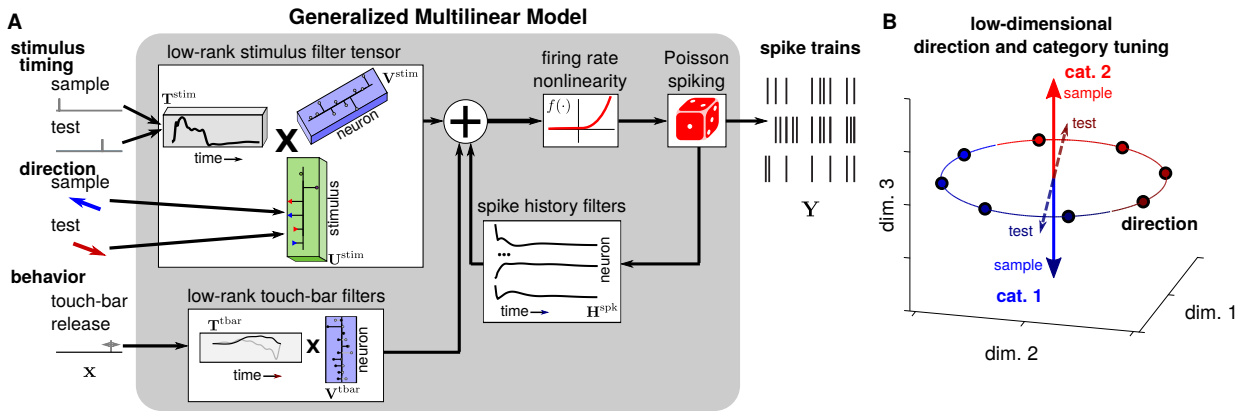

Figure 2: The generalized multilinear model for dimensionality reduction of neural populations. (**A**) Diagram of the GMLM. Incoming stimuli are factorized into temporal events and stimulus weights that encode direction and category information (Fig. S1E-I). A set of temporal kernels and stimulus coefficients filter the stimuli into a low-dimensional stimulus response space. The touch-bar release event is similarly filtered using a low-dimensional set of temporal kernels. Each individual neuron's firing rate at each time bin is a nonlinear function (here, $f(\cdot) = \exp(\cdot)$) of the sum of a linear weighting of the low-dimensional stimulus subspace, a linear weighting of the touch-bar subspace, and recent spike history. Spiking is given as a Poisson process given the instantaneous rate. Because we do not include interactions between neurons here, this model can be applied to a set of single-neuron recordings. However, the model can readily be extended to include interactions in simultaneously recorded populations (Pillow et al., 2008; Pandit et al., 2020). (**B**) The model represents the different stimulus directions and categories — including whether it is the sample or test stimulus — as vectors in a low dimensional space where the dimensionality is the number of factors. The vectors change over time given the temporal kernels. The models we focus on here constrain the direction tuning, but not category, to be constant over the sample and test stimuli. The full GMLM allows for flexible direction tuning (i.e., each stimulus direction is a distinct point) and the cosine-tuned GMLM constrains the direction tuning to an ellipse. Thus, dimensionality reduction in the GMLM can take into account that temporal dynamics and stimulus tuning information can be shared across the two stimulus presentations. Individual neurons' stimulus tuning is a linear projection of the low-dimensional space.

sample category could be more reliably decoded from LIP responses irrespective of the test-stimulus direction in the DMC-only monkeys than in the pretrained monkeys. Behaviorally, the pretrained monkeys were more likely to make categorization errors when the sample and test stimuli were similar than the DMC-monkeys. Additionally, we introduce dynamic spike history within the GMLM which reveals a difference in oscillatory, single-trial dynamics during the delay period between the DMC-only and pretrained monkeys. Together these results suggest that different subjects may recruit distinct behavioral and neuronal strategies for performing the DMC task, and that long-term training history may play a role in shaping these differences. Low-dimensional encoding of the DMC task in LIP may therefore reflect training history or a particular task strategy (or both).

2 Results

2.1 Dimensionality reduction in LIP with the GMLM

We aimed to describe how task-related responses in LIP are shared across neurons in a population by reducing the dimensionality of the eight LIP populations using GMLMs. The GMLM uses a low-dimensional set of components to describe the population responses during each trial as a multilinear function of the task variables (Fig. S1). To place the DMC task into the GMLM framework, the motion stimuli were linearized as two sets of regressors: (1) timing events to indicate the onset of a stimulus (sample or test) and (2) stimulus identity regressors that encode direction, category, or if the stimulus is the sample or test. The model's parameters include a set of stimulus components, where each component contains a single temporal kernel (or linear filter) and a set of weights for the stimulus identity. Each component temporally filters the stimulus onset events, and weights the filtered stimuli linearly by stimulus identity. As a result, each individual component contributes to the population encoding for all stimuli (not just a single motion direction or sample/test presentation). Each individual neuron's tuning to the motion stimuli is a linear combination of the stimulus components. The model also includes a low-dimensional set of components to represent the touch-bar release event: a set of temporal kernels describe the population response to a touch-bar release such that each neuron's touch-bar tuning is a linear combination of those kernels. Each spike train is then defined as a Poisson process in which the instantaneous firing rate is given by the sum of the filtered stimuli and touch-bar release, plus a linear function of recent spike history. The set of stimulus components is a low-rank tensor that represents population tuning to the motion stimuli in a low-dimensional subspace which captures shared response dynamics across neurons in the population. The factorized representation of the stimulus into temporal and identity weights captures shared temporal dynamics between different directions or between sample and test stimuli. As the number of components (i.e., rank) in the kernel tensor increases, the model approaches a GLM fit to each cell individually.

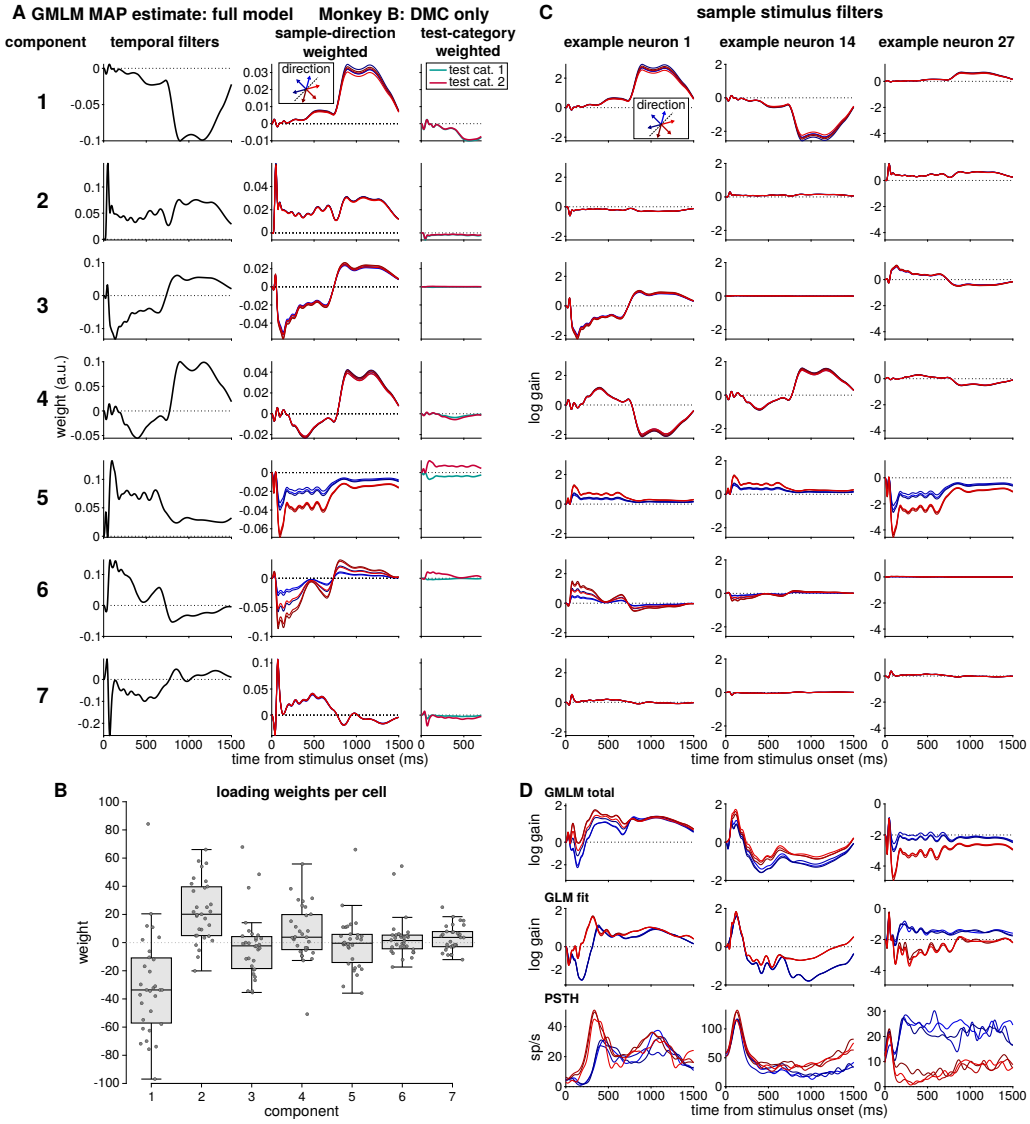

Figure 3: GMLM with seven stimulus components fit to the LIP neurons recorded from monkey B during the DMC task. (**A**) The seven GMLM stimulus components that define the population responses to the motion stimuli are given in each row. (left) The temporal kernels for each component (normalized). (middle) The temporal kernels weighted for each sample stimulus direction. (right) The temporal kernels weighted by the test category (the test stimulus kernels are shown shorter than the sample stimulus kernels because trial ends after the test stimulus presentation). The total kernels for each test direction are computed by adding the test category kernels to the sample direction filter. As a result, the sample and test kernels have the same direction tuning, but the kernels may have different category tuning (e.g., component 5). (**B**) The distributions of weights for each component across all neurons. Each dots is the weight for one neuron for the given component. (**C**) Each column shows the model fit of the sample stimulus kernels for three example neurons. The first seven rows show the seven components scaled by each neuron's linear weighting of the component (the middle column of part A with neuron-dependent scaling from B). (**D**) Total sample stimulus tuning for the three example cells. The total GMLM stimulus kernels (i.e., the sum of rows in C; top row), the GLM fit to the individual cell (middle), and the PSTH conditioned on sample stimulus direction for comparison (bottom). The low-dimensional set of kernels for the touch-bar release are shown in Fig. S2.

We designed a set of four nested models (i.e., different linearizations of the motion stimuli with increasing complexity) in order to assess what stimulus information is encoded by an LIP population (Fig. S1E-I). The simplest model was the no category or direction tuning model. In this model, the linear weights for the stimulus identity only defined whether the stimulus was the sample or test. This model can only capture the average trajectories in time during the task over all stimulus directions. The second model, the category only model, includes stimulus category weights, but does not consider specific motion direction. The category only model includes stimulus identity information for the category one and category two motion directions for both the sample and test stimuli. This way, the model can capture category tuning, which may be different for the two stimulus presentations. While the DMS task had no category, we still fit category weights to the DMS populations as a control (i.e., to ask what the model produces if category was not actually a behavioral factor in the task). The third model included cosine direction tuning and category. In addition to the category weights from the previous model, this GMLM included two coefficients for the sine and cosine of the motion direction. The cosine and sine weights were the same for both sample and test stimuli. Thus, this model constrains the geometry of direction tuning to lie on an ellipse in the low-dimensional space (Fig. 2B). The final model, the full model, extends the cosine direction tuning model by allowing different weights for each individual motion direction, rather than constraining the direction information to be cosine tuned.

The full GMLM fit to the LIP population from monkey B is shown in Fig. 3A,B. The model has seven stimulus tensor components, each with a temporal kernel (left column). The temporal kernel is scaled for each of the six directions in the task to give a temporal kernel for the sample stimulus (middle column). Next, we include additional test category kernels, which are added to the direction kernels in the middle column, to describe the response to the test stimulus (right column). Different components can have different temporal response dynamics and different stimulus tuning properties: for example, component 5 shows strong differentiation between the two stimulus categories (red and blue), while component 2 does not. Each cell's response tuning (linear kernels) is then a linear combination of the components. The tuning to the sample directions for individual cells is illustrated in Fig. 3C,D. The GLM fits to the individual cell and corresponding PSTH are shown below the GMLM fit for comparison. We note the total firing rates fit by the models are a combination of the stimulus filters, baseline rate, spike history effects, and firing rate nonlinearity. As a result, the PSTHs and filters do not match exactly.

The parameterization in the cosine-tuning and full models assumes the direction tuning (but not category tuning) is the same for both sample and test stimuli: that is, the difference between the kernels for two motion directions within the same category are the same for both the sample and test stimulus. Such direction tuning constancy would be consistent with common bottom-up, direction-tuned input from sensory areas such as MT for the two stimulus presentations. The model still includes test category filters, which allow for different category tuning or direction-independent gain differences between sample and test stimuli. We found that including separate sample-test direction tuning in either model did not improve the model fit (Fig. S3A). Additionally, comparing sample and test direction weighting in the low-dimensional GMLM component space showed similar direction preferences for the two stimuli (Fig. S3B). Thus, the GMLM framework can both constrain the parameters to enforce constant direction

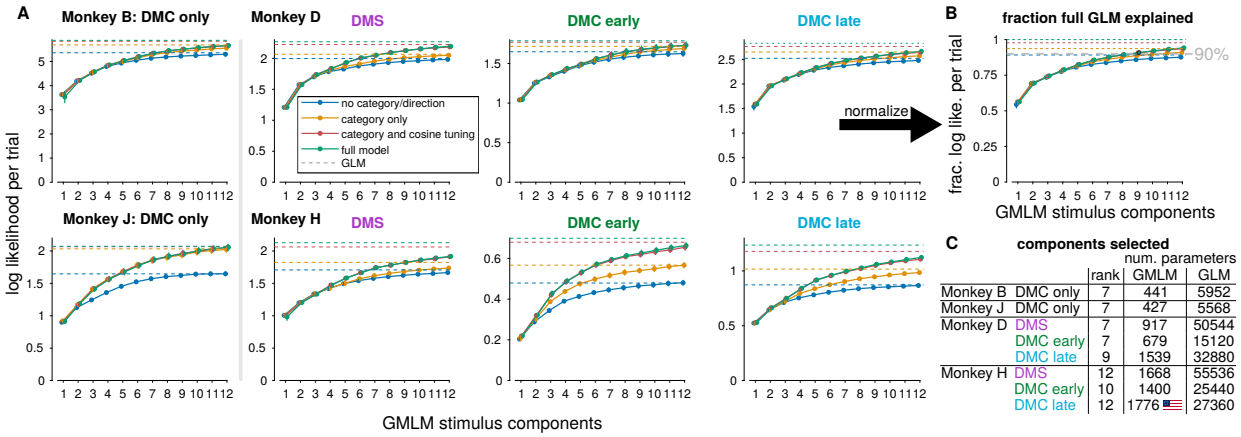

Figure 4: Rank selection in the GMLM. (**A**) The mean cross-validated log likelihood per trial averaged across neurons of the GMLM as a function of the stimulus kernel tensor rank relative to the model without any stimulus terms (i.e., the rank 0 model) for each of the eight LIP populations. Each trace shows the log likelihood for a single model parameterization: no stimulus information (i.e., only mean temporal dynamics; blue), category only (no specific direction information; yellow), category plus cosine direction tuning (red), and the full model with flexible direction tuning (green). The dashed lines show the mean cross-validated likelihood of the GLMs, which correspond to the full-rank model. (**B**) The fraction of log likelihood of the full GLM (fraction of the log likelihood that could be captured by the GMLM) was used to select the GMLM rank for each LIP population, shown here for monkey D, DMC late. This fraction is the cross-validated log likelihood divided by the log likelihood of the full GLM (dashed green line), and the threshold for rank selection was 90 % of the log likelihood per trial. (**C**) Number of stimulus components (rank) selected for the GMLM for each LIP population. The number of stimulus parameters in the low-rank GMLM (full model) is compared to total parameters in the equivalent single-cell GLM fits for all cells in each population.

tuning between the two stimuli, and test statistically whether that assumption holds.

2.2 Model selection and dimensionality

We first determined the dimensionality of stimulus-related activity in the LIP population responses. We varied the number of components to include in the GMLM (i.e., the rank of the stimulus kernel tensor). We compared the full GMLM to the corresponding single-cell GLM fits, where the GLMs represent the “full-rank” model. We computed the average likelihood per trial averaged over the neurons in each population for the GMLM fit, relative to the GMLM without any stimulus terms (Fig. 4A). We selected the rank by the number of components needed to explain, on average, 90 % of the explainable log likelihood per trial over all the neurons in each population (Fig. 4B). The GMLM required 7 to 12 stimulus components per population, thereby using only a fraction — less than 8 % — of the number of parameters compared to GLM fits to individual cells (Fig. 4C).

We then asked how much each task variable contributed to the low-dimensional LIP responses. To do so, we quantified model fit as we monotonically increased the complexity in the nested

models by including more information about the stimulus identity and category. A majority of the log likelihood was accounted for by the GMLM without category or direction tuning in all populations, which is consistent with many previous dimensionality reduction results (Kobak et al., 2016). The category-only GMLM captured a greater percentage of the log likelihood over the no-category model in the DMC late populations than in the DMS populations (Monkey B, DMC only 2.5 ± 0.3 %; Monkey J, DMC only 13.8 ± 0.9 %; Monkey D, DMS 2.1 ± 0.3 %, DMC early 1.4 ± 0.7 %, DMC late 2.8 ± 0.2 %; Monkey H, DMS 3.3 ± 0.2 %, DMC early 11.3 ± 0.7 %, DMC late 9.5 ± 0.3 %; mean percentage cross-validated log likelihood accounted for by the category-only GMLM minus the no-category GMLM where the errors are ±2 SEM over the cross-validation folds).

Cosine direction tuning during the DMC task accounted for a larger improvement of the model fit over the category only model for the pretrained monkeys than the DMC-only monkeys. (monkey B, DMC only 1.0 ± 0.4 %; monkey J, DMC only 0.2 ± 0.7 %; monkey D, DMS 4.1 ± 0.3 %, DMC early 1.6 ± 1.0 %, DMC late 1.9 ± 0.1 %; monkey H, DMS 8.2 ± 0.6 %, DMC early 11.9 ± 0.5 %, DMC late 9.9 ± 0.6 %; mean percentage cross-validated log likelihood accounted for by the cosine-tuned GMLM minus the category-only GMLM). Thus, direction-tuning played a stronger role in the pretrained monkeys' LIP activity than in the DMC-only monkeys.

We tested the adequacy of cosine parameterization of direction tuning by comparing the more flexible full model. The cosine direction tuning model was comparable to the full model for all populations (monkey B, DMC only 0.2 ± 0.2 %; monkey J, DMC only 0.3 ± 0.7 %; monkey D, DMS 0.2 ± 0.2 %, DMC early 0.2 ± 0.1 %, DMC late 0.2 ± 0.2 %; monkey H, DMS 0.3 ± 0.4 %, DMC early 1.1 ± 0.5 %, DMC late 1.3 ± 0.3 %; mean percentage cross-validated log likelihood accounted for by the full GMLM minus the cosine-tuned GMLM). For these tasks, the direction tuning in the population could therefore be approximated as an ellipse (and thus embedded within a plane).

2.3 Low-dimensional response to the sample stimulus

To gain intuition about how LIP dynamics may support the transformation of motion direction input into a representation of category, we visualized the low-dimensional population tuning to the sample stimuli. The top three dimensions of the trajectories show large, stimulus-independent transient responses (Fig. 5, inset; Fig. S4). This is consistent with the large fraction of the data explained by the GMLM without category or direction tuning. We therefore subtracted the mean response over stimuli and plotted the top three dimensions in the mean-removed responses (Seely et al., 2016). The two DMC-only LIP populations showed primarily two-dimensional responses with strong category separation (Fig. 5). For the pretrained monkeys, the trajectories during the DMS task reflected the stimulus geometry: the model shows two-dimensional transient activity organized by stimulus angle. The responses show little stimulus-specific persistent activity during the delay period (Sarma et al., 2016): the trajectories return to the origin after stimulus offset. During the DMC early phase, the low-dimensional LIP response reflects the stimulus geometry, but the top dimension is aligned to the task axis (i.e., blue and red directions are separated along dimension 1). The trajectories are still two-dimensional without clear delay period encoding. After training was completed,

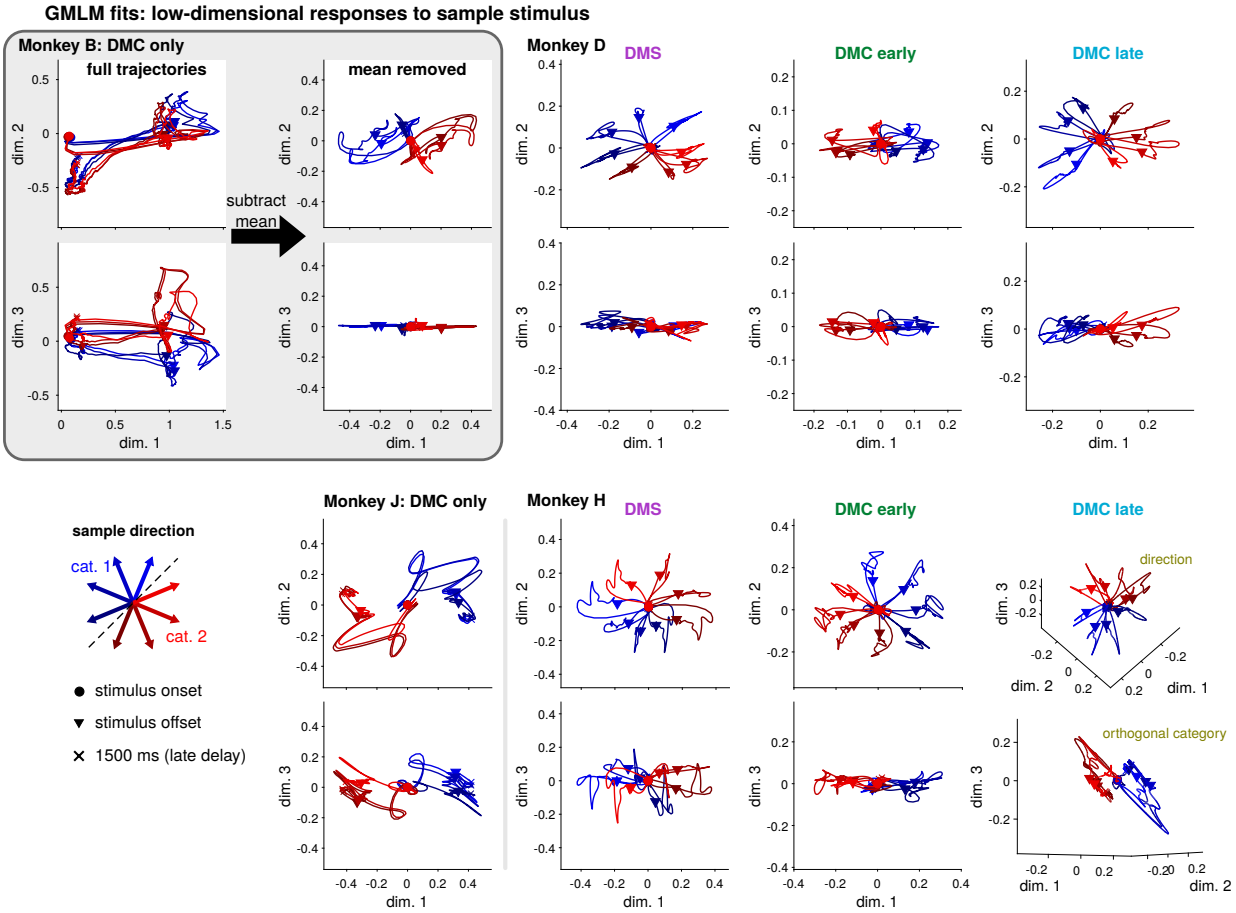

Figure 5: Low-dimensional representations of motion direction and category during the first 1500 ms of each trial (sample stimulus presentation and delay period). The top three dimensions of the GMLM's sample stimulus encoding for each of the eight LIP populations with the mean response over all directions removed. The inset for monkey B shows the top three dimensions including the mean (left) and the top three dimensions that remain after removing the mean (right). The two plots for each monkey show the top dimension on the x-axis plotted against the second or third dimensions on the y-axes (except for monkey H shown in the 3-D plots). The red and blue traces show the response to each motion direction from stimulus onset (circles), to stimulus offset (triangles), and into the delay period (x's denote 1500 ms after sample motion onset). The three-dimensional trajectories (of the cosine-tuned model) are shown as a function of time in Fig. S5.

monkey D's DMC late LIP activity showed strong direction tuning during the stimulus presentation which is elongated along the task axis (that is, the axis most oriented to category along the 135°-315° stimulus directions). In contrast, LIP in monkey H had a three-dimensional stimulus response in the late period: two dimensions reflecting the circular motion directions during the stimulus presentation and a third orthogonal axis for category that was sustained through the delay period. Similar orthogonal stimulus input and working memory representations have been observed in other decision making tasks (Aoi et al., 2020; Libby & Buschman, 2021). In summary, the low-dimensional stimulus components of the LIP activity differed across ani-

mals such that the pretrained monkeys' LIP showed strong, circular representations of motion direction, while the DMC-only monkeys had lower-dimensional responses that more strongly reflected category.

2.4 Quantifying the geometry of category and direction in the stimulus subspace

To go beyond visualization of the low-dimensional subspace, we wished to quantitatively assess the geometry of category- and direction-dependent responses in LIP. Here, we focused on the cosine-tuned GMLM. Given this choice of parameterization, direction and category could be decoupled, while still capturing a similar subspace to the full model (Fig. S5). At each point in time, category was encoded along a vector while direction (parameterized by angle) was encoded on an ellipse in the stimulus subspace (Fig. 6A). The ellipse could be circular, which would represent motion directions uniformly, or elongated so that the population representations are biased towards a preferred motion direction. We compared the norm of major and minor axes of the direction ellipse and the norm of the sample category vector (Fig. 6B) as a function of time relative to stimulus onset. We conducted a Bayesian analysis of the GMLM's sample stimulus subspace to take into account uncertainty in the model fit given the posterior distribution of the model parameters.

The DMC-only LIP population subspaces showed strong category tuning relative to direction tuning. The category vector in monkey B was of similar magnitude to the minor axis of the direction ellipse during stimulus presentation, and stronger during the delay period. The category vector in monkey J was larger than the direction tuning ellipse throughout the stimulus presentation and delay period. The direction tuning ellipses were elongated along a particular motion direction, rather than circular. Additionally, the direction ellipse aligned both with category and with the choice biases in monkeys B and J on trials where the sample motion direction was ambiguous (Fig. S6). The sample stimuli on ambiguous trials were placed on the category boundary, and the monkeys were rewarded randomly. The ambiguous trials were not used to fit the GMLM. Thus, the stimulus components in the DMC-only populations reflected category-specific input selection.

In the pretrained monkeys, the DMS populations showed strong direction tuning, which in monkey H was nearly circular or uniform (i.e., the major and minor axes of the direction ellipse were of similar length). During the late DMC sessions — but not during the DMS task — monkey D's direction ellipse was aligned with the task category (i.e., the major axis was along the 135°-315° angles; Fig. S6C). The same task-aligned direction encoding during the DMC task was not observed in monkey H. In both the monkey D late DMC and monkey H early DMC populations, we found stimulus offset activity in the direction ellipse, but not in the category vector. As a result, individual neurons may appear to respond more strongly for a particular category early in the delay period, but the model accounted for this as a direction-tuned response rather than category-specific encoding. The monkey H early and monkey D early and late DMC populations did not have large category vectors, and the low-dimensional activity instead reflects an elongated direction tuning ellipse (i.e., the major axis is larger than the minor axis) during stimulus presentation. In the monkey H late DMC population, we observed a

slow increase in the category vector length over time in the trial, which does not surpass the magnitude of direction tuning until the delay period.

We then asked how the subspace geometry could affect how decoding methods assess category selectivity in LIP. We applied a linear decoding technique previously proposed to reveal category representations independent of motion direction (Swaminathan et al., 2013; Sarma et al., 2016). The decoder classifies the sample category based on spike counts from pseudopopulation trials. We trained and validated the decoder on trials from orthogonal sets of stimulus directions (Fig. 6C). The logic of the decoder is that, if motion direction is represented in the population circularly without any additional category-specific responses, the decoder will not generalize across the training and validation conditions. The DMS populations provided a control for the method, because the monkey was not yet trained to classify motion category. We indeed found no significant category decoding in the two DMS populations (Fig. 6D).

The decoder performances during the DMC task were consistent with the GMLM stimulus subspaces. Category could be decoded with high accuracy in monkeys B and J early in the stimulus presentation and throughout the delay period (Fig. 6D). Similarly, the GMLM analysis had found strong category tuning beginning early in the stimulus period and continuing through the delay in those populations. The results were different in the pretrained animals. In both DMC early and monkey D's DMC late populations, we found decoding above chance during stimulus presentation, but not during the delay. The task-aligned, non-circular response to stimulus direction in monkey H DMC early and monkey D DMC early and late enabled the decoder to generalize across conditions due to over-representation of signal along the task dimension (135°-315°), rather than a category vector independent of direction. In contrast, in the DMC late population for monkey H, the decoder only found weak decoding late in the sample stimulus presentation, which became strong during delay period. The orthogonal category dimension of monkey H DMC late is only stronger than the circular direction coding during the delay period, corresponding to the onset of significant category decoding. The decoder's failure to generalize during the early sample period can therefore be explained by strong direction selectivity swamping the weaker, orthogonal category signal.

2.5 Comparing the sample and test stimuli

The DMC task requires different computations for the test and sample stimuli: the category of the sample stimulus must be computed and stored in short-term memory, while the test stimulus must be compared to the stored sample category. Recent work has suggested that LIP linearly integrates the test and sample stimuli during the test period of the DMC task while prefrontal cortex shows more nonlinear match/non-match selectivity (Zhou et al., 2021). We therefore compared the LIP responses to the test stimulus to the population responses to the sample stimulus. In the GMLM fits to the DMC populations, we found that test category tuning during the test stimulus presentation was weaker than sample category tuning during the sample presentation (Fig. S7). This can be seen in the low-dimensional subspace for monkey H (Fig. 7A). During the sample stimulus, the subspace reflected category tuning orthogonal to the motion direction response (Fig. 5, bottom right). However, we did not find the same category-selective response to the test stimulus in the stimulus subspace. Addition-

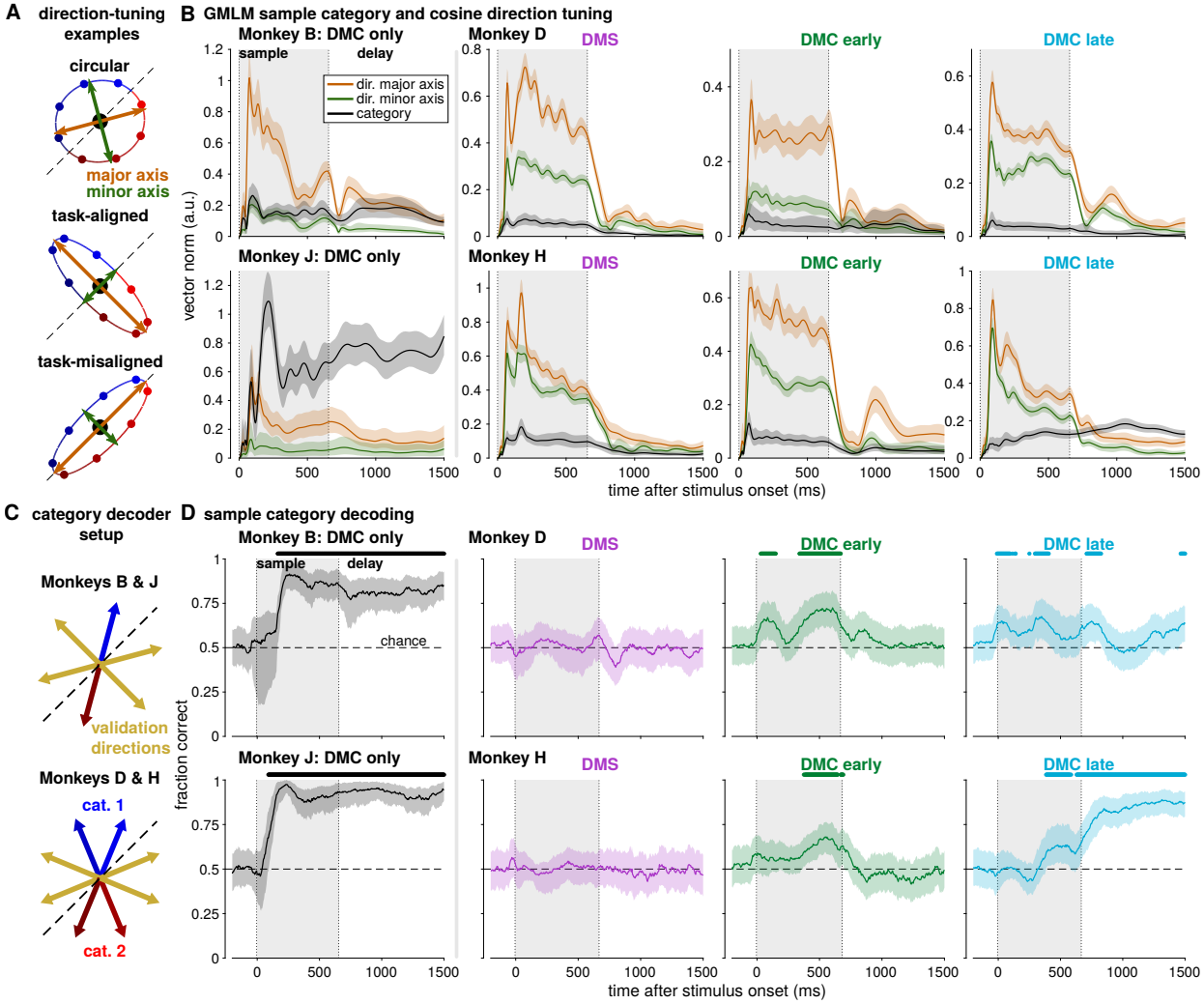

Figure 6: Quantification of category and direction encoding in LIP. **A** Diagram of direction encoding in the cosine-tuned GMLM. Motion stimulus direction is encoded as an ellipse in the low-dimensional stimulus space. The ellipse has major (orange arrows) and minor (green arrows) axes which define its shape. If the axes are of similar length, tuning is approximately circular and the motion directions are evenly distributed in the low-dimensional space (top). If the major axis is elongate compared to the minor axis, the population shows a preferred direction which may be aligned with category (middle) or the null direction (bottom) or anywhere in between. Category is encoded as a vector in addition to the direction ellipse, and the category vector is constant for all motion directions within a category (in contrast to the task-aligned direction tuning which places near-boundary directions closer together). **B** Bayesian estimate of the geometry of the sample stimulus tuning for the eight LIP populations. Each plot shows the norm of the sample category vector (black) and the norms of the major (orange) and minor (green) axes of the direction tuning ellipse for an LIP population as a function of time relative to the sample stimulus onset. The solid lines denote the posterior median at each time point, and the shaded regions denote 99 % credible intervals. If the major and minor axes have equal norms, then direction would follow a circle in the low-dimensional space. [Continued on next page]

Figure 6: [Continued] **C** The training and test scheme for a direction-independent category decoder. Psuedopopulation trials were created from 50 random trials sampled with replacement for each direction from each cell. For monkeys B and J, the decoder is trained using one direction from each category (180° apart; red and blue) and validated on the remaining four directions (yellow). For monkeys D and H, the decoder is trained using two adjacent directions from each category (again using opposite directions from each category; red and blue) and validated on the remaining four directions (yellow). **D** Median decoder generalization performance as a function of time for each of the eight LIP populations. The decoder was trained and tested using spike counts in a 200 ms window centered at the time relative to sample stimulus onset on the x-axis. The shaded regions denote a 99 % confidence interval over 1000 random pseudopopulations. Symbols denote decoding significantly greater than chance (50 %; $p < 0.01$ Benjamini-Hochberg corrected, one-sided bootstrap test).

ally, LIP population activity projected onto the touch-bar (motor response) subspace showed strong match/non-match separation with little category selectivity (Fig. S8). Thus, LIP does not appear simply to extract and sum the categories of the two stimuli to compute match or non-match.

We tested if the stored sample category and incoming test stimulus category were separable in the LIP population responses during the test stimulus presentation. We used linear classifiers to decode the sample or test category from pseudopopulation spike counts during the first 200 ms of the test stimulus presentation. The decoders were trained on trials of all stimulus directions. However, the training set consisted of only match (or non-match) trials, while the validation set included only non-match (or match trials). For the two DMC-only animals, monkeys B and J, the sample category decoder generalized across the two conditions (i.e., performed better than chance). The test category decoder, however, performed much worse than chance (Fig. 7B). Thus, the decoding axis for test category switched signs across match and non-match trials. We observed the opposite pattern in the pretrained monkeys: the decoders generalized to classify the test stimulus category, but not the sample. In the DMC-only animals, sample category can therefore be read out by a single linear decoder regardless of the test stimulus identity, which is consistent with stronger separability of the remembered sample category and the incoming test stimulus in the DMC-only monkeys than in the pretrained monkeys. Increased separability suggests a coding scheme that reduces interference between the stored sample stimulus category and the specific test stimulus direction (Libby & Buschman, 2021).

We then asked how the monkeys' performances depended on the similarity between the sample and test stimuli. We compared the monkeys' accuracy as a function of distance between test and sample directions (Fig. 7C). The pretrained monkeys showed a different pattern of accuracy than the DMC-only monkeys. At small sample-test differences, the pretrained animals showed better performance on match than non-match trials while the DMC-only monkeys perform similarly or better on non-match. Additionally, the pretrained animals showed greater dependence on distance. These effects were greatest during the early DMC training phase, but they persisted after extensive training on the order of several months (the total number DMC training sessions between the DMC early and DMC late periods was 78 for monkey D

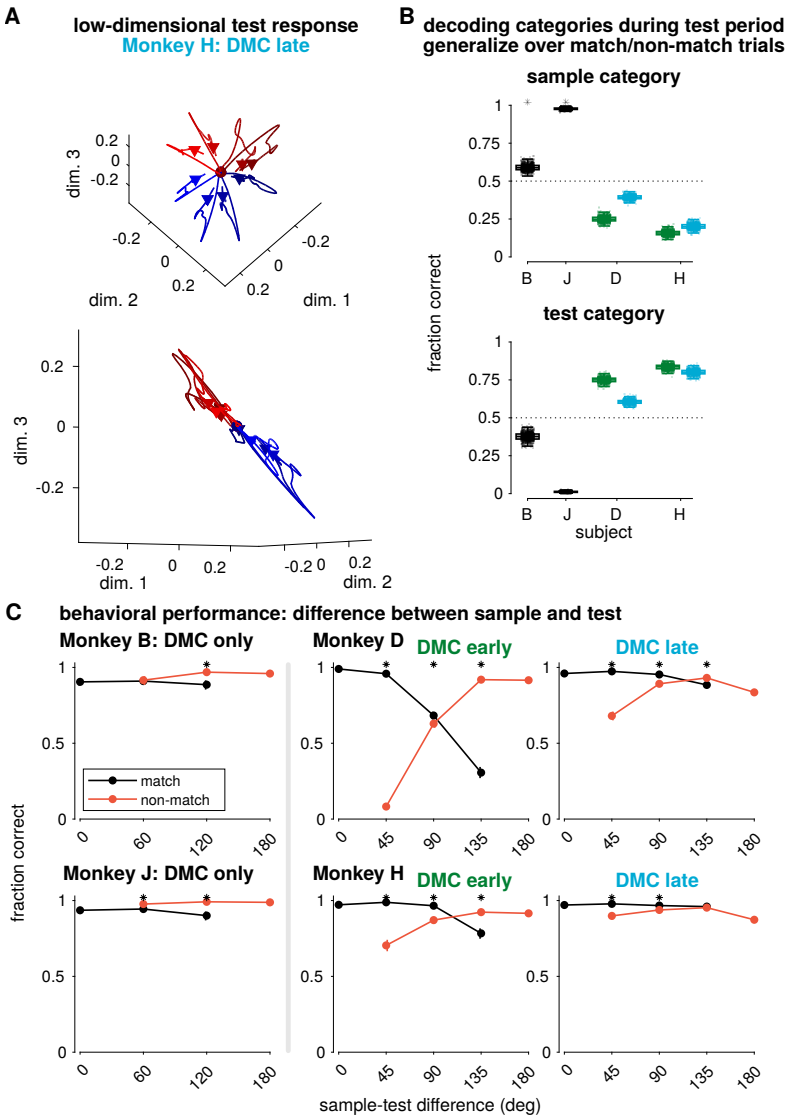

Figure 7: Matching the test stimulus to the stored sample in the DMC task. **A** The low-dimensional test stimulus response for each direction for monkey H, late DMC with the mean response removed projected into the same dimensions as in Fig. 5 (bottom right). **B** Decoding accuracy of sample (top) or test (bottom) category using the spike counts during the first 200 ms of the test stimulus (excluding the motor response for 95.7 % of match trials). The decoder was trained on trials from all stimulus directions, but only from match (or non-match) trials and then tested on non-match (or match) trials. All decoders generalized significantly different than chance (50 %; $p < 0.01$ Benjamini-Hochberg corrected, two-sided bootstrap test). **C** Average performance as a function of the difference in angle between the sample and test stimulus, sorted by match/non-match trials in the DMC task (error bars show a 99 % credible interval). Asterisks indicate match and non-match are significantly different ($p < 0.01$, two-sided rank sum test, Benjamini-Hochberg corrected).

and 65 for monkey H). Therefore, stimulus similarity — which was relevant in the DMS task — affects categorization behavior more strongly in the pretrained monkeys than the DMC-monkeys, reflecting the monkeys' strategy.

2.6 Single-trial dynamics during the delay period

Neural dynamics during single-trials may reflect aspects of sensory processing and working memory beyond the mean firing rate (Fontanini & Katz, 2008; Miller et al., 2018). For example, working memory may be supported by persistent activity (Constantinidis et al., 2018) or oscillatory bursts (Lundqvist et al., 2016) while stimulus-related activity may exhibit strong transient responses with quenched variability (Churchland et al., 2010). We therefore sought to characterize non-Poisson variability in single trials in LIP during the DMS and DMC tasks, which could reflect signatures of different strategies in performing the tasks. The GLM framework accounts for non-Poisson variability or single-trial dynamics by conditioning firing rates on recent spiking activity through a spike history filter, an autoregressive term which can reflect a combination of intrinsic (e.g., refractory periods) and network properties (e.g., oscillations) (Truccolo et al., 2005; Weber & Pillow, 2017; Zoltowski et al., 2019). Typically, the GLM assumes the spike history filter to be constant: that is, spike history has the same effect on spike rate throughout a trial. While fixed spike history effects may be an appropriate assumption in early sensory regions under stimulation with steady-state stimulus statistics, spiking dynamics in LIP might vary between the stimulus presentation and the delay period due to the transition in behavioral task demands between these two periods of the task (Hart & Huk, 2020). In order to quantify spike history effects in the DMS and DMC tasks, we extended the GMLM to include *dynamic spike history filters* which allows the autoregressive dynamics to change throughout a trial (Fig. 8A). The dynamic spike history in the GMLM was a low-rank tensor with a linear spike history component and a gain term relative to the stimulus timing (Harris et al., 2019). As a result, the model learns how each neuron's spike history filter changes during the course of a trial relative to task events, and can therefore capture distinct dynamics between stimulus-driven and delay periods of the trial.

We fit the GMLM with a single dynamic spike history component to the LIP populations. Including dynamic spike history improved the cross-validated model performance for all populations (Fig. S9A). The GMLM found similar dynamic spike history kernels for the two DMC-only monkeys, which showed oscillatory dynamics at approximately 12 to 14 Hz (low-beta; Fig. 8B). In contrast, the dynamic spike history kernel for the pretrained monkeys at all training stages was dominated by a faster timescale decay (time constants monkey D 10.2, 14.8 and 11.8 ms and monkey H 22.5, 9.0 and 6.6 ms DMS, DMC early, DMC late respectively), which suggests stronger gamma-frequency bursts. The stimulus-timing kernels showed that the dynamic spike history generally aligned with stimulus onset and offset (Fig. 8C). One notable exception was monkey J: the timing showed only a short transient gain after stimulus onset. The timing of this kernel corresponded with the strong category-dependent transient response in monkey J (Fig. 6B), and thus raises the possibility that the network enters a memory-storage state prior to stimulus offset.

Lastly, we examined population differences in the total spike history: the dynamic spike history filter (which depends on time in the trial) plus the individual neurons' fixed spike history filters. We computed the population mean spike history kernel at two different points (Fig. 8D): during stimulus-driven activity (100 ms after sample stimulus onset) and during the delay period (500 ms after sample offset). The mean spike history in the DMC-only monkeys showed a pronounced oscillatory-like trough during the delay period, compared to the pretrained monkeys

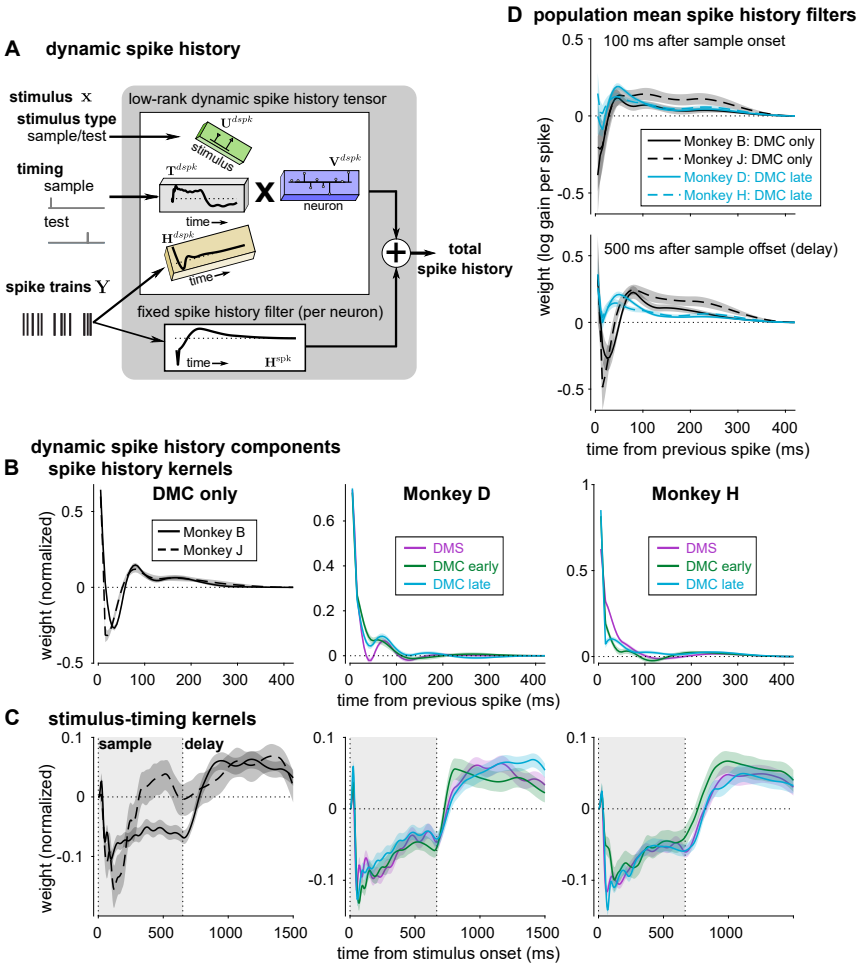

Figure 8: Dynamic spike history captures distinct stimulus-driven and delay-period dynamics. **A** The dynamic spike history filter is modeled as a low-rank, four-way tensor. The tensor includes two temporal kernels: one which filters spike history (gold) and a second which determines the weighting of the spike history component relative to the stimuli (gray). The spike history is scaled by stimulus identity (for simplicity, limited to sample or test stimulus weights only, without any category information). Each neuron adds the (weighted) dynamically filtered spike history to the neuron's constant spike history filter. The total spike history at any point in a trial is still a linear function of past spiking activity, but the effective linear kernel can change during a trial. The normalized rank-1 dynamic spike history components: **B** dynamic spike history kernels and **C** stimulus-timing kernels for each population (posterior median and 99 % credible interval). The left columns shows the two DMC-only monkeys, and the middle and right columns show all three training stages for monkeys D and H respectively. **D** The population mean effective spike history filters at two points in the task given the rank-1 dynamic spike history for the four fully trained DMC populations (mean of MAP estimate of filters ± 2 SEM). (top) The population mean spike history at 100 ms after sample stimulus onset. (bottom) The population mean spike history during the delay period (500 ms after sample stimulus offset). Positive weights indicate that a previous spike at the given lag increases a neuron's probability of firing, while negative weights indicate that spiking is suppressed.

(Fig. 8D bottom). Spike history differences between the populations were less evident during stimulus presentation. These results were consistent for higher-rank dynamic spike history tensors (Fig. S9). Thus, the structure of single-trial variability during the delay period differed between DMC-only and pretrained monkeys, but was similar within each pair, which suggests that the balance of beta- and gamma-frequency driven activity during the delay period differed between the animal pairs.

3 Discussion

Here we examined the low-dimensional geometry of task-related responses during a motion-direction categorization task in LIP in two pairs of monkeys performing the same motion categorization task, but with different training histories. In the monkeys that were pretrained on a motion-discrimination DMS task, we found similar direction-dependent activity in LIP activity during the DMS and DMC task: two-dimensional direction-encoding subspaces that reflected the stimulus geometry. Moreover, uniform direction tuning remained a dominant feature of this subspace after training on the DMC task. The common direction tuning observed across the sample and test stimuli could reflect cosine-like signals from sensory regions such as the area MT (Born & Bradley, 2005; Jazayeri & Movshon, 2006; Fanini & Assad, 2009). In contrast, the monkeys trained only on the categorization task showed stronger category tuning and category-aligned direction tuning in LIP compared to activity in animals first trained on the DMS task. Performing the categorization task may involve computations including input selection and local and/or top-down recurrent dynamics. Our findings indicate that differences in the sequences of tasks learned by the animals over long periods may result in different network configurations that perform the same task, perhaps manifesting in different behavioral strategies.

We hypothesize that these differences may be indicative of the pretrained monkeys still using computational strategies learned for the DMS task. Indeed, the pretrained monkeys' behavior showed greater dependence on the angular difference between sample and test stimuli than DMC-only monkeys, which was a key factor in solving the DMS task. Because the tasks used the same stimuli and shared many of the same correct or incorrect sample-test pairings, the same neural machinery and behavioral strategies could be recruited and maintained for the DMC task, despite extensive retraining. While many LIP neurons show delay period encoding of category during the DMC task, we did not see direction tuning in the average firing rates during the delay period in the DMC task (Fig. 1C, Fig. 5). It is possible that direction is maintained in working memory in LIP populations during the delay period by sparse bursting activity, but not by persistent firing, which cannot be seen by our analysis using single-neuron recordings (Miller et al., 2018). Additionally, previous theoretical work from our lab has demonstrated that recurrent neural networks performing the DMS task may recruit activity-silent computations to compare sample and test stimuli through short-term synaptic plasticity (Masse et al., 2019). In that study using recurrent neural networks trained on both DMS and DMC tasks, delay-period sustained activity was observed more often in tasks which required more complex manipulation of the sample stimulus information compared to the DMS. Our dynamic spike history analysis revealed single-trial dynamics with low-beta frequency os-

cillatory structure during the delay period in the DMC-only monkeys, but not the pretrained monkeys, which could reflect different working memory dynamics in the DMC-only pair (Miller et al., 2018). Together, this raises the possibility that computations learned during the DMS task which recruited activity-silent working memory during the delay period could explain the observed reduced separability of the sample and test stimulus in the neural subspace in the pretrained monkeys compared to the DMC-only monkeys (Orhan & Ma, 2019).

We also extended the GLM framework to perform dimensionality reduction on neural populations using a flexible tensor-regression model. In complex decision-making tasks, the trials may not be aligned such that common dimensionality reduction methods can be applied without artificially re-aligning single-trial firing rates by stretching or time-warping (Kobak et al., 2016). For example, the touch-bar release ended the trial early at a time determined by the animal in the DMS and DMC tasks. We applied the GMLM to perform dimensionality reduction to find task-relevant features in spike trains when the events in the task were not exactly aligned on every trial, without the need for aligned trial structure. Our approach is related to reduced-rank regression (Steinmetz et al., 2019; Stringer et al., 2019) and the recently proposed model-based targeted dimensionality reduction (Aoi et al., 2020) with two important distinctions: (1) our model is fit to spike trains through an autoregressive Poisson observation model and (2) we consider a more general tensor decomposition of task-related dynamics. The tensor decomposition is used to describe low-rank temporal dynamics in response to stimulus events, similar to low-rank receptive field models of early visual neurons (Ahrens et al., 2008; Park & Pillow, 2013; Elsayed et al., 2020), and those components are shared across all neurons in a population. Unlike PSTH-based dimensionality reduction methods, the Poisson spike generation mechanism accounts for discrete spiking observations and aspects of single-trial dynamics through the spike history kernel and tensor-based dynamic spike history (Park et al., 2014; Holbrook et al., 2017). In contrast to demixed principal components analysis (Kobak et al., 2016) which requires balanced conditions across all variables to recover task-relevant subspaces, the cosine-tuned GMLM takes into account cosine-like direction tuning observed in sensory regions in order to disentangle category and direction information even though direction and category are not separated in the task. Bayesian inference in this model allowed us to quantify uncertainty in the low-dimensional subspace and test hypotheses about the geometry of neural representations. This modeling framework could extend to many other tasks and questions, given appropriate linearizations of specific tasks. For instance, the tensor could be extended to model slow trial-to-trial changes in stimulus response within a recording session by including coefficients for weighting each trial, thereby generalizing applications of tensor component analysis as in Williams et al. (2018).

There are several important limitations about the inferred behavioral strategy and neural mechanisms in the present study. Primarily, this study included only a small number of animals, as is the norm in non-human primate experiments. Furthermore, multiple cortical and subcortical areas are involved in decision making, and our analysis only considered neural activity in LIP. Even within a single region, it is possible that our results could depend on differences in sampling within LIP between animals, or other factors not directly related to the animals' training history. LIP recording sessions were performed using different sets of motion directions for the two pairs (six directions for the DMC-only monkeys and eight for the pretrained pair). However, we do not believe this small difference in sample directions contributed to the observed differ-

ences in LIP because the monkeys were trained using many more motion directions (Sarma et al., 2016). We cannot exclude the possibility that animals could switch behavioral strategy with additional training such that, for example, both pairs of monkeys would perform similarly. Consequently, the possibility remains that LIP representations of the DMC task could change to match the currently adopted behavioral strategy, rather than purely reflecting training history. We think this is unlikely because the LIP recordings were made after all animals had received extensive training on the DMC task and their behavioral accuracy had appeared to asymptote at a high-level of accuracy (Sarma et al., 2016).

There are multiple ways that the brain could learn to perform the same task. Average results across animals may therefore fail to reflect the neural mechanisms of decision making in individual animals (Golowasch et al., 2002; Rahnev & Denison, 2018). Individual differences are a major focus of human decision-making research and have led to many insights into cognitive functions including working memory (Vogel & Awh, 2008; Luck & Vogel, 2013). Here, we explored between subject differences in the dimensionality and the relationship between direction and category tuning in LIP, and we found differences that correlated with long-term training history. Primates in particular may participate in many experiments and receive extensive training in multiple closely related tasks over the course of years. Experimenters should report and consider animals' training histories when interpreting such data and when comparing seemingly conflicting results from different labs. In conclusion, the low-dimensional dynamics that posterior parietal cortex enlists to support abstract visual categorization can manifest differently across subjects, and exploring long-term effects of training over more subjects can provide broader perspectives of the diverse neural computations that give rise to decision making skills.

4 Methods

4.1 Data

All datasets used for this study were previously published in Swaminathan & Freedman (2012) and Sarma et al. (2016).

4.1.1 Tasks

The details of the tasks have been described previously for monkeys B and J (DMC-only monkeys; Swaminathan & Freedman, 2012; Swaminathan et al., 2013) and monkeys D and H (pretrained monkeys; Sarma et al., 2016). For all animals, stimuli were high-contrast, 100 % coherent random dot motion stimuli with a dot velocity 12 °s^{-1}. The motion patch was 9.0° diameter, and the frame rate was 75 frames/s. Monkeys were required to keep fixation within a 2° radius of the fixation point during each trial.

For monkeys D and H, there were eight sample stimulus directions for both the DMS and DMC tasks, spaced 45° apart: {22.5°, 67.5°, 112.5°, 157.5°, 202.5°, 247.5°, 292.5° and 337.5°}.

The test stimuli for the DMS task were 45°, 60°, 75°, or 0° (match) away from the sample stimulus, giving a total of 24 possible motion directions in the task. Test stimuli for the DMC task were the same as the eight sample stimuli. The stimulus presentations were 667 ms and the delay period was 1013 ms.

The DMC task for monkeys B and J used directions spaced evenly in 60° intervals: {15°, 75°, 135°, 195°, 255° and 315°}. The stimulus presentations were 650 ms and the delay period was 1000 ms.

For all monkeys on the DMC task, the motion directions were split evenly into two categories separated by a constant boundary at 45° and 225°.

The DMC task for monkeys B and J included a set of null-direction trials (Fig. S6A). In these trials, the sample direction was along the category boundary (45° or 225°) and the test direction was either 135° or 315° (one direction from each category, furthest from the boundary). These trials were not used to fit neural models, but examined for behavior in Fig. S6B. The monkey's response was randomly rewarded at 50 % chance on these trials. We note that monkeys B and J were first trained on a simplified DMS task where the sample and test stimuli were either match or 180° opposite. This version of DMS task therefore did not require fine motion direction discrimination, and all correct sample-test response pairs in this task matched were consistent with the DMC task.

4.1.2 Electrophysiology

Neurons in LIP were recorded using single tungsten microelectrodes. During both the DMS and DMC tasks, the motion stimuli were placed inside an LIP cell's response field.

In this study, we included only cells with a mean firing rate of at least 2 sp/s, averaged from sample stimulus onset to test offset. We included $N = 31$ cells from 26 sessions for monkey b, and $N = 29$ from 27 sessions for monkey J. For monkey D, $N = 81$ cells from 39 sessions for the DMS task, $N = 63$ cells from 33 sessions for the DMC early period, and $N = 137$ cells from 59 sessions for the DMC late period. For monkey H, $N = 89$ cells from 55 sessions for the DMS task, $N = 106$ cells from 40 sessions for the DMC early period, and $N = 114$ cells from 50 sessions for the DMC late period.

4.1.3 Data used for modeling

For all the modeling and decoding analyses, we included only correct trials. The null-direction trials for DMC task for monkeys B and J were not included for model fitting.

We considered a window in each trial starting from sample stimulus onset until 50 ms after the touch-bar release (if a touch-bar release occurred) or 50 ms after the test motion offset. We discretized the spike trains during each trial into 5 ms bins. We note that on non-match trials the animal was required to hold the touch bar until a second test stimulus (which is always a match) appeared. However, the second test stimulus presentation was never included in our analysis.

We used 10 fold cross-validation to compare the models. The trials from each cell were divided into folds evenly by sample directions. For example, if there were 40 trials recorded with sample motion of 22.5° for one cell, these trials were divided into groups of four to make the folds.

For plotting the PSTHs in Fig. 1C, Fig. 3D, and Fig. S8, we smoothed the average spike rate over trials conditioned on motion direction using a Gaussian kernel with a 30 ms width.

4.1.4 Behavioral performance

For quantifying behavioral performance, we only analyzed behavior during the LIP recording sessions. In the behavioral analyses in Fig. 7A and Fig. S6B, we estimated the fraction correct (or fraction touch-bar releases) independently in each condition using a beta-binomial model. The prior parameters in the model were $\alpha = 1$ and $\beta = 1$ (for a beta distribution over the prior fraction correct or touch-bar released trials). In this model, the posterior over the fraction correct (or touch-bar released) is a beta distribution. The point estimate of the fraction correct was the posterior mean, and the error bars denote 99 % credible intervals over the posterior.

4.2 GLM for single cells (the full-rank model)

In this section, we define the generalized linear point-process model of single cells during the DMS and DMC tasks. This class of model for single neurons in decision-making tasks is defined in general in Park et al. (2014). The GLM defines the distribution of spike count at time t as a Poisson random variable with mean rate given as a linear function of external events (here, stimulus and touch-bar release) and previous spiking activity:

$$\lambda(t) = w + \left(\mathbf{h}^{\mathrm{spk}} * \mathbf{y}\right)(t) + \left(\mathbf{k}^{\mathrm{tbar}} * x^{\mathrm{tbar}}\right)(t) + \sum_{s \in \mathcal{S}} \left(\mathbf{k}^{(s)} * x^{(s)}\right)(t) \qquad \text{(log firing rate at time } t\text{)} \qquad (1)$$

$$\mathbf{y}(t) \sim \mathrm{Poisson}(f(\lambda(t))\Delta) \qquad \text{(spike count for bin } t\text{)} \qquad (2)$$

$$f(\cdot) = \exp(\cdot) \qquad \text{(inverse link function)} \qquad (3)$$

The $*$ operator denotes convolution. The bin width is Δ, and the log baseline firing rate parameter is w. Recent spiking activity, $\mathbf{y}$, affects the rate through the spike history kernel, $\mathbf{h}^{\mathrm{spk}}$.

The stimulus event regressors $x^{(s)}(t)$ are functions of at time representing information about the motion stimulus events. The set of all stimulus events is $\mathcal{S}$. The touch-bar event is x^{tbar}. The linear temporal kernels $\mathbf{k}^{(s)}$ and $\mathbf{k}^{\mathrm{tbar}}$ describe the cell's response to each external variable (stimulus or touch-bar, respectively) as a function of time. The stimulus events we consider encapsulate both sample and test stimuli, but the configuration and number of stimulus kernels depends on the specific model parameterization.

We parameterized the temporal kernels using raised cosine basis functions (Pillow et al., 2008). The stimulus kernel basis consisted of $P_{\mathrm{spk}} = 24$ functions with a nonlinear stretching parameter of 0.2 and peaks spanning 0 ms to 1500 ms (Fig. S1J, left). We aligned the

basis so that the first basis was zero at exact time of stimulus onset, giving peaks between 40 ms to 1540 ms relative to stimulus onset. The touch-bar basis was constructed using the first $P_{\text{tbar}} = 8$ functions of the stimulus basis (Fig. S1J, middle. The functions were reversed and shifted the basis so that the function peaks ranged from –235 ms to 25 ms relative to the touch-bar release and the fastest temporal resolution of the basis set was near the touch-bar release time. We used $P_{\text{stim}} = 24$ basis functions for the spike history (Fig. S1J, right). The first two basis functions were Kronecker delta functions to account for the first two bins (0 to 5 ms and 6 to 10 ms after a spike). The remaining eight functions were a raised cosine basis set with nonlinear stretching parameter of 0.05 and peaks from 10 ms to 20 ms post spike.

We define the kernels as the bases times a set of coefficients:

$$\begin{aligned} \mathbf{k}^{(s)} &= \mathbf{B}^{\text{stim}}\tilde{\mathbf{k}}^{(s)}, && \text{for } s \in \mathcal{S} \text{ where each } \tilde{\mathbf{k}}^{(s)} \text{ is a vector of length } P_{\text{stim}} \\ \mathbf{k}^{\text{tbar}} &= \mathbf{B}^{\text{tbar}}\tilde{\mathbf{k}}^{\text{tbar}}, && \text{where } \tilde{\mathbf{k}}^{\text{tbar}} \text{ is a vector of length } P_{\text{tbar}} \\ \mathbf{h}^{\text{spk}} &= \mathbf{B}^{\text{spk}}\tilde{\mathbf{h}}^{\text{spk}}, && \text{where } \tilde{\mathbf{h}}^{\text{spk}} \text{ is a vector of length } P_{\text{spk}}. \end{aligned} \tag{4}$$

The parameters that are fit to data are $\phi = \{w, \tilde{\mathbf{h}}^{\text{spk}}, \tilde{\mathbf{k}}^{\text{tbar}}, \tilde{\mathbf{k}}^{(s)} | s \in \mathcal{S}\}$. This choice of basis ensures that the stimulus kernels are causal: the stimulus filters only contribute to firing rate after stimulus onset. In contrast, the touch-bar release is acausal: touch-bar release can contribute to the spike rate before the behavior to reflect buildup to the match decision.

We linearized the task events as point events in time. The touch-bar release is given as

$$x^{(\text{tbar})}(t) = \begin{cases} 1 & \text{if } t = t_{\text{tbar}} \\ 0 & \text{otherwise.} \end{cases}.$$

where the time the monkey released the touch-bar to signal a match response is t_{tbar} (if the touch-bar was release in the trial). We similarly consider the stimulus onsets (both sample and test) as point events. The sample stimulus duration is constant across all trials, and although the test stimulus is terminated early by a touch-bar release, this does not factor into the window of the trial we model. However, the model can extend to tasks with variable stimulus duration, as has been shown previously (Park et al., 2014). Each GLM kernel gives a scalar contribution to firing rate of the relative time of the task event

We considered a set of four nested models of increasing complexity for the motion stimuli. For simplicity of notation, we present the linearization for a single trial. The sample stimulus onset time is $t_{\text{sample on}}$ and the test stimulus onset is $t_{\text{test on}}$. The sample and test stimulus directions are θ_{sample} and θ_{test} for $\text{sample}, \text{test} \in \{1, 2, \ldots, D\}$ where D is the total number of stimulus directions in the task. The stimulus directions belong to categories denoted $c_{\text{sample}}, c_{\text{test}} \in \{1, 2\}$.

1. **No category or direction model** (Fig. S1A). This model includes only two stimulus regressors/kernels: one for the sample stimulus onset and one for the test stimulus onset (Fig. S1A): $\mathcal{S} = \{x^{(\text{sample})}, x^{(\text{test})}\}$. The regressors are defined as point events

$$x^{(\text{sample})}(t) = \begin{cases} 1 & \text{if } t = t_{\text{sample on}}, \\ 0 & \text{otherwise.} \end{cases}$$

$$x^{(\text{test})}(t) = \begin{cases} 1 & \text{if } t = t_{\text{test on}}, \\ 0 & \text{otherwise.} \end{cases}$$

This model captures temporal dynamics in the mean response for each neuron across all stimuli.

2. **Category only model** (Fig. S1B). This model includes four stimulus kernels: two for each category for the sample stimulus ($x^{(\text{cs1})}$ and $x^{(\text{cs2})}$), and two separate kernels for the test stimulus categories ($x^{(\text{ct1})}$ and $x^{(\text{ct2})}$). The regressors are again point events, but the points are now conditioned on stimulus category (but not specific direction). For each category $k \in \{1, 2\}$:

$$x^{(\text{csk})}(t) = \begin{cases} 1 & \text{if } t = t_{\text{sample on}} \text{ and } c_{\text{sample}} = c_k, \\ 0 & \text{otherwise,} \end{cases}$$
$$x^{(\text{ctk})}(t) = \begin{cases} 1 & \text{if } t = t_{\text{test on}} \text{ and } c_{\text{test}} = c_k \\ 0 & \text{otherwise.} \end{cases}$$

Although the DMS task does not include category, we still applied this model to those data as if there was a category boundary at 45° and 225°.

3. **Cosine direction tuning model** (Fig. S1C). The cosine-tuned model includes both stimulus category and direction tuning, but direction encoding is constrained to a parametric form with cosine tuning. The model includes six stimulus events: two for each category for the sample stimulus ($x^{(\text{cs1})}$ and $x^{(\text{cs2})}$), two for the test stimulus categories ($x^{(\text{ct1})}$ and $x^{(\text{ct2})}$), and two for the sine and cosine of the direction ($x^{(\text{sin})}$ and $x^{(\text{sin})}$). The sample and test category events are defined as in the previous model. The direction regressors are weighted point events, which are shared for both sample and test stimuli:

$$x^{(\cos)}(t) = \begin{cases} \cos(\theta_{\text{sample}}) & \text{if } t = t_{\text{sample on}} \\ \cos(\theta_{\text{test}}) & \text{if } t = t_{\text{test on}} \\ 0 & \text{otherwise.} \end{cases}$$
$$x^{(\text{sin})}(t) = \begin{cases} \sin(\theta_{\text{sample}}) & \text{if } t = t_{\text{sample on}} \\ \sin(\theta_{\text{test}}) & \text{if } t = t_{\text{test on}} \\ 0 & \text{otherwise.} \end{cases}$$

4. **Full model** (Fig. S1D). The full model allows for general (non-cosine) direction tuning. However, we constrained the model to have the same direction tuning for both sample and test stimuli; that is, the difference in tuning between two directions within the same category was the same for both sample and test stimuli. The model included one stimulus regressor event for each directions plus two test category events (for the DMC task with $D = 6$ stimulus directions, there are eight kernels) For each trial and direction θ_d for $d \in \{1, 2, \ldots, D\}$, the stimulus regressors are

$$x^{(\theta_d)}(t) = \begin{cases} 1 & \text{if } t = t_{\text{sample on}} \text{ and } \theta_{\text{sample}} = \theta_d \\ 1 & \text{if } t = t_{\text{test on}} \text{ and } \theta_{\text{test}} = \theta_d \\ 0 & \text{otherwise.} \end{cases}$$

The two stimulus events that parameterize the test category responses ($x^{(\mathrm{ct1})}$, and $x^{(\mathrm{ct2})}$) are defined as before. This parameterization maintains identifiability: it would not be identifiable to directly expand the cosine model to have a kernel for each direction plus two sample and two test category kernels. As a result of the identifiability constraint, the interpretation of the corresponding kernels is different compared to the cosine tuning model: in this model, $\mathbf{k}^{(\theta_d)}$ is the kernel for a stimulus in the θ_d direction plus the response to a stimulus of the category of c_d. Therefore, we view $\mathbf{k}^{(\mathrm{ct1})}$ as the kernel to a test stimuli of category 1 minus the kernel for a category 1 sample stimulus (thereby subtracting the sample category tuning away from the direction kernel and adding the test category tuning).

We considered two additional models included in supplementary analyses that included independent sample and test direction tuning (Fig. S6). The independent direction cosine tuning model had eight kernels total: four for the sample and test category, four for the cosine and sine weights of the sample and test directions. The full independent direction model simply had one kernel for each sample direction and one kernel for each test direction These two models are defined analogously to the common direction tuning models.

4.2.1 Prior probabilities over model parameters

We defined zero-mean Gaussian priors over the stimulus kernels. The orthonormal basis functions controlled temporal smoothness of the kernels, and the prior distributions were independent over time. We defined i.i.d. priors for the ith coefficients of the stimulus kernels (i.e., a prior over the set of $\{\tilde{\mathbf{k}}_i^{(\cdot)}\}$ for each $i \in \{1, \ldots, P_{\mathrm{spk}}\}$). We describe the priors of the stimulus kernels for each of the nested models.

1. **No category or direction model**. We considered that the sample and test kernels would likely be correlated if they reflect the dynamics of common bottom-up sensory input. To construct a correlated prior, we assumed that the kernel could be constructed as the sum of stimulus-independent kernel (a response purely to contrast or motion in general), and a sample kernel or a test kernel (responses to the task epoch):

$$\alpha_{i,0} \sim \mathcal{N}(0, \psi_0^2), \qquad \alpha_{i,\mathrm{sample}} \sim \mathcal{N}(0, \psi_{\mathrm{s}}^2), \qquad \alpha_{i,\mathrm{test}} \sim \mathcal{N}(0, \psi_{\mathrm{s}}^2)$$

   such that

$$\tilde{\mathbf{k}}_i^{(\mathrm{sample})} = \alpha_{i,0} + \alpha_{i,\mathrm{sample}}, \qquad \tilde{\mathbf{k}}_i^{(\mathrm{test})} = \alpha_{i,0} + \alpha_{i,\mathrm{test}}.$$

   Using the rules of linear transformations of Gaussian variables, we obtained a correlated Gaussian prior over the original two sample and test kernels. The set of hyperparameters was $\mathcal{H}_{\mathrm{stim}} = \{\psi_0, \psi_{\mathrm{s}}\}$.

2. **Category only model**. For the category-dependent kernels, we made a similar Gaussian construction

$$\beta_{i,0} \sim \mathcal{N}(0, \psi_0^2), \qquad \beta_{i,\mathrm{csk}} \sim \mathcal{N}(0, \psi_{\mathrm{c}}^2), \qquad \beta_{i,\mathrm{ctk}} \sim \mathcal{N}(0, \psi_{\mathrm{c}}^2)$$

such that

$$\tilde{\mathbf{k}}_i^{(\mathrm{csk})} = \beta_{i,0} + \beta_{i,\mathrm{csk}}, \qquad \tilde{\mathbf{k}}_i^{(\mathrm{ctk})} = \beta_{i,0} + \beta_{i,\mathrm{ctk}}.$$

The set of hyperparameters was $\mathcal{H}_{\mathrm{stim}} = \{\psi_0, \psi_{\mathrm{c}}\}$.

3. **Cosine direction tuning model**. We used the same priors for the four category kernels as in the category-only model. We placed an independent Gaussian prior over the cosine and sine weights:

$$\tilde{\mathbf{k}}_i^{(\cos)} \sim \mathcal{N}(0, \psi_{\mathrm{d}}^2), \qquad \tilde{\mathbf{k}}_i^{(\sin)} \sim \mathcal{N}(0, \psi_{\mathrm{d}}^2).$$

The set of hyperparameters was $\mathcal{H}_{\mathrm{stim}} = \{\psi_0, \psi_{\mathrm{d}}, \psi_{\mathrm{c}}\}$.

4. **Full model**. For constructing the prior over individual direction kernels, we assumed that direction tuning should be smooth as a function of angle. We therefore used a Gaussian process prior over the direction weights. To nest the cosine-tuning model in the full model and provide regularization, we also included latent sine and cosine direction weighting. Sample category weights were included as before. We define the pieces of the prior as

$$\begin{aligned} \gamma_{i,0} &\sim \mathcal{N}(0, \psi_0^2), & \gamma_{i,\mathrm{csk}} &\sim \mathcal{N}(0, \psi_{\mathrm{c}}^2), \\ \gamma_{i,\cos} &\sim \mathcal{N}(0, \psi_{\mathrm{d}}^2), & \gamma_{i,\sin} &\sim \mathcal{N}(0, \psi_{\mathrm{d}}^2), \\ \gamma_{i,gp}(\theta) &\sim \mathcal{GP}\left(0, \psi_\theta^2 K\left(\theta, \theta'\right)\right), \end{aligned}$$

where the Gaussian process kernel over angle is (Padonou & Roustant, 2016)

$$\begin{aligned} K\left(\theta, \theta'\right) &= \left(1 + \frac{\tau + 4}{\pi} d(\theta, \theta')\right)\left(1 - \frac{1}{\pi} d(\theta, \theta')\right)^{\tau+4}, \\ d(\theta, \theta') &= \arccos(\cos(\theta - \theta')). \end{aligned}$$

The hyperparameter $\tau \geq 0$ determined the arc length over which similar directions are correlated, similar to a length scale in Gaussian process kernels on the real line. The complete direction plus sample category kernel was then defined as for each direction $d \in \{1, \ldots, D\}$

$$\tilde{\mathbf{k}}_i^{(\theta_d)} = \gamma_{i,0} + \gamma_{i,\mathrm{csk}} + \gamma_{i,\cos}\cos(\theta_d) + \gamma_{i,\sin}\sin(\theta_d) + \gamma_{i,gp}(\theta_d).$$

The test category prior was defined slightly differently than in the previous two models due to the identifiability constraints on our parameterization. We defined the prior using the construction

$$\begin{aligned} \gamma_{i,\mathrm{ctk}} &\sim \mathcal{N}(0, \psi_{\mathrm{c}}^2), \\ \tilde{\mathbf{k}}_i^{(\mathrm{ctk})} &= \gamma_{i,0} + \gamma_{i,\mathrm{ctk}} - \gamma_{i,\mathrm{csk}}. \end{aligned}$$

Because $\tilde{\mathbf{k}}_{\cdot}^{(\theta_d)}$ and $\tilde{\mathbf{k}}^{(\mathrm{ctk})}$ were again simply linear functions of Gaussian variables, we obtained a Gaussian prior with zero mean for the kernels depending on the hyperparameters set $\mathcal{H}_{\mathrm{stim}} = \{\psi_0, \psi_{\mathrm{d}}, \psi_{\mathrm{c}}, \psi_\theta, \tau\}$

For the supplementary models with independent sample and test direction tuning, the priors followed the same construction as above. The direction hyperparameters were shared for the sample and test direction kernels.

We placed an i.i.d. Gaussian prior on the spike history and touch-bar coefficients

$$\tilde{\mathbf{k}}_i^{\text{tbar}} \sim \mathcal{N}(0, \psi_{\text{tbar}}) \qquad \text{for } i \in \{1, \ldots, P_{\text{tbar}}\}$$
$$\tilde{\mathbf{h}}_j^{\text{spk}} \sim \mathcal{N}(0, \psi_{\text{spk}}) \qquad \text{for } j \in \{1, \ldots, P_{\text{spk}}\}$$

The prior over w was the improper uniform prior: $p(w) \propto 1$.

The complete set of hyperparameters was therefore $\mathcal{H} = \{\mathcal{H}_{\text{stim}}, \psi_{\text{tbar}}, \psi_{\text{spk}}\}$. We defined hyperpriors over each hyperparameter independently as half-t distributions (Gelman, 2006). For each $h \in \mathcal{H}$

$$p(h) \propto \left(1 + \frac{1}{\nu} h\right)^{-(\nu+1)/2} \tag{5}$$

where we set $\nu = 4$.

4.2.2 MAP estimation with evidence optimization

We fit the GLMs to each LIP cell using MAP estimation. To set the hyperparameters for the GLM, we used an approximate evidence optimization procedure (Sahani & Linden, 2003; Park & Pillow, 2013; Zoltowski & Pillow, 2018). We used a Laplace approximation of the posterior over model parameters to get likelihood of data given hyperparameters to estimate the log evidence (i.e., the marginal distribution of the data given the hyperparameters). We then optimized the log posterior over the hyperparameters. Because the hyperparameters are constrained to be positive, we optimized the log-transformed hyperparameters. We set the hyperparameters and parameters of the GLM independently for each fold of cross-validation.

Specifically, we maximized an approximation of the log posterior of the hyperparameters given the data. The posterior is

$$p(\mathcal{H}|\mathbf{y}, \mathbf{x}) \propto p(\mathbf{y}|\mathcal{H}, \mathbf{x}) p(\mathcal{H}) \tag{6}$$

and we want to find

$$\mathcal{H}_{MAP} = \arg\max_{\mathcal{H}} p(\mathcal{H}|\mathbf{y}, \mathbf{x}).$$

The evidence term can be written using Bayes' rule as

$$p(\mathbf{y}|\mathcal{H}, \mathbf{x}) = \frac{p(\mathbf{y}|\phi, \mathbf{x}) p(\phi|\mathcal{H})}{p(\phi|\mathcal{H}, \mathbf{y}, \mathbf{x})} \tag{7}$$

where ϕ denotes the model parameters. The posterior over the parameters ($p(\phi|\mathcal{H}, \mathbf{y}, \mathbf{x})$) is only given up to an intractable normalizing constant. We therefore took a Laplace approximation of the posterior distribution over parameters. The Laplace approximation was a Gaussian distribution centered around the MAP estimate of the parameters

$$\begin{aligned} p(\phi|\mathcal{H}, \mathbf{y}, \mathbf{x}) &\approx \mathcal{N}\left(\phi; \phi_{MAP}, \Sigma_{MAP}\right), \\ \phi_{MAP} &= \arg\max_{\phi} p(\mathbf{y}|\phi, \mathbf{x}) p(\phi|\mathcal{H}) \\ \Sigma_{MAP}^{-1} &= -\frac{d^2}{d\phi^2} \log p(\phi|\mathcal{H}, \mathbf{y}, \mathbf{x}) \Big|_{\phi=\phi_{MAP}}. \end{aligned} \tag{8}$$

The MAP estimate given the hyperparameters, ϕ_{MAP}, was found numerically (the log posterior over the parameters is log concave). Given this approximation, we evaluated the right side of Eq. 7 at ϕ_{MAP}. We then maximized the log posterior over the hyperparameters (Eq. 6) numerically to find $\mathcal{H}_{MAP}$. The final MAP estimate of the models parameters was ϕ_{MAP} given $\mathcal{H}_{MAP}$.

4.3 GMLM definition

The GMLM is a special case of the GLM in which the linear kernels in a population of neurons are assumed to share low-dimensional structure, rather than being modeled independently. In general, the model is a GLM in which the regressors and parameters (or a subset thereof) from all the neurons in a population can be expressed as tensors (or multi-way arrays). The parameters (or a subset of the parameters) are then assumed to have a low-rank structure: the parameter tensor can be decomposed into a small number of components. We emphasize that the neurons need not be simultaneously recorded for this model: we can still fit the parameters if only one neuron is observed at each time point.

Here, we define the GMLM for the DMC task. Introducing an index for neuron $n \in \{1, 2, \ldots, N\}$, we defined the model for the spike count in bin $t \in \{1, 2, \ldots, T\}$ for neuron n as

$$\begin{aligned} \lambda_n(t) = \mathbf{w}_n &+ \left(\mathbf{H}_n^{\text{spk}} * \mathbf{Y}_n\right)(t) + \sum_{q=1}^{R_l} \left(\mathbf{T}_q^{\text{tbar}} * x^{\text{tbar}}\right)(t) \mathbf{V}_{n,q}^{\text{tbar}} \\ &+ \sum_{s \in \{\text{sample,test}\}} \sum_{r=1}^{R_s} \mathbf{Z}_r^{(s)}(t) \mathbf{V}_{n,r}^{\text{stim}} \end{aligned}, \tag{9}$$

$$\mathbf{Z}_r^{(s)}(t) = \left(\mathbf{x}^{(\text{direction},s)} \cdot \mathbf{U}_r^{\text{stim}}\right) \cdot \left(\mathbf{T}_r^{\text{stim}} * x^{(\text{timing},s)}\right)(t), \tag{10}$$

$$Y_n(t) \sim \text{Poisson}(f(\lambda_n(t))\Delta).$$

The matrices $\mathbf{H}_n^{\text{spk}}$, $\mathbf{T}^{\text{tbar}}$, and $\mathbf{T}^{\text{stim}}$ denote matrices whose columns are temporal kernels for the stimulus, touch-bar and spike history respectively. Subscripts of those matrices indicate a particular column or kernel. Similarly, $\mathbf{H}^{\text{spk}}$ contains the spike history kernels for each cell. The length N vector $\mathbf{w}$ contains the baseline firing rates for each neuron. The baseline firing rates and spike history kernels are equivalent to the single-cell GLM.

We note that in this model, both the regressors and the parameters are decomposed into components (for the stimulus parameters $\mathbf{T}^{\text{stim}}$, $\mathbf{U}^{\text{stim}}$, and $\mathbf{V}^{\text{stim}}$ and regressors $\mathbf{x}^{(\text{direction},s)}$ and $x^{(\text{timing},s)}$), and thus we have a simple multilinear form for the stimulus tuning rather than writing out a dense tensor.

The touch-bar kernels are parameterized as low-rank matrix factorization where $\mathbf{T}^{\text{tbar}}$ contains R_l temporal kernels and $\mathbf{V}^{\text{tbar}}$ is a matrix of neuron loading weights of size $N \times R_l$. In our notation, $\mathbf{T}^{\text{tbar}}_q$ denotes the qth column or kernel, and $\mathbf{V}^{\text{tbar}}_{n,q}$ is the element in the nth row, qth column of $\mathbf{V}^{\text{tbar}}$. The touch-bar subspace is the span of the columns of $\mathbf{V}^{\text{tbar}}$. Thus, the model effectively approximates the GLM touch-bar kernel for neuron n as $\mathbf{k}^{\text{tbar}}_n \approx \mathbf{T}^{\text{tbar}}\mathbf{V}^{\text{tbar}\top}_{n,\cdot}$. The touch-bar function, x^{tbar}, is the same as in the GLM.

The stimulus kernels were parameterized as a tensor factorization of rank R_s. As we did for the GLM, we defined a matching set of nested models to parameterize the DMC task. As with the GLM definitions, we defined the regressors for a single trial for simplicity of notation. The notation for the stimulus timing and directions are the same as in the GLM definition.

The set of temporal kernels, $\mathbf{T}^{\text{stim}}$, did not depend on the stimulus direction or category. The temporal regressors were point events representing the stimulus onset time for each $s \in \{\text{sample}, \text{test}\}$. These were the same for all GMLM parameterizations (Fig. S1E):

$$x^{(\text{timing,sample})}(t) = \begin{cases} 1 & \text{if } t = t_{\text{sample on}} \\ 0 & \text{otherwise.} \end{cases},$$

$$x^{(\text{timing,test})}(t) = \begin{cases} 1 & \text{if } t = t_{\text{test on}} \\ 0 & \text{otherwise.} \end{cases}.$$

The set of stimulus weights, $\mathbf{U}^{\text{stim}}$, was a matrix $S \times R_s$ of coefficients for the particular stimulus identity (for example, weights to encode sample, test, direction, and category) where S is the same as the number of stimulus kernels in the matching GLM. Each observation had two stimulus direction regressor vectors: $\mathbf{x}^{(\text{direction,sample})}$ and $\mathbf{x}^{(\text{direction,test})}$. The entries of the stimulus direction regressors ($\mathbf{x}^{(\text{direction},s)}$) mirrored the kernel structure in the GLM parameterizations (this vector is constant for all t in a single trial). The stimulus direction coefficients depended on the model.

1. **No category or direction model** (Fig. S1F This model contained two stimulus regressor elements indexed by $\{\text{sample}, \text{test}\}$, As with the GLM, these elements represent the identity of a stimulus event as sample or test, but does not include category or direction information.

$$\mathbf{x}^{(\text{direction,sample})}_{\text{sample}} = 1, \qquad \mathbf{x}^{(\text{direction,sample})}_{\text{test}} = 0,$$
$$\mathbf{x}^{(\text{direction,test})}_{\text{sample}} = 0, \qquad \mathbf{x}^{(\text{direction,test})}_{\text{test}} = 1.$$

2. **Category only model** (Fig. S1G) This model includes four stimulus direction regressors representing the stimulus category and whether it is sample or test. For the indices $\{\text{cs1}, \text{cs2}, \text{ct1}, \text{ct2}\}$, the regressors are

$$\mathbf{x}^{(\text{direction,sample})}_{\text{csk}} = \begin{cases} 1 & \text{if } c_{\text{sample}} = c_k, \\ 0 & \text{otherwise.} \end{cases} \qquad \mathbf{x}^{(\text{direction,sample})}_{\text{ctk}} = 0$$

$$\mathbf{x}_{\mathrm{ctk}}^{(\mathrm{direction,test})} = \begin{cases} 1 & \text{if } c_{\mathrm{test}} = c_k, \\ 0 & \text{otherwise.} \end{cases} \qquad \mathbf{x}_{\mathrm{csk}}^{(\mathrm{direction,test})} = 0,$$

for category $k \in \{1, 2\}$.

3. **Cosine direction tuning model** (Fig. S1H). The cosine-tuning model included six stimulus regressors representing the identity of a stimulus event as sample or test, the motion category, and the sine and cosine of the direction. The regressors are indexed by $\{\mathrm{cs1}, \mathrm{cs2}, \mathrm{ct1}, \mathrm{ct2}, \cos, \sin\}$. The category terms are the same as in the above model. The direction tuning components are defined as cosine and sine weights of the direction:

$$\mathbf{x}_{\cos}^{(\mathrm{direction,sample})} = \cos(\theta_{\mathrm{sample}}), \qquad \mathbf{x}_{\sin}^{(\mathrm{direction,sample})} = \sin(\theta_{\mathrm{sample}}),$$
$$\mathbf{x}_{\cos}^{(\mathrm{direction,test})} = \cos(\theta_{\mathrm{test}}), \qquad \mathbf{x}_{\sin}^{(\mathrm{direction,test})} = \sin(\theta_{\mathrm{test}}).$$

4. **Full model** (Fig. S1I). This model includes one regressor for each stimulus direction and two for the test stimulus category indexed by the $D+2$ coefficients in $\{\mathrm{ct1}, \mathrm{ct2}, \theta_1, \ldots, \theta_D\}$. For each $d \in \{1, \ldots, D\}$

$$\mathbf{x}_{\theta_d}^{(\mathrm{direction,sample})} = \begin{cases} 1 & \text{if } \theta_{\mathrm{sample}} = \theta_d \\ 0 & \text{otherwise.} \end{cases}$$
$$\mathbf{x}_{\theta_d}^{(\mathrm{direction,test})} = \begin{cases} 1 & \text{if } \theta_{\mathrm{test}} = \theta_d \\ 0 & \text{otherwise.} \end{cases}$$

The test category regressors (indexed by $\mathrm{ct1}$ and $\mathrm{ct2}$) are the same as in the previous two models. As with the full GLM, the specific model construction does not include additional weights for the sample category for identifiability. The coefficients in $\mathbf{U}^{\mathrm{stim}}(\theta_i)$ represent the tuning strength for direction θ_i plus the tuning for the category of θ_i. Therefore, the coefficients in $\mathbf{U}^{\mathrm{stim}}(\mathrm{ctk})$ represent the tuning for a test stimulus of category k minus the tuning for sample stimulus of category k.

Together, the matrices $\mathbf{T}^{\mathrm{stim}}$, $\mathbf{U}^{\mathrm{stim}}$, and $\mathbf{V}^{\mathrm{stim}}$ define a CP or PARAFAC decomposition of the GLM stimulus kernels over a population of cells (Kolda & Bader, 2009). That is, the sth stimulus kernel at time t for neuron n is approximated as the low-rank decomposition

$$\mathcal{K}(t, s, n) = \sum_{r=1}^{R_s} \mathbf{T}_r^{\mathrm{stim}}(t) \mathbf{U}_{s,r}^{\mathrm{stim}} \mathbf{V}_{n,r}^{\mathrm{stim}} \tag{11}$$

Another way to view the dimensionality reduction is that the values of $\mathbf{Z}_{\cdot}^{(s)}(t)$ give an R_s dimensional representation of the response to each stimulus over time. Each neuron's response to the stimulus is given as a linear projection of that low-dimensional stimulus with weights defined as the rows of the matrix $\mathbf{V}^{\mathrm{stim}}$ so that $\mathbf{V}^{\mathrm{stim}}$ defines the stimulus subspace.

The temporal kernels of the GMLM are parameterized using the same basis set as the GLM:

$$\mathbf{H}^{\mathrm{spk}} = \mathbf{B}^{\mathrm{spk}} \tilde{\mathbf{H}}^{\mathrm{spk}}, \qquad \text{where } \tilde{\mathbf{H}}^{\mathrm{spk}} \text{ is a matrix of size } P_{\mathrm{spk}} \times N, \tag{12}$$

$$\mathbf{T}^{\text{tbar}} = \mathbf{B}^{\text{tbar}}\tilde{\mathbf{T}}^{\text{tbar}}, \qquad \text{where } \tilde{\mathbf{T}}^{\text{tbar}} \text{ is a matrix of size } P_{\text{tbar}} \times R_l,$$
$$\mathbf{T}^{\text{stim}} = \mathbf{B}^{\text{stim}}\tilde{\mathbf{T}}^{\text{stim}}, \qquad \text{where } \tilde{\mathbf{T}}^{\text{stim}} \text{ is a matrix of size } P_{\text{stim}} \times R_s.$$

We set $R_l = 3$ for all GMLM fits and we selected R_s using cross-validation (see Section 4.3.3). The set of parameters that are fit to the data from all the trials in an LIP population is $\phi = \left\{\mathbf{w}, \mathbf{V}^{\text{stim}}, \mathbf{V}^{\text{tbar}}, \mathbf{U}^{\text{stim}}, \tilde{\mathbf{H}}^{\text{spk}}, \tilde{\mathbf{T}}^{\text{tbar}}, \tilde{\mathbf{T}}^{\text{stim}}\right\}$.

For Bayesian inference, we defined prior distributions the same way we did for the GLM. The prior for the stimulus kernels was defined independently for each component r of the stimulus direction regressor matrix (i.e., each column $\mathbf{U}^{\text{stim}}_{\cdot,r}$). The vectors $\mathbf{U}^{\text{stim}}_{\cdot,r}$ and corresponding GLM kernel parameters $\tilde{\mathbf{k}}_i^{(\cdot)}$ are the same length and are indexed by the same set of stimulus events The prior over the vector $\mathbf{U}^{\text{stim}}_{\cdot,r}$ was therefore the same as the prior over the $\tilde{\mathbf{k}}_i^{(\cdot)}$ for the corresponding GLM. The same stimulus hyperparameter set ($\mathcal{H}_{\text{stim}}$) was used for each model parameterization. However, unlike the single-cell GLM fits, the hyperparameters were shared across all neurons in each LIP population.

The prior for the entries of the spike history kernels, $\tilde{\mathbf{H}}^{\text{spk}}$, was i.i.d. normal with zero mean and variance ψ^2_{spk}. Similarly, the prior for the entries of the touch-bar kernels, $\tilde{\mathbf{T}}^{\text{tbar}}$, was i.i.d. normal with zero mean and variance ψ^2_{tbar}. The prior distribution on the entries of neuron loading matrices ($\mathbf{V}^{\text{stim}}$ and $\mathbf{V}^{\text{tbar}}$) and $\mathbf{T}^{\text{stim}}$ was i.i.d. standard normal. We again used an improper uniform prior on $\mathbf{w}$.

The complete hyperparameters set was $\mathcal{H} = \{\mathcal{H}_{\text{stim}}, \psi_{\text{tbar}}, \psi_{\text{spk}}\}$. The hyperpriors were the same half-t distributions used for the GLM (Eq. 5). We note that these priors only affected the GMLM in the MCMC analysis, as rank selection used the maximum likelihood estimate.

4.3.1 Dynamic spike history

We augmented the log rate in the GMLM with low-rank dynamic spike history components to allow spike history to change over time relative to task events:

$$h_n(t) = \sum_{q=1}^{R_{bh}} \mathbf{Z}_q^{(\text{bdspk})}(t)\mathbf{V}^{\text{bdspk}}_{n,q} + \sum_{s\in\{\text{sample,test}\}}\sum_{r=1}^{R_h} \mathbf{Z}_r^{(\text{dspk},s)}(t)\mathbf{H}_q^{\text{bdspk}} \tag{13}$$
$$\mathbf{Z}_q^{(\text{bdspk})}(t) = \left(\mathbf{T}_q^{\text{bdspk}} * x^{\text{tbar}}\right)(t) \cdot \left(\mathbf{H}_q^{\text{bdspk}} * \mathbf{Y}_n\right)(t),$$
$$\mathbf{Z}_r^{(\text{dspk},s)}(t) = \left(\mathbf{x}^{(\text{direction},s)^*} \cdot \mathbf{U}_r^{\text{dspk}}\right) \cdot \left(\mathbf{T}_r^{\text{dspk}} * x^{(\text{timing},s)}\right)(t) \cdot \left(\mathbf{H}_r^{\text{dspk}} * \mathbf{Y}_n\right)(t),$$
$$\lambda_n^*(t) = \lambda_n(t) + h_n(t) \qquad (\log \text{ rate})$$

where $\lambda_n(t)$ is given by Eq. 9. For completeness, we included two dynamic spike history tensors to mirror the mean-rate filter terms: one for the motion stimuli and a second for the touch-bar release. However, we found including the touch-bar filters provided little improvement to the model's performance.

The dynamic spike history kernels $\mathbf{H}^{\text{dspk}}$ for the stimulus-dependent spike history (or $\mathbf{H}^{\text{bdspk}}$ for the touch-bar kernel) are shared across all neurons in a population. The stimulus kernels

$\mathbf{T}^{\text{dspk}}$ (or $\mathbf{T}^{\text{bdspk}}$ for the touch-bar release kernel) control the contribution of the dynamic spike history kernel relative to stimulus onset (or touch-bar release) . As with the stimulus filter tensor, we allow the dynamic spike history kernels to depend on stimulus information through the stimulus scaling terms $\mathbf{U}^{\text{dspk}}$. For simplicity, we limited the stimulus scaling for the dynamic spike history in $\mathbf{x}^{(\text{direction},s)^*}$ to include only sample or test information as defined for the "No category or direction model" in the previous section. We found that including category or direction information did not significantly affect our results (results not shown). Each neuron weights the dynamic spike history components by the loading matrices $\mathbf{V}^{\text{dspk}}$ and $\mathbf{V}^{\text{bdspk}}$.

At any given time in the trial, the spike history for a neuron is still a linear function of past spiking. The "effective" spike history kernel for a neuron n at time t can be computed by rearranging the terms in Eq. 13 and adding the constant spike history kernel:

$$\mathbf{H}_{n,t}^{\text{spk eff}}(s) = \mathbf{H}_{n}^{\text{spk}}(s) + \sum_{q=1}^{R_{bh}} \left(\mathbf{V}_{n,q}^{\text{bdspk}} \left(\mathbf{T}_{q}^{\text{bdspk}} * x^{\text{tbar}}\right)(t)\right) \cdot \mathbf{H}_{q}^{\text{bdspk}}(s) + \sum_{s\in\{\text{sample,test}\}} \sum_{r=1}^{R_h} \left(\mathbf{V}_{n,r}^{\text{dspk}} \cdot \left(\mathbf{x}^{(\text{direction},s)^*} \cdot \mathbf{U}_{r}^{\text{dspk}}\right) \cdot \left(\mathbf{T}_{q}^{\text{bdspk}} * x^{\text{tbar}}\right)(t)\right) \cdot \mathbf{H}_{r}^{\text{dspk}}(s). \tag{14}$$

The temporal kernels were parameterized using the same basis set as before:

$$\begin{aligned}
\mathbf{H}^{\text{bdspk}} &= \mathbf{B}^{\text{spk}}\tilde{\mathbf{H}}^{\text{spk}}, && \text{where } \tilde{\mathbf{H}}^{\text{bdspk}} \text{ is a matrix of size } P_{\text{spk}} \times R_{bh}, \\
\mathbf{T}^{\text{bdspk}} &= \mathbf{B}^{\text{tbar}}\tilde{\mathbf{T}}^{\text{tbar}}, && \text{where } \tilde{\mathbf{T}}^{\text{bdspk}} \text{ is a matrix of size } P_{\text{tbar}} \times R_{bh}, \\
\mathbf{H}^{\text{dspk}} &= \mathbf{B}^{\text{spk}}\tilde{\mathbf{H}}^{\text{spk}}, && \text{where } \tilde{\mathbf{H}}^{\text{dspk}} \text{ is a matrix of size } P_{\text{spk}} \times R_{h}, \\
\mathbf{T}^{\text{dspk}} &= \mathbf{B}^{\text{stim}}\tilde{\mathbf{T}}^{\text{stim}}, && \text{where } \tilde{\mathbf{T}}^{\text{dspk}} \text{ is a matrix of size } P_{\text{stim}} \times R_{h}.
\end{aligned} \tag{15}$$

The parameter set for the dynamic spike history model was $\phi^* = \phi \bigcup \left\{\tilde{\mathbf{H}}^{\text{bdspk}}, \tilde{\mathbf{T}}^{\text{bdspk}}, \mathbf{V}^{\text{bdspk}}, \tilde{\mathbf{H}}^{\text{dspk}}, \tilde{\mathbf{T}}^{\text{dspk}}, \mathbf{U}^{\text{dspk}}, \mathbf{V}^{\text{dspk}}\right\}$.

We set i.i.d. standard normal priors for $\mathbf{V}^{\text{dspk}}$, $\mathbf{V}^{\text{bdspk}}$, $\tilde{\mathbf{H}}^{\text{bdspk}}$, $\tilde{\mathbf{H}}^{\text{dspk}}$, and $\tilde{\mathbf{T}}^{\text{dspk}}$. The prior for $\tilde{\mathbf{T}}^{\text{bdspk}}$ was i.i.d. zero-mean normal with variance ψ^2_{bdspk}. The Gaussian prior for $\mathbf{U}^{\text{dspk}}$ was defined analogously to the stimulus term (for the no category model) with hyperparameters ψ_0^* and ψ_{s}^*. The complete hyperprior set was then $\mathcal{H}^* = \mathcal{H} \bigcup \{\mathcal{H}, \psi_{\text{bdspk}}, \psi_0^*, \psi_{\text{s}}^*\}$.

For all dynamic spike history models here, we set $R_{bh} = 1$ and we varied R_h from 1 to 2. For the stimulus filter tensor, we used the cosine direction tuning model with the rank selected in Fig. 4.

4.3.2 Model inference

We performed rank selection in the GMLM by testing cross-validated model performance of the maximum likelihood fit. We used gradient-descent methods to numerically maximize the log likelihood for fold. We initialized the GMLM components randomly. The entries of $\mathbf{V}^{\text{stim}}$, $\mathbf{V}^{\text{tbar}}$,

and $\tilde{\mathbf{H}}^{\text{spk}}$ were generated as standard normal. The matrices $\tilde{\mathbf{T}}^{\text{stim}}$ and $\tilde{\mathbf{T}}^{\text{tbar}}$ were random orthonormal matrices. The baseline firing rate parameters, w, were drawn independently from a normal distribution.

For the MAP estimates shown in Fig. 3 and Fig. 5, we set the hyperparameters to the marginal posterior medians of each hyperparameter estimated using Markov chain Monte Carlo methods (described below). We then maximized the posterior log likelihood given those hyperparameters.

For the Bayesian analyses of the GMLM, we used MCMC to generate samples from the posterior distribution of the model parameters and hyperparameters given all data from an LIP population. We used Hamiltonian Monte Carlo (HMC) to sample jointly from the posterior of the parameters and the log-transformed hyperparameters. The log transform on the hyperparameters ensures that the hyperparameters are positive. A detailed description of the HMC sampling algorithm is given in (Neal, 2011). The Hamiltonian equations were solved numerically using a leap-frog integrator with step size ϵ for S steps. We set $S = \min\left(100, \left\lceil \frac{1}{\epsilon} \right\rceil\right)$ where $\lceil \cdot \rceil$ denotes the ceiling operator. The maximum number of steps was 100 to limit computational costs per sample. However, after tuning the sampler during warmup, we found that $S < 100$.

We denote the vectorized set of all parameters and log-transformed hyperparameters for sample s as $\Phi^{(s)} = \{\phi, \mathcal{H}\}$. We initialize the model parameters for the sampler ($s = 1$) by initializing the parameters randomly as we did for the maximum likelihood estimation. The log hyperparameters were initialized as i.i.d. draws from a standard normal distribution. The HMC sampler requires specifying a $P \times P$ mass matrix, $\mathbf{M}$. Because the model is high dimensional, we assume $\mathbf{M}$ is diagonal.

We tuned the parameters of the sampler (ϵ and $\mathbf{M}$) by generating 25 000 warmup samples (also known as "burn-in"). The initial value of the step size was $\epsilon = 0.01$. We used the dual-averaging algorithm of Nesterov (2009) to adapt ϵ at each step for the first 24 000 warmup samples (and fixed for the last 1000 warmup samples) to achieve a desired acceptance rate. We used the parameters given in Hoffman & Gelman (2014) to control the learning rate and target sample acceptance rate (80 %). The mass matrix $\mathbf{M}$ was set at three steps.

1. Sample 1: $\mathbf{M}$ is initialized as identity matrix.
2. Sample 4001: $\mathbf{M} = \text{Diag}\left(\text{cov}(\Phi_i^{(2001:4000)})^{-1}\right)$. The diagonal is the inverse empirical variance of each parameter given samples 2001 to 4000.
3. Sample 19 001: $\mathbf{M} = \text{Diag}\left(\text{cov}(\Phi_i^{(4001:19000)})^{-1}\right)$.

After warmup, we generated 50 000 HMC samples. These sampled were used as the estimate of the posterior distribution of the model parameters and hyperparameters.

One source of autocorrelation in the HMC sampler that could reduce the quality of inference is that the GMLM tensor components could be re-scaled without changing the likelihood. For any $a, b \neq 0$, the rth component of the GMLM stimulus kernel tensor can be rescaled

$$\mathbf{U}^{\text{stim}}_{\cdot,r} \longleftarrow a\mathbf{U}^{\text{stim}}_{\cdot,r}, \qquad \mathbf{T}^{\text{stim}}_{\cdot,r} \longleftarrow b\mathbf{T}^{\text{stim}}_{\cdot,r}, \qquad \mathbf{V}^{\text{stim}}_{\cdot,r} \longleftarrow \frac{1}{ab}\mathbf{V}^{\text{stim}}_{\cdot,r}$$

without changing the resulting kernel tensor. Thus, the log likelihood remains constant. One way to is to constrain fix the norm of two of those vectors, and thereby disallowing re-scaling. Inference can then be performed for those parameters on an appropriate manifold (product of sphere manifolds) using geodesic Monte Carlo methods (Byrne & Girolami, 2013; Holbrook et al., 2016). Instead, we took a different approach by including an efficient Metropolis-Hastings (MH) step for rapidly traversing the locally flat region of the likelihood without additional constraints on the model parameters. The MH step was performed independently for each component r. We define for the current sample s

$$
\begin{aligned}
u^{(s)} &= \|\mathbf{U}^{\text{stim}}_{\cdot,r}\|, & t^{(s)} &= \|\mathbf{T}^{\text{stim}}_{\cdot,r}\|, & v^{(s)} &= \|\mathbf{V}^{\text{stim}}_{\cdot,r}\|, \\
\mathbf{u}^{(s)} &= \frac{1}{u^{(s)}}\mathbf{U}^{\text{stim}}_{\cdot,r}, & \mathbf{t}^{(s)} &= \frac{1}{t^{(s)}}\mathbf{T}^{\text{stim}}_{\cdot,r}, & \mathbf{v}^{(s)} &= \frac{1}{v^{(s)}}\mathbf{V}^{\text{stim}}_{\cdot,r}, \\
\zeta^{(s)} &= u^{(s)}t^{(s)}v^{(s)}.
\end{aligned}
$$

The prior probabilities for each $\mathbf{U}^{\text{stim}}_{\cdot,r}$, $\mathbf{T}^{\text{stim}}_{\cdot,r}$, and $\mathbf{V}^{\text{stim}}_{\cdot,r}$ are multivariate Gaussian with zero mean. Therefore, the prior probability of the vector lengths $p(u^{(s)}, t^{(s)}, v^{(s)}|\mathbf{u}^{(s)}, \mathbf{t}^{(s)}, \mathbf{v}^{(s)}, \mathcal{H}^{(s)}_{\text{stim}})$ can be factorized into independent chi distributions:

$$
\begin{aligned}
p(u^{(s)}|\mathbf{u}^{(s)}, \mathcal{H}^{(s)}_{\text{stim}}) &= \frac{\eta_u^S}{2^{S/2-1}\Gamma\left(S/2\right)}(u^{(s)})^{S-1}\exp\left(-(\eta_u u^{(s)})^2/2\right) \\
p(t^{(s)}|\mathbf{t}^{(s)}, \mathcal{H}^{(s)}_{\text{stim}}) &= \frac{\eta_t^{P_{\text{stim}}}}{2^{P_{\text{stim}}/2-1}\Gamma\left(P_{\text{stim}}/2\right)}(t^{(s)})^{P_{\text{stim}}-1}\exp\left(-(\eta_t t^{(s)})^2/2\right) \\
p(v^{(s)}|\mathbf{v}^{(s)}, \mathcal{H}^{(s)}_{\text{stim}}) &= \frac{\eta_v^N}{2^{N/2-1}\Gamma\left(N/2\right)}(v^{(s)})^{N-1}\exp\left(-(\eta_v v^{(s)})^2/2\right)
\end{aligned}
$$

$$
\eta_u = \left(\mathbf{u}^{(s)\top}\Sigma_u^{-1}\mathbf{u}^{(s)}\right)^{1/2}, \qquad \eta_t = \left(\mathbf{t}^{(s)\top}\mathbf{t}^{(s)}\right)^{1/2}, \qquad \eta_v = \left(\mathbf{v}^{(s)\top}\mathbf{v}^{(s)}\right)^{1/2}
$$

where Σ_u is the prior covariance matrix for $\mathbf{U}^{\text{stim}}_{\cdot,r}$ given the hyperparameters $\mathcal{H}^{(s)}_{\text{stim}}$ (the Gaussian priors for the other two vectors have identity covariance). Our goal is to construct a MH proposal to focus on the case where the total component norm $\zeta^{(s)}$ is constant. Therefore, we perform a change of variables on the prior over $p(u^{(s)}, t^{(s)}, v^{(s)}|\mathbf{u}^{(s)}, \mathbf{t}^{(s)}, \mathbf{v}^{(s)}, \mathcal{H}^{(s)}_{\text{stim}})$ to $p(u^{(s)}, t^{(s)}, \zeta^{(s)}|\mathbf{u}^{(s)}, \mathbf{t}^{(s)}, \mathbf{v}^{(s)}, \mathcal{H}^{(s)}_{\text{stim}})$ in order to compute

$$
\begin{aligned}
p(u^{(s)}, t^{(s)}|\zeta^{(s)}, \mathbf{u}^{(s)}, \mathbf{t}^{(s)}, \mathbf{v}^{(s)}, \mathcal{H}^{(s)}_{\text{stim}}) &\propto p(u^{(s)}, t^{(s)}, \zeta^{(s)}|\mathbf{u}^{(s)}), \mathbf{t}^{(s)}, \mathbf{v}^{(s)}, \mathcal{H}^{(s)}_{\text{stim}}) \\
&= \frac{1}{u^{(s)}t^{(s)}}p(u^{(s)}|\mathbf{u}^{(s)}, \mathcal{H}^{(s)}_{\text{stim}})p(t^{(s)}|\mathbf{t}^{(s)}, \mathcal{H}^{(s)}_{\text{stim}})p(v^{(s)}|\mathbf{v}^{(s)}, \mathcal{H}^{(s)}_{\text{stim}})
\end{aligned}
$$

We can then generate independent scaling factors to perform a random walk on the scaling factors:

$$
\begin{aligned}
s_u &\sim \text{Lognormal}(0, \omega^2), & s_t &\sim \text{Lognormal}(0, \omega^2) \\
u^* &= s_u u^{(s)}, & t^* &= s_t t^{(s)}, & v^* &= \frac{1}{s_u s_t}v^{(s)}
\end{aligned}
$$

We then accept the proposal u^*, t^*, v^* with the MH acceptance probability

$$A(\{u^*, t^*, v^*\}, \{u^{(s)}, t^{(s)}, v^{(s)}\}) = \min\left[1, \frac{p(u^*, t^*|\zeta^{(s)}, \mathbf{u}^{(s)}, \mathbf{t}^{(s)}, \mathbf{v}^{(s)}, \mathcal{H}^{(s)}_{\text{stim}}) q\left(s_u^{-1}\right) q\left(s_t^{-1}\right)}{p(u^{(s)}, t^{(s)}|\zeta^{(s)}, \mathbf{u}^{(s)}), \mathbf{t}^{(s)}, \mathbf{v}^{(s)}, \mathcal{H}^{(s)}_{\text{stim}}) q\left(s_u\right) q\left(s_t\right)}\right] \tag{16}$$

where $q(s) = \text{Lognormal}(s; 0, \omega^2)$. Because the likelihood remains constant in this proposal, we only need to compute the prior to determine the MH acceptance probability. As a result, this step is very fast to compute. We applied the same class of MH proposal to the touch-bar components. We set $\omega = 0.2$ and interleaved 10 MH steps for each tensor component between every HMC step.

4.3.3 Rank selection

We applied cross-validation to select the stimulus kernel tensor rank (R_s) for the GMLM. To do so, we computed the mean test log likelihood per trial per cell. For neuron n,

$$\text{lp}_{\mathcal{M}}(n) = \frac{1}{M_n} \sum_{k=1}^{K} \sum_{l=1}^{M_n^k} \log p(\mathbf{y}^*_{n,k,l} | \phi_k, \mathbf{x}^*_{n,k,l}, \mathcal{M}) \tag{17}$$

where K is the number of folds ($K = 10$ for all the analyses conducted here). The trials in the test set are given as $\mathbf{y}^*_{n,k,l}$ and $\mathbf{x}^*_{n,k,l}$ which represent the spike train and regressors respectively for test trial l in fold k. The number of test trials in fold k for the neuron is M_n^k, and the total number of trials is $M_n = \sum_{k=1}^{K} M_n^k$. The model parameters for the model $\mathcal{M}$ fit to the training data for fold k is ϕ_k .

We then took the average across all cells

$$\overline{\text{lp}}_{\mathcal{M}} = \frac{1}{N} \sum_{n=1}^{N} \text{lp}_{\mathcal{M}}(n). \tag{18}$$

For normalization, we subtracted the $\overline{\text{lp}}_{\mathcal{M}}$ of the GMLM without any stimulus terms (the "rank 0" model):

$$\Delta\overline{\text{lp}}_{\mathcal{M}} = \overline{\text{lp}}_{\mathcal{M}} - \overline{\text{lp}}_{\mathcal{M}_{R_s=0}}. \tag{19}$$

Fig. 4A shows the $\Delta\overline{\text{lp}}_{\mathcal{M}}$ for each GLM and GMLM of from $r = 1$ to 12.

The fraction of log likelihood explained by the GMLM was computed relative to the full GLM (the "full rank" model, denoted $\mathcal{M}_{\text{GLM}}$):

$$\text{frac}(\mathcal{M}) = \frac{\Delta\overline{\text{lp}}_{\mathcal{M}}}{\Delta\overline{\text{lp}}_{\mathcal{M}_{\text{GLM}}}} \tag{20}$$

Fig. 4B shows the $\text{frac}(\mathcal{M})$ for each GLM and GMLM of from $r = 1$ to 12 for the monkey D, DMC late population. We selected the rank r for the full GMLM by selecting the smallest r

for which $\mathrm{frac}(\mathcal{M}_{\mathrm{GMLM},R_s=r}) > 0.9$ (i.e., the number of model components needed to explain 90 % of the likelihood that could be explained by this GLM framework).

The error bars over the cross-validated log likelihood in Fig. 4A were computed by computing $\overline{\mathrm{lp}}_{\mathcal{M}}$ on the test trials for each fold separately (instead of averaging over all K folds). The error bars show two standard errors of $\Delta\overline{\mathrm{lp}}_{\mathcal{M}}$ over the folds.

For the dynamic spike history, we compared model predictive performance with leave-one-out cross-validation estimated with Pareto-smoothed importance sampling using the MCMC samples (Vehtari et al., 2017). The leave-one-out cross-validated log likelihoods were computed for each trial, and then we computed the mean cross-validated log likelihood for each neuron.

4.3.4 Visualizing the GMLM parameters

Fig. 3A shows the individual components of the MAP fit of the full GMLM, which included kernels for each stimulus direction and two kernels for the test stimulus category. For scale, we normalized each component by placing the magnitude of each tensor component in the neuron loading dimension:

$$\mathbf{V}^{\mathrm{stim}}_{\cdot,r} \longleftarrow \|\tilde{\mathbf{T}}^{\mathrm{stim}}_{\cdot,r}\|\|\mathbf{U}^{\mathrm{stim}}_{\cdot,r}\|\mathbf{V}^{\mathrm{stim}}_{\cdot,r}, \tag{21}$$

$$\tilde{\mathbf{T}}^{\mathrm{stim}}_{\cdot,r} \longleftarrow \frac{1}{\|\tilde{\mathbf{T}}^{\mathrm{stim}}_{\cdot,r}\|}\tilde{\mathbf{T}}^{\mathrm{stim}}_{\cdot,r},$$

$$\mathbf{U}^{\mathrm{stim}}_{\cdot,r} \longleftarrow \frac{1}{\|\mathbf{U}^{\mathrm{stim}}_{\cdot,r}\|}\mathbf{U}^{\mathrm{stim}}_{\cdot,r}. \tag{22}$$

The rth row of the left column of Fig. 3A shows the re-scaled $\mathbf{B}^{\mathrm{stim}}\tilde{\mathbf{T}}^{\mathrm{stim}}_{\cdot,r}$. The middle column shows those temporal kernels scaled by the direction weights: the rth row plots $\mathbf{U}^{\mathrm{stim}}_{\theta_d,r}\mathbf{T}^{\mathrm{stim}}_{r}(t)$ for each direction d. The right columns shows the temporal kernels scaled by the additional category weights for the test stimulus: thr rth row plots for $\mathbf{U}^{\mathrm{stim}}_{\mathrm{ctk},r}\mathbf{T}^{\mathrm{stim}}_{r}(t)$ for both categories k. The loading weights in the box plot of Fig. 3B show the elements $\mathbf{V}^{\mathrm{stim}}_{\cdot,r}$ for each component r.

The sample stimulus kernels for the example cells in Fig. 3C are the sample direction kernels scaled by the neuron's loading weights for each component. The r row for example neuron n shows $\mathbf{V}^{\mathrm{stim}}_{n,r}\mathbf{U}^{\mathrm{stim}}_{\theta_d,r}\mathbf{T}^{\mathrm{stim}}_{r}(t)$ for each direction d. The total GMLM tuning (Fig. 3D top row) was the sum over the r components.

To visualize the subspaces, we projected the components of the full GMLM into the top three dimensions. The loading weights of the tensor decomposition used to define the model ($\mathbf{V}^{\mathrm{stim}}$) are not constrained to be orthonormal (as is standard for the PARAFAC decomposition). Therefore, we applied a Tucker decomposition (i.e., higher-order singular value decomposition) to find the three-dimensional subspace that captures most of the population's stimulus tuning structure. We took the stimulus kernel tensor of $\mathcal{K}(t,\theta_d,n)$ (Eq. 11) for all sample stimulus directions. We then took the Tucker decomposition of the stimulus kernel tensor such that

$$\mathcal{K}(t,\theta_d,n) \approx \hat{\mathcal{K}}(t,\theta_d,n) = \mathcal{T} \times_1 \hat{\mathbf{T}} \times_2 \hat{\mathbf{U}} \times_3 \hat{\mathbf{V}} \tag{23}$$

$$\mathcal{K}^*(t, \theta_d, i) = \frac{1}{\sqrt{N}} \mathcal{T} \times_1 \hat{\mathbf{T}} \times_2 \hat{\mathbf{U}}$$

where $\mathcal{T}$ is the core tensor of size $R_s \times D_{\text{sample}} \times 3$, and $\hat{\mathbf{T}}$, $\hat{\mathbf{U}}$, and $\hat{\mathbf{V}}$ are orthonormal matrices. The filter tensor projected into the top three subspace dimensions ($i \in \{1, 2, 3\}$) for each direction over time is then $\mathcal{K}^*(t, \theta_d, i)$.

To find the mean-removed space, we took

$$\overline{\mathcal{K}}(t, s, \theta_d) = \mathcal{K}(t, s, \theta_d) - \frac{1}{D_{sample}} \sum_{j \in \text{sample directions}} \mathcal{K}(t, s, \theta_j) \tag{24}$$

we performed the Tucker decomposition on $\overline{\mathcal{K}}(t, s, \theta_d)$ to obtain the mean-removed subspace.

For visualizing the rank-1 dynamic spike history components in Fig. 8B, we plot the posterior median and pointwise 99 % credible intervals computed using MCMC for the normalized temporal filters, $\mathbf{T}^{\text{dspk}}/||\mathbf{T}^{\text{dspk}}||$ and $\mathbf{H}^{\text{dspk}}/||\mathbf{H}^{\text{dspk}}||$. Because the sign of individual components in the PARAFAC decompositions is not identifiable, we set the sign of the posterior median components with the following transformation in order to better compare across populations:

$$\begin{aligned}
\mathbf{T}^{\text{dspk}} &\longleftarrow \text{mode}(\text{sign}(\mathbf{V}^{\text{dspk}})) \cdot \text{sign}(\mathbf{H}^{\text{dspk}}(1)) \cdot \text{sign}(\mathbf{U}^{\text{dspk}}(1)) \cdot \mathbf{T}^{\text{dspk}}, \\
\mathbf{H}^{\text{dspk}} &\longleftarrow \text{sign}(\mathbf{H}^{\text{dspk}}(1)) \cdot \mathbf{H}^{\text{dspk}}.
\end{aligned}$$

To quantify the timescales of the dynamic spike history kernels (for the pretrained monkeys only), we fit the MAP estimate of the rank-1 dynamic spike history kernel with an exponential function with a least-squares fit.

4.3.5 Bayesian analysis of subspace geometry

We defined tuning metrics in the low-dimensional space estimated by the GMLM with cosine direction tuning to analyze the geometry of task encoding. The metrics were constant over rotations and translations of the latent subspace. We used the posterior distribution of the model parameters estimated using MCMC to establish credible intervals over the metrics.

At each time point t relative to stimulus onset, the cosine-tuned GMLM defines direction tuning in the population as an ellipse embedded in $\mathbb{R}^{R_s}$ parameterized by angle as

$$\begin{aligned}
\mathbf{E}_t(\theta) &= \sum_{r=1}^{R_s} \mathbf{T}_r^{\text{stim}}(t) \left(\mathbf{U}_{r,\cos}^{\text{stim}} \cos(\theta) + \mathbf{U}_{r,\sin}^{\text{stim}} \sin(\theta) \right) \mathbf{R}_{\cdot,r}, \\
\mathbf{R} &= \frac{1}{\sqrt{N}} \text{orth}(\mathbf{V}^{\text{stim}})^\top \mathbf{V}^{\text{stim}},
\end{aligned} \tag{25}$$

where $\text{orth}(\mathbf{V}^{\text{stim}})$ denotes a matrix whose columns contain an orthonormalized basis for the span of the columns of $\mathbf{V}^{\text{stim}}$. Here, the orthogonalized R_s-dimensional output space, $\mathbf{R}$, is normalized by the number of cells. We computed the angle of the major axis of the ellipse as $\theta_{\max}$ where

$$\theta_{\max} = \arg\max_{\theta} D_t(\theta), \tag{26}$$

$$
\begin{aligned}
D_t(\theta) &= \|\mathbf{E}_t(\theta) - \mathbf{E}_t(\theta + 180°)\|, \\
t_0 &= \frac{1}{2}\text{arccot}\left(\frac{\vec{f_1}\cdot\vec{f_1} - \vec{f_2}\cdot\vec{f_2}}{2}\right), \qquad \vec{f_1} = \mathbf{E}_t(0), \qquad \vec{f_2} = \mathbf{E}_t(90°), \\
\implies \theta_{\max} &= t_0 \text{ or } t_0 + 90.°
\end{aligned}
$$

Because $\theta_{\max}$ is identifiable only up to a factor of $180°$, we added the constraint $\theta_{\max} \in [45°, 225°]$ to relate the angle to category in the task. The norms of the major and minor axes are $D_t(\theta_{\max})$ and $D_t(\theta_{\max} + 90°)$ respectively.

The category tuning vector is the difference in the low-dimensional tuning space between the category one and category two kernels:

$$\mathbf{F}_t = \sum_{r=1}^{R_s} \mathbf{T}_r^{\text{stim}}(t)\left(\mathbf{U}_{r,\text{cs1}}^{\text{stim}} - \mathbf{U}_{r,\text{cs2}}^{\text{stim}}\right)\mathbf{R}_{\cdot,r}. \tag{27}$$

Category tuning norm at each time t relative to stimulus onset is then the norm of the vector, $\|\mathbf{F}_t\|$.

For the Bayesian analysis, we computed $\theta_{\max}$, $D_t(\theta_{\max})$, $D_t(\theta_{\max} + 90°)$, and $\|\mathbf{F}_t\|$ for each sample from the posterior distribution of the model parameters. We then computed the posterior median and a 99 % credible interval covering 0.5 % to 99.5 % of the posterior for each time t.

For the supplementary analyses in Fig. S3 and Fig. S7, we performed component-wise analyses of the GMLM fits. We note that the GMLM posterior has multiple modes: the order of the components can be permuted or a sign flip could occur between $\mathbf{U}^{\text{stim}}$ and $\mathbf{T}^{\text{stim}}$. These modes define equivalent subspaces and kernel tensors, and the prior distributions are the same at each mode. We did not find that the HMC sampler jumped between these modes, and thus we could simply analyze the individual components of the GMLM tensor. For the component-wise analysis, we looked at each $r \in \{1, \ldots, R_s\}$ individually. The direction tuning for the component was

$$
\begin{aligned}
\theta^{(r)} &= \arctan 2\left(\mathbf{U}_{r,\text{sin}}^{\text{stim}}, \mathbf{U}_{r,\text{cos}}^{\text{stim}}\right), && \text{(angle)} \qquad (28)\\
a^{(r)} &= \sqrt{{\mathbf{U}_{r,\text{sin}}^{\text{stim}}}^2 + {\mathbf{U}_{r,\text{cos}}^{\text{stim}}}^2}. && \text{(direction magnitude)}
\end{aligned}
$$

The sample and test category tuning for the component was

$$
\begin{aligned}
C_{\text{sample}}^{(r)} &= |\mathbf{U}_{r,\text{sample}}^{\text{stim}}|, && \text{(sample category magnitude)} \qquad (29)\\
C_{\text{test}}^{(r)} &= |\mathbf{U}_{r,\text{test}}^{\text{stim}}|. && \text{(test category magnitude)} \qquad (30)
\end{aligned}
$$

4.4 Decoding analyses

All decoders were linear, binary classifiers on pseudopopulation trials spike counts fit with logistic regression in MATLAB using the `fitclinear` function. The training set spike counts were z-scored and the decoder was fit with ridge regression with penalty 0.1. Because the

neurons were recorded independently, we constructed pseudopopulation of 50 trials per stimulus. Each pseudopopulation trial consisted of one randomly sampled (with replacement) trial from each neuron in the recorded population for a particular stimulus direction. We repeated the decoding analysis on 1000 random pseudopopulations to obtain bootstrapped confidence intervals.

To decode sample category as a function of time from stimulus onset, we fit and tested decoders using spike counts in a sliding 200 ms window (centered at the decoding time). To test for direction-independent category tuning, training and validation conditions were trials from different directions to test direction-independent category encoding (Sarma et al., 2016). We therefore fit two decoders, each using trials only from a subset of motion directions. Generalization was evaluated using the withheld directions for each decoder, and the total generalization performance was averaged across the two decoders. The two sets of monkeys had a different set of sample directions, and thus different train/validation conditions For monkeys B and J, each training set contained two motion directions, spaced $180°$ apart: {15° and 195°} and {75° and 225°}. Test sets were then the four remaining motion directions in each condition (135° and 315° trials were in the validation set for both decoders). For monkeys D and H, the training sets were {67.5°, 112.5°, 247.5° and 292.5°} or {157.5°, 202.5°, 337.5° and 22.5°}.

For decoding category decoding during the test stimulus, we used pseudopopulation spike counts in a window from 0 to 200 ms after test motion onset. For these decoders, the training and validation sets included pseudopopulation trials from all motion directions. The decoders were trained using only match (or non-match) trials and tested for generalization on non-match (or match). The total performance was the average across the match-trained and non-match-trained decoders. We trained separate decoders for sample and test category. The DMS populations were excluded from this analysis, because the test stimulus directions depended on the sample stimulus.

4.5 Modeling software

All GLM and GMLM analyses were performed using custom software for MATLAB (MathWorks) and CUDA (Nvidia). The GMLM tools are available publicly at `https://github.com/latimerk/GMLM_dmc`. Tucker decomposition for visualizing the subspaces was performed with Tensor Toolbox for MATLAB (Bader et al., 2019).

Acknowledgements

This work was supported by a Chicago Fellowship (KWL) and grants NIH R01 EY019041 and DOD VBFF (DJF). We thank Rheza Budiono, Jeffrey Johnston, Pantea Moghimi, Barbara Peysakhovich, Jonathan Pillow, Matthew Rosen, Jacob Yates, and Oliver Zhu for helpful comments and discussions.

References

Ahrens, M. B., Paninski, L., & Sahani, M. (2008). Inferring input nonlinearities in neural encoding models. *Network: Computation in Neural Systems*, *19*(1), 35–67.

Aoi, M. C., Mante, V., & Pillow, J. W. (2020). Prefrontal cortex exhibits multidimensional dynamic encoding during decision-making. *Nature Neuroscience*, *23*(11), 1410–1420.

Bader, B. W., Kolda, T. G., et al. (2019). Matlab tensor toolbox.
URL `https://www.tensortoolbox.org`

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

Byrne, S., & Girolami, M. (2013). Geodesic monte carlo on embedded manifolds. *Scandinavian Journal of Statistics*, *40*(4), 825–845.

Chowdhury, S. A., & DeAngelis, G. C. (2008). Fine discrimination training alters the causal contribution of macaque area mt to depth perception. *Neuron*, *60*(2), 367–377.

Churchland, A. K., & Kiani, R. (2016). Three challenges for connecting model to mechanism in decision-making. *Current Opinion in Behavioral Sciences*, *11*, 74–80.

Churchland, M. M., Byron, M. Y., Cunningham, J. P., Sugrue, L. P., Cohen, M. R., Corrado, G. S., Newsome, W. T., Clark, A. M., Hosseini, P., Scott, B. B., et al. (2010). Stimulus onset quenches neural variability: a widespread cortical phenomenon. *Nature Neuroscience*, *13*(3), 369–378.

Constantinidis, C., Funahashi, S., Lee, D., Murray, J. D., Qi, X.-L., Wang, M., & Arnsten, A. F. (2018). Persistent spiking activity underlies working memory. *Journal of Neuroscience*, *38*(32), 7020–7028.

Elsayed, G., Ramachandran, P., Shlens, J., & Kornblith, S. (2020). Revisiting spatial invariance with low-rank local connectivity. In *International Conference on Machine Learning*, (pp. 2868–2879). PMLR.

Fanini, A., & Assad, J. A. (2009). Direction selectivity of neurons in the macaque lateral intraparietal area. *Journal of Neurophysiology*, *101*(1), 289–305.

Fontanini, A., & Katz, D. B. (2008). Behavioral states, network states, and sensory response variability. *Journal of Neurophysiology*, *100*(3), 1160–1168.

Freedman, D. J., & Assad, J. A. (2016). Neuronal mechanisms of visual categorization: an abstract view on decision making. *Annual Review of Neuroscience*, *39*, 129–147.

Freedman, D. J., & Ibos, G. (2018). An integrative framework for sensory, motor, and cognitive functions of the posterior parietal cortex. *Neuron*, *97*(6), 1219–1234.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian analysis*, *1*(3), 515–534.

Goldstone, R. L., & Byrge, L. A. (2015). Perceptual learning. *Oxford Handbook of the Philosophy of Perception*, (pp. 812–32).

Golowasch, J., Goldman, M. S., Abbott, L., & Marder, E. (2002). Failure of averaging in the construction of a conductance-based neuron model. *Journal of Neurophysiology*, *87*(2), 1129–1131.

Harris, K. D., Aravkin, A., Rao, R., & Brunton, B. W. (2019). Time-varying autoregression with low rank tensors. *arXiv preprint arXiv:1905.08389*.

Hart, E., & Huk, A. C. (2020). Recurrent circuit dynamics underlie persistent activity in the macaque frontoparietal network. *Elife*, *9*, e52460.

Hoffman, M. D., & Gelman, A. (2014). The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, *15*(1), 1593–1623.

Holbrook, A., Vandenberg-Rodes, A., Fortin, N., & Shahbaba, B. (2017). A bayesian supervised dual-dimensionality reduction model for simultaneous decoding of lfp and spike train signals. *Stat*, *6*(1), 53–67.

Holbrook, A., Vandenberg-Rodes, A., & Shahbaba, B. (2016). Bayesian inference on matrix manifolds for linear dimensionality reduction. *arXiv preprint arXiv:1606.04478*.

Huk, A. C., Katz, L. N., & Yates, J. L. (2017). The role of the lateral intraparietal area in (the study of) decision making. *Annual Review of Neuroscience*, *40*, 349–372.

Jazayeri, M., & Movshon, J. A. (2006). Optimal representation of sensory information by neural populations. *Nature Neuroscience*, *9*(5), 690–696.

Kobak, D., Brendel, W., Constantinidis, C., Feierstein, C. E., Kepecs, A., Mainen, Z. F., Qi, X.-L., Romo, R., Uchida, N., & Machens, C. K. (2016). Demixed principal component analysis of neural population data. *eLife*, *5*, e10989.

Kolda, T. G., & Bader, B. W. (2009). Tensor decompositions and applications. *SIAM review*, *51*(3), 455–500.

Kossaifi, J., Lipton, Z. C., Kolbeinsson, A., Khanna, A., Furlanello, T., & Anandkumar, A. (2020). Tensor regression networks. *Journal of Machine Learning Research*, *21*, 1–21.

Li, S., Zhou, X., Constantinidis, C., & Qi, X. (2020). Plasticity of persistent activity and its constraints. *Frontiers in Neural Circuits*, *14*, 15.

Libby, A., & Buschman, T. J. (2021). Rotational dynamics reduce interference between sensory and memory representations. *Nature Neuroscience*, *24*(5), 715–726.

Liu, L. D., & Pack, C. C. (2017). The contribution of area mt to visual motion perception depends on training. *Neuron*, *95*(2), 436–446.

Luck, S. J., & Vogel, E. K. (2013). Visual working memory capacity: from psychophysics and neurobiology to individual differences. *Trends in Cognitive Sciences*, *17*(8), 391–400.

Lundqvist, M., Rose, J., Herman, P., Brincat, S. L., Buschman, T. J., & Miller, E. K. (2016). Gamma and beta bursts underlie working memory. *Neuron*, *90*(1), 152–164.

Masse, N. Y., Yang, G. R., Song, H. F., Wang, X.-J., & Freedman, D. J. (2019). Circuit mechanisms for the maintenance and manipulation of information in working memory. *Nature Neuroscience*, *22*(7), 1159–1167.

Miller, E. K., Lundqvist, M., & Bastos, A. M. (2018). Working memory 2.0. *Neuron*, *100*(2), 463–475.

Mohamed, S., Ghahramani, Z., & Heller, K. A. (2008). Bayesian exponential family pca. In *Advances in Neural Information Processing Systems*, vol. 21, (pp. 1089–1096). Citeseer.

Neal, R. M. (2011). Mcmc using hamiltonian dynamics. In S. Brooks, A. Gelman, G. L. Jones, & X.-L. Meng (Eds.) *Handbook of Markov Chain Monte Carlo*, chap. 5, (pp. 113–162). Boca Raton: Chapman and Hall–CRC Press.

Nesterov, Y. (2009). Primal-dual subgradient methods for convex problems. *Mathematical Programming*, *120*(1), 221–259.

Okazawa, G., Hatch, C. E., Mancoo, A., Machens, C. K., & Kiani, R. (2021). Representational geometry of perceptual decisions in the monkey parietal cortex. *Cell*, *184*(14), 3748–3761.

Orhan, A. E., & Ma, W. J. (2019). A diverse range of factors affect the nature of neural representations underlying short-term memory. *Nature Neuroscience*, *22*(2), 275–283.

Padonou, E., & Roustant, O. (2016). Polar gaussian processes and experimental designs in circular domains. *SIAM/ASA Journal on Uncertainty Quantification*, *4*(1), 1014–1033.

Pandit, P., Sahraee-Ardakan, M., Amini, A. A., Rangan, S., & Fletcher, A. K. (2020). Generalized autoregressive linear models for discrete high-dimensional data. *IEEE Journal on Selected Areas in Information Theory*, *1*(3), 884–896.

Park, I. M., Meister, M. L., Huk, A. C., & Pillow, J. W. (2014). Encoding and decoding in parietal cortex during sensorimotor decision-making. *Nature Neuroscience*, *17*(10), 1395–1403.

Park, M., & Pillow, J. W. (2013). Bayesian inference for low rank spatiotemporal neural receptive fields. In *Advances in Neural Information Processing Systems*, (pp. 2688–2696).

Pillow, J. W., Shlens, J., Paninski, L., Sher, A., Litke, A. M., Chichilnisky, E., & Simoncelli, E. P. (2008). Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature*, *454*(7207), 995–999.

Qi, X., & Constantinidis, C. (2013). Neural changes after training to perform cognitive tasks. *Behavioural Brian Research*, *241*, 235–243.

Rahnev, D., & Denison, R. N. (2018). Suboptimality in perceptual decision making. *Behavioral and Brain Sciences*, *41*.

Robinson, B. S., Berger, T. W., & Song, D. (2016). Identification of stable spike-timing-dependent plasticity from spiking activity with generalized multilinear modeling. *Neural Computation*, *28*(11), 2320–2351.

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

Sahani, M., & Linden, J. F. (2003). Evidence optimization techniques for estimating stimulus-response functions. In *Advances in Neural Information Processing Systems*, (pp. 317–324). MIT; 1998.

Sarma, A., Masse, N. Y., Wang, X.-J., & Freedman, D. J. (2016). Task-specific versus generalized mnemonic representations in parietal and prefrontal cortices. *Nature Neuroscience*, *19*(1), 143–149.

Seely, J. S., Kaufman, M. T., Ryu, S. I., Shenoy, K. V., Cunningham, J. P., & Churchland, M. M. (2016). Tensor analysis reveals distinct population structure that parallels the different computational roles of areas m1 and v1. *PLoS Computational Biology*, *12*(11), e1005164.

Steinmetz, N. A., Zatka-Haas, P., Carandini, M., & Harris, K. D. (2019). Distributed coding of choice, action and engagement across the mouse brain. *Nature*, *576*(7786), 266–273.

Stringer, C., Pachitariu, M., Steinmetz, N., Carandini, M., & Harris, K. D. (2019). High-dimensional geometry of population responses in visual cortex. *Nature*, *571*(7765), 361–365.

Summerfield, C., & De Lange, F. P. (2014). Expectation in perceptual decision making: neural and computational mechanisms. *Nature Reviews Neuroscience*, *15*(11), 745–756.

Swaminathan, S. K., & Freedman, D. J. (2012). Preferential encoding of visual categories in parietal cortex compared with prefrontal cortex. *Nature Neuroscience*, *15*(2), 315–320.

Swaminathan, S. K., Masse, N. Y., & Freedman, D. J. (2013). A comparison of lateral and medial intraparietal areas during a visual categorization task. *Journal of Neuroscience*, *33*(32), 13157–13170.

Tang, H., Riley, M. R., Singh, B., Qi, X., Blake, D. T., & Constantinidis, C. (2020). Training-induced prefrontal neuronal changes transfer between tasks. *bioRxiv*.

Truccolo, W., Eden, U. T., Fellows, M. R., Donoghue, J. P., & Brown, E. N. (2005). A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects. *Journal of Neurophysiology*, *93*(2), 1074–1089.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and computing*, *27*(5), 1413–1432.

Vogel, E. K., & Awh, E. (2008). How to exploit diversity for scientific gain: Using individual differences to constrain cognitive theory. *Current Directions in Psychological Science*, *17*(2), 171–176.

Weber, A. I., & Pillow, J. W. (2017). Capturing the dynamical repertoire of single neurons with generalized linear models. *Neural Computation*, *29*(12), 3260–3289.

Williams, A. H., Kim, T. H., Wang, F., Vyas, S., Ryu, S. I., Shenoy, K. V., Schnitzer, M., Kolda, T. G., & Ganguli, S. (2018). Unsupervised Discovery of Demixed, Low-Dimensional Neural Dynamics across Multiple Timescales through Tensor Component Analysis. *Neuron*, *98*(6), 1099–1115.e8.

Zhou, H., Li, L., & Zhu, H. (2013). Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association*, *108*(502), 540–552.

Zhou, Y., Rosen, M. C., Swaminathan, S. K., Masse, N. Y., Zhu, O., & Freedman, D. J. (2021). Distributed functions of prefrontal and parietal cortices during sequential categorical decisions. *eLife*, *10*, e58782.

Zoltowski, D. M., Latimer, K. W., Yates, J. L., Huk, A. C., & Pillow, J. W. (2019). Discrete stepping and nonlinear ramping dynamics underlie spiking responses of lip neurons during decision-making. *Neuron*, *102*(6), 1249–1258.

Zoltowski, D. M., & Pillow, J. W. (2018). Scaling the poisson glm to massive neural datasets through polynomial approximations. In *Advances in Neural Information Processing Systems*, vol. 31, (p. 3517). NIH Public Access.

Supplementary Information

Low-dimensional encoding of decisions in parietal cortex reflects long-term training history

Authors: Kenneth W. Latimer[1]* & David J. Freedman[1]

[1]**Department of Neurobiology, University of Chicago**

*__Correspondence; E-mail: latimerk@uchicago.edu.__

Supplementary figures 1-9

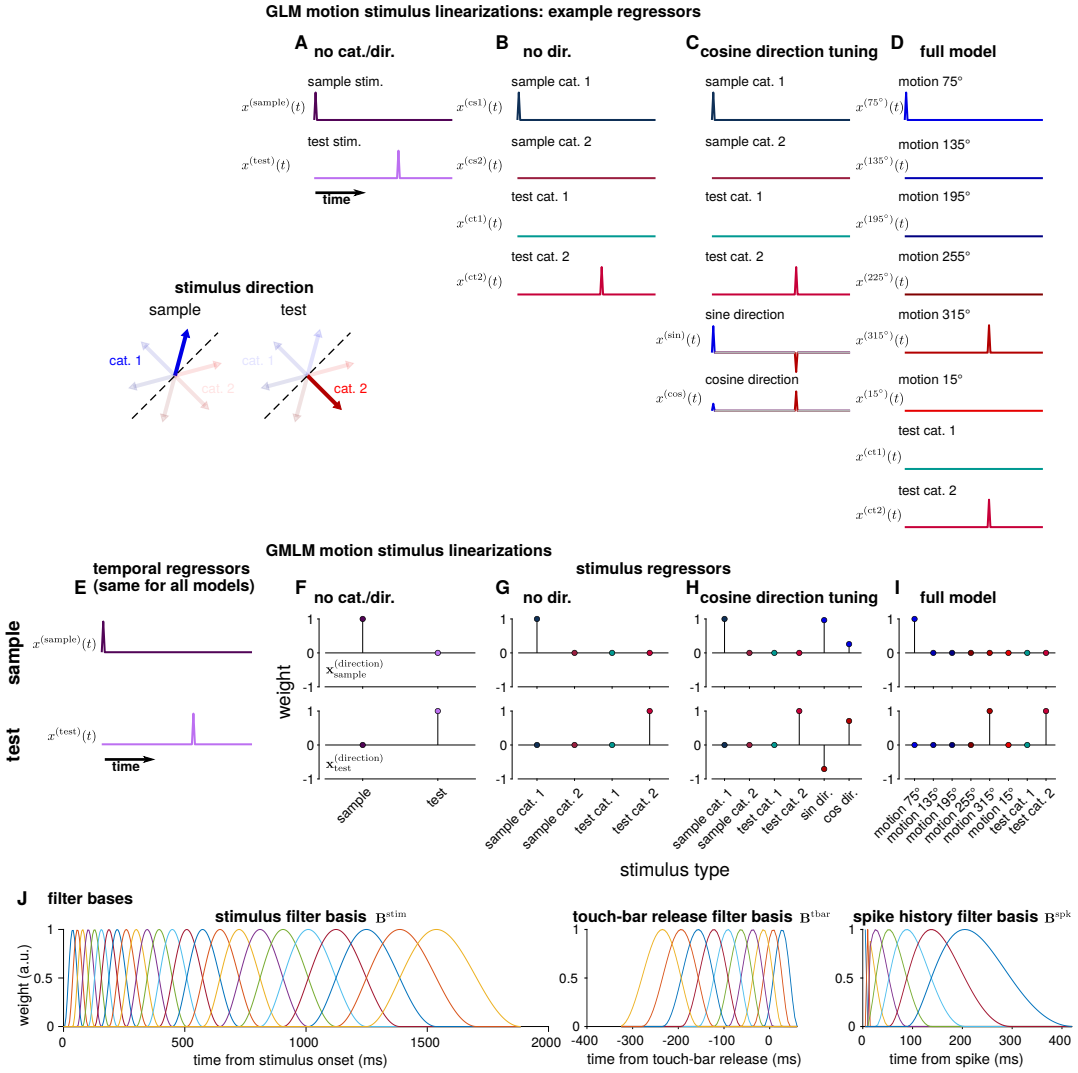

Figure S1: Linearizations of the DMS or DMC tasks in the GMLM. (**A-D**) The temporal event regressors for the four GLM types for an example trial with a sample stimulus 75° (category one) and test stimulus of 315° (category two). (**A**) The two stimulus events for the no category or direction tuning model. The top event is 1 at the sample stimulus onset time and 0 elsewhere, and the bottom event is 1 at the test stimulus onset time and 0 elsewhere. (**B**) The stimulus events for the no direction tuning model. The two sample (or test) category events encode the onset time of a sample (or test) stimulus only for a specific category (the sample category two event is 0 for this trial because the sample stimulus is category one). (**C**) The stimulus events for the cosine tuning model. The category events are the same as the category events in B. The sine (or cosine) event is equal to the sine (or cosine) of the stimulus direction at the onset of either stimulus. (**D**) The stimulus events for the full tuning model. The first six events are 1 at the onset time of a specific stimulus direction (sample or test). The two category events are the same as before. This configuration is identifiable while allowing the category tuning to be different between the sample and test period, while keeping the direction tuning constant. (**E-F**) The event regressors for the four GMLM types for the same trial configuration. (**E**) The temporal events for the sample (top) and test (bottom) stimulus onset times. (**F-I**) The GMLM stimulus weightings for the four model configurations for the sample (top) and test (bottom) stimulus correspond to the weight of the stimulus events in A-D. The complete temporal kernels in the corresponding GLMs are thus the outer product of the temporal regressors in E and the weights of the weights in F-I, summed over the sample and test events. (**J**) The three bases for the temporal kernels used in both the GLM and GMLM: the stimulus event bases (left), the touch-bar release basis (middle), and the spike history basis (right). The bases were orthonormalized for model fitting.

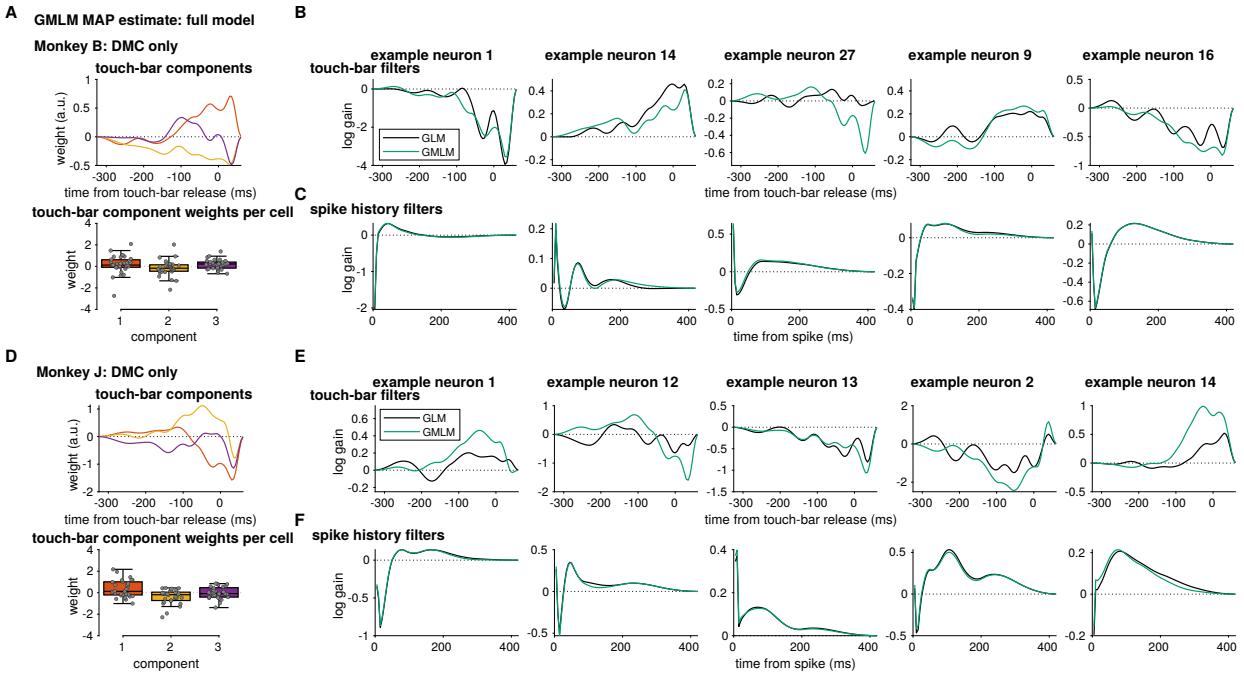

Figure S2: Touch-bar and spike-history kernels from the GMLM (full model) and the GLM fits. (**A**) The low-dimensional touch-bar release components for monkey B. (top) The three temporal kernels. (bottom) The loading weights for each touch-bar release component for each cell (points). (**B**) Example touch-bar filters for five cells. The GLM touch-bar filters (black) are compared to the GMLM fit (cyan). (**C**) Spike-history filters fit to the same cells in B. The GLM spike-history filters (black) are nearly identical to the to the GMLM fit (cyan). (**D-F**) Same as A-C for monkey J.

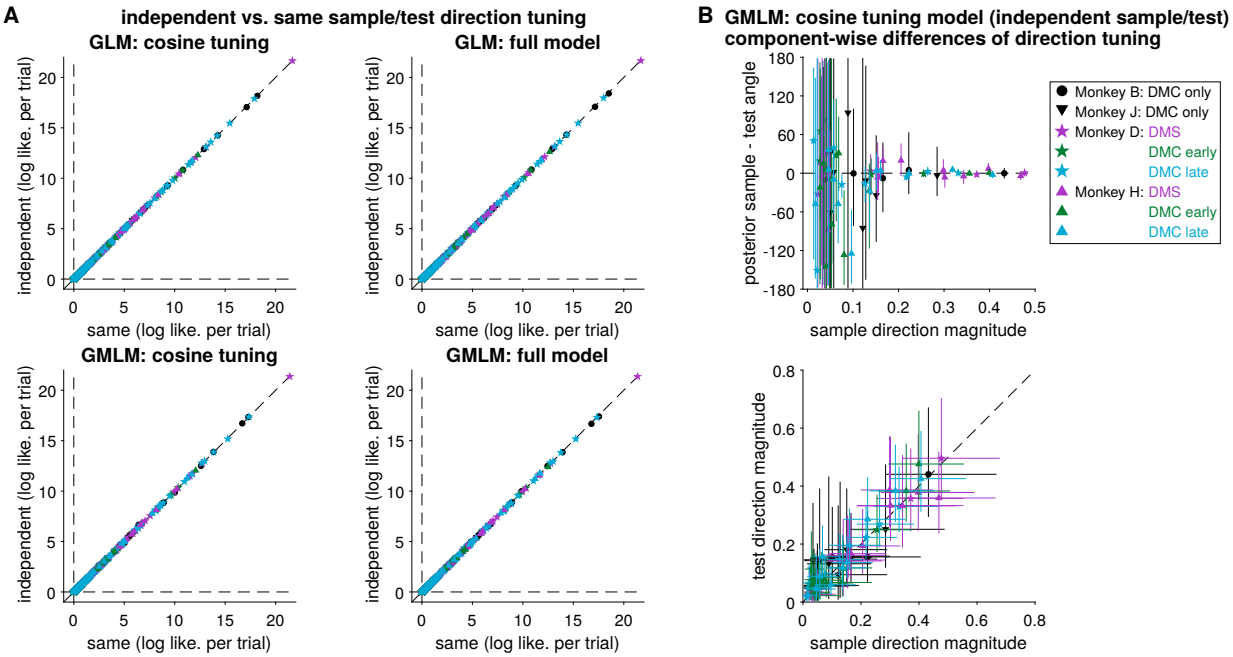

Figure S3: The GMLM and GLM find similar direction tuning across the sample and test stimuli. (**A**) cross-validated per-trial log likelihoods for each cell (relative to the GMLM without any stimulus terms, $R_s = 0$). The left column shows comparisons using the cosine tuning model and the right column shows the full model. The top row compares the GLM fits to single cells and the bottom row shows the GMLM fits for the same model configurations. No population showed a significant improvement including the independent directions (for each model and population $p > 0.8$, one-sided Wilcoxin signed rank test with Benjamini-Hochberg correction). Several populations indicated that overfitting occurred with the GLM with independent directions: the same direction model was on average better. (**B**) Bayesian analysis of the cosine-tuned GMLM with independent sample and test direction parameters. Each point represents a single GMLM stimulus component for one population (i.e., there are seven points for monkey B because we selected seven GMLM stimulus components). (top) The posterior difference in preferred angle between the sample and test stimuli as a function of the magnitude of sample direction tuning. The angle is $\theta^{(r)}$ and the magnitude is $a^{(r)}$ in Eq. 28 (see Methods). As the magnitude increases, the test and sample directions tend towards zero. At lower magnitudes, the preferred angle is difficult to estimate (undetectable) and therefore the difference shows high uncertainty. (bottom) The magnitude of the sample and test direction tuning for each component. Error bars show 99 % credible intervals.

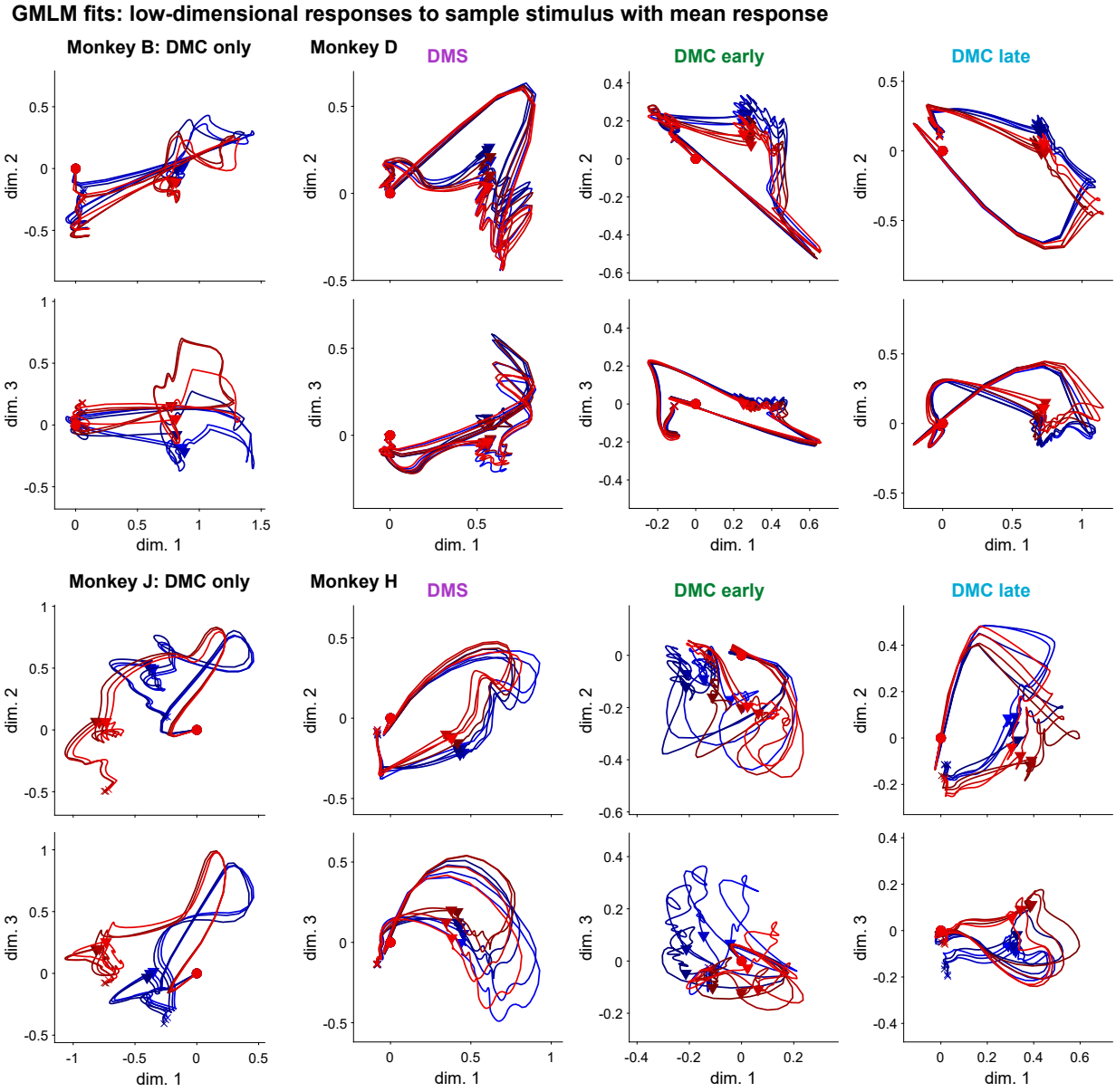

Figure S4: The top three dimensions of the GMLM subspaces (full model) in response to the sample stimulus for each animal and recording epoch without removing the mean over motions directions. Fig. 5 shows the subspaces and trajectories after removing the mean.

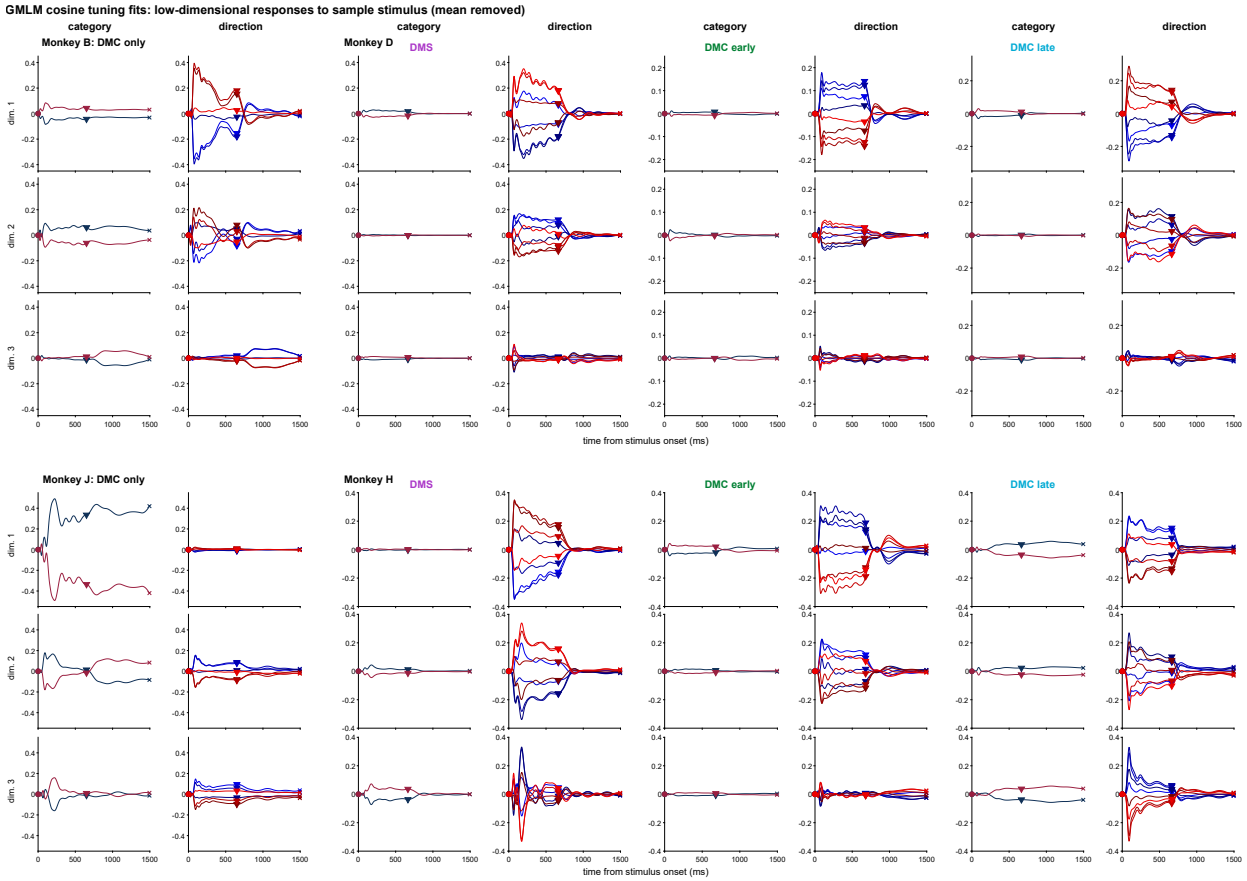

Figure S5: Low-dimensional subspaces of the GMLM with cosine direction tuning with the mean removed for each LIP population. Each of the top three dimensions are shown as a function of time relative to sample stimulus onset (this is similar to Fig. 5, but with the dimensions plotted separately relative to time). For the cosine model, we can separate the direction and category components. The left column for each population shows the sample category trajectories in the three dimensions. The right columns shows the direction trajectories, decoupled from category.

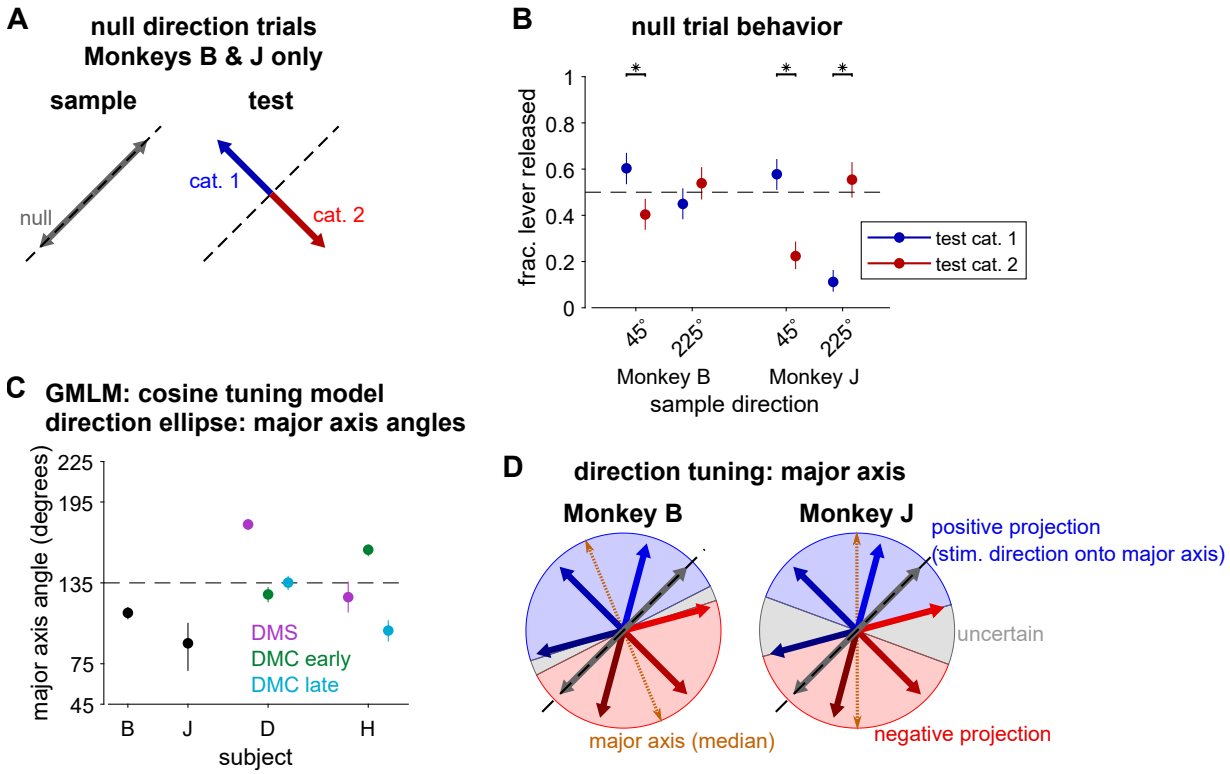

Figure S6: Analysis of the direction tuning of the low-dimensional GMLM cosine tuning model components. (**A**) The sample and test directions selected on "null" direction trials for monkeys B and J. The sample directions lie on the category boundary. These trials were not included in the GMLM analysis and were rewarded randomly. (**B**) Behavior for the four possible combinations of angles on the null-direction trials. The color indicates the test direction/category. The point shows the posterior mean estimate of the fraction of touch-bar released during the test stimulus presentation and the error bars denote a 99 % credible interval. Asterisks indicate that the response proportion for the two test directions was different for a given null sample direction ($p < 0.01$; two-sided Wilcoxon rank sum test, Holm-Bonferroni corrected). (**C**) Bayesian analysis of the cosine tuned GMLM. The GMLM defines the direction tuning as an ellipse in a low-dimensional space. We computed the angle of the major axis of the ellipse: the angle with the most modulation in the low-dimensional space (see Methods Eq. 26). The angle is only identifiable up to 180°. Therefore, we placed it within 45° to 225° to align with the task. If the axis aligned exactly with the categorization task, the angle would be 135°. The ellipse depends on time relative to stimulus onset, and so we took the mean angle during the first 650 ms of stimulus presentation. The points show the posterior median and the error bars denote a 99 % credible interval. (**D**) Illustration of how the direction ellipse's major axis aligns with the task directions. The blue region shows where motion-direction angles project positively along the major axis vector (generally overlapping with category one). The red region shows where motion-direction angles project negatively along the major axis vector. The gray region shows angles that are within the 99 % credible region of the posterior (from C) and cannot be classified. We note that the regions do not exactly align with the category bounds. However, they do correlate with the monkeys' choice biases for the null directions: for monkey B, the 45° null-direction (up and to the right) is in the blue region and the monkey was more likely to release the touch-bar on 45° trials when the test stimulus was in category one than for a category two test stimulus.

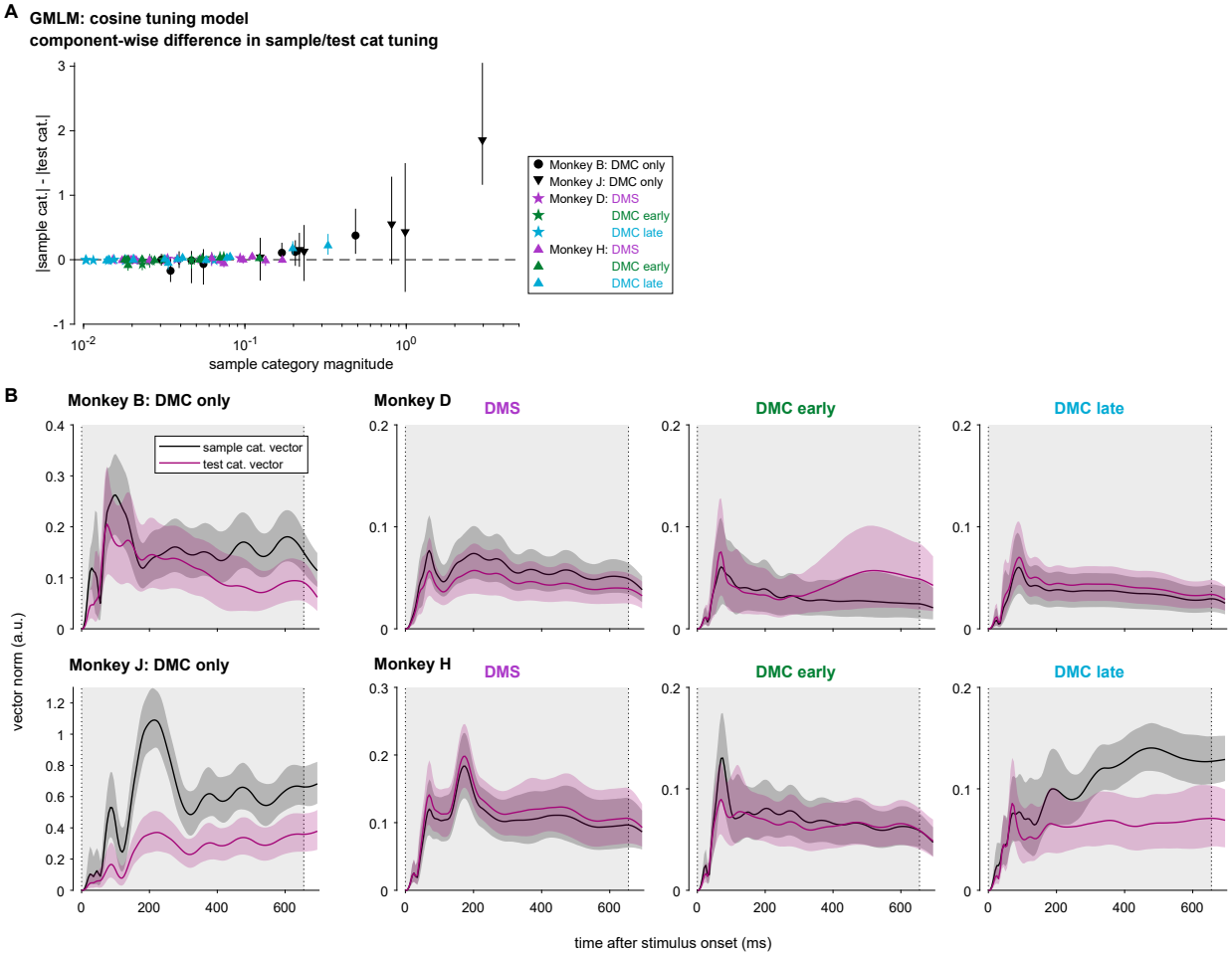

Figure S7: Analysis of the sample and test category tuning of the low-dimensional GMLM cosine tuning model components. (**A**) The difference in the magnitude of coefficients for the sample and test categories in the individual GMLM components (one point per each GMLM stimulus component per population). The component-wise sample category magnitude is computed as $C^{(r)}_{\text{sample}}$ and the difference is $C^{(r)}_{\text{sample}} - C^{(r)}_{\text{test}}$ in Eq. 29 (see Methods). The points show the posterior median and the error bars denote a 99 % credible interval. (**B**) The norm of the category tuning vector as a function of stimulus onset time for the sample and test stimuli in each of the eight LIP populations. The category vector norm is given in Eq. 27 (see Methods). The traces show the posterior median and the shaded regions denote a pointwise 99 % credible interval.

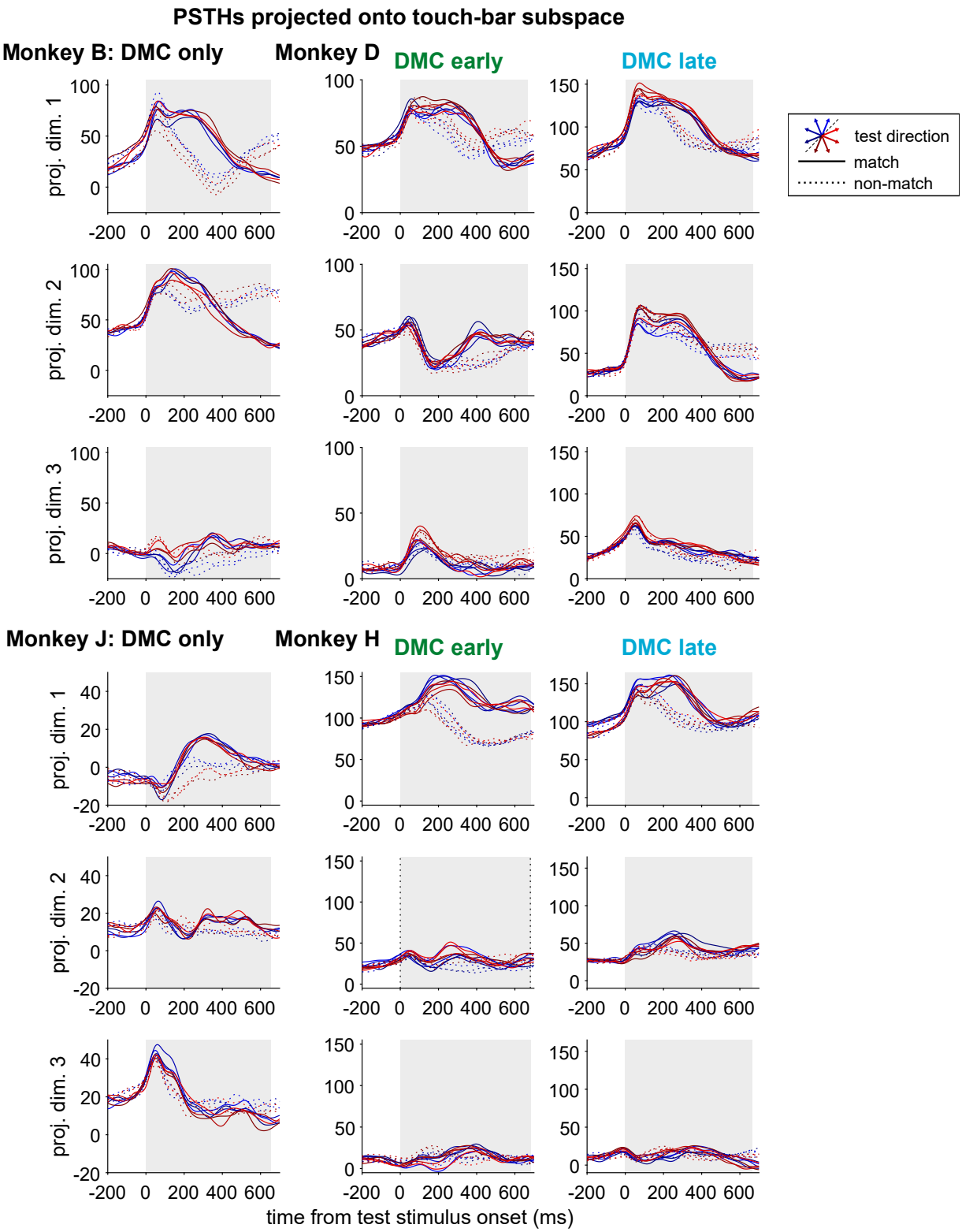

Figure S8: The PSTHs of the six DMC populations projected onto the three-dimensional touch-bar subspace fit by the full GMLM (the subspace given by $\mathrm{orth}(\mathbf{V}^{\mathrm{tbar}})$, see Methods). The PSTHs are conditioned by both test stimulus direction (color) and by match (solid lines) or non-match (dotted lines) trials. The gray region denotes the stimulus presentation period (although it is terminated early on match trials by the touch-bar release).

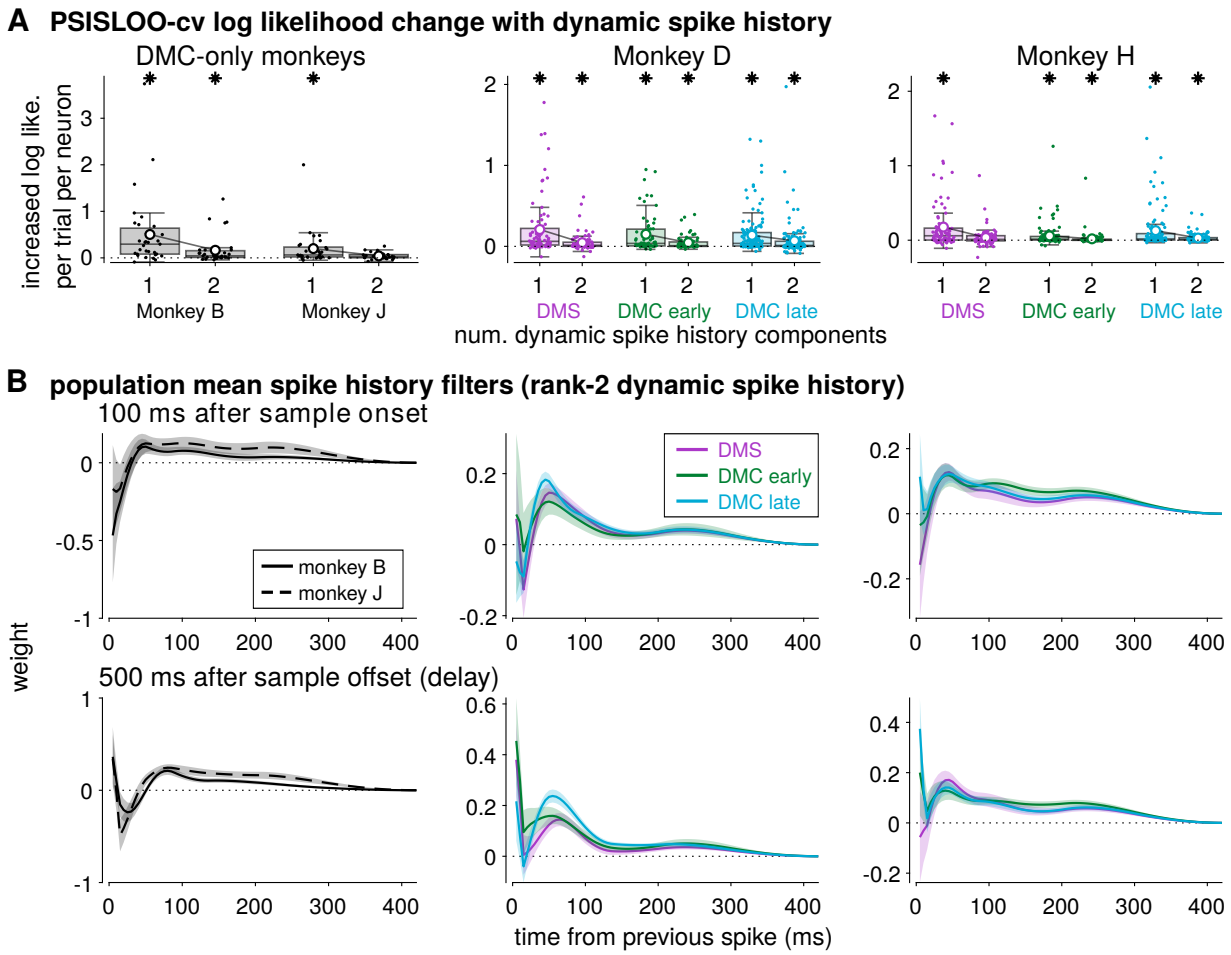

Figure S9: Including a dynamic spike history filter tensor improves model fit. (**A**) The mean change in cross-validated log likelihood per-trial for each neuron as a function of the number of components (i.e., the average improvement in predictive performance for adding an additional dynamic spike history component). The log likelihood for the rank-1 dynamic spike history is relative to the GMLM without any dynamic spike history (but still includes each individual neuron's static spike history filters). Leave-one-out cross-validation was estimated for each trial using Pareto-smoothed importance sampling (PSISLOO-cv). Stars denote a statistically significant improvement after including the dynamic spike history component ($p < 10^{-4}$, paired, one-sided Wilcoxin signed-rank test). The lines with white circle markers denote the average log likelihood per trial across neurons. While the improvement after including two dynamic spike history components was often statistically significant, it was less dramatic than the gain from a single component. (**B**) The mean population mean effective spike history filters for all eight LIP populations during sample stimulus presentation (top; 100 ms after stimulus onset) and during the delay period (bottom; 500 ms after stimulus offset). The spike history filters were computed as the MAP estimate of the GMLM with rank-2 dynamic spike history ($R_h = 2$). Error regions denote ± 2 SEM.