# An enhanced inverted encoding model for neural reconstructions

Paul S. Scotti, Jiageng Chen, & Julie D. Golomb

Department of Psychology, The Ohio State University, Columbus, Ohio, USA

Please address correspondence to:
Julie D. Golomb
The Ohio State University
Department of Psychology
1835 Neil Avenue
Columbus, OH 43210
Email: golomb.9@osu.edu

Abstract word count: 150
Main text word count: 4484
Figures: 5

## Abstract

Inverted encoding models (IEMs) have recently become a popular method for investigating neural representations by reconstructing the contents of perception, attention, and memory from neuroimaging data. However, the standard IEM procedure can produce spurious results and interpretation issues. Here we present a novel modification to IEMs ("enhanced inverted encoding modeling," eIEM) that addresses key issues inherent in the standard IEM procedure, improves the flexibility and interpretability of stimulus reconstructions, and provides trial-by-trial stimulus predictions and goodness-of-fit estimates. Our modifications are advantageous due to our decoding metric taking into account the choice of population-level tuning functions and employing a prediction error-based metric directly comparable across experiments. Our modifications also allow trial-by-trial confidence estimates independent of prediction error which can be used to threshold reconstructions and improve neural decoding performance and brain-behavior correlations. We validate the improved utility of eIEM across three fMRI datasets and offer a Python package for easy implementation.

## 1. Introduction

A mental representation can be defined as the "systematic relationship between features of the natural world and the activity of neurons in the brain"[1]. An increasingly common approach to study mental representations using neuroimaging data is to employ encoding models, which describe this relationship computationally, typically by reducing the complexity of the input data with a set of functions that, when combined, roughly approximate the neural signal. Encoding and decoding models (aka voxelwise modeling or stimulus-model based modeling) have become a standard method for investigating neural representational spaces and predicting stimulus-specific information from brain activity[2–4]. The key advantages of such models over other computational approaches such as multivariate pattern classification or representational similarity analysis are typically touted as the following: (1) Encoding models can take inspiration from single-unit physiology by consisting of tuning functions in stimulus space (aka feature space), allowing both the maximally receptive feature and the precision/sensitivity of the tuning to be estimated across a population of neurons; (2) The encoding model that transforms stimuli into brain activity can be inverted into a decoding model capable of predicting stimuli given a pattern of neural activity; and (3) The decoding model can predict novel stimuli or experimental conditions not used in the training of the model. The features of an encoding model can be anything from Gabor filters[5,6] to perceptual colors[7] to acoustic musical features[8] and even human faces[9].

The inverted encoding model (IEM) is one example of an encoding and decoding model that uses simple linear regression and a basis set representing the hypothesized population-level tuning functions, consisting of several channels that are modeled as

cosines (or von Mises) equally separated across stimulus space (e.g., orientation, color, spatial location). Due to its simplicity, robust performance, and grounding in single-unit physiology principles, inverted encoding models have quickly risen to prominence in the cognitive neuroscience community[10–39]. The basic idea behind IEMs is that each channel in the basis set can be assigned a weight per voxel (we adopt fMRI nomenclature here, but IEMs can be applied to any modality like EEG, MEG, etc.) and hence a model can be trained to predict the activity of each voxel using the weights of each channel as predictors (i.e., the regressors in a linear regression). Then, this trained encoder is inverted such that it becomes a decoder capable of predicting, or *reconstructing*, a trial's stimulus when provided with a novel set of voxel activations.

Here we present a novel modification to IEMs that improves the interpretability of stimulus reconstructions, addresses some key issues inherent in the standard IEM procedure, and provides trial-by-trial stimulus predictions and goodness-of-fit estimates. In the sections that follow, we briefly review the standard procedure used for implementing IEMs, present an overview of our modified procedure, and discuss methodological issues and limitations with the standard procedure that our procedure addresses. We then present results comparing the standard procedure to our procedure using three existing fMRI datasets. These results validate our approach and highlight its practical advantages in terms of improved flexibility and interpretability. We also offer a publicly available Python package for researchers to easily implement our IEM approach, which we subsequently refer to as "enhanced inverted encoding modeling" (eIEM).

## 1.1 Standard procedure for inverted encoding models

The standard IEM procedure is illustrated in Figure 1A-B using a toy example reconstructing stimulus colors.

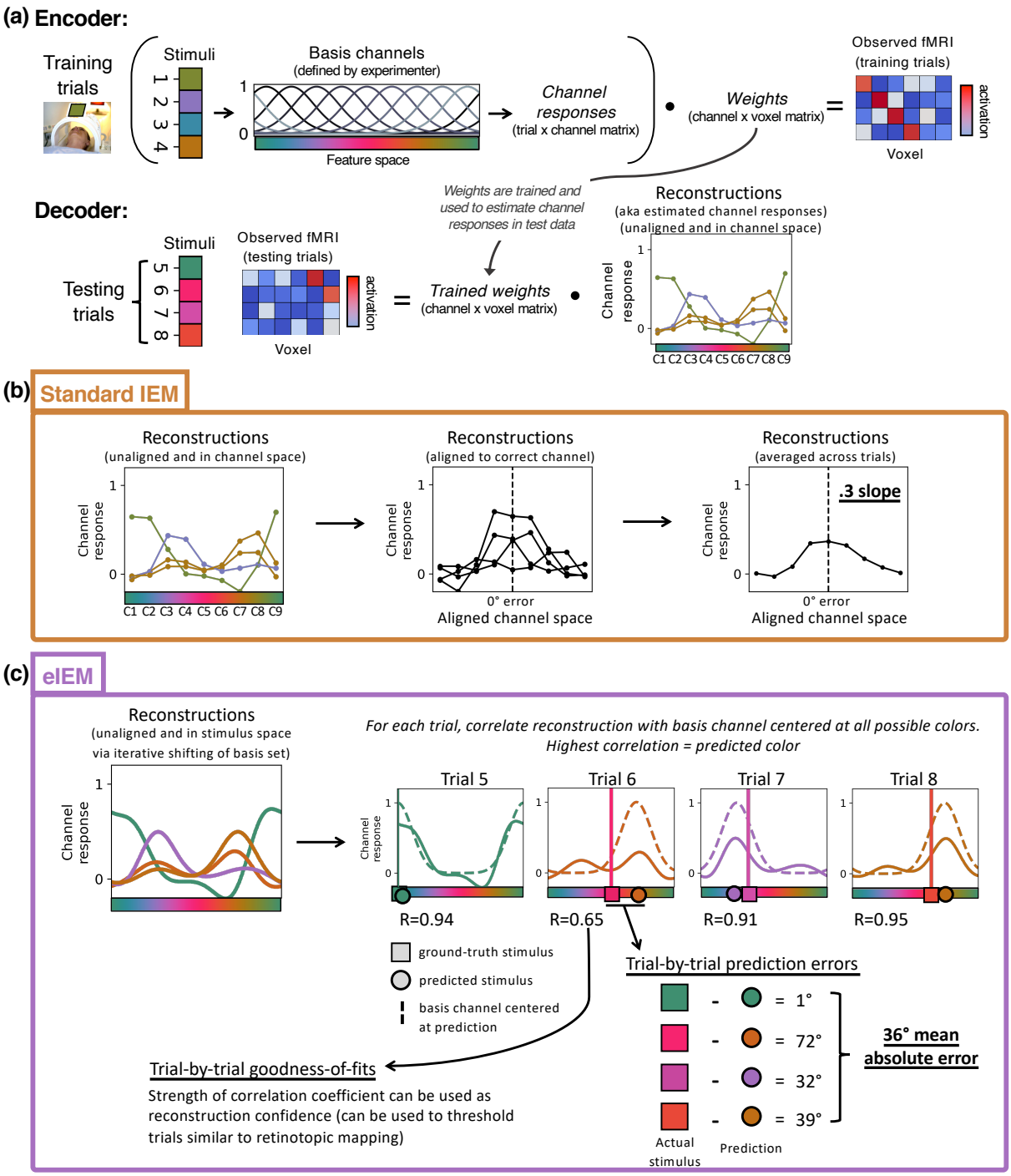

Figure 1. Simulated example of an fMRI experiment using the standard IEM procedure and our eIEM procedure, where a participant was shown eight trials of colored squares and the researcher used an IEM to decode the presented colors with a six-voxel brain region. The standard procedure is our depiction of the typical steps used to implement IEMs, but there may be nuances particular to specific papers that deviate from this procedure. (a) The top row depicts the encoding model where the weights matrix is estimated (via linear regression) and the second row depicts the decoding model where the channel responses for the test data are estimated using the trained weights from the encoding model. (b) The standard procedure involves aligning and averaging reconstructions and measuring the result according to a variety of possible metrics (e.g., amplitude, slope). (c) eIEM deviates from the standard procedure by evaluating prediction errors rather than an averaged reconstruction. We correlate the basis channel with reconstructions to estimate predicted stimuli, use iterative shifting of the basis set to allow channel space to equal stimulus space, and estimate goodness-of-fits for each trial reconstruction which can be used as a measure of confidence for each trial's predicted stimulus. For simplicity, this example shows the encoder trained on the first half of trials and the decoder used to predict the color of the remaining trials, but in most applications cross-validation should be used such that every trial may be decoded while avoiding circularity/double-dipping.

The prerequisites for implementing an IEM are (1) the feature value of the stimulus for every trial, (2) a trial-by-voxel matrix of brain activations[a] for every voxel per trial, and (3) a basis set representing hypothesized population-level tuning functions. Typically, researchers use a basis set of several equidistant channels modeled as cosines raised to the number of channels minus 1 (see Methods). The specification of the number of channels is mostly arbitrary (more channels are typically chosen if suspecting narrow tuning and vice-versa for broad tuning). The encoder models each voxel's response as the weighted sum of the channels. Given that the trial-by-voxel matrix and the basis set are already given, the weights matrix can be estimated via least-squares linear regression.

Once the weights matrix is estimated from the training dataset, it can be inverted such that the encoder becomes a decoder for the test dataset. Now, instead of

[a]Observed brain activations could be beta weights from general linear model estimation[40] or raw BOLD signal from a block or slow-event related design.

estimating the weights matrix via least-squares linear regression, the weights matrix and the trial-by-voxel matrix are given and the channel responses (i.e., reconstructions) are estimated. The resulting estimated channel responses, or simply *reconstructions*, is a trial-by-channel matrix where each trial has its own reconstruction composed of weighted cosines.

In the standard procedure, all the trial-by-trial reconstructions are then circularly shifted along the x-axis such that the channel that *should* have been maximally responsive on every trial (i.e., the channel closest to the ground truth stimulus feature) is aligned to the center of the x-axis (Figure 1B). The aligned reconstructions are then averaged together to result in a single reconstruction. Assuming that the model performs well, the averaged reconstruction should resemble the shape of the original basis channel with the highest point centered on the aligned location. The averaged and aligned reconstruction is then typically assessed using a number of possible metrics (see Figure 2), and then subjected to statistical analysis.

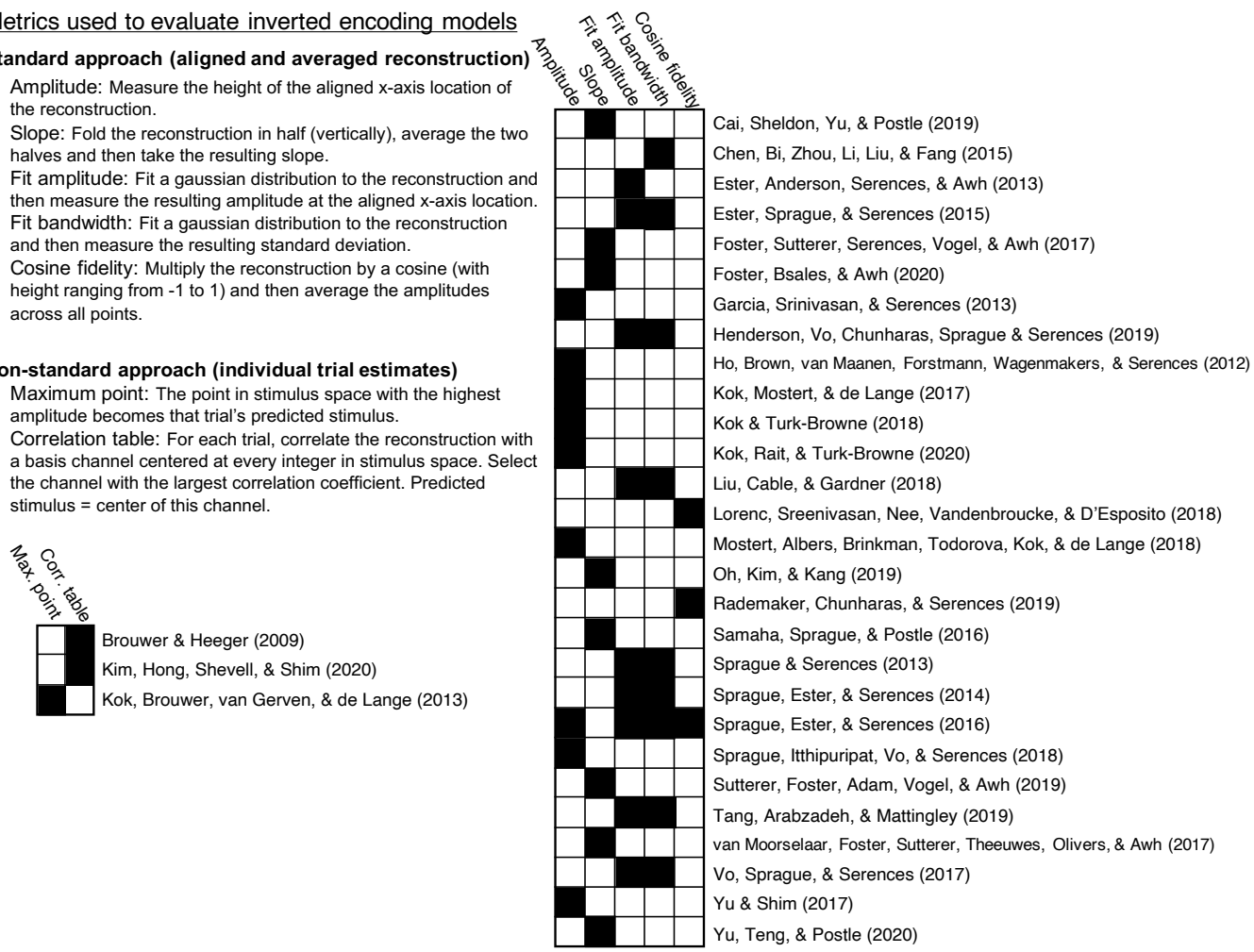


Figure 2. Summary of metrics used to evaluate IEM reconstructions in a sampling of published papers[10–39]. Note that the methodological concerns about spurious conclusions raised in Section 1.3.1 apply to the metrics labeled under the standard approach, although our proposed modifications pose improvements over typical applications of "maximum point" and "correlation table" approaches as well. The non-standard approaches can be quantified with a single value (like the standard approach metrics) by taking the mean absolute error between predicted and actual stimuli.

## 1.2 Overview of our enhanced inverted encoding modeling procedure

The core steps depicted in Figure 1A – using least-squares linear regression for estimating the channel weights (encoder) and estimating the channel responses (decoder), as well as the use of a basis set of hypothesized population-level tuning functions – remain the same between the standard procedure and our eIEM procedure.

eIEM (Figure 1C) differs from the standard procedure in three ways. As a brief overview, our first modification is to repeat the entire IEM procedure multiple times with slightly shifted basis sets such that reconstructions are in stimulus space rather than channel space. This iterative shifting modification has been employed in a few previous papers[18,38], however, it is not common practice. This step is important because otherwise stimulus predictions will be biased by the arbitrary placement of the basis channels, as described in section 1.3.2.

Second, and most critically: we evaluate reconstructions in terms of average *prediction error* instead of the aligned and averaged reconstruction metrics described above. eIEM obtains trial-by-trial stimulus predictions using the correlation table approach (the only approach for obtaining a decoding metric in Figure 2 that adapts to the shape of the basis channel, see section 1.3.1) and calculates trial-by-trial prediction error; this error is averaged across trials to obtain mean absolute error (MAE), which we recommend as a simple and interpretable metric. (Note that other error metrics are also possible in this framework, including signed error metrics if there is reason to expect asymmetric reconstructions.) This modification is critical because it resolves several methodological concerns inherent in the standard approach, described more below, and prediction error is an easily interpretable metric.

Our final modification takes the trial-by-trial prediction approach described above one step further. Using the correlation table approach to determine the best-fitting basis channel, the center of that best-fitting channel is taken as the stimulus prediction, but the goodness-of-fit (correlation coefficient) values themselves can also be optionally leveraged to estimate trial-by-trial confidence of predictions (see section 1.3.3). This

modification adds substantial flexibility to the IEM procedure; e.g., allowing for thresholding reconstructions to potentially increase statistical power and performance, as we demonstrate in the Results. We discuss these modifications in more depth in the subsequent sections, while highlighting the advantages of eIEM over the standard procedure.

### 1.3 Value of our eIEM approach over the standard IEM approach

As summarized above, the eIEM approach that we present here is a combination of several modifications and improvements on the standard IEM procedure. The value of these modifications is primarily in terms of evaluating IEM results: We propose that eIEM is better than the standard IEM approach in terms of improved interpretability, flexibility, and robustness to methodological concerns. We also offer eIEM as a standardized set of “best practices”. As depicted in Figure 2, various approaches exist for evaluating IEMs and often researchers report several decoding metrics due to ambiguity over which metric is best. eIEM is the combination of specific practices (some previously employed, some novel) intended to offer a preferred solution.

Below we describe several methodological concerns and limitations of the standard IEM procedure that are addressed by eIEM. (In the Results section, we further demonstrate the appeal of our approach in terms of improved flexibility and interpretability.)

#### 1.3.1 Standard procedure can produce misleading or difficult to interpret results

The standard procedure is susceptible to inappropriate decoding evaluations, largely due to the align-and-average step. Aligning and averaging across trial reconstructions loses information that is important for evaluating decoding performance and can be prone to heavy bias from outliers. Moreover, the metrics used to evaluate the aligned and averaged reconstructions are not easily interpretable.

As depicted in Figure 3, averaging can obscure important information present in trial reconstructions. Panels 3a and 3b would be interpreted identically according to the standard procedure even though one example shows every channel correctly predicted (i.e., predicting the correct stimulus feature with minimal error) and the other example shows every channel incorrectly predicted (large errors). eIEM using MAE would correctly identify the first case as demonstrating superior decoding performance. The takeaway here is that averaging across prediction errors, and not across trial reconstructions, avoids the pitfall of interpreting Figures 3a and 3b as reflecting the same level of stimulus-specific brain signal despite clear support for Figure 3a demonstrating improved decoding on a trial-by-trial level.

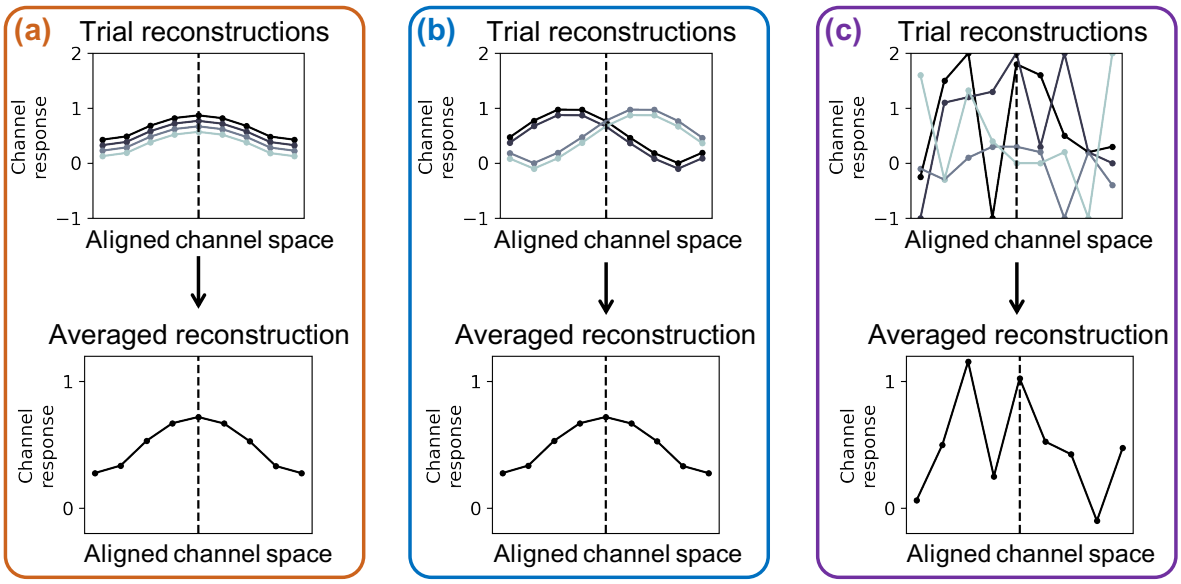


Figure 3. Cartoon example depicts some problems with the standard procedure of evaluating the aligned and averaged reconstruction and using a decoding metric that does not consider the shape of the basis channel or the variability of trial reconstructions. For each of the 3 simulated data examples, the top row depicts four single-trial reconstructions, and the bottom row depicts the aligned-and-averaged reconstruction. In (a) each individual trial's reconstruction accurately predicts the correct channel (i.e., the correct stimulus feature), appropriately reflected in the averaged reconstruction. In (b) each individual trial's reconstruction predicts an incorrect channel. Averaging across trials leads to a misleading result, i.e., the standard approach would consider (b) to reflect the same level of decoding performance as (a). In (c) each individual trial's reconstruction is essentially noise, such that the averaged reconstruction results in a false peak around the aligned point; the standard procedure using align-and-average metrics would result in spuriously superior decoding performance than both (a) and (b), with (c) having a higher amplitude, steeper slope, and narrower standard deviation when fit with a gaussian distribution. The eIEM procedure, calculating MAE from trial-wise prediction error, would correctly conclude that case (a) shows the best decoding performance.

MAE is also less prone to bias from outlier reconstructions compared to any of the align-and-average metrics. In the standard procedure, a single outlier reconstruction can disproportionately bias the averaged reconstruction, potentially completely flipping the averaged reconstruction in the most extreme cases. In contrast, the influence of an outlier is naturally capped for MAE: Consider an experiment composed of 300 trials where 299 trials predict the correct stimulus and one trial predicts the stimulus 180° away (assuming 360° stimulus space); the outlier could only increase MAE by a maximum 0.6°.

1.3.2 Standard procedure does not account for the shape of the basis channels

IEMs produce reconstructions that depend on the choice of basis set[15,41]. The decoding metrics commonly used to evaluate averaged reconstructions in the standard IEM procedure, however, do not take this observation into account. That is, intuition – and standard practice – wrongly assume that a monotonic relationship exists between decoding metrics such as slope, amplitude, and bandwidth and a greater amount of stimulus-specific information in the brain signal. If the basis set consists of identical basis channels, then a perfect reconstruction returns the shape of the basis channel, and so it makes sense to compare the shape of the reconstruction to the shape of the basis channel to make predictions and evaluations. The correlation table approach employed in eIEM leverages this observation to provide the most direct relationship between IEM performance and stimulus-specific brain signal.

The correlation table approach operates as follows and has been previously used in a small number of papers[7,38]. For each trial, compute a set of correlation coefficients, each reflecting the correlation between that trial's reconstruction and a canonical basis channel (i.e., a perfect reconstruction) centered at every integer in stimulus space (e.g., resulting in 360 correlation coefficients for a stimulus space ranging from 0-359°). The highest of these correlation coefficients is determined to be the best fit for that trial, and the predicted stimulus feature for that trial is simply the center of that best-fitting basis channel. In this manner, the predictions obtained from the correlation table metric

automatically adjust to consider the shape of the basis channel because it is the basis channel itself that is being used to obtain predictions.

Simply put, amplitude, slope, bandwidth, etc. are inferior metrics compared to the correlation table metric because they do not adapt to the choice of basis set. This is true even if a researcher were to skip the align-and-average step by evaluating reconstructions at the trial-by-trial level. For instance, using the amplitude metric, a higher amplitude at the aligned point is thought to reflect improved performance. If the basis channel ranges from 0 to 1, a perfect reconstruction should have an amplitude of exactly 1 at the aligned point, but reconstructions can feasibly have amplitudes far greater than 1. Such a problem is demonstrated visually in Figure 3 where it is clear that Figure 3c looks to be a worse reconstruction than Figure 3a, but standard IEM metrics would produce spuriously high values for this case.

It is possible to partially account for the shape of the basis channel by, for example, fitting the reconstruction with a gaussian distribution[13]. However, such fitting procedures may be problematic because such a procedure forces the reconstruction to appear to be a reasonable gaussian shape regardless of the data (e.g., fitting Figure 3c with a gaussian distribution would still lead to the same incorrect conclusion of superior decoding performance compared to Figure 3a).

Given the various ways to measure IEM performance listed in Figure 2, the correlation table approach best takes the shape of the basis set into account, but one limitation is that the basis channels are in stimulus space, but the reconstructions are in channel space. That is, to properly correlate the shape of the basis channel to the reconstruction, one must linearly interpolate between points in channel space. eIEM

solves this limitation by employing iterative shifting[18,38]. By repeatedly fitting the encoding model with every possible (circular) shift of the basis set and then combining all of these iterations together, a fuller reconstruction is obtained that is no longer impoverished by a limited number of *num_channels* points (i.e., the range of channel space becomes equal to the range of stimulus space).

The iterative shifting procedure also aids more generally in producing more interpretable and less biased reconstructions, as illustrated in Figure 4. Iterative shifting of the basis set is especially important because our decoding model must be capable of predicting any possible feature in stimulus space (that is, not solely the stimuli that are located at the centers of the basis channels) if we want to obtain accurate trial-by-trial stimulus predictions. Note that iterative shifting does not change the fact that different basis sets result in different reconstructions, rather, it simply allows for the most accurate reconstruction given a set number of channels with defined bandwidths.

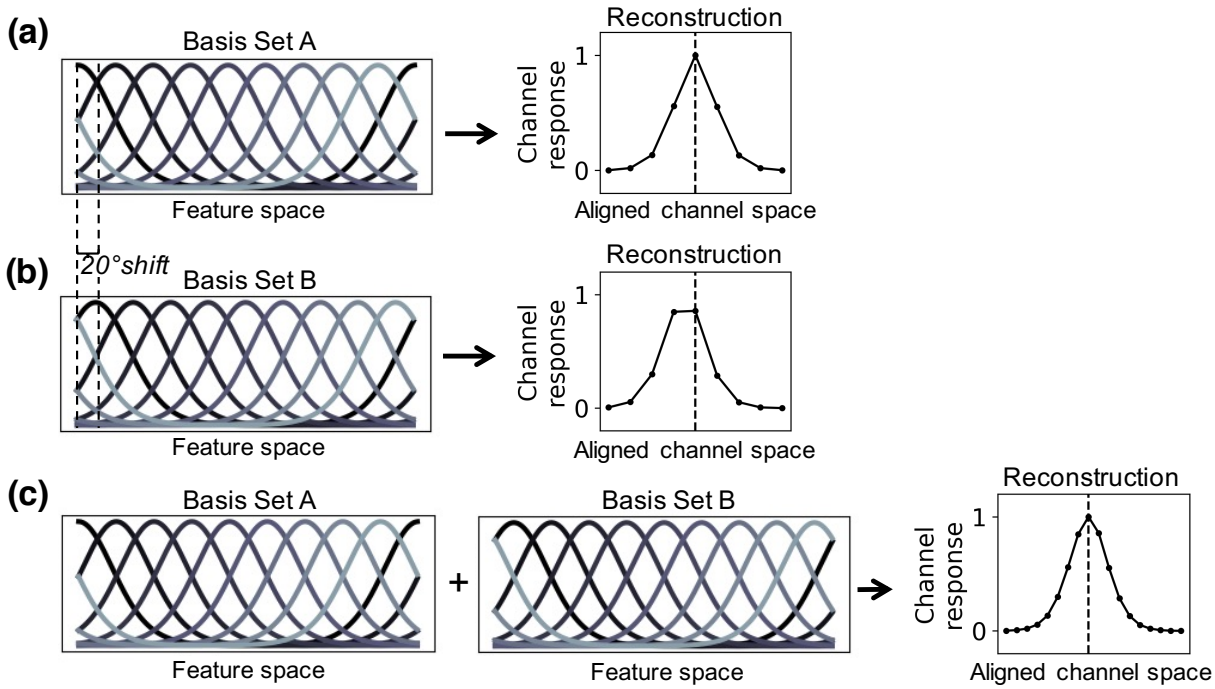


Figure 4. Simulated data depicting how a slightly altered basis set (means shifted by 20°) can alter reconstructions, even if the same signal is present in all cases. Here the trial by voxel activations reflect perfect (zero noise) information with identical train and test sets, such that the resulting reconstructions *should* also be perfect. (a) Basis set perfectly reflects the underlying voxel tuning functions (simulated ground truth). (b) Reconstruction of the same data, now with basis set of channels circularly shifted 20°. (c) By combining the results of both basis sets, the channel space changes from *num_channels* to *num_channels**2, leading to a fuller reconstruction. Iterative shifting in eIEM repeats this procedure for all possible shifts of the basis set to make the channel space equal the stimulus space, decreasing variation and allowing the correlation table metric to be optimally applied. The code to reproduce this figure from simulated data can be found at https://osf.io/et7m2/ (also contains code for reproducing Figures 1 and 3).

### 1.3.3 Standard procedure lacks a measure of decoding uncertainty

Another limitation of the standard approach resolved by eIEM is that the standard IEM procedure does not incorporate uncertainty into decoding performance. Individual trials can vary substantially in signal quality, driven by factors including attentional fluctuations, alertness, head motion, and scanner noise. Noisier trials could potentially obscure an underlying signal, but as exemplified in Figure 3, highly variable trial reconstructions are not weighted differently from robust reconstructions according to the

standard procedure. The lack of uncertainty information has been noted in other contexts, with some recent alternatives to IEM proposed to incorporate uncertainty[42,43]. However, here we demonstrate that eIEM can easily and automatically produce a trial-by-trial measure of prediction uncertainty within the IEM framework itself, which can then be used in flexible and accessible ways.

The correlation table approach produces a best-fitting stimulus prediction—and associated goodness-of-fit value (correlation coefficient)—for each trial. We propose that the correlation coefficient of the best-fitting basis channel can be used as a proxy to estimate the confidence, or reliability, of trial-by-trial predictions. It is important to emphasize that the correlation coefficient reflects the degree to which the reconstruction matches the *best-fitting* basis channel, not the basis channel centered on the correct stimulus. In other words, this goodness-of-fit information is obtained independently and prior to any calculation of prediction error.

This trial-by-trial prediction uncertainty information could be used in several ways. One suggestion we put forth is that goodness-of-fit can be used to threshold reconstructions, such that worse-fitting trials may be excluded from analysis. This principle is analogous to the phase-encoded retinotopic mapping and population receptive field modeling techniques, where a set of models spanning the full stimulus space is evaluated for every voxel, and the parameters of the best-fitting model are selected as that voxel's preferred stimulus, with the goodness-of-fit values then used to threshold the results[44–46].

In the Results section we provide a proof of concept using real fMRI data to demonstrate the utility of using trial-by-trial goodness-of-fit values to threshold IEMs

based on confidence. Note that although r-squared is the more commonly used statistic for goodness-of-fit using regression, squaring the correlation coefficient is not preferred here because the sign of the correlation coefficient is informative (e.g., a perfectly inverted reconstruction should *not* be assigned equal confidence as a perfect reconstruction), so we recommend the use of the r-statistic.

## 2. Results

To validate eIEM and demonstrate its practical advantages, we implemented standard IEM and eIEM across three real fMRI datasets. The three datasets[13,18,47] span the research topics of perception, attention, and memory. We demonstrate how eIEM improves flexibility (potential to exclude low-confidence trial reconstructions) and interpretability (reconstruction performance in terms of prediction error) while avoiding the methodological pitfalls discussed in the previous section. See the Online Methods for information regarding each dataset and how data were processed.

### 2.1 Validating eIEM on real fMRI data

Figure 5 shows results from the standard IEM and eIEM procedures. First, to validate our method, we confirmed that across all three datasets eIEM replicated the overall pattern of results obtained with the standard procedure (for consistency, we reanalyzed all datasets ourselves using a standard procedure quantified by slope, even if that was not the metric used in the original paper). In the Perception dataset[13], we used both techniques to decode the horizontal position of a stimulus in V1, V4, and IPS. The standard procedure revealed significant decoding performance in all 3 ROIs, with the strongest decoding (greatest slope) in V1, followed by V4, and then IPS. eIEM

replicated this pattern. In the Attention dataset[47], we used both techniques to decode the attended orientation within a multi-item, multi-feature stimulus array, in the same three ROIs. The standard IEM procedure revealed significant decoding in V1 and V4, but not IPS. eIEM again replicated this pattern. Finally, in the Memory dataset[18], we used both techniques to decode the remembered orientation of a stimulus over two types of working memory delays: blank delay and distractor delay. With the standard procedure, the remembered orientation could be successfully decoded in V1, V4, and IPS, with significantly greater decoding in V1 and V4 during the blank delay compared to the distractor delay. With eIEM, we replicated each of those results, with the additional finding of significantly greater decoding during the blank vs distractor delay in IPS.

The important takeaways from these validations are that (1) eIEM is not susceptible to the methodological concerns raised earlier that plague the standard procedure, and yet (2) the results from the two techniques produce largely consistent patterns across these datasets. We note that these three datasets were useful validation cases because they contained robust results (as may be more likely with published, publicly available datasets); however, we would not necessarily expect the eIEM and standard IEM procedures to always produce consistent patterns, especially in cases where data are less robust and therefore more susceptible to the aforementioned methodological concerns. In those cases, we argue that eIEM offers a more accurate reflection of decoding performance, as illustrated in Figure 3.

### 2.2 Demonstrating the improved flexibility and interpretability of eIEMs

Having validated our method across three diverse fMRI datasets, we next use these same datasets to illustrate the practical advantages of eIEM. First, we obtained metrics that are easily interpretable and comparable across datasets due to decoding performance being measured in terms of prediction error rather than arbitrary units. For example, in the Memory dataset, the standard procedure results in an average slope of .006 for the blank delay condition and .004 for the distractor delay condition in V1 (or cosine fidelity values of .100 and .098 as reported in the original paper[18]); eIEM replicates this pattern, but now with a more interpretable and meaningful metric: orientation can be decoded with an average error of 27.9 degrees in the blank delay and 32.5 degrees in the distractor condition.

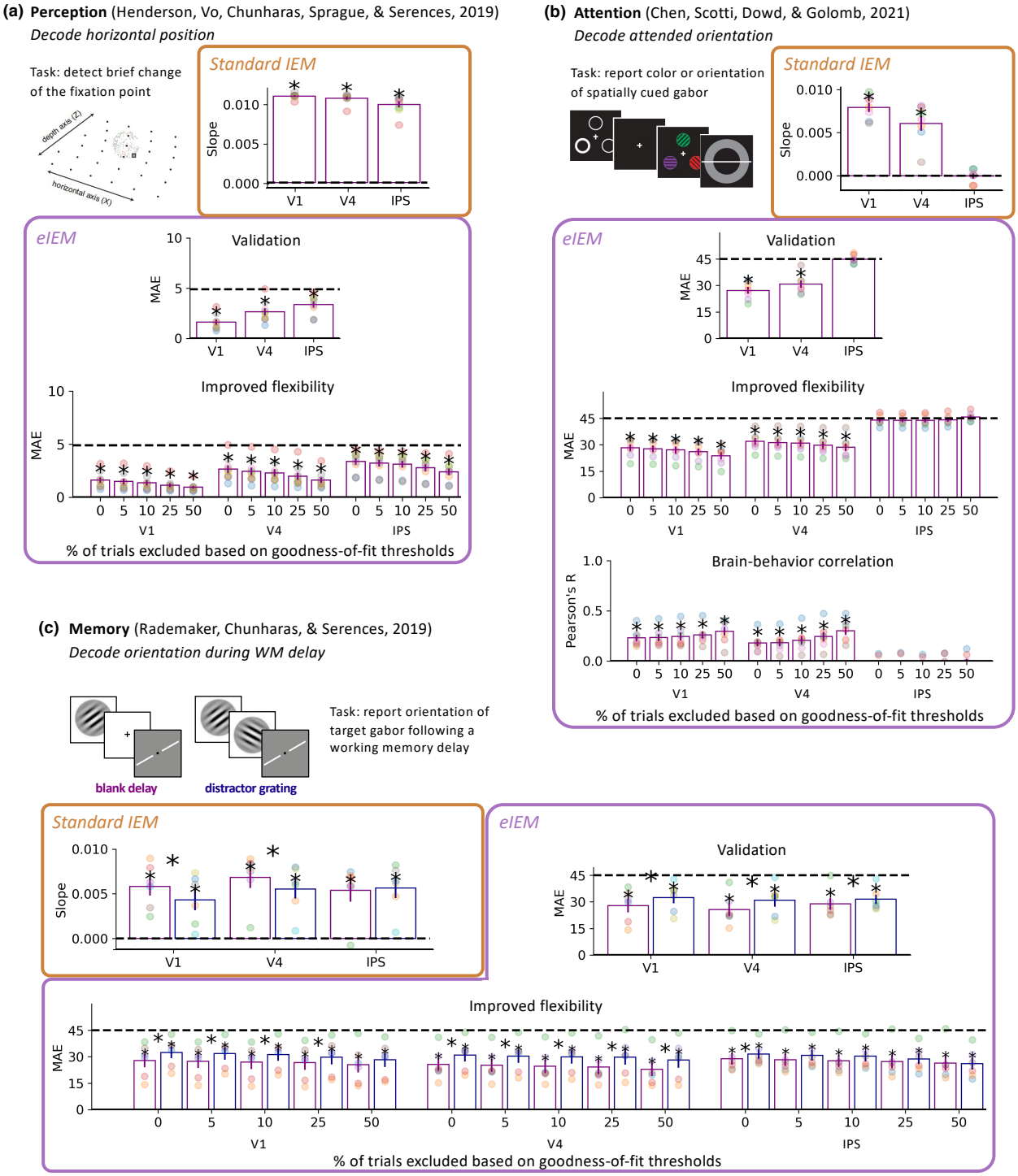


Figure 5. Results of the standard IEM procedure and our eIEM procedure across three real fMRI datasets spanning the topics of perception (a), attention (b), and memory (c). For each dataset, results are plotted obtained from the standard procedure (orange boxes), eIEM with full data (purple boxes, validation plots), and eIEM using increasingly stringent cutoffs based on goodness-of-fit (purple boxes, improved flexibility plots). Bar plots depict the average slope (standard procedure; higher values are better) and MAE (eIEM; lower values are better) across

subjects, with individual subjects overlaid as colored dots. For the Attention dataset, the brain-behavior correlation plot additionally plots the trial-by-trial correlation between absolute behavioral error and absolute decoding error for each ROI and goodness-of-fit threshold. Error bars depict standard error of the mean, dotted black lines represent chance decoding, and asterisks represent statistically significant decoding ($p<.05$). Overall results show that conclusions are similar between the standard IEM and eIEM, but MAE is more interpretable (not based in arbitrary units) and not prone to methodological concerns discussed in the Introduction. In addition, each dataset showed that MAE consistently improved with increasing exclusion thresholds, demonstrating the flexibility of goodness-of-fit to exclude noisy trials. Increasing exclusion thresholds also appeared to strengthen brain-behavior correlations in the Attention dataset. See Online Methods for additional information.

Next, we tested the flexibility of eIEM to make use of the trial-by-trial goodness-of-fit information. For each trial, eIEM produces a predicted stimulus, associated prediction error, and a goodness-of-fit value. The goodness-of-fit value is a measure of how well the predicted stimulus fits an ideal basis function centered *at that predicted value*. That is, it is a measure of the confidence of that prediction, not the accuracy of the prediction, and so is obtained independently of prediction error. To test the impact of using goodness-of-fit information on decoding performance, we performed an analysis where we excluded trials with the lowest 5%, 10%, 25%, and 50% of goodness-of-fit values (Figure 5). This resulted in visible improvements in MAE with increasing exclusion thresholds for all three datasets (linear regression revealed significant negative slope in all cases except for IPS in the Attention dataset). Notably, in the Attention dataset, MAE improved with increasing thresholds in V1 and V4 (where decoding was significant in the unthresholded analysis) but not in IPS (where decoding was at chance in the unthresholded analysis). Thus, the goodness-of-fit information can be used to improve decoding performance when a brain region contains reliable information about a stimulus, but does not produce false positives in the absence of observable stimulus-specific brain activity.

Finally, having trial-by-trial prediction error and goodness-of-fit values lends itself to analyses correlating neural measures with behavior. We demonstrate this in the Attention dataset (the Perception dataset did not collect behavioral responses, and in the Memory dataset behavioral performance was too close to ceiling [~3° avg. error]). In V1 and V4, we observed a significant correlation between a trial's behavioral error magnitude and neural prediction error. Moreover, the strength of these correlations increased with higher goodness-of-fit thresholds (Figure 5). We note that behavioral error itself did not noticeably change across these thresholds, suggesting that the goodness-of-fit information seemed to be reflecting noise at the level of the fMRI signal, not simply fluctuations in behavior or cognitive focus.

Altogether, these findings suggest that not only does eIEM produce more interpretable and robust results, but the trialwise goodness-of-fit values offer increased flexibility to improve both neural decoding power and brain-behavior correlations.

## 3. Discussion

Inverted encoding modeling has become a popular method for predicting stimuli and investigating neural representations because of its robust performance, simplicity of linear modeling, ability to predict untrained classes, and grounding in single-unit physiology. eIEM improves the flexibility and interpretability of results while fixing important methodological concerns surrounding the standard IEM procedure, namely how the standard procedure ignores trial-by-trial variability, does not account for the fact that a perfect reconstruction returns the basis channel, and cannot leverage uncertainty in its evaluations. The practical advantages of our method are made tangible by

comparing the results of the standard IEM and eIEM across three fMRI datasets, highlighting the wide range of applications intended for eIEMs.

Importantly, our method can increase statistical power and IEM performance by leveraging uncertainty in model fits. Researchers have the flexibility to exclude trials with noisier reconstructions based on how similar in shape each reconstruction is to the basis channel at the predicted stimulus. Note that we do not prescribe a specific cutoff for determining goodness-of-fit thresholds in this paper, rather, we simply offer that such an approach is possible for improving IEM performance. For example, a researcher could weight trials with higher confidences more heavily or simply decide to exclude the noisiest 20% of trials.

Our method also improves interpretability by evaluating reconstructions in terms of prediction error. For example, “V1 showed 10° average prediction error and V4 showed 20° average prediction error” is more interpretable than “V1 showed .02 amplitude and V4 showed .01 amplitude” because the latter is in arbitrary units, whereas MAE is in meaningful units. Further, unlike amplitude or slope, the magnitude of prediction error is not dependent on the choice of basis set and can be directly compared to other experiments using the same stimulus space.

We demonstrated the above two advantages using three fMRI datasets. Our validations further demonstrated how our IEM approach can be applied to both circular and non-circular stimulus spaces, is sensitive to variations in decoding performance across brain regions and experimental conditions, can be used to accurately decode the contents of perception, attention, and working memory, and can be used to meaningfully link neural reconstructions with behavior. Our modifications allowed for the

decoding performance of each dataset to be directly compared to each other and demonstrated how uncertainty, measured via goodness-of-fit, can be leveraged to increase statistical power. Note that just because these three datasets produced consistent overall results (in terms of significance testing) across procedures does not ensure this will always be the case—for less reliable results, the methodological pitfalls discussed in the Introduction become increasingly problematic.

In this paper we have referred to IEMs as a specific kind of encoding and decoding model that involves simple linear regression with population-level tuning functions. There are more complex neuroimaging methods that can similarly be used to produce reconstructions via hypothesized tuning functions. For instance, Kay et al. (2008) decoded natural images from brain activity via voxel-level receptive field models that describe tuning functions across space, orientation, and spatial frequency. Naselaris et al. (2009) further produced Bayesian reconstructions of natural images via the combination of encoding models meant to estimate structural and semantic content. Van Bergen and colleagues[43,48,49] introduced models where voxels with similar tuning account for shared noise and which produce trial-by-trial probability distributions such that uncertainty can be obtained similarly to our procedure (although the researchers discuss this in terms of testing Bayesian theories of neural computation rather than trial thresholding). An advantage of eIEMs is that improvements to the standard IEM approach are accomplished without sacrificing simplicity—the encoding model weights and the decoding model channel responses are simply estimated via least-squares estimation.

Inverted encoding modeling has become increasingly popular in recent years, and yet the proper method for evaluating IEMs has become increasingly uncertain. As depicted in Figure 2, researchers often report IEM performance according to several metrics because of a lack of consensus regarding the “correct” way to evaluate reconstructions. Other decoding techniques in neuroimaging, such as support vector machines or neural networks, use the easily interpretable metric of classification performance (% correct), but IEMs are typically evaluated in terms of arbitrary units that are abstracted away from the stimulus space they were intended to predict. We demonstrate a clear and practical advantage for evaluating reconstructions according to our method: researchers can increase their statistical power via thresholding, compare decoding performance across varying basis sets, evaluate performance in stimulus space, and obtain concrete stimulus predictions (with corresponding goodness-of-fits) for every trial rather than rely on a summary statistic based in arbitrary units. While there already exist approaches capable of addressing some of these concerns, our approach represents a suite of best practices that can be adopted by researchers in future work. Researchers can easily implement our procedure with one line of code using our accessible Python package (https://pypi.org/project/inverted-encoding; see Online Methods).

## Funding

This work was funded by the National Institutes of Health (R01-EY025648 to JDG) and the National Science Foundation (NSF DGE-1343012 to PSS, NSF BCS-1848939 to JDG).

## References

1. Poldrack, R. A. The physics of representation. *Synthese* 1–19 (2020).
2. Naselaris, T., Kay, K. N., Nishimoto, S. & Gallant, J. L. Encoding and decoding in fMRI. *Neuroimage* **56**, 400–410 (2011).
3. Naselaris, T. & Kay, K. N. Resolving ambiguities of MVPA using explicit models of representation. *Trends Cogn. Sci.* **19**, 551–554 (2015).
4. Popov, V., Ostarek, M. & Tenison, C. Practices and pitfalls in inferring neural representations. *NeuroImage* **174**, 340–351 (2018).
5. Kay, K. N., Naselaris, T., Prenger, R. J. & Gallant, J. L. Identifying natural images from human brain activity. *Nature* **452**, 352–355 (2008).
6. Naselaris, T., Prenger, R. J., Kay, K. N., Oliver, M. & Gallant, J. L. Bayesian reconstruction of natural images from human brain activity. *Neuron* **63**, 902–915 (2009).
7. Brouwer, G. J. & Heeger, D. J. Decoding and reconstructing color from responses in human visual cortex. *J. Neurosci.* **29**, 13992–14003 (2009).
8. Casey, M., Thompson, J., Kang, O., Raizada, R. & Wheatley, T. Population codes representing musical timbre for high-level fMRI categorization of music genres. in *Machine Learning and Interpretation in Neuroimaging* 34–41 (Springer, 2012).

9. Lee, H. & Kuhl, B. A. Reconstructing perceived and retrieved faces from activity patterns in lateral parietal cortex. *J. Neurosci.* **36**, 6069–6082 (2016).

10. Ester, E. F., Sprague, T. C. & Serences, J. T. Parietal and frontal cortex encode stimulus-specific mnemonic representations during visual working memory. *Neuron* **87**, 893–905 (2015).

11. Ester, E. F., Anderson, D. E., Serences, J. T. & Awh, E. A neural measure of precision in visual working memory. *J. Cogn. Neurosci.* **25**, 754–761 (2013).

12. Foster, J. J., Bsales, E. M. & Awh, E. Covert spatial attention speeds target individuation. *J. Neurosci.* **40**, 2717–2726 (2020).

13. Henderson, M., Vo, V., Chunharas, C., Sprague, T. & Serences, J. Multivariate analysis of BOLD activation patterns recovers graded depth representations in human visual and parietal cortex. *Eneuro* **6**, (2019).

14. Kok, P., Rait, L. I. & Turk-Browne, N. B. Content-based dissociation of hippocampal involvement in prediction. *J. Cogn. Neurosci.* **32**, 527–545 (2020).

15. Liu, T., Cable, D. & Gardner, J. L. Inverted encoding models of human population response conflate noise and neural tuning width. *J. Neurosci.* **38**, 398–408 (2018).

16. Oh, B.-I., Kim, Y.-J. & Kang, M.-S. Ensemble representations reveal distinct neural coding of visual working memory. *Nat. Commun.* **10**, 1–12 (2019).

17. Yu, Q., Teng, C. & Postle, B. R. Different states of priority recruit different neural representations in visual working memory. *PLoS Biol.* **18**, e3000769 (2020).

18. Rademaker, R. L., Chunharas, C. & Serences, J. T. Coexisting representations of sensory and mnemonic information in human visual cortex. *Nat. Neurosci.* **22**, 1336–1344 (2019).

19. Sprague, T. C. & Serences, J. T. Attention modulates spatial priority maps in the human occipital, parietal and frontal cortices. *Nat. Neurosci.* **16**, 1879–1887 (2013).

20. Sutterer, D. W., Foster, J. J., Adam, K. C., Vogel, E. K. & Awh, E. Item-specific delay activity demonstrates concurrent storage of multiple active neural representations in working memory. *PLoS Biol.* **17**, e3000239 (2019).

21. Cai, Y., Sheldon, A. D., Yu, Q. & Postle, B. R. Overlapping and distinct contributions of stimulus location and of spatial context to nonspatial visual short-term memory. *J. Neurophysiol.* **121**, 1222–1231 (2019).

22. Chen, N. *et al.* Sharpened cortical tuning and enhanced cortico-cortical communication contribute to the long-term neural mechanisms of visual motion perceptual learning. *Neuroimage* **115**, 17–29 (2015).

23. Foster, J. J., Sutterer, D. W., Serences, J. T., Vogel, E. K. & Awh, E. Alpha-band oscillations enable spatially and temporally resolved tracking of covert spatial attention. *Psychol. Sci.* **28**, 929–941 (2017).

24. Garcia, J. O., Srinivasan, R. & Serences, J. T. Near-real-time feature-selective modulations in human cortex. *Curr. Biol.* **23**, 515–522 (2013).

25. Ho, T. *et al.* The optimality of sensory processing during the speed–accuracy tradeoff. *J. Neurosci.* **32**, 7992–8003 (2012).

26. Kok, P. & Turk-Browne, N. B. Associative prediction of visual shape in the hippocampus. *J. Neurosci.* **38**, 6888–6899 (2018).

27. Lorenc, E. S., Sreenivasan, K. K., Nee, D. E., Vandenbroucke, A. R. & D'Esposito, M. Flexible coding of visual working memory representations during distraction. *J. Neurosci.* **38**, 5267–5276 (2018).

28. Mostert, P. *et al.* Eye movement-related confounds in neural decoding of visual working memory representations. *Eneuro* **5**, (2018).

29. Samaha, J., Sprague, T. C. & Postle, B. R. Decoding and reconstructing the focus of spatial attention from the topography of alpha-band oscillations. *J. Cogn. Neurosci.* **28**, 1090–1097 (2016).

30. Sprague, T. C., Ester, E. F. & Serences, J. T. Reconstructions of information in visual spatial working memory degrade with memory load. *Curr. Biol.* **24**, 2174–2180 (2014).

31. Sprague, T. C., Ester, E. F. & Serences, J. T. Restoring latent visual working memory representations in human cortex. *Neuron* **91**, 694–707 (2016).

32. Sprague, T. C., Itthipuripat, S., Vo, V. A. & Serences, J. T. Dissociable signatures of visual salience and behavioral relevance across attentional priority maps in human cortex. *J. Neurophysiol.* **119**, 2153–2165 (2018).

33. van Moorselaar, D. *et al.* Spatially selective alpha oscillations reveal moment-by-moment trade-offs between working memory and attention. *J. Cogn. Neurosci.* **30**, 256–266 (2018).

34. Vo, V. A., Sprague, T. C. & Serences, J. T. Spatial tuning shifts increase the discriminability and fidelity of population codes in visual cortex. *J. Neurosci.* **37**, 3386–3401 (2017).

35. Yu, Q. & Shim, W. M. Temporal-order-based attentional priority modulates mnemonic representations in parietal and frontal cortices. *Cereb. Cortex* **29**, 3182–3192 (2019).

36. Kok, P., Mostert, P. & De Lange, F. P. Prior expectations induce prestimulus sensory templates. *Proc. Natl. Acad. Sci.* **114**, 10473–10478 (2017).

37. Tang, M. F., Arabzadeh, E. & Mattingley, J. B. Forward modelling reveals dynamics of neural orientation tuning to unconscious visual stimuli during binocular rivalry. *bioRxiv* (2019) doi:10.1101/574905.

38. Kim, I., Hong, S. W., Shevell, S. K. & Shim, W. M. Neural representations of perceptual color experience in the human ventral visual pathway. *Proc. Natl. Acad. Sci.* **117**, 13145–13150 (2020).

39. Kok, P., Brouwer, G. J., van Gerven, M. A. & de Lange, F. P. Prior expectations bias sensory representations in visual cortex. *J. Neurosci.* **33**, 16275–16284 (2013).

40. Mumford, J. A., Turner, B. O., Ashby, F. G. & Poldrack, R. A. Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. *Neuroimage* **59**, 2636–2643 (2012).

41. Sprague, T. C., Boynton, G. M. & Serences, J. T. The importance of considering model choices when interpreting results in computational neuroimaging. *Eneuro* **6**, (2019).

42. Li, H.-H., Sprague, T. C., Yoo, A., Ma, W. J. & Curtis, C. E. Joint representation of working memory and uncertainty in human cortex. *bioRxiv* (2021) doi:10.1101/2021.04.05.438511.

43. van Bergen, R. S. & Jehee, J. F. TAFKAP: An improved method for probabilistic decoding of cortical activity. *bioRxiv* (2021) doi:10.1101/2021.03.04.433946.

44. Dumoulin, S. O. & Wandell, B. A. Population receptive field estimates in human visual cortex. *Neuroimage* **39**, 647–660 (2008).

45. Engel, S. A., Glover, G. H. & Wandell, B. A. Retinotopic organization in human visual cortex and the spatial precision of functional MRI. *Cereb. Cortex N. Y. NY 1991* **7**, 181–192 (1997).

46. Sereno, M. I. *et al.* Borders of multiple visual areas in humans revealed by functional magnetic resonance imaging. *Science* **268**, 889–893 (1995).

47. Chen, J., Scotti, P. S., Dowd, E. W. & Golomb, J. D. Neural Representations of Task-relevant and Task-irrelevant Features of Attended Objects. *bioRxiv* (2021) doi:10.1101/2021.05.21.445168.

48. Van Bergen, R. S., Ma, W. J., Pratte, M. S. & Jehee, J. F. Sensory uncertainty decoded from visual cortex predicts behavior. *Nat. Neurosci.* **18**, 1728–1730 (2015).

49. Van Bergen, R. S. & Jehee, J. F. Modeling correlated noise is necessary to decode uncertainty. *Neuroimage* **180**, 78–87 (2018).

50. Brouwer, G. J. & Heeger, D. J. Cross-orientation suppression in human visual cortex. *J. Neurophysiol.* **106**, 2108–2119 (2011).

51. Wandell, B. A., Dumoulin, S. O. & Brewer, A. A. Visual field maps in human cortex. *Neuron* **56**, 366–383 (2007).

52. Destrieux, C., Fischl, B., Dale, A. & Halgren, E. Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *Neuroimage* **53**, 1–15 (2010).

## Online Methods

We performed analyses on two publicly available published datasets[13,18] and one unpublished dataset from our lab[47]. Note that we only analyzed a subset of the data from each dataset, analyzing one or two conditions across three brain regions for the sake of simplicity. The experimental paradigms and conditions / regions chosen are described more in each dataset's respective subsection below.

### Inverted encoding model procedures

For all datasets, we performed a set of analyses using both the standard IEM and eIEM procedures, as described in the Introduction, with the exception that we used iterative shifting for both the standard IEM and eIEM. For both procedures, each basis channel is modeled as

$$\cos\left((\theta - \mu)\frac{\pi}{stimulus_range}\right)^{num_channels-1}$$

where θ is degrees in stimulus space, μ is the center of each channel, and *stimulus_range* is the range of stimulus space (e.g., 360° hues on a color wheel). The reasoning behind raising cosines to the *num_channels*-1 is to make the tuning curves narrower and more comparable to physiological findings[50]. For the encoder, each voxel's response is modeled as the weighted sum of the channels such that the observed trial-by-voxel fMRI activation matrix is equal to the dot product of the basis set and the weight matrix,

$$basis_set[trial_features, :] \cdot channel_by_voxel_weights = trial_by_voxel_activation$$

where *trial_features* is the feature (e.g., orientation) of the stimulus and *basis_set* is the matrix of channels with shape (*stimulus_range, num_channels*).

Once the weights matrix is estimated from the training dataset, it is inverted such that the weights matrix and the trial-by-voxel matrix are given and the channel responses (i.e., reconstructions) are estimated.

$$trial_by_voxel_activation \cdot channel_by_voxel_weights^{-1} = reconstructions$$

For all datasets, we used a basis set composed of nine equidistant channels each modeled as $\cos\left((\theta - \mu)\frac{\pi}{180}\right)^8$. We used 10-fold cross-validation, such that each iteration trained the model on 90% of the data and tested the model on the remaining 10%, repeated such that all trials were at one point decoded as part of the testing set.

For the standard IEM procedure, we aligned and averaged the single trial reconstructions into an average reconstruction and calculated slope as a traditional decoding metric. For the eIEM procedure, we calculated absolute prediction error for each trial via the correlation table metric and then calculated MAE. We performed these steps for each subject, ROI, and condition, and then we calculated the average slopes and MAEs across subjects. For each condition and ROI, we assessed significance via permutation testing. Significance tests were one-sided and uncorrected, calculated by comparing the t-statistic calculated from the actual data against the permuted null distribution of t-statistics (one t-statistic per each of 5,000 permutations).

For eIEMs, we also repeated this analysis pipeline using varying levels of goodness-of-fit thresholds. That is, we discarded a certain percent of trials based on the worst goodness-of-fits and then calculated MAE using the remaining trials.

**Perception dataset: Henderson, Vo, Chunharas, Sprague, and Serences (2019)**

Data were obtained by downloading post-processed fMRI data associated with Henderson et al (2019)[13], publicly available on OSF (https://osf.io/j7tpf/). In this experiment, nine participants attended to a central fixation while a sphere (multicolored flickering dots positioned on the shell of a 3D sphere with radius 3.4°) was presented at varying positions along the horizontal and depth axes (depth achieved through stereoscopic MR-compatible goggles). The task was to detect a brief luminance change of the fixation point. Participants completed between 7 and 21 runs, where each run of 36 trials began with a sphere presented for 3s followed by a jittered intertrial interval (2-6s). There were also runs where participants covertly attended to the sphere, but we did not include these runs in the analysis. We only reconstructed horizontal position for simplicity and because position-in-depth was only sampled across six unique locations (varied sampling across the entire stimulus space is more appropriate for inverted encoding models) whereas horizontal position was sampled across 36 unique locations (from 0.9° to 9.8° eccentricity in both directions, collapsing across position-in-depth). We analyzed V1, V4, and IPS regions of interest which were defined via retinotopic mapping protocols where participants viewed rotating wedges and bowtie stimuli[51] while performing a covert attention task of detecting contrast dimming on a row of the checkerboard for the rotating wedge stimulus. We applied IEMs (following the procedures outlined earlier) to the post-processed data conducted by the authors of the original paper: Single-trial activation estimates consisted of averaged z-scored BOLD signal of the 3rd and 4th TRs following stimulus presentation. For more methods information, please refer to the original paper[13].

**Attention dataset: Chen, Scotti, Dowd, & Golomb (2021)**

Data were previously collected in our lab for another study[47]. In this experiment, seven participants completed a visual attention task. Each trial started with a central fixation cross. After 700ms, three circle outlines were displayed at equidistant locations surrounding the fixation cross for 200ms. One outline was thicker than the others, representing the spatial cue. Participants were instructed to covertly attend to the spatial cue location while maintaining fixation at the fixation cross. After 1100ms, three colored and oriented gratings were briefly displayed for 100ms, followed by a 200ms mask and a continuous orientation report. Participants were instructed to report the orientation of the grating that appeared at the location of the spatial cue. There were also trials where participants were asked to shift attention to a different spatial location before the onset of the gratings, and entire runs where participants were asked to attend and report the color of the grating (instead of orientation), but we did not include these in our analysis. Participants completed at least 440 trials of each condition across multiple runs and sessions. We analyzed V1, V4, and IPS regions of interest: V1 and V4 were defined via retinotopic mapping protocols where participants viewed rotating wedges and bowtie stimuli[51], while IPS was defined from the Destrieux atlas[52] in Freesurfer (parcel labelled "S_intrapariet_and_P_trans"). To obtain single-trial neural activations for IEM, we modified a commonly used single-trial general linear model (GLM) approach[40] to improve the model sensitivity and account for the large number of trials. Specifically, we conducted 40 GLMs per subject, where each GLM includes one regressor per run for one of the 40 trials in that run and one regressor per run for all the other remaining trials

in that run. In this way, across the 40 GLMs, each trial in the experiment had an estimated single-trial beta weight. For more methods information, please refer to the original paper[47].

**Memory dataset: Rademaker, Chunharas, and Serences (2019)**

Data were obtained by downloading post-processed fMRI data associated with Rademaker et al (2019)[18], publicly available on OSF (https://osf.io/dkx6y). We reanalyzed Experiment 1, where six participants underwent a visual working memory task. For each trial, a cue indicating the distractor condition was shown for 1.4s, followed by a target grating shown for .5s where participants were instructed to memorize its orientation, followed by a 1s blank delay, and then an 11s delay where 3 possible distractor conditions were possible: blank delay, Fourier-filtered noise, or distractor grating of a pseudo-random orientation. Following an additional 1s blank delay, participants had 3s to report the orientation of the target grating, and finally a variable intertrial interval (3/5/8s). Each participant completed 108 trials per distractor condition. We only reconstructed the blank delay and distractor grating conditions for simplicity. We analyzed V1, V4, and IPS regions of interest which were defined via retinotopic mapping protocols where participants viewed rotating wedges and bowtie stimuli[51]. We applied IEMs to the post-processed data conducted by the authors of the original paper: Single-trial activation estimates consisted of averaged BOLD signal between 5.6-13.6s (7-17 TRs) after target onset. For more methods information, please refer to the original paper[18].

**Python package: inverted-encoding**

We have released the Python 3 package “inverted-encoding” on PyPi (https://pypi.org/project/inverted-encoding/) and GitHub (https://github.com/paulscotti/inverted_encoding) for easy implementation of our eIEM procedure. The package contains two main functions, “IEM” and “permutation.”

For the “IEM” function, the only necessary inputs are an array of the stimulus features for every trial and a trial by voxel activations matrix (note: inputs other than voxels may be used for other modalities). The basis set can be specified as an optional parameter and will otherwise default to a basis set composed of nine equidistant channels each modeled as $\cos\left((\theta - \mu)\frac{\pi}{180}\right)^8$. The stimulus space defaults to a circular 0-179° range but can be optionally set to other ranges. Non-circular stimulus spaces can be set by the Boolean parameter “is_circular.” The IEM procedure defaults to a 10-fold cross-validation procedure but can be optionally specified. The final outputs are an array of each trial’s predicted stimulus and an array of each trial’s corresponding goodness-of-fit. The user can then compute MAE themselves by averaging the (circular) absolute error between the predicted stimulus features and the actual stimulus features. The user can decide whether they want to threshold any trials using the provided goodness-of-fit values prior to calculating MAE.

For the “permutation” function, the only necessary input is an array of the actual stimulus features. For each iteration, the stimulus features are randomly shuffled and used as the predicted stimuli to compute the MAE. The function outputs a null distribution of MAE values for the user to compare against the MAE obtained from the “IEM” function. A more exact and computationally intensive method would be to rerun

the entire IEM pipeline with shuffled stimulus labels on every iteration to obtain the null distribution. This can also be performed using our package by simply repeating the IEM function with a different shuffling of the stimulus features for every iteration. Our exploratory comparisons of null distributions obtained using both approaches across the three fMRI datasets discussed in the main text yielded no obvious differences.