# FROM INFORMATION TO CONDITIONAL CONTRIBUTION: RESIDUAL-INFORMED SPARSE DECODING FOR TRACEABLE NEURAL REPRESENTATION ANALYSIS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Neural representations are organized patterns of population activity that carry information about external stimuli. Large-scale recordings make such information decodable, but it remains unclear which neural responses, at which times and temporal scales, support a given readout, or whether their conditional discriminative value changes with population context, stimulus, and decoding objective. Latent-component methods often form mixed coordinates that are difficult to trace to the original responses. Information-theoretic methods preserve traceability, but static rankings do not capture a response's value for the current readout. We introduce Residual-Informed Sparse Decoding (RISD), a residual-conditioned framework that constructs task-specific neural representations from traceable multiscale response coordinates. RISD first identifies an information core using conditional mutual information, then uses grouped out-of-fold residuals to add responses not captured by the current readout, yielding sparse, testable representations selected using training data only. We evaluate RISD on image-identity and semantic decoding using recordings from marmoset retina, the mouse visual pathway, and macaque visual cortex. Under matched support budgets, RISD outperforms baselines with compact response-coordinate sets and traces selected coordinates to specific neural regions, neurons, time windows, and scales. It reveals stimulus-specific response combinations and partially recoverable readout support in retinal activity, while showing across datasets that a response's discriminative value depends on selected support, stimulus, region, and decoding objective rather than its overall statistical information.

## 1 INTRODUCTION

Visual stimuli are represented by population activity distributed across neurons, brain regions, and processing stages. A neural representation is an organized pattern of neural activity that carries information about an external stimulus. In this work, we operationalize this concept by identifying which traceable response coordinates support stimulus decoding under a given readout. Large-scale spike recordings can support reliable decoding, but experiments often require a more specific answer: which neurons provide discriminative evidence, when that evidence appears, at what temporal scale it is expressed, and how the relevant responses work together. Because population activity is high-dimensional, redundant, and time-varying, decodability does not by itself imply interpretability. Extracting stimulus-related neural signals with clear physiological coordinates therefore remains a central challenge (Harris, 2005; Quiroga & Panzeri, 2009; Rigotti et al., 2013; Stringer et al., 2019).

Existing methods only partially address this problem. Low-dimensional representation methods can reveal population structure, but their latent coordinates are often difficult to trace back to individual neurons and their original firing times (Cunningham & Yu, 2014; Kobak et al., 2016; Zimnik et al., 2026). Traceable information-theoretic methods can select raw response components associated with a stimulus and reduce redundancy among candidates (Jia et al., 2022a; Jia & Yu, 2022), but their priorities are driven mainly by statistical associations with the label. When the representation is limited in size, however, the readout value of an individual response also depends on the current stimulus, the neural evidence already selected, and the decoding objective. A response with modest overall information may be valuable precisely because it resolves samples that the current model



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

still confuses. It therefore remains unclear whether stimulus-related neural signals are carried by a fixed set of highly informative responses or by neuron–time–scale combinations that change with the existing support and the readout objective.

We introduce RISD, which first establishes an information core in a traceable multiscale neural-response space. It then uses grouped out-of-fold prediction residuals from the core model to add responses with incremental value for the current readout. We evaluate RISD on complementary tasks spanning the retina, the early mouse visual pathway, and the macaque ventral visual pathway (Zhang et al., 2020; Zheng et al., 2021; Chen et al., 2024; Yu et al., 2025). The retinal data test the spatial and temporal physiological meaning of the selected responses; the mouse data test the region–time organization of image-identity information; and the macaque data test how the response coordinates selected by RISD change when the readout objective switches from image identity to cross-image semantic-category decoding. Under matched feature budgets, RISD improves decoding relative to the information-theoretic baseline and exposes the conditional role of neural responses.

Our contributions are as follows:

1. We introduce RISD, which combines an information core with residual-conditioned support completion to construct a sparse, traceable neural readout whose feature-set size is determined adaptively using only the training data.

2. We map the selected support back to individual neurons, post-stimulus time windows, and temporal scales, linking sparse feature selection to testable physiological interpretations.

3. Across visual levels and decoding objectives, we show that the discriminative role of a neural response is not a fixed statistical property. It is jointly conditioned by the stimulus, the responses already selected, the region and processing stage, and the decoding objective.

## 2 Related Work

Population neuroscience has studied the distribution of visual information through cell assemblies, mixed selectivity, and representational geometry (Harris, 2005; Rigotti et al., 2013; Stringer et al., 2019; Kriegeskorte et al., 2008). Principal-component, factor, and neural-manifold methods extract low-dimensional population structure and characterize relationships across stimuli or brain regions (Cunningham & Yu, 2014; Kobak et al., 2016; Onken et al., 2016).

Recent sparse component analysis (SCA) seeks separable computations in neural population activity using compact latent components (Zimnik et al., 2026). Like other latent-component methods, SCA can combine activity across neurons and time. RISD instead operates on original neuron–time–scale coordinates for a supervised readout.

Stimulus information can be expressed not only in mean firing rate but also in spike patterns at different response stages and temporal scales (Zheng et al., 2021; Zhang et al., 2023). Wavelet information (WI) methods use wavelet decompositions to obtain localized multiscale components (dos Santos et al., 2015; 2018) and select them independently according to mutual information with the stimulus label. They therefore do not explicitly account for redundancy or joint discriminative effects among candidates. Wavelet conditional mutual information (WCMI) instead builds a joint support according to the conditional information added beyond the responses already selected, reducing redundancy and capturing higher-order interactions (Jia et al., 2022a; Jia & Yu, 2022). Its priorities nevertheless remain driven by label-information gain rather than by the out-of-sample errors left by the current readout.

Residual-driven forward selection is also used in orthogonal matching pursuit and boosting to correct the current fitting error step by step (Pati et al., 1993; Friedman, 2001). Mutual-information feature selectors likewise balance relevance and redundancy or use conditional information during forward selection (Peng et al., 2005; Fleuret, 2004). RISD differs from generic feature selection in two ways: its candidates are traceable neuron–time–scale coordinates, and it first builds an information-theoretic core before reranking the remaining coordinates using grouped out-of-fold residuals.

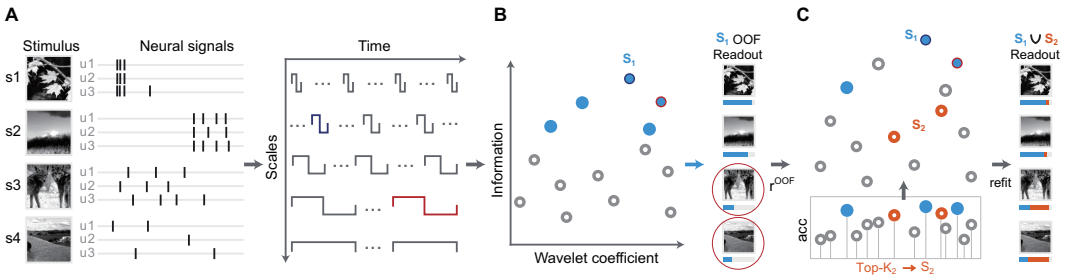



Figure 1: **RISD framework.** **A**, A fixed discrete wavelet decomposition converts each neuron's post-stimulus response into candidate coordinates indexed by neuron, time, and scale. **B**, Within each outer training fold, WCMI conditionally ranks the candidates. The first $K_1$ coordinates form the base support $S_1$, and grouped cross-fitting produces the out-of-fold (OOF) residuals $r^{\mathrm{OOF}}$. **C**, The remaining coordinates are scored by their first-order alignment with $r^{\mathrm{OOF}}$. The top $K_2$ coordinates form $S_2$, after which the decoder is refitted on $S_1 \cup S_2$. Support selection and hyperparameter tuning use only the outer training fold; the outer test fold is evaluated after the model is frozen.

## 3 RISD: Residual-Informed Support Completion

RISD constructs a support-conditioned sparse readout in a traceable multiscale neural-response space. It first applies a fixed discrete wavelet decomposition. WCMI then recursively ranks candidate responses by the conditional information they add to the current support (Jia et al., 2022a). An early high-information prefix forms the information core $S_1$ and defines a base readout. RISD next uses grouped out-of-fold errors from that readout to score the remaining responses by their first-order ability to correct the current predictions. The top residual-informed responses form the complementary support $S_2$, and the final decoder is jointly fitted on $S = S_1 \cup S_2$. Figure 1 summarizes the procedure.

### 3.1 Traceable multiscale response space

Stimulus information can appear in overall firing level or in a particular post-stimulus stage and temporal scale. To preserve both types of structure, we apply a fixed Haar wavelet decomposition independently along the time dimension of each neuron, without mixing responses across neurons (dos Santos et al., 2015; Zhang et al., 2023). Each coordinate $f$, instantiated as $z_{if}$ for sample $i$, retains a unique neuron, post-stimulus interval, and temporal scale, yielding a traceable multiscale response space.

### 3.2 Information-core construction

Within each outer training split, WCMI starts from an empty support and recursively ranks response coordinates using an interaction-aware conditional-information score (Jia et al., 2022a; Fleuret, 2004). At each step, it reevaluates all unselected coordinates according to the label information they add beyond the current support and then adds the highest-scoring coordinate. Repeating this process produces the ranking

$$\pi = (f_1, \ldots, f_M). \tag{1}$$

The early prefix concentrates the clearest label information while accounting for redundancy and complementarity with previously selected coordinates.

Conditional-information rank need not coincide with the out-of-sample marginal value of adding a coordinate to a finite-sample decoder. That value also depends on the current prediction state and the readout model. RISD therefore uses WCMI to establish an information core before evaluating which remaining coordinates can complement the current readout.

Within each outer training split, grouped inner validation compares only prespecified early prefixes of the WCMI ranking. A task-specific validation rule selects $K_1$, giving the information core

$$S_1 = \pi_{1:K_1}. \tag{2}$$

The grouped out-of-fold predictions of the readout fitted on $S_1$ define the residual prediction structure that remains unexplained.

### 3.3 Out-of-fold residual-informed support completion and final decoding

After fixing $S_1$, we cross-fit the preprocessing and probabilistic decoder, keeping samples from the same task group in the same fold. Let $v(i)$ denote the held-out fold containing sample $i$, let $P$ be the candidate response pool, and $C(P) = P \setminus S_1$. For class $c$ and candidate coordinate $f$, we compute

$$r_{ic}^{\mathrm{OOF}} = \mathbb{1}\{y_i = c\} - \hat{q}_{ic}^{\mathrm{OOF}}, \tag{3}$$

$$a_{fc} = \frac{1}{n-1} \sum_{i=1}^{n} (z_{if} - \bar{z}_f) \left( r_{ic}^{\mathrm{OOF}} - \bar{r}_c^{\mathrm{OOF}} \right), \tag{4}$$

$$s_f = \max_c |a_{fc}|, \tag{5}$$

$$S_2 = \mathrm{TopK}_{K_2}(C(P); s_f), \qquad S = S_1 \cup S_2. \tag{6}$$

Here, $\hat{q}_{ic}^{\mathrm{OOF}}$ is the out-of-fold predicted probability that sample $i$ belongs to class $c$, $r_{ic}^{\mathrm{OOF}}$ is the corresponding class-prediction error not explained by $S_1$, and $a_{fc}$ measures how a candidate response aligns with that error. For a linear correction to the current logits under a cross-entropy surrogate, $a_{fc}$ points in the same direction as the corresponding local negative loss gradient. Thus, $s_f$ provides a first-order criterion for completing the current support. The score is computed once after $S_1$ has been fixed.

Candidate configurations of $K_1$, the response pool $P$, and $K_2$ are prespecified. Grouped validation within the outer training data selects among these configurations. After a configuration is selected, the cross-fitted submodels are discarded, the selected configuration is re-derived on the full outer training set, and the final decoder is jointly fitted on the resulting support. The outer test set is evaluated once after all choices have been frozen. The size-matched WCMI baseline uses the first $|S|$ coordinates in the WCMI ranking. This baseline therefore serves as a continue-ranking ablation: it uses the same conditional-information front end and the same final support size, but replaces residual-informed completion with continued static ranking. Every final coordinate retains its neuron-time-scale identity for held-out decoding, support tracing, and deletion-refit analyses against equal-count random-deletion controls.

### 3.4 Datasets, tasks, and evaluation protocol

We evaluate RISD on three complementary visual neural-recording settings. The marmoset retinal ganglion-cell dataset (Karamanlis et al., 2025; Karamanlis & Gollisch, 2023) records responses to natural images and is used for image-identity decoding, enabling spatial validation through receptive fields. The Allen mouse visual-pathway dataset (Siegle et al., 2021) records natural-image responses across the lateral geniculate nucleus (LGN), primary visual cortex (V1), lateral visual area (VISl), anteromedial visual area (VISam), and laterointermediate visual area (VISli), and is used to examine the region-time organization of image-identity decoding. The macaque ventral-visual-pathway dataset (Li et al., 2026) contains recordings from V1, secondary visual cortex (V2), body-selective inferotemporal cortex (BodyIT), and face-selective inferotemporal cortex (FaceIT), spanning early visual cortex to high-level category-selective cortex. It is used for both image-identity decoding and five-way semantic decoding. For semantic decoding, training, validation, and test images are strictly separated by stimulus identifier (StimID), requiring category generalization across images.

All experiments use nested group-held-out evaluation. The grouping unit is defined by the task-specific stimulus structure; in particular, semantic decoding is grouped by StimID. Support selection, hyperparameter tuning, and decoder fitting use only the outer training data; the selected support and decoder are frozen before evaluation on the outer test set. RISD and information-theoretic baselines are compared at matched support sizes unless otherwise stated.

Table 1: Overall decoding performance across datasets and tasks. Image-identity columns report balanced accuracy (%); the semantic column reports StimID voting accuracy (%).

| Method | Marmoset retina | Allen mouse | Macaque identity | Semantic |
|---|---|---|---|---|
| PCA | 51.30 | 26.88 | 4.98 | 65.83 |
| SCA | 80.39 | 38.80 | 9.56 | 68.44 |
| WI | 95.19 | 64.13 | 4.74 | 65.73 |
| WCMI | 96.43 | 71.00 | 12.28 | 67.81 |
| **RISD** | **97.34** | **72.90** | **13.35** | **70.31** |

# 4 RESULTS

## 4.1 RISD IMPROVES DECODING ACROSS DATASETS AND TASKS

Table 1 compares RISD with Principal Component Analysis (PCA), SCA, WI, and WCMI under matched support budgets. RISD outperformed every baseline across all four aggregate settings. In particular, RISD reached balanced accuracies of $97.34\%$, $72.90\%$, and $13.35\%$ for marmoset retinal, mouse visual-pathway, and macaque image-identity decoding, respectively, and achieved $70.31\%$ StimID voting accuracy for macaque semantic decoding. Relative to the size-matched WCMI baseline, RISD improved the four tasks by 0.91, 1.90, 1.07, and 2.50 percentage points, respectively. Notably, RISD also outperformed SCA in every setting while retaining direct neuron–time–scale correspondence, rather than representing population activity through latent components.

These results show that residual-conditioned support completion improves decoding across distinct datasets and readout objectives. We next examine the neural-response supports identified by RISD.

## 4.2 RISD IDENTIFIES COMBINATORIAL IMAGE REPRESENTATIONS AND PARTIALLY RECOVERABLE READOUT SUPPORT IN RETINAL ACTIVITY

Retinal neurons encode natural images through spatially localized receptive fields and time-varying spike responses (Yan et al., 2022; Yu et al., 2025; Peng et al., 2025). Information about an image can therefore be distributed across several cells and across responses at different temporal scales (Pillow et al., 2008; Shlens et al., 2006). We first used a marmoset retinal ganglion-cell dataset recorded under natural-image stimulation to ask whether RISD could identify which cells, response intervals, and multicell combinations support the representation and discrimination of individual images.

Two image-level examples illustrate this combinatorial structure. For Image 149, readouts based on C17 or C51 alone each identified only $1/7$ trials correctly, whereas the joint C17+C51 readout reached $7/7$. For Image 173, the best single-cell or two-cell subset identified only $2/7$ trials, whereas the full three-cell combination reached $6/7$ (Figure 2A). Some image discriminations therefore depend on a joint multineuron response and cannot be recovered from any lower-order subset, consistent with broader evidence for synergistic and combinatorial neural codes (Brenner et al., 2000; Schneidman et al., 2011; Tang et al., 2025).

We next removed cells near the top of the RISD ranking to test their role in the current readout. With the original readout frozen, removing these cells caused substantially greater performance loss than removing equal numbers of random cells matched for cell type, receptive-field stability, and RISD feature count. Removing 10 cells reduced accuracy by more than 50 percentage points (Figure 2B). Refitting the readout on the remaining responses greatly reduced the loss. RISD therefore identifies cells on which the current readout depends, while the remaining population retains information that can support a partially recovered readout after refitting. This functional emphasis complements earlier work that inferred retinal substructure and circuit components from spike-triggered factorizations and deep response models (Liu et al., 2017; Jia et al., 2022b; Yan et al., 2022).

To examine the source of this recovery, we compared the deletion loss of other cells before and after a key cell was removed and the readout was refitted. We defined a substitution relationship when a cell produced a larger deletion loss after the key cell had been removed. The training data yielded 21 candidate relationships, 13 of which retained a positive substitution effect on independent test data (Figure 2C). For example, after C51 was removed, C17 assumed a larger readout role within a

nearby 282–376 ms window and an overlapping receptive-field region. Other substitutions occurred between nearby or distant receptive fields. Of the 87 trials that became correct after refitting, only 18 could be attributed to a single substituting cell, suggesting that the remaining recovery depended on joint responses not captured by single-cell substitution links.

At the population level, RISD identified 119 two-cell and 66 three-cell joint events. Relative to the best lower-order subset, the median information gains were 0.768 and 0.541 bits/trial, respectively (Figure 2D). The gains were image-selective: among 59 cell pairs involved in judgments for at least five images, 51 produced positive gain for only a subset of images. Among 230 cell-pair–image relationships observed again in an independent repeat, 153 preserved the original gain direction. Randomly permuting repeat correspondences across cells left the joint gains largely unchanged, indicating that the complementarity was supported mainly by stable image-response patterns in individual cells rather than by fixed trial pairings across cells. Pairwise and population-level retinal analyses provide complementary precedents for interpreting such joint events (Shlens et al., 2006; Pillow et al., 2008).

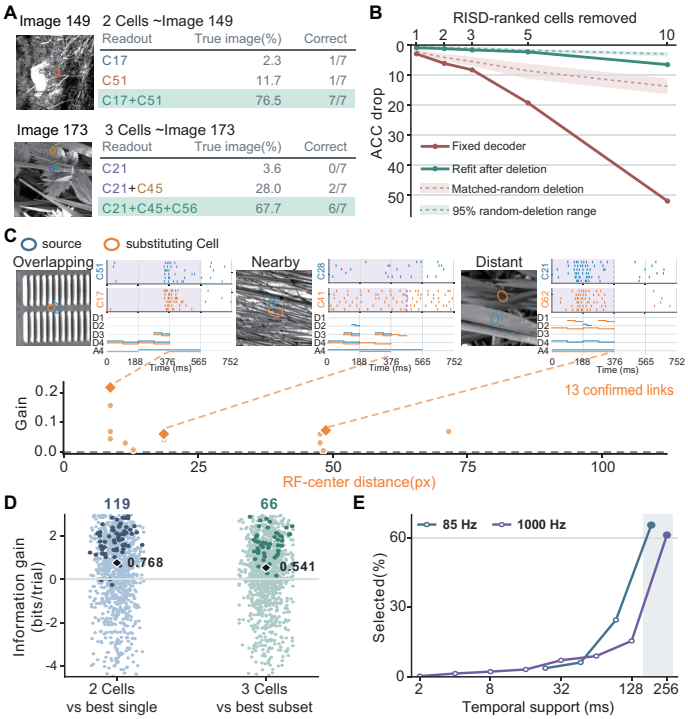



Figure 2: **RISD identifies combinatorial image representations and partially recoverable readout support in retinal population activity.** **A**, Complete cell combinations outperform lower-order subsets for two images. **B**,Targeted deletion impairs the frozen readout, whereas refitting partially restores performance; dashed lines and bands show matched random-deletion medians and 95% ranges. **C**,Thirteen held-out source–substituting-cell links across receptive-field distances. **D**,Information gain of two- and three-cell combinations over their best lower-order subsets. **E**, Selected temporal supports are stable across 85 and 1000 Hz.

Finally, we tested whether the temporal supports identified by RISD depended on binning resolution. After rebinning the same raw spikes from 85 to 1000 Hz, RISD continued to select supports on similar physical timescales, despite the change in Haar level (Figure 2E). Temporal localization by RISD is therefore not tied to a particular binning resolution, in line with multiscale spike representations that preserve temporal order across wavelet coordinates (dos Santos et al., 2015; Zhang et al., 2023).

### 4.3 RISD IDENTIFIES WHEN DIFFERENT REGIONS IN THE MOUSE VISUAL PATHWAY SUPPORT IMAGE REPRESENTATIONS

We next analyzed the Allen mouse visual-pathway dataset, which records activity from multiple visual nodes during natural-image stimulation. Large-scale surveys of this system have revealed a functional hierarchy across thalamic and cortical areas (Siegle et al., 2021), while recent decoding analyses have emphasized how dynamic visual information changes across processing stages (Chen et al., 2024). This dataset allows RISD to extend the analysis from cell combinations to region–time organization. RISD produced more correct decoding decisions than WCMI in every region (Figure 3A). Among single-region readouts, LGN and V1 reached accuracies of 88.1% and 83.9%, respectively, clearly exceeding the other visual regions. This pattern is consistent with their early position in visual processing and their rapid encoding of local contrast, edges, texture, and spatial

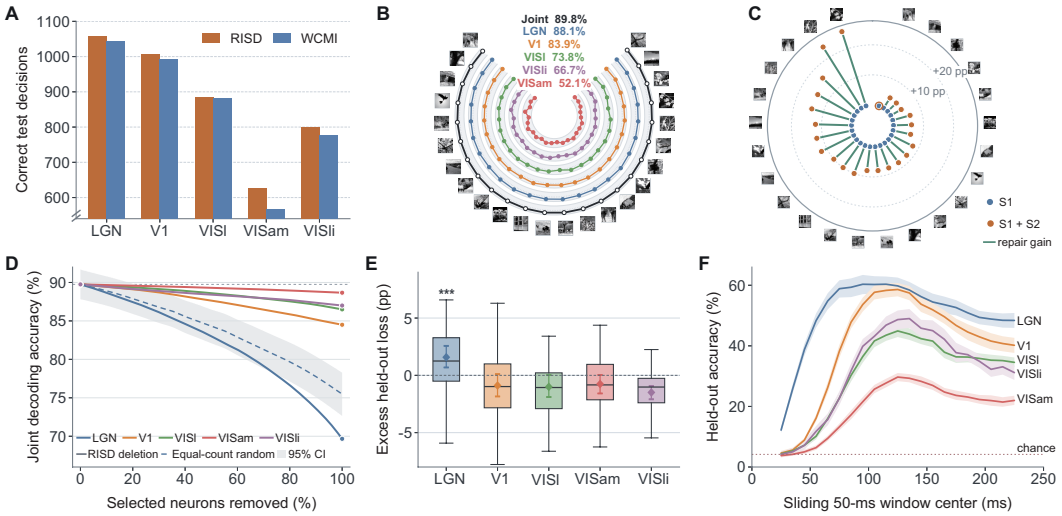



Figure 3: **RISD reveals how different stages of the mouse visual pathway jointly support image representation.** **A**, RISD and WCMI performance by region. **B**, Image-identity decoding by individual regions and their joint readout. **C**, Images repaired after extending $S_1$ to $S_1 \cup S_2$, together with their decoding gains. **D**, Change in joint decoding accuracy when RISD-selected neurons are progressively removed, relative to equal-count random deletions. **E**, Excess held-out loss by region after equal-count deletion and refitting. **F**, Held-out image-decoding accuracy from raw responses in sliding 50 ms windows. CI denotes confidence interval, and pp denotes percentage points.

layout. Combining all five regions yielded $89.8\%$ accuracy, exceeding every single-region readout (Figure 3B), which indicates that cross-region integration further improves image-identity decoding.

Having established where image-identity information is strongest across regions, we next asked whether residual completion improves the particular images that remain difficult for the information core. Extending $S_1$ to $S_1 \cup S_2$ repaired decoding for 23 images, increasing their mean accuracy from $79.83\%$ to $89.75\%$ (Figure 3C). Thus, $S_2$ does not simply duplicate evidence already contained in $S_1$; it adds discriminative responses for images that the core readout still confuses.

To test whether the responses localized by RISD were genuinely important to the joint readout, we progressively removed selected neurons from each region. Removing RISD-selected neurons generally caused a larger performance drop than removing equal numbers of randomly chosen neurons (Figure 3D), showing that the selected responses were not arbitrary. After the decoder was refitted on the remaining responses, deleting RISD-selected LGN neurons produced a substantially greater excess loss than equal-count random deletion (Figure 3E). We then decoded directly from raw spike responses in sliding 50 ms windows. Image information appeared first in LGN; V1 and the other visual regions subsequently reached higher decoding levels and retained discriminative information at later times (Figure 3F). The deletion–refit analysis shows that LGN has the strongest population-level effect on the joint image-identity readout. The sliding-window analysis further shows that image information emerges earliest in LGN, while V1 and the other regions retain complementary discriminative information at later times (Siegle et al., 2021; Zheng et al., 2021; Chen et al., 2024).

### 4.4 RISD IDENTIFIES OBJECTIVE-CONDITIONED RESPONSES FOR SEMANTIC DECODING IN THE VENTRAL VISUAL PATHWAY

The preceding analyses fixed image identity as the readout objective and examined how cell combinations and cross-region temporal responses support the discrimination of particular images. The macaque ventral visual pathway allows a further test: as visual processing shifts from local image structure toward object and category selectivity (DiCarlo & Cox, 2007; Hong et al., 2016; Kriegeskorte et al., 2008), do the responses that support image identity also support semantic categorization in the same way? In the Triple-N recordings, we changed the readout objective for the same natural images to five semantic categories. Training, validation, and test sets were strictly grouped by StimID,

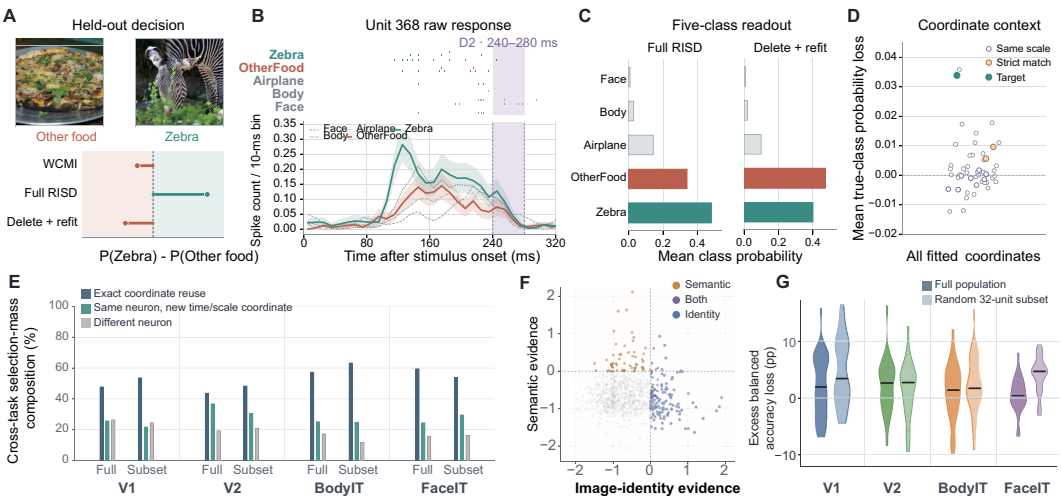



Figure 4: **RISD identifies objective-conditioned response organization for image-identity and semantic decoding in the macaque visual pathway. A**, Held-out Zebra decision before and after target-coordinate deletion and refitting. **B**, Raw responses of Unit 368; purple marks the selected 240–280 ms window. **C**, Five-class probabilities before and after deletion. **D**, Target-coordinate loss relative to matched coordinates. **E**, Cross-task coordinate reuse. **F**, Coordinate evidence for image identity and semantics. **G**, Excess deletion loss by region and population context.

so no test image appeared in the training data. The model therefore had to extract category information across different images, rather than reconstructing or recognizing a previously seen stimulus (Zhang et al., 2020). This design tests whether image-identity and semantic decoding depend on the same neuron–time–scale responses or on different responses under the current objective.

A held-out example illustrates this objective dependence. For a Zebra image, size-matched WCMI favored the class Other food, whereas full RISD increased the probability of Zebra relative to Other food (Figure 4A). RISD localized the D2 response of Unit 368 at 240–280 ms after stimulus onset, a window in which the unit exhibited stronger raw firing for Zebra (Figure 4B). After this coordinate was deleted and the model was refitted on the remaining responses, the five-class readout shifted back toward Other food (Figure 4C). Across the full set of fixed coordinates, deleting this coordinate caused a larger true-class probability loss than deleting other coordinates from the same unit or similar candidate coordinates (Figure 4D). RISD therefore localizes not merely a response correlated with category statistics, but discriminative evidence used by the current semantic readout. This model-conditioned interpretation is related in spirit to work that dissects computational components of spiking responses rather than treating all responsive coordinates as equivalent (Jia et al., 2021).

This dependence is not equivalent to static reuse of the image-identity support. Across four regions, semantic and image-identity tasks reused only some of the exact same neuron–time–scale coordinates. Other selected responses came from the same unit but a different time–scale coordinate, or from a different unit (Figure 4E). Coordinate priorities for image identity and semantics were also clearly separated: some coordinates primarily supported semantic decoding, others primarily supported image identity, and only a small subset supported both tasks (Figure 4F). A shared neural substrate therefore does not imply a fixed set of effective response coordinates, consistent with the broader view that mixed selectivity and explicit object information depend on the current task geometry (Rigotti et al., 2013; Hong et al., 2016).

Finally, we compared the excess deletion loss of RISD-localized units in the full population and in random 32-unit subpopulations (Figure 4G). Loss distributions differed across regions and population contexts, showing that the contribution of a unit to the current semantic readout depends on the population responses decoded alongside it. Semantic-category decoding does not simply reuse a fixed support for image identity. Instead, RISD reveals a partly shared but objective-conditioned readout organization: the effective coordinates depend on temporal scale, population context, and the decoding objective.

Table 2: Residual-completion ablation under fixed information-ranking front ends. Values are mean gains over the corresponding front end alone, in percentage points.

| Method | Image identity $\Delta$ balanced accuracy | Semantic decoding $\Delta$ accuracy |
|---|---|---|
| $\mathrm{RISD}_{\mathrm{WCMI}}$ | $+1.39$ | $+1.77$ |
| $\mathrm{RISD}_{\mathrm{MI}}$ | $+1.74$ | $\mathbf{+3.85}$ |
| $\mathrm{RISD}_{\mathrm{mRMR}}$ | $+1.86$ | $+3.65$ |
| $\mathrm{RISD}_{\mathrm{CMIM}}$ | $\mathbf{+2.35}$ | $+2.60$ |

### 4.5 Feature efficiency and ablations of residual completion

To test whether RISD maintains effective decoding with a compact response set, we compared it with the full Haar representation and the first 400 WCMI features under the same data splits and decoder. This sparsity-performance comparison is distinct from the size-matched baseline comparison in Table 1. In the Allen mouse and macaque image-identity tasks, RISD used approximately 68 and 114 features on average and achieved balanced accuracies of $72.90\%$ and $13.35\%$, respectively. These values exceeded both the full Haar representation ($58.25\%$ and $7.10\%$) and WCMI top-400 ($72.13\%$ and $12.00\%$). In semantic decoding, RISD used approximately 36 features and achieved a StimID voting accuracy of $70.31\%$, which is $94.8\%$ of the $74.17\%$ achieved by WCMI top-400 while using approximately $91\%$ fewer features. RISD therefore achieved higher performance with fewer features in the image-identity tasks and traded a modest semantic-performance loss for a substantial reduction in support size, complementing compact factorization and sparse-component approaches (Onken et al., 2016; Zimnik et al., 2026).

We next tested whether the benefit of RISD depended on the default conditional-mutual-information implementation by pairing residual completion with mutual information (MI), minimum-redundancy maximum-relevance (mRMR), and conditional mutual information maximization (CMIM) ranking front ends (Peng et al., 2005; Fleuret, 2004). Each front end built its own information core and candidate pool, and the support size was determined using training data only. With the feature count and decoder held fixed, adding RISD increased mean balanced accuracy across 13 image-identity configurations by 1.74, 1.86, and 2.35 percentage points for MI, mRMR, and CMIM, respectively (Table 2). Residual completion can therefore recover discriminative responses missing from different information cores rather than relying on a single ranking implementation.

## 5 Conclusion

RISD combines an information-theoretic core with grouped out-of-fold residual completion, adaptively determines the feature-set size using only the training data, and identifies traceable neuron–time–scale response coordinates that, for a particular readout, add discriminative evidence not captured by the current information core. Results from the retina, the mouse visual pathway, and the macaque ventral visual pathway show that the discriminative role of a neural response is not a fixed property; rather, it is conditioned by the stimulus, the responses already available to the population readout, the region and processing stage, and the decoding objective. RISD therefore turns the broad question of what can be decoded into the testable question of which responses support decoding under a given set of conditions, extending the shift from static feature description toward explicit dissection of neural computations (Jia et al., 2021; 2022b).

Although we evaluated RISD on visual spike recordings, its central idea is not restricted to spikes or visual tasks. Whenever a neural signal can be represented by traceable recording channels, temporal locations, and scale coordinates, the framework can in principle be applied to time-varying signals such as calcium imaging, local field potentials, electrocorticography (ECoG), and electroencephalography (EEG) (Wei et al., 2020; An et al., 2020; Zheng et al., 2020). It may also support decoding of nonvisual stimuli, behavioral state, or motor intention. For retinal translation, the same emphasis on interpretable temporal computation may be useful for neural prostheses and artificial photoreceptor systems (Yu et al., 2020; Yang et al., 2024). Future work should test its robustness across recording days, under input failure, and in closed-loop neural interfaces.

## AI USE STATEMENT

Generative AI tools were used in this work to assist with language editing and improve the readability of the manuscript. All AI-assisted content was reviewed and verified by the authors. The authors take full responsibility for the final content of this work.

## REFERENCES

Lingling An, Yuanhong Tang, Doudou Wang, Shanshan Jia, Qingqi Pei, Quan Wang, Zhaofei Yu, and Jian K. Liu. Intrinsic and synaptic properties shaping diverse behaviors of neural dynamics. *Frontiers in Computational Neuroscience*, 14:26, 2020. doi: 10.3389/fncom.2020.00026.

Naama Brenner, Steven P. Strong, Roland Koberle, William Bialek, and Rob R. de Ruyter van Steveninck. Synergy in a neural code. *Neural Computation*, 12(7):1531–1552, 2000. doi: 10.1162/089976600300015259.

Ye Chen, Peter Beech, Ziwei Yin, Shanshan Jia, Jiayi Zhang, Zhaofei Yu, and Jian K. Liu. Decoding dynamic visual scenes across the brain hierarchy. *PLOS Computational Biology*, 20(8):e1012297, 2024. doi: 10.1371/journal.pcbi.1012297.

John P. Cunningham and Byron M. Yu. Dimensionality reduction for large-scale neural recordings. *Nature Neuroscience*, 17(11):1500–1509, 2014. doi: 10.1038/nn.3776.

James J. DiCarlo and David D. Cox. Untangling invariant object recognition. *Trends in Cognitive Sciences*, 11(8):333–341, 2007. doi: 10.1016/j.tics.2007.06.010.

Vítor Lopes dos Santos, Stefano Panzeri, Christoph Kayser, Mathew E. Diamond, and Rodrigo Quian Quiroga. Extracting information in spike time patterns with wavelets and information theory. *Journal of Neurophysiology*, 113(3):1015–1033, 2015. doi: 10.1152/jn.00380.2014.

Vítor Lopes dos Santos, Hernan G. Rey, Joaquin Navajas, and Rodrigo Quian Quiroga. Extracting information from the shape and spatial distribution of evoked potentials. *Journal of Neuroscience Methods*, 296:12–22, 2018. doi: 10.1016/j.jneumeth.2017.12.014.

François Fleuret. Fast binary feature selection with conditional mutual information. *Journal of Machine Learning Research*, 5:1531–1555, 2004. URL `https://www.jmlr.org/papers/v5/fleuret04a.html`.

Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001. doi: 10.1214/aos/1013203451.

Kenneth D. Harris. Neural signatures of cell assembly organization. *Nature Reviews Neuroscience*, 6(5):399–407, 2005. doi: 10.1038/nrn1669.

Ha Hong, Daniel L. K. Yamins, Najib J. Majaj, and James J. DiCarlo. Explicit information for category-orthogonal object properties increases along the ventral stream. *Nature Neuroscience*, 19(4):613–622, 2016. doi: 10.1038/nn.4247.

Shanshan Jia and Zhaofei Yu. Zooming in on the brain via data science. *Patterns*, 3(3):100463, 2022. doi: 10.1016/j.patter.2022.100463.

Shanshan Jia, Dajun Xing, Zhaofei Yu, and Jian K. Liu. Dissecting cascade computational components in spiking neural networks. *PLOS Computational Biology*, 17(11):e1009640, 2021. doi: 10.1371/journal.pcbi.1009640.

Shanshan Jia, Xingyi Li, Tiejun Huang, Jian K. Liu, and Zhaofei Yu. Representing the dynamics of high-dimensional data with non-redundant wavelets. *Patterns*, 3(3):100424, 2022a. doi: 10.1016/j.patter.2021.100424.

Shanshan Jia, Zhaofei Yu, Arno Onken, Yonghong Tian, Tiejun Huang, and Jian K. Liu. Neural system identification with spike-triggered non-negative matrix factorization. *IEEE Transactions on Cybernetics*, 52(6):4772–4783, 2022b. doi: 10.1109/TCYB.2020.3042513.

Dimokratis Karamanlis and Tim Gollisch. Dataset - marmoset and mouse retinal ganglion cell responses to natural stimuli and supporting data. G-Node, 2023. URL https://doi.org/10.12751/g-node.ejk8kx.

Dimokratis Karamanlis, Mohammad H. Khani, Helene M. Schreyer, Sören J. Zapp, Matthias Mietsch, and Tim Gollisch. Nonlinear receptive fields evoke redundant retinal coding of natural scenes. *Nature*, 637(8045):394–401, 2025. doi: 10.1038/s41586-024-08212-3.

Dmitry Kobak, Wieland Brendel, Christos Constantinidis, Claudia E. Feierstein, Adam Kepecs, Zachary F. Mainen, Xue-Lian Qi, Ranulfo Romo, Naoshige Uchida, and Christian K. Machens. Demixed principal component analysis of neural population data. *eLife*, 5:e10989, 2016. doi: 10.7554/eLife.10989.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A. Bandettini. Representational similarity analysis—connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008. doi: 10.3389/neuro.06.004.2008.

Yipeng Li, Xieyi Liu, Wanru Li, Jia Yang, Baoqi Gong, Wei Jin, Zhengxin Gong, Kesheng Wang, Jingqiu Luo, Zishuo Zhao, and Pinglei Bao. Triple-N dataset: Large-scale fMRI-guided dense recordings of nonhuman primate neural responses to natural scenes. *Nature Neuroscience*, 29(8): 1999–2011, 2026. doi: 10.1038/s41593-026-02322-z.

Jian K. Liu, Helene M. Schreyer, Arno Onken, Fernando Rozenblit, Mohammad H. Khani, Vidhyasankar Krishnamoorthy, Stefano Panzeri, and Tim Gollisch. Inference of neuronal functional circuitry with spike-triggered non-negative matrix factorization. *Nature Communications*, 8(1): 149, 2017. doi: 10.1038/s41467-017-00156-9.

Arno Onken, Jian K. Liu, P. P. Chamanthi R. Karunasekara, Ioannis Delis, Tim Gollisch, and Stefano Panzeri. Using matrix and tensor factorizations for the single-trial analysis of population spike trains. *PLOS Computational Biology*, 12(11):e1005189, 2016. doi: 10.1371/journal.pcbi.1005189.

Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Proceedings of the 27th Asilomar Conference on Signals, Systems and Computers*, pp. 40–44, 1993. doi: 10.1109/ACSSC.1993.342465.

Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, 2005. doi: 10.1109/TPAMI.2005.159.

Jing Peng, Shanshan Jia, Jiyuan Zhang, Yongxing Wang, Zhaofei Yu, and Jian K. Liu. Decoding natural visual scenes via learnable representations of neural spiking sequences. *Neural Networks*, 192:107863, 2025. doi: 10.1016/j.neunet.2025.107863.

Jonathan W. Pillow, Jonathon Shlens, Liam Paninski, Alexander Sher, Alan M. Litke, E. J. Chichilnisky, and Eero P. Simoncelli. Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature*, 454(7207):995–999, 2008. doi: 10.1038/nature07140.

Rodrigo Quian Quiroga and Stefano Panzeri. Extracting information from neuronal populations: Information theory and decoding approaches. *Nature Reviews Neuroscience*, 10(3):173–185, 2009. doi: 10.1038/nrn2578.

Mattia Rigotti, Omri Barak, Melissa R. Warden, Xiao-Jing Wang, Nathaniel D. Daw, Earl K. Miller, and Stefano Fusi. The importance of mixed selectivity in complex cognitive tasks. *Nature*, 497 (7451):585–590, 2013. doi: 10.1038/nature12160.

Elad Schneidman, Jason L. Puchalla, Ronen Segev, Robert A. Harris, William Bialek, and Michael J. Berry. Synergy from silence in a combinatorial neural code. *The Journal of Neuroscience*, 31(44): 15732–15741, 2011. doi: 10.1523/JNEUROSCI.0301-09.2011.

Jonathon Shlens, Greg D. Field, Jeffrey L. Gauthier, Matthew I. Grivich, Dumitru Petrusca, Alexander Sher, Alan M. Litke, and E. J. Chichilnisky. The structure of multi-neuron firing patterns in primate retina. *The Journal of Neuroscience*, 26(32):8254–8266, 2006. doi: 10.1523/JNEUROSCI.1282-06.2006.

Joshua H. Siegle, Xiaoxuan Jia, Séverine Durand, Sam Gale, Corbett Bennett, et al. Survey of spiking in the mouse visual system reveals functional hierarchy. *Nature*, 592(7852):86–92, 2021. doi: 10.1038/s41586-020-03171-x.

Carsen Stringer, Marius Pachitariu, Nicholas Steinmetz, Matteo Carandini, and Kenneth D. Harris. High-dimensional geometry of population responses in visual cortex. *Nature*, 571(7765):361–365, 2019. doi: 10.1038/s41586-019-1346-5.

Yuanhong Tang, Shanshan Jia, Tiejun Huang, Zhaofei Yu, and Jian K. Liu. Implementing feature binding through dendritic networks of a single neuron. *Neural Networks*, 189:107555, 2025. doi: 10.1016/j.neunet.2025.107555.

Ziqiang Wei, Bei-Jung Lin, Tsai-Wen Chen, Kayvon Daie, Karel Svoboda, and Shaul Druckmann. A comparison of neuronal population dynamics measured with calcium imaging and electrophysiology. *PLOS Computational Biology*, 16(9):e1008198, 2020. doi: 10.1371/journal.pcbi.1008198.

Qi Yan, Yajing Zheng, Shanshan Jia, Yichen Zhang, Zhaofei Yu, Feng Chen, Yonghong Tian, Tiejun Huang, and Jian K. Liu. Revealing fine structures of the retinal receptive field by deep-learning networks. *IEEE Transactions on Cybernetics*, 52(1):39–50, 2022. doi: 10.1109/TCYB.2020.2972983.

Ruyi Yang, Peng Zhao, Liyang Wang, Chenli Feng, Chen Peng, Zhexuan Wang, Yingying Zhang, Minqian Shen, Kaiwen Shi, Shijun Weng, Chunqiong Dong, Fu Zeng, Tianyun Zhang, Xingdong Chen, Shuiyuan Wang, Yiheng Wang, Yuanyuan Luo, Qingyuan Chen, Yuqing Chen, Chengyong Jiang, Shanshan Jia, Zhaofei Yu, Jian Liu, Fei Wang, Su Jiang, Wendong Xu, Liang Li, Gang Wang, Xiaofen Mo, Gengfeng Zheng, Aihua Chen, Xingtao Zhou, Chunhui Jiang, Yuanzhi Yuan, Biao Yan, and Jiayi Zhang. Assessment of visual function in blind mice and monkeys with subretinally implanted nanowire arrays as artificial photoreceptors. *Nature Biomedical Engineering*, 8(8):1018–1039, 2024. doi: 10.1038/s41551-023-01137-8.

Zhaofei Yu, Jian K. Liu, Shanshan Jia, Yichen Zhang, Yajing Zheng, Yonghong Tian, and Tiejun Huang. Toward the next generation of retinal neuroprosthesis: Visual computation with spikes. *Engineering*, 6(4):449–461, 2020. doi: 10.1016/j.eng.2020.02.004.

Zhaofei Yu, Tong Bu, Yijun Zhang, Shanshan Jia, Tiejun Huang, and Jian K. Liu. Robust decoding of rich dynamical visual scenes with retinal spikes. *IEEE Transactions on Neural Networks and Learning Systems*, 36(2):3396–3409, 2025. doi: 10.1109/TNNLS.2024.3351120.

Jiyuan Zhang, Shanshan Jia, Zhaofei Yu, and Tiejun Huang. Learning temporal-ordered representation for spike streams based on discrete wavelet transforms. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(1):137–147, 2023. doi: 10.1609/aaai.v37i1.25085.

Yichen Zhang, Shanshan Jia, Yajing Zheng, Zhaofei Yu, Yonghong Tian, Siwei Ma, Tiejun Huang, and Jian K. Liu. Reconstruction of natural visual scenes from neural spikes with deep neural networks. *Neural Networks*, 125:19–30, 2020. doi: 10.1016/j.neunet.2020.01.033.

Yajing Zheng, Shanshan Jia, Zhaofei Yu, Tiejun Huang, Jian K. Liu, and Yonghong Tian. Probabilistic inference of binary markov random fields in spiking neural networks through mean-field approximation. *Neural Networks*, 126:42–51, 2020. doi: 10.1016/j.neunet.2020.03.003.

Yajing Zheng, Shanshan Jia, Zhaofei Yu, Jian K. Liu, and Tiejun Huang. Unraveling neural coding of dynamic natural visual scenes via convolutional recurrent neural networks. *Patterns*, 2(10):100350, 2021. doi: 10.1016/j.patter.2021.100350.

Andrew J. Zimnik, Xinyue An, K. Cora Ames, Andrew Ulmer, Antonio H. Lara, Abigail A. Russo, Vladislav Susoy, Laura Driscoll, John P. Cunningham, Liam Paninski, Mark M. Churchland, and Joshua I. Glaser. Sparse component analysis: A method that uncovers separable computations within neural population activity. *Neuron*, 2026. doi: 10.1016/j.neuron.2026.05.022.