

Decomposing stimulus-specific sensory neural information via diffusion models

Steeve Laquittaine* Simone Azeglio* Carlo Paris Ulisse Ferrari† Matthew Chalk‡
 Institut de la Vision
 Sorbonne Université, Paris 75014

Abstract

To understand sensory coding, we must ask not only how much information neurons encode, but also what that information is about. This requires decomposing mutual information into contributions from individual stimuli and stimulus features—a fundamentally ill-posed problem with infinitely many possible solutions. We address this by introducing three core axioms—*additivity*, *positivity*, and *locality*—that any meaningful stimulus-wise decomposition should satisfy. We then derive a decomposition that meets all three criteria and remains tractable for high-dimensional stimuli. Our decomposition can be efficiently estimated using diffusion models, allowing for scaling up to complex, structured and naturalistic stimuli. Applied to a model of visual neurons, our method quantifies how specific stimuli and features contribute to encoded information. Our approach provides a scalable, interpretable tool for probing representations in both biological and artificial neural systems.

1 Introduction

A central question in sensory neuroscience is *how much*, but also *what* information neurons transmit about the world. While Shannon’s information theory provides a principled framework to quantify the amount of information neurons encode about *all* stimuli, it does not reveal *which* stimuli contribute most, or *what* stimulus features are encoded [Shannon, 1948, Brenner et al., 2000, DeWeese and Meister, 1999]. As a concrete example, it is known that neurons in the early visual cortex are ‘sensitive’ to stimuli in a small region of space (their receptive field). However, it is not clear how such simple intuitions carry to more complex scenarios, e.g. with large, noisy & non-linear population of neurons and high-dimensional stimuli.

Several previous measures of neural sensitivity have been proposed. For example, the Fisher information quantifies the sensitivity of neural responses to infinitesimal stimulus perturbations [Rao, 1992, Brunel and Nadal, 1998, Yarrow et al., 2012, Kriegeskorte and Wei, 2021, Ding et al., 2023]. However, as the Fisher is not a valid decomposition of the mutual information it cannot say how different stimuli contribute to the total encoded information. On the other hand, previous works have proposed stimulus dependent decompositions of mutual information, which define a function $I(x)$ such that $I(R; X) = \mathbb{E}[I(x)]$ [DeWeese and Meister, 1999, Butts, 2003, Butts and Goldman, 2006, Kostal and D’Onofrio, 2018]. However, this decomposition is inherently ill-posed: infinitely many functions $I(x)$ satisfy the constraint, with no principled way to select among them. Further, different decompositions behave in qualitatively different ways, making it hard to interpret what are they are telling us. Finally, most proposed decompositions are computationally intractable for the high-dimensional stimuli and non-linear encoding models relevant for neuroscience.

*Equal contribution.

†Equal senior contribution.

‡Corresponding author: matthew.chalk@inserm.fr

To resolve these limitations, we propose a set of axioms that any stimulus specific and feature-specific information decomposition should satisfy in order to serve as a meaningful and interpretable measure of neural sensitivity. These axioms formalize intuitive desiderata: that the information assigned to each stimulus, and stimulus feature, should be non-negative, and additive with respect to repeated measurements. We also require the decomposition to respect a form of *locality*: changes in how a neuron responds to a stimulus x should not affect the information attributed to a distant stimulus x' . Finally, the attribution must be insensitive to irrelevant features, which do not contribute to the total information. Together, these constraints ensure that the decomposition is both interpretable and theoretically grounded.

We show that existing decompositions violate one or more of these axioms, limiting their interpretability and use as information theoretic measures of neural sensitivity. We then introduce a novel decomposition that satisfies all of our axioms. It generalizes Fisher information by capturing neural sensitivity to both infinitesimal and finite stimulus perturbations. Moreover, it supports further decomposition across individual stimulus features (e.g., image pixels), enabling fine-grained analysis of neural representations.

Beyond satisfying our theoretical axioms, our decomposition is computationally tractable for large neural populations and high-dimensional naturalistic stimuli, through the use of diffusion models. We demonstrate the power of our method by quantifying the information encoded by a model of visual neurons about individual images and pixels. Our approach uncovers aspects of the neural code that are not picked up by standard methods, such as the Fisher information, and opens the door to similar analyses in higher-order sensory areas, and artificial neural networks.

2 Desired properties of stimulus-specific decomposition of information

We aim to decompose the mutual information $I(R; X)$ between a stimulus X and a neural response R into local attributions $I(x)$, assigning to each stimulus x a measure of its contribution to the total information. By construction, such a decomposition must satisfy:

- **Axiom 1: Completeness.** The average of local attributions must recover the total mutual information:

$$I(R; X) = \mathbb{E}_X[I(x)].$$

This constraint alone does not uniquely determine the function $I(x)$, as many decompositions satisfy completeness. To further constrain the attribution, we impose additional desiderata that reflect desirable properties of local information measures.

First, for $I(x)$ to serve as an *interpretable*, stimulus-specific measure of neural sensitivity, it should satisfy a *locality principle*: perturbations to the likelihood &/or prior in a neighbourhood of x_0 should have a vanishing influence on $I(x)$ for distant stimuli $x \neq x_0$. Without this property, changes in $I(x)$ may reflect changes in neural sensitivity to any stimuli, undermining interpretability.

- **Axiom 2: Locality.** Let $p(R, X)$ and $\tilde{p}(R, X)$ be two joint distributions on $\mathcal{R} \times \mathbb{R}^d$, and let $I(x)$ and $\tilde{I}(x)$ denote the corresponding local information decompositions derived from p and \tilde{p} , respectively. Suppose there exists $\epsilon > 0$ and $x_0 \in \mathbb{R}^d$ such that:

$$p(R, x) = \tilde{p}(R, x) \quad \text{for all } x \notin B_\epsilon(x_0),$$

where $B_\epsilon(x_0)$ is the open ball of radius ϵ centered at x_0 . Then we require that there exists some value $d > 0$, such that:

$$|I(x) - \tilde{I}(x)| \rightarrow 0 \quad \text{when} \quad \|x - x_0\| \gg d$$

That is, local perturbations to the likelihood or prior near x_0 should not affect the information assigned to distant stimuli, x .

Mutual information is globally non-negative: $I(R; X) \geq 0$. To preserve this property in our decomposition, we require the pointwise contributions $I(x)$ to be non-negative as well. Intuitively, observing a neural response can only refine our beliefs about the stimulus—it cannot undo information. In addition, negative attributions can harm interpretability, since they can cancel out, obscuring how different stimuli contribute to the total information.

- **Axiom 3: Positivity.** Local information attributions must be non-negative:

$$I(x) \geq 0 \quad \forall x \in X.$$

Finally, Shannon 1948 posited additivity as a fundamental property of mutual information: the total information from multiple sources should equal the sum of their individual contributions. We extend this principle pointwise, requiring that information combine additively across measurements.

- **Axiom 4: Additivity.** For two responses R and R' , the local attribution should decompose as:

$$I_{R,R'}(x) = I_R(x) + I_{R'|R}(x), \quad \forall x \in X$$

where we have renamed $I(x)$ as $I_R(x)$ here, to make explicit its dependence on the response, R . $I_{R'|R}(x)$ is the conditional pointwise information from R' given R , and $I_{R,R'}(x)$ denotes the pointwise information from observing both responses. By construction, we require: $I(R'; X|R) = E_x [I_{R'|R}(x)]$ and $I(R, R'; X) = E_x [I_{R,R'}(x)]$.

Remark 1: Local data processing inequality. Any decomposition that fulfils both additivity and positivity, as stated above, also obeys a local form of the data processing inequality. That is, post-processing should not increase information, even at the level of the individual attributions. Formally,

$$\text{If } X \rightarrow R \rightarrow R', \quad \text{then } I_R(x) \geq I_{R'}(x) \quad \forall x \in X$$

To prove this we use additivity to write $I_{R',R}(x)$ in two ways:

$$I_R(x) + I_{R'|R}(x) = I_{R'}(x) + I_{R|R'}(x)$$

Positivity gives $I_{R|R'}(x) \geq 0$ for all x , so:

$$I_R(x) + I_{R'|R}(x) \geq I_{R'}(x)$$

Now, if R' is independent of X given R , then $I(R'; X|R) = \mathbb{E}_X [I_{R'|R}(x)] = 0$. Since if $I_{R'|R}(x) \geq 0$ pointwise, this implies $I_{R'|R}(x) = 0$ for all x . Substituting into the inequality above yields the desired result: $I_R(x) \geq I_{R'}(x)$.

Remark 2: Invariance to invertible transformations. The data processing inequality implies that local information attributions are invariant under invertible transformations of the response variable. That is, for any invertible function $\phi(R)$, we have:

$$I_{\phi(R)}(x) = I_R(x).$$

This follows from the fact that $I(R; X) = I(\phi(R); X)$, and hence $\mathbb{E}_X [I_{\phi(R)}(x)] = \mathbb{E}_X [I_R(x)]$. Supposing, for contradiction, that $I_{\phi(R)}(x) < I_R(x)$ for some x , equality of expectations would require $I_{\phi(R)}(x) > I_R(x)$ for some other x , violating the pointwise data processing inequality. This highlights why our axioms are important for interpretability: they imply that $I(x)$ quantifies how information is transmitted through the system $X \rightarrow R$, independently of how the responses are parameterized.

3 Previous stimulus-dependent decompositions

Several previous works have proposed decompositions of mutual information into stimulus-specific contributions, such as the stimulus-specific information I_{SSI} [Butts, 2003], the specific information I_{sp} , the stimulus-specific surprise I_{surp} , [DeWeese and Meister, 1999], and the coordinate-invariant stimulus-specific information I_{CISSI} [Kostal and D'Onofrio, 2018]:

$$I_{sp}(x) = H(R) - H(R|x) \tag{1}$$

$$I_{SSI}(x) = H(X) - \mathbb{E}_{R|x} [H(X|R=r)] \tag{2}$$

$$I_{surp}(x) = D_{KL}(p(R|x) || p(R)) \tag{3}$$

$$I_{CISSI}(x) = \mathbb{E}_{R|x} [D_{KL}(p(X|R=r) || p(X))] \tag{4}$$

While these decompositions satisfy some of our proposed axioms (Table 1), none satisfy locality. This is because they all depend on global terms such as the marginal response distribution $p(r) = \int p(r|x')p(x') dx'$, or the posterior $p(x|r) = p(r|x)p(x) / \int p(r|x')p(x') dx'$, both of which can be

Table 1: Satisfaction of axioms by different information-theoretic measures of neural sensitivity. Only our proposed decomposition, I_{local} , fulfils all the axioms.

Axiom	$\mathcal{J}(x)$	$I_{sp}(x)$	$I_{SSI}(x)$	$I_{surp}(x)$	$I_{CiSSI}(x)$	$I_{local}(x)$
Completeness	X	✓	✓	✓	✓	✓
Locality	✓	X	X	X	X	✓
Positivity	✓	X	X	✓	✓	✓
Additivity	✓	✓	✓	✓	✓	✓

influenced by changes to the likelihood &/or prior for any stimulus. As discussed above, this limits their interpretability as measures of neural sensitivity, since a non-zero attribution at x could be due to changes in neural sensitivity anywhere in the stimulus space.

Unlike the above decompositions, the Fisher information is inherently local. However, while previous authors found a relation between the Fisher and mutual information [Brunel and Nadal, 1998, Wei and Stocker, 2016], this only holds approximately, and in certain limits (e.g. low-noise). Therefore, the Fisher information cannot be used to quantify how different stimuli contribute to the total encoded information (i.e. it fails the **completeness** axiom).

Recently Kong et al. 2024 used diffusion models to decompose the mutual information into contributions from both the stimulus, x , and the response, r . Two different decompositions, $I(r, x)$, were proposed. However, averaging these over $p(R|x)$ does not give stimulus-wise decompositions that fulfil our axioms (Appendix B). In one case, we obtain $I_{surp}(x)$ (Eqn 3), which is non-local; in the other case, the decomposition is not additive, which is a fundamental information theoretic constraint.

In the following we propose a new stimulus-wise decomposition of the mutual information which fulfils all our axioms, combining the advantages of the Fisher information (locality, positivity & additivity) while being a valid decomposition of the mutual information.

4 Diffusion-based information decomposition

We first derive an expression for the mutual information, $I(R; X)$, that can be used to construct a stimulus-dependent decomposition that fulfils all of the above axioms.

4.1 Exact relation between Shannon information and Fisher information

Consider the information between the response of a population of neurons, R , and a noise-corrupted variable, $X_\gamma = X + \sqrt{\gamma}Z$, where $Z \sim \mathcal{N}(0, I)$. From the fundamental theorem of calculus:

$$-\int_0^\infty \frac{dI(X_\gamma; R)}{d\gamma} d\gamma = I(R; X_{\gamma=0}) - \lim_{\gamma \rightarrow \infty} I(R; X_\gamma) = I(R; X),$$

since, $I(R; X_{\gamma=0}) = I(R; X)$, while $\lim_{\gamma \rightarrow \infty} I(R; X_\gamma) = 0$. It follows that,

$$I(R; X) = -\int_0^\infty \frac{dI(X_\gamma; R)}{d\gamma} d\gamma = \int_0^\infty \left(\frac{dH(X_\gamma|R)}{d\gamma} - \frac{dH(X_\gamma)}{d\gamma} \right) d\gamma$$

where $H(X_\gamma)$ and $H(X_\gamma|R)$ are the marginal and conditional entropy of X_γ , respectively. Then, from de Bruijn’s identity [Rioul, 2010],

$$\begin{aligned} I(R; X) &= -\frac{1}{2} \text{Trace} \int_0^\infty \mathbb{E}_{X_\gamma, R} \left[\nabla_{x_\gamma}^2 \log p(X_\gamma|R) - \nabla_{x_\gamma}^2 \log p(X_\gamma) \right] d\gamma \\ &= -\frac{1}{2} \text{Trace} \int_0^\infty \mathbb{E}_{X_\gamma, R} \left[\nabla_{x_\gamma}^2 \log p(R|X_\gamma) \right] d\gamma \\ &= \frac{1}{2} \text{Trace} \int_0^\infty \mathbb{E}_{X_\gamma} [\mathcal{J}(X_\gamma)] d\gamma \end{aligned} \tag{5}$$

where $\mathcal{J}(x_\gamma)$ is the Fisher information with respect to a noise-corrupted stimulus, X_γ .

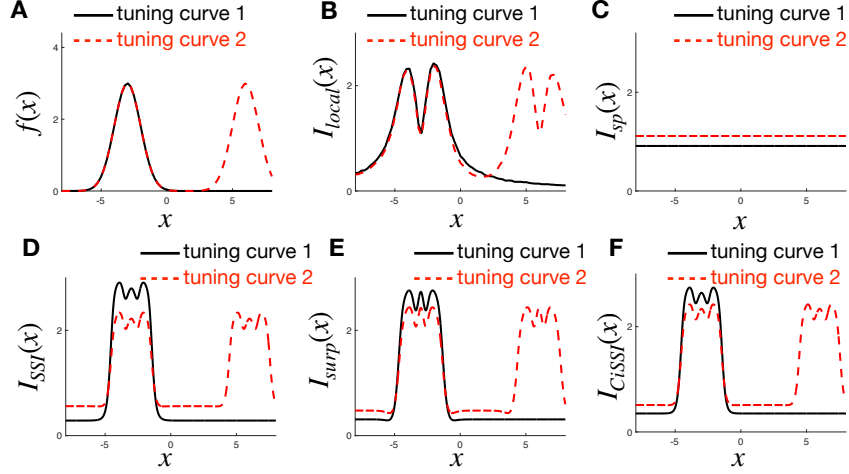


Figure 1: **Demonstration of locality.** (A) A neuron responds to a 1D stimulus x with tuning curve $f(x)$. We compare two cases: (i) tuning changes only near $x = -3$ (black), and (ii) additional changes near $x = 6$ (red dashed). (B) Our local information measure $I_{\text{local}}(x)$ converges for both cases away from the modified region. (C–F) Other decompositions (Eqs. 1–4) show non-local changes, altering information attribution across all x .

This identity provides a direct relation between the mutual information $I(R; X)$ and the Fisher information, $\mathcal{J}(x_\gamma)$. Further, it also admits a natural interpretation in terms of de-noising diffusion models trained to predict X from a noisy observation X_γ . To see this, first we use an alternative formulation of the Fisher information, in terms of the mean-squared score:

$$\begin{aligned} I(R; X) &= \frac{1}{2} \int_0^\infty \mathbb{E}_{X_\gamma, R} \left[\|\nabla_{x_\gamma} \log p(R|X_\gamma)\|^2 \right] d\gamma \\ &= \frac{1}{2} \int_0^\infty \mathbb{E}_{X_\gamma, R} \left[\|\nabla_{x_\gamma} \log p(X_\gamma|R) - \nabla_{x_\gamma} \log p(X_\gamma)\|^2 \right] d\gamma \end{aligned} \quad (6)$$

Next, from Tweedie’s formula [Meng et al., 2021, Kadkhodaie and Simoncelli, 2020] we have:

$$I(R; X) = \int_0^\infty \frac{1}{2\gamma^2} \mathbb{E}_{X_\gamma, R} \left[\|\hat{x}(X_\gamma, R) - \hat{x}(X_\gamma)\|^2 \right] d\gamma, \quad (7)$$

where $\hat{x}(x_\gamma) = \mathbb{E}[X|x_\gamma]$ and $\hat{x}(x_\gamma, r) = \mathbb{E}[X|x_\gamma, r]$. These conditional means can be approximated using denoising diffusion models trained to sample from the prior, $p(X)$, and posterior $p(X | R)$, respectively [Sundararajan et al., 2017].

4.2 Local stimulus-specific information

Building on the integral representation of mutual information in Eq. (5), we define a stimulus-specific decomposition:

$$I_{\text{local}}(x) = \frac{1}{2} \sum_i \int_0^\infty \mathbb{E}_{X_\gamma|x} [\mathcal{J}_{ii}(X_\gamma)] d\gamma, \quad (8)$$

where $X_\gamma = x + \sqrt{\gamma}Z$, with $Z \sim \mathcal{N}(0, I)$, and $\mathcal{J}_{ii}(x_\gamma) = \mathbb{E}_{R|x_\gamma} \left[-\nabla_{(x_\gamma)_i}^2 \log p(R|x_\gamma) \right]$ is the i^{th} diagonal element of the Fisher information matrix, evaluated at x_γ . This expression parallels Eq. (5), with the key distinction that the expectation is now conditioned on a fixed stimulus x , rather than averaging over the full stimulus distribution. Consequently, the decomposition satisfies the **completeness axiom**, since by construction, $I(R; X) = \mathbb{E}_X [I_{\text{local}}(x)]$.

Next we outline why our decomposition fulfils the **locality axiom** (for the proof, see Appendix A). Recall that the Fisher information matrix $\mathcal{J}(x)$ characterizes the local curvature of the log-likelihood, $\log p(R|x)$, and thus quantifies the local sensitivity of the response to changes in the stimulus [Rao, 1992]. The term $\mathbb{E}_{X_\gamma|x} [\mathcal{J}(X_\gamma)]$ generalizes this notion, measuring neural sensitivity

to noise-perturbed versions of the stimulus, $X_\gamma \sim \mathcal{N}(x, \gamma I)$. For finite γ , this term is dominated by values of X_γ close to x , and thus, it depends only on the local shape of the likelihood and prior around x . It receives a vanishingly small contribution from changes to the likelihood and/or prior for distant stimuli x' , if $\|x' - x\| \gg \gamma$. Meanwhile, as $\gamma \rightarrow \infty$, X_γ becomes pure noise and thus $E_{X_\gamma|x}[\mathcal{J}(X_\gamma)] \rightarrow 0$ for all x . Taken together, this implies that our decomposition, obtained by integrating $E_{X_\gamma|x}[\mathcal{J}(X_\gamma)]$ over all $\gamma > 0$, satisfies the **locality axiom**: local perturbations to $p(R, x)$ affect $I_{local}(x')$ only for nearby x' , while their influence vanishes as $\|x' - x\| \rightarrow \infty$.

The remaining axioms follow directly from standard properties of the Fisher information matrix. **Positivity** follows from the fact that the Fisher information is positive semi-definite. **Additivity** follows from the identity $\mathcal{J}_{R',R}(x_\gamma) = \mathcal{J}_R(x_\gamma) + \mathcal{J}_{R'|R}(x_\gamma)$ [Zamir, 1998]. Both properties are preserved when we average the Fisher information over $p(X_\gamma|x)$ and integrate over $\gamma > 0$, to obtain $I_{local}(x)$.

4.3 Feature-Wise Decomposition

We assume the stimulus x is a vector of image features x_i (e.g. image pixels). Given that the local stimulus information (Eqn. 8) is expressed as a sum over diagonal elements of the Fisher information matrix, it is natural to decompose $I_{local}(x)$ into feature-wise contributions $I_i(x)$.

As with the stimulus-wise decomposition, this problem is ill-posed: infinitely many decompositions exist in theory. However, the same axioms constrain the feature-wise decomposition. To satisfy **additivity**, $I_i(x)$ must be a linear combination of Fisher diagonal terms: $I_i(x) = \frac{1}{2} \sum_j a_{ij} \int_0^\infty \mathbb{E}_{X_\gamma|x}[\mathcal{J}_{jj}(X_\gamma)] d\gamma$. **Completeness** requires $\sum_i a_{ij} = 1$, while **positivity** enforces $a_{ij} \geq 0$. Finally, to fully specify the weights, a_{ij} , we need to introduce one further axiom, which ensures that the attribution $I_i(x)$ is zero for irrelevant stimulus features, X_i , which are independent of the response, R .

- **Axiom 5: Insensitivity to irrelevant features.** If X_i is independent of R , then $I_i(x) = 0 \forall x \in X$.

For this axiom to hold, we need to set $a_{ij} = \delta_{ij}$. To see this, we observe that if, on the contrary, $a_{ij} \neq \delta_{ij}$, then a neuron’s sensitivity to other features (i.e. $\mathcal{J}_{jj}(x) > 0$) could ‘leak over’ to make $I_i(x) > 0$ even when X_i is independent of R , violating the axiom.

5 Results

5.1 Locality

To illustrate the implication of the locality axiom, we analyzed the responses of a model neuron to a one-dimensional stimulus drawn from a Gaussian prior. The neuron’s response was modeled as a Gaussian random variable with mean $f(x)$ and fixed standard deviation. We compared two tuning curves: one with a single peak at $x = -3$ (Fig. 1A, black), and another with an additional peak at $x = 6$ (Fig. 1A, red).

In both cases, $I_{local}(x)$ peaked where the tuning curves were steepest, echoing the behaviour of the Fisher information, which (under Gaussian noise) scales with $f'(x)^2$ (Fig. 1B). Crucially, the difference in $I_{local}(x)$ between the two tuning curves vanished outside the region where they differ, consistent with the locality axiom. In contrast, existing attribution methods (Fig. 1C–F) showed global sensitivity: adding a second peak at $x = 6$ altered the attributed information across the entire stimulus space, including far from the added feature.

5.2 Effect of Prior

We next examined how $I_{local}(x)$ responds to changes in the shape of the stimulus prior. For this, we considered neural responses to stimuli drawn from a bimodal prior with peaks at $x = -1$ and $x = 1$ (Fig. 2A, upper panel). To isolate the effect of the prior, we used a simple linear-Gaussian likelihood model: $r \sim \mathcal{N}(x, \sigma^2)$. Under this model, the neuron’s sensitivity is uniform across all stimuli, so any variation in attributed information must arise solely from the prior.

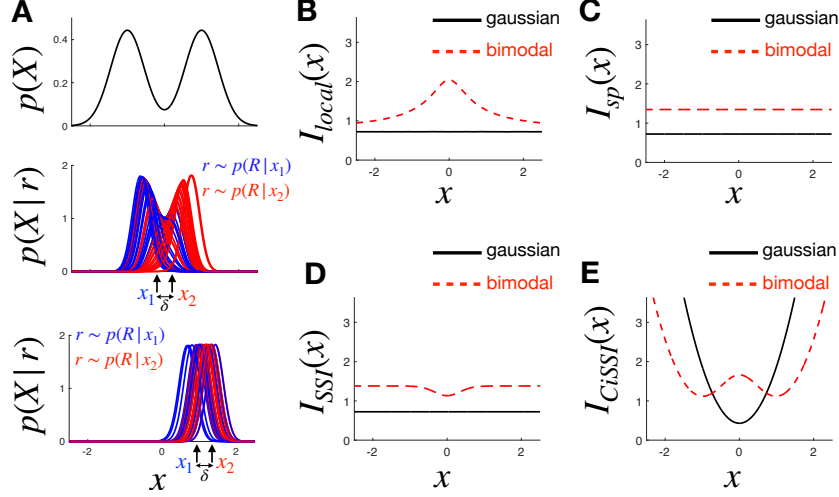


Figure 2: **Effect of prior.** (A) A bimodal prior (top) with a neuron responding as $r = x + \text{noise}$. Middle and bottom: posterior $p(X | r)$ given $r \sim p(R | x_1)$ (blue) or $r \sim p(R | x_2)$ (red), for stimuli x_1, x_2 separated by δ . The posterior is more sensitive to x near 0 (middle) than near 1 (bottom). (B) Our information measure $I_{\text{local}}(x)$ peaks near $x = 0$ for the bimodal prior (red dashed), where the posterior is most sensitive, and is flat for a Gaussian prior (black), where posterior shape is constant. (C–E) Other decompositions behave differently: (C) I_{sp} is flat for both priors; (D) I_{SSI} is minimal near $x = 0$ for the bimodal prior; (E) I_{Cissi} is quadratic under a Gaussian prior. I_{surp} (not shown) behaves similarly to I_{Cissi}

Intuitively, one can assess neural sensitivity by measuring how much the posterior distribution $p(X | r)$ changes, on average, in response to small perturbations in the stimulus x . In the bimodal case, the posterior is highly sensitive near $x = 0$, where the two modes compete (Fig. 2A, middle panel), and relatively stable near the modes themselves, e.g., around $x = 1$ (Fig. 2A, lower panel). Our attribution measure $I_{\text{local}}(x)$ reflects this structure, peaking at $x = 0$, and decaying elsewhere. For the Gaussian prior, where posterior sensitivity is constant, $I_{\text{local}}(x)$ is flat. In contrast, previously proposed attribution methods behave inconsistently: some remain constant across both priors (Fig. 2C), others respond in the opposite direction (Fig. 2D), and some varied strongly even under a Gaussian prior, where the posterior sensitivity is uniform (Fig. 2E).

5.3 Data Processing Inequality

Next we illustrate how $I_{\text{local}}(x)$ respects the data processing inequality while certain other attribution methods do not. For this, we used a bimodal prior (Fig. 3A) and compared a linear-Gaussian neuron ($r \sim \mathcal{N}(x, \sigma_r^2)$) to a downstream neuron with response $r' = |r|$. Since this transformation is non-invertible, information must be lost. Consistent with the inequality, $I_{\text{local}}(x)$ decreased at every x (Fig 3C). In contrast, both I_{SSI} and I_{sp} increased at some x and decreased at others, violating the pointwise data processing inequality (Fi 3D-E). I_{surp} , by comparison, respected the inequality (Fig 3F).

5.4 Scaling to high-dimensions

We can use Eqn 7 to write the feature-wise decomposition in terms of the outputs of an unconditional and conditional diffusion model, trained to output $\hat{x}(x_\gamma) = E[X|x_\gamma]$ and $\hat{x}(x_\gamma, r) = E[X|x_\gamma, r]$, respectively:

$$I_i(x) = \int_0^\infty \frac{1}{2\gamma^2} \mathbb{E}_{X_\gamma|x, R|X_\gamma} \left[(\hat{x}_i(X_\gamma, R) - \hat{x}_i(X_\gamma))^2 \right] d\gamma \quad (9)$$

To obtain a Monte-carlo approximation of this expression, we need to sample from $x_\gamma \sim p(X_\gamma|x)$, followed by, $r \sim p(R|x_\gamma)$. Sampling from $p(R|x_\gamma)$ is prohibitively expensive (since it requires first sampling from $x \sim p(x_0|x_\gamma)$, which requires a full backward pass of the diffusion model). To get around this, we adopt an approximation used by Chung et al. 2023, instead approximating $I_i(x)$ using samples $r \sim p(R|\hat{x}(x))$, which can be computed efficiently using one pass through the

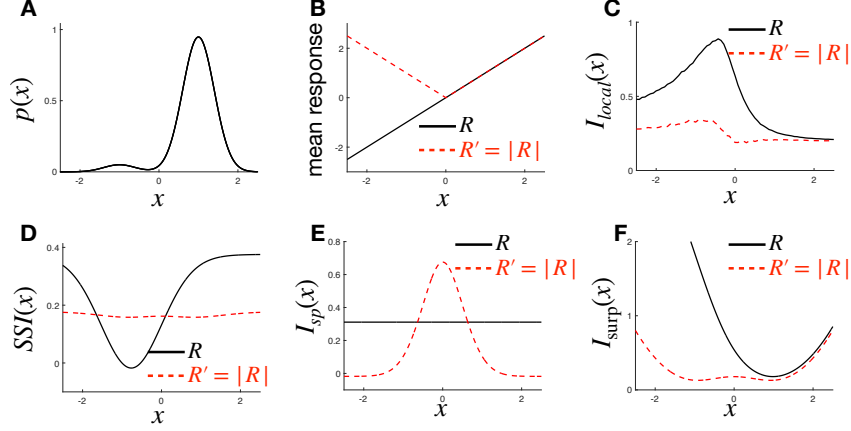


Figure 3: **Local data processing inequality.** (A) Bimodal prior. (B) Neural responses: $r = x + \text{noise}$; transformed response $r' = |r|$. The non-invertible transform reduces information. (C) Our decomposition satisfies the pointwise data processing inequality (DPI): $I_{R'}(x)$ (red) is always less than or equal to $I_R(x)$ (black). (D–F) I_{surp} also satisfies pointwise DPI, while I_{SSI} and I_{sp} do not. I_{C+SSI} (not shown) behaves similarly to I_{surp} .

de-noising network. The integral over γ was approximated numerically with evenly spaced γ (see Appendix C). Since the Fisher decays to zero for large γ , truncating the integral has little effect on our approximation of the integral.

5.5 Pixel-wise decomposition of encoded information

We applied our method to identify which regions of an image contribute most to the total information encoded by a population of visual neurons. For illustrative purposes, we used stimuli from the MNIST dataset and modelled a simple population of neurons with mean responses given by $Af(\mathbf{w} \cdot \mathbf{x} + b)$, where A and b are constants, $f(\cdot)$ is a sigmoid nonlinearity, and \mathbf{w} is a linear filter representing the neuron’s receptive field (RF). Neural responses were corrupted by Poisson noise. We simulated 49 neurons with RFs arranged in a uniform grid (Fig. 4A).

As a baseline, we first evaluated neural sensitivity using the diagonal of the Fisher information matrix (Fig. 4B–D, E–F). In this model, the Fisher information reduces to a weighted sum of squared RFs, where each neuron’s contribution is scaled by its activation level. This yields characteristic “blob-like” patterns, with each blob centered on the neuron’s RF.

We then used a diffusion model trained on MNIST to estimate the pixel-wise information decomposition, $I_{\text{local}}(\mathbf{x})$ (Fig 4D, G; additional images are shown in Supp Fig 2). This decomposition revealed that information was concentrated along object edges—regions where the decoded images (i.e. samples from the posterior) are most sensitive to small changes in the presented stimulus (cf. Fig. 2A). Unlike the Fisher information, our measure integrates how both the local sensitivity of neurons (via their RFs) and the statistical structure of the input (captured by the diffusion model), contribute towards the total encoded information.

6 Discussion

We introduced a principled, information-theoretic measure of neural sensitivity to stimuli. This measure satisfies a core set of axioms that ensure interpretability and theoretical soundness. Crucially, the measure can be estimated using diffusion models, making it scalable to high-dimensional inputs and complex, non-linear neural populations. We empirically demonstrated how each axiom shapes interpretability through simple, illustrative examples. Finally, we show how the method can be applied in a high-dimensional setting to quantify the information encoded by a neural population about visual stimuli.

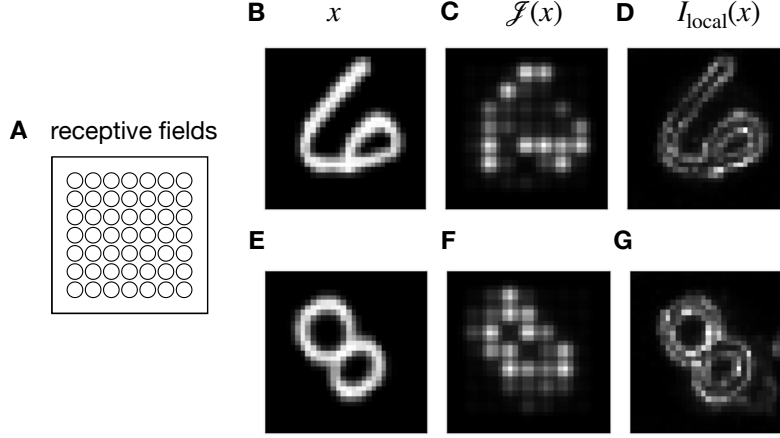


Figure 4: **Pixel-wise information decomposition with MNIST stimuli.** (A) Simulated population of linear–nonlinear–Poisson (LNP) neurons with circular receptive fields arranged in a grid. (B) Presented visual stimulus. (C) Diagonal of the Fisher information, given by a weighted sum of neural receptive fields. (D) Pixel-wise decomposition of mutual information, $I_{\text{local}}(x)$. E–G Same as panels A–B, but for a different stimulus.

Kong et al. 2024 recently proposed two decompositions of the mutual information between visual stimuli x and text prompts y , $I(x, y)$, which can be efficiently estimated using diffusion models. These were used to identify image regions most informative about accompanying text. Dewan et al. 2024 extended this approach to assess pixel-wise redundancy and synergy with respect to the prompt. However, these frameworks do not directly yield a stimulus-wise neural sensitivity measure $I(x)$, which was our goal (Appendix B). Moreover, simply averaging the decompositions of Kong et al. over neural responses does not produce stimulus-wise and feature-wise decompositions that satisfy our axioms: in one case the resulting decomposition is non-local and can be negative; in the other case, the decomposition is not additive, and thus violates a key information theoretic property pointwise.

Our work extends a classical result from Brunel & Nadal 1998, who showed that mutual information can be approximated in the low-noise limit using Fisher information [Wei and Stocker, 2016]. This approximation was later used by Wei & Stocker 2015 to explain a wide range of perceptual phenomena under the efficient coding hypothesis [Morais and Pillow, 2018]. However, their approximation, which depends on the log determinant of the Fisher, becomes very inaccurate in certain cases, such as at high noise, or where there are more stimulus dimensions than neurons (in which case it returns minus infinity) [Bethge et al., 2002, Yarrow et al., 2012, Huang and Zhang, 2018]. Here, we instead identify an *exact* relation between mutual information and Fisher information with respect to a noise-corrupted stimulus. This opens new potential to test theories of efficient coding of high-dimensional stimuli, and in the presence of noise.

In the future, we will use our method to investigate more realistic neural models, fitted on biological data, as well as diffusion models trained on natural image datasets. Here, to further improve efficiency we could use zero-shot methods that sample directly from the posterior, without requiring a trained conditional diffusion model [Peng et al., 2024, Chung et al., 2023]. Such approaches have recently achieved state-of-the-art performance in decoding visual scenes from retinal ganglion cell responses [Wu et al., 2022], and could enable rapid assessment of how changes to the neural model affect encoded stimulus information.

Finally, our axiomatic framework has close parallels with integrated gradients [Sundararajan et al., 2017], a method developed to attribute the output of deep neural networks to individual input features. Both our method and integrated gradients address ill-posed attribution problems by enforcing natural axioms. However, one limitation of our method, in contrast to integrated gradients, is that we do not prove the uniqueness of our attribution method, as following from our axioms. This will be interesting to investigate in the future. There are also key differences between both approaches: for example, our attribution explicitly accounts for noisy responses and does not require specification of an arbitrary baseline. These distinctions make our method particularly well-suited to neural data,

and suggest potential utility as a principled attribution tool for analyzing deep networks—identifying which stimuli, or stimulus features, different units or layers are sensitive to.

References

- Matthias Bethge, David Rotermund, and Klaus Pawelzik. Optimal short-term population coding: When fisher information fails. *Neural computation*, 14(10):2317–2351, 2002.
- 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.
- Nicolas Brunel and Jean-Pierre Nadal. Mutual information, fisher information, and population coding. *Neural computation*, 10(7):1731–1757, 1998.
- Daniel A Butts. How much information is associated with a particular stimulus? *Network: Computation in Neural Systems*, 14(2):177, 2003.
- Daniel A Butts and Mark S Goldman. Tuning curves, neuronal variability, and sensory coding. *PLoS biology*, 4(4):e92, 2006.
- Hyungjin Chung, Jeongsol Kim, Michael Thompson Mccann, Marc Louis Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=OnD9zGAGT0k>.
- Shaurya Rajat Dewan, Rushikesh Zawar, Prakanshul Saxena, Yingshan Chang, Andrew Luo, and Yonatan Bisk. Diffusion PID: Interpreting diffusion via partial information decomposition. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL <https://openreview.net/forum?id=aBpxukZS37>.
- Michael R DeWeese and Markus Meister. How to measure the information gained from one symbol. *Network: Computation in Neural Systems*, 10(4):325, 1999.
- Xuehao Ding, Dongsoo Lee, Joshua Melander, George Sivulka, Surya Ganguli, and Stephen Baccus. Information geometry of the retinal representation manifold. *Advances in Neural Information Processing Systems*, 36:44310–44322, 2023.
- Sylvain Gugger, Lysandre Debut, Thomas Wolf, Philipp Schmid, Zachary Mueller, Sourab Mangrulkar, Marc Sun, and Benjamin Bossan. Accelerate: Training and inference at scale made simple, efficient and adaptable. <https://github.com/huggingface/accelerate>, 2022.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- Wentao Huang and Kechen Zhang. Information-theoretic bounds and approximations in neural population coding. *Neural computation*, 30(4):885–944, 2018.
- Zahra Kadhodaie and Eero P Simoncelli. Solving linear inverse problems using the prior implicit in a denoiser. *arXiv preprint arXiv:2007.13640*, 2020.
- Xianghao Kong, Ollie Liu, Han Li, Dani Yogatama, and Greg Ver Steeg. Interpretable diffusion via information decomposition. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=X6tNkN6ate>.
- Lubomir Kostal and Giuseppe D’Onofrio. Coordinate invariance as a fundamental constraint on the form of stimulus-specific information measures. *Biological Cybernetics*, 112:13–23, 2018.
- Nikolaus Kriegeskorte and Xue-Xin Wei. Neural tuning and representational geometry. *Nature Reviews Neuroscience*, 22(11):703–718, 2021.
- Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- Chenlin Meng, Yang Song, Wenzhe Li, and Stefano Ermon. Estimating high order gradients of the data distribution by denoising. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL <https://openreview.net/forum?id=YTkQqRqSyE1>.

- Michael Morais and Jonathan W Pillow. Power-law efficient neural codes provide general link between perceptual bias and discriminability. *Advances in Neural Information Processing Systems*, 31, 2018.
- Xinyu Peng, Ziyang Zheng, Wenrui Dai, Nuoqian Xiao, Chenglin Li, Junni Zou, and Hongkai Xiong. Improving diffusion models for inverse problems using optimal posterior covariance, 2024. URL <https://openreview.net/forum?id=9mX0AZVEet>.
- C Radhakrishna Rao. Information and the accuracy attainable in the estimation of statistical parameters. In *Breakthroughs in Statistics: Foundations and basic theory*, pages 235–247. Springer, 1992.
- Olivier Rioul. Information theoretic proofs of entropy power inequalities. *IEEE transactions on information theory*, 57(1):33–55, 2010.
- Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.
- Claude E Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.
- Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.
- Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. <https://github.com/huggingface/diffusers>, 2022.
- Xue-Xin Wei and Alan A Stocker. A bayesian observer model constrained by efficient coding can explain ‘anti-bayesian’ percepts. *Nature neuroscience*, 18(10):1509–1517, 2015.
- Xue-Xin Wei and Alan A Stocker. Mutual information, fisher information, and efficient coding. *Neural computation*, 28(2):305–326, 2016.
- Eric Wu, Nora Brackbill, Alexander Sher, Alan Litke, Eero Simoncelli, and EJ Chichilnisky. Maximum a posteriori natural scene reconstruction from retinal ganglion cells with deep denoiser priors. *Advances in Neural Information Processing Systems*, 35:27212–27224, 2022.
- Stuart Yarrow, Edward Challis, and Peggy Seriès. Fisher and shannon information in finite neural populations. *Neural computation*, 24(7):1740–1780, 2012.
- Ram Zamir. A proof of the fisher information inequality via a data processing argument. *IEEE Transactions on Information Theory*, 44(3):1246–1250, 1998.

A Locality

Let $p(R, X)$ and $\tilde{p}(R, X)$ be two joint distributions on $\mathcal{R} \times \mathbb{R}^d$. Suppose there exists $\epsilon > 0$ and $x_0 \in \mathbb{R}^d$ such that:

$$p(R, x) = \tilde{p}(R, x) \quad \text{for all } x \notin B_\epsilon(x_0),$$

where $B_\epsilon(x_0)$ is the open ball of radius ϵ centered at x_0 .

We analyze the effect of this perturbation on the local information decomposition,

$$I_{local}(x) = \frac{1}{2} \text{Trace} \int_0^\infty \mathbb{E}_{X_\gamma|x} [\mathcal{J}(X_\gamma)] d\gamma, \quad (10)$$

where $X_\gamma = x + \sqrt{\gamma}Z$, $Z \sim \mathcal{N}(0, I)$, and $\mathcal{J}(x_\gamma) = \mathbb{E}_{R|x_\gamma} \left[-\nabla_{x_\gamma}^2 \log p(R | x_\gamma) \right]$ denotes the Fisher information matrix at x_γ .

We first consider the behaviour of the integrand at large $\gamma > d$. From Eqn 7, we can write

$$\mathcal{J}(x_\gamma) = \frac{1}{\gamma^2} \|\mathbb{E}[X|x_\gamma, R] - \mathbb{E}[X|x_\gamma]\|^2 \quad (11)$$

In the limit of large d , we have

$$\mathcal{J}(x_\gamma) \approx \frac{1}{2d^2} \|\mathbb{E}[X|R] - \mathbb{E}[X]\|^2 = \frac{1}{d^2} \Sigma_{X|R} \quad (12)$$

where $\Sigma_{X|R}$ is the posterior covariance in X given R . Thus, it follows that for some large but finite d , the local information converges to:

$$I_{local}(x) \approx \frac{1}{2} \text{Trace} \int_0^d \mathbb{E}_{X_\gamma|x} [\mathcal{J}(X_\gamma)] d\gamma + \frac{1}{2d} \text{Trace} (\Sigma_{X|R}). \quad (13)$$

Next, we consider the behaviour of the integrand at $\gamma \leq d$. First, recall that we can write the conditional log-likelihood as:

$$\log p(r | x_\gamma) = \log \frac{\int p(r, x') e^{-\frac{\|x_\gamma - x'\|^2}{2\gamma^2}} dx'}{\int p(x') e^{-\frac{\|x_\gamma - x'\|^2}{2\gamma^2}} dx'}, \quad (14)$$

Because this is a convolution with a Gaussian, with width γ , any perturbation to $p(R, X)$ localized around x_0 will have a vanishing influence on $\log p(R|x_\gamma)$ (and hence its derivatives) when $\|x_\gamma - x_0\|^2 \gg \gamma$, and thus

$$\mathbb{E}_{R|x_\gamma} \left[-\nabla_{x_\gamma}^2 \log p(R | x_\gamma) \right] - \mathbb{E}_{R|x_\gamma} \left[-\nabla_{x_\gamma}^2 \log \tilde{p}(R | x_\gamma) \right] \rightarrow 0 \quad \text{when } \|x_\gamma - x_0\|^2 \gg \gamma. \quad (15)$$

Averaging over $X_\gamma \sim p(X_\gamma | x) (= \mathcal{N}(X_\gamma | x, \gamma I))$ preserves this approximation, provided $\|x - x_0\|^2 \gg \gamma$. Hence,

$$\mathbb{E}_{X_\gamma|x} [\mathcal{J}(X_\gamma)] - \mathbb{E}_{X_\gamma|x} [\tilde{\mathcal{J}}(X_\gamma)] \rightarrow 0, \quad \text{when } \|x - x_0\|^2 \gg \gamma, \quad (16)$$

where $\tilde{\mathcal{J}}$ denotes the Fisher information computed under the perturbed distribution.

Combining Eqn 16, and Eqn 13, we arrive at the desired relation:

$$I_{local}(x) - \tilde{I}_{local}(x) \rightarrow 0, \quad \text{when } \|x - x_0\|^2 \gg d, \quad (17)$$

where \tilde{I}_{local} is computed using the perturbed distribution. That is, local perturbations to $p(R, X)$ at x_0 have a vanishing effect on the local information decomposition $I_{local}(x)$ for stimuli far from x_0 .

B Comparison with previous decompositions using diffusion models

Recently Kong et al. 2024 proposed two different decompositions of the mutual information into terms that depend on both the response and stimulus, which can be computed using diffusion models. Their decompositions take the following form:

$$I_{Kong,1}(x, r) = \int_0^\infty \frac{1}{2\gamma^2} E_{X_\gamma|x} [\|x - \hat{x}(x_\gamma, r)\|^2 - \|x - \hat{x}(x_\gamma)\|^2] d\gamma \quad (18)$$

$$I_{Kong,2}(x, r) = \int_0^\infty \frac{1}{2\gamma^2} E_{X_\gamma|x} [\|\hat{x}(x_\gamma) - \hat{x}(x_\gamma, r)\|^2] d\gamma \quad (19)$$

To obtain stimulus-wise decompositions from these expressions, that we can compare with our work, we averaged both of these decompositions with respect to $p(R|x)$, to obtain: $I_{Kong,1}(x) \equiv \mathbb{E}_{R|x} [I_{Kong,1}(x, R)]$ and $I_{Kong,2}(x) \equiv \mathbb{E}_{R|x} [I_{Kong,2}(x, R)]$.

Despite the apparent similarity with our proposed decomposition, neither $I_{Kong,1}(x)$ or $I_{Kong,2}(x)$ fulfill our axioms, limiting their potential use as principled & interpretable measures of neural sensitivity.

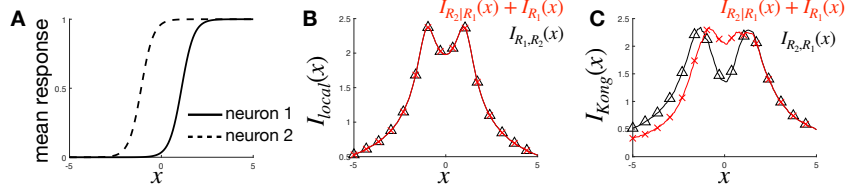


Figure 5: **Additivity, with respect to measurements from multiple neurons.** (A) We considered two neurons, with sigmoidal tuning curves, $f(x)$, and a 1-d stimulus with gaussian prior. (B) We confirmed that $I_{local}(x)$ satisfies additivity, by checking that the point-wise information from both neurons together ($I_{R_1, R_2}(x)$, red) equals $I_{R_1}(x) + I_{R_2}(x)$. (C) For the measure derived from the work of Kong et al., which does not satisfy additivity, the curves are non-overlapping.

In their paper, Kong et al. show that $I_{Kong,1}(x, r) = \log \frac{p(r|x)}{p(r)}$. Taking the average over $p(R|x)$, gives: $I_{Kong,1}(x) \equiv \mathbb{E}_{R|x} [I_{Kong,1}(x, r)] = D_{KL}(p(R|x) \| p(R))$. This is identical to the expression for the specific surprise, $I_{surp}(x)$, in the main text (Eqn 3). As shown in the main text (Fig 1E), this measure does not fulfill our locality axiom. This is because $I_{surp}(x)$ depends on $p(r) = \int p(r|x')p(r')dx'$, which can be influenced by changes to the likelihood or prior with respect to any stimulus, x' . Further, following the same prescription as in the main text to obtain a feature-wise decomposition of $I_{Kong,1}(x)$, results in feature-wise attributions that can be negative.

To understand the second decomposition, we use Tweedie’s law to write:

$$I_{Kong,2}(x) \equiv \mathbb{E}_{R|x} [I_{Kong,2}(x, r)] = \frac{1}{2} \int_0^\infty \mathbb{E}_{X_\gamma, R|x} [\|\nabla_{x_\gamma} \log p(R|X_\gamma)\|^2] d\gamma \quad (20)$$

This differs from our expression, due to the fact that it involves taking the average over $p(R|x)$, rather than $p(R|x_\gamma)$. However, while this difference may seem subtle it has important consequences for the resulting stimulus-wise decomposition.

To see that there are large qualitative differences between $I_{Kong,2}(x)$ and $I_{local}(x)$ we first consider their behavior in the case where $p(R, X)$ is jointly gaussian. Here, $I_{local}(x)$ is constant across x (Fig 2B, black), reflecting that the fact that the posterior is also independent of x (up to a linear translation). In contrast, $I_{Kong,2}(x)$ scales quadratically with the distance of x from the mean (similar to $I_{surp}(x)$, Fig 2E).

More importantly, $I_{Kong,2}(x)$ violates **additivity** point-wise. Supp Fig 1 illustrates this in a simple 1-d example. To see why it is the case, we can expand the expression for the local information from two neurons, R and R' :

$$I_{R, R'}^{Kong,2}(x) = \frac{1}{2} \int_0^\infty \mathbb{E}_{X_\gamma, R, R'|x} [\|\nabla_{x_\gamma} \log p(R'|R, X_\gamma) + \nabla_{x_\gamma} \log p(R|X_\gamma)\|^2] d\gamma \quad (21)$$

Now when we expand the square we see that the cross-terms *don’t cancel out*. As a result, $I_{Kong,2}$ is not linear in $\log p(R'|R, X_\gamma)$ and $\log p(R|X_\gamma)$, violating additivity.

This contrasts with the behaviour of I_{local} , which can be expressed as follows:

$$I_{R, R'}^{local}(x) = -\frac{1}{2} \int_0^\infty \mathbb{E}_{X_\gamma|x}, \left\{ \mathbb{E}_{R', R|X_\gamma} \left[\nabla_{x_\gamma}^2 \log p(R'|R, X_\gamma) + \nabla_{x_\gamma}^2 \log p(R|X_\gamma) \right] \right\} d\gamma \quad (22)$$

This equation is linear in $\log p(R'|R, X_\gamma)$ and $\log p(R|X_\gamma)$, and thus we can easily confirm that it is additive with respect to repeated measurements.

Additivity is a fundamental property that underpins the definition of mutual information [Shannon, 1948]. It requires that information from multiple measurements (e.g., different neurons) combine linearly, such that their individual contributions sum to the total local information (Fig. 5). A point-wise measure that doesn’t respect additivity could thus lead to misleading attributions as they don’t respect how different measurements (or neurons) combine additively to generate the total mutual information.

C Diffusion model details

C.1 Dataset Preparation

We utilized the MNIST dataset LeCun et al. [1998], which consists of 60,000 grayscale images of handwritten digits for training and 10,000 for testing. Each image, originally 28×28 pixels, was preprocessed by normalizing to the range $[-1, 1]$ and then rescaled to 32×32 pixels to align with the architectural requirements of our diffusion models.

C.2 Neural Encoding Model

We simulated a population of 49 neurons with spatially localized receptive fields (RFs) arranged in a 7×7 grid, mimicking the retinotopic organization of early visual cortex. Image pixels were mapped to 2D coordinates in visual space, $(x, y) \in [-1, 1]^2$, forming a uniform grid. Each neuron’s RF was modeled as a 2D Gaussian filter centered at (x_i, y_i) :

$$w_i(x, y) = A \cdot \exp\left(-\frac{(x - x_i)^2 + (y - y_i)^2}{2\sigma^2}\right), \quad (23)$$

where $A = 2$ is the amplitude, $\sigma = 0.2$ defines the spatial extent of the RF, and (x_i, y_i) specifies the center of the i -th neuron’s RF. These centers were evenly spaced across the image grid.

Neural firing rates were computed by projecting the input image I onto the receptive field weights w_i , followed by a sigmoid nonlinearity:

$$f_i = \frac{1}{1 + \exp(-g \cdot (w_i^\top I - \theta))}, \quad (24)$$

where $g = 1.0$ is the gain and $\theta = 0.5$ is the threshold. This yields the mean firing rate of neuron i in response to image I .

To capture neural variability, we modeled spike counts as samples from a Poisson distribution with mean f_i .

C.3 Diffusion Models

We trained two denoising diffusion probabilistic models (DDPMs) based on the *UNet-2D* architecture Ronneberger et al. [2015], using the implementation provided by the HuggingFace *diffusers* library von Platen et al. [2022].

The first model was trained to generate MNIST digits from standard Gaussian noise $Z \sim \mathcal{N}(0, I)$. Following the DDPM framework Ho et al. [2020], we simulated a forward diffusion process that progressively adds Gaussian noise to an image x_0 , producing a sequence of noisy images $\{x_t\}$. The process is defined as:

$$x_t = \sqrt{1 - \beta_t} \cdot x_{t-1} + \sqrt{\beta_t} \cdot z_t, \quad z_t \sim \mathcal{N}(0, I), \quad (25)$$

where β_t denotes the noise schedule. By recursively applying this equation, one obtains:

$$x_t \sim \mathcal{N}(\sqrt{\bar{\alpha}_t} \cdot x_0, (1 - \bar{\alpha}_t)I), \quad \text{with} \quad \bar{\alpha}_t = \prod_{s=1}^t (1 - \beta_s). \quad (26)$$

Although this differs from the isotropic Gaussian noise model used in the main text, $x_t = x_0 + \sqrt{\gamma}Z$, the two formulations are equivalent up to reparameterization, with $\gamma = \frac{1 - \bar{\alpha}_t}{\bar{\alpha}_t}$. Thus, both can be used interchangeably to estimate the information decomposition $I_{\text{local}}(x)$.

We used a linear noise schedule with β_t increasing from 1×10^{-4} to 0.02 over 1,000 time steps. Given a noisy image x_t , the model was trained to predict the noise component $\hat{z}_t(x_t)$, which can be used to reconstruct an estimate of the clean image:

$$\hat{x}(x_t) = \frac{x_t - \sqrt{1 - \bar{\alpha}_t} \cdot \hat{z}_t(x_t)}{\sqrt{\bar{\alpha}_t}}. \quad (27)$$

A second DDPM was trained using the same architecture, but conditioned on simulated neural responses r (see previous section). This model learned to predict the noise given both the noisy image and response:

$$\hat{x}(x_t, r) = \frac{x_t - \sqrt{1 - \bar{\alpha}_t} \cdot \hat{z}_t(x_t, r)}{\sqrt{\bar{\alpha}_t}}. \quad (28)$$

We used these models to estimate the pixel-wise information decomposition described in the main text (Eq. 9). To approximate the integral over γ , we applied additive Gaussian noise according to the diffusion schedule and evaluated $\hat{x}(x_t)$ and $\hat{x}(x_t, r)$ over 24 diffusion steps (indexed as [40:40:1000]) using 50 noisy samples of x_t and r at each t . On average, approximating $I_{local}(x)$ for one image took approximately 1 minute on an NVIDIA GeForce RTX 4080 GPU.

C.4 Training Details

Both models were implemented in PyTorch (version 2.7.0) with Cuda 12.6 and trained using the HuggingFace Diffusers library von Platen et al. [2022] and Accelerate library Gugger et al. [2022] for distributed training on 7 NVIDIA A100 GPUs (40 GB VRAM).

The training procedure for both models was as follows:

- Optimizer: AdamW with learning rate 1×10^{-4} , $\beta_1 = 0.9$, $\beta_2 = 0.999$
- Learning rate scheduler: Cosine schedule with warmup (500 steps)
- Batch size: 256
- Training duration: 150 epochs
- Loss function: Mean squared error (MSE) between predicted and true noise
- Precision: Mixed precision training with bfloat16
- Training Time: 25/30 minutes per model

C.5 Decomposition for additional digits

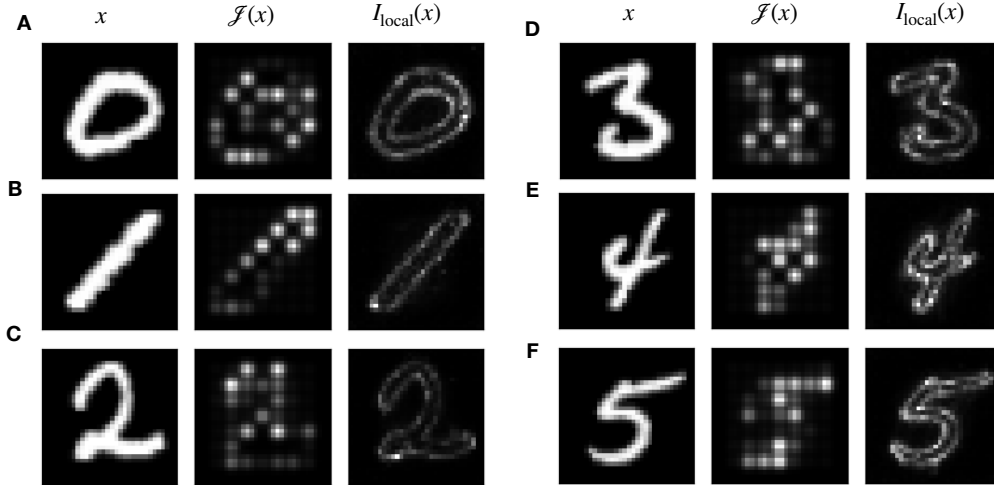


Figure 6: **Information decomposition across additional digits.** In the main text, we present pixel-wise information decomposition for two example digits. Here, we provide several additional examples to illustrate the decomposition patterns across other digits.

The code repository to reproduce figure 4 and 6 can be found at [neural-info-decomp](https://github.com/neural-info-decomp).