Multi-Modal Latent Variables for Cross-Individual Primary Visual Cortex Modeling and Analysis

Yu Zhu[1,2*], **Bo Lei**[4], **Chunfeng Song**[3†], **Wanli Ouyang**[3], **Shan Yu**[1†], **Tiejun Huang**[4]

[1]Institute of Automation, Chinese Academy of Sciences
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences
[3]Shanghai Artificial Intelligence Laboratory
[4]Beijing Academy of Artificial Intelligence
zhuyu2022@ia.ac.cn, songchunfeng@pjlab.org.cn, shan.yu@nlpr.ia.ac.cn

Abstract

Elucidating the functional mechanisms of the primary visual cortex (V1) remains a fundamental challenge in systems neuroscience. Current computational models face two critical limitations, namely the challenge of cross-modal integration between partial neural recordings and complex visual stimuli, and the inherent variability in neural characteristics across individuals, including differences in neuron populations and firing patterns. To address these challenges, we present a multi-modal identifiable variational autoencoder (miVAE) that employs a two-level disentanglement strategy to map neural activity and visual stimuli into a unified latent space. This framework enables robust identification of cross-modal correlations through refined latent space modeling. We complement this with a novel score-based attribution analysis that traces latent variables back to their origins in the source data space. Evaluation on a large-scale mouse V1 dataset demonstrates that our method achieves state-of-the-art performance in cross-individual latent representation and alignment, without requiring subject-specific fine-tuning, and exhibits improved performance with increasing data size. Significantly, our attribution algorithm successfully identifies distinct neuronal subpopulations characterized by unique temporal patterns and stimulus discrimination properties, while simultaneously revealing stimulus regions that show specific sensitivity to edge features and luminance variations. This scalable framework offers promising applications not only for advancing V1 research but also for broader investigations in neuroscience.

Introduction

The primary visual cortex (V1) plays a fundamental role in hierarchical visual information processing, making its functional characterization essential to systems neuroscience. Recent advances in calcium imaging technology (Sofroniew et al. 2016) have accelerated V1 research by enabling recording of large neuronal populations in multiple subjects. A central goal in utilizing this technology is to identify distinct patterns of neural activity and establish their relationships with specific visual stimuli (Stringer et al. 2019).

Current approaches to understanding V1 visual information processing primarily follow two paradigms. The first develops encoding models (Sinz et al. 2018; Wang et al. 2023) that map visual stimuli to neural activity, simulating V1 function. The second focuses on decoding neural activity to understand the representation of visual features such as edges and motion (Yoshida and Ohki 2020). However, these approaches face fundamental limitations due to the restricted shared subspace between sparse, locally-recorded neural activity and complex visual features, making it challenging to fully capture brain-vision relationships. Furthermore, these models often make the oversimplified assumption that locally recorded V1 populations can fully encode complete visual stimuli, despite evidence that visual processing is distributed across the extensive V1 area according to retinotopic mapping (Figure 1.A).

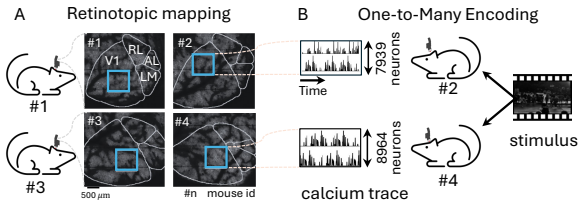

Figure 1: **Challenges in modeling V1.** (A) Retinotopic mapping (RL, AL and LM refers to other visual cortex) and the partially recorded field of view of calcium imaging (blue rectangle). (B) One-to-many challenge with cross-individual heterogeneity, especially neuron counts and firing patterns.

Additionally, cross-individual neural heterogeneity presents an additional significant challenge. Neural responses to identical stimuli can vary substantially among individuals (Guntupalli et al. 2016; Haxby et al. 2001), complicating cross-individual modeling efforts (Figure 1.B). Addressing this variability requires zero-shot or few-shot domain adaptation capabilities, a challenge that remains significant even in contemporary machine learning. Previous approaches (Wang et al. 2023) requiring individual-specific fine-tuning have not adequately addressed this cross-individual variability.

To address these challenges, we introduce a **m**ulti-modal **i**dentifiable **v**ariational **a**uto**e**ncoder (miVAE) featuring two-level latent space disentanglement. For neural activity, we separately model idiosyncratic and preserved latent variables, capturing both subject-specific characteristics and function-

*This work was done during his internship at Shanghai AI Lab
†Corresponding Author

ally consistent information across individuals. For visual stimuli, we distinguish between neural activity-related and unrelated latent variables, refining the preserved variables to capture highly relevant neural correlates. We further enhance these highly correlated variables through latent modeling and introduce a novel score-based attribution strategy for comprehensive data interpretation. Validated on a mouse V1 dataset (Turishcheva et al. 2023) using two-photon calcium imaging (Grienberger and Konnerth 2012), miVAE demonstrates superior cross-individual latent coding without requiring individual-specific fine-tuning, showing scalable performance with increasing dataset size and remarkable consistency across individuals exposed to identical visual stimuli. This cross-individual approach significantly enhances the scalability of data-driven research. Our attribution analysis successfully identifies key neuronal subpopulations with distinct stimulus-related responses and superior discriminative capabilities, while highlighting V1's sensitivity to edge and luminance patterns. This framework presents a powerful tool for V1 research with potential applications across various sensory cortices.

Our primary contributions include:

- **Multi-modal identifiable modeling with latent space refinement.** miVAE effectively captures highly correlated variables between visual stimuli and neural responses, enhanced through pair-wise modeling in the latent space.
- **Scalable cross-individual performance.** Our approach achieves state-of-the-art modeling and cross-individual alignment without subject-specific fine-tuning, demonstrating improved performance with increased data size.
- **Score-based attribution analysis with biological insights.** This novel method maps latent variables to original data, identifying key neurons with distinct temporal patterns and stimulus discriminative capabilities, while revealing visual regions sensitive to edges and luminance.

Related Work

Latent Variable Modeling. The advent of calcium imaging has provided unprecedented access to neural data while simultaneously introducing significant analytical challenges. A prominent approach to managing this complexity involves mapping high-dimensional neural signals to low-dimensional latent spaces. Early research focused on extracting latent variables from neural population activities using simple, prior-driven designs (Yu et al. 2008; Zhao and Park 2017). This evolved into more sophisticated methods employing recurrent neural network (RNN)-based explicit dynamics modeling (Pandarinath et al. 2018; Keshtkaran et al. 2022; Zhu et al. 2022), which yielded highly interpretable results. The increasing scale of neural activity data led to the development of more efficient approaches, including Transformer-based methods with implicit latent representation learning (Ye and Pandarinath 2021; Liu et al. 2022; Le and Shlizerman 2022; Ye et al. 2024; Antoniades et al. 2024). While these demonstrated superior decoding performance, they often lacked interpretability and identifiability. To address these limitations, interpretable approaches such as pi-VAE (Zhou and Wei 2020) and CEBRA (Schneider, Lee, and Mathis 2023) introduced auxiliary variables for identifiable latent variables, building upon identifiable variational autoencoders (iVAE) (Khemakhem et al. 2020) and nonlinear independent component analysis (ICA) (Hyvarinen, Sasaki, and Turner 2019). However, these methods prove inadequate for direct cross-individual analysis and fail to account for individual variability. Our approach differs by explicitly modeling cross-individual variability and separating neural activity components into stimulus-related and unrelated elements. Through multi-modal and bi-directed generative modeling, we effectively isolate highly relevant latent variables between neural activity and visual stimuli, ensuring interpretability while enabling sophisticated latent space modeling and analysis.

Multi-Modal VAEs. In machine learning, multi-modal VAEs aim to learn joint posterior approximations across modalities (Wu and Goodman 2018; Shi et al. 2019; Sutter, Daunhawer, and Vogt 2021). Traditional approaches often rely on restrictive assumptions, particularly regarding latent space aggregation (Sutter, Daunhawer, and Vogt 2021). Recent developments (Palumbo, Daunhawer, and Vogt 2023) have introduced modality-specific latent subspaces to enhance generative quality, while the latest MMVE (Sutter et al. 2024) implements data-dependent prior distributions for improved posterior approximation regulation. In neuroscience applications, emerging models like MM-GPVAE (Gondur et al. 2024) share our miVAE's latent space partition strategy. However, while these methods aggregate latent variables of neural activity and behavior, our approach implements directed generative modeling where neural activity and visual stimuli serve as mutual priors, avoiding fusion operations and emphasizing cross-individual modeling.

Functional Modeling and Stimuli Reconstruction of V1. Understanding V1 requires both functional modeling and neural activity decoding. Neural networks have successfully mapped visual stimuli to neural activity, revealing encoding mechanisms (Sinz et al. 2018; Ecker et al. 2019; Bashiri et al. 2021; Lurz et al. 2020; Ma et al. 2024). Wang et al. (Wang et al. 2023) proposed a shared feature extractor with separate readouts for V1 modeling, partially addressing the one-to-many challenge, though still requiring individual-specific fine-tuning. Current models often incorrectly assume complete visual information encoding in local V1 activity (Hubel and Wiesel 1977; Tootell et al. 1982). Decoding models focus on visual stimuli reconstruction from neural activity (Cobos et al. 2022; Ellis and Michaelides 2018; Yoshida and Ohki 2020), revealing interpretable variance (Berens et al. 2012; Eichhorn et al. 2003; Froudarakis et al. 2014; Garasto, Bharath, and Schultz 2018). Recent algorithms employ both linear (Ellis and Michaelides 2018; Yoshida and Ohki 2020) and nonlinear approaches (Cobos et al. 2022), yet maintain the assumption of complete visual information encoding in local V1 activity. Our approach acknowledges that recorded local V1 activity contains only partial visual information (Olshausen and Field 2004; Roe et al. 2012) and utilizes relevant latent variables from multi-modal modeling for both encoding and decoding. Additionally, we introduce a novel score-based attribution strategy for mapping latent variables to original data, providing new insights into neural processing mechanisms.

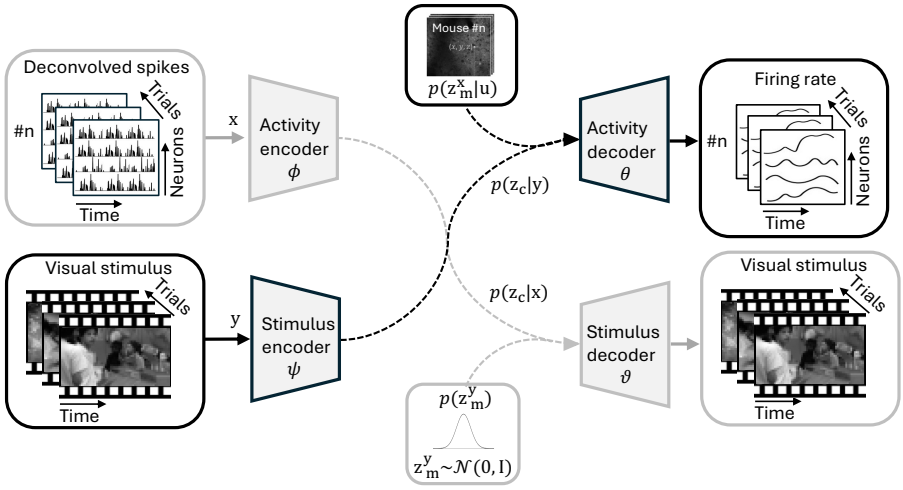

Figure 2: **miVAE for modeling V1 neural activity $\mathbf{x}$ and the corresponding visual stimulus $\mathbf{y}$.** $\mathbf{x}$ and $\mathbf{y}$ are decomposed in the latent space. For $\mathbf{x}$, it has the idiosyncratic latent variable $\mathbf{z^x_m}$ and the cross-subject and functionally relevant preserved latent variable $\mathbf{z^x_c}$. For $\mathbf{y}$, it has the neural activity-relevant variable $\mathbf{z^y_c}$ and the irrelevant variable $\mathbf{z^y_m}$. Our method employ generative modeling to derive these variables, with special focus on acquiring the most correlated variable $\mathbf{z^x_c}$, $\mathbf{z^y_c}$.

Method

Our method comprises three components, specifically, **Cross-Individual Multi-Modal and Cross-Modal Modeling** to extract relevant hidden variables between neural activity and visual stimulation, **Neural Encoding and Decoding in Latent Space.** for further representation alignment, and **Score-based Attribution Analysis** for interpretative analysis.

Cross-Individual Multi-Modal Modeling

We present a framework for modeling multi-modal data consisting of neural activity $\mathbf{x} \in \mathbb{R}^{N\times T}$ and visual stimulus $\mathbf{y} \in \mathbb{R}^{H\times W\times T}$ across individuals. The framework addresses both modalities through complementary approaches.

For neural activity $\mathbf{x}$, we derive bi-identifiable latent variables $\mathbf{z^x} \in \mathbb{R}^{d\times T}$ ($d \leq N$), comprising idiosyncratic latent variables $\mathbf{z^x_m} \in \mathbb{R}^{d/2\times T}$ and cross-individual preserved latent variables $\mathbf{z^x_c} \in \mathbb{R}^{d/2\times T}$. Individual-specific information $\mathbf{u}$ enhances the identifiability of $\mathbf{z^x_m}$, while visual stimulus $\mathbf{y}$ serves as a functional prior for $\mathbf{z^x_c}$.

For visual stimulus $\mathbf{y}$, we construct semi-identifiable latent variables $\mathbf{z^y} \in \mathbb{R}^{d\times T}$, separated into activity-related latent variables $\mathbf{z^y_c} \in \mathbb{R}^{d/2\times T}$ and activity-irrelevant latent variables $\mathbf{z^y_m} \in \mathbb{R}^{d/2\times T}$. Thus, the latent variables $\mathbf{z^y_c}$ and $\mathbf{z^x_c}$ are highly correlated, which means there exists the ideal $\mathbf{z_c}$ derived either from $\mathbf{x}$ or $\mathbf{y}$.

The multi-modal generative latent modeling for $p(\mathbf{x}, \mathbf{y}|\mathbf{u})$ employs two complementary generative models, [1] **activity-based** and [2] **stimulus-based**:

$$\begin{aligned} &p(\mathbf{x}, \mathbf{y}, \mathbf{z^x_m}, \mathbf{z_c}, \mathbf{z^y_m}|\mathbf{u}) \\ &= p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z^x_m}, \mathbf{z_c}) p_{\mathbf{T},\boldsymbol{\lambda}}(\mathbf{z^x_m}|\mathbf{u}) p_{\boldsymbol{\psi}}(\mathbf{z_c}|\mathbf{y}) p(\mathbf{y}, \mathbf{z^y_m}) \text{ [1]} \\ &= p_{\boldsymbol{\vartheta}}(\mathbf{y}|\mathbf{z_c}, \mathbf{z^y_m}) p_{\boldsymbol{\phi}}(\mathbf{z_c}|\mathbf{x}) p(\mathbf{z^y_m}) p(\mathbf{x}, \mathbf{z^x_m}|\mathbf{u}) \text{ [2]} \end{aligned} \quad (1)$$

As shown in Figure 2, neural activity $\mathbf{x}$ and visual stimuli $\mathbf{y}$ serve as mutual priors, connected through $\mathbf{z_c}$.

In the **activity-based** model, we first define the prior for the idiosyncratic latent variable $\mathbf{z^x_m}$. Drawing from the universal approximation capabilities of exponential families (Sriperumbudur et al. 2017), we employ a factorized exponential family distribution conditioned on $\mathbf{u}$ (Khemakhem et al. 2020):

$$p_{\mathbf{T},\boldsymbol{\lambda}}(\mathbf{z^x_m}|\mathbf{u}) = \prod_{i=1}^{m} \frac{Q_i(\mathbf{z^x_{m}}_i)}{Z_i(\mathbf{u})} \exp[\sum_{j=1}^{k} T_{i,j}(\mathbf{z^x_{m}}_i)\lambda_{i,j}(\mathbf{u})] \quad (2)$$

where Q_i denotes the base measure, $Z_i(\mathbf{u})$ represents the normalizing constant, $\mathbf{T}_i = (T_{i,1}, \ldots, T_{i,k})$ comprises sufficient statistics, and $\boldsymbol{\lambda}_i(\mathbf{u}) = (\lambda_{i,1}(\mathbf{u}), \ldots, \lambda_{i,k}(\mathbf{u}))$ represents the corresponding parameters. The dimension k of each sufficient statistic is predefined, with $\boldsymbol{\lambda}(\cdot)$ parameterized through a neural network.

Then, the prior for the key latent variable $\mathbf{z^x_c}$ reflects the functional uniformity of V1 activity across individuals (Safaie et al. 2023). Given that recorded V1 activity is stimulus-evoked and functionally consistent across individuals, we employ a factorized exponential family distribution $p_{\boldsymbol{\psi}}(\mathbf{z_c}|\mathbf{y})$, with statistics inferred through network $\boldsymbol{\psi}$.

For the **stimulus-based** modeling, we first introduce the prior of the activity-related latent variable, $\mathbf{z^y_c}$. Due to retinotopic mapping, only part of the visual stimulus information is related to the recorded local V1 activity (Hubel and Wiesel 1977; Tootell et al. 1982). Therefore, in this stimulus-based modeling, we aim to extract the latent variable most correlated between neural activity $\mathbf{x}$ and visual stimulus $\mathbf{y}$. We use $p_{\boldsymbol{\phi}}(\mathbf{z_c}|\mathbf{x})$ as the prior, and parameterize it with network $\boldsymbol{\phi}$.

Furthermore, we account for the prior of the activity-irrelevant latent variable, $\mathbf{z^y_m}$. Given that receptive fields cannot be treated as specific, easily processable data (such as identity-related information) that neural networks can readily handle, we instead introduce a standard normal distribution prior for $p(\mathbf{z^y_m})$, specifically $\mathbf{z^y_m} \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$, as a substitute.

Note that we emphasize that if more tractable prior information regarding receptive fields becomes available, the current modeling framework can be entirely restructured into a learnable, parameterized prior, analogous to the idiosyncratic latent variable $\mathbf{z}_{\mathbf{m}}^{\mathbf{x}}$.

Directed Cross-Modal Generative Modeling

The intrinsic correlation between the recorded neural activity $\mathbf{x}$ and visual stimulus $\mathbf{y}$ motivates our cross-modal modeling approach. Building on established neurophysiological principles (Hubel and Wiesel 1962; Niell and Stryker 2008), we model the stimulus-evoked V1 neural activity as:

$$p(\mathbf{x}|\mathbf{u},\mathbf{y}) = p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}_{\mathbf{m}}^{\mathbf{x}},\mathbf{z}_{\mathbf{c}})p_{\mathbf{T},\boldsymbol{\lambda}}(\mathbf{z}_{\mathbf{m}}^{\mathbf{x}}|\mathbf{u})p_{\boldsymbol{\psi}}(\mathbf{z}_{\mathbf{c}}|\mathbf{y}) \quad (3)$$

Conversely, acknowledging that neural activity encodes visual stimulus information, we model this relationship in the latent space while accounting for activity-independent stimulus components:

$$p(\mathbf{y}|\mathbf{x}) = p_{\boldsymbol{\vartheta}}(\mathbf{y}|\mathbf{z}_{\mathbf{c}},\mathbf{z}_{\mathbf{m}}^{\mathbf{y}})p_{\boldsymbol{\phi}}(\mathbf{z}_{\mathbf{c}}|\mathbf{x})p(\mathbf{z}_{\mathbf{m}}^{\mathbf{y}}) \quad (4)$$

This bidirectional modeling approach, expressed through Equations 3 and 4, reinforces the correlation between neural activity and visual stimuli, particularly in extracting the shared components $\mathbf{z}_{\mathbf{c}}$ in the latent space, as validated by experimental results.

Tackling Dual Modeling via Variational Inference

We adopt variational inference (Kingma and Welling 2014) to approximate the intractable true posteriors $p(\mathbf{z}_{\mathbf{m}}^{\mathbf{x}}|\mathbf{x},\mathbf{u})$, $p(\mathbf{z}_{\mathbf{c}}|\mathbf{x},\mathbf{y})$, and $p(\mathbf{z}_{\mathbf{m}}^{\mathbf{y}}|\mathbf{y})$ with $q(\mathbf{z}_{\mathbf{m}}^{\mathbf{x}}|\mathbf{x},\mathbf{u})$, $q(\mathbf{z}_{\mathbf{c}}|\mathbf{x},\mathbf{y})$, and $q(\mathbf{z}_{\mathbf{m}}^{\mathbf{y}}|\mathbf{y})$ respectively. To optimize both the model and the approximate posteriors, we maximize the combined variational evidence lower bound (ELBO) of $p(\mathbf{x},\mathbf{y}|\mathbf{u})$, $p(\mathbf{x}|\mathbf{y},\mathbf{u})$, and $p(\mathbf{y}|\mathbf{x})$, which is equivalent to minimizing the following:

$$\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\phi},\boldsymbol{\vartheta},\boldsymbol{\psi}) = \mathcal{L}_{\text{MM}} + \mathcal{L}_{\text{CM}} \quad (5)$$

where $\mathcal{L}_{\text{MM}}$ represents the multi-modal loss and $\mathcal{L}_{\text{CM}}$ represents the cross-modal loss. Specifically, $\mathcal{L}_{\text{MM}}$ is as follows:

$$\begin{aligned}\mathcal{L}_{\text{MM}} &= -\mathbb{E}_{q(\mathbf{z}_{\mathbf{m}}^{\mathbf{x}}|\mathbf{x},\mathbf{u})q(\mathbf{z}_{\mathbf{c}}|\mathbf{x},\mathbf{y})}[\log p(\mathbf{x}|\mathbf{z}_{\mathbf{m}}^{\mathbf{x}},\mathbf{z}_{\mathbf{c}})] \\ &+ KL[q(\mathbf{z}_{\mathbf{m}}^{\mathbf{x}}|\mathbf{x},\mathbf{u})||p(\mathbf{z}_{\mathbf{m}}^{\mathbf{x}}|\mathbf{u})] + KL[q(\mathbf{z}_{\mathbf{c}}|\mathbf{x},\mathbf{y})||p(\mathbf{z}_{\mathbf{c}}|\mathbf{y})] \\ &- \mathbb{E}_{q(\mathbf{z}_{\mathbf{c}}|\mathbf{x},\mathbf{y})q(\mathbf{z}_{\mathbf{m}}^{\mathbf{y}}|\mathbf{y})}[\log p(\mathbf{y}|\mathbf{z}_{\mathbf{c}},\mathbf{z}_{\mathbf{m}}^{\mathbf{y}})] \\ &+ KL[q(\mathbf{z}_{\mathbf{m}}^{\mathbf{y}}|\mathbf{y})||p(\mathbf{z}_{\mathbf{m}}^{\mathbf{y}})] + KL[q(\mathbf{z}_{\mathbf{c}}|\mathbf{x},\mathbf{y})||p(\mathbf{z}_{\mathbf{c}}|\mathbf{x})]\end{aligned} \quad (6)$$

For the cross-modal loss $\mathcal{L}_{\text{CM}}$, which is responsible for enhancing correlation modeling, we have:

$$\begin{aligned}\mathcal{L}_{\text{CM}} &= -\mathbb{E}_{p(\mathbf{z}_{\mathbf{m}}^{\mathbf{x}}|\mathbf{u})p(\mathbf{z}_{\mathbf{c}}|\mathbf{y})}[\log p(\mathbf{x}|\mathbf{z}_{\mathbf{m}}^{\mathbf{x}},\mathbf{z}_{\mathbf{c}})] \\ &+ KL[p(\mathbf{z}_{\mathbf{m}}^{\mathbf{x}}|\mathbf{u})||q(\mathbf{z}_{\mathbf{m}}^{\mathbf{x}}|\mathbf{x},\mathbf{u})] + KL[p(\mathbf{z}_{\mathbf{c}}|\mathbf{y})||q(\mathbf{z}_{\mathbf{c}}|\mathbf{x},\mathbf{y})] \\ &- \mathbb{E}_{p(\mathbf{z}_{\mathbf{c}}|\mathbf{x})p(\mathbf{z}_{\mathbf{m}}^{\mathbf{y}})}[\log p(\mathbf{y}|\mathbf{z}_{\mathbf{c}},\mathbf{z}_{\mathbf{m}}^{\mathbf{y}})] \\ &+ KL[p(\mathbf{z}_{\mathbf{c}}|\mathbf{x})||q(\mathbf{z}_{\mathbf{c}}|\mathbf{x},\mathbf{y})]\end{aligned} \quad (7)$$

Detailed derivations are shown in Supplementary Materials. And we especially include the discussion of avoiding spillover effect for latent disentanglement in the Supplementary Materials. Next, we introduce the approximate posteriors for each latent variables.

Approximate posterior of idiosyncratic latent variable $q(\mathbf{z}_{\mathbf{m}}^{\mathbf{x}}|\mathbf{x},\mathbf{u})$. For $\mathbf{z}_{\mathbf{m}}^{\mathbf{x}}$, we decompose the approximate posterior using a factorized exponential family distribution, following (Johnson et al. 2016; Zhou and Wei 2020)

$$q(\mathbf{z}_{\mathbf{m}}^{\mathbf{x}}|\mathbf{x},\mathbf{u}) \propto q_{\boldsymbol{\phi}}(\mathbf{z}_{\mathbf{m}}^{\mathbf{x}}|\mathbf{x})p_{\mathbf{T},\boldsymbol{\lambda}}(\mathbf{z}_{\mathbf{m}}^{\mathbf{x}}|\mathbf{u}) \quad (8)$$

where $q_{\boldsymbol{\phi}}(\mathbf{z}_{\mathbf{m}}^{\mathbf{x}}|\mathbf{x})$ is assumed to be conditionally independent Gaussian distribution, i.e., $q_{\boldsymbol{\phi}}(\mathbf{z}_{\mathbf{m}}^{\mathbf{x}}|\mathbf{x}) = \prod_{c=1}^{m} q(\mathbf{z}_{\mathbf{m}_c}^{\mathbf{x}}|\mathbf{x})$. Additionally, we assume that $q_{\boldsymbol{\phi}}(\mathbf{z}_{\mathbf{m}}^{\mathbf{x}}|\mathbf{x})$ and $p_{\mathbf{T},\boldsymbol{\lambda}}(\mathbf{z}_{\mathbf{m}}^{\mathbf{x}}|\mathbf{u})$ are independent distributions.

Approximate posterior of preserved latent variable $q(\mathbf{z}_{\mathbf{c}}|\mathbf{x},\mathbf{y})$. Similar to $\mathbf{z}_{\mathbf{m}}^{\mathbf{x}}$, the approximate posterior for $\mathbf{z}_{\mathbf{c}}$ is modeled using a factorized exponential family distribution. We assume conditionally independent Gaussian distributions $q_{\boldsymbol{\phi}}(\mathbf{z}_{\mathbf{c}}|\mathbf{x})$ and $p_{\boldsymbol{\psi}}(\mathbf{z}_{\mathbf{c}}|\mathbf{y})$.

$$q(\mathbf{z}_{\mathbf{c}}|\mathbf{x},\mathbf{y}) \propto \begin{cases} q_{\boldsymbol{\phi}}(\mathbf{z}_{\mathbf{c}}|\mathbf{x})p_{\boldsymbol{\psi}}(\mathbf{z}_{\mathbf{c}}|\mathbf{y}), & \text{for } \mathbf{x} \\ q_{\boldsymbol{\psi}}(\mathbf{z}_{\mathbf{c}}|\mathbf{y})p_{\boldsymbol{\phi}}(\mathbf{z}_{\mathbf{c}}|\mathbf{x}), & \text{for } \mathbf{y} \end{cases} \quad (9)$$

It is important to note that for neural activity $\mathbf{x}$, the auxiliary variable is the visual stimulus $\mathbf{y}$, with $p_{\boldsymbol{\psi}}(\mathbf{z}_{\mathbf{c}}|\mathbf{y})$ as the prior. Conversely, for the visual stimulus $\mathbf{y}$, the auxiliary variable is the neural activity $\mathbf{x}$, with $p_{\boldsymbol{\phi}}(\mathbf{z}_{\mathbf{c}}|\mathbf{x})$ as the prior. Consequently, $q(\mathbf{z}_{\mathbf{c}}|\mathbf{x},\mathbf{y})$ serves as the shared approximate posterior for both $p(\mathbf{z}_{\mathbf{c}}|\mathbf{y})$ and $p(\mathbf{z}_{\mathbf{c}}|\mathbf{x})$. This consistency ensures that the latent variable $\mathbf{z}_{\mathbf{c}}$ captures the shared information between neural activity $\mathbf{x}$ and visual stimulus $\mathbf{y}$. By aligning the latent variables produced by both encoders within a unified latent space, it facilitates the learning of robust and interpretable latent variables, which is crucial for subsequent encoding-decoding modeling and analysis.

Approximate posterior of neural activity-irrelevant variable $q(\mathbf{z}_{\mathbf{m}}^{\mathbf{y}}|\mathbf{y})$. The approximate posterior $q(\mathbf{z}_{\mathbf{m}}^{\mathbf{y}}|\mathbf{y})$ follows the same approach as the original VAE (Kingma and Welling 2014). It is a Gaussian distribution inferred by the visual stimulus encoding network $\boldsymbol{\phi}$ with input $\mathbf{y}$.

Neural Encoding and Decoding in the Latent Space

The above modeling enable us to acquire the highly related variable $\mathbf{z}_{\mathbf{c}}^{\mathbf{x}}$, $\mathbf{z}_{\mathbf{c}}^{\mathbf{y}}$ between the neural activity $\mathbf{x}$ and visual stimulus $\mathbf{y}$. Here, we further propose to modeling the encoding and decoding in the latent space, based on $\mathbf{z}_{\mathbf{c}}^{\mathbf{x}}$, $\mathbf{z}_{\mathbf{c}}^{\mathbf{y}}$.

Given the pretrained miVAE, we have two encoders for neural activity $\mathbf{x}$ and visual stimulus $\mathbf{y}$ with parameters $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$, respectively. Therefore, we have $\mathbf{z}_{\mathbf{c}}^{\mathbf{x}} \sim p_{\boldsymbol{\phi}}(\mathbf{z}_{\mathbf{c}}|\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)$ and $\mathbf{z}_{\mathbf{c}}^{\mathbf{y}} \sim p_{\boldsymbol{\psi}}(\mathbf{z}_{\mathbf{c}}|\mathbf{y}) = \mathcal{N}(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)$ [1]. Here, we transform the original neural encoding and decoding into a distribution matching problem in the latent space. For neural encoding, we map $\mathbf{z}_{\mathbf{c}}^{\mathbf{y}} \sim p_{\boldsymbol{\psi}}(\mathbf{z}_{\mathbf{c}}|\mathbf{y})$ to $\mathbf{z}_{\mathbf{c}}^{\mathbf{x}} \sim p_{\boldsymbol{\phi}}(\mathbf{z}_{\mathbf{c}}|\mathbf{x})$, and similarly for neural decoding, as follows:

$$\begin{cases} \mathbf{z}_{\mathbf{c}}^{\mathbf{x}} = \mathbf{A}_1^T\mathbf{z}_{\mathbf{c}}^{\mathbf{y}} + \mathbf{b}_1, \mathbf{A}_1^* = \boldsymbol{\Sigma}_1^{1/2}\boldsymbol{\Sigma}_2^{-1/2}, \mathbf{b}_1^* = \boldsymbol{\mu}_1 - \mathbf{A}_1^*\boldsymbol{\mu}_2, \\ \mathbf{z}_{\mathbf{c}}^{\mathbf{y}} = \mathbf{A}_2^T\mathbf{z}_{\mathbf{c}}^{\mathbf{x}} + \mathbf{b}_2, \mathbf{A}_2^* = \boldsymbol{\Sigma}_2^{1/2}\boldsymbol{\Sigma}_1^{-1/2}, \mathbf{b}_2^* = \boldsymbol{\mu}_2 - \mathbf{A}_2^*\boldsymbol{\mu}_1, \end{cases} \quad (10)$$

[1]Note that the approximating posterior q in miVAE are currently the known/prior distribution p in latent coding modeling. Thus, for ease of illustration, we denote all distributions with p in this latent coding section and following attribution analysis section.

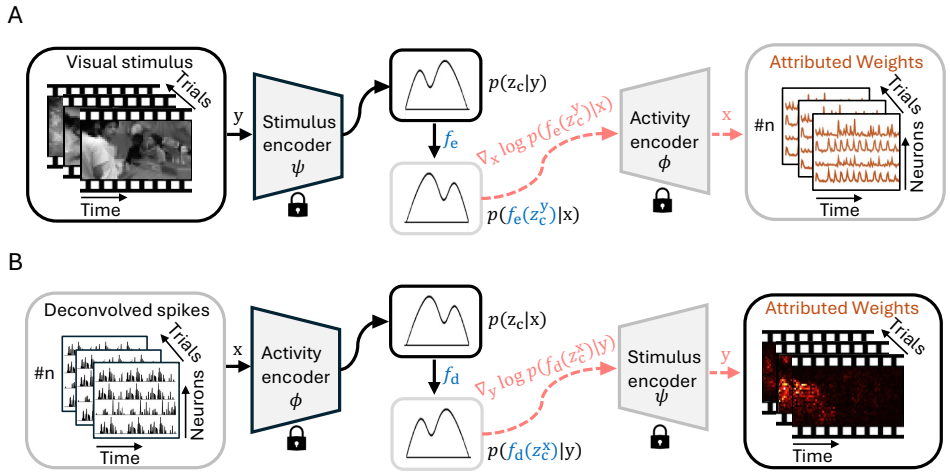

Figure 3: **Neural coding in the latent space with following score-based Attribution Analysis.** (A) The latent encoding function f_e leverages pretrained miVAE encoders to align $\mathbf{z}_c^x$ and $\mathbf{z}_c^y$ in the shared latent space. Subsequently, the pretrained neural activity encoder ϕ functions as a *decoder*, enabling our score-based attribution analysis to identify salient neurons at each temporal point. (B) For latent decoding, an analogous approach using f_d reveals key visual regions through the same attribution framework.

where $\mathbf{A}_1^*$, $\mathbf{b}_1^*$, $\mathbf{A}_2^*$ and $\mathbf{b}_2^*$ represent the optimal transform solutions for single paired ($\mathbf{z}_c^x$, $\mathbf{z}_c^y$). For cross-individual validation, these parameters are learned using either linear or nonlinear functions. We shown such encoding and decoding functions in the middle of Figure 3. During training, we use the KL-divergence that matching the transformed distribution and target distribution as the loss function. Modeling in latent space with unified latent size enable us to easily extend the cross-individual modeling and following analysis.

Score-based Attribution Analysis

Encoding and decoding in the latent space can further improve the correlations across $\mathbf{z}_c^x$ and $\mathbf{z}_c^y$. However, this does not yet provide a direct interpretation of the original data, which is critical for pellucid analysis. To solve this problem, we propose an new algorithm to trace back to the original data, based on the better-aligned latent variables, as shown in Figure 3.

Specifically, we present an attribution strategy based on a score function, treating the neural activity encoder as a *decoder*. For $\mathbf{z}_c^y \rightarrow \mathbf{z}_c^x$, this attribution reveals the preferences of neuron subpopulation within a neural population for specific visual stimuli or measures the importance of these subpopulation for given visual stimuli at any given timestep. Similarly, for $\mathbf{z}_c^x \rightarrow \mathbf{z}_c^y$, we treat the visual stimulus encoder as a *decoder* to analyze which parts of the visual stimulus are critical for the given neural activity.

To illustrate our attribution strategy, we illustrate the principles behind the latent encoding $\mathbf{z}_c^y \rightarrow \mathbf{z}_c^x$ attribution strategy. We have the transformed distribution for $\mathbf{x}$ given by $p(f_e(\mathbf{z}_c^y)|\mathbf{y}) \approx p(f_e(\mathbf{z}_c^y)|\mathbf{x})$. By applying Baysian rule, $p(f_e(\mathbf{z}_c^y)|\mathbf{x}) = \frac{p(\mathbf{x}|f_e(\mathbf{z}_c^y))p(f_e(\mathbf{z}_c^y))}{p(\mathbf{x})}$, then we have following score function:

$$\begin{aligned}\nabla_{\mathbf{x}} \log p(f_e(\mathbf{z}_c^y)|\mathbf{y}) &\approx \nabla_{\mathbf{x}} \log p(f_e(\mathbf{z}_c^y)|\mathbf{x}) \\ &= \nabla_{\mathbf{x}} \log p(\mathbf{x}|f_e(\mathbf{z}_c^y)) - \nabla_{\mathbf{x}} \log p(\mathbf{x})\end{aligned} \tag{11}$$

where $\nabla_{\mathbf{x}} \log p(f_e(\mathbf{z}_c^y)|\mathbf{x})$ represents the score (Fisher 1970) of the transformed latent distribution $f_e(\mathbf{z}_c^y)$ comprises two key components. The first term, $\nabla_{\mathbf{x}} \log p(\mathbf{x}|f_e(\mathbf{z}_c^y))$, represents the gradient of the log-likelihood with respect to observed neural activity $\mathbf{x}$, given the transformed latent variable (distribution). The second term, $\nabla_{\mathbf{x}} \log p(\mathbf{x})$, acts as a correction factor through the marginal log-likelihood gradient. This formulation enables $\nabla_{\mathbf{x}} \log p(f_e(\mathbf{z}_c^y)|\mathbf{x})$ to quantify neuronal importance relative to $\mathbf{z}_c^y$ at any timestep. The same principles extend to latent decoding ($\mathbf{z}_c^x \rightarrow \mathbf{z}_c^y$), providing a comprehensive framework for attribution analysis in both encoding and decoding processes.

In summary, the latent variables $\mathbf{z}_c^x$, $\mathbf{z}_c^y$ and learned f_e and f_d serve as bridges connecting neural activity $\mathbf{x}$ and visual stimuli $\mathbf{y}$. The proposed score function analysis enable us get rid of the prior of true distributions of $\mathbf{z}_c^x$ and $\mathbf{z}_c^y$, and can generate new insights into the original data space, potentially serving as a new alternative analysis tool for neural activity.

Experiments

We evaluated our approach using single-trial calcium imaging data from the Sensorium 2023 dataset (Turishcheva et al. 2023), comprising neural recordings from 10 head-fixed mice paired with corresponding visual stimuli. The dataset included 5 distinct mouse pairs, where each pair viewed identical visual sequences. Neural activity was recorded via two-photon calcium imaging (Sofroniew et al. 2016) and subsequently deconvolved into spike trains (Turishcheva et al. 2023), yielding recordings from 78,853 neurons. All recordings were temporally aligned at 30Hz sampling rate, with each trial spanning approximately 10 seconds. The visual stimuli consisted of movies at 36×64 pixel resolution.

Data from 7 mice were used for training, with the remaining 3 mice reserved for validation. All results are reported on previously unseen mice. Detailed experimental protocols are provided in the Supplementary Materials.

Methods	Stage 1	Stage 2			
		Latent Encoding		Latent Decoding	
		Linear	nonLinear	Linear	nonLinear
V1-FM$^*_e$	0.2169*	-	-	-	-
V1-FM$^*_d$	0.3626*	-	-	-	-
VAE†	0.0004	0.1869	0.1989	0.0074	0.0331
iVAE†	0.1677	0.6416	0.8343	0.4852	0.8341
MVAE	0.7992	0.8023	0.8393	0.7983	0.8641
MMVAE	0.0067	0.0048	0.0339	0.0294	0.0930
MoPoE	-0.0093	0.0123	0.0563	0.0257	0.1215
MEME	0.7746	0.7821	0.7963	0.7785	0.8692
MMVE	0.7160	0.7306	0.7538	0.7206	0.8304
miVAE	**0.8694**	**0.8809**	**0.9149**	**0.8770**	**0.9094**

Table 1: **Comparisons of neural encoding and decoding.** We report correlation on the latent variables $\mathbf{z}^{\mathbf{x}}_{\mathbf{c}}$ and $\mathbf{z}^{\mathbf{y}}_{\mathbf{c}}$ generated by the two encoders in VAE models and further coding models. * indicates the coding models are in the original data space. † indicates the VAE in Stage 1 are separately trained.

Baseline and Evaluation Metric

Baseline. We considered two types of baseline models. The first type codes in the original data space, using a core and linear readout module (Lurz et al. 2020). Following (Wang et al. 2023), we pre-trained the encoding model on data from seven mice, then fixed the core and fine-tuned the readout on data from three mice, referred to as V1-FM$_e$. The same method was used for decoding, referred to as V1-FM$_d$.

The second type codes in the latent space using a two-stage process. First, we applied self-supervised models to neural activity or visual stimuli using VAE (Kingma and Welling 2014) and iVAE (Zhou and Wei 2020). We also consider multi-modal VAEs in machine learning, including MVAE (Wu and Goodman 2018), MMVAE (Shi et al. 2019), MoPoE (Sutter, Daunhawer, and Vogt 2021), MEME (Joy et al. 2022) and MMVE (Sutter et al. 2024). Second, we performed linear or nonlinear coding in the latent space with $d = 16, T = 256$. All models share identical architectures.

Metric. Following Wang et al. (Wang et al. 2023), we employed the single-trial Pearson Correlation Coefficient (R) as our evaluation metric. For the reference V1-FM models, we measure the similarity between predicted and real neural activity for the encoding model, and between decoded and actual visual stimuli for the decoding model. In our latent space encoding-decoding framework, we first compare latent variables $\mathbf{z_c}$ from the visual stimulus encoder and the neural activity encoder for VAE models. In the second step, involving latent coding models, we fix the latent variable from one encoder as the reference and calculate the correlation of the coding model's output with this reference. We use this setting for all quantitative results and ablation studies.

Quantitative Results for Neural Coding

miVAE and neural latent modeling achieve remarkable coding performance. Table 1 shows the performance of encoding and decoding models. Though coding in the original data space cannot be directly compared to latent space coding, our approach demonstrates more effective coding than similar baselines. miVAE and subsequent latent coding surpasses all baseline models, highlighting the effectiveness of our whole method. The validation was performed on $\mathbf{z_c}$ of $\mathbf{x}$ and $\mathbf{y}$.

Methods	Stage 1	Latent Decoding (Stage 2)	
		Linear	NonLinear
VAE†	0.0822±0.0094	0.0371±0.0110	0.1490±0.0285
iVAE†	0.7449±0.0111	0.3345±0.0241	0.7855±0.0098
MVAE	0.7316±0.0091	0.7286±0.0095	0.7908±0.0027
MMVAE	0.1417±0.0242	0.1502±0.0227	0.1902±0.0654
MoPoE	0.1708±0.0312	0.1738±0.0366	0.2043±0.0366
MEME	0.6505±0.0100	0.6526±0.0088	0.7398±0.0086
MMVE	0.6598±0.0038	0.6651±0.0037	0.7594±0.0027
miVAE	**0.8984**±0.0159	**0.9091**±0.0133	**0.9635**±0.0067

Table 2: **Correlation between paired mice with the same visual stimuli.** We report the mean and standard deviation of the correlation of $\mathbf{z}^{\mathbf{x}}_{\mathbf{c}}$ across all 5 pairs of mice.

miVAE and latent modeling shows outstanding cross-individual alignment. Table 2 shows correlations between pairs of mice (5 pairs) under the same visual stimuli. Previous multi-modal VAEs achieve partial alignment. In contrast, our miVAE achieves state-of-the-art alignment performance in the first stage compared to former all results, and following decoding alignment further enhanced the performance, indicating high similarity in neural activity latent variables under the same stimuli (Safaie et al. 2023).

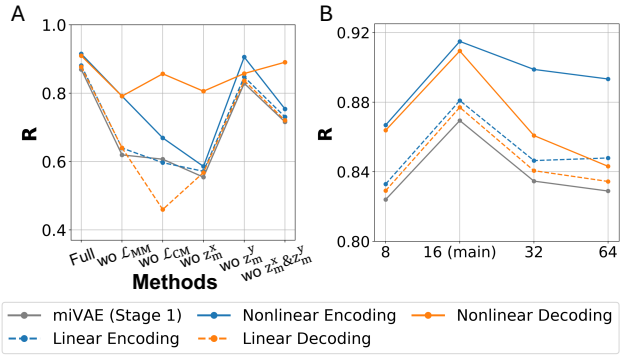

Figure 4: **Ablations on modeling methods.** (A) Ablations on losses and latent variables. (B) Ablations on Latent size. *R refers to the correlation of $\mathbf{z}^{\mathbf{x}}_{\mathbf{c}}$ and $\mathbf{z}^{\mathbf{y}}_{\mathbf{c}}$.

Ablation Studies

Ablations on modeling methods. In Figure 4.A, we show that combining $\mathcal{L}_{\text{CM}}$ with $\mathcal{L}_{\text{MM}}$ yields the best performance. Using only $\mathcal{L}_{\text{CM}}$ generally surpasses using only $\mathcal{L}_{\text{MM}}$, indicating better regularization on latent variables.

Ablation studies on latent variables demonstrate that omitting $\mathbf{z}^{\mathbf{x}}_{\mathbf{m}}$ while retaining $\mathbf{z}^{\mathbf{y}}_{\mathbf{m}}$, or excluding both, substantially impairs performance, highlighting the critical role of cross-

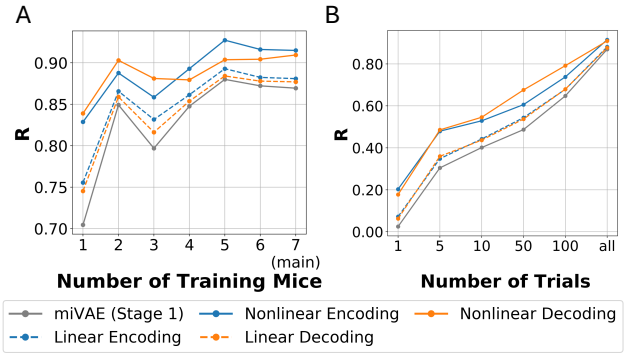

Figure 5: **Ablations studies on data scale.** (A) R over increasing training mice number. (B) R over increasing training trials. *R refers to the correlation of $\mathbf{z}_{\mathbf{c}}^{\mathbf{x}}$ and $\mathbf{z}_{\mathbf{c}}^{\mathbf{y}}$.

individual modeling. Notably, retaining $\mathbf{z}_{\mathbf{m}}^{\mathbf{x}}$ while omitting $\mathbf{z}_{\mathbf{m}}^{\mathbf{y}}$ maintains robust performance, highlighting the exceptional efficacy of $\mathbf{z}_{\mathbf{m}}^{\mathbf{x}}$ in capturing cross-individual variations. We suppose that the minimal impact of $\mathbf{z}_{\mathbf{m}}^{\mathbf{y}}$ stems from no priors for its modeling. In Figure 4.B, we demonstrate that a latent size of 16 is optimal, with the best generalization error, and used in our main method.

Ablations on data scale. We further considered the impact of the training data scale. As illustrated in Figure 5.A, increasing the number of mice in the training set correlates positively with performance. Similarly, Figure 5.B shows that increasing the number of training trials yields improved correlations. These findings indicate the beneficial effect of increased data size on cross-individual generalization.

Attribution Analysis Based on Latent Modeling

We conducted score-based attribution analysis using the high-quality latent variable $\mathbf{z}_{\mathbf{c}}$ with encoding or decoding functions f_{e},f_{d}. For the neural activity of Mouse 8 in the validation set (Figure 6.A), we attributed the encoded visual stimuli $\mathbf{z}_{\mathbf{c}}^{\mathbf{y}}$, obtaining the neuron weights shown in Figure 6.B, which range from 0 to 1 at any given time. We then identified important neurons as those with weights greater than 0.55 at all timesteps. By sorting the activities of these important neurons, we observed the significant pattern shown in Figure 6.C, indicating the temporal patterns of important neurons responsive to visual stimuli.

We further validated the classification of visual stimuli using the identified important neurons from three mice in the validation set. We use 7 classical classifiers, including Linear SVM, RBF SVM, Polynomial SVM, KNN (with n_neighbors=2), Decision Tree, Random Forest, Naïve Bayes, and MLP, and report their averaged performance. Figure 6.D shows that the neural activity of these important neuron subpopulations (91.29% with about 700 neurons) significantly outperformed that of the non-important neurons (82.92% with more than 7000 neurons) and even exceeded the results of the full population (87.24%). This confirms the significance of the identified important neurons. Detailed implementations are provided in Supplementary Materials.

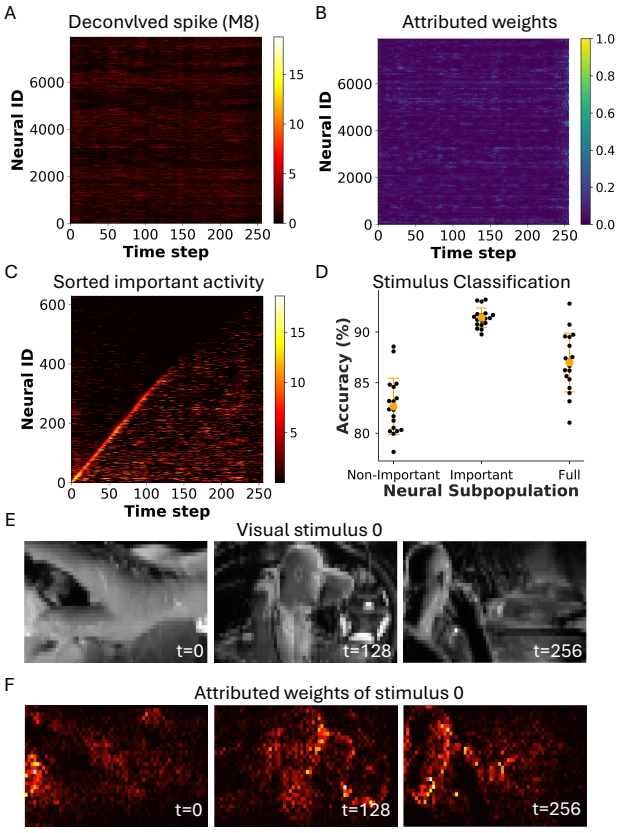

Figure 6: **Bidirectional Attribution analysis.** (A) Recorded neural activity of Mouse 8 (M8), $\mathbf{x}$. (B) Attributed weights of $\mathbf{x}$ based on $\mathbf{z}_{\mathbf{c}}^{\mathbf{y}}$. (C) Sorted important activity shows noteworthy temporal pattern. (D) Stimulus classification shows the identified key subpopulations best at distinguishing visual stimuli. (E) Visual stimulus samples, $\mathbf{y}$. (F) Attributed weights of $\mathbf{y}$ based on $\mathbf{z}_{\mathbf{c}}^{\mathbf{x}}$ shows the recorded activity has high correlation with edge and luminance.

Lastly, we performed attribution analysis of the decoded latent variable $\mathbf{z}_{\mathbf{c}}^{\mathbf{x}}$ to the visual stimuli (Figure 6.E). This analysis revealed that the neural activity exhibited a stronger correlation with edge and luminance (Figure 6.F).

Conclusions

We present a multi-modal identifiable VAE, miVAE, and latent modeling, acquiring the highly correlated variable between neural activity and visual stimulus. Our joint modeling achieves state-of-the-art performance in latent encoding-decoding and demonstrates superior cross-individual neural representation alignment. Through the proposed score-based attribution analysis, we enable fine-grained analysis by identifying key neurons and critical visual regions in V1. With its high precision and interpretability, our approach contributes to understanding V1 information processing and offers methodological insights for analyzing neural data in different sensory cortices, such as the auditory cortex.

Acknowledgements

This work was supported by the National Science and Technology Major Project (2022ZD01163013) and the Strategic Priority Research Program of Chinese Academy of Sciences (XDB1010302). We especially thank Shanghai Artificial Intelligence Laboratory for providing GPU resources.

References

Antoniades, A.; Yu, Y.; Canzano, J.; Wang, W.; and Smith, S. L. 2024. Neuroformer: Multimodal and Multitask Generative Pretraining for Brain Data. In *The Eighteenth International Conference on Learning Representations*.

Bashiri, M.; Walker, E.; Lurz, K.-K.; Jagadish, A.; Muhammad, T.; Ding, Z.; Ding, Z.; Tolias, A.; and Sinz, F. 2021. A flow-based latent state generative model of neural population responses to natural images. *Advances in Neural Information Processing Systems*, 34: 15801–15815.

Berens, P.; Ecker, A. S.; Cotton, R. J.; Ma, W. J.; Bethge, M.; and Tolias, A. S. 2012. A fast and simple population code for orientation in primate V1. *Journal of Neuroscience*, 32(31): 10618–10626.

Cobos, E.; Muhammad, T.; Fahey, P. G.; Ding, Z.; Ding, Z.; Reimer, J.; Sinz, F. H.; and Tolias, A. S. 2022. It takes neurons to understand neurons: Digital twins of visual cortex synthesize neural metamers. *bioRxiv*, 2022–12.

Ecker, A. S.; Sinz, F. H.; Froudarakis, E.; Fahey, P. G.; Cadena, S. A.; Walker, E. Y.; Cobos, E.; Reimer, J.; Tolias, A. S.; and Bethge, M. 2019. A rotation-equivariant convolutional neural network model of primary visual cortex. In *The Thirteenth International Conference on Learning Representations*.

Eichhorn, J.; Tolias, A.; Zien, A.; Kuss, M.; Weston, J.; Logothetis, N.; Schölkopf, B.; and Rasmussen, C. 2003. Prediction on spike data using kernel algorithms. *Advances in Neural Information Processing Systems*, 16.

Ellis, R. J.; and Michaelides, M. 2018. High-accuracy decoding of complex visual scenes from neuronal calcium responses. *bioRxiv*, 271296.

Fisher, R. A. 1970. Statistical methods for research workers. In *Breakthroughs in statistics: Methodology and distribution*, 66–70. Springer.

Froudarakis, E.; Berens, P.; Ecker, A. S.; Cotton, R. J.; Sinz, F. H.; Yatsenko, D.; Saggau, P.; Bethge, M.; and Tolias, A. S. 2014. Population code in mouse V1 facilitates readout of natural scenes through increased sparseness. *Nature neuroscience*, 17(6): 851–857.

Garasto, S.; Bharath, A. A.; and Schultz, S. R. 2018. Visual reconstruction from 2-photon calcium imaging suggests linear readout properties of neurons in mouse primary visual cortex. *bioRxiv*, 300392.

Gondur, R.; Sikandar, U. B.; Schaffer, E.; Aoi, M. C.; and Keeley, S. L. 2024. Multi-modal Gaussian Process Variational Autoencoders for Neural and Behavioral Data. In *The Eighteenth International Conference on Learning Representations*.

Grienberger, C.; and Konnerth, A. 2012. Imaging calcium in neurons. *Neuron*, 73(5): 862–885.

Guntupalli, J. S.; Hanke, M.; Halchenko, Y. O.; Connolly, A. C.; Ramadge, P. J.; and Haxby, J. V. 2016. A model of representational spaces in human cortex. *Cerebral cortex*, 26(6): 2919–2934.

Haxby, J. V.; Gobbini, M. I.; Furey, M. L.; Ishai, A.; Schouten, J. L.; and Pietrini, P. 2001. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539): 2425–2430.

Hubel, D. H.; and Wiesel, T. N. 1962. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1): 106.

Hubel, D. H.; and Wiesel, T. N. 1977. Ferrier lecture-Functional architecture of macaque monkey visual cortex. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 198(1130): 1–59.

Hyvarinen, A.; Sasaki, H.; and Turner, R. 2019. Nonlinear ICA using auxiliary variables and generalized contrastive learning. In *The International Conference on Artificial Intelligence and Statistics*, 859–868. PMLR.

Johnson, M. J.; Duvenaud, D. K.; Wiltschko, A.; Adams, R. P.; and Datta, S. R. 2016. Composing graphical models with neural networks for structured representations and fast inference. *Advances in Neural Information Processing Systems*, 29.

Joy, T.; Shi, Y.; Torr, P. H.; Rainforth, T.; Schmon, S. M.; and Siddharth, N. 2022. Learning multimodal VAEs through mutual supervision. In *The Sixteenth International Conference on Learning Representations*.

Keshtkaran, M. R.; Sedler, A. R.; Chowdhury, R. H.; Tandon, R.; Basrai, D.; Nguyen, S. L.; Sohn, H.; Jazayeri, M.; Miller, L. E.; and Pandarinath, C. 2022. A large-scale neural network training framework for generalized estimation of single-trial population dynamics. *Nature Methods*, 19(12): 1572–1577.

Khemakhem, I.; Kingma, D.; Monti, R.; and Hyvarinen, A. 2020. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, 2207–2217. PMLR.

Kingma, D. P.; and Welling, M. 2014. Auto-encoding variational bayes. In *The Eighth International Conference on Learning Representations*.

Le, T.; and Shlizerman, E. 2022. Stndt: Modeling neural population activity with spatiotemporal transformers. *Advances in Neural Information Processing Systems*, 35: 17926–17939.

Liu, R.; Azabou, M.; Dabagia, M.; Xiao, J.; and Dyer, E. 2022. Seeing the forest and the tree: Building representations of both individual and collective dynamics with transformers. *Advances in Neural Information Processing Systems*, 35: 2377–2391.

Lurz, K.-K.; Bashiri, M.; Willeke, K.; Jagadish, A. K.; Wang, E.; Walker, E. Y.; Cadena, S. A.; Muhammad, T.; Cobos, E.; Tolias, A. S.; et al. 2020. Generalization in data-driven models of primary visual cortex. *bioRxiv*, 2020–10.

Ma, G.; Jiang, R.; Yan, R.; and Tang, H. 2024. Temporal Conditioning Spiking Latent Variable Models of the Neural

Response to Natural Visual Scenes. *Advances in Neural Information Processing Systems*, 36.

Niell, C. M.; and Stryker, M. P. 2008. Highly selective receptive fields in mouse visual cortex. *Journal of Neuroscience*, 28(30): 7520–7536.

Olshausen, B. A.; and Field, D. J. 2004. Sparse coding of sensory inputs. *Current opinion in neurobiology*, 14(4): 481–487.

Palumbo, E.; Daunhawer, I.; and Vogt, J. E. 2023. MMVAE+: Enhancing the generative quality of multimodal VAEs without compromises. In *The Eleventh International Conference on Learning Representations*.

Pandarinath, C.; O'Shea, D. J.; Collins, J.; Jozefowicz, R.; Stavisky, S. D.; Kao, J. C.; Trautmann, E. M.; Kaufman, M. T.; Ryu, S. I.; Hochberg, L. R.; et al. 2018. Inferring single-trial neural population dynamics using sequential auto-encoders. *Nature methods*, 15(10): 805–815.

Roe, A. W.; Chelazzi, L.; Connor, C. E.; Conway, B. R.; Fujita, I.; Gallant, J. L.; Lu, H.; and Vanduffel, W. 2012. Toward a unified theory of visual area V4. *Neuron*, 74(1): 12–29.

Safaie, M.; Chang, J. C.; Park, J.; Miller, L. E.; Dudman, J. T.; Perich, M. G.; and Gallego, J. A. 2023. Preserved neural dynamics across animals performing similar behaviour. *Nature*, 623(7988): 765–771.

Schneider, S.; Lee, J. H.; and Mathis, M. W. 2023. Learnable latent embeddings for joint behavioural and neural analysis. *Nature*, 617(7960): 360–368.

Shi, Y.; Paige, B.; Torr, P.; et al. 2019. Variational mixture-of-experts autoencoders for multi-modal deep generative models. *Advances in Neural Information Processing Systems*, 32.

Sinz, F.; Ecker, A. S.; Fahey, P.; Walker, E.; Cobos, E.; Froudarakis, E.; Yatsenko, D.; Pitkow, Z.; Reimer, J.; and Tolias, A. 2018. Stimulus domain transfer in recurrent models for large scale cortical population prediction on video. *Advances in Neural Information Processing Systems*, 31.

Sofroniew, N. J.; Flickinger, D.; King, J.; and Svoboda, K. 2016. A large field of view two-photon mesoscope with subcellular resolution for in vivo imaging. *Elife*, 5: e14472.

Sriperumbudur, B.; Fukumizu, K.; Gretton, A.; Hyv, A.; Kumar, R.; et al. 2017. Density estimation in infinite dimensional exponential families. *Journal of Machine Learning Research*, 18(57): 1–59.

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

Sutter, T. M.; Daunhawer, I.; and Vogt, J. E. 2021. Generalized multimodal ELBO. In *The Fifteenth International Conference on Learning Representations*.

Sutter, T. M.; Meng, Y.; Fortin, N.; Vogt, J. E.; and Mandt, S. 2024. Unity by Diversity: Improved Representation Learning in Multimodal VAEs. *arXiv preprint arXiv:2403.05300*.

Tootell, R. B.; Silverman, M. S.; Switkes, E.; and De Valois, R. L. 1982. Deoxyglucose analysis of retinotopic organization in primate striate cortex. *Science*, 218(4575): 902–904.

Turishcheva, P.; Fahey, P. G.; Hansel, L.; Froebe, R.; Ponder, K.; Vystrčilová, M.; Willeke, K. F.; Bashiri, M.; Wang, E.; Ding, Z.; et al. 2023. The Dynamic Sensorium competition for predicting large-scale mouse visual cortex activity from videos. *ArXiv*.

Wang, E. Y.; Fahey, P. G.; Ponder, K.; Ding, Z.; Chang, A.; Muhammad, T.; Patel, S.; Ding, Z.; Tran, D.; Fu, J.; et al. 2023. Towards a foundation model of the mouse visual cortex. *bioRxiv*.

Wu, M.; and Goodman, N. 2018. Multimodal generative models for scalable weakly-supervised learning. *Advances in Neural Information Processing Systems*, 31.

Ye, J.; Collinger, J.; Wehbe, L.; and Gaunt, R. 2024. Neural data transformer 2: multi-context pretraining for neural spiking activity. *Advances in Neural Information Processing Systems*, 36.

Ye, J.; and Pandarinath, C. 2021. Representation learning for neural population activity with Neural Data Transformers. *arXiv preprint arXiv:2108.01210*.

Yoshida, T.; and Ohki, K. 2020. Natural images are reliably represented by sparse and variable populations of neurons in visual cortex. *Nature communications*, 11(1): 872.

Yu, B. M.; Cunningham, J. P.; Santhanam, G.; Ryu, S.; Shenoy, K. V.; and Sahani, M. 2008. Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. *Advances in Neural Information Processing Systems*, 21.

Zhao, Y.; and Park, I. M. 2017. Variational latent gaussian process for recovering single-trial dynamics from population spike trains. *Neural computation*, 29(5): 1293–1316.

Zhou, D.; and Wei, X.-X. 2020. Learning identifiable and interpretable latent models of high-dimensional neural activity using pi-VAE. *Advances in Neural Information Processing Systems*, 33: 7234–7247.

Zhu, F.; Grier, H. A.; Tandon, R.; Cai, C.; Agarwal, A.; Giovannucci, A.; Kaufman, M. T.; and Pandarinath, C. 2022. A deep learning framework for inference of single-trial neural population dynamics from calcium imaging with subframe temporal resolution. *Nature neuroscience*, 25(12): 1724–1734.