# AVM: Towards Structure-Preserving Neural Response Modeling in the Visual Cortex Across Stimuli and Individuals

**Qi Xu**[1], **Shuai Gong**[1], **Xuming Ran**[2*], **Haihua Luo**[1,3], **Yangfan Hu**[4]

[1]School of Computer Science and Technology, Dalian University of Technology,
[2]National University of Singapore,
[3]Faculty of Information Technology, University of Jyväskylä,
[4]School of Information Technology and Artificial Intelligence, Zhejiang University of Finance and Economics,
ranxuming@gmail.com

## Abstract

While deep learning models have shown strong performance in simulating neural responses, they often fail to clearly separate stable visual encoding from condition-specific adaptation, which limits their ability to generalize across stimuli and individuals. We introduce the **A**daptive **V**isual **M**odel (**AVM**), a structure-preserving framework that enables condition-aware adaptation through modular subnetworks, without modifying the core representation. AVM keeps a Vision Transformer-based encoder frozen to capture consistent visual features, while independently trained modulation paths account for neural response variations driven by stimulus content and subject identity. We evaluate AVM in three experimental settings, including stimulus-level variation, cross-subject generalization, and cross-dataset adaptation, all of which involve structured changes in inputs and individuals. Across two large-scale mouse V1 datasets, AVM outperforms the state-of-the-art V1T model by approximately 2% in predictive correlation, demonstrating robust generalization, interpretable condition-wise modulation, and high architectural efficiency. Specifically, AVM achieves a 9.1% improvement in explained variance (FEVE) under the cross-dataset adaptation setting. These results suggest that AVM provides a unified framework for adaptive neural modeling across biological and experimental conditions, offering a scalable solution under structural constraints. Its design may inform future approaches to cortical modeling in both neuroscience and biologically inspired AI systems.

## Introduction

Understanding the computational mechanisms of neurons in the visual system—particularly how they respond to natural image stimuli—remains a central challenge in sensory neuroscience (Carandini et al. 2005; Yamins et al. 2013; McIntosh et al. 2016; Yang and Wang 2020; Zhang et al. 2025a). Modeling the response patterns of the primary visual cortex (V1) has become an effective strategy, bridging biological vision with machine perception (Klindt et al. 2017; Sinz et al. 2019; Lurz et al. 2020; Li et al. 2023).

Recent advances in deep learning have significantly improved our ability to predict V1 responses. Traditional methods, including generalized linear models and shallow multi-layer networks, have been surpassed by CNNs, which extract nonlinear stimulus features with greater fidelity (Yamins et al. 2014; Cadena et al. 2019). More recently, Vision Transformer (ViT) models have shown superior representational power, with architectures like V1T (Li et al. 2023) setting new benchmarks for mouse V1 modeling by coupling a ViT backbone with behavioral feature modules.

However, these models tend to conflate stable representation learning with response adaptation, making them brittle under condition changes such as stimulus shifts, individual variability, or environmental perturbations. Once trained, such monolithic models require full retraining to adjust to new conditions, limiting their generalization and interpretability. In contrast, biological visual systems exhibit a compelling duality: they preserve a stable structural organization while flexibly modulating responses based on context, internal state, or inter-subject variability (Franke et al. 2022; Cheng et al. 2022). This motivates a critical modeling challenge: **how to reconcile structural stability with functional flexibility** in neural response prediction.

To address this, we propose the Adaptive Visual Model (AVM)—a framework grounded in the principle of **structure-function decoupling**. Unlike prior approaches that entangle representation and modulation, AVM introduces an explicit architectural separation: a shared, frozen encoder captures invariant structural representations, while lightweight, condition-aware modules enable flexible modulation of neural responses. This design not only reflects biological organization but also supports **scalable and efficient adaptation** to diverse conditions—input shifts, individual differences, and environmental changes—without altering the core encoding structure. We evaluate AVM on two large-scale mouse V1 datasets and test its generalization ability across three major challenges in cortical modeling: stimulus changes, inter-individual differences, and cross-dataset shifts. In all scenarios, AVM consistently delivers accurate and interpretable neural predictions, achieving high performance through localized adaptation while maintaining a stable representational backbone. These results suggest that AVM provides a robust and biologically grounded solution for condition-aware neural response modeling.Our contribu-

*Corresponding author:Xuming Ran

tions are threefold:

- We propose AVM, a condition-aware cortical modeling framework grounded in structure-function decoupling, which separates stable representation encoding from dynamic response modulation;
- AVM introduces modular control subnetworks that enable localized response adaptation under frozen representational backbones, satisfying biological constraints of structural stability and contextual flexibility;
- AVM demonstrates scalable and interpretable generalization across input domains, individual anatomy, and environmental conditions, without requiring full model retraining.

## Related Work

The modeling of neural responses in the visual cortex dates back to Hubel and Wiesel's discovery of simple and complex cells (Hubel and Wiesel 1962; Carandini, Heeger, and Anthony Movshon 1999; Carandini et al. 2005; Batty et al. 2017; Ponce et al. 2019; Billeh et al. 2020; Bashivan, Kar, and DiCarlo 2019; Bao et al. 2020; Zhang et al. 2021). Early models were linear, describing neuronal tuning to visual features like orientation and contrast (Jones and Palmer 1987; Olshausen and Field 1996), later extended by nonlinear models such as the energy model (Adelson and Bergen 1985), LN model (Jones and Palmer 1987), and LN-LN cascades (Heeger 1992).

Traditional models, while useful, struggle to generalize to complex stimuli. With increased computational power and large datasets, deep learning approaches, especially CNNs and ViTs, have become dominant for modeling visual cortex responses (Margalit et al. 2023; Klindt et al. 2017; Lurz et al. 2020; Cotton, Sinz, and Tolias 2020; Ran et al. 2021; Franke et al. 2022; Li et al. 2023; Du et al. 2024; Deng, Schwendeman, and Guan 2024; Zhang et al. 2025c). These models typically follow two paradigms: task-driven models use pretrained networks for object recognition and readout modules to predict neural responses (Yamins et al. 2014; Cadieu et al. 2014; Cadena et al. 2019), while data-driven models train directly from neural recordings without supervision, learning shared representations across animals (Klindt et al. 2017; Lurz et al. 2020; Cadena et al. 2019; Franke et al. 2021, 2022; Li et al. 2023; Zhang et al. 2025b).

Additionally, Franke et al. (Franke et al. 2022) integrated behavioral information with visual stimuli, demonstrating improved neural response prediction by combining behavioral data with visual inputs to capture dynamic neural fluctuations.

## Method

### Overview: Structure-Function Decoupling in AVM

We propose AVM (Adaptive Visual Modeling), a structure-function decoupled framework for cortical response modeling. It comprises a frozen Transformer-based visual encoder and a set of condition-aware modulation modules inserted in parallel, forming a dual-path structure:

- **Shared Representation Path:** encodes invariant visual features common across subjects, stimuli, and domains;
- **Condition-Specific Modulation Path:** applies flexible, lightweight transformations that adjust neural responses under different conditions without altering the core encoder.

This design supports interpretable, scalable adaptation while preserving representation consistency.

### Core Encoder: Stable Visual Representation

As shown in Fig. 1A, the main network is a 4-layer ViT backbone (from V1T(Li et al. 2023)), frozen throughout training. The model receives visual input $\mathbf{x}$ and behavior variables $\mathbf{b}$. The behavior signal is first embedded via a B-MLP and added to the input before processing with attention and MLP layers.

Each block contains two standard Transformer components: multi-head attention (MHA) and MLP. Their outputs form the base visual representation $\mathbf{f}$ for each location, computed as:

$$b \leftarrow b_{prev} + \mathrm{MLP}_{\text{behavior}}(b), \tag{1}$$

$$a_i \leftarrow \mathrm{MHA}(x_i + b) + x_i, \tag{2}$$

$$f_i \leftarrow \mathrm{MLP}(a_i) + a_i, \tag{3}$$

where $\mathbf{x}_i$ is the patch embedding of the image input, $\mathbf{b}$ is the behavior vector, and $\mathbf{f}_i$ is the resulting visual representation at each layer. This frozen backbone ensures a stable and interpretable representational structure across all adaptation settings.

### Condition-Aware Modulation Unit (CAMU)

To achieve condition-specific flexibility, AVM introduces **Condition-aware Modulation Units** (Fig. 1B) into each Transformer block. These lightweight modules reshape activations via bottleneck-style feedforward layers:

$$\mathrm{CAMU}(x) = x + \mathrm{Up}(\mathrm{ReLU}(\mathrm{Down}(x))) \tag{4}$$

Each ViT block integrates three modulation modules placed at key positions: after the attention output, the MLP output, and the block output. These modules enable localized context-sensitive adjustments under a shared structure. The revised computation becomes:

$$b \leftarrow b_{prev} + \mathrm{MLP}_{\text{behavior}}(b), \tag{5}$$

$$a_i \leftarrow \mathrm{MHA}(x_i + b) + x_i + \mathrm{CAMU}_1(x_i), \tag{6}$$

$$f_i \leftarrow \mathrm{MLP}(a_i) + a_i + \mathrm{CAMU}_2(a_i), \tag{7}$$

$$f \leftarrow f_i + \mathrm{CAMU}_3(x_i), \tag{8}$$

These independently trainable controllers modulate neural responses without disrupting the stable encoder stream, enabling modular, interpretable adjustments.

### Readout: Neuron-Wise Prediction

We employ a Gaussian readout module as proposed by Lurz et al.(Lurz et al. 2020). Each neuron is represented by a learned 2D Gaussian position $(\mu, \sigma)$, which defines a

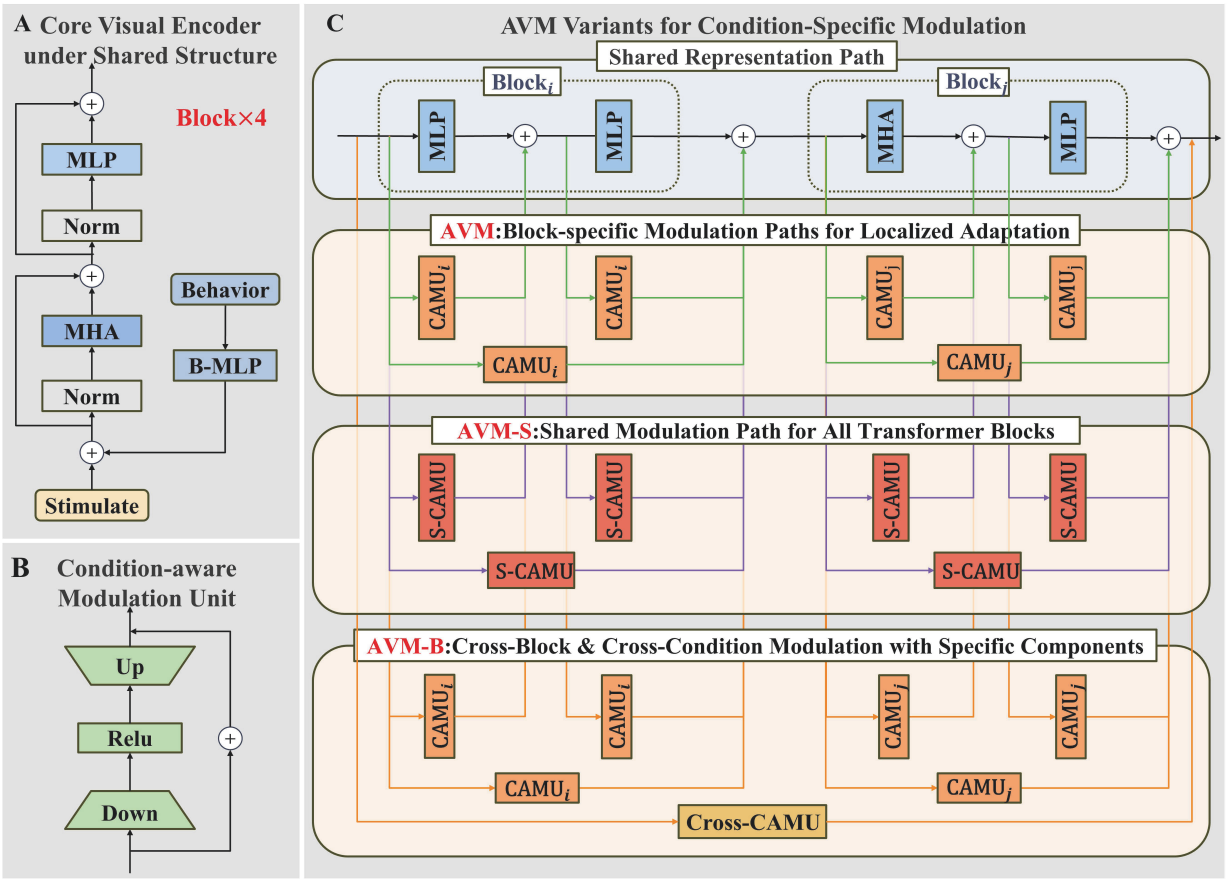



Figure 1: AVM model architecture and condition-specific modulation variants. (A) The main network encodes stable visual representations under a consistent architecture. (B) Condition-aware Modulation Unit(CAMU): A lightweight feedforward module with a bottleneck structure, serving as the basic modulation component across AVM variants. (C) Condition-Specific Modulation Variants: AVM Employs block-specific modulation paths for localized response adaptation. AVM-S Shares a single modulation path across all Transformer blocks, enabling parameter-efficient tuning. AVM-B Introduces Cross-CAMU to support cross-block and cross-condition transfer, modeling higher-level adaptation interactions.

location-sensitive sampling from the visual feature map $\mathbf{f}$. A linear layer projects sampled features to neural response space. An ELU activation with offset ensures positivity:

$$\hat{y}_n = \text{ELU}\left(\mathbf{w}_n^\top \cdot f\left(\mu_n\right)\right) + 1 \tag{9}$$

## Modulation Variants: Adaptation Strategies

As shown in Fig. 1C, AVM supports three structural variants to flexibly support various adaptation needs:

- **AVM**: Each block contains a unique modulation subnetwork, enabling fine-grained, layer-wise response control for highly heterogeneous input or subject conditions.

$$h_i = \text{Block}_i(h_{i-1}) + \text{ModPath}_i(h_{i-1}) \tag{10}$$

- **AVM-S**: All blocks share the same modulation subnetwork, enforcing a consistent adaptation rule across layers. This design favors parameter efficiency and coherence in response shifts.

$$h_i = \text{Block}_i(h_{i-1}) + \text{SharedModPath}(h_{i-1}) \tag{11}$$

- **AVM-B**: Beyond intra-layer adaptation, this variant adds inter-layer modulation to propagate context across hierarchical stages. Suitable for multi-level adaptation tasks such as cross-domain generalization.

$$h_i = \text{Block}_i(h_{i-1}) + \text{ModPath}_i(h_{i-1}) \tag{12}$$
$$h_{ij} = \text{Block}_j(h_i) + \text{ModPath}_j(h_i) \tag{13}$$
$$h_j = h_{ij} + \text{CrossTaskModPath}(h_{i-1}) \tag{14}$$

These variants enable AVM to flexibly adjust its adaptation strategy under different degrees of condition complexity while maintaining a stable representational backbone.

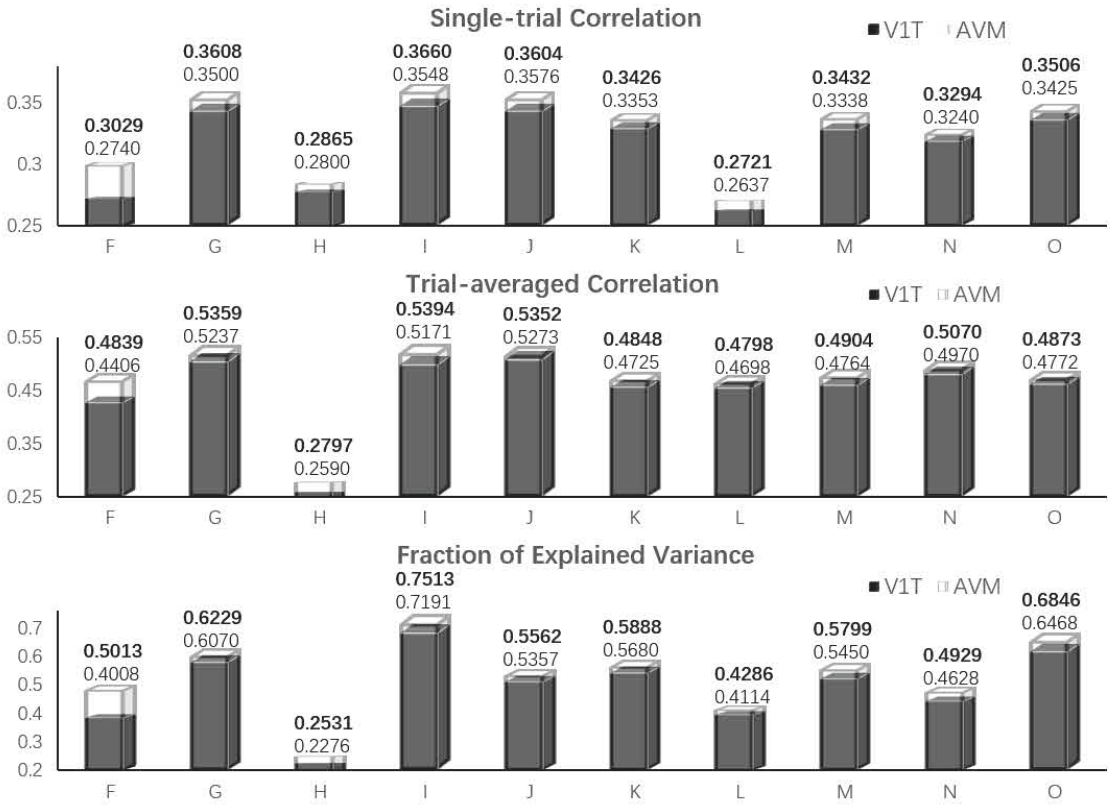



Figure 2: AVM consistently improves individual-level neural prediction. Evaluation results on Dataset-F for each individual mouse (F–O), comparing AVM and baseline V1T. Three metrics are reported: single-trial correlation (*top*), trial-averaged correlation (*middle*), and fraction of explained variance (FEVE, *bottom*). AVM achieves consistent gains across all individuals.

## Experimental Settings

### Datasets

This study uses two datasets for validation: the Sensorium dataset (Willeke et al. 2022) and the dataset of mouse primary visual cortex responses (Franke et al. 2022).

The Sensorium dataset (denoted as Dataset S) records neural activity from over 7,000 neurons in the V1 region of five mice (labeled A–E) using two-photon calcium imaging. Each mouse was presented with grayscale natural images from ImageNet ($x_{\text{image}} \in \mathbb{R}^{1\times h\times w}$), totaling 25,100 unique images. Among these, 5,000 images were used for training, while 100 images were each repeated 10 times to construct the test set. For each mouse, the neural responses of $m_i$ neurons were recorded across $n$ repetitions, resulting in $m_i \times n$ total neural response samples. In addition to neuronal data, Dataset S includes anatomical coordinates for each neuron and four behavioral variables: pupil diameter, its temporal derivative, the 2D pupil center coordinates, and locomotion speed.

The Franke dataset (denoted as Dataset F) captures V1 responses from ten mice (labeled F–O) to both grayscale and color images from ImageNet. The training set consists of 4,500 unique color images ($x_{\text{image}} \in \mathbb{R}^{2\times h\times w}$) and 750 grayscale images ($x_{\text{image}} \in \mathbb{R}^{1\times h\times w}$), while the test set comprises 100 color and 50 grayscale images. Neural recordings were obtained from 1,000 neurons across the ten mice. For consistency with our model's input requirements, all color images from Dataset F were converted to grayscale during preprocessing.

### Training Protocol

The AVM training process follows a two-phase strategy that reflects the model's core design philosophy: separating stable encoding and adaptive modulation.

**Phase 1 (Backbone Pretraining)**: The main network (ViT-based backbone) and readout module are jointly trained on a designated pretraining dataset (e.g., Dataset S) using all available parameters. This phase models stable visual feature encoding shared across conditions.

**Phase 2 (Subnetwork Adaptation)**: The backbone is frozen, and a condition-specific sub-network (adapter modules) is trained on a new condition (e.g., different dataset or subject). This reflects the AVM design for localized response modulation without altering the shared representation.

We use the AdamW optimizer and the Poisson loss function:

$$\mathcal{L}_m^{\text{Poisson}}(r, o) = \sum_{t=1}^{n_t} \sum_{i=1}^{n_m} \left( o_{i,t} - r_{i,t} \log o_{i,t} \right), \tag{15}$$

where $r_{i,t}$ and $o_{i,t}$ are the true and predicted responses for neuron $i$ in trial $t$.

All models are trained with a batch size of 16, an initial learning rate of 0.0016, and up to 400 epochs. We apply early stopping and learning rate decay (factor 0.3) when validation loss plateaus for 10 epochs. The same training schedule is used in both phases, except that in Phase 2 only the sub-network parameters are updated.

## Evaluation Metrics

The predictive performance of our model is evaluated using three metrics: single-trial correlation, average trial correlation, and the model's Fraction of Explained Variance (FEVE) (Willeke et al. 2022).

Single-trial correlation considers the inter-experiment variability of the same visual stimulus presented multiple times in the test set, providing a detailed measure of the model's prediction accuracy:

$$\rho_{trial}(r, o) = \frac{\sum_{i,j} \left( r_{i,j} - \bar{r} \right) \left( o_{i,j} - \bar{o} \right)}{\sqrt{\sum_{i,j} \left( r_{i,j} - \bar{r} \right)^2 \sum_{i,j} \left( o_{i,j} - \bar{o} \right)^2}}, \tag{16}$$

where $r_{i,j}$ represents the true response of a single neuron to a single stimulus presentation, $\bar{r}$ is the average response across all repetitions of the image, $o_{i,j}$ is the predicted response of the neuron to the same stimulus, and $\bar{o}$ is the average predicted response across all repetitions of the image.

Average trial correlation is calculated by considering the neuron responses $r_{ij}$ for image $i$ and repetition $j$ , and their predicted values $o_i$. The correlation between the predicted responses and the average neuronal response $r_i$ for image $i$ is computed as follows:

$$\rho_{avg}(r, o) = \frac{\sum_{i,j} \left( \bar{r_i} - \bar{r} \right) \left( o_i - \bar{o} \right)}{\sqrt{\sum_{i,j} \left( \bar{r_i} - \bar{r} \right)^2 \sum_{i,j} \left( o_i - \bar{o} \right)^2}}, \tag{17}$$

where $\bar{r_i}$ denotes the average response of $r_{i,j}$ over the $j$ repeated experiments.

FEVE (Fraction of Explained Variance) measures the proportion of variance in the neural responses explained by the model relative to the total variance, assessing the model's ability to explain neural activity while excluding the influence of random noise or unexplained variability:

$$\text{FEVE } = 1 - \frac{\frac{1}{N} \sum_{i,j} \left( r_{i,j} - o_i \right)^2 - \sigma_\varepsilon^2}{\text{Var}[\mathbf{r}] - \sigma_\varepsilon^2}. \tag{18}$$

where $\text{Var}[\mathbf{r}]$ is the total response variance computed across all $N$ trials and $\sigma_\epsilon^2 = E_i[Var_j[r|x]]$ is the observation noise variance computed as the average variance across responses to repeated presentations of the stimulus $x$.

# Results

## Condition-Driven Local Response Adjustment

**AVM can capture the response shift caused by different input changes in the same individual.** To evaluate AVM's capability to capture condition-specific shifts in neural responses under input variation, we conducted experiments simulating visual state transitions within individual animals. Instead of retraining the entire system, AVM introduces condition-sensitive modulation paths embedded within each representational block. These paths locally reshape output activations in response to new stimuli while preserving a stable structural stream, supporting structure-function decoupling in cortical modeling.

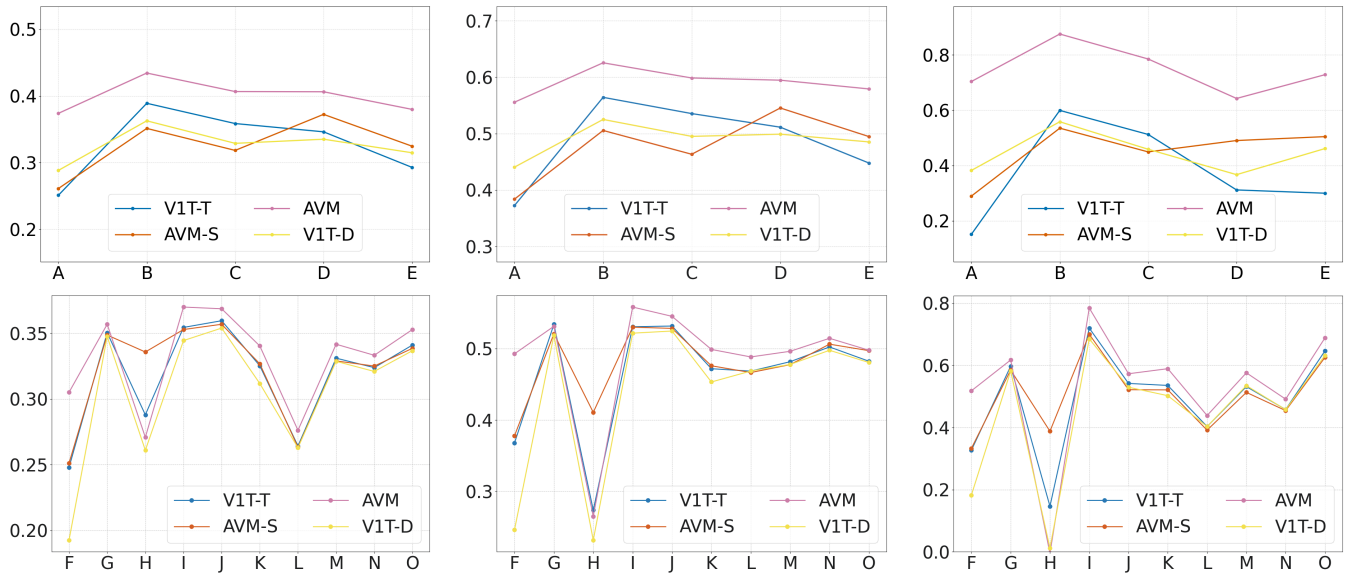
As shown in Fig 2, AVM consistently outperformed the baseline V1T across all mice in Dataset-F, achieving higher single-trial correlation ($\rho_{\text{trial}}$), trial-averaged correlation ($\rho_{\text{avg}}$), and fraction of explained variance (FEVE). These improvements demonstrate the model's ability to fine-tune response predictions without altering the core representation pathway. Furthermore, when deploying a shared representation across mice in Dataset-F and Dataset-S (Fig 3), AVM maintained robust generalization performance and surpassed all comparative variants, including the pre-trained fine-tuning model (V1T-T), non-modulated baseline (V1T-D), and shared modulation version (AVM-S). Notably, AVM achieved a >30% improvement in FEVE on Dataset-S, indicating that localized, condition-specific paths enable more expressive regulation of stimulus-driven variability.To ensure the stability of these results, we ran experiments using five different random seeds, and the standard deviations of the key metrics were consistently less than 0.001, demonstrating the robustness of AVM's performance.

To assess architectural efficiency, we further compared the number of trainable parameters required by AVM variants and the V1T baseline (Fig 4). While V1T involves over 2.46M trainable parameters, AVM and its variants drastically reduce this cost to as low as 0.03M for AVM-S and 0.11M for AVM, with minimal compromise in predictive accuracy.

## Adaptive Modeling Under Individual Variation

**AVM enables condition-aware generalization across individuals with minimal structural adjustment.** To assess how well AVM models subject-specific response patterns while maintaining representational consistency, we design an individual generalization task in which neural responses from held-out subjects must be predicted based on knowledge learned from others. Specifically, for each subject in the dataset, we train the model using data from all remaining individuals and adapt it to the target subject using a small amount of subject-specific tuning.

This setting reflects a biologically plausible scenario where cortical encoding remains largely invariant, while individual-level variability is captured through lightweight, context-dependent adaptation. The evaluation is conducted on both Dataset-S and Dataset-F, and compared against three baselines: V1T-D (training from scratch), V1T-T (pretrain-




(a) Single-Trial Correlation (b) Average Correlation (c) FEVE

Figure 3: The tuning ability of the AVM model under different input conditions. The top three figures show the results for dataset S, and the bottom three figures show the results for dataset F. From left to right, these figures show the single-trial correlation, average correlation, and explained variance, respectively. Each figure includes four structures: V1T-D, V1T-T, AVM-S, and AVM. The x-axis represents each mouse, and the y-axis represents the predicted value.

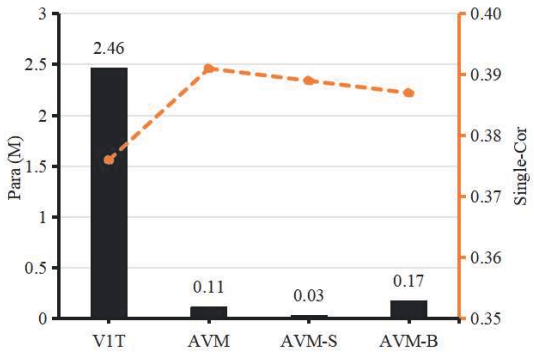



Figure 4: The number of trainable parameters. Comparison of the number of trainable parameters of our proposed AVM core and V1T core.

ing and full-model fine-tuning), and AVM-S (shared adaptation across subjects without conditioning).

As shown in Table 1 and Table 2, AVM consistently outperforms all baselines across all three metrics—single-trial correlation, average correlation, and explained variance (FEVE). The gains are particularly pronounced on Dataset-F, with average improvements ranging from 1.5% to 3.0%. Moreover, AVM achieves this performance with minimal subject-specific parameterization, highlighting the efficiency and scalability of condition-aware response modulation. The comparison with AVM-S further demonstrates the importance of individual conditioning: adaptive tuning that accounts for subject identity outperforms shared adjustment strategies, underscoring the utility of contextual modulation in capturing biological variability.

## Adaptive Modeling Under Environmental Shift

**AVM enables robust neural response prediction across environmental distribution shifts.** To assess its generalization ability, we conduct a cross-dataset adaptation experiment. The model is trained on Dataset-S to obtain stable visual representations and then adapted to Dataset-F using lightweight contextual modulation. This simulates adapting to a new stimulus distribution while preserving core representational consistency. We compare AVM against several baselines, including a linear-nonlinear (LN) model, a convolutional predictor (Lurz), a method from Deng et al., and two Transformer architectures—V1T-D (trained from scratch) and V1T-T (fully fine-tuned). We also evaluate three AVM structural variants to assess different modulation mechanisms.

Notably, despite being trained solely on Dataset-S, AVM achieves strong performance on Dataset-F after lightweight adaptation—outperforming models trained directly on Dataset-F, such as Lurz and Deng. As shown in Table 3, AVM achieves 0.3906 in single-trial correlation, 0.6114 in average correlation, and 0.7536 in FEVE, with a 9.1% improvement in explained variance over the V1T-T baseline. AVM improves single-trial correlation by over 8% compared to Lurz, and FEVE by more than 19 percentage points compared to Deng. These results demonstrate that AVM achieves high generalization performance in new environments through structure-preserving learning and condition-aware modulation. Ablation studies (see ap-

| Core | F | G | H | I | J | K | L | M | N | O |
|---|---|---|---|---|---|---|---|---|---|---|
| $\rho_{trial}$ | | | | | | | | | | |
| V1T-D | 0.3153 | 0.3815 | 0.2743 | 0.3859 | 0.3856 | 0.3713 | 0.2895 | 0.3688 | 0.3510 | 0.3682 |
| V1T-T | 0.3189 | 0.3790 | 0.3628 | 0.3817 | 0.3895 | 0.3481 | 0.2859 | 0.3629 | 0.3501 | 0.3637 |
| **AVM-S** | **0.3252** | **0.3862** | **0.3560** | **0.3883** | **0.3917** | **0.3666** | **0.2936** | **0.3703** | **0.3512** | **0.3738** |
| **AVM** | **0.3264** | **0.3860** | **0.3673** | **0.3898** | **0.3918** | **0.3694** | **0.2951** | **0.3691** | **0.3527** | **0.3748** |
| $\rho_{avg}$ | | | | | | | | | | |
| V1T-D | 0.5141 | 0.5605 | 0.2303 | 0.5650 | 0.5718 | 0.5251 | 0.5120 | 0.5254 | 0.5398 | 0.5120 |
| V1T-T | 0.5307 | 0.5692 | 0.4752 | 0.5828 | 0.5828 | 0.5245 | 0.5237 | 0.5356 | 0.5453 | 0.5287 |
| **AVM-S** | **0.5322** | **0.5795** | **0.4600** | **0.5803** | **0.5855** | **0.5361** | **0.5232** | **0.5398** | **0.5487** | **0.5306** |
| **AVM** | **0.5341** | **0.5794** | **0.4711** | **0.5841** | **0.5854** | **0.5345** | **0.5256** | **0.5382** | **0.5526** | **0.5317** |
| **FEVE** | | | | | | | | | | |
| V1T-D | 0.5305 | 0.6685 | -0.0037 | 0.8202 | 0.6198 | 0.6827 | 0.4829 | 0.6648 | 0.5423 | 0.7413 |
| V1T-T | 0.5615 | 0.6851 | 0.4798 | 0.8372 | 0.6371 | 0.6199 | 0.4821 | 0.6449 | 0.5461 | 0.7305 |
| **AVM-S** | **0.5766** | **0.6992** | **0.4635** | **0.8464** | **0.6419** | **0.6696** | **0.4943** | **0.6736** | **0.5444** | **0.7627** |
| **AVM** | **0.5837** | **0.7201** | **0.4661** | **0.8608** | **0.6425** | **0.6811** | **0.4999** | **0.6649** | **0.5476** | **0.7692** |

Table 1: Experimental results of regulatory ability across individuals in dataset F. Four models are compared in the table: V1T-D, V1T-T, AVM-S, and AVM. The three rows in the table correspond to three indicators, respectively.

| Core | A | B | C | D | E |
|---|---|---|---|---|---|
| $\rho_{trial}$ | | | | | |
| V1T-D | 0.3607 | 0.4176 | 0.3947 | 0.4132 | 0.3382 |
| V1T-T | 0.3787 | 0.4522 | 0.4124 | 0.4145 | 0.3833 |
| **AVM-S** | **0.3849** | **0.4575** | **0.4157** | **0.4258** | **0.3897** |
| **AVM** | **0.3855** | **0.4588** | **0.4199** | **0.4264** | **0.3916** |
| $\rho_{avg}$ | | | | | |
| V1T-D | 0.5343 | 0.6072 | 0.5812 | 0.6061 | 0.5091 |
| V1T-T | 0.5669 | 0.6519 | 0.6181 | 0.6117 | 0.5896 |
| **AVM-S** | **0.5744** | **0.6574** | **0.5185** | **0.6230** | **0.5969** |
| **AVM** | **0.5751** | **0.6585** | **0.6233** | **0.6230** | **0.5972** |
| **FEVE** | | | | | |
| V1T-D | 0.6467 | 0.7927 | 0.7223 | 0.6366 | 0.5551 |
| V1T-T | 0.7269 | 0.9483 | 0.8157 | 0.6551 | 0.7520 |
| **AVM-S** | **0.7553** | **0.9711** | **0.8271** | **0.6936** | **0.7699** |
| **AVM** | **0.7500** | **0.9755** | **0.8424** | **0.6916** | **0.7752** |

Table 2: Experimental results of regulatory ability across individuals in dataset S.

| Core | $\rho_{trial}$ | $\rho_{avg}$ | **FEVE** |
|---|---|---|---|
| DataSet F | | | |
| LN | 0.2230 | — | — |
| Lurz | 0.3090 | — | — |
| Deng | — | 0.600 | 0.558 |
| V1T-D | 0.3607 | 0.5388 | 0.6363 |
| V1T-T | 0.3761 | 0.5954 | 0.6662 |
| **AVM-S** | **0.3893** | **0.6104** | **0.7515** |
| **AVM-B** | **0.3873** | **0.6076** | **0.7467** |
| **AVM** | **0.3906** | **0.6114** | **0.7536** |

Table 3: Prediction performance under different environmental distribution conditions. This experiment compares four baseline models, namely LN linear model, CNN convolutional model, V1T-D and V1T-T. The experiments are conducted on dataset F. AVM, AVM-S and AVM-B represent three different frameworks proposed in this paper.

pendix) further show that tuning modulation strength and bottleneck dimensions is crucial for robust adaptation. The AVM-B variant, however, suffers from mild oversharing and reduced layer specialization, leading to performance saturation and a decline in overall effectiveness.

## Conclusion

This work introduces the Adaptive Visual Model (AVM), a structure-function decoupled framework for neural response modeling under variable biological and environmental conditions. By separating stable sensory encoding from flexible condition-driven modulation, AVM enables generalizable cortical response prediction without mixing representational learning and context adaptation. We evaluate AVM across three generalization settings: stimulus variation, cross-individual transfer, and domain adaptation between environments using two large-scale mouse V1 datasets. In all scenarios, AVM outperforms strong baselines while maintaining efficiency and interpretability. These results show that condition-aware modulation atop a shared scaffold enables robust adaptation to input and subject variability without compromising structural stability. This study highlights the importance of modeling functional modulation as an explicit computational goal in neuroscience-inspired architectures. Future work will extend this to higher cortical regions and explore real-time or closed-loop applications, advancing biologically grounded neural prediction.

## Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant (No. 62476035, 62206037, and U24B20140), and the Young Elite Scientists Sponsorship Program by CAST under Grant 2024QNRC001.

## References

Adelson, E. H.; and Bergen, J. R. 1985. Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America A*, 2(2): 284–299.

Bao, P.; She, L.; McGill, M.; and Tsao, D. Y. 2020. A map of object space in primate inferotemporal cortex. *Nature*, 583(7814): 103–108.

Bashivan, P.; Kar, K.; and DiCarlo, J. J. 2019. Neural population control via deep image synthesis. *Science*, 364(6439): eaav9436.

Batty, E.; Merel, J.; Brackbill, N.; Heitman, A.; Sher, A.; Litke, A.; Chichilnisky, E.; and Paninski, L. 2017. Multilayer recurrent network models of primate retinal ganglion cell responses. In *Proceedings of International Conference on Learning Representations*.

Billeh, Y. N.; Cai, B.; Gratiy, S. L.; Dai, K.; Iyer, R.; Gouwens, N. W.; Abbasi-Asl, R.; Jia, X.; Siegle, J. H.; Olsen, S. R.; et al. 2020. Systematic integration of structural and functional data into multi-scale models of mouse primary visual cortex. *Neuron*, 106(3): 388–403.

Cadena, S. A.; Denfield, G. H.; Walker, E. Y.; Gatys, L. A.; Tolias, A. S.; Bethge, M.; and Ecker, A. S. 2019. Deep convolutional models improve predictions of macaque V1 responses to natural images. *PLoS Computational Biology*, 15(4): e1006897.

Cadieu, C. F.; Hong, H.; Yamins, D. L.; Pinto, N.; Ardila, D.; Solomon, E. A.; Majaj, N. J.; and DiCarlo, J. J. 2014. Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLoS Computational Biology*, 10(12): e1003963.

Carandini, M.; Demb, J. B.; Mante, V.; Tolhurst, D. J.; Dan, Y.; Olshausen, B. A.; Gallant, J. L.; and Rust, N. C. 2005. Do we know what the early visual system does? *Journal of Neuroscience*, 25(46): 10577–10597.

Carandini, M.; Heeger, D. J.; and Anthony Movshon, J. 1999. Linearity and gain control in V1 simple cells. *Models of Cortical Circuits*, 401–443.

Cheng, S.; Butrus, S.; Tan, L.; Xu, R.; Sagireddy, S.; Trachtenberg, J. T.; Shekhar, K.; and Zipursky, S. L. 2022. Vision-dependent specification of cell types and function in the developing cortex. *Cell*, 185(2): 311–327.

Cotton, R. J.; Sinz, F.; and Tolias, A. 2020. Factorized neural processes for neural processes: K-shot prediction of neural responses. *Advances in Neural Information Processing Systems*, 33: 11368–11379.

Deng, K.; Schwendeman, P. S.; and Guan, Y. 2024. Predicting Single Neuron Responses of the Primary Visual Cortex with Deep Learning Model. *Advanced Science*, 11(15): 2305626.

Du, F.; Núñez-Ochoa, M. A.; Pachitariu, M.; and Stringer, C. 2024. Towards a simplified model of primary visual cortex. *bioRxiv*, 2024–06.

Franke, K.; Willeke, K. F.; Ponder, K.; Galdamez, M.; Muhammad, T.; Patel, S.; Froudarakis, E.; Reimer, J.; Sinz, F.; and Tolias, A. S. 2021. Behavioral state tunes mouse vision to ethological features through pupil dilation. *bioRxiv*, 2021–09.

Franke, K.; Willeke, K. F.; Ponder, K.; Galdamez, M.; Zhou, N.; Muhammad, T.; Patel, S.; Froudarakis, E.; Reimer, J.; Sinz, F. H.; et al. 2022. State-dependent pupil dilation rapidly shifts visual feature selectivity. *Nature*, 610(7930): 128–134.

Heeger, D. J. 1992. Half-squaring in responses of cat striate cells. *Visual Neuroscience*, 9(5): 427–443.

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.

Jones, J. P.; and Palmer, L. A. 1987. The two-dimensional spatial structure of simple receptive fields in cat striate cortex. *Journal of Neurophysiology*, 58(6): 1187–1211.

Klindt, D.; Ecker, A. S.; Euler, T.; and Bethge, M. 2017. Neural system identification for large populations separating "what" and "where". *Advances in Neural Information Processing Systems*, 30.

Li, B. M.; Cornacchia, I. M.; Rochefort, N. L.; and Onken, A. 2023. V1T: large-scale mouse V1 response prediction using a Vision Transformer. *arXiv preprint arXiv:2302.03023*.

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.

Margalit, E.; Lee, H.; Finzi, D.; DiCarlo, J. J.; Grill-Spector, K.; and Yamins, D. L. 2023. A unifying principle for the functional organization of visual cortex. *bioRxiv*.

McIntosh, L.; Maheswaranathan, N.; Nayebi, A.; Ganguli, S.; and Baccus, S. 2016. Deep learning models of the retinal response to natural scenes. *Advances in neural information processing systems*, 29.

Olshausen, B. A.; and Field, D. J. 1996. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583): 607–609.

Ponce, C. R.; Xiao, W.; Schade, P. F.; Hartmann, T. S.; Kreiman, G.; and Livingstone, M. S. 2019. Evolving images for visual neurons using a deep generative network reveals coding principles and neuronal preferences. *Cell*, 177(4): 999–1009.

Ran, X.; Zhang, J.; Ye, Z.; Wu, H.; Xu, Q.; Zhou, H.; and Liu, Q. 2021. Deep auto-encoder with neural response. *arXiv preprint arXiv:2111.15309*.

Sinz, F. H.; Pitkow, X.; Reimer, J.; Bethge, M.; and Tolias, A. S. 2019. Engineering a less artificial intelligence. *Neuron*, 103(6): 967–979.

Willeke, K. F.; Fahey, P. G.; Bashiri, M.; Pede, L.; Burg, M. F.; Blessing, C.; Cadena, S. A.; Ding, Z.; Lurz, K.-K.; Ponder, K.; et al. 2022. The sensorium competition on

predicting large-scale mouse primary visual cortex activity. *arXiv preprint arXiv:2206.08666*.

Yamins, D. L.; Hong, H.; Cadieu, C.; and DiCarlo, J. J. 2013. Hierarchical modular optimization of convolutional networks achieves representations similar to macaque IT and human ventral stream. *Advances in neural information processing systems*, 26: 3093–3101.

Yamins, D. L.; Hong, H.; Cadieu, C. F.; Solomon, E. A.; Seibert, D.; and DiCarlo, J. J. 2014. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23): 8619–8624.

Yang, G. R.; and Wang, X.-J. 2020. Artificial neural networks for neuroscientists: a primer. *Neuron*, 107(6): 1048–1070.

Zhang, M.; Luo, X.; Wu, J.; Belatreche, A.; Cai, S.; Yang, Y.; and Li, H. 2025a. Toward Building Human-Like Sequential Memory Using Brain-Inspired Spiking Neural Models. *IEEE transactions on neural networks and learning systems*.

Zhang, M.; Wang, J.; Wu, J.; Belatreche, A.; Amornpaisannon, B.; Zhang, Z.; Miriyala, V. P. K.; Qu, H.; Chua, Y.; Carlson, T. E.; et al. 2021. Rectified linear postsynaptic potential function for backpropagation in deep spiking neural networks. *IEEE transactions on neural networks and learning systems*, 33(5): 1947–1958.

Zhang, M.; Wang, S.; Wu, J.; Wei, W.; Zhang, D.; Zhou, Z.; Wang, S.; Zhang, F.; and Yang, Y. 2025b. Toward Energy-Efficient Spike-Based Deep Reinforcement Learning With Temporal Coding. *IEEE Computational Intelligence Magazine*, 20(2): 45–57.

Zhang, M.; Wei, W.; Zhou, Z.; Liu, W.; Zhang, J.; Belatreche, A.; and Yang, Y. 2025c. Spike-Driven Lightweight Large Language Model With Evolutionary Computation. *IEEE Transactions on Evolutionary Computation*.