IMAGE2MUA: GENERATIVE MODELING OF TRIAL-LEVEL MACAQUE MUA WITH AREA STRUCTURED VISUAL EXPERT ROUTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding how visual stimuli shape neural population activity is a central goal of visual neuroscience. Visual encoding models predict mean responses to visual stimuli accurately, but mean responses alone do not capture response variability across repeated presentations of the same stimulus. Capturing these trial-level responses across cortical areas requires modeling correlated fluctuations while accommodating different visual feature preferences. We introduce IMAGE2MUA, a conditional rectified flow model that learns to generate joint multi-unit activity (MUA) in macaque primary visual cortex (V1), visual area V4, and inferior temporal (IT) cortex from one response per training image. IMAGE2MUA combines joint response generation with visual expert routing, which learns how individual channels draw on different visual representations. Across two macaques, its samples capture both trial to trial variability within individual channels and pooled residual correlations between channels, while their averages outperform the evaluated deterministic encoding baselines in mean response prediction. Learned visual expert preferences differ across cortical areas, even when area labels are not provided during training. Together, these results show that a single model can capture population variability across trials while making its use of visual representations explicit for each channel.

1 INTRODUCTION

Given an image, we seek to model and generate the distribution of simultaneous neural population responses across the primate ventral stream. Primary visual cortex (V1), visual area V4, and inferior temporal (IT) cortex span successive stages of visual processing, from local visual attributes in V1 through shape sensitive representations in V4 to tolerant object codes in IT (Felleman & Van Essen, 1991; Pasupathy & Connor, 2002; Yamins et al., 2014). Repeated presentations of the same image nevertheless elicit variable multi-unit activity (MUA) (Churchland et al., 2010). Predicting only the stimulus evoked mean is therefore insufficient; a complete model must also capture the response distribution of each channel and how channels fluctuate together across trials.

Visual encoding models based on features optimized for visual tasks, flexible readouts, and attention achieve strong mean prediction (Yamins et al., 2014; Schrimpf et al., 2018; Klindt et al., 2017; Lin et al., 2025; Adeli et al., 2025). Generative population models extend neural encoding to response distributions conditioned on images in mouse visual cortex (Bashiri et al., 2021) and macaque V1 (Shrinivasan et al., 2023); notably, the former learns these distributions without repeated training presentations. Building on these advances, we ask how a single generator can model joint trial-level MUA across V1, V4, and IT while making each channel's visual expert preferences explicit.

Matching individual channel distributions does not determine how channels fluctuate together. Different cortical channels may also favor different visual representations. We introduce IMAGE2MUA, a conditional rectified flow model that addresses both aspects by generating joint MUA responses while learning a mixture of visual experts for each channel. Population self-attention models interactions between channels, and a factorized visual expert router combines stable channel preferences with an adjustment driven by the image and shared across channels. The model learns from one response per

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

training image. Its routing weights let us test whether the learned visual expert preferences differ across cortical areas even when training excludes area labels.

Our main contributions can be summarized as follows:

- We introduce IMAGE2MUA, a conditional rectified flow model with factorized visual expert routing for joint MUA generation across V1, V4, and IT.
- We demonstrate that IMAGE2MUA learns response variability across trials from one response per training image in two macaques. Its samples capture individual channel distributions and pooled residual correlations between channels, while their averages outperform the evaluated deterministic encoding baselines in predicting mean responses.
- We find that learned visual expert preferences differ across cortical areas even when area labels are excluded during training, revealing area-related structure in how the model allocates visual representations to individual channels.

2 RELATED WORK

Image-to-neural encoding. Goal driven networks established that visual features optimized for task performance can predict responses across the visual cortex (Yamins et al., 2014; Schrimpf et al., 2018). Flexible readouts and contextual modulation through attention further improved neural response prediction from shared feature maps (Klindt et al., 2017; Willeke et al., 2022; Lin et al., 2025). The Transformer Brain Encoder (TBE) uses attention over retinotopic features to predict functional magnetic resonance imaging (fMRI) responses in high-level visual cortex (Adeli et al., 2025). The THINGS Ventral-stream Spiking Dataset (TVSD) provides simultaneous MUA recordings from V1, V4, and IT, together with array-specific Inception-v1 encoders (Papale et al., 2025). Together, these studies show how visual features can be selected and combined to predict conditional mean responses across recording modalities. Our focus is on joint distributions of responses across trials and the mixture of visual representations used by each channel.

Generative neural population models. Neural responses vary across repeated trials in a manner that depends on the stimulus (Churchland et al., 2010). A generative population model combines a normalizing flow for each neuron with a Gaussian factor analysis model of the population to capture activity conditioned on natural images and shared variability across mouse visual areas (Bashiri et al., 2021). Importantly, it learns response distributions without repeated presentations during training. Measured cortical coordinates guide its receptive field mapping, and its receptive field and latent factor analyses recover structure associated with cortical areas. Neural sampling is formulated with a joint generative model of images and responses, in which neural responses act as latent variables for image generation (Shrinivasan et al., 2023). The model uses an approximate posterior to infer the response distribution for a given image. Related models based on flows translate between image and fMRI latents (Mai et al., 2026) or forecast blood oxygen level dependent (BOLD) dynamics (Rogalla et al., 2026), rather than generate electrophysiological population responses for individual trials. We study joint trial-level MUA across V1, V4, and IT and ask whether learned mixtures of visual experts vary with cortical area in models trained without area labels. IMAGE2MUA uses rectified flow matching (Liu et al., 2023; Lipman et al., 2023) to learn a conditional velocity field over the simultaneous V1, V4, and IT MUA vector. Each channel is conditioned on a learned mixture of representations from frozen visual backbones.

Mixture-of-experts routing. Mixture of experts (MoE) models learn to combine specialized predictors using weights that depend on the input (Jacobs et al., 1991). Modern sparse MoEs increase Transformer capacity while activating only a subset of feedforward networks (FFNs) for each token (Shazeer et al., 2017; Fedus et al., 2022). Attention mixtures extend routing to attention heads or projections (Zhang et al., 2022; Csordás et al., 2024). In IMAGE2MUA, visual expert routing serves a different purpose: it determines how each neural channel combines representations from distinct visual sources. Each cross-attention expert processes one frozen feature source. All four visual experts are evaluated, and their outputs are mixed for each channel. The routing weights describe each channel's preferences among these experts. A separate FFN MoE retains the conventional role of increasing nonlinear capacity through sparse computation.

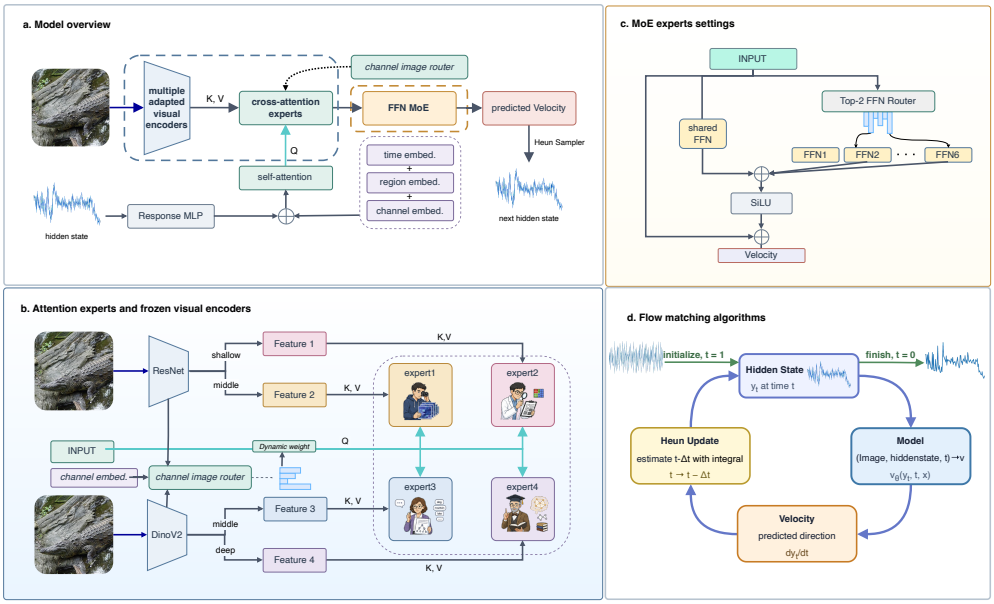

Figure 1: **IMAGE2MUA architecture and generation.** (a) Response backbone. (b) Dense routing over four source-specific cross-attention experts. (c) A shared FFN plus the top-2 of six routed FFNs. (d) Fifty-step Heun generation from Gaussian noise to simultaneous MUA.

3 METHOD

3.1 PROBLEM SETTING AND OVERVIEW

We model the response distribution of one fixed neural population for each animal. For an animal with C retained channels, the training set $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{y}_{0,i})\}$ contains pairs of an image $\boldsymbol{x}_i$ and one simultaneous V1/V4/IT response vector $\boldsymbol{y}_{0,i} \in \mathbb{R}^C$. Repeated responses do not enter model fitting or checkpoint selection, although they are used for reliability based channel screening and evaluation. From these single presentation pairs, we learn the conditional population distribution $p_\theta(\boldsymbol{y}_0 \mid \boldsymbol{x})$.

IMAGE2MUA predicts a joint velocity field with a population-response backbone conditioned by four visual experts (Figure 1). Each expert queries one frozen visual source with the current response state, and a factorized router mixes the expert outputs for each channel. Population self-attention represents cross-channel interactions, while a separate sparse FFN MoE provides nonlinear processing. Rectified flow training fits the velocity field; Heun integration transforms Gaussian noise into simultaneous population responses.

3.2 FACTORIZED VISUAL EXPERT ROUTING

To share image dependent adjustments without imposing identical expert mixtures across channels, we factorize the routing logits into a static channel specific term and a shared image dependent term. For channel c, image $\boldsymbol{x}_i$, and expert $e \in \{0, 1, 2, 3\}$, the mixing probability is

$$\pi_{i,c,e} = \mathrm{softmax}_e(a_{c,e} + g_{\phi,e}(\boldsymbol{x}_i)) \,. \tag{1}$$

The learned channel logits $a_{c,:} \in \mathbb{R}^4$ do not vary across images and are distinct from the channel identity embedding in the response tokens. The image logits $g_\phi(\boldsymbol{x}_i) \in \mathbb{R}^4$ supply the same adjustment to every channel before softmax normalization. Image dependence therefore enters the router through a shared logit adjustment rather than a separate channel–image interaction term.

To compute the image term, ResNet-18 conv1 and layer2 and DINOv2 blocks 5 and 11 provide four frozen feature sources, E0–E3 (Figure 2) (He et al., 2016; Oquab et al., 2024). A trainable source specific projection maps each bank to $F_{i,e} \in \mathbb{R}^{N_e \times d}$, where N_e is its token count and d is the common width. Four learned summary queries $Q_e^{\mathrm{sum}} \in \mathbb{R}^{4 \times d}$ attend only to that source, producing

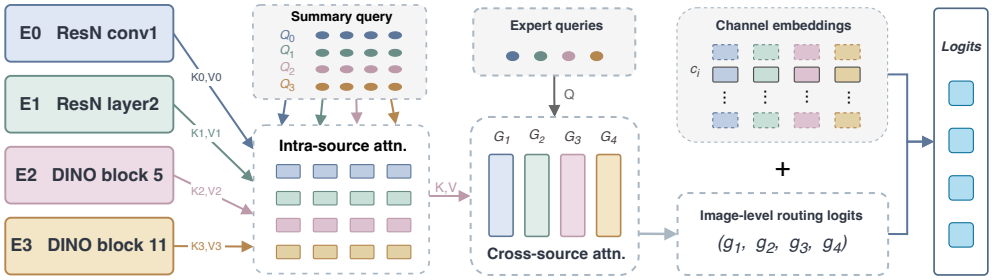

Figure 2: **Two-stage image router.** Per-source summary attention yields 16 tokens, to which four expert queries jointly attend to produce G_1–G_4. A shared readout maps them to g_1–g_4, which combine with channel routing logits to weight E0–E3.

$S_{i,e} \in \mathbb{R}^{4\times d}$. Concatenating the summaries gives $S_i = [S_{i,0}; S_{i,1}; S_{i,2}; S_{i,3}] \in \mathbb{R}^{16\times d}$. Four learned expert queries $q_e^{\text{exp}} \in \mathbb{R}^d$ then attend jointly to S_i as keys and values, producing $G_{i,e} \in \mathbb{R}^d$. A shared linear readout maps each output to $g_{\phi,e}(\boldsymbol{x}_i)$; Eq. 1 combines these image logits with the channel logits to obtain $\boldsymbol{\pi}_{i,c} \in \mathbb{R}^4$.

The routing weights modulate a separate visual path conditioned on the response. Response tokens are formed by summing embeddings of channel identity, the current flow state, flow time, and, in the main model, cortical area. Population self-attention updates these tokens to hidden states $h_{i,c,t}$. Each normalized hidden state queries all four visual sources through separate trainable cross-attention experts, producing $z_{i,c,e,t}$ for $e \in \{0, 1, 2, 3\}$. The four outputs are combined using $\boldsymbol{\pi}_{i,c}$ and added residually to $h_{i,c,t}$, yielding the visually conditioned state $\tilde{h}_{i,c,t}$.

The routing weights depend only on channel identity and the image; the response state, flow time, and area information affect the expert outputs through the response query. All four visual experts are evaluated, with no top k selection. After visual fusion, a separate FFN MoE combines an always-active shared FFN with a weighted top-2 mixture of six routed FFNs to expand nonlinear capacity. These FFN experts are not assigned to visual sources or cortical areas.

3.3 Rectified flow framework

To model the conditional distribution of simultaneous population responses rather than produce a single point prediction, we train IMAGE2MUA as a conditional rectified flow. For a recorded response $\boldsymbol{y}_0 \in \mathbb{R}^C$, we draw Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I_C)$ and a time $t \sim \mathcal{U}[0, 1]$. We denote the interpolated response state by $\boldsymbol{y}_t$ and its target velocity by $\boldsymbol{u}_t$:

$$\boldsymbol{y}_t = (1-t)\boldsymbol{y}_0 + t\boldsymbol{\epsilon}, \qquad \boldsymbol{u}_t = \boldsymbol{\epsilon} - \boldsymbol{y}_0. \tag{2}$$

The network predicts $v_\theta(\boldsymbol{y}_t, t, \boldsymbol{x}) \approx \boldsymbol{u}_t$. With C_r denoting the channels in cortical area r, we minimize squared velocity error, averaged within each area and then equally across the three areas:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}\left[\frac{1}{3}\sum_r \frac{1}{|C_r|}\left\|v_{\theta,C_r}(\boldsymbol{y}_t, t, \boldsymbol{x}) - \boldsymbol{u}_{t,C_r}\right\|_2^2\right]. \tag{3}$$

The expectation is over training pairs, Gaussian noise, and flow time. Equal area weighting prevents areas with more channels from dominating the objective.

At inference, a 50-step Heun solver integrates the learned velocity field from Gaussian noise at $t = 1$ to a simultaneous C-channel response at $t = 0$. Independent initial noise yields additional population samples. The joint velocity field allows coordinated variation across V1, V4, and IT; Appendix C gives the full update equations.

4 Experimental Setup

Dataset and preprocessing. We use MUA recordings from Monkeys F and N in TVSD (Papale et al., 2025), with modified data splits and response normalization. The single-presentation training

and validation sets share no classes with the repeated test set. For each recording day and channel, normalization statistics are estimated from the combined training and validation responses and applied to all three splits. Channels are retained if their mean pairwise reliability exceeds .3, leaving 606 channels for Monkey F and 722 for Monkey N. Full preprocessing and split details are provided in Appendix B.

Comparisons. Mean response baselines comprise ridge readouts on principal component analysis (PCA), DINOv2, and Contrastive Language-Image Pre-training (CLIP) features, a Sensorium convolutional neural network (CNN), and TBE. Stochastic training runs use common seeds; baselines with deterministic fitting are evaluated once. Table 1 reports two settings separately: our class disjoint split and the original TVSD split and preprocessing (606 channels for Monkey F and 749 for Monkey N). We train a separate IMAGE2MUA instance for each setting and compare methods only within the same block. As a generative baseline, we train a denoising diffusion probabilistic model (DDPM) with the same frozen visual backbones and trainable architecture, and generate samples using 50-step denoising diffusion implicit model (DDIM) sampling (DDPM–DDIM-50).

Control experiments. We train two routing controls: Global Static, with one expert mixture shared across channels and images, and a region blind model, with area derived inputs removed and the loss averaged over all channels. Area labels are attached to the region blind results only after inference. We further intervene on trained models without retraining by swapping the image dependent routing term, permuting channel specific preferences, or applying both interventions, while keeping the visual features unchanged. For distributional evaluation, we collapse generated samples to their image conditioned means or independently shuffle trials within each image and channel. We also compute a ground-truth leave-one-out (GT LOO) reference from recorded repetitions. Appendix B gives the configurations and seed counts.

Evaluation metrics. We evaluate mean responses with channel averaged Pearson correlation (MCP), marginal distributions with the continuous ranked probability score (CRPS), population dependence with noise-correlation (NC) similarity, and visual expert preferences with centered log ratio (CLR) profile correlation. Let K and M be the numbers of generated and recorded trials per image, and let $\mathcal{C}$ contain $C = |\mathcal{C}|$ evaluated channels. For image i and channel c, $\hat{y}_{i,c}^{(k)}$ and $y_{i,c}^{(m)}$ denote generated sample k and recorded trial m.

MCP correlates the generated and recorded means across images within each channel and then averages across channels:

$$\mathrm{MCP} = \frac{1}{C} \sum_{c \in \mathcal{C}} \mathrm{corr}_i \left(\frac{1}{K} \sum_{k=1}^{K} \hat{y}_{i,c}^{(k)}, \frac{1}{M} \sum_{m=1}^{M} y_{i,c}^{(m)} \right). \tag{4}$$

Higher MCP indicates better mean prediction. CRPS evaluates the empirical response distribution for each image and channel (Gneiting & Raftery, 2007):

$$\mathrm{CRPS}_{i,c} = \frac{1}{KM} \sum_{k,m} \left| \hat{y}_{i,c}^{(k)} - y_{i,c}^{(m)} \right| - \frac{1}{2K^2} \sum_{k,k'} \left| \hat{y}_{i,c}^{(k)} - \hat{y}_{i,c}^{(k')} \right|. \tag{5}$$

We average CRPS across images and channels; lower is better. The GT LOO reference scores each recorded trial against the other $M - 1 = 29$ repetitions.

NC similarity removes the stimulus mean before comparing dependence across channels. We center recorded and generated responses as $r_{i,c}^{(m)} = y_{i,c}^{(m)} - M^{-1} \sum_{m'} y_{i,c}^{(m')}$ and $\hat{r}_{i,c}^{(k)} = \hat{y}_{i,c}^{(k)} - K^{-1} \sum_{k'} \hat{y}_{i,c}^{(k')}$, respectively, and define

$$S_{\mathrm{NC}} = \mathrm{corr}_{c<c'} \left(\mathrm{corr}_{i,k}(\hat{r}_{i,c}^{(k)}, \hat{r}_{i,c'}^{(k)}), \mathrm{corr}_{i,m}(r_{i,c}^{(m)}, r_{i,c'}^{(m)}) \right). \tag{6}$$

Thus, S_{NC} is the Pearson similarity between the upper triangles of the generated and recorded residual correlation matrices pooled across images; higher is better.

For routing, let $p_{c,e} = 100^{-1} \sum_{i=1}^{100} \pi_{i,c,e}$ be channel c's mean probability for expert e. We apply the within-channel CLR transform $z_{c,e} = \log \tilde{p}_{c,e} - \frac{1}{4} \sum_{j=0}^{3} \log \tilde{p}_{c,j}$, where $\tilde{p}_{c,e} = \max(p_{c,e}, 10^{-12})$.

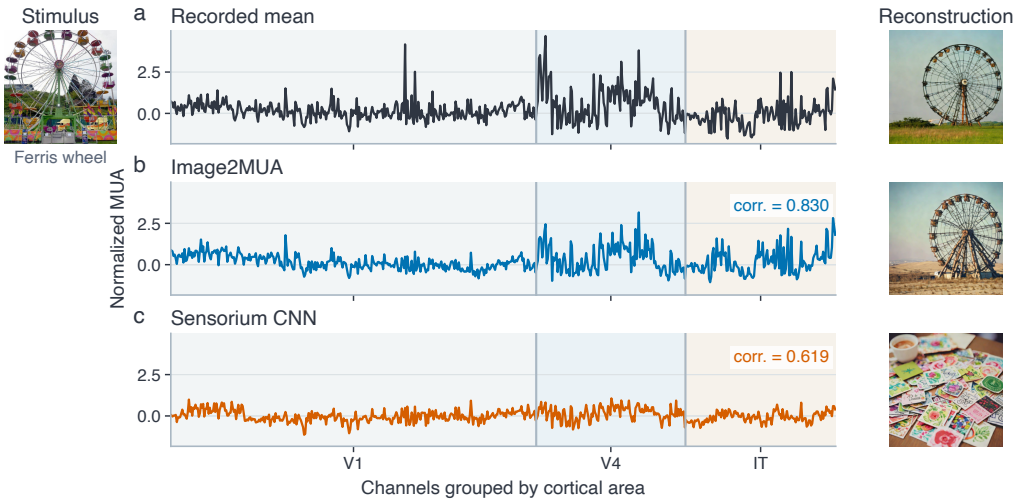

Figure 3: **Average neural responses and neural-to-image reconstructions for one image from Monkey F.** The left column shows the average response across all 606 channels, grouped by V1, V4, and IT: (a) the mean of 30 recorded trials; (b) the mean of 30 IMAGE2MUA samples generated from 15 sampled noise vectors and their negatives; and (c) the prediction from the Sensorium CNN checkpoint. In the right column, each response vector is passed through the same separately trained neural decoding model.

This expresses each expert probability relative to the geometric mean for that channel. We concatenate each channel's four values in a fixed channel order to form a length $4C$ vector and correlate these vectors between models. Higher correlation indicates greater agreement in these CLR profiles, which retain both population-wide expert preferences and channel specific deviations.

5 RESULTS

We evaluate the generated responses at three levels: stimulus driven means, per-channel trial distributions, and pooled residual correlations. We then examine how learned visual expert preferences relate to cortical areas and prediction.

5.1 AVERAGE-RESPONSE PREDICTION

On both our class disjoint split and the original TVSD split, averages of IMAGE2MUA samples achieve higher MCP than the evaluated deterministic encoding baselines, both overall and in every cortical area (Table 1). Internal visual-source and routing controls are compared separately in Appendix Table 6.

Figure 3 illustrates this result for the Ferris wheel test image. Across Monkey F's 606 channels, the generated mean correlates with the recorded mean at .830, compared with .619 for Sensorium CNN, the strongest convolutional baseline in this evaluation. The generated mean tracks the large V4 and IT responses more closely. Passing the three response vectors through the same separately trained, fixed neural-to-image model produces similar Ferris wheel structure from the recorded and generated means, but not from the Sensorium prediction. In this selected example, the reconstructions suggest that the generated responses retain stimulus specific information; Appendix Figure 6 gives additional examples for both animals.

Together, these results show that modeling the response distribution does not come at the expense of mean prediction. Averaging IMAGE2MUA samples yields a strong mean response predictor that outperforms the evaluated deterministic baselines in both animals and all three cortical areas.

Table 1: **Average-response MCP.** The top block uses our class-disjoint split; the bottom uses the original TVSD data split and preprocessing. Separate IMAGE2MUA models are trained. † Inception-v1 values are as originally reported.

	Monkey F				Monkey N			
Model	All	V1	V4	IT	All	V1	V4	IT
PCA Ridge	.26	.38	.12	.10	.17	.28	.08	−.02
DINOv2 Ridge (Oquab et al., 2024)	.46	.43	.44	.53	.49	.46	.49	.57
CLIP Ridge (Radford et al., 2021)	.48	.43	.49	.60	.52	.47	.54	.62
Sensorium CNN (Willeke et al., 2022)	.71	.83	.69	.47	.75	.85	.70	.54
TBE (fMRI transfer) (Adeli et al., 2025)	.84	.89	.82	.73	.75	.74	.75	.80
IMAGE2MUA	**.87**	**.92**	**.85**	**.79**	**.89**	**.92**	**.88**	**.84**
Inception-v1 [TVSD setting] (Papale et al., 2025)	.75†	.76	.74	.72	.76†	.75	.76	.79
IMAGE2MUA [TVSD setting]	**.87**	**.91**	**.85**	**.80**	**.89**	**.92**	**.88**	**.83**

5.2 TRIAL-LEVEL POPULATION RESPONSE GENERATION

Beyond mean prediction, generated samples capture per-channel response distributions and pooled cross-channel residual correlations (Figure 4). CRPS is $.434 \pm .006$ for Monkey F and $.400 \pm .001$ for Monkey N; variance ratios are $1.07 \pm .03$ and $1.05 \pm .02$, close to one; and NC similarities are $.814 \pm .009$ and $.775 \pm .017$, respectively (Table 2; Appendix Table 5).

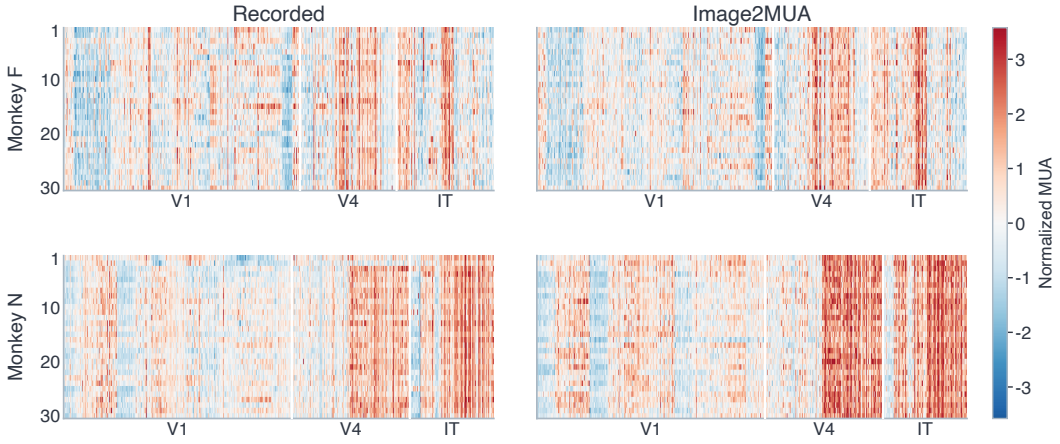

Figure 4: **Recorded and generated trial populations.** For one image (horse), all 30 recorded trials (left) and 30 independent samples (right) are shown for Monkey F (top) and Monkey N (bottom). White lines separate channels from different cortical areas.

Collapsing each generated ensemble to its mean raises CRPS from .434 to .605 in Monkey F and from .400 to .553 in Monkey N. Retaining the distribution reduces CRPS by 28.2% and 27.7%, respectively, with improvements for every image across all five seeds. Relative to GT LOO values of .380 and .342, IMAGE2MUA closes 75.9% and 72.6% of the empirical gap (Eq. 8). Independently shuffling trials within each image and channel leaves CRPS unchanged but reduces NC similarity to approximately zero. These contrasts separate marginal fidelity from population dependence.

Together, these results show that IMAGE2MUA captures trial variability at two complementary scales: it predicts each channel's response distribution across repeated presentations and preserves dependence among channels within simultaneous population samples. The model therefore recovers both channel level uncertainty and pooled residual correlations between channels.

5.3 ASSOCIATIONS BETWEEN VISUAL EXPERTS AND CORTICAL REGIONS

Visual expert preferences differ across cortical areas even in models trained without area labels (Figure 5). In the region blind controls, source centered area means show elevated allocation to E0 in V1 and to E1 and E2 in V4 in both animals. IT differs between animals: E1 and E3 are elevated in

Table 2: **Distribution fidelity, generative comparison, and controls.** Lower CRPS and higher pooled NC similarity are better. DDPM–DDIM-50 uses the same architecture with a different generative formulation. Values are reported as mean ± standard deviation (SD) across five seeds; GT LOO is a fixed data-derived reference. CW shuffle denotes channel-wise trial shuffling.

Animal	Metric	Entry	Overall	V1	V4	IT
Monkey F	CRPS ↓	IMAGE2MUA	.434 ± .006	.418 ± .006	.428 ± .005	.480 ± .009
		DDPM–DDIM-50	.462 ± .005	.454 ± .006	.453 ± .005	.490 ± .003
		GT LOO	.380	.382	.349	.405
		Mean collapse	.605 ± .011	.583 ± .012	.594 ± .010	.668 ± .016
	NC sim. ↑	IMAGE2MUA	.814 ± .009	.887 ± .010	.871 ± .010	.849 ± .008
		DDPM–DDIM-50	.770 ± .015	.844 ± .012	.868 ± .007	.841 ± .007
		CW shuffle	.003 ± .003	.002 ± .005	.001 ± .019	.000 ± .013
Monkey N	CRPS ↓	IMAGE2MUA	.400 ± .001	.381 ± .002	.415 ± .003	.431 ± .004
		DDPM–DDIM-50	.454 ± .002	.445 ± .003	.452 ± .003	.481 ± .004
		GT LOO	.342	.336	.354	.345
		Mean collapse	.553 ± .003	.526 ± .004	.575 ± .005	.596 ± .006
	NC sim. ↑	IMAGE2MUA	.775 ± .017	.874 ± .011	.743 ± .018	.863 ± .009
		DDPM–DDIM-50	.728 ± .006	.820 ± .007	.707 ± .014	.865 ± .006
		CW shuffle	.000 ± .002	−.001 ± .003	−.002 ± .009	−.007 ± .010

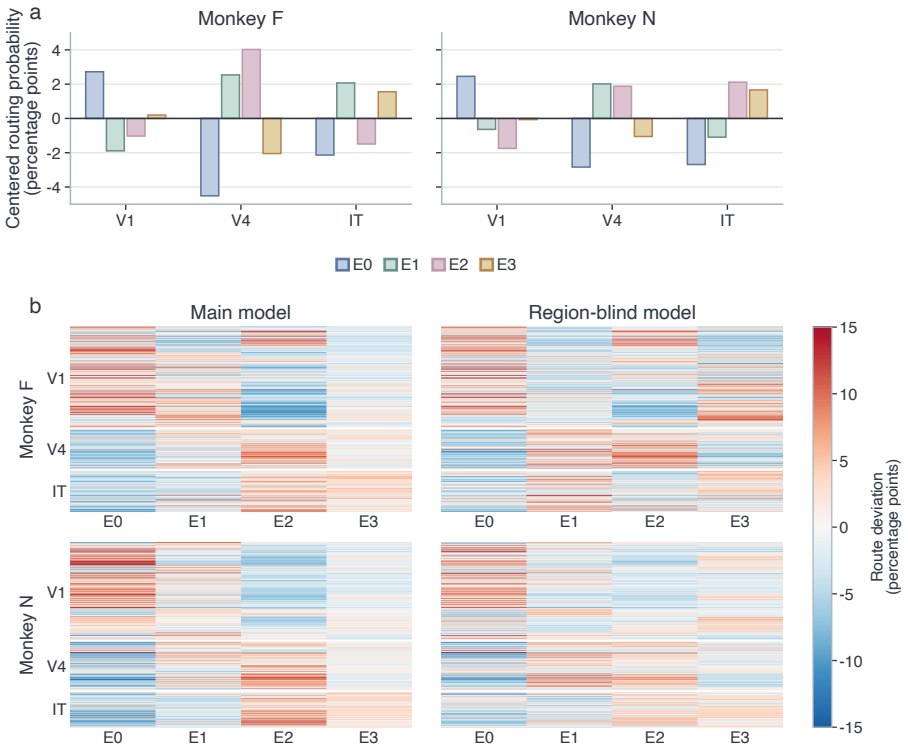

Figure 5: **Area-expert associations in region blind models.** Main and region blind checkpoints trained with seed 42 are evaluated on the same 100 images, with probabilities centered across channels per source: (a) region blind area means; (b) all-channel main and region blind profiles. Area labels are excluded from the region blind model and loss and used to group and order its inferred routing profiles. E0/E1: ResNet-18 conv1/layer2; E2/E3: DINOv2 blocks 5/11.

Monkey F, whereas E2 and E3 are elevated in Monkey N. Here, elevated allocation means above that expert's average routing probability across all channels, not a fixed expert-to-area assignment.

Cortical area R^2 measures the fraction of variance in the 100-image mean routing profiles associated with area membership (Appendix B). It is 24.08% in the main model and 18.13% in the region blind model for Monkey F; the corresponding values for Monkey N are 30.14% and 16.50%. Predictive

Table 3: **Routing interventions without retraining.** Five-seed paired changes from the unchanged model; negative ΔMCP and ΔNC similarity and positive ΔCRPS indicate degradation. Area-wise results are in Appendix Table 8.

Animal	Intervention	ΔMCP	ΔCRPS	ΔNC sim.
Monkey F	Image	−.0014	+.0003	+.00019
	Channel	−.0075	+.0040	−.00005
	Both	−.0092	+.0043	+.00012
Monkey N	Image	−.0006	−.0001	+.00067
	Channel	−.0048	+.0041	+.00229
	Both	−.0052	+.0040	+.00286

scores remain close. MCP is 0.872 versus 0.876 and CRPS is 0.429 for both Monkey F models; for Monkey N, MCP is 0.897 versus 0.888 and CRPS is 0.398 versus 0.399, with the main model listed first in each comparison.

The main and region blind CLR routing profiles correlate at $r = 0.655$ for Monkey F and $r = 0.960$ for Monkey N. This agreement includes population-wide expert preferences and therefore complements, rather than replaces, the source centered profiles in Figure 5(b) and cortical-area R^2.

Permuting channel specific expert preferences in the trained models degrades MCP and CRPS more than swapping the shared image term in both animals, while S_{NC} changes little (Table 3). The trained model therefore relies more on channel specific preferences for prediction. Across five seeds, the full router has higher mean MCP and lower mean CRPS than Global Static in Monkey N, but lower mean MCP and higher mean CRPS in Monkey F (Appendix Table 6).

Together, these results show that channel preferences for visual experts are systematically organized by cortical area. This organization remains visible when area labels are excluded from both model inputs and the training loss, indicating that it can emerge without explicit cortical area information.

6 Discussion and Conclusion

We argue for jointly modeling trial-level population responses and channel specific visual feature use in multi-area neural encoding. IMAGE2MUA combines these aims in one conditional generator: its samples capture per-channel variability and pooled cross-channel correlations, their averages predict stimulus driven means (Tables 1 and 2), and its routing weights specify each channel's mixture of visual expert outputs. Our single-presentation result extends the use of macaque ventral-stream recordings from fitting mean responses to training population generators. Building on earlier distributional models (Bashiri et al., 2021; Shrinivasan et al., 2023), IMAGE2MUA learns simultaneous V1/V4/IT MUA from one response per training image. Held-out repeated responses allow us to assess whether this training setting captures trial-level structure for unseen stimuli.

The region blind results further show that channel specific expert allocation can acquire cortical-area structure without an imposed expert-to-area mapping (Figure 5). We interpret the shared V1 and V4 preference patterns, together with the differing IT mixtures, as partially shared organization of visual feature use across the two animals. To test whether these preferences affect response prediction, we perturb the routing of trained models without retraining. Permuting channel specific expert preferences degrades prediction more than swapping the image dependent term, indicating that the generator relies on its learned channel preferences (Table 3).

Limitations and future directions. We study fixed neural populations in two macaques, using fixed visual sources and one region blind checkpoint per animal. Cross-animal comparisons concern area-level routing patterns, as channels and electrode placements are not aligned across animals. Evaluating more animals, recording configurations, and independently trained region blind models would test whether these patterns recur beyond the present recordings.

Taken together, our results show that visual encoding need not stop at the mean response: trial-to-trial population variation and channel specific visual expert preferences can be learned together.

AI USE STATEMENT

Required disclosure. We used generative AI tools to critique and refine hypotheses first formulated by the authors, provide feedback on methodology and experiments designed by the authors, assist with method implementation, and, to a limited extent, formulate mathematical claim and help interpret results. The tools did not propose hypotheses or design experiments. We did not use them to generate synthetic datasets; develop theoretical models or conceptual frameworks; provide critical ingredients for mathematical proofs; write proofs; translate text; clean or reformat datasets; or conduct qualitative or thematic data analysis.

Additional assistance. We also used generative AI tools to create or modify scientific figures, create or edit software code and other research artifacts, draft and edit parts of the manuscript, summarize and identify relevant literature, brainstorm presentation strategies, improve readability, format references, and suggest the manuscript structure.

Human verification and responsibility. We reviewed all AI-assisted work. AI-assisted code was executed and tested, reported quantitative values were checked against saved experimental outputs, figures were checked against their source data, and literature suggestions were checked against the original sources. We take responsibility for the final content of this work, including text, claims, code, figures, and other artifacts produced with the aid of generative AI.

REPRODUCIBILITY AND ETHICS STATEMENT

Model, hyperparameter, and checkpoint selection use only the class-disjoint validation data; repeated-image responses are used only for reliability-based channel selection and evaluation. Frozen evaluation uses predefined seeds and matched response and noise banks, while controls retain the stated splits, checkpoint rules, and evaluation procedure. Appendices A–D report the evaluation, control, and statistical details.

TVSD consists of previously released non-human-primate recordings, and this work introduces no new animal experiments. We treat results from the two animals as cross-animal replication rather than as support for universal claims about primate visual cortex.

REFERENCES

Hossein Adeli, Minni Sun, and Nikolaus Kriegeskorte. Transformer brain encoders explain human high-level visual responses. In *Advances in Neural Information Processing Systems*, volume 38, pp. 63288–63319, 2025. doi: 10.52202/085713-1902.

Mohammad Bashiri, Edgar Y. Walker, Konstantin-Klemens Lurz, Akshay Kumar Jagadish, Taliah Muhammad, Zhiwei Ding, Zhuokun Ding, Andreas S. Tolias, and Fabian H. Sinz. A flow-based latent state generative model of neural population responses to natural images. In *Advances in Neural Information Processing Systems*, volume 34, pp. 15801–15815, 2021.

Mark M. Churchland, Byron M. Yu, John P. Cunningham, Leo P. Sugrue, Marlene R. Cohen, Greg S. Corrado, William T. Newsome, Andrew M. Clark, Paymon Hosseini, Benjamin B. Scott, David C. Bradley, Matthew A. Smith, Adam Kohn, J. Anthony Movshon, Katherine M. Armstrong, Tirin Moore, Steve W. Chang, Lawrence H. Snyder, Stephen G. Lisberger, Nicholas J. Priebe, Ian M. Finn, David Ferster, Stephen I. Ryu, Gopal Santhanam, Maneesh Sahani, and Krishna V. Shenoy. Stimulus onset quenches neural variability: A widespread cortical phenomenon. *Nature Neuroscience*, 13(3):369–378, 2010. doi: 10.1038/nn.2501.

Róbert Csordás, Piotr Piękos, Kazuki Irie, and Jürgen Schmidhuber. SwitchHead: Accelerating transformers with mixture-of-experts attention. In *Advances in Neural Information Processing Systems*, volume 37, pp. 74411–74438, 2024. doi: 10.52202/079017-2368.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Daniel J. Felleman and David C. Van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, 1(1):1–47, 1991. doi: 10.1093/cercor/1.1.1-a.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Martin N. Hebart, Adam H. Dickter, Alexis Kidder, Wan Y. Kwok, Anna Corriveau, Caitlin Van Wicklin, and Chris I. Baker. THINGS: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PLOS ONE*, 14(10):e0223792, 2019. doi: 10.1371/journal.pone.0223792.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991. doi: 10.1162/neco.1991.3.1.79.

David A. Klindt, Alexander S. Ecker, Thomas Euler, and Matthias Bethge. Neural system identification for large populations separating "what" and "where". In *Advances in Neural Information Processing Systems*, volume 30, pp. 3506–3516, 2017.

Isaac Lin, Tianye Wang, Shang Gao, Shiming Tang, and Tai Sing Lee. Self-attention-based contextual modulation improves neural system identification. In *International Conference on Learning Representations*, 2025.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023.

Nikos K. Logothetis, Jon Pauls, Mark Augath, Torsten Trinath, and Axel Oeltermann. Neurophysiological investigation of the basis of the fMRI signal. *Nature*, 412(6843):150–157, 2001. doi: 10.1038/35084005.

Fernando Lopes da Silva. EEG and MEG: Relevance to neuroscience. *Neuron*, 80(5):1112–1128, 2013. doi: 10.1016/j.neuron.2013.10.017.

Weijian Mai, Mu Nan, Yu Zhu, Jiahang Cao, Rui Zhang, Yuqin Dai, Chunfeng Song, Andrew Luo, and Jiamin Wu. NeuroFlow: Toward unified visual encoding and decoding from neural activity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12966–12976, June 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Paolo Papale, Feng Wang, Matthew W. Self, and Pieter R. Roelfsema. An extensive dataset of spiking activity to reveal the syntax of the ventral stream. *Neuron*, 113(4):539–553.e5, 2025. doi: 10.1016/j.neuron.2024.12.003.

Anitha Pasupathy and Charles E. Connor. Population coding of shape in area V4. *Nature Neuroscience*, 5(12):1332–1338, 2002. doi: 10.1038/nn972.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Nicole Rogalla, Yuzhen Qin, Mario Senden, Ahmed El-Gazzar, and Marcel van Gerven. Probabilistic prediction of neural dynamics via autoregressive flow matching. *arXiv preprint arXiv:2604.11178*, 2026. doi: 10.48550/arXiv.2604.11178.

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, Kailyn Schmidt, Daniel L. K. Yamins, and James J. DiCarlo. Brain-Score: Which artificial neural network for object recognition is most brain-like? *bioRxiv*, 2018. doi: 10.1101/407007.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017.

Suhas Shrinivasan, Konstantin-Klemens Lurz, Kelli Restivo, George H. Denfield, Andreas S. Tolias, Edgar Y. Walker, and Fabian H. Sinz. Taking the neural sampling code very seriously: A data-driven approach for evaluating generative models of the visual system. In *Advances in Neural Information Processing Systems*, volume 36, pp. 21945–21959, 2023. doi: 10.52202/075280-0962.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

Konstantin F. Willeke, Paul G. Fahey, Mohammad Bashiri, Laura Pede, Max F. Burg, Christoph Blessing, Santiago A. Cadena, Zhiwei Ding, Konstantin-Klemens Lurz, Kayla Ponder, Taliah Muhammad, Saumil S. Patel, Alexander S. Ecker, Andreas S. Tolias, and Fabian H. Sinz. The sensorium competition on predicting large-scale mouse primary visual cortex activity. *arXiv preprint arXiv:2206.08666*, 2022. doi: 10.48550/arXiv.2206.08666. NeurIPS 2022 Competition Track.

Daniel L. K. Yamins, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014. doi: 10.1073/pnas.1403112111.

Xiaofeng Zhang, Yikang Shen, Zeyu Huang, Jie Zhou, Wenge Rong, and Zhang Xiong. Mixture of attention heads: Selecting attention heads per token. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 4150–4162. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.278.

A Reproducibility and Ethics Details

Evaluation settings. Frozen evaluation uses seeds 41–45, 100 repeated images, 30 observed responses, and 30 generated responses. MCP uses the registered antithetic $15{+}15$ bank, whereas CRPS and variance use 30 independent and identically distributed generated samples (IID30). GT LOO cycles through all 30 held-out trials and uses the other 29 repetitions of the same image–channel pair as the empirical forecast in each fold.

Frozen controls. The identical-mean, dependence, and routing interventions use frozen outputs or checkpoints from the main models. Figure 5 uses matched seed-42 main and region blind validation-selected checkpoints for each animal and the same 100-image evaluation set. In the region blind configuration, neither the model nor the loss consumes true area labels, the loss averages over the batch and all channels, and recorded V1/V4/IT labels are joined only after frozen routing inference.

Ethics and scope. TVSD consists of previously released non-human-primate recordings; this work introduces no new animal experiments.

B EXPERIMENTAL DETAILS

Dataset construction. TVSD contains trial-level MUA responses from Monkeys F and N to 22,248 images presented once and 100 images repeated 30 times (Papale et al., 2025). We remove all single-presentation images whose THINGS classes overlap with the repeated set (Hebart et al., 2019), leaving 21,048 images. We split these by class into 18,948 training images from 1,579 classes and 2,100 validation images from 175 classes; neither split overlaps the repeated classes. Only the test set contains repeated trials. Its repeated responses are used for reliability-based channel selection and evaluation, but not for model fitting, hyperparameter selection, or checkpoint selection.

Response preprocessing and normalization. Intracortical MUA measures local population activity at high temporal resolution, complementing the coarser spatial localization of electroencephalography (EEG) and magnetoencephalography (MEG) and the temporally indirect hemodynamic signal of fMRI (Logothetis et al., 2001; Lopes da Silva, 2013; Papale et al., 2025). MUA is averaged within area-specific post-stimulus windows: 25–125 ms for V1, 50–150 ms for V4, and 75–175 ms for IT. For each recording day and channel, the mean and standard deviation are estimated from the combined training and validation responses and used to z-score the training, validation, and repeated test responses. The model reads these stored values directly and applies no additional response normalization. Reliability is computed after this scaling, and channels with reliability greater than .3 are retained, yielding 606 channels for Monkey F (333/136/137 in V1/V4/IT) and 722 for Monkey N (383/198/141).

Image preprocessing. Images are resized to 224×224 and normalized with the ImageNet channel means and standard deviations expected by the frozen visual backbones.

Hyperparameter selection and optimization. The class-disjoint validation split is the only data used for model and hyperparameter selection. A prespecified three-seed grid selects peak learning rate 5×10^{-4} and batch size 64 by validation mean squared error (MSE) before seeds 41–45 and the full comparison matrix are launched. The response model has width 256, four attention heads, one response block, and dropout .05. We optimize with AdamW using a 5% linear warmup followed by cosine decay. Checkpoint selection, image splits, and training settings then remain unchanged across all main models and controls; repeated-image responses are not used for selection.

Visual source controls. Single-source controls retain only one of E0, E1, E2, or E3. All-ResNet and all-DINO controls instead replace the heterogeneous source set with four depths from one backbone family. These controls retain the fixed data splits, optimization procedure, checkpoint rule, and evaluation procedure. They compare source availability or composition rather than isolate one router parameter.

Routing controls. Global Static changes the routing rule by learning one four-source mixture shared across all images and channels; its response backbone and training settings otherwise match IMAGE2MUA. The three shuffle conditions operate on a trained IMAGE2MUA model without retraining (Figure 2). They swap the image term, reassign channel preferences, or perform both interventions while retaining the correct visual features. These are inference-time routing interventions rather than retrained one-factor ablations.

Area information and loss controls. The no-region-embedding control is a one-factor ablation that removes only the region vector from the response-token sum. It retains channel embeddings, the factorized router, source experts, splits, optimizer, checkpoint rule, area-balanced loss, and evaluation noise banks. The region-embedding parameter remains in the checkpoint only for key compatibility and is absent from the forward graph. An additional loss-only control retains the complete forward graph, including the cortical-area embedding and all routing inputs, and replaces the equally weighted mean of the V1, V4, and IT MSEs with an MSE averaged over the batch and all selected channels. It is evaluated on matched seeds 42–43 in both animals.

Figure 5 uses a separate seed-42 region blind control. This control removes the region vector and all other brain-region structural inputs and replaces the area-balanced loss with a mean over the batch and all channels. True area labels enter neither the model nor the loss and are attached only for post

hoc analysis. Because both the area-derived inputs and loss aggregation change, the region blind control is not a one-factor ablation.

Mean prediction baselines. Table 1 additionally includes ridge readouts on 512 grayscale PCA components and on final DINOv2 and CLIP embeddings, together with a four-layer Sensorium-style CNN using a full Gaussian readout. Ridge strengths and neural checkpoints are selected on the class-disjoint validation split; test responses are used only for reliability-based channel selection and frozen evaluation. The TBE-style fMRI transformer (Adeli et al., 2025) is ported to each animal as a deterministic-output reference under the same split and validation-only checkpoint rule. Table 1 uses its runs for seeds 42 and 43 in both animals. Because TBE predicts only a mean response, no CRPS result is reported. These independently trained models are comparison baselines rather than ablations of IMAGE2MUA.

Diffusion formulation control. The auxiliary DDPM condition replaces rectified flow training with a 1,000-step epsilon-prediction objective (Ho et al., 2020). Its linear noise schedule runs from $\beta_1 = 10^{-4}$ to $\beta_T = .005$, and training uses batch size 64 and learning rate 5×10^{-4}. At evaluation, it uses deterministic DDIM sampling (Song et al., 2021) with $\eta = 0$ and 50 steps. Because the objective, noise construction, and sampler all change, this condition evaluates the diffusion formulation as a whole rather than one factor. Equal step counts do not imply equal sampling cost.

Table 4 reports MCP for the common seed pair 42–43 and CRPS and NC similarity across all five frozen runs. Table 5 reports the corresponding variance ratio and correlation-matrix mean absolute error (MAE). For comparability, the condition retains the same hybrid visual source bank, dense visual expert router, response width, and frozen backbones as IMAGE2MUA, and its distributional evaluation uses the same IID30 noise bank.

Mean response metric. MCP is computed using Eq. 4. It uses 30 generated responses arranged as 15 Gaussian initial states and their exact negatives, reducing Monte Carlo error in the generated mean.

Marginal distribution metrics. CRPS uses 30 independent generated responses and all 30 recorded repetitions in the empirical plug-in statistic of Eq. 5. For seed s and region r, the average variance ratio is

$$R_{\mathrm{var}}^{(s,r)} = \frac{\mathrm{mean}_{(i,c)\in r}\, \widehat{\mathrm{Var}}_k(\hat{y}_{i,c}^{(k)})}{\mathrm{mean}_{(i,c)\in r}\, \widehat{\mathrm{Var}}_m(y_{i,c}^{(m)})}. \tag{7}$$

Here, a value of one is ideal, and both variances are unbiased sample estimates from the IID30 generations and recorded repetitions. For both CRPS and the variance ratio, we first average over images and channels in the requested region, compute the result separately for each seed, and report the five-seed mean and sample standard deviation.

CRPS controls and references. The mean-collapse control replaces each image–channel ensemble with a point mass at the mean of the same 30 samples. It therefore preserves the predicted mean exactly, and its CRPS reduces to the mean absolute error against the 30 recorded responses.

For the GT LOO reference, each recorded trial is held out in turn and scored using Eq. 5 against the other 29 repetitions. We average over the 30 held-out folds and then over images and channels using the same regional aggregation as model CRPS. This fixed, data-derived reference has no model-seed standard deviation. Because it uses repeated responses from the evaluated image, it is neither an image-only prediction baseline nor a formal performance bound. Using full-precision, overall channel-weighted aggregates, we summarize the model's position between the two references as

$$G = \frac{C_{\mathrm{collapse}} - C_{\mathrm{model}}}{C_{\mathrm{collapse}} - C_{\mathrm{LOO}}}. \tag{8}$$

Here, C_{model} and C_{collapse} are five-seed means, whereas C_{LOO} is the fixed recorded-repeat value.

The primary model evaluation uses 30 forecast samples, whereas LOO uses 29. For sensitivity, let L_{fair} be half the mean absolute difference over distinct recorded-trial pairs within each image–channel pair, followed by the same overall aggregation. The corresponding finite-ensemble plug-in references

are $L_{V,29} = 30L_{\text{fair}}/29$ and $L_{V,30} = 31L_{\text{fair}}/30$; the latter is the expected, bias-matched reference for an independent 30-member forecast. Replacing C_{LOO} with $L_{V,30}$ changes the overall reference CRPS by less than 4.3×10^{-4} and G by at most .0015.

Population dependence metrics. We subtract each image–channel's 30-trial mean and pool the residuals across the 100 repeated images. Noise correlation is the cross-channel Pearson correlation matrix of these pooled residuals. We report the Pearson similarity and mean absolute error between the upper triangles of the recorded and generated matrices, $E_{\text{NC}} = \text{mean}_{c<c'} |\widehat{R}_{c,c'} - R_{c,c'}|$. The channel-wise trial-shuffle control independently permutes trial indices within every image–channel pair. It preserves the 30 generated values of every empirical marginal, including its CRPS and variance, while removing trial alignment across channels.

Routing visualization. Effective four-source probabilities are averaged over the same 100 images for each checkpoint. Figure 5 shows every retained channel from the seed-42 main and region blind models. Within each animal and model, the all-channel mean for each source is subtracted from each channel's probability and the difference is expressed in percentage points. The heatmaps therefore show relative channel preferences rather than global source prevalence. Channels are grouped by cortical area only after route extraction. Panel (a) averages the region blind profiles within each area; panel (b) displays all source centered channel profiles for both models.

Routing statistics. Main–region blind profile similarity uses the within-channel CLR transform defined in Section 4, not the across-channel source centering used for display. We concatenate the four CLR values for each channel in a fixed order and correlate the resulting vectors. The statistic retains population-wide expert preferences; it is not a correlation of the source centered profiles shown in Figure 5.

For cortical-area organization, let $\boldsymbol{w}_c = \boldsymbol{p}_c^{(42)}$ denote channel c's untransformed seed-42 probability profile. Multivariate area variance is

$$R^2_{\text{area}} = \frac{\sum_a n_a \|\bar{\boldsymbol{w}}_a - \bar{\boldsymbol{w}}\|_2^2}{\sum_c \|\boldsymbol{w}_c - \bar{\boldsymbol{w}}\|_2^2}. \tag{9}$$

C Architecture Details

Visual token banks. ResNet feature maps receive learned two-dimensional positions and DINOv2 patch tokens retain their token order. Each source has an independent projection and cross-attention module. The E0–E3 order is an empirically chosen model basis rather than a prespecified or universal neurobiological depth axis.

Visual expert routing. The channel router is a directly indexed $C \times 4$ parameter table. The image router pools visual tokens and returns four image logits. "Dense visual expert routing" refers specifically to evaluating and mixing all four source-specific cross-attention experts described in Section 3.2; this is the router analyzed in the paper.

Response backbone and FFN MoE. The response backbone contains global self-attention, the dense four-source visual mixture, and a feed-forward mixture with six routed FFN experts, top-2 selection, and one always-active shared FFN. This internal FFN MoE supplies capacity, is fixed across all visual-source comparisons, and is distinct from visual expert routing. Its auxiliary loss-free bias is updated from accumulated expert-selection counts and affects top-2 selection, but not mixture weights or the training objective.

Training, control, and sampling. The response backbone and routers are trained from scratch; only the ResNet-18 and DINOv2 visual backbones are pretrained and frozen. In the five-seed embedding-only control, response-token construction skips the region-embedding addition; the factorized channel–image routing rule is unchanged and never reads the stored region-index buffer. Sampling uses a 50-step Heun solver.

At inference, we integrate the learned velocity field backward in time. We use $\boldsymbol{s}_k$ to denote the numerical sampling trajectory, distinguishing it from the training state $\boldsymbol{y}_t$. Starting from $\boldsymbol{s}_0 \sim$

$\mathcal{N}(0, I_C)$ at $t_0 = 1$, we set $t_k = 1 - k/50$ and $\Delta t = t_{k+1} - t_k = -1/50$ for $k = 0, \ldots, 49$. Each Heun step applies the predictor and corrector

$$\begin{aligned} \widehat{\boldsymbol{s}}_{k+1} &= \boldsymbol{s}_k + \Delta t\, v_\theta(\boldsymbol{s}_k, t_k, \boldsymbol{x}), \\ \boldsymbol{s}_{k+1} &= \boldsymbol{s}_k + \frac{\Delta t}{2}\left[v_\theta(\boldsymbol{s}_k, t_k, \boldsymbol{x}) + v_\theta(\widehat{\boldsymbol{s}}_{k+1}, t_{k+1}, \boldsymbol{x})\right]. \end{aligned} \tag{10}$$

After 50 steps, $\boldsymbol{s}_{50}$ is one simultaneous C-channel response sample at $t_{50} = 0$. Repeating this procedure with independent initial noise produces additional trials. Because the velocity field is predicted jointly across channels using population self-attention, the generated trials can express coordinated variability across V1, V4, and IT.

D Expanded Five-Seed Results

Table 4: **Same-architecture generative-formulation ablation.** The DDPM condition retains the visual-source bank, routing module, response backbone, data split, and evaluation banks, but jointly changes the generative objective, noise construction, and sampler. It is therefore an auxiliary formulation ablation rather than an external baseline or a one-factor objective ablation. MCP entries are means over the common seeds 42–43; CRPS and NC similarity are mean ± SD over seeds 41–45.

Animal	Metric	Formulation	Overall	V1	V4	IT
Monkey F	MCP ↑	Rectified flow	.87	.92	.85	.79
		DDPM–DDIM-50	.84	.88	.82	.75
	CRPS ↓	Rectified flow	.434 ± .006	.418 ± .006	.428 ± .005	.480 ± .009
		DDPM–DDIM-50	.462 ± .005	.454 ± .006	.453 ± .005	.490 ± .003
	NC sim. ↑	Rectified flow	.814 ± .009	.887 ± .010	.871 ± .010	.849 ± .008
		DDPM–DDIM-50	.770 ± .015	.844 ± .012	.868 ± .007	.841 ± .007
Monkey N	MCP ↑	Rectified flow	.89	.92	.88	.84
		DDPM–DDIM-50	.83	.86	.82	.75
	CRPS ↓	Rectified flow	.400 ± .001	.381 ± .002	.415 ± .003	.431 ± .004
		DDPM–DDIM-50	.454 ± .002	.445 ± .003	.452 ± .003	.481 ± .004
	NC sim. ↑	Rectified flow	.775 ± .017	.874 ± .011	.743 ± .018	.863 ± .009
		DDPM–DDIM-50	.728 ± .006	.820 ± .007	.707 ± .014	.865 ± .006

Table 5: **Variance ratios and residual-correlation errors.** Correlation MAE compares the same recorded and generated residual-correlation entries as noise-correlation similarity; lower is better. The ideal variance ratio is one. DDPM–DDIM-50 is the same auxiliary formulation ablation as in Table 4. CW shuffle has the same variance ratio as IMAGE2MUA by construction, whereas mean collapse has zero generated variance, so these redundant entries are omitted. Values are mean ± SD over five seeds.

Animal	Metric	Entry	Overall	V1	V4	IT
Monkey F	Corr. MAE ↓	IMAGE2MUA	.051 ± .002	.050 ± .002	.076 ± .008	.053 ± .006
		DDPM–DDIM-50	.059 ± .002	.062 ± .002	.079 ± .003	.056 ± .002
		CW shuffle	.083 ± <.001	.098 ± <.001	.125 ± <.001	.098 ± <.001
	Var. ratio (ideal 1)	IMAGE2MUA	1.07 ± .03	1.04 ± .04	1.13 ± .03	1.10 ± .03
		DDPM–DDIM-50	.97 ± .04	.94 ± .04	1.04 ± .05	.99 ± .04
Monkey N	Corr. MAE ↓	IMAGE2MUA	.081 ± .003	.081 ± .005	.096 ± .006	.105 ± .011
		DDPM–DDIM-50	.097 ± .001	.106 ± .003	.110 ± .002	.106 ± .005
		CW shuffle	.127 ± <.001	.153 ± <.001	.151 ± <.001	.197 ± <.001
	Var. ratio (ideal 1)	IMAGE2MUA	1.05 ± .02	1.02 ± .03	1.08 ± .02	1.06 ± .05
		DDPM–DDIM-50	.94 ± .04	.93 ± .04	.95 ± .03	.95 ± .03

Table 6: **Overall five-seed visual-source and routing controls.** MCP uses antithetic 15+15 sample averaging, whereas CRPS uses IID30 generation banks. Values are mean ± sample SD over five seeds; higher MCP and lower CRPS are better.

	Monkey F		Monkey N	
Model	MCP ↑	CRPS ↓	MCP ↑	CRPS ↓
E0: ResNet conv1	.8379 ± .0013	.4348 ± .0007	.8598 ± .0020	.4067 ± .0010
E1: ResNet layer2	.8475 ± .0033	.4403 ± .0020	.8801 ± .0017	.4029 ± .0008
E2: DINO block 5	.8569 ± .0032	.4373 ± .0021	.8842 ± .0028	.4033 ± .0016
E3: DINO block 11	.6956 ± .0041	.4895 ± .0029	.7185 ± .0032	.4689 ± .0025
All DINO	.8626 ± .0020	.4373 ± .0027	.8841 ± .0055	.4036 ± .0065
All ResNet	.8624 ± .0018	.4345 ± .0014	.8893 ± .0051	.3990 ± .0044
Global Static	.8746 ± .0026	.4311 ± .0019	.8881 ± .0029	.4048 ± .0033
No area embedding (area-balanced loss)	.8721 ± .0023	.4306 ± .0015	.8847 ± .0056	.4038 ± .0039
IMAGE2MUA	.8708 ± .0032	.4341 ± .0059	.8917 ± .0034	.4001 ± .0014

Table 6 compares the complete model with the visual source and routing controls under the same evaluation settings. The separate five-seed embedding-only control remains close to the main model on MCP, CRPS, and NC similarity while retaining the area-balanced loss. Its MCP and CRPS values are reported in Table 6. For Monkey F, its IID30 NC similarity is .8167 ± .0098, compared with .8140 ± .0092 for IMAGE2MUA. For Monkey N, it is .7770 ± .0217, compared with .7750 ± .0172 for IMAGE2MUA. All values are mean ± sample SD over seeds 41–45.

Relative to Global Static, the full factorized router lowers MCP by .0038 and raises CRPS by .0030 in Monkey F, but raises MCP by .0036 and lowers CRPS by .0047 in Monkey N. Across the five paired seeds, the routing effect is not statistically significant for Monkey F but is significant for Monkey N.

Supplementary discussion. The retained channels in Monkey F have lower reliability (mean .5530, median .5426) than those in Monkey N (mean .6278, median .6482), indicating lower data quality. One hypothesis is that uncertainty in electrode placement and localization following surgery produced incorrect cortical area labels for some Monkey F channels, which could weaken the learned area-related routing pattern. This explanation remains to be tested.

E PAIRED TWO-SEED LOSS-AGGREGATION CONTROL

Table 7: **Loss-only aggregation control.** The all-channel mean-loss condition retains the complete IMAGE2MUA architecture and training settings, including the cortical-area embedding, and changes only the flow-matching loss aggregation from the equally weighted mean of the V1, V4, and IT MSEs to an MSE averaged over the batch and all selected channels. Evaluation uses the same frozen observations and noise banks. Values are mean ± sample SD over matched seeds 42–43. Δ is all-channel mean loss minus the matched main model; negative ΔMCP and positive ΔCRPS indicate degradation.

Animal	Metric	Main model	All-channel mean loss	Δ
Monkey F	MCP ↑	.8719 ± .0007	.8670 ± .0035	−.0049
	CRPS ↓	.4287 ± .0007	.4319 ± .0020	+.0032
Monkey N	MCP ↑	.8930 ± .0058	.8843 ± .0026	−.0087
	CRPS ↓	.3986 ± .0007	.4025 ± .0040	+.0039

For each of the four animal–seed combinations, the matched main checkpoint has higher MCP and lower CRPS. With only two seeds, this controlled loss ablation suggests a consistent direction of change but does not establish statistical significance.

F ADDITIONAL VISUALIZATIONS AND ROUTING INTERVENTIONS

This section preserves the complete versions of results compressed in the main paper. Figure 6 extends the fixed-decoder visualization across stimuli and animals; Figure 7 visualizes pooled response distributions alongside complete trial heatmaps; and Table 8 expands the frozen routing interventions by cortical area.



Figure 6: **Qualitative downstream readout of generated population means.** Nine repeated images shared by Monkey F and Monkey N. The top row shows each stimulus, followed by reconstructions from recorded and generated MUA for Monkey F (606 channels) and Monkey N (722 channels). Recorded responses are averaged over 30 trials; generated responses are the mean of 30 seed-42 IID samples. Each animal has a fixed reconstruction pipeline, with decoder and stochastic reconstruction settings matched only within each recorded–generated pair. Examples were selected post hoc for qualitative visualization.

Table 8: **Full area-wise frozen routing interventions.** Five-seed paired changes (shuffled minus full); negative ΔMCP and ΔNC similarity and positive ΔCRPS indicate degradation.

Animal	Shuffled	Metric	Overall	V1	V4	IT
Monkey F	Image	ΔMCP	$-.0014$	$-.0006$	$-.0019$	$-.0030$
		ΔCRPS	$+.0003$	$+.0000$	$+.0006$	$+.0005$
		ΔNC sim.	$+.00019$	$+.00009$	$+.00018$	$+.00024$
	Channel	ΔMCP	$-.0075$	$-.0092$	$-.0041$	$-.0067$
		ΔCRPS	$+.0040$	$+.0044$	$+.0027$	$+.0044$
		ΔNC sim.	$-.00005$	$-.00002$	$-.00023$	$+.00056$
	Both	ΔMCP	$-.0092$	$-.0099$	$-.0068$	$-.0097$
		ΔCRPS	$+.0043$	$+.0045$	$+.0036$	$+.0047$
		ΔNC sim.	$+.00012$	$+.00008$	$-.00008$	$+.00078$
Monkey N	Image	ΔMCP	$-.0006$	$+.0001$	$+.0000$	$-.0035$
		ΔCRPS	$-.0001$	$+.0000$	$-.0004$	$+.0004$
		ΔNC sim.	$+.00067$	$+.00007$	$+.00149$	$+.00066$
	Channel	ΔMCP	$-.0048$	$-.0059$	$-.0061$	$+.0001$
		ΔCRPS	$+.0041$	$+.0035$	$+.0070$	$+.0014$
		ΔNC sim.	$+.00229$	$+.00312$	$-.00182$	$-.00023$
	Both	ΔMCP	$-.0052$	$-.0057$	$-.0061$	$-.0027$
		ΔCRPS	$+.0040$	$+.0035$	$+.0069$	$+.0016$
		ΔNC sim.	$+.00286$	$+.00316$	$-.00078$	$+.00038$

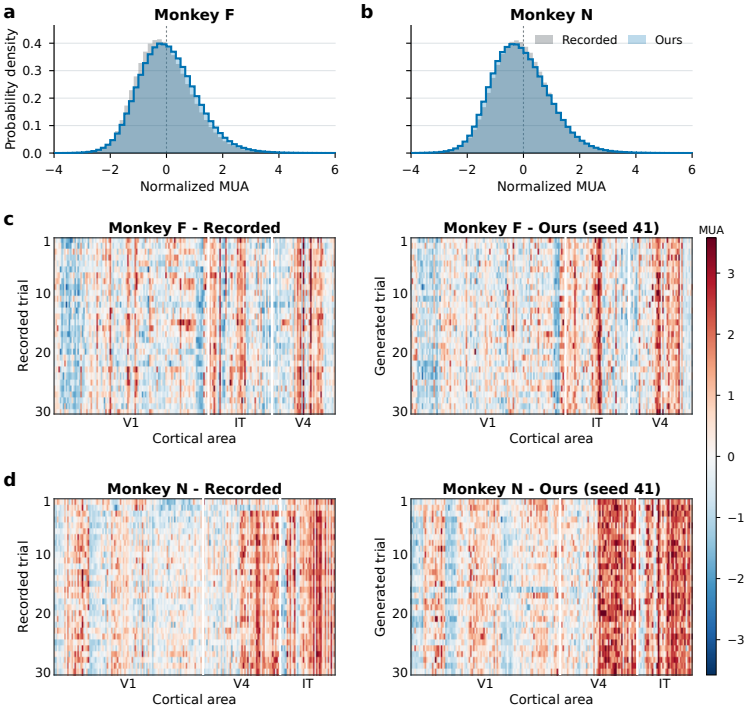

Figure 7: **Full population samples and pooled response distributions.** (a,b) Recorded and seed-41 generated MUA pooled for Monkeys F and N. (c,d) All 30 recorded trials and 30 seed-41 IID samples for the same horse stimulus (zero-based index 050); the independent rows show both high-response and low or suppressed channels. Each histogram combines 100 images, 30 trials, and all retained channels; separately normalized densities use shared bins over $[-4, 6]$, contain more than 99.98% of values, and have histogram intersections ($\sum_b \min(p_b, q_b)\Delta_b$) of .948 and .974. Channels retain native array order (F: V1–IT–V4; N: V1–V4–IT), white lines mark area boundaries, and all heatmaps use the joint 99.5th percentile limit of ± 3.577 without trial centering.

G Complementary Single-Source Results

The following single-source comparisons are complementary to, but not used to construct, the channel-routing analysis in Figure 5. They are paired by training seed. For MCP, the preferred-source margins are E0−E2 in V1 and E2−E0 in V4 and IT. For Monkey F, the MCP margins in (V1, V4, IT) order are $(.0600, .0593, .1713)$, and the corresponding CRPS benefits are $(.0264, .0177, .0357)$. For Monkey N, the MCP margins in the same order are $(.0369, .0630, .1364)$, and the corresponding CRPS benefits are $(.0190, .0190, .0426)$. All six comparisons have the preferred sign in every one of the five seeds.