SAMPLE-ADAPTIVE EXPERT ROUTING FOR EEG VISUAL DECODING

Anonymous authors
Paper under double-blind review

ABSTRACT

EEG visual decoding analyzes the relationship between EEG and visual stimuli by aligning EEG to image. Because cortical responses are biased toward coarse, low-spatial-frequency content, recent work aligns EEG to blurred images that carry more of the visual information reflected in the EEG, but fixes the blur level and representation depth to a single global setting shared across all samples. This ignores the variability of EEG responses across subjects, stimuli, and samples, and hence the ideal degree of blur and level of abstraction is inherently sample-dependent. In this paper, we propose a novel sample-adaptive visual decoding method that composes the alignment target by conditioning on each EEG embedding. Building on a mixture-of-experts formulation, dynamic routers select, per sample, the blur level, encoder layer, and backbone of the visual representation, while two auxiliary objectives ground each expert in its own visual content and supervise the routers with the similarity between the EEG and each expert, respectively. Evaluated by zero-shot retrieval on two benchmarks, the proposed method achieves state-of-the-art performance, and our ablations confirm that the conditional routing and the two auxiliary objectives each contribute to the gains.

1 INTRODUCTION

Decoding perceived visual content from brain signals *i.e.*, visual neural decoding serves both as a means of understanding the representational mechanisms of the human visual system and as a major application of brain-computer interfaces (BCIs) Wolpaw et al. (2002); Palazzo et al. (2021). Among the non-invasive brain signals used for BCIs, electroencephalography (EEG) is portable and offers high temporal resolution, making it well suited for real-world deployment Wolpaw et al. (2002). Early studies Spampinato et al. (2017) formulated EEG-based visual decoding as a classification problem over a predefined set of categories. However, such approaches are restricted to the categories seen during training and are therefore difficult to scale to real-world settings given the diversity of visual concepts.

Recent studies Chen et al. (2024); Li et al. (2024) have explored the task as a zero-shot retrieval problem, in which EEG is aligned to an image and the most similar stimulus is retrieved from a candidate set of images. Since retrieval does not rely on a predefined label set, this formulation can handle the diverse visual concepts encountered in practical scenarios. To this end, the contrastive learning that aligns EEG to the features of a pretrained visual encoder has been widely adopted Zhang et al. (2026); Zheng et al. (2026); Tang et al. (2026); Liu et al. (2026); Wu et al. (2026). The choice of visual representation used as the alignment target governs retrieval performance. Early studies Song et al. (2024); Wei et al. (2024) aligned EEG to the final feature of the encoder. However, these features are extracted from sharp images and therefore align poorly with EEG responses, which reflect visual processing that initially prioritizes low-spatial-frequency (LSF) information Bar (2003); Kauffmann et al. (2014). To alleviate this, Liu et al. (2026) recently proposed extracting features from progressively blurred images and aligning EEG to a representation that combines them through learned weights. This exploits the high neural visibility of LSF content, which reflects how strongly a representation appears in the EEG response. This method, however, shares the combination weights over blur levels globally across all samples. In a separate direction, some studies fuse multiple pretrained encoders to combine complementary visual representations Zheng et al. (2026), and others select intermediate encoder layers to match the representation depth reflected in EEG Tang et al. (2026). Yet each of these methods addresses only one such choice and still fixes

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.

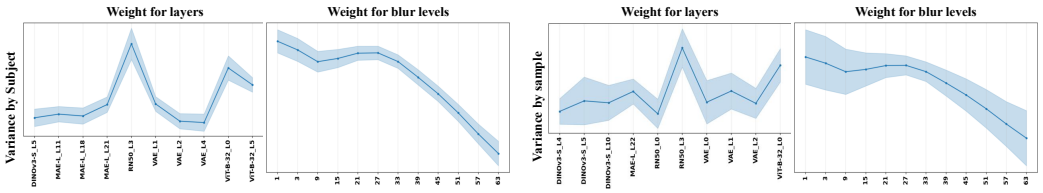

Figure 1: Weight variability in experts. Each line denotes the mean routing weight and the shaded band denotes the standard deviation. The top-10 image experts with the largest variance are shown.

the combination to a global setting shared across all samples. Because EEG responses vary across subjects Lotte et al. (2018), stimuli Cichy et al. (2014), and trials Blankertz et al. (2011), the ideal blur level and level of abstraction are inherently sample-dependent, and a global choice is therefore mismatched to many samples. As shown in Figure 1, the router indeed assigns non-uniform weights that vary across subjects and samples along both the layer and blur axes, showing that a single global weight cannot capture this variation.

To address this limitation, we propose Sample-AdaptiVe Expert Routing (SAVER) that selects the blur level, encoder layer, and backbone conditioned on each EEG sample. The suitable target depends on several factors, none of which can directly serve as a condition. Signal quality is unlabeled, and the stimulus is what must be retrieved. Subject identity misses stimulus-level variation. We therefore condition the target on the EEG, the only input that captures all of them. To this end, SAVER builds on a mixture-of-experts (MoE) structure Shazeer et al. (2017), in which each expert is defined as the image feature extracted at one combination of backbone, encoder layer, and blur level. Dynamic routers that take the EEG embedding as input predict the contribution of each expert in place of fixed weights. Because backbone, layer, and blur are unified into a single expert pool, the router jointly composes complementary visual factors such as semantics, texture, and spatial frequency without any change to the formulation. To keep each expert grounded in the visual content it represents, we further introduce expert alignment, which aligns every EEG expert with the visual content at its own blur level. In addition, to prevent routing from being learned only implicitly through the contrastive loss, we introduce router guidance, which uses the cosine similarity between the EEG embedding and each expert as the learning target for the router.

The main contributions of this work are as follows:

1. We propose a sample-adaptive visual decoding method that selects the alignment target for each sample. Conditioned on the EEG embedding, dynamic routers select the blur level, encoder layer, and backbone of the visual representation to which the EEG is aligned. Unlike existing methods that rely on a fixed weighting across samples, this draws each sample's target from a pool of hierarchical and complementary visual representations.
2. We introduce two auxiliary objectives, expert alignment and router guidance. Expert alignment ties each EEG expert to the visual content at its own blur level, while router guidance directly supervises the routers with the similarity between the EEG and each expert, together guiding stable adaptive selection.
3. Comprehensive experiments demonstrate that SAVER achieves state-of-the-art zero-shot retrieval performance. Our ablations verify the contribution of each component and show that the enlarged expert pool alone does not account for the gains, highlighting the necessity of dynamic routers that adapt the target to each sample.

2 Related Work

2.1 EEG Visual Decoding.

Decoding perceived visual stimuli from EEG is a central problem both for understanding the representational mechanisms of the human visual system and for realizing BCIs. Recent studies have addressed this problem as a retrieval task, where EEG is matched against a candidate set specified at inference time rather than a fixed label set, thereby covering diverse visual concepts. NICE Song et al. (2024) aligned EEG to the features of a frozen CLIP encoder with an InfoNCE objective,

enabling retrieval of unseen stimuli, and Li et al. (2024) combined this with guided diffusion to jointly perform retrieval and image reconstruction. Building on this framework, subsequent work has largely focused on refining the alignment mechanism itself. NeuroBridge Zhang et al. (2026) proposed a self-supervised approach using cognitive prior augmentation together with bidirectional alignment, while ATS Wu et al. (2026) focused on the fidelity and semantic gaps arising from the asymmetry between EEG and vision and shrank the vision teacher to the capacity of the EEG student. However, these methods have the limitation of fixing the alignment target to the final semantic feature of the pretrained model.

It is therefore sought to redesign the alignment target itself, questioning which visual representation the EEG should be aligned to. MBP Zheng et al. (2026) hierarchically fused multiple pretrained visual encoders to form an alignment target containing both high-level semantics and low-level detail, and HCF Tang et al. (2026) introduced the notion of neural visibility to align EEG to intermediate layer with the final layer. In particular, Liu et al. (2026) aligned EEG to features of multi-level Gaussian-blurred images, exploiting the coarse-to-fine nature of visual perception, in which low-spatial-frequency content is processed first. Nevertheless, they fix it to a global setting shared across all samples. In this work, we integrate blur level, encoder layer, and backbone into a single expert pool and dynamically select the alignment target for each EEG sample.

2.2 Mixture-of-Experts.

A mixture-of-experts (MoE) routes each input to different experts through a learned gating function, expanding a model's representational capacity via conditional computation. Jacobs et al. (1991) proposed an early form in which a softmax gate assigns inputs to local experts. Shazeer et al. (2017) later introduced top-k sparse routing with an auxiliary load-balancing loss, greatly expanding model capacity while keeping per-example computation bounded. This line was scaled further by the top-1 routing of Fedus et al. (2022), and in the vision domain Riquelme et al. (2021) replaced ViT FFN blocks with sparse MoE. More recently, studies have improved the balance and differentiability of routing. Zhou et al. (2022) let experts select tokens, achieving load balance without an auxiliary balancing loss, and Puigcerver et al. (2024) proposed a fully differentiable soft routing based on weighted combinations of tokens. In the multimodal setting, Mustafa et al. (2022) contrastively trained image and text within a single MoE.

These MoE models all aim to expand model capacity and encourage uniform expert usage through load balancing. In contrast, we use MoE not for parameter expansion but as a mechanism for per-sample selection of the visual representation across blur level, encoder layer, and backbone. To this end, we introduce router guidance, which uses the similarity between the EEG and each expert as the learning target of router, selectively activating experts conditioned on each EEG embedding.

3 Method

We propose Sample-AdaptiVe Expert Routing (SAVER), an EEG visual decoding method that recovers the visual information from EEG sample. SAVER aligns each EEG sample with a visual representation of that stimulus in a shared embedding space. As shown in Figure 2, rather than aligning every sample to a single fixed target, SAVER employs sample conditioned alignment target. To realize this adaptive selection, both the visual and EEG sides are represented as sets of experts. On the visual side, each image is expanded over combinations of backbone, layer, and blur, providing visual experts as candidate targets that differ in spatial frequency and representation depth. On the EEG side, the embedding is projected into multiple blur views corresponding to different levels of visual detail. These two sets are then combined by two EEG-conditioned routers. The image router produces a per-sample distribution over the visual experts, determining which combination of backbone, layer and blur the sample should align to. It then forms the alignment target by weighting the experts accordingly. The EEG router produces a distribution over the EEG experts and fuses them into a single EEG representation. The resulting EEG-image pair is aligned by contrastive learning.

3.1 Visual experts.

The visual experts form a set of features that represent each stimulus image under different perceptual conditions. Following Liu et al. (2026), who represent visual information hierarchically across

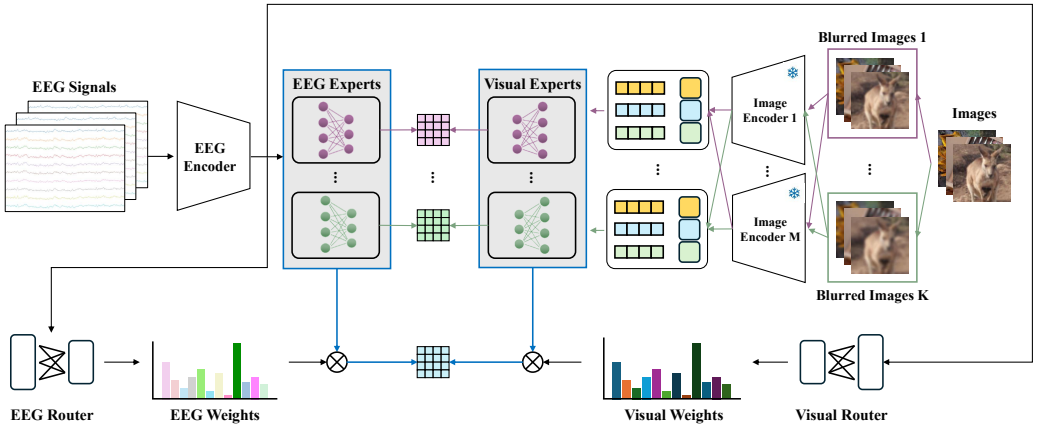

Figure 2: Overview of the proposed method. Each blurred image is expanded over combinations of backbone, layer, and blur to form a pool of visual experts. The EEG embedding is projected into a set of EEG experts for K blur levels. Two routers take the EEG embedding and produce distributions over the visual and EEG experts, whose weighted combinations form the alignment target and the fused EEG representation. The fused EEG representation and the alignment target are then matched by contrastive learning. Each EEG expert is aligned with the visual content at its own blur level.

multiple blur levels, we render each image at K Gaussian-blur levels, where we define I_b as the image at blur level $b \in \{1, \ldots, K\}$. Specifically, each I_b is passed through one of M backbones, and we define $g^{m,l}(I_b)$ as the feature obtained at layer $l \in \{1, \ldots, L_m\}$ of backbone $m \in \{1, \ldots, M\}$. Each combination of backbone m, layer l, and blur level b becomes a single expert indexed by $n \in \{1, \ldots, N\}$, where $N = K \sum_m L_m$, so the pool gathers complementary factors such as semantics, texture, and spatial frequency. Each feature is then projected into a shared d-dimensional space by a linear map $P_{m,l}$ as follows: $v_n = P_{m,l}\, g^{m,l}(I_b) \in \mathbb{R}^d$, and stacking the experts of each stimulus image gives $V = [v_1, \ldots, v_N] \in \mathbb{R}^{N \times d}$. V thus spans spatial frequency, representation depth, and backbone, offering representations from which to compose target of each sample.

3.2 EEG EXPERTS.

The EEG experts form a set of representations, each intended to capture the visual information reflected in the EEG at a different level of visual detail. An image I evokes an EEG sample $x \in \mathbb{R}^{C \times T}$ of C channels and T time points, from which we obtain an embedding $e = f_\theta(x) \in \mathbb{R}^d$ through an encoder f_θ. The visual information reflected in the EEG varies primarily along spatial frequency, since how clearly a stimulus is perceived determines how much high-spatial-frequency detail appears in the response Bar (2003); Kauffmann et al. (2014). To account for this variation, we map the EEG embedding into K EEG experts, each corresponding to a blur level b, using a separate projection head Q_b, which we define as $u_b = Q_b(e) \in \mathbb{R}^d$.

3.3 EEG-CONDITIONED ROUTING.

The two expert pools are recombined into a sample-specific pair of representations by two routers that take the same EEG embedding. To prevent the routers from disturbing the encoder and to make their decisions depend on the direction rather than the magnitude of e, each router takes the ℓ_2-normalized and stop-gradient embedding $\bar{e}$, where $\bar{\cdot}$ denotes ℓ_2 normalization throughout. An image router r^{I} produces a distribution over the N visual experts, and an EEG router r^{E} produces a distribution over the K EEG experts, which we define as

$$p = \mathrm{softmax}\big(r^{\mathrm{I}}(\bar{e})/\tau\big) \in \Delta^N, \quad w = \mathrm{softmax}\big(r^{\mathrm{E}}(\bar{e})/\tau\big) \in \Delta^K. \tag{1}$$

Each router is a two-layer MLP of hidden width $d/2$ with a GELU nonlinearity, and the two routers share the temperature τ. Both routers assign dense soft weights over all experts, so every expert contributes to representation of each sample in proportion to its weight.

We consider a batch of B pairs of EEG and image, where i indexes the query EEG sample and j the candidate image within the batch. The EEG side is combined into a single fused embedding $\hat{e} = \sum_{b=1}^{K} w_b u_b$, and the visual side is combined by the image routing into a single representation, which is then passed through a nonlinear adapter A_ψ and normalized. Since the routing is computed from the query, the visual target is query-specific. We therefore cross-route by applying the image routing of query i identically to every image in the batch, which we define as

$$a_{ij} = \sum_{n=1}^{N} p_n^{(i)} v_n^{(j)}, \quad h_{ij} = A_\psi(a_{ij}) / \|A_\psi(a_{ij})\|. \tag{2}$$

3.4 Contrastive retrieval objective.

The primary signal aligns each EEG sample with its matched image. Specifically, we pair fused EEG representation of each query with the cross-routed visual representation and score the pair by their inner product $s_{ij} = \langle \hat{e}^{(i)}, h_{ij} \rangle$. Taking the diagonal of the matrix as positives, we minimize InfoNCE loss in the EEG-to-image direction,

$$\mathcal{L}_{\mathrm{NCE}} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp(s_{ii})}{\sum_{j=1}^{B} \exp(s_{ij})}. \tag{3}$$

The reverse direction is not used, as each column of the matrix would compare representations formed under different routings.

3.5 Expert alignment.

The EEG experts are supervised to match the visual content at their own blur level. Averaging the visual experts over backbone and layer yields one image target per blur level, $\tilde{v}_b = \frac{1}{|\mathcal{N}_b|} \sum_{n \in \mathcal{N}_b} v_n$, where $\mathcal{N}_b$ is the set of N/K experts at blur level b. At each blur level, we score every EEG expert against the ℓ_2-normalized targets over the batch $\sigma_{ij}^b = \langle u_b^{(i)}, \bar{\tilde{v}}_b^{(j)} \rangle$. We align each EEG expert u_b to its own target $\bar{\tilde{v}}_b$ with an InfoNCE loss over the batch,

$$\ell^b = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp(\sigma_{ii}^b)}{\sum_{j=1}^{B} \exp(\sigma_{ij}^b)}. \tag{4}$$

The expert-alignment loss averages this over blur levels $\mathcal{L}_{\mathrm{EA}} = \frac{1}{K} \sum_{b=1}^{K} \ell^b$. This ties every EEG blur expert to the visual content at its blur level, regardless of how the routers later weight it.

3.6 Router guidance.

The contrastive loss scores only the final fused EEG and image representations, not the routing distributions p and w that produce them. As a result, the routers are supervised only indirectly, receiving a gradient diluted by the weighted combination rather than a direct signal on which experts to select. To address this, we add a direct target to each router, which we construct from the cross-branch cosine similarity between detached representations. The image router is guided toward the similarity of the fused EEG to each visual expert, and the EEG router toward the similarity of each EEG expert to the query's own fused image h_{ii}, which we define as

$$t^{\mathrm{I},(i)} = \mathrm{softmax}_n\big(\langle \bar{\hat{e}}^{(i)}, \bar{v}_n^{(i)} \rangle / \tau_{\mathrm{RG}}\big) \in \Delta^N, \quad t^{\mathrm{E},(i)} = \mathrm{softmax}_b\big(\langle \bar{u}_b^{(i)}, h_{ii} \rangle / \tau_{\mathrm{RG}}\big) \in \Delta^K, \tag{5}$$

where τ_{RG} is the target temperature and all quantities on the right-hand side are detached. Each router is then trained to match its target through a KL divergence over the batch, which we define as

$$\mathcal{L}_{\mathrm{RG}}^{\mathrm{I}} = \tfrac{1}{B} \sum_{i=1}^{B} \mathrm{KL}\big(t^{\mathrm{I},(i)} \| p^{(i)}\big), \quad \mathcal{L}_{\mathrm{RG}}^{\mathrm{E}} = \tfrac{1}{B} \sum_{i=1}^{B} \mathrm{KL}\big(t^{\mathrm{E},(i)} \| w^{(i)}\big), \tag{6}$$

and the two terms are summed into $\mathcal{L}_{\mathrm{RG}} = \mathcal{L}_{\mathrm{RG}}^{\mathrm{I}} + \mathcal{L}_{\mathrm{RG}}^{\mathrm{E}}$.

3.7 Overall objective.

We define the full training objective as the weighted sum $\mathcal{L} = \mathcal{L}_{\mathrm{NCE}} + \gamma_{\mathrm{EA}} \mathcal{L}_{\mathrm{EA}} + \gamma_{\mathrm{RG}} \mathcal{L}_{\mathrm{RG}}$.

4 EXPERIMENTAL SETTINGS

4.1 DATASET AND PREPROCESSING.

We evaluate on two datasets, THINGS-EEG2 Gifford et al. (2022) and THINGS-MEG Hebart et al. (2023), both recording neural responses to natural object images. For both, we epoch each sample to the 0 to 1000 ms post-stimulus window and average the repeated presentations of the same image. The training set holds images from 1,654 concepts, and the test set holds 200 images of novel concepts that form the retrieval gallery. Unless stated otherwise we operate within subject, holding out 5% of the training images per subject as a validation split and evaluating on that subject's test set under the 200-way zero-shot retrieval protocol. THINGS-EEG2 records 64-channel EEG from ten subjects resampled to 250 Hz with four repetitions per image in the training set and eighty in the test set. THINGS-MEG records 271-channel MEG from four subjects resampled to 200 Hz with one repetition per image in the training set and twelve in the test set. We reuse the visual feature pipeline unchanged across both and the neural encoder is identical apart from its spatial convolution whose kernel spans all input channels. Following Liu et al. (2026), each image is rendered at K=12 Gaussian-blur levels before feature extraction. The blur is a Gaussian kernel applied to the RGB image, with kernel sizes $\{1, 3, 9, 15, 21, 27, 33, 39, 45, 51, 57, 63\}$.

4.2 IMPLEMENTATION DETAILS.

We employ five pretrained image backbones, namely CLIP RN50 and ViT-B-32 Radford et al. (2021), DINOv3-S Siméoni et al. (2025), MAE-L He et al. (2022), and SDXL-VAE Podell et al. (2024), and extract ℓ_2-normalized features at multiple depths from each of them. For CLIP ViT, we take the patch-token mean of every residual block, and for CLIP ResNet, the pooled output of each of the four stages, adding the final image embedding in both cases. For DINOv3 and MAE, we take the class token of every transformer block, and for the VAE, we take the pooled down and mid blocks together with the flattened latent, resizing its input to 128×128. This yields 60 features per blur level and thus 720 experts in total, each of which is linearly projected to a shared dimension of 1024. For the EEG encoder, we adopt the same architecture as Liu et al. (2026) to demonstrate the effectiveness of our routing. Both routers are two linear layers of hidden width 512 with a GELU nonlinearity, the image router producing a distribution over the 720 visual experts and the EEG router over the 12 blur-level experts. The adapter A_ψ consists of two linear layers that project the 1024-dimensional input to a 768-dimensional hidden layer and back to 1024 dimensions, with an ELU function. We train with AdamW Loshchilov & Hutter (2019) using a weight decay of 0.01, a learning rate of 10^{-3}, and a batch size of 512. We first train the EEG encoder and both experts for 100 epochs under uniform averaging with expert alignment, and then train only the routers, EEG experts, and adapter for 50 epochs to learn a per-sample composition of the fixed representations. For evaluation, we select the checkpoint with the best validation Top-1 accuracy. We set $\tau = \tau_{RG} = 0.1$ and $\gamma_{EA} = \gamma_{RG} = 1.0$. All experiments were conducted on a single NVIDIA RTX A5000 GPU. Each experiment was repeated three times with different random seeds, reporting the mean accuracy.

4.3 BASELINES AND EVALUATION PROTOCOL.

We compare against 11 EEG visual decoding methods such as BraVL Du et al. (2023), NICE Song et al. (2024), ATM-S Li et al. (2024), MB2C Wei et al. (2024), CognitionCapturer Zhang et al. (2025), UBP Wu et al. (2025), ATS Wu et al. (2026), NeuroBridge Zhang et al. (2026), MBP Zheng et al. (2026), Liu et al. (2026) and HCF Tang et al. (2026). We report zero-shot retrieval accuracy over the 200-image test gallery. Given a query EEG, the gallery images are ranked by the cross-routed similarity of Eq. 2, applying the query's own routing to every gallery image, and we measure Top-1 and Top-5 accuracy.

5 EXPERIMENTAL RESULTS

5.1 MAIN RESULTS.

Table 1 reports 200-way zero-shot retrieval on THINGS-EEG2 in the intra-subject setting. SAVER outperforms all comparison methods at both Top-1 and Top-5. SAVER attains the best Top-1 on nine

Table 1: 200-way zero-shot retrieval performance on THINGS-EEG2 in intra-subject setting.

Method		Sub1	Sub2	Sub3	Sub4	Sub5	Sub6	Sub7	Sub8	Sub9	Sub10	Avg
BraVL	Top-1	6.1	4.9	5.6	5.0	4.0	6.0	6.5	8.8	4.3	7.0	5.8
	Top-5	17.9	14.9	17.4	15.1	13.4	18.2	20.4	23.7	14.0	19.7	17.5
NICE	Top-1	13.3	12.1	15.3	15.9	9.8	14.2	17.9	18.2	14.4	16.0	14.7
	Top-5	40.2	36.1	39.6	49.0	34.4	42.4	43.6	50.2	38.7	42.8	41.7
ATM-S	Top-1	25.6	22.0	25.0	31.4	12.9	21.3	30.5	38.8	34.4	29.1	27.1
	Top-5	60.4	54.5	62.4	60.9	43.0	51.1	61.5	72.0	51.5	63.5	58.1
MB2C	Top-1	23.7	22.7	26.3	34.8	21.3	31.0	25.0	39.0	27.5	33.2	28.4
	Top-5	56.3	50.5	60.2	67.0	53.0	62.3	54.8	69.3	59.3	70.8	60.4
CognitionCapturer	Top-1	27.2	28.7	37.2	37.7	21.8	31.6	32.8	47.6	33.4	35.1	33.3
	Top-5	59.5	57.0	66.1	63.2	47.8	58.1	59.6	73.5	57.6	63.6	60.6
UBP	Top-1	41.2	51.2	51.2	51.1	42.2	57.5	49.0	58.6	45.1	61.5	50.9
	Top-5	70.5	80.9	82.0	76.9	72.8	83.5	79.9	85.8	76.2	88.2	79.7
ATS	Top-1	53.0	62.0	61.5	57.0	55.0	68.0	53.0	66.5	58.5	67.5	60.2
	Top-5	79.0	87.5	89.0	86.5	84.0	90.5	84.0	91.0	86.0	89.0	86.7
NeuroBridge	Top-1	50.5	63.2	61.6	61.4	54.8	69.7	62.7	71.2	64	73.6	63.2
	Top-5	77.6	90.6	91.1	90.0	85.0	92.9	88.8	95.1	91.0	97.1	89.9
MBP	Top-1	64.3	76.3	74.0	67.0	68.0	81.5	76.8	84.8	76.8	87.3	75.7
	Top-5	88.8	95.3	95.0	91.8	91.5	96.3	96.8	98.5	95.8	99.3	94.6
Liu et al.	Top-1	81.9	81.7	78.3	76.9	71.4	84.3	78.2	84.5	78.7	84.1	80.0
	Top-5	96.8	96.8	96.0	97.2	93.7	98.7	96.7	98.1	96.8	98.1	96.9
HCF	Top-1	81.9	86.0	84.3	82.2	74.5	90.4	82.8	**90.9**	79.7	93.1	84.6
	Top-5	98.3	98.9	97.7	**98.8**	94.1	98.9	97.2	99.4	98.4	99.8	98.2
SAVER	Top-1	**91.0**	**93.0**	**90.0**	**87.0**	**86.2**	**93.8**	**90.7**	90.3	**91.0**	**93.5**	**90.7**
	Top-5	**98.3**	**99.8**	**99.2**	98.7	**98.2**	**100.0**	**99.7**	**100.0**	**99.7**	**100.0**	**99.4**

of the ten subjects and trails only on Sub8. The gains over HCF are concentrated on the subjects for which HCF performs worst, improving by 11.7 and 11.3 points on Sub5 and Sub9, respectively, and the standard deviation across subjects decreases from 5.4 to 2.4. This suggests that dynamic routers primarily benefit the cases where a single global target is mismatched. Across all $2,000$ test queries from the ten subjects, only average 13 fail to recover the correct stimulus within the Top-5, showing that when the correct stimulus is not ranked first it is almost always found within the Top-5.

Table 2 evaluates the inter-subject setting under leave-one-subject-out validation. Absolute accuracy falls sharply for all methods, reflecting the large cross-subject distribution shift of EEG. SAVER again achieves the highest average accuracy at both Top-1 and Top-5, and it holds the best Top-1 on nine of the ten subjects. Relative to the strongest baseline HCF, the improvement ratio at Top-1 rises from 7.2% in the intra-subject setting to 23.1% here. This indicates that conditioning helps more when the responses differ from those seen in training, since a single global target cannot accommodate responses drawn from a different distribution.

5.2 Evaluation on different modality.

Table 3 evaluates the same method on THINGS-MEG in the intra-subject setting, testing whether the routing formulation generalizes to another neuroimaging modality. SAVER outperforms all baselines at both Top-1 and Top-5, and it attains the best Top-1 on three of the four subjects. The improvement ratio over the strongest baseline Liu et al. (2026) at Top-1 is 10.9%, obtained without modality-specific tuning, as the visual expert pool and all hyperparameters are reused and only the spatial convolution of the neural encoder is adapted to the channel count. This suggests that the benefit of conditioning arises from the sample-to-sample variability of neural responses in general rather than from properties specific to EEG.

5.3 Effect of Conditional Routing.

Table 4 isolates the effect of making the alignment target conditional on the EEG by varying only the router while holding the expert pool fixed, so that all three schemes draw on the same large expert pool. Note that both uniform and static schemes also use the expert alignment. Router guidance does not apply to either, since uniform combination has no router and a single global weight cannot represent a per-sample target. Uniform averaging and static global weighting show similar performance, indicating that a fixed combination gains little from learning a single shared weight. These two fixed schemes underperform the strongest fixed-target baseline HCF, even though they

Table 2: 200-way zero-shot retrieval performance on THINGS-EEG2 in inter-subject setting.

Method		Sub1	Sub2	Sub3	Sub4	Sub5	Sub6	Sub7	Sub8	Sub9	Sub10	Avg
BraVL	Top-1	2.3	1.5	1.4	1.7	1.5	1.8	2.1	2.2	1.6	2.3	1.8
	Top-5	8.0	6.3	5.9	6.7	5.6	7.2	8.1	7.6	6.4	8.5	7.0
NICE	Top-1	7.6	5.9	6.0	6.3	4.4	5.6	5.6	6.3	5.7	8.4	6.2
	Top-5	22.8	20.5	22.3	20.7	18.3	22.2	19.7	22.0	17.6	28.3	21.4
ATM-S	Top-1	10.5	7.1	11.9	14.7	7.0	11.1	16.1	15.0	4.9	20.5	11.8
	Top-5	26.8	24.8	33.8	39.4	23.9	35.8	43.5	40.3	22.7	46.5	33.7
MB2C	Top-1	10.5	11.3	8.8	13.7	10.7	12.2	11.5	12.0	12.2	16.2	11.9
	Top-5	28.2	32.8	27.7	33.5	27.5	33.2	31.8	32.2	31.3	42.2	32.0
UBP	Top-1	11.5	15.5	9.8	13.0	8.8	11.7	10.2	12.2	15.5	16.0	12.4
	Top-5	29.7	40.0	27.0	32.3	33.8	31.0	23.8	32.2	40.5	43.5	33.4
NeuroBridge	Top-1	23.2	21.2	13.2	17.0	14.5	25.0	15.3	20.1	13.7	27.2	19.0
	Top-5	52.4	49.3	36.5	45.3	37.7	55.0	45.1	44.9	36.5	56.3	45.9
Liu et al.	Top-1	27.3	32.0	11.1	18.6	16.9	16.1	16.0	18.3	13.6	30.1	20.0
	Top-5	54.9	63.0	34.5	46.0	40.7	43.2	45.4	49.0	41.2	62.1	48.0
HCF	Top-1	28.5	22.5	20.0	20.5	18.5	**29.5**	20.0	20.0	21.0	34.0	23.4
	Top-5	61.0	56.5	45.5	54.0	46.5	**60.5**	53.5	53.5	48.5	69.0	54.9
SAVER	Top-1	**37.3**	**32.3**	**21.7**	**26.0**	**20.3**	26.5	**33.3**	**24.7**	**23.5**	**42.3**	**28.8**
	Top-5	**73.2**	**68.7**	**50.2**	**56.3**	**50.7**	58.3	**67.8**	**54.5**	**50.8**	**78.2**	**60.9**

Table 3: 200-way zero-shot retrieval performance on THINGS-MEG in intra-subject setting.

Method		Sub1	Sub2	Sub3	Sub4	Avg
NICE	Top-1	9.6	18.5	14.2	9.0	12.8
	Top-5	27.8	47.8	41.6	26.6	36.0
UBP	Top-1	15.0	46.0	27.3	18.5	26.7
	Top-5	38.0	80.5	59.0	43.5	55.2
NeuroBridge	Top-1	16.5	53.7	40.4	18.1	32.2
	Top-5	41.6	85.3	73.2	43.1	60.8
MBP	Top-1	14.0	63.8	41.0	17.0	33.9
	Top-5	31.8	91.8	78.3	41.0	60.7
Liu et al.	Top-1	25.4	78.1	37.3	**35.3**	44.0
	Top-5	-	-	-	-	72.0
SAVER	Top-1	**27.2**	**81.5**	**63.8**	22.5	**48.8**
	Top-5	**50.8**	**98.2**	**90.0**	**49.8**	**72.2**

Table 4: Ablation for routing scheme.

Routing	Top-1	Top-5
Uniform	79.5	95.9
Static	80.1	97.0
Dynamic	**90.7**	**99.4**

Table 5: Ablation for loss functions.

$\mathcal{L}_{\text{RG}}$	$\mathcal{L}_{\text{EA}}$	Top-1	Top-5
		69.2	92.7
✓		74.9	95.6
	✓	83.5	97.7
✓	✓	**90.7**	**99.4**

use a much larger expert pool, highlighting that the larger pool alone does not improve performance. Thus, conditioning the combination on each sample is what makes the pool useful. Under the same supervision as the fixed schemes, dynamic routing already improves over both (83.5 in Table 5), though it stays slightly below HCF, since the contrastive objective supervises the routing distribution only indirectly. With router guidance, the same routing reaches 90.7 and clearly surpasses HCF. The improvement over HCF thus comes from conditioning the target on each sample together with the effective supervision that makes such routing effective, not from the size of the expert pool.

5.4 Effect of Auxiliary Losses.

Table 5 adds the two auxiliary losses to the contrastive objective. Either loss alone improves over the contrastive-only model, and combining them yields the best result, showing that the two forms of supervision are complementary. Expert alignment provides significant gain by aligning each EEG blur expert to the matched visual content and router guidance contributes a further gain by supervising the routing distribution directly.

5.5 Visualization of Dynamic Routing Weights.

Figure 3 shows the visual routing weights across the expert pool as they vary with blur level, subject, and sample. Across all three factors, the weights are not uniform but shift with the conditioning signal, confirming that the router assigns different experts to different inputs. The variation is most pronounced on the RN50 and ViT-B-32 backbones, while MAE-L shows mild variation across subjects and DINOv3-S shows mild variation across samples.

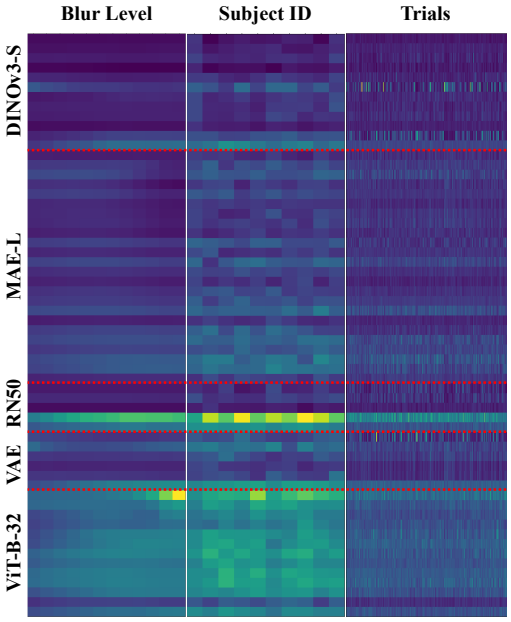

Figure 3: Visual routing weights across the expert pool, shown as they vary with blur level, subject, and sample.



Figure 4: Retrievals missing the correct stimulus at Top-1 but recovering it within Top-5.



Figure 5: Retrievals missing the correct stimulus at Top-5.

5.6 Failure Case Analysis.

Figure 4 shows retrievals in which the correct stimulus is missed at Top-1 but recovered within Top-5. In these cases the competing candidates share a similar foreground color with the target, so the residual errors are confined to a coherent visual neighborhood rather than unrelated confusions. To examine the source of these errors, we compare the EEG router weights of samples that succeed and fail, and observe that the failed ones place less weight on the mid-blur experts and shift their peak toward the blurriest ones. This shift leaves the coarse color of the object intact but removes its shape, which explains these confusions.

Figure 5 collects the harder cases in which the correct stimulus is missed even at Top-5, where the shift is stronger. Here the retrieved candidates instead share a similar background color or texture with the target, as a target composed almost entirely of coarse structure is dominated by the background, which occupies the larger image area. This indicates a need for approaches that focus more on foreground content.

6 Conclusion

In this paper, we propose a novel sample-adaptive visual decoding method that composes the alignment target by conditioning on each EEG embedding, overcoming the limitation of existing methods that fix this target to a single global setting shared across all samples. By recasting the choice of alignment target as an EEG-conditioned mixture-of-experts problem, dynamic routers select, per sample, the blur level, encoder layer, and backbone of the visual representation, while expert alignment and router guidance ground each expert in its own visual content and supervise the router with the measured similarity between the EEG and each expert. Extensive experiments on THINGS-EEG2 and THINGS-MEG demonstrate state-of-the-art zero-shot retrieval, and our ablations confirm that each component contributes to the gains and that the improvement stems from the per-sample conditioning rather than from the enlarged expert pool. Ultimately, our method underscores the importance of conditioning the alignment target on each sample rather than fixing it.

AI USE STATEMENT

In this work, we used generative AI tools for implementing methods. We have not used generative AI tools for proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, assisting with translation, cleaning and reformatting datasets, or interpreting results, and generating synthetic datasets, developing theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, and supporting qualitative and thematic data analysis are not applicable to this work. Additionally, we used generative AI tools for editing the paper to improve readability. We have reviewed all AI-assisted work. AI-modified code was verified and tested for correctness by the authors, and AI-suggested grammar corrections were reviewed by the authors to ensure the intended meaning was preserved. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We have made several efforts to ensure the reproducibility of our work. The full formulation of SAVER is given in Section 3, including the construction of the visual and EEG experts (Sections 3.1–3.2), the EEG-conditioned routing and cross-routing (Eqs. 1–2), the contrastive retrieval objective (Eq. 3), expert alignment (Eq. 4), router guidance (Eqs. 5–6), and the overall objective (Section 3.7). Section 4.1 describes the datasets and preprocessing, including the epoching window, repetition averaging, validation split, and the exact Gaussian-blur kernel sizes. Both datasets, THINGS-EEG2 and THINGS-MEG, are publicly available. Section 4.2 specifies the five pretrained image backbones and the layers from which features are extracted, the resulting number of experts, the architectures of the routers and the adapter, the EEG encoder adopted from Liu et al. (2026), the two-stage training schedule, the optimizer and all hyperparameters, the checkpoint selection criterion, and the hardware used. Section 4.3 details the evaluation protocol. All reported results are averaged over three random seeds. To further facilitate reproduction, we will publicly release the code upon publication.

REFERENCES

Moshe Bar. A cortical mechanism for triggering top-down facilitation in visual object recognition. *Journal of cognitive neuroscience*, 15(4):600–609, 2003.

Benjamin Blankertz, Steven Lemm, Matthias Treder, Stefan Haufe, and Klaus-Robert Müller. Single-trial analysis and classification of erp components—a tutorial. *NeuroImage*, 56(2):814–825, 2011.

Hongzhou Chen, Lianghua He, Yihang Liu, Longzhen Yang, Shaohua Shang, and MengChu Zhou. Visual neural decoding via improved visual-eeg semantic consistency. *arXiv preprint arXiv:2408.06788*, 2024.

Radoslaw Martin Cichy, Dimitrios Pantazis, and Aude Oliva. Resolving human object recognition in space and time. *Nature Neuroscience*, 17(3):455–462, 2014.

Changde Du, Kaicheng Fu, Jinpeng Li, and Huiguang He. Decoding visual neural representations by multimodal learning of brain-visual-linguistic features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10760–10777, 2023.

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.

Alessandro T Gifford, Kshitij Dwivedi, Gemma Roig, and Radoslaw M Cichy. A large and rich eeg dataset for modeling human visual object recognition. *NeuroImage*, 264:119754, 2022.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Martin N Hebart, Oliver Contier, Lina Teichmann, Adam H Rockter, Charles Y Zheng, Alexis Kidder, Anna Corriveau, Maryam Vaziri-Pashkam, and Chris I Baker. Things-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *Elife*, 12:e82580, 2023.

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.

Louise Kauffmann, Stephen Ramanoël, and Carole Peyrin. The neural bases of spatial frequency processing during scene perception. *Frontiers in integrative neuroscience*, 8:37, 2014.

Dongyang Li, Chen Wei, Shiying Li, Jiachen Zou, and Quanying Liu. Visual decoding and reconstruction via eeg embeddings with guided diffusion. *Advances in Neural Information Processing Systems*, 37:102822–102864, 2024.

Wenchao Liu, Hongwei Li, Zhouyang Xu, Lin Ma, and Haifeng Li. Leveraging visual blur perception characteristics for eeg decoding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 17580–17588, 2026.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Fabien Lotte, Laurent Bougrain, Andrzej Cichocki, Maureen Clerc, Marco Congedo, Alain Rakotomamonjy, and Florian Yger. A review of classification algorithms for eeg-based brain–computer interfaces: a 10 year update. *Journal of Neural Engineering*, 15(3):031005, 2018.

Basil Mustafa, Carlos Riquelme, Joan Puigcerver, Rodolphe Jenatton, and Neil Houlsby. Multimodal contrastive learning with limoe: the language-image mixture of experts. In *Advances in Neural Information Processing Systems*, volume 35, pp. 9564–9576, 2022.

Simone Palazzo, Concetto Spampinato, Isaak Kavasidis, Daniela Giordano, Joseph Schmidt, and Mubarak Shah. Decoding brain representations by multimodal learning of neural activity and visual features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):3833–3849, 2021.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, volume 2024, pp. 1862–1874, 2024.

Joan Puigcerver, Carlos Riquelme, Basil Mustafa, and Neil Houlsby. From sparse to soft mixtures of experts. In *International Conference on Learning Representations*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. In *Advances in Neural Information Processing Systems*, volume 34, pp. 8583–8595, 2021.

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

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Yonghao Song, Bingchuan Liu, Xiang Li, Nanlin Shi, Yijun Wang, and Xiaorong Gao. Decoding natural images from eeg for object recognition. In *The Twelfth International Conference on Learning Representations*, 2024.

Concetto Spampinato, Simone Palazzo, Isaak Kavasidis, Daniela Giordano, Nasim Souly, and Mubarak Shah. Deep learning human mind for automated visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6809–6817, 2017.

Jingyi Tang, Shuai Jiang, Fei Su, and Zhicheng Zhao. Aligning what eeg can see: Structural representations for brain-vision matching. *arXiv preprint arXiv:2603.07077*, 2026.

Yayun Wei, Lei Cao, Hao Li, and Yilin Dong. Mb2c: Multimodal bidirectional cycle consistency for learning robust visual neural representations. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 8992–9000, 2024.

Jonathan R Wolpaw, Niels Birbaumer, Dennis J McFarland, Gert Pfurtscheller, and Theresa M Vaughan. Brain-computer interfaces for communication and control. *Clinical neurophysiology*, 113(6):767–791, 2002.

Haitao Wu, Qing Li, Changqing Zhang, Zhen He, and Xiaomin Ying. Bridging the vision-brain gap with an uncertainty-aware blur prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2246–2257, 2025.

Lukun Wu, Jie Li, Ziqi Ren, Kaifan Zhang, and Xinbo Gao. Shrinking the teacher: An adaptive teaching paradigm for asymmetric eeg-vision alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 17859–17867, 2026.

Kaifan Zhang, Lihuo He, Xin Jiang, Wen Lu, Di Wang, and Xinbo Gao. Cognitioncapturer: Decoding visual stimuli from human eeg signal with multimodal information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 14486–14493, 2025.

Wenjiang Zhang, Sifeng Wang, Yuwei Su, Xinyu Li, Chen Zhang, and Suyu Zhong. Neurobridge: Bio-inspired self-supervised eeg-to-image decoding via cognitive priors and bidirectional semantic alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 18028–18036, 2026.

Jiawen Zheng, Haonan Jia, MING LI, Yuhui Zheng, Yufeng Zeng, Yang Gao, and Chen Liang. Learning brain representation with hierarchical visual embeddings. In *The Fourteenth International Conference on Learning Representations*, 2026.

Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Zhao, Andrew M Dai, Zhifeng Chen, Quoc V Le, and James Laudon. Mixture-of-experts with expert choice routing. In *Advances in Neural Information Processing Systems*, volume 35, pp. 7103–7114, 2022.