PAIRCAL: TRAINING-FREE PAIRED-SUPPORT CALIBRATION FOR CROSS-SUBJECT EEG-TO-IMAGE RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-subject EEG-to-image retrieval remains highly challenging due to substantial inter-subject variability. Existing methods primarily address this issue by learning subject-agnostic representations for cross-modal alignment. In contrast, few studies have explored how to adapt pretrained retrieval models to unseen subjects using limited examples. We propose PairCal, a training-free calibration framework for adapting retrieval to unseen subjects with only a small set of paired EEG–image support examples. PairCal calibrates both EEG queries and image candidates through two complementary components. Support-Guided Calibration (SGC) uses EEG–image residuals from the support set to derive query-adaptive cross-modal corrections. Meta-Distractor Calibration (MDC) constructs pseudo-unseen queries from the same support set via a leave-one-out procedure to estimate and suppress candidate-specific hubness bias. Together, these components enable effective few-shot adaptation to unseen subjects without updating the pretrained encoders. Extensive experiments across held-out subjects, encoder architectures, training schemes, and support-set sizes demonstrate consistent retrieval gains.

1 INTRODUCTION

Decoding visual information from brain activity is a fundamental problem in computational neuroscience and brain–computer interfaces. Owing to its non-invasive acquisition and high temporal resolution, electroencephalography (EEG) has been widely used for visual recognition, retrieval, and reconstruction (Song et al., 2024; Li et al., 2024). Among these tasks, EEG-to-image retrieval evaluates whether neural representations preserve sufficient visual-semantic information to identify the corresponding image stimulus from a candidate gallery. Recent approaches typically address this problem by aligning EEG signals with visual embeddings through contrastive learning objectives (Song et al., 2024; Li et al., 2024; Chen et al., 2024).

However, retrieval models trained on source subjects often generalize poorly to unseen individuals because EEG responses exhibit substantial inter-subject variability. Even when viewing the same visual stimulus, different subjects may produce markedly different neural responses. Consequently, the EEG representation of an unseen subject can be misaligned with the visual embedding space learned during training, degrading retrieval performance (Tran et al., 2026; Huang & Zhu, 2026).

Existing methods primarily address this challenge by improving cross-subject generalization during training. They enhance EEG representations, incorporate subject-aware modeling, or introduce richer cross-modal supervision. For example, SAMGA employs subject-aware supervision (Jiang et al., 2026), while EEGiT transfers pretrained visual priors to improve EEG representations (Zhou et al., 2026). Although these approaches strengthen source-trained representations, they do not directly address how to calibrate a frozen retrieval model using a small number of paired examples from a new subject.

We therefore study a few-shot cross-subject EEG-to-image retrieval setting. Before retrieval begins, a small set of paired EEG–image examples is collected from an unseen target subject and retained as a fixed support set. The pretrained EEG and image encoders remain frozen, no target-side parameters are optimized, and each test EEG query is processed independently using this support set and the candidate gallery, without access to other target test queries. The support images are disjoint from

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.

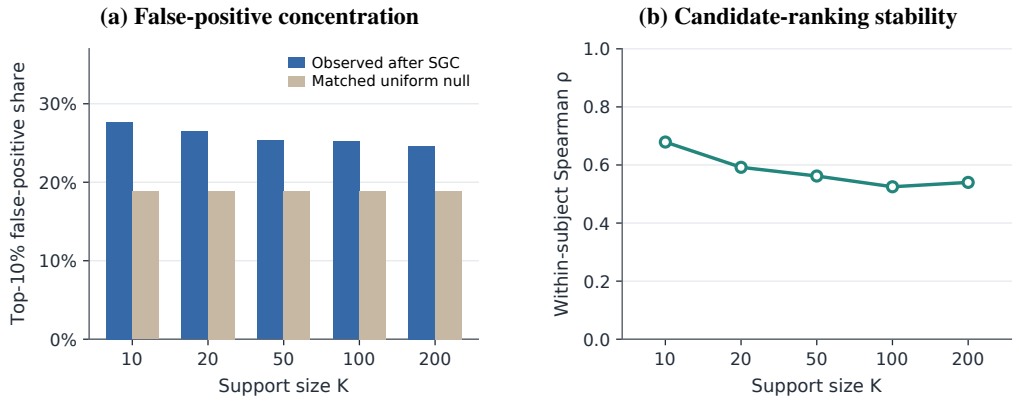

Figure 1: Candidate-side error patterns after SGC, evaluated within the training-image pool. (a) Share of false-positive Top-5 occurrences assigned to the 10% most frequently retrieved incorrect candidates, compared with a uniform-allocation null matched for the number of error events. (b) Within-subject Spearman correlation of candidate false-positive frequency rankings across support draws. Results are averaged across ten subjects, four EEG backbones, and three support draws.

the evaluation gallery. This setting raises a central question: *how can a few paired support examples calibrate both query representations and candidate rankings for an unseen subject?*

To address this challenge, we propose **PairCal**, a training-free framework that leverages a paired support set for two complementary forms of calibration. To directly correct cross-modal misalignment, we introduce **Support-Guided Calibration (SGC)**, which utilizes the embedding discrepancies of the support pairs as a residual dictionary to predict query-dependent corrections based on EEG-space similarities, enabling the dynamic adaptation of the query representation with a frozen EEG encoder. However, query-side calibration alone is insufficient. Our preliminary analysis demonstrates that false positives disproportionately concentrate on a small subset of candidate images (Fig. 1a). Candidate rankings by false-positive frequency also show consistency across support draws within subjects (Fig. 1(b)).This phenomenon reveals that certain gallery items act as universal distractors with inherent baseline attractions, motivating an explicit mechanism to estimate and suppress them. To resolve this, we propose **Meta-Distractor Calibration (MDC)** to isolates and penalizes this spurious attraction using only the paired support set and candidate gallery. MDC simulates the behavior of unseen queries by applying a leave-one-out SGC procedure to the support EEG data. By measuring how these pseudo-unseen reference queries interact with the gallery, MDC effectively quantifies the baseline attractiveness of each candidate image and discounts it during the final retrieval ranking. Together, SGC and MDC calibrate query representations and candidate rankings through closed-form computations, without target-side parameter updates or access to other target test queries.

We evaluate PairCal under a strict leave-one-subject-out protocol across ten unseen subjects, multiple EEG backbones and training realizations, and support budgets of 10–200 paired examples. PairCal consistently improves the frozen retrieval model. In our main evaluation with 200 support pairs, SGC raises Top-1 accuracy from 20.50% to 33.44%, while the full framework reaches 37.90%. Comparisons with alternative few-shot adaptation and candidate-calibration methods, together with component ablations and candidate-level diagnostics, support the effectiveness of the framework and the complementary benefits of its two components.

Our main contributions are summarized as follows:

- We propose PairCal, a few-shot paired-support framework for cross-subject EEG-to-image retrieval that calibrates query representations and candidate rankings without updating the pretrained encoders or accessing other target test queries.
- We introduce two complementary calibration components: SGC transfers support EEG–image residuals to incoming queries through query-dependent correction, while MDC constructs leave-one-out support references to estimate and discount candidate-specific attraction.

- We conduct systematic evaluations across ten unseen subjects, multiple EEG backbones, training realizations, and support budgets. Training-pool diagnostics reveal recurring candidate-side error concentration after query calibration, while held-out-subject evaluations and mechanism analyses support the effectiveness of paired-support calibration.

2 Related Work

2.1 EEG-to-Image Retrieval

EEG-to-image retrieval typically aligns neural signals with visual representations to identify the viewed stimulus (Song et al., 2024; Li et al., 2024; Chen et al., 2024). Recent methods address the modality gap in different ways: UBP suppresses high-frequency visual information (Wu et al., 2025); Jo et al. (Jo et al., 2026) use hyperbolic feature interpolation; ATS adapts semantically dense visual features to EEG (Wu et al., 2026); and hierarchical visual fusion combines pixel-level detail with high-level semantics (Zheng et al., 2026). The choice of image representation also matters (Rajabi et al., 2025). EEGiT instead converts EEG into image-like patches for a pretrained Vision Transformer (Zhou et al., 2026; Dosovitskiy et al., 2021). Despite these advances, inter-subject variability can impair retrieval for unseen subjects (Tran et al., 2026; Huang & Zhu, 2026).

Cross-Subject Transfer and Calibration. Classical approaches reduce subject differences through explicit alignment. Euclidean alignment normalizes trials using subject-level covariance (He & Wu, 2020), Riemannian Procrustes analysis aligns covariance representations geometrically (Rodrigues et al., 2019), and fMRI hyperalignment maps individual responses into a shared space (Haxby et al., 2011). Their assumptions and supervision differ: covariance normalization, for example, need not rely on shared stimuli.

Recent models address variability during representation learning. CroSS combines soft contrastive alignment with subject-aware normalization (Kim et al., 2026); SAMGA combines a shared visual granularity prior with subject-specific residuals while retaining subject-agnostic inference (Jiang et al., 2026); and NEED learns transferable representations across datasets and image and video reconstruction tasks (Huang et al., 2025). These approaches leave room for calibration using evidence from a new subject at deployment.

Retrieval also depends on the gallery: *hubness* makes some candidates the nearest neighbors to disproportionately many queries (Radovanović et al., 2010). Local similarity scaling (Lample et al., 2018), query-bank normalization in QB-Norm (Bogolin et al., 2022), dual-bank normalization in DBNorm (Wang et al., 2023), and candidate-specific bias estimation in Nearest Neighbor Normalization (Chowdhury et al., 2024) address related effects. In EEG-to-image retrieval, SATTC combines label-free subject adaptation and ranking calibration with frozen encoders (Huang & Zhu, 2026). PairCal instead uses a small set of *paired* target examples to calibrate both EEG queries and gallery attraction. It constructs the latter's references from support EEGs while excluding each EEG's own pair, without requiring access to other target test queries.

2.2 Few-Shot Target Adaptation

Limited target data can personalize shared-subject fMRI decoders (Scotti et al., 2024) or support an explicit transfer from a new subject to known subjects (Dai et al., 2025). Other approaches meta-optimize conditioning on image–brain examples without target-side fine-tuning (Nan et al., 2026), or train a residual adapter over a frozen backbone using shared-stimulus supervision (Liu et al., 2026). PairCal instead calibrates an existing EEG retrieval model from paired target examples without changing source training or optimizing a target-specific neural adapter.

Related vision-language methods adapt pretrained predictions using learned residual adapters (Gao et al., 2021), a non-parametric support cache (Zhang et al., 2022), or proximal kernel ridge regression (Bendou et al., 2025). Whereas Tip-Adapter and ProKeR adapt class predictions, PairCal's SGC transfers corrections in the cross-modal embedding space: each support pair supplies the residual between its EEG and image embeddings, and a test EEG retrieves relevant residuals according to support-EEG similarity. PairCal also uses leave-one-out support predictions to estimate gallery at-

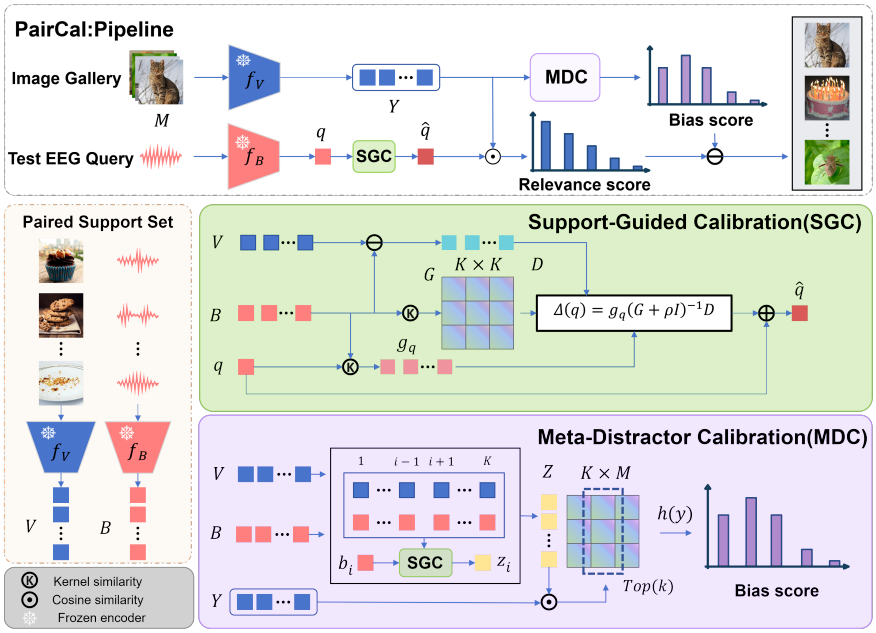

Figure 2: Overview of PairCal. SGC transfers cross-modal residuals from paired support to the target EEG query. MDC uses leave-one-out support queries to estimate candidate attraction, which is subtracted from query–image relevance.

traction. Thus, the same paired support set corrects query representations and calibrates ranking while the source model remains fixed.

3 PairCal

3.1 Problem Setting

We study EEG-to-image retrieval for a target subject excluded from source training. Both the pre-trained EEG encoder f_B and image encoder f_V remain frozen. For an incoming EEG trial x_q and a gallery of M candidate images, let $\mathbf{q} = f_B(x_q)/\|f_B(x_q)\|_2$ and $\mathbf{y}_j = f_V(I_j)/\|f_V(I_j)\|_2$ denote their normalized embeddings. The frozen model ranks candidates by $\mathbf{q}^\top \mathbf{y}_j$. Before retrieval begins, we observe a fixed paired support set $\mathcal{S} = \{(x_i^s, I_i^s)\}_{i=1}^K$ from the target subject. We write $\mathbf{b}_i$ and $\mathbf{v}_i$ for the normalized EEG and image embeddings of its i-th pair. The support images are disjoint from the evaluation gallery. PairCal uses the paired support in two complementary ways: SGC predicts how to correct the representation of the current EEG, while MDC estimates which gallery candidates are broadly attractive to calibrated target EEGs, as summarized in Figure 2. Support-dependent quantities are prepared before inference, and each test EEG is processed without accessing any other target test query.

3.2 Support-Guided Calibration

The source-trained EEG embedding already contains information useful for identifying visual stimuli. A target-side mapping learned from only a few pairs may alter this transferable structure while trying to close the remaining EEG–image gap. SGC therefore starts from the frozen EEG embedding and estimates a correction rather than a replacement. Each paired support example reveals an observed discrepancy in the shared embedding space. With $B = [\mathbf{b}_1, \ldots, \mathbf{b}_K]^\top$ and $V = [\mathbf{v}_1, \ldots, \mathbf{v}_K]^\top$, these discrepancies form the residual dictionary

$$\mathbf{d}_i = \mathbf{v}_i - \mathbf{b}_i, \qquad D = [\mathbf{d}_1, \ldots, \mathbf{d}_K]^\top = V - B. \tag{1}$$

Applying the same average residual to every new EEG would treat this discrepancy as a global translation. In practice, the useful correction may depend on the stimulus represented by the query:

support EEGs near a query can carry more relevant evidence than distant ones. We express this locality using an exponential cosine kernel on normalized EEG embeddings,

$$\kappa(\mathbf{a}, \mathbf{b}) = \exp\left(\frac{\mathbf{a}^\top \mathbf{b} - 1}{\tau}\right), \qquad \tau > 0. \tag{2}$$

The bandwidth τ controls how quickly affinity decreases as two embeddings diverge. Let $G_{ij} = \kappa(\mathbf{b}_i, \mathbf{b}_j)$ be the support Gram matrix and let $\mathbf{g}_q = [\kappa(\mathbf{q}, \mathbf{b}_i)]_{i=1}^K$ contain the affinities of the current query to all support EEGs. The paired images enter through D, whereas the choice of relevant residuals depends on the target EEG geometry. We estimate the query-specific displacement by kernel ridge regression, a form of support-conditioned adaptation also used with pretrained vision-language models (Bendou et al., 2025):

$$\Delta(\mathbf{q}) = \mathbf{g}_q (G + \rho I)^{-1} D, \qquad \rho > 0. \tag{3}$$

The ridge coefficient ρ stabilizes the support solve and limits the influence of noisy or redundant pairs. Unlike a fixed subject-level shift, $\Delta(\mathbf{q})$ varies with the incoming EEG. We retain the frozen query as an explicit anchor, scale the predicted displacement by γ, and return to the unit sphere:

$$\widehat{\mathbf{q}} = \mathrm{Norm}\big(\mathbf{q} + \gamma \Delta(\mathbf{q})\big), \qquad \gamma \geq 0. \tag{4}$$

Here $\mathrm{Norm}(\mathbf{u}) = \mathbf{u}/\|\mathbf{u}\|_2$. The residual dictionary and regularized solve are shared across all queries from the target subject; only $\mathbf{g}_q$ is recomputed when a query arrives. This formulation assumes that proximity between target EEG embeddings is informative about their cross-modal discrepancies. It uses paired support images to establish the corrections, but requires neither a paired image for a test EEG nor any update to the pretrained encoders.

3.3 META-DISTRACTOR CALIBRATION

A corrected query can still favor gallery candidates that are similar to many target EEGs, as in hubness-aware retrieval (Radovanović et al., 2010; Bogolin et al., 2022). This candidate-side attraction cannot be estimated from a batch of target test queries in our inductive setting. MDC instead reuses the paired support to approximate how independently calibrated target queries interact with the gallery. Directly calibrating a support EEG with the complete support set would let its own paired image contribute to its predicted residual, making the resulting reference overly optimistic. We therefore withhold pair i when constructing a reference from $\mathbf{b}_i$. The leave-one-out predictions can be recovered from the SGC kernel system without solving K separate regression problems. Writing $C = (G + \rho I)^{-1}$ and $A = CD$, the kernel-ridge identity gives

$$\Delta_{-i}(\mathbf{b}_i) = \mathbf{d}_i - \frac{A_{i:}}{C_{ii}}, \tag{5}$$

where $A_{i:}$ is the i-th row of A. This is the residual predicted for $\mathbf{b}_i$ using all support pairs except its own. We then apply the same correction strength and normalization used for an incoming query:

$$\mathbf{z}_i = \mathrm{Norm}\big(\mathbf{b}_i + \gamma \Delta_{-i}(\mathbf{b}_i)\big). \tag{6}$$

Each $\mathbf{z}_i$ uses $K - 1$ support pairs, whereas a test query uses all K; the reference bank is therefore a proxy rather than an exact sample from deployment-time predictions. Its role is to expose candidate-level attraction, not to retrieve the held-out support images.

For each gallery candidate $\mathbf{y}_j$, let $a_{ij} = \mathbf{z}_i^\top \mathbf{y}_j$ denote its cosine similarity to reference $\mathbf{z}_i$. A candidate may strongly attract a subset of reference queries even when its average similarity across the entire bank is moderate. We therefore estimate its upper-tail candidate attraction, focusing on the k largest reference similarities, where $k \in \{1, \ldots, K\}$. Let t_j be the k-th largest value among $\{a_{ij}\}_{i=1}^K$. We define

$$h(\mathbf{y}_j) = t_j + \frac{1}{k} \sum_{i=1}^{K} \big[a_{ij} - t_j\big]_+, \qquad [u]_+ = \max(u, 0). \tag{7}$$

Here, t_j marks the boundary of the candidate's upper similarity tail, and the positive-part operator $[\cdot]_+$ retains only similarities exceeding this boundary. The second term sums these excess similarities and scales them by the tail size k. Together, the two terms are exactly equal to the mean of the k

largest reference similarities, including when similarities are tied. Thus, $h(\mathbf{y}_j)$ summarizes attraction within the candidate's most similar reference neighborhood rather than across the entire bank. The threshold t_j is determined by the reference similarities and does not introduce an additional hyperparameter.

A large $h(\mathbf{y}_j)$ indicates strong upper-tail attraction to calibrated reference EEGs; it is an unlabeled attraction proxy, not proof that the candidate is incorrect for any particular query. MDC discounts this shared attraction when ranking images for the current EEG:

$$s_{\mathrm{PairCal}}(\mathbf{q}, \mathbf{y}_j; \mathcal{S}) = \widehat{\mathbf{q}}^{\top}\mathbf{y}_j - \lambda h(\mathbf{y}_j), \qquad \lambda \geq 0. \tag{8}$$

The first term preserves query-specific relevance after SGC, while the second adjusts a candidate property estimated before the query arrives. For a fixed gallery, all $h(\mathbf{y}_j)$ values are reusable. Constructing the support kernel, solving the regularized system, and scoring the leave-one-out references against the gallery require $O(K^2d + K^3 + KMd)$ operations once. Each subsequent EEG adds $O(Kd)$ support calibration before the frozen model's $O(Md)$ gallery scoring. We select τ, ρ, γ, k, and λ exclusively on source-side validation subjects and keep them fixed when evaluating a held-out target subject.

4 Experiments

In this section, We first describe the experimental setup and then compare PairCal with target-adaptation and candidate-calibration baselines across support budgets, analyze the contributions of SGC and MDC, and assess generalization across EEG encoders. We further present post-hoc diagnostics of residual transfer and candidate-bias estimation. Unless explicitly marked as an offline diagnostic, all experiments use frozen encoders and process each target test query without access to other test queries.

4.1 Implementation Details

All experiments run on a single NVIDIA RTX 3090 GPU. Following Brain-hive, we use Brain-MLP as the EEG encoder and a visual feature fusion encoder that combines multiple visual features into an ℓ_2-normalized, 1024-dimensional image representation. The EEG encoder, visual feature extractors, and visual fusion module remain frozen during evaluation.

All calibration hyperparameters are selected exclusively on source-side validation subjects. For SGC, we use $\tau = 0.3$, $\rho = 0.01$, and $\gamma = 1$. For MDC, we set $\lambda_{\mathrm{MDC}} = 0.55$ and use neighborhood sizes $k = 5, 10, 20, 25, 30$ for support budgets $K = 10, 20, 50, 100, 200$, respectively.

4.2 Dataset

We evaluate on THINGS-EEG (Gifford et al., 2022), which contains recordings from 10 participants. The training split comprises 1,654 concepts with 10 images per concept and four presentations per image; the test split contains 200 disjoint concepts with one image per concept and 80 presentations per image. EEG signals are filtered, baseline-corrected, and downsampled, retaining 17 occipital and parietal channels. We average repeated trials for each image, yielding 16,540 training and 200 test examples per participant.

We follow a leave-one-subject-out protocol: each participant is held out in turn as the target subject, while the remaining participants are used for source-model training and source-only validation. No target-subject data are used to train the encoders or visual fusion module. Each target subject's 200 test EEG queries are ranked against the complete gallery of 200 test images.

4.3 Support Set Construction

For each target subject, we sample $K \in \{10, 20, 50, 100, 200\}$ paired EEG–image examples from the official training-image split. Each support EEG averages four repeated trials, so these budgets require $4K$ raw trials and use at most 1.21% of the available image-level training examples. Each pair is represented by normalized EEG and image embeddings.

Support budgets are sampled independently using prespecified seeds, and all paired-support methods receive the same support set within each evaluation episode. Support and test data are disjoint in EEG trials, image identities, and object concepts. SGC and MDC use support pairs only for closed-form calibration, without gradient-based updates or access to other target test queries or aggregate target-test statistics.

4.4 Evaluation Metrics

We report Top-1 accuracy, Top-5 accuracy, and mean reciprocal rank (MRR). All methods share the same frozen checkpoints, test queries, and image gallery. For each held-out subject, we evaluate three independently trained EEG–visual model checkpoints, each with three fixed support seeds per support budget. Results are averaged first across support seeds, then across checkpoints, and finally across the 10 held-out subjects.

4.5 Comparison with Baselines and Effect of Support-Set Size

To evaluate PairCal's ability to adapt to unseen subjects, we compare it with the frozen base retrieval model BrainHIVE, the transductive test-time calibration method SATTC, and several classical target-domain adaptation methods. The results are reported in Table 1. All methods use the same pretrained EEG and image encoders and are evaluated using the same complete test-query set and candidate gallery. Methods requiring paired target-subject support use identical support samples within each evaluation episode.

Table 1: Top-1 / Top-5 accuracy (%) of zero-shot, transductive, and paired-support adaptation methods on THINGS-EEG2 under different calibration budgets K.

Method	$K = 10$	$K = 20$	$K = 50$	$K = 100$	$K = 200$
	Zero-shot cross-subject retrieval baseline				
Brainhive	20.50/44.92				
	Transductive calibration with unlabeled test queries				
SATTC	21.31/49.94	20.38/49.61	22.10/50.63	22.81/50.33	22.92/50.91
	Few-shot paired-support adaptation methods				
Ridge mapping	3.23/12.22	5.82/19.72	12.72/36.37	20.12/49.24	28.11/59.82
Linear adapter	19.53/45.72	20.06/46.30	21.73/49.01	22.83/50.68	23.66/51.45
Orthogonal Procrustes	21.72/48.48	22.49/49.72	24.18/53.88	27.56/58.26	32.12/64.01
PairCal (ours)	**25.98/55.62**	**28.46/59.24**	**31.59/63.44**	**35.08/66.50**	**37.90/69.81**

The meaning of K differs between the two calibration paradigms. For SATTC, K unlabeled queries are randomly sampled from the target test set to estimate its whitening and similarity-calibration statistics. Their labels are never used, and SATTC is evaluated on the same complete 200-query test set as all other methods, including the queries used to estimate its calibration statistics. For the few-shot methods, K denotes the number of paired EEG–image support examples sampled from the target subject's training-image split, which is disjoint from the evaluation set. SATTC is therefore included as a transductive reference rather than a resource-matched baseline.

The frozen model performs zero-shot cross-subject retrieval without using target-subject calibration data. PairCal instead uses a fixed set of paired support examples collected before retrieval and processes each test EEG independently, without accessing other test queries. We additionally consider Ridge Mapping, Orthogonal Procrustes, and Linear Adapter as paired-support adaptation baselines. Ridge Mapping learns a regularized linear mapping from target-subject EEG embeddings to image embeddings. Orthogonal Procrustes estimates an orthogonal transformation that reduces cross-modal misalignment while preserving the geometry of the EEG representations. Linear Adapter learns a linear predictor of the EEG–image residual and adds the predicted residual to the original EEG embedding. Implementation and hyperparameter-selection details are provided in the supplementary material.

We examine support budgets of $K \in \{10, 20, 50, 100, 200\}$. PairCal consistently outperforms the frozen model, with Top-1 accuracy increasing from 25.98% at $K = 10$ to 37.90% at $K = 200$, compared with the frozen baseline of 20.50%. The additional gain from MDC over SGC also grows with K, suggesting that larger support sets benefit both query calibration and candidate-attraction estimation.

4.6 Component Ablation

Table 2: Top-1 and Top-5 accuracy and MRR (%) for Brainhive, SGC, MDC, and PairCal across different values of K.

	Brainhive			SGC			MDC			PairCal		
K	T1	T5	MRR	T1	T5	MRR	T1	T5	MRR	T1	T5	MRR
10	20.50	44.92	33.11	24.99	53.74	38.74	23.69	51.91	37.15	**25.98**	**55.62**	**39.99**
20	20.50	44.92	33.11	26.79	56.73	40.96	24.61	52.31	37.90	**28.46**	**59.24**	**42.90**
50	20.50	44.92	33.11	29.48	61.29	44.24	24.07	51.46	37.30	**31.59**	**63.44**	**46.37**
100	20.50	44.92	33.11	31.78	63.04	46.35	23.17	49.43	36.02	**35.08**	**66.50**	**49.47**
200	20.50	44.92	33.11	33.44	65.55	48.17	22.79	49.27	35.71	**37.90**	**69.81**	**52.36**

As shown in Table 2, we conduct a component ablation study of PairCal. The results show that the performance gains provided by SGC and MDC are complementary.

Query-side calibration. To isolate the design choices underlying SGC, we compare it with three simpler paired-support estimators. Mean Residual applies the average EEG–image residual to every query and therefore models only a global subject-specific shift. Similarity-Weighted Residual directly averages support residuals using normalized query–support kernel affinities, without solving a kernel regression system. Direct KRR predicts the complete image embedding rather than preserving the pretrained EEG query and estimating only its residual. All methods use the same frozen encoders and paired support examples, with hyperparameters selected exclusively on source-validation subjects.

Table 3: Top-1 accuracy (%) for query-side calibration ablations.

K	Brainhive	Mean Res.	Weighted Res.	Direct KRR	SGC
10	20.50	24.43	17.64	2.34	**24.99**
20	20.50	25.93	18.03	3.94	**26.79**
50	20.50	27.01	19.73	9.31	**29.48**
100	20.50	27.39	20.66	14.83	**31.78**
200	20.50	27.44	20.78	20.53	**33.44**

As shown in Table 3, Mean Residual improves over the frozen model but saturates as K increases, suggesting that a single subject-level shift cannot fully capture query-dependent misalignment. Direct KRR performs poorly in the low-support regime, indicating that predicting the complete visual embedding from limited target data can discard useful structure in the pretrained EEG representation. Similarity-Weighted Residual is also consistently less effective than SGC: SGC improves Top-1 accuracy by 7.35 to 12.66 percentage points across the evaluated support budgets. At $K = 200$, SGC reaches 33.44% Top-1 accuracy, compared with 27.44% for Mean Residual, 20.78% for Similarity-Weighted Residual, and 20.53% for Direct KRR. The complete Top-5 and MRR results in Appendix A.4 exhibit the same overall pattern. These comparisons support both preserving the pretrained query through residual correction and modeling interactions among support examples through the regularized kernel system.

Candidate-side calibration The component ablation shows that MDC improves retrieval accuracy, but does not by itself establish whether the improvement arises from estimating candidate-specific hubness. We therefore directly examine whether the support-derived attraction score $h(y_j)$ predicts how frequently candidate y_j appears as a false positive on the disjoint test queries. For each

candidate, we define its false-positive frequency under SGC as

$$F_j^{(5)} = \sum_q \mathbb{I}\left[j \in \mathrm{Top5}(q) \ \wedge \ j \neq y_q^\star\right],$$

where $y_q^\star$ is the ground-truth image for test query q. Test labels are used only for this post-hoc analysis and are not available to either SGC or MDC during calibration.

We evaluate the association between $h(y_j)$ and $F_j^{(5)}$ using Spearman correlation. We further quantify false-positive concentration using

$$H_{10} = \frac{\sum_{j \in \mathcal{H}_{10}} F_j^{(5)}}{\sum_j F_j^{(5)}},$$

where $\mathcal{H}_{10}$ contains the 10% of candidates with the largest false-positive frequencies. A lower H_{10} indicates that retrieval errors are less concentrated on a small set of candidate hubs.

Table 4: Mechanism analysis of MDC. The Spearman correlation measures whether the support-derived attraction score h predicts candidate false-positive frequency $F^{(5)}$. ΔH_{10} denotes PairCal minus SGC, where negative values indicate reduced false-positive concentration. $\Delta\mathrm{T1}_{\mathrm{perm}}$ compares the correct candidate-specific h with the mean over 50 random candidate permutations, in percentage points.

K	$\rho(h, F^{(5)})$	SGC H_{10}	PairCal H_{10}	$\Delta\mathrm{T1}_{\mathrm{perm}}$
10	0.332	0.422	0.402	+2.87
20	0.387	0.393	0.365	+3.05
50	0.446	0.368	0.340	+3.02
100	0.501	0.358	0.320	+5.08
200	0.519	0.358	0.312	+7.24

As shown in Table 4, $h(y_j)$ is positively associated with the false-positive frequency of candidate y_j for every support-set size. The correlation increases from 0.332 at $K = 10$ to 0.519 at $K = 200$, and all subject-level bootstrap confidence intervals exclude zero. Candidates in the top 20% according to h are also selected incorrectly substantially more often than candidates in the bottom 20%; for example, the corresponding false-positive counts are 8.05 and 1.47 at $K = 200$.

Applying MDC reduces H_{10} relative to SGC for every K. The reduction ranges from 0.0199 at $K = 10$ to 0.0461 at $K = 200$, with bootstrap confidence intervals below zero in all settings. Moreover, randomly permuting h across candidates substantially degrades retrieval accuracy, although the marginal distribution of the penalties remains unchanged. This control indicates that the gain depends on assigning the estimated attraction penalty to the corresponding candidate, rather than merely changing the overall similarity distribution. Together, these results support the interpretation that MDC estimates candidate-specific attraction from the paired support set and reduces the concentration of false-positive retrievals. Additional statistics, confidence intervals, and comparisons of alternative reference-query constructions are provided in Appendix A.5.

5 Conclusion

We present PairCal, a paired-support calibration framework for cross-subject EEG-to-image retrieval. Unlike existing methods, PairCal adapts a frozen retrieval model to unseen subjects without target-side optimization or access to other test queries. It combines query-dependent residual correction with cross-fitted candidate calibration using a small paired support set. Experiments across subjects, support budgets, and EEG encoders validate the effectiveness of our approach.

AI use statement

In this work, we did not use generative AI tools for any tasks requiring disclosure. Additionally, we used generative AI tools to edit the manuscript for grammar, clarity, and readability. We have

reviewed all AI-assisted work and verified that the edits did not alter the scientific content or conclusions. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Yassir Bendou, Amine Ouasfi, Vincent Gripon, and Adnane Boukhayma. ProKeR: A Kernel Perspective on Few-Shot Adaptation of Large Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Bendou_ProKeR_A_Kernel_Perspective_on_Few-Shot_Adaptation_of_Large_Vision-Language_CVPR_2025_paper.html.

Simion-Vlad Bogolin, Ioana Croitoru, Hailin Jin, Yang Liu, and Samuel Albanie. Cross Modal Retrieval With Querybank Normalisation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Bogolin_Cross_Modal_Retrieval_With_Querybank_Normalisation_CVPR_2022_paper.html.

Hongzhou Chen, Lianghua He, Yihang Liu, and Longzhen Yang. Visual Neural Decoding via Improved Visual-EEG Semantic Consistency, 2024. URL https://arxiv.org/abs/2408.06788v1.

Neil Chowdhury, Franklin Wang, Sumedh Shenoy, Douwe Kiela, Sarah Schwettmann, and Tristan Thrush. Nearest Neighbor Normalization Improves Multimodal Retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024. doi: 10.18653/v1/2024.emnlp-main.1257. URL https://aclanthology.org/2024.emnlp-main.1257/.

Yuqin Dai, Zhouheng Yao, Chunfeng Song, Qihao Zheng, Weijian Mai, Kunyu Peng, Shuai Lu, Wanli Ouyang, Jian Yang, and Jiamin Wu. MindAligner: Explicit Brain Functional Alignment for Cross-Subject Visual Decoding from Limited fMRI Data. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025. URL https://proceedings.mlr.press/v267/dai25m.html.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. CLIP-Adapter: Better Vision-Language Models with Feature Adapters, 2021. URL https://arxiv.org/abs/2110.04544.

Alessandro T. Gifford, Kshitij Dwivedi, Gemma Roig, and Radoslaw M. Cichy. A large and rich EEG dataset for modeling human visual object recognition. *NeuroImage*, 2022. doi: 10.1016/j.neuroimage.2022.119754. URL https://doi.org/10.1016/j.neuroimage.2022.119754.

James V. Haxby, J. Swaroop Guntupalli, Andrew C. Connolly, Yaroslav O. Halchenko, Bryan R. Conroy, M. Ida Gobbini, Michael Hanke, and Peter J. Ramadge. A Common, High-Dimensional Model of the Representational Space in Human Ventral Temporal Cortex. *Neuron*, 2011. doi: 10.1016/j.neuron.2011.08.026. URL https://doi.org/10.1016/j.neuron.2011.08.026.

He He and Dongrui Wu. Transfer Learning for Brain-Computer Interfaces: A Euclidean Space Data Alignment Approach. *IEEE Transactions on Biomedical Engineering*, 2020. doi: 10.1109/TBME.2019.2913914. URL https://doi.org/10.1109/TBME.2019.2913914.

Qunjie Huang and Weina Zhu. SATTC: Structure-Aware Label-Free Test-Time Calibration for Cross-Subject EEG-to-Image Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Huang_SATTC_Structure-Aware_Label-Free_Test-Time_Calibration_for_Cross-Subject_EEG-to-Image_Retrieval_CVPR_2026_paper.html.

Shuai Huang, Huan Luo, Haodong Jing, Qixian Zhang, Litao Chang, Yating Feng, Xiao Lin, Chendong Qin, Han Chen, Shuwen Jia, Siyi Sun, and Yongxiong Wang. NEED: Cross-Subject and Cross-Task Generalization for Video and Image Reconstruction from EEG Signals. In *Advances in Neural Information Processing Systems*, 2025. URL https://papers.nips.cc/paper_files/paper/2025/hash/fca83589e85cb061631b7ebc5db5d6bd-Abstract-Conference.html.

Lin Jiang, Qingshan She, Jiale Xu, Haiqi Xu, Duanpo Wu, and Zhenzhong Kuang. Subject-Aware Multi-Granularity Alignment for Zero-Shot EEG-to-Image Retrieval, 2026. URL https://arxiv.org/abs/2604.17782.

Sangmin Jo, Wootaek Jeong, Da-Woon Heo, Yoohwan Hwang, and Heung-Il Suk. HyFI: Hyperbolic Feature Interpolation for Brain-Vision Alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026. URL https://ojs.aaai.org/index.php/AAAI/article/view/37476.

Jun Mo Kim, Woohyeok Choi, Yebin Choi, Songha Kim, and Tae Eui Kam. CroSS: Cross-subject Visual Brain Decoding via Soft Alignment and Subject-aware Batch Normalization. In *ACM Symposium on Applied Computing*, 2026. doi: 10.1145/3748522.3779936. URL https://doi.org/10.1145/3748522.3779936.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word Translation Without Parallel Data. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=H196sainb.

Dongyang Li, Chen Wei, Shiying Li, Jiachen Zou, and Quanying Liu. Visual Decoding and Reconstruction via EEG Embeddings with Guided Diffusion. In *Advances in Neural Information Processing Systems*, 2024. URL https://papers.nips.cc/paper_files/paper/2024/hash/ba5f1233efa77787ff9ec015877dbd1f-Abstract-Conference.html.

Jiaxiang Liu, Jiawei Du, Xupeng Chen, Guoqi Li, Jiang Cai, Simon Fong, and Mingkun Xu. MindAdapter: Few-Shot Parameter-Efficient Residual Calibration of Cross-Subject Brain-to-Visual Decoding Models, 2026. URL https://arxiv.org/abs/2605.24679.

Mu Nan, Muquan Yu, Weijian Mai, Jacob S. Prince, Hossein Adeli, Rui Zhang, Jiahang Cao, Benjamin Becker, John A. Pyles, Margaret M. Henderson, Chunfeng Song, Nikolaus Kriegeskorte, Michael J. Tarr, Xiaoqing Hu, and Andrew F. Luo. Meta-Learning In-Context Enables Training-Free Cross Subject Brain Decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Nan_Meta-Learning_In-Context_Enables_Training-Free_Cross_Subject_Brain_Decoding_CVPR_2026_paper.html.

Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. Hubs in Space: Popular Nearest Neighbors in High-Dimensional Data. *Journal of Machine Learning Research*, 2010. URL https://www.jmlr.org/papers/v11/radovanovic10a.html.

Nona Rajabi, Antonio H. Ribeiro, Miguel Vasco, Farzaneh Taleb, Mårten Björkman, and Danica Kragic. Human-Aligned Image Models Improve Visual Decoding from the Brain. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025. URL https://proceedings.mlr.press/v267/rajabi25a.html.

Pedro Luiz Coelho Rodrigues, Christian Jutten, and Marco Congedo. Riemannian Procrustes Analysis: Transfer Learning for Brain-Computer Interfaces. *IEEE Transactions on Biomedical Engineering*, 2019. doi: 10.1109/TBME.2018.2889705. URL https://doi.org/10.1109/TBME.2018.2889705.

Paul Steven Scotti, Mihir Tripathy, Cesar Torrico, Reese Kneeland, Tong Chen, Ashutosh Narang, Charan Santhirasegaran, Jonathan Xu, Thomas Naselaris, Kenneth A. Norman, and Tanishq Mathew Abraham. MindEye2: Shared-Subject Models Enable fMRI-To-Image With 1 Hour of Data. In *Proceedings of the 41st International Conference on Machine Learning*, 2024. URL https://proceedings.mlr.press/v235/scotti24a.html.

Yonghao Song, Bingchuan Liu, Xiang Li, Nanlin Shi, Yijun Wang, and Xiaorong Gao. Decoding Natural Images from EEG for Object Recognition. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/d0865cbe51d35ec322f9af9db7806fc7-Abstract-Conference.html.

Xuan-The Tran, Thien-Nhan Vo, Son-Tung Vu, Thoa-Thi Tran, Manh-Dat Nguyen, Thomas Do, and Chin-Teng Lin. Inter- and Intra-Subject Variability in EEG: A Systematic Survey, 2026. URL https://arxiv.org/abs/2602.01019.

Yimu Wang, Xiangru Jian, and Bo Xue. Balance Act: Mitigating Hubness in Cross-Modal Retrieval with Query and Gallery Banks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. doi: 10.18653/v1/2023.emnlp-main.652. URL https://aclanthology.org/2023.emnlp-main.652/.

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*, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wu_Bridging_the_Vision-Brain_Gap_with_an_Uncertainty-Aware_Blur_Prior_CVPR_2025_paper.html.

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*, 2026. URL https://ojs.aaai.org/index.php/AAAI/article/view/38844.

Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-Adapter: Training-Free Adaption of CLIP for Few-Shot Classification. In *European Conference on Computer Vision*, 2022. URL https://www.ecva.net/papers/eccv_2022/papers_ECCV/html/154_ECCV_2022_paper.php.

Jiawen Zheng, Haonan Jia, Ming Li, Yuhui Zheng, Yufeng Zeng, Yang Gao, and Chen Liang. Learning Brain Representation with Hierarchical Visual Embeddings. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/c5506f154e5f2203a7dca2bee2faeced-Abstract-Conference.html.

Jiahao Zhou, Chenghao Xu, Wei Wang, Erkun Yang, and Cheng Deng. EEGiT: Teaching Vision Transformers to Understand the EEG signal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Zhou_EEGiT_Teaching_Vision_Transformers_to_Understand_the_EEG_signal_CVPR_2026_paper.html.

A APPENDIX

A.1 PAIRCAL ALGORITHM

Algorithm 1: SGC + MDC (PairCal) for EEG-to-image retrieval

Input: Paired support set $\mathcal{S} = \{(b_i, v_i)\}_{i=1}^{K}$ ($K \geq 2$), with EEG embeddings b_i and image embeddings v_i; test EEG query q; candidate image set $\mathcal{Y} = \{y_j\}_{j=1}^{M}$; fixed hyperparameters $\tau, \rho, \gamma, k, \lambda$ ($1 \leq k, \tau, \rho > 0$)

Output: Candidates ranked by calibrated score; Top-1, Top-5, and MRR over the test queries

Normalize b_i, v_i, q, y_j according to the original retrieval protocol;

Step 1: Support-guided query correction (SGC);

Function $\texttt{SGC}(z, \mathcal{T})$ **:**

$D_{\mathcal{T}} \leftarrow [\, v - b \,]_{(b,v)\in\mathcal{T}}$;

$G_{\mathcal{T}} \leftarrow [\exp((b^\top b' - 1)/\tau)\,]_{(b,v),(b',v')\in\mathcal{T}}$;

$A_{\mathcal{T}} \leftarrow G_{\mathcal{T}} + \rho I_{|\mathcal{T}|}$;

$g_z \leftarrow [\exp((z^\top b - 1)/\tau)\,]_{(b,v)\in\mathcal{T}}$;

$u_z \leftarrow \mathrm{solve}(A_{\mathcal{T}}, g_z^\top)$; // stable linear solve

$\Delta_z \leftarrow u_z^\top D_{\mathcal{T}}$;

return $\mathrm{normalize}(z + \gamma\Delta_z)$;

$\hat{q} \leftarrow \texttt{SGC}(q, \mathcal{S})$;

Step 2: Leave-one-out reference bank for MDC;

for $i \leftarrow 1$ **to** K **do**

$\mathcal{S}_{-i} \leftarrow \mathcal{S} \setminus \{(b_i, v_i)\}$;

$r_i \leftarrow \texttt{SGC}(b_i, \mathcal{S}_{-i})$; // exclude pair i

end

$R \leftarrow [\, r_1; \ldots; r_K \,]$;

Step 3: Candidate hubness estimation;

for $j \leftarrow 1$ **to** M **do**

$s_{ij} \leftarrow r_i^\top y_j \quad (i = 1, \ldots, K)$;

$\mathcal{I}_j \leftarrow \mathrm{TopKIndices}(\{s_{ij}\}_{i=1}^{K}, \min(k, K))$;

$h(y_j) \leftarrow |\mathcal{I}_j|^{-1} \sum_{i\in\mathcal{I}_j} s_{ij}$;

end

Step 4: Calibrated retrieval;

for $j \leftarrow 1$ **to** M **do**

$s_{\mathrm{query}}(j) \leftarrow \hat{q}^\top y_j$;

$\mathrm{score}(j) \leftarrow s_{\mathrm{query}}(j) - \lambda h(y_j)$;

end

ranked_candidates $\leftarrow \mathrm{SortDescending}(\{(y_j, \mathrm{score}(j))\}_{j=1}^{M})$;

Step 5: Evaluation across N test queries;

Let p_t be the one-based rank of the matching image for test query t;

Top-1 $\leftarrow N^{-1} \sum_{t=1}^{N} \mathbf{1}[p_t = 1]$, Top-5 $\leftarrow N^{-1} \sum_{t=1}^{N} \mathbf{1}[p_t \leq 5]$;

MRR $\leftarrow N^{-1} \sum_{t=1}^{N} p_t^{-1}$;

A.2 HYPERPARAMETER SELECTION

All hyperparameters are selected exclusively on source-validation episodes that exclude the current held-out target subject. For SGC, we search

$$\begin{aligned} &\tau \in \{0.03, 0.07, 0.15, 0.3\}, \\ &\rho \in \{0.001, 0.01, 0.1, 1.0\}, \\ &\gamma \in \{0.05, 0.1, 0.25, 0.5, 1.0\}. \end{aligned} \tag{9}$$

After fixing SGC, we select the MDC coefficient from

$$\lambda_{\mathrm{MDC}} \in \{0.1, 0.25, 0.5, 0.75, 1.0\}. \tag{10}$$

The candidate-neighborhood grid depends on the support budget:

$$\begin{aligned} K = 10 &: \{1, 2, 5, 10\}, \\ K = 20 &: \{1, 5, 10, 20\}, \\ K = 50 &: \{2, 10, 12, 20, 25, 50\}, \\ K = 100 &: \{5, 10, 20, 25, 50, 100\}, \\ K = 200 &: \{10, 20, 30, 50, 100, 200\}. \end{aligned} \tag{11}$$

The selection objective is

$$\frac{1}{2}\left(\text{Top-1} + \text{Top-5}\right) \tag{12}$$

on source-validation episodes. Ties are resolved by prioritizing Top-1, followed by Top-5, and then the predefined conservative parameter ordering. The held-out target subject's test results never enter parameter selection.

A.3 Preprocessing of the THINGS-EEG Dataset

The THINGS-EEG dataset contains EEG recordings from 10 participants, acquired using a 64-channel system. Following the BrainHive protocol, the raw signals were filtered to 0.1–$100\,\mathrm{Hz}$, yielding 63 channels sampled at $1000\,\mathrm{Hz}$. Each trial was segmented from 0 to $1000\,\mathrm{ms}$ after stimulus onset and baseline-corrected using the mean signal over the preceding $200\,\mathrm{ms}$.

The signals were then downsampled to $250\,\mathrm{Hz}$. Seventeen occipital and parietal channels associated with visual processing were retained: P7, P5, P3, P1, Pz, P2, P4, P6, P8, PO7, PO3, POz, PO4, PO8, O1, Oz, and O2. To improve the signal-to-noise ratio, repetitions of the same image were averaged within each participant. Each training image had four repetitions, whereas each test image had 80 repetitions. This procedure yielded 16,540 training samples and 200 test samples per participant.

A.4 Full Query-Side Calibration Results

Table 5: Full query-side calibration ablation across support-set sizes K. All values are percentages. Mean Res. and Weighted Res. denote Mean Residual and Similarity-Weighted Residual, respectively.

Top-1 accuracy (%)					
K	Frozen	Mean Res.	Weighted Res.	Direct KRR	SGC
10	20.50	24.43	17.64	2.34	**24.99**
20	20.50	25.93	18.03	3.94	**26.79**
50	20.50	27.01	19.73	9.31	**29.48**
100	20.50	27.39	20.66	14.83	**31.78**
200	20.50	27.44	20.78	20.53	**33.44**
Top-5 accuracy (%)					
K	Frozen	Mean Res.	Weighted Res.	Direct KRR	SGC
10	44.92	52.40	43.43	9.17	**53.74**
20	44.92	54.81	44.28	14.91	**56.73**
50	44.92	57.38	46.93	28.88	**61.29**
100	44.92	57.80	48.22	40.41	**63.04**
200	44.92	57.98	48.58	50.56	**65.55**
MRR (%)					
K	Frozen	Mean Res.	Weighted Res.	Direct KRR	SGC
10	33.11	37.95	30.43	7.56	**38.74**
20	33.11	40.24	31.06	11.20	**40.96**
50	33.11	41.30	33.05	19.71	**44.24**
100	33.11	41.68	34.09	27.77	**46.35**
200	33.11	41.76	34.33	34.81	**48.17**

Table 5 reports the complete Top-1, Top-5, and MRR results for the query-side calibration ablations summarized in the main paper. All methods use the same frozen encoders, paired support sets, test queries, and candidate galleries. Method-specific hyperparameters are selected exclusively on source-validation subjects. Mean Residual applies a single support-averaged residual to every query. Similarity-Weighted Residual directly averages support residuals using normalized query–support kernel affinities, without solving the regularized kernel system. Direct KRR predicts the complete visual embedding from the EEG embedding, whereas SGC preserves the pretrained query and predicts only its residual correction.

The full results exhibit the same pattern across all three retrieval metrics. Mean Residual improves over the frozen model but approaches a plateau as the support budget increases. Direct KRR is particularly ineffective in the low-support regime, consistent with the difficulty of estimating a complete cross-modal mapping from limited target-subject data. Similarity-Weighted Residual also remains below SGC across all support budgets and metrics. These results indicate that SGC's gain is not explained by a global residual shift, direct kernel mapping, or normalized local residual averaging. Instead, the results support preserving the pretrained EEG representation and estimating its correction through the regularized support kernel system.

We also evaluated an anchored variant of Direct KRR that interpolates between the pretrained EEG query and the directly predicted visual embedding. Source-only validation selected an interpolation coefficient of $\gamma = 1$ for every support budget, making this variant identical to Direct KRR; we therefore do not report it as a separate row.

A.5 Additional MDC Mechanism Analysis

This section provides additional evidence for the mechanism underlying Meta-Distractor Calibration (MDC). Unless otherwise specified, statistics are first averaged across checkpoints or support-set realizations within each held-out subject. Confidence intervals are then computed over subjects using $10{,}000$ bootstrap resamples, treating the held-out subject as the statistical unit.

Does the support-derived score predict false positives? For each candidate y_j, we compute its false-positive frequency $F_j^{(5)}$ from the Top-5 retrieval results of SGC on the disjoint target-subject test queries. We then measure the Spearman correlation between $F_j^{(5)}$ and the MDC attraction score $h(y_j)$. Test labels are used only to compute the diagnostic statistic $F_j^{(5)}$; they are not used to construct h, select hyperparameters, or calibrate retrieval scores.

Table 6: Association between the support-derived attraction score and false-positive retrievals. The last two columns report the mean false-positive frequency among candidates in the top and bottom 20% according to h, respectively.

K	$\rho(h, F^{(5)})$	95% CI	Top-20% h	Bottom-20% h
10	0.332	[0.274, 0.397]	7.75	2.47
20	0.387	[0.329, 0.448]	7.66	2.16
50	0.446	[0.411, 0.481]	7.63	1.84
100	0.501	[0.460, 0.547]	7.89	1.67
200	0.519	[0.492, 0.546]	8.05	1.47

Table 6 shows a positive correlation for every support-set size, with all confidence intervals excluding zero. Moreover, candidates assigned larger attraction scores are substantially more likely to appear as false positives. The separation between the high- and low-attraction groups becomes more pronounced as K increases. These results indicate that h, although estimated exclusively from the paired support set, provides a meaningful ranking of candidate-specific attraction on unseen test queries.

Does MDC reduce false-positive concentration? We measure hubness using H_{10}, the fraction of all false-positive Top-5 occurrences attributable to the 10% most frequently retrieved incorrect candidates. We additionally report the skewness of the candidate-level false-positive frequency dis-

Table 7: Effect of MDC on the concentration of false-positive retrievals. ΔH_{10} and $\Delta\mathrm{Skew}$ denote PairCal minus SGC. Negative values indicate a reduction in false-positive concentration.

K	SGC H_{10}	PairCal H_{10}	ΔH_{10}	95% CI	ΔSkew
10	0.422	0.402	-0.0199	[-0.0273, -0.0132]	-0.087[†]
20	0.393	0.365	-0.0282	[-0.0358, -0.0203]	-0.182
50	0.368	0.340	-0.0274	[-0.0314, -0.0236]	-0.208
100	0.358	0.320	-0.0382	[-0.0467, -0.0305]	-0.273
200	0.358	0.312	-0.0461	[-0.0535, -0.0397]	-0.350

[†]The bootstrap confidence interval for the skewness difference at $K = 10$ includes zero. Exact confidence intervals for the remaining skewness differences are omitted here because only their point estimates are reported.

Table 8: Spearman correlation between candidate-attraction estimates constructed from the support set and the offline oracle attraction statistic computed from the complete target test-query set. The oracle diagnostic is available only for $K \in \{50, 100, 200\}$.

K	Raw support	Full support	Leave-one-out
50	0.276	0.411	**0.486**
100	0.155	0.465	**0.566**
200	0.132	0.462	**0.604**

tribution. For both measures, a negative difference between PairCal and SGC indicates reduced concentration.

MDC reduces H_{10} for every K, and all corresponding confidence intervals exclude zero. The reduction generally becomes larger with additional support examples, reaching 0.0461 at $K = 200$. The mean skewness difference follows the same direction and increases in magnitude from -0.087 to -0.350. However, because the skewness confidence interval includes zero at $K = 10$, we do not claim a reliable reduction in this statistic for the smallest support set. Overall, H_{10} provides the more stable evidence that MDC distributes false-positive retrievals across a broader set of candidates rather than allowing a small set of hubs to dominate the rankings.

Why use leave-one-out pseudo-unseen queries? MDC estimates candidate attraction using leave-one-out (LOO) pseudo-unseen queries. To assess whether this construction is necessary, we compare three reference-query banks:

- **Raw support:** the original support EEG embeddings without query calibration;
- **Full support:** support queries calibrated using all support pairs, including their own paired examples;
- **LOO:** each support query is calibrated while excluding its own EEG–image pair.

We compare the attraction score produced by each construction with an offline oracle candidate-attraction statistic computed from the complete target test-query set. The oracle is used only for this diagnostic analysis and is never used for calibration, hyperparameter selection, or retrieval.

As shown in Table 8, the LOO construction is more strongly correlated with the oracle candidate-attraction statistic than either raw or full-support reference queries. The difference becomes larger as the support set grows. Calibrating a support query with its own pair can make the reference query unrealistically easy and therefore less representative of an unseen test query. The LOO construction avoids this self-pair information and more closely approximates the deployment condition in which the paired image of a test EEG is unknown.

Candidate-permutation control. A reduction in hubness could, in principle, arise merely because MDC changes the marginal distribution of similarity scores. To distinguish this explanation from a genuinely candidate-specific correction, we randomly permute the estimated h values across candidates while keeping the same set of penalty values. We use 50 permutations for every experimental setting and do not retune the penalty strength.

Table 9: Retrieval advantage of the correctly assigned attraction score over randomly permuted candidate assignments. Values are percentage-point differences between PairCal with the correct h and the mean over 50 permutations.

K	ΔTop-1	ΔTop-5
10	+2.87	+3.96
20	+3.05	+3.88
50	+3.02	+3.32
100	+5.08	+5.06
200	+7.24	+7.11

The correctly assigned attraction scores outperform the permutation control for every K. Because permutation preserves the empirical distribution and scale of h, this result cannot be explained solely by applying a generic similarity offset or by changing the overall score variance. Instead, the retrieval gain depends on associating each attraction penalty with the corresponding image candidate.

Taken together, the analyses establish three aspects of the proposed mechanism: (i) the support-derived attraction score predicts candidate false-positive frequency on disjoint test queries; (ii) applying MDC reduces the concentration of false-positive retrievals; and (iii) both the leave-one-out construction and the candidate-specific assignment of h contribute to this behavior. These findings support the interpretation of MDC as a support-based estimator and suppressor of candidate-specific attraction, while avoiding the stronger claim that it eliminates hubness entirely.

A.6 Subject-Wise Comparison

We compare SATTC and PairCal across all ten held-out subjects at $K = 200$, reporting Top-1 accuracy in Figure 3 and Top-5 accuracy and mean reciprocal rank (MRR) in Figure 4. For SATTC, K denotes the number of unlabeled target test queries used jointly for transductive calibration; for PairCal, it denotes the number of paired support examples drawn from the target subject's training-image split. Both methods are evaluated on the same complete 200-query test set and 200-image gallery. SATTC is therefore a transductive reference rather than a resource-matched baseline.

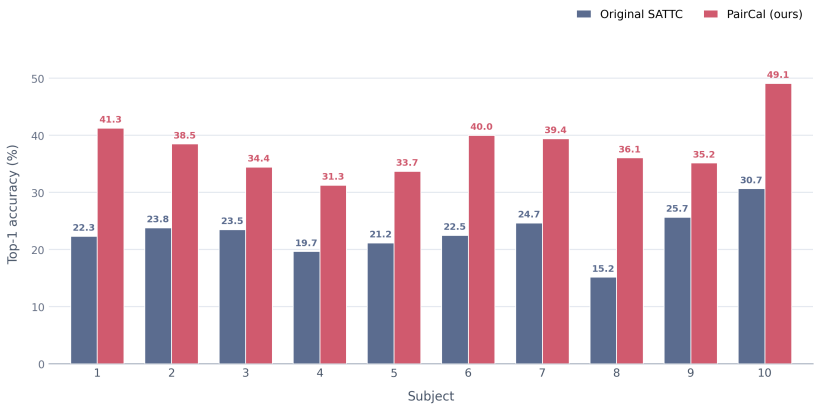

Figure 3: Subject-wise Top-1 accuracy (%) of SATTC and PairCal at $K = 200$. PairCal improves accuracy for all ten held-out subjects.

PairCal improves Top-1 accuracy for every subject, increasing the cross-subject average from 22.93% to 37.90%, a gain of 14.97 percentage points. Subject-level gains range from 9.5 to 20.9 percentage points, indicating that the overall improvement is not driven by a small subset of subjects.

The same pattern holds for Top-5 accuracy and MRR. PairCal increases the cross-subject average Top-5 accuracy from 49.87% to 69.80%, a gain of 19.93 percentage points, and MRR from 35.86% to 52.36%, a gain of 16.50 percentage points. Both metrics improve for every held-out subject.

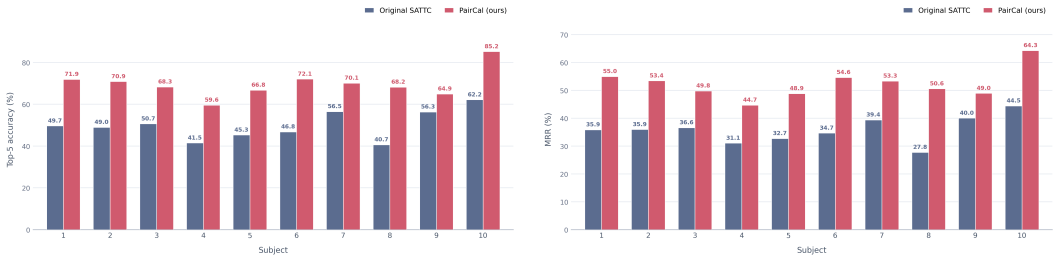

Figure 4: Subject-wise Top-5 accuracy (%, left) and MRR (%, right) of SATTC and PairCal at $K = 200$. PairCal improves both metrics for all ten held-out subjects.

These consistent gains are particularly relevant to cross-subject EEG retrieval, where substantial inter-subject variability can cause aggregate results to obscure failures on individual subjects. Despite differences in baseline retrieval difficulty, PairCal provides improvements across all subjects and all three retrieval metrics.

A.7 Generalization Across EEG Encoders

To examine whether the performance gains of PairCal depend on a particular EEG encoder, we repeat the complete cross-subject calibration experiments using multiple EEG backbones. These encoders represent the spatiotemporal structure of EEG signals in different ways, and their frozen models exhibit clear differences in both absolute retrieval performance and cross-subject generalization. Therefore, if PairCal were effective only with a specific base encoder, its applicability would be substantially limited.

Table 10: **Generalization across EEG encoders on THINGS-EEG2.** Each entry reports Top-1 / Top-5 / MRR (%).

EEG Encoder	K	Brainhive	SGC	**PairCal**
Brain-MLP	100	20.50 / 44.92 / 33.11	31.78 / 63.04 / 46.35	**35.08 / 66.50 / 49.47**
	200	20.50 / 44.92 / 33.11	33.44 / 65.55 / 48.17	**37.90 / 69.81 / 52.36**
BrainProjection	100	21.85 / 48.30 / 34.72	30.94 / 62.08 / 45.41	**33.89 / 65.52 / 48.41**
	200	21.85 / 48.30 / 34.72	32.59 / 64.57 / 47.15	**36.47 / 68.83 / 51.14**
NICE	100	24.18 / 52.85 / 37.98	26.02 / 57.68 / 40.80	**29.14 / 61.73 / 44.10**
	200	24.18 / 52.85 / 37.98	28.45 / 59.92 / 43.12	**32.16 / 64.76 / 47.06**
TS-Conv	100	23.13 / 53.57 / 37.50	25.53 / 56.67 / 40.14	**28.12 / 60.68 / 43.10**
	200	23.13 / 53.57 / 37.50	27.32 / 59.14 / 42.12	**31.14 / 63.78 / 46.12**

As shown in Table 10, PairCal consistently improves retrieval performance across different EEG encoders. These results indicate that the calibration mechanism of PairCal is not tailored to a specific EEG encoder and can provide consistent relative improvements across encoders with different representation characteristics.

A.8 Extension to MEG-to-Image Retrieval

We additionally evaluate whether the paired-support calibration formulation extends beyond EEG to magnetoencephalography (MEG). This experiment is intended as a cross-modality generalization study rather than a primary benchmark: we retain the same support-based calibration construction, but select its hyperparameters using MEG source-side validation episodes.

Experimental protocol. We conduct the experiment on THINGS-MEG with four held-out subjects. Visual representations are extracted by a frozen BrainCLIP encoder. For each target subject,

we evaluate support budgets $K \in \{10, 20, 50, 100, 200\}$ and compare the frozen retrieval model, source-selected SGC, and SGC followed by MDC. The hyperparameters of both SGC and MDC are selected exclusively on MEG training-image episodes that exclude the current outer target subject. The official target test set is loaded only after hyperparameter selection is complete. Thus, no target-test labels or target-test query statistics are used for model selection or calibration.

For aggregation, support-set seeds and independently trained realizations are first averaged within each target subject, after which the four held-out subjects receive equal weight. Importantly, the MEG experiment performs modality-specific source-side hyperparameter selection; it therefore evaluates whether the calibration *formulation* transfers from EEG to MEG, rather than whether EEG-selected hyperparameters transfer zero-shot across modalities.

Table 11: **Extension to MEG-to-image retrieval on THINGS-MEG.** Each entry reports Top-1 / Top-5 / MRR (%). Results are averaged equally over four held-out subjects after averaging support-set seeds and training realizations within each subject. Δ reports the improvement of PairCal over SGC in Top-1 / Top-5 (%)accuracy, in percentage points.

K	Brainhive	SGC	PairCal(ours)	Δ vs. SGC
10	5.00 / 15.21 / 11.72	4.83 / 15.49 / 11.69	4.96 / 16.18 / 12.04	+0.12 / +0.69
20	5.00 / 15.21 / 11.72	5.03 / 16.37 / 12.25	5.18 / 17.32 / 12.65	+0.15 / +0.94
50	5.00 / 15.21 / 11.72	5.26 / 17.47 / 12.79	5.78 / 18.15 / 13.37	+0.51 / +0.68
100	5.00 / 15.21 / 11.72	5.44 / 18.54 / 13.36	5.71 / 18.64 / 13.57	+0.26 / +0.10
200	5.00 / 15.21 / 11.72	5.37 / 19.12 / 13.66	5.96 / 19.97 / 14.24	+0.58 / +0.85

Results. Table 11 provides evidence that the same paired-support calibration formulation remains useful when the neural modality changes from EEG to MEG. The effect is more modest than in the EEG experiments, particularly in the lowest-support regime. SGC does not improve Top-1 accuracy at $K = 10$ and yields only small Top-1 gains over the frozen model at larger support budgets. Its Top-5 accuracy, however, improves over the frozen model for all evaluated values of K.

Adding MDC on top of SGC yields a positive aggregate improvement in Top-1, Top-5, and MRR at every support budget. The Top-1 gain over SGC ranges from 0.12 to 0.58 percentage points, while the Top-5 gain ranges from 0.10 to 0.94 points. At $K = 200$, SGC + MDC reaches 5.96% Top-1, 19.97% Top-5, and 14.24 MRR, compared with 5.00%, 15.21%, and 11.72 for the frozen model.

The candidate-side gain is therefore retained under the MEG retrieval geometry even though the query-side calibration benefit is substantially smaller than in the EEG experiments. We view these results as evidence that the support-based calibration construction can extend beyond EEG, rather than as evidence of modality-independent performance. In particular, the experiment uses only four MEG subjects and performs MEG-specific source-side hyperparameter selection, so broader cross-modality conclusions would require evaluation on additional datasets and neural recording settings.

A.9 Qualitative retrieval examples.

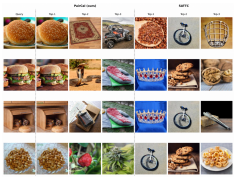

Figure 5: Qualitative retrieval examples. Each row shows a query image and the top three retrieved images from PairCal (ours) and SATTC.

In addition, Figure 5 presents several retrieval examples together with the top three candidate images.