PERCEPTUAL-PRIOR ALIGNMENT WITH SUBJECT-ADAPTIVE TEMPERATURE FOR EEG-TO-IMAGE RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

EEG-to-image retrieval aims to identify perceived visual stimuli from non-invasive neural recordings, but remains challenging because EEG is noisy, spatially coarse, and highly variable across trials and subjects. Existing contrastive approaches commonly align EEG with a single, fixed visual representation using one temperature shared across all subjects. This formulation overlooks both trial-level variation in the visual information represented by EEG and subject-level heterogeneity in neural separability. We propose Perceptual-Prior Alignment with Subject-Adaptive Temperature (PPA-SAT), a retrieval-oriented framework that calibrates cross-modal alignment at two complementary resolutions. At the trial level, Perceptual-Prior Alignment (PPA) retains a heterogeneous bank of raw, blurred, noisy, low-resolution, and mosaic visual candidates. It constructs an EEG-conditioned target through residual attention fusion. At the source-subject level, Subject-Adaptive Temperature (SAT) complements this alignment by learning lightweight subject-specific offsets to a shared contrastive temperature, enabling subject-aware calibration with negligible parameter overhead. Experiments on THINGS-EEG 200-way zero-shot retrieval demonstrate consistent improvements under both intra- and inter-subject protocols. PPA-SAT achieves 91.7% Top-1 and 99.0% Top-5 intra-subject accuracy, exceeding the strongest prior method by 9.1 and 1.3 percentage points, respectively. Under leave-one-subject-out evaluation, it achieves 45.8% Top-1 and 72.7% Top-5 accuracy, outperforming the strongest baseline by 21.8 and 17.1 points. These results demonstrate the complementary value of trial-adaptive visual alignment and subject-aware contrastive calibration for robust EEG-to-image retrieval.

1 INTRODUCTION

Visual perception is represented in the brain at multiple levels, from local structure and spatial frequency to object-level semantics. Recovering such information from electroencephalography (EEG) has become an important problem in neural decoding and brain–computer interfaces. EEG-to-image retrieval is particularly challenging because scalp EEG has limited spatial resolution, relatively low signal-to-noise ratio, and substantial variation across trials and subjects. The THINGS-EEG benchmark has enabled large-scale evaluation of visual decoding from EEG Grootswagers et al. (2022), while recent methods have increasingly formulated the task as cross-modal representation learning Song et al. (2024); Li et al. (2024).

A common formulation maps an EEG trial and its corresponding image into a shared representation space and optimizes a contrastive objective. This strategy is effective, but it makes an implicit assumption: the visual representation used as the alignment target is fixed for every EEG trial. Recent work challenges this assumption. NeuroBridge uses perceptually motivated transformations to provide alternative visual representations Zhang et al. (2026b), UBP uses uncertainty-aware visual processing Wu et al. (2025), and Shallow Alignment shows that intermediate visual representations can retain information that is less accessible in final semantic embeddings Du et al. (2026). These results indicate that visual granularity matters for neural–visual alignment. However, existing approaches generally use a fixed representation, a predefined transformation, or visual-side weighting

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.

rather than asking a more direct question: *which combination of visual representations is most compatible with a particular EEG query?*

A second source of mismatch occurs across subjects. Contrastive learning commonly uses a shared temperature to control logit sharpness. Individualized temperature has been investigated in general contrastive learning Qiu et al. (2023), but EEG–vision models generally retain a common scale despite subject-dependent differences in neural response quality and cross-modal separability. A single scale can therefore be unnecessarily restrictive when multiple subjects are jointly optimized.

We address these two forms of heterogeneity with **Perceptual-Prior Alignment with Subject-Adaptive Temperature (PPA-SAT)**. PPA treats the original image and four perceptually motivated transformations as a heterogeneous candidate bank. For each EEG query and candidate image, it predicts weights over these representations and constructs a query-conditioned visual target. SAT complements PPA by parameterizing the contrastive temperature as a shared global component plus a source-subject-specific offset. The two mechanisms therefore operate at different resolutions: PPA adapts *what visual information is aligned* at the trial level, whereas SAT adapts *how sharply the alignment is optimized* at the subject level.

Our contributions are:

- We formulate **Perceptual-Prior Alignment (PPA)** as a query-conditioned visual fusion mechanism over heterogeneous perceptual representations, allowing the effective visual target to vary with each EEG query.
- We introduce **Subject-Adaptive Temperature (SAT)**, a low-capacity subject-conditioned temperature parameterization that adapts contrastive sharpness across source subjects without introducing subject-specific encoders.
- We establish a two-level calibration perspective for EEG–vision alignment and evaluate PPA and SAT independently and jointly under intra-subject and leave-one-subject-out protocols.
- On THINGS-EEG, PPA-SAT reaches 91.7%/99.0% Top-1/Top-5 intra-subject retrieval and 45.8%/72.7% Top-1/Top-5 leave-one-subject-out retrieval under the reported evaluation protocol.

2 Related Work

2.1 EEG–Vision Representation Learning

Early EEG decoding studies demonstrated that visual categories and semantic information can be recovered from neural activity Spampinato et al. (2017). More recent work formulates neural decoding as cross-modal representation learning. BRAVL jointly aligns brain, visual, and linguistic representations Du et al. (2023); NICE applies contrastive EEG–image alignment for zero-shot recognition Song et al. (2024); ATM extends EEG–vision learning toward retrieval and generation Li et al. (2024); and Neural-MCRL develops multimodal contrastive representation learning Li et al. (2025). MB2C reduces the neural–visual modality gap through bidirectional cycle consistency Wei et al. (2024). These approaches establish the effectiveness of shared representation spaces, but typically rely on a fixed visual target for each image.

2.2 Perceptual Priors and Visual Granularity

Several studies show that visual representations at different granularities can be useful for neural alignment. UBP introduces uncertainty-aware blurring Wu et al. (2025), NeuroBridge constructs multiple perceptual alternatives using blur, noise, low resolution, and mosaic transformations Zhang et al. (2026b), and Liu et al. learn selection over multiple blur levels Liu et al. (2026). CognitionCapturerPro combines complementary visual modalities through uncertainty-weighted fusion Zhang et al. (2026a). Shallow Alignment further demonstrates the usefulness of intermediate visual features Du et al. (2026), while ViEEG decomposes images into contour, foreground, and contextual representations Liu et al. (2025).

The common theme is that a single semantic embedding may not be the optimal neural alignment target. PPA differs in treating multiple visual representations as a candidate bank whose contribution

is conditioned jointly on the EEG query and candidate image. Thus, visual augmentation becomes part of the learned retrieval function rather than only a data-level transformation.

2.3 Contrastive Temperature Calibration

Temperature controls the concentration of the softmax distribution in contrastive learning Oord et al. (2018). CLIP uses a global logit scale Radford et al. (2021), while individualized temperature parameterizations have been proposed to account for sample-level differences Qiu et al. (2023). PPA-SAT adapts this idea to heterogeneous EEG subjects by introducing one scalar offset per source subject while retaining a shared global temperature. This design separates subject-level calibration from trial-level visual-target selection.

3 Method

3.1 Problem Formulation

Let $\mathcal{D} = \{(x_{E,n}, x_{I,n}, s_n)\}_{n=1}^{N}$ denote paired EEG–image observations, where $x_{E,n} \in \mathbb{R}^{C\times T}$ is an EEG trial, $x_{I,n}$ is the corresponding image, and $s_n \in \{1, \dots, S\}$ is the source-subject identity. An EEG encoder f_E maps a trial to $h_{E,n} = f_E(x_{E,n})$. For a batch of B paired observations, i indexes EEG queries and j indexes candidate images; $i = j$ denotes the paired image.

We focus on 200-way zero-shot retrieval. The concepts in the evaluation set are disjoint from training concepts, and a test EEG query is ranked against the complete candidate gallery.

3.2 Perceptual-Prior Candidate Bank

Following the intermediate-feature motivation of Shallow Alignment Du et al. (2026), we use a frozen InternViT-6B visual encoder and extract an intermediate representation at layer 28. As shown in Fig. 1, for each image, we construct five views: the identity image, Gaussian blur, Gaussian noise, low-resolution degradation, and mosaic degradation. Let $t_0(x) = x$ and $t_1, \dots, t_4$ denote the four transformations. We compute

$$h_{I,n,k} = \mathrm{Pool}\Big(f_I^{(28)}(t_k(x_{I,n}))\Big) \in \mathbb{R}^{d_V}, \quad k = 0, \dots, 4, \tag{1}$$

where $d_V = 3200$. Visual features are precomputed and frozen during EEG model training.

The EEG and visual features are projected into a common space $\mathbb{R}^{d_C}$ with $d_C = 512$:

$$z_{E,i} = p_E(h_{E,i}), \qquad z_{I,j,k} = p_I^{(k)}(h_{I,j,k}). \tag{2}$$

3.3 Query-Conditioned Perceptual-Prior Alignment

The central idea of PPA is to treat the five visual views as candidate representations rather than fixed augmentations. Lightweight projections produce query and candidate keys,

$$q_i = g_E(z_{E,i}), \qquad r_{j,k} = g_I(z_{I,j,k}). \tag{3}$$

The relevance of view k for EEG query i and candidate image j is

$$\alpha_{i,j,k} = \frac{\exp(\langle q_i, r_{j,k}\rangle/\sqrt{d_C})}{\sum_{\ell=0}^{4}\exp(\langle q_i, r_{j,\ell}\rangle/\sqrt{d_C})}. \tag{4}$$

The query-conditioned representation is

$$\widetilde{z}_{I,i,j} = \sum_{k=0}^{4} \alpha_{i,j,k} z_{I,j,k}. \tag{5}$$

To stabilize early training, we interpolate this representation with the uniform prior mean:

$$z_{I,i,j} = (1-\lambda)\bar{z}_{I,j} + \lambda\widetilde{z}_{I,i,j}, \qquad \bar{z}_{I,j} = \frac{1}{5}\sum_{k=0}^{4} z_{I,j,k}, \quad \lambda = \sigma(\eta), \tag{6}$$

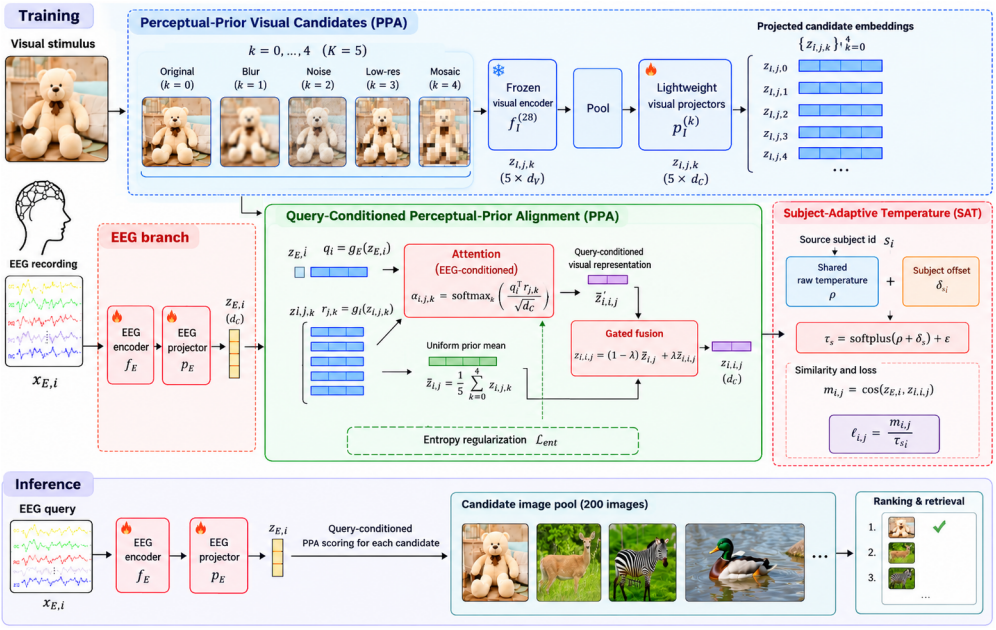


Figure 1: Overview of PPA-SAT. A frozen visual encoder produces five perceptual-prior representations for each candidate image. PPA uses the EEG query to obtain a query-conditioned visual representation for each candidate. SAT supplies a source-subject-conditioned temperature to the contrastive objective. During leave-one-subject-out inference, the held-out subject uses the shared global temperature because no target-specific offset is learned.

where η is learnable and λ is initialized to 0.05. The query–candidate similarity is cosine similarity:

$$m_{i,j} = \left\langle \frac{z_{E,i}}{\|z_{E,i}\|_2}, \frac{z_{I,i,j}}{\|z_{I,i,j}\|_2} \right\rangle. \tag{7}$$

Importantly, PPA is a *query-conditioned scorer*, not a conventional fixed-image encoder. At inference, the same operation is performed for every EEG–image pair in the gallery. The ground-truth identity is never used to compute $\alpha_{i,j,k}$.

We regularize the ground-truth attention weights with a negative-entropy term,

$$\mathcal{L}_{\text{ent}} = \frac{1}{B}\sum_{i=1}^{B}\sum_{k=0}^{4} \alpha_{i,i,k}\log(\alpha_{i,i,k} + \epsilon), \tag{8}$$

with a small coefficient. Since $\mathcal{L}_{\text{ent}} = -H(\alpha)$, minimizing this term discourages premature collapse to a single view.

3.4 Subject-Adaptive Temperature

PPA addresses trial-level visual-target variation. SAT addresses a different source of variation: source-subject heterogeneity. We parameterize the temperature for subject s as

$$\tau_s = \text{softplus}(\rho + \delta_s) + \epsilon, \tag{9}$$

where ρ is a shared raw-temperature parameter and δ_s is a learned scalar offset for source subject s. All offsets are initialized at zero and the shared temperature is initialized to 0.07.

For an EEG query i from source subject s_i, the retrieval logit is

$$\ell_{i,j} = \frac{m_{i,j}}{\tau_{s_i}}. \tag{10}$$

The induced conditional distribution over candidate images is

$$p(j \mid i) = \frac{\exp(\ell_{i,j})}{\sum_{r=1}^{B} \exp(\ell_{i,r})}. \tag{11}$$

Thus, τ_s controls the sharpness of the candidate distribution while the shared representation space remains unchanged.

For leave-one-subject-out evaluation, the held-out subject has no learned δ_s. We therefore use the shared temperature, i.e., $\delta_{\text{test}} = 0$, and do not perform target-specific optimization. This distinction is important: SAT is a source-subject training calibration mechanism, not test-time subject adaptation.

3.5 Contrastive Objective

The primary retrieval objective is the EEG-to-image contrastive loss

$$\mathcal{L}_{E\to I} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(m_{i,i}/\tau_{s_i})}{\sum_{j=1}^{B} \exp(m_{i,j}/\tau_{s_i})}. \tag{12}$$

For the bidirectional auxiliary objective used in the reported experiments, the image-to-EEG term is written with the temperature associated with each source subject in the corresponding candidate row:

$$\mathcal{L}_{I\to E} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(m_{i,i}/\tau_{s_i})}{\sum_{j=1}^{B} \exp(m_{j,i}/\tau_{s_j})}. \tag{13}$$

The complete objective is

$$\mathcal{L} = \frac{1}{2}\left(\mathcal{L}_{E\to I} + \mathcal{L}_{I\to E}\right) + \beta_{\text{ent}}\mathcal{L}_{\text{ent}}, \qquad \beta_{\text{ent}} = 10^{-3}. \tag{14}$$

The EEG encoder, projection heads, PPA modules, fusion coefficient, global temperature, and source-subject offsets are optimized jointly, while the visual backbone remains frozen.

3.6 Why Two-Level Calibration?

PPA and SAT address distinct axes of variation. PPA operates within a subject and across trials: two EEG responses can contain different mixtures of visual detail even when generated by the same subject. SAT operates across source subjects: the same representation-learning objective may have different effective sharpness for subjects with different neural–visual separability. Their combination therefore targets two separate components of the retrieval function,

$$\ell_{i,j} = \frac{\text{sim}\left(z_{E,i}, z_I(x_{I,j} \mid x_{E,i})\right)}{\tau_{s_i}}. \tag{15}$$

This decomposition is the central design principle of PPA-SAT.

4 Experiments

4.1 Dataset and Evaluation Protocol

We evaluate on THINGS-EEG Grootswagers et al. (2022), containing 10 subjects, 1,654 training concepts, and 200 disjoint test concepts. Following prior work Song et al. (2024); Wu et al. (2025); Zhang et al. (2026b); Du et al. (2026), we use 200-way zero-shot image retrieval. For intra-subject evaluation, each subject is trained and evaluated independently. For leave-one-subject-out (LOSO) evaluation, the model is trained using source subjects and evaluated on the held-out subject.

We report Top-1 and Top-5 retrieval accuracy. For each EEG query, all 200 candidate images are scored and ranked. Top-1 is the fraction of queries for which the paired image receives rank one; Top-5 is the fraction for which it appears among the five highest-scoring candidates.

Table 1: Intra-subject Top-1 and Top-5 image retrieval accuracy (%) on THINGS-EEG. Each model is trained and evaluated independently for each subject.

Method	Metric	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Avg.
ATM Li et al. (2024)	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
Neural-MCRL Li et al. (2025)	Top-1	27.5	28.5	37.0	35.0	22.5	31.5	31.5	42.0	30.5	37.5	32.4
	Top-5	64.0	61.5	69.0	66.0	51.5	61.0	62.5	74.5	59.5	71.0	64.1
UBP Wu et al. (2025)	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
NeuroBridge Zhang et al. (2026b)	Top-1	50.0	63.2	61.6	61.4	54.8	69.7	62.7	71.2	64.0	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
HyFI Jo et al. (2026)	Top-1	60.6	65.9	69.5	66.5	55.0	74.4	68.4	78.9	66.0	77.0	68.2
	Top-5	85.3	94.0	93.9	89.8	86.0	95.0	91.3	96.9	90.6	96.4	91.9
AVDE Dai et al. (2026)	Top-1	25.0	24.1	27.5	29.8	25.4	33.5	27.4	41.7	26.1	39.5	30.0
	Top-5	55.2	51.0	58.6	54.7	50.3	60.3	55.2	71.3	52.1	73.0	58.2
EEGiT Zhou et al. (2026)	Top-1	82.3	73.7	67.5	64.9	55.9	66.8	70.2	84.8	51.7	86.5	70.4
	Top-5	97.6	97.9	95.7	93.5	91.3	92.0	98.1	98.6	87.3	99.4	95.1
Shallow Du et al. (2026)	Top-1	75.0	87.5	83.2	79.5	74.6	89.9	78.5	86.9	81.3	89.3	82.6
	Top-5	94.3	98.9	98.2	96.1	96.4	99.3	97.3	99.4	97.8	99.1	97.7
PPA-SAT (Ours)	Top-1	**95.0**	**90.0**	**90.0**	**92.0**	**88.0**	**92.5**	**90.5**	**96.0**	**87.5**	**95.5**	**91.7**
	Top-5	**99.5**	98.0	**100.0**	**99.0**	**98.0**	99.0	99.0	**100.0**	98.0	**99.5**	**99.0**

4.2 Implementation Details

We use the EEGProject encoder used in the original experimental implementation. EEG is restricted to 17 occipital–parietal channels and a 0–250 ms post-stimulus window. Training uses AdamW with batch size 1024 and learning rate 10^{-4}. The contrastive dimension is 512. PPA uses five visual candidates and initializes λ at 0.05; $\beta_{\text{ent}} = 10^{-3}$. SAT offsets are initialized to zero and the shared temperature to 0.07.

For the visual pathway, we use the frozen InternViT-6B intermediate representation at layer 28. The extracted representation has dimension 3200. Visual features are precomputed before EEG model optimization. This avoids repeatedly executing the large visual backbone during contrastive training.

4.3 Compared Methods

We compare with representative EEG-based visual decoding methods, including Neural-MCRL Li et al. (2025), UBP Wu et al. (2025), NeuroBridge Zhang et al. (2026b), HyFI Jo et al. (2026), EEGiT Zhou et al. (2026), AVDE Dai et al. (2026), and Shallow Alignment Du et al. (2026). These methods cover contrastive representation learning, perceptual-prior modeling, transformer-based EEG encoding, and intermediate-layer alignment. Results reported below follow the evaluation protocol and values available in the current experimental record. Before submission, every baseline should be verified under an identical concept split and candidate-gallery definition; literature-reported and independently reproduced results should be explicitly distinguished in the final version.

4.4 Main Results

PPA-SAT reaches 91.7% Top-1 and 99.0% Top-5 averaged across the ten intra-subject evaluations. Relative to Shallow Alignment, the reported gain is 9.1 percentage points for Top-1 and 1.3 points for Top-5. The per-subject values show that the gain is not confined to a single subject: the reported PPA-SAT Top-1 value is higher than the listed baselines for each subject.

Under LOSO, PPA-SAT reaches 45.8% Top-1 and 72.7% Top-5. Relative to the listed EEGiT baseline, this corresponds to gains of 21.8 and 17.1 percentage points. The improvement is especially relevant to cross-subject generalization because the model is trained on source subjects and the held-out subject does not receive a learned subject-specific temperature offset.

Table 2: Leave-one-subject-out (LOSO) Top-1 and Top-5 image retrieval accuracy (%) on THINGS-EEG. Each column corresponds to the held-out subject.

Method	Metric	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Avg.
ATM Li et al. (2024)	Top-1	10.5	7.1	11.9	14.7	7.0	11.1	16.1	15.0	4.9	20.5	11.9
	Top-5	26.8	24.8	33.8	39.4	23.9	35.8	43.5	40.3	22.7	46.5	33.8
UBP Wu et al. (2025)	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
Neural-MCRL Li et al. (2025)	Top-1	13.0	12.0	14.5	12.5	11.5	13.5	14.0	18.5	13.5	17.0	14.0
	Top-5	31.5	30.5	35.5	35.5	29.0	35.5	36.0	38.5	32.5	39.0	34.4
HyFI Jo et al. (2026)	Top-1	16.2	20.0	7.5	18.8	9.7	15.5	11.0	13.2	13.3	25.3	15.1
	Top-5	35.8	47.7	26.7	41.3	27.5	33.8	34.5	30.3	38.8	55.7	37.2
AVDE Dai et al. (2026)	Top-1	14.1	17.0	9.1	15.2	12.5	17.3	7.4	18.5	13.2	18.0	14.3
	Top-5	32.2	38.4	21.8	32.5	32.4	38.6	20.4	40.1	33.6	39.3	32.9
NeuroBridge Zhang et al. (2026b)	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
Shallow Du et al. (2026)	Top-1	23.5	30.6	10.0	19.5	18.1	22.7	18.6	17.3	23.0	34.4	21.8
	Top-5	53.2	60.4	28.1	48.3	43.2	49.8	46.0	46.1	54.8	62.0	49.4
EEGiT Zhou et al. (2026)	Top-1	26.9	22.9	14.5	20.5	27.1	26.0	22.0	20.2	26.5	33.7	24.0
	Top-5	57.2	51.4	41.8	50.2	56.4	63.4	48.5	50.7	67.3	69.4	55.6
PPA-SAT (Ours)	Top-1	**65.0**	**57.0**	**20.5**	**51.0**	**58.5**	**40.0**	**31.5**	**32.5**	**41.5**	**60.0**	**45.8**
	Top-5	**86.5**	**79.0**	**47.0**	**74.5**	**84.5**	**68.5**	**59.0**	**64.5**	**75.0**	**88.0**	**72.7**

Table 3: Component ablation under LOSO. Values are averaged over the ten held-out subjects.

Configuration	Top-1	Top-5
Baseline	19.00	45.90
Baseline + SAT	22.30	53.00
Intermediate only	21.80	49.40
Intermediate + SAT	30.40	63.00
PPA	42.50	72.20
PPA-SAT	**45.75**	**72.65**

4.5 Ablation Studies

PPA and SAT. Table 3 isolates the contributions of intermediate alignment, PPA, and SAT under LOSO. The baseline obtains 19.0%/45.9% Top-1/Top-5. Adding SAT raises these values to 22.3%/53.0%. Intermediate-layer alignment reaches 21.8%/49.4%, and combining it with SAT reaches 30.4%/63.0%. PPA increases performance to 42.5%/72.2%, while the full PPA-SAT model reaches 45.75%/72.65%. The cross-product of these components supports their complementary roles.

Perceptual-prior removal. Table 4 reports the existing leave-one-prior-out analysis. Using all five candidates yields 79.20% Top-1 and 95.75% Top-5, compared with 64.90% and 88.35% when only the original representation is retained. Removing individual transformations decreases performance to varying degrees, indicating that the candidate bank contains complementary information.

Effect of Query-Conditioned Visual Fusion. Table 5 examines how different strategies for combining multiple perceptual visual representations affect EEG-to-image retrieval. Using only the original visual representation achieves 64.90% Top-1 and 88.35% Top-5 accuracy. Introducing multiple representations through uniform mean fusion improves performance to 76.80% and 95.65%, respectively, indicating that the complementary visual representations provide useful information beyond the original image representation. Residual mean fusion further improves Top-1 and Top-5 accuracy to 81.00% and 96.40%, respectively. Adding entropy regularization, however, reduces performance to 79.20% Top-1 and 95.75% Top-5 accuracy, suggesting that explicitly constraining the fusion-weight distribution does not consistently improve retrieval.

Table 4: Leave-one-prior-out ablation under intra-subject evaluation. Values are averaged across subjects.

Visual candidates	Top-1	Top-5
Original only	64.90	88.35
All except Gaussian blur	74.20	93.15
All except Gaussian noise	77.00	94.60
All except low resolution	77.05	94.60
All except mosaic	73.70	93.15
All five candidates	**79.20**	**95.75**

Table 5: Intra-subject image retrieval accuracy (%) on THINGS-EEG under different visual fusion strategies.

Fusion Strategy	**Top-1 (%)**	**Top-5 (%)**
Original Representation	64.90	88.35
Uniform Mean	76.80	95.65
Residual Mean	81.00	96.40
Residual Mean + Entropy	79.20	95.75
EEG-conditioned Attention (PPA)	**91.70**	**99.00**

In contrast, the proposed EEG-conditioned attention mechanism (PPA) achieves 91.70% Top-1 and 99.00% Top-5 accuracy. Compared with the original representation, PPA yields gains of 26.80 and 10.65 percentage points in Top-1 and Top-5 accuracy, respectively. More importantly, PPA improves Top-1 accuracy by 10.70 percentage points over residual mean fusion, demonstrating that the benefit does not arise solely from aggregating multiple visual representations. Instead, conditioning the fusion weights on the EEG–image pair enables the model to selectively emphasize the visual representation that is most compatible with the current neural query. These results support the central design of PPA: perceptual representations are more effective when their contributions are dynamically adapted to the individual EEG query and candidate image rather than combined using fixed or globally shared weights.

4.6 Qualitative Analysis

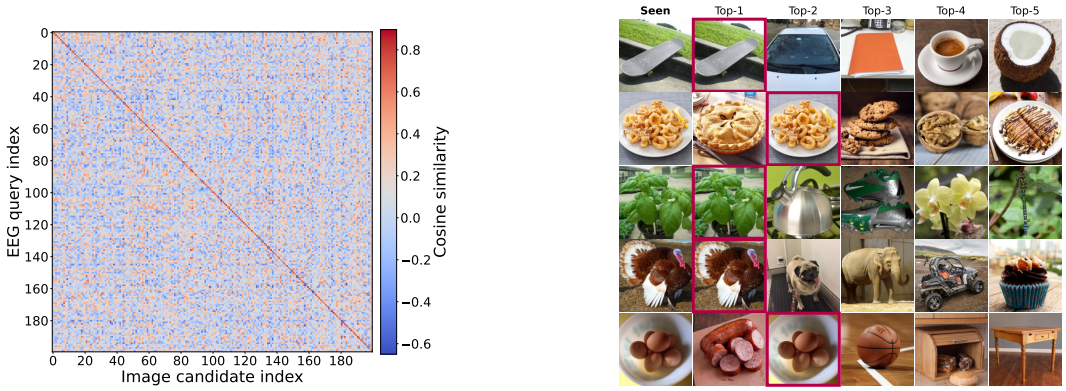
Figure 2 shows the existing qualitative analysis for Subject 08. The similarity matrix exhibits a diagonal structure, while the retrieval examples place the ground-truth image near the top of the ranking. These examples align with the quantitative retrieval results and show that the query-conditioned scoring function can distinguish the paired image from visually related alternatives.

5 Discussion

The results support a two-level view of neural–visual alignment. At the trial level, different EEG responses can retain different mixtures of visual information. PPA addresses this variation by treating perceptually motivated representations as candidate alignment targets and conditioning their contribution on the EEG query. At the source-subject level, neural–visual separability can differ across subjects. SAT addresses this variation by allowing the contrastive logit scale to vary across source subjects while preserving a shared representation space.

The ablation results indicate that these mechanisms contribute through different pathways. SAT improves both the baseline and intermediate-layer model, while PPA produces a substantially larger gain when multiple visual candidates are available. Their combination gives the strongest reported LOSO Top-1 result. This supports the central design principle that visual-target calibration and contrastive-temperature calibration should be treated as distinct problems.

A useful conceptual distinction is that PPA is query-conditioned whereas SAT is subject-conditioned. PPA changes the representation used to compare a particular EEG query with a particular image candidate. SAT changes the scale applied to the resulting similarity distribution for a source subject.



(a) Similarity matrix.

(b) Top-5 retrieval.

Figure 2: Qualitative EEG-to-image retrieval for Subject 08.

Keeping these levels separate makes the model lightweight while allowing heterogeneity to enter at the resolution where it occurs.

6 Limitations

First, PPA requires query–candidate fusion during retrieval and is therefore less computationally efficient than ranking fixed image embeddings. The large visual backbone is frozen and its features are precomputed, but the candidate-specific fusion still scales with the number of gallery images.

Second, the perceptual-prior bank is manually specified. The current five representations provide useful diversity, but they are not an exhaustive basis of visual variation. Automatically learned priors could reduce dependence on hand-designed transformations.

Third, SAT does not learn a target-specific offset for an unseen subject in LOSO evaluation. Instead, the held-out subject uses the shared global temperature. Test-time calibration or uncertainty-aware adaptation could potentially improve this setting, but such procedures are outside the current evaluation protocol.

Finally, the present evaluation focuses on THINGS-EEG and image retrieval. Broader datasets, larger galleries, and reconstruction or generation tasks are needed to establish how well the two-level calibration principle transfers beyond this benchmark.

7 Conclusion

We introduced PPA-SAT, a two-level calibration framework for EEG-to-image retrieval. PPA performs query-conditioned fusion over original and perceptually motivated visual representations, allowing the effective visual target to adapt to each EEG query. SAT complements this mechanism with a source-subject-conditioned contrastive temperature. The resulting formulation separates trial-level visual-target variation from subject-level contrastive calibration without requiring subject-specific encoders. On the reported THINGS-EEG experiments, PPA-SAT achieves strong intra-subject retrieval and substantial gains under leave-one-subject-out evaluation. More broadly, the results suggest that conditional representation construction and adaptive contrastive calibration are complementary tools for learning from heterogeneous neural–visual correspondences.

AI Assistance Disclosure

During preparation of this manuscript, large language models were used for language refinement and presentation editing. The authors remain responsible for the scientific claims, experimental results, implementation, and final manuscript.

REFERENCES

Sicheng Dai, Hongwang Xiao, Shan Yu, and Qiwei Ye. Autoregressive visual decoding from eeg signals. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 101075–101097, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/a4dfcbcfa1f0425cd18aafa35a68019a-Paper-Conference.pdf.

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.

Yang Du, Siyuan Dai, Yonghao Song, Paul M Thompson, Haoteng Tang, and Liang Zhan. Deep models, shallow alignment: Uncovering the granularity mismatch in neural decoding. *arXiv preprint arXiv:2601.21948*, 2026.

Tijl Grootswagers, Ivy Zhou, Amanda K Robinson, Martin N Hebart, and Thomas A Carlson. Human eeg recordings for 1,854 concepts presented in rapid serial visual presentation streams. *Scientific Data*, 9(1):3, 2022.

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*, volume 40, pp. 5575–5583, 2026.

Dongyang Li, Chen Wei, Shiying Li, Jiachen Zou, Haoyang Qin, and Quanying Liu. Visual decoding and reconstruction via EEG embeddings with guided diffusion. *arXiv preprint arXiv:2403.07721*, 2024.

Y. Li, Z. Kang, S. Gong, W. Dong, W. Zeng, H. Yan, W. T. Siok, and N. Wang. Neural-MCRL: Neural multimodal contrastive representation learning for EEG-based visual decoding. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2025.

Minxu Liu, Donghai Guan, Chuhang Zheng, Chunwei Tian, Jie Wen, and Qi Zhu. ViEEG: Hierarchical visual neural representation for EEG brain decoding. *arXiv preprint arXiv:2505.12408*, 2025.

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. doi: 10.1609/aaai.v40i21.38813.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Zi-Hao Qiu, Quanqi Hu, Zhuoning Yuan, Denny Zhou, Lijun Zhang, and Tianbao Yang. Not all semantics are created equal: Contrastive self-supervised learning with automatic temperature individualization. *arXiv preprint arXiv:2305.11965*, 2023.

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.

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.

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.

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. doi: 10.1145/3664647.3681292.

Table 6: Ablation of the prior-entropy regularization coefficient β on intra-subject THINGS-EEG retrieval. Results are averaged over all 10 subjects. H_{norm} denotes the normalized entropy of the learned perceptual-prior weights.

β	**Top-1 (%)**	**Top-5 (%)**	H_{norm}
10^{-4}	89.20	95.75	0.98842
10^{-3}	**91.35**	**96.70**	0.99591
10^{-2}	90.70	96.45	0.99558

H. Wu, Q. Li, C. Zhang, Z. He, and X. 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 (CVPR)*, pp. 2246–2257, 2025.

Kaifan Zhang, Lihuo He, Junjie Ke, Yuqi Ji, Lukun Wu, Lizi Wang, and Xinbo Gao. CognitionCapturerPro: Towards high-fidelity visual decoding from EEG/MEG via multi-modal information and asymmetric alignment. *arXiv preprint arXiv:2603.12722*, 2026a.

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, 2026b.

J. Zhou, C. Xu, W. Wang, E. Yang, and C. Deng. EEGiT: Teaching vision transformers to understand EEG signals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 40438–40447, 2026.

A Appendix

Sensitivity to prior-entropy regularization. Table 6 examines the sensitivity of PPA-SAT to the prior-entropy regularization coefficient β on intra-subject THINGS-EEG retrieval. The model achieves its best performance at $\beta = 10^{-3}$, obtaining 91.35% Top-1 and 96.70% Top-5 accuracy. Reducing β to 10^{-4} decreases Top-1 and Top-5 accuracy to 89.20% and 95.75%, respectively, while increasing it to 10^{-2} results in a smaller decrease to 90.70% and 96.45%. These results indicate that the regularization strength affects retrieval performance, with an intermediate value providing the most favorable trade-off.

The normalized entropy remains high across all settings (0.98842–0.99591), indicating that the learned perceptual-prior weights remain relatively distributed rather than collapsing onto a single representation. Interestingly, the highest retrieval performance at $\beta = 10^{-3}$ coincides with the highest normalized entropy (0.99591). This suggests that, for the evaluated setting, maintaining sufficiently diverse contributions from the perceptual-prior representations is beneficial for retrieval. Overall, the relatively small performance variation across the tested coefficients also indicates that PPA-SAT is not highly sensitive to the precise choice of β within this range.