REPRESENTATION, NORMALIZATION, AND SEQUENTIAL CALIBRATION FOR CROSS-SUBJECT EEG-TO-IMAGE RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

EEG-to-image retrieval is commonly formulated as a cross-modal representation-learning problem, but cross-subject deployment introduces an additional train–test mismatch: subject-specific normalization statistics available during source training are unavailable for an unseen subject at cold start. We present the Representation–Normalization–Sequential Calibration (RNSC) framework, which treats shared representation learning, source-subject normalization, and unseen-subject calibration as coordinated stages of a single retrieval space. RNSC couples a shared, subject-agnostic visual reference with an EEG encoder that factorizes temporal structure before channel-semantic mixing and aggregates repeated trials after latent encoding. During source training, EEG latents are optimized in subject-standardized coordinates; at deployment, target moments are estimated sequentially from a finite source prior and past unlabeled target EEG under a strict predict-before-update protocol. On THINGS-EEG2, the shared representation achieves state-of-the-art intra-subject 200-way retrieval, with 93.67% Top-1 and 99.36% Top-5, while the complete RNSC framework reaches 48.10% Top-1 and 76.34% Top-5 under inter-subject evaluation. Controlled ablations, latent-geometry diagnostics, and cold-start analyses disentangle the complementary roles of representation learning, training-time normalization, and sequential target-statistic estimation.

1 INTRODUCTION

EEG-to-image retrieval connects the study of human visual representations with non-invasive brain–computer interfaces by matching a visually evoked EEG response to its eliciting image (Du et al., 2023; Song et al., 2024; Jiang et al., 2026; Tang et al., 2026). Moving from subject-specific models to deployment on subjects absent from training changes the nature of the problem: performance depends not only on the discriminative quality of EEG and visual features, but also on substantial subject-dependent distribution shift (Jayaram et al., 2016; Zanini et al., 2018; Huang & Zhu, 2026). Cross-subject retrieval therefore asks how a shared retrieval space can remain stimulus-discriminative while transferring reliably to a new subject.

Existing work mainly advances along two directions. Representation-oriented methods improve what visual information EEG should align with, using intermediate layers, hierarchical visual embeddings, region-level sub-features, or multi-granularity supervision (Du et al., 2026; Zheng et al., 2026; Tang et al., 2026; Zhu et al., 2026; Jiang et al., 2026), while stronger EEG encoders further improve neural representations (Zhou et al., 2026). Complementary approaches address subject variation through subject-aware supervision or test-time calibration: SAMGA introduces subject-dependent visual routing during training, whereas SATTC applies subject-adaptive whitening and similarity-structure calibration to frozen embeddings (Jiang et al., 2026; Huang & Zhu, 2026). Together, these studies show that both representation quality and latent-statistic handling matter, motivating a coupled question: if an EEG representation is learned in normalized coordinates, how should the same coordinate system be instantiated for an unseen subject?

This issue is most acute at cold start. The first target queries provide too little history to estimate reliable subject-specific moments, so normalization based only on target observations is initially

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.

ill-conditioned. Permanently retaining source statistics, however, prevents the model from adapting to the target subject. Cross-subject EEG retrieval is therefore not only about *what* representation is learned, but also about the statistical coordinate system in which it is learned and how those coordinates are estimated as target evidence accumulates.

Our central view is to preserve a common visual reference while handling subject-dependent distributional variation on the EEG branch. The shared representation should retain stimulus-discriminative geometry; training-time normalization should control subject-dependent location and scale; and deployment-time calibration should estimate the moments required to apply the same normalization to an unseen subject. Under this view, representation learning, source-subject normalization, and target-subject calibration form consecutive stages of a common retrieval coordinate system.

We instantiate this principle with the Shared Temporal-basis Aggregated Representation encoder (STAR-EEG) and the Unified Multi-level Visual Anchor (UMVA) for the shared representation. We then use Training-Compatible Subject-Adaptive Whitening (TC-SAW) during source training and Source-Prior Prequential Calibration (SPPC) for unseen-subject deployment. STAR-EEG factorizes shared temporal structure before channel-semantic mixing and postpones repeat aggregation until after complete latent encoding, while UMVA provides a common multi-depth InternViT candidate space. During inter-subject training, TC-SAW places EEG latents in subject-standardized coordinates inside the contrastive objective. At deployment, SPPC initializes target normalization from source moments and updates it only from previously observed target EEG latents, following a strict predict-before-update protocol. As target history grows, the source contribution decays and normalization becomes increasingly target-specific. Together, these components form our Representation–Normalization–Sequential Calibration (RNSC) framework for cross-subject EEG-to-image retrieval.

The paper makes three contributions:

- **A representation–normalization view of cross-subject retrieval.** We separate the roles of a fixed visual reference, subject-dependent EEG statistics, and deployment-time moment estimation, providing a single coordinate-system perspective for source training and unseen-subject deployment.
- **Shared representation learning in subject-standardized coordinates.** STAR-EEG combines channel-shared temporal factorization with representation-space repeat aggregation, while TC-SAW optimizes source-subject EEG latents directly in subject-standardized coordinates against the same visual reference.
- **Sequential cold-start estimation for unseen subjects.** SPPC treats target normalization as an online moment-estimation problem: a finite source prior stabilizes early queries and automatically yields to target-specific statistics as observations accumulate. Five independent runs over ten leave-one-subject-out folds, paired ablations, and cold-start analyses characterize these complementary effects.

2 Related Work

EEG–vision representation and visual targets. BraVL and NICE established multimodal and contrastive brain–vision alignment (Du et al., 2023; Song et al., 2024), while ATM connects EEG-to-CLIP alignment with diffusion-based reconstruction (Li et al., 2024). Later work broadens supervision through semantic consistency, multimodal transfer, uncertainty-aware targets, and cognitive priors (Chen et al., 2024a; Zhang et al., 2025; Wu et al., 2025; Zhang et al., 2026). Motivated by hierarchical cortical and deep-network representations (Felleman & Van Essen, 1991; Yamins et al., 2014; Cichy et al., 2016), Shallow Alignment, hierarchical visual embeddings, HCF, RAHS, and SAMGA study intermediate layers, complementary encoders, EEG-visible depths, region-level sub-features, and subject-aware multi-granularity targets (Du et al., 2026; Zheng et al., 2026; Tang et al., 2026; Zhu et al., 2026; Jiang et al., 2026). EEGNet, EEG Conformer, EEGiT, and more recent staged, multiview, and subject-aware encoders strengthen the neural branch (Lawhern et al., 2018; Song et al., 2023; Zhou et al., 2026; Gao et al., 2026; Yadav et al., 2026; Gong et al., 2026). These works primarily improve what is represented and aligned. Our focus is complementary: how unseen-subject latent statistics should be instantiated in the same shared space.

Cross-subject normalization for EEG retrieval. Cross-subject transfer is difficult because EEG distributions vary across subjects and sessions (Jayaram et al., 2016; Zanini et al., 2018). EEGiT transfers visual priors into the EEG encoder (Zhou et al., 2026); NeuroBridge uses augmentation and bidirectional semantic alignment (Zhang et al., 2026); SUP-MCRL combines multi-scale enhancement with inter-band interaction (Gong et al., 2026); and SAMGA introduces subject dependence into visual supervision (Jiang et al., 2026). SATTC targets unseen-subject embedding shift through adaptive whitening and similarity-structure calibration (Huang & Zhu, 2026). These approaches show that subject variation can be addressed through the encoder, supervision target, embedding statistics, or similarity geometry. We focus on learning the EEG representation in subject-standardized coordinates during source training, then estimating the corresponding normalization statistics for an unseen subject at deployment.

Online and test-time calibration under domain shift. AdaBN replaces source normalization moments with target statistics (Li et al., 2018); TENT adapts normalization-related parameters at test time (Wang et al., 2021); and TTN interpolates source and test statistics (Lim et al., 2023). EATA, CoTTA, and related methods further study reliability, forgetting, small batches, and temporally correlated streams (Niu et al., 2022; Wang et al., 2022; Niu et al., 2023; Tomar et al., 2024). SATTC brings label-free calibration to unseen-subject EEG retrieval (Huang & Zhu, 2026). Our evaluation adopts a prequential protocol (Dawid, 1984): before predicting target query t, calibration may use only source-training statistics and target queries $1, \dots, t-1$. This defines sequential calibration separately from transductive access to the complete target set.

3 METHOD

3.1 CROSS-SUBJECT RETRIEVAL UNDER SEQUENTIAL TARGET ACCESS

EEG-to-image retrieval aims to identify the visual stimulus that elicited a neural response. Let $\mathcal{D}_{\mathrm{src}} = \{(e_i, x_i, s_i)\}$ contain source-training EEG e_i, paired images x_i, and subject identities s_i, and let $\mathcal{C} = \{x^{(1)}, \dots, x^{(M)}\}$ be the evaluation candidate bank. EEG and visual encoders map the two modalities into a shared embedding space, and each EEG query ranks candidate images by cosine similarity. The training objective preserves stimulus-discriminative EEG–vision correspondence so that the paired image is ranked ahead of the remaining candidates.

Intra-subject evaluation trains and tests on the same subject and primarily measures representation quality. Inter-subject evaluation trains on source subjects and deploys the model to an unseen target subject $s^\star$, whose EEG representation may exhibit subject-dependent distribution shift. We use a strict predict-before-update protocol: before ranking query t, calibration may use source-training statistics and target observations from queries $1, \dots, t-1$ only; the current target latent $\mathbf{u}_t^e$ enters the calibration state only after its prediction. This follows the prequential evaluation principle (Dawid, 1984). We write $\mathcal{N}(\mathbf{a}) = \mathbf{a}/\|\mathbf{a}\|_2$ for ℓ_2 normalization, $\mathbf{u}$ for a pre-normalization latent, and $\mathbf{z} = \mathcal{N}(\mathbf{u})$ for a retrieval embedding.

3.2 RNSC OVERVIEW: FIXED VISUAL REFERENCE AND SUBJECT-STANDARDIZED EEG COORDINATES

Figure 1 summarizes the RNSC framework. We keep one shared visual candidate space for all subjects and place representation learning, normalization, and calibration on the EEG branch. STAR-EEG models channel-shared temporal structure and aggregates repeated trials after complete latent encoding, while UMVA maps frozen InternViT features into a fixed visual reference. During source training for inter-subject evaluation, TC-SAW standardizes source-subject EEG latents before the contrastive objective. During deployment, SPPC estimates the moments required for the same EEG-side normalization from a source prior and the accumulated target history. The three stages therefore operate in one retrieval space, with subject dependence confined to EEG-side statistics.

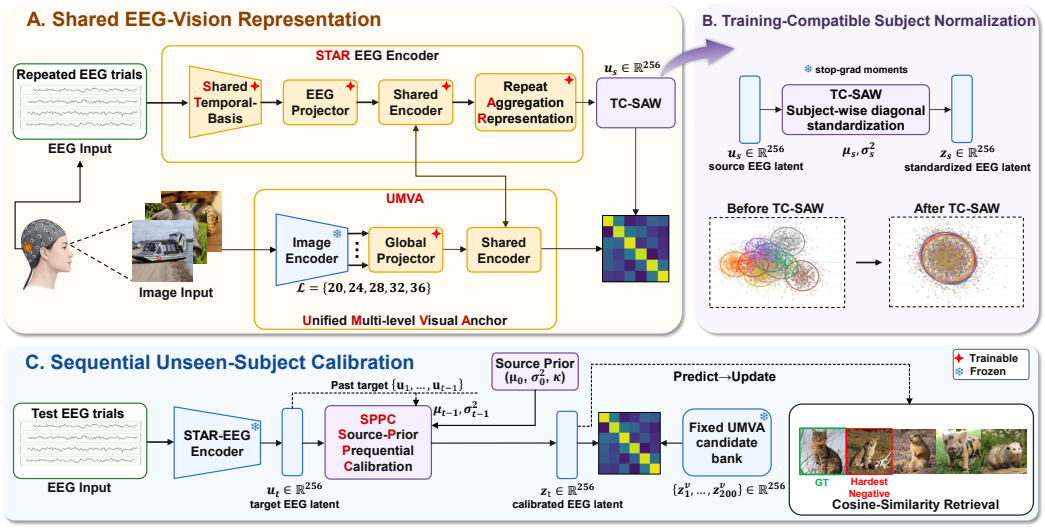

Figure 1: Overview of the RNSC framework. During training (top), STAR-EEG and UMVA establish a shared EEG–vision retrieval space, while TC-SAW standardizes source-subject EEG latents before the contrastive objective. During evaluation (bottom), SPPC combines a fixed source prior with past target-query statistics to normalize each incoming query and updates the target state only after prediction. The insets illustrate the temporal-basis encoder and the effect of training-time normalization.

3.3 Shared EEG–Vision Representation: UMVA and STAR-EEG

UMVA. For image x_i, frozen InternViT (Chen et al., 2024b) features from depths $\mathcal{L} = \{20, 24, 28, 32, 36\}$ are concatenated and mapped once into a subject-agnostic candidate space:

$$\mathbf{v}_i = \mathrm{Concat}_{\ell \in \mathcal{L}}\, \mathbf{v}_i^{\ell}, \qquad \mathbf{z}_i^{v} = \mathcal{N}(h(p_v(\mathbf{v}_i)))\,. \tag{1}$$

The resulting visual bank is shared by all subjects throughout training and retrieval. Its role is to provide a stable stimulus reference rather than to encode subject identity, so cross-subject variation is handled on the EEG branch.

STAR-EEG. For repetition r of object i, let $\mathbf{E}_{i,r} \in \mathbb{R}^{C\times T}$ with $C = 17$ and $T = 250$. Instead of learning a fully unconstrained map from flattened channel–time samples, STAR-EEG first projects every channel through a shared temporal basis $\mathbf{A} \in \mathbb{R}^{K\times T}$ and then mixes the resulting channel–basis coefficients with $\mathbf{B} \in \mathbb{R}^{D\times CK}$:

$$\begin{aligned} \mathbf{q}_{i,r} &= \mathbf{B}\ \mathrm{vec}\big(\mathbf{E}_{i,r}\mathbf{A}^{\top}\big) + \mathbf{b}, \\ \mathbf{f}_{i,r} &= \mathrm{LN}(\mathbf{q}_{i,r} + \mathrm{Drop}(\mathbf{W}_{\mathrm{res}}\,\mathrm{GELU}(\mathbf{q}_{i,r})))\,. \end{aligned} \tag{2}$$

We use $K = 192$ and $D = 256$. The bias-free $\mathbf{A}$ is shared across channels, so reusable temporal structure is extracted before channel- and output-specific semantic mixing. Each repetition then passes independently through feature dropout, the EEG projector p_e, and the shared residual encoder h. Only after this complete path are the R_i repetitions averaged in the pre-normalization representation space:

$$\mathbf{u}_i^{e} = \frac{1}{R_i}\sum_{r=1}^{R_i} h(p_e(\mathrm{Drop}(\mathbf{f}_{i,r})))\,, \qquad \mathbf{z}_i^{e} = \mathcal{N}(\mathbf{u}_i^{e}). \tag{3}$$

Thus, STAR-EEG separates shared temporal structure from channel-semantic mixing and postpones repeat aggregation from sensor space to the fully encoded latent representation.

Training minimizes one-way EEG-to-image InfoNCE (van den Oord et al., 2018; Radford et al., 2021; Chen et al., 2020) over a minibatch of size N_{b}:

$$\mathcal{L}_{\mathrm{ret}} = -\frac{1}{N_{\mathrm{b}}}\sum_{i=1}^{N_{\mathrm{b}}} \log \frac{\exp\big((\mathbf{z}_i^{e})^{\top}\mathbf{z}_i^{v}/\tau\big)}{\sum_{j=1}^{N_{\mathrm{b}}}\exp\big((\mathbf{z}_i^{e})^{\top}\mathbf{z}_j^{v}/\tau\big)}, \tag{4}$$

where τ is the contrastive temperature. This shared representation path is used in both intra-subject and inter-subject experiments; source training for inter-subject evaluation additionally applies the normalization below.

3.4 Learning in Subject-Standardized Coordinates: TC-SAW

The semantic EEG latent $\mathbf{u}^e$ can contain subject-dependent offsets and per-dimension scale differences. Subject-adaptive whitening has previously been applied to unseen-subject retrieval at test time (Huang & Zhu, 2026); here, TC-SAW instead places subject-wise standardization inside source-subject contrastive training so that the representation itself is optimized in standardized coordinates. In this work, TC-SAW uses diagonal coordinate-wise standardization: it estimates per-dimension means and variances and does not model off-diagonal covariance. For source subject s and its minibatch subset $\mathcal{B}_s$, we compute the diagonal minibatch mean $\widehat{\boldsymbol{\mu}}_s$ and variance $\widehat{\boldsymbol{\sigma}}_s^2$ and detach both moment estimators from the gradient graph:

$$\widetilde{\mathbf{u}}_i^e = \frac{\mathbf{u}_i^e - \widehat{\boldsymbol{\mu}}_{s_i}}{\sqrt{\widehat{\boldsymbol{\sigma}}_{s_i}^2 + \lambda}}, \qquad \mathbf{z}_i^e = \mathcal{N}(\widetilde{\mathbf{u}}_i^e), \tag{5}$$

with $\lambda = 10^{-3}$. The moment estimates are stop-gradient statistics, but $\widetilde{\mathbf{u}}_i^e$ remains differentiable; the retrieval loss therefore updates the EEG backbone through the standardized representation. The visual branch remains the same shared UMVA space for every source subject.

3.5 Cold-Start Statistics for Unseen Subjects: SPPC

At deployment, applying the same normalization requires target-subject moments, yet none are available for the first query and early estimates remain unstable. SPPC treats this as sequential moment estimation: a finite-strength prior from source subjects initializes the target statistics, and its contribution decays automatically as target observations accumulate.

Source prior and sequential merge. From deterministic source EEG, we estimate a mean $\boldsymbol{\mu}_s^{\mathrm{src}}$ and diagonal variance $\boldsymbol{\sigma}_{s,\mathrm{src}}^2$ for each source subject, then average subjects equally:

$$\boldsymbol{\mu}_0 = \frac{1}{S}\sum_{s=1}^{S} \boldsymbol{\mu}_s^{\mathrm{src}}, \qquad \boldsymbol{\sigma}_0^2 = \frac{1}{S}\sum_{s=1}^{S} \boldsymbol{\sigma}_{s,\mathrm{src}}^2. \tag{6}$$

The prior uses at most 2,048 deterministic source samples per subject and has strength $\kappa = 8$, with prior centered sum of squares $\mathbf{M}_{2,\mathrm{src}} = (\kappa - 1)\boldsymbol{\sigma}_0^2$. Before target query $t = n + 1$, let $\boldsymbol{\mu}_n = n^{-1}\sum_{j=1}^{n} \mathbf{u}_j^e$ and $\mathbf{M}_{2,n}^{\mathrm{tgt}} = \sum_{j=1}^{n} (\mathbf{u}_j^e - \boldsymbol{\mu}_n)^{\odot 2}$ be the mean and coordinate-wise centered sum of squares of the $n \geq 1$ previously observed target latents. Set $\mathbf{M}_{2,0}^{\mathrm{tgt}} = \mathbf{0}$. Defining $N_n = \kappa + n$ and $\boldsymbol{\delta}_n = \boldsymbol{\mu}_n - \boldsymbol{\mu}_0$ for $n \geq 1$, parallel moment merging (Chan et al., 1983) gives

$$\begin{aligned} \boldsymbol{\mu}_n^\star &= \frac{\kappa \boldsymbol{\mu}_0 + n \boldsymbol{\mu}_n}{N_n}, \\ (\boldsymbol{\sigma}_n^\star)^2 &= \frac{\mathbf{M}_{2,\mathrm{src}} + \mathbf{M}_{2,n}^{\mathrm{tgt}} + \frac{\kappa n}{N_n} \boldsymbol{\delta}_n^{\odot 2}}{N_n - 1}. \end{aligned} \tag{7}$$

where $\boldsymbol{\delta}_n^{\odot 2}$ denotes element-wise squaring of $\boldsymbol{\delta}_n$. At $n = 0$, the merged moments are initialized to $\boldsymbol{\mu}_0^\star = \boldsymbol{\mu}_0$ and $(\boldsymbol{\sigma}_0^\star)^2 = \boldsymbol{\sigma}_0^2$. Before predicting query t, only the source prior and target state through $t - 1$ are used:

$$\mathbf{z}_t^e = \mathcal{N}\left(\frac{\mathbf{u}_t^e - \boldsymbol{\mu}_{t-1}^\star}{\sqrt{(\boldsymbol{\sigma}_{t-1}^\star)^2 + \lambda}} \right). \tag{8}$$

After query t is ranked, $\mathbf{u}_t^e$ is added to the running target state. The source coefficient in the merged mean is $\kappa/(\kappa + n)$, producing a gradual handover from source-informed initialization to target-specific normalization.

Information regimes at evaluation. *History-only* estimates target moments from preceding target queries; SPPC uses the same target history together with fixed source-training statistics.

4 EXPERIMENTS

4.1 DATASET, PROTOCOLS, AND METRICS

THINGS-EEG2 contains EEG from ten subjects viewing natural object images (Gifford et al., 2022). We follow the common zero-shot split used by NICE and later work (Song et al., 2024): 1,654 training concepts contain ten images per concept, and 200 disjoint test concepts contain one image each. EEG is downsampled to 250 Hz; all experiments use 17 posterior P/PO/O channels and 250 post-stimulus samples. Each test query ranks the same 200 image candidates, and we report Top-1 and Top-5 accuracy.

For intra-subject evaluation, a separate model is trained for each subject. Inter-subject evaluation trains on nine source subjects and evaluates on the tenth, rotating the held-out subject across ten folds. Models are trained for at most 60 epochs with early-stopping patience 20 on the corresponding training split. The online calibration state is reset at the start of every inter-subject fold. In the main intra-subject and inter-subject systems, repeated trials remain separate through STAR-EEG and are averaged in the pre-ℓ_2 latent space defined in Section 3.3. For inter-subject evaluation, SPPC operates on the post-aggregation representation associated with each test concept, and the predict-before-update sequence is defined over the 200 concept-level test queries.

4.2 IMPLEMENTATION DETAILS

All experiments are implemented in PyTorch and run on a single NVIDIA GeForce RTX 4090 GPU with 24 GB memory. The visual branch uses frozen InternViT depths $\{20, 24, 28, 32, 36\}$, one linear image projector, and a shared residual encoder to produce 256-D retrieval embeddings. STAR-EEG uses $K = 192$ temporal coefficients per channel, EEGProject-style residual refinement, and 0.1 feature dropout. The main intra-subject and inter-subject systems use one-way EEG-to-image InfoNCE, batch size 512 with drop-last, and AdamW with learning rate 10^{-4}; the inter-subject configuration uses weight decay 10^{-2}. TC-SAW uses diagonal moments with $\lambda = 10^{-3}$ and detached minibatch statistics, while SPPC uses at most 2,048 source-training samples per source subject. SPPC fixes $\kappa = 8$ for target evaluation; source-only validation selects only $\kappa \in \{4, 8\}$ across outer folds (8 in 6/10). All main results aggregate five independent runs.

4.3 MAIN RESULTS

Intra-subject retrieval. Table 1 compares the UMVA+STAR-EEG representation with published intra-subject results. Across five independent runs, our model reaches $93.67 \pm 0.28\%$ Top-1 and $99.36 \pm 0.19\%$ Top-5. The larger gain at Top-1 indicates that channel-shared temporal structure and post-encoding repeat aggregation preserve highly discriminative stimulus information under the fixed visual reference.

Table 1: Intra-subject 200-way retrieval accuracy (%) on THINGS-EEG2. Subject columns report the corresponding subject result; our row reports per-subject means over five independent runs.

Method	Metric	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Avg.
NICE (Song et al., 2024)	Top-1	13.2	13.5	14.5	20.6	10.1	16.5	17.0	22.9	15.4	17.4	16.1
	Top-5	39.5	40.3	42.7	52.7	31.5	44.0	42.1	56.1	41.6	45.8	43.6
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
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
Hierarchical Visual Embeddings (Zheng et al., 2026)	Top-1	64.3	76.3	74.0	67.0	68.0	81.5	76.8	84.8	76.8	87.3	75.7
	Top-5	88.8	95.3	95.0	91.8	91.5	96.3	96.8	98.5	95.8	99.3	94.6
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 Alignment (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
HCF (Tang et al., 2026)	Top-1	81.9	86.0	84.3	82.2	74.5	90.4	82.8	90.9	79.7	93.1	84.6
	Top-5	98.3	98.9	97.7	98.8	94.1	98.9	97.2	99.4	98.4	99.8	98.2
SAMGA (Jiang et al., 2026)	Top-1	85.2	94.4	91.8	89.7	86.3	97.2	89.2	94.8	88.7	96.4	91.3
	Top-5	95.5	99.8	98.2	98.7	98.4	99.9	98.7	99.8	99.2	99.8	98.8
Ours	Top-1	93.9	95.4	94.8	92.7	90.2	97.1	88.9	95.4	91.4	96.9	$\mathbf{93.67 \pm 0.28}$
	Top-5	98.6	99.6	99.8	99.4	98.4	99.8	98.9	99.8	99.7	99.6	$\mathbf{99.36 \pm 0.19}$

Inter-subject retrieval. Table 2 reports held-out-subject performance and published inter-subject results. The complete RNSC framework reaches 48.10% Top-1 and 76.34% Top-5; subject-wise Top-1 ranges from 34.7% on S8 to 63.8% on S10, showing substantial residual heterogeneity.

Table 2: Inter-subject 200-way retrieval accuracy (%) for every held-out subject on THINGS-EEG2. Our row reports the per-subject mean over five runs for the complete UMVA+STAR-EEG+TC-SAW+SPPC system.

Method	Metric	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Avg.
NICE (Song et al., 2024)	Top-1	7.6	5.9	6.0	6.3	4.4	5.6	5.6	6.3	5.7	8.4	6.2
	Top-5	22.8	20.5	22.3	20.7	18.3	22.2	19.7	22.0	17.6	28.3	21.4
ATM (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
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
Shallow Alignment (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	45.2	49.8	46.0	46.1	54.8	62.0	49.4
HCF (Tang et al., 2026)	Top-1	28.5	22.5	20.0	20.5	18.5	29.5	20.0	20.0	21.0	34.0	23.4
	Top-5	61.0	56.5	45.5	54.0	46.5	60.5	53.5	53.5	48.5	69.0	54.9
SAMGA (Jiang et al., 2026)	Top-1	36.3	42.2	24.7	34.0	30.0	35.2	35.1	28.7	28.1	49.6	34.4
	Top-5	69.8	71.7	50.7	66.7	64.3	67.7	61.0	59.5	57.0	79.4	64.8
SATTC (Huang & Zhu, 2026)	Top-1	–	–	–	–	–	–	–	–	–	–	14.8
	Top-5	–	–	–	–	–	–	–	–	–	–	38.4
Ours (full RNSC)	Top-1	58.3	53.7	39.3	39.1	44.5	52.8	44.9	34.7	49.9	63.8	$\mathbf{48.10 \pm 0.67}$
	Top-5	85.5	81.4	71.5	66.6	71.8	80.4	73.7	64.3	80.3	87.9	$\mathbf{76.34 \pm 0.20}$

Published results in Table 2 follow their original method protocols. The matched ablations below isolate the contributions of shared representation learning, TC-SAW, and SPPC within RNSC.

4.4 Ablations of the Shared Representation

Temporal factorization and representation-space repeat aggregation. Table 3 forms a controlled 2×2 comparison between the channel-shared temporal factorization and the point at which repeated trials are averaged. Temporal factorization improves Top-1 under both pooling rules ($+1.13$ points with raw mean and $+0.86$ with encoded mean), while encoded-mean aggregation improves both EEGProject and the factorized encoder ($+0.76$ and $+0.49$ points, respectively). Their combination is strongest at $93.67 \pm 0.28\%$, indicating that both structural choices contribute beyond encoder capacity alone. Encoded mean averages repetitions in the pre-ℓ_2 latent space; Δ_{agg} and Δ_{STAR} denote the gains from representation-space repeat aggregation and STAR-EEG relative to EEGProject, respectively.

Does the shared visual reference use multiple depths? Table 4 tests whether the learned UMVA representation collapses onto a single InternViT depth. Across the 50 intra-subject checkpoints, projected activation share stays between 19.0% and 21.0% for every selected depth. Masking any one depth at the trained checkpoint lowers Top-1 by 14.21–18.94 percentage points, while exact Shapley contributions range from 16.01 to 20.08 points. The positive contributions across all five depths show that the learned visual reference uses distributed evidence rather than relying on a single selected depth; the largest Shapley contribution comes from depth 28.

Table 3: Component ablation of STAR-EEG for intra-subject retrieval.

EEG encoder	Raw mean	Encoded mean	Δ_{agg}
EEGProject	92.05 ± 0.36	92.81 ± 0.27	$+0.76$
STAR-EEG	93.18 ± 0.31	$\mathbf{93.67 \pm 0.28}$	$+0.49$
Δ_{STAR}	$+1.13$	$+0.86$	–

Table 4: UMVA multi-depth contribution analysis over 50 intra-subject checkpoints (mean $\pm$ SD).

Depth	Activation share (%)	Shapley Top-1 (pp)
20	19.04 ± 0.41	16.01 ± 1.46
24	20.33 ± 0.39	19.45 ± 1.61
28	21.04 ± 0.34	20.08 ± 1.08
32	19.52 ± 0.27	18.92 ± 1.57
36	20.07 ± 0.22	18.64 ± 2.13

4.5 Complementarity of Training-Time Normalization and Deployment Calibration

Table 5 provides a direct 2×2 design on the STAR-EEG representation: TC-SAW is switched on or off during source training, while deployment uses either History-only calibration or SPPC. Both mechanisms provide independent Top-1 gains, and their combination reaches $48.10 \pm 0.67\%$. The key distinction is functional: TC-SAW changes the coordinates in which the representation is learned, whereas SPPC changes how those coordinates are estimated for an unseen subject. Across five independent runs, SPPC improves Top-1 in all $5/5$ paired comparisons, with a mean gain of $+1.21 \pm 0.21$ percentage points.

Table 5: Ablation of training-time subject normalization and sequential calibration on the final STAR-EEG representation. History-only and SPPC follow the same predict-before-update protocol. Results are mean $\pm$ SD across five run-level ten-fold means.

TC-SAW	Calibration	Top-1	Top-5
×	History-only	46.00 ± 0.38	75.92 ± 0.47
×	SPPC	46.90 ± 0.61	76.02 ± 0.49
✓	History-only	46.89 ± 0.79	75.68 ± 0.34
✓	**SPPC**	$\mathbf{48.10 \pm 0.67}$	$\mathbf{76.34 \pm 0.20}$

4.6 Cross-Subject Latent Geometry and Cold-Start Dynamics

Does training-time normalization reduce subject-dependent latent mismatch? Across 50 paired run–fold comparisons, TC-SAW reduces the source-to-target pre-ℓ_2 mean-vector gap from 4.479 to 2.742 (49/50 improve) and increases mean-vector cosine from 0.332 to 0.991 (50/50 improve). Relative variance mismatch also decreases in most pairs. Standardized mean-shift RMS instead rises from 0.459 to 0.590 because TC-SAW strongly contracts the pre-normalization variance, which appears in the denominator of that statistic.

Figure 2 visualizes this effect for the sub-01 held-out fold. The pre-normalization geometry becomes substantially more aligned across source subjects under TC-SAW, and the final TC-SAW+SPPC mapping produces stronger source–target overlap after deployment normalization, consistent with the paired statistics and the retrieval ablation in Table 5.

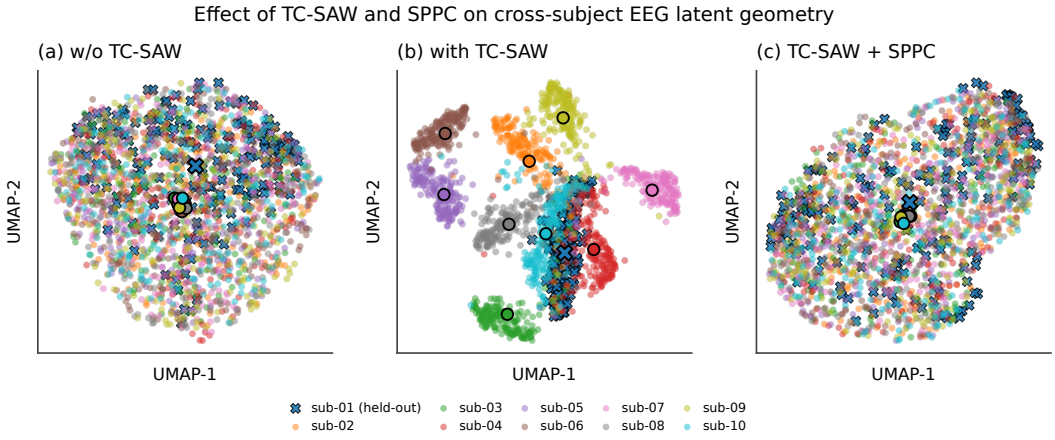

Figure 2: Cross-subject EEG latent geometry for the sub-01 held-out fold: (a) without TC-SAW, (b) with TC-SAW before deployment normalization, and (c) after TC-SAW+SPPC normalization. Colors denote subjects; outlined markers indicate subject centroids.

When is the source prior most useful? Figure 3(a) shows that the benefit of SPPC is concentrated in the low-history regime. At the first query, SPPC reaches 16.0% Top-1 while History-only has no usable target variance and remains at 0%. Across the first four predictions, the mean source coefficient is 85.4% and cumulative Top-1 is 35.0% versus 7.0% for History-only ($+28.0$ points). By 32 predictions, the coefficient falls to 41.5% and the gain narrows to $+5.8$ points; over all 200

predictions, the mean coefficient is 13.3% and the final gain is $+1.2$ points. The joint decay in source weight and retrieval gain reflects the intended handover from source-informed initialization to target-driven normalization.

Figure 3(b) characterizes the source–target trade-off through κ. Performance is stable for moderate priors, increasing from 47.86% at $\kappa = 2$ to 48.10% at the primary $\kappa = 8$ setting, but falls to 46.35% and 43.48% for $\kappa = 16$ and 32. The strongest prior is therefore worse than the 46.89% History-only baseline, indicating that an overly persistent source prior delays the transition to target-subject statistics.

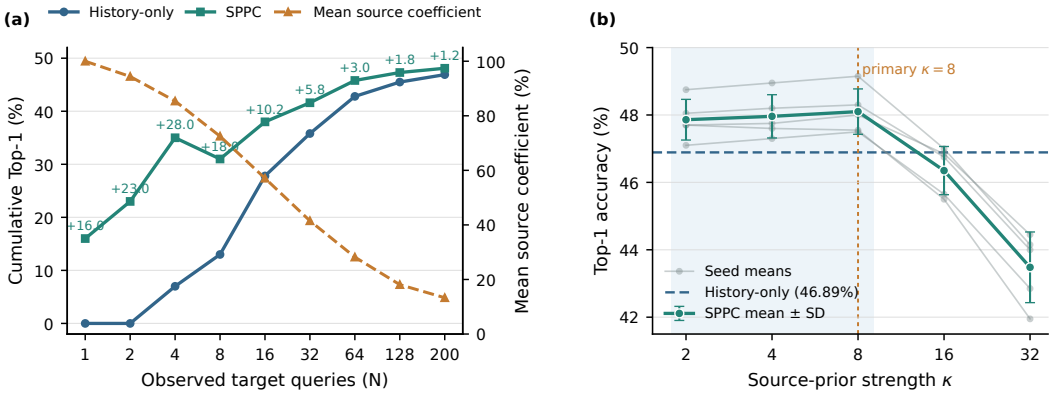

Figure 3: Cold-start behavior and source-prior strength for SPPC. (a) Cumulative Top-1 for History-only and SPPC across target prefixes; the dashed curve is the mean source coefficient and annotations show SPPC minus History-only. (b) Top-1 across κ values over ten held-out subjects and five runs; thin curves are run-level ten-subject means, markers/error bars show mean $\pm$ SD, and the dashed horizontal line is the History-only baseline (46.89%).

4.7 Discussion and Limitations

The results support a clear division of roles. STAR-EEG and the fixed multi-depth visual reference establish stimulus-discriminative EEG–vision geometry; TC-SAW changes how that geometry is learned across source subjects; and SPPC estimates the normalization statistics required to deploy the same coordinate system to an unseen subject. This decomposition evaluates representation quality, training-time subject normalization, and deployment-time moment estimation as distinct factors.

Held-out-subject Top-1 ranges from 34.7% to 63.8%, indicating that diagonal location and scale account for only part of cross-subject heterogeneity; richer cross-dimensional structure is a natural next step.

5 Conclusion

Cross-subject EEG-to-image retrieval requires not only a stimulus-discriminative shared representation, but also consistent treatment of EEG latent statistics between source training and unseen-subject deployment. We keep the visual candidate space fixed and concentrate subject-dependent variation on the EEG branch: STAR-EEG builds the shared representation, TC-SAW learns it in subject-standardized coordinates, and SPPC resolves the missing-target-statistics cold start before gradually yielding to target-specific moments. On THINGS-EEG2, the STAR-EEG-based model establishes state-of-the-art intra-subject retrieval, reaching 93.67% Top-1 and 99.36% Top-5, while the complete RNSC framework reaches $48.10 \pm 0.67\%$ Top-1 under inter-subject evaluation. Controlled ablations separate representation, training-time normalization, and deployment-time estimation, while latent-geometry and cold-start analyses show where each mechanism acts. The remaining subject-to-subject variation suggests extending the current diagonal moment model toward richer cross-dimensional distribution structure; future analyses may further characterize this residual mismatch with higher-order distributional metrics.

AI USE STATEMENT

Generative AI tools were used for literature retrieval and discovery and to aid and polish the writing of the manuscript. All AI-assisted content was reviewed and verified by the authors, who take full responsibility for the final manuscript.

REPRODUCIBILITY STATEMENT

Sections 3 and 4 specify the model, training objective, data protocol, preprocessing, optimization, and sequential evaluation procedure. The appendix provides detailed configurations, multi-run results, ablations, and calibration analyses. The supplementary code package contains the implementation, experiment configurations, evaluation procedures, and analysis scripts used to produce the reported results.

ETHICS STATEMENT

This work performs secondary analysis of the existing THINGS-EEG2 dataset and collects no new human-subject data (Gifford et al., 2022). The study is intended for methodological research on non-invasive neural representation learning. Any deployment involving inference from individuals' neural data would require appropriate consent, privacy safeguards, and separate validation.

REFERENCES

Tony F. Chan, Gene H. Golub, and Randall J. LeVeque. Statistical computing: Algorithms for computing the sample variance: Analysis and recommendations. *The American Statistician*, 37 (3):242–247, 1983. doi: 10.2307/2683386.

Hongzhou Chen, Lianghua He, Yihang Liu, and Longzhen Yang. Visual neural decoding via improved visual–EEG semantic consistency, 2024a. arXiv preprint arXiv:2408.06788.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607, 2020.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. InternVL: Scaling up vision foundation models and aligning for generic visual–linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24185–24198, 2024b.

Radoslaw Martin Cichy, Aditya Khosla, Dimitrios Pantazis, Antonio Torralba, and Aude Oliva. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, 6:27755, 2016. doi: 10.1038/srep27755.

Zhenyao Cui, Siyuan Kan, Siyang Li, Ziwei Wang, and Dongrui Wu. SCORE: Subject coordinate recovery for label-free cross-subject EEG-to-image retrieval, 2026. arXiv preprint arXiv:2608.19134.

A. P. Dawid. Present position and potential developments: Some personal views: Statistical theory: The prequential approach. *Journal of the Royal Statistical Society. Series A (General)*, 147(2): 278–290, 1984. doi: 10.2307/2981683.

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. doi: 10.1109/TPAMI.2023.3263181.

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, 2026. arXiv preprint arXiv:2601.21948.

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

Xiang Gao, Hui Tian, Yanming Zhu, Xuefei Yin, and Alan Wee-Chung Liew. Neuroscience-inspired staged representation learning with disentangled coarse- and fine-grained semantics for EEG visual decoding, 2026. arXiv preprint arXiv:2605.16923.

Alessandro T. Gifford, Kshitij Dwivedi, Gemma Roig, and Radoslaw M. Cichy. A large and rich EEG dataset for modeling human visual object recognition. *NeuroImage*, 264:119754, 2022. doi: 10.1016/j.neuroimage.2022.119754.

Shengyu Gong, Weiming Zeng, Yueyang Li, Zijian Kang, Hongjie Yan, Wai Ting Siok, and Nizhuan Wang. SUP-MCRL: Subject-aware unified pseudo-feature coded multimodal contrastive representation learning for EEG visual decoding, 2026. arXiv preprint arXiv:2606.16615.

Martin N. Hebart, Oliver Contier, Lina Teichmann, Adam H. Rockter, Charles Y. Zheng, Alexis Kidder, Anna Corriveau, Maryam Vaziri-Pashkam, and Chris I. Baker. THINGS-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *eLife*, 12:e82580, 2023. doi: 10.7554/eLife.82580.

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*, pp. 16887–16896, 2026.

Vinay Jayaram, Morteza Alamgir, Yasemin Altun, Bernhard Schölkopf, and Moritz Grosse-Wentrup. Transfer learning in brain–computer interfaces. *IEEE Computational Intelligence Magazine*, 11(1):20–31, 2016. doi: 10.1109/MCI.2015.2501545.

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. arXiv preprint arXiv:2604.17782.

Vernon J. Lawhern, Amelia J. Solon, Nicholas R. Waytowich, Stephen M. Gordon, Chou P. Hung, and Brent J. Lance. EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of Neural Engineering*, 15(5):056013, 2018. doi: 10.1088/1741-2552/aace8c.

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*, volume 37, 2024. doi: 10.52202/079017-3266.

Yanghao Li, Naiyan Wang, Jianping Shi, Xiaodi Hou, and Jiaying Liu. Adaptive batch normalization for practical domain adaptation. *Pattern Recognition*, 80:109–117, 2018. doi: 10.1016/j.patcog.2018.03.005.

Hyesu Lim, Byeonggeun Kim, Jaegul Choo, and Sungha Choi. TTN: A domain-shift aware batch normalization in test-time adaptation. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=EQfeudmWLQ.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16888–16905, 2022.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=g2YraF75Tj.

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

Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. EEG Conformer: Convolutional transformer for EEG decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:710–719, 2023. doi: 10.1109/TNSRE.2022.3230250.

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://openreview.net/forum?id=dhLIno8FmH.

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

Devavrat Tomar, Guillaume Vray, Jean-Philippe Thiran, and Behzad Bozorgtabar. Un-mixing test-time normalization statistics: Combatting label temporal correlation. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=xyxU99Nutg.

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

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=uXl3bZLkr3c.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7201–7211, 2022.

Haitao Wu, Qing Li, Changqing Zhang, Zhen He, and Xiaomin Ying. Bridging the vision-brain gap with an uncertainty-aware blur prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2246–2257, 2025. doi: 10.1109/CVPR52734.2025.00215.

Salini Yadav, Taveena Lotey, Pravendra Singh, and Partha Pratim Roy. What does the brain see? multiview neural representations to demystify the brain–visual alignment, 2026. arXiv preprint arXiv:2606.25718.

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

Paolo Zanini, Marco Congedo, Christian Jutten, Salem Said, and Yannick Berthoumieu. Transfer learning: A riemannian geometry framework with applications to brain–computer interfaces. *IEEE Transactions on Biomedical Engineering*, 65(5):1107–1116, 2018. doi: 10.1109/TBME.2017.2742541.

Kaifan Zhang, Lihuo He, Xin Jiang, Wen Lu, Di Wang, and Xinbo Gao. CognitionCapturer: Decoding visual stimuli from human EEG signal with multimodal information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39(13), pp. 14486–14493, 2025. doi: 10.1609/aaai.v39i13.33587.

Wenjiang Zhang, Sifeng Wang, Yuwei Su, Xinyu Li, Chen Zhang, and Suyu Zhong. NeuroBridge: Bio-inspired self-supervised EEG-to-image decoding via cognitive priors and bidirectional semantic alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 18028–18036, 2026. doi: 10.1609/aaai.v40i21.38863.

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.

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*, pp. 40438–40447, 2026.

Yanan Zhu, Ziwei Xiang, Jiamin Wu, Jinyang Guo, Hongyuan Zhang, Chunfeng Song, Hongjian Fang, Yufei Guo, and Xianglong Liu. Region-aware hierarchical sub-feature alignment for robust EEG-based visual decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Findings*, pp. 6667–6676, 2026.

A Additional Experimental Details

A.1 Configuration Summary

Table 6 brings together the settings for the three experiment families used in the paper. The first two columns describe the headline intra-subject and complete inter-subject systems; the third records the EEGProject inter-subject ablation recipe. The ablation family differs in encoder, repeat aggregation, batch size, drop-last handling, learning rate, and its TC-SAW on/off comparison, so its absolute scores are not a controlled comparison with the complete STAR-EEG system. The matched calibration comparisons instead hold each family's trained checkpoints fixed and change the evaluation schedule. The table also makes clear that the intra-subject family has no cross-subject calibration schedule, while the inter-subject families report History-only, SPPC, and transductive evaluation. The shared 256-D embedding and EEG-to-image loss direction aid comparison, but do not remove the training-recipe differences listed in the table.

Table 6: Configuration of the three experimental result families. "Diagonal" denotes coordinate-wise mean and variance normalization in TC-SAW; it does not model off-diagonal covariance. "Source prior" indicates that SPPC's source-derived prior is available in the evaluation family.

Choice	Main intra	Complete inter-subject	EEGProject inter-subject ablation
EEG encoder	STAR-EEG ($K = 192$)	STAR-EEG ($K = 192$)	EEGProject
Repeat aggregation	pre-ℓ_2 representation mean	pre-ℓ_2 representation mean	raw EEG mean
Embedding dim.	256	256	256
Loss direction	EEG-to-image	EEG-to-image	EEG-to-image
Batch / drop-last	512 / yes	512 / yes	256 / no
Base LR	10^{-4}	10^{-4}	5×10^{-5}
Train TC-SAW	no	diagonal	off / diagonal
Test schedules	None	History-only / SPPC / Transductive	History-only / SPPC / Transductive
Source prior	no	yes	yes
Training	max 60 epochs; patience 20	max 60 epochs; patience 20	max 60 epochs; patience 20
Independent runs	five	five	five

Evaluation-information schedules. Figure 4 summarizes the four information schedules used in the calibration analyses. Source-only uses fixed statistics from source training and does not use target EEG history. History-only estimates target moments from the previously observed unlabeled target queries $u_1, \ldots, u_{t-1}$ and predicts query t before incorporating u_t. SPPC follows the same causal predict-before-update order, but combines a finite source prior with the accumulated target moments so that the source contribution decays as target history grows. Transductive evaluation instead uses the complete target set to estimate target moments before scoring all queries; it is therefore non-causal and is reported only as an information-access reference rather than a deployment protocol.

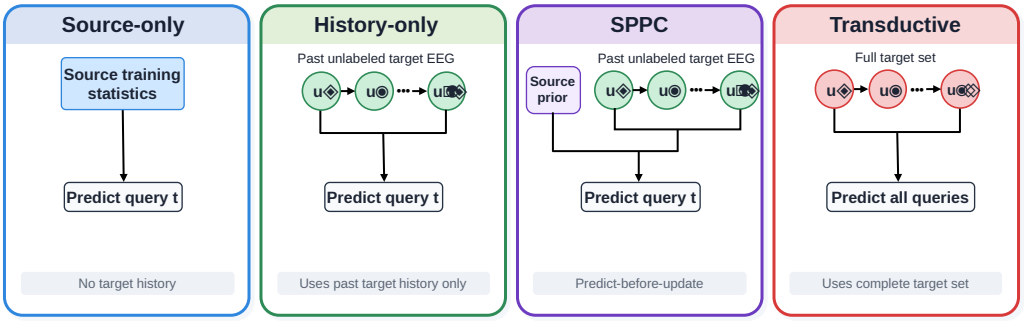

Figure 4: Information available under the four evaluation schedules. Source-only uses source-training statistics only; History-only and SPPC are causal and use no future target EEG when predicting query t; Transductive uses the complete target set before prediction and serves only as a non-causal reference.

A.2 Main Intra-Subject UMVA + STAR-EEG Results

The main intra-subject configuration contains 50 trained models from five independent runs across ten subjects. Every run uses UMVA, STAR-EEG with $K = 192$, and representation-space repeat averaging. Table 7 provides the run-level Top-1 results, while Table 8 gives per-subject statistics across runs.

Table 7: Main UMVA+STAR-EEG intra-subject Top-1 (%) for five independent runs and ten subjects.

Run	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Avg.
Run 1	93.5	95.0	95.5	94.0	90.5	96.0	91.0	96.0	92.0	97.0	94.05
Run 2	94.5	95.5	94.0	91.0	89.0	97.0	88.5	95.5	90.5	97.5	93.30
Run 3	94.0	96.0	95.0	92.5	90.0	97.5	88.0	95.0	92.5	96.5	93.70
Run 4	93.0	95.0	94.0	93.5	90.0	97.0	88.5	95.0	92.0	97.5	93.55
Run 5	94.5	95.5	95.5	92.5	91.5	98.0	88.5	95.5	90.0	96.0	93.75
Mean	93.9	95.4	94.8	92.7	90.2	97.1	88.9	95.4	91.4	96.9	93.67

Table 8: Per-subject mean $\pm$ sample SD over five independent intra-subject runs (%).

Subject	Top-1	Top-5	Top-10
S1	93.90 ± 0.65	98.60 ± 0.96	99.90 ± 0.22
S2	95.40 ± 0.42	99.60 ± 0.42	99.90 ± 0.22
S3	94.80 ± 0.76	99.80 ± 0.27	100.00 ± 0.00
S4	92.70 ± 1.15	99.40 ± 0.42	100.00 ± 0.00
S5	90.20 ± 0.91	98.40 ± 0.65	99.90 ± 0.22
S6	97.10 ± 0.74	99.80 ± 0.27	100.00 ± 0.00
S7	88.90 ± 1.19	98.90 ± 0.22	99.70 ± 0.27
S8	95.40 ± 0.42	99.80 ± 0.27	99.90 ± 0.22
S9	91.40 ± 1.08	99.70 ± 0.45	100.00 ± 0.00
S10	96.90 ± 0.65	99.60 ± 0.22	99.70 ± 0.27
Run-mean aggregate	93.67 ± 0.28	99.36 ± 0.19	99.90 ± 0.05

A.3 EEGProject + Raw-Mean UMVA Control

The EEGProject+raw-mean UMVA control contains 50 trained models from five independent runs across ten subjects. Each value is evaluated from the selected checkpoint. Table 9 reports run-level Top-1, and Table 10 reports per-subject statistics.

Table 9: EEGProject + raw-mean intra-subject control Top-1 (%) for five independent runs and ten subjects.

Run	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Avg.
Run 1	88.5	93.0	92.0	90.5	86.0	97.0	89.0	95.5	88.5	97.0	91.70
Run 2	90.0	93.5	94.5	91.0	85.5	96.5	87.0	96.0	88.0	97.5	91.95
Run 3	89.5	94.5	92.5	91.5	88.0	97.5	86.5	95.0	87.5	96.5	91.90
Run 4	88.0	94.0	94.0	94.0	87.0	97.5	89.5	96.0	89.0	97.5	92.65
Run 5	91.0	94.0	93.5	92.5	87.0	97.0	87.0	96.0	85.5	97.0	92.05
Mean	89.4	93.8	93.3	91.9	86.7	97.1	87.8	95.7	87.7	97.1	92.05

Table 10: Per-subject mean $\pm$ sample SD across five independent EEGProject + raw-mean intra-subject control runs (%).

Subject	Top-1	Top-5	Top-10
S1	89.40 ± 1.19	97.00 ± 0.61	98.20 ± 0.45
S2	93.80 ± 0.57	99.80 ± 0.27	100.00 ± 0.00
S3	93.30 ± 1.04	99.70 ± 0.45	100.00 ± 0.00
S4	91.90 ± 1.39	99.30 ± 0.76	99.90 ± 0.22
S5	86.70 ± 0.97	98.00 ± 0.71	99.30 ± 0.45
S6	97.10 ± 0.42	99.80 ± 0.27	100.00 ± 0.00
S7	87.80 ± 1.35	98.40 ± 0.42	99.50 ± 0.50
S8	95.70 ± 0.45	99.50 ± 0.35	99.80 ± 0.27
S9	87.70 ± 1.35	97.90 ± 0.89	99.90 ± 0.22
S10	97.10 ± 0.42	99.90 ± 0.22	99.90 ± 0.22
Run-mean aggregate	92.05 ± 0.36	98.93 ± 0.08	99.65 ± 0.06

A.4 UMVA Multi-Depth Contribution Diagnostic

We audit the 50 main intra-subject checkpoints by masking visual-depth subsets at the trained checkpoint and measuring projected activation. Table 11 reports the contribution statistics; the positive Shapley contributions across all five depths support distributed use of the trained UMVA representation.

Table 11: UMVA depth-contribution diagnostics over the 50 main intra runs (mean $\pm$ SD). Weight share is the normalized norm share of projection weights associated with each depth block; activation share is the normalized projected contribution-norm share. LOO ΔTop-1 is full-checkpoint accuracy minus single-depth-masked accuracy. Shapley values use exact coalition masking. All diagnostics use the trained checkpoint without retraining.

Depth	Weight share (%)	Activation share (%)	LOO ΔTop-1 (pp)	Shapley Top-1 (pp)
20	18.89 ± 0.22	19.04 ± 0.41	15.98 ± 4.16	16.01 ± 1.46
24	19.38 ± 0.13	20.33 ± 0.39	16.55 ± 4.51	19.45 ± 1.61
28	19.69 ± 0.05	21.04 ± 0.34	15.88 ± 4.32	20.08 ± 1.08
32	20.47 ± 0.09	19.52 ± 0.27	14.21 ± 4.26	18.92 ± 1.57
36	21.58 ± 0.30	20.07 ± 0.22	18.94 ± 4.49	18.64 ± 2.13

A.5 Complete STAR-EEG + TC-SAW + SPPC Inter-Subject Replication

The complete inter-subject configuration contains 50 trained models from five independent runs across ten held-out-subject folds. Each run uses STAR-EEG with $K = 192$, representation-space repeat aggregation, UMVA, TC-SAW during source training, and SPPC with $\kappa = 8$ at deployment. Table 12 reports run-level Top-1; the main-text table gives the aggregate retrieval metrics.

The selected checkpoints are also re-evaluated with the source prior removed while retaining the preceding target-query history. Table 13 shows a Top-1 gain for all five runs and reports the paired run-level differences.

Table 12: Complete STAR-EEG+TC-SAW+SPPC inter-subject Top-1 (%).

Run	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Avg.
Run 1	59.5	53.0	39.5	40.5	45.0	55.0	47.0	35.5	52.0	64.5	49.15
Run 2	55.0	55.0	39.5	39.0	43.0	52.5	45.0	33.0	49.0	64.5	47.55
Run 3	60.5	52.0	39.0	38.0	45.0	51.0	44.5	34.5	47.5	63.0	47.50
Run 4	58.0	54.0	38.5	39.5	43.0	54.0	42.5	36.5	49.0	65.0	48.00
Run 5	58.5	54.5	40.0	38.5	46.5	51.5	45.5	34.0	52.0	62.0	48.30
Mean	58.3	53.7	39.3	39.1	44.5	52.8	44.9	34.7	49.9	63.8	48.10

Table 13: Run-level comparison of History-only and SPPC deployment for the complete inter-subject system (%).

	History-only		SPPC		Δ SPPC−History	
Run	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
Run 1	48.00	75.90	49.15	76.40	+1.15	+0.50
Run 2	46.45	75.20	47.55	76.05	+1.10	+0.85
Run 3	45.95	76.10	47.50	76.35	+1.55	+0.25
Run 4	46.75	75.55	48.00	76.60	+1.25	+1.05
Run 5	47.30	75.65	48.30	76.30	+1.00	+0.65
Mean ± SD	46.89 ± 0.79	75.68 ± 0.34	48.10 ± 0.67	76.34 ± 0.20	$+1.21 \pm 0.21$	$+0.66 \pm 0.31$

A.6 Unified Inter-Subject Ablation Metrics

Table 14 reports the full Top-1/Top-5/Top-10 summary corresponding to the main-paper inter-subject ablation. STAR-EEG × denotes the EEGProject/raw-mean representation family; STAR-EEG ✓ denotes the STAR-EEG with $K = 192$ family. None reports the checkpoints without target-subject normalization. TC-SAW on/off comparisons use the corresponding run and fold pairs within each representation family.

Table 14: Unified five-run inter-subject ablation with detailed retrieval metrics (mean ± SD).

STAR-EEG	TC-SAW	Calibration	Top-1	Top-5	Top-10
×	×	None	25.69 ± 1.02	54.32 ± 1.44	67.96 ± 0.87
×	×	History-only	42.64 ± 0.50	72.33 ± 0.83	83.95 ± 0.53
×	×	SPPC	44.25 ± 0.48	72.96 ± 0.90	84.35 ± 0.45
×	×	Transductive	44.48 ± 0.59	74.36 ± 1.05	85.90 ± 0.43
×	✓	History-only	45.26 ± 0.55	74.39 ± 0.60	84.47 ± 0.72
×	✓	SPPC	46.71 ± 0.51	75.24 ± 0.87	85.75 ± 0.66
×	✓	Transductive	47.62 ± 0.63	77.15 ± 0.52	87.18 ± 0.52
✓	×	None	26.63 ± 0.61	55.11 ± 0.75	67.93 ± 1.10
✓	×	History-only	46.00 ± 0.38	75.92 ± 0.47	86.52 ± 0.13
✓	×	SPPC	46.90 ± 0.61	76.02 ± 0.49	86.47 ± 0.33
✓	×	Transductive	47.72 ± 0.35	77.96 ± 0.63	88.13 ± 0.30
✓	✓	History-only	46.89 ± 0.79	75.68 ± 0.34	85.44 ± 0.49
✓	✓	**SPPC**	$\mathbf{48.10 \pm 0.67}$	$\mathbf{76.34 \pm 0.20}$	86.15 ± 0.54
✓	✓	Transductive	49.42 ± 0.46	78.33 ± 0.47	87.96 ± 0.35

For the final STAR-EEG family, Table 15 gives run-level metrics for the TC-SAW on/off comparison. The five run-level Top-1 changes are positive under both History-only and SPPC deployment, with the strongest and most consistent effect at Top-1.

Table 15: Run-level inter-subject metrics for the final STAR-EEG family under TC-SAW on/off training and the two sequential deployment schedules. Each entry averages ten held-out-subject folds (%).

Run	TC-SAW	History-only			SPPC		
		Top-1	Top-5	Top-10	Top-1	Top-5	Top-10
Run 1	×	46.15	75.50	86.60	47.25	76.15	86.30
Run 2	×	46.25	75.45	86.40	46.20	75.20	86.25
Run 3	×	45.65	76.20	86.45	46.50	76.50	86.30
Run 4	×	46.40	75.90	86.70	47.75	76.25	87.05
Run 5	×	45.55	76.55	86.45	46.80	76.00	86.45
Run 1	✓	48.00	75.90	85.45	49.15	76.40	86.40
Run 2	✓	46.45	75.20	85.60	47.55	76.05	86.15
Run 3	✓	45.95	76.10	85.75	47.50	76.35	86.75
Run 4	✓	46.75	75.55	84.60	48.00	76.60	85.30
Run 5	✓	47.30	75.65	85.80	48.30	76.30	86.15
Mean ± SD	×	46.00 ± 0.38	75.92 ± 0.47	86.52 ± 0.13	46.90 ± 0.61	76.02 ± 0.49	86.47 ± 0.33
Mean ± SD	✓	46.89 ± 0.79	75.68 ± 0.34	85.44 ± 0.49	48.10 ± 0.67	76.34 ± 0.20	86.15 ± 0.54

A.7 EEGPROJECT INTER-SUBJECT CONTROL WITHOUT TC-SAW

The combined no-TC-SAW experiments contain 50 trained models from five independent runs across ten held-out subjects using EEGProject, 256-D embeddings, UMVA, and the inter-subject optimizer recipe. Each selected checkpoint is evaluated under History-only, SPPC, and transductive calibration. Table 16 reports schedule-level results, and Table 17 gives subject-wise results. The comparison shows that the source-prior gain also persists in the TC-SAW-off representation family.

Table 16: No-TC-SAW inter-subject Top-1 under the three information schedules. Each row averages ten held-out subjects for one run (%). The final row reports mean ± sample SD across run-level means.

Run	History-only	SPPC	Transductive
Run 1	43.25	44.70	45.45
Run 2	42.65	44.10	44.35
Run 3	42.15	44.35	44.10
Run 4	43.00	44.60	44.55
Run 5	42.15	43.50	43.95
Mean ± SD	42.64 ± 0.50	44.25 ± 0.48	44.48 ± 0.59

For Top-5, the corresponding five-run means are $72.33 \pm 0.83/72.96 \pm 0.90/74.36 \pm 1.05\%$ for History-only/SPPC/transductive. For Top-10, they are $83.95 \pm 0.53/84.35 \pm 0.45/85.90 \pm 0.43\%$.

Table 17: No-TC-SAW inter-subject Top-1 (%) under SPPC deployment for five independent runs and ten held-out subjects.

Run	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Avg.
Run 1	49.0	50.0	34.0	38.5	39.5	53.5	43.0	34.5	46.0	59.0	44.70
Run 2	49.0	49.0	37.5	36.5	39.0	54.5	43.0	32.5	43.0	57.0	44.10
Run 3	50.0	49.0	33.5	40.5	43.0	51.0	40.0	32.5	46.5	57.5	44.35
Run 4	50.0	50.0	37.0	38.0	41.0	52.5	41.0	33.0	46.0	57.5	44.60
Run 5	48.5	47.5	31.0	38.0	42.5	51.0	40.5	32.5	47.5	56.0	43.50
Mean	49.3	49.1	34.6	38.3	41.0	52.5	41.5	33.0	45.8	57.4	44.25

A.8 EEGPROJECT INTER-SUBJECT TC-SAW REPLICATION

The TC-SAW-on experiments contain 50 trained models over five independent runs and ten held-out subjects. Table 18 summarizes the three evaluation schedules, while Table 19 gives detailed TC-SAW-on subject results. All five run-level Top-1 changes are positive under each schedule.

For SPPC deployment, the Top-5 and Top-10 changes are $+2.28 \pm 1.45$ and $+1.40 \pm 0.87$ points. Under History-only they are $+2.06 \pm 1.21$ and $+0.52 \pm 1.17$, while under transductive evaluation

Table 18: Five-run inter-subject TC-SAW training ablation for Top-1 under each evaluation schedule (%). Δ is TC-SAW minus no-TC-SAW, computed per run before averaging.

Evaluation schedule	w/o TC-SAW	TC-SAW	Δ
History-only	42.64 ± 0.50	45.26 ± 0.55	$+2.62 \pm 0.98$
SPPC	44.25 ± 0.48	46.71 ± 0.51	$+2.46 \pm 0.89$
Transductive	44.48 ± 0.59	47.62 ± 0.63	$+3.14 \pm 0.92$

they are $+2.79 \pm 1.53$ and $+1.28 \pm 0.89$. Run-level values for all schedules and metrics are provided in the accompanying evidence files.

Table 19: TC-SAW-on inter-subject Top-1 (%) under SPPC deployment for five independent runs and ten held-out subjects. The final row gives per-subject means across runs.

Run	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	Avg.
Run 1	52.5	52.0	34.5	37.0	42.5	51.5	45.5	34.5	45.5	67.0	46.25
Run 2	54.0	52.5	37.0	37.0	42.5	51.0	43.5	32.5	48.5	63.5	46.20
Run 3	56.0	54.0	37.0	40.0	44.0	51.0	44.0	33.5	49.5	61.0	47.00
Run 4	55.5	54.0	34.0	39.0	44.0	49.0	45.5	33.5	47.5	65.0	46.70
Run 5	57.0	53.0	37.5	38.5	43.5	50.0	45.5	37.0	49.0	63.0	47.40
Mean	55.0	53.1	36.0	38.3	43.3	50.5	44.8	34.2	48.0	63.9	46.71

The corresponding per-subject five-run mean Top-1 change from TC-SAW off to on is $+5.7, +4.0, +1.4, 0.0, +2.3, -2.0, +3.3, +1.2, +2.2, +6.5$ points for S1–S10. Eight subjects improve, S4 is effectively unchanged, and S6 declines. This subject-wise pattern is consistent with a substantial role for location/scale variation in cross-subject transfer.

A.9 Multi-run Paired TC-SAW Latent Geometry Analysis

For the final STAR-EEG inter-subject family, the TC-SAW-on and TC-SAW-off analyses pair the same runs and held-out-subject folds, with the training-time TC-SAW switch as the only changed factor. We inspect the pre-ℓ_2 EEG latent before deployment-time calibration. At each run's selected checkpoint, we compare source-average and held-out-target latent statistics. Table 20 summarizes all 50 run–fold pairs.

Table 20: Five-run paired pre-ℓ_2 latent diagnostics on the final STAR-EEG inter-subject family.

Statistic	w/o TC-SAW	TC-SAW	Change	Improved
Mean-vector ℓ_2 gap $\downarrow$	4.479	**2.742**	-38.8%	49/50
Mean-vector cosine $\uparrow$	0.332	**0.991**	$+0.659$	50/50
Standardized mean-shift RMS $\downarrow$	**0.459**	0.590	$+28.7\%$	1/50
Relative variance ℓ_2 gap $\downarrow$	0.644	**0.571**	-11.3%	46/50
Absolute log variance ratio $\downarrow$	1.119	**0.975**	-12.9%	45/50

TC-SAW substantially reduces absolute mean mismatch, while the mean marginal variance contracts from 0.378 to 0.092 for source subjects and from 0.135 to 0.036 for held-out targets. Because standardized mean-shift RMS divides the remaining mean difference by the source standard deviation, this variance contraction increases the normalized quantity even as the absolute mean gap falls. Together, these diagnostics show that TC-SAW reshapes first-order cross-subject latent geometry and reduces several variance-mismatch measures.

A.10 Final STAR-EEG Cold-Start Analysis

The prefix analysis uses the final STAR-EEG+TC-SAW inter-subject checkpoint family over five independent runs and ten held-out-subject folds. Table 21 reports cumulative History-only and SPPC retrieval values together with the mean source coefficient over each target prefix. The source coefficient is computed as $\kappa/(\kappa+n)$, and gains are measured relative to History-only normalization. The final row recovers the main 200-query results, while the early rows isolate the cold-start regime.

Table 21: Final STAR-EEG+TC-SAW cold-start prefix analysis.

N	Mean src. coef. (%)	History-only Top-1	History-only Top-5	SPPC Top-1	SPPC Top-5	Δ Top-1/Top-5 (pp)
1	100.0	0.0	0.0	16.0	32.0	+16.0 / +32.0
2	94.4	0.0	0.0	23.0	41.0	+23.0 / +41.0
4	85.4	7.0	25.0	35.0	58.0	+28.0 / +33.0
8	72.5	13.0	34.8	31.0	60.0	+18.0 / +25.2
16	57.1	27.8	54.5	38.0	66.9	+10.2 / +12.4
32	41.5	35.8	64.3	41.6	70.2	+5.8 / +5.9
64	28.2	42.8	72.2	45.8	74.9	+3.0 / +2.7
128	18.1	45.5	74.9	47.3	76.1	+1.8 / +1.3
200	13.3	46.9	75.7	48.1	76.3	+1.2 / +0.7

The prefix results show that the advantage over History-only is largest at the beginning of the target sequence and narrows as the sequence grows. History-only lacks enough observations to estimate a variance for the first two predictions; Table 21 reports the schedule gap and source-coefficient trajectory across prefixes.

A.11 Source-Only Prior-Strength Validation

To check whether fixed $\kappa = 8$ lies in a source-supported regime, we perform source-only pseudo-target validation within each outer fold over $\kappa \in \{2, 4, 8, 16, 32\}$. Each of the nine source subjects serves once as a pseudo-target, with the other eight forming the source prior. Candidates are replayed under the predict-before-update protocol defined in Section 3.1 and scored by predictive NLL; the fold choice minimizes mean NLL over its nine pseudo-targets. Table 22 reports the fold selections and comparison statistics. The results favor a moderate-prior range, rather than identifying a unique global value.

Table 22: Source-only pseudo-target validation of the SPPC prior strength. Avg. rank and Pseudo wins are computed over 90 pseudo-target comparisons; Outer selected is the number of outer folds in which each candidate minimizes mean predictive NLL; mean NLL regret is the fold-level excess over the best candidate, averaged across folds.

κ	Avg. rank	Outer selected	Pseudo wins	Mean NLL regret
2	3.5	0/10	0/90	0.004265
4	1.6	4/10	34/90	0.000308
8	1.4	6/10	50/90	0.000096
16	3.5	0/10	6/90	0.003993
32	5.0	0/10	0/90	0.013687

A.12 Coordinate-Recovery-Free CSLS Diagnostic

All headline RNSC results in the main paper use cosine similarity. Causal CSLS is included here only as an optional retrieval-time neighborhood correction for diagnostic reranking; it is excluded from the 48.10% main system so that the primary experiments isolate representation learning, normalization, and sequential calibration without an additional neighborhood-correction stage. Table 23 compares causal and full-set CSLS, with and without full-set target moments. The strongest in-house result comes from adding target moments to full-set CSLS. SCORE provides a separate reference based on an orthogonal target-to-image map estimated from unlabeled target landmarks (Cui et al., 2026).

Table 23: Coordinate-recovery-free CSLS diagnostics and the reported SCORE result on THINGS-EEG2.

Method / diagnostic	Coord. recovery	Top-1	Top-5
RNSC + causal CSLS (sequential)	no	51.15	80.55
RNSC + full-set CSLS	no	53.15	81.20
RNSC + transductive norm + full CSLS	no	**54.60**	82.70
SCORE (Cui et al., 2026)	orthogonal	53.23 ± 1.62	$\mathbf{83.55 \pm 1.13}$

A.13 Additional Intra-Subject Calibration Analysis

The main intra-subject result is aggregated over five independent UMVA+STAR-EEG runs. This supplementary comparison examines the schedules on corresponding selected checkpoints and their within-subject behavior; Table 24 gives the per-subject results. These schedules are applied post hoc to the same selected checkpoints as an information-access diagnostic, rather than defining the headline intra-subject protocol, which uses no sequential calibration.

Table 24: Intra-subject retrieval under different calibration schedules for one evaluation run.

	History-only			SPPC			Transductive		
Subject	Top-1	Top-5	Top-10	Top-1	Top-5	Top-10	Top-1	Top-5	Top-10
S1	88.5	96.0	98.0	92.0	98.5	99.5	92.5	97.5	99.0
S2	89.0	98.5	99.0	91.5	98.5	99.5	92.0	98.5	100.0
S3	90.0	97.5	99.0	94.5	100.0	100.0	95.0	100.0	100.0
S4	87.5	98.0	98.5	92.5	99.5	100.0	92.5	100.0	100.0
S5	84.5	96.0	97.0	87.5	96.0	98.5	89.5	98.5	99.0
S6	94.0	99.5	99.5	97.0	100.0	100.0	97.5	100.0	100.0
S7	90.0	98.0	98.0	91.5	99.5	99.5	93.0	99.5	100.0
S8	90.5	97.0	99.0	93.5	99.0	100.0	94.5	99.0	100.0
S9	87.0	97.5	98.0	91.5	99.5	100.0	91.0	99.5	100.0
S10	93.0	98.5	99.0	97.0	100.0	100.0	97.5	99.5	100.0
Mean	89.40	97.65	98.50	92.85	99.05	99.70	93.50	99.20	99.80

A.14 THINGS-MEG Intra-Subject Results

We additionally evaluate the representation on THINGS-MEG (Hebart et al., 2023) with five independent runs for four subjects (20 trained models). Table 25 summarizes the corresponding intra-subject retrieval results across subjects. This experiment is used only as a cross-modality transfer check of the representation recipe rather than as a cross-subject calibration benchmark; all reported values are intra-subject and aggregated over the five runs for each participant.

Table 25: THINGS-MEG intra-subject retrieval accuracy (mean $\pm$ sample SD, %).

Subject	Top-1	Top-5	Top-10
S1	41.50 ± 1.12	72.50 ± 2.98	82.10 ± 1.88
S2	86.90 ± 1.29	98.60 ± 0.42	99.90 ± 0.22
S3	67.20 ± 1.44	89.90 ± 1.34	95.50 ± 0.79
S4	45.80 ± 1.15	73.80 ± 1.64	85.20 ± 1.48
Overall	60.35 ± 0.70	83.70 ± 0.85	90.68 ± 0.58

A.15 Qualitative Inter-Subject Retrieval Examples with Sequential Calibration

Figure 5 complements the quantitative inter-subject evaluation with retrieval examples, using intra-subject performance as a subject-specific reference for the held-out-subject cases.

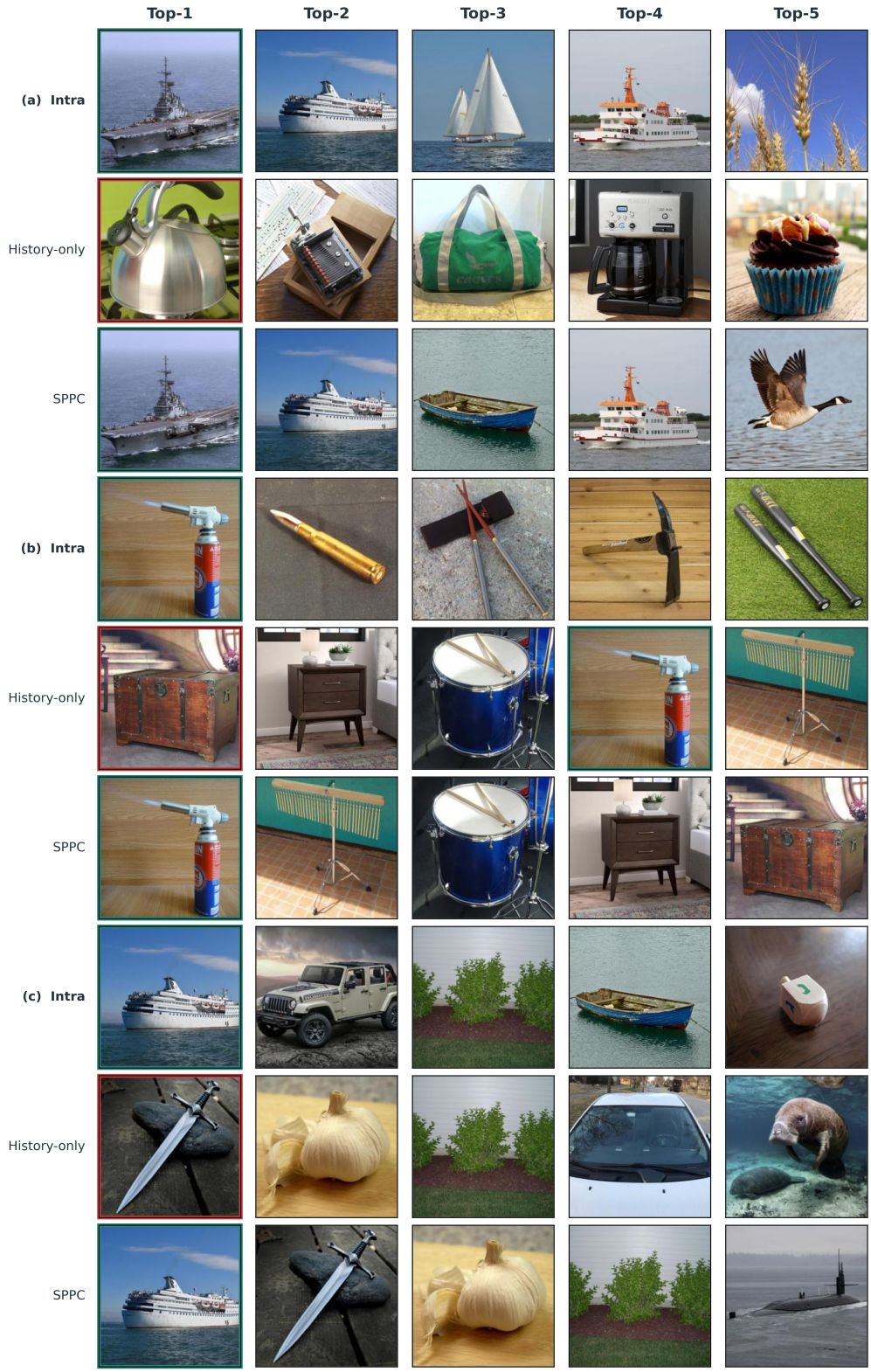

Figure 5: Qualitative examples of inter-subject retrieval with sequential calibration.