SAUNA: SOURCE-ANCHORED UNSUPERVISED NEURAL ADAPTATION FOR CROSS-SUBJECT EEG DECODING

Anonymous authors
Paper under double-blind review

ABSTRACT

Decoding seen images from EEG has advanced rapidly, but much of that progress has been measured *within* subjects, with a decoder trained and evaluated on the same person. However, practical adoption will require a more challenging cross-subject setting, in which the decoder must generalize to a person whose labeled data was never observed. We ask what actually changes when a decoder trained on some brains is applied to an unseen brain. A contrastive decoder aligns two representation spaces, an image side and a signal side, and we answer this question for each side by direct measurement. On the image side, we show that the depth at which a vision backbone should supply its alignment target is a property of the visual hierarchy rather than of the individual. The alignment target can therefore be *anchored*: estimated once from source subjects and shared by every new subject. Implemented as per-dimension selection over the depth of a multi-backbone feature pool, the anchored target yields a zero-shot decoder that attains **37.72%** 200-way top-1 accuracy on THINGS-EEG2, surpassing the strongest published result of 24.0%. On the signal side, we show that about half of the attainable accuracy is subject-specific and beyond the reach of source data alone. To recover this component without labels, we introduce SAUNA (Source-Anchored Unsupervised Neural Adaptation), which uses only *unlabeled* EEG from the target subject and never observes the trial-to-image pairing. SAUNA reads three properties of the subject's signal from such data: its position relative to the image manifold, its distribution shape, and its trial noise. One objective is instantiated for each property and optimized jointly with the anchor, so that the representation adapts without drifting from image space. SAUNA raises accuracy from 37.25% for a matched control to **47.77%**, a gain that is positive for all ten held-out subjects, recovers roughly a third of the subject-specific gap, and is, to our knowledge, the strongest label-free cross-subject result reported to date.

1 INTRODUCTION

Reconstructing or retrieving the image a person is viewing from their EEG has become an active research problem. NICE (Song et al., 2024) established contrastive alignment of a shallow convolutional EEG encoder to CLIP embeddings as a strong and reproducible baseline, and two years of successors have refined the encoder, the objective, and the image-side representation. Much of this progress has been measured *within* subjects: a decoder is fit to one subject's recordings and evaluated on held-out images from the same subject, a protocol that deliberately isolates the representational problem from the transfer problem. Under this protocol the field has advanced quickly, with the best reported accuracy on THINGS-EEG2 (Gifford et al., 2022) climbing from 13.8% (Song et al., 2024) to 82.6% (Du et al., 2026) in two years.

A decoder intended for practical deployment, however, must generalize to a subject it has never seen. The corresponding evaluation, which most published methods also report, is leave-one-subject-out (LOSO): training uses one set of subjects and testing another. The two frontiers have not moved together: over the same period, published LOSO accuracy on the identical recordings rose only from 6.2% (Song et al., 2024) to 11.8% (Li et al., 2024), 14.0% (Li et al., 2025), 21.8% (Du et al., 2026) and 24.0% (Zhou et al., 2026). The field is learning to read the brain's response to an image

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.

far faster than it is learning to read a new brain. This divergence motivates the central question of this paper:

When a decoder trained on some brains is applied to an unseen brain, what actually changes?

The question can be made precise by considering the structure of the decoder. A contrastive decoder aligns two representation spaces: an *image side*, which fixes the representation of the stimulus that the EEG must be mapped to, and a *signal side*, which supplies the EEG to be mapped. Any failure on an unseen subject must therefore originate on one of these two sides, and each side can be interrogated separately by direct measurement. We perform both measurements in this paper and obtain opposite answers: the image side is subject-invariant, whereas the signal side is not. Each finding contributes one component of our method. Fixing the alignment target on source subjects alone lifts our zero-shot accuracy on THINGS-EEG2 to 37.72%, against 24.0% for the strongest published result; adapting on the target's unlabeled EEG lifts it further to **47.77%**.

Reading depth is a property of the visual hierarchy, not of the subject. Within subjects, recent gains share one organizing principle: final-layer embeddings are optimized for semantic invariance while neural signals span several levels of abstraction, a discrepancy Du et al. (2026) name the granularity mismatch. Methods in this line blur the target (Wu et al., 2025), decompose it (Liu et al., 2026a), shrink it (Wu et al., 2026), or widen it across backbones and intermediate layers (Zheng et al., 2026; Jiang et al., 2026). If the appropriate reading depth were subject-specific, the alignment target itself would need to be re-estimated for every new person, and no decoder could fix it in advance. We measure this question directly: in a variance decomposition of the (subject $\times$ depth) encoding profile across ten subjects, the depth profile shared by all subjects carries roughly half the variance, while the subject-by-depth interaction accounts for under 1% (§4). The depth at which the visual hierarchy should be read is therefore a property of the hierarchy rather than of the individual. This invariance permits the alignment target to be *anchored*: estimated once from source subjects, implemented as per-dimension selection across the full depth of a multi-backbone feature pool, and fixed for every new subject (§4).

Signal geometry is a property of the subject, and is readable from unlabeled EEG. An oracle that additionally trains on the target subject's own labeled data, otherwise identical to the anchored decoder and evaluated on the same unseen concepts, reaches 73.53% against 37.72% (§6). About half of the attainable accuracy is therefore subject-specific and cannot be recovered from any quantity of source data. This deficit is one of data rather than of modelling, and it can be closed only with data from the target subject. Labeled data, however, is expensive to acquire from a new subject, requiring a synchronized and annotated recording session (Lotte et al., 2018), whereas unlabeled EEG carries no such burden. We show that unlabeled EEG reveals the *geometry* of a subject's signal through three readable properties, and SAUNA (Source-Anchored Unsupervised Neural Adaptation) instantiates one objective for each.

- **Position relative to the image manifold**: each target trial should match *some* training image confidently while different trials match *different* images; an InfoMax objective enforces both.
- **Distribution shape**: the target subject's embeddings should occupy the same region of the shared space as the source population; a moment-matching term aligns their first and second moments.
- **Noise geometry**: disjoint halves of a trial's repetitions are two views of one neural event; a repetition-consistency term makes their embeddings agree, using the grouping of repetitions and never the images.

The three act *jointly* with the source-side anchor, so that the shared encoder remains tied to image space at every step. Individually they are worth $+7.2$, $+1.5$ and $+1.8$ points; jointly $+10.5$, slightly more than their sum (§6). In summary, the reading depth of the visual hierarchy is shared across subjects and is therefore anchored from source data, whereas the geometry of the neural signal is specific to each subject and is adapted from that subject's unlabeled EEG.

The result. The two halves stack, and every step is reported: the feature pool read at its final layer scores 23.9%; anchoring it across depth reaches 37.72% with nothing from the target subject; and adaptation lifts a matched control from 37.25% to **47.77%** (§6). The improvement is positive for all

ten held-out subjects, and the labeled-data oracle of 73.53% bounds what remains to be recovered. SAUNA thereby closes roughly one third of the subject gap without using any label or test trial from the target subject, and most of this gain is already obtained from approximately one hour of unlabeled recording (§6.3).

2 RELATED WORK

EEG-to-image retrieval. The contrastive formulation we adopt was established by NICE (Song et al., 2024), which aligns a shallow convolutional EEG encoder to CLIP (Radford et al., 2021) image embeddings with a symmetric InfoNCE objective (van den Oord et al., 2018). Successors have varied every component of this recipe. On the EEG side, ATM (Li et al., 2024) replaced the encoder with a transformer over channel tokens. On the image side, Zheng et al. (2026) showed that fusing several vision backbones (CLIP, DINOv2 (Oquab et al., 2024), a latent diffusion autoencoder (Rombach et al., 2022; Podell et al., 2024)) substantially outperforms any single one. On the objective side, a broad line refines the loss or the supervision (Wei et al., 2024; Li et al., 2025; Wu et al., 2025; Liu et al., 2026b; Wu et al., 2026; Jo et al., 2026; Zhang et al., 2026; Liu et al., 2026a; Zhou et al., 2026). All report leave-one-subject-out accuracy, collected in Table 2. Our image side follows the fusion lineage, and our contribution is orthogonal to these lines and composes with them.

Intermediate representations. That intermediate layers of vision models align better with brain responses than final layers is a classical finding of the encoding-model literature (Yamins et al., 2014; Güçlü & van Gerven, 2015; Schrimpf et al., 2018), and decoding work has begun to exploit it by selecting or mixing intermediate layers (Benchetrit et al., 2024; López-Cardona et al., 2025; Skean et al., 2025; Yu et al., 2025), including the granularity-mismatch analysis of Du et al. (2026). What this literature leaves open is whether the best-aligned stage is shared across individuals or must be estimated per subject. We contribute a ten-subject variance decomposition that answers this question directly (§4).

Cross-subject transfer in BCI. Covariance whitening is the standard subject aligner in BCI (He & Wu, 2020), and unsupervised domain adaptation contributes adversarial alignment (Ganin & Lempitsky, 2015), moment matching (Sun & Saenko, 2016), and information maximization (Shi & Sha, 2012; Liang et al., 2020). SAUNA draws on the latter two. The closest work is SATTC (Huang & Zhu, 2026), a label-free calibration sharing our premise that a held-out subject's unlabeled signal carries usable alignment information, but under a different access model: it calibrates at test time on the target's test queries, jointly over the candidate set, whereas SAUNA adapts during training on the target's unlabeled training split, never its test split, and scores every query independently. Our access model is strictly more restrictive and more deployable, and we quote SATTC's 14.8% average top-1 as context rather than as a controlled comparison. Our contribution is not a new alignment primitive but the observation that the useful unlabeled signal is tripartite (position relative to the image manifold, distribution shape, noise geometry) and must act jointly with the supervised objective.

3 SETTING AND PROTOCOL

Data and task. We evaluate on THINGS-EEG2 (Gifford et al., 2022), which records ten subjects with 63-channel EEG at 250 Hz, and we use the standard public preprocessing of the NICE/ATM lineage (Song et al., 2024; Li et al., 2024). The training split contains $16{,}540$ images ($1{,}654$ concepts with 10 exemplars each), presented four times per image. The test split contains 200 images from 200 held-out concepts, presented 80 times each. The task is 200-way top-1 retrieval from the repetition-averaged test EEG: each query is scored by cosine similarity against the 200-image gallery and ranked by per-query `argmax`.

Decoder. We build on the most common decoder in this literature: the shallow convolutional EEG encoder of NICE (Song et al., 2024), applied to all 63 channels and trained to align each trial's embedding with the embedding of the image that evoked it, using symmetric InfoNCE with a learned temperature. On the image side we construct a feature pool from five frozen backbones (CLIP ViT-B/32 and ViT-L/14, DINOv2-B, an SDXL autoencoder, and CLIP-RN50), fused by one trained linear map per backbone into a shared 1024-dimensional space. Training uses AdamW (Loshchilov

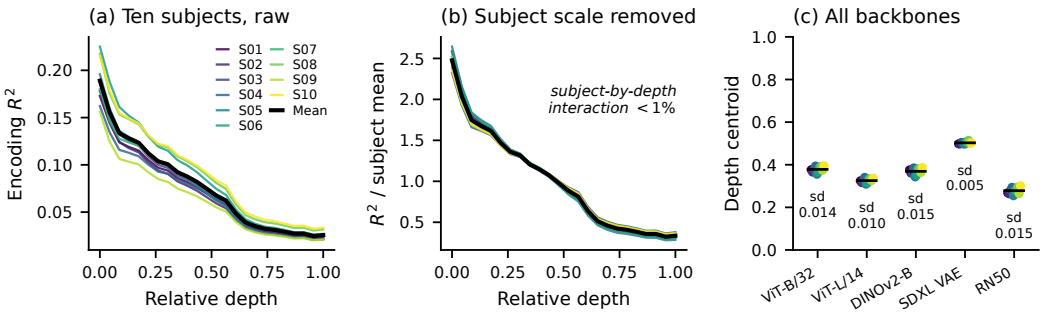

Figure 1: The (subject $\times$ depth) encoding profile is a shared shape with a per-subject amplitude. (a) Held-out encoding R^2 from each subject's EEG to every block of CLIP ViT-L/14. (b) The same profiles divided by their subject means collapse onto one curve. (c) The R^2-weighted depth centroid per subject and backbone. The across-subject standard deviation is 0.005 to 0.015 in relative depth.

& Hutter, 2019), 25 epochs, batch 1000, learning rate 2×10^{-4}, and a cosine schedule, with three seeds per subject. Appendix A.1 gives full details. Results are reported under two protocols, which share the decoder above and the reporting conventions of §6. They differ only in what the decoder may observe.

- **Zero-shot leave-one-subject-out (LOSO).** The decoder is trained on nine subjects and evaluated on the held-out subject's test split, from which it receives only the forward pass. Every statistic the encoder applies at evaluation is estimated on the nine training subjects only.
- **Unsupervised adaptation (UDA).** The decoder additionally observes the held-out subject's *unlabeled* training-split EEG, with the trial-to-image mapping discarded, together with the training image set, which is public and identical for every subject. The grouping of repeated presentations is retained. It is side information beyond the raw signal, and §5 states which objective relies on it. At no point does the decoder observe which image any target trial came from, or the target subject's test split.

A UDA number solves an easier problem than a zero-shot number and the two are not interchangeable. Every adaptation result we report is therefore accompanied by its own matched control, trained on identical data with the adaptation terms disabled, in the same job on the same GPU.

4 Reading depth transfers across subjects

Most EEG-to-image decoders align neural activity to the *final* layer of a vision backbone, the embedding optimized for semantic invariance (Song et al., 2024; Li et al., 2024). A growing body of work reads intermediate layers instead: neural responses span several levels of visual abstraction, and mid-depth features align with them better than the final embedding does (Benchetrit et al., 2024; Du et al., 2026; Zheng et al., 2026). This section establishes that the depth at which the hierarchy is best read is a property of the hierarchy rather than of the subject.

Is the optimum subject-specific? The literature disagrees: Du et al. (2026) conclude that "subject differences do not strongly alter the preferred representational stage", while SAMGA (Jiang et al., 2026) formulates a subject-dependent granularity mismatch and learns a per-subject residual on its layer router. The question matters here because a subject-specific optimum would be one of the few things a cross-subject method could recover from unlabeled data.

We measure this on the encoding side, where a per-subject quantity is available: for each subject and each backbone block we compute the held-out R^2 of a ridge map from that subject's EEG to the block's activations, giving a (subject $\times$ depth) profile per backbone. In raw form the ten profiles differ visibly in amplitude (Figure 1a); once each is divided by its own subject mean, they collapse onto a single shape (Figure 1b). The R^2-weighted depth centroid varies across subjects by a standard deviation of only 0.005 to 0.015 in relative depth, about one block, in all five backbones (Figure 1c), and in each of the four recognition backbones the subject-by-depth interaction absorbs under 1% of

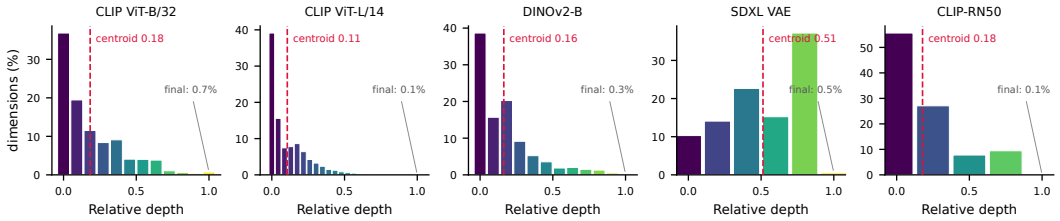

Figure 2: Where each dimension is best read. Distribution over relative depth of the per-dimension best layer (the layer of highest encoding R^2, averaged over the ten subjects; Appendix A.1), for each of the five backbones. The four recognition backbones are strongly early-weighted, and the final layer, the reading point of the standard lineage, retains almost no dimensions (0.1–0.7%); the SDXL autoencoder instead peaks mid-depth.

the variance, against roughly half for the shape of the depth profile itself (Appendix C). Subjects differ in *how well* their EEG is predicted, not in *where* along the hierarchy the correspondence peaks. Consequently a single fixed reading depth is appropriate for everyone, and adapting depth per subject has no headroom. This is a useful negative: depth is one of the few quantities that could otherwise be chosen from a new subject's unlabeled data.

The same measurement then chooses the depth. For each dimension of the feature pool (the five backbones shown in Figure 2), the anchor keeps the single layer with the highest encoding R^2 for each dimension, estimated on the source subjects only. Figure 2 shows the resulting per-dimension best-layer distribution. Anchored this way, the zero-shot decoder reaches **37.72%** 200-way top-1 (LOSO for ten subjects), against 23.90% for the same pool read at every backbone's final layer. What varies across subjects is overall alignment rather than preferred depth, and recovering that subject-specific component without labels is the task of the next section.

5 SAUNA: SOURCE-ANCHORED UNSUPERVISED NEURAL ADAPTATION

The previous section established what a new subject does not change, and hence what may be anchored; what a new subject does change must be recovered from the only data they can cheaply provide, unlabeled EEG. The organizing principle of the method is:

> *A subject's unlabeled EEG reveals the geometry of their signal: its position relative to the image manifold, its distribution, and its trial noise.*

Three quantities are separately readable without labels, and SAUNA instantiates exactly those three.

Overview. SAUNA trains a single EEG encoder f_θ, which maps a trial to a point (an *embedding*) in a shared representation space, together with an image encoder g_ϕ, which maps an image's features into the same space (Figure 3). Retrieval succeeds when a trial's embedding lies closest to the embedding of the image that evoked it. Training minimizes one objective with four parts,

$$\mathcal{L} = \underbrace{\mathrm{NCE}\big(f_\theta(x_s),\, g_\phi(v)\big)}_{\text{source anchor, always active}} + \lambda_{\text{im}}\mathcal{L}_{\text{im}} + \lambda_{\text{coral}}\mathcal{L}_{\text{coral}} + \lambda_{\text{rep}}\mathcal{L}_{\text{rep}},$$

optimized from scratch in a single stage. Subscripts s and t distinguish source from target throughout: x_s is a source-subject trial, v holds the features of the image that evoked it, and x_t is a target-subject trial. The anchor is the supervised term, possible only for the labeled source subjects. The three $\mathcal{L}$ terms are unsupervised, one per readable quantity, applied to the target subject's unlabeled trials and weighted by $\lambda_{\text{im}} = 0.5$, $\lambda_{\text{coral}} = 0.1$ and $\lambda_{\text{rep}} = 0.5$. Both the anchor and $\mathcal{L}_{\text{rep}}$ use the symmetric InfoNCE contrastive loss (van den Oord et al., 2018), written $\mathrm{NCE}(a, b)$.

InfoMax ($\mathcal{L}_{\text{im}}$): position relative to the image manifold. A well-aligned embedding of a visual trial should lie close to the image manifold: every target trial should match *some* training image confidently, and different trials should match *different* images (Shi & Sha, 2012; Liang et al., 2020). This is the one term that ties the target subject to image space, and the largest contributor (§6). For

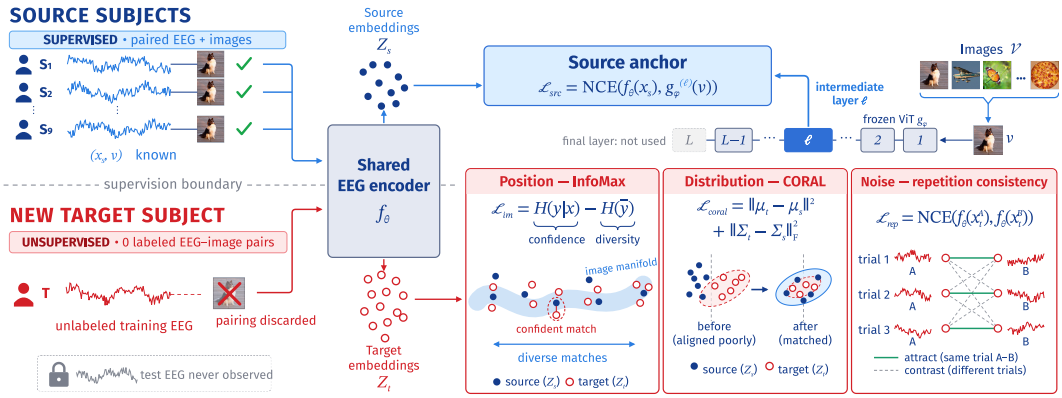

Figure 3: SAUNA overview. *Top, supervised*: source subjects $S_1, \ldots, S_9$ contribute paired EEG–image trials (x_s, v) from THINGS-EEG2; the shared encoder f_θ maps their trials to embeddings Z_s, which the source anchor $\mathcal{L}_{\text{src}}$ aligns with image embeddings read at intermediate depth. *Bottom, unsupervised*: the target subject T contributes unlabeled training EEG only; the pairing is discarded, and test EEG is never observed. The same encoder produces Z_t, shaped jointly with the anchor by InfoMax ($\mathcal{L}_{\text{im}}$: confident, diverse matches to the training set $\mathcal{V}$), moment matching ($\mathcal{L}_{\text{coral}}$: mean and covariance of Z_t to those of Z_s), and repetition consistency ($\mathcal{L}_{\text{rep}}$: disjoint repetition halves A, B attracted, other trials contrasted).

each target trial i in a batch of B trials, a softmax over similarities to a fixed gallery of training-image embeddings gives a probability distribution p_i over the gallery: the model's soft guess, involving no label, of which gallery image the trial resembles. With $H(p) = -\sum_k p_k \log p_k$ the entropy of a distribution, k indexing the gallery, the loss is

$$\mathcal{L}_{\text{im}} = \underbrace{\frac{1}{B}\sum_i H(p_i)}_{\text{confidence}} - \underbrace{H(\bar{p})}_{\text{diversity}}, \qquad \bar{p} = \frac{1}{B}\sum_i p_i.$$

The confidence term is small when every trial's guess commits to *some* image; the diversity term, the entropy of the batch-averaged guess, is large when different trials commit to *different* images. The name refers to the classical principle of information maximization (Linsker, 1988; Gomes et al., 2010): $H(\bar{p}) - \frac{1}{B}\sum_i H(p_i)$ estimates the mutual information between trials and assignments (Appendix D). The diversity term is load-bearing rather than cosmetic: confidence alone is minimized by collapsing every target trial onto a single gallery image, a failure that reduces the loss while destroying the representation.

Moment matching ($\mathcal{L}_{\text{coral}}$): distribution shape. With μ and Σ the mean and covariance of a batch of embedded trials, computed separately for the target (μ_t, Σ_t) and the source (μ_s, Σ_s), target embeddings should have source-like first and second moments (Sun & Saenko, 2016):

$$\mathcal{L}_{\text{coral}} = \left\|\mu_t - \mu_s\right\|^2 + \left\|\Sigma_t - \Sigma_s\right\|_F^2,$$

where $\|\cdot\|$ is the Euclidean norm and $\|\cdot\|_F$ the Frobenius norm: the new subject's embedding cloud should occupy the same region of the shared space, with the same spread and orientation, as the source population.

Repetition consistency ($\mathcal{L}_{\text{rep}}$): noise geometry. Each target trial is recorded several times. Two *disjoint* subsets of a trial's repetitions are two views of the same underlying event, so their embeddings should agree. This uses only the fact that repetitions are grouped; it never uses which image any trial came from. The grouping is nonetheless side information, since it identifies trials as sharing a stimulus, though never which; of the three terms this one is therefore the furthest from purely unsupervised, which is why we order it last. Split a target trial's repetitions into two disjoint halves and average each, giving two noisy views $x_t^{(1)}$ and $x_t^{(2)}$, whose embeddings the loss matches:

$$\mathcal{L}_{\text{rep}} = \text{NCE}\big(f_\theta(x_t^{(1)}),\ f_\theta(x_t^{(2)})\big).$$

Because the two views differ only in trial noise, matching them teaches the encoder to cancel this subject's noise, while the batch's other trials, serving as negatives, keep different neural events apart.

Table 1: Contribution of each objective, each arm paired against a matched control trained under identical conditions. Top-1 and top-5 are accuracies in %, mean ± standard deviation over the three per-seed ten-subject means, each followed by its gain Δ over the control. The sign column reports the paired per-subject outcome, (subjects improved)/(subjects worsened) out of ten, computed from each subject's seed-averaged top-1 difference against the control. Ten subjects, three seeds.

Configuration	top-1	Δ	top-5	Δ	sign
Control (no adaptation)	37.25±0.65	—	71.68±0.43	—	—
+ $\mathcal{L}_{\text{im}}$ only	44.43±0.58	+7.18	76.15±0.63	+4.47	10/0
+ $\mathcal{L}_{\text{coral}}$ only	38.75±1.19	+1.50	73.47±0.78	+1.78	9/1
+ $\mathcal{L}_{\text{rep}}$ only	39.00±0.54	+1.75	73.05±0.74	+1.37	7/2
SAUNA (all three)	**47.77**±0.92	+10.52	**80.70**±0.46	+9.02	10/0

Why the objectives must be joint. The anchor is the only term that contains ground truth, and it alone defines what the shared space means. The three target terms carry no per-trial labels, so optimized on their own they admit degenerate solutions in which the encoder satisfies them while drifting out of the image-aligned space. With the anchor active at every gradient step, no such drift is available. The natural alternative, fine-tuning a finished zero-shot decoder on the three target objectives with g_ϕ frozen and the anchor removed, is measurably worse: post-hoc adaptation raises top-1 from 37.35% to 40.90% ($+3.55$, $8/10$ subjects, ten subjects $\times$ three seeds; the base is a separate run within the seed noise of Appendix A.4), against $+10.52$ for the same objectives applied jointly. Bolting adaptation onto a trained decoder recovers about a third of what training with it recovers. Screening indicated the target signal is *data*-starved rather than under-weighted (Appendix D), so the final configuration matches target and source batch sizes; $\lambda_{\text{im}} \geq 2$ destabilizes training.

6 EXPERIMENTS

We report the final training epoch throughout; no epoch, schedule, or loss weight in this paper is chosen by comparing test numbers on the subjects we report. Configurations were screened on four subjects ($\{01, 03, 06, 08\}$), and the winning configuration was then run on all ten subjects with three seeds. Differences are assessed paired across subjects with a sign test. Normalization provenance, the selection history of the image feature pool, measured noise floors, and the protocol differences across published baselines are catalogued in Appendix A.

6.1 CONTRIBUTION OF EACH OBJECTIVE

Table 1 builds SAUNA objective by objective against its matched control, trained on identical data with the adaptation objectives disabled. The headline gain is $+10.52$ points, from 37.25% to **47.77%**, and the three objectives are complementary rather than redundant: individually they are worth $+7.18$, $+1.50$ and $+1.75$, and jointly $+10.52$, slightly exceeding their sum. The contributions are, however, distinctly unequal. InfoMax carries most of the effect and is the only objective that is individually significant at $10/10$ subjects; moment matching is smaller but consistent ($9/1$, $p = 0.021$); repetition consistency alone does not reach significance ($7/2$, $p = 0.18$) despite contributing within the combination. This ordering supports the design argument: the two EEG-internal signals say nothing about where images live, and supplying that constraint is what most improves alignment for a new subject.

6.2 COMPARISON WITH PUBLISHED RESULTS

Table 2 places SAUNA and its control in the field. The grouping does not move an inconvenient number out of the way: the two strongest 63-channel baselines, EEGiT at 24.0% and ViEEG at 22.9%, sit in our own block. Our depth-selected row and the matched control of Table 1 differ by 0.47, which is seed noise between two implementations of the same recipe (Appendix A.4); every adaptation claim is made against the matched control. The table also measures the electrode restriction itself: under our identical recipe, the 17-channel subset that helps within subjects costs 3.35 points across them (34.37% against 37.72%, negative for $9/10$ subjects, Appendix A.5).

Table 2: **200-way top-1/top-5 retrieval accuracy (%) on THINGS-EEG2, cross-subject.** Zero-shot LOSO methods are grouped by channel setup, the full 63-channel montage (left) and the 17-channel occipito-parietal subset (right); the final block reports label-free adaptation to the held-out subject. Chance is $0.5\%/2.5\%$; our rows are ten subjects $\times$ three seeds, final epoch.

Method	Top-1	Top-5
Zero-shot LOSO, 63 *channels*		
BraVL (Du et al., 2023)	1.8	7.0
NICE (Song et al., 2024)	6.2	21.4
ATM-S (Li et al., 2024)	11.8	33.7
MB2C (Wei et al., 2024)	11.9	32.0
Neural-MCRL (Li et al., 2025)	14.0	34.3
NeuroBridge (Zhang et al., 2026)	19.0	45.9
Liu et al. (2026b)	20.0	48.0
ViEEG (Liu et al., 2026a)	22.9	51.4
EEGiT (Zhou et al., 2026)	24.0	55.6
Ours (final layer)	23.90	54.05
Ours (depth-selected)	**37.72**	**71.52**

Method	Top-1	Top-5
Zero-shot LOSO, 17 *occ.-parietal ch.*		
UBP (Wu et al., 2025)	12.4	33.4
ATS (Wu et al., 2026)	14.0	35.8
HyFI (Jo et al., 2026)	15.1	37.2
Zheng et al. (2026)	20.0	44.1
Du et al. (2026)	21.8	49.4
Ours (final layer, 17 ch.)	21.35	49.30
Ours (depth-selected, 17 ch.)	**34.37**	**67.42**
Label-free adaptation, 63 *channels*		
SATTC (Huang & Zhu, 2026)	14.8	38.4
SAUNA	**47.77**	**80.70**

Two comparisons deserve care. The closest reported label-free cross-subject result, 14.8% average top-1 (Huang & Zhu, 2026), was not re-run and is obtained under a different access model (§2), so we quote it as context rather than as a controlled comparison; our own matched control (the control row of Table 1) is the comparison the claim rests on. Separately, in the strict zero-shot setting, a harder problem that is *not* interchangeable with SAUNA's, our pipeline reaches 37.72%, against 24.0% for the strongest peer-reviewed result (Zhou et al., 2026). We report this for completeness and do not present SAUNA's 47.77% as a zero-shot result.

Headroom. To bound what remains, we additionally trained an oracle on all ten subjects, including the held-out subject's own training split, otherwise identical to the zero-shot decoder and evaluated on the same unseen test concepts (ten subjects, three seeds, final epoch). It reaches 73.53%, against 37.72% for the zero-shot decoder. The two differ only in having seen the target subject, so the 35.8-point difference between them, the *subject gap*, is the value of that exposure, and SAUNA's $+10.52$ recovers roughly a third of it without labels.

6.3 PER-SUBJECT BEHAVIOUR AND THE UNLABELED-DATA BUDGET

Figure 4a shows the per-subject picture, with the final-layer baseline and the labeled-data oracle bracketing the paired comparison. SAUNA's gain over its matched control is **positive for every one of the ten held-out subjects** (sign test, $10/10$) and is largest where the zero-shot decoder is worst (S03 gains 14.0 points from a base of 25.7%), which is the behaviour one wants from an adaptation method. Top-5 moves in step, from 71.68% to 80.70%.

Zero-shot decoding and full adaptation are endpoints of a continuum, indexed by how much unlabeled EEG the target subject provides; Figure 4b varies that amount with everything else fixed, at the paper's full statistics. Three features matter for deployment. First, a little data is worse than none: at about two minutes of recording, adaptation loses 4.5 points and is negative for all ten subjects, as the objectives overfit an unrepresentative sample of the signal; break-even comes at roughly seven minutes ($+0.2$, $4/10$ subjects). Second, the curve then rises quickly: twenty-two minutes buys $+4.0$ ($9/10$) and about one hour buys $+8.7$ ($10/10$), 83% of the full-session gain. Third, the benchmark's remaining 2.6 hours add only $+1.8$ more. A practical calibration session is therefore closer to one hour than to 3.7 hours, and adaptation from recordings under ten minutes should not be attempted.

7 LIMITATIONS

The adaptation setting has requirements. Two of the three objectives constrain the calibration session: $\mathcal{L}_{\text{im}}$ assumes the target subject viewed the training stimulus set, and $\mathcal{L}_{\text{rep}}$ requires repeated presentations. The deployment scenario is thus a calibration session over the same stimulus set with repetitions, in which one never needs to record or align which stimulus was shown when.

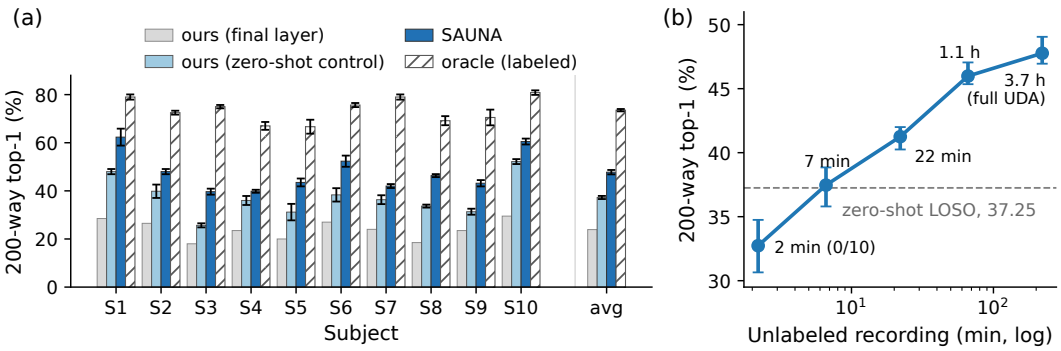

Figure 4: (a) Per-subject and average 200-way top-1 accuracy (%): our feature pool read at its final layer, our matched zero-shot control, SAUNA, and the labeled-data oracle. Bars are means over three seeds and error bars their standard deviations. (b) Accuracy against the amount of unlabeled target-subject EEG available to SAUNA (ten subjects, three seeds, final epoch; error bars span the three per-seed ten-subject means). One row of target data is one image condition with its four repetitions, 0.8 s of presentation at the dataset's 200 ms stimulus onset asynchrony, so the horizontal axis converts directly to recording time. The two-minute point is below the zero-shot baseline for all ten subjects.

This removes the labelling and synchronization burden, not the session: millisecond-level EEG–stimulus alignment is the costly, failure-prone component of calibration, and it is precisely what our objectives never use (Appendix D). $\mathcal{L}_{\text{coral}}$ carries neither assumption and would apply to arbitrary EEG from a new user, worth $+1.5$ on its own (Table 1). Finally, the session has a measured floor: below roughly ten minutes of recording, the adaptation objectives overfit and fall below the zero-shot baseline (§6.3), so adaptation should simply be disabled for very short sessions.

Scale. THINGS-EEG2 has ten subjects, and participant count is the one lever we cannot push. Table 4 (Appendix B.4) varies the number of training subjects with the number of gradient steps held constant. A log-linear fit gives $+11.1$ top-1 per doubling ($r = 0.99$) with no visible flattening at nine, which suggests that participant count remains the binding constraint and that substantially higher cross-subject accuracy is reachable primarily by recording more people. We report this as suggestive rather than established: the curve is two held-out subjects at one seed, and extrapolating it past $k = 9$ is exactly what this dataset cannot test.

8 Conclusion

Cross-subject EEG-to-image decoding lags within-subject decoding for a reason that is measurable and specific: 35.8 points of the deficit are attributable to never having seen the target subject. Given that diagnosis, we introduced SAUNA, which extracts from a target subject's unlabeled EEG exactly the three quantities such data can legitimately reveal (position relative to the image manifold, distribution shape, and noise geometry) and optimizes them jointly with the source-side objective so that the representation is shaped without being unmoored. The result improves 200-way top-1 from 37.25% to 47.77%, positive for all ten held-out subjects. We hope the diagnostic framing is as useful as the method: in this problem, the question of *what unlabeled data can tell you* turned out to be more productive than the question of which architecture to use next.

AI use statement

In this work, we used generative AI tools for manuscript writing and editing, for implementing analysis and experiment-orchestration code, and for preparing figures. We have not used generative AI tools to generate, impute, or modify experimental data, and AI involvement in data collection is not applicable to this work, since all recordings come from the public THINGS-EEG2 benchmark. Additionally, we used generative AI tools for literature search and formatting assistance. We have reviewed all AI-assisted work: AI-drafted text was revised and approved by the authors, AI-written

code was inspected and its outputs verified against the underlying run artifacts, reported numbers were recomputed from the raw result files, and AI-suggested citations were checked against the original papers. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Reproducibility statement

An anonymous repository containing the training and evaluation code, the run configuration of every experiment, and the scripts that generate each figure and table is available at `https://anonymous.4open.science/r/SAUNA-4DCC`. All experiments use the public THINGS-EEG2 dataset with the standard public preprocessing of the NICE/ATM lineage, named in Appendix A.2. The decoder and its training configuration are summarized in §3 and fully specified in Appendix A.1, including the per-dimension depth selection that defines the anchor and the exact definition of every reported arm. The evaluation protocol (what the zero-shot and adaptation settings may observe, the per-query scoring rule, and the provenance of every normalization statistic) is fixed in §3 and Appendix A, and the measured noise floors and the comparability of published baseline numbers are discussed in Appendices A.4 and A.5. Every reported number is the final training epoch, with no checkpoint selected on test data, and each table caption states its subject and seed counts.

References

Yohann Benchetrit, Hubert Banville, and Jean-Rémi King. Brain decoding: Toward real-time reconstruction of visual perception. In *International Conference on Learning Representations (ICLR)*, 2024.

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. *Transactions on Machine Learning Research (TMLR)*, 2026. ISSN 2835-8856. URL `https://openreview.net/forum?id=SzDHkICZpm`. arXiv:2601.21948.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning (ICML)*, volume 37 of *PMLR*, pp. 1180–1189, 2015.

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.

Ryan Gomes, Andreas Krause, and Pietro Perona. Discriminative clustering by regularized information maximization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 23, 2010.

Umut Güçlü and Marcel A. J. van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27): 10005–10014, 2015. doi: 10.1523/JNEUROSCI.5023-14.2015.

Matthias Guggenmos, Philipp Sterzer, and Radoslaw Martin Cichy. Multivariate pattern analysis for MEG: A comparison of dissimilarity measures. *NeuroImage*, 173:434–447, 2018. doi: 10.1016/j.neuroimage.2018.02.044.

He He and Dongrui Wu. Transfer learning for brain-computer interfaces: A Euclidean space data alignment approach. *IEEE Transactions on Biomedical Engineering*, 67(2):399–410, 2020. doi: 10.1109/TBME.2019.2913914.

Qunjie Huang and Weina Zhu. SATTC: Structure-aware label-free test-time calibration for cross-subject EEG-to-image retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. arXiv:2603.20738.

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. *arXiv preprint arXiv:2604.17782*, 2026. v2; code at `https://github.com/LinJiang8/SAMGA`.

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 (AAAI)*, volume 40, 2026. doi: 10.1609/aaai.v40i7.37476. arXiv:2603.22721.

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 (NeurIPS)*, volume 37, 2024.

Yueyang Li, Zijian Kang, Shengyu Gong, Wenhao Dong, Weiming Zeng, Hongjie Yan, Wai Ting Siok, and Nizhuan Wang. Neural-MCRL: Neural multimodal contrastive representation learning for EEG-based visual decoding. In *IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2025. doi: 10.1109/ICME59968.2025.11210130. arXiv:2412.17337.

Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? Source hypothesis transfer for unsupervised domain adaptation. In *International Conference on Machine Learning (ICML)*, volume 119 of *PMLR*, pp. 6028–6039, 2020.

Ralph Linsker. Self-organization in a perceptual network. *Computer*, 21(3):105–117, 1988. doi: 10.1109/2.36.

Minxu Liu, Donghai Guan, Chuhang Zheng, Chunwei Tian, Jie Wen, and Qi Zhu. ViEEG: Hierarchical visual neural representation for EEG brain decoding. In *International Conference on Machine Learning (ICML)*, 2026a. arXiv:2505.12408.

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 (AAAI)*, volume 40, pp. 17580–17588, 2026b. doi: 10.1609/aaai.v40i21.38813.

Ángela López-Cardona, Sebastián Idesis, Mireia Masias-Bruns, Sergi Abadal, and Ioannis Arapakis. Brain–language model alignment: Insights into the platonic hypothesis and intermediate-layer advantage. In *UniReps Workshop, Advances in Neural Information Processing Systems (NeurIPS)*, volume 322 of *PMLR*, pp. 81–101, 2025.

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

Fabien Lotte, Laurent Bougrain, Andrzej Cichocki, Maureen Clerc, Marco Congedo, Alain Rakotomamonjy, and Florian Yger. A review of classification algorithms for EEG-based brain–computer interfaces: a 10 year update. *Journal of Neural Engineering*, 15(3):031005, 2018. doi: 10.1088/1741-2552/aab2f2.

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

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations (ICLR)*, 2024.

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 *International Conference on Machine Learning (ICML)*, volume 139 of *PMLR*, pp. 8748–8763, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022.

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

Yuan Shi and Fei Sha. Information-theoretical learning of discriminative clusters for unsupervised domain adaptation. In *International Conference on Machine Learning (ICML)*, 2012.

Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. Layer by layer: Uncovering hidden representations in language models. In *International Conference on Machine Learning (ICML)*, 2025.

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 (ICLR)*, 2024.

Baochen Sun and Kate Saenko. Deep CORAL: Correlation alignment for deep domain adaptation. In *European Conference on Computer Vision (ECCV) Workshops*, pp. 443–450, 2016.

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

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 (ACM MM)*, pp. 8992–9000, 2024. doi: 10.1145/3664647.3681292.

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

Lukun Wu, Jie Li, Ziqi Ren, Kaifan Zhang, and Xinbo Gao. Shrinking the teacher: An adaptive teaching paradigm for asymmetric EEG-vision alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 40, pp. 17859–17867, 2026. doi: 10.1609/aaai.v40i21.38844.

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 (PNAS)*, 111(23):8619–8624, 2014. doi: 10.1073/pnas.1403112111.

Zony Yu, Yuqiao Wen, and Lili Mou. Revisiting intermediate-layer matching in knowledge distillation: Layer-selection strategy doesn't matter (much). In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2025*, pp. 1686–1694, Mumbai, India, 2025. doi: 10.18653/v1/2025.findings-ijcnlp.105.

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 (AAAI)*, volume 40, pp. 18028–18036, 2026. doi: 10.1609/aaai.v40i21.38863. arXiv:2511.06836.

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 (ICLR)*, 2026. arXiv:2602.07495.

Jiahao Zhou, Chenghao Xu, Wei Wang, Erkun Yang, and Cheng Deng. EEGiT: Teaching vision transformers to understand the EEG signal. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 40438–40447, 2026.

A PROTOCOL DETAILS

A.1 IMPLEMENTATION DETAILS

EEG encoder. The encoder is the official NICE architecture (Song et al., 2024): a shallow convolutional patch embedding over the 63-channel, 250 Hz input followed by a residual projection head, producing one embedding per trial in the shared space. No architectural changes are made; our cross-subject use differs from the within-subject original only in the training data.

Image-side feature pool. Five frozen backbones supply candidate features: CLIP ViT-B/32, CLIP ViT-L/14, DINOv2-B, the SDXL autoencoder, and CLIP-RN50. Each ViT backbone contributes activations from blocks spaced across its depth, so the pool spans the visual hierarchy rather than final layers alone. Every pooled dimension is scored on the source (training) subjects only, per-backbone quantile thresholds determine which dimensions are retained, and the retained features are fused by one trained linear map per backbone into the shared 1024-dimensional space. The selection history of the pool's composition is disclosed in Appendix A.3.

Training. Symmetric InfoNCE with a learned temperature; AdamW (Loshchilov & Hutter, 2019), 25 epochs, batch 1000, learning rate 2×10^{-4}, cosine schedule; three seeds per subject, final epoch, seed means reported.

The named arms. Three configurations of this one pipeline recur throughout the paper. *Final layer* (23.90%): the endpoint of the depth sweep, reading every backbone at its final block with everything else unchanged; this is our own ablation, not a reproduction of any published method, and its numerical proximity to EEGiT's 24.0% is coincidental. *Ours, depth-selected* (37.72%): the full anchored target, with per-dimension selection across all depths, trained in the plain harness. *Matched control* (37.25%): the identical zero-shot recipe executed inside the adaptation harness with the three adaptation terms disabled; it differs from the plain harness only in implementation, the difference is inside seed noise (Appendix A.4), and every adaptation claim is paired against it.

A.2 PREPROCESSING AND NORMALIZATION PROVENANCE

The standard public preprocessing (`Preprocessed_data_250Hz`) applies multivariate noise normalization (Guggenmos et al., 2018) per subject, using statistics estimated from that subject's training split only. We note this explicitly because it is not universal: UBP, HyFI and the ATS lineage instead train on a separately whitened variant of the same recordings, and numbers are routinely tabulated across the two pipelines as if they were interchangeable.

In the zero-shot setting, every statistic the encoder applies to the held-out subject's forward pass is estimated on the nine training subjects only: the per-channel mean and scale, and the BatchNorm running statistics, which are frozen at evaluation so the 200 queries never share batch statistics. The one exception is the dataset-standard MVNN whitening above, which is applied per subject from that subject's own *training* split before any model sees the data. This is a property of the public preprocessing and is shared by every method in Table 2, but it is not nothing, and we state it rather than claim a stricter protocol than we run.

A.3 MODEL SELECTION AND THE FEATURE POOL'S HISTORY

The stopping epoch is never selected; the 25-epoch schedule is justified by validation-subject curves, and the reading depth by held-out training concepts, in both cases without reading test data. The four screening subjects $\{01, 03, 06, 08\}$ have a zero-shot mean of 37.67%, matching the ten-subject mean of 37.72%. Screening runs are marked as such wherever they appear, and are single-seed unless stated.

One component does carry a selection history, and we state it rather than let the paragraph above imply otherwise. The composition of the five-backbone feature pool was chosen during preliminary within-subject experimentation on this dataset, by adding backbones and observing within-subject retrieval accuracy. The per-dimension selection *within* the pool is clean, being scored on training subjects only, and the reading depth was re-justified on held-out training concepts, but the choice

of which backbones to include was informed by benchmark accuracy, as is standard practice in this literature. We therefore treat the zero-shot pipeline as a strong baseline rather than as a controlled demonstration that five backbones are optimal, and every adaptation claim in this paper is made against our own matched control rather than against that number.

A.4 Measured noise floors

We measured the relevant noise floors directly by repeating identical configurations, and they differ by an order of magnitude depending on what is being compared. Across seeds, the standard error of a ten-subject three-seed mean is 0.3–0.8 points; the headline paired difference of §6 reproduces at $+11.80$, $+10.50$ and $+9.25$ across the three seeds ($\mathrm{SE} = 0.74$), $10/10$ subjects positive in each seed independently. A *single-subject* figure is far noisier: 200 test queries carry a binomial floor of roughly ± 3.5 points, and the epoch-to-epoch standard deviation within one run's plateau is 1.47. We therefore interpret ten-subject paired means and treat per-subject cells as illustrative. Our depth-selected row of Table 2 and the matched control of Table 1, produced by two implementations of the same recipe (the adaptation harness with its terms disabled, and the plain harness) run at the same schedule on the same data, differ by 0.47, inside these floors.

A.5 Comparability of published numbers

We report every baseline as its authors reported it and did not re-run any method; each cell of Table 2 was verified against the originating paper's own table rather than against another paper's comparison table. Cross-subject numbers on this benchmark are therefore not produced by a single pipeline, and four differences are large enough to matter. *First, the electrode set is not standard.* NICE, ATM, MB2C, Neural-MCRL, ViEEG, EEGiT, NeuroBridge and Liu et al. (2026b) use all 63 channels for their cross-subject runs, while UBP, ATS, HyFI, Zheng et al. (2026) and Du et al. (2026) use a 17-channel occipito-parietal subset. Liu et al. (2026b) do not state the count in the paper; their released preprocessing selects the full 63-channel montage, and we group them accordingly. ViEEG states "all 64 channels", which we take to mean the full montage (63 EEG channels in this dataset). We use all 63, and Table 2 is grouped accordingly. Two observations about this split are worth recording. First, Du et al. (2026) is the only study we found that measures the choice, and it reports that frontal, central and temporal channels carry almost no decodable signal and that including them costs 12–18 points within subjects, which is a solid justification for the subset in that setting. Second, the two methods we found that evaluate both settings (NeuroBridge in peer-reviewed work, SAMGA in a preprint) adopt the subset *within* subjects and the full montage *across* them, without stating a reason in either case. Our own measurement supplies one: rerunning our zero-shot recipe with the single-factor change to the 17-channel subset (ten subjects, three seeds, final epoch, otherwise identical to Appendix A.1) yields 34.37% against 37.72% for the full montage, a loss of 3.35 points that is negative for $9/10$ subjects (two-sided sign test, $p = 0.021$); at the final layer the same restriction costs 2.55 points (21.35% against 23.90%, one seed). The restriction that is worth 12–18 points within subjects therefore reverses sign across them, which explains why cross-subject work routinely drops it. *Second, model selection differs.* ATM, Neural-MCRL and NeuroBridge select the reported epoch on the test split, which we verified in their released training code; none of the three constructs a validation split. We report the final epoch. *Third, the preprocessing lineage differs*, as noted in Appendix A.2. *Fourth, evaluation is otherwise uniform*: to our knowledge every method above averages the 80 test repetitions into one query and scores by per-query `argmax` over the 200-image gallery, with the exception of SATTC, whose calibration is fitted over the full candidate set so that its scores are not independent across queries. One further reading note: BraVL's leave-one-subject-out trains on the other nine subjects' *averaged* data rather than their trials. We make no claim that the remaining differences are negligible, and we regard our own matched control (the control row of Table 1) as the comparison our claims actually rest on.

B Additional analyses

B.1 Most of the apparent headroom in epoch selection is not real

It is natural to suppose that a decoder's best epoch is meaningfully better than its last, and therefore that a better stopping rule is worth points. On this benchmark that supposition is mostly false, and

Table 3: Scoring rules applied to identical similarity matrices from our zero-shot recipe (ten subjects $\times$ three seeds); top-1 in %. The 37.15% base of this run set differs from the 37.72% headline within the seed noise of Appendix A.4. Only per-query `argmax` is used anywhere in this paper; the others are reported to quantify what is being declined.

Rule	Uses information across queries?	top-1	Δ
Per-query `argmax`	no	37.15	—
CSLS (hubness correction)	yes, gallery popularity	47.45	$+10.30$
Hungarian (exact assignment)	yes, one-to-one constraint	56.42	$+19.27$
Sinkhorn (entropic assignment)	yes, one-to-one constraint	**56.98**	$+19.83$

the reason is measurable. Over 30 runs of our zero-shot recipe, the best test epoch averages 40.52% and the final epoch 37.72%, an apparent best-versus-final gap of 2.80 points. The accuracy curve, however, does not climb over that range: it plateaus at 37.63% with a within-run standard deviation of 1.47 points across epochs, which is what 200 test queries produce. Simulating the maximum of 15 independent draws of that noise around the plateau predicts a best-versus-plateau gap of $+2.56$; the observed best-versus-plateau gap is $+2.88$. So roughly 89% of the apparent headroom is the maximum of a noisy metric rather than signal that a stopping rule could locate.

This is consistent with a rule we can test directly. Choosing the stopping epoch on a held-out *validation subject*, a principled criterion that never touches the test subject, yields 36.82%, below the final epoch of the same configuration (37.40%); both numbers sit under the nine-subject 37.72% because holding out a validation subject removes one training subject. A better rule cannot recover what is not there. We accordingly report the final epoch throughout, and we suggest that best-epoch figures on this benchmark be read with the 200-query noise floor in mind.

What *does* help is reducing that variance rather than exploiting it. Averaging the similarity matrices of three independently seeded models (an ordinary ensemble, with per-query `argmax` preserved) raises zero-shot top-1 from 37.15%, the per-run base of the matrices in §B.2, to 39.20% ($+2.05$, $10/10$ subjects, $p = 0.002$). Averaging model *weights* over the plateau (SWA) gives $+0.15$ and is not significant, which is what the analysis above predicts: the plateau's level, not its variance, is what limits a single model.

B.2 What the gallery structure is worth, measured

Our evaluation scores each query independently, with no batch-level or gallery-level operation (§3). Since that decision costs us accuracy, it is worth quantifying rather than asserting. The 200 test queries pair one-to-one with the 200 gallery images, and a decoder that exploits the pairing is solving an assignment problem instead of a retrieval problem. We applied three progressively stronger cross-query rules to the *same* similarity matrices that produce our reported numbers, so the decoder is held fixed and only the scoring changes (Table 3).

Exact assignment is worth $+19.27$ points, positive for all ten subjects. This is not a small effect that only appears at high accuracy: regressing the gain on base accuracy over the thirty individual matrices, which span 24.0% to 53.0% top-1, gives a slope of $+0.12$ points of gain per point of base accuracy; the gain is essentially flat over the entire range in which published cross-subject numbers sit. A method that adopted Hungarian scoring without comment would post a cross-subject number in the mid-fifties while decoding no better than one in the high thirties. We note this because nothing in the benchmark prevents it, and because the size of the effect is large enough that scoring rules deserve to be stated explicitly in this literature rather than assumed.

B.3 A self-supervised auxiliary helps only where labels are absent

The repetition-consistency objective is worth $+1.75$ points when applied to the unlabeled target subject (Table 1), and considerably more in combination. Applied instead to the nine *labeled* source subjects, a manipulation that requires no protocol relaxation whatsoever, the identical loss costs 8.4 points at $\lambda = 0.5$ and 10.8 at $\lambda = 1.0$ (four subjects, one seed, paired against a control in the same job; both effects are far outside the ± 1.5-point resolution of that screen, and monotone in λ).

The asymmetry has a natural reading. On the source subjects the image-alignment loss already supplies a supervisory signal, and the auxiliary term competes with it, spending encoder capacity on trial-invariance at the expense of image alignment. On the target subject no supervision exists, so the same term is the only structure available and is purely additive. We suggest this is worth bearing in mind more generally when adding self-supervised auxiliaries to partially supervised pipelines: their sign can depend on the presence of labels rather than on the objective itself.

B.4 SUBJECT-COUNT SCALING

Table 4 accompanies the scale discussion in §7.

Table 4: Zero-shot top-1 accuracy (%) against the number of training subjects, with gradient steps held constant across k. Two held-out subjects (01, 06), one seed, final epoch; the $k = 9$ column is the ten-subject recipe restricted to the same two subjects. Thin, and reported as such: the trend is clean, but two subjects cannot establish the absence of saturation.

Training subjects k	1	2	4	6	8	9
top-1	7.8	22.0	27.2	36.5	41.5	45.3

C DEPTH MEASUREMENTS

Table 5 sweeps the relative depth at which every backbone of the feature pool is read, under the zero-shot LOSO protocol with everything else held fixed. The result is a pronounced inverted-U: mid-depth is worth $+12.5$ points over either the first block or the final layer, which are equally poor, and the optimum lies at the same relative depth as the within-subject optimum.

Table 5: Cross-subject top-1 accuracy (%) as a function of the relative depth at which every backbone is read. Ten subjects, one seed; the 0.4-versus-0.6 ordering is inside the noise, the inverted-U far outside it.

Relative depth	0.0	0.2	0.4	0.6	0.8	1.0
top-1	23.7	29.0	**36.4**	35.3	30.8	23.9

Table 6 reports the full variance decomposition summarized in §4 and visualized in Figure 1: for each backbone, the share of variance in the (subject $\times$ depth) encoding-R^2 profile carried by the common depth shape, by a per-subject offset, and by the subject-by-depth interaction, together with the across-subject standard deviation of the R^2-weighted depth centroid. In every backbone the interaction term absorbs under 1% of the variance.

Table 6: Variance decomposition of the (subject $\times$ depth) encoding-R^2 profile, ten subjects. "Shape" is the common depth profile, "offset" a per-subject scale, and "interaction" the subject-specific departure from the common shape. The last column is the across-subject standard deviation of the R^2-weighted depth centroid, in units of relative depth.

Backbone	blocks	shape	subject offset	interaction	sd of centroid
CLIP ViT-B/32	12	47.6%	4.1%	0.7%	0.014
CLIP ViT-L/14	24	48.6%	2.0%	0.8%	0.010
DINOv2-B	12	48.3%	2.6%	0.8%	0.015
CLIP-RN50	5	49.5%	0.8%	0.1%	0.015

D PASSAGES COMPRESSED OUT OF THE MAIN TEXT

This appendix preserves, essentially verbatim, explanatory passages compressed out of the main text for space; nothing here is load-bearing for the claims.

The InfoNCE loss (§5). Given a batch of paired embeddings, $\mathrm{NCE}(a, b)$ is in effect a softmax classification over the batch, run in both directions. It does not merely pull true pairs together; it demands that each item be closer to its partner than to all competitors, which is what retrieval at test time requires.

The InfoMax objective (§5). The entropy $H(p)$ is small when the distribution commits to one option and large when it is spread out. The mutual information $H(\bar{p}) - \frac{1}{B}\sum_i H(p_i)$ measures how much the model's guesses reveal about which trial produced them. Guesses that are individually decisive and collectively varied are maximally informative, which is how a correct decoder's embeddings behave; guesses that collapse onto a single image, or that stay uniformly uncertain, reveal nothing.

The unconsumed mapping (§7). One might press further: a system that controls the stimuli holds the trial-to-image mapping in principle, so in such a system the mapping is not absent but unconsumed. What the setting removes is the requirement that the mapping be recorded, synchronized, and trusted. Millisecond-level EEG–stimulus alignment is the costly and failure-prone component of a calibration session, and it is precisely the information our objectives never use: stimulus-set membership and repetition grouping survive coarse timing, whereas the trial-to-image mapping does not.

The weight-versus-batch screening (§5). In screening we found the target signal to be data-starved rather than under-weighted: raising λ_{rep} from 2.0 to 4.0 costs 2.1 points, while enlarging the target batch at fixed loss weights helps, and the best arm of the sweep was the one with the smallest λ_{rep} and the largest target batch. We report the direction rather than a single number because the sweep varied weight and batch size together, and their contributions cannot be cleanly separated from it.