WHAT DOES COMPARING SUPERVISION TARGETS IDENTIFY? SEPARATING SUPERVISED VECTORS FROM CANDIDATE WEIGHTING

Anonymous authors
Paper under double-blind review

ABSTRACT

When one supervision target outperforms another, what does the comparison establish about its supervised vectors? In sample-coupled objectives, changing a target can also change the candidate weights governing its learning updates. We test the adoption consequence of separating these roles: keep paired and candidate vectors fixed and replace their weighting input with reference supervision. In controlled electroencephalography (EEG)-to-visual learning, this substitution turns a coordinate-rescaled target from harmful to beneficial on a cross-image concept-ranking task with model-specific neural templates. A control matching actual AdamW step lengths while retaining its own image-parameter update direction remains harmful. A prospective five-crop test also reproduces a natural target's preference for a reference-informed temperature projection with matched update lengths and positive adoption in the same setting. These interventions retain reference information in weighting; they establish task benefits conditional on that treatment. A negative ordinary comparison therefore need not justify discarding the supervised vectors: their adoption verdict can reverse against the same reference after replacing only the weighting input. Target comparisons support configuration selection, while content attribution requires evidence separating these roles.

1 INTRODUCTION

When one supervision target outperforms another, what does the comparison establish about the information it supplies? Target design adapts teacher outputs to student capacity (Qian et al., 2025), filters image detail, or compresses visual features (Wu et al., 2025; 2026). Yet changing a target can change both what a learner matches and how examples shape its updates. A gain supports selecting the resulting training configuration; attributing it to the supervised vectors requires separating these roles. We test this distinction in controlled electroencephalography (EEG)-to-visual learning: retaining the vectors while replacing only their weighting input reverses adoption against the same reference, beyond a control matching actual AdamW step lengths.

In *sample-coupled objectives*, one example's loss depends on other examples. Targets supply both supervised vectors and the similarities that determine *candidate weights*, the coefficients combining compared vectors in an update. Holding the loss formula fixed therefore need not hold its induced weights fixed. We call improvement over reference supervision under a specified treatment *positive adoption*; its practical value also depends on effect size and treatment cost.

Prior work already separates supervision and weighting. DKD uncouples non-target knowledge transfer from teacher confidence (Zhao et al., 2022); CSD matches logit differences under vocabulary-pair weights (Kim et al., 2026); contrastive coordinate-wise optimization separates pair weights from representation updates (Tian, 2022). Coupling and temperature controls (Yeh et al., 2022; Huang et al., 2023), teacher similarity distributions (Zheng et al., 2021; Denize et al., 2023), and teacher-guided triplet reweighting (Ye et al., 2020) establish related weighting mechanisms.

Our distinguishing contrast keeps a new target's paired and candidate vectors, restores reference supervision only in the weighting input, and compares both treatments with the *same reference*. Improving on ordinary weighting could merely reduce harm; our primary result crosses that common baseline from harmful to beneficial. It is a controlled counterexample to discarding supervised

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

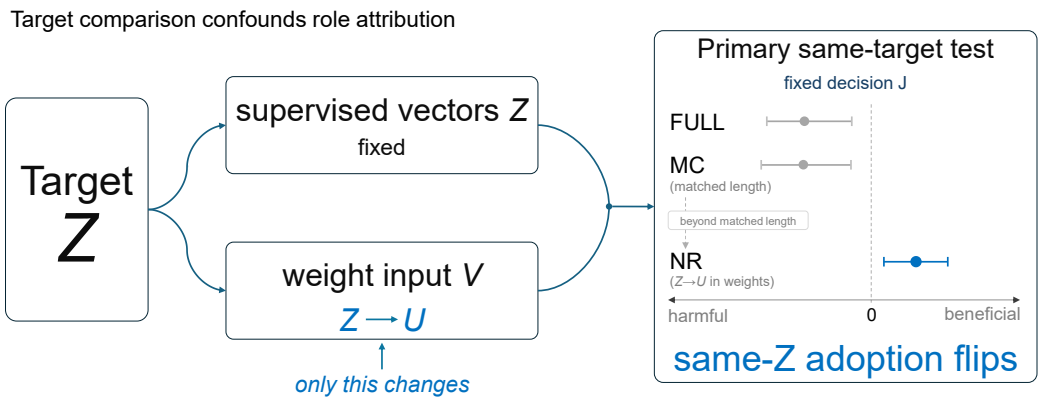


Figure 1: **Separating target roles can reverse adoption.** Ordinary comparison changes supervised vectors and their input to candidate weights. FULL uses Z in both roles; NR (*weight-input replacement*) retains paired and candidate vectors Z but computes weights from reference U. MC retains its own FULL image-parameter update direction while matching NR's actual AdamW step length. FULL and MC are harmful whereas NR is beneficial relative to the same reference supervision. The scoring rule J is shared, with model-specific neural templates (Section 3). Effects are normalized rank $\times 10^3$; bars are participant-bootstrap 95% intervals ($n = 20$). NR retains reference information in weighting. Figures 2–4 quantify the intervention and its target/state boundaries.

vectors on the basis of their ordinary adoption effect: the negative verdict belongs to the tested vector–weighting combination.

Our task ranks 166 concepts by comparing EEG-derived image outputs with templates from other images' EEG outputs. The rule is fixed, but each model supplies its own templates. The intervention keeps Z in paired attraction and candidate vectors and uses U to compute weights from the receiver's current outputs (Figure 1). It identifies a task benefit conditional on reference-informed weighting, not the value of Z alone.

The preferred treatment also depends on the target. Coordinate rescaling favors input replacement; an unblurred-image target favors a reference-informed temperature projection with matched step lengths. This discovery motivates a prospective five-crop test, where preference and positive adoption recur in the same teacher, population and task setting (Figure 3). A factorial test with unmatched step lengths establishes dependence on the state provider (Figure 4). Together these results motivate testing a target's weighting role before attributing its disadvantage to its vectors.

2 Separating Target Roles and Measuring Adoption

For n paired examples with d-dimensional features, let $\mathbf{R}, Z, U, V \in \mathbb{R}^{n\times d}$ collect learner outputs, supervised targets, reference targets and weighting inputs, respectively, in the same batch order. Only learner outputs must have unit length; targets are fixed during an update. With rows $r_i^\top, v_j^\top$ and fixed scale $a > 0$, define similarities $\ell_{ij} = a r_i^\top v_j$ and candidate weights

$$W_{ij}(\mathbf{R}, V) = \frac{1}{2}\left(\frac{e^{\ell_{ij}}}{\sum_{k=1}^{n} e^{\ell_{ik}}} + \frac{e^{\ell_{ij}}}{\sum_{k=1}^{n} e^{\ell_{kj}}}\right). \tag{1}$$

The fractions normalize similarities across candidate targets and learner outputs, respectively; both include the paired entry $i = j$. The symmetric image log-partition P_a averages row/column log-sum-exp terms divided by a. Its ordinary-training (FULL) output gradient is $W(\mathbf{R}, Z)Z/n$ (Appendix C). We replace only the weighting input:

$$g_P(\mathbf{R}, V; Z) = \frac{1}{n} W(\mathbf{R}, V)Z, \qquad V \in \{Z, U\}. \tag{2}$$

FULL sets $V = Z$; NR sets $V = U$, retaining Z in candidate vectors and *paired attraction*, the pull toward the matched target. NR recomputes weights from its own current outputs and stops their gradient during backpropagation; evolving outputs still affect later weights. Image-loss scaling, normalization and network derivatives map g_P to parameter gradients. When $Z = U$, NR matches FULL. Section 7 separately replaces the state supplying weights while retaining the receiver's own network and normalization derivatives.

Let $J(f)$ be the score of the model(s) trained under one condition, with higher values better. For EEG, three seed-specific models are scored jointly within each participant. Write $f_{\lambda,Z}$ for target-Z models with the image partition omitted ($\lambda = 0$) or included ($\lambda = 1$), with architecture, data and training budget matched. Let Z_0 be the reference ($Z_0 = U$ in the main EEG tests), and $f_{\mathrm{NR},Z}$ use NR. Define

$$\begin{aligned} D_\lambda &= J(f_{\lambda,Z}) - J(f_{\lambda,Z_0}), \\ I &= D_1 - D_0, \\ R &= J(f_{\mathrm{NR},Z}) - J(f_{1,Z}), \\ D_{\mathrm{NR}} &= J(f_{\mathrm{NR},Z}) - J(f_{1,Z_0}). \end{aligned} \tag{3}$$

D_λ measures adoption under each loss; I measures its change on partition insertion. R is NR's gain over FULL, while D_{NR} is its gain over reference supervision. The identity $D_{\mathrm{NR}} = D_1 + R = D_0 + I + R$ shows why $R > 0$ can coexist with negative adoption. It compares final scores, not causal shares along different trajectories. In D_{NR}, both learners use the rule $W(\mathbf{R}, U)$, but their realized weights can differ as their outputs evolve. Thus $D_1 < 0 < D_{\mathrm{NR}}$ identifies benefit under reference-informed weighting, not a treatment-independent value of Z.

The matched-length control (MC) rescales its own FULL AdamW image-parameter update to NR's actual step length while retaining its direction. Define $S = J(f_{\mathrm{NR},Z}) - J(f_{\mathrm{MC},Z})$ and, for treatment m, $D_m(Z) = J(f_{m,Z}) - J(f_{1,Z_0})$. The temperature control (TC) projects reference weights onto separately scaled target-similarity distributions for each row and column, then matches NR's actual step length (Section 6). Preference $H(Z) = D_{\mathrm{NR}}(Z) - D_{\mathrm{TC}}(Z) > 0$ favors NR over TC; $D_m(Z) > 0$ instead denotes positive adoption.

3 EEG LEARNING TASK AND EVALUATION

Alljoined (Xu et al., 2025) provides 20 participants, 1,488 fitting concepts and 166 evaluation concepts, each with ten images. Evaluation concepts were exposed during earlier model development; Table 2 records each comparison's history. New interventions are fixed before their new training runs.

Participant-specific multilayer perceptrons (MLPs) map repeat-averaged EEG to independent visual and text branches (Liu et al., 2026), with equally weighted losses and unchanged text supervision. Visual targets use 1,024-dimensional features from a fixed OpenCLIP ViT-H/14 encoder (Cherti et al., 2023). For each participant/seed, reference U unit-normalizes each image's mean target from an earlier baseline run through its selected stopping epoch. Three seeds share initialization, batches, dropout streams and that inherited epoch across treatments. Training and CNN-extension details are in Appendices B and G.1.

For each seed and image position, a concept template is the unit-normalized mean output for its other nine images. The held-out position is excluded from every candidate template. Cosine scores are averaged across seeds *before* ranking 166 candidates. Correct-concept rank r gives $1 - (r-1)/165$, with 1 best, 0 worst, chance 0.5 and average ranks for ties. J averages over concepts and images. Each model supplies both queries and templates: adoption measures cross-image ranking under this fixed rule on exposed concepts. Section 8 separately tests a fixed external decision model.

A condition qualifies when intervals for its mean advantages over chance and a cyclic wrong-label control both exclude zero positively. The control shifts assigned concepts by one place without changing scores or models. Qualification establishes usable performance, not target adoption. Participants are independent units; seeds, concepts and images are combined within participant first. Later temperature/state tests use 19 participants after predictor development on one participant. We retain prescribed participant-bootstrap or paired-participant t intervals and stated multiplicity adjustments; visual tests use paired seeds. An interval excluding zero resolves direction (Appendix B).

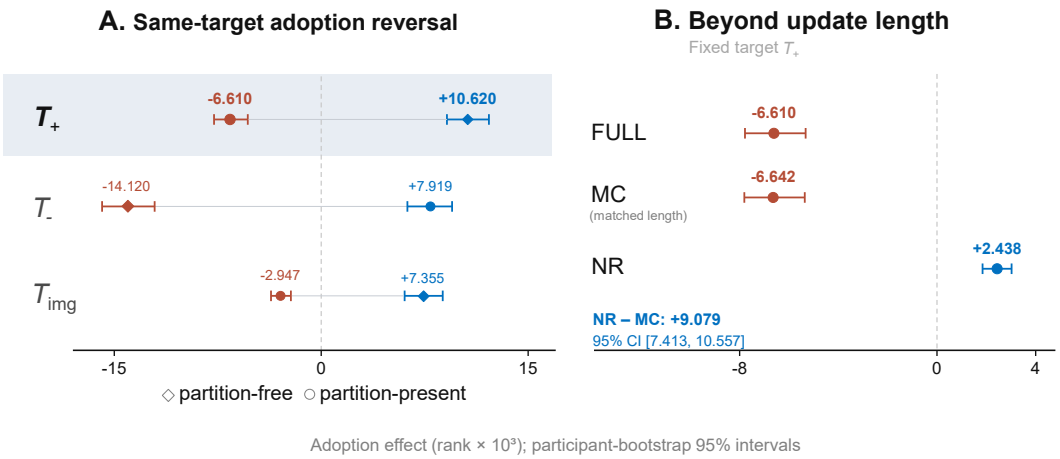


Figure 2: **Same-target reversal and gain beyond matched update length. A:** T_+ changes from beneficial without the log-partition to harmful with it; T_- is the reciprocal companion. T_{img} repeats the direction under an image-based coordinate rule and a disclosed pre-receiving revision of the success criterion. Diamonds/circles denote partition-free/present; blue/rust denotes positive/negative effects. **B:** FULL and MC remain harmful for T_+; NR restores positive adoption. MC preserves its own FULL direction and AdamW moments while matching NR's actual image-parameter step norm. NR–MC is $+9.079$ ($[7.413, 10.557]$). Effects are normalized rank $\times 10^3$, with participant-bootstrap 95% intervals ($n = 20$). This tests weighting-input contribution, not unique mediation or a generally superior loss (Appendix D.1).

4 Adding the Log-Partition Term Reverses Target Adoption

We construct T_+ and T_- by coordinate-rescaling reference U. Using fitting images only, we rank coordinates of a partition-trained reference model's outputs by between-concept relative to within-concept variance. For T_+, the higher-ranked half receives twice the multiplier of the other half; T_- reverses this assignment. A global rescaling matches U's total squared energy, without subsequently normalizing individual vectors. Identical target tensors are fixed before receiving training and used in both loss conditions (Appendix C).

4.1 A Controlled Comparison with the Same Targets

We keep target vectors, unit-output normalization and paired attraction fixed, adding only the log-partition. Within each loss, the target is compared with a U reference trained using that loss:

$$L_{\text{img},\lambda} = \frac{1}{2n} \sum_{i=1}^{n} \|r_i - z_i\|_2^2 + \lambda P_a(\mathbf{R}, Z), \qquad \lambda \in \{0, 1\}. \tag{4}$$

Here z_i is the transpose of row i of Z, and $\|\cdot\|_2$ is the Euclidean norm. Through unit normalization, the squared term has the same parameter gradient as paired attraction. With $\lambda = 1$, the gradient equals that of symmetric cross-entropy divided by a. The partition includes paired entries, so their net contribution may change although the separate attraction coefficient stays fixed. Scale $a = 87.2614$ is fixed by the baseline initialization rule; batches, stopping and text supervision remain paired, without rescaling the total gradient.

The prespecified primary requires all reference/T_+ conditions to qualify and participant-bootstrap 95% intervals supporting $D_0 > 0$, $D_1 < 0$ and $I < 0$. T_- is a prespecified reciprocal companion (Appendix D).

The primary criterion passes: T_+ benefits without the partition and harms with it; $I < 0$ in $20/20$ participants. T_- reverses in the opposite direction (Figure 2A). The full partition is therefore sufficient to reverse adoption under this matched protocol. Section 5 next retains candidate weighting and replaces only the target input used to compute its coefficients.

4.2 Recurrence with coordinates selected from image features

T_{img} applies the same ranking rule to fixed image-encoder features, retaining the 2:1 ratio and total-energy rule. After reference results were known, a disclosed criterion revision separated reversal from quantitative prediction; targets and forecasts were fixed before receiving training (Appendix D.2). Adoption is positive without the partition in $20/20$ participants, negative with it in $19/20$, and $I < 0$ in $20/20$ (Figure 2A). The reversal recurs, but partition-present forecasts do not beat zero change at the participant-mean level and have higher concept-level error.

5 Intervene: change adoption through the target input

Can changing only the target input to candidate weights change adoption? NR retains Z in paired attraction and candidate vectors, substituting unmodified Z_0 (U here) inside weighting:

$$g_P^{\text{NR}}(\mathbf{R}; Z, Z_0) = \frac{1}{n} W(\mathbf{R}, Z_0) Z. \tag{5}$$

NR implements this prescribed backward rule by recomputing weights from its current outputs and holding them constant in backpropagation, with normalization and network derivatives retained. At fixed $\mathbf{R}$, the FULL-minus-NR output-gradient difference is $[W(\mathbf{R}, Z) - W(\mathbf{R}, Z_0)]Z/n$.

MC first computes its complete FULL AdamW image-parameter update, including its own gradients, moments and weight decay, then rescales it to the paired NR step's Euclidean length. It preserves its own direction and moments; the text branch stays paired and unchanged (Appendix D.1). FULL, NR and MC share the partition-present U reference. FULL and MC remain harmful for T_+, while NR has positive adoption in $19/20$ participants (Figure 2B). The NR–MC advantage excludes the tested update-length change alone as an explanation of the gain.

Size and meaning of the adoption effect. Mean normalized ranks are 0.605576 for U, 0.598966 for FULL, 0.598935 for MC and 0.608014 for NR. Adoption $D_{\text{NR}} = +0.002438$ has participant-bootstrap 95% interval $[+0.001847, +0.003024]$. Multiplying by 165 translates this to approximately 0.40 correct-concept rank positions among 166 candidates; NR–MC translates to 1.50 positions. These are rank translations, not top-1 gains. The small but resolved adoption effect establishes the attribution counterexample, not a cost-adjusted deployment recommendation.

The gradient contribution changed by the intervention. At fixed unit outputs $\mathbf{R}$, let $\Delta Z = Z - U$, $W_0 = W(\mathbf{R}, U)$ and $W_Z = W(\mathbf{R}, Z)$. The partition-present image-loss output-gradient change from U to Z is exactly

$$\Delta g_{\text{img}}(\mathbf{R}) = \frac{1}{n} \left[\underbrace{-\Delta Z}_{\text{attraction}} + \underbrace{W_0 \Delta Z}_{\text{candidate vectors}} + \underbrace{(W_Z - W_0) Z}_{\text{candidate weights}} \right]. \tag{6}$$

NR removes the last difference at the same outputs. During training, normalization, loss scaling and the optimizer map these fields into updates, and receivers follow different trajectories. This identity alone does not predict adoption; the after-training comparison supplies that evidence (Appendices E.3, C).

6 Weighting preference depends on the target

6.1 Reference-informed temperature projection as a comparator

TC fits a separate nonnegative inverse temperature in each row and column of target similarities, minimizing Kullback–Leibler (KL) divergence from reference weights to the rescaled target distribution at the receiver's current outputs. KL divergence measures distribution mismatch; larger scales concentrate weights on higher similarities, while zero gives uniform weights. The fitted distributions are averaged as in Equation 1, recomputed each step and held constant in backpropagation. TC retains Z in attraction and candidate vectors, then matches the paired NR run's actual AdamW

image-parameter step length. Reported TC effects therefore combine a reference-informed projection with step-length matching (Appendix C).

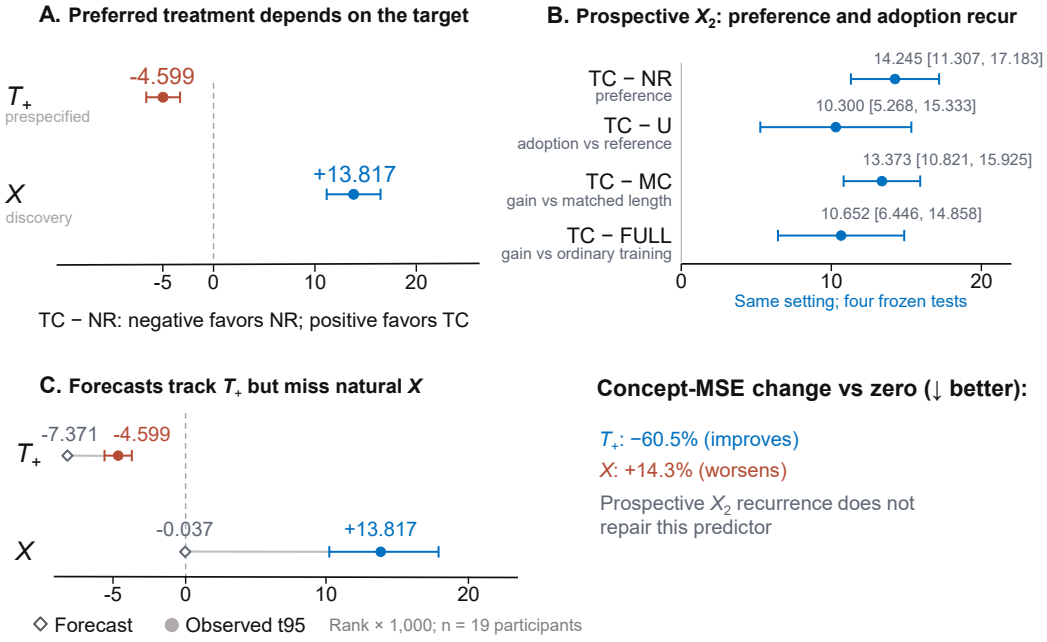


Figure 3: **Target-dependent preference and prospective recurrence, with a prediction boundary.** TC projects reference weights onto row/column temperature families and matches NR's actual AdamW step length. **A:** TC–NR equals $-H$; negative favors NR. The prespecified T_+ test favors NR; TC preference for natural X was discovered after the TC runs. **B:** For prespecified five-crop X_2, ordered comparisons favor TC over NR, U, MC and FULL in the same teacher, population and task setting. **C:** Pre-TC-run forecasts (diamonds; no calibrated intervals) miss the sign for X. Concept-MSE improves by 60.5% over zero change for T_+ but worsens by 14.3% for X. Observed bars are paired-participant 95% t intervals ($n = 19$). X_2 does not repair this predictor; its FULL adoption remains unresolved (Appendices E, E.2).

6.2 The original natural target favors TC

X is each unblurred image's unit feature from the same fixed encoder. Two comparisons on 19 participants were fixed before new TC runs: $H(T_+) > 0$, then $H(T_+) - H(X) > 0$. Both pass. The unpredicted $H(X) < 0$ shows that X instead favors TC (Figure 3A). TC adoption is $+0.008648$ (participant $t95$ $[0.003530, 0.013766]$; $15/19$ positive), and TC exceeds FULL and MC. These latter gains were discovered after the TC runs, with nominal, unadjusted intervals; FULL and NR adoption remain unresolved (Table 4).

6.3 Preference and adoption recur on a prespecified new target

X_2 unit-normalizes the mean of five unit ViT-H features, from the four corner crops and center crop of each fitting image. Targets and test order are fixed before new FULL/NR/MC/TC runs on the same 19 participants and three paired seeds, retaining the evaluation, U reference and budget (Appendix E.2).

All four ordered participant $t95$ tests favor TC over NR, U, MC and FULL; all conditions pass chance/wrong-label qualification (Figure 3B). TC adoption is $+0.010300$, approximately 1.70 rank positions among 166 candidates. FULL and NR adoption remain unresolved. This is prospective recurrence of an empirically motivated treatment choice within the same teacher, population and task setting.

Prediction boundary. The T_+/X comparisons were informed by a first-order predictor through the full U-reference AdamW trajectory (Maclaurin et al., 2015; Franceschi et al., 2017). It extrapolates the local response to the full target change and scores predicted outputs with the same template rule. Predicted step-length response and the measured NR lengths used by actual TC training are distinct information sources (Appendix E.3). The predictor misses $H(X) < 0$: concept-MSE improves over zero by 60.5% for T_+ but worsens by 14.3% for X, with the respective error direction in $19/19$ participants. Positive natural-target adoption and the X_2 result do not resolve this quantitative failure (Figure 3C).

7 THE TARGET-INPUT EFFECT DEPENDS ON THE STATE PROVIDER

Holding $Z = X_2$ in attraction and candidate vectors, we cross the two inputs to the original W in Equation 1. A (FULL) and B (NR) use each receiver's current outputs with weighting inputs Z and U; C and D use a synchronized U-reference learner's outputs with those same inputs. The reference is paired by initialization, batch, dropout and step; each receiver retains its own trajectory and normalization derivatives.

The design followed the observed X_2 A/B outcomes; C/D were fixed before their new runs. A/B are reused after exact implementation checks, and the reference and all four receiving conditions qualify. Let A–D also denote their task scores. With no prespecified direction, the primary test compares the weighting-input effects under reference and receiving outputs:

$$\Gamma = (D - C) - (B - A) = (A - B) - (C - D). \tag{7}$$

The participant-mean interaction is positive (Figure 4B): replacing Z by U has a larger, resolved benefit under reference outputs; the receiving-state effect remains unresolved. C and D both remain below U for adoption. All four conditions use native AdamW without step-length matching, leaving magnitude and optimizer effects as possible mechanisms (Appendix F).

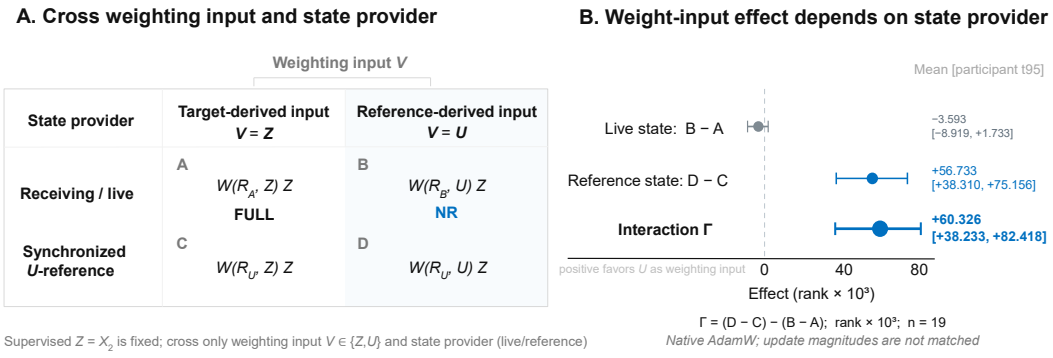


Figure 4: **The weighting-input effect depends on the state provider. A:** Supervised $Z = X_2$ stays fixed. A/B use receiving outputs and C/D synchronized U-reference outputs, each with weighting inputs Z/U. Each receiver retains its own trajectory. **B:** Positive effects favor U as weighting input: $B - A = -3.593$ ($[-8.919, +1.733]$) is unresolved; $D - C = +56.733$ ($[+38.310, +75.156]$) and $\Gamma = +60.326$ ($[+38.233, +82.418]$) are positive. $\Gamma > 0$ in 18/19 participants. Units are normalized rank $\times 10^3$; bars are paired-participant $t95$. C/D were specified after A/B outcomes, without a directional forecast. Both have negative adoption. Native AdamW step lengths are unmatched (Appendix F).

A separate TC companion replaces receiving outputs by synchronized U-reference outputs when fitting temperature weights, retaining X and the same NR step-length control. Adoption changes from positive to negative; all 15 previously positive participants cross below reference supervision. Only the reference-output arm is new. This matched-length TC experiment is distinct from the original-W factorial (Appendix F).

Table 1: **Intervention gain does not imply adoption.** $S = \mathrm{NR} - \mathrm{MC}$, $R = \mathrm{NR} - \mathrm{FULL}$ and $D_{\mathrm{NR}} = \mathrm{NR} - \mathrm{reference}$. Prototype P is a unit-normalized class mean; its reference is X, versus U elsewhere. Rows retain their own metric, interval method and independent unit and are not pooled. Bold estimates have positive 95% intervals; intervals containing zero do not establish equivalence.

	Intervention gain		**Target adoption**	**Scope / interpretation**
Condition / unit	S NR–MC	R NR–FULL	D_{NR} NR–reference	
I *Same EEG task: targets and encoders*				
Neural-template rank $\times 10^3$; participant-bootstrap 95% intervals				
T_{img}/MLP 20 participants	$\mathbf{+2.929}$ $[+2.108, +3.728]$	$\mathbf{+2.896}$ $[+2.075, +3.691]$	-0.051 $[-0.663, +0.583]$	New coordinate rule; adoption unresolved
T_{img}/CNN 20 participants	$\mathbf{+2.262}$ $[+1.357, +3.252]$	$\mathbf{+2.680}$ $[+1.709, +3.749]$	$\mathbf{+1.788}$ $[+0.598, +3.185]$	CNN extension; adoption positive; capacity not matched
Natural X/MLP 20 participants	-0.013 $[-2.922, +2.803]$	-3.373 $[-8.137, +1.113]$	-5.382 $[-11.328, +0.011]$	Natural target; NR gains and adoption unresolved
II *New EEG participants*				
Neural-template rank $\times 10^3$; participant 95% t intervals; shared stimuli				
THINGS-EEG2 6 participants	$\mathbf{+1.249}$ $[+0.453, +2.045]$	$\mathbf{+1.275}$ $[+0.492, +2.058]$	-0.412 $[-0.998, +0.174]$	New participants; intervention gains recur; adoption unresolved
III *Fixed external decision model*				
Top-1 accuracy in percentage points; participant-bootstrap 95% intervals				
Prototype P/DINO 20 participants	$\mathbf{+0.093}$ $[+0.030, +0.154]$	$\mathbf{+0.096}$ $[+0.027, +0.163]$	$+0.009$ $[-0.108, +0.123]$	Fixed DINO interface; top-1 adoption unresolved; rank adoption negative
IV *Visual-only classification*				
Accuracy in percentage points; five paired seeds; paired-seed 95% t intervals				
CIFAR-100 5 paired seeds	-0.130 $[-0.302, +0.042]$	$+0.004$ $[-0.326, +0.334]$	-0.094 $[-0.474, +0.286]$	Qualified classifier; intervention gains and adoption unresolved

8 Intervention gains and target adoption across settings

Table 1 separates intervention gains S (NR–MC) and R (NR–FULL) from adoption D_{NR} (NR–reference) as targets, encoders, participants and decision models change. Transfer means recurrence of the specified effect, not automatic transfer of adoption.

For T_{img}, NR improves on both controls with the MLP but adoption is unresolved; the CNN has positive adoption, with architecture and capacity changing together (Appendix G.1). The natural-X row retains the earlier 20-participant NR test, distinct from the later 19-participant TC comparison.

On six new THINGS-EEG2 participants, $S, R > 0$ in $6/6$ but adoption is unresolved. Participants are disjoint from Alljoined; stimuli and prior information for target design are shared. A separate NR–TC test establishes mean-score equivalence at the prespecified precision: its 90% t interval lies within ± 0.000606061 normalized rank (Appendix G.2).

The fixed external decision test replaces X by class prototype P, holding the DINOv2 mapping (Oquab et al., 2024) and 166-concept gallery fixed. Against X supervision, top-1 gains are positive but adoption unresolved; reference/NR accuracy is about 1.4% (chance 0.6%), with ties sharing credit. The rank companion has negative adoption despite positive intervention gains (Appendix G.3; Table 6).

Visual classification with different training and scoring rules. The supervised CIFAR-100 experiment uses five paired seeds as independent units; test images do not supply independent training replicates. All arms pass prespecified development-accuracy gates, but gains and adoption remain unresolved (Table 1). A subsequent fixed-prototype readout of the same checkpoints also leaves the change in S unresolved (Appendix G.4).

A separately frozen follow-up removes training cross-entropy (CE), which fits predictions to class labels, retaining the fixed prototypes and matching MC to the new NR step lengths. Labels still construct prototypes and evaluate the task. Reference competence passes, but $S = -0.014$ percentage points (pp; paired-seed $t95$ $[-0.362, +0.334]$); the direct change in S versus fixed-consumer CE is also unresolved. NR–FULL is positive in $5/5$ seeds, $+0.140$ pp $[-0.039, +0.319]$, while adoption spans zero. This direction signal establishes neither gain beyond MC, equivalence nor a complete magnitude explanation; positive visual-only transfer remains open (Appendix G.4).

9 Scope and limitations

Evaluation and prediction. The main task uses model-specific templates and EEG averaged over repeated presentations, on concepts exposed during earlier development. Generalization to unseen concepts, single-trial EEG and a fixed external decision model requires separate evidence. The DINO test supplies one fixed-interface comparison, with low absolute accuracy and unresolved top-1 adoption. The original X adoption result is a discovery; X_2 adds one prospective construction in the same setting. Failed natural-target forecasts leave prediction of adoption on new targets open.

Training mechanisms. The partition reversal is established under the specified scale, batch and matched training conditions. NR exceeds the tested control of actual update lengths, while the original-W interaction uses unmatched lengths. Fixing supervised vectors still allows the learners to follow different training trajectories. NR and TC retain reference information from U in weighting; their effects do not isolate the usefulness of Z without that reference. Output-provider replacement tests the specified reference learner and, for TC, the chosen NR step-length control. These interventions do not isolate one unique training mechanism or establish that receiving-state feedback is necessary across settings.

Transfer and statistical scope. The study does not test every target with every encoder, and encoder changes are not capacity-matched. New EEG participants share stimuli and prior source information with the original setting. The EEG scope tests require population advantages over chance and wrong-label controls; they do not inherit the earlier protocol's full controls using independent EEG repeats (Appendix B). The visual tests leave positive transfer unresolved, without establishing that the effect is specific to EEG. Reported intervals describe individual contrasts. Ordered testing governs claims, not simultaneous interval coverage; family adjustments are stated where applied. Further boundaries are recorded in Appendices C and H.

10 Conclusion

In the sample-coupled objective studied here, an ordinary target comparison changes both supervised vectors and their input to candidate weighting. Our controlled EEG experiments supply a counterexample to interpreting a negative outcome as a reason to discard the vectors: keeping them fixed and replacing only their weighting input reverses adoption against the same reference, beyond the tested control of actual AdamW step lengths. This is a small but resolved benefit on cross-image neural-template ranking, with each model supplying its own templates and reference information retained in weighting.

The target-dependent treatment preference, prospective five-crop recurrence and state-provider interaction delimit this result. They support an intervention-based interpretation of target comparisons: a disadvantage under ordinary weighting can motivate changing the target's weighting role while retaining its supervised vectors. Choosing a complete training configuration and attributing its performance to vector content are distinct scientific decisions; the latter requires evidence that separates the two roles at the intended task and consumer.

AI USE STATEMENT

Generative AI tools assisted hypothesis refinement, experimental design, mathematical derivation checks, code implementation and validation, analysis and interpretation of results, literature search, preliminary figure drafts, manuscript drafting and language editing, and internal manuscript review. The authors directed the research, made all key scientific decisions, reviewed and verified AI-assisted outputs, and take full responsibility for the methods, results, and final manuscript.

REFERENCES

Xinyang Chen, Sinan Wang, Jianmin Wang, and Mingsheng Long. Representation subspace distance for domain adaptation regression. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 1749–1759. PMLR, 2021. URL https://proceedings.mlr.press/v139/chen21u.html.

Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. DarkRank: Accelerating deep metric learning via cross sample similarities transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. URL https://ojs.aaai.org/index.php/AAAI/article/view/11783.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2818–2829, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Cherti_Reproducible_Scaling_Laws_for_Contrastive_Language-Image_Learning_CVPR_2023_paper.html.

Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4794–4802, 2019. URL https://arxiv.org/abs/1910.01348.

Julien Denize, Jaonary Rabarisoa, Astrid Orcesi, Romain Hérault, and Stéphane Canu. Similarity contrastive estimation for self-supervised soft contrastive learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2706–2716, 2023. URL https://openaccess.thecvf.com/content/WACV2023/html/Denize_Similarity_Contrastive_Estimation_for_Self-Supervised_Soft_Contrastive_Learning_WACV_2023_paper.html.

Luca Franceschi, Michele Donini, Paolo Frasconi, and Massimiliano Pontil. Forward and reverse gradient-based hyperparameter optimization. In *International Conference on Machine Learning*, pp. 1165–1173. PMLR, 2017. URL https://proceedings.mlr.press/v70/franceschi17a.html.

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.

Sarah E. Harvey, David Lipshutz, and Alex H. Williams. What representational similarity measures imply about decodable information. In *Proceedings of UniReps: the Second Edition of the Workshop on Unifying Representations in Neural Models*, volume 285, pp. 140–151. PMLR, 2024. URL https://proceedings.mlr.press/v285/harvey24a.html.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.

Martin N. Hebart, Adam H. Dickter, Alexis Kidder, Wan Y. Kwok, Anna Corriveau, Caitlin Van Wicklin, and Chris I. Baker. THINGS: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PLOS ONE*, 14(10):e0223792, 2019. doi: 10.1371/journal.pone.0223792.

Zizheng Huang, Haoxing Chen, Ziqi Wen, Chao Zhang, Huaxiong Li, Bo Wang, and Chunlin Chen. Model-aware contrastive learning: Towards escaping the dilemmas. In *International Conference on Machine Learning*, pp. 13774–13790, 2023. URL https://proceedings.mlr.press/v202/huang23c.html.

JaeSeong Kim and Suan Lee. Same benchmark, same subspace: Task-selective convergence in LLM representations. In *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337, pp. 3101–3116. PMLR, 2026. URL https://proceedings.mlr.press/v337/kim26g.html.

Yeongmin Kim, Donghyeok Shin, Mina Kang, Byeonghu Na, and Il-Chul Moon. Distillation of large language models via concrete score matching. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/1bfc9f74afa91b9b8add5a97a97001a1-Abstract-Conference.html.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Yazhe Li, Jorg Bornschein, and Marcus Hutter. Evaluating representations with readout model switching. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Fsd-6ax4T1m.

Dongjun Liu, Weichen Dai, Jingsheng Qian, Honggang Liu, Hangjie Yi, and Wanzeng Kong. Linguistic priors for visual decoupling: Towards symmetric vision-brain alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7869–7878, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Liu_Linguistic_Priors_for_Visual_Decoupling_Towards_Symmetric_Vision-Brain_Alignment_CVPR_2026_paper.html.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1711.05101.

Dougal Maclaurin, David Duvenaud, and Ryan P. Adams. Gradient-based hyperparameter optimization through reversible learning. In *International Conference on Machine Learning*, pp. 2113–2122. PMLR, 2015. URL https://proceedings.mlr.press/v37/maclaurin15.html.

Shunsuke Onoo, Yoshihiro Nagano, and Yukiyasu Kamitani. Readout representation: Redefining neural codes by input recovery. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/5f326d2e040a20d52438873f5d727e58-Abstract-Conference.html.

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*, 2024. URL https://arxiv.org/abs/2304.07193.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3967–3976, 2019. URL https://arxiv.org/abs/1904.05068.

Chengyao Qian, Trung Le, and Mehrtash Harandi. A good teacher adapts their knowledge for distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1239–1248, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Qian_A_Good_Teacher_Adapts_Their_Knowledge_for_Distillation_ICCV_2025_paper.html.

Yuandong Tian. Understanding deep contrastive learning via coordinate-wise optimization. In *Advances in Neural Information Processing Systems*, volume 35, pp. 19511–19522, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/7b5c9cc08960df40615c1d858961eb8b-Abstract-Conference.html.

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

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

Jonathan Xu, Ugo Bruzadin Nunes, Wangshu Jiang, Samuel Ryther, Jordan Pringle, Paul S. Scotti, Arnaud Delorme, and Reese Kneeland. Alljoined-1.6m: A million-trial EEG-image dataset for evaluating affordable brain-computer interfaces. *arXiv preprint arXiv:2508.18571*, 2025. URL https://arxiv.org/abs/2508.18571.

Han-Jia Ye, Su Lu, and De-Chuan Zhan. Distilling cross-task knowledge via relationship matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12396–12405, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Ye_Distilling_Cross-Task_Knowledge_via_Relationship_Matching_CVPR_2020_paper.html.

Chun-Hsiao Yeh, Cheng-Yao Hong, Yen-Chi Hsu, Tyng-Luh Liu, Yubei Chen, and Yann LeCun. Decoupled contrastive learning. In *European Conference on Computer Vision (ECCV)*, pp. 668–684, 2022. URL https://www.ecva.net/papers/eccv_2022/papers_ECCV/html/5549_ECCV_2022_paper.php.

Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11953–11962, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Zhao_Decoupled_Knowledge_Distillation_CVPR_2022_paper.html.

Mingkai Zheng, Shan You, Fei Wang, Chen Qian, Changshui Zhang, Xiaogang Wang, and Chang Xu. ReSSL: Relational self-supervised learning with weak augmentation. In *Advances in Neural Information Processing Systems*, volume 34, pp. 2543–2555, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/14c4f36143b4b09cbc320d7c95a50ee7-Abstract.html.

A EVIDENCE MAP AND SELECTION HISTORY

The appendix answers four questions: what is held fixed, which comparisons support the reported adoption effects, where prediction fails, and how far those effects transfer. Table 2 links these questions to the main paper. Training and evaluation are specified once in Appendices B–C; result sections report departures from that shared protocol.

Table 2: Main-paper evidence and its exposure history. Prospective status applies to the indicated new outcomes; the source concepts and prior experiments were already exposed. EEG units are participants, with three nested seeds; visual units are paired training seeds.

Main anchor	Evidence role and exposure	Details
Figure 2: component	$n = 20$; prescribed T_+ test and reciprocal companion. Image-selected recurrence followed a disclosed pre-fit criterion amendment.	D
Figure 2: NR/MC	$n = 20$; fixed-target receiving interventions specified after FULL was known; natural X extrapolation unresolved.	D.1
Figure 3: T_+/X	$n = 19$; two ordered TC contrasts prospective. Negative $H(X)$ and positive TC adoption discovered afterward.	E
Figure 3: X_2	$n = 19$; five-crop construction and four tests frozen after the X discovery, before new receiving fits.	E.2
Figure 4: state input	$n = 19$; A/B outcomes known; new C/D cells prospective, without a directional forecast.	F
Table 1: EEG scope	$n = 20$ for architecture/consumer changes; six new THINGS participants share stimuli and source-informed target design.	G
Section 8: visual scope	Five seeds; later readout/no-CE diagnostics informed by earlier benchmark outcomes, not independent replications.	G.4

The sequence developed adaptively; freezing a new comparison does not make the project outcome-naive. Throughout, $S = \mathrm{NR} - \mathrm{MC}$, $R = \mathrm{NR} - \mathrm{FULL}$ and $D_{\mathrm{NR}} = \mathrm{NR} - \mathrm{reference}$ retain their distinct meanings. Intervals spanning zero are unresolved, including when all units share a sign. Ordered testing governs claims, not simultaneous interval coverage. The sole mean-equivalence claim has an explicit margin (Appendix G.2).

B SHARED EXPERIMENTAL AND STATISTICAL PROTOCOL

B.1 DATA, PREPROCESSING AND SOURCE EXPOSURE

Alljoined (Xu et al., 2025) provides 20 participants, 1,488 fitting concepts and 166 selection concepts, with ten images per concept (14,880/1,660 images). EEG repeats are averaged per image. Selection concepts were used in source-model development; this is an exposed-concept task.

The raw-EDF adapter takes 256 Hz, retains a fixed 32-channel montage, epochs $[-0.2, 1.0]$ s, baseline-corrects through zero, converts volts to microvolts and resamples to 200 Hz. Samples `40:240` give the nominal $[0, 1)$ s window. Participant/session-specific multivariate noise normalization is fitted using source trials only and applied in 32-channel space before selecting P7, P5, P3, P1, Pz, P2, P4, P6, P8, PO7, PO3, POz, PO4, PO8, O1, Oz, O2. Inputs are 17×200; no receiving outcome selects normalization. THINGS-EEG2 (Gifford et al., 2022; Hebart et al., 2019) retains its four-repeat support and 250-point input: acquisition, preprocessing and input length change with population, while stimuli are shared.

B.2 MODELS, STOPPING AND PAIRED TRAINING

The BVSA `EEGProject` MLP (Liu et al., 2026) has separate image and text branches: flattened EEG maps linearly to 1,024 dimensions, followed by a residual GELU–linear–dropout block and LayerNorm (dropout 0.3). Image outputs are unit-normalized. Branch independence lets the image intervention leave text-parameter updates unchanged.

Each participant/seed (2025–2027) inherits a source guide using BVSA's nine-variant dynamic target rule and common text targets. Its stopping epoch minimizes unblurred-image selection loss over at

most 50 epochs, with earliest tie-breaking. References and receiving arms inherit this epoch; they do not select their own best downstream score. AdamW (Loshchilov & Hutter, 2019) uses learning rate and decay 10^{-4}, batch 1,024 with a 544-image tail, and deterministic float32 execution. Arms pair initialization, batches, dropout, text targets and epoch count. Loss weights are one half for image and one half for unchanged symmetric text contrastive loss. The image scale is fixed below; the text scale remains trainable. Receiving scores do not retune these choices. CNN and visual departures are specified in Appendix G.

B.3 Decision rule and inference

Let q_{sci} be the image-head output for seed s, concept c and image position i. For each fold i, construct all concept templates as

$$t_{sc}^{(-i)} = \mathrm{unit}\left(\frac{1}{9}\sum_{j\neq i}\mathrm{unit}(q_{scj})\right), \qquad C_{sicd} = \mathrm{unit}(q_{sci})^{\top} t_{sd}^{(-i)}.$$

Position i is excluded from every candidate template, not only the correct one. No teacher image/text features enter these scores. Average cosine matrices across seeds before ranking: $\bar{C}_{icd} = \frac{1}{3}\sum_s C_{sicd}$. With $N_>$ incorrect scores strictly above the correct score and $N_=$ tied,

$$r = 1 + N_> + N_=/2, \qquad \rho = 1 - (r-1)/165, \qquad J = \mathrm{mean}_{c,i}\,\rho_{ci}.$$

Chance is 0.5. Images and concepts have equal weight; neither per-seed rank averaging nor cross-seed coordinate averaging implements this estimand. Each model supplies its own queries and templates, so the fixed rule tests cross-image neural ranking rather than a fixed external consumer.

Each participant supplies one paired effect. Early component, NR/MC, CNN and DINO comparisons use 20,000 paired participant-bootstrap resamples and percentile 95% intervals. Campaign-specific frozen streams explain small endpoint differences for reused means. Later TC/X_2/state tests use 19 participants after predictor development on one participant, with $t95$ intervals (df 18). THINGS uses six participants (df 5); visual experiments use five paired seeds (df 4), conditional on the benchmark. Seeds, images and trials do not increase the participant sample size.

A required EEG cell qualifies if population intervals for rank above chance and correct-minus-cyclic-wrong assignment both exclude zero positively. The wrong control shifts concept assignments by one without changing scores or models. All planned participants remain; qualification does not require positive adoption or every participant to have a positive individual effect. Family corrections and ordered tests are stated with each result.

Qualification boundary. Earlier objective–readout work used a separate 200-concept gallery, independent-repeat margins/ranks and eight reverse-control families. That stronger qualification is not inherited by the current 166-concept task. Those historical outcomes are not pooled with the present effects or used to increase their sample size.

C Target construction and intervention implementation

C.1 Targets fixed before receiving training

Reference U. For a guide stopping at epoch E, let V_e contain the unit targets actually used for all fitting images at epoch e. Row-normalize $M = E^{-1}\sum_{e=1}^{E} V_e$ to obtain U. Each participant/seed uses its own selected-prefix history, not a terminal marginal or another guide. This normalization changes the first moment, so raw-mean/U comparisons do not isolate variance alone.

Diagonal targets. For fitting outputs q_{cij} of the partition-trained U model, rank coordinates using

$$F_j = \frac{\mathrm{Var}_c(\bar{q}_{cj})}{\mathbb{E}_{c,i}(q_{cij} - \bar{q}_{cj})^2 + 10^{-12}}, \qquad D_{\pm} = \mathrm{diag}\Big(e^{\pm(\log 2)h_j/2}\Big), \qquad Z_{\pm} = UD_{\pm}\frac{\|U\|_F}{\|UD_{\pm}\|_F}.$$

Stable sorting breaks ties by coordinate index; $h_j = +1$ for the top 512 coordinates and -1 for the other 512. The resulting T_+, T_- have a 2:1 multiplier ratio and match total squared energy,

but not individual row or gradient norms. Store float32 targets without row renormalization. T_{img} applies the same Fisher ranking to unit, unblurred fitting-image teacher features, then applies the same diagonal/energy rule to each U. Its direction uses no receiving outcome; its pre-fit criterion amendment is reported in Appendix D.2.

Natural targets. X is the unit unblurred OpenCLIP ViT-H/14 feature. For X_2, bicubically resize RGB images to shorter side 256, take four corner and one center 224×224 crops, unit-normalize their embeddings, average, then normalize again. Only fitting images enter construction. The recipe was chosen after the X discovery and frozen before X_2 fits. Prototype P is the unit-normalized mean of a fitting concept's ten unit X vectors, shared by its images. Its adoption reference is X; other main EEG interventions use U.

C.2 Matched loss and prescribed fields

For unit student rows r_i, fixed target rows z_i and $s_{ij} = r_i^\top z_j$,

$$P_a(\mathbf{R}, Z) = \frac{1}{2na}\left[\sum_i \log \sum_j e^{a s_{ij}} + \sum_j \log \sum_i e^{a s_{ij}}\right].$$

The image loss is $L_\lambda = (2n)^{-1}\sum_i \|r_i - z_i\|^2 + \lambda P_a(\mathbf{R}, Z)$. Through unit normalization, the squared term has the same parameter gradient as attraction $-n^{-1}\sum_i r_i^\top z_i$. For $\lambda = 1$, this equals the symmetric cross-entropy gradient divided by a. Both row/column softmaxes include the paired entry, so fixed explicit attraction does not fix the net positive-pair gradient. The image scale is $a = \log(1 + 1/0.07)\sqrt{1024} = 87.2613979330$, an initialization-based rule, not an exact match to all historical raw logits. Partition insertion does not match gradient norms.

FULL and NR. FULL uses its own current weights $W(\mathbf{R}_t, Z)$ from Equation 1. NR substitutes U only in that input:

$$g_{\text{image}}^{\text{NR}} = \frac{\mathbf{R} - Z + W(\mathbf{R}, U)Z}{n}.$$

The radial $\mathbf{R}$ term vanishes through the unit-output Jacobian. Weights are recomputed at every step and held constant during backpropagation; normalization and network derivatives remain. The receiver's evolving outputs still change later weights. This is a prescribed gradient field, without a claim of a conservative replacement loss. At $Z = U$, NR reproduces FULL, enabling reference reuse. Equation 6 identifies its same-state change; it does not decompose different finite trajectories into causal shares or predict adoption.

MC: match actual parameter displacement. MC computes a FULL AdamW proposal using its own moments, preconditioning and decoupled decay, then applies

$$\Delta\theta_{\text{MC}} = \frac{\|\Delta\theta_{\text{NR}}\|_2}{\|\Delta\theta_{\text{FULL}}^{\text{MC}}\|_2}\Delta\theta_{\text{FULL}}^{\text{MC}}.$$

The donor is the paired NR step, not a fitted schedule. MC retains its own moments and image-update direction. Text/scale updates remain paired. With independent neural branches, matching image norm also matches total norm, but does not preserve the concatenated image-plus-text direction. A zero proposal with nonzero donor is an error, not an epsilon fallback. Traces check float32 rounding. BatchNorm buffers are outside the match; visual experiments match backbone and both heads together.

TC: reference-informed temperature projection. For each row and column separately, with reference probabilities p_0 and target logits ℓ evaluated at the receiver's current outputs, fit

$$\beta^* = \arg\min_{\beta \geq 0} \mathrm{KL}(p_0 \,\|\, \mathrm{softmax}(\beta\ell)).$$

Use zero for a boundary solution; otherwise bracket and bisect, checking the probability-moment residual. Failure to resolve a finite bracket is an error. Logit centering/scaling conditions the solve without changing the family. Average projected row/column distributions to form $\Pi(\mathbf{R}, Z; U)$. TC backpropagates $(\mathbf{R} - Z + \Pi Z)/n$ with stopped weights and then matches its actual image-parameter displacement to NR. Scalars change concentration but cannot arbitrarily reorder logits. TC is a combined reference-informed projection and step-length treatment.

D FIGURE 2: DOES THE SAME TARGET CHANGE ADOPTION?

Table 3 gives the complete component and NR/MC contrasts. The tests hold the target tensors and paired training fixed; only the stated loss component or weight input changes.

Table 3: **Support for Figure 2.** Panel (a): normalized-rank effects $\times 10^3$, with frozen participant-bootstrap 95% intervals. Panel (b): unscaled normalized rank for the separate NR/MC experiments. All use 20 participants and three nested seeds. Bootstrap streams differ across campaigns, so reused FULL means have slightly different interval endpoints. Natural X in panel (b) is the original 20-participant NR test, not the later 19-participant TC discovery.

(a) Matched component insertion: T_+ primary; T_- companion; T_{img} new direction

Effect	T_+	T_-	T_{img}
D_0	$+10.620\ [9.114, 12.157]$	$-14.120\ [-15.884, -12.075]$	$+7.355\ [6.046, 8.808]$
D_1	$-6.610\ [-7.768, -5.323]$	$+7.919\ [6.247, 9.487]$	$-2.947\ [-3.639, -2.198]$
$I = D_1 - D_0$	$-17.230\ [-19.261, -14.980]$	$+22.039\ [18.613, 25.001]$	$-10.302\ [-11.905, -8.686]$

(b) Direct target-input intervention and its natural-target boundary

Contrast	T_+	Natural $X - U$
D_{FULL}	$-0.006610\ [-0.007783, -0.005325]$	$-0.002009\ [-0.004452, +0.000466]$
D_{NR}	$+0.002438\ [+0.001847, +0.003024]$	$-0.005382\ [-0.011328, +0.000011]$
D_{MC}	$-0.006642\ [-0.007809, -0.005360]$	$-0.005369\ [-0.009175, -0.001669]$
$R = \text{NR} - \text{FULL}$	$+0.009048\ [+0.007364, +0.010538]$	$-0.003373\ [-0.008137, +0.001113]$
$S = \text{NR} - \text{MC}$	$+0.009079\ [+0.007413, +0.010557]$	$-0.000013\ [-0.002922, +0.002803]$

D.1 MATCHED COMPONENT AND DIRECT INTERVENTION

The sole component primary was specified before new reference training: $D_0(T_+) > 0$, $D_1(T_+) < 0$ and $I(T_+) < 0$, with two-sided participant-bootstrap 95% intervals and qualified reference/T_+ cells. Partition-free fits are reused; new partition-present references qualify before transformed fits. T_- and teacher readouts are companions, not replacement primaries. The primary passes, with signs 20/20, 19/20 and 20/20; T_- reverses each direction in 20/20. All six neural and 18 teacher cells qualify. This establishes component sufficiency under the matched protocol, not its necessity or a universal disadvantage of contrastive training.

NR beyond matched step length. After FULL was known, the T_+ intervention required qualified reference, FULL, NR and MC cells, and positive intervals for NR adoption, NR–FULL and NR–MC. All pass, each with positive signs in 19/20. MC need not itself have resolved negative adoption. Its failure to reproduce the gain excludes the tested length change alone, without holding future optimizer states or trajectories fixed. The natural $X - U$ extrapolation was specified before these fits, conditional on their success. It qualifies but all three effects remain unresolved (Table 3b); this is distinct from the later 19-participant TC discovery.

D.2 IMAGE-SELECTED RECURRENCE AND FAILED QUANTITATIVE PREDICTION

T_{img} changes the coordinate-selection rule while retaining neural source history in U and stopping. The original gate required three resolved forecast directions; it failed because partition-forward $F_1 = +0.000323\ [-0.000995, +0.001645]$ spanned zero. No transformed fit ran under that gate. After reference outcomes, but before receiving fits, an amendment retained targets/forecasts and separated actual reversal from quantitative prediction. The reversal passes with signs 20/20, 19/20 and 20/20 (Table 3a).

Forecasts apply forward/inverse diagonal maps to individual reference outputs, normalize, then rebuild templates. The ordered inverse-forecast test does not establish lower participant-mean error than zero or forward following: MSE is 2.4864183×10^{-5}, with reductions $-1.3469472 \times 10^{-5}\ [-3.5376650 \times 10^{-5}, +2.4113652 \times 10^{-6}]$ and $-5.9504772 \times 10^{-6}\ [-2.3610719 \times 10^{-5}, +8.8976856 \times 10^{-6}]$, respectively. Concept MSEs are 0.001252107 (inverse) and 0.001141560 (forward), versus

0.000243550 for zero; both are worse in 20/20. Directional recurrence therefore does not validate a finite-training prediction law. An earlier qualified natural-X matched-component test also left its neural interaction unresolved; the present component reversal is specific to the tested rescaled targets.

E FIGURE 3: TARGET-DEPENDENT TREATMENT AND PREDICTION LIMITS

All comparisons below use 19 participants, three nested seeds and participant $t95$ intervals. The prospective contrasts, discovery contrasts and quantitative forecast check answer different questions.

Table 4: **Support for Figure 3.** Panel (a) uses normalized rank $\times 10^3$; only its first two rows are ordered prospective tests. Panels (b,c) use unscaled normalized rank. Natural X adoption is discovery evidence; the first four X_2 rows are ordered prospective tests. Remaining intervals are nominal. $H = \mathrm{NR} - \mathrm{TC}$, so negative H favors TC.

(a) Frozen response discrimination; forecasts precede new TC fits

Contrast	Forecast	Observed	95% interval
$H(T_+)$	$+7.371$	$+4.599$	$[+3.443, +5.755]$
$H(T_+) - H(X)$	$+7.335$	$+18.416$	$[+14.995, +21.836]$
$H(X)$	$+0.037$	-13.817	$[-16.472, -11.161]$

(b) Adoption discovery and control increments for natural X

Contrast	Mean	95% interval
D_{FULL}	-0.002075	$[-0.004933, +0.000782]$
D_{NR}	-0.005169	$[-0.011775, +0.001438]$
D_{MC}	-0.005238	$[-0.009649, -0.000828]$
D_{TC}	$+0.008648$	$[+0.003530, +0.013766]$
TC–FULL	$+0.010723$	$[+0.006282, +0.015164]$
TC–NR	$+0.013817$	$[+0.011161, +0.016472]$
TC–MC	$+0.013886$	$[+0.011099, +0.016674]$

(c) Frozen five-crop X_2 recurrence

Contrast	Mean	95% interval	Positive
$H = \mathrm{NR} - \mathrm{TC}$	-0.01424492	$[-0.01718284, -0.01130700]$	0/19
D_{TC}	$+0.01030015$	$[+0.00526775, +0.01533254]$	15/19
TC–MC	$+0.01337273$	$[+0.01082061, +0.01592484]$	19/19
TC–FULL	$+0.01065218$	$[+0.00644586, +0.01485850]$	18/19
D_{FULL}	-0.00035203	$[-0.00303138, +0.00232732]$	8/19
D_{NR}	-0.00394477	$[-0.01051473, +0.00262519]$	10/19
D_{MC}	-0.00307258	$[-0.00726686, +0.00112170]$	6/19

E.1 PROSPECTIVE CONTRASTS AND NATURAL-TARGET DISCOVERY

After development on one participant, the response formula was fixed for 19 whose FULL/NR/MC outcomes were already exposed. Its retrospective gate required positive concept-MSE gains over zero for both historical FULL responses and $\widehat{H}(T_+) > \delta$, $\widehat{H}(T_+) - \widehat{H}(X) > \delta$, with $\delta = 1/(10 \times 165)$. This passed before new TC fits; it was not independent predictor validation.

The ordered new tests $H(T_+) > 0$ and $H(T_+) - H(X) > 0$ pass, with signs 18/19 and 19/19. All required cells qualify. The unpredicted $H(X) < 0$ holds in 19/19; positive TC adoption (15/19) and its control increments are discoveries with nominal intervals. On T_+, NR adoption is $+0.002408$ $[+0.001733, +0.003082]$, while TC adoption is -0.002191 $[-0.003014, -0.001368]$. Thus treatment preference crosses, but unresolved FULL adoption on X precludes a strict FULL-to-TC reversal.

E.2 PROSPECTIVE FIVE-CROP RECURRENCE

After the X discovery, the X_2 recipe and four tests were fixed before new FULL/NR/MC/TC fits. The targets were sealed before training and all fits before endpoint reduction. Teacher, participants, seeds, stopping, text branch and decision stay fixed; U references are reused and TC uses the actual NR norm donor. Ordered tests $H(X_2) < 0$, $D_{\mathrm{TC}} > 0$, TC–MC> 0 and TC–FULL> 0 all pass (Table 4c); all cells qualify with positive chance/wrong advantages in 19/19. FULL/NR adoption remain unresolved. This extends construction within the same exposed setting, not population, teacher or task; the crop recipe was not derived from the predictor.

E.3 WHAT THE PREDICTOR COMPUTES AND WHERE IT FAILS

Along $Z(\alpha) = U + \alpha(Z - U)$ without intermediate renormalization, differentiate parameters and both AdamW moments at $\alpha = 0$:

$$\dot{s}^a_{t+1} = D_s\Phi_a(s^0_t, U)\dot{s}^a_t + D_Z\Phi_a(s^0_t, U)[Z - U].$$

Φ_a is the complete arm update. FULL includes all three forcing terms in Equation 6; NR omits only direct target reweighting. MC/TC propagate their optimizer responses and the derivative of the NR norm donor. For TC at the identity, $p = \mathrm{softmax}(\ell)$ and interior $\beta = 1$ give $\dot{\beta} = -\operatorname{Cov}_p(\ell, \dot{\ell})/\operatorname{Var}_p(\ell)$ when logit variance is positive. Here $\dot{\ell}$ is the row/column direct perturbation $a\mathbf{R}^0_t(Z - U)^\top$; state effects enter $D_s\Phi_a$. The predictor models donor response; it does not read the actual receiving NR norms later supplied to TC.

Extrapolate with $\widehat{q}_a = \mathrm{unit}(q_0 + \dot{q}_a)$, rebuild query-excluded templates, then average seed scores before ranking. No amplitude or postrun coefficient is fitted. Finite-difference checks and exact reference replay validate implementation, not finite-intervention prediction. Differentiating through training itself is established (Maclaurin et al., 2015; Franceschi et al., 2017).

Concept-level H MSE is 0.000064482 versus zero-change 0.000163233 for T_+ (60.5% lower; 19/19 improve), but 0.001686457 versus 0.001474973 for X (14.3% higher; 19/19 worsen). Participant-mean aggregation does not replace this prespecified finer-grained check. Postrun residuals $e_a = D_a - \widehat{D}_a$ locate the missed arm: $e_{\mathrm{NR}} = +0.002275\,[-0.003141, +0.007691]$ and $e_{\mathrm{TC}} = +0.016129\,[+0.011076, +0.021181]$; this is not an operational repair or a unique causal explanation.

Postrun response repairs remain insufficient. Finite-forcing/one-step-AdamW variants retained the reference-state Jacobian. The optimizer-aware exact-forward variant predicts $H(X) = -0.014381$ but TC adoption -0.078262, versus the observed positive value. A common adoption error -0.087192 (19/19 negative) cancels in the difference; concept H MSE is still 11.33 times zero for X and 56.66 times for T_+. Recovering preference direction therefore does not explain adoption or repair quantitative prediction.

F FIGURE 4: DOES THE STATE SUPPLYING WEIGHTS MATTER?

For fixed supervised $Z = X_2$, the original operator W receives two inputs:

	Z as weight input	U as weight input
receiving state	$A:\ W(\mathbf{R}^A_t, Z)Z$	$B:\ W(\mathbf{R}^B_t, U)Z$
reference state	$C:\ W(\mathbf{R}^U_t, Z)Z$	$D:\ W(\mathbf{R}^U_t, U)Z.$

The synchronized U learner has paired initialization, batches, dropout and steps but its own trajectory. It supplies weights only; each receiver retains its attraction, parameters and normalization Jacobian. This changes the numerical state input, not merely gradient detachment. All four cells use native AdamW with unmatched step lengths, not TC's projected operator.

The design followed the X_2 A/B outcomes. Exact parameter, optimizer-state and trajectory equivalence permits their reuse; C/D are new, with fixed training before their outcomes. The nondirectional primary tests $\Gamma = (A - B) - (C - D)$ by participant $t95$. A–C and B–D use $t97.5$ for their two-test family; simple/adoption effects are nominal $t95$. All five required cells qualify; C's wrong-assignment advantage is positive in 18/19, with a positive population interval.

Table 5: **Support for Figure 4.** Original-W factorial in normalized rank. $D_a = J_a - J_U$. Positive counts use the displayed orientation. The interaction is tested directly; it does not require two individually resolved simple effects of opposite sign.

Contrast	Mean	Interval	t coverage	Positive
$A - B$	$+0.003593$	$[-0.001733, +0.008919]$	95	11/19
$C - D$	-0.056733	$[-0.075156, -0.038310]$	95	0/19
$\Gamma = (A - B) - (C - D)$	$+0.060326$	$[+0.038233, +0.082418]$	95	18/19
$A - C$	$+0.075901$	$[+0.048600, +0.103203]$	97.5	19/19
$B - D$	$+0.015576$	$[+0.010501, +0.020650]$	97.5	18/19
D_A	-0.000352	$[-0.003031, +0.002327]$	95	8/19
D_B	-0.003945	$[-0.010515, +0.002625]$	95	10/19
D_C	-0.076253	$[-0.100701, -0.051806]$	95	1/19
D_D	-0.019520	$[-0.028090, -0.010951]$	95	2/19

The direct interaction resolves (18/19 positive), while $A - B$ remains unresolved and $C - D$ is negative in 19/19. Figure 4 uses the opposite simple-effect orientation so positive values favor U. Both C/D have negative adoption: D improves on C without exceeding reference supervision. This identifies input interaction, leaving magnitude and optimizer dynamics as possible mechanisms.

Separate TC state substitution. For known T_+ and X, replace $\Pi(\mathbf{R}_t, Z; U)Z$ by $\Pi(\mathbf{R}_t^U, Z; U)Z$, retaining attraction and the same NR norm donor. Only reference-state outcomes are new. The two X co-primary $t97.5$ intervals give live-minus-reference-state $K_X = +0.070302$ $[+0.048529, +0.092075]$ and reference-state adoption -0.061654 $[-0.087218, -0.036090]$. All 19 lose performance; the 15 previously positive adopters cross below reference, not 19 individual sign reversals. The T_+ companion has a larger state increment $+0.084276$ $[+0.062733, +0.105819]$ and reference-state adoption -0.086467 $[-0.108416, -0.064518]$ (nominal $t95$); live adoption is already negative. Both new cells qualify. Thus state contribution is neither X-specific nor sufficient for adoption; this matched-length TC comparison is not a cell of the original-W factorial or an identification of the predictor's unique error cause.

G TABLE 1: SCOPE AND ADOPTION BOUNDARIES

Main Table 1 reports the S/R/adoption contrasts for these settings. We give the additional design information and boundary contrasts here, without repeating that table. These changes in target, architecture, population and consumer are not interchangeable replications.

G.1 TARGET SELECTION AND ARCHITECTURE

After T_{img} FULL and the T_+ intervention were known, NR/MC were fixed under the unchanged MLP protocol. S/R resolve (18/20 positive each), but adoption does not; all task cells qualify. This extends intervention gain within the diagonal family, not positive adoption.

CNN extension. ShallowNetBVSA receives identical targets, 17×200 inputs and stopping budgets. Its existing projection maps 1,040 features to 1,024; convolution, BatchNorm, ELU, pooling and dropout 0.5 differ from the MLP. Capacity is not matched. References qualify before receiving fits; the S→R→adoption sequence passes, with positive counts 18/20, 18/20 and 15/20. All four neural and 12 teacher cells qualify. FULL adoption -0.000892 $[-0.002023, +0.000518]$ and MC adoption -0.000475 $[-0.001648, +0.000990]$ remain unresolved, precluding a strict FULL reversal. BatchNorm buffers are not norm-matched.

G.2 NEW PARTICIPANTS AND THE TEMPERATURE COMPARISON

THINGS-EEG2 participants 05–10 are disjoint from Alljoined but share stimuli and source-informed target design. Each uses its own guide, stopping and fitting-derived diagonal; all cells qualify. The order is S$>$ 0, R$>$ 0, I$<$ 0, then adoption$>$ 0. S/R resolve (6/6 positive), but $I = +0.000531$ $[-0.001019, +0.002081]$, so adoption is nominal and unresolved. Partition-free adoption is

-0.002218 $[-0.003369, -0.001067]$ and FULL adoption -0.001687 $[-0.002299, -0.001074]$. Thus intervention gain transfers, while component reversal and positive adoption do not resolve.

A separately specified TC test gives $H = \mathrm{NR} - \mathrm{TC} = +0.000106$ $[-0.000487, +0.000700]$ ($t95$). The prespecified equivalence margin is $\pm 1/(10 \times 165) = \pm 0.000606061$; $t90$ interval $[-0.000358560, +0.000571534]$ lies inside it. This establishes mean-score equivalence at that precision, not equivalent representations or temperature rules generally. TC adoption is -0.000518 $[-0.001062, +0.000025]$; TC–MC is $+0.001142$ $[+0.000463, +0.001821]$ and TC–FULL $+0.001168$ $[+0.000526, +0.001811]$. The transport result does not identify a unique NR advantage over TC.

G.3 Fixed external consumer and prototype targets

The $X \to P$ intervention keeps the 20-participant MLP protocol. Before receiving outputs, an orthogonal map from source teacher features to DINOv2-L (Oquab et al., 2024) and a 166-concept DINO gallery are fixed. The map uses 1,654 source concepts including exposed selection concepts; this is not unseen-concept evaluation. Seed cosine matrices are averaged before top-1, with fractional tie credit. Candidate vectors are external to the receivers.

All cells qualify. Top-1 S/R resolve but adoption does not; reference/NR accuracies are 1.38554%/1.39458% (chance $1/166 = 0.6024\%$). The rank companion gives S $+0.005938$ $[+0.005020, +0.006887]$, R $+0.005213$ $[+0.004352, +0.006100]$, but adoption -0.002739 $[-0.004880, -0.000474]$. A control masks off-diagonal same-class candidates; identical within-class prototypes make FULL's attraction agree with the multi-positive attraction gradient. NR exceeds masking by $+0.12651$ pp $[+0.06325, +0.19277]$, nominal because unresolved adoption blocked its ordered test.

A later exposed-condition diagnostic adds partition-free fits. Top-1 $D_0 = -0.003012$ pp $[-0.054217, +0.045181]$ and $I = -0.084337$ pp $[-0.195783, +0.030120]$ are unresolved; rank $D_0 = -0.005673$ $[-0.006919, -0.004422]$ and $I = -0.002279$ $[-0.003929, -0.000581]$ are negative. Table 6 shows why positive R need not offset the reference-relative deficit; its identity is arithmetic, not causal shares.

Table 6: Endpoint accounting, $D_{\mathrm{NR}} = D_0 + I + R$, using the frozen means. Neural/DINO-rank rows are $\times 10^3$; top-1 is pp. This table organizes the outcomes, not mediation shares or new hypothesis tests. Component and intervention intervals are in Table 3, main Table 1 and the prototype paragraph above.

Target / endpoint	D_0	I	R	D_{NR}
T_+ / neural rank	$+10.620$	-17.230	$+9.048$	$+2.438$
T_{img} / neural rank	$+7.355$	-10.302	$+2.896$	-0.051
P / fixed-DINO top-1 pp	-0.003	-0.084	$+0.096$	$+0.009$
P / fixed-DINO rank	-5.673	-2.279	$+5.213$	-2.739

G.4 Visual classification: competence, readout and training CE

CIFAR-100 (Krizhevsky, 2009) uses 40,000 fitting, 10,000 development and the official 10,000 test images. Frozen ViT-H supplies unit original-image U and normalized mean original/flipped features Z. Five seeds 2025–2029 pair initialization, order and augmentation. Final checkpoints precede each terminal test reduction, but later designs inherit benchmark exposure. Table 7 retains all stages and direct moderator tests.

Training and competence. The initial ResNet-20 fixed-ridge experiment used seven arms and 60 epochs. Teacher development accuracy was 89.32%; reference mean 10.472% passed its original mean-10%/each-seed-1% gate. Its unresolved effects remain valid under that weak gate.

The later CIFAR ResNet-18 (He et al., 2016), selected before X_2/factorial outcomes, uses a 3×3, stride-one stem without initial maxpool, 512-dimensional pooled features, a biased linear 1,024-dimensional unit feature head and a separate biased 100-class head. Its loss combines label CE, half mean squared feature error and symmetric partition at scale 10. Training uses 200 epochs, batch 128

Table 7: The bounded visual sequence: classification-accuracy differences in pp, paired-seed $t95$ (df 4). All S, R and adoption intervals span zero. Later rows are exposed-benchmark diagnostics, not independent replications. Signs in parentheses count positive seeds. Moderator intervals are direct paired contrasts, not comparisons of significance labels.

Apparatus / consumer	S	R	D_{NR}
Initial fixed ridge	-0.026	$+0.066$	$+0.044$
	$[-.494, +.442]$	$[-.252, +.384]$	$[-.317, +.405]$
Competent learned head	-0.130 (1/5)	$+0.004$ (2/5)	-0.094 (2/5)
	$[-.302, +.042]$	$[-.326, +.334]$	$[-.474, +.286]$
Same fits, fixed prototypes	-0.310 (1/5)	-0.106 (2/5)	-0.166 (2/5)
	$[-.876, +.256]$	$[-.375, +.163]$	$[-.599, +.267]$
No CE, fixed prototypes	-0.014 (2/5)	$+0.140$ (5/5)	$+0.096$ (4/5)
	$[-.362, +.334]$	$[-.039, +.319]$	$[-.211, +.403]$
Fixed minus learned: M_S		-0.180 (2/5)	$[-.638, +.278]$
No-CE minus CE: M		$+0.296$ (3/5)	$[-.424, +1.016]$

(tail 64), AdamW learning rate .001, decay .0001, betas $(.9, .999)$, epsilon 10^{-8}, epochwise cosine decay without warmup, float32, padded random crops and horizontal flips. Cached teacher targets stay attached to image IDs; no best-checkpoint selection occurs. MC matches backbone and both heads' actual NR displacement, including label-loss and decay contributions, while retaining its own moments.

Teacher development must reach 80%; reference and receiving arms must have mean accuracy at least 50%, each seed at least 40%, and finite outputs. All pass; reference development mean is 73.50%. Test means are 73.752% (reference), 73.654% (FULL), 73.658% (NR), 73.788% (MC). The ordered S→R→adoption sequence stops at unresolved S.

Fixed-prototype readout. A subsequent frozen diagnostic reads the same feature-head checkpoints with unit class prototypes formed from the 400 fitting U images per class. No development/test features, receiving outputs or new probe fit the consumer; maximum inner product predicts class. Reference development mean 68.952% (minimum 68.670%) passes its mean-20%/each-seed-10% gate. Both $S_{\mathrm{fixed}} > 0$ and direct $M_S = S_{\mathrm{fixed}} - S_{\mathrm{learned}} > 0$ remain unresolved. The heads access different interfaces; this is a compatibility test, not independent information recovery.

Separately frozen no-CE test. The preceding failed gate closed the conditional follow-up. A separate post-stop diagnostic was then frozen before new no-CE fits: remove training CE while retaining targets, split, network (including the unused class head), optimizer, 200 epochs and fixed prototypes. Labels still construct prototypes and evaluate the task. New reference development mean 69.972% (minimum 69.590%) passes mean 63.952% (previous mean minus five pp) and each-seed 10% thresholds. MC uses this new NR donor.

The order is S, $M = S_{\mathrm{noCE}} - S_{\mathrm{CE}}$, R, then adoption, with fixed-consumer CE $S = -0.310$ pp as comparator, not learned-head -0.130 pp. All four intervals span zero individually. R nevertheless has seed effects $+0.09, +0.16, +0.02, +0.38, +0.05$ pp (5/5 positive); descriptive MC–FULL is $+0.154$ pp. Neither positive gain beyond MC, equivalence nor a complete magnitude explanation follows. No further target, seed or consumer was selected after this result. These tests leave positive visual transfer open without establishing EEG specificity or general CE irrelevance.

H Closest prior, reproducibility and ethics

H.1 What the contribution adds to existing weighting methods

DKD, CSD and coordinate-wise contrastive optimization already separate supervision and weighting (Zhao et al., 2022; Kim et al., 2026; Tian, 2022). DCL and MACL study coupling and temperature (Yeh et al., 2022; Huang et al., 2023); SCE/ReSSL supply teacher similarity distributions (Denize et al., 2023; Zheng et al., 2021); RKD/DarkRank transfer relational structure (Park et al., 2019; Chen et al., 2018). The present contribution is the same-target adoption counterexample: retain

paired/candidate vectors, replace only the weight input, and cross the same reference baseline beyond an actual-step-length control. NR/TC retain reference information in weighting, so their benefits are conditional on that treatment. It is not novelty for weighting, relational supervision or adaptive temperature.

Teacher–student compatibility and scaling already affect distillation (Qian et al., 2025; Cho & Hariharan, 2019; Chen et al., 2021); visual-target adaptation is established in neural decoding (Wu et al., 2025; 2026; Liu et al., 2026). Readout switching and recoverability also constrain representation comparisons (Li et al., 2023; Kim & Lee, 2026; Onoo et al., 2026; Harvey et al., 2024). Our fixed decision rule still uses model-specific templates; the fixed DINO test supplies a separate, low-accuracy boundary. The matched component, NR–MC contrast and unmatched original-W factorial answer different questions and cannot be combined into unique mediation. Natural-target prediction fails despite established differentiation-through-training tools (Maclaurin et al., 2015; Franceschi et al., 2017); neither a general adoption law nor a universally superior NR/TC policy is established.

H.2 Reproducibility materials

The anonymous supplementary package links main figures and scope tests to frozen summaries, participant/seed score panels, selected actual training traces and production source. Its independent reducer checks effect orientation, seed-score averaging before rank, the original interval conventions and figure values. A real score-to-rank witness and visual test predictions permit endpoint inspection. Synthetic gradient, identity, optimizer-displacement and training-interface checks are explicitly separate from scientific results.

Full retraining additionally requires public datasets, upstream models and frozen initial states/large target tensors not bundled in the package. The study records retain all TC-state endpoint scores/traces but only selected image checkpoints; model checkpoints are not bundled. The included hashes establish file integrity, not independent authentication of original data acquisition or execution. Data contracts and replay requirements are listed in the package's `DATA_AND_REPLAY.md`; `PAPER_MAP.md` maps current claims to evidence. Earlier exploratory records do not confer confirmation status or additional sample size on this paper's tests.

H.3 Ethics

This work reanalyzes existing de-identified EEG datasets under their respective licenses. It makes no clinical inference and does not attempt to identify participants. Neural decoding can create privacy and agency risks outside controlled research. The conclusions concern group-level evaluation and training interactions for visual-decoding representations, not individual intent or unrestricted mental-state decoding.