HIDDEN LABEL DISTINCTIONS CHANGE THE GAIN FROM REPRESENTATION RATE

Anonymous authors
Paper under double-blind review

ABSTRACT

How much accuracy does a fixed increase in representation rate buy under ambiguous supervision? We study whether this marginal gain changes when candidate-label channels have equal ambiguity and spectrum but hide different class distinctions. Using frozen ImageNet features, we compare a fixed increase from 32 to 64 nominal bits per feature vector. Separate DESIGN measurements select the class pairs with the largest and smallest discrimination gains, fixing two matched hiding interventions and their expected gain ordering before downstream fitting and TEST. Across 68 parent groups and 340 classes, hiding the largest-gain pairs yields 0.702 percentage points less mean fine-class gain (95% interval [0.088, 1.345]); the fine-minus-coarse contrast is 0.752 points. The interaction persists with each high-rate-trained classifier held fixed; fixed-low persistence remains unresolved. The higher final-accuracy arm has the smaller marginal gain. Difficulty and headroom remain possible explanations; the contrast does not isolate a unique mechanism. Concept-preserving EEG correspondence controls provide supporting discovery. In the tested system, matched channel ambiguity and spectrum do not equate the marginal value of the same representation increment.

1 INTRODUCTION

What determines the marginal value of a fixed representation-rate increment? We measure this value as the increase in task accuracy under a specified supervision relation and reader-fitting rule. The return on an upgrade concerns the improvement over the existing representation; it need not follow the ranking of final accuracies. With candidate-label supervision, two channels can each supply two labels including the truth, yet leave different class pairs indistinguishable. Our question is whether the same representation increment yields different gains when these channels have equal ambiguity and spectrum but leave different class distinctions unresolved. We call the identity of these hidden distinctions the channel's *orientation* relative to the fixed representation. The comparison holds the task, representation pair, and reader family fixed while changing this orientation.

Information-bottleneck formulations trade compression against predictive information (Tishby et al., 1999; Alemi et al., 2017). Predictive and supervised compression evaluate downstream utility (Dubois et al., 2021; Matsubara et al., 2022); nested representations support multiple resource levels (Kusupati et al., 2022). Cross-domain and multi-task compression address domain shifts and task dependencies (Nguyen et al., 2026; Guo et al., 2025). Usable-information theory (Xu et al., 2020), task-compatible coding (de Andrade & Bajić, 2024), and representation-size analysis (Njaradi et al., 2026) relate useful resources to the task and reader. We compare the same input-rate increment and task across supervision relations matched in ambiguity and spectrum, asking which properties of the learning relation must accompany a rate–utility measurement.

Supervision-dependent capacity effects are known: label noise changes generalization curves as model width varies (Xue et al., 2024). Partial-label learning studies learnability (Cour et al., 2011), consistent risk estimation (Feng et al., 2020), and label identification (Lv et al., 2020; Tian et al., 2024). Representation learning (Wang et al., 2022; Xia et al., 2023) and candidate pruning (He et al., 2024) exploit feature–label geometry; noisy-correspondence methods address imperfect matches (Huang et al., 2021; Jia et al., 2021). Our test concerns the interaction between input representation rate and matched channel orientation, with reader family and width held fixed. The target is a difference between two low-to-high accuracy gains.

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.

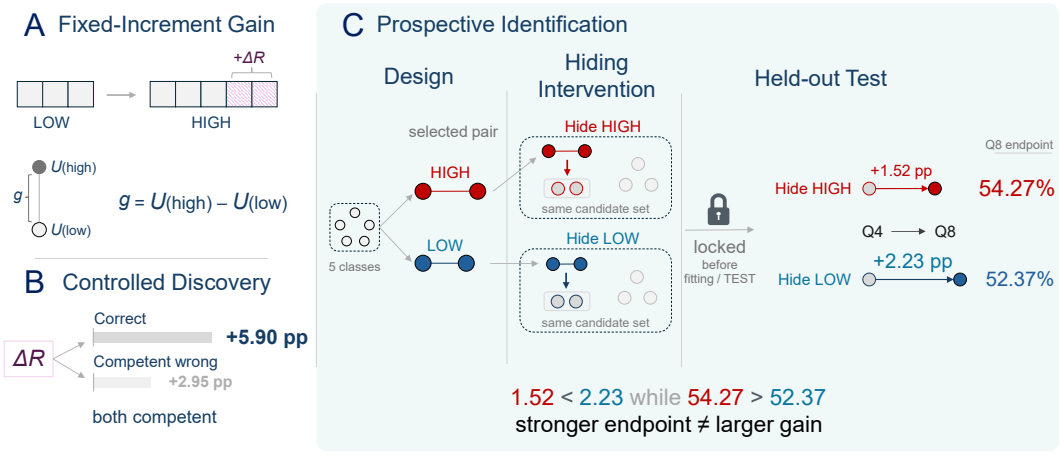

Figure 1: **Measurement → controlled discovery → prospective test.** (a) Marginal gain is the downstream return from one fixed low-to-high representation increment. (b) In EEG, correct and within-concept image-swap supervision are both task-competent but gain by different amounts. (c) In ImageNet, a separate DESIGN split ranks fine-class pairs by their Q4-to-Q8 discrimination gain; matched supervision channels then hide either the HIGH or LOW pair by mapping its two classes to the same candidate set. The map, channels, and directional hypothesis are locked before downstream-reader fitting and TEST. Held-out TEST confirms the gain ordering, while Q8 endpoint accuracy orders the arms oppositely. Figures 2–3 and Table 2 report uncertainty; physical rate units differ across settings and are not pooled.

Our primary experiment uses frozen pretrained ImageNet features (Russakovsky et al., 2015) and a Q4-to-Q8 increase from 32 to 64 nominal bits per feature vector. Both rates decode to the same feature dimension, keeping the downstream classifier's parameter count fixed. Separate DESIGN measurements construct the intervention: binary probes select the largest-gain pair HIGH and smallest-gain pair LOW within each five-class parent. In either arm, the selected pair maps both true classes to the same two-label candidate set; the remaining three classes form the same symmetric candidate-pair construction. The channels match in candidate size, true-label inclusion, conditional uncertainty, rank, spectrum, objective, and exposure. They are label permutations of the same rank-four construction, with one bit of conditional label uncertainty under the balanced class prior. Changing the hidden pair changes this channel's orientation relative to the fixed features. If supervision hides a distinction newly accessible at higher rate, its discrimination benefit may fail to translate into global task gain. We predeclare a smaller gain for HIDE_HIGH and lock the channels before fitting global 340-way classifiers and before TEST. Their objective and endpoint differ from the binary probes; the directional contrast is not an algebraic consequence of pair selection.

Held-out TEST confirms the predeclared contrast. HIDE_HIGH gains 1.524 percentage points and HIDE_LOW gains 2.225, a paired difference of 0.702 points (95% interval [0.088, 1.345]). Parent groups are paired inference units, with three fitting seeds averaged within each group. We also compare the gain difference across fine and coarse classification, testing whether it exceeds a shift common to both resolutions. The fine-minus-coarse contrast is 0.752 points, and the interaction remains 0.431 points with each high-rate-trained reader held fixed (Figure 3; Table 2). The latter comparison changes only the representation input within each arm, showing that this interaction does not require refitting at each evaluation rate.

The matched channel summaries assign both arms the same description, yet the same representation upgrade has different measured value. They specify how much candidate ambiguity remains and the channel's algebraic structure, while leaving its relation to the fixed feature geometry unspecified. HIDE_HIGH also has higher Q8 accuracy but smaller gain, separating final performance from marginal return. Thus comparing final systems and comparing the return on upgrading them can order the same arms differently. This reversal does not establish a better allocation policy.

The evidence is conditional on one pretrained geometry, rate increment, and shared reader system: both primary effects are positive in 42 of 68 groups, and fixed-low persistence remains unresolved.

The groups share the pretrained backbone and fitted readers; they are not 68 independent replications of the full system. HIGH pairs have lower DESIGN Q4 accuracy, while HIDE_LOW has lower global Q4 accuracy; neither difficulty nor headroom was isolated. The test measures the total contrast between the constructed channels, without isolating incremental discrimination from baseline difficulty. Thus the independent value of the DESIGN score and the internal explanation of the contrast remain open. EEG correspondence controls preserving concept labels provide supporting discovery, with a 2.95-point rate-gain contrast (Figure 2). Candidate-set studies retain gains but leave a scalar ordering unresolved (Figure 4); these are distinct evidence roles, not replications of the ImageNet orientation test.

We characterize the supervision-conditioned marginal value of a fixed representation increment through a prospective intervention. In the tested system, matching ambiguity and channel spectrum does not equate this value: which distinctions supervision leaves unresolved matters. Rate–utility comparisons should specify the learning relation and reader state, alongside the representation resources and task.

Table 1: **Experimental contracts and evidence roles of the two scoped settings.** Both settings measure low-to-high marginal gain, but their tasks, physical rate coordinates, and population inference units differ; comparisons are paired within setting and never pooled across settings.

	EEG discovery	**ImageNet prospective test**
Representation	Reconstructed EEG from frozen residual vector quantizer (RVQ)	Frozen ResNet18 features with nested RVQ
Low → high	800 → 1600 nominal bit/s	Q4 → Q8: 32 → 64 nominal bits/vector
Task	144-image Top-1 retrieval	340-class fine / 68-class coarse classification
Learning relation	Correct vs. within-concept wrong-image supervision	HIDE_HIGH vs. HIDE_LOW candidate supervision
Reader fitting	Matched target-labelled updates	Matched true-containing candidate-set objective
Evaluation	Held-out concepts	Held-out TEST images from selected classes
Inference unit	Participant ($n = 8$); three fitting seeds nested	Parent group ($n = 68$); three fitting seeds nested
Evidence role	Controlled discovery	Prospective intervention test

2 Marginal Value: Supervision and Reader State

For the two scoped settings in Table 1, we separate the representation rate used for reader training from that used for evaluation.

Let Z_L and Z_H denote two fixed members of the same nested representation family, with Z_H using the larger resource budget. For learning relation r, let $U_r(t, e)$ denote task accuracy when the reader is trained at $t \in \{L, H\}$ and evaluated using representation input $e \in \{L, H\}$. The diagonal marginal gain and its contrast across two learning relations are

$$g_r = U_r(H, H) - U_r(L, L), \qquad J_{a,b} = g_a - g_b. \tag{1}$$

Thus g_r compares two rate-specific systems, changing both the input representation and fitted reader.

To measure the response to the same increment while holding reader weights fixed, define

$$D_{r,t} = U_r(t, H) - U_r(t, L), \qquad K_{t;a,b} = D_{a,t} - D_{b,t}. \tag{2}$$

$D_{r,t}$ holds fitted weights constant while changing input rate; $K_{t;a,b}$ compares that response across learning relations. A nonzero $K_{t;a,b}$ shows that per-rate refitting is unnecessary for the corresponding interaction. The relations retain separate readers, so K is not a causal supervision effect under common model parameters.

The crossed design also defines the complementary *reader-training contrast*,

$$\begin{aligned} T_{r,e} &= U_r(H,e) - U_r(L,e), & T_{e;a,b} &= T_{a,e} - T_{b,e}, \\ g_r &= D_{r,H} + T_{r,L} = D_{r,L} + T_{r,H}, & & \\ J_{a,b} &= K_{H;a,b} + T_{L;a,b} = K_{L;a,b} + T_{H;a,b}. & & \end{aligned} \tag{3}$$

These exact identities traverse the four train/evaluate cells in opposite orders; they imply neither a continuous optimization trajectory nor causal mediation shares. $T_{r,e}$ absorbs every consequence of fitting different readers at the two rates, so the accounting does not uniquely partition information, optimization, or supervision effects.

Experimental contract. Within each experiment, the compared learning relations receive the same low-to-high representation increment, data identities, training exposure, reader architecture, and evaluation support. They also share the quantizer, normalization, paired fitting seeds, and fixed final-checkpoint rule; no held-out TEST outcome selects a rate, checkpoint, or supervision channel. The learning relation changes which label, match, or candidate set the training objective rewards. The ImageNet intervention further matches the training objective, candidate structure, exposure, and scalar channel spectrum while changing which fine-class distinction is hidden.

The representation increment uses an exact nested prefix of a residual vector quantizer (RVQ) (Barnes et al., 1996). With M active codebooks $\{\mathcal{C}_\ell\}_{\ell=1}^{M}$, the nominal resource coordinate is

$$R_{\mathrm{vector}}(M) = \sum_{\ell=1}^{M} \log_2 |\mathcal{C}_\ell|, \qquad R_{\mathrm{time}}(M) = f_z R_{\mathrm{vector}}(M). \tag{4}$$

EEG uses four or eight 256-entry codebooks at $f_z = 25$ Hz, giving 800 or 1600 nominal bit/s. Vision uses the same four/eight-stage prefix once per feature vector, giving 32 or 64 nominal bits/vector. These physical rate coordinates differ across settings and are never pooled numerically. Entropy-model-based codecs optimize rate–distortion objectives (Ballé et al., 2017; 2018); our coordinates count code indices before entropy coding or model overhead. The intervention changes representation input resources rather than reader width.

We treat the rate intervention as qualified only when the low-rate representation is an exact prefix of the high-rate representation, outputs are finite, the added stages are nontrivially used, reconstruction improves, and the relevant downstream endpoints remain task-competent. Code-usage entropy is reported only as a diagnostic and is not interpreted as achieved bitrate.

Utility is task accuracy, with uncertainty attached to the paired population unit appropriate to each setting. EEG uses Top-1 retrieval accuracy in a closed image gallery. ImageNet fine utility is 340-way classification accuracy; coarse utility sums the five fine-class probabilities within each parent before selecting among 68 parents. Conditional five-way accuracy is used only as a competence check and does not replace the global fine endpoint.

Seeds are averaged within each participant or parent before population inference. EEG therefore treats participants as the inferential units, whereas ImageNet treats the 68 parent groups as the inferential units. We use paired unit-level bootstrap intervals (Efron, 1979) together with the prespecified study-specific sign tests and multiplicity families; repeated queries and fitting seeds do not create additional population units. Detailed EEG protocols appear in Appendix B, and the complete prospective ImageNet contract appears in Appendix A.

3 Controlled Discovery in EEG

Competent-control discovery. EEG provides the controlled discovery motivating the prospective test. Correct supervision (A) rewards the true image; a within-concept swap (S) rewards another training image of the same concept, preserving concept membership and marginal image exposure while altering correspondence. The test gallery contains one image per concept, so success requires concept discrimination rather than within-concept image identity. S therefore corrupts image correspondence while retaining the semantic labels needed by this endpoint. Both arms use the same frozen 800-to-1600 nominal-bit/s increment, matched reader architectures, and equal training exposure. The comparison uses target-labelled training in eight reused Alljoined-1.6M participants

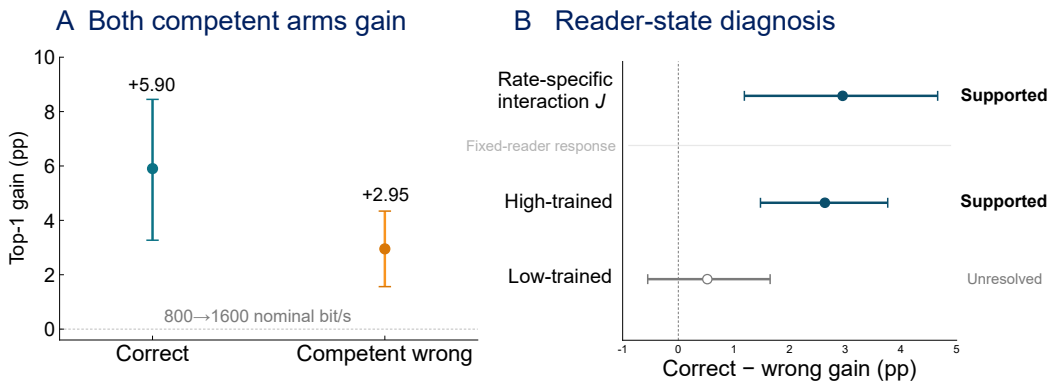

Figure 2: **Controlled EEG discovery and reader-state diagnosis.** (a) Correct and within-concept wrong supervision are both task-competent and gain from the same frozen 800-to-1600 increment, with a larger gain under correct supervision. (b) The correct-minus-wrong interaction is supported with rate-specific readers and with high-trained readers held fixed; the low-trained fixed-reader contrast remains unresolved. Effects are in percentage points; intervals are 95% participant intervals after averaging three seeds within each of eight participants. Fixed comparisons retain separate correct/wrong readers and diagnose reader-state persistence rather than a common-parameter causal effect.

(Xu et al., 2025), with evaluation on held-out concepts and three seeds averaged within each person (Appendix B.2).

Both learning relations remain competent at both rates and benefit from added rate, but correct supervision gains more (Figure 2a). The paired interaction is $J_{A,S} = 2.95$ pp (95% interval $[1.19, 4.66]$; Holm $p = .03125$), positive in seven of eight participants after averaging seeds. The smaller gain under S occurs within a competent task, so task collapse cannot explain the contrast.

Reader-state decomposition. Crossed evaluation asks whether this interaction requires fitting a new reader at each representation rate. We hold each fitted model's weights constant and change only its input representation rate, retaining separate A and S readers. The interaction persists for high-rate-trained readers, while the low-rate-trained contrast remains unresolved (Figure 2b). Per-rate refitting is therefore unnecessary for the observed high-trained interaction, but reader-independent persistence is not established. For the A–S interaction, Eq. 3 gives two exact accounting paths, with saved estimates rounded to two decimals:

$$\begin{aligned} J_{A,S} &= K_{H;A,S} + T_{L;A,S} = 2.63 + 0.32 \\ &= K_{L;A,S} + T_{H;A,S} = 0.52 + 2.43 = 2.95 \text{ pp}. \end{aligned} \tag{5}$$

Along the high-trained path, most of the observed interaction appears in the fixed-reader response; along the low-trained path, the larger remainder lies on the reader-training side. This path dependence makes reader state diagnostically important without assigning causal mediation shares. The T terms remain descriptive secondary contrasts; their uncertainty and the complete four-cell accounting appear in Appendix B.3.

A true-containing ambiguous objective tests whether a marginal gain can persist without selecting the exact positive image. The set-positive objective (MASS) rewards the true image together with its partner and retains a 7.29-pp gain (95% interval $[3.65, 11.02]$) in a matched single-seed study. Its gain also exceeds S, but the prespecified equivalence to exact supervision remains unresolved. This supports retained marginal gain under ambiguity; objective identities and the complete comparison appear in Appendix C.1.

The EEG study establishes different marginal gains under competent supervision. Section 4 examines this supervision dependence with matched ImageNet channels, testing a predeclared gain contrast after changing which DESIGN-selected distinctions they hide.

4 PROSPECTIVE TEST UNDER MATCHED SUPERVISION

Prospective design. Building on the EEG discovery in Section 3, our primary test asks whether the marginal value of one fixed Q4-to-Q8 increment changes when matched supervision channels hide different distinctions. Separate DESIGN measurements choose the two channel orientations and lock a directional hypothesis; TEST evaluates their marginal-gain contrast.

We test this question in ImageNet (Russakovsky et al., 2015). An outcome-independent rule on the WordNet hierarchy (Miller, 1995) selects 68 disjoint five-class parent groups (340 fine classes), and each class is partitioned into separate FIT, DESIGN, and TEST roles. FIT supplies representation and reader training data; DESIGN fixes the intervention; TEST evaluates the locked objects. Features are frozen 512-dimensional pooled outputs of an ImageNet-pretrained ResNet18 (He et al., 2016), so the experiment concerns a fixed pretrained feature geometry rather than unseen-class or end-to-end ImageNet generalization.

For every parent, DESIGN measures *added accessibility*: the change in binary discrimination of each fine-class pair when moving from Q4 to Q8. The pair with the largest Q8-minus-Q4 improvement is frozen as HIGH and the pair with the smallest improvement as LOW. These selections are made before reader fitting and before TEST. The predeclared hypothesis is that hiding HIGH yields a smaller mean fine-task Q4-to-Q8 marginal gain than hiding LOW. The full role separation, duplicate exclusions, probe specification, and lock order appear in Appendix A.

DESIGN constructs the intervention; its ranking does not determine the TEST outcome. On a split disjoint from TEST, two-fold cross-fitted binary ridge probes optimize squared error to rank pairwise discrimination gains. After the map and channels are locked, global 340-way readers are fitted on FIT, independently of those probes, with the candidate-supervision objective below. The directional hypothesis precedes these fits, and all final checkpoints lock before TEST. TEST measures their global fine-class marginal gain, not binary accuracy on the selected pair. This representation-dependent construction chooses the intervention without TEST outcomes; no algebraic identity forces its predeclared mean ordering. The selections also differ in baseline pair difficulty: mean DESIGN Q4 accuracy is 80.49% for HIGH and 91.04% for LOW. Thus this test does not isolate added accessibility from baseline difficulty or establish superiority over a difficulty-based selection rule. Appendix A reports the saved selection summaries.

Matched supervision intervention. We next change *which* fine-class distinction is unidentifiable from the candidate-supervision channel while holding its scalar structure fixed. Within each five-class parent, a channel hiding classes i and j returns the same candidate set $\{i, j\}$ for either true class. The remaining three classes form a symmetric candidate-pair triangle. Every candidate set contains the true class and remains within its parent. HIDE_HIGH and HIDE_LOW use the same construction, differing only in whether the frozen HIGH or LOW pair is placed in the unidentifiable direction.

The two channels are permutations of the same rank-four channel and therefore share singular values

$$\left\{\sqrt{2}, 1, \tfrac{1}{2}, \tfrac{1}{2}, 0\right\}$$

and one bit of conditional label uncertainty under the balanced class prior. They also share the same candidate-set size, true-label inclusion, exposure, reader architecture, initialization schedule, and probability-mass objective

$$\ell(x, S) = -\log \sum_{y \in S} p_\theta(y \mid Z(x)).$$

For this symmetric channel, the objective agrees with the corresponding forward likelihood up to a parameter-independent constant. Thus the intervention matches the amount and algebraic structure of candidate ambiguity while rotating *which* fine-class distinction is hidden. This does not make the two hidden directions equally difficult relative to the fixed image-feature geometry; that geometry is part of the context being tested. Nor does scalar matching control semantic identity, local feature margins, optimization geometry, or historical pretrained exposure.

All required diagonal and crossed reader states pass the prespecified competence checks. No TEST outcome selects the HIGH/LOW map, representation edge, reader regularization, checkpoint, or fitting seed.

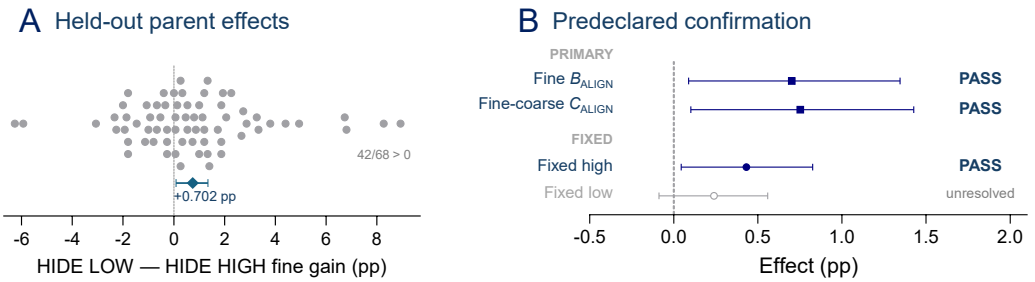

Figure 3: **Held-out confirmation of the predeclared ImageNet contrast.** (a) The 68 held-out parent-level HIDE_LOW-minus-HIDE_HIGH fine marginal-gain contrasts are heterogeneous: 42 are positive, 25 negative, and one is zero; the diamond marks the mean. (b) The predeclared fine (B_{ALIGN}) and fine-minus-coarse (C_{ALIGN}) primary contrasts are shown with the fixed-reader contrasts. Both primaries and fixed-high satisfy the predeclared criteria; fixed-low remains unresolved. Parents are the inference units; intervals are the original 95% parent-bootstrap intervals after averaging three fitting seeds within parent. Table 2 reports exact effects, intervals, and adjusted tests.

Held-out confirmation. Let $g_{r,e}$ denote the diagonal Q4-to-Q8 marginal gain under supervision relation r for endpoint $e \in \{\mathrm{fine}, \mathrm{coarse}\}$. The prospective test uses two predeclared primary contrasts:

$$B_{\mathrm{ALIGN}} = g_{\mathrm{HIDE_LOW,fine}} - g_{\mathrm{HIDE_HIGH,fine}}, \tag{6}$$

$$C_{\mathrm{ALIGN}} = B_{\mathrm{ALIGN}} - \Big(g_{\mathrm{HIDE_LOW,coarse}} - g_{\mathrm{HIDE_HIGH,coarse}}\Big). \tag{7}$$

$B_{\mathrm{ALIGN}} > 0$ asks whether hiding HIGH reduces the mean fine marginal gain relative to hiding LOW. $C_{\mathrm{ALIGN}} > 0$ asks whether that difference exceeds the corresponding coarse-task interaction, limiting an explanation based on a common shift in accuracy across fine and coarse task resolutions. Fixed-reader contrasts $K_{L,\mathrm{ALIGN}}$ and $K_{H,\mathrm{ALIGN}}$ apply the same arm difference to the fixed-reader response in Eq. 2.

The confirmation rule was frozen before TEST: both primary contrasts had to meet the prespecified effect-size and uncertainty criteria, together with persistence in at least one competent fixed-reader comparison. Inference uses the 68 parent groups as paired population units, averaging the three fitting seeds within parent before bootstrap and sign-based inference. Parent-level inference uses working exchangeability and sign-flip assumptions, conditional on the shared pretrained backbone, taxonomy, representation family, and reader system; parents are not independent system replications. The two primary contrasts form one Holm family and the two fixed-reader contrasts another. Full thresholds and inferential assumptions are given in Appendix A.

On held-out TEST, both primary contrasts satisfy the predeclared criteria (Figure 3a,b; Table 2): $B_{\mathrm{ALIGN}} = +0.702$ pp and $C_{\mathrm{ALIGN}} = +0.752$ pp. The fixed-high reader contrast also satisfies its predeclared criterion, whereas fixed-low persistence remains unresolved. All three fitting seeds give positive primary and fixed-high contrasts; at the parent level, however, each primary effect is positive in 42 of 68 groups, indicating heterogeneity rather than a per-category law.

HIDE_HIGH has higher accuracy at both Q4 and Q8, while HIDE_LOW gains more from the increment. The opposite endpoint and gain rankings distinguish final performance from the return to an upgrade; they do not show that selecting HIDE_HIGH for final accuracy is a worse decision. Under the tested frozen ImageNet feature geometry and matched candidate-supervision family, hiding the DESIGN-selected HIGH pair yields the predeclared smaller mean Q4-to-Q8 marginal gain than hiding LOW. The fine-minus-coarse contrast limits an account based on a common shift across task resolutions, while fixed-high persistence shows that the interaction does not require refitting the reader at each evaluation rate.

The intervention supports the predeclared mean ordering when the hidden distinction changes under the tested geometry and supervision family. It does not uniquely identify the internal mechanism producing that ordering. Feature geometry, finite-sample difficulty, and optimization can still mediate how a hidden distinction affects the fitted reader. Section 5 therefore separates the supported structural diagnosis from stronger mechanistic interpretations.

Table 2: **ImageNet held-out TEST: endpoint means and predeclared gain contrasts.** Part (a) reports descriptive Q4/Q8 endpoint means and marginal gains. Part (b) reports paired parent-level HIDE_LOW-minus-HIDE_HIGH gain contrasts after averaging three fitting seeds within parent. The two primary contrasts and the two fixed-reader contrasts are corrected in separate Holm families; the coarse interaction is descriptive. Accuracy is percent; gains, effects, and intervals are percentage points.

(a) Rate-specific endpoint means and gains (descriptive)

Channel	Fine Q4	Fine Q8	Fine gain	Coarse gain
HIDE_HIGH	52.7451	54.2686	1.5235	1.6980
HIDE_LOW	50.1402	52.3657	2.2255	1.6480

(b) Paired HIDE_LOW minus HIDE_HIGH gain contrasts

Contrast	Effect (pp)	95% interval	Holm p	Role / status
Fine gain (B_{ALIGN})	$+0.7020$	$[0.0882, 1.3451]$	0.02977	Primary / PASS
Fine $-$ coarse (C_{ALIGN})	$+0.7520$	$[0.1020, 1.4265]$	0.02977	Primary / PASS
Fixed high ($K_{H,\mathrm{ALIGN}}$)	$+0.4314$	$[0.0441, 0.8255]$	0.03461	Fixed / PASS
Fixed low ($K_{L,\mathrm{ALIGN}}$)	$+0.2392$	$[-0.0882, 0.5588]$	0.07881	Fixed / unresolved
Coarse interaction	-0.0500	$[-0.2569, 0.1578]$	—	Descriptive

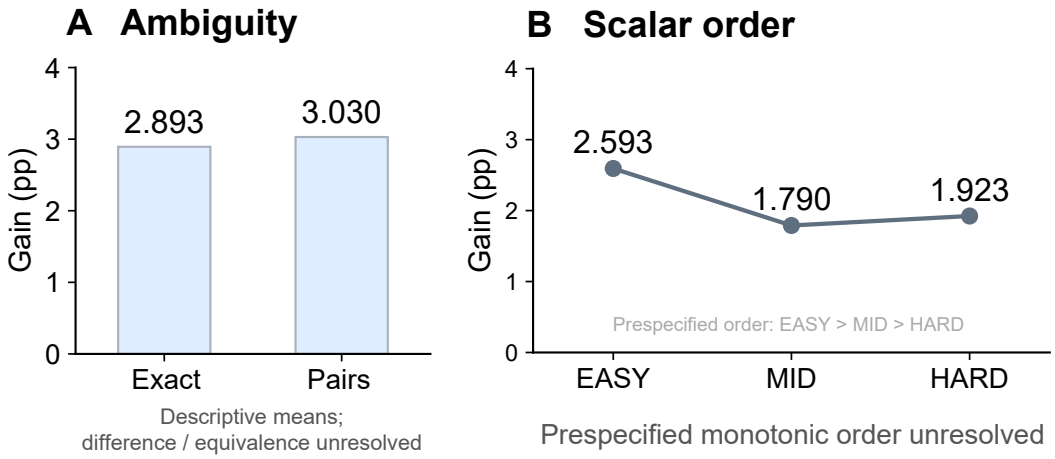

Figure 4: **Boundary calibration: ambiguity and scalar order.** (a) True-containing CIFAR pairs retain positive marginal gain, but neither the exact-minus-pair difference nor equivalence is established. (b) EASY/MID/HARD fine gains violate the prespecified monotonic order at MID-to-HARD, leaving the full scalar-order hypothesis unresolved. Panels show descriptive means over 20 parents after averaging fitting seeds.

5 Boundaries and Scope

Three boundaries limit simpler accounts. True-containing ambiguity can retain substantial marginal gain; the stronger Top-1-specific EEG hypothesis is unsupported; and the predeclared EASY/MID/HARD study does not establish a monotonic scalar-recoverability ordering (Figure 4). Appendix C gives the supporting protocols and unresolved contrasts behind these ambiguity and scalar-order boundaries. These findings do not refute scalar accounts in general or eliminate feature-relative difficulty. They motivate the narrower intervention on which distinction supervision hides at a fixed channel spectrum.

The ImageNet result characterizes supervision-conditioned marginal value for a fixed representation increment. HIDE_HIGH and HIDE_LOW are isospectral and share the training objective; their matched channel summaries give the same description, yet their gains differ. The contrast concerns the channel's relation to the fixed feature geometry, which these summaries do not specify.

Matching this summary does not equate feature-relative difficulty or isolate a unique internal mechanism. Feature geometry, finite-sample difficulty, and optimization can still mediate how that orientation affects the fitted reader. In particular, selected HIGH pairs start at lower DESIGN Q4 accuracy than LOW pairs; this difficulty difference is not removed by the fine-minus-coarse or fixed-high contrasts. Global Q4 accuracy is also lower for HIDE_LOW, leaving endpoint headroom as a possible explanation. These unmatched properties limit the interpretation of how the orientations produce different marginal gains; the comparison does not isolate the incremental-discrimination score's contribution beyond baseline difficulty. Constructing the intervention uses 34,000 clean-labeled DESIGN images and 2,720 ridge solves.

Reader-state boundary. Fixed-reader evaluation measures the response to changing input rate while holding each reader's weights constant. In both EEG and ImageNet, the interaction persists for readers trained at the high rate, so fitting a new model at each evaluation rate is unnecessary for the observed effect. The low-rate-trained contrasts remain unresolved in both central settings (Figures 2b and 3b). A low-trained reader may not have learned weights that exploit distinctions exposed at the higher rate, but this is a possible explanation, not an identified mechanism. Different significance statuses do not establish reader-state modulation or prove high-rate training necessary.

Scope and measurement consequence. EEG discovery and the ImageNet intervention provide complementary evidence across two scoped settings. EEG establishes the phenomenon under competent learning relations and examines its fixed-reader decomposition; ImageNet tests a predeclared contrast under matched channels. Their physical rate coordinates, tasks, and inferential units differ and are never pooled (Table 1); the shared object is the paired low-to-high marginal gain. ImageNet evidence is conditional on frozen pretrained features, one hierarchy, and one Q4-to-Q8 increment; EEG uses heavily averaged recordings and one principal 800-to-1600 nominal-bit/s increment. Independent EEG-acquisition confirmation remains limited: NOD supervision experiments did not establish jointly competent endpoints, and the stronger adapted-reader qualification failed its measurement criterion (Zhang et al., 2025; Wu et al., 2026). These settings serve distinct evidence roles rather than interchangeable replications.

A rate–utility curve should be reported with its supervision relation and reader state. The ImageNet arms differ in marginal gain despite sharing the stated scalar channel properties. The endpoint reversal illustrates a different reporting question: HIDE_HIGH has the better final accuracy, whereas HIDE_LOW gains more from the upgrade. A larger gain can matter when assigning equal-cost upgrades to already fixed systems, but this study evaluates no such allocation policy. It establishes neither optimal-rate selection nor a universal scaling rule. A complete coarse-to-fine-to-instance hierarchy also remains untested.

6 Conclusion

We study the supervision-conditioned marginal value of a fixed representation increment. In the tested ImageNet system, two channels matched in ambiguity, rank, spectrum, objective, and exposure yield different held-out marginal gains when they hide different DESIGN-selected distinctions. The predeclared fine gain contrast exceeds the coarse-task contrast and persists with high-rate-trained readers held fixed; fixed-low persistence remains unresolved. EEG correspondence controls provide supporting discovery, not independent replication of the same orientation test.

Matching scalar channel summaries does not equate the measured return on this increment under the tested pretrained geometry and reader system. Endpoint accuracy and marginal return also rank the two arms differently. Feature-relative difficulty, headroom, and optimization remain possible explanations within the compared learning relations. The evidence supports reporting the supervision relation and reader state with a rate–utility measurement; it establishes neither the DESIGN score's independent value nor an optimal allocation policy or universal scaling law.

REPRODUCIBILITY STATEMENT

The appendices specify splits, rate coordinates, reader components, training budgets, statistical units and principal unresolved outcomes. The accompanying anonymous artifact supplies the exact 68-parent hierarchy, HIGH/LOW map and DESIGN counts, fixed candidate assignments, configurations, saved TEST counts, and scripts for statistical verification and export of the supplied figure layouts. Its manifest binds the files by hash. Verification from saved artifacts is supported; full acquisition and training replay is outside this package's scope (Appendix D).

AI USE STATEMENT

Generative AI tools assisted with conceptual framing and hypothesis refinement, mathematical formulation and algebraic derivations, experimental design, method implementation and data preprocessing, code debugging and testing, and interpretation of results. They also assisted with literature search and checking, manuscript drafting and editing, figure preparation, and internal critical review. The authors directed the research, reviewed the AI-assisted contributions, made the final scientific decisions, and take full responsibility for the manuscript and supporting artifacts.

REFERENCES

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=HyxQzBceg.

Alljoined. Alljoined-1.6M dataset card, 2025. URL https://huggingface.co/datasets/Alljoined/Alljoined-1.6M/blob/main/README.md. Accessed September 26, 2026.

Johannes Ballé, Valero Laparra, and Eero P. Simoncelli. End-to-end optimized image compression. In *International Conference on Learning Representations*, 2017. URL https://www.cns.nyu.edu/pub/lcv/balle17a-final.pdf.

Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rkcQFMZRb.

C. F. Barnes, S. A. Rizvi, and N. M. Nasrabadi. Advances in residual vector quantization: A review. *IEEE Transactions on Image Processing*, 5(2):226–262, 1996. doi: 10.1109/83.480761. URL https://doi.org/10.1109/83.480761.

Timothee Cour, Ben Sapp, and Ben Taskar. Learning from partial labels. *Journal of Machine Learning Research*, 12(42):1501–1536, 2011. URL https://jmlr.org/papers/v12/cour11a.html.

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

Anderson de Andrade and Ivan V. Bajić. Towards task-compatible compressible representations. In *IEEE International Conference on Multimedia and Expo Workshops*, pp. 1–6, 2024. doi: 10.1109/ICMEW63481.2024.10645459. URL https://arxiv.org/abs/2405.10244.

Yann Dubois, Benjamin Bloem-Reddy, Karen Ullrich, and Chris J. Maddison. Lossy compression for lossless prediction. In *Advances in Neural Information Processing Systems*, volume 34, pp. 14014–14028, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/7535bbb91c8fde347ad861f293126633-Abstract.html.

B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979. doi: 10.1214/aos/1176344552. URL https://doi.org/10.1214/aos/1176344552.

Lei Feng, Jiaqi Lv, Bo Han, Miao Xu, Gang Niu, Xin Geng, Bo An, and Masashi Sugiyama. Provably consistent partial-label learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 10948–10960, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/7bd28f15a49d5e5848d6ec70e584e625-Abstract.html.

Sha Guo, Jing Chen, Zixuan Hu, Zhuo Chen, Wenhan Yang, Yu Lin, Xing Jiang, and Ling-Yu Duan. Which tasks should be compressed together? a causal discovery approach for efficient multi-task representation compression. In *International Conference on Learning Representations*, pp. 15264–15287, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/27a61686fc5928a43c8211f7da70ac75-Abstract-Conference.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.

Shuo He, Chaojie Wang, Guowu Yang, and Lei Feng. Candidate label set pruning: A data-centric perspective for deep partial-label learning. In *International Conference on Learning Representations*, pp. 27189–27206, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/72fb9ab442fc60b7ae5d53bf6b478273-Abstract-Conference.html.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

Zhenyu Huang, Guocheng Niu, Xiao Liu, Wenbiao Ding, Xinyan Xiao, Hua Wu, and Xi Peng. Learning with noisy correspondence for cross-modal matching. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29406–29419, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/f5e62af885293cf4d511ceef31e61c80-Abstract.html.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4904–4916, 2021. URL https://proceedings.mlr.press/v139/jia21b.html.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 18661–18673, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/d89a66c7c80a29b1bdbab0f2a1a94af8-Abstract.html.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi. Matryoshka representation learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 30233–30249, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/c32319f4868da7613d78af9993100e42-Abstract-Conference.html.

Jiaqi Lv, Miao Xu, Lei Feng, Gang Niu, Xin Geng, and Masashi Sugiyama. Progressive identification of true labels for partial-label learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6500–6510, 2020. URL https://proceedings.mlr.press/v119/lv20a.html.

Yoshitomo Matsubara, Ruihan Yang, Marco Levorato, and Stephan Mandt. Supervised compression for resource-constrained edge computing systems. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2685–2695, 2022. URL https://openaccess.thecvf.com/content/WACV2022/html/Matsubara_Supervised_Compression_for_Resource-Constrained_Edge_Computing_Systems_WACV_2022_paper.html.

George A. Miller. WordNet: A lexical database for english. *Communications of the ACM*, 38(11): 39–41, 1995. doi: 10.1145/219717.219748. URL https://doi.org/10.1145/219717.219748.

Nam Nguyen, Thinh Nguyen, and Bella Bose. Cross-domain lossy compression via rate- and classification-constrained optimal transport. In *International Conference on Learning Representations*, pp. 57877–57926, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/5f0e54e34ba4c30fbc6f229986dbffb4-Abstract-Conference.html.

Valentina Njaradi, Clémentine Dominé, Rachel Swanson, Marco Mondelli, and Andrew Saxe. Optimal representation size: High-dimensional analysis of pretraining and linear probing. *arXiv preprint arXiv:2605.20105v1*, 2026. URL https://arxiv.org/abs/2605.20105v1.

Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1944–1952, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Patrini_Making_Deep_Neural_CVPR_2017_paper.html.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115 (3):211–252, 2015. doi: 10.1007/s11263-015-0816-y. URL https://link.springer.com/article/10.1007/s11263-015-0816-y.

Shiyu Tian, Hongxin Wei, Yiqun Wang, and Lei Feng. CroSel: Cross selection of confident pseudo labels for partial-label learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19479–19488, 2024. doi: 10.1109/CVPR52733.2024.01842. URL https://openaccess.thecvf.com/content/CVPR2024/html/Tian_CroSel_Cross_Selection_of_Confident_Pseudo_Labels_for_Partial-Label_Learning_CVPR_2024_paper.html.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. In *Proceedings of the 37th Annual Allerton Conference on Communication, Control, and Computing*, pp. 368–377, 1999. URL https://arxiv.org/abs/physics/0004057.

Haobo Wang, Ruixuan Xiao, Yixuan Li, Lei Feng, Gang Niu, Gang Chen, and Junbo Zhao. PiCO: Contrastive label disambiguation for partial label learning. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2201.08984v1.

Lukun Wu, Jie Li, Ziqi Ren, Kaifan Zhang, and Xinbo Gao. Shrinking the teacher: An adaptive teaching paradigm for asymmetric EEG-vision alignment. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(21):17859–17867, 2026. doi: 10.1609/aaai.v40i21.38844. URL https://ojs.aaai.org/index.php/AAAI/article/view/38844.

Shiyu Xia, Jiaqi Lv, Ning Xu, Gang Niu, and Xin Geng. Towards effective visual representations for partial-label learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15589–15598, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Xia_Towards_Effective_Visual_Representations_for_Partial-Label_Learning_CVPR_2023_paper.html.

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.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. A theory of usable information under computational constraints. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=r1eBeyHFDH.

Yihao Xue, Kyle Whitecross, and Baharan Mirzasoleiman. Investigating the impact of model width and density on generalization in presence of label noise. In *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, volume 244 of *Proceedings of Machine Learning Research*, pp. 3912–3935, 2024. URL https://proceedings.mlr.press/v244/xue24a.html.

Guohao Zhang, Ming Zhou, Shuyi Zhen, Shaohua Tang, Zheng Li, and Zonglei Zhen. A large-scale MEG and EEG dataset for object recognition in naturalistic scenes. *Scientific Data*, 12:857, 2025. doi: 10.1038/s41597-025-05174-7. URL https://www.nature.com/articles/s41597-025-05174-7.

A IMAGENET INTERVENTION AND CONFIRMATION

This appendix specifies the intervention, matching conditions and inference behind Section 4 and Table 2. The test concerns one fixed representation increment under two supervision relations; it is conditional on the pretrained geometry and shared reader system, not an end-to-end ImageNet benchmark.

A.1 DATA ROLES AND REPRESENTATION INSTRUMENT

An outcome-independent hierarchy rule selects 68 disjoint five-class parents from winter-2021 ImageNet, giving 340 classes. Each class contributes 134/100/100 FIT/DESIGN/TEST images: 45,560/34,000/34,000 in total. Classes are shared across roles; the split is constructed for this experiment, not the official ILSVRC validation split. Known previously used image IDs and exact encoded-byte duplicates are excluded before allocation. Roles share neither IDs nor encoded-byte hashes. Near duplicates, re-encoded duplicates, unknown historical exposure and backbone-pretraining exposure are not excluded. FIT determines normalization, the codec and reader fits; DESIGN constructs the intervention; TEST evaluates locked models. No TEST outcome selects classes, rates, direction, hyperparameters, checkpoints or seeds.

Frozen torchvision ResNet18 `IMAGENET1K_V1` supplies 512-dimensional average-pooled features. Features are row-normalized, then centered and scaled using FIT statistics. FIT trains one eight-stage RVQ with 256-entry codebooks, a retained zero entry and 20 Lloyd iterations. Q4 and Q8 share their first four stages exactly; reconstructed vectors are normalized before fitting the reader. Finite outputs, prefix identity, nonzero added-stage contribution and improved reconstruction qualify the instrument. Activity and entropy are diagnostics, not outcome-selective effective-code-count gates. The increment is 32 to 64 nominal bits per vector, excluding entropy coding and model overhead; both inputs retain dimension 512.

A.2 DESIGN SELECTION AND PROSPECTIVE LOCKS

Each parent's ten unordered class pairs are measured at Q4 and Q8 using balanced two-fold cross-fitted binary ridge probes. The ridge constant is one, the intercept is unpenalized, and solves use FP64. Each probe trains on one DESIGN fold and evaluates on the other, giving 2,720 solves across parents, pairs, rates and folds. TEST is not used.

The Q8-minus-Q4 difference in integer correct counts defines added accessibility. Its within-parent maximum and minimum select HIGH and LOW; a frozen hash resolves ties. No parent is dropped, and the realized map has no coincident HIGH/LOW selections. The verification package supplies all 680 pair records, counts, denominators and tie hashes, plus the 68-row map. Selected HIGH pairs have mean DESIGN Q4/Q8 accuracies 80.4853/85.3015% (gain 4.8162 pp); LOW pairs have 91.0441/90.3971% (gain -0.6471 pp). These are descriptive selected-sample means, affected by extreme selection and finite-sample noise. They expose a baseline-difficulty difference. No comparison against a baseline-difficulty or margin-based selector was conducted; the gain score's independent value remains untested.

The sequence is FIT representation $\rightarrow$ DESIGN map $\rightarrow$ map/channel lock $\rightarrow$ all 12 reader fits and final-checkpoint locks $\rightarrow$ TEST decoding and scoring. The directional hypothesis and confirmation rule are fixed before TEST. Binary DESIGN probes construct the channels; the global 340-way readers use a different objective and endpoint. Pair selection therefore does not algebraically fix the global TEST gain ordering.

A.3 MATCHED CHANNEL AND OBSERVATION LIKELIHOOD

Number the five classes so that 1 and 2 form the hidden pair. With candidate rows $\{1,2\},\{3,4\},\{4,5\},\{3,5\}$, the channel is

$$Q_{S,y} = \Pr(S \mid y) = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \end{pmatrix}. \tag{8}$$

Every column sums to one. The hidden classes always produce their shared pair; each remaining class produces either incident triangle edge with probability one half. HIDE_HIGH and HIDE_LOW permute this construction so that the selected HIGH or LOW distinction is hidden. The pair block has singular values $\sqrt{2}, 0$ and the triangle block has $1, 1/2, 1/2$: both channels have rank four and the same spectrum over five label coordinates. Under balanced priors, the two labels in every candidate set are equally likely, giving one bit of conditional label uncertainty. Candidate size, true-label inclusion, ambiguity, rank and spectrum are therefore matched.

Both readers minimize

$$\ell(x, S) = -\log \sum_{y \in S} p_\theta(y \mid Z(x)). \tag{9}$$

Within an observed pair, the two nonzero channel entries are equal: one for the hidden pair and one half for a triangle edge. Its forward likelihood is thus the same probability sum times a parameter-independent constant. Equation 9 is the corresponding negative log-likelihood up to that constant. This matches the observation model and loss; it does not match feature-relative difficulty or isolate a unique internal mechanism. Forward loss correction also composes predictions with an observation channel (Patrini et al., 2017); here the observations are candidate sets and the channel has a deliberately hidden distinction.

A.4 Matched fitting and fixed-reader evaluation

Each classifier is a linear 512-to-340 map with bias (174,420 parameters). Two arms, two training rates and seeds 31901/31902/31903 give 12 fits, each using 30 epochs, 67 batches per epoch, batch size 680 and 2,010 updates. AdamW uses learning rate 10^{-3}, weight decay 0.01, temperature one and gradient clipping at five. Initialization, image order and exposure are paired across arms and rates; the frozen FP32 recipe disables TF32.

Candidate assignments are fixed per image. Each minibatch has two true-label and two sibling occurrences per class, matching exposure between arms. Every fit uses its final checkpoint, without early stopping or best-seed selection. Diagonal and crossed evaluation produce 24 TEST banks. Crossing changes only the reconstructed input; each arm retains its own fitted weights and bias. Fixed-reader persistence is consequently not a comparison under common weights across arms.

A.5 Competence and statistical decision rule

Fine accuracy is global 340-way classification; coarse accuracy sums fine probabilities within each parent before choosing among 68 parents. Conditional five-way accuracy uses the true parent only for competence checks. For every seed and required diagonal/crossed state, fine/coarse accuracy must reach $10/340$ and $4/68$, and conditional accuracy must reach 0.30. Parent-bootstrap lower bounds must exceed chance; at least 55/68 parents must reach conditional accuracy 0.25, with a noncollapse check. All eight states pass, with 68/68 parent coverage. These checks do not select or refit models.

Each parent contains 500 TEST images. Three fitting seeds are averaged before inference, giving an integer denominator of 1,500 decisions per parent. Intervals use 50,000 paired parent-bootstrap draws; exact integer sign-flip tails retain tied statistics. $B_{\mathrm{ALIGN}}, C_{\mathrm{ALIGN}}$ form the PRIMARY Holm family (Holm, 1979) and the two fixed-reader contrasts form a separate FIXED family. The bootstrap assumes approximately independent, exchangeable parent effects within the shared system. Sign-flip tails assume joint invariance under independent parent sign flips; marginal symmetry alone is insufficient. Neither assumption follows from randomized channel assignment. Shared readers, backbone and taxonomy prevent interpreting 34,000 images, 204 parent/seed combinations or 68 groups as independent system replications.

Confirmation requires both primaries to have means at least 0.5 pp, positive pointwise 95% lower bounds and PRIMARY Holm $p \leq .05$, plus at least one competent fixed-reader contrast with positive mean and lower bound and FIXED Holm $p \leq .05$. Both primaries and fixed-high pass; fixed-low remains unresolved (Table 2). The rule does not require both fixed contrasts to pass, and a 0.5-pp point estimate is not a 0.5-pp confidence lower bound. Both primary directions are positive in 42/68 groups, not in every group.

B EEG DISCOVERY AND READER-STATE DEPENDENCE

These controls support Section 3. They use the same eight Alljoined participants and source objects, so follow-up fits do not constitute independent EEG replications. Their bit/s coordinate is not pooled with the ImageNet bits/vector coordinate.

B.1 INPUTS, SOURCE OBJECTS AND MATCHED FITTING

The target participants are sub-01, sub-02, sub-04, sub-09, sub-12, sub-15, sub-19 and sub-20 (Xu et al., 2025). Training has 15,361 images from 1,654 official concepts with four repeats per image; the disjoint test concepts provide a 144-image gallery with 80 repeats per image. One image per test concept makes this a concept-discrimination endpoint, not a within-concept identity test. Correct target labels are used for training concepts only; this is not source-only or zero-target deployment.

Alljoined inputs have 32 channels and 250 samples. A frozen temporal codec (hidden width 192, latent width 128) reconstructs waveforms for the reader. Its 25-Hz latent sequence and 256-entry RVQ codebooks give Q4/Q8 coordinates 800/1600 nominal bit/s; Q8 extends the same four-stage prefix. These exclude entropy coding and model overhead. The source normalizer is frozen. Test trials are reconstructed separately before averaging their 80 repeats; target-training reconstruction is trialwise. Source training instead uses four-repeat averages before augmentation. Distortion is measured on source-standardized waveforms before test averaging.

Twelve distinct source participants supply the codec, normalization and rate-specific reader initializers. The reader combines an EEG encoder, learned visual projection and layer router: five InternViT layers, visual width $3200 \to 1024$, EEG width 1024 and alignment width 512. The qualified SCORE-derived source recipe (Cui et al., 2026) uses multi-positive fitting and source recovery over 50 epochs. Target fits start from these source checkpoints, not target-TEST-selected checkpoints.

Each target fit uses 2,080 updates, AdamW learning rate and weight decay 10^{-4}, gradient clipping one, and MMD weight 0.1 with bandwidths $[0.1, 0.2, 0.5, 1, 2]$. The inherited symmetric contrastive target objective and cosine evaluation are fixed across comparisons. Each step has 512 target rows and 96 source forward rows. Source-task gradients have zero weight; matched source forwards retain normalization and random-number execution. Final checkpoints, samples and initialization schedules are paired; there is no target normalization adjustment, test adaptation or test-informed selection.

B.2 COMPETENCE, INTERACTION AND SPECIFICITY

Each batch draws 256 concepts and two distinct images per concept. Correct supervision A uses the true image; S swaps the pair, preserving concept and marginal visual exposure. The stepwise pair schedule is shared across arms and rates, and reverse loss uses its inverse. Seeds 31001/31002/31003 are averaged within each of the eight participant units.

Top-1/Top-5 chance levels are $0.694/3.472\%$. All required endpoints must have above-chance participant-bootstrap intervals and shared-gallery permutation $p \leq .05$. There are 50,000 paired bootstrap draws, 10,000 joint gallery permutations and all 2^8 sign assignments; boundary ties are misses. All A/S endpoints qualify. Their absolute values, including all crossed cells, appear once in Table 5. Intervals are pointwise and conditional on shared source objects; paired participant resampling and sign exchangeability are working assumptions, not guarantees from randomized treatment assignment.

Let J_1, J_5 be A-minus-S rate-gain interactions at Top-1/Top-5 and $H = J_1 - J_5$. Here H is a contrast, not the high-rate coordinate. The primary Holm family is $\{J_1, H\}$; each positive test requires mean at least 0.5 pp, a positive 95% lower bound and adjusted one-sided $p \leq .05$. The stronger specificity hypothesis also requires the full J_5 interval inside $[-1, 1]$ pp and positive lower bounds for both A/S Top-5 gains. It fails through both H and the equivalence condition; only the narrower competent Top-1 interaction is supported.

Top-1 and Top-5 threshold the same score matrix; their contrast does not test different semantic resolutions. The failed specificity hypothesis therefore does not substitute for the ImageNet fine/coarse test. Population competence also does not imply competence or equal accuracy for every participant.

Table 3: EEG interaction and the unsupported specificity hypothesis (pp). Only J_1, H form the Holm family. Failure of interval containment does not establish equivalence.

Contrast	Mean	95% interval	p_{Holm}	Registered role
J_1	$+2.95$	$[+1.19, +4.66]$	0.03125	Primary: passes
$H = J_1 - J_5$	-2.00	$[-5.12, +0.81]$	0.859375	Primary: fails
J_5	$+4.95$	$[+1.62, +8.13]$	—	± 1 pp equivalence fails

Table 4: All eight participant effects (pp), averaging three seeds. $\tau_{A,1}, \tau_{S,1}$ are the Top-1 gains g_A, g_S. No participant is excluded by effect direction; J_1 is positive in seven of eight.

Subject	$\tau_{A,1}$	$\tau_{S,1}$	J_1	J_5	H
sub-01	$+8.102$	$+3.704$	$+4.398$	$+4.861$	-0.463
sub-02	$+11.574$	$+6.250$	$+5.324$	$+2.778$	$+2.546$
sub-04	$+6.481$	$+0.463$	$+6.019$	$+5.556$	$+0.463$
sub-09	-0.231	$+0.926$	-1.157	-1.389	$+0.231$
sub-12	$+5.787$	$+4.861$	$+0.926$	$+9.028$	-8.102
sub-15	$+4.630$	$+2.315$	$+2.315$	$+11.343$	-9.028
sub-19	$+1.157$	$+0.694$	$+0.463$	-2.315	$+2.778$
sub-20	$+9.722$	$+4.398$	$+5.324$	$+9.722$	-4.398

B.3 Fixed-reader crossing

Each existing reader is evaluated at both input rates with its EEG encoder, visual projection and router fixed. The same participants, seeds, source normalizer and reconstruction-before-averaging rule are used; all original diagonal scores reproduce bitwise. The complete cells and contrast intervals support the accounting in Eq. 5.

Table 5: Complete EEG crossing, in percent. Coordinates index reader-training and evaluation-input rate; $L/H = 800/1600$ nominal bit/s. All Top-1 and Top-5 population cells meet competence. Diagonal cells also give the discovery endpoints.

Pairing	Metric	$U(L,L)$	$U(L,H)$	$U(H,L)$	$U(H,H)$
A	Top-1	5.32	7.15	6.19	11.23
A	Top-5	17.01	24.57	21.35	31.66
S	Top-1	3.47	4.77	4.02	6.42
S	Top-5	13.45	19.07	15.74	23.15

Table 6 uses $K_t = K_{t;A,S}$ and $T_e = T_{e;A,S}$ from Section 2. Saved $D_{t,r}, T_{e,r}$ reverse the paper's subscript order $D_{r,t}, T_{r,e}$. The same bootstrap, permutation and sign enumeration as above applies. Each fixed-reader primary requires mean at least 0.5 pp, a positive 95% lower bound and Holm $p \leq .05$; the stronger conjunction requires both to pass. K_H passes and is positive for all eight participants; K_L is unresolved and positive for five. Thus the stronger conjunction remains unresolved. The secondary $K_H - K_L$ interval does not replace this unmet primary conjunction. Differing significance statuses alone do not establish reader-state modulation or necessity of high-rate training. The T terms combine source initialization, target fitting and visual-head changes; the exact accounting is not causal mediation.

Table 6: Fixed-reader interactions and reader-training contrasts. Only K_L, K_H form the primary Holm family; all intervals are pointwise.

Top-1 contrast	Mean (pp)	95% interval	p_{Holm}	Role
K_L	$+0.52$	$[-0.55, +1.65]$	0.22656	Primary: unresolved
K_H	$+2.63$	$[+1.48, +3.76]$	0.00781	Primary: passes
T_L	$+0.32$	$[-0.90, +1.39]$	—	Secondary
T_H	$+2.43$	$[+0.90, +4.05]$	—	Secondary
$K_H - K_L$	$+2.11$	$[+0.87, +3.36]$	—	Secondary

C AMBIGUOUS SUPERVISION AND SCALAR-CHANNEL LIMITS

These studies support the boundaries in Section 5: ambiguity need not remove marginal gain, while equivalence to exact labels and a monotonic scalar-recoverability ordering remain unestablished.

C.1 EEG AMBIGUITY AND UNRESOLVED EQUIVALENCE

For contrastive logits a, b of the true image and its concept mate, denominator Z, and $\delta = a - b$, the executed set-positive objective obeys

$$L_{\text{MASS}} = \log Z - \log(e^a + e^b) + \log 2, \tag{10}$$

$$E = \log\cosh(\delta/2), \qquad M = \delta/2,$$

$$L_A = L_{\text{MASS}} + E - M, \qquad L_S = L_{\text{MASS}} + E + M. \tag{11}$$

Losses average both directions and rows. The constant $\log 2$ has no gradient. MASS permits unequal within-pair probability; equal-weight log-probability aggregation instead gives $L_{\text{MASS}} + E$. These are established aggregation choices (Khosla et al., 2020). The identities hold at the same model state, not as a causal decomposition of separately trained readers.

MASS reuses the EEG participants, codec, source initializers, samples and 2,080-update recipe in Appendix B. It uses the preselected single seed 31001, not the three-seed discovery average. New fits start from source checkpoints and lock before TEST; the same-seed A/S anchors reproduce bitwise. All required A/S/MASS Top-1/Top-5 diagonal cells pass competence. Inference uses eight participants, 50,000 paired bootstrap draws, 10,000 gallery permutations and all 2^8 sign assignments.

Table 7: MASS conjunctive primary (pp). The two directional conditions pass; the interval for $g_{\text{MASS}} - g_A$ is not contained in the equivalence band.

Primary Top-1 contrast	Mean (pp)	95% interval	One-sided p	Condition
g_{MASS}	$+7.29$	$[+3.65, +11.02]$	0.01172	Met
$g_{\text{MASS}} - g_S$	$+3.73$	$[+0.87, +6.77]$	0.03125	Met
$g_{\text{MASS}} - g_A$	$+1.22$	$[-0.87, +3.73]$	—	Unresolved

The conjunction requires g_{MASS} and $g_{\text{MASS}} - g_S$ each to have mean at least one pp, a positive 95% lower bound and one-sided $p \leq .05$, and the full $g_{\text{MASS}} - g_A$ interval inside $[-1, 1]$ pp. Table 7 supports retained value under ambiguity but leaves this conjunction unresolved. It establishes neither equivalence nor inequivalence to A, despite the larger point estimate.

C.2 CIFAR AMBIGUITY VERSUS RECOVERABILITY

Both CIFAR-100 studies use frozen pretrained ResNet18 features, FIT-only RVQ, and a 32-to-64-nominal-bit/vector Q4/Q8 increment. They are development evidence on a previously used test set, not independent confirmation. Three seeds are averaged within each of 20 five-class parent groups. Fine accuracy is global 100-way classification; coarse accuracy sums fine probabilities within a parent.

With matched fitting budgets, exact labels and true-containing pairs give fine gains 2.8933 and 3.0300 pp. The exact-minus-pair contrast is -0.1367 pp (95% interval $[-0.5833, 0.3267]$), and

the fine-minus-coarse contrast is -0.3733 pp ($[-0.9733, 0.2567]$). Neither the proposed difference nor equivalence is established. The balanced pair channel can select any of four siblings and has full column rank. Under its class-conditional assumptions, two-label ambiguity per observation therefore need not imply global class nonidentifiability; this does not prove that the fitted reader recovered every distinction.

C.3 Unresolved scalar ordering

EASY/MID/HARD mix five-cycle and disjoint pair-plus-triangle channels with their inverses, at frozen mixing weights $112/225$, $22/225$ and zero. The ranks are $5/5/4$ and smallest singular values approximately $0.217804/0.048655/0$. Candidate size, probability-mass loss and one-bit conditional uncertainty are shared; class order is fixed before outcomes. Pair frequencies and optimization curvature change, so this does not isolate a causal effect of one singular value.

Fine gains are $2.5933/1.7900/1.9233$ pp. The predeclared hypothesis requires both pooled EASY–MID–HARD ordering and paired support for a score comparing all three within-parent gains. MID/HARD violates the pooled order, and the three-level trend remains unresolved. Positive EASY–HARD contrasts do not replace this conjunction. Both CIFAR studies retain their inconclusive outcomes; they neither establish equivalence nor refute scalar accounts in general. This is the boundary motivating the fixed-spectrum intervention in Appendix A.

D External EEG scope and reproducibility

The NOD studies constrain the external-EEG claim in Section 5. Failure to qualify their endpoints is not evidence against a supervision interaction within a competent task.

D.1 Supervision comparison on NOD

NOD (Zhang et al., 2025) uses twelve source and seven target participants, with a fixed 800-TRAIN/200-HELD concept split. Each target has 200 single-trial queries and one gallery image per HELD concept. Inputs have 62 channels, 120 samples over 0–480 ms, source-only normalization and a frozen 768-dimensional OpenCLIP layer. This differs from Alljoined's visual interface and 80-repeat averaging.

Target arms A (true singleton), S (same-concept other singleton) and MASS (true-containing pair) share within-rate source initialization, samples and updates; source supervision is exact. All fits and exposure identities lock before HELD scoring. Source-recovery opportunities differ across rates, so equal sample/update counts do not isolate usable auxiliary supervision.

Out of 1,400 queries, A has 9/9 low/high hits, S 9/5 and MASS 11/17. Only high-rate MASS passes the competence rule requiring an above-chance participant-bootstrap lower bound and shared-gallery permutation test. The seven participants, not pooled queries, are the inference units. Joint endpoint competence and the structural conjunction remain unresolved; this is not a qualified external replication.

D.2 Reader qualification and its limit

A separate study partitions the original 800 TRAIN concepts into 600 FIT and 200 DEV concepts, leaving HELD unused. DEV is excluded from all new normalization, codec and reader fits. The same seven target participants are reused. Existing and ATS-adapted readers receive RAW/800/1600 inputs under correct supervision only. All 49 fits finish before the 42 DEV score banks are evaluated, using final checkpoints. The ATS adaptation retains its temporal encoder, residual-free adapter and bidirectional contrastive core (Wu et al., 2026), adapting the input shape, visual interface and source-to-target fitting to NOD. This compares reader systems, not isolated encoders or ATS benchmark accuracy.

Table 8: NOD qualification: all six endpoints and original 95% participant intervals. Chance is 0.5%; coverage counts participants at 1.5% accuracy. None meets the full absolute/coverage rule.

Reader	Input	Hits/1400	Top-1 (%)	95% interval	Coverage
Existing	RAW	17	1.214	[0.857, 1.571]	3/7
Existing	800	9	0.643	[0.357, 1.000]	1/7
Existing	1600	17	1.214	[0.643, 1.786]	4/7
ATS-adapted	RAW	17	1.214	[0.643, 1.786]	3/7
ATS-adapted	800	12	0.857	[0.429, 1.357]	2/7
ATS-adapted	1600	13	0.929	[0.643, 1.214]	2/7

Each endpoint must reach mean accuracy 3%, a lower bound above 0.5%, permutation $p \leq .05$, and at least five of seven participants at 1.5%. The rate-gain interval must have half-width at most two pp. Both systems pass precision but all six endpoints fail absolute/coverage qualification (Table 8). RAW permutation p is 0.000700 for the existing reader and 0.001000 for ATS-adapted: detectable signal does not imply task qualification. No MASS/S interaction is fitted here, so qualification failure cannot replace a test of $J_{A,S}$ or $J_{M,S}$. These frozen thresholds are operational, not universal task requirements.

D.3 Data ethics and verification scope

This work analyzes existing recordings and recruits no new participants. The Alljoined publication reports informed consent and its dataset card specifies data-use terms (Xu et al., 2025; Alljoined, 2025); the NOD publication documents its acquisition (Zhang et al., 2025). These describe original acquisitions, not new institutional approval for this analysis. Closed-gallery averaged retrieval does not establish unrestricted mental decoding or clinical performance. Alljoined follow-ups reuse eight people; NOD follow-ups reuse seven. New fits and endpoints do not add participants.

The anonymous package supplies the ImageNet hierarchy, role identifiers and encoded-image hashes, DESIGN counts and HIGH/LOW map, candidate assignments, configurations, saved TEST counts and statistical verification scripts. It reproduces the primary and fixed-reader estimates, intervals and adjusted tests from saved counts, and exports the supplied figure layouts. Supporting EEG/CIFAR summaries are not full experiment replays. The manifest establishes file integrity; author-supplied lock records are not external timestamp certification. Original EEG, photographs, participant metadata and weights are excluded, and source materials retain their access terms. Broader local checks included saved-score reconstruction and bounded original-reader/query and selected feature/codec/probe replay. Neither these checks nor the release constitute complete acquisition and training replay or independent confirmation of experimental provenance.