BEYOND FULL-CONTEXT ACCURACY: CONTEXT-MARGINALIZED EVALUATION OF TEST-TIME STRUCTURE-AWARE INFERENCE IN EEG-TO-IMAGE RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Test-time structure-aware inference can improve cross-subject EEG-to-image retrieval by exploiting relations among unlabeled test queries, but it also makes predictions depend on query-set composition. Existing evaluations typically report accuracy under a single full-query context, leaving unclear whether gains and comparative conclusions remain reliable when the available context changes. We introduce Context-Marginalized Retrieval Evaluation (CMRE), which repeatedly evaluates each target query within sampled subsets of companion queries, applying each inference rule to the corresponding rows of a fixed query–gallery similarity matrix. CMRE distinguishes structural exploitability, quantified by gains over independent retrieval, from context robustness, measured by same-query Top-1 context consistency. Under leave-one-subject-out evaluation on THINGS-EEG2 across six frozen representations and six structure-aware inference rules, structure-aware inference improves full-context Top-1 accuracy by up to 20.40 percentage points over independent retrieval. However, the rule selected under full-context evaluation is no longer optimal in 17 of 24 representation–context-size conditions (70.8%). Selecting the full-context winner incurs an average loss of 2.70 percentage points and a maximum loss of 8.29 percentage points under reduced contexts. Representation rankings are likewise inference- and context-dependent: the ordering under independent retrieval is not consistently preserved under structure-aware inference. Moreover, larger retrieval gains do not necessarily imply greater context robustness, showing that retrieval effectiveness and prediction stability provide complementary views of structure-aware inference. Together, these results show that full-context accuracy represents only one operating point of context-dependent inference. CMRE complements conventional evaluation by characterizing structural gain, context consistency, and the context dependence of conclusions about inference rules and frozen EEG-to-image representations.

1 INTRODUCTION

EEG-to-image retrieval seeks to identify visually corresponding images in a candidate gallery by aligning EEG representations with visual embedding spaces. Recent advances in neural representation learning and cross-modal alignment have substantially improved retrieval performance (Song et al., 2024; Wu et al., 2025; Zhang et al., 2025; Li et al., 2024; Zheng et al., 2026; Jo et al., 2026). Despite this progress, cross-subject generalization remains challenging because of pronounced inter-subject variability in EEG signals. This challenge has motivated leave-one-subject-out (LOSO) evaluation, in which models must generalize to previously unseen subjects without subject-specific supervised retraining. Beyond representation learning, recent work has begun to exploit information from unlabeled target-subject queries at test time to improve cross-subject retrieval without labeled calibration (Huang & Zhu, 2026; Cui et al., 2026). However, once test-time inference exploits relations among multiple target queries, retrieval becomes context-dependent: the prediction for the same EEG query may change with the composition of the other queries processed alongside it.

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.

This coupling creates an evaluation problem. For a fixed gallery, independent retrieval predicts each query from its own similarity vector and is therefore unaffected by the other test queries. By contrast, inference rules based on set-level statistics or global assignment constraints couple each prediction to the remaining test queries. Their measured performance therefore reflects both the quality of the frozen representation and the particular query context used for evaluation. A single full-context operating point can consequently conceal sensitivity to query-set composition. Existing cross-subject EEG-to-image studies that exploit unlabeled target-side information at test time do not systematically examine how retrieval performance and same-query prediction stability vary with query context (Huang & Zhu, 2026; Cui et al., 2026). Related sensitivity to the composition and size of unlabeled test sets has also been reported in transductive learning (Veilleux et al., 2021; Martin et al., 2022; Yuan et al., 2023). These observations motivate two complementary questions. First, how much additional retrieval performance can structure-aware inference extract from a fixed similarity matrix? Second, how stable are the predictions for the same EEG query when the available companion-query context changes? We refer to these properties as **structural exploitability** and **context robustness**, respectively. Beyond these two properties, we further examine whether comparative conclusions about inference rules and frozen EEG-to-image representations remain stable across test-time inference conditions.

To address these questions, we introduce Context-Marginalized Retrieval Evaluation (CMRE). CMRE repeatedly evaluates each target query across sampled query contexts at predefined sizes while ensuring that every target is evaluated equally often. For each frozen representation, all inference rules operate on context-specific rows drawn from the same full query–gallery similarity matrix. CMRE reports context-marginalized retrieval performance, quantifies structural exploitability through gains over independent retrieval, and measures context robustness through the consistency of same-query Top-1 predictions across contexts. Under LOSO cross-subject protocols on THINGS-EEG2 (Gifford et al., 2022), we apply CMRE to six frozen EEG-to-image representations and six structure-aware inference rules spanning set-level and assignment-based inference, with independent retrieval serving as the context-invariant reference. Within each frozen representation, this controlled design isolates differences attributable to test-time inference by providing all inference rules with identical query–gallery similarity evidence.

Our results reveal several limitations of relying on a single full-context evaluation. Across six frozen EEG-to-image representations and six structure-aware inference rules, test-time structure improves full-context Top-1 accuracy by up to 20.40 percentage points over independent retrieval. However, the inference rule ranked first under full-context evaluation is no longer optimal in 17 of 24 representation–context-size conditions (70.8%). Selecting the full-context winner incurs an average performance loss of 2.70 percentage points and a maximum loss of 8.29 percentage points under reduced context. The apparent ranking of frozen representations is also inference- and context-dependent: the ordering observed under independent retrieval is not consistently preserved under structure-aware inference. Moreover, larger gains over independent retrieval do not necessarily coincide with greater same-query context consistency, showing that retrieval effectiveness and context robustness capture distinct properties of structure-aware inference. Together, these findings indicate that full-context accuracy represents only one operating point of context-dependent inference and can provide an incomplete basis for comparing both inference strategies and frozen EEG-to-image representations. Our main contributions are summarized as follows:

- We formalize test-time EEG-to-image retrieval under fixed query–gallery similarity scores and distinguish three inference regimes: independent retrieval, set-level inference, and assignment-based inference. This formulation enables controlled comparisons of inference rules while holding the underlying representation fixed.
- We identify **structural exploitability** and **context robustness** as two complementary properties of test-time structure-aware inference and introduce **Context-Marginalized Retrieval Evaluation (CMRE)** to quantify both. CMRE repeatedly evaluates each target query under randomly sampled query contexts of predefined sizes while balancing target-query evaluation frequency. It measures gains over independent retrieval and same-query Top-1 consistency across contexts.
- Through LOSO cross-subject evaluation across six frozen EEG-to-image representations and six structure-aware inference rules, we show that larger structural gains do not necessarily imply greater context robustness. More importantly, the inference rule ranked first

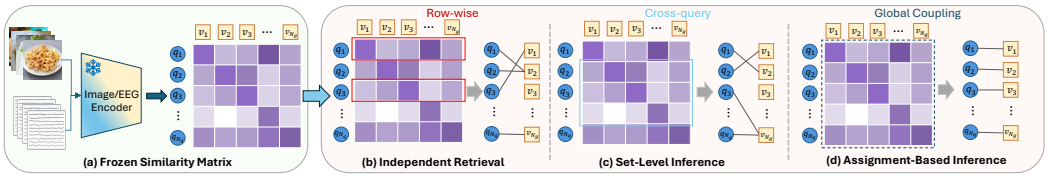


Figure 1: Overview of the frozen-similarity retrieval setting and the three test-time inference regimes considered in this work. (a) Frozen EEG–image representations are used to construct the similarity matrix $\mathbf{S}$. (b) Independent retrieval performs row-wise ranking without cross-query interaction. (c) Set-level inference exploits cross-query statistics to recalibrate similarity scores while retaining query-wise prediction. (d) Assignment-based inference jointly couples query–gallery correspondences through global coupling.

under full-context evaluation is no longer optimal in 17 of 24 representation–context-size conditions (70.8%), incurring an average performance loss of 2.70 percentage points and a maximum loss of 8.29 percentage points under reduced context. We further show that the apparent ranking and relative performance margins of frozen representations depend on how test-time structure is exploited and can vary with the available query context.

2 Problem Formulation

Let $\mathcal{Q} = \{q_i\}_{i=1}^{N_q}$ denote the EEG queries and $\mathcal{G} = \{g_j\}_{j=1}^{N_g}$ the image gallery. Frozen EEG and image encoders produce normalized embeddings $\mathbf{z}_i$ and $\mathbf{v}_j$, from which we construct the query–gallery similarity matrix

$$\mathbf{S} \in \mathbb{R}^{N_q \times N_g}, \qquad S_{ij} = \mathbf{z}_i^\top \mathbf{v}_j. \tag{1}$$

All inference regimes considered in this work operate on the same frozen similarity matrix and differ only in how they use information across queries, as illustrated in Fig. 1.

Under independent retrieval (Fig. 1(b)), each query is predicted using only its own similarity row:

$$\hat{y}_i^{\text{ind}} = \arg\max_j S_{ij}. \tag{2}$$

Because $\hat{y}_i^{\text{ind}}$ depends only on $\mathbf{S}_{i,:}$, it is invariant to the presence or composition of the other test queries. Let $\mathcal{B} \subseteq \mathcal{Q}$, with $|\mathcal{B}| = m$, denote a query context processed jointly at inference. A context-dependent inference rule Φ operates on the corresponding rows of the frozen similarity matrix:

$$\hat{\mathbf{y}}^{(\mathcal{B})} = \Phi(\mathbf{S}_{\mathcal{B},:}). \tag{3}$$

For two contexts $\mathcal{B}$ and $\mathcal{B}'$ that both contain q_i, the resulting prediction for q_i may differ even though its similarity row, the image gallery, and the underlying representations remain fixed.

As illustrated in Fig. 1(c,d), set-level inference uses cross-query statistics to recalibrate similarity scores while retaining query-wise prediction, whereas assignment-based inference jointly couples query–gallery correspondences. This distinction describes how cross-query information enters inference rather than an ordering of method strength; stronger coupling does not necessarily yield better retrieval performance.

3 Context-Marginalized Retrieval Evaluation

A single full-context evaluation captures only one operating point of context-dependent inference. We therefore introduce **Context-Marginalized Retrieval Evaluation (CMRE)**, which marginalizes evaluation over repeatedly constructed query contexts while keeping the full query–gallery similarity matrix fixed (Fig. 2). Each inference rule is applied independently to every context, and the resulting per-query predictions and ranks are aggregated across context evaluations. CMRE reports context-marginalized retrieval performance, quantifies **structural exploitability** (SE) through gains over independent retrieval, and measures **context robustness** using Context Consistency (CC), the pairwise agreement of each query's Top-1 predictions across contexts.

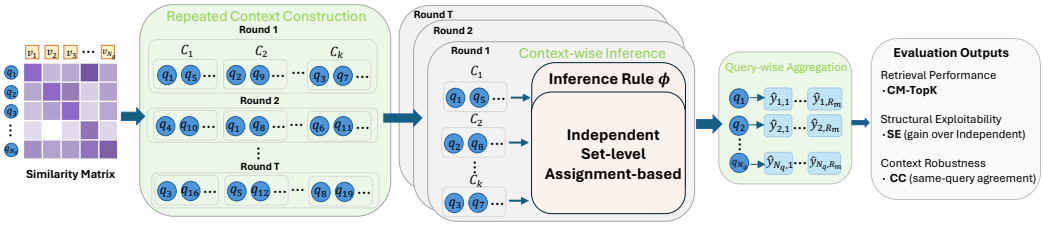


Figure 2: Overview of Context-Marginalized Retrieval Evaluation (CMRE). Repeated context construction and context-wise inference generate per-query observations from a fixed query–gallery similarity matrix; query-wise aggregation yields CM-TopK, structural exploitability relative to independent retrieval, and same-query Context Consistency (CC).

Repeated Context Construction. For a context size m, CMRE repeatedly constructs query contexts at the test-set level according to a balanced evaluation schedule, and every query contained in a constructed context is evaluated. For a given query q_i, let $\mathcal{B}_{i,r}^{(m)}$ denote the r-th constructed context containing q_i, where $r \in \{1, \ldots, R_m\}$ and

$$q_i \in \mathcal{B}_{i,r}^{(m)}, \qquad \left|\mathcal{B}_{i,r}^{(m)}\right| = m. \tag{4}$$

We evaluate reduced query-context sizes $m \in \{25, 50, 100, 150\}$ using $T = 10$ construction rounds. For $m \in \{25, 50, 100\}$, the $N_q = 200$ test queries are randomly permuted in each round and partitioned into N_q/m non-overlapping groups of size m, denoted by $\{C_k\}_{k=1}^{N_q/m}$. Thus, every query appears in exactly one context per round. For $m = 150$, the queries are randomly partitioned into four disjoint groups of 50 in each round, and four 150-query contexts, denoted by $\{C_k\}_{k=1}^{4}$, are constructed as their complements. Every query therefore appears in three contexts per round. Consequently, each query receives $R_{25} = R_{50} = R_{100} = 10$ evaluations and $R_{150} = 30$ evaluations. The schedule equalizes the number of evaluations per query at each context size, while leaving pairwise companion-query co-occurrence unconstrained. The conventional full-context case, $m = N_q = 200$, is evaluated separately. The complete CMRE evaluation procedure and context-sampling schedule are provided in Appendix B.4.

Context-Marginalized Retrieval Performance. Let y_i denote the index of the ground-truth image for query q_i, and let $\rho_{i,r}^{(m,\Phi)}$ denote its rank under inference rule Φ in the r-th context evaluation at context size m. We define

$$\text{CM-TopK}_\Phi(m) = \frac{1}{N_q R_m} \sum_{i=1}^{N_q} \sum_{r=1}^{R_m} \mathbb{I}\left[\rho_{i,r}^{(m,\Phi)} \le K\right], \tag{5}$$

Setting $K = 1$ yields $\text{CM-Top1}_\Phi(m)$. This metric averages Top-K retrieval success across all query–context evaluations. Let ρ_i^{ind} denote the rank of the ground-truth image for query q_i under independent retrieval. The independent-retrieval reference at rank K is

$$\text{TopK}_{\text{ind}} = \frac{1}{N_q} \sum_{i=1}^{N_q} \mathbb{I}\left[\rho_i^{\text{ind}} \le K\right]. \tag{6}$$

We quantify SE as the gain over independent retrieval at the corresponding rank cutoff K:

$$\text{SE}_\Phi(m) = \text{CM-TopK}_\Phi(m) - \text{TopK}_{\text{ind}}. \tag{7}$$

A positive value indicates that the inference rule benefits from cross-query structure relative to independent query-wise retrieval. For the full-context case, we set $R_{N_q} = 1$, such that $\text{CM-TopK}_\Phi(N_q)$ reduces to conventional full-context Top-K accuracy. The same definition therefore applies to both reduced and full contexts.

Same-Query Context Consistency. Context-marginalized retrieval performance does not indicate whether the same query receives stable Top-1 predictions across contexts. We therefore define CC.

Let $n_{i,c}^{(m,\Phi)}$ denote the number of context evaluations in which inference rule Φ predicts gallery candidate c as the Top-1 match for query q_i. We define

$$\mathrm{CC}_{\Phi}(m) = \frac{1}{N_q} \sum_{i=1}^{N_q} \frac{\sum_{c=1}^{N_g} n_{i,c}^{(m,\Phi)} \left(n_{i,c}^{(m,\Phi)} - 1 \right)}{R_m(R_m - 1)}. \tag{8}$$

For each query, CC measures the fraction of pairs of distinct context evaluations that yield the same Top-1 prediction, and then averages this agreement across queries. It therefore measures context robustness rather than retrieval correctness. Independent retrieval is context-invariant and therefore attains $\mathrm{CC}_{\mathrm{ind}}(m) = 100\%$. In the full-context case, each query is evaluated only once ($R_{N_q} = 1$), so CC cannot be computed and is therefore not reported for $m = N_q$.

4 Experiments

4.1 Experimental Setup

Dataset and Retrieval Protocol. We evaluate THINGS-EEG2 (Gifford et al., 2022) under the standard 200-way EEG-to-image retrieval setting using a leave-one-subject-out (LOSO) cross-subject protocol, with results averaged across ten held-out subjects. We additionally perform cross-modal validation on THINGS-MEG (Hebart et al., 2023) under a LOSO protocol across four held-out subjects.

Representations. We evaluate six frozen EEG-to-image representations: NICE (Song et al., 2024), ATM (Li et al., 2024), UBP (Wu et al., 2025), BLUR (Liu et al., 2026), HyFI (Jo et al., 2026), and BrainHIVE (Zheng et al., 2026). These methods cover diverse neural–visual alignment strategies and representation strengths, allowing us to assess whether test-time structure-aware inference behaves consistently across different EEG–image similarity spaces.

Compared Inference Methods. We compare independent retrieval with three set-level methods: CSLS (Lample et al., 2018), IS (Smith et al., 2017), and GC (Dinu et al., 2015) and three assignment-based methods: Sinkhorn (Cuturi, 2013), Hungarian matching (Kuhn, 1955), and Partial OT (Chapel et al., 2020). For each representation, all inference methods operate on the same frozen similarity matrix, with independent retrieval serving as the context-invariant reference.The fixed configurations and test-time complexity of all inference methods are provided in Appendix B.3.

CMRE Evaluation. We apply CMRE following the protocol in Section 3, using the same constructed contexts for all inference methods. The conventional full-context setting ($m = 200$) is evaluated separately. We report CM-Top1 and CM-Top5 for all methods. SE is measured by the corresponding Top-K gain over independent retrieval, whereas CC is reported only for reduced contexts with $R_m > 1$. We use $T = 10$ construction rounds in the main experiments. Analysis up to $T = 50$ shows that the estimates at $T = 10$ are close to those obtained with more rounds (Appendix F).

4.2 Full-Context Structural Exploitability

We first evaluate structural exploitability at the conventional full-context operating point ($m = N_q = 200$), where all test queries are processed jointly. For each frozen representation, every inference rule receives the same full query–gallery similarity matrix. The gain over independent retrieval therefore measures how much additional retrieval performance each rule obtains by exploiting cross-query structure at test time.

Table 1 reports full-context Top-1 accuracy for seven inference strategies across six frozen EEG-to-image representations. All six structure-aware inference methods improve accuracy over independent retrieval for every representation, although the gains vary substantially across representations. The largest gain is obtained by Hungarian matching on HyFI, which increases Top-1 accuracy from 13.70% to 34.10%, an improvement of 20.40 percentage points (pp). Averaged across representations, Hungarian achieves the highest Mean SE (+12.96 pp), followed by Sinkhorn (+10.71 pp) and Partial OT (+8.68 pp). The preferred inference rule is representation-dependent. Among the set-level methods, IS performs best on five representations, whereas GC performs best on ATM. Among

Table 1: Full-context Top-1 retrieval accuracy (%) across six representations. Mean SE denotes the average gain (pp) over independent retrieval.

Type	Method	NICE	ATM	UBP	BLUR	HyFI	BrainHIVE	Mean SE
Independent	Independent	6.95	11.50	10.45	20.05	13.70	22.25	–
Set-level	CSLS	7.65	14.50	17.20	26.95	20.90	30.35	5.44
	IS	8.15	14.25	21.65	30.20	27.70	35.50	8.76
	GC	7.10	15.50	20.95	27.20	26.05	30.90	7.13
Assignment	Sinkhorn	9.05	15.50	20.15	35.25	26.90	42.30	10.71
	Hungarian	7.85	15.35	23.60	40.15	34.10	41.60	12.96
	Partial OT	8.20	15.55	18.75	30.90	23.40	40.20	8.68

the assignment-based methods, Hungarian performs best on UBP, BLUR, and HyFI; Sinkhorn performs best on NICE and BrainHIVE; and Partial OT is marginally best on ATM. Representation comparisons are also inference-dependent. Independent retrieval ranks ATM above UBP, whereas every structure-aware inference rule ranks UBP above ATM. Thus, even at full context, the preferred inference rule depends on the representation, while the apparent representation ordering depends on the inference rule. These ranking dependencies are even more pronounced in the corresponding full-context Top-5 results (Appendix C.2).

These results establish that structural exploitability depends jointly on the frozen representation and the inference rule. To characterize the source of these gains, Appendix C.1 analyzes Top-1 wrong-to-correct and correct-to-wrong transitions, together with the initial ground-truth ranks of recovered queries. However, full-context results cannot determine whether either the observed gains or comparative rankings remain stable as the available query context changes. We therefore next evaluate structure-aware inference under reduced query contexts using CMRE.

4.3 Structural Exploitability under Reduced Query Contexts

The full-context results quantify structural exploitability at a single operating point, but do not show whether retrieval performance, prediction stability, and comparative rankings remain stable as query context changes. We therefore apply CMRE at reduced query-context sizes $m \in \{25, 50, 100, 150\}$. Across these evaluations, the image gallery and frozen full query–gallery similarity matrix remain fixed, while only the size and composition of the constructed query contexts vary.

Table 1 identifies Hungarian as the full-context winner on BLUR, achieving 40.15% Top-1 accuracy at $m = 200$. However, Fig. 3(a) shows that Sinkhorn ranks first at every reduced context size, whereas Hungarian declines from 33.36% at $m = 150$ to 22.02% at $m = 25$. Retaining the full-context winner therefore incurs losses of 1.30, 5.71, 8.29, and 8.11 pp at $m = 150$, 100, 50, and 25, respectively. Independent retrieval remains context-invariant at 20.05%. Thus, on BLUR, reducing the query context reverses the preference between Hungarian and Sinkhorn.

Table 2 summarizes mean CM-Top1 structural exploitability across the four reduced context sizes. All 36 representation–rule combinations yield positive mean gains over independent retrieval, ranging from 0.02 pp for GC on NICE to 12.82 pp for Sinkhorn on BLUR. Thus, useful cross-query structure remains exploitable even when inference operates on only a subset of the test queries. To assess whether the observed gains were consistent across subjects, we averaged each subject's gain across the six representations and four reduced context sizes. All six structure-aware inference rules yielded positive mean subject-level gains, and the corresponding subject-level gains differed significantly from zero under two-sided Wilcoxon signed-rank tests with Holm correction across methods (Appendix D.2). However, both the magnitude of the gain and the ranking of inference rules depend strongly on the frozen representation. Sinkhorn performs best on NICE (+1.32 pp), BLUR (+12.82 pp), and BrainHIVE (+10.02 pp), whereas CSLS ranks first on ATM (+2.51 pp) and IS ranks first on UBP (+10.34 pp) and HyFI (+12.69 pp). No single inference rule is therefore uniformly optimal across representations.The same conclusion extends beyond Top-1 retrieval: CM-Top5 also shows positive structural gains across frozen representations, while the preferred inference rule remains representation-dependent (Appendix D.3).

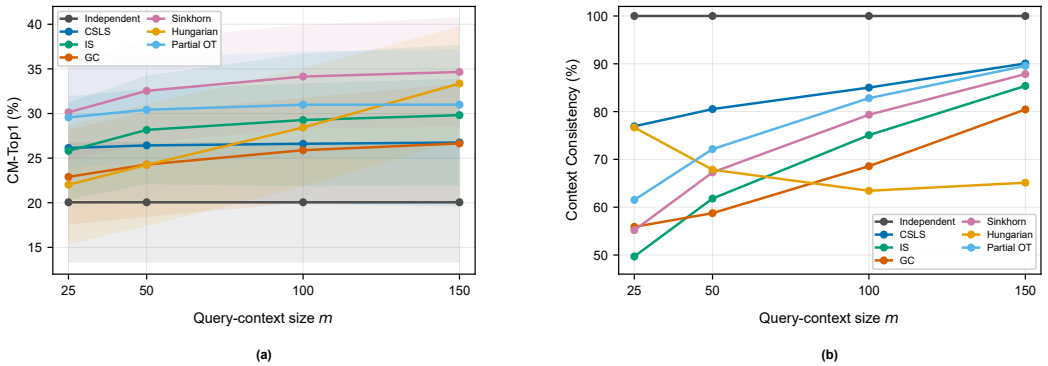


Figure 3: Context-marginalized retrieval performance and context robustness on BLUR across query-context sizes m. (a) Context-marginalized Top-1 retrieval accuracy (CM-Top1); shaded bands indicate standard deviation across subjects. (b) Same-query Context Consistency (CC)

Table 2: CM-Top1 structural exploitability (SE, pp), averaged over the reduced query-context sizes. The final column reports the average across the six frozen representations.

Type	Method	NICE	ATM	UBP	BLUR	HyFI	BrainHIVE	Average
Set-level	CSLS	0.68	**2.51**	6.34	6.43	7.45	7.46	5.15
	IS	0.36	1.14	**10.34**	8.22	**12.69**	5.62	6.40
	GC	0.02	2.28	7.37	4.88	7.91	5.22	4.61
Assignment	Sinkhorn	**1.32**	2.28	9.38	**12.82**	12.14	**10.02**	**7.99**
	Hungarian	0.46	2.36	6.78	6.96	8.37	7.01	5.32
	Partial OT	0.99	2.40	8.08	10.44	9.75	9.47	6.85

Averaged across the six representations, Sinkhorn achieves the largest mean reduced-context gain (+7.99 pp), followed by Partial OT (+6.85 pp) and IS (+6.40 pp). Across the 24 representation–context combinations, however, the full-context winner is no longer optimal in 17 cases (70.8%). Selecting the full-context winner incurs an average Top-1 loss of 2.70 pp relative to the best method at each reduced context (Appendix D.1). Thus, changes in query context can alter not only retrieval performance but also the ranking of inference rules. We next examine whether same-query predictions remain consistent across contexts.

4.4 Context Robustness under Reduced Query Contexts

Retrieval performance alone does not reveal whether the same EEG query retains its Top-1 prediction as the companion-query composition changes. We therefore quantify context robustness using Context Consistency (CC), which measures pairwise agreement among the Top-1 predictions produced for the same query across its constructed context evaluations at a given context size.

Figure 3(b) reports the CC profiles on BLUR across the four reduced context sizes. Independent retrieval maintains $\mathrm{CC} = 100\%$ because its predictions are context-invariant by construction. Among the structure-aware inference rules, CSLS is comparatively stable, with CC increasing from 76.92% at $m = 25$ to 90.07% at $m = 150$. By contrast, Hungarian's CM-Top1 increases from 22.02% to 33.36% over the same range, while its CC decreases from 76.69% to 65.12%, reaching a minimum of 63.44% at $m = 100$. This divergence shows that higher retrieval performance does not necessarily imply greater context robustness.

Table 3 shows that the pattern extends beyond BLUR. CSLS achieves the highest mean CC on each of the six frozen representations and the highest overall CC (85.04%). The other inference rules attain lower overall CC values, ranging from 61.70% to 69.44%, and vary more strongly across representations. To assess structural exploitability and context robustness jointly, Fig. 4(a) plots each inference rule's mean reduced-context Top-1 gain against its mean CC. CSLS achieves the

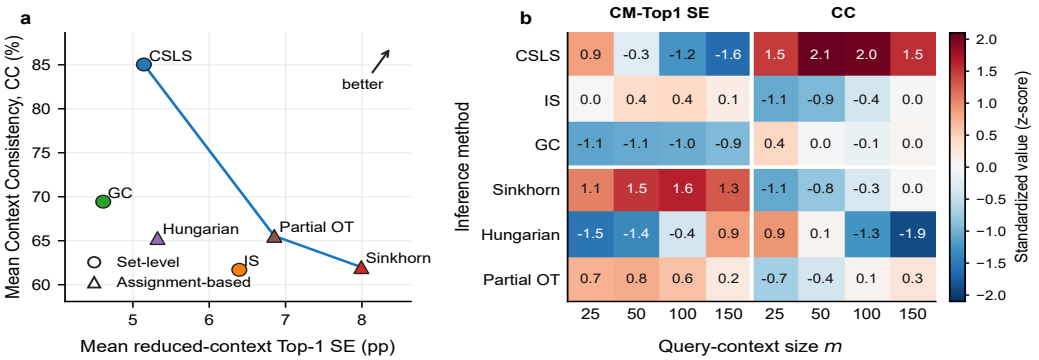


Figure 4: Joint profiles of structural exploitability and context robustness. (a) Mean reduced-context Top-1 SE versus mean Context Consistency (CC), across six frozen representations and four reduced context sizes. (b) Context-size-specific SE and CC, averaged across representations and standardized across rules within each metric and context size.

Table 3: Average same-query context consistency (CC,%) across reduced query contexts. Final column reports the average across the six frozen representations.

Type	Method	NICE	ATM	UBP	BLUR	HyFI	BrainHIVE	Average
Set-level	CSLS	**83.81**	**84.81**	**85.79**	**83.14**	**85.78**	**86.89**	**85.04**
	IS	54.78	47.32	76.23	67.99	74.72	49.17	61.70
	GC	65.28	77.24	63.07	65.92	66.34	78.81	69.44
Assignment	Sinkhorn	58.06	43.34	77.32	72.42	75.64	45.24	62.00
	Hungarian	65.69	64.66	62.03	68.27	64.01	67.17	65.31
	Partial OT	60.90	46.28	81.04	76.51	80.19	48.45	65.56

highest CC (85.04%) with a moderate gain (+5.15 pp), whereas Sinkhorn provides the largest gain (+7.99 pp) but substantially lower CC (62.00%). Partial OT occupies an intermediate operating point, combining a relatively large gain (+6.85 pp) with higher CC (65.56%) than Sinkhorn. These three methods form the non-dominated Pareto frontier; IS, GC, and Hungarian are dominated by methods with both higher gain and higher CC.

Figure 4(b) resolves these joint profiles by context size. For each m, structural exploitability and CC are averaged across the six representations and standardized across inference rules within each metric. Sinkhorn remains above average in structural exploitability but below average in CC, whereas CSLS exhibits the opposite profile. Hungarian's relative gain improves as the context size increases, while its relative CC declines. Thus, no inference rule dominates both dimensions, and method selection depends on the relative priority assigned to retrieval gain and context robustness.

Detailed CC results for all six representations and individual context sizes are provided in Appendix E.1. Because CC measures agreement irrespective of correctness, Appendix E.2 further decomposes same-query behavior into Stable-Correct, Stable-Wrong, and Context-Sensitive cases. Averaged across representations and context sizes, CSLS exhibits both the highest Stable-Correct fraction (14.90%) and the largest Stable-Wrong fraction (48.14%). High context consistency can therefore reflect both beneficial and erroneous stability and should be interpreted jointly with retrieval performance. Beyond query-context variation, we also evaluated robustness to gallery expansion by adding unmatched distractors. On BLUR, structure-aware inference remained beneficial as the gallery grew although the gains diminished(Appendix G).

4.5 Validation on THINGS-MEG

Table 4 and Figure 5 extend the CMRE analysis to THINGS-MEG. All six structure-aware inference rules outperform independent retrieval under both full and reduced query contexts, indicating

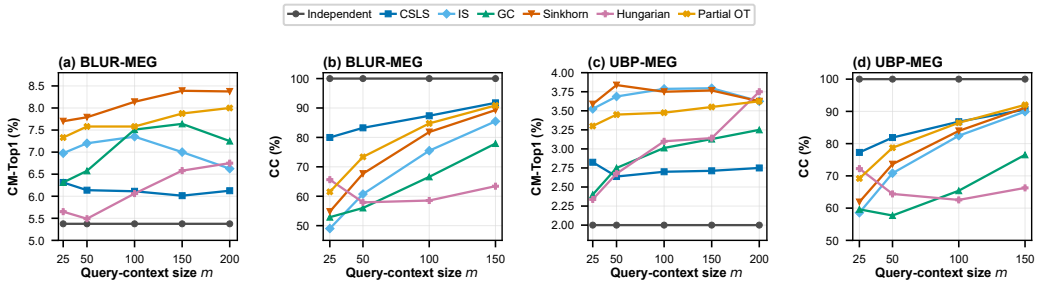


Figure 5: CM-Top1 and CC across query-context sizes for BLUR (a,b) and UBP (c,d).

Table 4: Validation on THINGS-MEG. Full-context Top-1 is averaged across four held-out subjects. Reduced-context results are averaged over $m \in \{25, 50, 100, 150\}$. SE denotes Top-1 structural exploitability relative to independent retrieval.

Representation	Method	Full Context		Reduced Contexts		
		Top-1 (%)	SE (pp)	CM-Top1 (%)	SE (pp)	CC (%)
BLUR	Independent	5.38	–	–	–	100.00
	CSLS	6.13	+0.75	6.14	+0.76	**85.58**
	IS	6.63	+1.25	7.13	+1.75	67.67
	GC	7.25	+1.88	7.01	+1.63	63.35
	Sinkhorn	**8.38**	**+3.00**	**8.01**	**+2.63**	73.39
	Hungarian	6.75	+1.38	5.95	+0.57	61.38
	Partial OT	8.00	+2.63	7.59	+2.21	77.62
UBP	Independent	2.00	–	–	–	100.00
	CSLS	2.75	+0.75	2.72	+0.72	**84.15**
	IS	3.63	+1.63	3.70	+1.70	75.43
	GC	3.25	+1.25	2.82	+0.82	64.80
	Sinkhorn	3.63	+1.63	**3.74**	**+1.74**	77.63
	Hungarian	**3.75**	**+1.75**	2.82	+0.82	66.35
	Partial OT	3.63	+1.63	3.45	+1.45	81.60

that cross-query structure remains exploitable beyond EEG. On BLUR, Sinkhorn performs best in both the full-context setting (8.38%, +3.00 pp) and the reduced query contexts (8.01%, +2.63 pp). On UBP, the preferred rule changes with context: Hungarian performs best under full context (3.75%,+1.75 pp), whereas Sinkhorn achieves the highest mean CM-Top1 under reduced query contexts (3.74%, +1.74 pp). Figure 5 further shows that context-response profiles vary across inference rules. For most methods, CC increases with query-context size, but the ranking by CC differs from that by SE. CSLS achieves the highest mean CC on BLUR (85.58%) and UBP (84.15%), despite modest SE values of +0.76 and +0.72 pp, respectively. These results show that the separation between structural exploitability and context robustness persists in MEG, and that the preferred inference rule depends on both query-context size and representation.

5 Conclusion

We studied how test-time structure-aware inference in EEG-to-image retrieval depends on query-context composition. Using fixed EEG–image similarity matrices, we introduced CMRE to measure structural exploitability and Context Consistency across sampled contexts. The preferred inference rule varied across representations and context sizes, while representation rankings were themselves inference- and context-dependent. Larger structural gains did not necessarily imply greater context robustness, and this separation also persisted on THINGS-MEG. These findings show that structural exploitability and context robustness are complementary properties and that full-context accuracy alone can provide an incomplete evaluation of context-dependent test-time inference. Accordingly, CMRE evaluates inference across sampled query contexts by jointly considering context-marginalized retrieval performance, structural gains, and same-query prediction consistency.

AI USE STATEMENT

In this work, generative AI tools were used to assist with language editing, LaTeX formatting, visualization refinement, and limited code development. All AI-assisted outputs were reviewed and validated by the authors. The authors independently verified all analyses, results, figures, and scientific claims and take full responsibility for the final manuscript.

REPRODUCIBILITY STATEMENT

All experiments use a fixed evaluation pipeline in which each inference rule operates on frozen EEG-to-image query–gallery similarity matrices, without updating either encoder. Section 3 and Appendix B.4 specify the CMRE formulation, balanced context-construction schedule, inference rules, and evaluation metrics.

REFERENCES

Simion-Vlad Bogolin, Ioana Croitoru, Hailin Jin, Yang Liu, and Samuel Albanie. Cross modal retrieval with querybank normalisation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5194–5205, 2022.

Malik Boudiaf, Imtiaz Masud Ziko, Jérôme Rony, José Dolz, Pablo Piantanida, and Ismail Ben Ayed. Transductive information maximization for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Laetitia Chapel, Mokhtar Z. Alaya, and Gilles Gasso. Partial optimal transport with applications on positive-unlabeled learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Chi-Sheng Chen and Chun-Shu Wei. Mind's eye: Image recognition by EEG via multimodal similarity-keeping contrastive learning, 2024.

Zexuan Chen, Sichao Liu, Runhao Lu, Huichao Qi, Alexandra Woolgar, Xi Vincent Wang, and Lihui Wang. Mindalign: Bridging eeg, vision, and language for zero-shot visual decoding, 2026.

Xing Cheng, Hezheng Lin, Xiangyu Wu, Fan Yang, and Dong Shen. Improving video-text retrieval by multi-stream corpus alignment and dual softmax loss, 2021. URL https://arxiv.org/abs/2109.04290.

Yuhao Cheng, Xiaoguang Zhu, Jiuchao Qian, Fei Wen, and Peilin Liu. Cross-modal graph matching network for image-text retrieval. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 18(4):1–23, 2022. doi: 10.1145/3499027.

Lénaïc Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Unbalanced optimal transport: Dynamic and kantorovich formulations. *Journal of Functional Analysis*, 274(11): 3090–3123, 2018.

Woohyeok Choi, Jun Mo Kim, Songha Kim, Yebin Choi, and Tae Eui Kam. Subject-invariant EEG embeddings via mixup and adversarial learning for semantic retrieval. In *Proceedings of the 41st ACM Symposium on Applied Computing*, pp. 1960–1961. Association for Computing Machinery, 2026. doi: 10.1145/3748522.3779935.

Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9):1853–1865, 2017. doi: 10.1109/TPAMI.2016.2615921.

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

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, pp. 2292–2300, 2013.

Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. Improving zero-shot learning by mitigating the hubness problem. In *International Conference on Learning Representations Workshop*, 2015.

Alessandro T. Gifford, Kshitij Dwivedi, Gemma Roig, and Radoslaw M. Cichy. A large and rich eeg dataset for modeling human visual object recognition. *NeuroImage*, 264:119754, 2022.

Taesik Gong, Jongheon Jeong, Taewon Kim, Yewon Kim, Jinwoo Shin, and Sung-Ju Lee. NOTE: Robust continual test-time adaptation against temporal correlation. In *Advances in Neural Information Processing Systems*, 2022.

Tim Hagen, Harrisen Scells, and Martin Potthast. Revisiting query variation robustness of transformer models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 4283–4296, 2024.

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

Qunjie Huang and Weina Zhu. Sattc: Structure-aware label-free test-time calibration for cross-subject eeg-to-image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16887–16896, 2026.

Lin Jiang, Qingshan She, Jiale Xu, Haiqi Xu, Duanpo Wu, and Zhenzhong Kuang. Subject-aware multi-granularity alignment for zero-shot EEG-to-image retrieval, 2026.

Sangmin Jo, Wootaek Jeong, Da-Woon Heo, Yoohwan Hwang, and Heung-Il Suk. Hyfi: Hyperbolic feature interpolation for brain-vision alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 5575–5583, 2026.

H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97, 1955.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. In *International Conference on Learning Representations*, 2018.

Chang Min Lee, Jun Mo Kim, Yebin Choi, Woo Hyeok Choi, and Tae Eui Kam. Define-align-fuse: Dual-stream EEG-vision alignment using hierarchical CLIP representation. In *2026 14th International Conference on Brain-Computer Interface (BCI)*. IEEE, 2026. doi: 10.1109/BCI69045.2026.11435103.

Dongyang Li, Chen Wei, Shiying Li, Jiachen Zou, and Quanying Liu. Visual decoding and reconstruction via EEG embeddings with guided diffusion. In *Advances in Neural Information Processing Systems*, volume 37, pp. 102822–102864, 2024. doi: 10.52202/079017-3266.

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

Zengrong Lin, Zheng Wang, Tianwen Qian, Pan Mu, Sixian Chan, and Cong Bai. Neighborretr: Balancing hub centrality in cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9263–9273, 2025.

Wenchao Liu, Hongwei Li, Zhouyang Xu, Lin Ma, and Haifeng Li. Leveraging visual blur perception characteristics for eeg decoding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 17580–17588, 2026.

Ségolène Martin, Malik Boudiaf, Emilie Chouzenoux, Jean-Christophe Pesquet, and Ismail Ben Ayed. Towards practical few-shot query sets: Transductive minimum description length inference. In *Advances in Neural Information Processing Systems*, volume 35, pp. 34677–34688, 2022.

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

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. In *International Conference on Learning Representations*, 2023.

Zhengxin Pan, Haishuai Wang, Fangyu Wu, Peng Zhang, and Jiajun Bu. Hubness reduction with dual bank sinkhorn normalization for cross-modal retrieval. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 6153–6162, 2025.

Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4938–4947, 2020.

Feixue Shao, Guangze Shi, Xueyu Liu, Yongfei Wu, Mingqiang Wei, Jianan Zhang, Jianbo Lu, Guiying Yan, and Weihua Yang. Brain-inspired capture: Evidence-driven neuromimetic perceptual simulation for visual decoding, 2026.

Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *International Conference on Learning Representations*, 2017.

Yonghao Song, Bingchuan Liu, Xiang Li, Nanlin Shi, Yijun Wang, and Xiaorong Gao. Decoding natural images from eeg for object recognition. In *International Conference on Learning Representations*, 2024.

Yonghao Song, Yijun Wang, Huiguang He, and Xiaorong Gao. Recognizing natural images from EEG with language-guided contrastive learning. *IEEE Transactions on Neural Networks and Learning Systems*, 36(9):15896–15910, 2025. doi: 10.1109/TNNLS.2025.3562743.

Shitong Sun, Qilei Li, Shaogang Gong, Weitong Cai, Philip Torr, and Jindong Gu. Benchcir: Benchmarking robustness in composed image retrieval across modalities. *Pattern Recognition*, 179: 113724, 2026.

Olivier Veilleux, Malik Boudiaf, Pablo Piantanida, and Ismail Ben Ayed. Realistic evaluation of transductive few-shot learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 9290–9302, 2021.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021.

Jiyuan Wang, Li Zhang, Haipeng Lin, Qile Liu, Gan Huang, Ziyu Li, Zhen Liang, and Xia Wu. Neuroclip: Brain-inspired prompt tuning for EEG-to-image multimodal contrastive learning, 2025a.

Junhan Wang and Kani Chen. Cortiva: Candidate-score fusion of complementary visual teachers for eeg- and meg-to-image retrieval, 2026.

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

Yimu Wang, Xiangru Jian, and Bo Xue. Balance act: Mitigating hubness in cross-modal retrieval with query and gallery banks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 10542–10567, 2023.

Zixin Wang, Yadan Luo, Liang Zheng, Zhuoxiao Chen, Sen Wang, and Zi Huang. In search of lost online test-time adaptation: A survey. *International Journal of Computer Vision*, 133(3): 1106–1139, 2025b.

Yayun Wei, Lei Cao, Hao Li, and Yilin Dong. Mb2c: Multimodal bidirectional cycle consistency for learning robust visual neural representations. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 8992–9000, 2024. doi: 10.1145/3664647.3681292.

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.

Haopei Xu and Ni Lin. Neurovista: A bidirectional masked cross-modal fusion network for robust EEG-to-image decoding. *Neural Networks*, 195:108297, 2026. doi: 10.1016/j.neunet.2025.108297.

Longhui Yuan, Binhui Xie, and Shuang Li. Robust test-time adaptation in dynamic scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15922–15932, 2023.

Kaifan Zhang, Lihuo He, Xin Jiang, Wen Lu, Di Wang, and Xinbo Gao. Cognitioncapturer: Decoding visual stimuli from human eeg signal with multimodal information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 14486–14493, 2025.

Kaifan Zhang, Lihuo He, Yuqi Ji, Junjie Ke, Lukun Wu, Tianhao You, and Xinbo Gao. Eeg-editbench: Probing visual information in eeg–image retrieval models with controlled image edits, 2026a. URL https://arxiv.org/abs/2607.27857.

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

Hao Zhao, Yuejiang Liu, Alexandre Alahi, and Tao Lin. On pitfalls of test-time adaptation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 42058–42080, 2023.

Jiawen Zheng, Haonan Jia, Ming Li, Yuhui Zheng, Yufeng Zeng, Yang Gao, and Chen Liang. Learning brain representation with hierarchical visual embeddings. In *International Conference on Learning Representations*, 2026.

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

A Related Work

A.1 EEG-to-Image Retrieval

Most EEG-to-image retrieval methods align EEG embeddings with pretrained visual spaces. Early studies established zero-shot retrieval through contrastive neural–visual learning (Song et al., 2024; Li et al., 2024; Wei et al., 2024; Chen & Wei, 2024). Subsequent work added language-guided supervision, multimodal semantic bridging, and richer cross-modal objectives (Song et al., 2025; Li et al., 2025; Zhang et al., 2025; Chen et al., 2026; Wang & Chen, 2026). More recent methods explored prompt tuning, hierarchical visual representations, pretrained vision encoders, and subject-aware alignment (Wang et al., 2025a; Zhou et al., 2026; Lee et al., 2026; Jiang et al., 2026).

Perceptual and neuroscientific priors have also been introduced through uncertainty-aware supervision, multiscale targets, hyperbolic modeling, neuromimetic simulation, and masked cross-modal fusion (Wu et al., 2025; Zheng et al., 2026; Jo et al., 2026; Zhang et al., 2026b; Shao et al., 2026; Xu & Lin, 2026). Cross-subject retrieval remains challenging because EEG representations vary across individuals. Recent methods address this problem through subject-invariant learning, label-free calibration, similarity-geometry adaptation, and target-subject coordinate recovery (Choi et al., 2026; Huang & Zhu, 2026; Cui et al., 2026). A recent diagnostic study further showed that standard retrieval performance may not transfer to controlled evaluation settings (Zhang et al., 2026a). However, prior work generally evaluates each target under a fixed test context, leaving sensitivity to companion-query composition insufficiently characterized.

A.2 Test-Time Structural Inference for Retrieval

Post-hoc structural inference refines retrieval scores without changing the underlying representation. One line of work addresses hubness, where a small number of gallery items become nearest neighbors of many queries (Dinu et al., 2015; Smith et al., 2017; Lample et al., 2018; Bogolin et al., 2022; Wang et al., 2023; Lin et al., 2025; Pan et al., 2025). Global Correction (GC) re-ranks candidates using query-population preferences (Dinu et al., 2015); Inverted Softmax (IS) suppresses popular targets through reverse normalization (Smith et al., 2017); and Cross-domain Similarity Local Scaling (CSLS) corrects for local embedding density (Lample et al., 2018). Related matrix-normalization methods have also been applied to cross-modal retrieval (Cheng et al., 2021; Bogolin et al., 2022; Wang et al., 2023). We refer to these methods as set-level inference because predictions remain query-wise while using statistics from other queries or gallery items.

Joint matching introduces stronger structural coupling. Optimal transport provides soft global assignments (Cuturi, 2013; Courty et al., 2017; Chizat et al., 2018; Chapel et al., 2020), with entropic regularization enabling efficient Sinkhorn iterations (Cuturi, 2013). Unbalanced transport relaxes mass conservation (Chizat et al., 2018), while partial transport relaxes full matching (Chapel et al., 2020); the Hungarian algorithm instead imposes hard one-to-one assignment (Kuhn, 1955). Related global matching methods have also been used in vision and cross-modal correspondence (Sarlin et al., 2020; Cheng et al., 2022). These approaches span score recalibration, soft global coupling, and hard assignment. The effectiveness of post-hoc normalization can also depend on available query and gallery statistics. Querybank Normalisation uses an external query bank (Bogolin et al., 2022), whereas Dual Bank Normalization uses both query and gallery banks (Wang et al., 2023). Target-set structure has also been used for label-free cross-subject EEG-to-image calibration (Huang & Zhu, 2026; Cui et al., 2026). However, these studies optimize retrieval under prescribed test configurations. We instead hold the representation and inference rule fixed and vary only the available query context.

A.3 Evaluation of Transductive Inference

Transductive inference uses statistics from unlabeled test samples, making its performance sensitive to assumptions about the query set. Transductive few-shot methods benefit from joint reasoning over queries (Boudiaf et al., 2020), but evaluations show that class balance, class composition, and query-set structure can substantially alter measured performance (Veilleux et al., 2021; Martin et al., 2022). Test-set composition therefore forms part of the information available to an inference procedure rather than merely a benchmark detail.

Related concerns arise in test-time adaptation. Early methods adapt models using statistics or entropy objectives from unlabeled batches (Wang et al., 2021), while continual and source-free variants address changing test distributions (Wang et al., 2022; Niu et al., 2022). Test-time behavior is also affected by temporal correlation, batch size, class imbalance, and distribution shifts (Gong et al., 2022; Niu et al., 2023; Yuan et al., 2023). Benchmarks further show sensitivity to evaluation protocols, model selection, backbone quality, and shift type (Zhao et al., 2023; Wang et al., 2025b; Hagen et al., 2024; Sun et al., 2026). However, these studies primarily vary distributions or adaptation settings rather than repeatedly evaluating the same target under controlled changes to its companion-query context.

B Implementation Details

This section provides additional implementation details for the six EEG-to-image retrieval representations and the six test-time structural inference methods used in our experiments. For each representation, we preserve its original EEG–image representation pipeline and obtain a frozen query–gallery similarity matrix before applying any structural inference. The inference methods therefore modify only the test-time similarity structure and do not update the representation parameters or embeddings.

B.1 Representation Implementations

NICE. NICE (Song et al., 2024) represents the standard contrastive EEG–vision alignment paradigm. It learns an EEG encoder whose outputs are directly aligned with pretrained visual representations through a contrastive objective, providing a relatively direct neural-to-visual embedding space.

ATM. ATM (Li et al., 2024) introduces an Adaptive Thinking Mapper to transform EEG representations toward the pretrained visual embedding space. Compared with the direct contrastive alignment of NICE, ATM explicitly learns a mapping between the neural and visual representations, providing a different similarity geometry for retrieval. We use the retrieval branch of the original framework without involving its reconstruction components.

UBP. UBP (Wu et al., 2025) incorporates uncertainty-aware visual supervision into EEG–image alignment. Rather than treating every EEG–image pair as an equally reliable training target, it modulates visual supervision according to the estimated ambiguity of neural–visual correspondence. This produces a similarity space shaped by uncertainty-aware alignment rather than purely deterministic contrastive matching.

BLUR. BLUR (Liu et al., 2026) constructs a perceptually diversified visual space using multiple levels of image blur and learns to select informative visual evidence for EEG alignment. Its retrieval representation therefore reflects multi-view perceptual supervision rather than a single visual target, making it particularly useful for examining whether test-time structural inference behaves differently on a perceptually enriched similarity space.

HyFI. HyFI (Jo et al., 2026) models EEG–visual relations in a hyperbolic feature space, aiming to capture hierarchical semantic structure that may be difficult to represent in a purely Euclidean embedding. It therefore provides a geometrically distinct retrieval representation from the Euclidean contrastive and perceptual-alignment backbones considered above.

BrainHIVE. BrainHIVE (Zheng et al., 2026) employs hierarchical and multiscale visual supervision to align EEG representations with visual information at multiple semantic granularities. Unlike approaches based on a single visual embedding or a single perceptual transformation, BrainHIVE integrates complementary visual representations before retrieval, yielding a richer hierarchical EEG–image similarity structure.

The six representations are selected to span different representation families rather than merely different model capacities. NICE and ATM provide relatively direct neural–visual alignment, UBP and BLUR modify visual supervision through perceptual degradation, HyFI introduces a different embedding geometry, and BrainHIVE incorporates hierarchical multi-source visual representations. This diversity allows us to evaluate whether structural inference behavior persists across different frozen EEG–image similarity geometries.

B.2 Test-Time Inference Methods

All inference methods operate on the same frozen EEG–image similarity matrix $\mathbf{S} \in \mathbb{R}^{N_q \times N_g}$, where S_{ij} denotes the cosine similarity between EEG query q_i and gallery candidate g_j. We group the compared approaches into set-level methods, which recalibrate retrieval scores using statistics from the available query set, and assignment-based methods, which jointly couple queries and gallery candidates through global matching constraints.

CSLS. Cross-domain Similarity Local Scaling (CSLS) (Lample et al., 2018) mitigates hubness by recalibrating each pairwise similarity using local neighborhood statistics on both sides of the retrieval space. For query i and gallery candidate j, the adjusted similarity is

$$S_{ij}^{\mathrm{CSLS}} = 2S_{ij} - r_q(i) - r_g(j), \tag{9}$$

where

$$r_q(i) = \frac{1}{k} \sum_{j' \in \mathcal{N}_g^k(i)} S_{ij'}, \qquad r_g(j) = \frac{1}{k} \sum_{i' \in \mathcal{N}_q^k(j)} S_{i'j}. \tag{10}$$

Here, $\mathcal{N}_g^k(i)$ and $\mathcal{N}_q^k(j)$ denote the corresponding k-nearest neighborhoods. Retrieval is then performed using the recalibrated scores S_{ij}^{CSLS}.

Inverted Softmax. Inverted Softmax (IS) (Smith et al., 2017) suppresses gallery candidates that receive uniformly high similarity from many competing queries. For each query–gallery pair, we compute

$$S_{ij}^{\mathrm{IS}} = \frac{\exp(\beta S_{ij})}{\sum_{i' \neq i} \exp(\beta S_{i'j})}, \tag{11}$$

where β controls the sharpness of the normalization. The denominator captures the relative popularity of gallery candidate j across the available queries, while the target query itself is excluded from the normalization.

Global Correction. Global Correction (GC) (Dinu et al., 2015) corrects hubness by reversing the ranking perspective. For gallery candidate j, let

$$\rho_{ij} = 1 + \sum_{i' \neq i} \mathbb{I}[S_{i'j} > S_{ij}] \tag{12}$$

denote the rank of query q_i among all competing queries with respect to candidate g_j. GC predicts

$$\hat{y}_i^{\mathrm{GC}} = \arg\min_j \rho_{ij}, \tag{13}$$

with the original similarity S_{ij} used to resolve rank ties. Candidates that rank the target query unusually highly are therefore preferred over globally popular candidates.

Sinkhorn. Sinkhorn inference (Cuturi, 2013) performs globally coupled soft matching through entropy-regularized optimal transport. We define the transport cost as

$$C_{ij} = -S_{ij}, \tag{14}$$

and solve

$$\mathbf{P}^* = \arg\min_{\mathbf{P} \in \Pi(\mathbf{a}, \mathbf{b})} \langle \mathbf{P}, \mathbf{C} \rangle - \varepsilon H(\mathbf{P}), \tag{15}$$

where

$$\Pi(\mathbf{a}, \mathbf{b}) = \left\{ \mathbf{P} \geq 0 : \mathbf{P}\mathbf{1} = \mathbf{a}, \mathbf{P}^\top \mathbf{1} = \mathbf{b} \right\}, \tag{16}$$

and

$$H(\mathbf{P}) = -\sum_{i,j} P_{ij} \log P_{ij}. \tag{17}$$

We use uniform marginals $\mathbf{a} = \frac{1}{N_q}\mathbf{1}$ and $\mathbf{b} = \frac{1}{N_g}\mathbf{1}$. For a reduced context B with $|B| = m$, the marginals is defined as $\mathbf{a}_B = \frac{1}{m}\mathbf{1}_m$. The transport plan $\mathbf{P}^*$ is obtained using Sinkhorn iterations, and gallery candidates are ranked according to the corresponding row of $\mathbf{P}^*$.

Hungarian Matching. Hungarian matching (Kuhn, 1955) performs hard global assignment by maximizing the total similarity between queries and gallery candidates:

$$\mathbf{X}^* = \arg\max_{\mathbf{X}} \sum_{i=1}^{N_q} \sum_{j=1}^{N_g} X_{ij} S_{ij}, \tag{18}$$

subject to

$$X_{ij} \in \{0, 1\}, \qquad \sum_{j=1}^{N_g} X_{ij} = 1, \qquad \sum_{i=1}^{N_q} X_{ij} \leq 1. \tag{19}$$

Thus, every query is assigned exactly one gallery candidate, while each candidate can be assigned to at most one query. This formulation also applies to the expanded-gallery setting where $N_g > N_q$.

Table 5: Fixed configurations and test-time complexity of the compared inference methods. N_q and N_g denote the numbers of EEG queries and gallery candidates, respectively, and L denotes the number of transport iterations.

Method	Fixed configuration	Complexity
Independent	None	$O(N_q N_g)$
CSLS	$k = 10$	$O(N_q N_g)$
IS	$\beta = 30$	$O(N_q N_g)$
GC	None	$O(N_q N_g \log N_q)$
Sinkhorn	$\varepsilon = 0.05$	$O(L N_q N_g)$
Partial OT	$\tau = 0.90,\ \varepsilon = 0.05$	$O(L N_q N_g)$
Hungarian	None	$O(N_q^2 N_g)$

Partial Optimal Transport. Partial Optimal Transport (Partial OT) (Chapel et al., 2020) relaxes the full mass-conservation assumption of standard optimal transport and transports only a prescribed fraction of the available mass. Using the same cost matrix $\mathbf{C} = -\mathbf{S}$, we solve

$$\mathbf{P}^* = \arg\min_{\mathbf{P} \geq 0} \langle \mathbf{P}, \mathbf{C} \rangle - \varepsilon H(\mathbf{P}), \tag{20}$$

subject to

$$\mathbf{P}\mathbf{1} \leq \mathbf{a}, \qquad \mathbf{P}^\top \mathbf{1} \leq \mathbf{b}, \qquad \mathbf{1}^\top \mathbf{P} \mathbf{1} = \tau. \tag{21}$$

Here, τ controls the amount of transported mass. Unlike full transport, Partial OT does not require all query and gallery mass to participate in the coupling, making it more flexible in the presence of unmatched candidates.

B.3 Test-Time Inference Configurations and Complexity

To enable a controlled comparison, we fix each inference rule and its associated hyperparameters across all frozen representations and evaluation settings, including all context sizes. This design isolates the effect of available query context from context-specific retuning. Accordingly, our results characterize the behavior of each rule at a fixed operating point, rather than its optimally tuned performance for every representation or context size. Table 5 summarizes the corresponding configurations and asymptotic inference complexity after the $N_q \times N_g$ similarity matrix has been constructed.

Independent retrieval, CSLS, and IS have costs linear in the similarity-matrix size. GC additionally performs candidate-to-query sorting, resulting in $O(N_q N_g \log N_q)$ complexity. Sinkhorn and Partial OT require iterative global coupling with complexity $O(L N_q N_g)$, where L is the number of transport iterations. Transport solvers use $L_{\max} = 10{,}000$ and a convergence tolerance of 10^{-8}. Hungarian matching performs hard global assignment with complexity $O(N_q^2 N_g)$ for $N_q \leq N_g$. Importantly, these computations are performed only at test time on already extracted embeddings. No method requires retraining or fine-tuning the EEG or image encoders. Using the same fixed configurations across all representations also prevents specific hyperparameter tuning from confounding the comparison of inference behavior.

B.4 CMRE Evaluation Procedure

Algorithm 1 summarizes the complete CMRE evaluation procedure, including context construction, structure-aware inference, and metric aggregation. Table 6 further specifies the context-construction schedule used in our experiments and the resulting relationship between T and R_m.

C Additional Full-Context Results

C.1 Analysis of Full-Context Structural Gains

To characterize the full-context Top-1 gains, we analyze how predictions change relative to independent retrieval. We denote wrong-to-correct and correct-to-wrong transitions by W→C and C→W,

Algorithm 1 Context-Marginalized Retrieval Evaluation (CMRE)

Input: Frozen similarity matrix $\mathbf{S} \in \mathbb{R}^{N_q \times N_g}$, ground-truth targets $\mathbf{y}$, inference rule Φ, context size m, rank cutoff K, and construction rounds T.
Output: $\text{CM-TopK}_\Phi(m)$ and, when defined, $\text{CC}_\Phi(m)$.

// Query-balanced context construction
Initialize $\mathcal{R}_i \leftarrow [\,]$ and $\mathcal{P}_i \leftarrow [\,]$ for all $i = 1, \ldots, N_q$.
if $m = N_q$ **then**
  $T \leftarrow 1$
end if
for $t = 1, \ldots, T$ **do**
  Construct the query-balanced context collection $\mathcal{C}_t^{(m)}$.
  for each context $\mathcal{B} \in \mathcal{C}_t^{(m)}$ **do**
    $\mathbf{S}_\mathcal{B} \leftarrow \mathbf{S}[\mathcal{B}, :]$
    Obtain query-wise gallery rankings $\{\boldsymbol{\pi}_{i,\mathcal{B}}\}_{q_i \in \mathcal{B}}$ from $\Phi(\mathbf{S}_\mathcal{B})$.
    for each query $q_i \in \mathcal{B}$ **do**
      $\rho_{i,\mathcal{B}} \leftarrow \text{rank}_{\boldsymbol{\pi}_{i,\mathcal{B}}}(y_i)$
      $\hat{y}_{i,\mathcal{B}} \leftarrow \text{top1}(\boldsymbol{\pi}_{i,\mathcal{B}})$
      Append $\rho_{i,\mathcal{B}}$ to $\mathcal{R}_i$.
      Append $\hat{y}_{i,\mathcal{B}}$ to $\mathcal{P}_i$.
    end for
  end for
end for

// Metric aggregation
Compute $\text{CM-TopK}_\Phi(m)$ from $\{\mathcal{R}_i\}_{i=1}^{N_q}$.
$R_m \leftarrow |\mathcal{P}_1|$
if $R_m > 1$ **then**
  Compute $\text{CC}_\Phi(m)$ from $\{\mathcal{P}_i\}_{i=1}^{N_q}$.
else
  $\text{CC}_\Phi(m) \leftarrow$ undefined.
end if
return $\text{CM-TopK}_\Phi(m)$ and $\text{CC}_\Phi(m)$ when defined.

Table 6: Relationship between partition rounds and per-query context counts in CMRE.

Context size m	Contexts per round	Query appearances per round	T	R_m
25	8	1	10	10
50	4	1	10	10
100	2	1	10	10
150	4	3	10	30

respectively. For each structural inference method, we further examine the rank of the ground-truth (GT) image under independent retrieval for W→C cases.

Figure 6 reveals distinct prediction-transition patterns across inference rules. CSLS primarily corrects near-miss errors: in 89.56% of its W→C cases, the GT image was already within the independent Top-5, including 53.22% at rank 2. By contrast, 43.93% and 47.93% of the queries corrected by Sinkhorn and Hungarian, respectively, had initial GT ranks beyond 5. Hungarian produces the most W→C transitions but also the most C→W degradations. These patterns show that inference rules exploit the frozen similarity structure differently. Set-level score recalibration mainly resolves near-top ambiguities, whereas globally coupled assignment can recover targets from deeper ranks but may also introduce more errors.

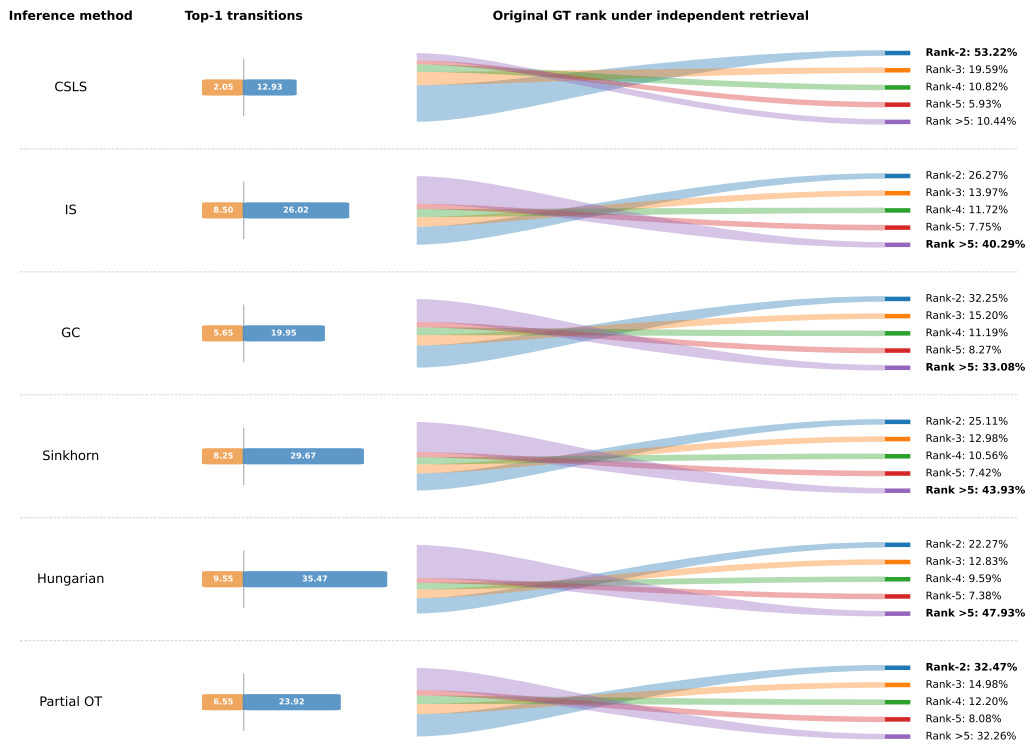


Figure 6: Origin of full-context Top-1 structural gains. For each inference method, the diverging bars show the average numbers of wrong-to-correct (W→C; blue) and correct-to-wrong (C→W; orange) transitions relative to independent retrieval. The flows show the ground-truth (GT) rank under independent retrieval for W→C queries. Results are averaged across six representations and ten LOSO folds.

Table 7: Full-context Top-5 retrieval accuracy (%) across six representations. Mean SE denotes the average gain (pp) over independent retrieval. Top-5 is not reported for Hungarian because hard one-to-one assignment produces a single discrete assignment rather than a ranked candidate list.

Type	Method	NICE	ATM	UBP	BLUR	HyFI	BrainHIVE	Mean SE
Independent	Independent	20.95	30.60	30.20	49.25	36.05	47.10	–
Set-level	CSLS	22.95	36.35	41.80	58.65	50.75	57.85	9.03
	IS	24.10	37.60	47.40	60.95	57.40	65.10	13.07
	GC	24.55	40.35	49.20	58.60	59.60	64.95	13.85
Assignment	Sinkhorn	27.90	41.30	46.50	66.90	56.65	73.15	16.38
	Hungarian	–	–	–	–	–	–	–
	Partial OT	26.30	41.20	42.15	61.95	52.80	69.55	13.30

C.2 Full-Context Top-5 Retrieval Results

To complement the full-context Top-1 results reported in the main paper, we evaluate Top-5 retrieval accuracy under the same frozen similarity matrices. Table 7 reports the results for all inference methods that produce a ranked candidate list.

The Top-5 results show that the benefit of full-context structural inference extends beyond the final Top-1 decision. All structural methods with ranked outputs improve over independent retrieval

across all six representations. Averaged across representations, Sinkhorn achieves the largest Top-5 gain (+16.38 pp), followed by GC (+13.85 pp), Partial OT (+13.30 pp), and IS (+13.07 pp). Sinkhorn obtains the highest Top-5 accuracy on NICE, ATM, BLUR, and BrainHIVE, whereas GC performs best on UBP and HyFI. ATM slightly outperforms UBP (30.60% vs. 30.20%), whereas UBP surpasses ATM under every structure-aware method that produces a ranked candidate list. Likewise, BLUR outperforms BrainHIVE under independent retrieval, BrainHIVE achieves markedly higher Top-5 accuracy under all structure-aware methods except CSLS. Under GC, HyFI also slightly surpasses BLUR (59.60% vs. 58.60%). Together, these results show that structural inference affects not only Top-5 retrieval performance but also the apparent ordering of frozen representations, providing further evidence that representation rankings are inference-dependent.

D Additional Reduced-Context Results

D.1 Detailed Results across Context Sizes

Table 8 reports the complete CM-Top1 accuracy at each reduced query-context size, complementing the context-averaged results in the main paper. Importantly, the inference rule selected as best under full-context evaluation does not remain optimal in 17 of the 24 representation–context-size conditions (70.8%). Relative to selecting the best rule at each reduced context size, using the full-context winner incurs an average performance loss of 2.70 pp and a maximum loss of 8.29 pp. This instability is observed across multiple representations: the full-context winner changes in $1/4$ reduced contexts for NICE, $3/4$ for ATM, $4/4$ for UBP, $4/4$ for BLUR, $3/4$ for HyFI, and $1/4$ for BrainHIVE. These results show that inference-rule selection based on a single full-context operating point does not reliably transfer to reduced-query contexts.

The detailed results further reveal context-response patterns that are obscured by averaging across context sizes. Some methods are comparatively insensitive to m; for example, CSLS on BLUR changes only from 26.13% at $m = 25$ to 26.75% at $m = 150$. In contrast, Hungarian on BLUR increases from 22.02% to 33.36%, Hungarian on HyFI from 16.79% to 28.53%, and Sinkhorn on BrainHIVE from 23.27% to 39.33%. Thus, similar context-averaged performance can mask substantially different sensitivity to the amount of available query context.

D.2 Statistical Significance of Reduced-Context Gains

To assess whether the reduced-context gains are consistent across held-out subjects, we conduct a subject-level significance analysis of CM-Top1 under the LOSO protocol. The ten held-out subjects serve as the units of analysis. Let $m = \{25, 50, 100, 150\}$ denote the set of reduced query-context sizes. For each subject s and inference method Φ, we average the retrieval gain over the six EEG–image representations and all context sizes in

$$\Delta_{s,\Phi} = \frac{1}{6}\sum_{b=1}^{6}\frac{1}{4}\sum\left[\text{CM-Top1}^{(b)}_{s,\Phi}(m) - \text{Top1}^{(b)}_{s,\text{ind}}\right], \tag{22}$$

where $\text{Top1}^{(b)}_{s,\text{ind}}$ denotes the corresponding independent-retrieval performance for subject s on representation b. This produces one aggregated gain for each subject and inference method.

We report the mean subject-level gain together with its 95% bootstrap confidence interval. To test whether the observed improvement differs significantly from zero, we apply a Wilcoxon signed-rank test to the ten subject-level gains. Since six structural inference methods are tested simultaneously, the resulting p-values are corrected using the Holm procedure.

As shown in Table 9, all six structural inference methods achieve positive CM-Top1 gains over independent retrieval. Importantly, the 95% bootstrap confidence intervals of all methods remain entirely above zero. The improvements also remain statistically significant after correcting for the six simultaneous comparisons, with $p_{\text{Holm}} = 0.0117$ for every method. Sinkhorn exhibits the largest mean subject-level improvement, followed by Partial OT and IS, while even the smallest average gain, obtained by GC, remains statistically reliable. These results provide subject-level evidence that the reduced-context gains reported in the main experiments are not attributable to a small subset of held-out subjects. Rather, the benefit of structural inference persists across held-out subjects under the LOSO protocol, with each held-out subject treated as an independent statistical unit.

Table 8: Detailed CM-Top1 accuracy (%) under reduced query contexts $m \in \{25, 50, 100, 150\}$. Each entry reports mean±standard deviation across all ten LOSO folds.

Representation	Method	$m = 25$	$m = 50$	$m = 100$	$m = 150$
NICE	CSLS	$\mathbf{7.67 \pm 2.31}$	7.63 ± 2.40	7.66 ± 2.50	7.58 ± 2.63
	IS	6.99 ± 1.37	7.29 ± 1.69	7.28 ± 1.97	7.67 ± 2.18
	GC	6.88 ± 2.16	7.07 ± 2.09	6.91 ± 2.00	7.01 ± 2.32
	Sinkhorn	7.61 ± 1.76	$\mathbf{8.28 \pm 1.82}$	$\mathbf{8.47 \pm 2.06}$	$\mathbf{8.71 \pm 2.18}$
	Hungarian	7.07 ± 2.24	7.42 ± 2.06	7.55 ± 2.12	7.61 ± 2.38
	Partial OT	7.53 ± 1.68	7.90 ± 1.93	8.14 ± 2.06	8.18 ± 2.29
ATM	CSLS	$\mathbf{13.61 \pm 3.16}$	$\mathbf{13.87 \pm 3.26}$	14.29 ± 3.42	14.24 ± 3.59
	IS	10.29 ± 2.05	12.47 ± 2.61	13.74 ± 2.94	14.04 ± 3.15
	GC	12.33 ± 3.53	13.38 ± 3.50	14.39 ± 3.50	15.01 ± 3.62
	Sinkhorn	11.06 ± 1.94	13.67 ± 2.81	14.99 ± 2.73	$\mathbf{15.40 \pm 2.95}$
	Hungarian	12.50 ± 3.32	13.29 ± 3.24	14.45 ± 3.34	15.20 ± 3.87
	Partial OT	11.30 ± 2.08	13.82 ± 2.71	$\mathbf{15.12 \pm 3.12}$	15.37 ± 3.15
UBP	CSLS	16.30 ± 4.06	16.66 ± 4.08	17.01 ± 4.45	17.19 ± 3.47
	IS	$\mathbf{19.79 \pm 4.07}$	$\mathbf{20.57 \pm 4.10}$	$\mathbf{21.23 \pm 4.39}$	$\mathbf{21.59 \pm 3.59}$
	GC	14.99 ± 3.80	16.79 ± 4.10	19.19 ± 5.57	20.29 ± 4.92
	Sinkhorn	19.04 ± 3.59	19.89 ± 3.44	20.15 ± 3.44	20.22 ± 2.65
	Hungarian	13.95 ± 3.19	15.57 ± 4.02	18.48 ± 5.18	20.91 ± 4.86
	Partial OT	18.11 ± 3.11	18.60 ± 3.15	18.69 ± 2.90	18.71 ± 2.64
BLUR	CSLS	26.13 ± 5.79	26.43 ± 6.32	26.61 ± 6.77	26.75 ± 6.80
	IS	25.81 ± 5.46	28.16 ± 6.09	29.27 ± 7.43	29.82 ± 7.49
	GC	22.90 ± 5.36	24.28 ± 5.86	25.89 ± 5.87	26.64 ± 6.35
	Sinkhorn	$\mathbf{30.13 \pm 5.35}$	$\mathbf{32.54 \pm 5.83}$	$\mathbf{34.14 \pm 5.79}$	$\mathbf{34.66 \pm 5.81}$
	Hungarian	22.02 ± 6.67	24.25 ± 6.85	28.43 ± 6.62	33.36 ± 6.07
	Partial OT	29.57 ± 5.64	30.42 ± 5.89	30.98 ± 5.98	31.00 ± 5.83
HyFI	CSLS	20.99 ± 4.37	21.21 ± 4.94	21.21 ± 5.31	21.18 ± 5.48
	IS	$\mathbf{24.71 \pm 4.27}$	$\mathbf{26.24 \pm 4.40}$	$\mathbf{27.16 \pm 4.78}$	27.44 ± 5.20
	GC	17.90 ± 4.73	20.65 ± 5.04	23.11 ± 5.98	24.80 ± 6.68
	Sinkhorn	24.12 ± 3.86	25.62 ± 4.02	26.71 ± 4.34	26.90 ± 4.58
	Hungarian	16.79 ± 4.41	19.23 ± 4.63	23.71 ± 4.85	$\mathbf{28.53 \pm 5.34}$
	Partial OT	22.95 ± 3.91	23.54 ± 3.89	23.71 ± 4.16	23.60 ± 4.44
BrainHIVE	CSLS	$\mathbf{29.16 \pm 6.47}$	29.59 ± 6.91	29.85 ± 7.24	30.22 ± 7.41
	IS	19.96 ± 3.53	26.30 ± 5.41	31.59 ± 7.02	33.65 ± 8.18
	GC	24.70 ± 4.85	26.29 ± 5.54	28.72 ± 5.85	30.15 ± 6.15
	Sinkhorn	23.27 ± 3.35	$\mathbf{30.65 \pm 5.25}$	$\mathbf{35.83 \pm 6.91}$	$\mathbf{39.33 \pm 8.63}$
	Hungarian	24.24 ± 4.94	26.11 ± 5.56	31.01 ± 7.01	35.67 ± 8.30
	Partial OT	23.48 ± 4.48	30.21 ± 5.53	35.27 ± 7.34	37.95 ± 8.49

D.3 Context-Marginalized Top-5 Results

Table 10 shows that the benefit of structural inference extends beyond Top-1 retrieval. Sinkhorn achieves the largest average CM-Top5 gain across representations (+13.73 pp), followed by Partial OT (+11.78 pp) and IS (+11.53 pp). The preferred method nevertheless varies across representations: Sinkhorn performs best on NICE, BLUR, and BrainHIVE, Partial OT on ATM, and IS on UBP and HyFI. These results indicate that structural inference can improve the broader candidate ranking, while the magnitude of the gain remains strongly dependent on the underlying EEG–image similarity space.

Table 9: Statistical significance of reduced-context CM-Top1 gains over independent retrieval. Each comparison uses the ten held-out subjects as independent observations after averaging over six representations and four reduced query context sizes. Reported p-values are Holm-adjusted across the six inference methods.

Method	Mean SE (pp)	95% bootstrap CI	p_{Holm}
CSLS	+5.14	[4.32, 5.92]	0.0117
IS	+6.39	[5.08, 7.56]	0.0117
GC	+4.61	[3.74, 5.45]	0.0117
Sinkhorn	+7.99	[6.38, 9.45]	0.0117
Hungarian	+5.32	[4.29, 6.33]	0.0117
Partial OT	+6.85	[5.32, 8.12]	0.0117

Table 10: CM-Top5 structural exploitability (SE, pp), averaged over the reduced query-context sizes.

Type	Method	NICE	ATM	UBP	BLUR	HyFI	BrainHIVE	Average
Set-level	CSLS	2.43	4.98	10.81	9.24	14.01	10.95	8.74
	IS	2.80	6.02	**16.56**	9.24	**20.53**	14.01	11.53
	GC	2.27	6.75	15.15	6.71	18.67	12.61	10.36
Assignment	Sinkhorn	**5.55**	7.41	15.52	**15.66**	19.80	**18.45**	**13.73**
	Hungarian	–	–	–	–	–	–	–
	Partial OT	4.43	**7.52**	11.85	12.58	16.91	17.41	11.78

E Additional Analysis of Context Consistency

E.1 Detailed Context Consistency Results

Table 11 reports the complete same-query Context Consistency (CC) results for each reduced query context size $m \in \{25, 50, 100, 150\}$. CSLS exhibits the strongest overall prediction stability, achieving the highest CC for nearly all representation–context combinations. Its consistency generally increases as more companion queries become available, for example from 76.92% to 90.07% on BLUR and from 81.21% to 92.92% on BrainHIVE as m increases from 25 to 150. Partial OT is frequently the second-most consistent method and becomes the best-performing method on HyFI at $m = 150$ (91.70%), slightly exceeding CSLS (91.67%). IS and Sinkhorn also show substantial increases in CC with larger query contexts on most representations.

Hungarian matching exhibits a qualitatively different context-response pattern. Rather than becoming consistently more stable as m increases, its CC often decreases at intermediate or large context sizes. On BLUR, for example, CC drops from 76.69% at $m = 25$ to 63.44% at $m = 100$ and recovers only slightly to 65.12% at $m = 150$. Similar non-monotonic behavior is observed on UBP and HyFI. GC also shows non-monotonic consistency on several representations, although the variation is generally less pronounced. These results show that the context-averaged CC values reported in the main paper can conceal markedly different stability trajectories across individual context sizes.

E.2 Decomposing Context Consistency into Prediction Outcomes

CC measures agreement among predictions for the same query across sampled contexts, but does not indicate whether a consistent prediction is correct. We therefore complement CC with a query-level decomposition of prediction stability. For each target query q_i, evaluated across R_m sampled contexts, we assign its predictions to one of three mutually exclusive categories. A query is Stable Correct if all R_m predictions identify its ground-truth image. It is Stable Wrong if all predictions identify the same non-ground-truth candidate, and Context-Sensitive otherwise. These categories are mutually exclusive and collectively exhaustive.

Table 11: Detailed Context Consistency (CC, %) under reduced query contexts m. Each entry reports mean±standard deviation across ten LOSO folds.

Representation	Method	$m = 25$	$m = 50$	$m = 100$	$m = 150$
NICE	CSLS	$\mathbf{76.95 \pm 2.08}$	$\mathbf{81.30 \pm 1.87}$	$\mathbf{86.36 \pm 1.10}$	$\mathbf{90.61 \pm 1.03}$
	IS	35.96 ± 0.85	46.01 ± 1.23	61.47 ± 1.40	75.67 ± 1.30
	GC	63.04 ± 3.10	58.07 ± 2.23	64.25 ± 2.52	75.74 ± 1.08
	Sinkhorn	37.42 ± 1.36	50.01 ± 1.89	65.68 ± 1.04	79.16 ± 1.30
	Hungarian	73.29 ± 4.58	64.66 ± 4.33	60.57 ± 2.57	64.26 ± 1.82
	Partial OT	40.60 ± 1.14	53.12 ± 1.97	68.93 ± 1.58	80.94 ± 1.40
ATM	CSLS	$\mathbf{78.90 \pm 1.32}$	$\mathbf{82.25 \pm 1.43}$	$\mathbf{86.89 \pm 0.90}$	$\mathbf{91.19 \pm 0.74}$
	IS	31.32 ± 1.38	38.84 ± 1.75	51.58 ± 1.83	67.54 ± 1.46
	GC	76.85 ± 2.09	73.90 ± 1.98	76.07 ± 1.70	82.16 ± 1.40
	Sinkhorn	25.35 ± 1.85	36.08 ± 1.30	48.46 ± 1.07	63.48 ± 1.03
	Hungarian	71.17 ± 3.10	62.87 ± 3.03	59.87 ± 1.54	64.72 ± 0.82
	Partial OT	28.98 ± 1.61	38.55 ± 1.25	51.36 ± 1.17	66.23 ± 1.15
UBP	CSLS	$\mathbf{80.40 \pm 2.43}$	$\mathbf{83.39 \pm 1.78}$	$\mathbf{87.29 \pm 1.36}$	$\mathbf{92.08 \pm 0.59}$
	IS	61.17 ± 2.56	71.94 ± 2.11	81.89 ± 1.76	89.90 ± 1.38
	GC	56.80 ± 2.80	55.10 ± 3.14	63.70 ± 2.20	76.67 ± 1.05
	Sinkhorn	63.27 ± 2.65	73.25 ± 2.51	82.68 ± 2.56	90.09 ± 1.30
	Hungarian	65.90 ± 2.10	60.31 ± 2.84	58.31 ± 2.53	63.61 ± 1.29
	Partial OT	68.77 ± 1.86	77.77 ± 1.36	85.61 ± 2.17	92.01 ± 0.81
BLUR	CSLS	$\mathbf{76.92 \pm 1.79}$	$\mathbf{80.53 \pm 1.32}$	$\mathbf{85.02 \pm 2.15}$	$\mathbf{90.07 \pm 1.83}$
	IS	49.69 ± 3.28	61.81 ± 3.68	75.07 ± 3.57	85.37 ± 2.50
	GC	55.89 ± 3.15	58.75 ± 2.27	68.58 ± 1.47	80.45 ± 1.17
	Sinkhorn	55.19 ± 2.48	67.29 ± 2.56	79.34 ± 1.56	87.85 ± 1.36
	Hungarian	76.69 ± 6.09	67.84 ± 4.29	63.44 ± 3.40	65.12 ± 1.74
	Partial OT	61.53 ± 2.19	72.15 ± 2.22	82.81 ± 1.77	89.57 ± 1.36
HyFI	CSLS	$\mathbf{80.34 \pm 1.98}$	$\mathbf{83.67 \pm 1.88}$	$\mathbf{87.45 \pm 1.35}$	91.67 ± 0.76
	IS	59.20 ± 1.89	69.87 ± 2.09	81.03 ± 1.68	88.78 ± 1.49
	GC	61.56 ± 3.34	59.01 ± 2.59	66.91 ± 1.61	77.90 ± 1.40
	Sinkhorn	60.56 ± 2.09	70.74 ± 1.93	81.41 ± 1.89	89.86 ± 1.22
	Hungarian	69.50 ± 3.54	62.57 ± 3.78	59.71 ± 2.18	64.25 ± 1.79
	Partial OT	67.18 ± 1.75	76.59 ± 1.87	85.30 ± 1.63	$\mathbf{91.70 \pm 1.10}$
BrainHIVE	CSLS	$\mathbf{81.21 \pm 2.02}$	$\mathbf{84.33 \pm 1.53}$	$\mathbf{89.09 \pm 1.35}$	$\mathbf{92.92 \pm 0.77}$
	IS	31.67 ± 1.28	40.18 ± 1.75	54.71 ± 2.38	70.12 ± 2.40
	GC	76.75 ± 1.95	74.57 ± 1.98	78.56 ± 1.67	85.38 ± 1.40
	Sinkhorn	26.18 ± 1.59	38.57 ± 1.41	51.06 ± 1.85	65.14 ± 1.50
	Hungarian	74.52 ± 4.95	66.08 ± 3.90	62.24 ± 2.43	65.86 ± 2.09
	Partial OT	29.71 ± 2.47	41.30 ± 2.65	54.22 ± 2.63	68.57 ± 1.88

Figure 7(a) shows the resulting composition, averaged across four reduced context sizes and ten held-out subjects under the LOSO protocol. High prediction consistency does not necessarily imply stable correctness. CSLS, for example, contains a large Stable-Wrong component across several representations despite achieving the highest average CC in the main experiments. Its Stable-Wrong proportion reaches 55.5% on NICE, 52.1% on ATM, and 50.8% on UBP, while the corresponding Stable-Correct proportions are substantially lower. Thus, much of the apparent stability of CSLS arises from repeatedly selecting the same incorrect candidate rather than consistently retrieving the ground truth.

Figure 7(b) further resolves these components by context size. For CSLS, IS, Sinkhorn, and Partial OT, increasing m generally raises the Stable-Correct proportion and reduces the Context-Sensitive proportion. Hungarian exhibits a different pattern: its Context-Sensitive proportion often increases with m, without a corresponding improvement in Stable-Correct predictions. This pattern is consis-

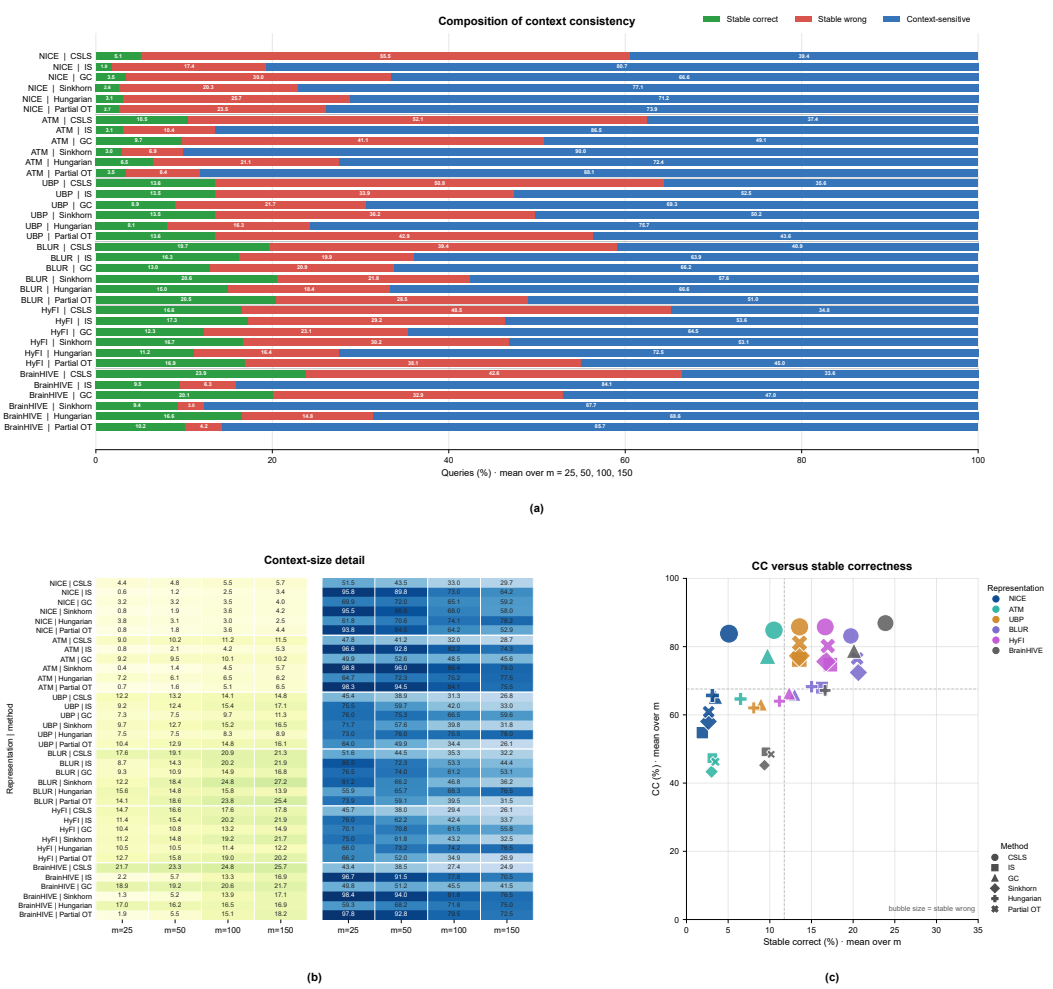


Figure 7: Decomposition of same-query Context Consistency. Queries are categorized as Stable Correct, Stable Wrong, or Context-Sensitive according to their CM-Top1 predictions across sampled query contexts. (a) Average proportions of the three categories across m. (b) Stable-Correct and Context-Sensitive proportions at each context size. (c) Average CC versus Stable-Correct rate, where color denotes the representation, marker shape denotes the inference method, and bubble size represents the Stable-Wrong rate.

tent with the non-monotonic or decreasing CC profiles of Hungarian observed in the main experiments.

Figure 7(c) relates average CC to the Stable-Correct proportion across representation–method combinations. Bubble size represents the Stable-Wrong proportion, allowing consistency, stable correctness, and persistent error to be examined jointly. The preferred operating region is the upper right with small bubbles, indicating high CC, a high Stable-Correct proportion, and little persistent error. Conversely, high CC combined with a low Stable-Correct proportion and large bubbles indicates repeated selection of the same incorrect prediction.

These results show that CC alone does not reveal the source of prediction stability across query contexts. Decomposing predictions into Stable-Correct, Stable-Wrong, and Context-Sensitive categories distinguishes consistent success, persistent error, and sensitivity to query composition.

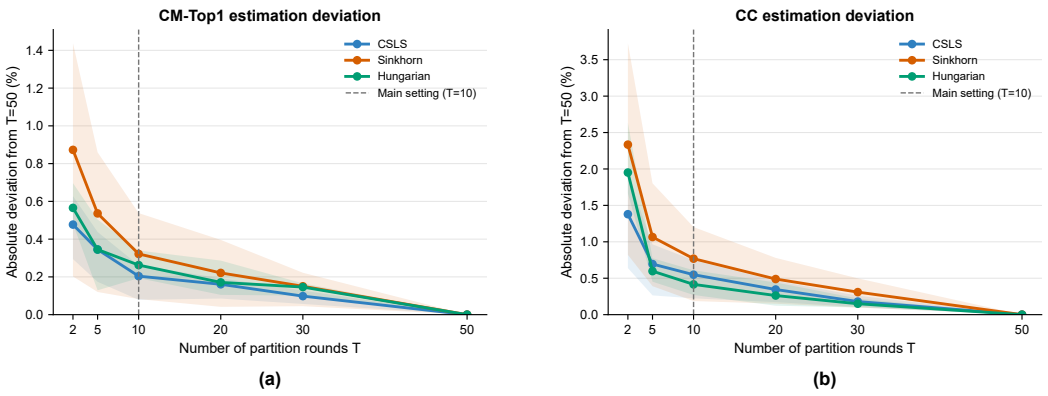


Figure 8: CMRE convergence over context-construction rounds on BLUR. Absolute deviations of (a) CM-Top1 and (b) Context Consistency (CC) from the $T = 50$ reference as the number of partition rounds increases. Curves correspond to CSLS, Sinkhorn, and Hungarian and are averaged across context sizes. The dashed line marks the $T = 10$ setting used in the main experiments. Both metrics are already close to the high-budget reference at $T = 10$.

F CMRE Convergence over Context-Construction Rounds

We conduct the convergence analysis on BLUR using three representative inference rules with distinct context-response characteristics: CSLS, which exhibits consistently high Context Consistency; Sinkhorn, which achieves the largest average reduced-context retrieval gain; and Hungarian, which shows pronounced context-dependent variation in both retrieval performance and prediction consistency. For each reduced context size $m \in \{25, 50, 100, 150\}$, we generate a fixed schedule of 50 random partition rounds and evaluate progressively larger prefixes with $T \in \{2, 5, 10, 20, 30, 50\}$, where T denotes the number of partition rounds. The estimate obtained with $T = 50$ is treated as a high-budget reference.

For inference method Φ, we measure the average absolute deviation of CM-Top1 from the reference as

$$D^{\Phi}_{\mathrm{Top1}}(T) = \frac{1}{4} \sum \left| \mathrm{CM\text{-}Top1}_{\Phi,T}(m) - \mathrm{CM\text{-}Top1}_{\Phi,50}(m) \right|, \tag{23}$$

and analogously define the deviation of Context Consistency as

$$D^{\Phi}_{\mathrm{CC}}(T) = \frac{1}{4} \sum \left| \mathrm{CC}_{\Phi,T}(m) - \mathrm{CC}_{\Phi,50}(m) \right|. \tag{24}$$

Figure 8 shows that both estimates rapidly approach the high-budget reference as the number of partition rounds increases. At the $T = 10$ setting used in the main experiments, the average CM-Top1 deviations are only 0.204, 0.322, and 0.263 pp for CSLS, Sinkhorn, and Hungarian, respectively. The corresponding CC deviations are 0.549, 0.769, and 0.415 pp. Thus, even for inference rules with substantially different context-response behaviors, the estimates obtained with 10 partition rounds are already close to those obtained with 50 rounds. These results support the partition-round budget used in the main experiments. Increasing the number of rounds beyond $T = 10$ produces only small changes in the reported CMRE estimates. This analysis evaluates prefix convergence within a shared 50-round partition schedule, rather than variability across independently resampled schedules.

G Robustness to Gallery Expansion

The standard THINGS-EEG2 retrieval protocol uses a closed gallery with $N_q = N_g = 200$, containing one ground-truth image for each held-out EEG query. To examine whether structural gains depend on this setting, we keep the 200 queries and their ground-truth images fixed while expanding the gallery to $N_g \in \{200, 500, 1000, 2000\}$. We construct a fixed pool of 1,800 distractors from unused images in the original THINGS collection. We exclude all images used in the THINGS-EEG2

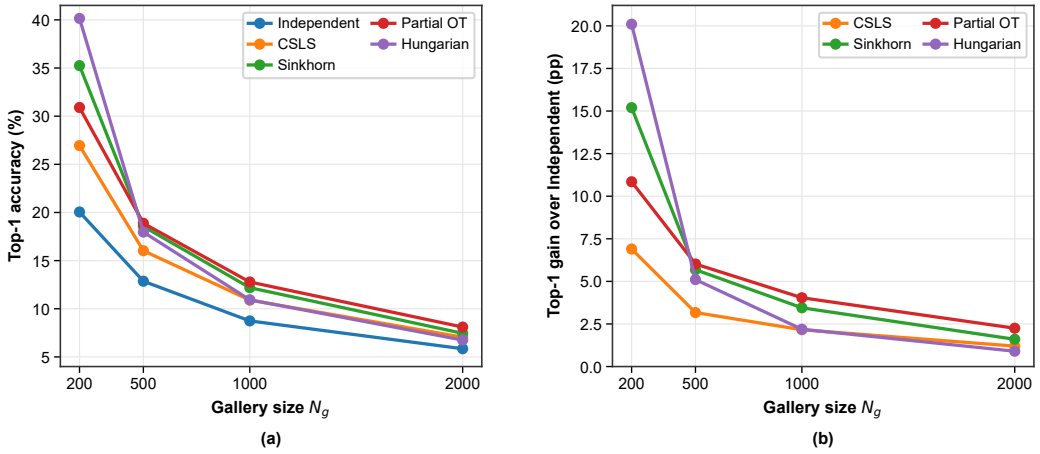


Figure 9: Robustness to gallery expansion on BLUR. (a) Top-1 retrieval accuracy as the gallery size increases from $N_g = 200$ to $N_g = 2000$ while keeping all 200 EEG queries fixed. (b) Corresponding Top-1 gain over independent retrieval.

training and test sets and all images belonging to the 200 test concepts, ensuring that the distractors neither duplicate evaluated concepts nor have corresponding EEG queries. Expanded galleries are formed by adding subsets of this pool; at $N_g = 2000$, all 1,800 distractors are included. The EEG and image representations remain frozen throughout, so only the test-time gallery composition changes.

Figure 9 summarizes the BLUR results. As expected, absolute Top-1 accuracy decreases for every method as N_g increases, reflecting greater retrieval difficulty (Figure 9(a)). Structure-aware inference nevertheless remains superior to independent retrieval across all expanded-gallery settings, although its gains decrease substantially with N_g. Hungarian achieves the largest gain in the original closed gallery, improving Top-1 accuracy by 20.10 percentage points at $N_g = 200$. Its gain falls to 5.11, 2.19, and 0.90 percentage points at $N_g = 500$, 1000, and 2000, respectively. By contrast, Partial OT retains the largest gain after expansion, achieving 6.02, 4.04, and 2.25 percentage points at the three larger gallery sizes, while Sinkhorn and CSLS show intermediate behavior.

These results indicate that the effectiveness of structural inference depends not only on the frozen similarity representation but also on the structural assumptions induced by the test gallery. Hard one-to-one assignment benefits strongly from the original closed-gallery setting, whereas partial transport is more robust when additional unmatched gallery candidates are introduced.

H Discussion and Limitations

CMRE characterizes how a fixed inference rule behaves when the query context available at test time changes. This perspective complements conventional full-context evaluation, but it does not model every plausible deployment setting. In particular, our contexts are generated through balanced random constructions. Real query streams may instead be temporally correlated, non-uniform, or task dependent. Although the balanced construction controls the number of appearances of each query, sampled contexts overlap and therefore do not constitute independent observations.

We intentionally keep each inference rule and its hyperparameters fixed across frozen representations and context sizes. This isolates the effect of available query context from context-specific retuning. Consequently, our results describe the behavior of the evaluated rules at shared fixed operating points, rather than their optimally tuned performance under every representation or context size. Moreover, Context Consistency measures agreement among predictions across contexts, not their correctness. A method can therefore be highly consistent while making consistently incorrect predictions.

Our empirical study focuses primarily on THINGS-EEG2. The complementary THINGS-MEG evaluation includes four participants and two frozen representations, and should therefore be interpreted as supporting evidence rather than a comprehensive cross-modal benchmark. The gallery-expansion analysis further shows that the gains of hard one-to-one assignment depend strongly on the closed, bijective gallery assumption. Its behavior is less favorable when unmatched distractors are introduced. Extending CMRE to larger cohorts, broader neural recording modalities, more realistic query streams, and non-bijective retrieval settings is an important direction for future work.