BIGEM: BIDIRECTIONAL EEG–IMAGE GENERATIVE MODELING FOR PROBING VISUAL REPRESENTATIONAL GEOMETRY

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding the correspondence between visual stimuli and brain activity requires models that characterize the representations linking them. However, the mappings between these two modalities are typically modeled independently. We introduce BiGEM, a bidirectional generative framework that connects EEG-to-image decoding and image-to-EEG generation through a shared visual latent space. On THINGS-EEG2, BiGEM achieves 49.3% Top-1 accuracy in 200-way visual retrieval from EEG and a Pearson correlation of 0.456 between generated and measured mean EEG responses. Bidirectional analysis reveals that stimulus identity is preserved most strongly in the shared latent representation, while joint refinement improves reciprocal consistency without improving direct decoding. The learned EEG representation aligns most strongly with intermediate stages of the visual hierarchy, with early EEG responses associated with spatial phase, texture, and edge orientation. Controlled perturbations further dissociate two complementary properties of generated neural responses: perturbation severity determines response magnitude, whereas the identity of perturbation families remains recoverable from spatiotemporal response organization after severity control. This study provides a unified framework for modeling and studying the shared representations linking visual and neural signals.

1 INTRODUCTION

Visual decoding and neural response modeling offer complementary approaches to understanding how the brain represents visual information. Decoding asks what perceptual information can be recovered from measured brain activity, whereas encoding and generative models characterize how visual stimuli are reflected in evoked neural responses (Gifford et al., 2022; Naselaris et al., 2011; Cohen, 2017). Yet predictive performance alone does not reveal how relationships among stimuli are organized in neural response space. Representational geometry captures this organization by describing the relational structure of responses evoked by stimuli and providing a common framework for comparing neural and computational representations (Kriegeskorte & Kievit, 2013; Kriegeskorte & Wei, 2021; Kriegeskorte et al., 2008). In the reverse direction, models that generate EEG from images have extended neural encoding from deterministic response prediction to generative synthesis of activity evoked by stimuli (Gifford et al., 2022; Xu et al., 2025). These models are typically optimized independently of decoding models and therefore do not explicitly characterize the representational relationship between the two directions.

To address these limitations, we introduce BiGEM, a bidirectional generative framework for EEG and images for probing visual representational geometry. First, a shared visual latent space links EEG-to-image reconstruction and EEG generation conditioned on images while retaining output models for each modality. Frozen bidirectional compositions determine where information is preserved or lost across reciprocal transformations, separating correspondence in the shared latent space from rendered image and EEG output fidelity. We find that stimulus identity is preserved substantially more strongly in the shared latent representation than after generation in each modality, while joint refinement improves reciprocal consistency without improving direct decoding accuracy. Second, we characterize visual geometry in learned EEG representations and measured neural responses. These analyses distinguish the organization learned by the EEG Encoder from visual

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.

structure expressed in its neural input. The learned EEG representation shows its strongest correspondence with intermediate visual hierarchy stages, whereas measured EEG exhibits early conditional associations with spatial phase, texture, and edge orientation. Finally, we probe the fixed generator using controlled perturbations of nine visual feature families. We trace changes in its native visual representation into generated EEG and distinguish response magnitude from organization across time and sensors. Across perturbation families, effective perturbation severity primarily determines how much the generated EEG response changes, whereas perturbation-family identity remains recoverable from how that change is organized across sensors and time after severity control. Together, these results show that bidirectional generative modeling can reveal where correspondence between visual and neural representations is preserved, what visual structure it contains, and how that structure is organized across representations, sensors, and time.

2 Related Work

Visual Decoding. NICE and ATM-S align EEG with pretrained visual representations for stimulus retrieval (Song et al., 2024; Li et al., 2024), while approaches based on diffusion extend these representations toward image reconstruction (Bai et al., 2024; Li et al., 2024; Ferrante et al., 2024). UBP reduces the mismatch between EEG and fine visual detail through a blur prior (Wu et al., 2025). These methods primarily optimize the neural-to-visual direction. **Modeling EEG from Visual Input**. Classical encoding models relate hierarchical visual features to measured neural activity (Güçlü & van Gerven, 2015; Yamins et al., 2014). Xu et al. introduce M/EEG generation conditioned on images with multimodal diffusion (Xu et al., 2025). Recent approaches further extend this direction toward stochastic and structured neural-response generation, but remain largely independent of visual decoding models. **Bidirectional Visual–Neural Modeling**. MB2C introduces cycle consistency between EEG and visual feature distributions (Wei et al., 2024), while D^2-FOSA reconstructs neural and visual embeddings in both directions (Yu et al., 2026). NeuroFlow models visual encoding and decoding as reversible transformations within a shared generative process in fMRI (Mai et al., 2026). Existing EEG approaches mainly use bidirectionality for feature alignment or reconstruction, rather than evaluating where information is preserved or lost across complete reciprocal mappings. **Visual Representational Geometry**. Lu and Golomb identify distinct temporal profiles for multiple visual dimensions in measured EEG (Lu & Golomb, 2025), and ReAlnet uses EEG prediction to align internal visual network representations with human neural responses (Lu et al., 2026). Bashivan et al. established model-guided image synthesis as a tool for probing neural population responses (Bashivan et al., 2019). These studies motivate moving from observational correspondence toward controlled probing of representational structure. BiGEM connects these perspectives through bidirectional generation, representational analysis, and controlled visual perturbation.

3 Method

BiGEM combines bidirectional generation, representational characterization, and controlled perturbation probing (Figures 1 and A1).

3.1 Bidirectional EEG and Image Generative Modeling

Shared Visual Latent Alignment. The frozen OpenCLIP ViT-H/14 encoder E_I defines a 1,024-dimensional visual space (Radford et al., 2021; Cherti et al., 2023), into which a subject-specific EEG encoder E_E maps EEG. Training adaptively selects each image's target from cached low-, medium-, and high-blur features based on EEG–image alignment scores (Wu et al., 2025). Image-to-EEG conditioning and the base retrieval gallery use a fixed medium-blur reference:

$$z_I = \mathcal{N}(E_I(B_{\text{med}}(I))), \qquad z_E = \mathcal{N}(E_E(X)) \tag{1}$$

Here, I is an image, X its EEG response, $\mathcal{N}$ denotes L2 normalization, and B_{med} applies medium blur. Training inputs average four recorded trials per image. In perturbation analyses, native visual representations are image-encoder outputs before and after perturbation. Selection, modulation, and implementation details appear in Appendix A.

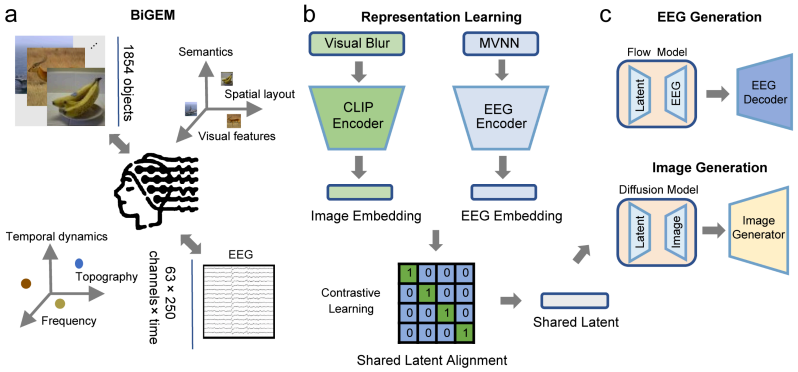

Figure 1: Overview of BiGEM. (a) BiGEM links natural visual stimuli and EEG responses across visual and neural representational dimensions. (b) In Representation Learning, the OpenCLIP image encoder (ViT-H/14) and EEG Encoder project image and EEG representations into a shared visual latent space through contrastive alignment. (c) The shared latent space supports bidirectional generation. In EEG Generation, it is transformed by the EEG Flow and reconstructed by the EEG Decoder. In Image Generation, it is passed through the Diffusion Model and Image Generator to synthesize images.

EEG-to-image Generation. A visual adapter A converts the shared EEG representation z_E into IP-Adapter conditioning tokens for frozen SDXL (Ye et al., 2023; Podell et al., 2024):

$$\widehat{I} = G_I(z_E) = \mathrm{SDXL}(A(z_E)) \tag{2}$$

Text conditioning and sampling noise are omitted from the notation. The adapter is pretrained by distillation to match native ViT-H/14 conditioning using denoising, representation, attention, and retrieval objectives. Afterward, the adapter and SDXL remain frozen during inference and joint refinement.

Image-to-EEG Generation. A subject-specific rectified-flow sampler F_s transforms Gaussian noise ϵ, conditioned on z_I, into 1,024 standardized principal component analysis (PCA) coefficients (Liu et al., 2023). A fixed subject-specific PCA decoder D_s reconstructs EEG:

$$\widehat{X} = G_E(z_I, \epsilon) = D_s(F_s(\epsilon; z_I)), \qquad \epsilon \sim \mathcal{N}(0, \mathbf{I}_{1024}). \tag{3}$$

Different noise draws produce stochastic trials for the same image. Training combines flow-velocity matching with waveform and semantic losses.

Bidirectional Composition and Joint Refinement. Frozen compositions evaluate the full image $\rightarrow$ EEG $\rightarrow$ image and EEG $\rightarrow$ image $\rightarrow$ EEG paths, separating latent correspondence from rendered output fidelity (Figure A1). Joint refinement instead optimizes latent-space cycles without backpropagating through image rendering:

$$z_I \rightarrow \widehat{X} \rightarrow \widehat{z}_I, \qquad X \rightarrow z_E \rightarrow \widehat{X} \tag{4}$$

The joint objective is:

$$\mathcal{L}_{\text{joint}} = \mathcal{L}_{\text{direct}} + \lambda_{\text{cyc}}(\mathcal{L}_{\text{image cycle}} + \mathcal{L}_{\text{EEG cycle}}) \tag{5}$$

Here, $\mathcal{L}_{\text{direct}}$ combines contrastive alignment and direct EEG-generation losses. The image cycle compares returned and conditioning visual representations using cosine, contrastive, and geometry terms, whereas the EEG cycle compares regenerated and measured trial groups through mean-response, covariance, and variability losses. Joint refinement updates the EEG Encoder and EEG generator, while the visual adapter and SDXL remain frozen outside the optimization graph. Independent training defines the principal BiGEM system; joint endpoints are evaluated separately under a common sampling manifest.

3.2 Visual Representational Geometry Characterization

We characterize visual geometry in learned EEG representations and measured EEG, and assess encoder input dependence. First, Spearman representational similarity analysis (RSA) compares representational dissimilarity matrices (RDMs) of normalized EEG embeddings and layerwise features

from a frozen ResNet-50 probe (He et al., 2016), separate from the ViT-H/14 alignment and image-evaluation encoders. Partial RSA adjusts for nine image-feature RDMs. Second, deterministic descriptors of unperturbed images cover nine families: luminance, contrast, chroma/opponent color, spatial frequency, spatial phase, edge orientation, texture, global layout, and clutter/complexity. Ridge regression jointly relates their RDMs to measured EEG RDMs across temporal windows and sensor groups. Upper-triangle RDM entries are rank-transformed and standardized before fitting; coefficients quantify associations conditional on the other families. Finally, occlusion probes the frozen EEG encoder by zeroing individual 40 ms windows or sensor groups without input renormalization. Spectral occlusion zeros Fourier coefficients in consecutive 5 Hz bands before inverse transformation. Retrieval degradation is the full-input score minus the occluded-input score. Appendix A.3 details feature definitions and statistical corrections.

3.3 Controlled Visual Perturbation Probing

The nine feature families are perturbed at three doses (Weak, Medium, Strong) in 200 test images. The frozen image-to-EEG pathway generates 80 trials per subject–image condition with matched noise across conditions. Native embedding displacement is:

$$D_{\text{vis}} = 1 - \cos(z_I, z_I^{\text{pert}}) \tag{6}$$

where z_I^{pert} is the perturbed-image embedding under the original preprocessing. Identity retention uses the original 200-image gallery; learned perceptual image patch similarity (LPIPS; Zhang et al., 2018) measures perceptual change. EEG change is the root mean square (RMS) difference between perturbed and original mean waveforms within six sensor groups $\times$ 25 consecutive 40 ms windows. Overall magnitude averages these 150 cells over 0–1000 ms.

For each subject and family, image-averaged cellwise changes are regressed on doses 0–3 (0: unperturbed), and slopes are L1-normalized across cells. Their across-family mean defines the shared map; subtracting the other eight families' mean from each map yields relative maps. Spearman matching to templates averaged from the other nine subjects identifies perturbation families. Full normalized slope maps are primary; early and relative maps are secondary. At Strong dose, within-subject, within-family regressions relate magnitude to standardized displacement and LPIPS across images. Across families and doses, nested magnitude models include these predictors and their separate quadratic terms, then add family identity and family-by-displacement interactions. Fivefold image-grouped cross-validation holds out entire image identities, uses training-only standardization, and computes predictive R^2 from pooled held-out errors. For strict severity-controlled identification, each outer fold fits cellwise ordinary least squares models to the other nine subjects using displacement, LPIPS, and their separate quadratic terms, without family labels. The same models residualize query and template responses; template preprocessing also excludes the query subject. Residuals are averaged within each subject and family over images and nonzero doses, then L2-normalized across cells. These residual-mean maps are distinct from the primary L1-normalized slope maps. Statistical procedures, permutation tests, and multiplicity correction appear in Appendix A.

4 Experiments

4.1 Dataset and Evaluation

Experiments were conducted on THINGS-EEG2 (Gifford et al., 2022), which contains EEG recordings from ten subjects viewing natural object images. The training partition comprises 1,654 object concepts with ten exemplars per concept (16,540 images), each presented four times to every subject. The test partition contains 200 held-out concepts with one image per concept, each presented 80 times. Training and test stimuli are kept strictly separate. EEG signals from 63 channels are bandpass filtered from 1 to 100 Hz, with a notch filter at 50 Hz, and baseline corrected over the 200 ms interval before stimulus onset. The 1 s interval after stimulus onset is retained, resampled to 250 Hz, and multivariate noise normalization (MVNN) is applied, yielding a 63$\times$250 matrix of channels by time. Alignment of EEG with visual representations is evaluated using 200-way retrieval. Test EEG responses are averaged over 80 repetitions, and retrieval is performed using the trained blur-aware scoring rule. Top-1, Top-5, and mean reciprocal rank (MRR) are reported, with 0.50% Top-1 chance. Blur-aware and nonblurred targets use the same protocol. Generated images are evaluated

using a separate frozen OpenCLIP image encoder (ViT-H/14) (Radford et al., 2021; Cherti et al., 2023). Top-1 measures stimulus identity recovery; CLIP cosine similarity, LPIPS (Zhang et al., 2018), and SSIM (Wang et al., 2004) measure feature, perceptual, and structural correspondence. Frozen cycles starting from images evaluate retrieval both in the returned latent and after rendering. EEG conditioned on images is evaluated by averaging K = 80 independently generated trials and comparing this response with the mean of 80 measured repetitions. Pearson correlation, cosine similarity, and mean squared error (MSE) quantify waveform agreement. The trial dispersion ratio is the ratio of mean pairwise Euclidean distances among generated and measured trials in the retained standardized PCA space.

4.2 Direct Retrieval and Generation

Table 1: 200-way visual retrieval from EEG.

Method	Ch.	Top-1 (%)	Top-5 (%)	Method	Ch.	Top-1 (%)	Top-5 (%)
BraVL	17	5.80	17.50	D^2-FOSA	63	37.96	70.67
NICE	63	16.10	43.60	VE-SDN	63	39.64	71.23
ATM-S	63	28.50	60.40	UBP	17	47.45	75.25
MB2C	63	28.45	60.37	**BiGEM**	63	**49.30**	**78.55**

BiGEM achieved 49.30% Top-1 accuracy using the full input from 63 channels (Table 1). This performance was closely associated with adaptive blur-aware alignment (Figure 1). BiGEM selects and modulates targets with low, medium, and high blur within the OpenCLIP ViT-H/14 space, attenuating fine details weakly recoverable from EEG. Replacing the blur-aware target with the nonblurred representation reduced Top-1 to 37.95% (Table 4a). These results indicate that adaptive control of visual detail strengthens correspondence at the stimulus level for downstream bidirectional generation. For EEG-to-image generation, BiGEM showed stronger paired reconstruction metrics in Table 2, while maintaining competitive image quality at the distribution level. Unlike approaches based primarily on direct latent regression, the visual adapter is designed to preserve the structure and semantic usability of the shared representation rather than minimize latent error at each coordinate (Table 4b). This result suggests that maintaining representational structure relevant to the generator is important for translating the EEG latent with high retrieval performance into visually faithful reconstructions. Full qualitative atlases are in Appendix B.2, Figures A3a and A3b. See Table A3b for extended component ablations.

Table 2: EEG-to-image reconstruction. PCC, Pearson correlation coefficient; SSIM, structural similarity; IS, Inception Score; FID, Fréchet Inception Distance. Sources: Appendix A.2.6.

Method	Top-1 (%) ↑	PCC ↑	SSIM ↑	IS ↑	FID ↓	CLIP cosine ↑
ATM-S	2.62	.160	.345	—	—	.567
MB2C-SDXL	—	.188	.333	10.19	163.94	—
D^2-FOSA	—	.193	.350	11.81	146.33	—
BiGEM	**7.10**	**.204**	**.382**	**13.36**	**147.49**	**.703**

In the reverse direction, BiGEM leverages the stimulus-discriminative shared representation supporting 49.30% Top-1 visual retrieval from measured EEG (Table 1) to condition EEG generation and provide latent-semantic supervision. Conditional flow combines this supervision with waveform constraints to generate variable, stimulus-dependent trials. Generated averages achieved $r = 0.4555$, with a near-unit trial-dispersion ratio (1.0101) in the retained PCA space. Deterministic prediction yielded higher mean-response correlation but no trial variability. An exploratory ATM-S readout not used in generator training recovered stimulus identity at 13.55%; mismatched conditioning reduced this to 0.25% (chance: 0.50%) and lowered r to 0.4233 (Table A2). These findings demonstrate stimulus-information retention alongside trial variability. EEG-Encoder retrieval measures preservation within the training-aligned latent space, not independent validation. Table 3 provides contextual comparisons with external methods.

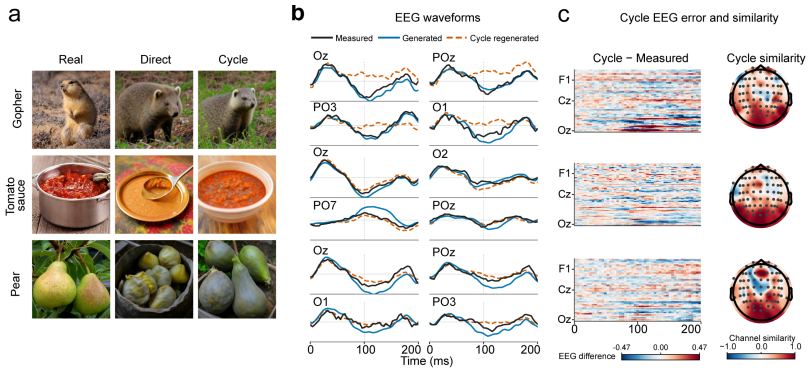

Figure 2: Qualitative bidirectional generation and cycle reconstruction in BiGEM. (a) Representative target images (Real) and EEG-to-image reconstructions from the independently trained endpoint (Direct) and the joint-refined endpoint (Cycle). (b) Measured EEG, image-conditioned generated EEG, and EEG regenerated through the cycle for the same three examples. Four posterior channels are shown for each example over 0–200 ms, selected by measured RMS amplitude. (c) Cycle reconstruction error and similarity for the corresponding examples. Heatmaps show Cycle − Measured EEG differences across channels over 0–200 ms, and scalp maps show channelwise similarity between cycle-regenerated and measured EEG over 0–1000 ms.

Table 3: Contextual results for image-to-EEG generation on THINGS-EEG2. Sources: Appendix A.2.6.

Method	**MSE ↓**	**Pearson r ↑**	**Cosine ↑**
Güçlü-style	.211	.178	.181
Yamins-style	.196	.203	.209
SynBrain	.156	.366	.347
Xu et al.	.109	.425	.402
BiGEM	**.04375**	**.4555**	**.4558**

Table 4: Ablation of core BiGEM components. (a) Blur-aware alignment. (b) visual adapter.

a Alignment targets

Targets	**Top-1 (%)**	**Top-5 (%)**	**MRR**
Nonblurred	37.95	69.10	.5223
Blur-aware	**49.30**	**78.55**	**.6250**

b Visual adapter

Adapter	**Top-1 (%)**	**LPIPS ↓**	**SSIM ↑**
Base MLP	4.35	.7663	.2773
Visual adapter	**7.10**	**.7402**	**.382**

4.3 Cycles Separate Latent Preservation from Output Fidelity

After establishing effective direct generation in both directions, we composed the independently pretrained translators without joint refinement to examine how information survives complete bidirectional translation (Figure A1; Table 5). Stimulus identity was strongly preserved in the returned visual latent (97.15% Top-1), whereas recovery after rendering decreased to 8.90%. EEG reconstructed through EEG → image → EEG retained r=0.4348, close to direct generation (r=0.4555). These results reveal a clear separation between latent correspondence and final output fidelity: generation in each modality introduces an additional information bottleneck. The joint stage endpoints showed convergent improvements in round-trip information preservation (Table 5). Consistency of cycles starting from images increased by 15.2% (q=0.0081), while consistency of cycles starting from EEG increased by 9.0% (q=0.0215). EEG cycle MSE decreased by 29.1%, and round-trip EEG Encoder Top-1 improved by 14.1%. These joint stage gains were not accompanied by improved direct visual retrieval from EEG. These joint endpoints show better compatibility between directions and round-trip recovery.

Table 5: Frozen cycle outputs and joint refinement. CI, confidence interval; BH, Benjamini–Hochberg.

Frozen cycle readout	**Returned latent**	**Rendered image**	**Returned EEG**		
Top-1 / Pearson	97.15%	8.90%	0.4348		
Joint stage metric	**Independent**	**Cycle-active**	**Paired Δ**	**95% CI**	**BH q**
Image cycle cosine	.2707	.3118	+.0412	[.0390, .0433]	.0081
EEG cycle Pearson r	.4014	.4376	+.0362	[.0185, .0529]	.0215
Direct Top-1 %	49.30	48.85	−.45 pp	[−1.25, .30]	.386719

4.4 Learned Representations and Measured EEG Reveal Complementary Visual Structure

Having examined information preservation across the two directions, we next asked what visual organization was captured by the learned EEG representation across the ten subjects. Within the independently probed ResNet-50 hierarchy, the EEG Encoder representation showed its highest mean correspondence with the intermediate stages (ρ=0.3265, 95% CI [0.3137, 0.3397]); this ordering was retained through joint refinement. After controlling the nine visual feature families, partial RSA remained strongest at layer3 (ρ=0.2437, 95% CI [0.2248, 0.2600]; Figure A6b). Thus, the learned EEG representation retained visual structure at intermediate levels beyond the variance captured by the manually defined feature set.

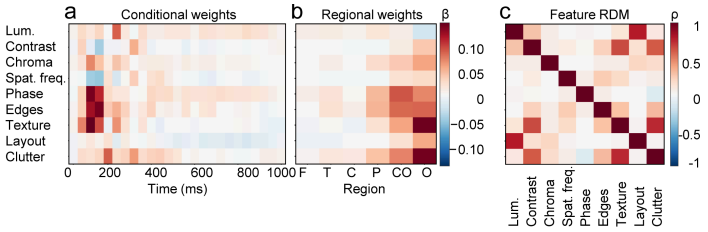

Figure 3: Conditional visual feature associations in measured EEG. (a,b) Mean ridge coefficients across ten subjects after jointly modeling all nine image feature RDMs against measured EEG RDMs, resolved over time and sensor region. (c) Pairwise Spearman correlations between image feature RDMs. Panels a and b share the coefficient scale on the right (β), c uses a separate correlation scale (ρ), with rows and columns following the feature order in a. F, frontal; T, temporal; C, central; P, parietal; CO, centro-occipital; O, occipital.

We next examined visual structure expressed directly in measured EEG. Because the visual descriptors were correlated, all nine feature RDMs were entered jointly into a ridge model after rank transformation and standardization. The strongest conditional associations were spatial phase at 80–120 ms (β=0.1339, q=0.0251), texture at 80–120 ms (β=0.1419, q=0.0176), and edge orientation at 120–160 ms (β=0.1335, q=0.0176; Figure 3). Regional analyses independently showed posterior associations (corrected q≈0.0083). Temporal and regional summaries are interpreted separately rather than as a single jointly localized spatiotemporal peak. Coefficients describe conditional associations rather than independent responses selective for features. Input occlusion showed that retrieval relied predominantly on posterior, temporally distributed, and low-frequency EEG information. The largest temporal losses followed removal of the 200–240 ms and 80–120 ms intervals. Occipital occlusion reduced Top-1 by 34.15 percentage points; removing 0–5 Hz and 5–10 Hz reduced it by 22.80 and 24.15 points. These analyses characterize learned representational geometry, recorded EEG structure, and model dependence without implying that a single visual feature propagates causally through the complete model.

4.5 Controlled Perturbations Dissociate Response Magnitude from Spatiotemporal Organization

Having characterized visual structure in learned representations and measured EEG, we next probed how controlled perturbations propagate from the native visual representation into generated EEG (Figures 4 and 5). Strong edge orientation and spatial phase manipulations produced large displacements and low identity retention, whereas spatial frequency and texture manipulations caused smaller changes. All 90 dose-response slopes for subject and feature combinations were positive. Across stimuli, native displacement was strongly associated with EEG RMS change for representative perturbations (ρ=0.837–0.942), after accounting for LPIPS, embedding coefficients were positive in all ten subjects for every feature family, adding fitted R^2 of 0.355–0.742.

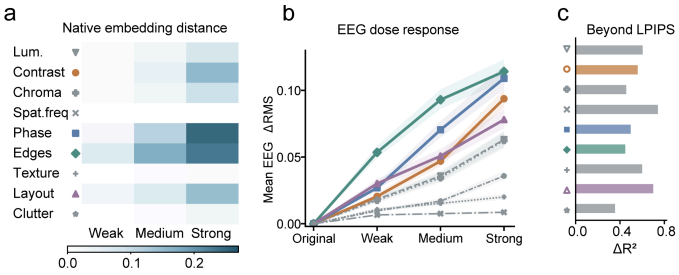

Figure 4: Controlled visual perturbations propagate through the native visual representation into generated EEG. (a) Native embedding displacement across perturbation families and doses. (b) EEG response magnitude increases with perturbation dose across all nine feature families. (c) Additional fitted R^2 from native displacement beyond LPIPS, with feature order as in a. All nine feature families are included in quantitative analyses.

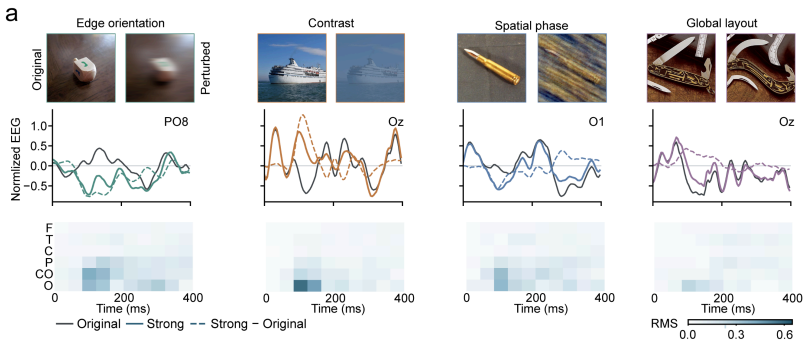

Figure 5: Generated EEG responses to Strong perturbations in individual examples. Columns show edge orientation, contrast, spatial phase, and global layout. Top: original and perturbed stimuli. Middle: K = 80 mean waveforms at the displayed peak channel, in preprocessed whitened EEG units: original (gray), Strong (colored), and Strong minus Original (dashed). Bottom: RMS differences across six sensor groups and 40 ms bins over 0–400 ms. Additional examples for all nine feature families appear in Figure A12.

Response magnitude, however, did not fully determine spatiotemporal organization (Figures 6 and 7). Normalized maps were highly similar across families (ρ=0.892), with a shared occipital peak at 80–120 ms, yet posterior relative deviations for each feature remained after global maxT correction. Identification across subjects reached 37.78% for the full map and 52.22% over 0–400 ms (chance 11.11%, p=0.0001).

Effective perturbation severity explained most of the variation in overall response magnitude. A cross-validated model based on LPIPS and native displacement explained 90.22% of held-out variance, increasing to 91.72% with perturbation-family identity and interactions (Appendix A.4.6). Severity matching supported 29 of 36 pairs, but none survived Holm correction; this null result does not establish equivalent effects across features. Critically, perturbation-family identity re-

mained recoverable after response components associated with severity were removed in a leave-one-subject-out severity-controlled analysis: 28.89% for the full map and 41.11% over 0–400 ms, both above 11.11% chance (p=0.0001; Figure 7b,c). L1-normalized dose-response slope maps and L2-normalized severity-residual maps are complementary tests rather than a direct difference between before and after. Effective perturbation severity primarily determines how much the EEG response changes, whereas perturbation-family identity remains expressed in how that change is organized across sensors and time.

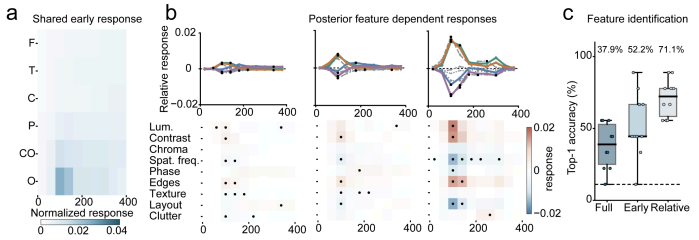

Figure 6: Shared and perturbation-family-specific spatiotemporal organization of generated EEG. (a) Shared L1-normalized dose-response slope map across the nine perturbation families over 0–400 ms and six sensor groups. (b) Posterior relative dose-response maps for the nine perturbation families in parietal, centro-occipital, and occipital regions from left to right. Temporal trajectories and corresponding feature-by-time maps show each family's normalized response relative to the mean of the other eight families. Dots mark global maxT $p < 0.05$. (c) Perturbation-family identification across subjects from full (0–1000 ms), early (0–400 ms), and full maps of relative feature responses: 37.78%, 52.22%, and 71.11%. The analysis of the full map is primary; the other two are secondary. Points show ten subjects; the dashed line marks 11.11% chance.

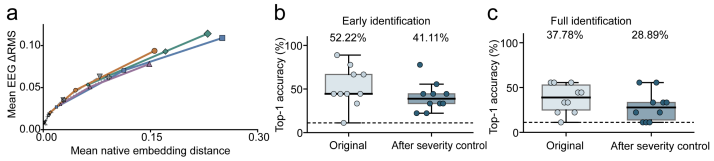

Figure 7: Severity predicts magnitude while perturbation-family identity remains recoverable. (a) Native displacement and mean EEG cell RMS difference; colors and symbols follow Figure 4a. (b,c) Top-1 accuracy for early and full maps for L1-normalized dose-response slope maps (Original) and L2-normalized severity-residual maps (After severity control). Points show ten subjects; labels give means; boxes show quartiles and medians, with whiskers using the 1.5 IQR rule. Dashed lines mark 11.11% chance.

5 Conclusion

BiGEM establishes a bidirectional framework for revealing how visual information is organized across visual representations and EEG. Frozen bidirectional compositions expose a clear separation between correspondence preserved in the shared latent space and information lost during generation in each modality. Controlled perturbations further show that perturbation severity primarily determines response magnitude, whereas perturbation-family identity is retained in the spatiotemporal organization of generated EEG. Together with complementary structure independently observed in measured EEG, these results reveal visual and neural representational geometry at multiple levels spanning shared latent representations, neural response organization, and outputs in each modality. BiGEM therefore extends bidirectional generation of EEG and images beyond reconstruction, providing a general framework for probing how visual information is preserved, transformed, and organized across brain and computational representations.

AI USE STATEMENT

During manuscript revision, generative AI tools assisted language condensation, feedback on scientific presentation and alignment between claims and evidence, allocation of figure panels, and Word and LaTeX typesetting. Reported experimental measurements were retained from the source manuscript and accompanying supplements.

ETHICS STATEMENT

Experiments use the publicly released THINGS-EEG2 dataset (Gifford et al., 2022). Controlled perturbations characterize the trained generative model and do not establish causal effects in human participants.

REPRODUCIBILITY STATEMENT

The main text defines the shared representation, generation pathways, evaluation readouts, and procedures for severity control. Appendix A gives the architecture, component losses, hyperparameters, evaluation manifests, and statistical specifications. Appendices B–E follow the main experiments: direct retrieval and generation, frozen cycles and joint refinement, visual representations and input dependence, and controlled perturbations. Figures, tables, and complete qualitative atlases are placed with their corresponding analyses.

REFERENCES

Y. Bai, X. Wang, Y.-P. Cao, Y. Ge, C. Yuan, and Y. Shan. DreamDiffusion: High-Quality EEG-to-Image Generation with Temporal Masked Signal Modeling and CLIP Alignment. In *Proceedings of the European Conference on Computer Vision*, pp. 472–488, 2024. doi: 10.1007/978-3-031-72751-1_27.

P. Bashivan, K. Kar, and J. J. DiCarlo. Neural Population Control via Deep Image Synthesis. *Science*, 364(6439):eaav9436, 2019. doi: 10.1126/science.aav9436.

H. Chen, L. He, Y. Liu, L. Yang, S. Shang, and M. Zhou. Visual Neural Decoding via Improved Visual-EEG Semantic Consistency. arXiv preprint arXiv:2408.06788v2, 2026. URL `https://arxiv.org/abs/2408.06788v2`.

M. Cherti, R. Beaumont, R. Wightman, M. Wortsman, G. Ilharco, C. Gordon, C. Schuhmann, L. Schmidt, and J. Jitsev. Reproducible Scaling Laws for Contrastive Language-Image Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2818–2829, 2023.

R. M. Cichy, D. Pantazis, and A. Oliva. Resolving Human Object Recognition in Space and Time. *Nature Neuroscience*, 17(3):455–462, 2014. doi: 10.1038/nn.3635.

M. X. Cohen. Where Does EEG Come From and What Does It Mean? *Trends in Neurosciences*, 40 (4):208–218, 2017. doi: 10.1016/j.tins.2017.02.004.

C. Du, K. Fu, J. Li, and H. He. Decoding Visual Neural Representations by Multimodal Learning of Brain-Visual-Linguistic Features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10760–10777, 2023. doi: 10.1109/TPAMI.2023.3263181.

M. Ferrante, T. Boccato, S. Bargione, and N. Toschi. Decoding Visual Brain Representations from Electroencephalography through Knowledge Distillation and Latent Diffusion Models. *Computers in Biology and Medicine*, 178:108701, 2024. doi: 10.1016/j.compbiomed.2024.108701.

A. T. Gifford, K. Dwivedi, G. Roig, and R. M. Cichy. A Large and Rich EEG Dataset for Modeling Human Visual Object Recognition. *NeuroImage*, 264:119754, 2022. doi: 10.1016/j.neuroimage.2022.119754.

U. Güçlü and M. A. J. van Gerven. Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. *Journal of Neuroscience*, 35(27):10005–10014, 2015. doi: 10.1523/JNEUROSCI.5023-14.2015.

K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://arxiv.org/abs/1706.08500.

N. Kriegeskorte and R. A. Kievit. Representational Geometry: Integrating Cognition, Computation, and the Brain. *Trends in Cognitive Sciences*, 17(8):401–412, 2013. doi: 10.1016/j.tics.2013.06.007.

N. Kriegeskorte and X. X. Wei. Neural Tuning and Representational Geometry. *Nature Reviews Neuroscience*, 22(11):703–718, 2021. doi: 10.1038/s41583-021-00502-3.

N. Kriegeskorte, M. Mur, and P. A. Bandettini. Representational Similarity Analysis: Connecting the Branches of Systems Neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008. doi: 10.3389/neuro.06.004.2008.

D. Li, C. Wei, S. Li, J. Zou, and Q. 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.

D. Li, K. Xie, M. Wu, Y. Kong, J. Tang, H. Qin, C. Wei, and Q. Liu. MindPilot: Closed-Loop Visual Stimulation Optimization for Brain Modulation with EEG-Guided Diffusion. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2602.10552.

X. Liu, C. Gong, and Q. Liu. Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2209.03003.

N. K. Logothetis. What We Can Do and What We Cannot Do with fMRI. *Nature*, 453(7197):869–878, 2008. doi: 10.1038/nature06976.

J. Loke, L. K. A. Sörensen, I. I. A. Groen, N. Cappaert, and H. S. Scholte. Shared Texture-Like Representations Underlie Deep Neural Network Alignment with Human Visual Processing. *Current Biology*, 36(18):4827–4834.e3, 2026. doi: 10.1016/j.cub.2026.08.008.

Z. Lu and J. D. Golomb. Human EEG and Artificial Neural Networks Reveal Disentangled Representations and Processing Timelines of Object Real-World Size and Depth in Natural Images. *eLife*, 13:RP98117, 2025. doi: 10.7554/eLife.98117.

Z. Lu, Y. Wang, and J. D. Golomb. Achieving More Human Brain-Like Vision via Human EEG Representational Alignment. *Communications Biology*, 9:463, 2026. doi: 10.1038/s42003-026-09685-w.

W. Mai, J. Wu, Y. Zhu, Z. Yao, D. Zhou, A. F. Luo, Q. Zheng, W. Ouyang, and C. Song. SynBrain: Enhancing Visual-to-fMRI Synthesis via Probabilistic Representation Learning. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://arxiv.org/abs/2508.10298.

W. Mai, M. Nan, Y. Zhu, J. Cao, R. Zhang, Y. Dai, C. Song, A. Luo, and J. Wu. NeuroFlow: Toward Unified Visual Encoding and Decoding from Neural Activity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12966–12976, 2026.

T. Naselaris, K. N. Kay, S. Nishimoto, and J. L. Gallant. Encoding and Decoding in fMRI. *NeuroImage*, 56(2):400–410, 2011. doi: 10.1016/j.neuroimage.2010.07.073.

D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=di52zR8xgf.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning Transferable Visual Models from Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763, 2021.

T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved Techniques for Training GANs. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://arxiv.org/abs/1606.03498.

A. Santos-Mayo, F. Gilbert, A. Mirifar, A. L. Tebbe, R. Fang, M. Ding, and A. Keil. Concept2Brain: An AI Model for Predicting Neurophysiological Responses to Text and Pictures. *Nature Communications*, 17:8961, 2026. doi: 10.1038/s41467-026-75653-x.

Y. Song, B. Liu, X. Li, N. Shi, Y. Wang, and X. Gao. Decoding Natural Images from EEG for Object Recognition. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=dhLIno8FmH.

Y. Song, Y. Wang, H. He, and X. 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.

Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861.

Y. Wei, L. Cao, H. Li, and Y. 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.

H. Wu, Q. Li, C. Zhang, Z. He, and X. 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.

G. Xu, Z.-R. Lai, Y. Tang, Y. Song, G. Zhou, B. Wang, J. Zhu, and J. Long. Image-to-Brain Signal Generation for Visual Prosthesis with CLIP Guided Multimodal Diffusion Models. arXiv preprint arXiv:2509.00787v5, 2025. URL https://arxiv.org/abs/2509.00787v5. revised 13 February 2026.

G. Xu, Z.-R. Lai, Y. Tang, Y. Song, S. Zhou, G. Zhou, B. Wang, J. Zhu, and J. Long. ViBE: Visual-to-M/EEG Brain Encoding via Spatio-Temporal VAE and Distribution-Aligned Projection. arXiv preprint arXiv:2604.26218, 2026. URL https://arxiv.org/abs/2604.26218.

D. L. K. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. J. DiCarlo. Performance-Optimized Hierarchical Models Predict Neural Responses in Higher Visual Cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014. doi: 10.1073/pnas.1403112111.

H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang. IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models. arXiv preprint arXiv:2308.06721, 2023. URL https://arxiv.org/abs/2308.06721.

Z. Ye, L. Yao, Y. Zhang, and S. Gustin. Self-Supervised Cross-Modal Visual Retrieval from Brain Activities. *Pattern Recognition*, 145:109915, 2024. doi: 10.1016/j.patcog.2023.109915.

C. Yu, S. Shen, X. Li, and Y. Li. D^2-FOSA: Dual-Diffusion Guided EEG-to-Image Reconstruction with Frequency-Oriented Semantic Alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26698–26710, 2026.

R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

J. Zhou, C. Xu, W. Wang, E. Yang, and C. 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 IMPLEMENTATION AND STATISTICAL METHODS

A.1 DATASET AND PREPROCESSING

Dataset partitions and preprocessing follow Section 4.1 (63×250 channel-by-time matrices, a resolution of 4 ms); training and validation partitions for each implementation are specified in Appendix A.2.4.

A.2 CORE IMPLEMENTATION AND EVALUATION

A.2.1 EEG ENCODER ARCHITECTURE

The EEG Encoder with 5,490,177 parameters combines posterior and complementary channel branches, followed by residual transformation and layer normalization. The posterior branch projects 17×250 samples; the complementary branch uses learnable 46×64 channel weights and a temporal basis of rank 64. EEG Encoder inputs average four trials per image. For flow and joint stages, 66,160 trials from 16,540 images are split into 59,160 training and 7,000 validation trials.

A.2.2 EEG ENCODER TARGETS AND MODULATION

The frozen OpenCLIP image encoder (ViT-H/14) provides normalized features with medium blur for image-to-EEG conditioning and the base retrieval gallery. EEG Encoder training selects cached low, medium or high blur features, initially medium. Let h_i be the raw EEG Encoder output and $\mathrm{z}_i^{I,li}$ the matching normalized cached feature. Frozen scalar multilayer perceptrons (MLPs) U and B use 1,024 → 128 → 1 layers with ReLU and sigmoid or Softplus outputs, respectively. Target modulation is:

$$m_i = \big(1 - U(h_i)\big) z_i^{I,\ell_i}$$
$$\widetilde{z}_i = \frac{m_i}{\|m_i\|_2 + 10^{-6}} \frac{1}{1 + B(h_i)}$$

Symmetric contrastive training uses fixed modulation networks with input gradients. Their outputs modulate feature amplitude after selection of the blur level. Logits multiply the raw EEG output, modulated visual target and trainable a=Softplus(ρ), initialized with ρ=log(1/0.07). Retrieval and generation use EEG Encoder outputs with L2 normalization.

For the matching pair's logit di, each sampled image's state updates as e_i←$0.3d_\mathrm{i}$+$0.7e_\mathrm{i}$. Updated minibatch states define mean μ_e and sample standard deviation s_e. Next labels are low if d_i>μ_e+$1.96s_\mathrm{e}$, high if d_i<μ_e−$1.96s_\mathrm{e}$, and medium otherwise. Training uses EEG means over four trials, 50 epochs, batch size 1,024 and the fixed final checkpoint at 850 updates.

The EEG Encoder uses AdamW with learning rate 10^{-4} and weight decay 10^{-4}.

A.2.3 VISUAL ADAPTER AND DISTILLATION

The visual adapter maps OpenCLIP ViT-H/14 EEG latents to four IP-Adapter tokens of 2,048 dimensions. Adapter training and reported inference use fixed empty positive and negative text prompts; no class labels or text specific to each image are supplied. A compact MLP and a residual branch of rank 128 initialized to zero contain 4,741,632 trainable parameters. The adapter conditions SDXL; SDXL, its variational autoencoder (VAE) and text encoders remain frozen.

Distillation matches native IP-Adapter ViT-H/14 conditioning from the shared visual representation; the adapter is then frozen.

$$\begin{aligned} L_{\mathrm{visual}} = {} & \lambda_\varepsilon L_\varepsilon + \lambda_{\mathrm{x}_0} L_{\mathrm{x}_0} \\ & + \lambda_{\mathrm{KV}} L_{\mathrm{KV}} + \lambda_{\mathrm{attn}} L_{\mathrm{attn}} \\ & + \lambda_{\mathrm{token}} L_{\mathrm{token}} + \lambda_{\mathrm{retr}} L_{\mathrm{retr}} \\ & + \lambda_{\mathrm{norm}} L_{\mathrm{norm}} + \lambda_{\mathrm{Gram}} L_{\mathrm{Gram}} \end{aligned}$$

A.2.4 EEG CODEC AND DIRECT FLOW OBJECTIVE

The fixed PCA encoder C_s and EEG Decoder D_s use statistics of training EEG. In column vector notation, P_s is 1,024×15,750; PCA mean b_s has 15,750 entries; coefficient means a_s and standard

deviations d_s each have 1,024 entries, with d_s floored at $10{-}6$. Both global EEG mean μ_s and input X are $63{\times}250$. The function flatten yields a column vector with 15,750 entries; $\oslash$ and $\odot$ denote elementwise division and multiplication. Encoding and reconstruction are:

$$C_s(X) = \left[P_s\big(\mathrm{flatten}(X - \mu_s) - b_s\big) - a_s\right] \oslash d_s$$

$$D_s(u) = \mathrm{reshape}\big(P_s^T(d_s \odot u + a_s) + b_s\big) + \mu_s$$

Reconstruction yields the signal with 15,750 dimensions within the retained subspace. The flow predicts velocity from Gaussian noise to coefficients u, with condition dropout probability 0.10 during training. Its predicted clean endpoint is $u_t{+}(1{-}t)v\theta_s(u_t,t,\delta z_i{}^I)$, whereas generation decodes the integrated endpoint $\widehat{u}_1$. The waveform regularizer is MSE between decoded predicted endpoints and decoded target coefficients, divided by the population variance of target waveforms over the minibatch and all signal entries.

For the first min(8,batch size) samples, the semantic term is one minus the mean cosine between normalized EEG Encoder encodings of predicted waveforms and original image conditions before dropout. EEG Encoder parameters remain frozen while input gradients reach the flow. The direct objective is:

$$\begin{aligned} L_{\mathrm{flow}} &= L_{\mathrm{RF}} + \lambda_{\mathrm{wave}} L_{\mathrm{wave}} + \lambda_{\mathrm{sem}} L_{\mathrm{sem}} \\ \lambda_{\mathrm{wave}} &= 0.16 \\ \lambda_{\mathrm{sem}} &= 0.18 \end{aligned}$$

Flow training uses batch size 64 and AdamW (learning rate 10^{-4}, weight decay 10^{-4}) for 24,000 updates. Validation uses 1,750 images from 175 concepts, disjoint from 14,790 training images and 1,479 concepts. Checkpoint selection uses this validation split. Official generation uses 20 Euler steps with guidance 1.0.

A.2.5 Joint losses and calibration

Joint refinement for each subject updates both translators with the direct EEG Encoder and flow losses. The flow semantic term freezes EEG Encoder parameters but propagates gradients through generated EEG to the flow. Each translator uses AdamW (learning rate 10^{-5}, weight decay 10^{-4}), with gradient clipping at 1.0. Direct batches contain 1,024 image-level EEG means for the EEG Encoder and 64 trials for the flow; cycle batches contain eight conditions with four trials each. Cycle samples use differentiable Heun integration with four steps, sharing initial noise between the two cycles within a minibatch. The adapter and SDXL remain outside this graph.

The image-start cycle averages four generated trials before EEG Encoder encoding. Its raw loss sums cosine distance, 0.25 times contrastive loss and 0.10 times the mean squared difference between cosine similarity matrices within each batch; positives share image identity (temperature 0.07). The EEG-start cycle encodes the mean of four recorded trials, generates four trials and compares generated and recorded means using standardized PCA space, flattened waveform correlation and waveform error normalized by variance (weights 0.50, 0.30 and 0.20). Residual channel covariance MSE and the squared deviation of the ratio of generated to real residual variance from one each receive weight 0.05. Covariance pools trials and time after removing each condition's trial mean; the ratio uses population variances and denominator floor 10^{-8}.

Before optimization, 500 evaluations on training batches fix loss scales. Let m_{D} and m_{F} be median logged direct EEG Encoder and flow losses, and m_{I} and m_{E} median raw image-start and EEG-start cycle losses. The logged EEG Encoder loss includes the detached triplet diagnostic. Define $c_{\mathrm{I}}{=}\sqrt{(m_{\mathrm{D}}m_{\mathrm{F}})}/\max(m_{\mathrm{I}},10^{-8})$ and $c_{\mathrm{E}}{=}\sqrt{(m_{\mathrm{D}}m_{\mathrm{F}})}/\max(m_{\mathrm{E}},10^{-8})$. The joint objective for gradient computation is:

$$\begin{aligned} L_{\mathrm{joint}} = L_{\mathrm{E2V}} + L_{\mathrm{flow}} \\ + 0.02 w(n)\big(c_I L_{\mathrm{I}} + c_E L_{\mathrm{E}}\big) \end{aligned}$$

For optimization update n, w(n) is zero through 1,000, rises linearly to one by 2,000 and remains one thereafter. Comparisons use independent, warm-up (1,000) and cycle-active (2,000) endpoints; calibration scales remain fixed. The detached EEG Encoder diagnostic may appear in logged totals but does not alter the objective for gradient computation.

A.2.6 Main System and Evaluation Manifests

The independently pretrained EEG Encoder and flow provide the principal direct results. Retrieval from measured EEG is 49.30%. Re-encoding generated EEG yields 97.20% Top-1 under the same scoring rule; this value quantifies preservation of stimulus identity within the BiGEM shared latent geometry and is not treated as an independent validation of EEG realism. Cached OpenCLIP ViT-H/14 features with medium blur form the base gallery. Candidate scalars come from the two fixed modulation networks applied to EEG Encoder features of real test EEG means over 80 repetitions or generated EEG means over K samples, respectively. Modulation is applied twice before multiplying normalized query features by the transposed gallery. The gallery is fixed per subject and evaluation, with the same scoring operation for every query and candidate. Retained amplitude modulation distinguishes this EEG-assisted gallery from plain cosine similarity to cached image features.

Frozen latent composition reuses the directly generated EEG tensor and EEG Encoder checkpoint, averaging the same first K samples. Its 97.15% cycle retrieval in the returned latent space uses cosine similarity to the original normalized gallery with medium blur. Direct and cycle-rendered images share the frozen OpenCLIP image encoder (ViT-H/14), original stimulus gallery and cosine ranking. Preprocessing resizes the shorter edge to 224, crops the center to 224×224, converts to RGB and applies OpenCLIP normalization without blur; pixel metrics resize to 512×512. Deterministic predictions, stochastic means and conditioning controls share targets from measured responses and waveform metric definitions.

Waveform Pearson and MSE are computed per subject and stimulus against the mean measured response. Normalized MSE divides stimulus MSE by the population variance of the real target (epsilon $10-12$) before averaging.

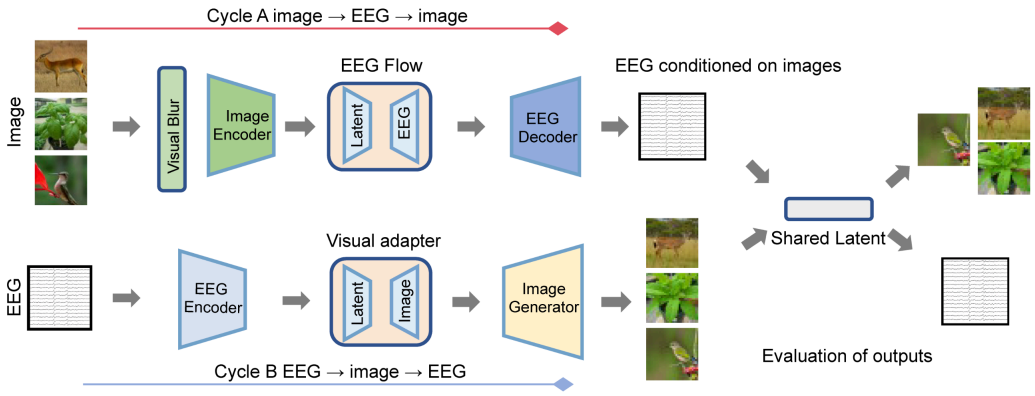

Figure A1: Frozen cycle evaluation: Cycle A composes image → EEG → image; Cycle B composes EEG → image → EEG. All modules are fixed. Joint refinement uses image latent → EEG → visual latent and EEG → visual latent → EEG losses, excluding the visual adapter and SDXL from the training graph.

A.3 Representation and Input Dependence

A.3.1 Visual Hierarchy and Measured Feature Associations

For each of the ten subjects and the independent, joint warm-up, and cycle-active endpoints, we encoded official EEG averaged over 80 repetitions into normalized latents with 1,024 dimensions and defined stimulus representational dissimilarity matrices (RDMs) as one minus cosine similarity. Frozen torchvision ResNet-50 features were globally averaged for the same 200 images. Low, mid, and high groups comprise conv1/layer1, layer2/layer3, and layer4/avgpool, respectively, corresponding to early, intermediate, and late groups in the main text; the third residual stage denotes layer3. Spearman RSA uses 19,900 upper triangle image pairs; grouped RSA averages correlations for individual layers. Partial RSA correlates EEG latent and ResNet layer RDM residuals after rank transformation and regression on the nine deterministic image feature RDMs below. Group infer-

ence uses exact two-sided sign flips at the subject level, subject bootstrap intervals, and BH false discovery rate (FDR) correction.

We extracted deterministic pixel descriptors from stimuli resized to 128×128 with Lanczos interpolation. One minus cosine RDMs represented nine families: luminance, contrast, chroma/opponent color, spatial frequency, spatial phase, edge orientation, texture, global layout, and clutter/complexity. For each subject, 100 fixed splits partitioned all 80 measured repetitions into nonoverlapping 40/40 halves. Symmetric cross half correlation distance EEG RDMs were averaged over splits for 25 nonoverlapping 40 ms windows, six fixed sensor groups, and all 150 region window cells. zero-order RSA is the Spearman correspondence between each feature RDM and the EEG RDM.

For conditional analysis, we rank transformed the upper triangle vectors of the measured EEG and nine visual predictor RDMs, then standardized them to zero mean and unit population standard deviation. A ridge model with fixed λ=1.0 and no intercept was fitted per subject and sensor time unit. Exact 2^{10} sign-flip tests and maximum cluster mass correction use subjects as inferential units. Coefficients quantify representational associations conditional on the other included families.

A.3.2 Measured EEG input occlusion

For each subject, the frozen independently trained EEG Encoder evaluated 200 official responses averaged over 80 repetitions using the direct evaluation's blur-aware gallery, normalization, and 200-way retrieval protocol. Occlusion at inference zeroed preprocessed whitened inputs without renormalization. Contribution was $\Delta\text{m} = \text{m(full)} - \text{m(occluded)}$, with MRR prespecified as primary. Temporal masks span 10 samples (40 ms); spectral masks zero Fourier coefficients in 5 Hz bands before inverse transformation; regional masks use six mutually exclusive sensor groups; sensor maps omit one of 63 channels at a time. Tables A7a–A7c report primary MRR results; Figure A8 visualizes Top-1 retrieval loss.

Inference uses subjects (n=10). Exact two-sided tests using sign flips enumerate all 2^{10} sign assignments at the subject level with a +1 correction. Temporal and spectral families use one dimensional maximum cluster mass correction; regional and sensor families use Benjamini–Hochberg FDR. Percentile confidence intervals use 10,000 subject bootstrap resamples. Effects quantify the frozen model's dependence on sensor inputs.

A.4 Controlled visual perturbations

A.4.1 Perturbation operators and image descriptors

The nine operators act on original RGB images scaled from zero to one. Weak, Medium, and Strong perturbation strengths were prespecified using 20 evenly spaced stimulus images before any generated EEG responses or downstream statistics were examined. Calibration used only perturbation behavior in images, after which all parameters were fixed across images and subjects (Table A1). Stochastic operators shared a fixed image seed across doses. Resulting RGB values were clipped to the valid range, rounded to integers, and saved as images.

The frequency operator normalized radial frequency by its maximum on the two dimensional sampling grid. The phase operator used real noise Fourier phase, preserving conjugate symmetry and original amplitude before clipping. The texture base used a bilateral filter with color scale 0.12 and spatial scale three times the shorter image side divided by 128. Global displacement used bilinear interpolation. Nominal dose denotes an operator setting; native displacement and LPIPS quantify effective severity separately.

Descriptor quality control used the measured EEG analysis's feature families and saved projection parameters. Images were resized to 128×128 with Lanczos interpolation and converted to RGB floating point values. Descriptors summarized luminance, contrast, chroma, Fourier amplitude and phase, gradient orientation, texture, spatial layout and clutter. Coordinates were standardized with discovery set means and population standard deviations, centered, and projected with saved family PCA loadings.

Table A1: Visual perturbation operators and fixed dose parameters. Parameters were calibrated before EEG responses were examined.

Family	Operation	Weak, Medium, Strong parameters
Luminance	Multiply RGB intensities by $1 - p$	$p = 0.15, 0.35, 0.60$
Contrast	Scale each RGB channel around its spatial mean by factor p	$p = 0.80, 0.50, 0.20$
Chroma	Interpolate between the RGB image and its grayscale image, retaining chroma fraction p	$p = 0.75, 0.45, 0.10$
Spatial frequency	Attenuate Fourier amplitudes above normalized radial frequency 0.25 with weight $\exp\left[-p\max(r - 0.25, 0)^2\right]$	$p = 2, 8, 24$
Spatial phase	Add p times the phase of a fixed real Gaussian noise field, with zero phase displacement at DC	$p = 0.10, 0.30, 0.65$
Edge orientation	Apply Gaussian smoothing along the horizontal image axis	$p = 1, 3, 7$, scaled by the shorter image side divided by 128
Texture	Interpolate towards a bilateral filtered base, attenuating the fine residual	$p = 0.25, 0.55, 0.85$
Global layout	Apply smooth sinusoidal displacements along both image axes, with reflected boundaries	$p = 0.025, 0.06, 0.12$ of the corresponding image dimension
Clutter	Add a fixed Gaussian spatial noise field shared by the RGB channels	SD $p = 0.015, 0.04, 0.09$

A.4.2 Native conditioning and image similarity

Native conditions used the EEG generator's OpenCLIP image encoder (ViT-H/14). Images were resized to 224 × 224, reproducing historical encoding/decoding in the original stimulus format. Exponential foveated blur ($g = 3$) preceded tensor conversion and channel normalization. The resulting 1024 dimensional vector retained the original normalization used to condition the generator.

Let z_i^0 and $z_i^{f,d}$ denote the original and perturbed conditions for image i, feature family f and dose d. Native displacement was

$$D_{i,f,d} = 1 - \frac{\langle z_i^0, z_i^{f,d} \rangle}{\|z_i^0\|_2 \|z_i^{f,d}\|_2}.$$

Stored generator conditions were used directly. Cosine similarity ranked each perturbed condition against a fixed gallery of 200 original conditions, breaking ties by gallery order. Identity retention was the proportion whose original image ranked first, assessing identity in the native conditioning representation.

AlexNet LPIPS used RGB images resized to 256 × 256 with Lanczos interpolation and scaled from −1 to 1. SSIM and a separate OpenCLIP ViT-H/14 image similarity measure provided image quality controls; the latter remained distinct from displacement in the native conditioning space.

A.4.3 Generated EEG and response construction

The trained flow and EEG codec for each subject generated 80 samples per condition with 20 Euler steps and guidance scale one. Subject, image and sample indices fixed noise across families and doses. Decoded waveforms were averaged into 63 channels, 250 sample responses; original responses used the corresponding averages of 80 previously generated samples. Signals use preprocessed whitened EEG units at 250 Hz.

For subject s, image i, feature f, dose d, sensor group r and time window t, the cell response was

$$R_{sifd}(r,t) = \sqrt{\frac{1}{|C_r||T_t|} \sum_{c \in C_r} \sum_{\tau \in T_t} \left[\overline{X}_{sifd}(c,\tau) - \overline{X}_{si0}(c,\tau)\right]^2}.$$

Here $\overline{X}$ is the average generated waveform; C_r contains sensors in group r; T_t contains ten samples in a 40 ms window. The 25 windows spanning 0–1000 ms and six fixed sensor groups yield 150

cells. Groups were frontal, temporal, central, parietal, centro-occipital and occipital; CO contained PO7, PO3, POz, PO4 and PO8. Memberships matched the measured EEG analysis.

Overall magnitude was the arithmetic mean of 150 RMS values, equally weighting sensor groups and time windows. Image responses were averaged within subjects before group summaries. Representative waveforms retained signed differences for the channel with the largest Strong versus Original RMS difference over 0–400 ms. This deterministic selection affected illustrations only; group analyses used all 150 cells.

A.4.4 DOSE-RESPONSE SLOPES NORMALIZED PATTERNS AND LOCAL INFERENCE

Let M_{sfdc} be the mean of $R_{sifd}(c)$ over the 200 images. The dose-response slope and normalized pattern were

$$b_{sfc} = \sum_{d=0}^{3} \frac{d - 1.5}{5} M_{sfdc}$$

$$P_{sfc} = \frac{b_{sfc}}{\sum_{c'=1}^{150} |b_{sfc'}| + 10^{-12}}.$$

The relative feature response was

$$Spec_{sfc} = P_{sfc} - \frac{1}{8} \sum_{j \neq f} P_{sjc}.$$

Shared maps averaged P over nine features within subjects, then over subjects. Pattern similarity used Spearman correlations over all 150 cells. The 0–400 ms displays retained ten early windows per group after normalization over all 150 cells. Relative values above/below zero indicate responses above/below the other eight families' mean in this representation.

For each of 1,350 combinations of features and cells, the statistic was mean relative response divided by (standard error + 10^{-12}). All $2^{10} = 1{,}024$ subject sign configurations applied the same subject sign across features and cells, recomputing statistics and standard errors. Adjusted two-sided p values counted configurations whose maximum absolute statistic over 1,350 cells equaled or exceeded the observed absolute statistic, divided by 1,024. Correction over 150 cells within each feature and directional results were secondary. Confidence intervals used 10,000 resamples of ten subjects.

A.4.5 PERTURBATION-FAMILY IDENTIFICATION AND UNCERTAINTY

For each query subject, nine feature templates averaged the other nine subjects' maps. Spearman correlation compared queries with every template, breaking ties by fixed feature order. Full L1-normalized dose-response slope maps were primary; early normalized, full relative and normalized Strong maps were secondary. Early maps retained normalization over the full map.

Each representation supplied 90 queries (nine features, ten subjects). Accuracy averaged subjects' correct fractions; chance was $1/9$. Matching margin was similarity to the correct template minus mean similarity to the other eight templates.

The null used 10,000 independent permutations of feature labels within subjects, rebuilding all templates and query similarities with shared assignments across representations. Matching margin separately used an exact test of subject means over 1,024 sign configurations. Subject bootstrap intervals used 10,000 resamples with predictions fixed, quantifying uncertainty conditional on fitted templates.

A.4.6 EMBEDDING ASSOCIATIONS AND MAGNITUDE PREDICTION

For each subject, feature and dose, Spearman correlations related native displacement to overall EEG response across 200 images. At Strong dose, ordinary least squares related standardized response to standardized native displacement and LPIPS, using means and population standard deviations for each subject and feature. Embedding coefficients and fitted R^2 increments over LPIPS were summarized across subjects; correlation/coefficient intervals used 10,000 subject bootstrap resamples. Scatter plots instead correlated 200 image summaries after averaging EEG responses across subjects within images.

Magnitude analysis used 5,400 nonzero image, feature and dose conditions and 54,000 subject responses. The initial model included an intercept, standardized LPIPS and native displacement, and their separate squares. The second added eight feature indicators; the third added their interactions with standardized native displacement. This sequence excluded feature by LPIPS and quadratic cross terms.

Fivefold image-grouped cross-validation split 200 image identities, keeping each image's 27 conditions together. Training images supplied standardization parameters; ordinary least squares fitted mean responses across ten subjects for each training condition, then predicted each subject's held-out responses. Pooled predictive R^2 was one minus total squared prediction error divided by total squared deviation of these observations from their grand mean. Displayed increments were differences between these pooled values. Supporting mixed models included subject and stimulus intercepts; their marginal and conditional R^2 are distinct summaries.

A.4.7 Severity overlap and matching support

Embedding overlap pooled 600 image and dose observations per feature, using 100 equal width bins spanning the minimum and maximum of all 5,400 embedding distances. Pairwise coefficients summed the smaller relative histogram frequency per bin; min/max intersection separately assessed common range. Joint support required points inside every feature's convex hull in standardized LPIPS and embedding coordinates, with a nearest observation within 0.25 SD on both axes. A common standardized effect required at least 30 distinct supported images.

Within each stimulus, matching selected the closest of nine nonzero dose combinations by Euclidean distance in standardized LPIPS and embedding coordinates, retaining pairs with both absolute coordinate differences $\leq$0.25 SD. Standardization used pooled population standard deviations of all 5,400 image conditions. Testing required at least 30 matched images per feature pair; a 0.50 SD caliper provided sensitivity analysis.

For each eligible pair, EEG differences between selected conditions were averaged over matched images per subject. Exact two-sided sign permutation tested the mean of ten subject differences, with Holm correction across eligible feature pairs at each caliper. Figure A17 (part 1) reports matching support and inferential resolution.

A.4.8 Severity-controlled perturbation-family identification

For each outer query subject and response cell, the nuisance design without feature labels included an intercept, standardized LPIPS and native displacement, and their separate squares. Ordinary least squares fitted responses averaged over nine training subjects at every image, feature and nonzero dose condition. The same fitted cell model was applied to the query and all nine template subjects in that outer fold.

Residuals were averaged over 200 images and three nonzero doses into maps for each subject and feature per outer fold, then divided by their L2 norm over 150 cells. Full maps were used directly; early maps retained 60 cells within 0–400 ms. Templates averaged nine training subject maps per feature. Assignment used the original analysis's Spearman similarity and fixed feature order. Feature label permutations rebuilt templates under the fixed severity fit without feature labels.

Original identification uses dose-response slopes with L1 normalization; strict identification uses mean residual responses across nonzero doses, with L2 normalization. Results describe these representations separately. Confusion matrices report the count and fraction of ten subject queries assigned from each true feature to each predicted feature.

A.4.9 Correspondence with measured EEG

The restricted analysis used exact memberships of corrected measured EEG clusters (Figure A19). Primary spatial phase, edge orientation and texture masks comprised 11, 27 and 14 cells. Masks were fixed before new perturbation statistics, retaining their full spatial and temporal extent. Restricted preference for each subject averaged relative dose-response slope responses over corresponding cluster cells. A secondary score weighted these responses by absolute measured ridge coefficients normalized within the mask.

Exact one-sided sign permutation tested positive preference across ten subjects for each score. Holm correction covered three primary features separately per score and map version. Two sided tests and the Strong representation were additional summaries; contrast (two cluster cells) was reported descriptively. Full map correspondence retained all nine features and original intervals; restricted scores quantified responses within particular measured feature masks.

B DIRECT RETRIEVAL AND GENERATION

B.1 DIRECT EEG GENERATION AND MATCHED CONTROLS

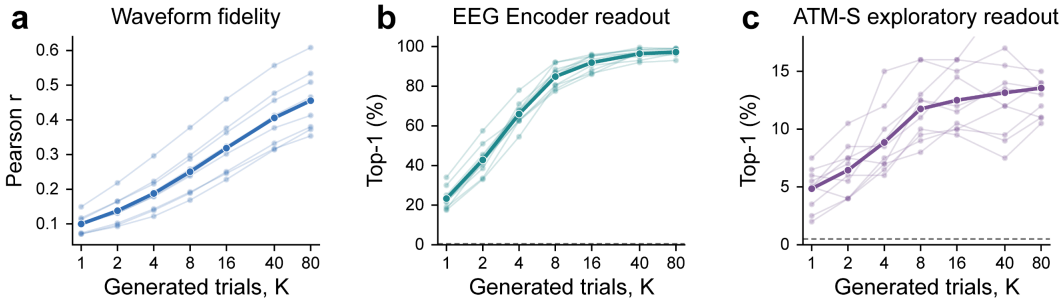

Figure A2: Direct EEG generation across subjects and K. Thin and thick trajectories show ten subjects and their means. (a) Waveform Pearson correlation. (b) Shared-latent identity Top-1. (c) ATM-S Top-1. Dashed lines mark 0.50% chance. At K=80, EEG Encoder Top-1 exceeds ATM-S by 83.65 percentage points (exact paired sign-flip p=0.002). The EEG Encoder shares the representation used by the flow semantic objective; ATM-S provides an exploratory alternative (Table A2).

Table A2: Matched controls for EEG generation (mean ± standard deviation (SD), n=10). Trial dispersion ratio is the ratio of generated to measured dispersion; ATM-S chance is 0.50%. Deterministic duplicate rate is one. Uncorrected paired comparisons: flow versus deterministic ATM-S, p=0.218750; correct versus fixed mismatched conditioning, ΔPearson=0.0322, 95% CI [0.0279, 0.0366], exact sign-flip p=0.001953, positive in all ten subjects.

Generator / conditioning	**Pearson r ↑**	**ATM-S Top-1 %**	**Trial dispersion ratio**
Deterministic mean	.5967 ± .0657	12.60 ± 3.77	.0000 ± .0000
Conditional EEG flow	.4555 ± .0797	13.55 ± 3.17	1.0101 ± .0522
Fixed mismatched pairing	.4233 ± .0765	.25 ± .35	1.0101 ± .0522

B.2 QUALITATIVE EXAMPLES AND SELECTION PROTOCOL

Figure 2a shows three qualitative examples from the same EEG Encoder architecture and pipeline for generating images from EEG used in the main quantitative evaluation, comparing endpoints from independent training and joint refinement. Both columns are forward image outputs under the same inference procedure. Figure 2b follows the separate frozen route for regenerating EEG and uses its own intermediate rendered image.

Figure A3a contains the qualitative atlas of EEG-to-image outputs for the same BiGEM EEG Encoder system used in the main evaluation, with groups of high, intermediate and low quality assigned separately within each subject according to the fixed ranking protocol. The Direct and C (joint-refined) rows are forward outputs from independent training and joint refinement, respectively; Target is the original stimulus. Figure A3b contains the corresponding quality strata for image-to-EEG outputs across all ten subjects.

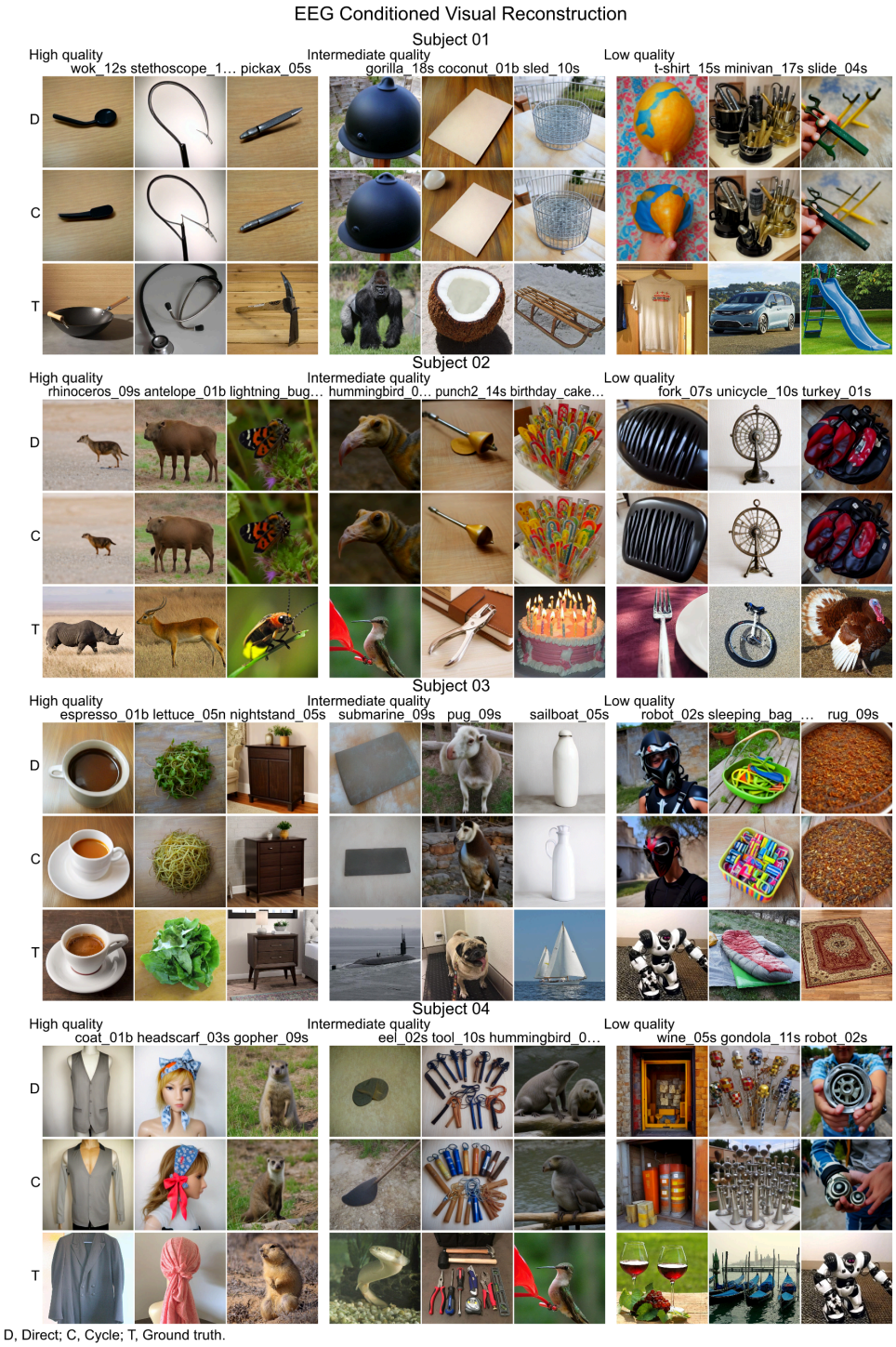

Figure A3a: (part 1 of 3) Examples of EEG-to-image outputs for subjects 01–04 from the same BiGEM EEG Encoder architecture used in the main quantitative evaluation. D, direct forward output from the independently trained endpoint; C, joint-refined forward output; T, target stimulus. Groups of high, intermediate and low quality follow the fixed ranking protocol within each subject.

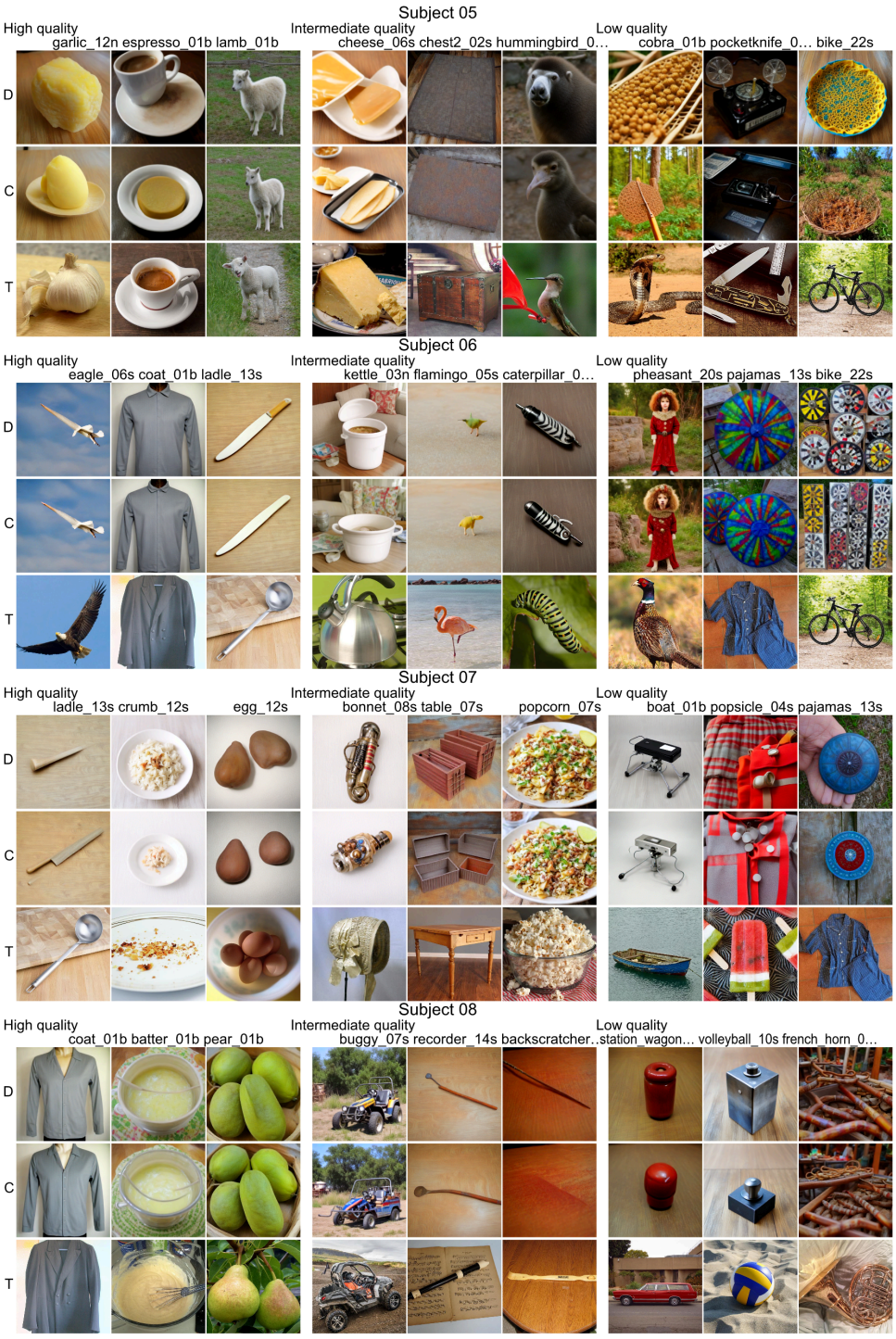

Figure A3a: (part 2 of 3) Examples of EEG-to-image outputs for subjects 05–08 from the same BiGEM EEG Encoder architecture used in the main quantitative evaluation. D, direct forward output from the independently trained endpoint; C, joint-refined forward output; T, target stimulus. Groups of high, intermediate and low quality follow the fixed ranking protocol within each subject.

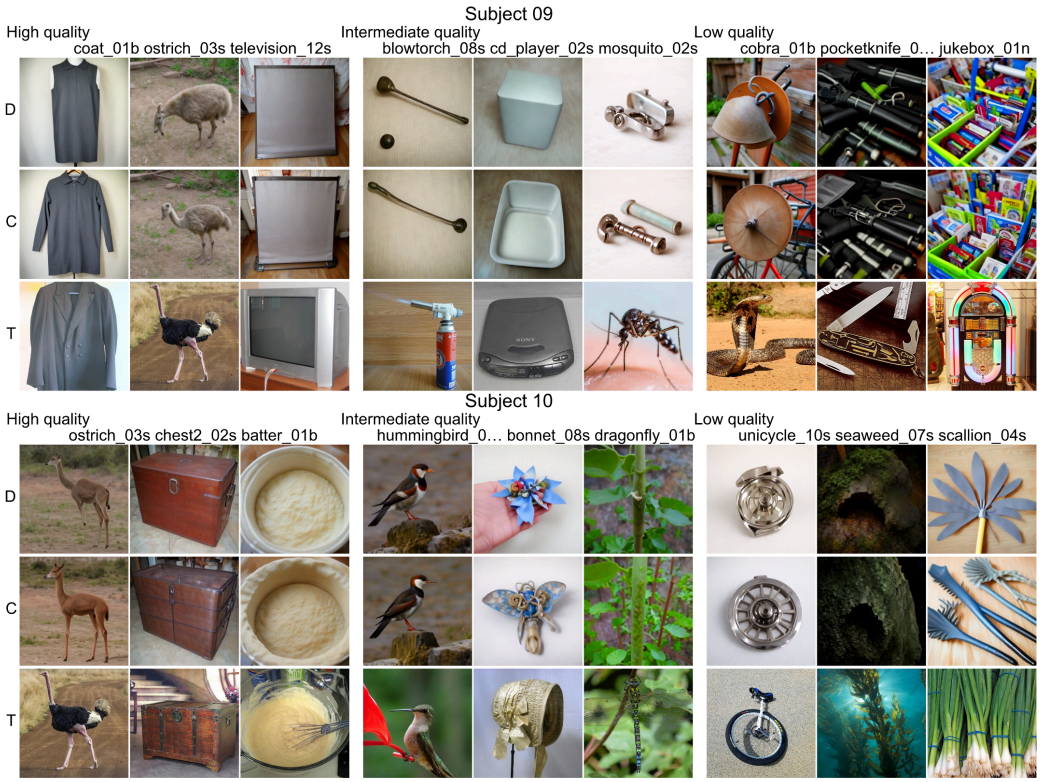

Figure A3a: (part 3 of 3) Examples of EEG-to-image outputs for subjects 09–10 from the same BiGEM EEG Encoder architecture used in the main quantitative evaluation. D, direct forward output from the independently trained endpoint; C, joint-refined forward output; T, target stimulus. Groups of high, intermediate and low quality follow the fixed ranking protocol within each subject.

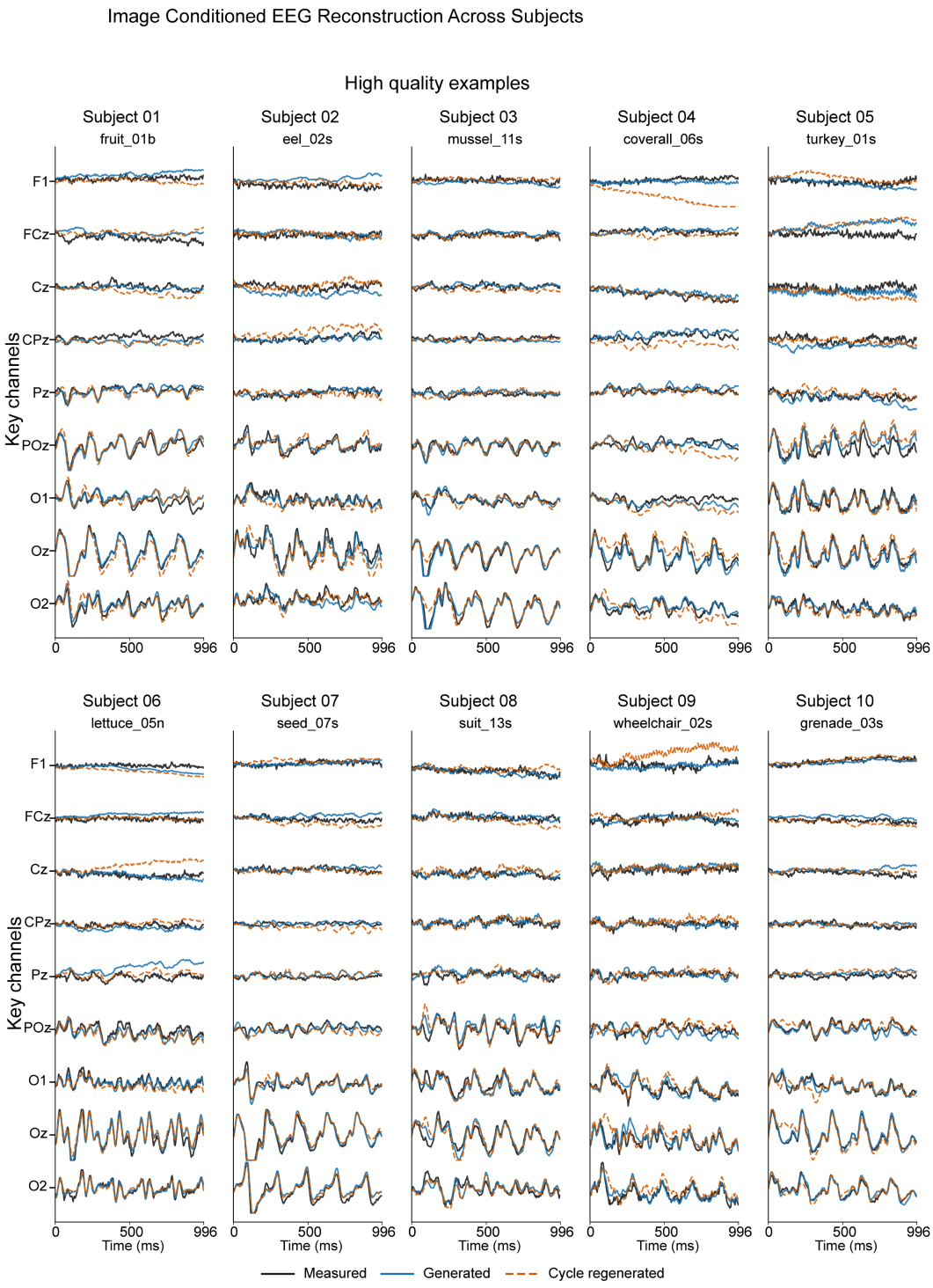

Figure A3b: (part 1 of 3) Examples of image-to-EEG outputs across ten subjects. High quality examples follow the fixed ranking within each subject described in the protocol for selecting qualitative examples in this appendix. Measured curves average 80 recorded repetitions; generated curves and curves regenerated through the cycle average 80 samples. Cycle regeneration first renders an image from EEG, then regenerates EEG from that image. Normalized curves use vertical offsets for display; amplitudes are not raw microvolts.

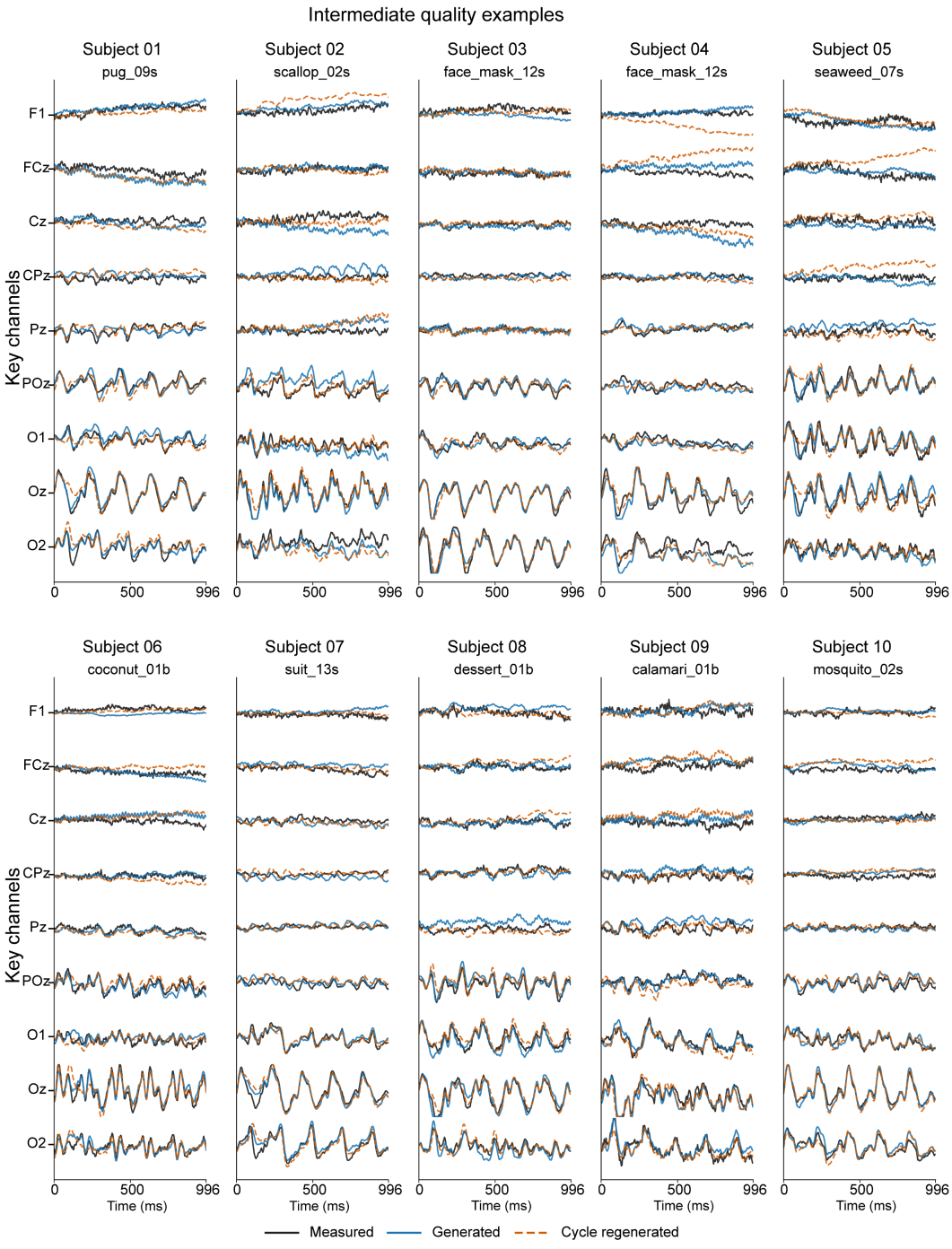

Figure A3b: (part 2 of 3) Examples of image-to-EEG outputs across ten subjects. Intermediate quality examples follow the fixed ranking within each subject described in the protocol for selecting qualitative examples in this appendix. Measured curves average 80 recorded repetitions; generated curves and curves regenerated through the cycle average 80 samples. Cycle regeneration first renders an image from EEG, then regenerates EEG from that image. Normalized curves use vertical offsets for display; amplitudes are not raw microvolts.

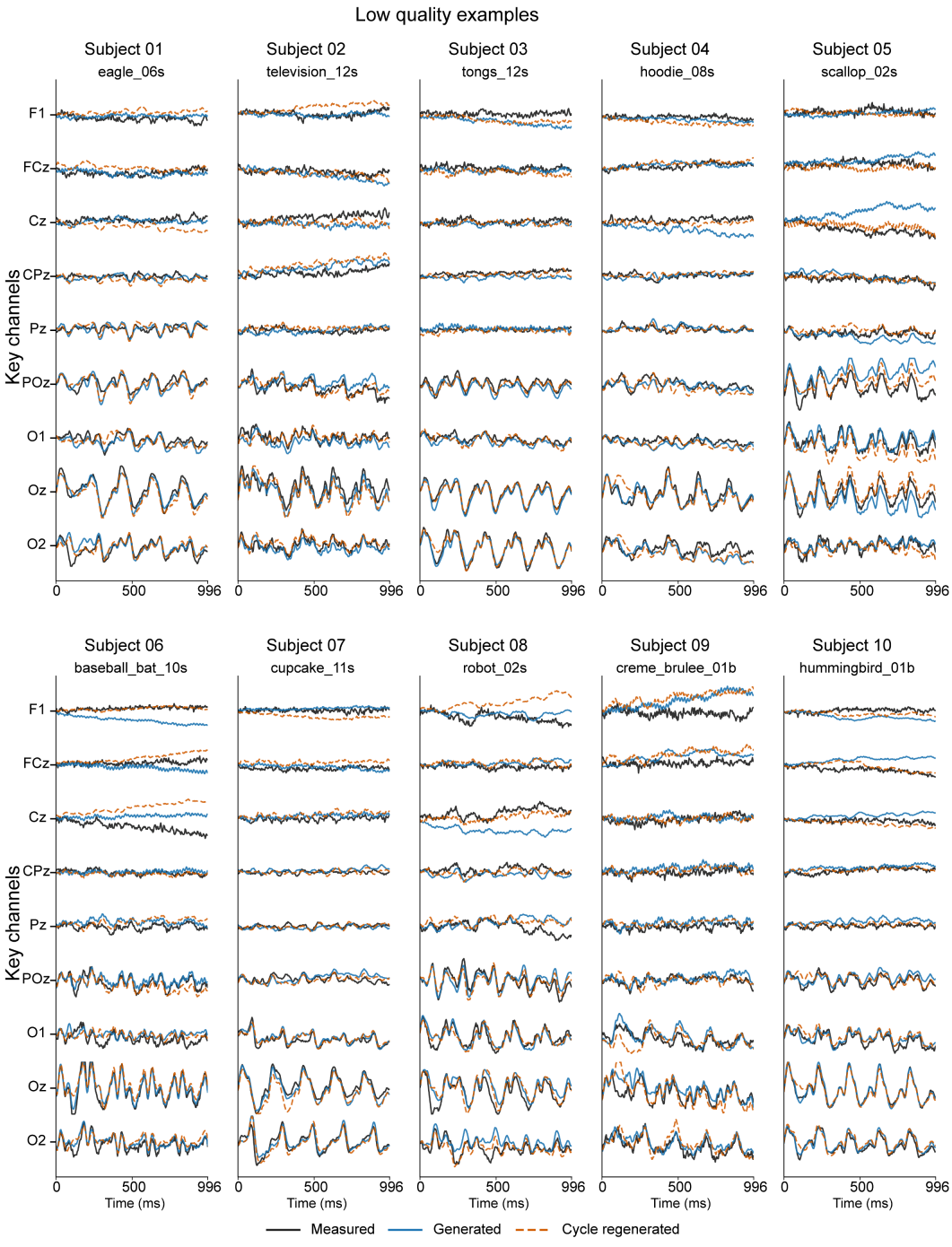

Figure A3b: (part 3 of 3) Examples of image-to-EEG outputs across ten subjects. Low quality examples follow the fixed ranking within each subject described in the protocol for selecting qualitative examples in this appendix. Measured curves average 80 recorded repetitions; generated curves and curves regenerated through the cycle average 80 samples. Cycle regeneration first renders an image from EEG, then regenerates EEG from that image. Normalized curves use vertical offsets for display; amplitudes are not raw microvolts.

B.3 CHANNEL AND VISUAL TARGET ABLATIONS

Table A3a: Channel and target ablations (mean, ten subjects). Channel and target ablations show that retrieval depends primarily on posterior EEG signals and blur-aware visual targets (Table A3a). Using all 63 channels with blur-aware targets achieved 49.30% Top-1 accuracy. Restricting the input to the posterior 17 channels retained most of this performance (47.93%). Replacing the blur-aware target with a nonblurred representation decreased Top-1 from 49.30% to 37.95% and Top-5 from 78.55% to 69.10%; the Top-1 reduction was observed in all ten subjects. These results indicate that the trained EEG Encoder relies predominantly on posterior signals and that suppressing weakly recoverable visual detail improves EEG–visual alignment.

Variant	**Top-1 (%)**	**Top-5 (%)**	**MRR**
All channels, blur-aware target	49.30	78.55	0.6250
Posterior 17 channels only	47.93	76.25	0.6221
All channels, nonblurred target	37.95	69.10	0.5223

Table A3b: Extended ablation of core BiGEM components. Extended ablations reveal complementary roles of the main BiGEM components (Table A3b). Adaptive blur-aware alignment improved retrieval over both nonblurred and fixed-medium targets, increasing Top-1 from 37.95% and 42.00% to 49.30%, respectively. For image generation, the full decoder-aware adapter achieved the highest SSIM (0.382), lowest LPIPS (0.7402), and highest CLIP similarity (0.7030), although removing decoder-aware supervision produced higher exact Top-1 accuracy. For EEG generation, velocity-only training slightly increased Pearson correlation but substantially reduced semantic readout accuracy, while removing the semantic objective similarly impaired stimulus identity. The full EEG Flow achieved MSE (0.04375), the highest Shared-latent Top-1 (97.20%), and a dispersion ratio closest to the measured-data target of one (1.0101), indicating the best overall balance between waveform fidelity, stimulus identity, and realistic trial variability.

(a) Blur-aware alignment

Variant	**Top-1 (%) ↑**	**Top-5 (%) ↑**	**MRR ↑**
Nonblurred	37.95	69.10	0.5223
Fixed medium	42.00	72.00	0.5944
Adaptive blur-aware	49.30	78.55	0.6250

(b) Decoder-aware visual adapter

Variant	**Top-1 (%) ↑**	**SSIM ↑**	**LPIPS ↓**	**CLIP ↑**
Base MLP	4.35	0.2773	0.7663	0.3333
w/o L_{DA}	9.50	0.2979	0.7677	0.3383
Full	7.10	0.382	0.7402	0.7030

(c) EEG flow objectives

Variant	**Pearson ↑**	**MSE ↓**	**Shared-latent Top-1 (%) ↑**	**ATM-S Top-1 (%)**	**Dispersion**
Velocity only	0.4658	0.0403	28.0	5.5	1.0610
w/o semantic	0.4243	0.0593	26.0	5.5	1.0547
w/o waveform	0.4517	0.0570	93.0	14.0	1.0671
Full	0.4555	0.0437	97.20	13.5	1.0101

C FROZEN CYCLES AND JOINT REFINEMENT

C.1 OUTPUT CYCLES ACROSS SAMPLING COUNTS

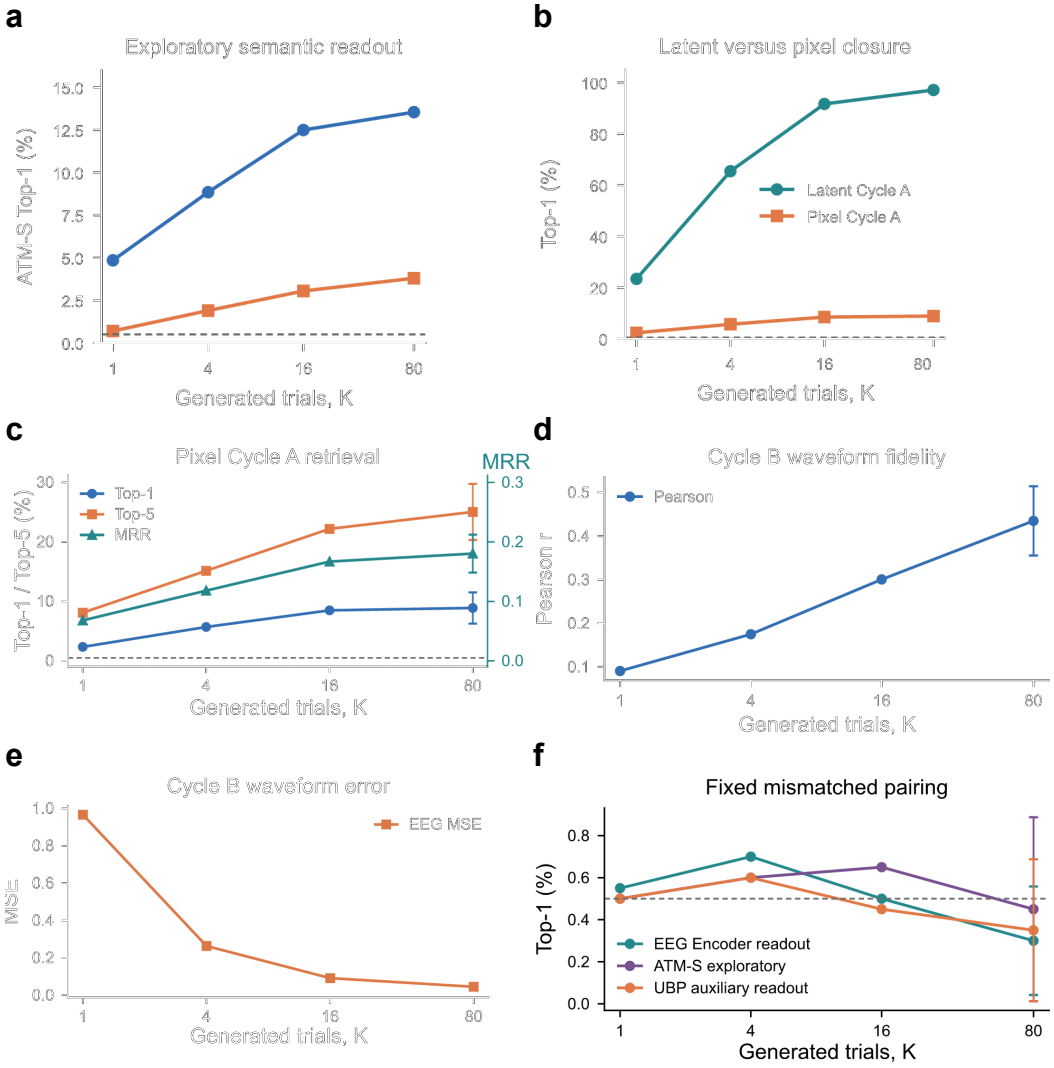

Figure A4: Frozen cycles across generated trial counts K=1, 4, 16 and 80 (means, ten subjects). (a) ATM-S Top-1 for direct image-to-EEG generation (blue) and Cycle B (orange). (b) Cycle A Top-1 for returned latents and rendered images. (c) Rendered Cycle A Top-1/Top-5 (left, %) and MRR (right, proportion). (d,e) Cycle B waveform Pearson r and MSE. (f) EEG Encoder, ATM-S and auxiliary UBP Top-1 under a fixed derangement of conditioning images. Dashed retrieval lines mark 0.50% chance; c–f show ± 1 SD only at K=80. Rendered Cycle A Top-1 rises from 2.10% to 8.90%; Cycle B Pearson rises from approximately 0.09 to 0.4348 as MSE decreases. Mismatched semantic readouts remain near chance.

C.2 TRAJECTORIES DURING JOINT REFINEMENT AND PAIRED INFERENCE

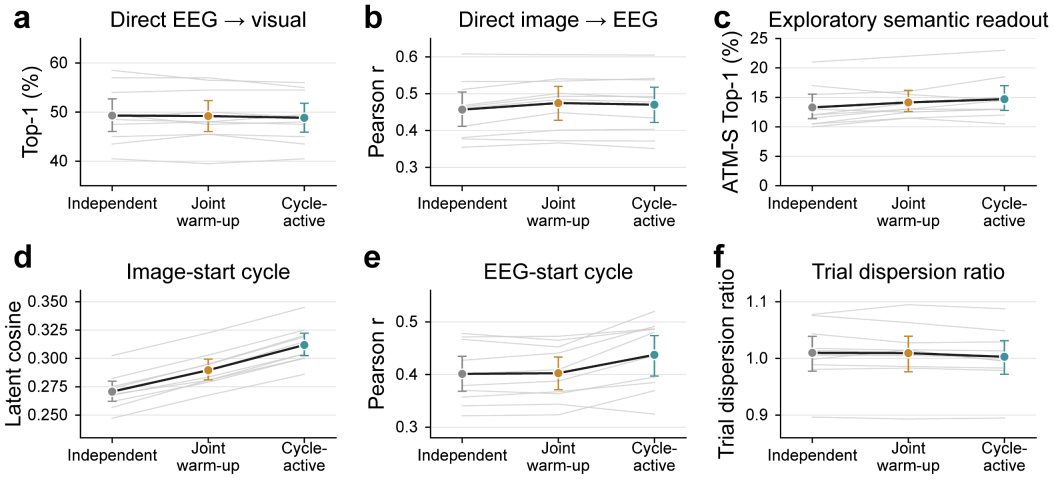

Figure A5: Joint refinement (n=10 subjects). Thin gray lines show paired trajectories; colored points show means and 95% intervals from 10,000 subject bootstrap resamples. (a) Top-1 for direct retrieval of visual stimuli from EEG. (b) Pearson r for direct image-to-EEG generation. (c) Exploratory ATM-S Top-1. (d) Latent cosine for cycles starting from images. (e) Pearson r for cycles starting from EEG. (f) Trial dispersion ratio of generated EEG at K=80. Cycle-active versus independent cosine improves in all ten subjects; Pearson r for cycles starting from EEG improves in nine.

Table A4: Readouts during joint training at K=80 (mean ± SD across ten subjects).

Readouts during joint training	**Independent**	**Joint warm-up**	**Cycle-active**
Image cycle cosine ↑	0.2707 ± 0.0150	0.2896 ± 0.0153	0.3118 ± 0.0166
EEG cycle Pearson r ↑	0.4014 ± 0.0570	0.4027 ± 0.0533	0.4376 ± 0.0652
Direct EEG → visual Top-1 (%) ↑	49.30 ± 5.77	49.20 ± 5.48	48.85 ± 5.13
Direct image → EEG Pearson r ↑	0.4555 ± 0.0792	0.4747 ± 0.0777	0.4702 ± 0.0805
Direct image → EEG MSE ↓	0.04364 ± 0.01049	0.03927 ± 0.00834	0.03998 ± 0.00885
Direct image → EEG ATM-S Top-1 (%)	13.30 ± 3.51	14.15 ± 3.15	14.70 ± 3.61
Trial dispersion ratio	1.0098 ± 0.0521	1.0093 ± 0.0531	1.0029 ± 0.0504

Joint stages share a sampling manifest; frozen endpoints appear in Table 5. Trial dispersion ratio is the ratio of mean pairwise distances among generated versus measured trials in standardized PCA space. Paired changes from the independent endpoint to the endpoint with active cycles: cosine for the image cycle, +0.0412, 95% CI [0.0390, 0.0433], exact p=0.001953, BH q=0.008057; Pearson r for the EEG cycle, +0.0362, 95% CI [0.0185, 0.0529], exact p=0.007812, BH q=0.021484; direct Top-1, −0.45 pp, 95% CI [−1.25, 0.30] pp, BH q=0.386719. Intervals use 10,000 subject bootstraps; exact paired tests enumerate all 2^{10} sign assignments; Benjamini–Hochberg correction covers 33 contrasts. Changes use original paired estimates, not rounded endpoint differences. Training duration and cycle activation covary; pp denotes percentage points.

D VISUAL REPRESENTATIONS AND EEG INPUT DEPENDENCE

D.1 LEARNED VISUAL HIERARCHY AND MEASURED EEG

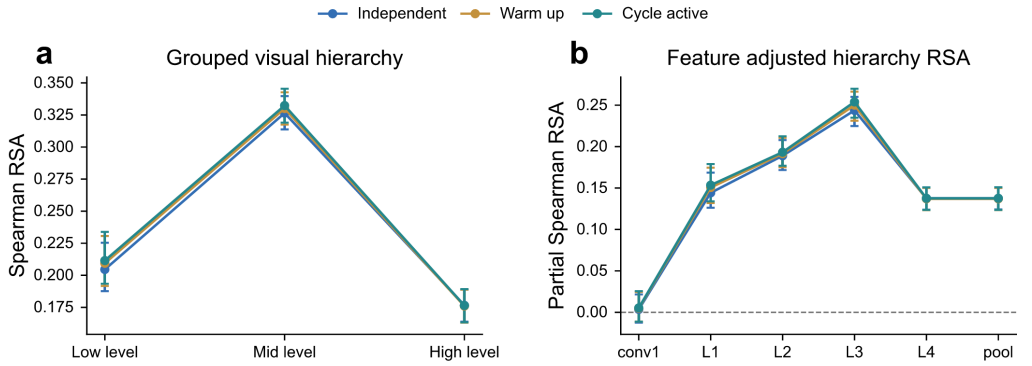

Figure A6: Visual hierarchy across training endpoints (n=10). (a) Spearman RSA grouped at low, mid and high levels. (b) Layerwise partial Spearman RSA controlling the nine RDMs of image features. Points and error bars show subject means and 95% subject bootstrap intervals. Mean correspondence ordering persists through joint warm-up and refinement with active cycles.

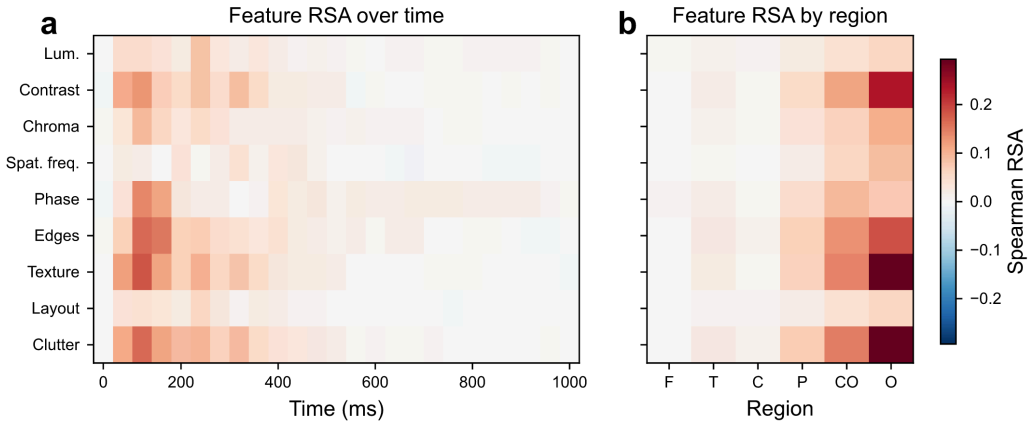

Figure A7: Zero-order Spearman RSA between RDMs of image features and RDMs of measured EEG. (a) Twenty-five nonoverlapping 40 ms windows. (b) Six sensor groups. Cells show mean correlations across ten subjects over the same 200 stimuli; both panels share a scale centered at zero. F, frontal; T, temporal; C, central; P, parietal; CO, centro-occipital; O, occipital. Texture, edge, contrast and clutter associations peak in early posterior windows. Figure 3 gives conditional associations from the joint feature model.

Table A5: Visual hierarchy RSA. Means and 95% bootstrap intervals across ten subjects.

Endpoint	Level	Mean	95% CI
Independent	Low	0.2045	[0.1876, 0.2254]
Independent	Mid	0.3265	[0.3137, 0.3397]
Independent	High	0.1766	[0.1639, 0.1889]
Warm-up	Low	0.2094	[0.1916, 0.2306]
Warm-up	Mid	0.3300	[0.3175, 0.3427]
Warm-up	High	0.1760	[0.1631, 0.1887]
Cycle-active	Low	0.2115	[0.1934, 0.2338]
Cycle-active	Mid	0.3323	[0.3190, 0.3454]
Cycle-active	High	0.1764	[0.1633, 0.1892]

Table A6: Variance inflation factors (VIFs) in the joint feature RDM model.

Feature	VIF
Luminance	2.920
Contrast	2.033
Chroma/opponent color	1.119
Spatial frequency	1.113
Spatial phase	1.008
Edge orientation	1.176
Texture	2.708
Global layout	2.870
Clutter/complexity	2.550

Conditional associations in measured EEG (Figure 3) peak for spatial phase at 80–120 ms (β=0.1339, 95% CI [0.1233, 0.1460], q=0.0251), texture at 80–120 ms (β=0.1419, 95% CI [0.1249, 0.1626], q=0.0176), and edge orientation at 120–160 ms (β=0.1335, 95% CI [0.1207, 0.1467], q=0.0176). Separate regional analyses show posterior associations (all corrected q$\approx$0.0083; Section 4.4). Figure 3c shows luminance/layout and texture/clutter overlap; VIFs span 1.008–2.920.

D.2 EEG INPUT OCCLUSION

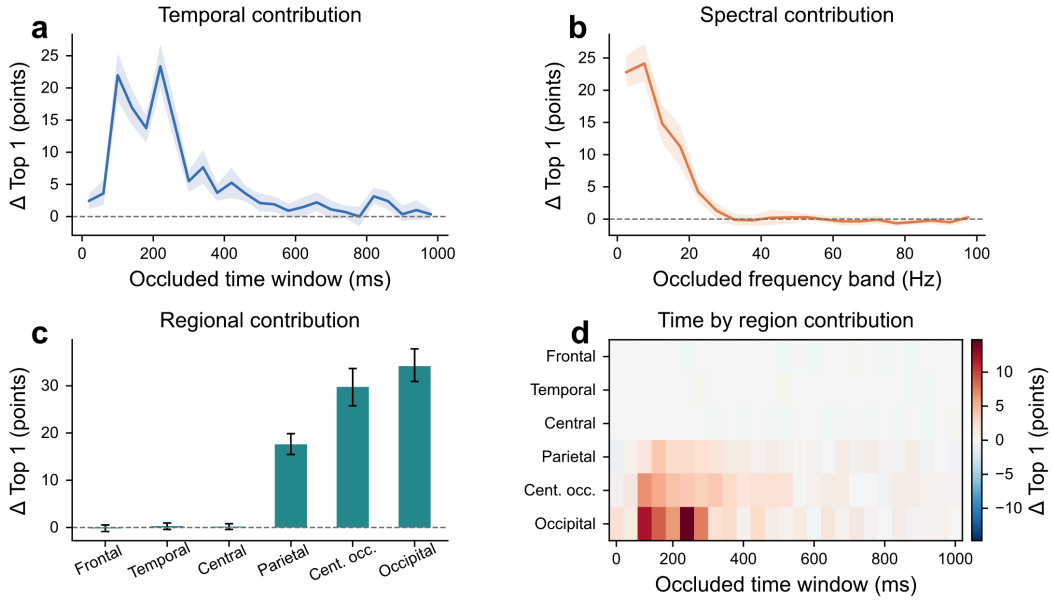

Figure A8: EEG input occlusion (ten subjects). Positive ΔTop-1 is the retrieval decrease in percentage points after removing an input; negative values indicate improvement. Lines and bars show means and 95% intervals from 10,000 subject bootstrap resamples. (a) Time, (b) frequency, (c) region and (d) time by region. Region abbreviations follow Figure 3. Posterior responses across early and later windows and activity at low frequencies contribute in all ten subjects, with varying magnitudes. Effects measure the frozen EEG Encoder's dependence on inputs at the sensor level.

Tables A7a–A7c give the prespecified MRR loss, full minus occluded input on the 0–1 scale, for every time window, frequency band and sensor region. Means and 95% percentile intervals use ten subjects and 10,000 subject bootstraps. Figure A8 shows complementary Top-1 losses.

Table A7a. Temporal occlusion across all 25 consecutive 40 ms windows. Uncorrected p values enumerate 2^{10} signs with the original +1 convention. Maximum cluster mass correction covers the temporal family. Cluster p values, repeated for constituent windows, test the cluster rather than individual windows; NA denotes no cluster.

Window (ms)	Mean ΔMRR	95% CI	Uncorrected p	Cluster p
0–40	0.0213	[0.0120, 0.0300]	0.004878	0.002927
40–80	0.0357	[0.0207, 0.0532]	0.002927	0.002927
80–120	0.2173	[0.1828, 0.2497]	0.002927	0.002927
120–160	0.1604	[0.1392, 0.1816]	0.002927	0.002927
160–200	0.1289	[0.1089, 0.1474]	0.002927	0.002927
200–240	0.2235	[0.1931, 0.2564]	0.002927	0.002927
240–280	0.1355	[0.1100, 0.1647]	0.002927	0.002927
280–320	0.0518	[0.0372, 0.0658]	0.002927	0.002927
320–360	0.0677	[0.0503, 0.0856]	0.002927	0.002927
360–400	0.0338	[0.0254, 0.0434]	0.002927	0.002927
400–440	0.0453	[0.0290, 0.0625]	0.002927	0.002927
440–480	0.0266	[0.0163, 0.0376]	0.002927	0.002927
480–520	0.0171	[0.0073, 0.0271]	0.014634	0.002927
520–560	0.0126	[0.0065, 0.0188]	0.006829	0.002927
560–600	0.0082	[0.0014, 0.0156]	0.067317	NA
600–640	0.0150	[0.0058, 0.0251]	0.022439	0.053659
640–680	0.0214	[0.0102, 0.0332]	0.008780	0.053659
680–720	0.0107	[0.0012, 0.0193]	0.059512	NA
720–760	0.0053	[−0.0015, 0.0121]	0.196098	NA
760–800	0.0035	[−0.0057, 0.0126]	0.490732	NA
800–840	0.0270	[0.0190, 0.0375]	0.002927	0.041951
840–880	0.0192	[0.0103, 0.0301]	0.002927	0.041951
880–920	0.0044	[−0.0029, 0.0123]	0.338537	NA
920–960	0.0088	[0.0001, 0.0174]	0.092683	NA
960–1000	0.0040	[−0.0011, 0.0091]	0.176585	NA

Table A7b. Spectral occlusion across all twenty 5 Hz bands. Values, intervals, exact sign configurations and +1 convention follow Table A7a. Maximum cluster mass correction covers the frequency family; NA denotes no cluster.

Band (Hz)	Mean ΔMRR	95% CI	Uncorrected p	Cluster p
0–5	0.2319	[0.2138, 0.2505]	0.002927	0.002927
5–10	0.2314	[0.2018, 0.2622]	0.002927	0.002927
10–15	0.1347	[0.1101, 0.1594]	0.002927	0.002927
15–20	0.1029	[0.0768, 0.1311]	0.002927	0.002927
20–25	0.0390	[0.0277, 0.0505]	0.002927	0.002927
25–30	0.0108	[0.0042, 0.0179]	0.018537	0.002927
30–35	0.0017	[−0.0034, 0.0069]	0.557073	NA
35–40	0.0002	[−0.0039, 0.0044]	0.920000	NA
40–45	0.0018	[−0.0051, 0.0089]	0.633171	NA
45–50	0.0008	[−0.0048, 0.0067]	0.822439	NA
50–55	0.0012	[−0.0025, 0.0047]	0.535610	NA
55–60	−0.0005	[−0.0028, 0.0019]	0.666341	NA
60–65	−0.0029	[−0.0059, 0.0011]	0.176585	NA
65–70	−0.0023	[−0.0054, 0.0014]	0.240976	NA
70–75	−0.0006	[−0.0038, 0.0026]	0.717073	NA
75–80	−0.0042	[−0.0068, −0.0018]	0.018537	0.225366
80–85	−0.0024	[−0.0053, 0.0001]	0.118049	NA
85–90	−0.0004	[−0.0033, 0.0024]	0.806829	NA
90–95	−0.0026	[−0.0057, 0.0005]	0.147317	NA
95–100	0.0005	[−0.0029, 0.0034]	0.793171	NA

Table A7c: Regional occlusion. All six mutually exclusive sensor groups are retained. The q values apply Benjamini–Hochberg FDR correction across these six regions to the uncorrected p values shown.

Region	Mean ΔMRR	95% CI	Uncorrected p	BH q
Frontal	−0.0006	[−0.0040, 0.0035]	0.775610	0.775610
Temporal	0.0033	[−0.0018, 0.0083]	0.278049	0.417073
Central	0.0017	[−0.0016, 0.0052]	0.447805	0.537366
Parietal	0.1703	[0.1515, 0.1887]	0.002927	0.005854
Centro-occipital	0.3008	[0.2644, 0.3355]	0.002927	0.005854
Occipital	0.3544	[0.3175, 0.3932]	0.002927	0.005854

E CONTROLLED PERTURBATIONS AND SEVERITY ANALYSES

E.1 PERTURBATION OPERATORS AND NATIVE VISUAL REPRESENTATIONS

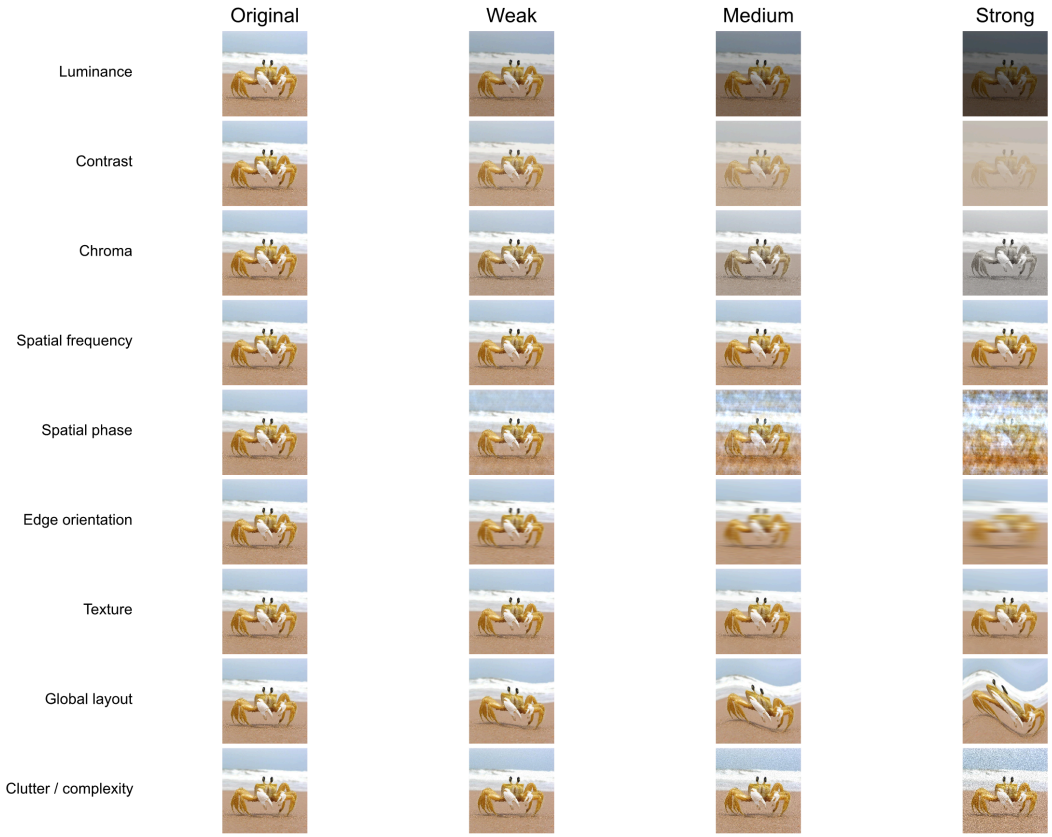

Figure A9: Complete operator examples for all nine families at Original, Weak, Medium and Strong doses. Operator quality control is reported in Figure A10. Fixed parameters remain in Table A1.

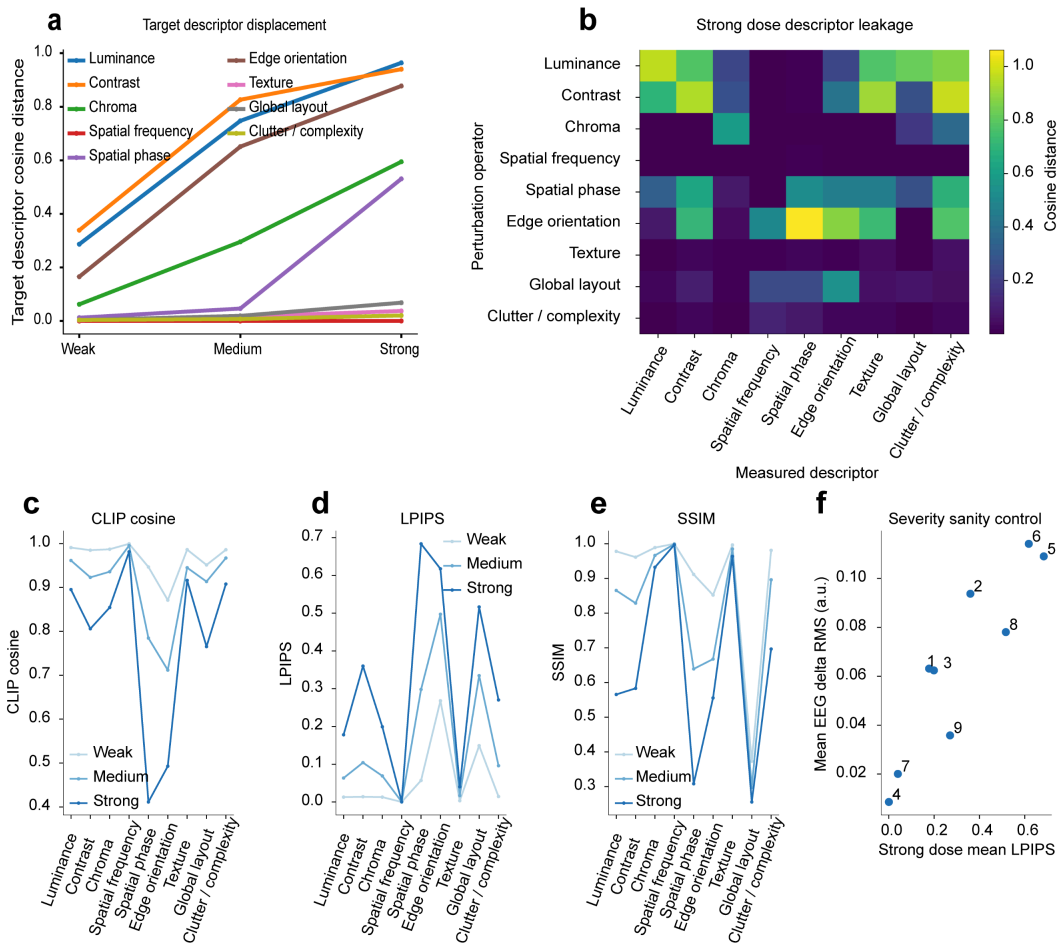

Figure A10: Operator quality control. (a) Displacement of the target descriptor across doses. (b) Changes in all nine descriptors at Strong dose. (c–e) CLIP similarity, LPIPS and SSIM. (f) Mean LPIPS at Strong dose versus EEG RMS change; labels 1–9 follow b. Descriptors match the analysis of measured EEG. Nominal dose is an operator setting; LPIPS and native displacement quantify effective severity. Figure A9 gives complete examples; Table A1 gives parameters.

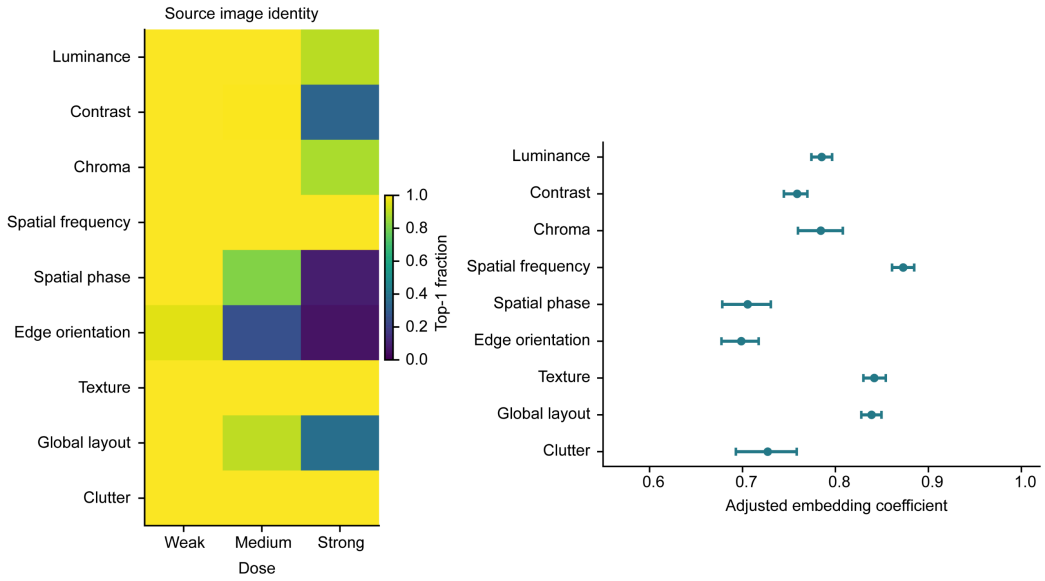

Figure A11: Native identity and associations between embeddings and EEG. Top: Top-1 retention of the source image in a gallery of 200 frozen original embeddings at Weak, Medium and Strong doses. Bottom: standardized coefficients for native displacement controlling LPIPS at Strong dose (means and 95% intervals, ten subjects); every family has ten positive coefficients. Figure 4a,c gives displacement and additional fitted R^2 (0.355380–0.741859 across families). These fitted gains within each feature differ from pooled R^2 for severity models evaluated by image-grouped cross-validation.

E.2 GENERATED RESPONSES ACROSS FEATURE FAMILIES

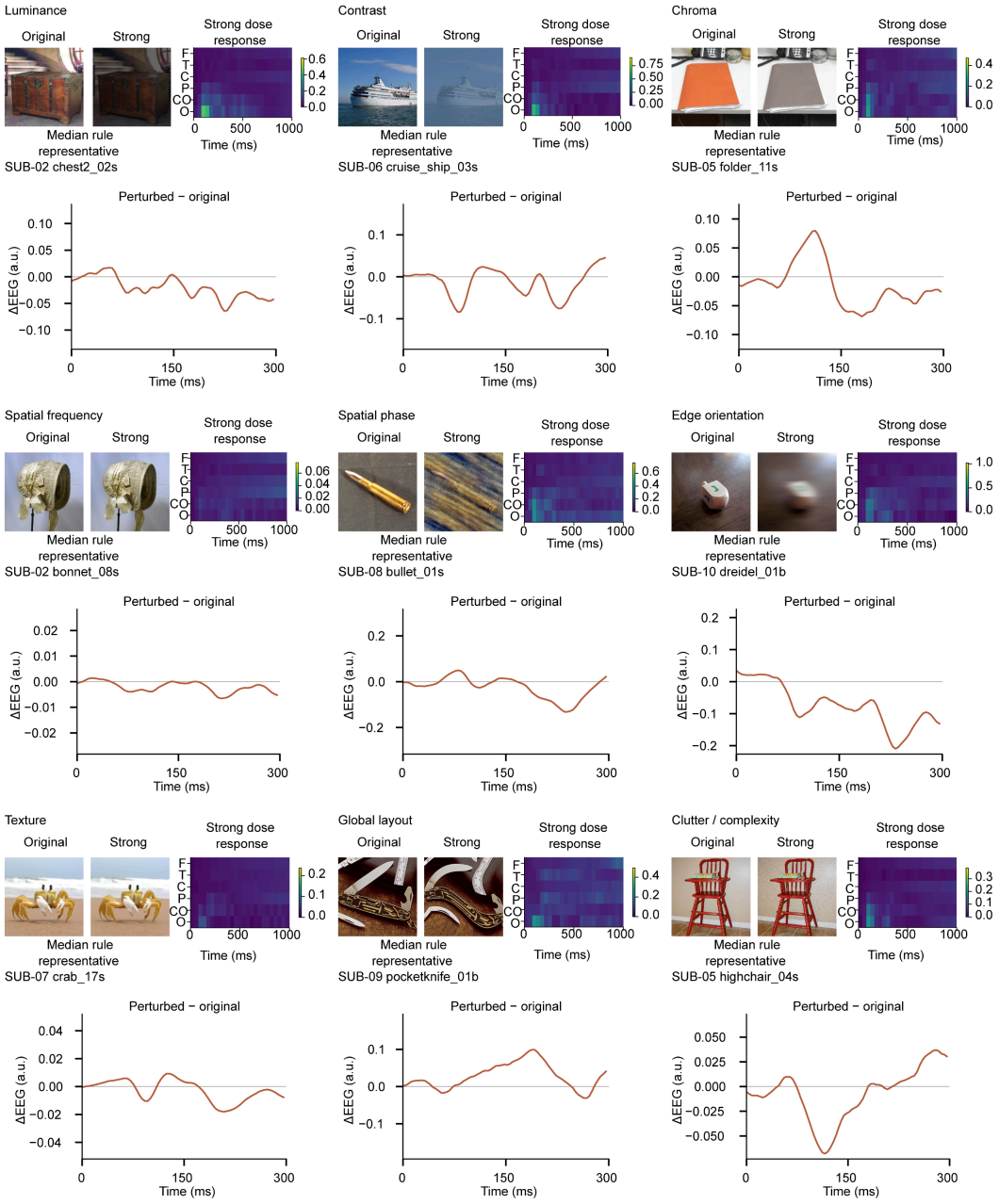

Figure A12: Representative perturbation outputs for all nine feature families. Each case selected by the median rule shows Original and Strong stimuli, the signed Strong minus Original waveform and its regional RMS map. K = 80; waveforms use whitened EEG units. F, frontal; T, temporal; C, central; P, parietal; CO, centro-occipital; O, occipital. Figure A3b provides the ten subject direct and cycle generation examples.

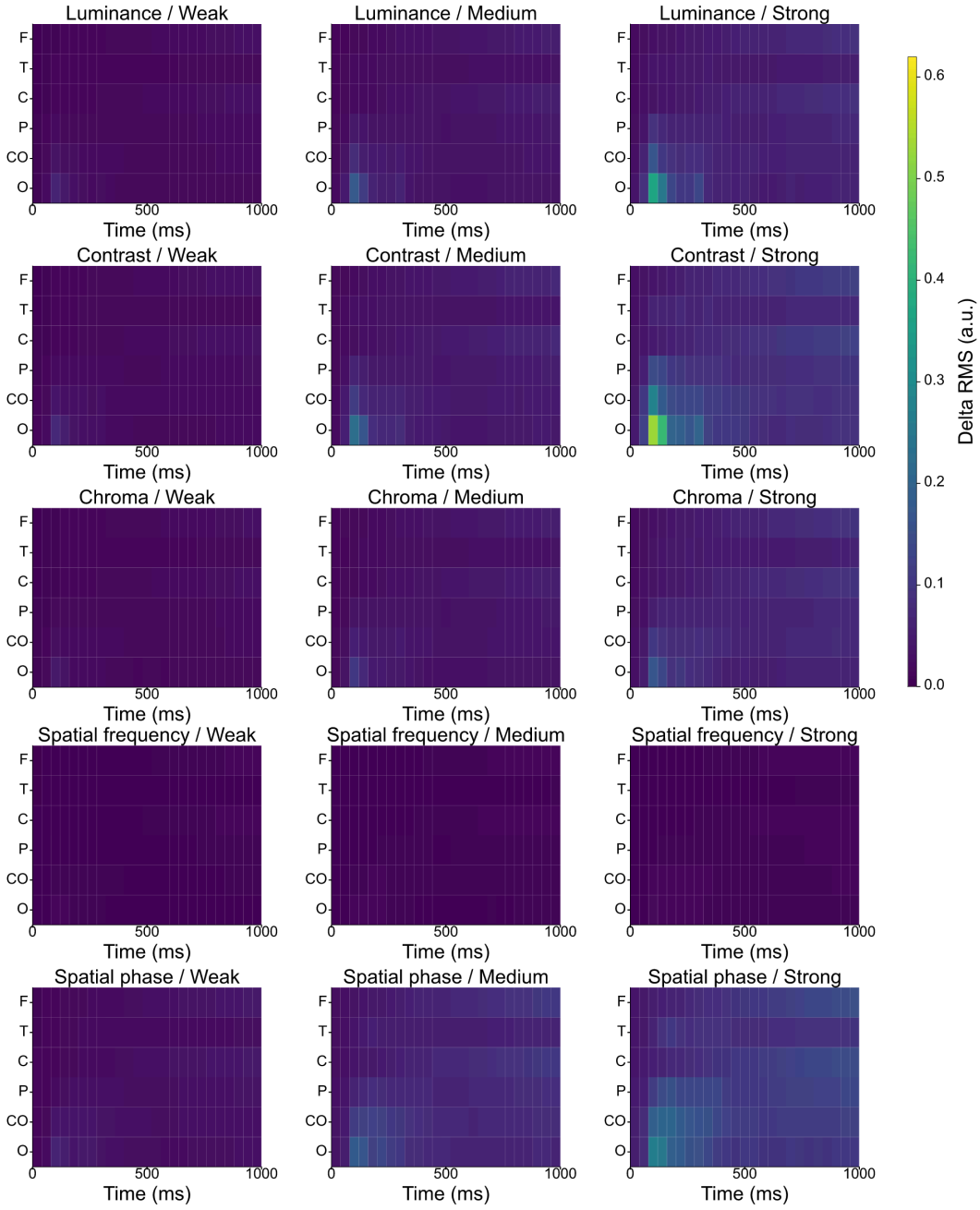

Figure A13: (part 1 of 2) Complete perturbation response atlas. Complete Weak, Medium and Strong regional RMS maps for Luminance, Contrast, Chroma, Spatial frequency and Spatial phase. Responses average 80 samples per condition, then 200 images within subject and ten subjects. All six sensor groups and the full 0–1000 ms epoch are retained. RMS is the magnitude of the difference from Original.

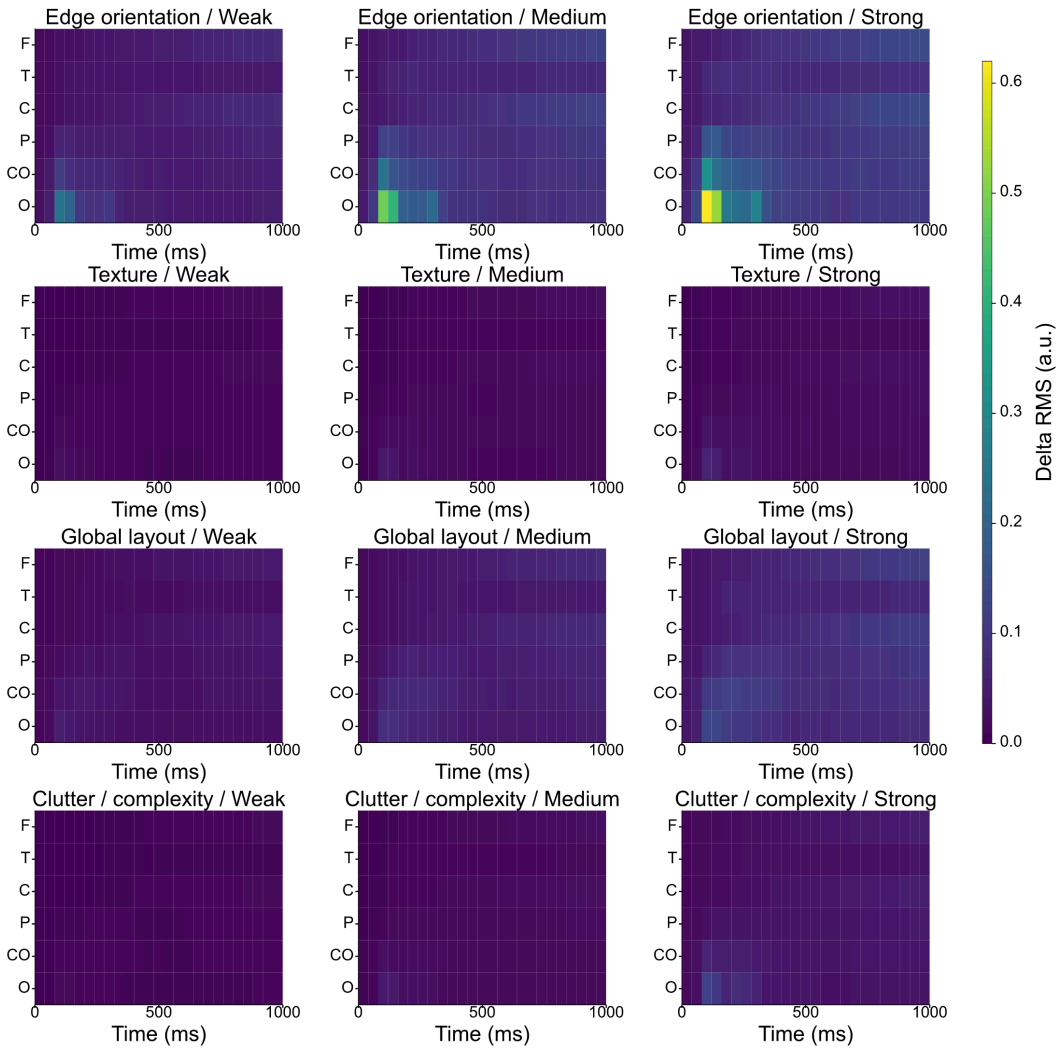

Figure A13: (part 2 of 2) Complete perturbation response atlas. Complete Weak, Medium and Strong maps for Edge orientation, Texture, Global layout and Clutter. Responses average 80 samples per condition, then 200 images within subject and ten subjects. All six sensor groups and the full 0–1000 ms epoch are retained. RMS is the magnitude of the difference from Original; panel color keys retain their original ranges.

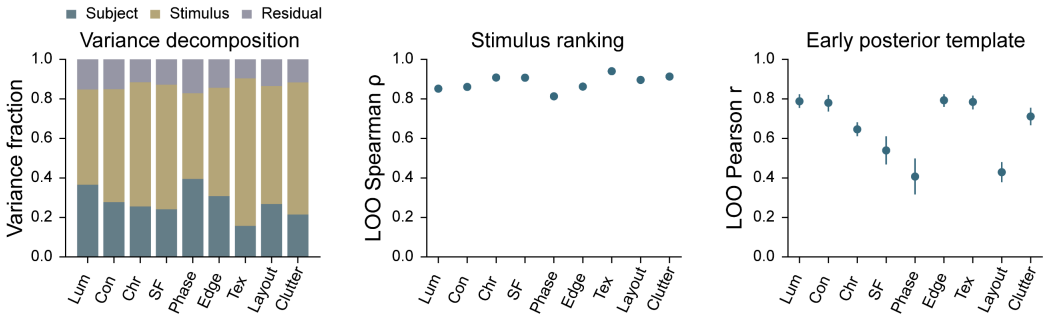

Figure A14: Heterogeneity at Strong dose (n=10). Left: subject, stimulus and residual variance fractions; residuals include nonadditivity between subjects and stimuli. Stimulus fractions are 43.3–74.7%. Center: Spearman correlation of stimulus rankings with each subject held out in turn (means 0.813–0.940). Right: early posterior template Pearson correlation with the other nine subjects' mean (means and 95% intervals). Feature order follows Table A1.

E.3 Organization that varies by feature and consistency across subjects

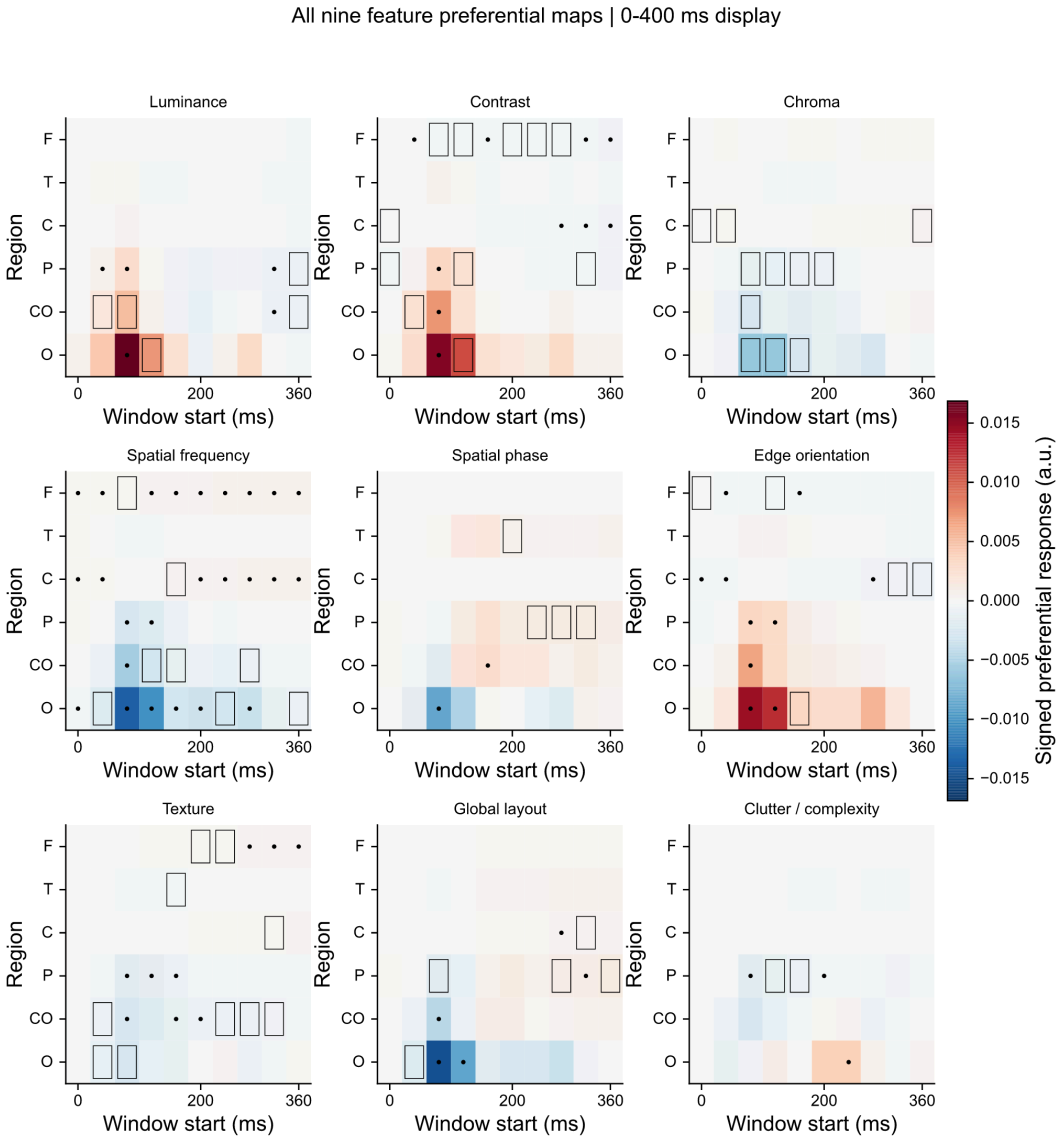

Figure A15: Relative response organization for nine families over 0–400 ms. Signed values compare each family's normalized dose response with the other eight families' mean. Dots mark global maxT p<0.05 across 1,350 cells (227 significant combinations of features and cells); outlines mark secondary correction within each family. Both corrections use the full epoch. Figure A16 gives complete identification metrics, secondary relative patterns and original full/early confusion matrices (b,d); Figure A18b gives strict residual matrices.

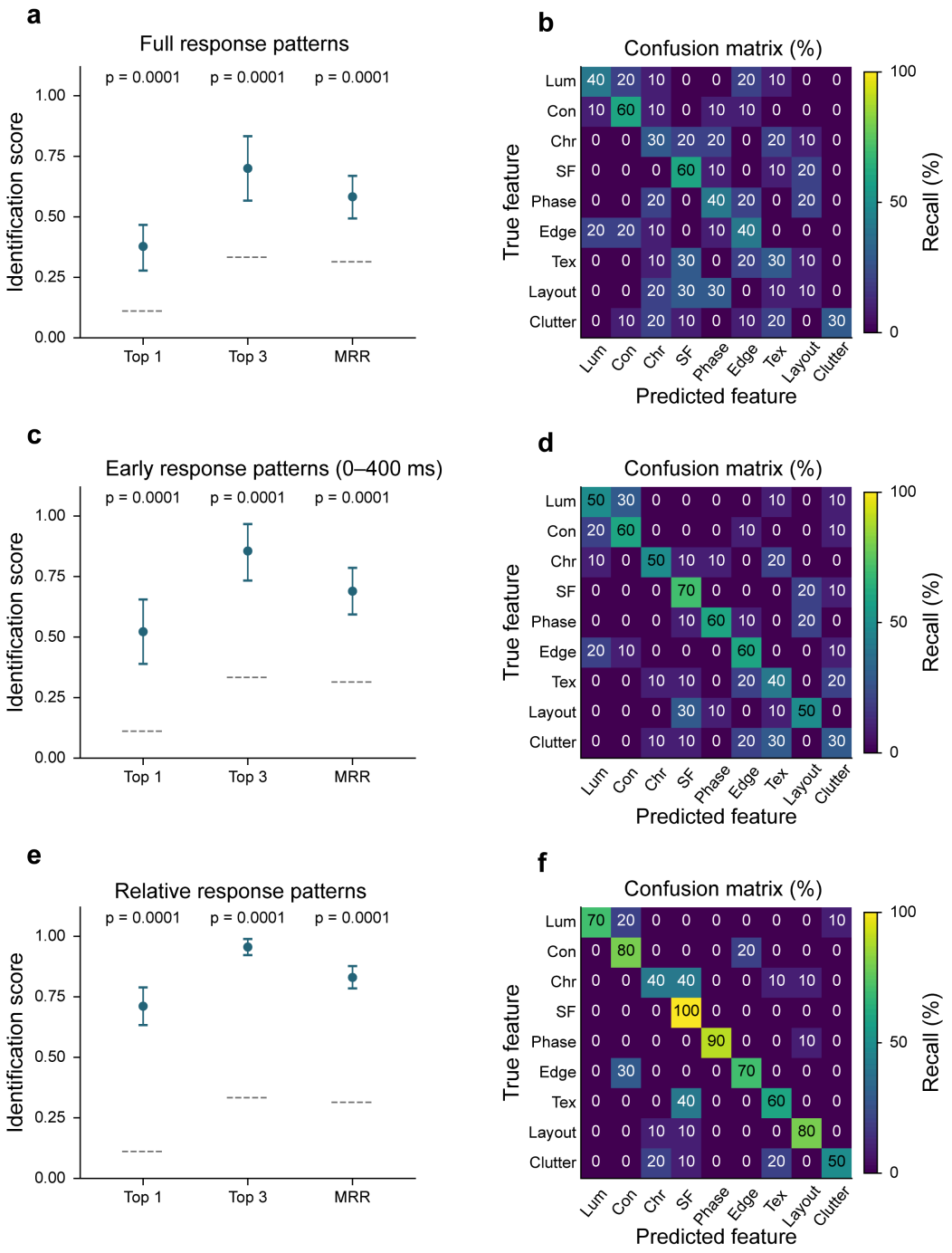

Figure A16: Perturbation-family identification. Rows show full (0–1000 ms), early (0–400 ms) and relative patterns. (a,c,e) Top-1, Top-3 and MRR: means and 95% intervals from 10,000 subject bootstraps (n=10); dashed segments mark chance. Tests use 10,000 label permutations within each subject. Original full/early dose-response slopes with L1 normalization have Top-1 chance 1/9 (11.11%) and p=0.0001 each. (b,d,f) Confusion matrices: true features in rows, predictions in columns, row percentages over ten queries (10% per query). Identification using the full map is primary; early and relative analyses are secondary. Global layout has no correct assignments for the full map. Figure A18b gives strict residual matrices; procedures appear in Appendices A.4.5 and A.4.8.

E.4 SEVERITY SUPPORT AND PERTURBATION-FAMILY IDENTIFICATION FROM RESIDUALS

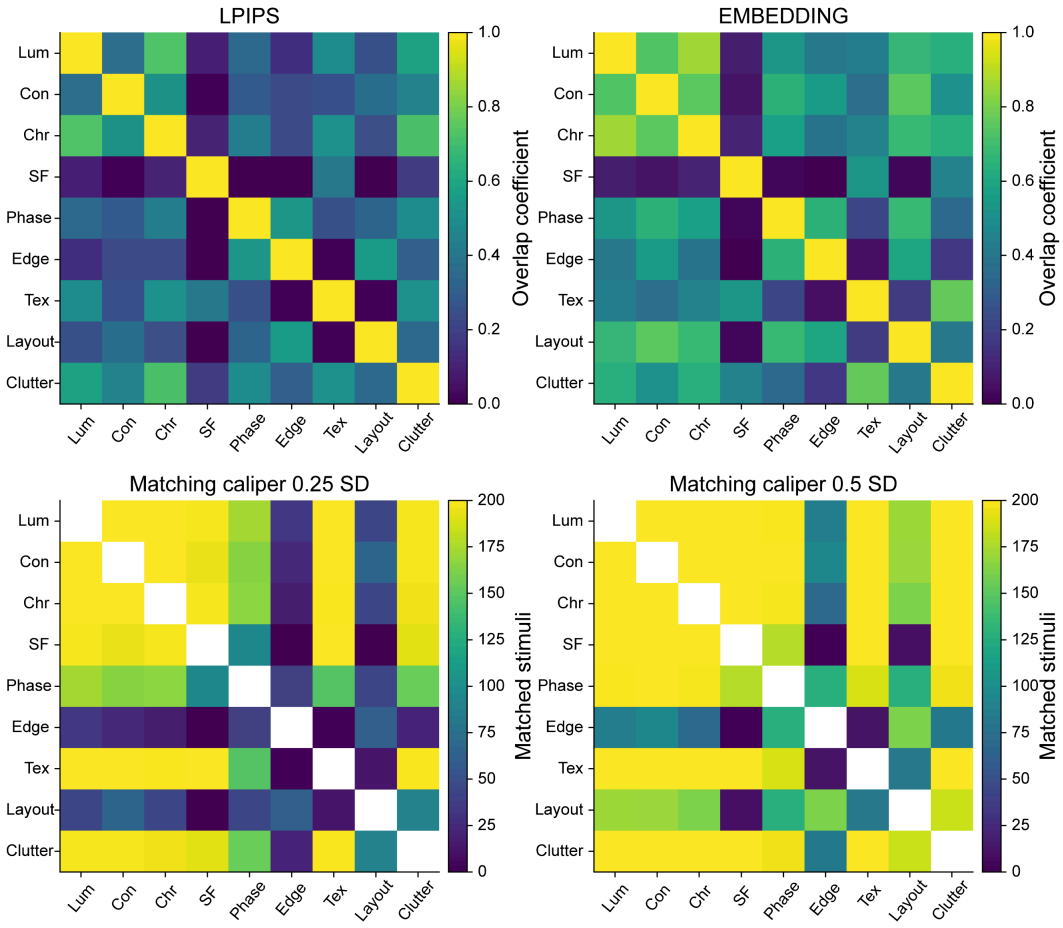

Figure A17: (part 1 of 2) Common LPIPS/embedding support and matching within each stimulus (Appendix A.4.7). Pairs require $\geq$30 matched images. Calipers of 0.25 and 0.50 SD support 29 of 36 and 33 pairs, respectively; seven pairs lack primary 0.25 SD support. With ten subjects, the minimum two-sided exact p is 2/1,024 and the minimum first p value after Holm adjustment is 0.056640625 across 29 tests. No eligible pair passes Holm 0.05 at either caliper; this does not establish equivalence.

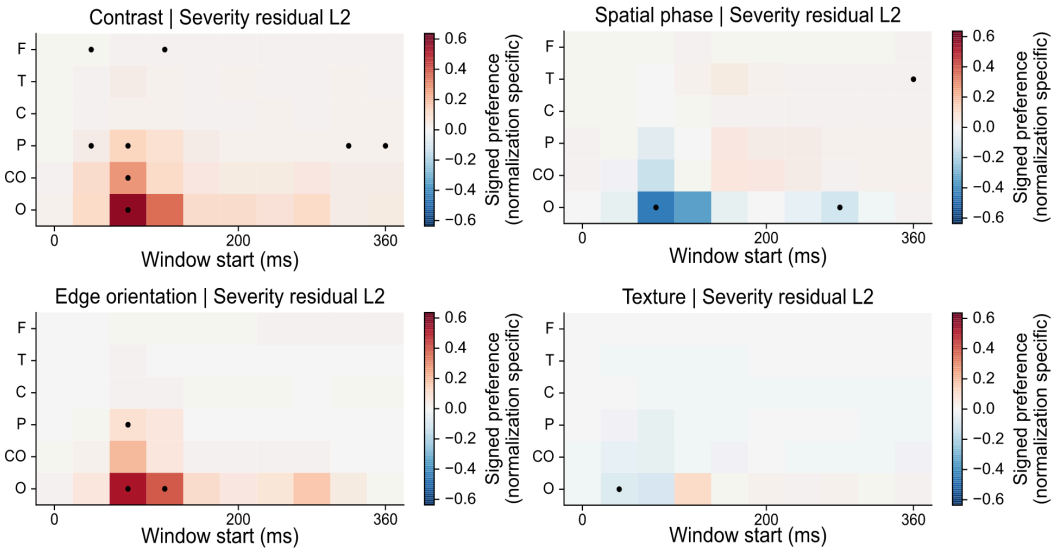

Figure A17: (part 2 of 2) Severity residuals relative to other features over 0–400 ms: contrast and spatial phase (top), edge orientation and texture (bottom). Maps are mean residuals with L2 normalization from fits of LPIPS, native displacement and separate quadratic terms without feature labels. Colors show signed preference relative to the other eight families; dots mark global maxT p<0.05 across 1,350 cells spanning the full epoch, conditional on fitted nuisance models. Original L1 dose-response slopes (Figure A15) have different units. Strict outer validation is separate (Appendix A.4.8; matrices in Figure A18b).

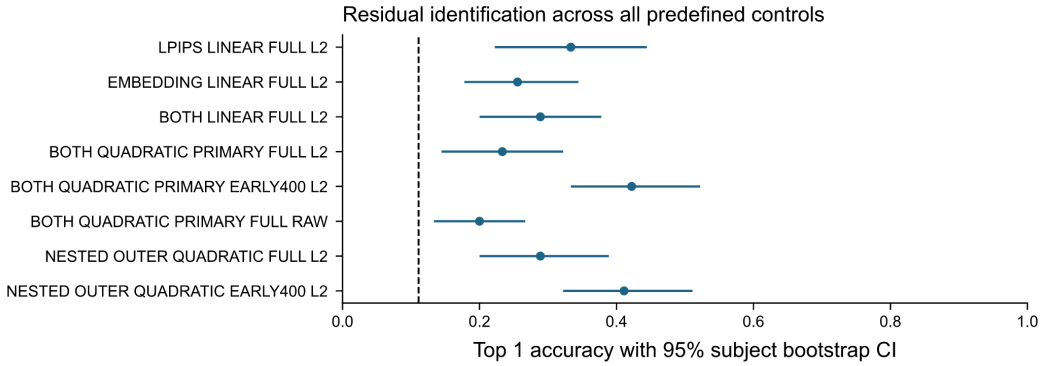

Figure A18a: Eight predefined severity controls, top to bottom: linear LPIPS; linear embedding; joint linear; joint quadratic, full maps; joint quadratic, early (0–400 ms) maps; joint quadratic, full raw residual maps; nested outer joint quadratic models, full maps; and nested outer joint quadratic models, early maps. Points show Top-1 and 95% subject bootstrap intervals; the dashed line marks 1/9 chance. L2 denotes normalized residual maps; raw denotes unnormalized maps. The last two controls exclude query subjects from severity fitting and template preprocessing. Original L1 dose-response slopes and L2 residual means are different representations. Figure A17 gives overlap, matching support, test resolution and residual maps.

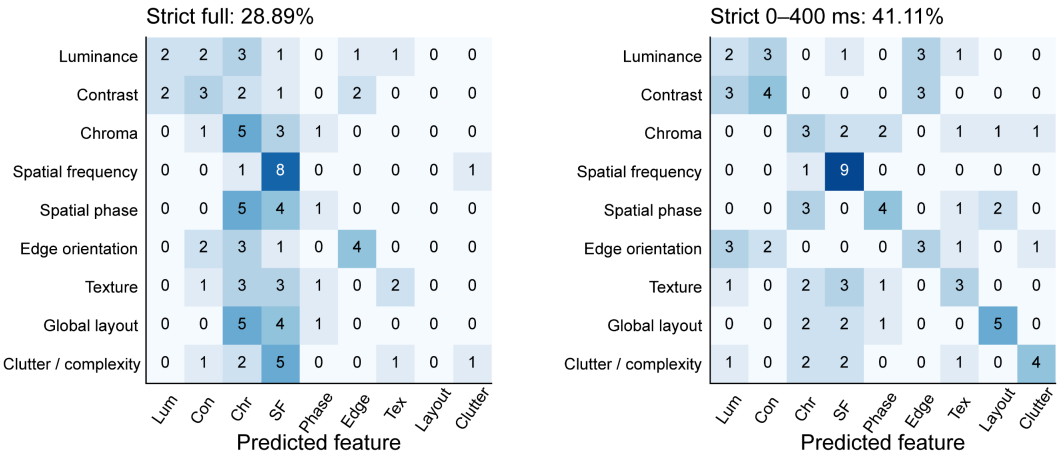

Figure A18b: Confusion matrices with strict severity control: full epoch and 0–400 ms. Rows are true families and columns predicted families. Each row contains ten subject queries; cell shading shows counts from 0 to 10. Full/early accuracy is 28.89%/41.11% (chance 11.11%; permutation p=0.0001 each). Original L1 dose-response slopes (Figure A16b,d) and strict L2 residual means differ in representation and preprocessing; their difference does not isolate severity adjustment.

E.5 Correspondence with measured EEG

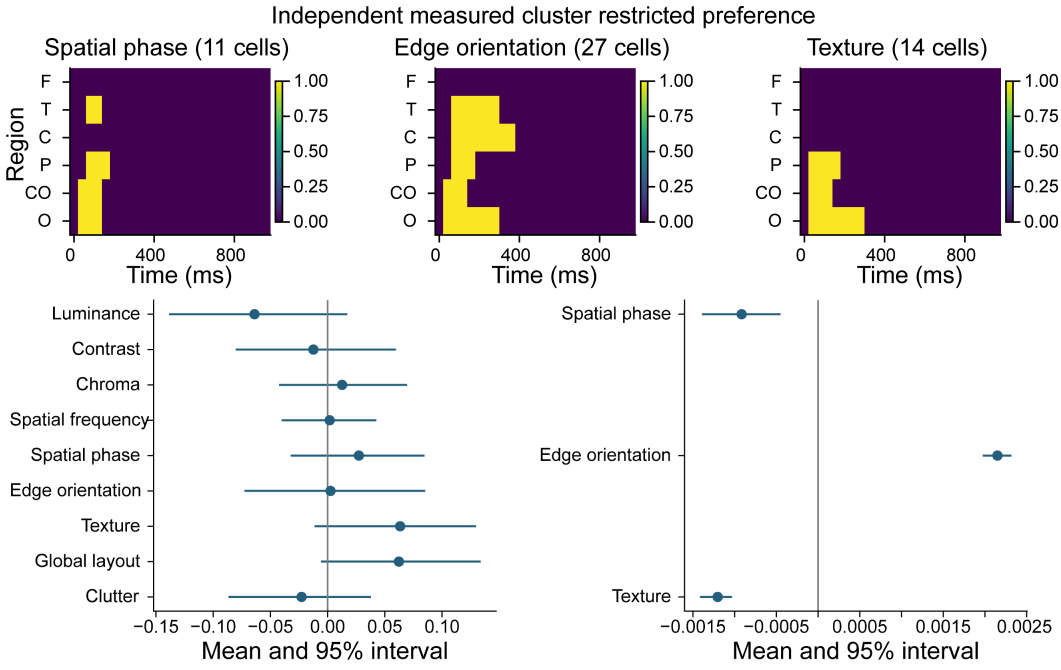

Figure A19: Global and local correspondence with measured EEG. Top: independent masks for spatial phase (11 cells), edge orientation (27) and texture (14). Bottom: global alignment advantage for the same feature (left; all nine 95% intervals include zero) and restricted preference (right; separate metric/axis). Points show means with 95% intervals. Within the edge-orientation mask, restricted preference was positive in all ten subjects (mean = 0.00215, 95% CI [0.00197, 0.00232]; one-sided exact (p=0.000977), Holm-adjusted (p=0.00293)). Spatial phase and texture did not show significant positive preference. Figure A20 gives extended maps.

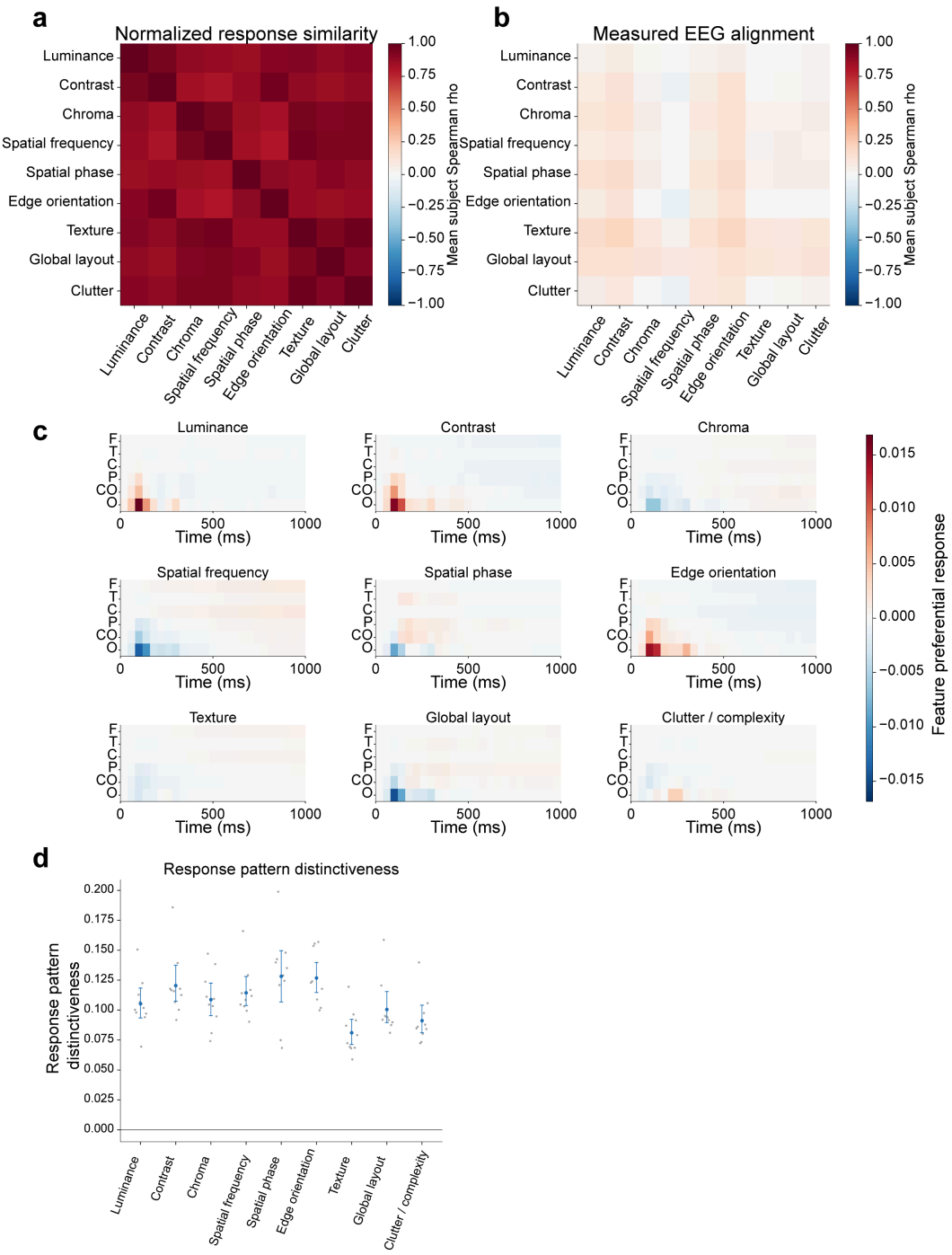

Figure A20: Global organization and correspondence with measured EEG. (a) Mean of the Spearman similarities computed within each subject for dose-response slopes with L1 normalization. (b) Mean Spearman correspondence between absolute ridge maps of measured EEG (rows) and perturbation maps (columns). (c) Mean relative responses across six groups and 0–1000 ms, with a shared diverging scale. (d) Distinctiveness: one minus mean correlation with the other eight patterns. Small points show ten subjects; large points/bars show means and 95% intervals from 10,000 subject bootstraps. Figure A19 gives global advantage for the same feature and localized preference.