RETHINKING EEG-TO-IMAGE DECODING: FROM FINE-GRAINED TEMPORAL ENCODING TO HIERARCHY-CONSISTENT DESIGN

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing EEG-to-image methods commonly use specialized encoders, cross-modal alignment, and generative decoders, but it remains unclear what each stage actually requires or whether added model complexity helps. In this work, we systematically revisit each stage of this common pipeline and introduce TRACE (**T**emporal **R**esolution-**A**ware **C**omplementary **E**EG Decoding) based on our analysis. At the **encoder level**, controlled comparisons reveal a counterintuitive trend where a minimal MLP outperforms Transformer-based and large pretrained baselines; our further analysis shows that decoding performance depends strongly on *fine-grained temporal encoding*, whereas increasing encoder capacity provides no systematic benefit. At the **alignment and decoder levels**, our analysis identifies unclear separation of level-specific evidence and hierarchical conditioning mismatches. We therefore propose *Complementary Dual-Subspace Alignment* to organize semantic and structural visual cues into complementary subspaces, alongside *Hierarchy-Consistent Injection* to route them through a coarse-to-fine visual autoregressive (VAR) generator. On THINGS-EEG, TRACE achieves state-of-the-art retrieval and reconstruction performance. Averaged across ten subjects, it achieves the highest Top-1/Top-5 retrieval accuracy (77.5%/96.2%) and improves six of seven reconstruction metrics over prior methods. Code will be released upon acceptance.

1 INTRODUCTION

Decoding perceived visual content from non-invasive brain activity is a central goal in brain–computer interfaces (BCIs). Electroencephalography (EEG) is particularly compelling for this task due to its millisecond temporal resolution, low cost, and portability. Current visual EEG decoding focuses on two primary objectives: image retrieval from a candidate pool (Song et al., 2024) and pixel-level reconstruction (Li et al., 2024; Bai et al., 2023). Both tasks remain challenging due to low signal-to-noise ratios and the substantial domain gap between EEG signals and natural image spaces. To bridge this gap, current methods have converged on a common pipeline (Fig. 1a): an EEG encoder produces a representation, an alignment stage maps it into an image-feature space, and a generative decoder renders the image from the aligned features.

One prominent trend pursues performance gains through increasingly complex architectures: convolutional backbones (Lawhern et al., 2018; Song et al., 2024), Transformer-based encoders (Li et al., 2024; Zhou et al., 2026), large pretrained EEG models (Jiang et al., 2024), and multi-scale generative decoders (Tian et al., 2024; Dai et al., 2026), implicitly assuming that higher capacity extracts more usable signal from noisy EEG. Module-replacement ablations are common in this literature (Li et al., 2024; Dai et al., 2026; Zheng et al., 2026; Zhang et al., 2026), but the comparison serves only to select the best-performing option; which property of the EEG representation drives the difference is left unanalyzed. We therefore systematically analyze each stage of the common pipeline.

We first analyze the **encoder**, where controlled replacement produces the largest performance changes. Benchmarks across representative architectures—MLP-based EEGProject (Wu et al., 2025), convolutional TSConv (Song et al., 2024), Transformer-based ATM (Li et al., 2024), and pretrained LaBraM (Jiang et al., 2024)—reveal a counterintuitive trend: the simplest MLP encoder

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

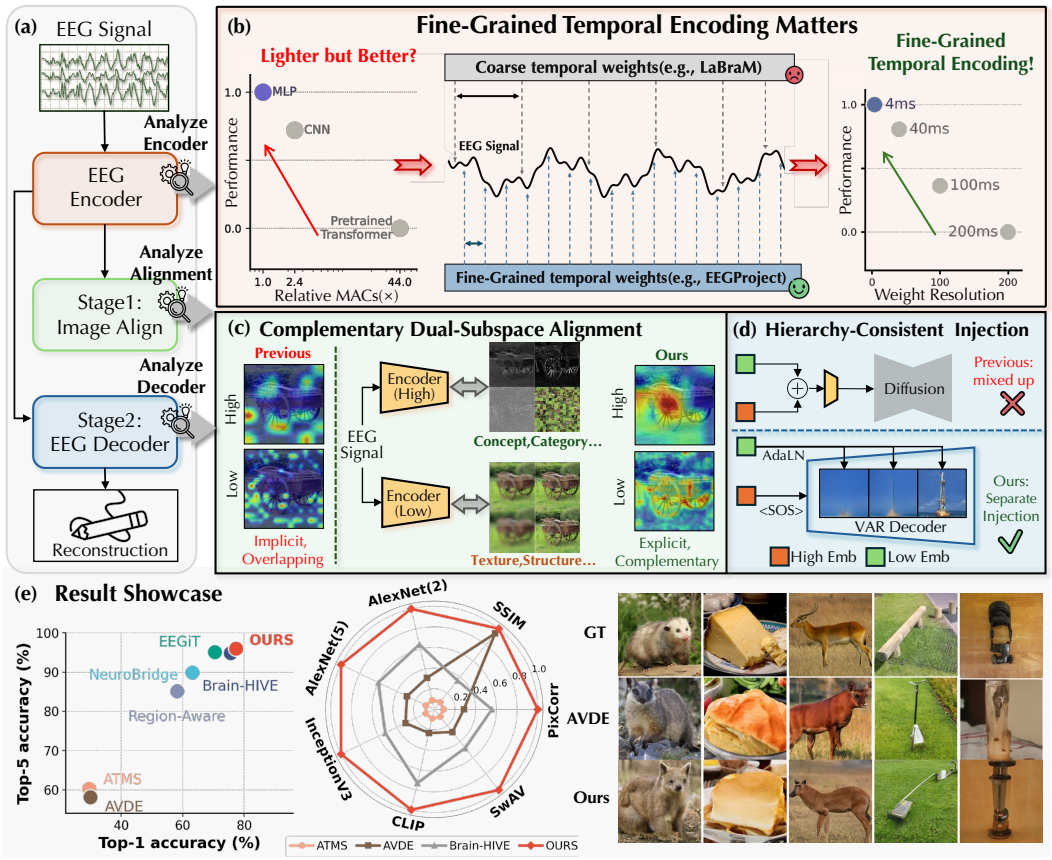

Figure 1: **From analysis to design.** (a) Common EEG-to-image pipeline, comprising encoder, alignment, and decoder stages. (b) Controlled encoder analysis shows that a simple MLP achieves higher performance than the tested CNN and pretrained Transformer encoders. Further analysis attributes this reversal to *fine-grained temporal encoding* rather than model capacity. (c) EEG-CAM comparisons reveal overlapping level-specific responses in Brain-HIVE (Zheng et al., 2026); Complementary Dual-Subspace Alignment yields a cleaner separation of semantic and structural evidence (Appendix A.10). (d) *Hierarchy-Consistent Injection* uses VAR's coarse-to-fine hierarchy to route each subspace to its corresponding decoder stage. (e) Retrieval and reconstruction comparisons with AVDE (Dai et al., 2026), ATM (Li et al., 2024), and Brain-HIVE (Zheng et al., 2026).

outperforms the heavier alternatives (Fig. 1b). In-depth analysis shows that *fine-grained temporal encoding*, the ability to resolve and exploit fine-grained temporal variations in EEG, strongly influences decoding performance.

At the **alignment level**, visual representations contain partially separable semantic and structural information (Cichy et al., 2014; DiCarlo et al., 2012). Our EEG-CAM comparison with Brain-HIVE (Zheng et al., 2026) reveals overlapping level-specific responses (Fig. 1c; Appendix A.10). This motivates Complementary Dual-Subspace Alignment, which aligns two independent EEG branches with level-specific augmented visual targets to obtain a cleaner separation of semantic and structural evidence.

At the **decoder level**, existing methods typically collapse multi-level features into a single condition (Zheng et al., 2026), failing to exploit their distinct roles. Although recent work such as AVDE (Dai et al., 2026) adopts a coarse-to-fine decoder (VAR (Tian et al., 2024)), it still lacks hierarchical conditioning. Resolving this mismatch calls for *Hierarchy-Consistent Injection* (Fig. 1d): semantic features initialize the coarse generation stage, while structural features guide subsequent refinement. Targeted injection experiments directly validate the effectiveness of this organization.

Based on this analysis, we introduce TRACE (**T**emporal **R**esolution-**A**ware **C**omplementary **E**EG Decoding), which instantiates the three findings as one pipeline: a temporally fine-grained encoder, Complementary Dual-Subspace Alignment for complementary semantic and structural subspaces, and *Hierarchy-Consistent Injection* that routes each subspace to its matching stage of a VAR decoder. On the THINGS-EEG benchmark (Gifford et al., 2022), TRACE achieves state-of-the-art retrieval and reconstruction performance (Fig. 1e). Averaged across ten subjects, it leads in Top-1/Top-5 retrieval accuracy and improves six of seven reconstruction metrics over prior methods.

Our contributions: **Mechanistic encoder analysis.** We show that fine-grained temporal encoding, rather than model capacity, strongly influences EEG encoder performance through temporal analyses and controlled interventions. **Analysis-informed design.** We present TRACE, which aligns independent EEG branches with complementary semantic and structural visual targets and routes them to matching stages of a VAR decoder through *Complementary Dual-Subspace Alignment* and *Hierarchy-Consistent Injection*. **Analysis-to-design validation.** Controlled branch-wise retrieval and routing ablations support the complementarity of the two subspaces and the benefit of hierarchy-consistent injection.

2 Related Work

Brain-to-visual alignment and reconstruction. Brain-to-visual decoding commonly begins with cross-modal alignment: an EEG encoder is trained to match the visual representation of the perceived image for retrieval (Song et al., 2024). Subsequent work enriches this bridge with softer visual targets or auxiliary priors, including blur-based supervision (Wu et al., 2025), complementary modalities (Zhang et al., 2025), region-aware matching (Zhu et al., 2026), and self-supervised cognitive priors (Zhang et al., 2026). For pixel-level reconstruction, diffusion-based systems use these aligned features to condition a powerful image generator (Li et al., 2024; Bai et al., 2023; Lu et al., 2023; Fu et al., 2025; Lan et al., 2023; Lopez et al., 2025; Yu et al., 2026; Takagi & Nishimoto, 2023; Ozcelik & VanRullen, 2023; Scotti et al., 2023; 2024), while retrieval-augmented methods provide additional visual support (Xie et al., 2024; Zhu et al., 2025). These approaches improve decoding quality, but often change encoder, alignment, and decoder choices together, making it difficult to determine which properties of the EEG representation are actually useful.

EEG encoder architectures and temporal assumptions. The EEG encoder has the least standardized design space in the pipeline: encoders differ not only in capacity but also in the effective temporal granularity imposed by smoothing, aggregation, shared filters, and output sampling. Convolutional models such as EEGNet (Lawhern et al., 2018) and TSConv (Song et al., 2024) use shared temporal filters and pooling; ATM (Li et al., 2024) and EEGiT (Zhou et al., 2026) introduce patching and attention; LaBraM (Jiang et al., 2024) combines temporal compression with large-scale pretraining; and the simple EEGProject baseline (Wu et al., 2025) retains position-specific time samples through an MLP. Existing studies typically evaluate these choices end to end, leaving temporal assumptions entangled with architecture and capacity. Neuroscience work shows that visual codes evolve over time: temporal generalization tests whether a code transfers across moments (King & Dehaene, 2014), while studies on brain signals link different temporal stages to different levels of visual representation (Cichy et al., 2014). These findings motivate treating *fine-grained temporal encoding* as a distinct design factor rather than a by-product of architecture or capacity.

Hierarchical representations and generation. Visual representations contain complementary semantic and structural information (DiCarlo et al., 2012; Yamins et al., 2014). MindDiffuser separates these roles in fMRI decoding (Lu et al., 2023), while Brain-HIVE extracts multi-level visual features but fuses them into a single alignment token (Zheng et al., 2026). On the generation side, AVDE adopts a coarse-to-fine VAR decoder (Dai et al., 2026; Tian et al., 2024), but its EEG input remains a single embedding from a heavily pretrained encoder. In contrast, guided by the encoder analysis, TRACE uses Complementary Dual-Subspace Alignment and assigns the resulting semantic and structural subspaces distinct roles in coarse-to-fine generation.

3 Method

On paired EEG–image observations (x_E, x_I), TRACE instantiates each pipeline stage from our analysis; retrieval evaluates the aligned representation, reconstruction the full pipeline.

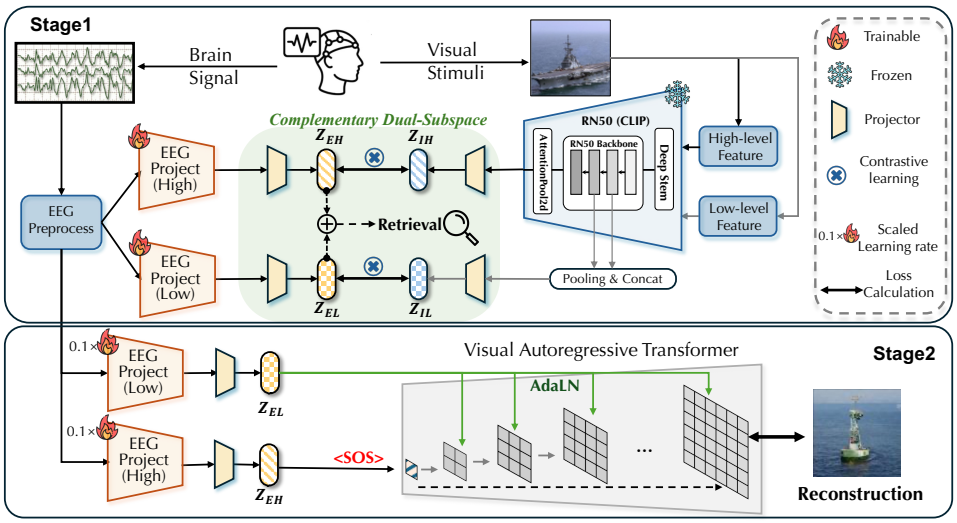

Figure 2: **Overview of TRACE.** In Stage 1, Complementary Dual-Subspace Alignment maps two EEG branches to complementary high- and low-level visual targets. In Stage 2, the semantic branch initializes the coarsest VAR scale and the structural branch modulates generation throughout refinement. The final reconstruction model uses the Stage-1 EEGProject encoder, fine-tuned with an encoder learning rate scaled by 0.1.

3.1 Encoder: Fine-Grained Temporal Encoding

To analyze encoder inductive bias, we vary only the EEG encoder while keeping the data, visual targets, alignment objective, and optimization protocol fixed. We compare four representative families: the position-preserving MLP EEGProject (Wu et al., 2025), convolutional TSConv (Song et al., 2024), attention-based ATM (Li et al., 2024), and pretrained LaBraM (Jiang et al., 2024). We define *fine-grained temporal encoding* as an encoder's ability to resolve and exploit fine-grained temporal variations in EEG. We test this capability using controlled capacity sweeps, temporal generalization and attribution, and bidirectional interventions that coarsen or restore effective temporal processing; their results and implementation details are reported in Sec. 4.2 and the appendix.

EEGProject encoder used by TRACE. The final system uses two independently instantiated EEGProject branches. Each branch receives the 17 selected EEG channels over 250 post-stimulus samples, flattens the input to $17 \times 250 = 4250$ dimensions, and applies

$$y = W_1 \,\mathrm{vec}(x_E) + b_1, \qquad f_E(x_E) = \mathrm{LN}(y + \mathrm{Dropout}_{0.3}(W_2\,\mathrm{GELU}(y) + b_2)), \tag{1}$$

where $W_1 \in \mathbb{R}^{D\times 4250}$ and $W_2 \in \mathbb{R}^{D\times D}$. The high-level and low-level branches use $D_H = 1024$ and $D_L = 1536$, respectively. There is no convolution, temporal pooling, or explicit permutation before flattening; thus each branch retains position-specific weights on the input sample grid. The two branches do not share encoder parameters; bias dimensions and parameter counts are given in Appendix A.6.

3.2 Alignment: Complementary Dual-Subspace Alignment

Guided by the encoder analysis, Complementary Dual-Subspace Alignment uses two independent EEGProject branches, whose outputs are mapped by separate linear projectors to a common $d = 512$ alignment space.

Hierarchical visual targets. Both targets are extracted from one frozen OpenCLIP RN50 image encoder (Ilharco et al., 2021; Radford et al., 2021; He et al., 2016):

$$v_{IH} = f_{IH}(x_I), \qquad v_{IL} = f_{IL}(x_I) = [\mathrm{GAP}(f_{\mathrm{layer2}}(x_I)); \mathrm{GAP}(f_{\mathrm{layer3}}(x_I))]. \tag{2}$$

Here f_{IH} returns RN50's 1024-dimensional image embedding; f_{IL} concatenates pooled layer-2 and layer-3 features ($512 + 1024 = 1536$ dimensions).

Augmentation-pooled alignment. For each level $* \in \{H, L\}$, training and retrieval use features averaged over level-specific image transformations:

$$v_{I*}^{\text{aug}} = \frac{1}{|\mathcal{A}_*|} \sum_{a \in \mathcal{A}_*} f_{I*}(a(x_I)) \,. \tag{3}$$

This feature pooling encourages each branch to retain information that is stable within its level: appearance-degrading transformations are used for the semantic branch, while edge- and texture-preserving transformations are used for the structural branch. The exact transformations and test-time pooling protocol are given in Appendices A.9 and A.2.

Independent affine projectors define $z_{E*} = P_{E*}(f_{E*}(x_E))$ and $z_{I*} = P_{I*}(v_{I*}^{\text{aug}})$ in 512 dimensions, where f_{E*} is the corresponding encoder in Eq. 1. Only image features are normalized: $\bar{z}_{I*} = z_{I*}/\|z_{I*}\|_2$; EEG features remain unnormalized. Each branch has its own parameters and optimizer and is trained with symmetric InfoNCE (Oord et al., 2018; Chen et al., 2020); the exact objective and fixed logit scale are given in Appendix A.2. Retrieval sums the branch scores:

$$s(x_E, x_I) = z_{EH}^{\mathsf{T}} \bar{z}_{IH} + z_{EL}^{\mathsf{T}} \bar{z}_{IL}. \tag{4}$$

EEG-CAM diagnostic. EEG-CAM visualizes branch-specific image evidence using Grad-CAM (Selvaraju et al., 2017) on the image-normalized score $S_*(I) = z_{E*}^{\mathsf{T}} \bar{z}_{I*}(I)$. It is used only for visualization; the image feature path and map normalization are detailed in Appendix A.10.

3.3 Decoder: Hierarchy-Consistent Injection

The semantic and structural subspaces produced by Complementary Dual-Subspace Alignment are mapped through separate linear bridges to different levels of a VAR decoder (Tian et al., 2024). The high-level (semantic) condition seeds the coarsest scale through the start (SOS) token, while the low-level (structural) condition modulates every VAR block through AdaLN (Peebles & Xie, 2023), guiding structure and local detail during refinement. This hierarchy-consistent routing preserves their distinct roles rather than collapsing them into an undifferentiated condition.

Stage 2 is trained with the standard next-scale token cross-entropy of VAR. The visual encoder and VQ tokenizer (Oord et al., 2017) remain frozen, while the final reconstruction model uses the Stage-1-trained EEGProject branches and fine-tunes their parameters with a $0.1\times$ learning-rate scale. Classifier-free guidance (Ho & Salimans, 2022) is used at inference. Detailed VAR dimensions, conditioning schedules, and optimization settings are reported in Appendix A.2.

3.4 Data and Training Protocol

We evaluate on THINGS-EEG (Gifford et al., 2022) using the standard concept split. Each model receives a 17×250 EEG segment, and the two alignment branches are trained independently before the VAR decoder is fine-tuned. Stage 2 uses one run per configuration. Dataset preprocessing, optimization, and evaluation details are reported in Appendix A.1 and Appendix A.2.

4 Results

We evaluate the complete system, then ask what makes an EEG encoder effective, whether the learned subspaces are complementary, and whether they should match the decoder hierarchy. Unless stated otherwise, retrieval is 200-way intra-subject retrieval on THINGS-EEG, and reconstruction and mechanism analyses use Subject 8, following prior EEG-to-image work that reports its main reconstruction comparison on this subject (Li et al., 2024; Dai et al., 2026). Per-subject results for all ten subjects are given in Appendix A.4; full settings and additional results are in the appendix.

4.1 Overall Performance and Component Ablations

For reconstruction, Table 1 summarizes Subject 8 and Figure 3 shows all ten subjects. TRACE leads every metric on Subject 8 and, against the best complete prior result per metric, leads PixCorr, AlexNet(2), AlexNet(5), Inception, and SwAV for all ten subjects; SSIM is mixed but slightly higher

Table 1: **Reconstruction against the best prior result per metric.** Each "Best prior" entry is the best value on that metric from the two strongest prior methods, Brain-HIVE (Zheng et al., 2026) and AVDE (Dai et al., 2026). Higher is better except SwAV (↓). All entries are evaluated on Subject 8.

	PixCorr	SSIM	Alex(2)	Alex(5)	Inc.	CLIP	SwAV↓
Best prior	.2190	.3960	.8760	.9260	.7880	.8270	.5310
TRACE	**.2907**	**.3994**	**.9430**	**.9697**	**.8636**	**.8426**	**.4625**

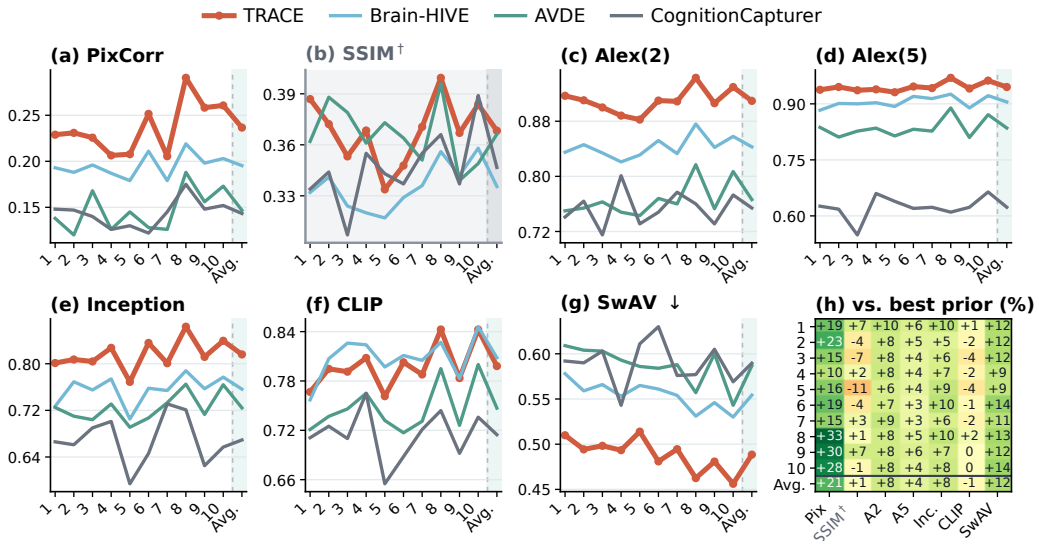

Figure 3: **Per-subject reconstruction on THINGS-EEG.** (a–g) TRACE against Brain-HIVE (Zheng et al., 2026), AVDE (Dai et al., 2026), and CognitionCapturer (Zhang et al., 2025); the †SSIM panel (b) is shaded because SSIM can increase despite degraded reconstruction under temporal shifts (Appendix A.12). (h) TRACE's relative difference (%) per subject and metric from the best prior result among methods reporting all ten subjects.

on average, while CLIP improves on three subjects but is slightly lower on average. Qualitative AVDE comparisons appear in Fig. 1e and Appendix A.5. **For retrieval,** TRACE reaches 77.5% Top-1 and 96.2% Top-5 averaged over ten subjects and five seeds, the strongest aggregate result; per-subject values are given in Appendix Table 6.

The module-wise ablation in Table 3 provides the starting point for our analysis. The "w/o EEGProject" row replaces EEGProject with the pretrained LaBraM encoder used by AVDE while retaining the remaining TRACE pipeline. The replacement causes by far the largest degradation in both retrieval and reconstruction. Using two high-level branches also lowers every reported metric, supporting complementary visual targets. Removing image augmentation pooling or replacing VAR with the Brain-HIVE diffusion decoder and injection design likewise reduces reconstruction quality. These component-level results motivate the following analyses of temporal encoding, complementary alignment, and decoder routing.

4.2 What Makes an EEG Encoder Effective?

A counterintuitive encoder ranking. We first replace only the encoder while holding the other components fixed. Table 4 compares four representative families. The minimal MLP-based EEGProject gives the strongest overall result across the reported retrieval and reconstruction metrics. It outperforms the convolutional TSConv, attention-based ATM, and pretrained LaBraM encoders on every reconstruction metric, while ATM is slightly higher on Top-5 retrieval. LaBraM's low-level branch is particularly weak, and combining the two branches does not improve Top-1/Top-5 retrieval

Table 2: **Ten-subject average intra-subject retrieval on THINGS-EEG.** Values are 200-way Top-1/Top-5 accuracy (%); TRACE is averaged over five random seeds and baseline results are as reported in the respective papers. Per-subject values are provided in Appendix Table 6.

	ATM (Li et al., 2024)	AVDE (Dai et al., 2026)	Region-Aware (Zhu et al., 2026)	NeuroBridge (Zhang et al., 2026)	EEGiT (Zhou et al., 2026)	Brain-HIVE (Zheng et al., 2026)	**TRACE**
Top-1	29.6	30.0	58.2	63.2	70.4	<u>75.7</u>	**77.5**
Top-5	60.3	58.2	85.2	89.9	<u>95.1</u>	94.9	**96.2**

Table 3: **Component ablation.** Each row removes or replaces one component while the remaining pipeline is fixed. "w/o EEGProject" uses the pretrained LaBraM encoder from AVDE in place of EEGProject. "w/o multi-level" retains two independent EEG branches but aligns both to high-level visual targets using different augmentation sets. "w/o VAR" follows Brain-HIVE's decoder and injection design. Reconstruction is evaluated at epoch 150; retrieval is 200-way Top-1/Top-5 where applicable.

Configuration	PixCorr	SSIM	Alex(2)	Alex(5)	Inc.	CLIP	SwAV↓	Top-1	Top-5
Full TRACE	**.2907**	**.3994**	**.9430**	**.9697**	**.8636**	**.8426**	**.4625**	**84.20**	**98.40**
w/o EEGProject (LaBraM)	.1693	.3461	.7905	.8311	.7034	.6814	.5738	23.50	57.50
w/o multi-level (two high branches)	.2722	.3979	.9399	.9671	.8567	.8305	.4677	72.00	93.50
w/o image augmentation pooling	.2788	.3976	.9251	.9504	.8385	.8259	.4707	76.00	96.50
w/o VAR (Brain-HIVE/diffusion)	.2436	.3462	.9015	.9426	.8448	.8076	.5004	–	–

accuracy over its high-level branch. The result is counterintuitive, but it does not yet explain why the simpler model wins.

Table 4: **Controlled encoder comparison.** Retrieval reports 200-way Top-1 and Top-5 accuracy (%); all retrieval entries are means over the same five fixed seeds. Reconstruction uses the same downstream VAR protocol at epoch 150. Higher is better except SwAV (↓).

Encoder	Top-1	Top-5	PixCorr	SSIM	Alex(2)	Alex(5)	Inc.	CLIP	SwAV↓
EEGProject (Wu et al., 2025)	**84.20**	98.40	**.2907**	**.3994**	**.9430**	**.9697**	**.8636**	**.8426**	**.4625**
ATM (Li et al., 2024)	81.20	**98.80**	.2625	.3959	.9252	.9553	.8400	.8251	.4796
TSConv (Song et al., 2024)	79.30	96.20	.2275	.3754	.9077	.9527	.8218	.8090	.4874
LaBraM (Jiang et al., 2024)	23.50	57.50	.1693	.3461	.7905	.8311	.7034	.6814	.5738

Capacity sweep. Across the tested widths (64–4096; 0.72M–78.89M parameters), retrieval accuracy shows no monotonic improvement (Fig. 4a). This exploratory sweep does not identify an optimal width; its configuration is detailed in Appendix A.6.

The task contains rapidly evolving temporal evidence. We next examine the information structure independently of a particular encoder. In Figure 4b, we independently train a Stage-1 EEG–image alignment model at each temporal location and evaluate it at every other location on a 16-ms grid. The model uses a single EEG sample at each location. The diagonal concentration is the signature of a succession of changing neural codes in the temporal-generalization framework (King & Dehaene, 2014), and is consistent with rapid representational dynamics reported for visual recognition (Cichy et al., 2014). This creates a task-level demand for an encoder that can distinguish nearby temporal variations. The signed EEGProject occlusion curve in Figure 4b provides model-level support: short-window masking produces localized, non-uniform changes in the matched score; occlusion is defined as the difference between the original and masked matched scores. We do not interpret its negative values as polarity reversal, nor use this analysis to infer a sub-16-ms neural timescale; the protocol is in Appendix A.7.

Bidirectional interventions link performance to effective temporal granularity. Figure 4c tests the explanation in both directions. For EEGProject, we introduce two TSConv-inspired forms of temporal coarsening. Sharing first-layer weights within 40-ms blocks forces samples in each block to contribute through an aggregate response, simulating non-overlapping temporal aggregation without altering the stored input. A second control applies 204-ms overlapping input smoothing while retaining a dense 4-ms grid. Across five seeds, either change lowers mean Top-1 by about 12 points, and both degrade all reconstruction metrics. Conversely, reducing TSConv's pooling window limits

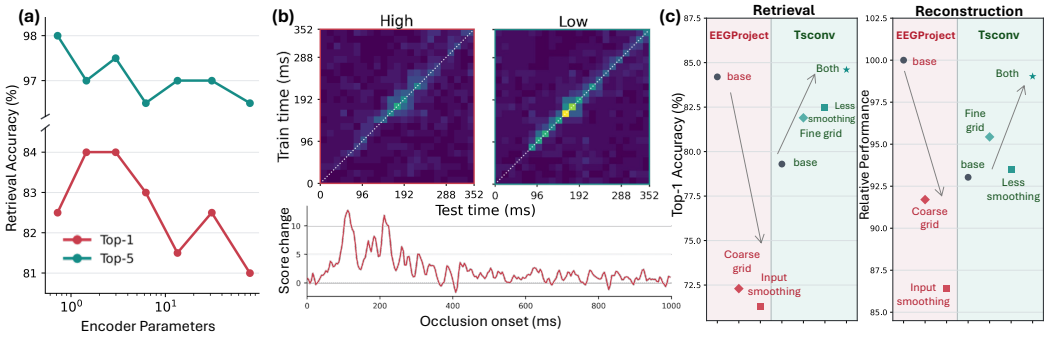

Figure 4: **Controlled tests of fine-grained temporal encoding.** (a) EEGProject capacity sweep shows no monotonic improvement in Top-1/Top-5 retrieval across tested widths. (b) High- and low-level temporal-generalization matrices (top) and the signed occlusion curve (bottom) characterize time-dependent EEG evidence. (c) Retrieval Top-1 (left) and the relative reconstruction index (right) compare EEGProject and TSConv under different smoothing, weight-sharing, pooling-window, and stride settings.

temporal aggregation, while reducing its stride exposes a denser set of temporal positions to the downstream readout. Combining both changes raises mean Top-1 from 79.3% to 84.6%, reaching a similar level to the five-seed EEGProject reference while improving every reconstruction metric over the original TSConv. Together, these interventions support effective temporal granularity as a performance-relevant property across both encoder families. To confirm consistency across subjects, we repeat these experiments on Subjects 5 and 9 and reach the same conclusions (Appendix A.8).

Temporal alignment versus architectural shift sensitivity. Test-time temporal shifts degrade EEGProject's reconstruction performance, but its position-specific weights make shift sensitivity alone insufficient to establish the importance of fine temporal alignment. We therefore train the same architecture with random temporal jitter. Wider jitter windows improve tolerance to test-time shifts while reducing five-seed mean zero-shift retrieval accuracy, with larger costs emerging at jitter half-widths of 20 ms and above. These results support the relevance of temporal alignment at the tens-of-milliseconds scale. Detailed reconstruction stress tests and jitter-training controls are provided in Appendices A.12 and A.13, respectively.

ATM shares EEGProject's readout head. ATM retains EEGProject's projection-head form without temporal pooling, preserving its fine-grained readout and a similar early-time attribution profile. This shared readout explains why its retrieval and reconstruction performance stays close to EEGProject. The added iTransformer and convolutional patch embedding target cross-channel structure, but this inductive bias is mismatched to MVNN-normalized inputs: the learned channel attention remains nearly uniform. Stacking these components in front of the shared projection head therefore brings no performance gain over EEGProject; detailed attention diagnostics and the MVNN interpretation are provided in Appendix A.6.

4.3 Do the Two Subspaces Provide Complementary Evidence?

We test whether the two branches of Complementary Dual-Subspace Alignment provide complementary evidence. Fusion raises five-seed mean Top-1 retrieval accuracy to 84.20%, compared with 71.20% for the high-level branch and 77.60% for the low-level branch (Appendix Table 20). It also outperforms two independently trained high-level branches (Table 3), supporting the benefit of complementary high/low targets over a same-level ensemble.

The retrieval examples and EEG-CAM maps in Figure 5 illustrate complementary semantic and structural emphasis. The cross-method CAM comparison in Appendix A.10 illustrates more distinct branch-specific emphasis, while further augmentation controls are reported in Appendix A.9.

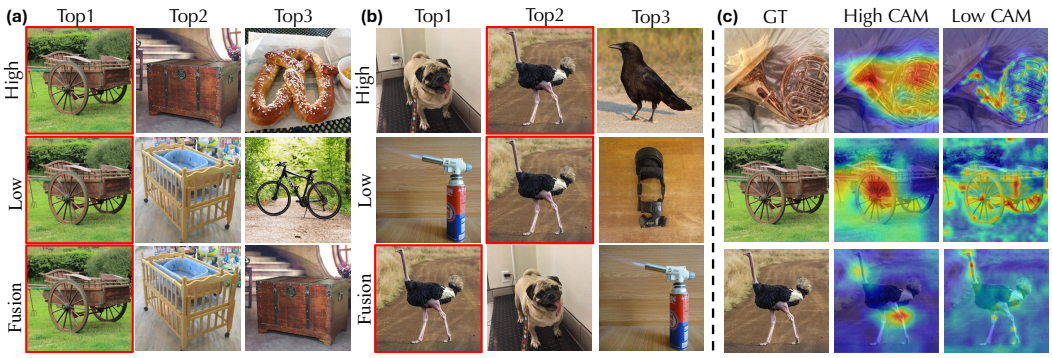

Figure 5: **Complementary level-specific evidence from Complementary Dual-Subspace Alignment.** (a) Both branches retrieve the ground-truth image at Top-1, but rank different subsequent candidates: the high-level branch favors semantically related images, whereas the low-level branch favors images sharing local structure such as fences and wheels. (b) The high-level branch favors semantically related animals, while the low-level branch follows the target's elongated, curved structure; combining these complementary cues recovers the ground truth at Top-1. (c) EEG-CAM visualizations show distinct image-space emphasis for the two branches.

Table 5: **Decoder injection consistency.** Reconstruction metrics for the proposed hierarchy-consistent routing and three controls: single-embedding, swapped, and undifferentiated conditioning.

Routing	PixCorr	SSIM	Alex(2)	Alex(5)	Inc.	CLIP	SwAV↓
High→SOS, Low→AdaLN	**.2907**	.3994	**.9430**	**.9697**	**.8636**	**.8426**	**.4625**
High→SOS+AdaLN	.2673	.3912	.9271	.9655	.8448	.8054	.4693
Swapped routes	.2658	.3967	.9353	.9673	.8418	.8194	.4634
High+Low→both	.2552	**.4001**	.9370	.9661	.8463	.8306	.4635

4.4 Does Hierarchy-Consistent Injection Matter?

Prior decoders often collapse multi-level features into one condition. We test whether hierarchy-matched routing improves reconstruction. Table 5 compares the proposed assignment with single-embedding, swapped-routing, and undifferentiated-fusion controls; the first two rows directly compare complementary dual conditions with one high-level embedding reused at both routes.

Proposed *Hierarchy-Consistent Injection* gives the strongest overall reconstruction result and leads on six of seven metrics; the "High+Low→both" control is slightly higher only on SSIM. This control mirrors the undifferentiated fusion used by prior multi-level methods and yields weaker overall performance. Reusing one high-level embedding or swapping the routes also underperforms the complementary assignment. These results support preserving the distinct roles of the two conditions through hierarchy-matched routing. Training ablations are reported in Appendix A.11.

5 Conclusion

In this work, we conduct a systematic rethinking of visual EEG-to-image decoding as an analysis-to-design problem. By revisiting the encoder, alignment, and decoder stages separately, we find that fine-grained temporal encoding, rather than model size, is central to the encoder comparison. Controlled interventions show that performance degrades when aggregation, smoothing, or coarse output sampling reduces effective temporal granularity. Guided by this diagnosis, we present TRACE, which aligns semantic and structural EEG evidence in complementary subspaces and routes them to matching stages of a VAR decoder. Across ten subjects, TRACE achieves the best average retrieval and strongest overall reconstruction in our reported comparisons.

REFERENCES

Yunpeng Bai, Xintao Wang, Yan-pei Cao, Yixiao Ge, Chun Yuan, and Ying Shan. Dreamdiffusion: Generating high-quality images from brain eeg signals. *arXiv preprint arXiv:2306.16934*, 2023.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Radoslaw Martin Cichy, Dimitrios Pantazis, and Aude Oliva. Resolving human object recognition in space and time. *Nature neuroscience*, 17(3):455–462, 2014.

Sicheng Dai, Hongwang Xiao, Shan Yu, and Qiwei Ye. Autoregressive visual decoding from eeg signals. *arXiv preprint arXiv:2602.22555*, 2026.

James J DiCarlo, Davide Zoccolan, and Nicole C Rust. How does the brain solve visual object recognition? *Neuron*, 73(3):415–434, 2012.

Honghao Fu, Hao Wang, Jing Jih Chin, and Zhiqi Shen. Brainvis: Exploring the bridge between brain and visual signals via image reconstruction. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

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

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Gabriel Ilharco, Mitchell Wortsman, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, et al. Openclip. *Zenodo*, 2021.

Wei-Bang Jiang, Liming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous eeg data in bci. In *International Conference on Learning Representations*, volume 2024, pp. 16405–16426, 2024.

Jean-Rémi King and Stanislas Dehaene. Characterizing the dynamics of mental representations: the temporal generalization method. *Trends in cognitive sciences*, 18(4):203–210, 2014.

Yu-Ting Lan, Kan Ren, Yansen Wang, Wei-Long Zheng, Dongsheng Li, Bao-Liang Lu, and Lili Qiu. Seeing through the brain: image reconstruction of visual perception from human brain signals. *arXiv preprint arXiv:2308.02510*, 2023.

Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of neural engineering*, 15(5):056013, 2018.

Dongyang Li, Chen Wei, Shiying Li, Jiachen Zou, Haoyang Qin, and Quanying Liu. Visual decoding and reconstruction via eeg embeddings with guided diffusion. *arXiv preprint arXiv:2403.07721*, 2024.

Minxu Liu, Donghai Guan, Chuhang Zheng, Chunwei Tian, Jie Wen, and Qi Zhu. Vieeg: Hierarchical visual neural representation for eeg brain decoding. *arXiv preprint arXiv:2505.12408*, 2025.

Eleonora Lopez, Luigi Sigillo, Federica Colonnese, Massimo Panella, and Danilo Comminiello. Guess what i think: Streamlined eeg-to-image generation with latent diffusion models. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Yizhuo Lu, Changde Du, Qiongyi Zhou, Dianpeng Wang, and Huiguang He. Minddiffuser: Controlled image reconstruction from human brain activity with semantic and structural diffusion. In *Proceedings of the 31st ACM international conference on multimedia*, pp. 5899–5908, 2023.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

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

Furkan Ozcelik and Rufin VanRullen. Natural scene reconstruction from fmri signals using generative latent diffusion. *Scientific Reports*, 13(1):15666, 2023.

Juhyeon Park, Peter Yongho Kim, Jiook Cha, Shinjae Yoo, and Taesup Moon. SEED: Towards more accurate semantic evaluation for visual brain decoding. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2503.06437.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021.

Paul Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Aidan Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth Norman, et al. Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors. *Advances in Neural Information Processing Systems*, 36:24705–24728, 2023.

Paul S Scotti, Mihir Tripathy, Cesar Kadir Torrico Villanueva, Reese Kneeland, Tong Chen, Ashutosh Narang, Charan Santhirasegaran, Jonathan Xu, Thomas Naselaris, Kenneth A Norman, et al. Mindeye2: Shared-subject models enable fmri-to-image with 1 hour of data. *arXiv preprint arXiv:2403.11207*, 2024.

Ramprasaath R Selvaraju, Abhishek Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626, 2017.

Yonghao Song, Bingchuan Liu, Xiang Li, Nanlin Shi, Yijun Wang, and Xiaorong Gao. Decoding natural images from eeg for object recognition. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 47648–47665, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/d0865cbe51d35ec322f9af9db7806fc7-Paper-Conference.pdf.

Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14453–14463, 2023.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37:84839–84865, 2024.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004.

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

Dian Xie, Peiang Zhao, Jiarui Zhang, Kangqi Wei, Xiaobao Ni, and Jiong Xia. Brainram: Cross-modality retrieval-augmented image reconstruction from human brain activity. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, pp. 3994–4003, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400706868.

Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23):8619–8624, 2014.

Chenglong Yu, Shuai Shen, Xiangsheng Li, and Yang 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 (CVPR)*, pp. 26698–26710, June 2026.

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

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

Jiawen Zheng, Haonan Jia, Ming Li, Yuhui Zheng, Yufeng Zeng, Yang Gao, and Chen Liang. Learning brain representation with hierarchical visual embeddings. *arXiv preprint arXiv:2602.07495*, 2026.

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

Shuqi Zhu, Ziyi Ye, Yi Zhong, Qingyao Ai, Yujia Zhou, and Yiqun Liu. Brain image reconstruction with retrieval-augmented diffusion. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, pp. 296–305, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400715921.

Yanan Zhu, Ziwei Xiang, Jiamin Wu, Jinyang Guo, Hongyuan Zhang, Chunfeng Song, Hongjian Fang, Yufei Guo, and Xianglong Liu. Region-aware hierarchical sub-feature alignment for robust eeg-based visual decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 6667–6676, June 2026.

A Additional Experimental Details and Results

A.1 Dataset and EEG–Image Preprocessing

We use the publicly released THINGS-EEG EEG–image pairs under the standard subject-wise concept split: 1,654 training concepts and 200 held-out test concepts for each of ten subjects. Each training image is presented four times and each test image 80 times; repetitions of the same image are averaged before training and evaluation, yielding 16,540 training pairs and 200 test pairs per subject. Following the NeuroBridge preprocessing protocol, EEG is band-pass filtered to 0.1–100 Hz, baseline-corrected with the preceding 200 ms, segmented from 0 to 1,000 ms after stimulus onset, and downsampled from 1,000 to 250 Hz (Zhang et al., 2026). We use MVNN normalization and the 17 posterior channels P7, P5, P3, P1, Pz, P2, P4, P6, P8, PO7, PO3, POz, PO4, PO8, O1, Oz, and O2. Thus, each model input is a 17×250 EEG segment paired with its corresponding ground-truth image.

A.2 Implementation Details

EEGProject maps the flattened $17 \times 250 = 4250$-dimensional input to 1024 dimensions for the high-level branch and 1536 dimensions for the low-level branch; both branches are then projected to the shared 512-dimensional alignment space. The VAR decoder uses width 1024, depth 16, and 16 attention heads. Its two condition bridges map to 1024 dimensions, with the high-level condition injected at SOS and the low-level condition injected through AdaLN. The frozen VQVAE tokenizer uses a downsampling factor of 16, and the VAR decoder generates 256×256 RGB reconstructions. Stage 1 uses two independent AdamW optimizers (one per branch), learning rate 10^{-4}, betas $(0.9, 0.999)$, weight decay 10^{-4}, batch size 1024, and the fixed logit scale defined below. The high- and low-level checkpoints are fixed at epochs 60 and 100, respectively. Stage 2 uses AdamW with betas $(0.9, 0.95)$, learning rate 2×10^{-4}, weight decay 0.05, batch size 32, and linear warmup over 20 epochs followed by cosine decay; the Stage-1 encoders use a $0.1\times$ learning-rate scale. We report the epoch-150 reconstruction checkpoint. Classifier-free condition dropout is 0.1, maximum inference guidance scale is 1.5, and VAR token loss uses label smoothing 0.1. Image-feature pooling averages four high-level augmentations (Gaussian blur, Gaussian noise, low resolution, and mosaic) or four low-level augmentations (Gabor maximum response, gradient X, LBP texture, and patch shuffling).

Stage-1 alignment and retrieval. Stage-1 EEG augmentation applies a five-sample moving average independently to each channel with probability 0.3, once after trial averaging at dataset initialization; boundary windows are truncated. No EEG augmentation is applied at test time or in Stage 2.

The four projectors P_{E*} and P_{I*} are independent affine layers with bias: both high-level projectors map $1024 \to 512$, and both low-level projectors map $1536 \to 512$. They contain no activation, residual block, or normalization. After projection, only image features are L2-normalized; EEG features remain unnormalized in both training and retrieval. For a batch of B paired samples, define the branch logits $Q_{*,ij} = \gamma z_{E*,i}^{\top} \bar{z}_{I*,j}$. The symmetric objective is

$$\mathcal{L}_* = -\frac{1}{2B} \sum_{i=1}^{B} \left[\log \frac{\exp Q_{*,ii}}{\sum_{j=1}^{B} \exp Q_{*,ij}} + \log \frac{\exp Q_{*,ii}}{\sum_{j=1}^{B} \exp Q_{*,ji}} \right], \qquad * \in \{H, L\}. \tag{5}$$

The diagonal entries are paired positives and the off-diagonal entries are in-batch negatives. Both branches use the same fixed multiplier $\gamma = \mathrm{softplus}(\log(1/0.07)) \approx 2.727$, equivalent to a temperature of $1/\gamma \approx 0.367$. Thus, 0.07 is the initialization argument, not the effective temperature; the logit-scale parameter is not learned. Each branch optimizes its own objective with a separate optimizer. There is no auxiliary MSE or text-alignment loss.

Training and test-time retrieval both use the augmentation-pooled image features in Eq. 3: four transformations per branch are averaged with equal weights, without an identity (untransformed-image) term. The transformation parameters and cached random draws are described in Appendix A.9. Test evaluation enables `--image_test_aug`; omitting this non-default flag selects the untransformed image features instead. Retrieval uses the direct score sum in Eq. 4, without additional branch cali-

bration. Since EEG features are unnormalized, their norms implicitly affect the relative contributions of the two branches to the fused score.

Reconstruction metrics. Following prior EEG-to-image studies (Li et al., 2024; Dai et al., 2026; Zheng et al., 2026), we evaluate the similarity between each reconstructed image and its corresponding ground-truth stimulus using seven complementary metrics. PixCorr measures pixelwise correlation, while SSIM measures structural similarity (Wang et al., 2004); these serve as low-level fidelity measures. AlexNet(2), AlexNet(5), InceptionV3, and CLIP measure feature-level or semantic similarity using pretrained visual networks. Following the two-way identification protocol used in prior work, these feature-based scores are computed as bidirectional retrieval scores between reconstructed and ground-truth images. SwAV is reported as the average correlation distance between SwAV-ResNet50 features (Caron et al., 2020), for which lower values indicate better agreement. Thus, higher values are better for PixCorr, SSIM, AlexNet(2), AlexNet(5), InceptionV3, and CLIP, whereas lower values are better for SwAV. These reconstruction metrics are distinct from 200-way retrieval accuracy: reconstruction compares each generated image with its paired ground-truth stimulus, while retrieval evaluates whether the correct concept is ranked among a 200-image candidate set.

Classifier-free guidance. The high- and low-level conditions are independently zeroed with probability 0.1 during training. At inference, the conditional and all-zero unconditional predictions are combined at scale i as

$$\mathbf{u}_i^{\mathrm{CFG}} = (1 + t_i)\mathbf{u}_i^{\mathrm{cond}} - t_i\mathbf{u}_i^{\mathrm{uncond}}, \qquad t_i = 1.5\,\frac{i}{K-1}, \tag{6}$$

where $\mathbf{u}_i$ denotes token logits, K is the number of VAR scales, and $i = 0, \ldots, K-1$. Guidance therefore increases linearly from 0 at the coarsest scale to 1.5 at the final scale.

Fixed Stage-1 checkpoints. The dataset provides training and test concepts but no validation split. We therefore fixed the Stage-1 training budgets from the training-loss curves rather than selecting epochs using test performance. The high-level and low-level branches use fixed training budgets of 60 and 100 epochs, respectively, by which point their training losses have largely plateaued. These budgets are fixed across subjects and random seeds. Epochs were fixed based on the convergence of the training losses; test performance was not used for checkpoint selection to avoid information leakage. The curves used to set the budgets are shown in Fig. 6.

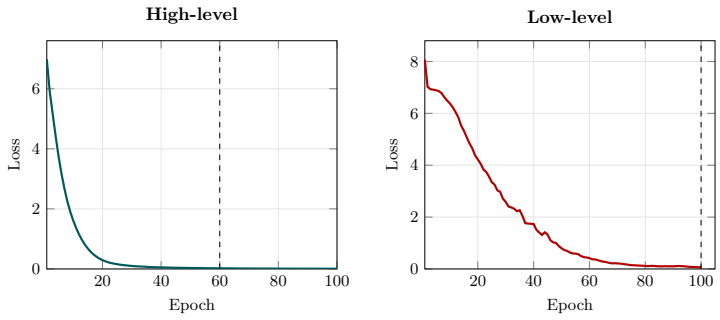

Figure 6: **Stage-1 training-loss curves.** The vertical markers denote the fixed checkpoints used for the high-level (60 epochs) and low-level (100 epochs) branches.

A.3 Complete Retrieval Results

A.4 All-Subject Reconstruction Results

Tables 9–13 report the per-subject reconstruction metrics for TRACE and four baseline methods. Brain-HIVE (Zheng et al., 2026) results are taken from its Table 13 (H14+B32+VAE configuration). AVDE (Dai et al., 2026) values are from its Table 13. CognitionCapturer values are from Zhang et

Table 6: **Per-subject intra-subject retrieval on THINGS-EEG.** Results are 200-way Top-1/Top-5 accuracy (%). Each subject is evaluated independently; TRACE entries are averaged over five random seeds and baseline results are as reported in the respective papers. Best per column is shown in **bold** and second best is underlined.

Method		Sub 1	Sub 2	Sub 3	Sub 4	Sub 5	Sub 6	Sub 7	Sub 8	Sub 9	Sub 10	Avg.
ATM	Top-1	24.0	27.0	34.0	28.0	23.5	26.0	28.0	39.5	30.0	35.5	29.6
(Li et al., 2024)	Top-5	57.0	55.0	59.5	61.5	51.0	60.0	63.0	70.5	55.5	70.0	60.3
AVDE	Top-1	25.0	24.1	27.5	29.8	25.4	33.5	27.4	41.7	26.1	39.5	30.0
(Dai et al., 2026)	Top-5	55.2	51.0	58.6	54.7	50.3	60.3	55.2	71.3	52.1	73.0	58.2
Region-Aware	Top-1	61.0	56.0	56.5	57.0	46.5	56.5	58.5	75.0	50.0	65.0	58.2
(Zhu et al., 2026)	Top-5	87.0	81.5	87.5	84.5	74.5	87.0	86.0	91.5	80.0	92.0	85.2
NeuroBridge	Top-1	50.0	63.2	61.6	61.4	54.8	69.7	62.7	71.2	64.0	73.6	63.2
(Zhang et al., 2026)	Top-5	77.6	90.6	91.1	90.0	85.0	92.9	88.8	95.1	91.0	97.1	89.9
EEGiT	Top-1	**82.3**	73.7	67.5	64.9	55.9	66.8	70.2	**84.8**	51.7	86.5	70.4
(Zhou et al., 2026)	Top-5	**97.6**	97.9	95.7	93.5	91.3	92.0	**98.1**	**98.6**	87.3	**99.4**	95.1
Brain-HIVE	Top-1	64.3	76.3	**74.0**	67.0	**68.0**	81.5	76.8	**84.8**	**76.8**	87.3	75.7
(Zheng et al., 2026)	Top-5	88.8	95.3	95.0	91.8	**91.5**	96.3	96.8	98.5	95.8	99.3	94.9
TRACE	Top-1	73.6	**77.0**	73.0	**74.6**	65.9	**84.2**	**77.9**	84.2	75.0	**89.2**	**77.5**
	Top-5	93.5	**98.4**	**96.8**	**95.5**	90.3	**97.9**	95.9	98.4	**96.7**	98.8	**96.2**

al. (2025) Table S1. All baselines report results on the same THINGS-EEG ten-subject split used by TRACE.

For ATM (Li et al., 2024), the Subject-8 comparison in Table 7 uses the results reported in Table 1 of the original paper. The per-subject results in Table 12 are taken from Table 14(a) of AVDE (Dai et al., 2026), and the ATM averages in Table 8 are computed from those ten entries. The Subject-8 values differ between these sources; we retain the values as reported in each source.

Table 7: **Complete Subject-8 reconstruction comparison.** Baseline results are as reported in the respective papers. Higher is better except SwAV (↓).

Method	PixCorr	SSIM	Alex(2)	Alex(5)	Inc.	CLIP	SwAV↓
ATM (Li et al., 2024)	.1600	.3450	.7760	.8660	.7340	.7860	.5820
CognitionCapturer (Zhang et al., 2025)	.1750	.3660	.7600	.6100	.7210	.7440	.5770
ViEEG (Liu et al., 2025)	–	.3560	.7870	.8810	.7480	.7980	.5840
AVDE (Dai et al., 2026)	.1880	.3960	.8170	.8890	.7650	.7950	.5570
Brain-HIVE (Zheng et al., 2026)	.2190	.3560	.8760	.9260	.7880	.8270	.5310
TRACE	**.2907**	**.3994**	**.9430**	**.9697**	**.8636**	**.8426**	**.4625**

Table 8: **Ten-subject average reconstruction comparison.** Values averaged across Sub-01 through Sub-10. ATM averages are computed from the per-subject values reported in AVDE (Dai et al., 2026), Table 14(a). Higher is better except SwAV (↓).

Method	PixCorr	SSIM	Alex(2)	Alex(5)	Inc.	CLIP	SwAV↓
ATM (Li et al., 2024)	.130	.339	.740	.822	.685	.716	.599
CognitionCapturer (Zhang et al., 2025)	.143	.347	.754	.623	.669	.715	.590
AVDE (Dai et al., 2026)	.147	.366	.766	.835	.724	.747	.587
Brain-HIVE (Zheng et al., 2026)	.195	.336	.843	.905	.756	**.808**	.554
TRACE	**.237**	**.368**	**.910**	**.945**	**.816**	.798	**.488**

Table 8 summarizes the ten-subject average for each method. TRACE improves PixCorr, SSIM, AlexNet(2), AlexNet(5), Inception, and SwAV compared to all four baselines. For CLIP, TRACE is below Brain-HIVE (0.798 vs 0.808) but above AVDE (0.747), ATM (0.716), and CognitionCapturer (0.715).

At the per-subject level, TRACE improves PixCorr, SSIM, AlexNet(2), AlexNet(5), Inception, and SwAV for all ten subjects compared to Brain-HIVE; CLIP improves on three subjects. Against AVDE, TRACE improves PixCorr, AlexNet(2), AlexNet(5), Inception, CLIP, and SwAV for all ten subjects, and improves SSIM on six. Compared to ATM and CognitionCapturer, TRACE improves every metric on all subjects except SSIM, where it improves on eight of ten subjects for each baseline.

Table 9: **TRACE** Reconstruction on THINGS-EEG All Subjects.

Subject	PixCorr	SSIM	Alex(2)	Alex(5)	Inc.	CLIP	SwAV↓
Sub-01	.2289	.3869	.9169	.9377	.8012	.7667	.5099
Sub-02	.2309	.3722	.9104	.9452	.8073	.7947	.4943
Sub-03	.2257	.3532	.9000	.9364	.8039	.7913	.4982
Sub-04	.2064	.3684	.8882	.9388	.8273	.8080	.4934
Sub-05	.2077	.3338	.8824	.9311	.7689	.7615	.5139
Sub-06	.2515	.3478	.9100	.9464	.8357	.8027	.4810
Sub-07	.2055	.3706	.9088	.9422	.8009	.7879	.4945
Sub-08	.2907	.3994	.9430	.9697	.8636	.8426	.4625
Sub-09	.2583	.3670	.9062	.9411	.8118	.7835	.4808
Sub-10	.2607	.3838	.9296	.9623	.8394	.8425	.4562

Table 10: **Brain-HIVE** (Zheng et al., 2026) Reconstruction on THINGS-EEG All Subjects.

Subject	PixCorr	SSIM	Alex(2)	Alex(5)	Inc.	CLIP	SwAV↓
Sub-01	.1930	.3320	.8350	.8830	.7270	.7570	.5780
Sub-02	.1880	.3410	.8460	.9010	.7690	.8070	.5590
Sub-03	.1960	.3240	.8340	.9000	.7550	.8260	.5660
Sub-04	.1870	.3200	.8210	.9030	.7740	.8240	.5530
Sub-05	.1790	.3170	.8310	.8930	.7050	.7970	.5650
Sub-06	.2110	.3290	.8520	.9200	.7580	.8110	.5610
Sub-07	.1790	.3360	.8330	.9140	.7540	.8050	.5540
Sub-08	.2190	.3560	.8760	.9260	.7880	.8270	.5310
Sub-09	.1980	.3420	.8420	.8890	.7570	.7830	.5460
Sub-10	.2030	.3580	.8580	.9220	.7770	.8460	.5300

Table 11: **AVDE** (Dai et al., 2026) Reconstruction on THINGS-EEG All Subjects.

Subject	PixCorr	SSIM	Alex(2)	Alex(5)	Inc.	CLIP	SwAV↓
Sub-01	.1380	.3620	.7500	.8370	.7250	.7210	.6090
Sub-02	.1200	.3880	.7540	.8110	.7100	.7370	.6040
Sub-03	.1680	.3790	.7630	.8270	.7040	.7460	.6030
Sub-04	.1260	.3610	.7480	.8350	.7310	.7650	.5930
Sub-05	.1450	.3730	.7430	.8140	.6910	.7320	.5860
Sub-06	.1280	.3640	.7680	.8320	.7070	.7170	.5840
Sub-07	.1260	.3510	.7600	.8270	.7330	.7310	.5880
Sub-08	.1880	.3960	.8170	.8890	.7650	.7950	.5570
Sub-09	.1560	.3390	.7530	.8100	.7130	.7260	.6000
Sub-10	.1730	.3490	.8070	.8710	.7640	.8000	.5430

Table 12: **ATM** (Li et al., 2024) Reconstruction on THINGS-EEG All Subjects. Values are taken from Table 14(a) of AVDE (Dai et al., 2026).

Subject	PixCorr	SSIM	Alex(2)	Alex(5)	Inc.	CLIP	SwAV↓
Sub-01	.1360	.3540	.7200	.8310	.7100	.7280	.5890
Sub-02	.1080	.3010	.6880	.7750	.6570	.7050	.6180
Sub-03	.1450	.3400	.7470	.8580	.7510	.7590	.5900
Sub-04	.0750	.3050	.7290	.7820	.6940	.6880	.6230
Sub-05	.1330	.3520	.7380	.7910	.6370	.7070	.6120
Sub-06	.1400	.3630	.7530	.8340	.6860	.7420	.5950
Sub-07	.1550	.3600	.7690	.8510	.6840	.6970	.6080
Sub-08	.1630	.3450	.7860	.8680	.7300	.7700	.5750
Sub-09	.1340	.3380	.7360	.7880	.6260	.6520	.6060
Sub-10	.1120	.3290	.7350	.8430	.6740	.7080	.5700

Table 13: **CognitionCapturer** (Zhang et al., 2025) Reconstruction on THINGS-EEG All Subjects.

Subject	PixCorr	SSIM	Alex(2)	Alex(5)	Inc.	CLIP	SwAV↓
Sub-01	.1480	.3340	.7410	.6260	.6660	.7110	.5920
Sub-02	.1470	.3440	.7640	.6180	.6610	.7250	.5900
Sub-03	.1400	.3070	.7150	.5490	.6900	.7100	.6030
Sub-04	.1260	.3550	.8010	.6600	.7010	.7650	.5430
Sub-05	.1300	.3430	.7310	.6390	.5940	.6550	.6110
Sub-06	.1220	.3370	.7480	.6200	.6460	.6880	.6300
Sub-07	.1450	.3550	.7770	.6230	.7310	.7210	.5760
Sub-08	.1750	.3660	.7600	.6100	.7210	.7440	.5770
Sub-09	.1480	.3370	.7310	.6230	.6250	.6920	.6050
Sub-10	.1520	.3890	.7730	.6640	.6570	.7360	.5690

Qualitative reconstructions across subjects. Figures 7 and 8 show selected test examples for all ten subjects. Each column pairs a ground-truth image with the corresponding subject-specific reconstructions.

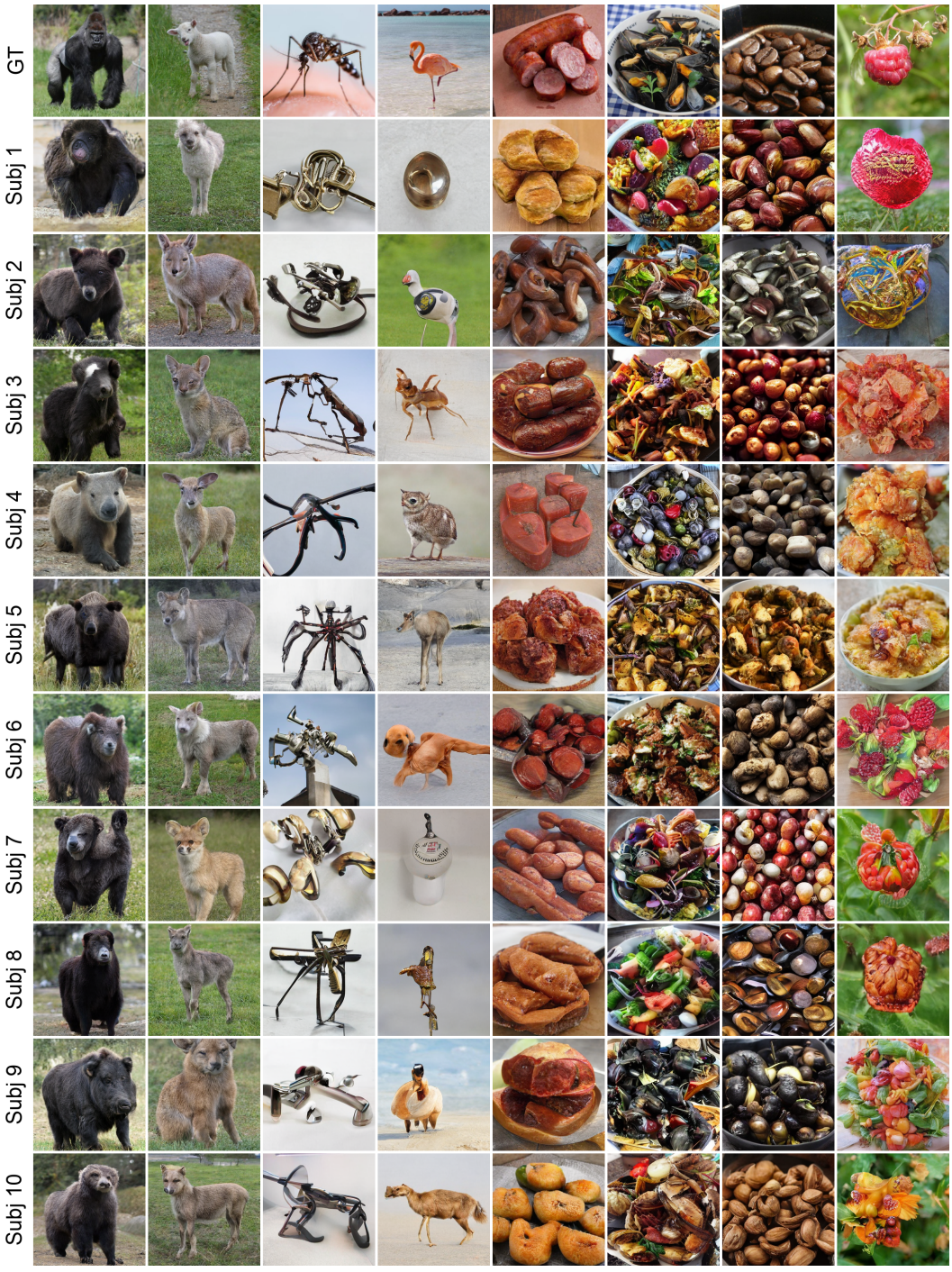

Figure 7: **All-subject reconstruction showcase (1/2).** The top row shows the ground-truth (GT) images; the remaining rows show TRACE reconstructions for Subjects 1–10. Each column corresponds to the same test image across subjects.

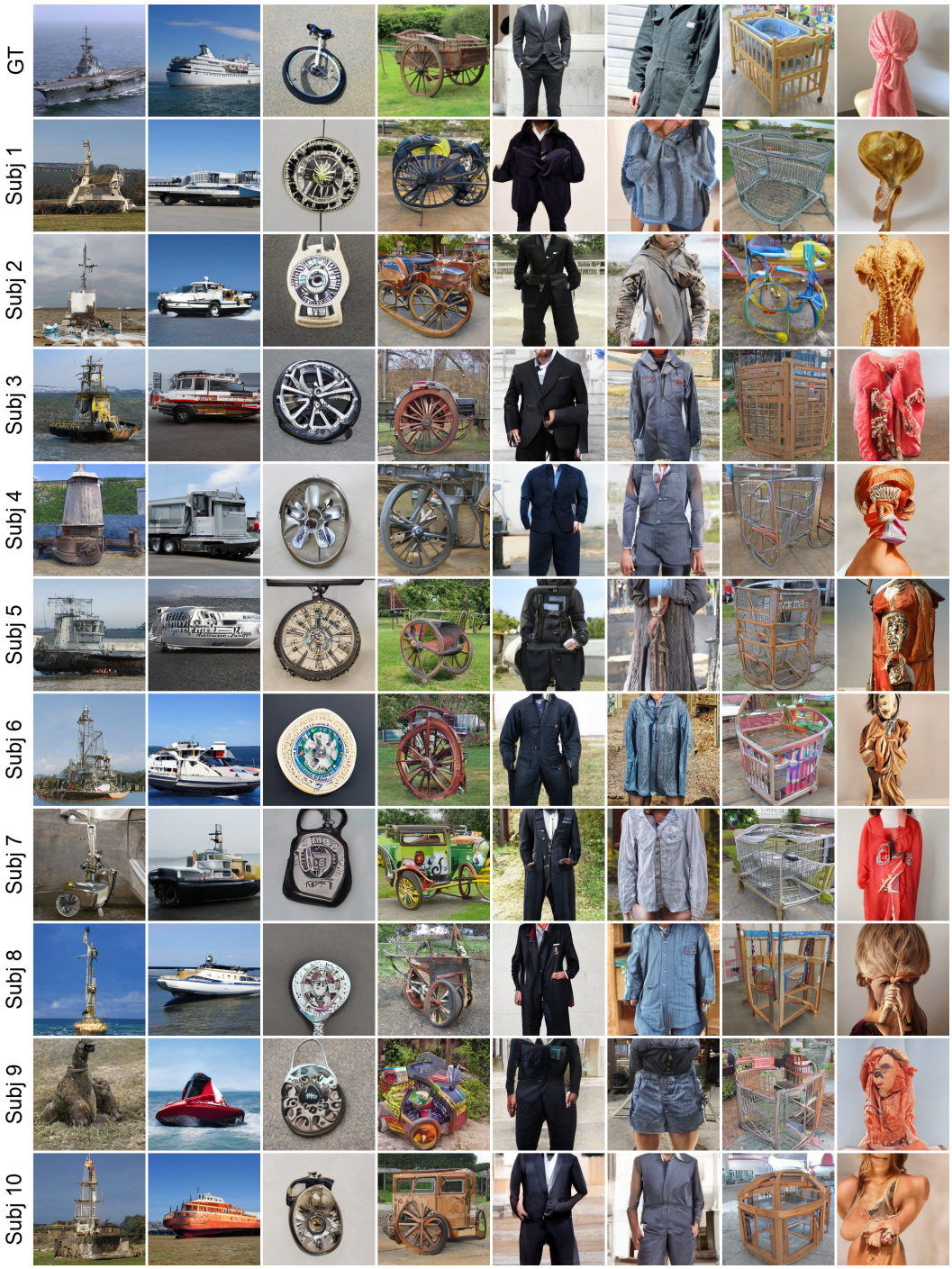

Figure 8: **All-subject reconstruction showcase (2/2).** Eight additional test images are shown in the same layout as Figure 7: ground truth (GT) in the top row, followed by TRACE reconstructions for Subjects 1–10.

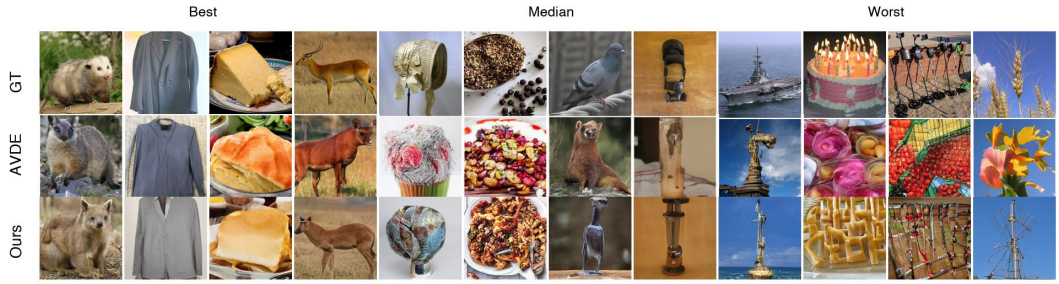

Figure 9: **Cross-paper qualitative comparison on Subject 8.** Columns are grouped by the quality categories reported in AVDE (Best, Median, and Worst); rows show the ground truth (GT), AVDE, and TRACE reconstructions.

A.5 Cross-Paper Qualitative Comparison with AVDE

For an additional visual reference, Figure 9 compares our reconstructions with the Subject-8 examples reported by AVDE. The same ground-truth stimuli are used in each column; the AVDE examples are reproduced from its published Figure 7, with four examples selected from each reported quality group.

A.6 Encoder Comparisons and Capacity Controls

Comparison encoder structures. In Eq. 1, $b_1, b_2 \in \mathbb{R}^D$, and LayerNorm has learned scale and bias. For the deployed widths $D_H = 1024$ and $D_L = 1536$, the two EEGProject encoders contain $\sum_{D\in\{1024,1536\}}(4250D + D^2 + 4D) = 14{,}298{,}112$ trainable parameters. Parameter counts in Table 14 include all trainable parameters of both EEG encoders, including biases and, for EEGProject, LayerNorm, but exclude the alignment projectors. The corresponding Linear count is $\sum_{D\in\{1024,1536\}}(4250D + D) = 10{,}882{,}560$.

The controlled comparison uses the following implementations. Linear flattens the EEG input and applies a single affine projection, without a nonlinearity, normalization, or residual block. All encoder comparisons use the fixed Stage-1 training protocol described in Appendix A.2. TSConv applies temporal convolution followed by average pooling and spatial convolution; its reference configuration uses a pooling window of $W = 51$ samples and stride $S = 5$, corresponding to a 204-ms smoothing window and a 20-ms output grid. ATM places an iTransformer and convolutional patch embedding before a projection head with the same linear–residual–normalization form as EEGProject (dropout 0.5 rather than 0.3). LaBraM uses the pretrained LaBraM Transformer encoder; its input path resamples the EEG to 200 Hz, uses a single patch over the 1-s temporal window, and applies stride-8 temporal convolutions before the Transformer. These structural details are reported here for completeness; the main text focuses on the controlled performance and temporal analyses. All LaBraM results reported in the paper use this pretrained initialization.

ATM matched-head comparison. ATM and EEGProject use the same form of EEG readout head; ATM additionally places an iTransformer and convolutional patch embedding before it, while its readout dropout is 0.5 rather than 0.3. The complete reconstruction comparison appears in Table 4 and uses the fixed Stage-1 checkpoints followed by the common Stage-2 protocol at epoch 150. ATM is lower on every retained reconstruction metric.

The retrieval comparison uses five seeds for each encoder. High-level Top-1 is effectively tied, whereas ATM has lower low-level retrieval and lower overall retrieval accuracy.

For temporal attribution, 25.8%/29.0% of EEGProject's positive high-/low-level contribution and 26.2%/31.0% of ATM's fall in the 60–160 ms interval. ATM therefore preserves a similar broad early-time profile despite its low-level performance loss. Its attention matrix, averaged over 200 examples and four heads, has normalized entropy 0.9843, mean off-diagonal weight 0.0548 (uniform baseline: $1/18 = 0.0556$), and mean per-element L_1 distance 0.0063 from the uniform matrix. Excluding interactions with the subject token, most off-diagonal entries lie between 0.057 and 0.061.

These statistics establish that the learned channel attention is nearly uniform. The connection to MVNN whitening is an architectural interpretation, not a direct causal test.

Linear capacity control. Table 14 compares EEGProject with its purely linear counterpart under the same five-seed Stage-1 protocol. Linear remains close to EEGProject on both branches and overall retrieval accuracy despite removing the residual nonlinear block. This lower-bound control complements the width sweep without establishing that capacity has no effect.

Table 14: **Linear capacity control on Subject 8.** Mean ± standard deviation over five Stage-1 seeds using the fixed branch-specific epochs described in Appendix A.6.

Encoder	Parameters (M)	High Top-1	Low Top-1	Fusion Top-1	Fusion Top-5
Linear	10.88	73.60 ± 2.22	76.60 ± 1.75	82.60 ± 2.43	99.00 ± 0.35
EEGProject	14.30	71.20 ± 2.05	77.60 ± 1.52	84.20 ± 0.91	98.40 ± 0.42

EEGProject width sweep. Table 15 gives the retrieval accuracies summarized in Figure 4a. The sweep varies the hidden width of the EEGProjectWidth variant. Its $D = 1024$ entry has 13.433M parameters and does not represent the deployed 14.30M model in Table 14, whose branch widths are $D_H = 1024$ and $D_L = 1536$ (Eq. 1). Each width is evaluated in one run with seed 2025; these entries are not five-seed means. Increasing hidden width over the 110-fold parameter range does not yield a monotonic improvement in either Top-1 or Top-5. This exploratory sweep does not establish an optimal width or a statistically reliable ordering among widths.

Table 15: **EEGProject capacity sweep on Subject 8.** One run per width (seed 2025); retrieval reports 200-way Top-1/Top-5 accuracy (%).

Hidden width	Parameters (M)	Top-1	Top-5
64	0.719	82.5	**98.0**
128	1.452	**84.0**	97.0
256	2.967	**84.0**	97.5
512	6.194	83.0	96.5
1024	13.433	81.5	97.0
2048	31.058	82.5	97.0
4096	78.892	81.0	96.5

A.7 Temporal Generalization, Attribution, and Intervention Protocols

Temporal-generalization and signed attribution. For temporal generalization, we independently train a Stage-1 EEG–image alignment model at each temporal location and evaluate it at every other location on a 16-ms grid from 0 to 352 ms. We use $\mathrm{win} = 1$, so the input at each location contains a single sample from the 17 selected channels and has shape $[B, 17, 1]$, with no within-window temporal axis or relative temporal position. The high-level and low-level branches align the EEG representation with the RN50 high-level and RN50-LOW image features, respectively. Each branch uses a symmetric EEG–image InfoNCE objective with the image-only L2 normalization and fixed logit scale specified in Appendix A.2. The positive label for each EEG sample is its paired image within the batch, while the other in-batch images are negatives. Evaluation uses 200-way image-retrieval Top-1 accuracy. The main-text matrices report the high- and low-level targets separately. The signed occlusion curve in Figure 4b spans the 0–1000 ms input window and serves as a model-level diagnostic: an 8-ms EEG window is masked at 4-ms steps and the difference between the original and masked matched EEG–image scores is averaged over 200 tests. Negative values are not interpreted as a reversed neural code.

Complete retrieval and reconstruction interventions. Table 16 pairs Stage-1 retrieval with the associated Stage-2 reconstruction results for the configurations summarized in Figure 4c. For EEGProject, we introduce two TSConv-inspired controls. First-layer weights are shared within 40-ms blocks, so all samples in a block enter the linear readout through their aggregate response; this

simulates non-overlapping temporal aggregation without resampling the stored input. The second control applies a $W = 51$ overlapping average to the input while retaining a dense 4-ms grid, matching TSConv's 204-ms pooling extent but not its stride. We evaluate a 2×2 factorial design over pooling window W and stride S. These operations vary two architectural aspects of the same effective temporal-granularity axis: aggregation extent and output-grid density. Every reconstruction row in Table 16 uses the same Stage-2 setup: the corresponding fixed Stage-1 checkpoints initialize the branches, the EEG encoder remains unfrozen with a $0.1\times$ learning-rate scale, the VAR decoder is trained for 150 epochs, and the visual encoder and VQ tokenizer remain frozen. All reconstruction metrics are evaluated at epoch 150. Both EEGProject coarsening controls degrade all seven reconstruction metrics. For TSConv, reducing either the pooling window or the stride improves mean retrieval accuracy. Stride refinement improves all seven reconstruction metrics, whereas window reduction alone improves five. Combining both refinements improves all seven metrics over the original TSConv and brings its retrieval and reconstruction performance close to EEGProject. In Table 16, "Smoothing" denotes an explicit input or pooling operation, while "Weight grid" denotes the interval between positions exposed to distinct downstream weights; the block-aggregation effect of EEGProject weight sharing is described above.

Table 16: **Bidirectional temporal interventions on Subject 8.** Retrieval reports the mean $\pm$ std over the same five fixed seeds of 200-way Top-1/Top-5 accuracy (%); reconstruction is evaluated at epoch 150. Higher is better except SwAV (↓).

	Temporal setup		Retrieval		Reconstruction						
Encoder	Smoothing	Weight grid	Top-1	Top-5	PixCorr	SSIM	Alex(2)	Alex(5)	Inc.	CLIP	SwAV↓
EEGProject	None	4 ms	84.20 ± 0.91	98.40 ± 0.42	.2907	.3994	.9430	.9697	.8636	.8426	.4625
EEGProject	None	40 ms	72.30 ± 2.17	95.70 ± 1.04	.2077	.3748	.8851	.9422	.8353	.8136	.4843
EEGProject	204 ms	4 ms	71.30 ± 3.05	97.00 ± 0.71	.1703	.3770	.8314	.9069	.7995	.7830	.5040
TSConv	204 ms	20 ms	79.30 ± 1.68	96.20 ± 0.57	.2275	.3754	.9077	.9527	.8218	.8090	.4874
TSConv	204 ms	4 ms	81.90 ± 2.33	97.70 ± 0.45	.2408	.3904	.9196	.9540	.8320	.8339	.4751
TSConv	20 ms	20 ms	82.50 ± 1.00	98.00 ± 0.94	.2188	.3943	.9147	.9530	.8363	.8107	.4888
TSConv	20 ms	4 ms	84.60 ± 1.39	98.70 ± 0.84	.2691	.4086	.9387	.9668	.8590	.8409	.4618

The full EEGProject weight-sharing sweep progressively enlarges the non-overlapping blocks over which first-layer weights are tied. Because all samples within a block share the same coefficient, the readout depends on their blockwise aggregate. Top-1/Top-5 retrieval is stable through 20 ms, decreases sharply at 40 ms, and collapses at 100–200 ms (Table 17).

For compact visualization only, Figure 4c summarizes reconstruction with an index relative to the controlled EEGProject reference r:

$$I(x) = 100 \exp\left[\frac{1}{7}\left(\sum_{m \in \mathcal{M}_+} \log \frac{m(x)}{m(r)} + \log \frac{\mathrm{SwAV}(r)}{\mathrm{SwAV}(x)}\right)\right], \tag{7}$$

where $\mathcal{M}_+$ contains PixCorr, SSIM, AlexNet(2), AlexNet(5), InceptionV3, and CLIP (Radford et al., 2021); SwAV follows Caron et al. (2020). This geometric mean gives equal multiplicative weight to the six higher-is-better metrics and the inverted SwAV ratio. It is not introduced as an additional evaluation metric; all claims are based on the individual values in Table 16.

Table 17: **EEGProject temporal weight-sharing sweep on Subject 8.** The block length is the smallest interval over which first-layer weights are forced to be identical. 200-way Top-1/Top-5 retrieval accuracy (%) is reported as mean ± std over the same five fixed seeds as Table 16.

Metric	4 ms	20 ms	40 ms	100 ms	200 ms
Top-1	84.20 ± 0.91	83.80 ± 1.35	72.30 ± 2.17	48.10 ± 2.04	27.70 ± 1.60
Top-5	98.40 ± 0.42	98.70 ± 0.76	95.70 ± 1.04	82.10 ± 1.08	59.60 ± 1.43

A.8 Cross-Subject Replication on Subjects 5 and 9

To check that the encoder ranking of Table 4 and the bidirectional interventions of Table 16 do not depend on Subject 8, we repeat both on Subjects 5 and 9 under the same Stage-1 and Stage-2 protocols: retrieval is averaged over five seeds, and reconstruction is one Stage-2 run per configuration evaluated at epoch 150. Table 18 reports the encoder comparison. EEGProject gives the highest Top-1 and Top-5 and the best PixCorr, AlexNet(2), AlexNet(5), Inception, and SwAV on both subjects; ATM is higher on SSIM and CLIP on Subject 5 and on CLIP on Subject 9, TSConv has the highest SSIM on Subject 9, and LaBraM remains far weakest. Table 19 reports the two endpoint interventions. Sharing EEGProject's first-layer weights within 40-ms blocks lowers Top-1 by 15.8 and 18.2 points on Subjects 5 and 9 and degrades six and seven of the seven reconstruction metrics, respectively; the exception is SSIM on Subject 5, consistent with the SSIM diagnostic in Appendix A.12. Reducing TSConv's pooling window to 20 ms and its stride to 4 ms raises Top-1 by 5.5 and 4.4 points and improves seven and six of the seven metrics; the exception is SSIM on Subject 9. Both intervention directions therefore reproduce the Subject-8 pattern.

Table 18: **Controlled encoder comparison on Subjects 5 and 9.** Same protocol as Table 4: retrieval reports 200-way Top-1/Top-5 accuracy (%) averaged over five seeds, and reconstruction uses one Stage-2 run per configuration at epoch 150. Higher is better except SwAV (↓); best per subject and column in **bold**.

Subject	Encoder	Top-1	Top-5	PixCorr	SSIM	Alex(2)	Alex(5)	Inc.	CLIP	SwAV↓
Sub-05	EEGProject	**65.90**	**90.30**	**.2077**	.3338	**.8824**	**.9311**	**.7689**	.7615	**.5139**
	ATM	63.80	85.40	.2018	**.3674**	.8635	.9084	.7666	**.7661**	.5221
	TSConv	56.40	83.70	.1879	.3488	.8427	.8929	.7014	.7447	.5375
	LaBraM	11.90	26.80	.0564	.2493	.5038	.5084	.4950	.5081	.7259
Sub-09	EEGProject	**75.00**	**96.70**	**.2583**	.3670	**.9062**	**.9411**	**.8118**	.7835	**.4808**
	ATM	74.20	95.70	.2398	.3828	.8994	.9377	.7957	**.7849**	.4990
	TSConv	64.80	92.10	.2251	**.3855**	.8621	.9067	.7573	.7698	.5082
	LaBraM	20.20	42.80	.0711	.2886	.5245	.5218	.4882	.5067	.7145

Table 19: **Bidirectional temporal interventions on Subjects 5 and 9.** Rows are the reference and endpoint configurations of Table 16. Retrieval reports 200-way Top-1/Top-5 accuracy (%) averaged over five seeds; reference rows repeat the corresponding entries of Table 18. Reconstruction is one Stage-2 run per configuration at epoch 150. Higher is better except SwAV (↓).

		Temporal setup		Retrieval		Reconstruction						
Subject	Encoder	Smoothing	Weight grid	Top-1	Top-5	PixCorr	SSIM	Alex(2)	Alex(5)	Inc.	CLIP	SwAV↓
Sub-05	EEGProject	None	4 ms	65.90	90.30	.2077	.3338	.8824	.9311	.7689	.7615	.5139
	EEGProject	None	40 ms	50.10	79.60	.1582	.3385	.7892	.8506	.6861	.7320	.5447
	TSConv	204 ms	20 ms	56.40	83.70	.1879	.3488	.8427	.8929	.7014	.7447	.5375
	TSConv	20 ms	4 ms	61.90	91.20	.1972	.3712	.8835	.9192	.7361	.7670	.5144
Sub-09	EEGProject	None	4 ms	75.00	96.70	.2583	.3670	.9062	.9411	.8118	.7835	.4808
	EEGProject	None	40 ms	56.80	89.50	.1542	.3578	.8404	.9015	.7735	.7760	.5087
	TSConv	204 ms	20 ms	64.80	92.10	.2251	.3855	.8621	.9067	.7573	.7698	.5082
	TSConv	20 ms	4 ms	69.20	94.70	.2509	.3788	.8978	.9332	.7967	.8003	.4866

A.9 Alignment and Augmentation Controls

We report branch-wise retrieval and additional controls for the visual alignment targets.

Table 20: **Branch-wise retrieval on Subject 8.** Mean ± std over five Stage-1 seeds.

	High	Low	Fusion
Top-1 (%)	71.20 ± 2.05	77.60 ± 1.52	84.20 ± 0.91
Top-5 (%)	93.80 ± 0.84	97.70 ± 0.97	98.40 ± 0.42

Branch-wise retrieval. Table 20 reports the complete five-seed results summarized in Sec. 4.3.

We categorized the applied image augmentations into two tiers:

- **High-level:** Gaussian blur, Gaussian noise, low resolution, and mosaic. Inspired by Zhang et al. (2026), these choices aim to capture semantic-level concepts such as subject position, general shape, and color tone.
- **Low-level:** Gabor max response, gradient X, local binary patterns (LBP) texture, and patch shuffling. These are designed to model fine-grained details, including edges, textures, and patterns.

For the high-level targets, Gaussian blur uses a 15×15 kernel with OpenCV's automatically determined sigma, and Gaussian noise has mean 0 and standard deviation 25 on the $[0, 255]$ pixel scale. Low-resolution augmentation downsamples each image to 10% of its original size and restores it with bilinear interpolation. Mosaic downsamples by a factor of 5 and restores the image with nearest-neighbor interpolation. Blur and noise parameters use zero fluctuation range.

For the low-level targets, Gabor maximum response uses kernel size 31, $\sigma = 4$, orientations $\{0, \pi/4, \pi/2, 3\pi/4\}$, wavelength 10, aspect ratio 0.5, and phase 0. Gradient X uses a Sobel kernel of size 3. LBP uses radius 1, eight sampling points, and the uniform-pattern encoding. Patch shuffling resizes images to 512×512 and permutes the 16×16 grid of 32×32 patches.

Each transformation is applied to every image when its feature set is extracted, without an additional application probability. Noise and patch permutations are sampled once during offline feature extraction and cached. For each level, the four resulting feature sets are averaged with equal weights, without an identity transformation. These cached pooled features are used for both training and test-time retrieval as specified in Appendix A.2.

Table 21 compares the proposed augmentation set with no augmentation and additional transformations of similar types. The proposed set improves retrieval and reconstruction over no augmentation; adding more types lowers Top-1 from 84.20% to 73.50% and degrades most reconstruction metrics. The tested benefit therefore saturates at the proposed set.

The "w/o multi-level" control in Table 3 retains two independent EEG branches, both aligned to high-level targets with different augmentation sets. Its lower retrieval and reconstruction scores support the benefit of complementary high/low targets over this same-level ensemble. This differs from reusing one high-level embedding at both decoder routes in Table 5.

Table 21: **Image augmentation controls on Subject 8.** Reconstruction is reported at epoch 150; retrieval is 200-way Top-1/Top-5 accuracy.

Configuration	PixCorr	SSIM	Alex(2)	Alex(5)	Inc.	CLIP	SwAV↓	Top-1	Top-5
TRACE	**.2907**	.3994	**.9430**	**.9697**	**.8636**	**.8426**	.4625	**84.20**	**98.40**
w/o image augmentation	.2788	.3976	.9251	.9504	.8385	.8259	.4707	76.00	96.50
w/ additional image augmentation	.2565	**.4034**	.9225	.9605	.8556	.8264	**.4613**	73.50	97.00

A.10 Visualizing Level-Specific Evidence

EEG-CAM extraction. For TRACE, CAM uses the live RN50 feature of the displayed image I, with the image encoder's standard preprocessing, rather than the cached augmentation-pooled retrieval target. Specifically, $z_{I*}(I) = P_{I*}(f_{I*}(I))$ and $\bar{z}_{I*}(I) = z_{I*}(I)/\|z_{I*}(I)\|_2$. We backpropagate $S_*(I) = z_{E*}^{\top} \bar{z}_{I*}(I)$ through the frozen RN50 image encoder, treating the EEG embedding as a constant. Image normalization is part of the differentiated computation; the EEG embedding

is not L2-normalized. Grad-CAM uses layer4 for the high-level branch and layers2 and 3 for the low-level branch. Each map is formed by gradient-weighted channel aggregation, ReLU, and bilinear upsampling to the image size. The high-level map is min–max normalized. For the low-level branch, the two upsampled maps are separately min–max normalized, averaged with equal weights, and min–max normalized again. The resulting maps are overlaid on the input image. EEG-CAM introduces no additional training objective.

Qualitative comparison. Previous methods (Zheng et al., 2026) use different visual models to capture different levels of visual information, but this does not necessarily yield a clear division between semantic and structural evidence. Figure 10 compares our two branches with Brain-HIVE's SynCLR, CLIP, and VAE pathways. In these examples, the latter exhibit overlapping responses or emphasize background regions, making their level-specific roles less clear. This motivates our use of different feature depths within one visual backbone, paired with level-specific augmentation sets. Our high-level branch emphasizes broad object-related regions, while the low-level branch highlights local details such as raspberry surface patterns and wheelchair components. These visualizations qualitatively support a clearer division of semantic and structural evidence.

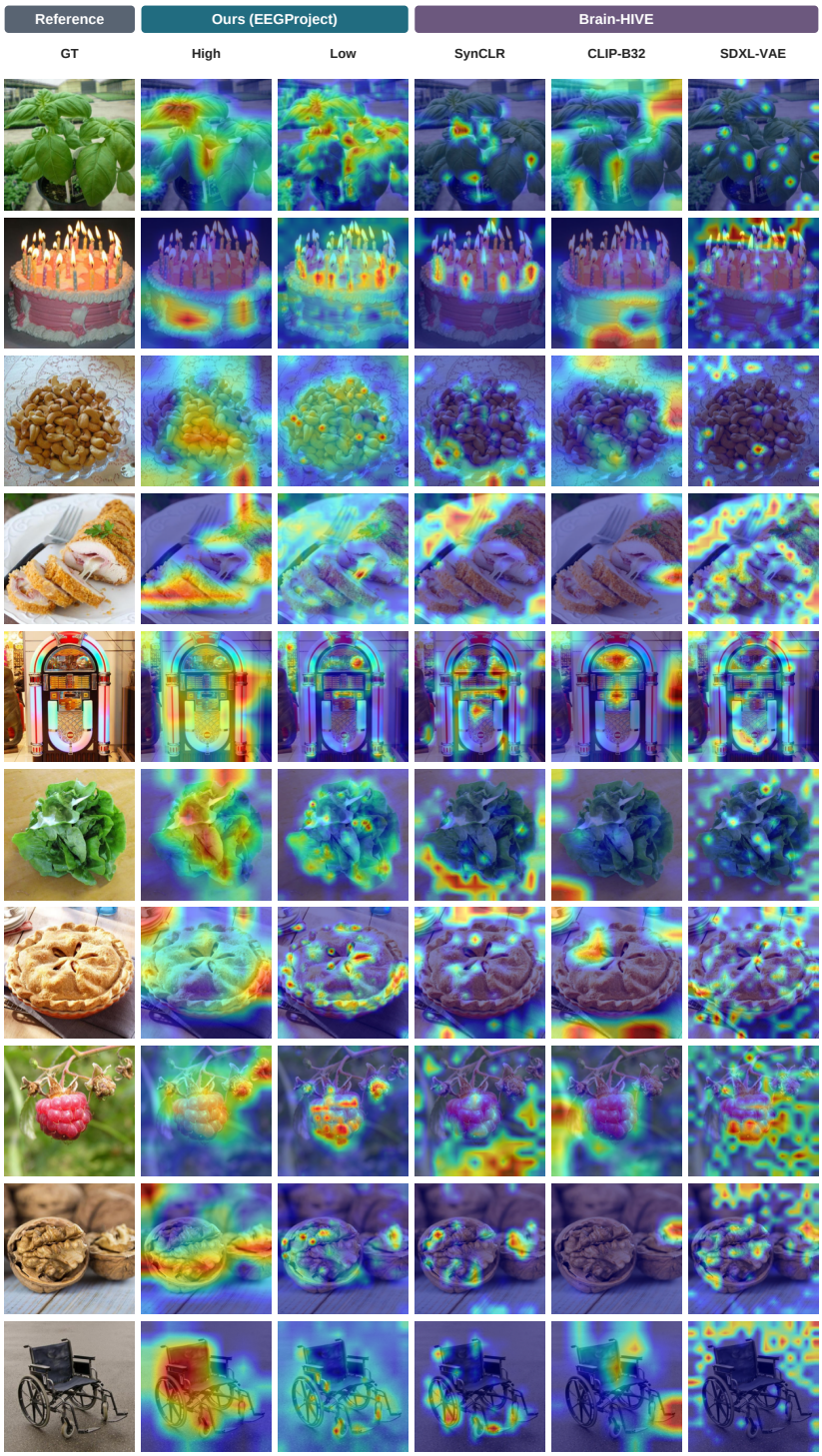

Figure 10: **EEG-CAM comparison on Subject 8.** Each row shows the same stimulus: ground truth, our high- and low-level branches, and Brain-HIVE's SynCLR, CLIP-B32, and SDXL-VAE pathways. Our branches show a clearer distinction between broad object-related evidence and local structural details in these selected examples.

A.11 Training Strategy Ablations

Training strategy. Our final protocol fine-tunes the Stage-1-trained encoders at a 0.1× learning rate during Stage-2 generation training. Table 22 compares this with three alternatives. Applying the same learning rate to all modules ("Same LR, encoder unfrozen") substantially lowers all metrics in this experiment, suggesting that the aligned encoders benefit from gentler adaptation. Freezing the encoders entirely ("Encoder frozen") also degrades performance, though less severely, indicating that some encoder adaptation is beneficial. Joint training—optimizing both alignment and generation objectives simultaneously—degrades all metrics further (PixCorr drops to .1998, SSIM to .3697); under this setup, the two objectives do not optimize stably together. Together, these results support the two-stage protocol and the 0.1× learning-rate choice for the Stage-1-trained components.

A.12 Test-Time Temporal Shift and Metric Diagnostics

Protocol. We use test-time temporal shifts as a downstream stress test of the temporal information retained by each encoder. For Subject 8, we keep the Stage-1 alignment checkpoints and the Stage-2 VAR checkpoint at epoch 150 fixed. We evaluate signed shifts $\delta \in \{-24, -20, -16, -12, -8, -4, 0, 4, 8, 12, 16, 20, 24\}$ ms with zero padding on the 250-Hz EEG input. Positive δ shifts the signal toward later times, whereas negative δ shifts it toward earlier times. The same 200 test images, generation configuration, and random seed are used for all shifts.

The resulting reconstruction curves are shown in Figure 11. The plot reports the raw metric values across all tested shifts; higher is better except for SwAV. For visualization, results at equal-magnitude positive and negative shifts are averaged, so the horizontal axis reports shift magnitude.

Fine-grained temporal encoding under shift. At zero shift, EEGProject is the strongest of the four encoders on all seven reconstruction metrics when the metric directions are taken into account. Increasing the shift produces the opposite ordering in temporal sensitivity: EEGProject has the largest directional deterioration in PixCorr and in the five feature-based metrics (AlexNet(2), AlexNet(5), InceptionV3, CLIP, and SwAV). TSConv overtakes EEGProject on most of these metrics at approximately 8–12 ms, and at the largest tested shift EEGProject is the weakest or near-weakest encoder. The clean-performance ranking and the shift-sensitivity ranking therefore move in opposite directions. Together with the weight-sharing and pooling/stride interventions in Appendix A.7, this stress test provides downstream evidence that maintaining fine-grained temporal encoding is useful for the encoder's reconstruction performance. It should be read as controlled supporting evidence on one subject and one reconstruction run, rather than as a standalone causal test. Appendix A.13 isolates the contribution of the position-specific parameterization to this pattern.

Figure 12 visualizes this pattern with selected test examples. The zero-shift reconstructions preserve more of the target content for EEGProject, while increasing the shift progressively replaces stimulus-specific content with unrelated but visually plausible structures. Figure 13 shows the same diagnostic at the maximum shift for additional LaBraM examples.

An unexpected SSIM diagnostic. SSIM (Wang et al., 2004) does not follow the degradation pattern above. It is non-monotonic under the same temporal perturbation: relative to its own zero-shift value, SSIM increases by 9.3% for TSConv and 35.1% for the diagnostic LaBraM run at 24 ms, even though PixCorr and the feature-based metrics deteriorate. EEGProject's SSIM also recovers after an initial decrease, but remains below its zero-shift value at the endpoint. This divergence is consistent with SSIM's emphasis on local luminance, contrast, and structural statistics: an output can retain locally natural-looking statistics without preserving the stimulus-specific content. Similar discrepancies have been reported in visual brain decoding: Park et al. (2026) show that SSIM can overrate semantically incorrect reconstructions, while Scotti et al. (2023) report higher SSIM for blurry reconstructions lacking high-level semantic content than for their full model. We therefore report SSIM together with pixel- and feature-based criteria and do not use it alone as evidence of reconstruction quality.

Table 22: **Training strategy ablations on Subject 8.** Reconstruction at epoch 150. All rows use the same conditioning configuration (High→SOS, Low→AdaLN). "Same LR" applies uniform learning rates; "Encoder frozen" keeps encoders fixed in Stage 2; "Joint training" optimizes alignment and generation losses simultaneously.

Training strategy	PixCorr	SSIM	Alex(2)	Alex(5)	Inc.	CLIP	SwAV↓
TRACE (0.1× encoder LR)	**.2907**	**.3994**	**.9430**	**.9697**	**.8636**	**.8426**	**.4625**
Encoder frozen	.2572	.3991	.9141	.9523	.8225	.8090	.4730
Joint training	.1998	.3697	.8821	.9300	.8102	.7769	.4922
Same LR, encoder unfrozen	.2544	.3617	.9069	.9483	.8255	.8068	.4792

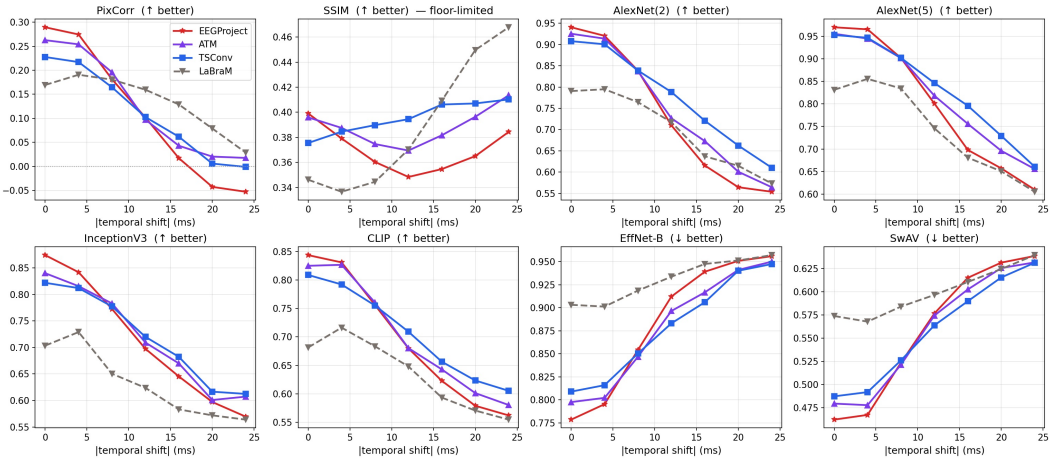

Figure 11: **Reconstruction metrics under test-time temporal shift.** Curves show the seven reconstruction metrics for EEGProject, ATM, TSConv, and LaBraM on Subject 8 at epoch 150 as the magnitude of the zero-padded input shift increases from 0 to 24 ms. Each point averages equal-magnitude earlier-time and later-time shifts. Higher is better except SwAV. EEGProject starts highest across the metrics and undergoes the steepest deterioration in PixCorr and the feature-based metrics, whereas SSIM is non-monotonic and is therefore interpreted as a diagnostic rather than in isolation.

A.13 Shift-Augmented Training: The Cost of Temporal Invariance

EEGProject uses position-specific temporal weights, raising the question of whether its test-time shift sensitivity merely reflects its parameterization or a genuine dependence on precise temporal alignment. To distinguish these explanations, we apply training-time temporal jitter while keeping the architecture unchanged, and examine whether shift tolerance improves without sacrificing clean decoding performance.

On Subject 8, we use circular shifts with half-widths of $\pm\{8, 20, 40, 80\}$ ms and a jitter-free control. We compare EEGProject with TSConv and assess EEGProject's clean retrieval, test-time shift tolerance, and reconstruction. Clean fused retrieval is reported over five seeds; cross-encoder comparisons, shift probes, and reconstruction use one run per condition. Reconstruction follows the Stage-2 protocol of Table 16 and is evaluated at epoch 150.

EEGProject outperforms TSConv on both retrieval branches at every jitter level (Table 23). Its high-level advantage increases from 3.5 percentage points without jitter to 24.0 points at ± 80 ms, showing that its advantage persists under training-time temporal perturbations.

Jitter training improves shift tolerance at the cost of clean performance (Tables 24 and 25). At ± 40 ms, retrieval is nearly constant across test-time shifts of 0–24 ms, but clean accuracy drops from $85.6 \pm 1.14\%$ to $67.1 \pm 2.92\%$. Mild jitter (± 8 ms) costs only 1.9 percentage points, yet retains just 10.2% of its zero-shift accuracy at a 24 ms test-time shift. Reconstruction also degrades: ± 8 ms jitter lowers six of seven metrics, and jitter of ± 20 ms or more yields worse values than the



Figure 12: **Qualitative reconstructions under test-time temporal shift.** The left column shows the ground truth, followed by EEGProject, TSConv, and LaBraM reconstructions at 0, 8, 16, and 24 ms input shift. Each row is a selected test example; the labels below the reconstructions indicate the shift magnitude.

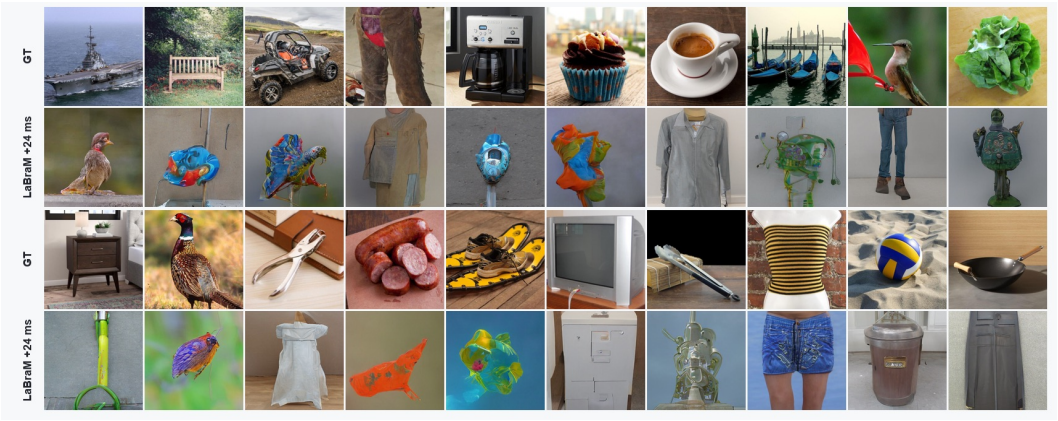

Figure 13: **LaBraM reconstructions at the maximum shift.** The four rows alternate ground truth and LaBraM reconstruction after a 24 ms (six-sample) zero-padded input shift. The upper and lower pairs of rows show selected examples from different portions of the test set. The qualitative mismatch illustrates why a higher SSIM under shift need not imply better stimulus correspondence.

control on all seven metrics (Table 26). We also tried alternative optimization strategies, but these did not improve performance under jitter-augmented training.

These results argue against a purely parameterization-based explanation of test-time shift sensitivity: training-time jitter substantially reduces this sensitivity, but at the cost of clean retrieval and reconstruction performance. Thus, precise temporal alignment at a scale of tens of milliseconds contributes information used for decoding, rather than merely reflecting an incidental sensitivity to fixed sample positions. Architecture and parameter count remain unchanged, and circular shifts preserve input samples.

A.14 Generator Choice and Computational Cost

We use VAR because its explicit coarse-to-fine stages provide distinct locations for Hierarchy-Consistent Injection. All routing controls in Table 5 use the same VAR backbone and differ only in the conditioning features and their routes. Differences within that table therefore isolate hierarchy-consistent conditioning rather than a change of generator family.

For computational context, TRACE contains 328 M trainable parameters out of 437 M total; the remaining parameters belong to the frozen VAE tokenizer. On one NVIDIA H100 (80GB) with batch size 32, a training batch requires 37.14 TFLOPs for the combined forward and backward passes (1.16 TFLOPs per sample), and 150 epochs take approximately 8.5 hours. Our adapted

Table 23: **Cross-encoder retrieval under training-time temporal jitter on Subject 8.** Entries are zero-shift 200-way Top-1 accuracy (%) from one run per condition for the high- and low-level branches.

Training jitter	EEGProject		TSConv	
	High	Low	High	Low
None (control)	72.0	79.0	68.5	71.0
$\pm$8 ms	75.0	78.5	66.0	66.0
$\pm$20 ms	69.0	69.0	56.5	58.0
$\pm$40 ms	59.5	60.0	43.5	45.5
$\pm$80 ms	51.0	47.5	27.0	30.0

Table 24: **Cost of shift-augmented training on Subject 8.** Zero-shift 200-way fused Top-1 retrieval accuracy (%), mean $\pm$ std over five fixed seeds. "Ret.@24" is the 24 ms value relative to the same arm's zero-shift value in the single-run shift probe of Table 25.

Training jitter	Top-1	Δ	Ret.@24
None (control)	85.6 ± 1.14	—	3.4%
$\pm$8 ms	83.7 ± 0.84	-1.9	10.2%
$\pm$20 ms	79.4 ± 1.75	-6.2	63.1%
$\pm$40 ms	67.1 ± 2.92	-18.5	100.4%
$\pm$80 ms	56.3 ± 2.68	-29.3	92.5%

diffusion generator contains 236 M trainable parameters and requires 21.87 TFLOPs per denoiser pass for the same batch size (0.68 TFLOPs per sample step).

At inference, VAR follows a fixed token-upscaling chain requiring 31.82 TFLOPs per batch (0.99 TFLOPs per sample); generating the 200 test images takes about 20 seconds. The adapted diffusion generator uses 50 denoising steps, requiring 693.11 TFLOPs per batch (21.66 TFLOPs per sample) and about 40 minutes for the same 200 images. In our setting, VAR therefore provides the coarse-to-fine structure required by Hierarchy-Consistent Injection while substantially reducing inference time relative to the adapted diffusion decoder.

A.15 Limitations

Our study has three main limitations. First, most mechanistic encoder analyses are conducted on Subject 8. The encoder comparison and the bidirectional interventions are replicated on Subjects 5 and 9 (Appendix A.8), but the temporal-generalization, attribution, test-time shift, and jitter analyses remain single-subject, and the ten-subject evaluation does not by itself establish that the same mechanism is invariant across individuals. Second, our experiments are conducted on THINGS-EEG only, so cross-dataset generalization is not evaluated. Finally, EEG-CAM provides qualitative evidence about image-space emphasis and does not by itself establish a causal neural mechanism.

B AI Use Statement

We used generative AI tools to assist with translation and to edit the manuscript to improve readability.

The authors reviewed all AI-assisted text. We checked translations and language edits against the original text for accuracy in terminology, numerical values, and meaning. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the assistance of generative AI.

Table 25: **Test-time shift probe for shift-augmented arms on Subject 8.** Rows are training jitter half-widths; columns are test-time zero-padded shifts. Entries are 200-way fused Top-1 retrieval accuracy (%), single run, so the zero-shift column differs from the five-seed means in Table 24.

Training jitter	0	4	8	12	16	20	24 ms
None (control)	88.00	73.25	42.50	17.00	8.25	4.00	3.00
$\pm$8 ms	86.00	84.00	79.75	65.25	41.75	18.50	8.75
$\pm$20 ms	78.50	79.00	78.00	75.00	72.00	65.50	49.50
$\pm$40 ms	67.50	67.50	67.75	68.50	66.75	67.25	67.75
$\pm$80 ms	56.50	55.25	55.25	53.25	52.75	52.25	52.25

Table 26: **Reconstruction under shift-augmented training on Subject 8.** One Stage-2 run per arm, evaluated at epoch 150. The control row is the TRACE entry of Table 1. Higher is better except SwAV (↓).

Training jitter	PixCorr	SSIM	Alex(2)	Alex(5)	Inc.	CLIP	SwAV↓
None (control)	.2907	.3994	.9430	.9697	.8636	.8426	.4625
$\pm$8 ms	.2596	.3903	.9346	.9648	.8428	.8286	.4557
$\pm$20 ms	.1925	.3842	.8812	.9323	.8194	.8097	.4850
$\pm$40 ms	.1889	.3841	.8884	.9337	.7768	.7911	.5022
$\pm$80 ms	.1776	.3701	.8798	.9125	.7796	.7760	.5018