VISUALIZING THE INVISIBLE: GENERATIVE VISUAL GROUNDING EMPOWERS UNIVERSAL EEG UNDERSTANDING IN MLLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

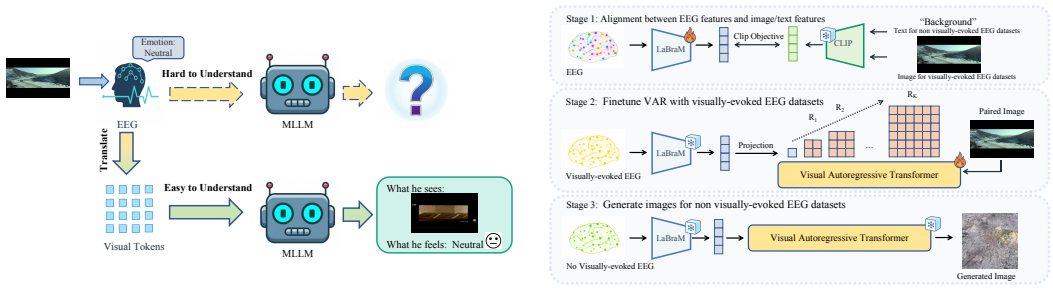
Leveraging the universal representations of pre-trained Large Language Models (LLMs) and Multimodal Large Language Models (MLLMs) has emerged as a promising paradigm for enhancing the capability of brain foundation models. Constrained by the severe scarcity of visually-evoked electroencephalography (EEG) datasets, existing foundation models predominantly focus on LLMs and compromise by aligning neural signals solely with abstract text—a lossy translation that inevitably discards the fine-grained, perceptual details encoded in brain activity. In this work, we propose **G**enerative **V**isual **G**rounding (**GVG**), a framework that visualizes the invisible. Rather than forcing neural signals into text, we employ an EEG-to-Image generative model as a "visual translator" to hallucinate instance-specific proxy images for non-visual EEG. This strategy provides structured visual contexts that allow MLLMs to apply their visual priors to interpret clinical states. To validate this core hypothesis, we first establish image-only alignment with GVG-X-Omni and GVG-Janus. GVG-X-Omni approaches the average performance of the 1.7B-parameter NeuroLM-XL while tuning 170M parameters over a frozen 7B backbone. Image-only GVG-Janus improves over text-only alignment on all six primary tasks, raising average balanced accuracy from 45.89% to 53.46% and surpassing NeuroLM-XL on four tasks. Building on this result, trimodal EEG–image–text alignment further raises GVG-Janus's average to 60.05% under full adaptation, the highest among the compared multi-task models. Both low-rank adaptation (LoRA) and full adaptation exceed NeuroLM-XL on all six tasks. Predicted native visual tokens also support EEG-based reconstruction through frozen decoders, while clinical analyses identify physiological variation retained in aligned EEG features. These results support visual proxy grounding as an effective complement to textual alignment for universal EEG understanding.

1 INTRODUCTION

The emergence of Large Language Models (LLMs) and Multimodal Large Language Models (MLLMs) has revolutionized AI (Achiam et al., 2023; Team et al., 2023; Wang et al., 2024b), inspiring the development of foundation models to learn universal representations for non-invasive brain signals such as electroencephalography (EEG) (Yi et al., 2023; Jiang et al., 2024b; Wang et al., 2024a; Fang et al., 2025). Studies of image decoding from non-invasive brain recordings show that performance improves with the amount of data recorded per participant (Banville et al., 2025). However, high-quality EEG data are expensive to collect and subject to privacy constraints, limiting the data available to train more capable EEG foundation models.

Recent approaches have connected EEG to language models by expressing EEG statistics as text (Kim et al., 2024) or learning language-compatible neural representations (Jiang et al., 2024a; Lu et al., 2025; Liu et al., 2025a). Aligning neural and language representations supports multi-task learning, allowing a unified decoder to learn from labels across tasks. Such a shared space also enables transfer of pretrained semantic knowledge, motivated by the hypothesis that representations learned across modalities can converge (Huh et al., 2024).

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.



(a) Core idea of GVG

(b) AVDE training and proxy-image generation

Figure 1: **Core idea and clinical proxy generation.** (a) GVG connects EEG to an MLLM's native visual vocabulary for understanding and reconstruction. (b) Before GVG training, AVDE (Dai et al., 2026) is adapted and frozen to generate proxy images from clinical EEG for Stage 1 alignment.

Nevertheless, the current paradigm of compressing neural signals solely into highly abstracted textual labels is inherently lossy. While text provides explicit semantic anchors and high-level categorical boundaries, it inevitably discards the fine-grained, perceptual details and spatial topologies encoded in dense brain activity. Because human visual perception is closely coupled with complex cognitive and emotional responses, we hypothesize that augmenting text alignment with the dense visual representation space of MLLMs may provide a useful complementary signal. While textual representations provide explicit semantic anchors, the continuous visual space may offer a dense perceptual canvas. This complementarity can help preserve the fine-grained spatial topologies and low-level neural dynamics that are often weakened during pure textual compression. By combining abstract textual concepts with visual grounding, we aim to leverage the multimodal (visual-linguistic) priors encoded in MLLMs for both discriminative understanding and generative interpretability.

Despite the appeal of this visually-augmented paradigm, operationalizing this strategy faces a severe bottleneck: its strict reliance on paired visual stimuli. First, compared to the billions of text-image pairs used to pre-train MLLMs, visually-evoked EEG datasets are inherently scarce and small-scale. More critically, the vast majority of real-world clinical applications—such as sleep staging and epilepsy detection—are completely blind (recorded purely physiologically without any external visual input). This creates a fundamental dilemma: researchers are often forced to rely on textual alignment for broad clinical applicability, while visual alignment remains largely confined to the small subset of stimulus-driven datasets.

Furthermore, even if visually-evoked EEG can be aligned well, we still face a substantial scale mismatch. Unlike the massive corpora available for text or images, EEG datasets remain too limited for MLLMs to acquire fluent EEG-specific representations from scratch. To mitigate this data bottleneck, our core insight is to translate neural dynamics into a modality that MLLMs already model well: vision. By mapping raw brain signals into recognizable visual formats, we reduce the burden of learning a new neural language from limited data and instead reuse established visual priors to interpret EEG.

Building on this insight, we propose **Generative Visual Grounding** (GVG), a framework that visualizes the invisible. For clinical recordings without paired stimuli, we use AVDE (Dai et al., 2026) to generate EEG-conditioned proxy images, bringing these datasets into the same visual-grounding pipeline as stimulus-paired EEG. To connect the aligned EEG features to pretrained MLLMs, a lightweight Transformer predicts tokens from the native visual vocabularies of X-Omni (Geng et al., 2025) and Janus (Wu et al., 2024). These EEG-predicted tokens complement continuous EEG features for task understanding and can also be decoded into images through the frozen native visual decoders. This connects EEG understanding and visual reconstruction within one framework, without requiring images at inference.

We first evaluate image-only grounding to examine the effectiveness of visual supervision. GVG-X-Omni approaches the average performance of NeuroLM-XL while tuning only 170M parameters over a frozen 7B backbone, compared with 1.7B trainable parameters for NeuroLM-XL. Image-only GVG-Janus improves over its text-only counterpart on all six primary tasks and outperforms NeuroLM-XL on four. Notably, this reverses the ranking observed with text-only Janus, whose

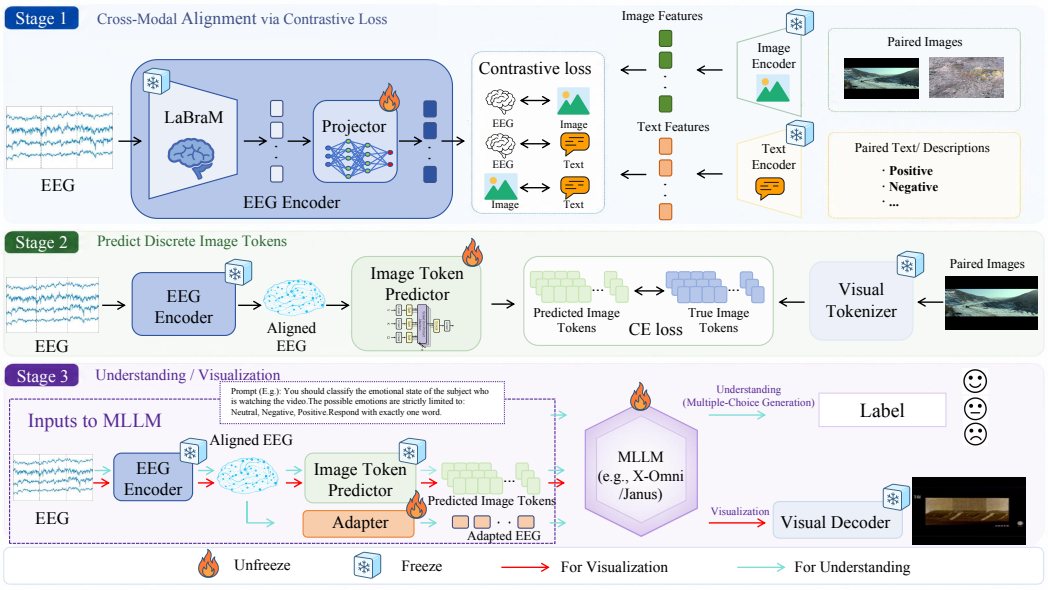


Figure 2: **The three-stage GVG pipeline.** Stage 1 aligns EEG with visual targets, with text added for trimodal alignment. Stage 2 learns native visual-token prediction from real EEG–stimulus pairs using cross-entropy (CE) loss. Stage 3 combines predicted tokens, continuous EEG features, and task prompts for understanding. The predicted tokens also support reconstruction through the frozen visual decoder.

average performance falls below NeuroLM-XL. We then combine visual and textual anchors through trimodal alignment, further strengthening Janus under both low-rank adaptation (LoRA) and full adaptation. Both configurations outperform NeuroLM-XL on all six primary tasks, with Full reaching 60.05% average balanced accuracy, the highest among the compared multitask models. These results support visual grounding as an effective complement to conventional EEG–text alignment.

Our main contributions are summarized as follows:

- **Generative Visual Grounding for Non-Visual EEG:** We use a generative model as a visual translator, extending visual grounding from stimulus-paired EEG to clinical recordings through training-only proxy images.
- **Modality Translation via Discrete Visual Tokens:** We map aligned EEG features into the native visual vocabularies of off-the-shelf MLLMs and combine them with a continuous EEG prefix for EEG-only task prediction.
- **Parameter-Efficient Adaptation and Competitive Understanding:** We evaluate image-only grounding with GVG-X-Omni and GVG-Janus, then extend Janus with trimodal alignment. The two backbones cover parameter-efficient and full-adaptation settings, with evaluation across six benchmarks and additional tasks from the TUSL and Workload.
- **Unified Understanding and Visual Reconstruction:** We connect EEG understanding and native visual decoding in one framework. Reconstruction examines the decoded visual content; additional controls, category geometry, cross-modal correspondence, and physiological readouts provide complementary evidence about the learned EEG features.

2 Related Work

Large Brain Foundation Models. Brain foundation models for EEG (Kostas et al., 2021; Zhang et al., 2022; Yang et al., 2023b; Cai et al., 2023; Jiang et al., 2024b; Wang et al., 2024a; Cui et al., 2024; Dimofte et al., 2025), intracranial recordings (Wang et al., 2023), and functional magnetic resonance imaging (fMRI) (Caro et al., 2023; Yang et al., 2024) use masked, contrastive,

and autoregressive objectives. Cross-modal systems, including NeuroLM (Jiang et al., 2024a), UniMind (Lu et al., 2025), E^2-LLM (Ma et al., 2026), and Brant-X (Zhang et al., 2024), connect heterogeneous tasks through language or physiological interfaces. These systems motivate shared task decoding. GVG asks a complementary question about the representation supplied to such a decoder: whether supervision through a native visual vocabulary can organize EEG features and connect them to an existing multimodal interface.

Cross-Modal Alignment between EEG and Vision. Prior work maps visually evoked EEG to image representations or generative latents, including Brain2Image (Kavasidis et al., 2017), early EEG–image alignment (Spampinato et al., 2017), DreamDiffusion (Bai et al., 2023), NEUROIMAGEN (Lan et al., 2023), EEG2Video (Liu et al., 2024), MINDEV (Huang et al., 2025), and EEGMirror (Liu et al., 2025b). Related fMRI reconstruction methods include Brain-Diffuser (Ozcelik & VanRullen, 2023) and Mind-Video (Chen et al., 2023). GVG shares the objective of connecting EEG to visual representations, but uses the resulting discrete tokens as inputs to downstream task understanding. This role also motivates clinical pseudo-targets where stimulus images are unavailable. Reconstruction remains useful as a readout of the learned interface, alongside rather than in place of task and representation evaluation.

3 METHOD

3.1 OVERVIEW AND NATIVE VISUAL INTERFACE

GVG uses visual supervision during training to connect EEG to a pretrained MLLM (Figure 2). Real stimulus images provide targets for visually evoked EEG, while an adapted AVDE teacher supplies proxy images for clinical recordings. The GVG pipeline then learns cross-modal feature alignment, native visual-token prediction, and prompted task understanding in three stages. At inference, the task model combines EEG-predicted visual tokens with a continuous EEG prefix; the same tokens can also be read by the frozen visual decoder for reconstruction.

We instantiate this interface with X-Omni (Geng et al., 2025) and Janus-1.3B (Wu et al., 2024). X-Omni uses SigLIP2-based visual features (Tschannen et al., 2025), a Qwen2.5-7B language backbone (Qwen Team, 2024), and a FLUX.1-dev diffusion decoder (Black Forest Labs, 2024). Janus separates its SigLIP-Large-Patch16-384 understanding encoder (Zhai et al., 2023) from its VQ-VAE generation tokenizer, with DeepSeek-LLM-1.3B as the language backbone. Both visual vocabularies contain 16,384 codes. GVG uses continuous image features for alignment and native tokenizer outputs as discrete prediction targets, keeping these two forms of visual supervision distinct.

3.2 TRAINING-ONLY VISUAL TARGETS

To supply visual targets for clinical EEG without paired stimuli, we adapt AVDE (Dai et al., 2026) before the three GVG training stages (Figure 1b). The teacher's EEG encoder A_ψ learns a shared CLIP conditioning space by aligning paired visual EEG with image features using contrastive InfoNCE and mean squared error (MSE) objectives, and clinical EEG with text features of a fixed task-label description $D_d(y)$. Its VAR decoder G_ψ (Tian et al., 2024) is trained on real EEG–stimulus pairs to predict vector-quantized (VQ) image tokens from EEG features. After adaptation, we freeze the teacher and pass each clinical EEG recording X_i through A_ψ and G_ψ offline to obtain its proxy image $I_i^p = G_\psi(A_\psi(X_i))$. GVG's Stage 1 then uses the real stimulus I_i for paired visual EEG and the proxy I_i^p for clinical EEG as its visual target I_i^*. The subsequent stages of native visual-token prediction and task adaptation support EEG-based task understanding and visual reconstruction (Section A.2).

3.3 CROSS-MODAL FEATURE ALIGNMENT

Pretrained EEG and multimodal encoders learn different feature spaces, so we first establish a shared space for cross-modal supervision. Previous EEG–language approaches map neural signals into language-compatible representations for multi-task decoding (Jiang et al., 2024a; Lu et al., 2025). In our setting, class names provide shared category anchors, while real stimuli and EEG-conditioned proxy images supply visual targets for individual recordings. We therefore align EEG, image, and text features jointly in Stage 1, combining categorical semantics with visual supervision

to prepare EEG features for subsequent native visual-token prediction. We initialize the EEG encoder with pretrained LaBraM (Jiang et al., 2024b), which represents channel–time patches using temporal and spatiotemporal processing. For each training sample, let $\mathbf{x}^i_{eeg} = X_i$, $\mathbf{x}^i_{img} = I^*_i$, and $\mathbf{x}^i_{text} = \ell_i = L_d(y_i)$, where $L_d(y_i)$ is the canonical class-name string. Encoders and modality-specific projectors map these inputs into a shared D-dimensional space:

$$\mathbf{h}^i_a = \phi_a\big(\mathrm{Enc}_a(\mathbf{x}^i_a)\big) \in \mathbb{R}^D, \qquad a \in \{\mathrm{eeg}, \mathrm{img}, \mathrm{text}\}. \tag{1}$$

Stage 1 trains the modality-specific projectors and logit scales while keeping the pretrained EEG, visual, and text encoders frozen. Following pairwise contrastive alignment (Radford et al., 2021; Tschannen et al., 2025; Zhai et al., 2023), we combine the three modality pairs:

$$\mathcal{L}_{\mathrm{tri}} = \lambda_{ei}\mathcal{L}_{ei} + \lambda_{et}\mathcal{L}_{et} + \lambda_{it}\mathcal{L}_{it}. \tag{2}$$

For normalized features, $\mathcal{L}_{NCE}(a, b)$ is the mean of the $a \to b$ and $b \to a$ cross-entropies. The EEG–image and EEG–text terms combine this contrastive objective with a feature-matching loss:

$$\mathcal{L}_{ab} = \alpha\mathcal{L}_{NCE}(\mathbf{h}_a, \mathbf{h}_b) + (1-\alpha)\|\mathbf{h}_a - \mathrm{sg}(\mathbf{h}_b)\|^2_2, \tag{3}$$

where $a = eeg$, $b \in \{img, text\}$, and $\mathrm{sg}(\cdot)$ denotes stop-gradient. The image–text term uses contrastive alignment alone, $\mathcal{L}_{it} = \mathcal{L}_{NCE}(\mathbf{h}_{img}, \mathbf{h}_{text})$.

3.4 Native Visual-Token Prediction

Stage 1 establishes cross-modal correspondence in a continuous feature space. To reuse the MLLM's pretrained visual decoder, the aligned EEG features must be mapped to its native visual tokens. The same tokens also provide visual embeddings for downstream task understanding, linking both uses through a shared representation. Stage 2 therefore trains a Transformer token predictor only on real EEG–stimulus pairs, with target sequences $s_i = T(I_i)$ produced by the frozen native tokenizer. Stage 1 modules and the visual towers remain frozen, and $M = 256$ learned queries use non-causal self-attention to predict all token positions in parallel.

The output head follows the chosen native interface. Janus predicts vocabulary logits with a learned classification head. For X-Omni, the predictor outputs a vector $\mathbf{z}_{im}$ in the codebook embedding space and computes its similarity to each fixed code vector $\mathbf{v}_k$:

$$p_\theta(s_{im} = k \mid \mathbf{h}^i_{\mathrm{eeg}}) = \frac{\exp(\mathbf{z}_{im} \cdot \mathbf{v}_k/\tau)}{\sum_{j=1}^{K}\exp(\mathbf{z}_{im} \cdot \mathbf{v}_j/\tau)}, \tag{4}$$

where $K = |\mathcal{V}| = 16{,}384$, τ is the temperature, and $\mathbf{v}_k \in \mathbb{R}^{D_v}$ is a code embedding. Both predictors minimize token cross-entropy averaged over samples and token positions. Main-model optimization settings appear in Section A.1.

3.5 EEG-Only Understanding and Visual Readout

Stage 3 initializes from the selected Stage 2 checkpoint, freezes the aligned EEG encoder and token predictor, and jointly learns task responses across the six primary datasets by adapting the EEG-to-LLM adapter and language model. We formulate classification as multiple-choice generation: the task prompt lists every candidate label, and training minimizes cross-entropy on the correct option-letter token. The input combines the prompt, 256 EEG-predicted visual-token embeddings, and a continuous EEG prefix. Constrained decoding selects an answer from the task's candidate set:

$$\hat{y} = f_{\mathrm{MLLM}}\big(q_d, \mathrm{Embed}(\hat{s}(X)), e(X); \mathcal{C}_d\big), \tag{5}$$

where q_d is the prompt, $\mathcal{C}_d$ is the complete candidate set, $\hat{s}(X)$ contains the argmax token predictions, and $e(X)$ projects continuous EEG features into the language backbone's input space. The candidate set is identical for every sample of a task, and both sample-dependent inputs are computed from EEG. Visual tokens provide the native interface, while the continuous prefix bypasses the codebook, allowing task prediction to access EEG features without requiring every relevant attribute to be expressed as a visual code. The computation and information flow are detailed in Algorithm 1, Table 7, and Sections A.2 and A.3.

The predicted tokens also support a separate visual readout, $\hat{I} = D_{\mathrm{vis}}(\hat{s}(X))$, using the corresponding frozen image decoder. This branch uses neither the language model's task response nor the continuous EEG prefix, and requires no additional decoder training. Reconstruction therefore examines the visual content that can be decoded from the predicted tokens. Its role as an interface probe is evaluated alongside task performance and representation analyses in Section 5.

Table 1: Balanced accuracy (BAcc, %) across eight EEG datasets. **Bold** marks the best per dataset; underlining marks the best among NeuroLM and GVG. Train.: trainable parameters.

Model	Train.	SEED	SEED-IV	SEED-VII	TUEV	TUAB	HMC	TUSL	Workload
Specialist models									
SPaRCNet	0.79M	55.96	29.88	25.12	41.61	78.69	47.56	41.85	59.77
ContraWR	1.6M	61.06	38.38	31.32	43.84	80.17	42.42	58.57	69.66
CNN-Transformer	3.2M	61.61	35.21	24.93	40.87	79.53	65.73	35.75	57.93
FFCL	2.4M	58.08	37.81	29.30	39.79	78.19	44.27	38.19	**70.69**
ST-Transformer	3.5M	54.79	36.93	21.56	39.84	79.66	25.59	40.00	61.03
BIOT	3.2M	70.97	36.19	32.63	52.81	79.59	68.62	57.58	66.55
LaBraM	5.8M	**73.18**	**47.63**	33.91	**64.09**	**81.40**	**72.86**	**76.25**	66.09
Multi-task baselines									
NeuroLM-B	254M	55.54	29.64	29.67	45.60	78.26	67.37	67.34	61.72
NeuroLM-L	500M	60.06	31.98	30.18	41.32	78.76	66.58	53.14	63.11
NeuroLM-XL	1.7B	60.34	32.30	30.03	46.79	79.69	57.61	68.45	63.45
GVG configurations									
GVG-X-Omni	170M	60.73	30.25	27.17	44.15	74.45	61.20	52.59	64.94
GVG-Janus (image)	1.7B	57.29	42.02	26.64	48.38	80.27	66.18	58.07	66.67
GVG-Janus (LoRA)	49M	65.35	44.80	32.10	52.44	80.75	62.21	70.36	67.59
GVG-Janus (full)	1.7B	68.70	45.35	**34.45**	60.20	80.08	71.51	73.78	69.31

4 EXPERIMENTS

4.1 DATASETS AND EVALUATION PROTOCOL

The primary evaluation covers SEED (Duan et al., 2013; Zheng & Lu, 2015) and its SEED-IV (Zheng et al., 2019) and SEED-VII (Jiang et al., 2025) extensions for emotion recognition, along with HMC for sleep staging (Alvarez-Estevez & Rijsman, 2021), TUAB for abnormality detection (Lopez et al., 2015), and TUEV for event classification (Harati et al., 2015). For each configuration and seed, a single jointly trained Stage 3 checkpoint serves all six tasks. To assess transfer to electrode configurations and tasks unseen during GVG pretraining, we further evaluate on TUSL (von Weltin et al., 2017) and Workload (Zyma et al., 2019).

Our baselines include multi-task models (Jiang et al., 2024a) and specialist models (Jing et al., 2023; Yang et al., 2023b; Peh et al., 2022; Li et al., 2022; Song et al., 2021; Yang et al., 2023a; Jiang et al., 2024b). We report balanced accuracy (BAcc), the mean class recall within each dataset, and Macro_6, its unweighted average over the six primary datasets. Dataset splits and training details appear in Sections A, A.1 and C.1.

4.2 PERFORMANCE ON UNDERSTANDING TASKS

Image-only GVG's results against NeuroLM-XL demonstrate the effectiveness of visual supervision (Table 1). GVG-X-Omni trains 170M parameters over a frozen 7B backbone and reaches 49.66% Macro_6, approaching the 51.13% of NeuroLM-XL, a 1.7B-parameter model pretrained on approximately 25,000 hours of EEG (Jiang et al., 2024a). It also exceeds NeuroLM-XL on SEED and HMC. Image-only GVG-Janus reaches 53.46% and outperforms NeuroLM-XL on four of six tasks, with gains of 9.72 and 8.57 percentage points on SEED-IV and HMC.

Notably, text-only Janus reaches 45.89% Macro_6, below NeuroLM-XL. Using visual targets for Stage 1 alignment reverses this ranking and improves Janus on all six tasks (Table 2), including the three clinical tasks whose visual targets are generated proxies. These gains support the effectiveness of visual supervision for EEG understanding beyond text alignment alone.

Building on this result, trimodal alignment combines visual and textual anchors to further improve GVG-Janus. Full reaches 60.05% Macro_6, exceeding image-only grounding on five of six tasks (Table 2), while LoRA reaches 56.28% with 49M trainable parameters. Both exceed NeuroLM-XL on all six primary tasks, and Full surpasses every NeuroLM variant on each task. Exploratory dataset-level paired tests for Full yield Holm-adjusted $p = 0.0335$ against each NeuroLM variant

Table 2: Text alignment and visual grounding (BAcc, %). Image-only and Full match Table 1.

Configuration	SEED	SEED-IV	SEED-VII	HMC	TUAB	TUEV	Macro_6
Text-only	54.85	34.06	19.03	42.89	80.11	44.41	45.89
Image-only	57.29	42.02	26.64	66.18	**80.27**	48.38	53.46
Full (trimodal)	**68.70**	**45.35**	**34.45**	**71.51**	80.08	**60.20**	**60.05**

Table 3: Visual-target controls (Macro_6, %). Per-task scores in Section C.

Visual target	Macro_6 (%)
Full: EEG-conditioned proxy	60.05
Class prototype	55.58
Random image	47.18
Within-class shuffled proxy	40.46
Dataset-global shuffled proxy	39.67

Table 4: Clinical proxies: EEG adjusted Rand index (ARI) and class-centered correspondence.

Dataset	EEG ARI		Residual	
	Full	w/o S1	$\Delta\cos$	p
HMC	0.4832	0.0860	.0226	$< 10^{-4}$
TUAB	0.3451	0.0034	$-.0037$	.232
TUEV	0.4163	0.0560	.0335	$< 10^{-4}$

(Section B.2). Against the specialist models, GVG-Janus (Full) ranks in the top two on four of the six primary datasets and third on SEED and TUAB, showing competitive performance with a shared multi-task checkpoint (Table 1).

To assess performance on datasets unseen during GVG pretraining, we evaluate on TUSL and Workload. Full is the strongest GVG configuration on both tasks, reaching 73.78% and 69.31%, compared with image-only Janus's 58.07% and 66.67%. Relative to LaBraM, Full is 2.47 points lower on TUSL and 3.22 points higher on Workload. These results show that the GVG paradigm generalizes effectively to montages and tasks unseen during pretraining. These tasks are excluded from Macro_6; adaptation protocols and seed results appear in Section A.4.

5 Analysis

5.1 Visual Alignment and Token Prediction

The image-only and trimodal results in Table 2 show the value of visual supervision alongside textual category anchors. Beyond feature alignment, GVG combines continuous EEG features with predicted native visual tokens for task understanding. A continuous-EEG-only model retains Stage 1/2 pretraining but reaches 55.45% Macro_6, compared with 60.05% for Full (Table 12). The predicted visual-token branch therefore complements continuous EEG features in downstream understanding.

The native visual vocabulary also supplies pretraining targets for the EEG-to-token mapping. Stage 2 learns this mapping from real EEG–stimulus pairs while keeping the aligned EEG features frozen. Omitting Stage 2 reduces Macro_6 to 51.76% (Table 8), showing the value of learning this visual mapping before task adaptation.

5.2 EEG–Proxy Correspondence

The gains from visual supervision leave open what information EEG-conditioned proxies contribute as Stage 1 targets for clinical EEG. Replacing these proxies with one fixed image per class tests whether category information alone is sufficient to preserve their benefit. These class prototypes yield 55.58% Macro_6, compared with 60.05% for Full; random images yield 47.18% (Table 3). Category-level visual anchors therefore retain substantial utility, while EEG-conditioned targets provide an additional benefit.

The gain over class prototypes could reflect richer visual targets, their correspondence with individual EEG recordings, or both. Shuffling preserves the image pool while breaking its original EEG assignments: global shuffling reduces Macro_6 to 39.67%, and within-class shuffling reaches only 40.46% despite preserving category identity. The latter result shows that category matching alone cannot recover the benefit of the original EEG–proxy pairs. Gaussian noise provides a complementary control, degrading the generated proxy images while leaving their EEG assignments intact. Performance



(a) GVG-Janus (b) GVG-Janus

(c) GVG-X-Omni (d) GVG-X-Omni

Figure 3: **EEG-based visual reconstruction.** Each panel shows the stimulus (GT), its ground-truth-token reconstruction (GT-tokens), and the EEG prediction (Pred), from left to right. Both models use their frozen native visual decoders.

declines as noise increases (Table 16), showing that the visual content of correctly paired targets also contributes to downstream understanding.

5.3 Visual Reconstruction

The reconstruction experiment tests whether EEG-predicted visual tokens carry information about the corresponding stimulus. We decode these tokens with the frozen native decoder, using EEG alone without decoder fine-tuning. On stimulus-paired datasets, peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), and learned perceptual image patch similarity (LPIPS) assess pixel agreement, structural similarity, and perceptual distance, respectively (Table 21). On SEED-VII, GVG-Janus improves all three over AVDE, while GVG-X-Omni achieves the highest PSNR and SSIM. On SEED, both GVG variants improve PSNR, whereas AVDE leads SSIM and LPIPS. These results show that the native visual-token interface supports competitive reconstruction through existing decoders.

The examples in Figure 3 show partial recovery of broad appearance despite changes in depicted objects. Each EEG prediction is compared with its stimulus and ground-truth-token reconstruction, which provides a reference for the native image coding path. In panel (a), the prediction retains the yellow-green palette without recovering the person; panel (b) retains the predominantly dark appearance.

To assess whether this visual correspondence extends beyond the displayed examples, we use GPT-6 to score 100 randomly sampled GVG-Janus test reconstructions from each of SEED, SEED-IV, and SEED-VII (Section E.1). Color and spatial layout receive higher mean scores than semantic category in all three datasets, with pooled scores of 2.58, 2.59, and 2.12 out of 5, respectively. The agreement between these ratings and the displayed examples supports partial recovery of overall appearance and composition, with weaker category correspondence.

5.4 Clinical Representation Analysis

The preceding experiments show that visual proxies and trimodal alignment benefit downstream EEG understanding. Clearer organization of clinical task categories in the aligned EEG features would help account for these gains. On identical subsets of HMC, TUAB, and TUEV, K-means clusters from Full agree more closely with clinical labels than those from Stage-1-ablated checkpoints, as measured by adjusted Rand index (ARI; Table 4). The downstream improvement is thus accompanied by a representation that groups clinical states more consistently.

Category organization alone leaves open whether a recording's departure from its category prototype is accompanied by a corresponding departure in its proxy representation. For each modality, subtracting the class centroid isolates these deviations; cosine similarity between paired EEG and proxy residuals then measures their correspondence. Full exceeds the Stage-1-ablated model on HMC and TUEV ($\Delta = 0.0226$ and 0.0335), with both effects significant after multiple-testing correction; TUAB shows no significant change (Table 4 and Section D). The within-class shuffling control tests the

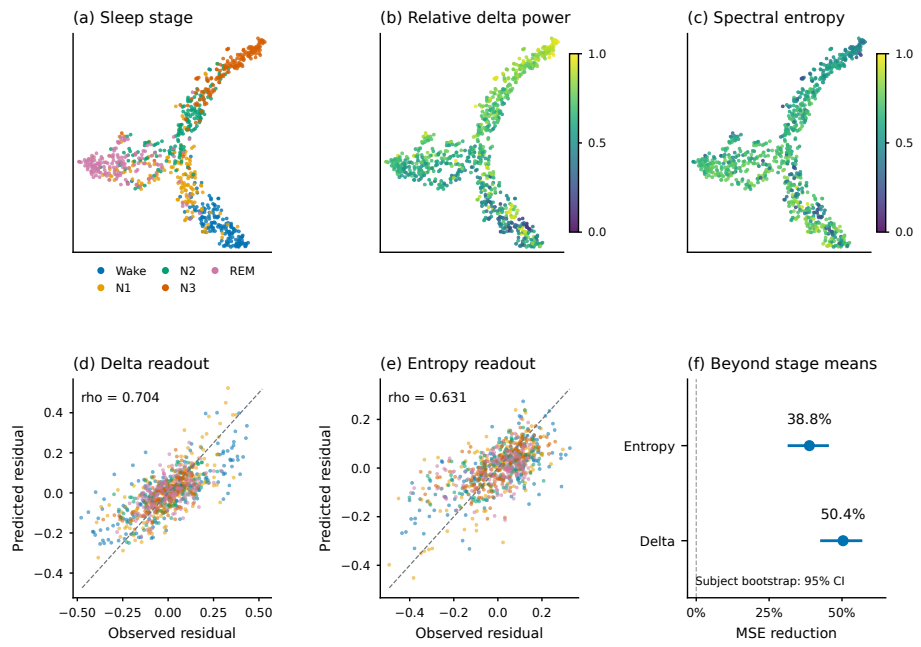


Figure 4: **Physiological information retained in aligned EEG features.** (a–c) Identical t-distributed stochastic neighbor embedding (t-SNE) coordinates colored by sleep stage (REM: rapid eye movement), relative delta power, and spectral entropy; visualization only. (d–e) Stage-centered predictions on held-out subjects. (f) Subject-weighted mean squared error (MSE) reductions over stage-only baselines with 95% subject-bootstrap confidence intervals (CI). Linear probes use frozen features and training-fold preprocessing.

downstream value of preserving individual pairings; this analysis locates that correspondence in the learned feature space.

Cross-modal correspondence does not specify which physiological properties remain in the EEG features. Relative delta power and normalized spectral entropy provide interpretable targets on HMC, describing low-frequency activity and the distribution of spectral power (Figure 4). Because both quantities vary continuously within a sleep stage, their prediction from frozen Full EEG features tests for physiological content beyond categorical task labels. Ridge probes predict variation remaining after training-fold stage centering, using subject-held-out evaluation over 750 windows from 26 subjects. The stage-only baseline already captures differences between sleep stages, so improving on it requires recovering variation around those stage means. Relative to these baselines, the probes reduce mean squared error averaged equally across subjects by 50.4% and 38.8%, respectively (Holm-adjusted $p = .004$ for both; Section D.2). These gains show that the aligned EEG features retain continuous physiological variation within sleep stages, extending the evidence from category structure to information beyond the task labels.

6 Conclusion

We introduced GVG to complement textual alignment with visual grounding for heterogeneous EEG, using training-only proxies for clinical recordings. Full trimodal GVG-Janus achieves 60.05% Macro_6, the highest among the compared multi-task models; both LoRA and full adaptation outperform NeuroLM-XL on all six primary tasks. Controls support the value of EEG–proxy pairing beyond category matching. Token-prediction pretraining and visual-token inputs improve task understanding, while the same tokens support reconstruction through frozen native decoders. Frozen-feature probes on HMC further reveal physiological variation within sleep stages, showing that the aligned EEG representations retain information beyond task labels.

AI USE STATEMENT

In this work, we used generative AI tools to polish the manuscript, generate selected images, score model-generated images, and cross-check experimental data. The authors reviewed the AI-assisted work and take full responsibility for the final content of this paper, including its text, claims, results, and artifacts.

ETHICS STATEMENT

This work analyzes previously collected EEG datasets under their respective access and usage terms. EEG is sensitive biometric and health-related data; released artifacts should exclude identifying metadata and follow each dataset's governance requirements. Generated clinical pseudo-targets are research supervision signals, not patient-facing explanations or clinical measurements. Clinical deployment would require dedicated validation, governance, and human oversight.

REPRODUCIBILITY STATEMENT

The model architectures and hyperparameters are described in this paper. We will publicly release the model implementations, training code, model checkpoints, and generated images upon acceptance.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Diego Alvarez-Estevez and Roselyne M Rijsman. Inter-database validation of a deep learning approach for automatic sleep scoring. *PloS one*, 16(8):e0256111, 2021.

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.

Hubert Banville, Yohann Benchetrit, Stéphane d'Ascoli, Jérémy Rapin, and Jean-Rémi King. Scaling laws for decoding images from brain activity. *arXiv preprint arXiv:2501.15322*, 2025.

Black Forest Labs. FLUX.1 [dev]. Model release, 2024. URL `https://huggingface.co/black-forest-labs/FLUX.1-dev`.

Donghong Cai, Junru Chen, Yang Yang, Teng Liu, and Yafeng Li. Mbrain: A multi-channel self-supervised learning framework for brain signals. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 130–141, 2023.

Josue Ortega Caro, Antonio H de O Fonseca, Christopher Averill, Syed A Rizvi, Matteo Rosati, James L Cross, Prateek Mittal, Emanuele Zappala, Daniel Levine, Rahul M Dhodapkar, et al. BrainLM: A foundation model for brain activity recordings. *bioRxiv*, 2023. doi: 10.1101/2023.09.12.557460.

Zijiao Chen, Jiaxin Qing, and Juan Helen Zhou. Cinematic mindscapes: High-quality video reconstruction from brain activity. *Advances in Neural Information Processing Systems*, 36:24841–24858, 2023.

Wenhui Cui, Woojae Jeong, Philipp Thölke, Takfarinas Medani, Karim Jerbi, Anand A Joshi, and Richard M Leahy. Neuro-gpt: Towards a foundation model for eeg. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5. IEEE, 2024.

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

Alexandru Dimofte, Glenn Anta Bucagu, Thorir Mar Ingolfsson, Xiaying Wang, Andrea Cossettini, Luca Benini, and Yawei Li. Cerebro: Compact encoder for representations of brain oscillations using efficient alternating attention. *arXiv preprint arXiv:2501.10885*, 2025.

Ruo-Nan Duan, Jia-Yi Zhu, and Bao-Liang Lu. Differential entropy feature for EEG-based emotion classification. In *6th International IEEE/EMBS Conference on Neural Engineering (NER)*, pp. 81–84. IEEE, 2013.

Zitao Fang, Chenxuan Li, Hongting Zhou, Shuyang Yu, Guodong Du, Ashwaq Qasem, Yang Lu, Jing Li, Junsong Zhang, and Sim Kuan Goh. Neuript: Foundation model for neural interfaces. *arXiv preprint arXiv:2510.16548*, 2025.

Zigang Geng, Yibing Wang, Yeyao Ma, Chen Li, Yongming Rao, Shuyang Gu, Zhao Zhong, Qinglin Lu, Han Hu, Xiaosong Zhang, et al. X-omni: Reinforcement learning makes discrete autoregressive image generative models great again. *arXiv preprint arXiv:2507.22058*, 2025.

Amir Harati, Meysam Golmohammadi, Silvia Lopez, Iyad Obeid, and Joseph Picone. Improved eeg event classification using differential energy. In *2015 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, pp. 1–4. IEEE, 2015.

Shuai Huang, Yongxiong Wang, Huan Luo, Haodong Jing, Chendong Qin, and Jingqun Tang. Mindev: Multi-modal integrated diffusion framework for video reconstruction from eeg signals. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 3350–3359, 2025.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024.

Wei-Bang Jiang, Yansen Wang, Bao-Liang Lu, and Dongsheng Li. Neurolm: A universal multi-task foundation model for bridging the gap between language and eeg signals. *arXiv preprint arXiv:2409.00101*, 2024a.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous eeg data in bci. *arXiv preprint arXiv:2405.18765*, 2024b.

Wei-Bang Jiang, Xuan-Hao Liu, Wei-Long Zheng, and Bao-Liang Lu. Seed-vii: A multimodal dataset of six basic emotions with continuous labels for emotion recognition. *IEEE Transactions on Affective Computing*, 16(2):969–985, 2025. doi: 10.1109/TAFFC.2024.3485057.

Jin Jing, Wendong Ge, Shenda Hong, Marta Bento Fernandes, Zhen Lin, Chaoqi Yang, Sungtae An, Aaron F Struck, Aline Herlopian, Ioannis Karakis, et al. Development of expert-level classification of seizures and rhythmic and periodic patterns during eeg interpretation. *Neurology*, 100(17): e1750–e1762, 2023.

Isaak Kavasidis, Simone Palazzo, Concetto Spampinato, Daniela Giordano, and Mubarak Shah. Brain2image: Converting brain signals into images. In *Proceedings of the 25th ACM international conference on Multimedia*, pp. 1809–1817, 2017.

Jonathan W Kim, Ahmed Alaa, and Danilo Bernardo. Eeg-gpt: exploring capabilities of large language models for eeg classification and interpretation. *arXiv preprint arXiv:2401.18006*, 2024.

Demetres Kostas, Stephane Aroca-Ouellette, and Frank Rudzicz. Bendr: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of eeg data. *Frontiers in Human Neuroscience*, 15:653659, 2021.

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.

Hongli Li, Man Ding, Ronghua Zhang, and Chunbo Xiu. Motor imagery eeg classification algorithm based on cnn-lstm feature fusion network. *Biomedical signal processing and control*, 72:103342, 2022.

Chenyu Liu, Yuqiu Deng, Tianyu Liu, Jinan Zhou, Xinliang Zhou, Ziyu Jia, and Yi Ding. Echo: Toward contextual seq2seq paradigms in large eeg models. *arXiv preprint arXiv:2509.22556*, 2025a.

Xuan-Hao Liu, Yan-Kai Liu, Yansen Wang, Kan Ren, Hanwen Shi, Zilong Wang, Dongsheng Li, Bao-Liang Lu, and Wei-Long Zheng. Eeg2video: Towards decoding dynamic visual perception from eeg signals. *Advances in Neural Information Processing Systems*, 37:72245–72273, 2024.

Xuan-Hao Liu, Bao-Liang Lu, and Wei-Long Zheng. Eegmirror: Leveraging eeg data in the wild via montage-agnostic self-supervision for eeg to video decoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18273–18283, 2025b.

S. Lopez, G. Suarez, D. Jungreis, I. Obeid, and J. Picone. Automated identification of abnormal adult EEGs. In *2015 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, pp. 1–5. IEEE, 2015. doi: 10.1109/SPMB.2015.7405423.

Weiheng Lu, Chunfeng Song, Jiamin Wu, Pengyu Zhu, Yuchen Zhou, Weijian Mai, Qihao Zheng, and Wanli Ouyang. Unimind: Unleashing the power of llms for unified multi-task brain decoding. *arXiv preprint arXiv:2506.18962*, 2025.

Fei Ma, Han Lin, Yifan Xie, Hongwei Ren, Xiaoyu Shen, Wenbo Ding, and Qi Tian. Eˆ 2-llm: Bridging neural signals and interpretable affective analysis. *arXiv preprint arXiv:2601.07877*, 2026.

Furkan Ozcelik and Rufin VanRullen. Natural scene reconstruction from fMRI signals using generative latent diffusion. *Scientific Reports*, 13(1):15666, 2023. doi: 10.1038/s41598-023-42891-8.

Wei Yan Peh, Yuanyuan Yao, and Justin Dauwels. Transformer convolutional neural networks for automated artifact detection in scalp eeg. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 3599–3602. IEEE, 2022.

Qwen Team. Qwen2.5: A party of foundation models, September 2024. URL `https://qwenlm.github.io/blog/qwen2.5/`.

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*, pp. 8748–8763. PmLR, 2021.

Yonghao Song, Xueyu Jia, Lie Yang, and Longhan Xie. Transformer-based spatial-temporal feature learning for eeg decoding. *arXiv preprint arXiv:2106.11170*, 2021.

Concetto Spampinato, Simone Palazzo, Isaak Kavasidis, Daniela Giordano, Nasim Souly, and Mubarak Shah. Deep learning human mind for automated visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6809–6817, 2017.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 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.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Eva von Weltin, Tameem Ahsan, Vinit Shah, Dawer Jamshed, Meysam Golmohammadi, Iyad Obeid, and Joseph Picone. Electroencephalographic slowing: A primary source of error in automatic seizure detection. In *2017 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, pp. 1–5. IEEE, 2017.

Christopher Wang, Vighnesh Subramaniam, Adam Uri Yaari, Gabriel Kreiman, Boris Katz, Ignacio Cases, and Andrei Barbu. Brainbert: Self-supervised representation learning for intracranial recordings. *arXiv preprint arXiv:2302.14367*, 2023.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. Cbramod: A criss-cross brain foundation model for eeg decoding. *arXiv preprint arXiv:2412.07236*, 2024a.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024b.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *arXiv preprint arXiv:2410.13848*, 2024.

Chaoqi Yang, M Westover, and Jimeng Sun. Biot: Biosignal transformer for cross-data learning in the wild. *Advances in Neural Information Processing Systems*, 36:78240–78260, 2023a.

Chaoqi Yang, Cao Xiao, M Brandon Westover, and Jimeng Sun. Self-supervised electroencephalogram representation learning for automatic sleep staging: model development and evaluation study. *JMIR AI*, 2(1):e46769, 2023b.

Yifan Yang, Yutong Mao, Xufu Liu, and Xiao Liu. Brainmae: a region-aware self-supervised learning framework for brain signals. *arXiv preprint arXiv:2406.17086*, 2024.

Ke Yi, Yansen Wang, Kan Ren, and Dongsheng Li. Learning topology-agnostic eeg representations with geometry-aware modeling. *Advances in Neural Information Processing Systems*, 36:53875–53891, 2023.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11975–11986, 2023.

Daoze Zhang, Zhizhang Yuan, Junru Chen, Kerui Chen, and Yang Yang. Brant-X: A unified physiological signal alignment framework. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4155–4166. ACM, 2024. doi: 10.1145/3637528.3671953.

Xiang Zhang, Ziyuan Zhao, Theodoros Tsiligkaridis, and Marinka Zitnik. Self-supervised contrastive pre-training for time series via time-frequency consistency. *Advances in neural information processing systems*, 35:3988–4003, 2022.

Wei-Long Zheng and Bao-Liang Lu. Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. *IEEE Transactions on Autonomous Mental Development*, 7(3):162–175, 2015. doi: 10.1109/TAMD.2015.2431497.

Wei-Long Zheng, Wei Liu, Yifei Lu, Bao-Liang Lu, and Andrzej Cichocki. EmotionMeter: A multimodal framework for recognizing human emotions. *IEEE Transactions on Cybernetics*, 49 (3):1110–1122, 2019. doi: 10.1109/TCYB.2018.2797176.

Igor Zyma, Sergii Tukaev, Ivan Seleznov, Ken Kiyono, Anton Popov, Mariia Chernykh, and Oleksii Shpenkov. Electroencephalograms during mental arithmetic task performance. *Data*, 4(1):14, 2019. doi: 10.3390/data4010014.

A Main Experimental Protocol and Training Details

SEED, SEED-IV, and SEED-VII use 4-second windows sampled at 200Hz. Paired video frames supervise the visual-dataset alignment stages; inference operates from EEG and a task prompt. Stage 1 uses clinical sampling weights 0.3 and visual-dataset weights 1.0; Stage 2 draws targets from real paired visual datasets; Stage 3 jointly optimizes task-conditioned understanding across all six primary datasets. Representation analyses pin EEG encoder, EEG projector, and image projector tensors by SHA-256 and reuse identical sample manifests across compared checkpoints.

Dataset splits. SEED, SEED-IV, and SEED-VII all use subject-dependent train/validation/test splits at the trial level, with 9/3/3, 16/4/4, and 12/4/4 trials, respectively. GVG and all reproduced baselines share the same SEED-VII partition.

TUAB and TUEV use their official subject-independent train/test partitions with an 80/20 train/validation split inside the official training set. HMC uses subjects 1–100/101–126/127–154 for training/validation/test. SEED-VII baseline results in Table 1 are reproduced as described below. The TUSL and Workload splits appear in Section A.4.

SEED-VII baseline reproductions. The seven specialist baselines and all three NeuroLM variants use the SEED-VII partition specified above. These reproductions share a data manifest and use validation-BAcc checkpoint selection. NeuroLM instruction tuning starts from released pretrained checkpoints without OpenWebText replay. Each variant trains for ten epochs with seed 0, learning rate 5×10^{-5} (minimum 5×10^{-6}), and effective batch size 256. For B, validation BAcc selects this rate from $\{10^{-5}, 5 \times 10^{-5}, 10^{-4}\}$; L and XL use the same rate. The reported test scores use validation-selected checkpoints from epochs 7, 8, and 5 for B, L, and XL, respectively. Other benchmark cells retain the published reference values (Jiang et al., 2024a).

Evaluation metrics. Balanced accuracy (BAcc) is mean class recall within each dataset. The primary benchmark aggregate covers $\mathcal{D}_6 = \{\text{SEED}, \text{SEED-IV}, \text{SEED-VII}, \text{HMC}, \text{TUAB}, \text{TUEV}\}$:

$$\text{MacroBAcc}_6 = \frac{1}{6} \sum_{d \in \mathcal{D}_6} \text{BAcc}(d). \tag{6}$$

Each dataset has equal weight, and rounding follows averaging. Full uses seeds 42, 1, and 2; within each seed, the same jointly trained checkpoint supplies all six task scores. Dataset means are then averaged across seeds. The same six-task aggregate is used wherever all six results are available. Main-model training settings and seed-wise results appear in Sections A.1 and C.1.

TUSL and Workload use separate supervised task-specific adaptations and do not enter the six-task average, main-task ablations, or representation-test family.

A.1 Input Tensors and Main-Model Hyperparameters

Model configurations. Text-only Janus uses the Stage 1 EEG–text alignment procedure, without visual-target alignment, native visual-token training, or visual-token inputs. Image-only and Full learn the native visual-token route; Full also incorporates textual category anchors. Image-only grounding is instantiated with X-Omni's Qwen2.5-7B-based backbone and Janus-1.3B. The former tunes 170M parameters over a frozen language backbone. Janus supports both image-only and trimodal alignment, with LoRA and full adaptation for the trimodal configuration. The full configuration uses $(\lambda_{ei}, \lambda_{et}, \lambda_{it}) = (0.4, 0.4, 0.2)$ and $\alpha = 0.8$; image-only variants optimize $\mathcal{L}_{ei}$. Janus LoRA uses rank 32, and X-Omni uses rank 8. These settings examine visual grounding across backbones and the added value of text alignment within Janus. The trainable-parameter counts in Table 1 describe the primary-benchmark adaptation budgets; inference costs are reported in Section C.2.

For the reported tasks, the model consumes tensors arranged as channels $\times$ temporal patches $\times$ 200 samples: SEED/SEED-IV/SEED-VII use $62 \times 4 \times 200$, HMC uses $4 \times 16 \times 200$, TUAB uses $23 \times 10 \times 200$, and TUEV uses $23 \times 5 \times 200$. Each patch spans one second at 200Hz. GVG performs reshaping and dataset-specific truncation but does not repeat the upstream signal filtering used to construct the released tensors. Two deterministic text lookups are distinct: $D_d(y)$ returns the descriptive sentence used to adapt AVDE on clinical EEG, whereas $L_d(y)$ returns the canonical class-name string consumed by GVG Stage 1. Both mappings are fixed before training.

Stage 3 uses hard argmax to select token IDs from the frozen predictor, including its learned queries and output head. Task-response gradients update the EEG-to-LLM adapter and the language-model parameters selected by the full or LoRA configuration. Table 18 reports the model's resource requirements.

All stages use AdamW, weight decay 0.05, seed 42 for the reference run, gradient clipping 1.0, and 10% warmup. Stage 1 uses learning rate 10^{-4}. Stage 2 uses 2×10^{-6} (classifier), 10^{-6} (decoder), and 2×10^{-7} (other trainable modules), label smoothing 0.05, non-causal token queries, and the

Table 5: Clinical text targets. $L_d(y)$ is the canonical class string for GVG Stage 1; $D_d(y)$ is used only when adapting the AVDE teacher.

Dataset	y	$L_d(y)$	$D_d(y)$
TUEV	0	spsw	An EEG recording showing spike and sharp wave patterns, abnormal transient waveforms indicating cortical irritability.
	1	gped	An EEG recording showing generalized periodic epileptiform discharge, repetitive sharp discharges seen bilaterally.
	2	pled	An EEG recording showing periodic lateralized epileptiform discharge, repetitive sharp discharges lateralized to one hemisphere.
	3	eyem	An EEG recording showing eye movement, ocular artifact from saccades or blinks.
	4	artf	An EEG recording showing artifact, non-cerebral signal contamination.
	5	bckg	An EEG recording showing background, normal ongoing cerebral activity.
TUAB	0	normal	A normal EEG recording within normal physiological range.
	1	abnormal	An abnormal EEG recording with pathological patterns such as slowing, epileptiform discharges, or asymmetry.
HMC	0	Wake	An EEG recording during the wake stage, alert state with alpha or beta rhythms and high muscle tone.
	1	N1	An EEG recording during light sleep N1, drowsy transitional stage with theta waves and slow eye movements.
	2	N2	An EEG recording during light sleep N2, stable light sleep marked by sleep spindles and K-complexes.
	3	N3	An EEG recording during deep sleep N3, slow-wave sleep dominated by high-amplitude delta waves.
	4	REM	An EEG recording during REM sleep, rapid eye movement sleep with low-amplitude mixed-frequency activity.

Table 6: Optimization schedule of Full GVG-Janus. Total params. includes frozen components; the final column lists the modules updated in each stage. Batch is per GPU; Stage 1/2 use 8 GPUs.

Stage	Schedule	Batch $\times$ accumulation	Total params.	Trainable modules
1: alignment	5 epochs	256×1	2.15B	EEG/image/text projectors; logit scales; pretrained encoders frozen
2: visual tokens	max 100; stopped at 26 (patience 5)	128×2	2.15B	token classifier/decoder; learned queries/positions; adapter; Stage 1 and visual towers frozen
3: understanding	10 epochs	8×2	2.15B	LLM; EEG-to-LLM adapter; EEG encoder/projector and token predictor frozen

best validation-accuracy checkpoint. Stage 3 initializes from that checkpoint and uses 5×10^{-5} (EEG-to-LLM adapter) and 10^{-6} (LLM), with label smoothing 0.1. The main configurations use seeds 42, 1, and 2, with seed-wise results in Section C.1.

A.2 Training and Inference Data Flow

Algorithm 1 specifies teacher preparation, offline proxy generation, the three GVG training stages, and inference. Here, $F_\theta = \phi_{\text{eeg}} \circ \text{Enc}_{\text{eeg}}$ combines the frozen LaBraM backbone with the EEG projector learned in Stage 1.

The task prompt q_d lists the complete candidate set $\mathcal{C}_d$, shared by all samples in dataset d. The strings $q_1 =$ "Image reconstructed from EEG:", $q_2 =$ "EEG Features:", and $q_3 =$ "Answer:" delimit embedding blocks. Multiple-choice training applies loss to the response tokens, and inference constrains the answer to $\mathcal{C}_d$.

Algorithm 1 Teacher preparation, GVG training, and inference

Require: Training tuples (X_i, y_i, I_i) when a stimulus image I_i exists; otherwise (X_i, y_i); dataset prompt q_d and complete candidate set $\mathcal{C}_d$

A. Adapt and freeze the AVDE teacher (training data only)

for all paired visual samples (X_i, I_i) **do**
  align $A_\psi(X_i)$ with the frozen CLIP image feature $E_{img}(I_i)$
end for
for all clinical samples (X_i, y_i) **do**
  $d_i \leftarrow D_d(y_i)$
  align $A_\psi(X_i)$ with the frozen CLIP text feature $E_{txt}(d_i)$
end for
train the visual decoder G_ψ on paired visual EEG to predict image VQ tokens; freeze (A_ψ, G_ψ)

B. Materialize clinical pseudo-targets offline

for all clinical EEG samples X_i **do**
  $a_i \leftarrow A_\psi(X_i)$; $I_i^p \leftarrow G_\psi(a_i)$
end for

C1. GVG Stage 1 alignment

$I_i^* \leftarrow I_i$ for paired visual EEG; $I_i^* \leftarrow I_i^p$ for clinical EEG
$\ell_i \leftarrow L_d(y_i)$; keep pretrained encoders frozen
update projectors and logit scales using (X_i, I_i^*, ℓ_i)

C2. GVG Stage 2 token prediction (Stage 1 frozen; visual EEG only)

for all paired visual samples **do**
  $h_i \leftarrow F_\theta(X_i)$; $s_i \leftarrow T(I_i)$; update token decoder to predict s_i from h_i
end for

C3. GVG Stage 3 understanding (initialize from best Stage 2)

freeze F_θ and the token predictor, including learned queries and output head
$h_i \leftarrow F_\theta(X_i)$; $\hat{s}_i \leftarrow \arg\max \mathrm{TokProj}(\mathrm{TokDecoder}(Q, h_i))$
$u_i \leftarrow [q_d, \mathrm{Embed}(\hat{s}_i), \mathrm{EEGtoLLM}(h_i)]$
update EEG-to-LLM and the selected MLLM parameters on $\mathrm{MLLM}(u_i) \rightarrow y_i$

D. Inference for one test sample

Require: EEG X, task prompt q_d, and complete candidate set $\mathcal{C}_d$

$h \leftarrow F_\theta(X)$
$z \leftarrow \mathrm{TokDecoder}(Q, h)$; $\hat{s} \leftarrow \arg\max \mathrm{TokProj}(z)$
$e \leftarrow \mathrm{EEGtoLLM}(h)$
$\hat{y} \leftarrow \mathrm{ConstrainedMLLM}([q_d, q_1, \mathrm{Embed}(\hat{s}), q_2, e, q_3]; \mathcal{C}_d)$

Ensure: predicted class $\hat{y}$

A.3 Visual Tokens and Continuous EEG Features

Both native visual interfaces predict $M = 256$ tokens from a vocabulary of $|\mathcal{V}| = 16{,}384$ codes. The token embeddings express EEG through the MLLM's native visual vocabulary, while the continuous EEG prefix provides a direct path from aligned features to the language backbone. Both branches are computed from the same EEG input. The predicted tokens also feed the frozen native image decoder, connecting the task interface to the visual reconstructions in Section E.

A.4 TUSL and Workload Adaptation

TUSL uses 23-channel, 10-second windows for background, seizure, and slowing classification. Workload uses 19-channel, 4-second windows for low/high cognitive workload classification. At 200Hz, their tensors are $23 \times 10 \times 200$ and $19 \times 4 \times 200$. Channel identities are matched by name to the validated dataset order, including an explicit 19-channel mapping for Workload.

Both datasets are excluded from AVDE training and GVG Stage 1/2 pre-training. Each GVG configuration reuses its corresponding upstream initialization and adapts with target-task labels, without new target-task Stage 1/2 training or proxy generation. Checkpoints are selected by validation BAcc with validation-loss tie breaking. The adaptation columns in Table 1 report means over seeds 42, 1, and 2, evaluated on fixed test sets of 41 TUSL and 290 Workload examples. The TUSL partition contains subject overlap between splits.

Table 7: Information visible in each phase. A dash means that the quantity is absent from that computation.

Phase	EEG	Ground-truth label/description	Visual quantity	Role
AVDE alignment training	X_i	$D_d(y_i)$ for clinical EEG	real I_i for paired EEG	learn shared CLIP conditioning
AVDE decoder training	X_i	–	real I_i and its VQ targets	learn EEG-conditioned generation
Proxy materialization	X_i	–	generated I_i^p (output)	create frozen offline target
GVG Stage 1 training	X_i	$L_d(y_i)$ as supervision	real I_i or proxy I_i^p	cross-modal alignment
GVG Stage 2 training	paired X_i	–	$T(I_i)$ from real stimuli	learn EEG-to-token prediction
GVG Stage 3 training	X_i, q_d	y_i as answer target	$\hat{s}_i$ predicted from EEG	learn task response
Test inference	$X, q_d, \mathcal{C}_d$	–	$\hat{s}$ predicted from EEG	output $\hat{y}$

For TUSL/Workload respectively, mean $\pm$ population standard deviation (SD; `ddof=0`) is 73.78 ± 1.24/69.31 ± 2.25 for Janus Trimodal Full, 58.07 ± 4.40/66.67 ± 1.72 for Janus Image-only Full, 70.36 ± 3.04/67.59 ± 5.08 for Janus Trimodal LoRA, and 52.59 ± 1.43/64.94 ± 1.55 for X-Omni Image LoRA. The X-Omni adaptation uses the LoRA-r8 mix/retrieval implementation with 25.38M trainable parameters, distinct from the 170M primary-benchmark configuration. Janus LoRA uses rank 32.

Published TUSL/Workload reference means come from NeuroLM Table 4 (Jiang et al., 2024a); the LaBraM cells use its LaBraM-Base row.

B Taskwise Comparisons and Supporting Analyses

Table 8: Stage-removal comparison across six datasets (BAcc, %). Full corresponds to Table 1. Macro_6 is computed before rounding.

Setting	SEED	SEED-IV	SEED-VII	HMC	TUAB	TUEV	Macro_6
w/o Stage 1	34.60	26.22	14.57	30.58	67.23	20.75	32.32
w/o Stage 2	63.41	25.56	31.36	59.58	79.87	50.76	51.76
Full	**68.70**	**45.35**	**34.45**	**71.51**	**80.08**	**60.20**	**60.05**

Table 9: Taskwise BAcc differences corresponding to Table 2, in percentage points. The macro row averages all six datasets.

Dataset	Image − text	Full − image	Full − text
SEED	+2.44	+11.41	+13.85
SEED-IV	+7.96	+3.33	+11.29
SEED-VII	+7.61	+7.81	+15.42
HMC	+23.29	+5.33	+28.62
TUAB	+0.16	−0.19	−0.03
TUEV	+3.97	+11.82	+15.79
Macro_6	+7.57	+6.59	+14.16

B.1 Stage-Removal Results

Table 8 compares models trained with different subsets of the GVG pipeline. Removing Stage 1 reduces Macro_6 from 60.05% to 32.32%, with especially large drops on HMC and TUEV. This pattern

supports feature alignment as a foundation for the later token and task stages. Stage 2 provides an additional benefit after alignment: omitting token-prediction pretraining reduces Macro_6 to 51.76%, with the largest losses on SEED-IV (19.80 points) and HMC (11.93 points). Together, these results support complementary roles for aligning EEG with pretrained modality spaces and learning its mapping to native visual tokens.

B.2 Reference-System and Adaptation Comparisons

Table 10: Full GVG-Janus mean minus selected reference configurations across the six primary datasets in Table 1, in BAcc percentage points.

Dataset	Full − NeuroLM-XL	Full − LaBraM	Full − LoRA
SEED	+8.36	−4.48	+3.35
SEED-IV	+13.05	−2.28	+0.55
SEED-VII	+4.42	+0.54	+2.35
TUEV	+13.41	−3.89	+7.76
TUAB	+0.39	−1.32	−0.67
HMC	+13.90	−1.35	+9.30

Full exceeds NeuroLM-XL on all six primary tasks and LaBraM on SEED-VII; LaBraM scores higher on the remaining five tasks. Full and LoRA adapt approximately 1.66B and 49M parameters, respectively; deployment costs are reported separately in Table 18.

Using the six primary-dataset BAcc entries displayed to two decimals in Table 1, we pair Full with each NeuroLM variant by dataset. For differences $\Delta_d = \text{BAcc}_{\text{Full},d} - \text{BAcc}_{\text{NeuroLM},d}$, we test a zero mean with a two-sided paired t test, $t = \bar{\Delta}/(s_\Delta/\sqrt{6})$, with five degrees of freedom. Table 11 reports mean effects, unadjusted 95% confidence intervals, and Holm-adjusted p-values across the three comparisons.

Table 11: Exploratory paired comparisons over six primary datasets. Effects and confidence intervals are in BAcc percentage points.

Baseline	Mean Δ	95% CI	$t(5)$	p	p_{Holm}
NeuroLM-B	+9.04	$[2.62, 15.45]$	3.622	0.0152	0.0335
NeuroLM-L	+8.57	$[1.71, 15.43]$	3.211	0.0237	0.0335
NeuroLM-XL	+8.92	$[3.07, 14.77]$	3.921	0.0112	0.0335

This exploratory test summarizes cross-benchmark variation over six partly related datasets. Its t model assumes independent, approximately normal dataset-level differences.

B.3 Continuous EEG Input

Table 12: Input-route comparison across six datasets (BAcc, %). Full corresponds to Table 1; the continuous-only configuration supplies EEG features without a visual-token prefix.

Input route	SEED	SEED-IV	SEED-VII	HMC	TUAB	TUEV	Macro_6
Full (mix)	68.70	45.35	34.45	71.51	80.08	60.20	60.05
Continuous EEG only	66.76	32.34	34.48	68.75	80.78	49.61	55.45

Table 12 gives the taskwise results for the continuous-EEG-only comparison in Section 5.1. This model retains Stage 1/2 pretraining and uses the continuous EEG prefix for task adaptation; the largest gains from adding the visual-token branch occur on SEED-IV and TUEV.

B.4 Teacher Features and Alternative Interfaces

Direct teacher-feature distillation. The frozen AVDE Stage-1 EEG branch supplies its normalized 768-dimensional condition in CLIP space as an auxiliary training target. Starting from the base

Table 13: Teacher, context-length, and initialization controls across six datasets (BAcc, %). Macro_6 is computed from task scores before rounding.

Control	SEED	SEED-IV	SEED-VII	HMC	TUAB	TUEV	Macro_6
Direct teacher-feature distillation	33.62	26.29	14.30	29.84	67.78	22.01	32.30
Length-matched continuous prefix	56.32	32.35	23.04	66.37	81.18	45.36	50.77
Stage 3 only	34.28	26.33	14.97	25.18	58.36	22.03	30.19
Within-sample token shuffle	56.67	31.74	28.12	45.50	75.63	21.86	43.25

EEG encoder and Janus, the student minimizes classification cross-entropy plus cosine distillation through a learned 512-to-768 projection. Inference uses the student's continuous EEG branch. Direct distillation reaches 32.30% Macro_6. Full's stronger performance supports the complete GVG transfer pipeline, which learns Janus-specific alignment and token prediction before task adaptation.

Length-matched continuous prefix. Let U denote the existing continuous EEG sequence after the EEG-to-LLM adapter. Adaptive average pooling produces 256 continuous prefix positions, replacing the 256 visual-token embeddings while retaining U and the prompt layout. Pooling introduces no parameters, and the unused token predictor is frozen. This control retains the upstream initialization and the additional prefix length. At 50.77% Macro_6, it remains 9.28 points below Full, supporting the contribution of the learned visual-token representation beyond the additional prefix positions.

Initialization and token order. Training Stage 3 from the base initialization, omitting GVG Stage 1/2, yields 30.19% Macro_6. This agrees with the stage-removal results in Table 8: task adaptation benefits from both feature alignment and token-prediction pretraining. To examine the organization of the predicted tokens, we permute their positions within each sample while preserving its token multiset; the deterministic permutation is shared between training and evaluation. Macro_6 falls to 43.25%, indicating that the native positional organization helps the model use the predicted visual codes.

Table 14: Compact decoder alternatives and pretrained reference configurations (BAcc, %). Counts denote registered trainable parameters. Dashes indicate unmeasured combinations; Macro_6 requires all six datasets.

Configuration	Train.	SEED	SEED-IV	SEED-VII	HMC	TUAB	TUEV	Macro_6
Janus Retrieval	8.397M	61.27	30.66	23.07	39.95	80.79	34.61	45.06
Janus Retrieval + EEG adaptation	49.491M	67.85	32.74	29.43	49.50	81.17	41.09	50.30
Janus LoRA, pretrained reference	49.491M	65.35	44.80	32.10	62.21	80.75	52.44	56.28
X-Omni Transformer	–	–	25.00	–	41.76	71.46	26.86	–
GVG-X-Omni, pretrained reference	170M	60.73	30.25	27.17	61.20	74.45	44.15	49.66

Retrieval Transformers replace the language decoder with compact trainable modules. The two Janus variants freeze or adapt the continuous EEG path, respectively. EEG adaptation raises Macro_6 from 45.06% to 50.30%, showing that the compact decoder benefits from task-specific feature adjustment. At the same registered trainable-parameter count, pretrained Janus LoRA reaches 56.28%, with particularly large gains on HMC and TUEV. A similar pattern holds for X-Omni, whose pretrained reference is stronger on all four measured tasks. These comparisons support retaining the pretrained multimodal language interface for performance across heterogeneous EEG tasks.

C Visual-Target Controls and Sensitivity

All visual-target controls use Janus with the combined continuous EEG and visual-token interface. Table 15 expands the aggregate results analyzed in Section 5.2.

Random and globally shuffled targets alter the clinical proxy content or pairing, whereas the prototype supplies one image per class. Within-class shuffling uses a fixed-point-free permutation within each clinical dataset–label pool. The prototype's largest gaps to Full occur on SEED-IV and TUEV, showing that the benefit of EEG-conditioned targets varies across the jointly trained tasks. Both

Table 15: Per-task visual-target results (BAcc, %) corresponding to the aggregate comparison in Table 3.

Target recipe	SEED	SEED-IV	SEED-VII	HMC	TUAB	TUEV	Macro_6
Full: EEG-conditioned	68.70	45.35	34.45	71.51	80.08	60.20	60.05
Class prototype	68.31	33.58	33.53	71.68	80.89	45.48	55.58
Random image	63.07	30.03	30.00	48.65	77.00	34.31	47.18
Within-class shuffle	44.32	30.24	20.41	51.43	76.71	19.63	40.46
Dataset-global shuffle	40.27	30.54	27.69	36.55	73.36	29.62	39.67

shuffling controls reduce scores on all six tasks, consistent with the role of sample-level pairing discussed in Section 5.2.

Table 16: Gaussian target-corruption comparison (BAcc, %). The clean Full configuration corresponds to Table 1; σ denotes the target-noise level.

Noise level	SEED	SEED-IV	SEED-VII	HMC	TUAB	TUEV	Macro_6
Full (clean)	68.70	45.35	34.45	71.51	80.08	60.20	60.05
$\sigma = 0.1$	65.60	32.50	33.37	68.47	80.09	48.61	54.77
$\sigma = 0.3$	56.94	30.41	30.93	49.82	62.62	34.40	44.19
$\sigma = 0.5$	38.09	30.01	20.60	27.17	50.11	19.84	30.97

Gaussian noise perturbs the generated proxy images while keeping EEG recordings and their image assignments fixed. Scores decrease on all six tasks as σ increases from 0.1 to 0.5; relative to clean proxies, the largest drops at $\sigma = 0.5$ occur on HMC and TUEV. This complements the shuffling controls by showing that the visual content of correctly paired targets also matters for learning.

C.1 Seed-Wise Main-Model Results

For each model configuration and seed, one jointly trained model is evaluated across SEED, SEED-IV, SEED-VII, HMC, TUAB, and TUEV. Table 17 groups the recorded seed-wise scores by configuration and dataset. The six Full rows supply the corresponding main-table means; their aggregate is the six-task Macro_6 defined in Equation (6). Means and sample standard deviations are computed before display rounding, with SD denominator $n - 1$.

Table 17: Seed-wise main-model results (BAcc, %).

Configuration	Dataset	Seed 42	Seed 1	Seed 2	Mean $\pm$ SD
Janus Trimodal Full	SEED	68.92	68.41	68.78	68.70 ± 0.26
	SEED-IV	45.49	45.68	44.89	45.35 ± 0.42
	SEED-VII	34.87	34.16	34.32	34.45 ± 0.37
	HMC	71.61	71.35	71.56	71.51 ± 0.14
	TUAB	80.20	80.26	79.77	80.08 ± 0.27
	TUEV	59.81	60.33	60.46	60.20 ± 0.34
Janus Trimodal LoRA	SEED-IV	45.31	44.20	44.90	44.80 ± 0.56
Janus Image Full	SEED-IV	41.16	42.31	42.59	42.02 ± 0.76
X-Omni Image LoRA	SEED-IV	32.04	30.05	28.67	30.25 ± 1.70

Across the six Full tasks, sample SDs range from 0.14 to 0.42 percentage points, describing variation over three training seeds on fixed test partitions.

C.2 End-to-End Resource Accounting

Frozen components contribute to inference cost alongside the parameters updated during adaptation. Operation counts are expressed in billions of floating-point operations (GFLOPs). In the single-sample HMC setting of Table 18, Janus and X-Omni have 3.06 and 5.87 times NeuroLM-XL's latency and 3.31 and 10.09 times its GFLOPs.

Table 18: Single-sample HMC inference on an A100-80GB in fp16. Total counts include both trainable and frozen components. Memory (Mem.) and latency (Lat.) are measured per sample; floating-point operations are reported in billions (GFLOPs).

Model	Total	Train.	Mem.	Lat.	GFLOPs
GVG-Janus	2.15B	1.66B	4.6GB	0.095s	1740
GVG-X-Omni	11.0B	0.17B	23.1GB	0.182s	5307
NeuroLM-XL	1.74B	1.65B	3.6GB	0.031s	526

The Janus inference profile contains 2,146,154,613 total parameters, including the multimodal backbone and frozen tokenizer components. It records 4.428GB of parameter memory, 4.601GB peak inference memory, median/mean latency 0.0951/0.1002 seconds, and a 603-position prefill sequence in the stated HMC setting. The reported 1,740.15 GFLOPs comes from `calflops`.

The profiler records 1,657,907,200 trainable parameters for Janus under full adaptation, rounded to 1.66B in Table 18 and 1.7B in Table 1.

D Detailed Clinical Representation Results

The clinical geometry and correspondence analyses compare SHA-256-pinned Full and Stage-1-ablated final checkpoints on identical deterministic subsets: HMC 2,500; TUAB 1,000; and TUEV 2,829, with at most 500 examples per class. The physiology analysis uses a fixed 750-window HMC subset and the frozen Full representation. The group-level procedures are specified below.

Table 19: Clinical label-residual retrieval: Top-5, mean average precision (mAP), and paired cosine similarity. Parentheses contain one-sided sample-level label-preserving permutation p-values. Cluster-aware evidence is reported separately in Table 4.

	Top-5		mAP		Paired cosine	
Dataset	Full	w/o S1	Full	w/o S1	Full	w/o S1
HMC	.0144 (.006)	.0100 (.478)	.01836 (.001)	.01436 (.205)	.01984 (.001)	−.00219 (.990)
TUAB	.0110 (.395)	.0120 (.237)	.01356 (.517)	.01354 (.476)	−.00408 (.976)	.00023 (.438)
TUEV	.01449 (.005)	.01131 (.326)	.01810 (.001)	.01495 (.214)	.02973 (.001)	−.00025 (.617)

Clinical groups are identified from held-out filenames: HMC subject ID (26 groups), TUAB patient prefix (249 groups), and TUEV source-recording prefix (155 groups). We average paired-cosine differences within each group, resample group means for bootstrap intervals, and use group-level sign flips for a two-sided null test. Holm correction covers the six-dataset correspondence family: three clinical and three paired-visual datasets. HMC and TUEV each have $p < 10^{-4}$ and remain significant after correction; TUAB shows no significant change ($p = .232$). The 95% group-bootstrap intervals are $[0.0180, 0.0272]$ for HMC, $[0.0248, 0.0421]$ for TUEV, and $[-0.0095, 0.0024]$ for TUAB.

D.1 Residual Estimand and Cluster Weighting

For modality m and checkpoint v, let $h_i^{m,v}$ be the representation of sample i and let $\mu_c^{m,v}$ be its class centroid over the analyzed subset. Residual vectors are

$$r_i^{m,v} = \text{normalize}\left(h_i^{m,v} - \mu_{y_i}^{m,v}\right), \qquad c_i^v = \left\langle r_i^{\text{EEG},v}, r_i^{\text{visual},v} \right\rangle. \tag{7}$$

Class centroids are computed post hoc for each checkpoint and modality. The cosine then measures whether EEG and proxy features deviate from their respective category prototypes in corresponding directions.

The clinical comparison uses the full and Stage-1-ablated checkpoints. For group g, define

$$d_g = \frac{1}{|g|} \sum_{i \in g} \left(c_i^{\text{full}} - c_i^{\text{w/o S1}}\right), \qquad \Delta_{\text{group}} = \frac{1}{G} \sum_{g=1}^{G} d_g. \tag{8}$$

Each subject, patient, or recording contributes one equally weighted group mean. Table 20 reports the group sizes and the sample- and group-weighted differences.

Table 20: Clinical residual-analysis sampling units and weighting. Counts describe the analyzed subsets. Sample and group means use window and subject/patient/recording weights, respectively.

Dataset	Windows	Groups	Windows/group	Sample mean Δ	Group mean Δ
HMC	2500	26	51–138	0.02203	0.02261
TUAB	1000	249	1–18	-0.00431	-0.00365
TUEV	2829	155	1–167	0.02998	0.03348

The direction of each dataset's effect is unchanged by group weighting; for example, TUEV's difference rises from 0.02998 to 0.03348. The primary analysis uses the group mean in Equation (8), giving subjects or recordings with fewer selected windows equal representation.

Bootstrap intervals resample group means 10,000 times. The two-sided sign-flip test uses 100,000 random sign assignments to the group differences, comparing the absolute randomized mean with the observed absolute mean.

D.2 HMC Physiology-Probe Protocol

The physiology analysis uses a deterministic HMC subset: 150 samples per sleep stage selected by path-hash order. The 750 windows cover 26 subjects. Both endpoints and probe settings were fixed before extracting the physiological targets. The displayed t-SNE coordinates are fixed across all three colorings.

We use the first 16 seconds from channels F4, C4, O2, and C3 at 200Hz. Welch spectra use 400-sample Hann windows, 200-sample overlap, and constant detrending. Relative delta power is the integrated 0.5–4Hz power divided by integrated 0.5–30Hz power. Spectral entropy is the Shannon entropy of normalized power spectral density (PSD) bin probabilities over 0.5–30Hz, divided by the logarithm of the bin count. Both quantities are calculated per channel and then averaged across the four channels.

Five-fold GroupKFold partitions subjects. Within each training fold, embeddings and targets are centered by their sleep-stage means; principal component analysis (PCA) and component standardization are fitted only on training subjects. A fixed-regularization ridge probe with no intercept predicts target residuals, and the training-fold stage mean is added back. The same preprocessing and regularization settings are used for both endpoints and all null draws, with the exact configuration retained in the analysis artifacts. Let E_g^{probe} and E_g^{stage} be the mean squared errors for subject g. The reported score is

$$S_{\mathrm{phys}} = 1 - \frac{\sum_{g=1}^{G} E_g^{\mathrm{probe}}}{\sum_{g=1}^{G} E_g^{\mathrm{stage}}}, \qquad G = 26. \tag{9}$$

Each subject receives equal weight in the probe and baseline error averages. The score measures recovery of physiological variation beyond the training-fold sleep-stage means.

For each of 499 null draws, the two targets are shifted together by a nonzero circular offset within each subject–stage group, ordered by the original window index. This preserves the target distribution in each group while changing its correspondence with the embeddings. Eight singleton windows remain unchanged. The same folds and probe recipe are applied to every draw. Plus-one Monte Carlo p-values are Holm-adjusted across the two targets, separately from the six-dataset correspondence-test family.

Confidence intervals use 5,000 paired subject-bootstrap draws of the out-of-fold probe and baseline errors, holding the folds and fitted probes fixed. Relative-delta baseline/probe MSE is 0.02267838/0.01125464; spectral-entropy baseline/probe MSE is 0.01943035/0.01188691.

The MSE reductions are 50.4% for relative delta power (95% subject-bootstrap CI 43.0–56.5%) and 38.8% for spectral entropy (31.8–45.0%). Both exceed all 499 cyclic-shift null scores (one-sided $p = .002$; Holm-adjusted $p = .004$ across the two targets). Stage-centered out-of-fold Spearman correlations are .704 and .631, respectively.

E VISUAL RECONSTRUCTION

Table 21: Reconstruction results on paired visual datasets. Higher PSNR/SSIM and lower LPIPS are better.

Dataset	Method	PSNR	SSIM	LPIPS
SEED	AVDE	10.17	**.3162**	**.5072**
	GVG-X-Omni	**10.69**	.2943	.7201
	GVG-Janus	10.52	.2989	.6372
SEED-VII	AVDE	11.78	.3496	.6102
	GVG-X-Omni	**14.41**	**.4335**	.6700
	GVG-Janus	13.48	.4149	**.5751**

On SEED, both GVG variants improve PSNR over AVDE, while AVDE leads SSIM and LPIPS. On SEED-VII, Janus improves all three metrics over AVDE, and X-Omni achieves the highest PSNR and SSIM. These metrics describe pixel agreement, structural similarity, and learned perceptual distance; the content assessment below examines category, color, and layout directly.

E.1 MODEL-BASED CONTENT ASSESSMENT

We evaluate the main GVG-Janus model in Table 1 without fine-tuning its pretrained visual decoder. The sample contains 100 test-set reconstructions from each of SEED, SEED-IV, and SEED-VII (300 total). Within each dataset, we sort the generated image files by filename, shuffle the ordered list using random seed 42, and select the first 100 entries. The selected samples are evaluated in filename order. Each reference stimulus frame captures one view of the stimulus associated with a four-second EEG window. Each sample contains three panels: the original stimulus (GT), its VQ encoder–decoder reconstruction (GT-tokens), and the reconstruction from EEG alone (Pred). GPT-6 is instructed to evaluate only Pred–GT correspondence, excluding the middle panel from scoring.

We report three dimensions from the image-level assessment: semantic category, color, and spatial layout, each scored by the judge on an integer scale from 1 to 5. These dimensions describe coarse visual correspondence. We report arithmetic means of the per-image scores within each dataset and over all 300 samples. With equal sample counts, the pooled mean also equals the unweighted mean of the three dataset means.

Table 22: GPT-6 content ratings of GVG-Janus EEG-based reconstructions. Each dimension uses integer scores from 1 to 5; entries are mean scores, and higher is better. Only correspondence between Pred and GT is scored.

Dataset	n	Semantic category	Color	Spatial layout
SEED	100	2.48	3.08	2.92
SEED-IV	100	1.82	2.38	2.41
SEED-VII	100	2.07	2.28	2.44
Overall	300	2.12	2.58	2.59

Color and layout receive higher mean scores than semantic category on all three datasets (Table 22). The pooled color and layout means fall between weak and moderate correspondence under the rubric, while semantic category remains near weak correspondence. This pattern indicates partial recovery of broad appearance and composition, with limited recovery of semantic content. These ratings complement PSNR, SSIM, and LPIPS by characterizing coarse visual content.

E.1.1 EVALUATION PROMPT

An abridged version of the evaluation prompt is shown below for the three reported dimensions.

> Evaluate the quality of EEG-based image reconstructions on an image-by-image basis.
> Each sample is a three-panel image:

- Left (GT): the original visual stimulus viewed by the participant.
- Middle (GT-tokens): the original image reconstructed through the VQ encoder and decoder, serving as an upper-bound reference.
- Right (Pred): the image reconstructed from EEG signals alone.

Evaluate only the correspondence between Pred and GT. The middle panel must not contribute to the scores.

Focus on coarse-grained visual correspondence rather than exact instance-level reconstruction. Assign an integer score from 1 to 5 for each of the following dimensions:

1. Semantic category
   Whether Pred preserves the coarse semantic category or scene attributes of GT, such as natural scenes, people, food, textures, or indoor versus outdoor environments.
2. Color
   Whether Pred matches the dominant colors and overall color tone of GT.
3. Spatial layout
   Whether Pred preserves the arrangement of major visual regions, foreground–background structure, light–dark patterns, and overall composition of GT.

Use the following five-point scale:

1. No meaningful correspondence to GT.
2. Weak correspondence in low-level cues, such as color, texture, or coarse layout.
3. Moderate coarse-grained correspondence, with clear semantic or structural errors remaining.
4. Strong coarse-grained correspondence in category, color, or layout, although instance-level details are not faithfully recovered.
5. High consistency with GT; the original content is largely recognizable.

Process the images in filename order and provide one table row per image using the following columns:

Filename | Semantic category | Color | Spatial layout

After evaluating all images, report the mean score for each dimension.

F LIMITATIONS

GVG is constrained by the limited scale of available EEG training data, especially paired EEG–visual recordings, and the inherently low signal-to-noise ratio of EEG. These factors make learning detailed EEG-to-image correspondences challenging. Current reconstructions recover aspects of color and coarse layout, with limited semantic and instance-level fidelity. Improving reconstruction quality is therefore central to progressing from a shared interface toward reliable, unified EEG understanding and visual generation. Larger and more diverse paired datasets, together with representations that better tolerate EEG noise, are promising directions. The sensitivity of task predictions to prompt wording and answer-option ordering also requires systematic evaluation.