Cognitron: Thoughts from EEG via Caption-Guided Latent Alignment

Pranshu Jain Nilesh Mishra Sooryakiran B Mohsina Bilal

Advanced Deep Learning & Computer Vision
Project Report, 7 May 2025

Outline

Introduction

- **Motivation:** Decode visual content from neural signals for neuroscience, BCI, assistive tech.
- **fMRI vs. EEG:**
  - fMRI: high spatial resolution but low temporal resolution, costly.
  - EEG: high temporal resolution, portable, low cost, but noisy and low spatial resolution.
- **Challenge:** Direct EEG-to-image mapping $\rightarrow$ high-dimensional pixel space, heavy computation.
- **Key Insight:** Use natural-language captions as intermediate semantic space for efficient alignment.

Background

Neural Decoding Evolution

- **fMRI-based:** Kamitani & Tong (2005), Miyawaki *et al.* (2008), Lin *et al.* (2022).
- **EEG-based:** Early classification → modern deep learning (CNNs, transformers).
- **Generative Models:** GANs, VAEs, diffusion for image synthesis.

Literature Review I

- **DreamDiffusion** Introduces temporal masked signal modeling (TMSM) to pre–train an EEG encoder in a self-supervised fashion, then adapts a frozen Stable Diffusion model via CLIP alignment to generate high-fidelity images directly from EEG. Addresses EEG noise, limited information content, and inter-subject variability.
- **Guess What I Think (GWIT)** Employs a lightweight ControlNet adapter on top of a latent diffusion backbone to condition image generation on raw EEG, minimizing preprocessing and training overhead. Demonstrates lower LPIPS scores and real-time feasibility on benchmark EEG datasets.
- **NeuroGAN** Uses an attention-augmented GAN generator to focus on informative EEG channels, paired with a pretrained image classifier for perceptual and class-specific loss. Achieves state-of-the-art Inception Scores and Class Diversity Scores on the ThoughtViz dataset.

Literature Review II

- **MindDiffuser** Proposes dual semantic and structural diffusion pathways to control both content and composition in EEG-to-image reconstruction, improving visual fidelity and layout consistency.
- **BrainVis** Segments EEG into functional units and applies self-supervised learning to align time-frequency embeddings with coarse and fine-grained CLIP features, achieving strong performance with only 10% of typical training data.

Dataset Overview & Acquisition

Dataset Overview

Subjects	6 participants
Image Categories	40 ImageNet classes
Total Images	2,000 stimuli
Trials / Image	1 trial (0.5 s, 440 samples)

Acquisition Details

- **EEG hardware:** 128 scalp channels
- **Sampling rate:** 1 kHz
- **Preprocessing:** Band-pass filters at 5–95 Hz, 14–70 Hz, 55–95 Hz

Data Splits

Training / Validation / Test

	# samples	% of total
Train	7,959	67%
Validation	1,994	17%
Test	1,987	16%

Data Description: Relevance

Two–Stage Mapping

$$\underbrace{\text{EEG} \xrightarrow{\text{Encoder}} \text{Caption Embedding}}_{\text{semantic alignment}} \xrightarrow{\text{Decoder}} \underbrace{\text{Reconstructed Image}}_{\text{generation from caption}}$$

Why This Dataset?

- **Temporal richness:** Millisecond-level EEG vs. multi-second fMRI
- **Controlled stimuli:** Known image–caption pairs for precise supervision
- **Scalability:** Hundreds of trials per subject enable robust learning

Methodology Overview

1. **Model Architecture**
2. **Data Embedding Preparation**
3. **Training & Inference**

Model Architecture: Original Pipeline

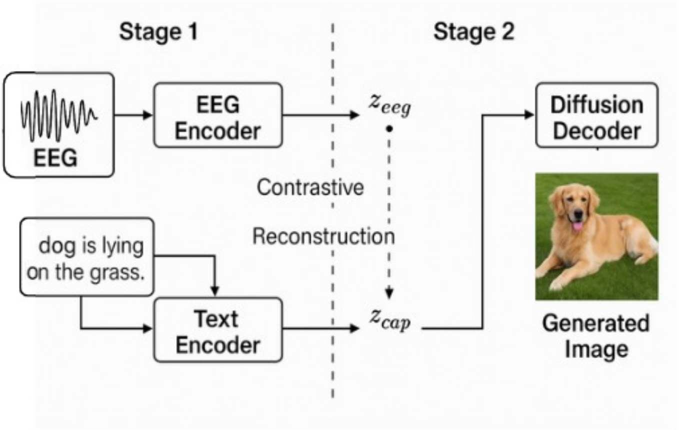


Figure: EEG signals → Encoder → Latent space (contrastive/MSE) → decoder → diffusion → Image.

Two-Stage EEG-to-Image Reconstruction I

Stage 1: Semantic Alignment

- **EEG Encoder:** Temporal–spatial CNN + channel-wise attention + MLP $\Rightarrow z_{\text{eeg}} \in \mathbb{R}^{512}$.
- **Text Encoder:** Pretrained BLIP/CLIP text encoder produces $z_{\text{cap}} \in \mathbb{R}^{512}$ from the image caption.
- **Joint Losses:**
  - *Contrastive (NT-Xent)* pulls true $(z_{\text{eeg}}, z_{\text{cap}})$ pairs together.
  - *Reconstruction (MSE)* penalizes $\|z_{\text{eeg}} - z_{\text{cap}}\|^2$.
- **Outcome:** After training, EEG alone yields embeddings $z_{\text{eeg}} \approx z_{\text{cap}}$ even on unseen trials.

Two-Stage EEG-to-Image Reconstruction II

Stage 2: Image Generation

- **Diffusion Decoder:** Frozen Stable Diffusion v1-5 model, conditioned on a 512-d embedding.
- **Inference:** Inject z_{eeg} into the U-Net's cross-attention in place of the text embedding.
- **Result:** High-fidelity "thought-to-image" output that semantically matches the original stimulus.

Data Embedding Preparation I

Caption Embeddings

- Generate natural-language descriptions of each image using the pretrained BLIP model.
- Embed each caption with the Universal Sentence Encoder (USE) to obtain a fixed 512-dim vector.
- Apply ℓ_2-normalization to place all caption vectors on the unit hypersphere.

Data Embedding Preparation II

EEG Embeddings

- Preprocess raw EEG signals: band-pass filter (5–95 Hz), drop first 20 ms, crop to 440 time-points.
- Use a SimCLR-style augment-and-contrast pretraining on a lightweight MLP encoder:
  - Create two augmented views per trial (e.g. time-masking, noise).
  - Train with NT-Xent loss (temperature $\tau = 0.07$) to pull same-trial views together.
- Extract a 768-dim feature and ℓ_2-normalize for downstream alignment.

EEG-to-Caption Encoder: Methodology I

Input & Output

- **Input:** Normalized 512-dim EEG embedding vector
- **Output:** 512-dim project EEG embedding vector

Model Layers

1. **Attention (Feature Gating)** Computes attention weights $\alpha \in \mathbb{R}^{512}$ via 1×1 Convs and softmax, then scales each input feature: $\hat{x}_i = \alpha_i x_i$.
2. **Feedforward Decoder** Three fully-connected layers with ReLU and dropout:
   - Linear($512\to1024$) $\to$ ReLU $\to$ Dropout(0.1)
   - Linear($1024\to1024$) $\to$ ReLU $\to$ Dropout(0.1)
   - Linear($1024\to512$)

EEG-to-Caption Encoder: Methodology II

Training & Metrics

- **Loss:** $\mathcal{L} = \lambda \|\hat{z} - z\|_2^2 + (1 - \lambda)\big(1 - \cos(\hat{z}, z)\big)$
- **Optimizer:** Adam, $\mathrm{lr} = 1 \times 10^{-3}$
- **Scheduler:** Cosine-annealing learning-rate
- **Regularization:** Dropout(0.1) to prevent overfitting
- **Evaluation:** Mean cosine similarity on validation set (target greater than 0.2)

EEG Decoder: Captions from EEG-Derived Embeddings

1. Build Reference Corpus

- Load BLIP captions and any additional caption sources
- Merge and de-duplicate into a single array of reference captions

2. Embed Normalize Corpus

- Convert each reference caption into a 512-dim BLIP embedding
- ℓ_2-normalize all embeddings for cosine similarity

3. Predict BLIP Embedding

- Feed new EEG trial into the trained encoder 512-dim embedding

4. Retrieve Top-k Captions

- Compute cosine similarities between the predicted embedding and all reference embeddings
- Sort and select the k highest-scoring captions

Caption Decoder: Methodology I

Input & Output

- **Input:** Blip Caption embeddings $\in \mathbb{R}^{B \times 512}$.
- **Output:** Generated token of natural-language caption.

Model Layers

1. **EmbeddingProjector:** Linear
2. **BART-base w/ LoRA:**
   - LoRA adapters have been utilized for fine-tuning
   - Remaining BART weights frozen.
3. **Beam Search:** To get next word prediction as a set of words num_beams $= 4$, length penalty $= 0.6$.

Caption Decoder: Methodology II

Training & Metrics

- **Loss:** Cross-entropy on true vs. predicted tokens

$$\mathcal{L} = -\sum_{t=1}^{L} \log p(y_t \mid y_{<t}, \mathrm{emb})$$

- **Optimizer:** AdamW $(\mathrm{lr} = 1\mathrm{e}{-4},\ \mathrm{wd} = 0.01)$
- **Scheduler:** StepLR$(\mathrm{step} = 5, \gamma = 0.5)$
- **Epochs:** 15
- **Val BLEU:** ≈ 28.5 (corpus BLEU on validation set)

Image Reconstruction via Stable Diffusion v1-5 I

Model Overview

- **Stable Diffusion v1-5** is a latent diffusion model pretrained on large-scale image–text pairs.
- Internally composed of:
  - A frozen VAE for mapping images latents
  - A U-Net denoiser with text-conditioning via cross-attention

Conditioning with Learned Embeddings

- $\mathbf{z}_{\mathrm{cap}}$ (caption decoder output) or $\mathbf{z}_{\mathrm{eeg}}$ (EEG encoder output) are injected in place of the usual text embeddings.
- Cross-attention layers attend to these vectors at every diffusion timestep.
- Since both embeddings occupy the same 512-D semantic space, the U-Net can interpret either for image synthesis.

Image Reconstruction via Stable Diffusion v1-5 II

Reconstruction Pipeline

1. **Latent Initialization:** Start from pure Gaussian noise in VAE latent space.
2. **Denoising Loop:** Iteratively apply the conditioned U-Net to remove noise across T timesteps.
3. **Decode to Image:** Use the VAE decoder to transform final latent into a 256×256 (or 512×512) RGB image.

Discussion I

EEG Encoder

Maps preprocessed EEG trials into a fixed-length semantic embedding. This vector captures the subject's visual or imagined content in a 512-dim latent space.

Caption Decoder

Takes the EEG-derived embedding and generates a natural-language caption. Demonstrates how well the embedding encodes semantic information about the stimulus.

Image Decoder

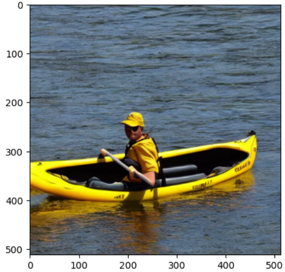
Feeds the same EEG-derived embedding into a frozen diffusion (or GAN) model to reconstruct an image. Shows the visual fidelity of "thought-to-image" generation from EEG.

Discussion II

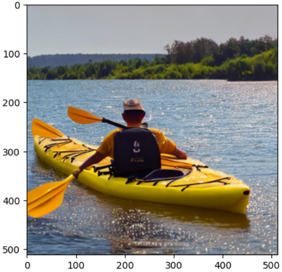
Evaluation

Visualize images generated from EEG embeddings against images produced from true caption embeddings. Assess semantic and perceptual alignment to see how closely EEG-based reconstructions match caption-based ones.

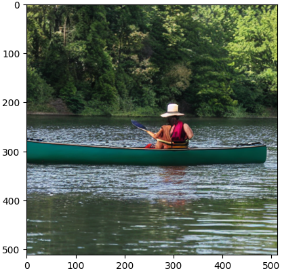
Decoder Result Comparison



(a) Original Image from Dataset



(b) Caption Decoder Output



(c) EEG Decoder Output

Figure: Comparison between the ground truth image, the image reconstructed from the BLIP caption embedding, and the image reconstructed from the EEG embedding.

Caption Generation Results

Original Caption

there is a man in a yellow kayak paddling through the water

Caption Decoder Output

there is a man in a kayak paddling through the river

EEG Decoder Output

there are two people in a canoe on the water

Results & Discussion

- EEG embeddings cluster by visual semantics.
- Caption reconstruction quality: human preference tests.
- Sample reconstructions show semantic fidelity.

Conclusion

- Introduced Cognitron: caption-mediated EEG$\rightarrow$image framework.
- Efficient semantic alignment reduces computational overhead.
- Demonstrated feasibility on public dataset.
- Future: quantitative evaluation, cross-subject generalization, imagined imagery.

References I

- Y. Kamitani and F. Tong, "Decoding the visual and subjective contents of the human brain," *Nat. Neurosci.*, vol. 8, no. 5, pp. 679–685, 2005.
- Y. Miyawaki *et al.*, "Visual image reconstruction from human brain activity," *Neuron*, vol. 60, no. 5, pp. 915–929, 2008.
- H. Lin *et al.*, "Dynamic natural scene reconstruction from human brain activity using GANs," *Nat. Commun.*, vol. 13, p. 420, 2022.
- A. Tagliasacchi *et al.*, "DreamDiffusion: High-Resolution image generation from EEG using latent diffusion," arXiv:2303.12548, 2023.
- C. Feng *et al.*, "MindDiffuser: Controlled image reconstruction from human brain activity," arXiv:2303.06540, 2023.
- M. Angrick *et al.*, "EEG2Speech: Direct reconstruction of audible speech from EEG," arXiv:2106.01933, 2021.
- L. S. Luigi, "EEG_Image_CVPR_ALL_subj," 2023. [Online]. Available: `https://huggingface.co/datasets/luigi-s/EEG_Image_CVPR_ALL_subj`
- E. Lopez *et al.*, "Guess What I Think: EEG-to-Image Generation with Latent Diffusion," *ICASSP*, 2025.

References II

E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2022. [Online]. Available: https://arxiv.org/abs/2106.09685

R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis With Latent Diffusion Models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10684–10695.