



KAUST Final Report

EEG-to-Image Reconstruction

Abdulaziz Alkhateeb **Amjad Alqasemi** **Husain Althagafi**

Abdulellah Mojallad **Musa Ibn Rashid**

Mentor: **Fares Fourati**

Abstract

Reconstructing visual experiences from EEG could enable assistive communication, cognitive assessment, and human–AI interaction, but the modality's low signal-to-noise ratio, coarse spatial resolution, and inter-/intra-subject variability make direct image decoding unstable. We introduce a **novel text-guided diffusion framework** that, reframes EEG-to-image reconstruction as a **semantic guidance problem**. In our method, an EEG encoder is aligned with a text retrieval backbone to generate class-level prompts, which then condition Stable Diffusion to produce reconstructions. This design leverages EEG's strength in encoding high-level semantics, avoiding brittle pixel-level mappings, and yields reconstructions that are both **more coherent and category-faithful**. To benchmark performance, we reproduced representative baselines (DreamDiffusion, NICE+CLIP, ATMS+Diffusion, Blur Prior) on AllJoined1 and THINGS-EEG-2, against which our approach demonstrates clear improvements. Finally, we report methodological insights—including negative results (spectrograms, topographic maps) and effective training strategies (supervised warm-start, contrastive learning, knowledge distillation, multi-label KL)—to ensure reproducibility and guide future work.

1 Introduction

1.1 Overview

Our project targets the challenging task of **EEG-to-image reconstruction**, aiming to generate accurate visual representations from human EEG signals. We began with the *AllJoined1* dataset [1], hypothesizing that its high quality multi-subject recordings could help our model outperform existing benchmarks.

1.2 Datasets

1.2.1 AllJoined1

The *AllJoined1* dataset [1] contains EEG recordings from 8 subjects with 64 channels, sampled at a sequence length of 334. It was initially selected for its broader subject coverage and claimed signal quality. However, baseline reproduction revealed underperformance and instability across models. We hypothesized that noise, preprocessing artifacts, and variability both *within* and *across* subjects degraded performance.

1.2.2 THINGS-EEG-2

Due to reproducibility issues with AllJoined1, we adopted the *THINGS-EEG-2* dataset [2], a widely used benchmark with consistent preprocessing, experimental controls, and robust baseline availability. This dataset provided a reliable platform for evaluating models and enabled cross-comparison with prior work.

Table 1: Dataset Comparison: AllJoined1 vs. THINGS-EEG-2

Property	AllJoined1	THINGS-EEG-2
Subjects	8	10
Channels	64	64
Sequence length	334	250
Preprocessing	Custom	Standardized
Performance stability	Low	High

2 Baseline Reproduction

2.1 DreamDiffusion

Our initial baseline was *DreamDiffusion* [3], which uses a Masked Autoencoder (MAE) for EEG feature extraction and a diffusion model for image generation. Reproduction attempts failed due to unreleased pre-trained weights and incomplete dataset specification. Training from scratch produced random-noise outputs.

2.2 Human-Aligned NICE + CLIP

Next, we built upon the code base from [4], which combines the NICE EEG encoder with CLIP-based contrastive learning. We reproduced only the non-human-aligned version of their pipeline; the full human-aligned model was not implemented.

In our experiments, we reused their codebase and integrated a high-quality NICE EEG encoder backbone with CLIP, aligning EEG features with image embeddings. Using the THINGS-EEG-2 dataset, we performed retrieval tasks to evaluate performance. Our reproduced pipeline achieved retrieval scores closely matching those reported in the original work, confirming the reproducibility of the method and providing a reliable benchmark for subsequent experiments.

2.3 EEG Embeddings with Guided Diffusion Baseline (ATMS + CLIP + Diffusion U-Net)

The primary reproduced baseline was from [5], featuring the ATMS EEG encoder (channel-wise Transformer + temporal-spatial convolutions), CLIP image alignment, and a two-stage image generation (VAE + diffusion U-Net). In our work, we did not incorporate the VAE component. The VAE branch in the original design is intended to preserve low-level structural features of the visual stimulus. While this can refine reconstruction fidelity, it does not alter the fundamental flow of the pipeline. Given our focus on validating and improving the core EEG-to-diffusion pathway, we prioritized stabilizing the ATMS encoder and the diffusion U-Net alignment with CLIP embeddings, leaving the VAE integration as potential future enhancement. Reproduction on THINGS-EEG-2 closely matched the reported results for the high-level pipeline, establishing a strong benchmark for subsequent experimentation.

2.4 Uncertainty-Aware Blur Prior (UBP)

Recognizing the significant impact of inter- and intra-subject variability in EEG decoding, we incorporated insights from "Bridging the Vision–Brain Gap with an Uncertainty-Aware Blur Prior (UBP)" [6]. This method explicitly addresses variability by introducing a blur-based uncertainty mechanism during training.

UBP achieves **top-1 = 50.9%** and **top-5 = 79.7%** retrieval accuracy in the zero-shot brain-to-image task under intra-subject conditions on THINGS-EEG—surpassing previous SOTA by 13.7% and 9.8% respectively. On inter-subject evaluation, though accuracy drops due to inherent variability, these results are still markedly higher than prior approaches, highlighting the strength of modeling uncertainty via blur augmentation.

2.5 Exploratory Baselines Beyond Literature

In addition to reproducing published methods, we experimented with several exploratory architectures that were not directly derived from existing papers. Notably, we evaluated the *ENIGMA* model on the AllJoined1.6M dataset. The motivation was to test whether scaling to a larger EEG dataset and leveraging a more complex encoder would yield better representations. While these trials did not reach the retrieval performance of the paper-based baselines, they provided useful diagnostic insights into dataset quality, preprocessing limitations, and the capacity of alternative model architectures. These experiments emphasized the importance of both reproducibility and theoretical grounding when selecting baselines for EEG-to-image reconstruction.

3 Proposed Method

3.1 Motivation and Approach

EEG-to-image reconstruction is fundamentally hindered by the noisy and unstable nature of brain signals. Directly mapping raw EEG features into images often produces inconsistent or low-quality results. Inspired by the strong retrieval performance of the BluePrior framework, our objective was to **transfer retrieval strengths into reconstruction**. Rather than treating EEG as a denoising challenge, we reformulated the problem into a **text-guided generative modeling task**, where reliable text cues derived from EEG signals could direct the diffusion model during image construction.

3.2 Pipeline Overview

Our proposed method builds upon and extends the Blur Prior codebase. The pipeline is composed of three main components:

1. **EEG Encoding:** Raw EEG signals are processed through specialized backbones for embedding extraction.
2. **Dual-Backbone Representation:** We employ separate pathways for image and text guidance, ensuring complementary strengths in alignment.
3. **Text-Guided Reconstruction:** Retrieved text classes are used as conditioning prompts in a diffusion model, enabling coherent and semantically accurate reconstructions.

3.3 Dual-Backbone Architecture

- **Image Backbone (ViT-H/14):** We adopted the Vision Transformer (ViT-H/14) for EEG-to-image embedding due to its proven robustness in high-dimensional representation learning. This encoder provided the structural alignment required for visual feature extraction.
- **Text Backbone (ResNet-50):** In parallel, we trained a ResNet-50 encoder on retrieval classes, generating text guidance aligned with EEG features. This pathway effectively produced linguistic descriptors of the stimulus, forming the basis for diffusion conditioning.

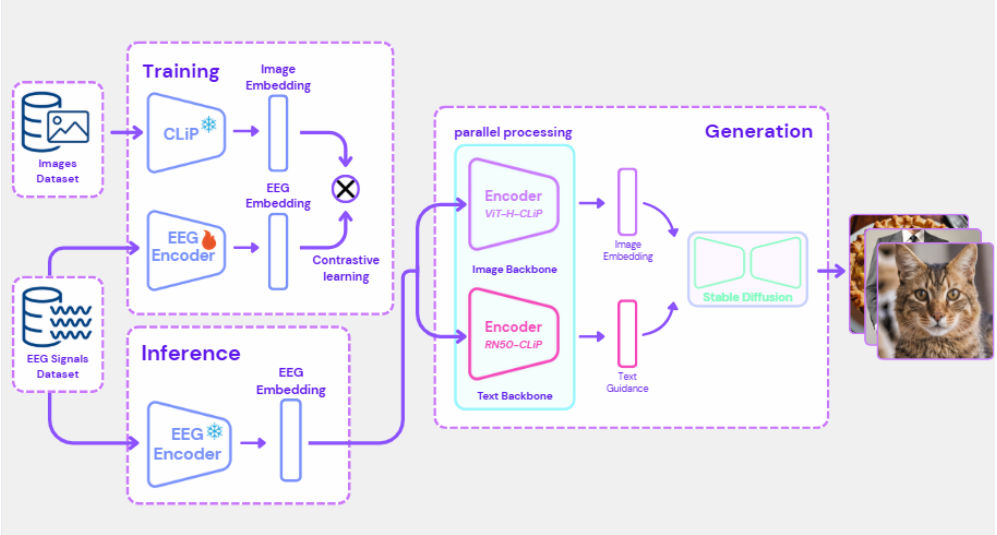


Figure 1: High-level view of our proposed architecture. EEG features are processed via dual backbones for image and text alignment, with text outputs guiding Stable Diffusion for reconstruction.

3.4 Text-Guided Reconstruction with Stable Diffusion

The text outputs from the ResNet-50 backbone were used as prompts for a Stable Diffusion model. This step transformed retrieval information into generative guidance, steering the diffusion process toward semantically faithful image reconstruction.

Strengths:

- When Top-1 retrieval was correct, reconstructions were highly coherent and visually accurate, please refer to Figure 4 for visuals.
- Text conditioning allowed the diffusion model to compensate for noise in raw EEG signals, effectively regularizing the generation process.

Limitations:

- Misretrievals occasionally misdirected the diffusion process, producing semantically off-target images.
- While the approach greatly improved fidelity, it relied heavily on retrieval accuracy.

3.5 Error Analysis and Robustness

We observed that retrieval errors did not catastrophically fail the pipeline; rather, they produced plausible but semantically shifted reconstructions. This demonstrates that the generative model retains robustness even when EEG retrieval is imperfect, an encouraging property for real-world deployment.

3.6 Novelty and Contribution

To the best of our knowledge, this is the first work that **reframes EEG-based image reconstruction as a guidance problem rather than a denoising task**. By coupling high-retrieval EEG embeddings with a text-to-image diffusion model, we:

- Constructed images that are not only visually coherent but also semantically faithful to the original stimulus.
- Opened a new research direction where future work can further integrate multimodal supervision for robustness.

4 Experiments prior to proposed method

EEG signals are inherently noisy and highly variable, limiting performance regardless of the encoder architecture. To address this, we hypothesized that transforming EEG into representations that could be processed by vision backbones might help. Vision models are strong feature extractors trained on large-scale data and may act as implicit regularizers by converting noisy EEG waveforms into structured visual patterns. This motivated a series of experiments where we reformulated EEG into vision-like modalities and explored how well such transformations could stabilize learning.

4.1 Representation Engineering

4.1.1 Spectrogram Transformation.

EEG signals were transformed into time-frequency spectrograms and embedded via ConvNeXt. We tested the hypothesis that the compression inherent in small image representations, such as spectrograms, might help regularize the noisy EEG signals. Contrary to our expectations, performance degraded sharply. A classification-only test revealed high intra-class variance in spectrogram amplitude patterns, indicating instability, this can be shown in Figure 2.

4.1.2 Topographical EEG Mapping.

After noticing the success of fMRI representation in retrieval and classification tasks, inspired by fMRI spatial mapping, we generated electrode-topography heatmaps, again with the idea that spatial compression could mitigate noise. These too suffered from instability, likely due to unresolved noise, this can be shown in Figure 3.

4.2 Noise Mitigation and Training Strategies

4.2.1 Supervised Pretraining → Contrastive Fine-tuning.

A short supervised phase shaped a separable latent space before contrastive alignment, which was our new approach. This approach allowed us to test and validate the supervised components independently, making experimentation faster and easier. Moreover, any success achieved in this supervised phase directly propagates to the contrastive latent space, since it serves as the initial representation for the main contrastive training.

4.2.2 Knowledge Distillation.

A CLIP teacher model provided soft targets to stabilize EEG representation learning. Minor gains were observed, but noise remained the dominant challenge.

4.2.3 Multi-label KL Divergence.

Fine-grained labels were replaced by superclass probability distributions, training with KL divergence to model uncertainty and reduce overfitting.

4.3 Architectural Enhancements

4.3.1 LEM Integration.

Large EEG Models (LaBraM) were integrated as pretrained feature extractors for benchmarking against our encoders. While LaBraM demonstrated strong performance in retrieval tasks, it exhibited limitations in reconstructing EEG signals. Specifically, its reconstructed outputs lacked fine-grained details, likely due to challenges in capturing the full spectrum of neural oscillations. These findings are consistent with the observations in the original LaBraM study, which highlighted the model's strengths in feature extraction but noted potential shortcomings in signal reconstruction. For more details, refer to the work by Jiang et al. (2024) [7].

4.3.2 Human-Aligned Image Encoders (DreamSim).

We replaced standard CLIP encoders with DreamSim for human-perceptual alignment. While theoretically promising, preliminary experiments showed lower retrieval accuracy than CLIP, likely due to suboptimal hyperparameters.

4.4 Multimodal and Text-based Integration

4.4.1 Text Prompts in Stable Diffusion.

Adding class names as prompts significantly improved reconstruction quality in many cases, especially for semantically distinct categories (e.g., animals, clothing, food).

4.4.2 EEG-to-Text Pipeline.

We explored a pipeline where RoBERTa-based caption embeddings were aligned with EEG features, followed by a text decoder to generate captions from EEG. In this setup, the text encoder is trained to be reused at inference time, generating captions that can guide the diffusion process. This component remains under development and is not yet fully implemented.

4.4.3 LLM-Generated Captions.

To bootstrap high-quality text supervision, we experimented with LLaVA, Gemini, and Qwen2.5-VL for generating rich captions. Qwen2.5-VL offered the best balance between speed and quality, although its outputs sometimes introduced very interesting but unexpected descriptive biases. The motivation was to integrate text alongside images in our contrastive learning framework, enabling the model to benefit from multi-modal alignment.

5 Results Visualization

To qualitatively evaluate our approach, we present side-by-side comparisons of reconstructed images against the baselines (Figure 4). Each row shows the ground truth stimulus, the output from the ATM retrieval baseline, the output from the Uncertainty-Aware Blur Prior (UBP) without diffusion prior, and our final reconstruction.

For the ATM baseline, we note that only the high-level retrieval pathway was executed; the low-level VAE branch was omitted, as our primary focus was on stabilizing and improving the core EEG-to-diffusion alignment. Similarly, for the UBP trials, both in our reimplementation and in the original paper, the diffusion prior was not employed in this setting, allowing for a fairer comparison at the same abstraction level.

Our method leverages text guidance in reconstruction and is evaluated using top-1 retrieval as the reference. As illustrated, the proposed pipeline produces reconstructions that are both more coherent and semantically faithful than those of the ATM and Blur Prior baselines, highlighting the added value of integrating text signals into EEG-to-image generation.

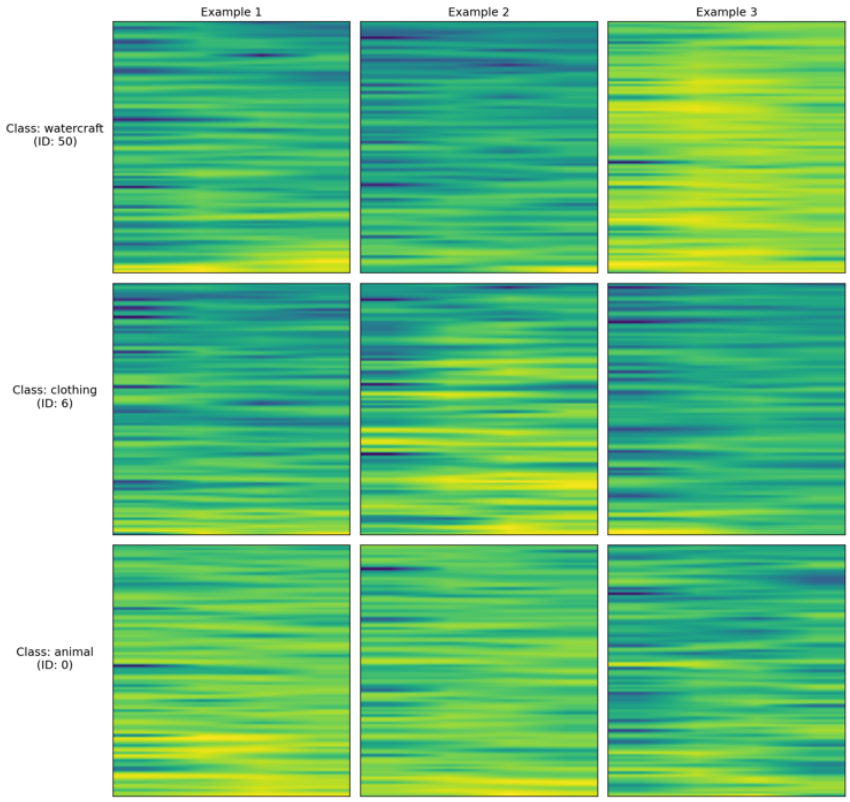


Figure 2: Example of spectrogram variability within the same class. The large differences between samples contribute to inconsistencies during classification and reconstruction.

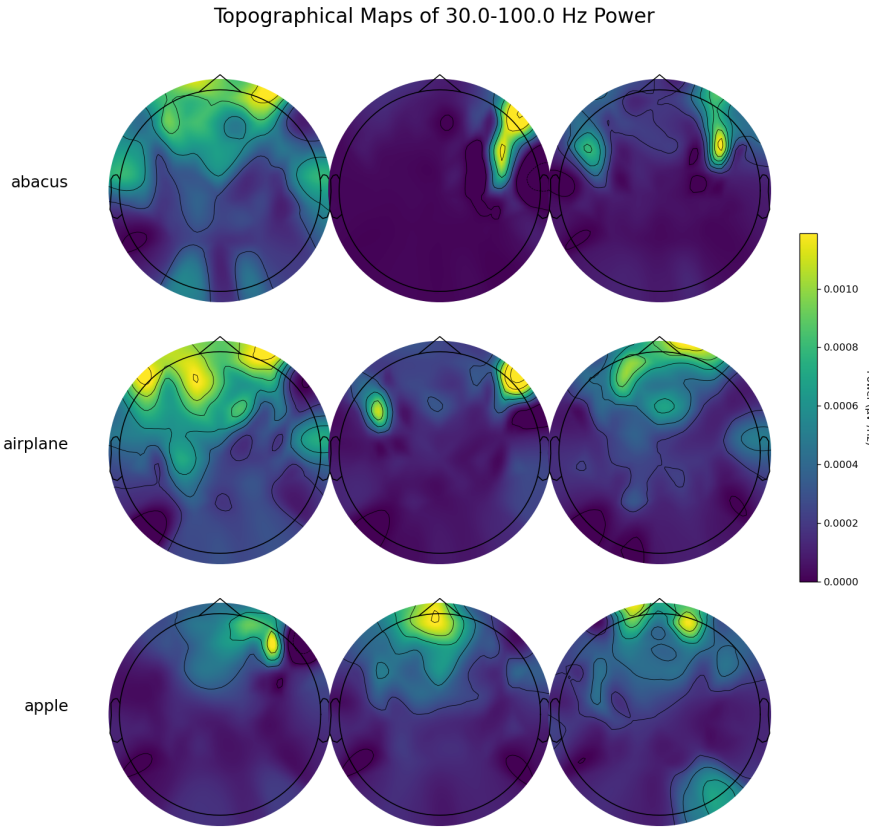


Figure 3: Illustration of topographic variability within the same class. Such large differences between samples can lead to inconsistencies in classification and reconstruction.



(a) Ground Truth (b) ATMs Retrieval (c) Blur Prior (d) Ours



(e) Ground Truth (f) ATMs Retrieval (g) Blur Prior (h) Ours



(i) Ground Truth (j) ATMs Retrieval (k) Blur Prior (l) Ours



(m) Ground Truth (n) ATMs Retrieval (o) Blur Prior (p) Ours

Figure 4: Comparison of EEG-to-image outputs for intra-subject training and inference on subject 8. Each row shows (left) the ground truth image, (second) ATM retrieval without low-level features, (third) Blur Prior without diffusion, and (right) our text-guided reconstruction.

Acknowledgments

We would like to express our sincere gratitude to KAUST Academy for providing this valuable opportunity. We are especially thankful to Prof. Naemullah Khan and Dr. Sultan Albarakati for their guidance and support throughout the program. Our heartfelt appreciation also goes to our amazing TAs for their constant assistance.

A special thanks to Mohammed Eltayeb and Lama Ayyash for their exceptional help and contributions, which greatly enhanced the progress and outcomes of our project.

References

[1] Jonathan Xu, Bruno Aristimunha, Max Emanuel Feucht, Emma Qian, Charles Liu, Tazik Shahjahan, Martyna Spyra, Steven Zifan Zhang, Nicholas Short, Jioh Kim, Paula Perdomo, Ricky Renfeng Mao, Yashvir Sabharwal, Michael Ahedor Moaz Shoura, and Adrian Nestor. Alljoined1 – a dataset for eeg-to-image decoding. *arXiv preprint arXiv:2404.05553*, 2024. URL https://arxiv.org/abs/2404.05553.

[2] Tijl Grootswagers, Alexandra K. Robinson, and Thomas A. Carlson. Things-eeg2: A dataset of multimodal neural responses for object recognition. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2022. URL https://arxiv.org/abs/2206.08384.

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

[4] Nona Rajabi, Antônio H. Ribeiro, Miguel Vasco, Farzaneh Taleb, Mårten Björkman, and Danica Kragic. Human-aligned image models improve visual decoding from the brain, 2025. URL https://arxiv.org/abs/2502.03081.

[5] Dongyang Li, Chen Wei, Shiying Li, Jiachen Zou, Haoyang Qin, and Quanying Liu. Visual decoding and reconstruction via eeg embeddings with guided diffusion, 2024. URL https://arxiv.org/abs/2403.07721.

[6] Runzhao Wu, Rui Zhang, Yuxuan Zhou, Qiongjie Guo, Haotian Zhang, Xudong Lin, Xiang Zhang, and Yang Zhang. Bridging the vision–brain gap with an uncertainty-aware blur prior. *arXiv preprint arXiv:2503.04207*, 2025. URL https://arxiv.org/abs/2503.04207.

[7] Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous eeg data in bci. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=QzTpTRVtrP.