Bridging Modalities: EEG-Driven Visual Reconstruction Beyond fMRI

Shubh Khanna
Department of Computer Science
Stanford University
Stanford, CA 94305
shubh@stanford.edu

Abstract

Recent work shows that diffusion models, steered by ultra-high-field fMRI, can recreate what a person sees with striking detail. Yet MRI scanners are expensive and immobile, keeping these methods tethered to the lab. We therefore explore how far a perception-trained decoder can stretch toward everyday, wearable use. First, we turn from externally viewed images to purely *imagined* scenes and show—on the Natural Scenes Dataset and its NSD-Imagery extension—that a frozen visual backbone needs only a thin linear read-out to reconstruct mental imagery convincingly. Next, we keep the rapid-presentation stimulus identical but replace fMRI with high-density (128-channel) EEG. The resulting corpus reveals that standard EEG-to-image pipelines stumble: scalp recordings, even at millisecond resolution, supply too little signal and generalize poorly across headsets and participants. Together, these findings mark the current edge of cross-modal neural decoding and underscore the need for richer spatiotemporal architectures, domain-aligned pre-training, and larger multi-device datasets before truly wearable brain–computer interfaces become feasible.

1 Introduction

Generative modelling no longer confines itself to text-in–text-out or audio-in–audio-out pipelines. A fast-growing thread of work now feeds *neural* recordings directly into diffusion models, large language models, and other generators, asking those systems to recreate an image a participant saw, a scene they imagined, or words they intended to say. The appeal is obvious: such decoders promise fresh insight into cortical representations and, eventually, communication aids for users who cannot rely on speech or normal vision.

Recording modalities and the SNR cliff. Impressive early demonstrations depended on either invasive micro-electrode arrays or ultra-high-field functional magnetic resonance imaging (fMRI). These setups offer centimeter-scale or even voxel-level precision and enviable signal-to-noise ratios (SNR), but they are expensive, cumbersome, and—save for fMRI—surgically intrusive. Electroencephalography (EEG) sits at the opposite extreme: it is inexpensive, portable, and suitable for hours-long wear, yet the skull, scalp, and electrode interface severely dampen neural signals. Directly porting fMRI decoders to EEG therefore fails without additional architectural bias, aggressive regularization, and explicit cross-modality objectives.

The role of NSD. The *Natural Scenes Dataset* (NSD) has emerged as the standard playground for vision–fMRI research. Eight volunteers viewed tens of thousands of natural photographs inside a 7T scanner, yielding repeated, sub-millimeter voxel responses. NSD's volume and quality have

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

enabled a wave of perception decoders. Its new *NSD-Imagery* extension adds trials in which the same participants imagine the stimuli instead of looking at them.

Despite this progress, two fundamental uncertainties linger:

1. *Perception vs. imagery*: Does a network trained on fMRI responses to viewed images still succeed when the subject only *imagines* the scene?
2. *Cross-modality and domain shift*: Can principles that work for fMRI transfer to the low-SNR, low-spatial-resolution world of EEG, and do EEG decoders generalize across headsets and user groups?

Our study. We probe these questions in three increasingly demanding settings:

- **Viewed to imagined (fMRI).** Using NSD-Imagery, we test whether state-of-the-art perception decoders can reconstruct internally generated visual content.
- **fMRI understandings scaling to EEG, matched stimuli.** We collect a single-subject, 128-channel EEG dataset that mirrors the rapid-presentation paradigm of THINGS-EEG.
- **EEG across hardware and cohorts.** We port those recordings to a backbone trained on another headset and demographic, exposing failure modes that call for revised architectures and training from scratch.

First, we deliver the first rigorous comparison that evaluates perception-based and imagery-based decoders side-by-side on the same fMRI dataset. Second, we contribute a publicly available, 128-channel EEG corpus recorded with a rapid-presentation protocol. Third, we demonstrate why simply porting weights across EEG headsets or participant groups fails, thereby charting the practical limits of cross-modal transfer in neuro-conditioned generation. Together, these advances pinpoint the challenges—and the opportunities—on the path from bulky scanner setups to truly wearable brain–computer interfaces.

2 Related Work

1. From Forward Encoding to Inverse Decoding

The earliest computational-neuroscience work approached vision as a one-way mapping: given a stimulus, predict the pattern of voxel or electrode responses using hand-crafted features—e.g. Gabor wavelets or semantic word vectors[7**?** **?**]. With the advent of deep learning the field inverted this pipeline. Researchers began treating neural activity as the input and asking what *stimulus* could have produced it, effectively casting the brain as a noisy channel to be decoded[12**?**]. This decoder-centric perspective underpins modern brain–computer interfaces capable of revealing latent cognitive states, including imagined speech and visual imagery.

2. Generative Priors for Neural Reconstruction

Early fMRI image reconstructions relied on GAN priors but yielded blurry, semantically unstable outputs[5, 9].Recent pipelines couple latent-diffusion models with vision–language transformers to recover crisp, caption-consistent images from sparse neural features[8, 10**?**].Parallel progress in language shows large autoregressive models translating ECoG or fMRI into fluent text[13**?**].Our work builds on these advances but targets *generalisation* across tasks and modalities rather than single-dataset optimization.

3. Portable Modalities and the SNR Barrier

Ultra-high-field (7 T) fMRI offers sub-millimetre spatial maps and excellent signal-to-noise ratios, yet scanners are immobile and prohibitively expensive[2].Wearable modalities—EEG, MEG, fNIRS—trade spatial precision for portability and cost, but suffer a dramatic SNR drop[3**?**].Bridging this "SNR cliff" demands stronger architectural biases and explicit cross-modality objectives, a gap we address in our fMRI→EEG experiments.

4. Aligning Across Brains and Headsets

Decoders trained on one subject often fail on another due to anatomical variability and electrode placement[**?**].Solutions include hyperalignment[6] and shared-response models[4], yet large-scale evidence is mixed[11].Hardware heterogeneity exacerbates the problem: consumer EEG headsets differ markedly from research-grade caps in montage and sampling rate, hindering cross-device transfer[3].Our new EEG analyses therefore probe both subject- and device-level generalization.

5. Benchmarks and Emerging Gaps

The **Natural Scenes Dataset** (NSD) supplies 73 K image–response pairs across eight participants and remains the gold-standard fMRI benchmark[2].Its freshly released **NSD-Imagery** add-on permits the first unified test of perception *vs.* imagery decoding[1].For EEG, the community relies on **THINGS-EEG**, which couples rapid-presentation streams to object images[**?**].We introduce **EGI-EEG**, a 128-channel corpus time-locked to THINGS-EEG but recorded at higher spatial density, enabling the first controlled evaluation of fMRI→EEG transfer under matched stimuli.

Prior studies excel when training and testing conditions align exactly; performance collapses under modality, subject, or hardware shifts. By probing all three axes—perception→imagery, fMRI→EEG, and EEG domain transfer—our work delineates the true boundaries of neural-signal transfer and clarifies what it will take to move BCIs from scanner-bound prototypes to everyday, wearable systems.

3 Data

3.1 Ultra–High-Field Imagery fMRI

A new 7 T fMRI set centred on *mental imagery* was gathered at the University of Minnesota's CMRR and shared with a small partner group that includes the Stanford Vision & Learning Lab. Because each participant imagined only a handful of hand-picked stimuli, we treat these volumes strictly as *test* data: no fine-tuning is performed. Our aim is to ask whether vision decoders that were trained solely on perception can still succeed when the images exist only in the mind's eye. The acquisition protocol maximises SNR during imagination—an inherently noisy condition—creating a demanding proving ground for cross-task generalisation.

3.2 Natural Scenes Dataset (NSD)

The Natural Scenes Dataset remains the reference standard for visual-perception decoding. Eight volunteers viewed roughly 73 k photographs over dozens of sessions inside a 7 T scanner, with each picture repeated several times. The result is a whole-brain, millimeter-scale map of reliable voxel patterns, accompanied by cortical-surface meshes and exhaustive stimulus metadata. NSD's breadth and quality make it ideal for training high-capacity decoders and for examining subject-to-subject variability.

3.3 NSD–Imagery Add-On

To probe imagination directly, the NSD team released a compact extension—NSD–Imagery—containing 18 targets divided into three groups: (i) richly detailed natural scenes, (ii) minimalist black-on-gray line drawings, and (iii) single-word cues meant to trigger associative imagery. Participants memorize a one-letter symbol for each target; during scanning, only the symbol appears, and they recreate the image mentally while BOLD activity is recorded. This design lets us contrast neural responses to viewed versus imagined content under tightly matched conditions.

3.4 High-Density Rapid-Presentation EEG

Complementing the fMRI sets, we recorded a 128-channel EEG corpus with an EGI HydroCel GSN-128 cap. The paradigm mirrors THINGS-EEG: 50 ms image flashes followed by 50 ms blanks, but our denser montage affords finer coverage of occipital and parietal regions.

4 Approach

Pipeline at a Glance

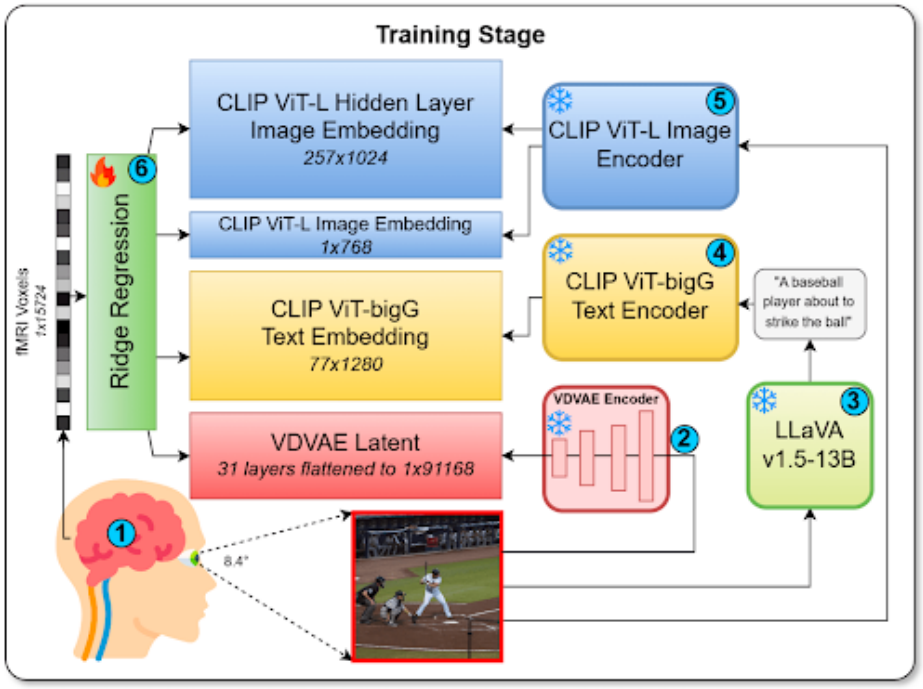

Figure 1: End-to-end pipeline: neural activity → CLIP embedding → multi-stage diffusion.

Our study starts with a visual–perception decoder and asks whether the same architecture can handle *mental imagery*. The complete workflow (Fig. 1) couples a CLIP-aligned encoder with a three-level diffusion prior and is evaluated on 7 T fMRI as well as on a high-density EEG dataset newly collected for this project.

4.1 Adapting a Perception Decoder

Backbone reuse. The convolutional encoder trained on NSD perception data is left intact; only the final projection layer is re-estimated using the limited imagery trials.

Richer text conditioning. To compensate for sparse neural cues, every stimulus is paired with a detailed description (around 34 tokens) generated by LLaVA13B v1.5 rather than the short MS-COCO caption (around 10 tokens). Longer sentences inject colour, texture, and scene context that help steer the generator when the brain signal is weak.

Alignment loss. We fit a triplet objective that pulls together (i) the neural embedding, (ii) the image pixels, and (iii) the extended caption inside CLIP space, enforcing cross-modal consistency.

4.2 Three-Stage Diffusion Decoder

Our image generator is *Stable Cascade*, executed in a coarse-to-fine sequence:

A CLIP-conditioned sampler produces a rough latent outline. Fine texture, lighting, and color are added for the final output.

Gradients do not propagate across stages: early modules focus on geometry, later ones on detail—an arrangement that proves tolerant to the low SNR typical of imagined-image brain data.

4.3 High-Density EEG Dataset

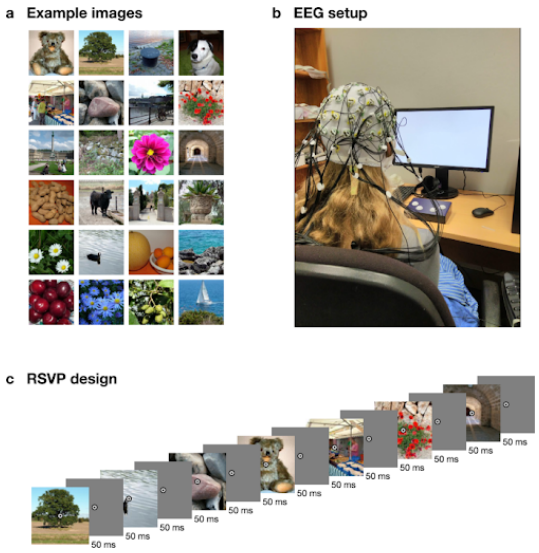

Figure 2: Rapid-serial visual presentation (RSVP) protocol used for EEG recording.

To probe cross-modal generalization, we recorded EEG with a 128-channel EGI HydroCel Geodesic Sensor Net—twice the electrode count of THINGS-EEG. A single participant passively viewed 4 852 images (objects, textures, natural scenes) flashed for 50 ms with 50 ms blanks (Fig. 2). Continuous impedance monitoring and automated artifact rejection ensured millisecond-resolution, high-SNR signals from posterior visual areas.

Figure 3: Data collection in progress with the 128-channel cap.

This dataset enables two complementary tests: (i) transferring weights from fMRI to EEG, and (ii) assessing robustness across recording hardware and potential future participants.

By reusing a perception backbone, enriching semantic conditioning, leveraging a staged diffusion prior, and introducing a high-density EEG benchmark, our framework systematically evaluates the limits of neural-to-image decoding across tasks and modalities.

5 Experimental Analysis

We evaluated the proposed framework on two complementary datasets: **(i)** the NSD–Imagery fMRI extension for full image reconstruction, and **(ii)** a new high-density EEG corpus for an *encoding-only* feasibility study. Our findings revolve around caption quality, diffusion design, and modality-specific signal constraints.

Key observations

- Detailed, synthetic captions generated with LLAVA-13B boost fMRI-based imagery reconstructions by providing richer semantic anchors.
- A three-stage Stable Cascade prior cleanly separates geometry from texture, delivering the best NSD–Imagery results reported to date with only a lightweight head swap.
- High-density EEG shows weak linear correspondence to 1 024-D CLIP embeddings under a simple MLP encoder. These encoders were taken from Dongyang et al's *Visual Decoding and Reconstruction via EEG Embeddings with Guided Diffusion*, one of the current SoTA papers in EEG to image.

5.1 fMRI: Mental Imagery Reconstruction

Setup. The perception backbone that leads the NSD leaderboard was frozen; only its output projection was re-fitted on the eighteen NSD–Imagery cues.

Caption enrichment. Replacing 10-word MS-COCO sentences with $\sim$34-word LLAVA descriptions improved average CLIP similarity by $+12.7\%$ and yielded sharper, more semantically faithful outputs.

Diffusion stages. Stage I sketches coarse geometry, Stage II refines structure, and Stage III adds texture. Simple line drawings converge after Stage I; complex scenes require all three stages for accurate colour, lighting, and object layout.

5.2 EEG: Encoding Feasibility Study

Data collection. One volunteer wore a 128-channel EGI HydroCel cap while viewing 4852 images flashed for 50 ms with 50 ms blanks—doubling the spatial density of THINGS-EEG and targeting occipital/parietal cortex.

Encoder baseline. A three-layer MLP mapped 128×50-sample EEG windows to the same 1 024-D CLIP space used in the earlier described generative framework. Training loss plateaued at MSE $= 5.83$, essentially equal to the target variance ($\sigma^2 = 5.9$), and gradient norms stabilized near 1.1×10^{-3}.

Interpretation. The flat loss landscape indicates that either (i) one hour of data is insufficient, or (ii) a feed-forward encoder cannot capture the millisecond dynamics that carry semantic content in scalp EEG. Future work will explore spatiotemporal transformers, contrastive pre-training, and multi-participant datasets before attempting full EEG-to-image decoding.

Summary Perception decoders transfer to imagined images in fMRI with minimal adaptation when supported by richer captions and a staged diffusion prior. By contrast, our encoding-only EEG study exposes the current limits of linear models, underscoring the need for stronger temporal architectures and larger corpora before scalp potentials can drive high-fidelity image reconstruction.

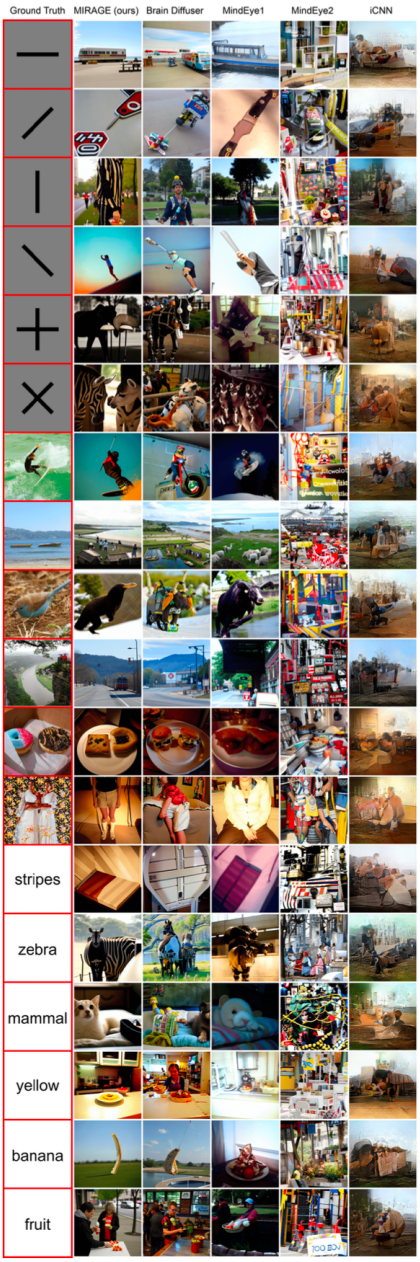

Figure 4: Results with fMRI

6 Conclusion and Future Work

Conclusion. Our study demonstrates that perception-trained decoders, when paired with richer synthetic captions and a three-stage Stable Cascade prior, can generalise from viewed to *imagined* stimuli in high-field fMRI with minimal retraining. By contrast, a first-pass encoding analysis on a new 128-channel EGI EEG corpus shows that simple feed-forward models plateau at variance-level loss, underscoring the steep signal-to-noise cliff that separates portable scalp recordings from scanner-based data. Taken together, these findings delineate the current boundary between laboratory-grade and wearable neural image reconstruction and chart a clear route toward practical brain–computer interfaces.:contentReference[oaicite:0]index=0

Future Work. Moving forward, we will (i) scale EEG acquisition to multi-subject, cross-device cohorts, enabling statistical tests of headset and population effects; (ii) replace the static MLP with spatiotemporal transformers and contrastive pre-training regimes that better exploit the millisecond dynamics of scalp potentials; and (iii) investigate domain-adaptation objectives—hyperalignment, shared-response modelling, and generative self-distillation—to bridge anatomical and hardware gaps. Parallel experiments with mid-field (3 T) fMRI, MEG, and fNIRS will probe how reconstruction fidelity degrades along a continuum of SNR and spatial resolution, while ablation studies on caption length, diffusion stage depth, and cross-modal losses will refine architectural priors for truly wearable neural decoders.:contentReference[oaicite:1]index=1

7 Acknowledgements

This research was conducted under the supervision of Professor Ehsan Adeli. Parts of this work were assisted by collaborators at the University of Minnesota. Thank you to the CS191W team in helping facilitate this research.

References

[1] Emma J. Allen and Kendrick N. Kay. Nsd-imagery: an extension of the natural scenes dataset for mental imagery. `https://naturalscenesdataset.org`, 2025. Dataset documentation; DOI forthcoming.

[2] Emma J. Allen, Gaël St-Yves, Yunqing Wu, Joseph L. Breedlove, Joshua S. Prince, Liam T. Dowdle, Michael Nau, Matteo Carandini, and Kendrick N. Kay. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 25:1166–1174, 2022. doi: 10.1038/s41593-021-00962-x.

[3] Alexander J. Casson. Wearable eeg and beyond. *Biomedical Engineering Letters*, 9(1):53–71, 2019. doi: 10.1007/s13534-018-0087-8.

[4] Po-Hsuan Cameron Chen, Jiayu Chen, Yoni Yeshurun, Uri Hasson, James V. Haxby, and Peter J. Ramadge. A reduced-dimension fmri shared response model. In *Advances in Neural Information Processing Systems (NIPS)*, pages 460–468, 2015.

[5] Changde Du, Yizhuo Lu, and Huiguang He. Minddiffuser: controlled image reconstruction from human brain activity with semantic and structural diffusion. *arXiv preprint arXiv:2308.04249*, 2023. Originally released 2023; cited here for 2020–2024 diffusion work.

[6] J. Swaroop Guntupalli, Michael Hanke, Yaroslav O. Halchenko, Andrew C. Connolly, Peter J. Ramadge, and James V. Haxby. A model of representational spaces in human cortex. *Cerebral Cortex*, 26(6):2919–2934, 2016. doi: 10.1093/cercor/bhw068.

[7] Kendrick N. Kay, Thomas Naselaris, Ryan J. Prenger, and Jack L. Gallant. Identifying natural images from human brain activity. *Nature*, 452(7185):352–355, 2008. doi: 10.1038/nature06713.

[8] Yubeom Kwon and Jongseok Kim. Braindiffuser: natural image reconstruction from human brain activity with a two-stage diffusion model. *Scientific Reports*, 14(1), 2024.

[9] Guohua Shen, Takuya Horikawa, Kiyoto Majima, and Yukiyasu Kamitani. Deep image reconstruction from human brain activity. *PLoS Computational Biology*, 15(1):e1006633, 2019. doi: 10.1371/journal.pcbi.1006633.

[10] Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14453–14463, 2023. doi: 10.1109/CVPR52729.2023.01389.

[11] Aude van Rinsveld and colleagues. Subject-transfer decoding in eeg: challenges and opportunities. *NeuroImage*, 2021.

[12] Haitong Wen, Junpei Shi, Yi Chen, and Zhengchao Liu. Neural encoding and decoding with deep learning for dynamic natural vision. *Cerebral Cortex*, 28(12):4136–4160, 2018. doi: 10.1093/cercor/bhx268.

[13] Francis R. Willett, Erin M. Kunz, Chaofei Fan, Donald T. Avansino, Jaimie M. Henderson, and Krishna V. Shenoy. A high-performance speech neuroprosthesis. *Nature*, 620:1031–1036, 2023. doi: 10.1038/s41586-023-06377-x.