Natural Scenes Dataset (NSD): Design and Data Properties

The Natural Scenes Dataset (NSD) is an ultra-high-field (7 T) fMRI dataset designed for deep sampling of visual experience. It comprises *whole-brain* 1.8 mm (isotropic) BOLD measurements (TR ≈1.6 s) from **8 healthy adult subjects** across ~30–40 sessions each. Each subject viewed on the order of 9,000–10,000 **unique color natural scene** images (drawn from the MS-COCO database) with three repetitions per image. In total NSD spans **~70,000 distinct images** (8 subjects × 9,000 unique + 1,000 shared images). The images were presented in a rapid event-related design (4 s trials) while subjects performed a **continuous recognition memory task** – indicating whether each scene had been seen previously during the experiment. This task ensured deep engagement over the year-long acquisition and made NSD the longest continuous-recognition memory fMRI study to date.

Stimuli and Task: The visual stimuli were naturalistic scenes from MS-COCO, center-cropped and matched in size, shown against a neutral gray background. Each run consisted of ~62 trials with occasional blank fixation periods, and each session contained 12 runs (~750 trials). The continuous recognition paradigm probed both short-term and long-term memory: images were scheduled such that repeats occurred with varying lags across and within sessions. This design created a rich set of regressors and controlled for arousal, while potentially introducing memory-related BOLD components.

Subjects and Variability: Eight carefully screened subjects (wide age range, normal vision) were scanned at ultra-high field, maximizing signal-to-noise (SNR). Notably, eyetracking showed some variability: most subjects maintained central fixation (>90% of the time within 1°), but one subject (Sub-5) had modest gaze excursions and another (Sub-8) exhibited apparent noise in pupil tracking. The NSD authors advise caution or exclusion for analyses requiring strict fixation for those subjects. In general, NSD prioritized **depth over breadth**: intensive sampling of few individuals, based on the rationale that a complete model of one brain can generalize to others.

Voxel-Level Data: NSD data are high-resolution (1.8 mm) whole-brain 7 T fMRI. Preprocessing included motion correction, surface-based alignment, and a sophisticated GLMsingle analysis to obtain single-trial response estimates. To capture fine spatial patterns, the data were upsampled to a 1 mm grid in post-processing. Early and higher visual regions (e.g. V1, V2, V3, V4, LOC, FFA, PPA, etc.) are densely sampled: typical ROI voxel counts are on the order of 10–20 thousand per subject (e.g. ~15–16k voxels in early visual areas, ~12–14k in ventral areas). The large number of trials per voxel (many repetitions across sessions) yields high reliability; indeed, simple inspection shows clear representational gradients along the ventral stream.

Known Limitations: Despite its unprecedented scale, NSD has some constraints. Only eight subjects were collected, so population variability beyond these individuals is limited (though the design rationale emphasizes “representational dynamics” that generalize). Each image is shown only three times per subject, which is enough for robust signals at 7 T but limits very low-noise averaging. The continuous recognition task may introduce cognitive or mnemonic confounds beyond pure visual encoding. Moreover, ultra-high-field 7 T scanning can produce susceptibility artifacts (especially near sinuses and ear canals) and variable sensitivity across cortex. In summary, NSD trades off breadth (few subjects) for depth (many trials), providing a high-quality but somewhat specialized dataset.

SynBrain vs. MindSimulator: Generative fMRI Encoding Models

SynBrain (Mai et al., NeurIPS 2025) and **MindSimulator** (Bao et al., ICLR 2025) are recent generative frameworks for mapping images to fMRI. Both are *encoding* models (predicting brain activity given stimuli) but make distinct assumptions and design choices:

- **Modeling assumptions:** SynBrain explicitly treats visual-to-fMRI mapping as *one-to-many*: identical images can evoke variable neural responses due to noise and individual state [1]. It assumes that observed trial-to-trial variability is structured and can be disentangled from core semantic content. MindSimulator also acknowledges variability ("noticeable differences in brain activity for the same stimulus") but primarily emphasizes learning the *distribution* of responses via generative modeling. Unlike many deterministic encoding models, both methods adopt generative probabilistic paradigms: SynBrain via a variational autoencoder (BrainVAE), MindSimulator via a diffusion-based pipeline.

- **Representation of neural activity:** SynBrain's **BrainVAE** learns a continuous latent space for fMRI activity. Given an image embedding (from CLIP), it learns a distribution over latent variables that generate voxel activations [2]. Thus each image is mapped to a Gaussian posterior in latent space, whose samples decode to realistic neural patterns. SynBrain's latent space is semantically grounded: a contrastive CLIP loss ties the fMRI latents to the corresponding image features. In effect, SynBrain models neural activity as a CLIP-conditioned probability distribution, preserving semantic consistency across trials [2] [3].

MindSimulator first learns an autoencoder for fMRI itself, projecting raw voxel patterns into a "brain latent" space. This latent space is explicitly aligned with a pretrained image representation space (such as CLIP) through a cross-modal (SoftCLIP) loss. In other words, the autoencoder ensures that the low-dimensional brain representation corresponds to semantic features of the stimulus. On top of this, a **diffusion model** is trained to sample from the conditional distribution of these fMRI latents given an image embedding [4]. Thus MindSimulator's generative process is: image → image embedding → conditional diffusion → brain latent → decoded voxel activity. Notably, MindSimulator's inference includes an "Inference Sampler" that generates multiple noisy fMRI samples (with correlated noise) and averages them, emphasizing reproducibility [5].

- **Training objectives and supervision:** SynBrain trains its BrainVAE by combining a voxel-wise reconstruction loss, a KL-divergence prior on the latent, and a contrastive CLIP loss to align latents with image semantics. The VAE is thus supervised both by real fMRI targets (for MSE reconstruction) and by semantic consistency constraints. A separate **Semantic-to-Neural (S2N) mapper** network (an MLP) learns to map CLIP embeddings to the BrainVAE latent distribution. Overall the training maximizes a variational lower bound on p(fMRI|image) plus cross-modal alignment.

MindSimulator's training is staged. First, the **fMRI autoencoder** is trained with an MSE loss on voxel reconstructions plus a SoftCLIP loss aligning brain latent to image features. Second, the **diffusion estimator** is trained to predict noise (or directly predict denoised latents) conditioned on image embeddings. The diffusion loss follows standard denoising objectives (predict Gaussian noise added to brain latents), effectively learning p(latent|image). During inference, the "Inference Sampler" draws multiple noise samples, feeds them through the diffusion model, and averages to produce a final latent (adding correlated Gaussian noise across "trials" to mimic real variability) [5]. This multi-trial strategy is a supervised design choice unique to MindSimulator, intended to boost SNR of the synthetic responses.

- **Strengths:** SynBrain's probabilistic design captures neural variability and yields faithful single-shot predictions. It **surpasses state-of-the-art** deterministic models on NSD: in comparisons it outperforms prior methods on both voxel-wise and semantic-level metrics while using only one sample per image. SynBrain also excels at cross-subject adaptation: starting from one subject it can rapidly adapt to another with only ~1 hour of data, thanks to its modeling of shared semantic subspaces [6]. Its interpretable latent space reveals that individual differences lie in low-dimensional subspaces orthogonal to semantics [6]. Importantly, SynBrain's generated fMRI can augment scarce real data and *improve decoding*: adding synthetic data significantly boosted image-reconstruction performance in low-data settings.

MindSimulator's strengths lie in its generative flexibility and data-driven exploration. By training on concept-oriented images, it can synthesize **vast quantities** of realistic fMRI for novel stimuli and concepts. It consistently outperforms linear and transformer baselines on NSD synthesis accuracy (both voxel and semantic scores), coming close to ground-truth performance when using multi-trial generation [7] [8]. Its innovations (multitrial averaging with correlated noise) measurably boost performance [9]. Crucially, MindSimulator allows exploration of concept-selective regions: synthetic fMRI can localize established ROIs (e.g. face-, body-, place-selective cortices) and even suggest new candidate areas [10] [11]. In short, MindSimulator is powerful for hypothesis generation in concept localization via synthetic data.

- **Weaknesses:** SynBrain's reliance on pretrained vision embeddings (e.g. CLIP) may introduce biases that imperfectly match actual neural codes [12]. It captures "bulk" variability but does not explicitly model factors like attention or arousal, leaving some variance unexplained [12]. Its VAE architecture can also be computationally heavy and may blur extremely fine-grained voxel patterns (though it generally preserves semantics).

MindSimulator, in turn, requires a complex multi-stage training and inference. Diffusion models involve iterative denoising and may be slower in practice, though its reported 300 ms per sample is fairly efficient [13]. Its performance depends on the assumption that the autoencoder's latent sufficiently captures neural semantics; mismatches here could limit fidelity. Furthermore, MindSimulator as presented is largely per-subject and concept-driven: it does not explicitly address cross-subject alignment, so its applicability to new subjects without retraining is unclear. Finally, its use of multitrial averaging means a single-shot synthetic pattern may be noisier, requiring post-processing to use for decoding.

- **Suitability for NSD:** Both models are directly demonstrated on NSD. SynBrain was evaluated on NSD's ventral visual areas, achieving superior encoding for individual NSD subjects and demonstrating few-shot transfer between them [6]. It naturally fits NSD's scale (73k images) and uses CLIP to condense image semantics. MindSimulator also leverages NSD (and fLoc-localizer data) for concept synthesis. The NSD images (COCO scenes) provide rich semantic variation, matching MindSimulator's concept framework [14]. In principle, SynBrain's subject-adaptation strength is ideal for NSD's multiple subjects, whereas MindSimulator's strength in out-of-distribution concept synthesis could complement NSD when exploring categories not explicitly labeled in the data.

Overall, SynBrain and MindSimulator share a generative philosophy but differ in implementation: **SynBrain** is a one-shot VAE-based encoder optimizing cross-modal consistency and subject transfer [2] [6]; **MindSimulator** is an autoencoder+diffusion pipeline emphasizing large-scale synthetic sampling and concept localization [15] [7]. Both produce semantically aligned synthetic fMRI, but SynBrain focuses on probabilistic semantics with cross-subject generality, while MindSimulator emphasizes data-driven exploration with multi-trial fidelity.

Hypotheses for Image-to-fMRI Synthesis on NSD

Based on the above analysis, we propose the following testable hypotheses about generative fMRI encoding using NSD:

1. **Probabilistic encoding outperforms deterministic encoding:** Modeling fMRI responses as *distributions* conditioned on image semantics (as in SynBrain/MindSimulator) will predict held-out NSD responses more accurately than deterministic regressions. In particular, incorporating trial-to-trial variability will improve both voxel-level correlation and semantic-level retrieval metrics. (Test by comparing a probabilistic model's likelihood or correlation to a linear or MLP baseline on NSD test data.)

2. **Low-dimensional subject subspaces:** For each NSD subject, the trial-to-trial variability of their voxel responses to repeated images lies predominantly in a low-dimensional subspace **orthogonal** to the semantic encoding subspace. That is, semantic content lives in a dominant latent subspace, while other variance (e.g. arousal, scanner noise) occupies a mostly orthogonal space. (Test by performing PCA or factor analysis on residuals after removing the mean response per image; assess alignment with semantic axes.)

3. **Cross-subject semantic alignment:** A model trained on one NSD subject and aligned via semantic latent space (e.g. CLIP embeddings) will require only a small affine or low-dimensional adaptation to fit another subject's data. Specifically, adding synthetic fMRI (via a generative model) from a source subject will significantly reduce the real-data calibration needed for a new subject. (Test by few-shot adaptation experiments: train on Sub-1 and adapt to Sub-2 with/without synthetic augmentation.)

4. **Hierarchical encoding constraints improve accuracy:** Incorporating neuroscientific structure – e.g. a layered CNN whose intermediate features align with successive visual areas – into the generative model will yield better NSD encoding than a flat model. For example, a model that explicitly maps lower-layer features to V1 voxels and higher-layer features to IT cortex should outperform a model that treats all voxels uniformly. (Test by building and comparing hierarchical vs. non-hierarchical generative models on NSD.)

5. **Synthetic fMRI enhances decoding:** Using synthetic NSD fMRI generated by a hybrid model as additional training data will improve the accuracy of downstream image reconstruction/decoding models, especially in the low-data regime. (Test by training a decoding model with and without synthetic augmentation, and measuring image-reconstruction accuracy on held-out NSD images.)

6. **Invertible mapping yields interpretability:** A flow-based (invertible) model that explicitly learns a bijective mapping between image features and fMRI voxels will allow recovering interpretable "basis" responses. For instance, we hypothesize that in such a model, restricting the latent noise inputs in specific ways will produce changes in fMRI that correspond to interpretable visual attributes (e.g. turning "face" or "food" features on/off). (Test by probing the invertible model: manipulate latent dimensions and verify cortical activation changes.)

Proposed Hybrid Flow-Based Model for Image-to-fMRI Generation

We propose **“CortexFlow”**, a novel hybrid generative model designed to synthesize fMRI patterns from images with strong cross-subject generalization, interpretability, and biological plausibility. CortexFlow combines invertible flow architectures with neuroscience-informed structure and multimodal alignment:

- **Model architecture (overview):** CortexFlow is a conditional normalizing flow that maps image embeddings to fMRI voxel activations. Its encoder is a pretrained vision backbone (e.g. ResNet or CLIP-ViT) that extracts multi-scale visual features from the input image. These features condition a series of invertible coupling layers (e.g. RealNVP-style blocks) that transform simple noise into a synthetic fMRI volume. Importantly, the flow is *hierarchically partitioned by visual area*: each stage of the flow is specialized for a cortical region (e.g. V1→V2→V4→IT). For example, an early flow module might output a latent map for V1-size voxels, which is then upsampled and fed (along with higher-level conditioning) into the next module for V2, etc. At each stage, the coupling layers obey spatial constraints (e.g. using convolutional subnetworks) to respect retinotopy.

- **Subject-conditioning:** To capture individual differences, the flow includes a learnable **subject embedding** that modulates affine couplings. For instance, a FiLM layer per coupling block injects a subject code that scales and shifts feature maps. This allows CortexFlow to adjust to each brain’s idiosyncrasies while sharing most weights across subjects. During training, all subjects’ data contribute to learning the shared flow, with only a small dimension of parameters per subject. This design encodes the hypothesis that individual variability is low-rank and separable from core semantics [6] .

- **Neuroscience constraints:** To enhance biological plausibility, CortexFlow explicitly incorporates known cortical hierarchies. For example, the noise input to each region’s flow block is shaped by a parametric *receptive-field map*: voxels receive only relevant spatial components (e.g. a Gabor-like prior for V1 couplings). The network also respects **representational similarity**: a loss term encourages the distance between synthetic response patterns to mirror that of real NSD data (an RSA-style loss). Another constraint enforces that the Jacobian of the flow has locally smooth structure, reflecting the smooth cortical mapping of nearby visual features.

- **Generative process:** Given an image, its feature vector z_{img} (from the vision backbone) conditions the flow. We sample a base noise tensor and apply the invertible transformations to produce a “clean” fMRI latent. The invertibility ensures that this mapping preserves information and allows exact likelihood training. Importantly, CortexFlow produces a *single high-quality sample* per image, but can also generate multiple samples by varying the input noise for stochasticity. Unlike diffusion, this sampling is fast (one forward pass) and yields a deterministic transform for a given noise.

- **Training objectives:** CortexFlow is trained via **maximum likelihood** on NSD: the negative log-likelihood of the real fMRI given the image conditions. Equivalently, it minimizes MSE between predicted and actual fMRI under the flow, plus a log-determinant regularizer. To ensure semantic alignment, we add a contrastive CLIP loss between the synthetic fMRI and the image embedding, similar to SynBrain. We also include an RSA loss on intermediate activations so that synthetic patterns preserve the similarity structure of real NSD responses. The subject embeddings are learned via a small supervised warm-start (e.g. ridge regression) before fine-tuning within the flow.

- **Interpretability and efficiency:** Because flows are invertible, one can map a synthetic fMRI back to its latent noise: by inspecting how perturbations in latent space affect the output, we can infer which features drive neural activity. For example, one could find the latent direction that maximally activates "food-selective" voxels, bridging to concept maps. Inference is efficient (a few coupling layers), allowing real-time generation of many synthetic patterns. The model can be extended to downstream decoding by inversion: the learned bijection implies a natural brain-to-image decoder.

This hybrid flow-based design optimizes the key desiderata: it is **probabilistic** (sampling-capable like SynBrain/MindSimulator), **interpretable** (invertible and structured by cortex), **biologically grounded** (hierarchical and receptive-field based), and **efficient** (one-shot generation). Its conditional normalization framework should generalize across NSD subjects via shared weights and low-rank subject codes, while accurately capturing voxel-level detail and semantic content. In summary, CortexFlow represents a new generative paradigm for NSD image-to-fMRI synthesis, blending cortical neuroscience principles with modern flow-based learning.

Sources: Detailed NSD design and statistics are described in Allen et al. (2022) and on the NSD website. SynBrain's framework and findings are summarized in Mai et al. (NeurIPS 2025) [2] [12]. MindSimulator's architecture and performance are described in Bao et al. (ICLR 2025) [14] [5]. These sources underpin the above analyses and proposals.

[1] [2] [3] SynBrain: Enhancing Visual-to-fMRI Synthesis via Probabilistic Representation Learning | OpenReview
https://openreview.net/forum?id=ZTHYaSxqmq

[4] [5] [7] [8] [9] [10] [11] [14] [15] [Quick Review] MindSimulator: Exploring Brain Concept Localization via Synthetic FMRI
https://liner.com/review/mindsimulator-exploring-brain-concept-localization-via-synthetic-fmri?entry-type=quick_review_hub

[6] [12] SynBrain.pdf
file://file_00000000c9f47206bd11ca8b6ee8d534

[13] [2503.02351] MindSimulator: Exploring Brain Concept Localization via Synthetic FMRI
https://ar5iv.labs.arxiv.org/html/2503.02351