CONCEPT GENERALIZATION IN VISUAL BRAIN DECODING VIA ENCODER-GENERATED FMRI

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual brain decoders are typically trained and evaluated on stimuli drawn from similar semantic distributions, leaving their ability to reconstruct unseen concepts poorly understood. We study this concept generalization problem by withholding semantic concepts from a target subject's training data. We find that three recent fMRI-to-image decoders degrade substantially for the held-out concepts, even when the concept remains available for multi-subject models. In contrast, a brain encoder trained to predict fMRI from images is markedly more robust to the same semantic holdout. Motivated by this asymmetry, we use the encoder to generate synthetic fMRI responses for new images and add the resulting image-fMRI pairs to decoder training. Across three decoders, four subjects, and twelve semantic concepts, synthetic data recovers a substantial fraction of the performance lost on unseen concepts without requiring additional, time-consuming fMRI acquisition. We further show that the same approach improves general decoding performance across a wide range of data budgets. Analyses of encoder training conditions and synthetic-data controls show that the gains depend on both encoder generalization and the semantic content of the synthetic supervision, rather than merely increasing the amount of training data. Together, our results show that encoder-generated fMRI can transfer semantic generalization to decoders.

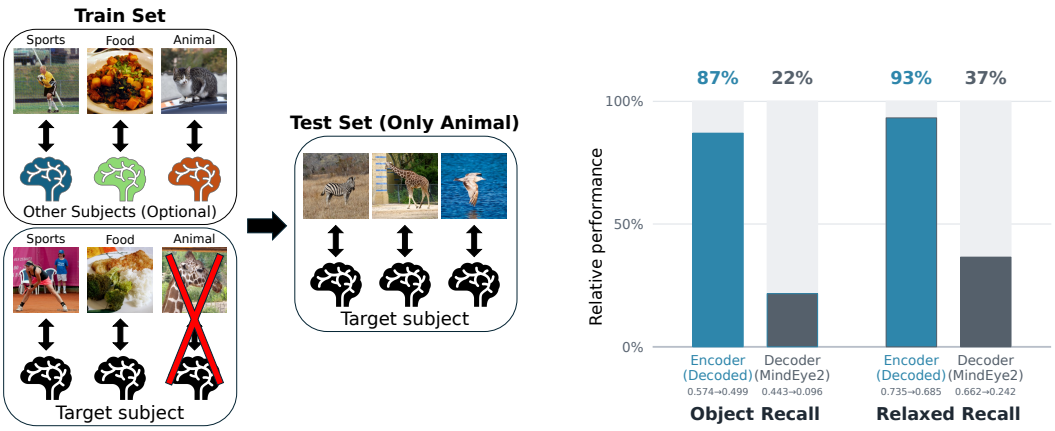
1 INTRODUCTION

Recent fMRI-to-image decoders produce reconstructions that are often semantically faithful to the viewed stimulus (Scotti et al., 2024; Beliy et al., 2026). These results are obtained under a protocol in which training and test images are drawn from the same pool of semantic concepts: a decoder asked to reconstruct a zebra at test time has already been trained on other zebras and the responses they evoked. The protocol therefore leaves a basic question unanswered: *Can a decoder reconstruct a concept it was never trained on?* A deployed decoder will inevitably face such concepts. Shirakawa et al. (2025) recently showed that the semantic overlap between training and test stimuli inflates apparent generalization, but the only remedy available so far is the one that limits brain decoding in the first place: collecting paired image-fMRI data for every new concept and every new subject, a time-consuming and expensive process.

We make this question measurable with a *concept generalization* setting (Figure 1a). We take a concept $\mathcal{C}$ as a COCO supercategory, remove every image containing $\mathcal{C}$ from the target subject's training set, and evaluate only on images containing $\mathcal{C}$. Because standard reconstruction metrics compare global embeddings and can score a reconstruction highly even when the object of interest is absent, we complement them with object-detection metrics that ask whether $\mathcal{C}$ itself is recovered. Under this setting, three recent decoders (Ozcelik & VanRullen, 2023; Scotti et al., 2024; Beliy et al., 2026) degrade sharply across four subjects and all twelve supercategories; *e.g.,* MindEye2 (Scotti et al., 2024) retains only 22% of its concept recall. Notably, the degradation persists even though $\mathcal{C}$ remains available in the data of the other subjects used for cross-subject pre-training or joint training. Decoders thus largely fail to transfer a concept seen in other brains to a subject who never saw it.

Brain encoders, which map images to fMRI, behave differently under the same holdout: when their predicted fMRIs are decoded, an encoder trained without $\mathcal{C}$ from the target subject retains 87% of its Object Recall, much higher than the case of the MindEye2 decoder (Figure 1b; Sec. 4.1). This asymmetry suggests using the encoder to transfer its robustness to the decoder through data. We

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) Concept generalization setting, $\mathcal{C}$ = *animal*.

(b) Recall retained when $\mathcal{C}$ is held out.

Figure 1: **The concept generalization setting and the encoder-decoder asymmetry.** (a) Images containing $\mathcal{C}$ are removed only from the target subject's training set and not from other subjects; the model is evaluated only on images containing $\mathcal{C}$. (b) The encoder retains most of its concept-specific recall under this holdout, while the MindEye2 decoder (Scotti et al., 2024) loses most of it. Encoder recall is measured by decoding its predicted fMRI with a MindEye2 trained with $\mathcal{C}$ present and with additional synthetic data (See Sec. 6 for more details).

generate synthetic fMRI responses for images containing $\mathcal{C}$ and add the resulting image-fMRI pairs to the decoder's training set. Because the intervention acts on the data alone, it applies to any decoder architecture or training strategy without modification, and it requires no additional scanning. Across the three decoders, the synthetic data recovers a large fraction of the generalization gap without any real fMRI for $\mathcal{C}$.

Since the encoder is trained on the same data as the decoder, why should its synthetic data help? By varying the encoder's training conditions and the composition of the synthetic data, we identify two sources of the gain. First, the encoder is inherently robust to unseen concepts: even an encoder that never saw $\mathcal{C}$ in any subject yields substantial gains, which we attribute to its pretrained visual backbone. Second, the encoder transfers exposure to $\mathcal{C}$ from other subjects, which decoders fail to exploit on their own. Controls further show that the gain is not a data-volume effect; it requires synthetic responses to images of $\mathcal{C}$ itself.

Finally, the use of encoder-generated synthetic data is not specific to concept generalization. When the synthesis pool consists of images unseen by the target subject, it improves all three decoders at every target-subject data budget from 225 to 9,000 images for most metrics, by an amount worth roughly a doubling of target-subject scan time (Sec. 5.3).

Our contributions are as follows. (i) We introduce the concept generalization setting, with concept-level recall metrics, and show that three recent decoders fail it even under cross-subject training. (ii) We show that brain encoders are far more robust to the same holdout, and that transferring this robustness through encoder-generated fMRI for images of the held-out concept recovers a large fraction of the gap across three decoders, four subjects, and twelve concepts. (iii) We show that the gain stems from the encoder's inherent robustness and from cross-subject transfer rather than from added data volume. (iv) We show that the same recipe improves decoding at various data budgets on most metrics, worth roughly a doubling of target-subject scan time.

2 Background and Related Work

2.1 Visual Brain Decoding and Encoding

From neuroimaging experiments that produce paired visual stimulus-fMRI response data, two complementary tasks can be defined: brain decoding and brain encoding. *Brain decoding* learns a map from fMRI to image, reconstructing the stimulus given the brain response. *Brain encoding* inversely learns a map from image to fMRI, predicting the brain response given the stimulus. One of the most

widely used datasets is the Natural Scenes Dataset (NSD) (Allen et al., 2022), where eight subjects viewed images drawn from COCO (Lin et al., 2014). NSD is the main dataset we use in this work.

By incorporating diffusion models, brain decoding models rapidly evolved, albeit requiring separately trained models for each subject (Ozcelik & VanRullen, 2023; Scotti et al., 2023). Since this paired data is costly, some recent decoders utilize cross-subject (pre-)training to help increase the performance for other subjects and therefore reduce the per-subject data burden (Scotti et al., 2024; Wang et al., 2024; Huo et al., 2024; Beliy et al., 2026).

On the encoding side, the Universal Brain Encoder (Beliy et al., 2024) learns voxel embeddings that attend to DINOv2 (Oquab et al., 2023) features with LoRA (Hu et al., 2022) tuning. The parameters are shared across voxels and subjects, which makes it applicable to different voxel counts and transferable between multiple subjects with minimal fine-tuning, and we use it as our brain encoding model for the experiments. Other recent encoders include SynBrain (Mai et al., 2025), which models the response probabilistically, and TRIBE v2 (d'Ascoli et al., 2026), a multimodal encoder which is also capable of predicting subject-agnostic cortical surface responses.

2.2 Semantic Generalization for Visual Brain Decoding

Whether a decoder can handle concepts absent from its training set is a question that has been raised before (Horikawa & Kamitani, 2017), but one that most recent decoding work does not engage with. Since NSD draws training and test images from the same pool, Shirakawa et al. (2025) argue that the resulting overlap inflates performance. They cluster NSD stimuli in CLIP text-feature space, withhold one cluster from training, and find that decoder performance drops and that the standard metrics overstate their generalization capabilities. Their proposed solutions concern dataset design and representation choice rather than a method that closes the gap for an existing dataset.

We later propose a *concept generalization* setting that builds on this diagnosis. We define the held-out concept from ground-truth annotations and evaluate decoders that train on multiple subjects (Sec. 3.1), use additional metrics to measure whether the concept itself is recovered (Sec. 5.1), and, unlike prior work, show a working solution that does not need to collect additional fMRI (Sec. 4).

2.3 Using Encoder-generated Synthetic fMRI for Decoding

To reduce the burden of data collection, a line of work produces synthetic neural responses for images that were never shown to the subject. Earlier work trains the decoder on such images through a fixed encoder, which supplies surrogate responses (St-Yves & Naselaris, 2018) or a cycle-consistency signal (Gaziv et al., 2022), the latter reporting gains on categories never seen in training. Their unpaired image pool is, however, not chosen with the test categories in mind: in Gaziv et al. (2022) the test categories are absent from the paired data and from the unpaired pool alike. More recently, brain encoders have been used to generate synthetic fMRI for decoder training: Brain-IT (Beliy et al., 2026) and SynBrain (Mai et al., 2025) add encoder-predicted responses to images drawn from the same distribution as the real stimuli, in full-data and few-shot settings, and Benchetrit et al. (2026) study synthetic cortical surface maps generated by TRIBE v2, focusing on common, subject-agnostic synthetic data and leaving subject-specific synthesis as an open question. None of these works selects the synthetic pool for the concept the decoder must generalize to, nor evaluates on held-out concepts, which is the gap we address.

3 Data-limited Settings for Brain Decoding

As paired image-fMRI training data is slow and expensive to acquire, the **data efficiency** of brain decoders is essential to gauging their practical utility. We use the term data efficiency to represent the ability to attain satisfactory performance, even when the training data is limited for the target subject (the subject it will be evaluated on). For decoders that train with multiple subjects (Scotti et al., 2024; Beliy et al., 2026), a common practice is to only limit the data from the target subject, not the other subjects. The reasoning is that in practice, data from other subjects is a one-time cost shared by every deployed decoder, whereas target-subject data must be collected for each subject.

Target-subject data can be restricted along two distinct axes. The commonly addressed one is quantity of the available data. The other is semantic coverage: which concepts those data depict. We

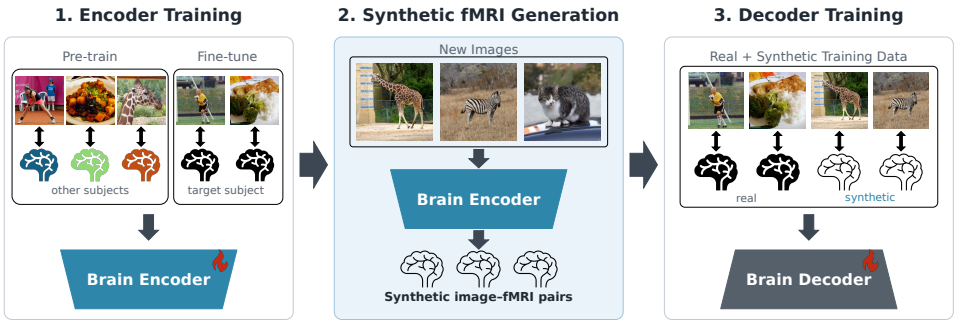


Figure 2: **Encoder-generated synthetic fMRI for decoder training.** The encoder is first pre-trained and fine-tuned to the target subject (step 1), applied to a new pool of images to generate synthetic fMRI (step 2), and the synthetic pairs are added to the decoder's training set (step 3).

introduce the semantic axis as the concept generalization setting (Sec. 3.1), and the quantity axis as the limited-data decoding setting (Sec. 3.2).

3.1 Concept Generalization: Limiting the Semantic Coverage

The concept generalization setting (Figure 1a) measures how a decoder behaves on a concept for which the target subject contributed no training data. Given a target concept $\mathcal{C}$, a concept which we want to measure the generalization capability for, we exclude every image containing $\mathcal{C}$ from the training set of the target subject. The model is then evaluated only on images containing $\mathcal{C}$.

Defining the concept. To introduce a large semantic gap between train and test time, we take $\mathcal{C}$ to be a COCO supercategory, for example *animal*. We filter images using their ground-truth COCO annotations, where an image is considered to contain $\mathcal{C}$ if its annotation includes at least one instance of any subcategory belonging to $\mathcal{C}$. We use ground-truth annotations rather than clusters in a learned feature space (Shirakawa et al., 2025) for two reasons. A cluster-based holdout retains training images in which the concept is present but not dominant, which leaks the concept into training; the annotation rule removes it entirely. It also keeps the holdout independent of any model's feature space, so that the holdout and the evaluation are not defined by the same representation. We use all 12 supercategories, each defining its own training set and separately trained encoders and decoders.

Can concept exposure from other subjects solve concept generalization? Given that recent decoding and encoding methods optionally allow multi-subject training, whether $\mathcal{C}$ is included for the other subjects must also be accounted for. For the main experiments we assume $\mathcal{C}$ remains available for the subjects used in pre-training, which we call the **transfer-concept setting**. For multi-subject models, this gives an opportunity to learn about $\mathcal{C}$ from other brains, a question that per-subject holdout protocols (Shirakawa et al., 2025) cannot pose. It also makes the setting deliberately favorable to existing models: if they still lose performance, concept generalization cannot be achieved by their inherent cross-subject transfer capabilities and a different solution is needed. We additionally explore settings where $\mathcal{C}$ is also absent for other subjects, as part of Sec. 6.

3.2 Limited-data Decoding: Limiting the Quantity

Most existing work on data efficiency instead limits the amount of data: the target subject's training set is subsampled to a small budget while the pre-training subjects retain their full data. We follow the same protocol and specify the exact budgets in Sec. 5.

We note that the two settings are not entirely independent. Subsampling a training set may underrepresent or even exclude some of the semantic content. For example, for our random 225-image training splits, equivalent to the *1 hour* splits from previous work (Scotti et al., 2024; Beliy et al., 2026), on average 31.5 out of the 80 COCO categories appear on 5 or fewer images from the training set. The more aggressively the training data is limited, the more of the concept space will be effectively unseen, and the more a few-shot problem also turns into a concept generalization problem.

4 GENERATING SYNTHETIC TRAINING DATA VIA BRAIN ENCODERS

4.1 MOTIVATION FOR USING BRAIN ENCODER

Using encoder-generated synthetic data for concept generalization is motivated by the asymmetry illustrated in Figure 1b. To compare the encoder and decoder on similar footing, we score the encoder by *decoded recall* (defined in Sec. 6): its predicted fMRI for images with $\mathcal{C}$ is reconstructed by a fixed MindEye2 readout trained with $\mathcal{C}$ present, and scored with the concept metrics given in Sec. 5.1. Under this holdout, the encoder retains 87% of Object Recall and 93% of Relaxed Recall, whereas MindEye2 retains 22% and 37%, respectively. The encoder is therefore more robust than the decoder, which motivates using its predictions to train decoders; Sec. 6 further examines why.

4.2 RECIPE FOR GENERATING SYNTHETIC TRAINING DATA

The cost of paired image-fMRI data lies in acquiring the fMRI, not the image. A brain encoder can therefore supply the missing half by predicting responses to novel images, and the resulting pairs can be used to train brain decoders. The procedure is summarized in Figure 2.

The recipe. Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ denote the target subject's training set, where x_i is an image and y_i the corresponding measured fMRI response. We train a brain encoder $E_\phi : x \mapsto \hat{y}$ on $\mathcal{D}$, initializing it from an encoder pre-trained on other subjects, so the encoder sees no target-subject fMRI beyond what the decoder already has. Given a new pool of images $\mathcal{U} = \{u_j\}$, we form a synthetic training set $\mathcal{S} = \{(u_j, E_\phi(u_j)) : u_j \in \mathcal{U}\}$ and train the decoder on $\mathcal{D} \cup \mathcal{S}$. The choice of $\mathcal{U}$ is the one component that varies across our experiments, and we specify it for each setting in Sec. 5. Because this is a data-level intervention, the synthetic pairs can train any decoding architecture without modification, regardless of its input shape or training strategy.

5 EXPERIMENTAL RESULTS

5.1 EXPERIMENTAL SETTING

All experiments are conducted on the Natural Scenes Dataset (NSD) (Allen et al., 2022). Following common practice, we experiment with four subjects (1, 2, 5 and 7) who completed all scanning sessions, and report the averaged result. We experiment with three image decoders: Brain-Diffuser (Ozcelik & VanRullen, 2023), MindEye2 (Scotti et al., 2024), and Brain-IT (Beliy et al., 2026). Decoders are trained with multiple subjects if possible: Brain-Diffuser can only be trained with a single subject, MindEye2 is first pre-trained with other subjects and then fine-tuned, and Brain-IT is jointly trained with all subjects. Refer to Appendix A for more details.

Note that although Brain-IT includes external synthetic data training in the original paper, similar in spirit to our synthetic data training, we consider this process to be a training technique independent of the core Brain-IT architecture. To keep the training process fair and consistent with other decoding models, and to measure the precise effect of synthetic data, we use the base Brain-IT model without this external data augmentation, just like the other models.

We use the Universal Brain Encoder (Beliy et al., 2024) as our synthetic-data generator. The encoder is pre-trained on other subjects and adapted to the target subject, using the same training data.

Evaluation. We report the standard decoding metrics: low-level (Pixelwise correlation, SSIM, Alex(2), Alex(5)) and high-level (Inception, CLIP, EffNet and SwAV). For concept generalization experiments, we additionally report *Object Recall* and *Relaxed Object Recall*, adapted from SEED (Park et al., 2026): using object detectors, we measure the ratio of reconstructions that exactly contain the object (Recall) or at least contain the same supercategory (Relaxed Recall), limited to objects that belong to $\mathcal{C}$. For example, if $\mathcal{C}$ is *animal*, and the ground truth includes a zebra, Recall counts the reconstruction correct if it also includes a zebra, while Relaxed Recall counts it correct if it contains any animal. Identification metrics compare global embeddings against a distractor and can remain above chance even when the held-out concept is absent from the reconstruction; Shirakawa et al. (2025) report exactly this discrepancy under a semantic holdout. We therefore aim to measure concept recovery directly through these metrics. Refer to Appendix B for more details.

Table 1: **Concept generalization for decoders.** Decoders are trained with/without $\mathcal{C}$ for the target subject and evaluated on images containing $\mathcal{C}$. Adding synthetic data improves generalization towards $\mathcal{C}$. Bold marks the better of the two unseen variants, with or without synthetic data.

		High-level				Object Detection	
Decoder	Concept $\mathcal{C}$	Incep ↑	CLIP ↑	EffNet ↓	SwAV ↓	Recall ↑	Relaxed ↑
Brain-Diffuser	Seen	90.2%	91.7%	0.761	0.397	0.330	0.584
	Unseen	81.2%	82.5%	0.847	0.458	0.073	0.189
	Unseen + syn.	**88.7%**	**90.5%**	**0.781**	**0.408**	**0.273**	**0.532**
MindEye2	Seen	95.9%	93.0%	0.636	0.340	0.443	0.662
	Unseen	88.5%	83.5%	0.767	0.410	0.096	0.242
	Unseen + syn.	**93.4%**	**89.0%**	**0.697**	**0.371**	**0.288**	**0.482**
Brain-IT	Seen	97.3%	95.0%	0.600	0.331	0.495	0.672
	Unseen	94.9%	91.9%	0.659	0.360	0.312	0.490
	Unseen + syn.	**96.0%**	**93.5%**	**0.637**	**0.349**	**0.386**	**0.565**

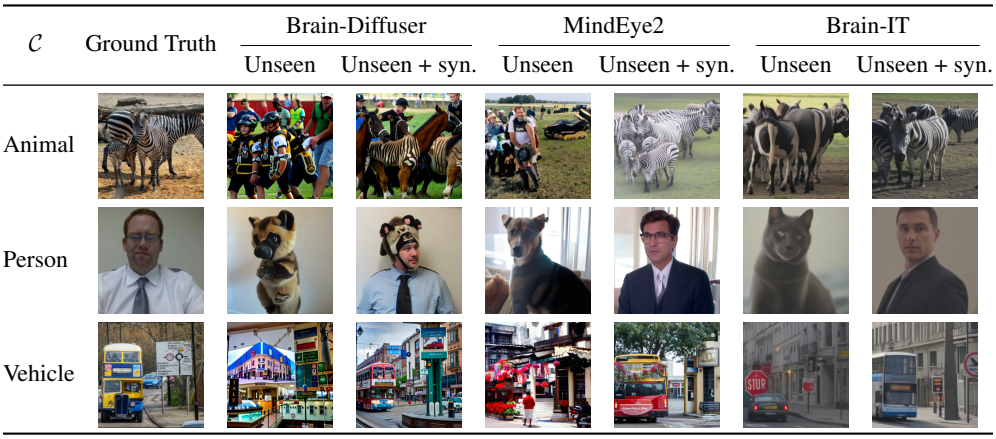


Figure 3: **Qualitative reconstruction highlights.** We show the ground truth image and the corresponding reconstructions for each decoder, before or after adding synthetic data to training. While the baseline model fails to accurately reconstruct $\mathcal{C}$, the synthetic variant is able to recover it.

5.2 Using Synthetic Data for Concept Generalization

Table 1 reports the performance of each decoder depending on whether $\mathcal{C}$ was included in the target subject's training data and whether synthetic data was added. To generate synthetic fMRI, we use the images that are removed from the training set, to directly match the baseline condition (unseen) and the synthetic condition (unseen + syn.). Low-level metrics, which the semantic holdout affects far less, are reported in Appendix C.2.

All three decoders show a concept generalization gap, largest on the recall metrics. This includes MindEye2 and Brain-IT, which were trained with $\mathcal{C}$ available from other subjects; as anticipated in Sec. 3.1, cross-subject training alone does not transfer $\mathcal{C}$ to the target subject. The gap is specific to $\mathcal{C}$: on test images without $\mathcal{C}$, the same models differ by an order of magnitude less (Appendix C.1).

When synthetic data is supplied, it closes a substantial portion of the gap. Weaker decoders recover more of the lost performance: Brain-Diffuser recovers 76-87%, MindEye2 recovers 53-66%, and Brain-IT recovers 37-53% of the lost performance across high-level and recall metrics.

Figure 3 shows a few qualitative examples on how the baseline decoder fails to reconstruct the object belonging to $\mathcal{C}$ and how the synthetic-trained decoder is able to reconstruct the object.

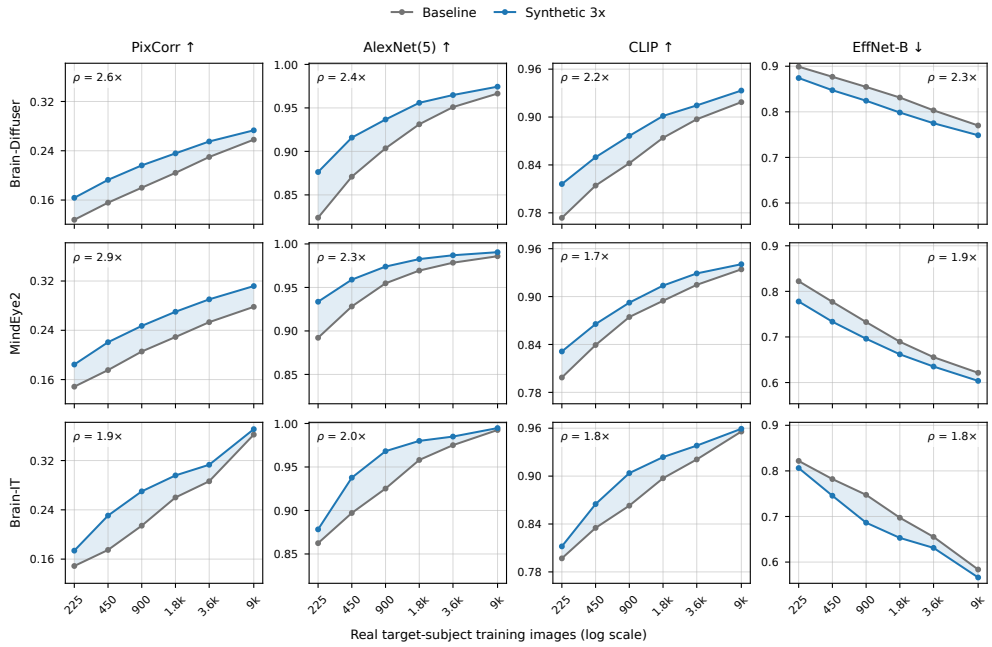


Figure 4: **Effect of synthetic data across data scales**. The x axis is in log scale. ρ denotes the fitted data-efficiency gain. We show 4 metrics spanning the metric groups; the rest are in Appendix E.3.

5.3 Using Synthetic Data Across Data Scales

We now turn to the quantity axis of Sec. 3.2, training decoders on $x \in \{225, 450, 900, 1800, 3600, 9000\}$ real target-subject images, with and without additional synthetic data. The synthesis pool $\mathcal{U}$ consists of NSD images shown to other subjects but never to the target subject, with $|\mathcal{U}| = 3x$ so that real and synthetic pairs contribute equal training volume: each real image carries three trials, while the encoder produces one deterministic response per image. Appendix E.4 varies this ratio between real and synthetic data.

Synthetic data outperforms the baseline at every budget on every metric in Figure 4, and both curves keep rising up to the full 9,000-image budget, indicating that the decoders are not yet saturated by target-subject data. The remaining metrics behave the same way (Appendix E.3), with the exception of SSIM, which varies little with budget and synthetic data does not consistently help it.

Quantifying the data-efficiency gain. The benefit can equivalently be read as the real data collection it saves. For each metric we fit the single factor ρ that aligns the baseline curve with the synthetic one, $\hat{\rho} = \arg\min_{\rho} \sum_x \big(s(x) - \hat{b}(\rho\, x)\big)^2$, where $b(x)$ and $s(x)$ are the baseline and synthetic scores at x real target-subject images and $\hat{b}$ is a monotone interpolant of the six measured budgets, held at its 9,000-image value beyond the measured range. The resulting $\hat{\rho}$, annotated in each panel of Figure 4, reads directly as the amount of real data the synthetic data is worth: $\hat{\rho} = 2$ means the same performance would otherwise have required twice as many scans of the target subject. Across the four metrics in Figure 4 and all three decoders, $\hat{\rho}$ ranges from 1.7 to 2.9 (the remaining metrics fall in the same range; confidence intervals in Appendix G.3).

Extension to video decoding. To test whether the recipe transfers beyond static images, we apply it to fMRI-to-video decoding with NEURONS (Wang et al., 2025) on the cc2017 dataset (Wen et al., 2018), training on 10% or the full amount of data. The same Universal Encoder, jointly pre-trained on NSD and the other cc2017 subjects, predicts a response from a single frame of each clip. Synthetic data improves the video decoder at both budgets, with the largest absolute gains on the semantic metrics: at the 10% budget, video-level 50-way accuracy rises from 0.185 to 0.235, recovering 68% of the gap to the full-data model, and at the full budget the gains remain significant on most metrics (Appendix F). That a still-image encoder improves a video decoder suggests the recipe applies beyond image decoding.

Table 2: **Encoder variants as synthetic-data sources (supercategory generalization).** Variants differ in whether $\mathcal{C}$ was available during cross-subject pre-training and target-subject adaptation (Target). *Probe decoder*: a from-scratch MindEye2 trained on real data plus that encoder's synthetic pairs, evaluated on images containing $\mathcal{C}$; the two oracle variants see real responses to $\mathcal{C}$ and cannot be used for synthesis. *Encoder*: Pearson r of the predicted fMRI, and decoded recall / decoded relaxed recall (D.Rec. / D.Rel.) of images reconstructed from it.

	$\mathcal{C}$ incl. for		Probe decoder trained with synthetic data						Encoder		
Encoder variant	Other	Target	Incep ↑	CLIP ↑	EffNet ↓	SwAV ↓	Rec. ↑	Rel. ↑	r ↑	D.Rec. ↑	D.Rel. ↑
Scratch-oracle	–	✓	*n/a: trained on real fMRI for $\mathcal{C}$*						0.303	0.497	0.686
Transfer-oracle	✓	✓							0.350	0.574	0.735
Scratch-blind	–	✗	91.4%	86.9%	0.729	0.390	0.218	0.427	0.277	0.290	0.492
Transfer-blind	✗	✗	91.4%	87.2%	0.723	0.387	0.229	0.441	0.337	0.379	0.585
Transfer-concept	✓	✗	92.8%	88.5%	0.709	0.381	0.279	0.486	0.344	0.499	0.685
No synthetic	–		87.2%	82.5%	0.777	0.418	0.087	0.228	*n/a: no encoder*		
Real fMRI for $\mathcal{C}$	–		95.3%	92.1%	0.647	0.346	0.411	0.643			

6 Why does Encoder-generated Data Help?

The encoder is trained on the same image-fMRI pairs available to the decoder, thus it uses no additional information from the data, yet its synthetic responses help the decoder. While related work attributes this to data augmentation or data enlargement (Mai et al., 2025; Beliy et al., 2026), the large improvement for concept generalization suggests the synthetic data contain additional information for $\mathcal{C}$. We first test whether this concept-specific information indeed drives the improvement, then vary the encoder's training conditions to locate where its knowledge of $\mathcal{C}$ comes from, and finally ask whether standard encoder metrics can predict the value of its synthetic data.

Concept-specific content of synthetic data matters. A simpler explanation of the gains in Sec. 5.2 is that any additional training pairs help. We test this on subject 1 with two controls that add the same number of pairs as the main condition, using encoder predictions for images without $\mathcal{C}$: *non-$\mathcal{C}$ images* pairs these predictions with the images they were predicted for, adding data but no information about $\mathcal{C}$; *wrong fMRI* pairs them with the $\mathcal{C}$ images instead, so the decoder sees $\mathcal{C}$ as a target but never with the response it elicits. The controls recover 2% and 12% of the Object Recall gap between the unseen decoder and one trained on real fMRI for $\mathcal{C}$, against 64% for correctly paired synthetic data (Appendix D.3). Neither more data nor exposure to $\mathcal{C}$ images alone suffices; the gain requires the encoder's response to $\mathcal{C}$ itself.

Comparing encoder variants. Table 2 defines five variants along two factors: whether $\mathcal{C}$ was available during cross-subject pre-training and during target-subject adaptation. The oracle variants observe real target-subject responses to $\mathcal{C}$ and cannot serve as synthesis sources; they are reported as upper bounds. Each of the three remaining variants is evaluated by a *probe decoder*, a MindEye2 trained from scratch (to prevent cross-subject information leak from decoder pre-training) on real data plus that variant's synthetic pairs. Transfer-concept is the variant used throughout Sec. 5.

Evaluating encoders. We score each encoder in two ways. *Fidelity* is the Pearson r between predicted and measured fMRI. *Decoded recall* measures how semantically decodable a prediction is: we reconstruct images from the encoder's predictions with a decoder, a MindEye2 trained with $\mathcal{C}$ present from the full budget plus synthetic data, and apply the concept metrics of Sec. 5.1. Since the decoder has seen $\mathcal{C}$, it can detect and recover the semantic content from the encoder's prediction. As the decoder is not a perfect judge, we intend to use this only to compare encoders (Appendix D.1).

Where does the encoder's knowledge of $\mathcal{C}$ come from? We read Table 2 first by the probe decoder, since that is the quantity we ultimately care about. Scratch-blind, which never sees fMRI for $\mathcal{C}$ from any subject, already recovers 41% of the Object Recall gap; cross-subject transfer is therefore not required, and this baseline robustness must come from the encoder itself, which we trace to its pretrained backbone below. At the full budget, pre-training on other subjects without $\mathcal{C}$ (transfer-blind) adds almost nothing: Object Recall moves from 0.218 to 0.229, and the two probe decoders are statistically equivalent (Appendix G.2). Exposure to $\mathcal{C}$ in the pre-training subjects

(transfer-concept) is what adds: the probe decoder improves on every metric in Table 2, by $+0.049$ in Object Recall, and reaches 59% of the gap. The gain thus has two sources: the encoder's inherent robustness to unseen concepts, and its transfer of concept exposure from other brains (Sec. 5.2). At the finer granularity of single COCO categories (category generalization), every effect shrinks with the semantic gap and the concept-specific transfer is no longer significant (Appendix D.2).

Encoder metrics do not predict the value of its synthetic data. One might expect a more accurate encoder to yield better synthetic data. The two transfer steps above show otherwise: the step that barely helps the probe decoder, pre-training without $\mathcal{C}$, raises Pearson r by 0.060, while the step that helps on every metric, exposure to $\mathcal{C}$, raises it by only 0.007. Pre-training makes the predicted response more accurate overall; seeing $\mathcal{C}$, from the target subject or from others, is what puts $\mathcal{C}$ into it. Decoded recall is designed to measure the latter and does recover this ordering (scratch-oracle exceeds transfer-blind by 0.118 and matches transfer-concept), but it does not track the probe decoder either: it credits the two transfer steps about equally ($+0.089$ and $+0.119$), whereas the probe decoder credits only the second.

The encoder generalizes where the decoder does not. The scratch-blind result shows that the encoder predicts useful responses to a concept it never saw in any subject's fMRI, and that a decoder trained on those predictions inherits part of that ability. Since no fMRI for $\mathcal{C}$ enters that encoder, this robustness must come from its pretrained visual backbone: the DINOv2 representation already covers $\mathcal{C}$, and a per-voxel readout fit on other concepts extends to it. Consistent with this, freezing the backbone instead of LoRA-tuning it leaves the decoder gains unchanged (Appendix D.4). This alone does not explain the asymmetry, since the decoders also build on pretrained components (CLIP and diffusion priors) that have seen $\mathcal{C}$ and still fail. The difference is where that pretrained knowledge sits. For the encoder it sits at the input: $\mathcal{C}$ images already land in a representation that handles them, and only a smooth map from that representation to noisy voxels has to be learned. For the decoder it sits at the output: the decoder must learn which fMRI patterns to map there, and for $\mathcal{C}$ it is never shown any. Synthetic pairs supply exactly the missing supervision, with quality bounded by the encoder. Semantic coverage is therefore an axis along which the encoder degrades gracefully and the decoder does not, and encoder-generated data is a way to move the decoder along it.

7 Conclusion & Limitations

We showed that in the concept generalization setting, in which a semantic concept is withheld from the target subject's training data, three recent fMRI-to-image decoders substantially lose their ability to reconstruct the held-out concept, even when it remains available from other subjects. Brain encoders trained on the same data are far more robust to this holdout, and generating synthetic fMRI for images of the concept transfers part of that robustness to decoders: across three decoders, four subjects, and twelve concepts, it recovers 37-87% of the gap on high-level and concept-recall metrics without any additional scanning. Our analyses indicate that this gain stems from the encoder's ability to predict responses to concepts absent from its fMRI training and from its transfer of concept exposure across subjects, rather than from added data volume. Beyond the holdout, the same procedure improves decoding on nearly every metric at every data budget we tested, by an amount equivalent on average to roughly doubling target-subject scan time. More broadly, our results suggest that decoders should be evaluated on concepts outside their training distribution, and that encoders offer a practical route to covering them.

Limitations. Our suggested solution for concept generalization has two main limitations. First, synthetic neural responses are not a perfect substitute for real ones: under an equal training budget, real fMRI for $\mathcal{C}$ still outperforms encoder-generated fMRI. Synthetic data should therefore be used when collecting real responses to $\mathcal{C}$ is impractical, not as a replacement for them. Second, the method requires knowing $\mathcal{C}$ before training, since synthetic pairs must be generated for images of $\mathcal{C}$ and included in the training set; a concept that becomes relevant only after the decoder is trained cannot be accounted for without retraining or fine-tuning.

AI USE STATEMENT

In this work, we used generative AI tools to assist with refining research methodology, assist with implementing code for experiments and data analysis, and to assist and polish writing for the manuscript. We did not use generative AI tools for the remaining required disclosure categories.

We additionally used generative AI to write code that is used to generate figures and plots. All generated code and figures created by those code were carefully reviewed by the authors. We take full responsibility for the final content of this work, including all text, claims, code, figures, and other artifacts produced with the assistance of generative AI.

REPRODUCIBILITY STATEMENT

We provide detailed experimental settings in the main manuscript and in the appendix. We plan to release the code used for experiments as well as related data and results after the review process concludes.

REFERENCES

Emily J Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Jacob S Prince, Logan T Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, et al. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25(1):116–126, 2022.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1708–1718. IEEE, 2021.

Roman Beliy, Navve Wasserman, Amit Zalcher, and Michal Irani. The wisdom of a crowd of brains: A universal brain encoder. *arXiv preprint arXiv:2406.12179*, 2024.

Roman Beliy, Amit Zalcher, Jonathan Kogman, Navve Wasserman, and Michal Irani. Brain-it: Image reconstruction from fmri via brain-interaction transformer. In *International Conference on Learning Representations*, 2026.

Yohann Benchetrit, Marlène Careil, Simon Dahan, Hubert Banville, Stéphane d'Ascoli, and Jean-Rémi King. Boosting brain-to-image decoding with tribe v2 data augmentation. *arXiv preprint arXiv:2606.06345*, 2026.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

Stéphane d'Ascoli, Jérémy Rapin, Yohann Benchetrit, Teon Brooks, Katelyn Begany, Joséphine Raugel, Hubert Banville, and Jean-Rémi King. A foundation model of vision, audition, and language for in-silico neuroscience. *arXiv preprint arXiv:2605.04326*, 2026.

Guy Gaziv, Roman Beliy, Niv Granot, Assaf Hoogi, Francesca Strappini, Tal Golan, and Michal Irani. Self-supervised natural image reconstruction and large-scale semantic classification from brain activity. *NeuroImage*, 254:119121, 2022.

Alessandro T Gifford, Benjamin Lahner, Sari Saba-Sadiya, Martina G Vilas, Alex Lascelles, Aude Oliva, Kendrick Kay, Gemma Roig, and Radoslaw Martin Cichy. The algonauts project 2023 challenge: How the human brain makes sense of natural scenes. *arXiv preprint arXiv:2301.03198*, 2023.

Zixuan Gong, Guangyin Bao, Qi Zhang, Zhongwei Wan, Duoqian Miao, Shoujin Wang, Lei Zhu, Changwei Wang, Rongtao Xu, Liang Hu, et al. Neuroclips: Towards high-fidelity and smooth fmri-to-video reconstruction. *Advances in Neural Information Processing Systems*, 37:51655–51683, 2024.

Tomoyasu Horikawa and Yukiyasu Kamitani. Generic decoding of seen and imagined objects using hierarchical visual features. *Nature communications*, 8(1):15037, 2017.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Jingyang Huo, Yikai Wang, Yun Wang, Xuelin Qian, Chong Li, Yanwei Fu, and Jianfeng Feng. Neuropictor: Refining fmri-to-image reconstruction via multi-individual pretraining and multi-level modulation. In *European Conference on Computer Vision*, pp. 56–73. Springer, 2024.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

Chong Li, Xuelin Qian, Yun Wang, Jingyang Huo, Xiangyang Xue, Yanwei Fu, and Jianfeng Feng. Enhancing cross-subject fmri-to-video decoding with global-local functional alignment. In *European Conference on Computer Vision*, pp. 353–369. Springer, 2024.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Weijian Mai, Jiamin Wu, Yu Zhu, Zhouheng Yao, Dongzhan Zhou, Andrew Luo, Qihao Zheng, Wanli Ouyang, and Chunfeng Song. Synbrain: Enhancing visual-to-fmri synthesis via probabilistic representation learning. *Advances in Neural Information Processing Systems*, 38:29253–29284, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Furkan Ozcelik and Rufin VanRullen. Natural scene reconstruction from fmri signals using generative latent diffusion. *Scientific Reports*, 13(1):15666, 2023.

Juhyeon Park, Peter Kim, Jiook Cha, Shinjae Yoo, and Taesup Moon. Seed: Towards more accurate semantic evaluation for visual brain decoding. In *International Conference on Learning Representations*, 2026.

Paul Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Aidan Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth Norman, et al. Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors. *Advances in Neural Information Processing Systems*, 36:24705–24728, 2023.

Paul S Scotti, Mihir Tripathy, Cesar Kadir Torrico Villanueva, Reese Kneeland, Tong Chen, Ashutosh Narang, Charan Santhirasegaran, Jonathan Xu, Thomas Naselaris, Kenneth A Norman, et al. Mindeye2: shared-subject models enable fmri-to-image with 1 hour of data. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 44038–44059, 2024.

Ken Shirakawa, Yoshihiro Nagano, Misato Tanaka, Shuntaro C Aoki, Yusuke Muraki, Kei Majima, and Yukiyasu Kamitani. Spurious reconstruction from brain activity. *Neural Networks*, 190: 107515, 2025.

Ghislain St-Yves and Thomas Naselaris. Generative adversarial networks conditioned on brain activity reconstruct seen images. In *2018 IEEE international conference on systems, man, and cybernetics (SMC)*, pp. 1054–1061. IEEE, 2018.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pp. 6105–6114. PmLR, 2019.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.

Haonan Wang, Qixiang Zhang, Lehan Wang, Xuanqi Huang, and Xiaomeng Li. Neurons: Emulating the human visual cortex improves fidelity and interpretability in fmri-to-video reconstruction. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18367–18376. IEEE, 2025.

Shizun Wang, Songhua Liu, Zhenxiong Tan, and Xinchao Wang. Mindbridge: A cross-subject brain decoding framework. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11333–11342. IEEE, 2024.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Haiguang Wen, Junxing Shi, Yizhen Zhang, Kun-Han Lu, Jiayue Cao, and Zhongming Liu. Neural encoding and decoding with deep learning for dynamic natural vision. *Cerebral cortex*, 28(12): 4136–4160, 2018.

Qi Zhang, Zixuan Gong, Zhuojia Wu, and Duoqian Miao. Mindsimulator: Exploring brain concept localization via synthetic fmri. In *International Conference on Learning Representations*, 2025.

Xiangyu Zhao, Yicheng Chen, Shilin Xu, Xiangtai Li, Xinjiang Wang, Yining Li, and Haian Huang. An open and comprehensive pipeline for unified object grounding and detection. *arXiv preprint arXiv:2401.02361*, 2024.

Concept Generalization in Visual Brain Decoding via Encoder-Generated fMRI

Appendix

A Additional Experimental Details

Decoder training protocols. As in the original paper, MindEye2 is first pre-trained on seven other subjects and then fine-tuned to the target subject. We utilized the pre-trained MindEye2 checkpoints provided by the authors, and only fine-tuned the models ourselves. Brain-IT is jointly trained with all eight subjects. Brain-Diffuser is trained from scratch on the target subject as it offers no multi-subject training method.

Decoder input format and the corresponding synthetic fMRI. Depending on the decoding model, the input fMRI voxel space may differ. While all models use single-trial GLM beta responses as the input, MindEye2 and Brain-Diffuser use the *nsdgeneral* region supplied by the NSD authors which includes around 15k voxels, Brain-IT uses a larger region (Gifford et al., 2023), resulting in around 40k voxels. Therefore, separate versions of encoders were trained for each fMRI voxel format, and they were used to generate appropriate fMRI inputs for each decoder. For all evaluations of the encoder outputs such as Pearson correlation, the results for the *nsdgeneral* encoder (corresponding to MindEye2 and Brain-Diffuser inputs) were reported.

Additional details for evaluation metrics. The low-level and high-level evaluation metrics (Inception, CLIP, EfficientNet and SwAV) follow the protocol of previous brain decoding literature: AlexNet(2), AlexNet(5), Inception and CLIP are two-way identification metrics, where the respective embeddings for the ground truth image, the reconstruction, and another reconstruction are extracted, and the percentage chance of the reconstruction being closer to the ground-truth than to another reconstruction is reported. EfficientNet and SwAV report the average correlation distance, calculated using embeddings extracted from the respective models.

Data counts for each supercategory. Table 3 lists the number of training images containing $\mathcal{C}$ for each target subject (out of 9,000) as well as the number of test images containing $\mathcal{C}$ (out of 1,000). Images may contain more than one supercategory, so columns do not sum to the set size.

Subsampling method for limited data experiments. Unlike previous work, which subsamples by scan session and therefore retains an uneven number of repetitions per image, we subsample by unique image and keep all of its repetitions, giving cleaner control over the budget.

B Evaluation Metrics

We report the standard set of low-level and high-level reconstruction metrics used in prior work (Scotti et al., 2023; 2024; Beliy et al., 2026). *PixCorr* is the pixel-wise correlation between the ground-truth and the reconstructed image, and *SSIM* is the grayscale structural similarity index (Wang et al., 2004); both are low-level. *Alex(2)* and *Alex(5)* are two-way identification accuracies computed from the second and fifth layers of AlexNet, and *Incep* and *CLIP* are the corresponding accuracies from InceptionV3 pool features and the final layer of CLIP ViT-L/14, forming a sequence of increasing abstraction. *Eff* and *SwAV* report the average correlation distance in EfficientNet-B1 (Tan & Le, 2019) and SwAV-ResNet50 (Caron et al., 2020) feature space, for which lower is better.

All two-way metrics ask whether a reconstruction is closer, in the relevant feature space, to its own stimulus than to an another reconstruction. They are therefore comparative, and a reconstruction that is structurally plausible but semantically empty can pass them; this is the reason we adapt the object-detection metrics of Sec. 5.1 for the semantic-holdout experiments.

Table 3: **Size of each concept $\mathcal{C}$.** The 12 COCO supercategories used for supercategory generalization, each followed by the two categories (indented) used for category generalization. *Person* has no underlying category. Train counts are out of 9,000 images per subject and test counts out of 1,000. Note that each image may contain several categories.

	Train images containing $\mathcal{C}$					Test set	
Concept $\mathcal{C}$	Subj 1	Subj 2	Subj 5	Subj 7	Mean %	Count	%
Accessory	1,281	1,237	1,232	1,211	13.8%	97	9.7%
Backpack	393	411	413	404	4.5%	29	2.9%
Handbag	465	478	473	446	5.2%	43	4.3%
Animal	1,934	1,929	1,924	1,943	21.5%	266	26.6%
Bird	258	239	274	271	2.9%	39	3.9%
Giraffe	222	208	215	219	2.4%	44	4.4%
Appliance	633	625	613	615	6.9%	66	6.6%
Oven	233	229	216	214	2.5%	21	2.1%
Sink	381	379	375	386	4.2%	49	4.9%
Electronic	839	867	852	783	9.3%	81	8.1%
Cell phone	328	333	318	292	3.5%	31	3.1%
TV	290	292	298	284	3.2%	32	3.2%
Food	1,194	1,192	1,211	1,217	13.4%	103	10.3%
Broccoli	140	133	155	159	1.6%	22	2.2%
Orange	103	147	136	135	1.4%	21	2.1%
Furniture	2,121	2,070	2,106	2,078	23.3%	186	18.6%
Chair	896	824	831	907	9.6%	84	8.4%
Dining table	816	797	810	771	8.9%	75	7.5%
Indoor	1,163	1,149	1,160	1,215	13.0%	134	13.4%
Book	393	384	394	369	4.3%	40	4.0%
Clock	368	382	392	408	4.3%	46	4.6%
Kitchen	1,508	1,439	1,487	1,446	16.3%	125	12.5%
Bottle	613	633	627	606	6.9%	49	4.9%
Cup	653	657	658	625	7.2%	56	5.6%
Outdoor	959	1,051	1,006	1,002	11.2%	101	10.1%
Bench	402	466	427	413	4.7%	50	5.0%
Traffic light	301	339	307	322	3.5%	35	3.5%
Person	4,636	4,567	4,533	4,488	50.6%	438	43.8%
Sports	1,850	1,744	1,763	1,773	19.8%	200	20.0%
Skis	297	250	213	259	2.8%	43	4.3%
Surfboard	309	319	321	303	3.5%	48	4.8%
Vehicle	2,073	2,149	2,111	2,084	23.4%	211	21.1%
Car	876	927	919	884	10.0%	74	7.4%
Train	286	305	294	300	3.3%	36	3.6%

B.1 Object Recall and Relaxed Recall

In order to supplement standard high-level metrics that compare global embeddings and can score a reconstruction highly even when the object of interest is absent, we adapt metrics that focus on whether the target concept itself was recovered. We follow SEED (Park et al., 2026) and run an object detector, namely MM-Grounding-DINO (Zhao et al., 2024), on both the ground-truth image and the reconstruction, then ask whether the detected content matches. Formally, we index every quantity by the subject $s \in \mathcal{S}$ and the decoder variant $v \in \mathcal{V}$. Let $\mathcal{O}$ ($|\mathcal{O}| = 80$) be the set of COCO subcategories and $\mathcal{G}$ the set of 12 COCO supercategories, which are the same supercategories as the held-out concept $\mathcal{C}$ in the main text. Each subcategory has a parent supercategory given by the map $\pi : \mathcal{O} \to \mathcal{G}$, and for a set $X \subseteq \mathcal{O}$ we write $\pi(X) = \{\pi(o) : o \in X\}$.

We prompt the detector with all 80 COCO class names and, on both ground-truth and reconstructed images, discard detections with confidence below 0.3 or bounding-box area below 50 pixels, to filter noisy results.

Which model produces the reconstructions. Each holdout excludes an entire supercategory, so there are 12 separately trained models per subject and variant. Both metrics for a supercategory γ, including the recall of every subcategory o with $\pi(o) = \gamma$, are always computed from the model trained with $\mathcal{C} = \gamma$: for example, *zebra* is scored with the animal-excluded model, *bus* with the vehicle-excluded model, and so on. This keeps every number a concept measurement, and it is consistent with the test set restriction of Sec. 5, since the test set of that model consists precisely of images containing γ.

The qualifying sets. For a test image i, write $A_i \subseteq \mathcal{O}$ for object categories present in the ground-truth COCO annotation, $D_i \subseteq \mathcal{O}$ for the detected object categories on the ground-truth image, and $\widehat{D}_i \subseteq \mathcal{O}$ for the detected object categories on its reconstruction. We only score an image if the COCO annotation and the detector run on the *ground-truth* image agree that the object being measured is present. The annotation confirms the object is genuinely in the scene, and the ground-truth detection confirms the detector is able to find it without any detector-side error. For a subcategory o, the qualifying set $\mathcal{Q}_o$ therefore contains the test images for which both A_i and D_i include o. For a supercategory γ, the qualifying set $\mathcal{Q}_\gamma$ contains the test images for which both A_i and D_i include at least one object of γ.

This condition ensures that the trivial perfect reconstruction, which is a reconstruction identical to the ground truth, always gets a perfect score for both metrics. Otherwise, even the perfect reconstruction would not get a perfect score if any detector-side error occurs.

The two metrics. *Recall* is defined per subcategory: R_o is the fraction of images in $\mathcal{Q}_o$ whose reconstruction contains the exact subcategory o. If the ground-truth image shows a zebra, the reconstruction must contain a zebra. *Relaxed recall* is defined directly on the concept: $\widetilde{R}_\gamma$ is the fraction of images in $\mathcal{Q}_\gamma$ whose reconstruction contains any object of γ, so a reconstructed giraffe still counts even if the ground truth is a zebra, as a giraffe is still an animal just like a zebra. We define relaxed recall at the image level rather than per subcategory because its hit condition does not depend on which subcategory is present; computing it per subcategory would only reweight images, for instance counting an image that contains both a zebra and a giraffe twice with the same outcome. The relaxed variant is informative because a decoder that has never seen a concept often produces a semantically adjacent object rather than an unrelated one, and this near-miss behavior is invisible to exact recall. For any single image, an exact hit is also a relaxed hit, and $\mathcal{Q}_o \subseteq \mathcal{Q}_{\pi(o)}$; the two metrics are, however, averaged over different sets, so relaxed recall exceeding recall is an empirical observation rather than an identity. The gap between them indicates how much of the supercategory a decoder recovers without pinning down the specific object. Algorithm 1 states the full procedure.

Algorithm 1 Object recall and relaxed recall, for a fixed subject s and decoder variant v

Require: Test images $\mathcal{I}^\gamma$ and reconstruction detections $\widehat{D}_i^\gamma$ for every held-out supercategory $\gamma \in \mathcal{G}$; annotations A_i and ground-truth detections D_i.
Ensure: R_o for every $o \in \mathcal{O}$ and $\widetilde{R}_\gamma$ for every $\gamma \in \mathcal{G}$.

for all $o \in \mathcal{O}$ **do**
  $\gamma \leftarrow \pi(o)$ ▷ score o on the model that held out its parent
  $\mathcal{Q}_o \leftarrow \{\, i \in \mathcal{I}^\gamma : o \in A_i \ \wedge\ o \in D_i \,\}$ ▷ annotation and ground-truth detector agree
  $R_o \leftarrow \dfrac{1}{|\mathcal{Q}_o|} \sum_{i \in \mathcal{Q}_o} \mathbb{1}\left[o \in \widehat{D}_i^\gamma\right]$ ▷ Recall: reconstruction recovers o itself
end for
for all $\gamma \in \mathcal{G}$ **do**
  $\mathcal{Q}_\gamma \leftarrow \{\, i \in \mathcal{I}^\gamma : \gamma \in \pi(A_i) \ \wedge\ \gamma \in \pi(D_i) \,\}$ ▷ annotation and ground-truth detector agree
  $\widetilde{R}_\gamma \leftarrow \dfrac{1}{|\mathcal{Q}_\gamma|} \sum_{i \in \mathcal{Q}_\gamma} \mathbb{1}\left[\gamma \in \pi\big(\widehat{D}_i^\gamma\big)\right]$ ▷ Relaxed Recall: reconstruction recovers γ
end for
return $\{R_o\}_{o \in \mathcal{O}}$, $\{\widetilde{R}_\gamma\}_{\gamma \in \mathcal{G}}$

Aggregation. Recall is first macro-averaged over the subcategories within each supercategory; relaxed recall is already computed per supercategory. Both are then averaged across subjects, and

Table 4: **Concept generalization for decoders, evaluated on unrelated images.** Decoders are trained with/without $\mathcal{C}$ and evaluated on images that **do not** contain $\mathcal{C}$. Removing C's training data or adding synthetic data shifts these scores by an order of magnitude less than on images containing C; see Appendix C.1 and Table 12.

		Low-level				High-level			
Decoder	Concept $\mathcal{C}$	PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑	Incep ↑	CLIP ↑	EffNet ↓	SwAV ↓
Brain-Diffuser	Seen	0.260	0.358	94.5%	96.6%	88.0%	92.0%	0.772	0.423
	Unseen	0.255	0.357	94.1%	96.4%	87.6%	91.9%	0.773	0.425
	Unseen + syn.	0.260	0.357	94.6%	96.7%	87.9%	92.2%	0.770	0.421
MindEye2	Seen	0.281	0.378	95.8%	98.7%	94.9%	93.7%	0.619	0.345
	Unseen	0.276	0.379	95.2%	98.5%	94.6%	92.9%	0.628	0.350
	Unseen + syn.	0.285	0.380	95.7%	98.7%	95.0%	93.4%	0.618	0.344
Brain-IT	Seen	0.365	0.502	98.0%	99.3%	96.8%	95.7%	0.581	0.336
	Unseen	0.360	0.502	98.0%	99.2%	96.6%	95.7%	0.583	0.337
	Unseen + syn.	0.362	0.503	98.1%	99.2%	96.8%	95.7%	0.584	0.337

finally across the 12 supercategories with equal weight. Weighting supercategories equally prevents frequent concepts such as *person*, which appears in nearly half of the test images, from dominating the reported score, so that the metric reflects performance across concepts rather than across images.

Category generalization. For the 22 categories of Table 3 (two per non-person supercategory), the procedure differs in two respects. First, each category o is scored with the model trained with o itself held out, rather than with the model that held out $\pi(o)$; an image containing several of the 22 categories is therefore scored separately by each corresponding model. Second, both metrics use the exact-category qualifying set Q_o of Algorithm 1 (line 3), with I^o the test images containing o. Recall R_o is computed as before, while relaxed recall becomes

$$\widetilde{R}_o = \frac{1}{|Q_o|} \sum_{i \in Q_o} \mathbb{1}\left[\pi(o) \in \pi\big(\widehat{D}_i^{\,o}\big)\right], \tag{1}$$

so a reconstruction containing a sibling o' with $\pi(o') = \pi(o)$ is a relaxed hit. Unlike in the supercategory setting, such siblings remain in the target subject's training set, so at this granularity relaxed recall partly credits the decoder for substituting a seen, semantically adjacent object. Both metrics are averaged over subjects and then over the 22 categories with equal weight; since every parent contributes exactly two categories, this is equivalent to weighting the 11 parent supercategories equally.

C Additional Concept Generalization Results

C.1 Complementary Test Set Results

Table 4 shows performance of different decoders from the concept generalization experiments in Table 1, evaluated instead on the complementary test set that do not contain $\mathcal{C}$. In other words, while results from Table 1 are from images that contain $\mathcal{C}$, results from Table 4 are from the left-out images that do not contain $\mathcal{C}$. Results from this table show how adding or removing data from the training set in our experiments influences semantically unrelated images that do not contain $\mathcal{C}$.

Both manipulations have small but measurable effects on images that do not contain C (Table 12). Removing C's images from training lowers performance on unrelated images, which is expected since the Seen model has more training data (Table 3); the shifts are significant on all eight metrics for MindEye2, five for Brain-Diffuser, and three for Brain-IT. Adding synthetic data raises performance on unrelated images, consistent with the general benefit reported in Sec. 5.3; this is significant on all eight metrics for MindEye2, six for Brain-Diffuser, and only PixCorr for Brain-IT. (MindEye2's SSIM moves by 0.001 in the opposite direction under the holdout, at the resolution where that metric is insensitive to all of our manipulations.)

What matters for interpreting Table 1 is the magnitude of these shifts relative to those on images containing C. On the high-level metrics, the effect of the holdout on unrelated images is 1-8%

Table 5: **Concept generalization for decoders: low-level metrics.** Same models and evaluation images as Table 1: decoders are trained with/without $\mathcal{C}$ for the target subject and evaluated on images containing $\mathcal{C}$. Bold marks the better of the two unseen variants, with or without synthetic data.

		Low-level			
Decoder	Concept $\mathcal{C}$	PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑
Brain-Diffuser	Seen	0.245	0.352	93.8%	96.8%
	Unseen	0.221	0.350	91.7%	94.6%
	Unseen + syn.	**0.240**	**0.351**	**93.6%**	**96.7%**
MindEye2	Seen	0.255	0.358	95.1%	98.6%
	Unseen	0.217	0.353	92.3%	96.7%
	Unseen + syn.	**0.244**	**0.358**	**94.0%**	**98.1%**
Brain-IT	Seen	0.337	0.489	97.6%	99.2%
	Unseen	0.321	**0.490**	97.0%	98.7%
	Unseen + syn.	**0.327**	0.489	**97.4%**	**99.0%**

Table 6: **Encoder impact on decoding performance by training condition, concept generalization: low-level metrics.** Same models and evaluation images as Tables 2 and 8. Each row trains a MindEye2 decoder from scratch, varying the encoder that generated its synthetic pairs. The last row of each panel uses real fMRI for $\mathcal{C}$ as an upper bound.

	Low-level			
Training condition	PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑
Supercategory generalization (12 concepts)				
$\mathcal{C}$ unseen	0.209	0.351	91.3%	96.2%
+ syn. (scratch-blind)	0.232	**0.358**	93.0%	97.5%
+ syn. (transfer-blind)	0.231	0.357	93.1%	97.5%
+ syn. (transfer-concept)	**0.236**	0.358	**93.4%**	**97.7%**
Real fMRI for $\mathcal{C}$ (seen)	0.246	0.359	94.0%	98.3%
Category generalization (22 concepts)				
$\mathcal{C}$ unseen	0.235	0.348	93.3%	97.5%
+ syn. (scratch-blind)	0.248	0.350	94.1%	98.1%
+ syn. (transfer-blind)	0.249	0.351	**94.5%**	98.2%
+ syn. (transfer-concept)	**0.254**	**0.351**	94.3%	**98.3%**
Real fMRI for $\mathcal{C}$ (seen)	0.254	0.349	94.4%	98.4%

of its effect on concept images, and the effect of synthetic data is 3-14% for Brain-Diffuser and MindEye2 and indistinguishable from zero (not significant) for Brain-IT. The low-level metrics are less specific, as expected for a semantic manipulation: there the general effects reach around 10-30% of the concept-specific ones. We conclude that the gap in Table 1 is predominantly a failure to generalize to C, and that the recovery from synthetic data is predominantly the encoder introducing concept-specific knowledge, with a small general improvement on top.

C.2 Low-Level Metrics on the Target Set

Table 1 of the main text reports only high-level and object-detection metrics on the target set, since the semantic shift is by construction a high-level one. For completeness, Table 5 reports the low-level metrics for the same models.

Table 7: **Synthetic-data controls and frozen-backbone encoders.** MindEye2 probe decoder trained from scratch on subject 1, supercategory generalization (12 concepts). Both controls use identical synthetic fMRI, predicted for images not containing $\mathcal{C}$; they differ only in whether it is paired with those images (non-$\mathcal{C}$ images) or with the $\mathcal{C}$ images (wrong fMRI). *Frozen* encoders keep DINOv2 fixed instead of LoRA-tuning it. The last row uses real fMRI for $\mathcal{C}$ as an upper bound.

	Low-level				High-level				Object Detection	
Training condition	PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑	Incep ↑	CLIP ↑	EffNet ↓	SwAV ↓	Recall ↑	Relaxed ↑
$\mathcal{C}$ unseen	0.247	0.360	94.0%	97.1%	88.4%	82.9%	0.770	0.411	0.080	0.219
+ syn. (non-$\mathcal{C}$ images)	0.262	0.365	94.2%	97.6%	88.1%	83.7%	0.766	0.407	0.088	0.242
+ syn. (wrong fMRI)	0.238	0.356	92.5%	96.4%	87.7%	83.1%	0.769	0.415	0.122	0.298
+ syn. (scratch-blind)	0.269	0.367	95.1%	98.1%	91.6%	87.2%	0.718	0.383	0.228	0.435
+ syn. (transfer-blind)	0.272	0.364	95.2%	98.2%	92.1%	87.8%	0.718	0.381	0.222	0.432
+ syn. (transfer-blind, frozen)	0.272	0.361	95.2%	98.3%	92.5%	87.5%	0.717	0.379	0.239	0.447
+ syn. (transfer-concept)	0.276	0.365	95.5%	98.5%	93.2%	89.0%	0.698	0.371	0.301	0.519
+ syn. (transfer-concept, frozen)	0.275	0.364	95.4%	98.5%	93.1%	89.2%	0.702	0.374	0.287	0.515
Real fMRI for $\mathcal{C}$ (seen)	0.282	0.364	95.9%	98.9%	95.6%	92.7%	0.638	0.338	0.423	0.644

Table 8: **Encoder variants and their value as synthetic-data sources, category generalization (22 concepts).** Same layout as Table 2. Bold marks the best of the three synthesis encoders in each probe-decoder column.

	$\mathcal{C}$ incl. for		Probe decoder trained with synthetic data						Encoder		
Encoder variant	Other	Target	Incep ↑	CLIP ↑	EffNet ↓	SwAV ↓	Rec. ↑	Rel. ↑	r ↑	D.Rec. ↑	D.Rel. ↑
Scratch-oracle	–	✓	*n/a: trained on real fMRI for $\mathcal{C}$*						0.275	0.551	0.693
Transfer-oracle	✓	✓							0.322	0.623	0.738
Scratch-blind	–	✗	94.5%	91.9%	0.685	0.363	0.338	**0.582**	0.268	0.436	0.644
Transfer-blind	✗	✗	94.7%	**92.4%**	**0.675**	0.360	0.351	0.580	0.318	0.513	0.692
Transfer-concept	✓	✗	**94.8%**	92.1%	0.676	**0.358**	**0.365**	0.578	0.320	0.551	0.697
No synthetic			92.4%	90.2%	0.729	0.383	0.179	0.505			
Real fMRI (seen)			96.0%	93.2%	0.636	0.339	0.469	0.650			

D Additional Encoder-Variant Results

D.1 Decoded Recall for Encoders

Following prior work that evaluates encoders by decoding their predictions (Zhang et al., 2025; Mai et al., 2025), we score encoders with the concept metrics of Appendix B.1. For every test image containing $\mathcal{C}$, we predict the target subject's response, reconstruct images from the predictions with a fixed readout, and apply Algorithm 1 unchanged. The qualifying sets depend only on annotations and ground-truth detections and are therefore identical to those used for decoders. The readout is the full-data MindEye2 for the same subject, trained with $\mathcal{C}$ present and with encoder-generated synthetic data (Sec. 5.3), so it can read encoder outputs as well as any decoder we have. The encoder retention in Fig. 1b is the decoded recall of the transfer-concept encoder divided by that of the transfer-oracle encoder.

Because the readout's synthetic training data come from an encoder that saw $\mathcal{C}$, the readout is expected to favor the transfer-oracle variant over the held-out ones. The retention fractions in Fig. 1b are therefore conservative, and the decoded columns of Table 2 should be read for ordering rather than for the size of differences between the scratch and transfer families.

D.2 Category generalization

To probe a smaller semantic gap than a supercategory, we repeat the analysis of Sec. 6 with individual COCO categories as $\mathcal{C}$, taking the two most frequent categories of each supercategory among the 1,000 test set images; *person* is excluded as it has no subcategories, giving 22 categories (Table 3). Table 8 reports the same quantities as Table 2.

Every effect shrinks with the semantic gap. Removing $\mathcal{C}$ from all subjects costs the encoder 0.004 in r (1%) and 18% of decoded Object Recall, against 0.014 and 34% at supercategory granularity. The reversal between fidelity and decoded recall for the pre-training contrast (scratch-oracle vs. transfer-blind) shrinks to 0.038 in Object Recall and is not robust across tests (Appendix G.2). For the probe decoder, concept-specific transfer is significant on no metric at this granularity, whereas it is significant on every metric at supercategory granularity; the dependence on granularity is itself significant for Inception, CLIP and EffNet. The within-cell rank correlation between encoder r and probe-decoder Object Recall vanishes ($\rho = 0.10$, n.s.). Scratch-blind and transfer-blind still recover 55% and 59% of the Object Recall gap. Note that at this granularity sibling categories of $\mathcal{C}$ remain in training, so relaxed recall partly credits the decoder for substituting a seen, semantically adjacent object (Appendix B).

D.3 Synthetic-data controls

Sec. 6 attributes the benefit of synthetic data to the encoder's ability to predict responses to $\mathcal{C}$. A simpler explanation is that any additional training pairs help, as in conventional data augmentation. We test this with two controls that add the same number of synthetic pairs as the main condition and use identical synthetic responses, differing only in which images those responses are paired with.

Non-$\mathcal{C}$ images. The synthetic responses are predicted by the transfer-concept encoder for images that do not contain $\mathcal{C}$, and are paired with those same images. The pairs are therefore correctly matched but carry no information about $\mathcal{C}$, which isolates the effect of training-set size.

Wrong fMRI. The same responses are instead paired with the $\mathcal{C}$ images that the main condition uses. The decoder thus sees $\mathcal{C}$ as a reconstruction target, but never with the response that $\mathcal{C}$ elicits. Because the fMRI is identical to the non-$\mathcal{C}$ control and comes from the same encoder, any failure of this control cannot be attributed to the synthetic responses being off-distribution.

Results. Table 7 reports both controls for subject 1. Neither closes the gap: both fall well short of correctly paired synthetic data on Object Recall (-0.21 and -0.18) and Relaxed Recall (-0.28 and -0.22; all $p < 0.001$). The non-$\mathcal{C}$ control is statistically equivalent to the unseen baseline on every high-level and object-detection metric (Object Recall 0.080 to 0.088), yet significantly improves the low-level metrics (PixCorr $+0.015$, SSIM $+0.005$, Alex(5) $+0.4$ points), consistent with the general benefit of synthetic data in Sec. 5.3. Additional data therefore improves decoding in general but not on $\mathcal{C}$. Wrong fMRI instead lowers Alex(2), Alex(5) and SSIM significantly, as expected from mismatched pairs acting as label noise, and modestly raises Object Recall (0.080 to 0.122) and Relaxed Recall (0.219 to 0.298). This is consistent with the decoder's output prior drifting toward $\mathcal{C}$ once $\mathcal{C}$ images appear as targets, but it amounts to only 19% [12, 26] and 27% [19, 35] of the gain from correct pairing. The benefit of synthetic data therefore requires both that the images depict $\mathcal{C}$ and that they are paired with the encoder's prediction of the response to them. Neither additional data volume nor exposure to $\mathcal{C}$ images alone is sufficient.

D.4 Frozen encoder backbone

The Universal Encoder adapts its DINOv2 backbone with LoRA during pre-training. If the encoder's robustness to $\mathcal{C}$ came from this fMRI-driven adaptation rather than from the pre-trained representation, freezing the backbone should reduce it. We retrain the transfer-blind and transfer-concept encoders with DINOv2 frozen, so that fMRI training updates only the remaining encoder parameters, and train the probe decoder on their synthetic pairs (Table 7).

Freezing the backbone has no detectable effect on the decoder gains from either encoder. For transfer-blind, Object Recall is 0.222 with LoRA and 0.239 frozen (Relaxed Recall 0.432 and 0.447); for transfer-concept, it is 0.301 and 0.287 (0.519 and 0.515). Both frozen encoders are statistically equivalent to their LoRA counterparts on all four high-level metrics and both object-detection metrics, with 90% confidence intervals for the Object Recall difference within ± 0.033 (Appendix G.7). The low-level metrics are mostly inconclusive rather than different; the only significant difference is a small SSIM decrease for transfer-blind (-0.003). Since transfer-blind never observes $\mathcal{C}$ from any subject, its ability to produce useful responses to $\mathcal{C}$ must come from the pre-trained representation itself. The concept-specific advantage of transfer-concept over transfer-blind also persists with

Table 9: **Effect of synthetic data across data scales**. Numbers correspond to values from Figures 4 and 6. The better result within each decoder pair in **bold**.

		Low-level				High-level			
Decoder	Syn.	PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑	Incep ↑	CLIP ↑	EffNet ↓	SwAV ↓
Full data (9,000 images)									
Brain-Diffuser	✗	0.258	**0.357**	94.4%	96.6%	88.2%	91.9%	0.770	0.420
	✓	**0.273**	0.353	**95.5%**	**97.4%**	**90.1%**	**93.3%**	**0.749**	**0.401**
MindEye2	✗	0.278	0.376	95.6%	98.6%	94.9%	93.4%	0.621	0.344
	✓	**0.312**	**0.389**	**96.9%**	**99.1%**	**96.1%**	**94.1%**	**0.604**	**0.330**
Brain-IT	✗	0.362	**0.501**	98.0%	99.3%	96.8%	95.6%	0.584	0.336
	✓	**0.371**	0.494	**98.7%**	**99.5%**	**97.2%**	**95.9%**	**0.566**	**0.327**
Limited data (225 images)									
Brain-Diffuser	✗	0.128	0.322	77.7%	82.4%	71.8%	77.4%	0.899	0.533
	✓	**0.164**	**0.335**	**83.5%**	**87.6%**	**76.2%**	**81.6%**	**0.874**	**0.507**
MindEye2	✗	0.148	0.360	82.4%	89.2%	79.5%	79.9%	0.822	0.476
	✓	**0.184**	**0.375**	**88.3%**	**93.4%**	**84.0%**	**83.1%**	**0.778**	**0.442**
Brain-IT	✗	0.149	0.479	79.2%	86.2%	80.7%	79.7%	0.822	0.509
	✓	**0.174**	**0.485**	**81.1%**	**87.8%**	**82.3%**	**81.2%**	**0.806**	**0.495**

a frozen backbone (Object Recall $+0.049$ [0.028, 0.070], Relaxed Recall $+0.068$ [0.039, 0.097], both $p < 0.001$), so concept-specific transfer through other subjects does not require adapting DINOv2. Whether freezing modestly reduces this advantage is unresolved (the difference in advantage is 0.030 [0.000, 0.059] in Object Recall). Equivalence holds at the margin used throughout the paper but not at a stricter margin of half the concept-specific advantage, so we cannot exclude differences of that size. These results are for a single subject and do not account for variation across training runs.

E Additional Data-scaling Results

E.1 Limited and full data results written out

Table 9 lists part of the exact results for the experiment in Sec. 5.3, with values identical to Fig. 4.

E.2 Encoder pre-training across data budgets

Figure 5 investigates the effect of encoder pre-training across budgets, using the full MindEye2 of Sec. 5.3 with synthetic data from either a cross-subject pre-trained and fine-tuned encoder or one trained from scratch on the target subject. A scratch encoder delivers roughly a quarter of the pre-trained encoder's improvement at 225 real images but roughly three quarters at 9,000. The increase of this ratio with budget is significant on every metric, as is the encoder $\times$ log-budget interaction (Appendix G.4). Given enough target-subject data the encoder can be fit without a cross-subject prior; below that it cannot. The scratch encoder still improves the decoder at every budget, and at the full budget the pre-trained encoder's advantage is no longer significant on PixCorr, Inception, CLIP or EffNet, so pre-training sharpens the effect rather than causing it. The negligible effect of general transfer in Table 2, measured at the full budget, would not hold at small budgets.

E.3 Additional Results for Using Synthetic Data Across Data Scales

Figure 6 shows results for additional metrics that were not shown in figure 4. While AlexNet(2), Inception, and SwAV behave similarly, SSIM is a clear outlier. ρ is not reported for SSIM because its nearly flat baseline curve makes it unidentifiable for all three models (Appendix G.3). Also it is observed that SSIM lowers for the synthetic data model at the full 9,000-image budget for Brain-Diffuser and Brain-IT.

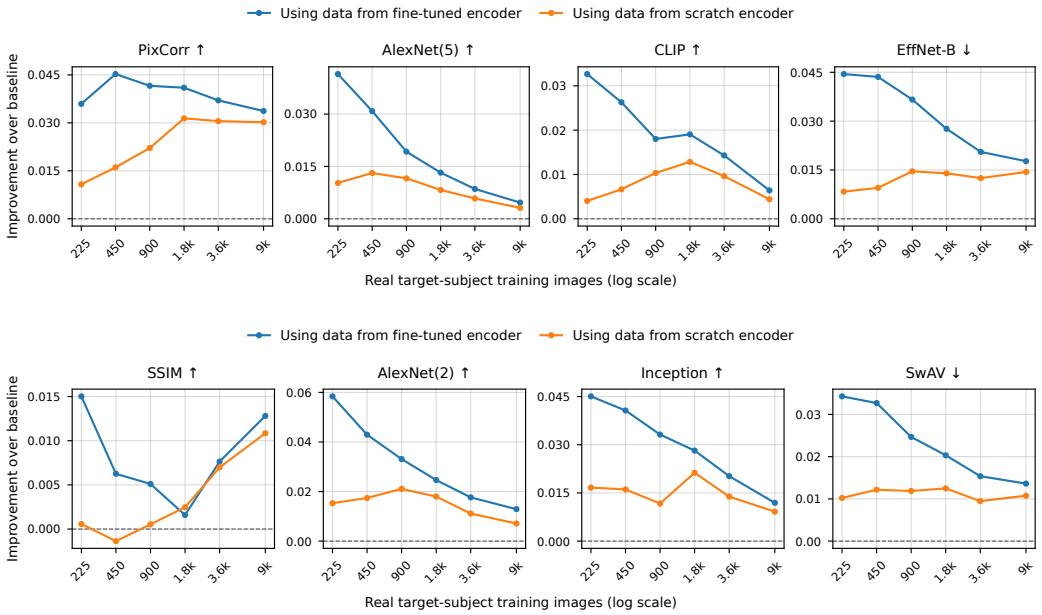


Figure 5: **Encoder pre-training matters most when target-subject data is scarce.** Improvement of synthetic data for MindEye2 over its real-data-only baseline, on all eight metrics, when the synthetic data come from a cross-subject pre-trained encoder (blue) versus an encoder trained from scratch on the target subject (orange). Both help at every budget, but the pre-trained encoder's advantage is largest at small budgets and shrinks as target-subject data grows (Appendix G.4). SSIM is the exception, with small gains and no consistent ordering between the two encoders.

E.4 Scaling the Amount of Synthetic Data

The experiments so far add a fixed amount of synthetic data. Here we vary it systematically, to establish whether the benefit is specific to one operating point and how it interacts with the amount of real data available for the target subject. Similar to Sec. 5.3, we fine-tune MindEye2 on $x \in \{225, 450, 900, 1800, 3600, 9000\}$ real images from the target subject, where $x = 9000$ corresponds to the full budget, and additionally vary the number of images that are used for synthetic data, $y = m \cdot x$ for $m \in \{0, 1, 2, 3, 4\}$. The synthetic pairs are generated in the same manner, and the $m = 0$ condition is the real-data-only baseline. Every configuration is trained separately for subjects 1, 2, 5 and 7, and all results are averaged over those four subjects. Figure 7 reports each metric as the improvement over that budget's own $m = 0$ baseline, so that all curves start at zero and lower-is-better metrics are sign-flipped.

Most of the benefit arrives by $2\times$. The gain is strongly concave in m. Averaged over the eight metrics, the first increment ($m = 0 \to 1$) delivers 63% (95% CI 60-66) of the total improvement measured at $m = 4$, and $m = 2$ delivers 84% (81-87). The remaining increments are small but not uniformly null: the $m = 1 \to 2$ step is significant on 32 of 48 cells and the $m = 2 \to 4$ step on 25 of 48, so a larger pool continues to help modestly; the only significant reversal is EffNet at the full budget from $m = 2$ to $m = 4$.

The benefit grows as real data becomes scarce. At $m = 4$, the improvement in AlexNet(2) falls monotonically from 6.1 points at $x = 225$ to 1.0 points at $x = 9000$, and AlexNet(5), Inception, CLIP, EffNet-B and SwAV show the same ordering without crossings; for all six of these metrics the $x = 225$ gain significantly exceeds the $x = 9000$ gain (Spearman $\rho \leq -0.94$ between budget and gain, all $p_{\text{Holm}} \leq 0.001$).

Low-level and high-level metrics behave differently. The budget ordering above is clean for the semantic and identification metrics but not for PixCorr and SSIM, where the curves are compressed

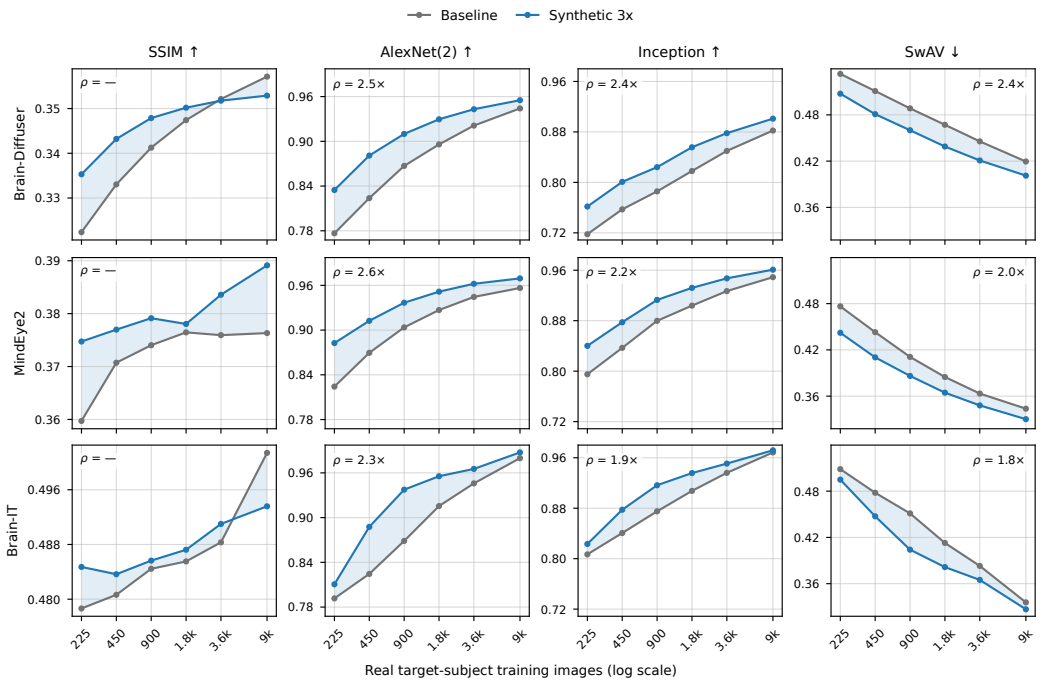


Figure 6: **Effect of synthetic data across data scales: remaining metrics.** Companion to Figure 4, showing SSIM, AlexNet(2), Inception and SwAV for each decoder.

and the ordering is scrambled: at $m = 4$, PixCorr improves by 0.047 at $x = 450$ against 0.041 at $x = 225$ and 0.035 at $x = 9000$, and neither metric shows a significant difference between the $x = 225$ and $x = 9000$ gains (PixCorr $p = 0.17$, SSIM $p = 0.59$).

F Extension to Video Decoding

We extend the *encoder-generated data for decoder* recipe to fMRI-to-video decoding to test if it could be extended to other modalities.

Setup. We trained and evaluated the NEURONS (Wang et al., 2025) model in the cc2017 dataset (Wen et al., 2018) under the limited data setting (10% of full data) as well as the full data setting.

The cc2017 dataset comprises three subjects who viewed 18 8-minute training clips and 5 8-minute test clips inside a 3-T scanner at a 2s temporal resolution. Each fMRI sample therefore corresponds to a 2s, 6-frame video clip, giving 4320 unique training clips presented twice and 1200 test clips presented ten times; following common practice, the test repetitions are averaged.

Following the convention of Sec. 5.3, we parameterize the limited-data budget by the number of unique clips and retain both repetitions of each. The target subject's training set is reduced to 432 unique clips, 10% of the full 4,320, corresponding to 864 recorded trials and roughly 29 minutes of scanning. The fMRI recorded for the remaining 3,888 clips is discarded; the video itself is kept.

Encoder adaptation and synthesis. The Universal Encoder (Beliy et al., 2024) is jointly pre-trained on the eight NSD subjects and the two non-target cc2017 subjects, and then adapted to the target subject. This joint pre-training spans two datasets, two field strengths and two stimulus modalities at once, which the encoder accommodates without architectural change: all weights other than the per-voxel embeddings are shared across voxels, subjects and datasets, so a new subject is absorbed simply by adding its voxels to the embedding table. Because the encoder predicts responses to still images, we sample a single frame from each 6-frame clip and treat it as the stimulus for the

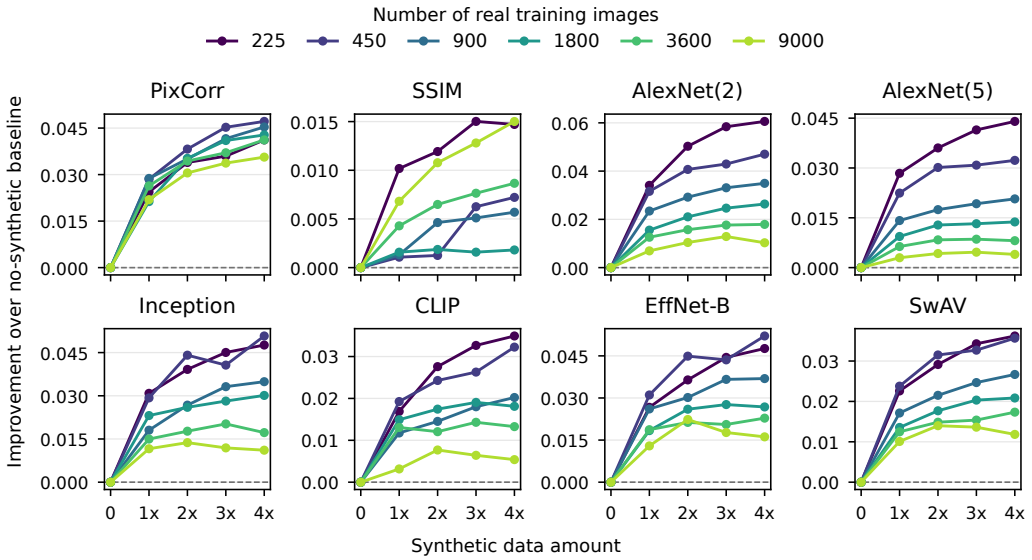


Figure 7: **Scaling the amount of synthetic data.** Improvement over the baseline for MindEye2, as a function of the number of synthetic image-fMRI pairs added per real training image. One line per real-data budget x; each curve is measured against its own x-matched $m{=}0$ baseline, so all curves start at zero. Lower-is-better metrics are sign-flipped, so *up is better in every panel*; note that vertical scales differ across panels. Gains are positive at every budget, and grow as the real-data budget shrinks.

Table 10: **Extension to video decoding**. A NEURONS model is trained with and without synthetic fMRI, for limited and full data settings. **Bold** marks the better of the two conditions for each budget.

		Video-based			Frame-based			
		Semantic-level		ST-level	Semantic-level		Pixel-level	
Training data	Syn.	2-way ↑	50-way ↑	CLIP-pcc ↑	2-way ↑	50-way ↑	SSIM ↑	PSNR ↑
Limited (432 clips)	✗	0.837	0.185	0.932	0.770	0.154	0.357	9.458
	✓	**0.855**	**0.235**	**0.932**	**0.797**	**0.187**	**0.383**	**9.669**
Full (4,320 clips)	✗	0.857	0.259	**0.937**	0.810	0.206	0.377	9.672
	✓	**0.865**	**0.275**	0.936	**0.816**	**0.221**	**0.397**	**9.896**

corresponding fMRI sample. Given the synthesis pool $\mathcal{U}$, we predict a response for each 2s window independently, reproducing the temporal structure of the real recordings.

For the limited data experiment, we replace all of the 3,888 discarded clips with synthetic data, while for the full data experiment we introduce additional video clips from the WebVid dataset (Bain et al., 2021). We use 1,200 8s clips from the dataset, using the same set of video clips selected in Li et al. (2024), which we split into 4,800 2s clips and generate synthetic fMRI based on them. As in the image experiments, real and synthetic pairs are mixed within each batch and are not distinguished by the decoder in any way.

Evaluation metrics. We report the standard cc2017 metrics (Wang et al., 2025; Gong et al., 2024). *Video-based* metrics evaluate the reconstruction as a clip: semantic-level 2-way and 50-way top-1 accuracy from a classifier based on VideoMAE (Tong et al., 2022) over the 400 Kinetics-400 classes (Kay et al., 2017), and spatiotemporal-level CLIP-pcc, the average cosine similarity between CLIP embeddings of adjacent frames, which measures temporal consistency. *Frame-based* metrics evaluate individual frames: semantic-level 2-way and 50-way top-1 accuracy from an ImageNet classifier, and pixel-level SSIM and PSNR.

Results. Table 10 shows that encoder-generated fMRI recovers most of the performance lost to the data limitation. Under the limited budget, synthetic data significantly improves every metric except CLIP-pcc, which the data limitation itself barely affects (a drop of 0.005). The limited-data model with synthetic data recovers 63-87% of the gap to the full-data model on the four semantic metrics and matches or exceeds it on the pixel-level metrics (SSIM 127% [107, 153], PSNR 99% [71, 139]). At the full budget the gains are smaller but remain significant on five of seven metrics (Appendix G.6).

Two aspects of this result are worth emphasizing. First, the encoder is a still-image model: it never observes motion, and each synthetic response is predicted from one frame of a six-frame clip. Such coarse supervision still improves a video decoder. Second, no component of either model was modified for the video case. This supports the claim of Sec. 4.2 that the intervention is at the level of the data rather than the model, and suggests the approach should carry to other stimulus and recording modalities for which a stimulus-to-signal encoder is available.

G Statistical Tests

In this section we provide the details for the statistical tests that are performed for the various quantitative experimental results presented in previous sections.

G.1 Concept generalization across decoders

For each decoder and metric, we compare conditions over the 48 subject–concept cells (4 subjects × 12 supercategories) after averaging within each cell. Differences are tested with two-sided Wilcoxon signed-rank tests, Holm-corrected across the 20 planned comparisons per decoder (10 metrics × 2 contrasts); paired t-tests and linear mixed models with crossed random intercepts for subject and concept reach the same conclusions. For Object Recall and Relaxed Recall we additionally fit logistic mixed models at the level of individual qualifying images, which likewise agree. Confidence intervals are 95% percentile bootstrap intervals over cells (10,000 resamples, stratified by subject). Recovery fractions are computed from aggregate means, with bootstrap CIs over cells and, more conservatively, over subjects. The two-way identification metrics are computed against the full pool of remaining test images rather than a random distractor, so no sampling enters the comparison. The same procedure is applied to the complementary test set, with Holm correction across the 16 comparisons per decoder (8 metrics × 2 contrasts). Full results are in Tables 11 and 12.

The concept generalization gap (Seen vs. Unseen) is significant for every decoder on every metric except SSIM, which the holdout does not change or the change is very small (Brain-Diffuser Δ = 0.002, p = 0.22; Brain-IT Δ = 0.000, p = 0.90; MindEye2 Δ = 0.004, p = 0.019). Adding synthetic data (Unseen + syn. vs. Unseen) significantly improves all ten metrics for MindEye2 and nine of ten for Brain-Diffuser and Brain-IT, again excepting SSIM; all corrected p-values are below 0.001 except MindEye2 SSIM (0.004), Brain-IT Alex(2) (0.005), and Brain-IT PixCorr (0.022), the last of which is not significant under the mixed model (p = 0.063). Object Recall improves for every one of the twelve concepts under every decoder.

G.2 Encoder variants

For Tables 2 and 8, we use the same cell-level paired procedure (48 subject-supercategory cells or 88 subject-category cells), with Holm correction within each granularity and test family. Equivalence is assessed with two one-sided tests (TOST) using a margin of 25% of the relevant oracle-to-blind gap for each metric: the unseen-to-real gap for probe-decoder contrasts, and the transfer-oracle to transfer-blind gap for encoder contrasts. The dependence of the concept-specific contrast on granularity is tested with a mixed model including a granularity $\times$ condition term and crossed random intercepts for subject and concept, Holm-corrected across metrics. Because no transfer variant exists in which $\mathcal{C}$ is exposed to the target subject but withheld from the other subjects, the two-factor design is incomplete, and we test the estimable contrasts directly. Agreement between an encoder-side metric and decoder benefit is measured as the Spearman rank correlation between that metric and probe decoder Object Recall across the three usable variants within each cell, averaged over cells with a bootstrap CI. For the decoded-recall metrics we additionally fit image-level logistic mixed models

Table 11: Fraction of the concept generalization gap recovered by synthetic data, (syn − unseen)/(seen − unseen) on aggregate means, with 95% bootstrap CIs over the 48 subject-supercategory cells. Left: the three decoders of Table 1. Right: the probe decoder of Table 2 under each synthesis encoder. SSIM is omitted because the holdout does not change it. Subject-level bootstrap intervals are similar and are provided with the released results.

	Decoders (Table 1)			Probe decoder by encoder (Table 2)		
Metric	Brain-Diffuser	MindEye2	Brain-IT	Scratch-blind	Transfer-blind	Transfer-concept
PixCorr	82% [77, 87]	71% [62, 81]	40% [16, 66]	61% [51, 74]	59% [49, 72]	73% [63, 86]
Alex(2)	92% [84, 102]	64% [54, 72]	64% [38, 89]	63% [51, 76]	67% [57, 79]	77% [67, 88]
Alex(5)	93% [89, 98]	73% [68, 79]	58% [46, 71]	63% [56, 72]	64% [57, 73]	74% [67, 81]
Incep	83% [79, 87]	66% [60, 71]	46% [33, 55]	51% [46, 56]	51% [46, 57]	69% [63, 74]
CLIP	87% [84, 90]	57% [53, 62]	53% [43, 62]	46% [41, 53]	49% [43, 56]	62% [57, 67]
EffNet	76% [74, 79]	53% [49, 58]	38% [30, 44]	37% [33, 41]	41% [37, 46]	52% [49, 56]
SwAV	81% [79, 83]	56% [51, 61]	37% [28, 44]	39% [35, 43]	43% [39, 48]	52% [48, 56]
Object Recall	78% [70, 85]	55% [50, 60]	41% [32, 48]	41% [36, 45]	44% [39, 49]	59% [54, 64]
Relaxed Recall	87% [83, 90]	57% [52, 62]	41% [32, 49]	48% [44, 52]	51% [47, 57]	62% [58, 67]

Table 12: Effect of the training-set manipulations on images that do *not* contain $\mathcal{C}$, compared with their effect on images that do. $\Delta_{\mathcal{C}}$ and $\Delta_{\neg\mathcal{C}}$ are mean paired differences over the 48 subject-concept cells on the concept test set (Table 1) and the complementary test set (Table 4); positive = improvement, EffNet and SwAV sign-flipped, identification metrics in percentage points. Ratio is $\Delta_{\neg\mathcal{C}}/\Delta_{\mathcal{C}}$. * marks $\Delta_{\neg\mathcal{C}}$ significant after Holm correction (Wilcoxon signed-rank, 16 comparisons per decoder).

		Seen vs. Unseen			Unseen + syn. vs. Unseen		
Decoder	Metric	$\Delta_{\mathcal{C}}$	$\Delta_{\neg\mathcal{C}}$	Ratio	$\Delta_{\mathcal{C}}$	$\Delta_{\neg\mathcal{C}}$	Ratio
Brain-Diffuser	Incep	9.0	0.4*	4%	7.5	0.2	3%
	CLIP	9.3	0.1	1%	8.1	0.3*	4%
	EffNet	0.086	0.001	1%	0.066	0.003*	4%
	SwAV	0.061	0.002*	4%	0.050	0.004*	8%
MindEye2	Incep	7.4	0.3*	5%	4.9	0.4*	8%
	CLIP	9.5	0.8*	8%	5.5	0.4*	8%
	EffNet	0.132	0.009*	7%	0.070	0.010*	14%
	SwAV	0.070	0.005*	8%	0.040	0.006*	14%
Brain-IT	Incep	2.5	0.2*	7%	1.1	0.2	15%
	CLIP	3.0	0.0	1%	1.6	0.0	0%
	EffNet	0.059	0.002	3%	0.022	0.000	-2%
	SwAV	0.029	0.001	2%	0.011	0.000	4%

on individual qualifying images (random intercepts for subject and concept) and report whether they agree with the cell-level test.

Fidelity. Removing $\mathcal{C}$ from all subjects reduces encoder Pearson r by 0.014 [0.011, 0.017] at supercategory and 0.004 [0.003, 0.005] at category granularity; the pre-training gap (transfer-blind vs. scratch-oracle) exceeds this by 0.020 [0.012, 0.026] and 0.039 [0.035, 0.043] respectively, all $p < 0.001$. Transfer-blind and scratch-blind are equivalent on nine of ten probe-decoder metrics at supercategory granularity and five of ten at category granularity; the granularity × concept-specific-transfer interaction is significant for Inception ($p = 0.005$), CLIP ($p = 0.002$) and EffNet ($p = 0.023$). The within-cell rank correlation between encoder r and decoder Object Recall is $\rho = 0.56$ [0.42, 0.70] at supercategory and 0.10 [-0.04, 0.24] at category granularity.

Decoded recall: holdout cost. Removing $\mathcal{C}$ from all subjects (transfer-oracle vs. transfer-blind) costs 0.195 [0.163, 0.229] in Object Recall and 0.150 [0.119, 0.184] in Relaxed Recall at supercategory granularity (34% [30, 38] and 20% [17, 25] of the oracle score), and 0.110 [0.085, 0.136] and 0.046 [0.030, 0.062] at category granularity (18% [14, 22] and 6% [4, 8]). Removing $\mathcal{C}$ from the target subject only (transfer-oracle vs. transfer-concept) costs 0.075 [0.063, 0.088] and 0.051 [0.040, 0.061] at supercategory granularity (13% [11, 15] and 7% [6, 8]) and 0.071 [0.052, 0.092] and 0.041

[0.025, 0.058] at category granularity. All are significant ($p_{\text{Holm}} < 10^{-4}$) and the image-level models agree.

Decoded recall: pre-training versus exposure. On decoded recall the pre-training contrast has the opposite sign to fidelity at supercategory granularity: scratch-oracle exceeds transfer-blind by 0.118 [0.084, 0.154] in Object Recall and 0.100 [0.065, 0.140] in Relaxed Recall ($p_{\text{Holm}} < 10^{-5}$, image-level models agree). At category granularity the Object Recall difference is 0.038 [0.011, 0.065] ($p_{\text{Holm}} = 0.003$ by Wilcoxon; not significant under the mixed model, $p = 0.59$) and Relaxed Recall does not differ. Transfer-concept and scratch-oracle are equivalent on Object Recall at both granularities (TOST $p_{\text{Holm}} < 10^{-6}$ and $= 0.017$) and on Relaxed Recall at supercategory granularity ($p_{\text{Holm}} < 10^{-5}$), but not at category granularity ($p = 0.19$); the supercategory Relaxed Recall contrast is null at the cell level ($p = 0.92$) but significant in the image-level model ($p = 0.004$), reflecting image rather than equal-cell weighting.

Decoded recall: transfer steps. Both transfer steps are significant on decoded recall at supercategory granularity: general transfer (transfer-blind vs. scratch-blind) 0.089 [0.069, 0.109] in Object Recall and 0.093 [0.074, 0.113] in Relaxed Recall, concept-specific transfer (transfer-concept vs. transfer-blind) 0.119 [0.089, 0.152] and 0.099 [0.071, 0.132], all $p_{\text{Holm}} < 10^{-6}$. Their difference is not significant (-0.030 [-0.072, 0.008], $p_{\text{Holm}} = 0.65$; -0.006 [-0.046, 0.030], $p_{\text{Holm}} = 0.65$). At category granularity general transfer is the larger step: 0.077 vs. 0.038 in Object Recall (difference 0.039 [0.004, 0.073], $p_{\text{Holm}} = 0.021$) and 0.048 vs. 0.005 in Relaxed Recall (0.043 [0.012, 0.075], $p_{\text{Holm}} = 0.027$; mixed model $p = 0.10$). Transfer-blind and scratch-blind are not equivalent on either decoded-recall metric at either granularity (TOST $p = 1.00$).

Decoded recall: agreement with the probe decoder. The within-cell rank correlation with probe decoder Object Recall is $\rho = 0.58$ [0.42, 0.72] for decoded Object Recall and 0.55 [0.40, 0.69] for decoded Relaxed Recall at supercategory granularity, and 0.14 [-0.01, 0.28] and 0.07 [-0.08, 0.22] at category granularity, matching the values for r above. The strict ordering scratch-blind $<$ transfer-blind $<$ transfer-concept holds in 79% [69, 90] of supercategory cells and 36% [27, 47] of category cells on decoded Object Recall, against 92% for r.

Encoder versus decoder retention (Fig. 1b). Per cell, encoder retention is transfer-concept over transfer-oracle and decoder retention is MindEye2 unseen over seen, on the 48 supercategory cells (no zero denominators). Aggregate encoder retention is 0.869 [0.847, 0.889] for Object Recall and 0.931 [0.917, 0.946] for Relaxed Recall; decoder retention is 0.216 [0.175, 0.260] and 0.365 [0.304, 0.427]. Encoder retention exceeds decoder retention in every cell, by 0.638 [0.587, 0.688] and 0.561 [0.505, 0.617] on average ($p_{\text{Holm}} < 10^{-13}$; subject-level bootstrap intervals [0.620, 0.670] and [0.525, 0.589]).

Sensitivity. Paired t-tests and linear mixed models agree with the Holm-corrected Wilcoxon conclusions on every contrast except three: the category Object Recall reversal above, the category r difference between scratch-oracle and scratch-blind (Wilcoxon $p_{\text{Holm}} = 0.031$, mixed model $p = 0.10$), and the category Relaxed Recall difference between the two transfer steps. Excluding the four subcategories with fewer than five qualifying images (toaster, parking meter, hair drier, toothbrush) from Object Recall changes no decision (encoder retention 0.875 vs. 0.869). Decoded-recall bootstrap intervals are over the same 48 and 88 cells as the other Table 2 columns; subject-level intervals are similar and are provided with the released results, as are all per-contrast estimates, CIs and p-values.

G.3 Data scaling

For the scaling experiments of Sec. 5.3 the four subjects share the same 1,000 test images, so we compare conditions at the image level: each of the $6 \times 8 = 48$ budget-metric contrasts per decoder is the mean paired difference between synthetic-data and baseline training over the 4,000 subject-image pairs, with a 95% percentile CI and a two-sided p-value from 10,000 bootstrap resamples of images within subject, Holm-corrected within decoder. Of the 144 contrasts, 137 show a significant improvement and 5 are not significant; the two significant reversals are SSIM at the full 9,000-pair budget for Brain-Diffuser ($\Delta = -0.004$) and Brain-IT ($\Delta = -0.008$), visible in Table 9. The remaining non-significant contrasts are SSIM at intermediate budgets and, for Brain-IT at the full

Table 13: Data-efficiency factor $\hat{\rho}$ for every metric and decoder, with 95% CIs from 10,000 subject-stratified image-level bootstraps in which the interpolant and $\hat{\rho}$ are refitted on each replicate. SSIM is omitted: its curves are nearly flat across budgets, so $\hat{\rho}$ is not identifiable for it (point estimates of 0.8 to 4.9 with CIs spanning an order of magnitude).

Metric	Brain-Diffuser	MindEye2	Brain-IT
PixCorr	2.64 [2.48, 2.77]	2.94 [2.67, 3.26]	1.94 [1.87, 2.02]
Alex(2)	2.55 [2.42, 2.69]	2.61 [2.45, 2.80]	2.27 [2.17, 2.38]
Alex(5)	2.39 [2.27, 2.53]	2.35 [2.20, 2.51]	2.03 [1.91, 2.19]
Incep	2.37 [2.20, 2.64]	2.16 [1.98, 2.36]	1.86 [1.72, 1.99]
CLIP	2.20 [2.08, 2.35]	1.74 [1.61, 1.87]	1.78 [1.66, 1.90]
EffNet	2.30 [2.21, 2.40]	1.87 [1.78, 1.96]	1.78 [1.72, 1.83]
SwAV	2.43 [2.35, 2.52]	1.99 [1.91, 2.08]	1.84 [1.79, 1.89]

budget, Inception and CLIP ($\Delta = +0.3$ points each). At the subject level, all four subjects improve in 46, 47 and 41 of the 48 contrasts for Brain-Diffuser, MindEye2 and Brain-IT respectively; with $n = 4$ a sign test cannot reach $p < 0.05$, so we report the image-level tests as primary. The rise of each curve from 3,600 to 9,000 real images is significant in 47 of 48 decoder-condition-metric combinations (the exception is SSIM for the MindEye2 baseline, $\Delta = 0.0004$).

For $\hat{\rho}$, the monotone interpolant $\hat{b}$ is a PCHIP fit to the six baseline points on $\log x$, and the fit, the interpolant and $\hat{\rho}$ are recomputed on each bootstrap replicate. Because $\hat{\rho}x$ exceeds 9,000 at the largest budget for every metric, $\hat{b}$ is held at its 9,000-image value there rather than extrapolated. Table 13 lists $\hat{\rho}$ with image-level CIs; subject-level bootstrap intervals are wider but overlap them for every entry and are provided with the released results. SSIM is excluded because its baseline curve is nearly flat, which makes $\hat{\rho}$ unidentifiable (Brain-IT 0.8 [0.7, 1.6]; MindEye2 4.9 [3.1, 23.7]). Finally, the improvement from synthetic data decreases monotonically with budget for five of eight metrics for Brain-Diffuser and for MindEye2, but for none for Brain-IT, whose gain is non-monotonic across budgets.

G.4 Encoder pre-training across budgets

For Figure 5 (full MindEye2, three conditions per budget: baseline, synthetic data from the pre-trained encoder, synthetic data from the scratch encoder) we use the image-level paired bootstrap of Appendix G.3, Holm-corrected within metric. Both encoders improve on the baseline at every budget: the pre-trained encoder significantly on 47 of 48 budget-metric contrasts (the exception is SSIM at 1,800 pairs) and the scratch encoder on 42 of 48 (the exceptions are SSIM at 225-900 pairs and CLIP at 225, 450 and 9,000 pairs). The pre-trained encoder's advantage over the scratch encoder is significant on 41 of 48 contrasts; it is significant on every metric except SSIM at 225-900 pairs, and at 9,000 pairs only on Alex(2), Alex(5), SSIM and SwAV. The ratio of scratch to pre-trained improvement, averaged over the upper four metrics of Figure 5, is 0.21 at 225 pairs and 0.77 at 9,000, and its increase from 225 to 9,000 is significant on all eight metrics (one-sided paired bootstrap of the difference of ratios). A mixed model on subject-level scores, with an encoder $\times \log x$ term and a random intercept per subject, confirms the interaction on all eight metrics ($p_{\mathrm{Holm}} \leq 0.04$).

G.5 Scaling the amount of synthetic data

For Figure 7 we use the image-level paired bootstrap of Appendix G.3, comparing each synthetic multiplier m against the x-matched $m = 0$ baseline and against the preceding multiplier, Holm-corrected within metric. The fraction of the $m = 4$ gain reached at $m = 1$ and $m = 2$ is computed per metric as the ratio of its gain averaged over the six budgets to its $m = 4$ gain averaged over the six budgets, then averaged over metrics, with a bootstrap CI. The dependence of the $m = 4$ gain on budget is tested two ways: a bootstrap Spearman correlation between x and the gain over the six budgets, and a one-sided paired bootstrap of the $x = 225$ gain minus the $x = 9000$ gain. Both are significant for AlexNet(2), AlexNet(5), Inception, CLIP, EffNet and SwAV ($p_{\mathrm{Holm}} \leq 0.001$) and for neither PixCorr nor SSIM.

G.6 VIDEO DECODING

For Table 10, the three subjects share the same 1,200 test clips, so we compare training with and without synthetic data at the clip level ($n = 3{,}600$ subject-clip pairs): two-sided Wilcoxon signed-rank tests, Holm-corrected within budget, with 95% CIs from 10,000 subject-stratified clip bootstraps; paired t-tests agree throughout. The identification metrics use one fixed distractor assignment for all four conditions. Under the limited budget, six of seven metrics improve significantly (all $p_{\text{Holm}} < 10^{-6}$); CLIP-pcc does not ($\Delta = 0.0001$ [-0.0012, 0.0016]). Under the full budget, Video 2-way ($p = 0.03$), Video 50-way (0.01), Frame 50-way (0.01), SSIM and PSNR (both $< 10^{-10}$) improve significantly, while Frame 2-way ($\Delta = 0.006$, $p = 0.06$) and CLIP-pcc ($\Delta = -0.001$, $p = 0.06$) do not. All three subjects improve individually on every metric except CLIP-pcc at both budgets and Frame 50-way at the full budget. Recovery fractions, $(\text{limited+syn} - \text{limited})/(\text{full} - \text{limited})$ with clip-bootstrap CIs, are Video 2-way 87% [56, 127], Video 50-way 68% [53, 85], Frame 2-way 69% [57, 82], Frame 50-way 63% [48, 80], SSIM 127% [107, 153] and PSNR 99% [71, 139]; CLIP-pcc is excluded because the full-minus-limited gap is negligible.

G.7 SYNTHETIC-DATA CONTROLS AND FROZEN BACKBONE

For Table 7, all models are trained on subject 1 and evaluated on the same test images, so we use a paired stratified bootstrap (10,000 resamples). Images are resampled within each subcategory's qualifying set Q_o for Object Recall, within each supercategory's qualifying set Q_γ for Relaxed Recall, and within each supercategory's test set for the remaining metrics, with the same resampled indices for every condition; stratum sizes and macro-averaging weights are held fixed. Qualifying sets depend only on annotations and ground-truth detections and are therefore identical across conditions.

We test each control against the unseen baseline on all ten metrics (Holm-corrected across 20 comparisons) and against transfer-concept on Object Recall and Relaxed Recall (4 comparisons). Equivalence of the non-$\mathcal{C}$ control to the unseen baseline, and of each frozen encoder to its LoRA counterpart, is assessed with TOST using the margin of Appendix G.2 (25% of the unseen-to-seen gap), Holm-corrected within each family; we report 90% confidence intervals. SSIM is excluded from equivalence claims because the holdout barely changes it, which makes its margin degenerate. The concept-specific advantage under a frozen backbone is tested on the two recall metrics.

Both controls are significantly below transfer-concept on both recall metrics ($p_{\text{Holm}} < 0.001$). The non-$\mathcal{C}$ control is equivalent to the unseen baseline on all four high-level metrics and both recall metrics ($p_{\text{Holm}} \leq 0.002$), and significantly improves PixCorr, SSIM and Alex(5) ($p_{\text{Holm}} \leq 0.017$). Wrong fMRI significantly raises Object Recall ($+0.042$ [0.025, 0.058]) and Relaxed Recall ($+0.080$ [0.054, 0.106]; both $p_{\text{Holm}} = 0.004$) and significantly lowers Alex(2), Alex(5) and SSIM ($p_{\text{Holm}} \leq 0.047$); its PixCorr decrease is not significant after correction ($p_{\text{Holm}} = 0.15$).

Both frozen encoders are equivalent to their LoRA counterparts on all four high-level metrics and both recall metrics ($p_{\text{Holm}} \leq 0.002$). On the low-level metrics, equivalence is established only for Alex(5) (both encoders) and PixCorr (transfer-blind); the only significant difference is SSIM for transfer-blind (-0.003 [-0.006, -0.001]). Under a stricter margin of half the concept-specific advantage with LoRA, equivalence holds only for Relaxed Recall of transfer-concept. The frozen transfer-concept encoder significantly exceeds the frozen transfer-blind encoder on Object Recall ($+0.049$ [0.028, 0.070]) and Relaxed Recall ($+0.068$ [0.039, 0.097]; both $p_{\text{Holm}} < 0.001$). The difference in this advantage between LoRA and frozen backbones is 0.030 [0.000, 0.059] for Object Recall and 0.019 [-0.020, 0.058] for Relaxed Recall. Excluding subcategories with fewer than five qualifying images (toaster, parking meter, hair drier and toothbrush) from Object Recall changes none of these decisions. These tests account for the sampling of test images but not for variation across training runs.