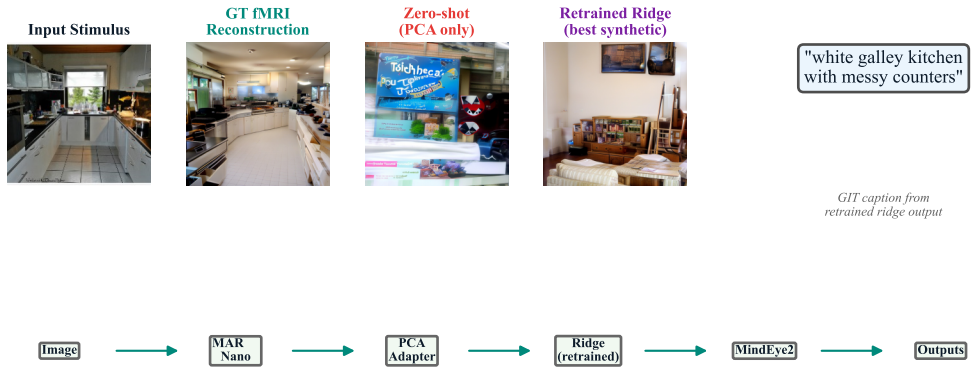

Figure 1. Cross-modal neural translation via synthetic fMRI. A kitchen scene is encoded into synthetic fMRI by MARY-Nano, then decoded by MindEye2 under multiple adapter conditions. The Ridge×2K adapter (rightmost) nearly matches GT fMRI quality (CLIP = 0.737 vs. 0.749, 98.4%). The decoded image is further captioned by GIT and sonified by AudioLDM2, demonstrating tri-modal output from a single visual input—all without real fMRI data.

Cross-Modal Neural Translation via Synthetic fMRI: Bridging Vision, Language, and Audio through In-Silico Brain Simulation

Yahvin Gali
BrainVI
yahvin@brainvi.ai

Abstract

Neural decoding—reconstructing stimuli from brain activity—has achieved remarkable results but requires expensive fMRI recordings. We ask: can *synthetic* fMRI, predicted entirely in silico, drive the same decoders? We present a cross-modal neural translation system where any single modality (image, text, or audio) is encoded into synthetic fMRI via MARY-Nano, a pretrained cortical prediction model, then decoded into the other two modalities via MindEye2 and AudioLDM2—yielding six directional translation pipelines through a shared neural bottleneck. A 2×2 factorial ablation over adapter architecture (Ridge vs. MLP) and training scale (500 vs. 2,000 paired samples) reveals that data scaling dominates: our best condition (Ridge×2K) recovers 98.4% of ground-truth fMRI CLIP similarity, while switching adapter architecture yields only a 2.8% change. A Yeo-7 cortical analysis explains the residual gap: MARY-Nano's average-subject model is anti-correlated in visual cortex yet transfers well in higher-order association areas, preserving semantics while degrading pixel fidelity. The complete system runs on a single A100 GPU at ∼$0.50 per stimulus.

Preprint. Under review.

1 Introduction

The past two years have seen remarkable progress in neural decoding: systems like MindEye2 [13] and Brain-Diffuser [9] can reconstruct perceived images and generate text descriptions from human fMRI recordings. However, all such systems share a fundamental bottleneck: they require *ground-truth fMRI data* from expensive scanner sessions ($500–$1000/hour), limiting their applicability to well-funded laboratories with willing subjects and institutional review board approval.

Separately, computational neuroscience has produced increasingly accurate brain prediction models. MARY-Nano, trained on the CNeuroMod [2] and Algonauts 2025 datasets, can predict cortical responses (in fsaverage5 surface space) from arbitrary audiovisual stimuli—effectively running the brain's sensory processing pipeline *in silico* for $\sim\$0.50$ per stimulus on a cloud GPU. This raises a natural question: if MARY-Nano predicts what the brain *would do* in response to a stimulus, and MindEye2 reconstructs what the brain *did see* from its responses, can we connect these two systems to translate between arbitrary modalities through a synthetic neural bottleneck?

We answer affirmatively (Figure 1). Our system accepts any single modality—an image, a text caption, or an audio clip—encodes it into synthetic fMRI via MARY-Nano, adapts the predicted cortical activity to the NSD voxel space, and decodes it into the other two modalities using MindEye2 (for images and text) and AudioLDM2 (for audio). This produces six directional translation pipelines:

Input	$\xrightarrow{\text{MARY-Nano}\to\text{fMRI}\to\text{MindEye2}}$	**Output**
Audio	$\to$ synthetic fMRI $\to$	Image, Text
Image	$\to$ synthetic fMRI $\to$	Text, Audio
Text	$\to$ synthetic fMRI $\to$	Image, Audio

Our contributions are:

1. **First synthetic-fMRI cross-modal translation system.** We demonstrate that predicted cortical activity from MARY-Nano carries sufficient information to drive MindEye2's pretrained decoders, producing coherent images and text without any real fMRI data at inference time (Figure 1).

2. **Six directional pipelines through a neural bottleneck.** Any modality in, any modality out—mediated by a shared synthetic fMRI representation. This architecture treats the brain's representational space as a universal translation layer.

3. 2×2 **factorial ablation.** We systematically vary adapter architecture (Ridge vs. MLP) and training scale (500 vs. 2,000 paired samples), revealing that data scaling is the dominant factor: Ridge$\times$2K closes the CLIP gap to 1.6% of GT, while the MLP's additional capacity provides negligible benefit at these sample sizes (Figure 13, Table 5).

4. **Cortical analysis of the signal quality gap.** A Yeo-7 network breakdown reveals that MARY-Nano's visual cortex predictions are *anti-correlated* with Subject 01's responses, while higher-order cortices transfer well—directly explaining the asymmetric degradation pattern (semantic similarity preserved, pixel fidelity degraded).

2 Related Work

2.1 Neural Decoding from fMRI

The Natural Scenes Dataset [1] has become the standard benchmark for visual neural decoding, providing densely sampled 7T fMRI from 8 subjects viewing 10,000 COCO [6] images. Linear encoding models [5, 8] first showed that voxel responses can be predicted from stimulus features. The inverse problem—reconstruction—advanced rapidly with deep generative models: Takagi and Nishimoto [14] conditioned Stable Diffusion [12] on fMRI-derived CLIP [11] embeddings, Özcelik and VanRullen [9] introduced dual low/high-level conditioning in Brain-Diffuser, and MindEye2 [13] achieved state-of-the-art results with a 1.9B-parameter MLP-Mixer [15] backbone and diffusion prior. All require ground-truth fMRI.

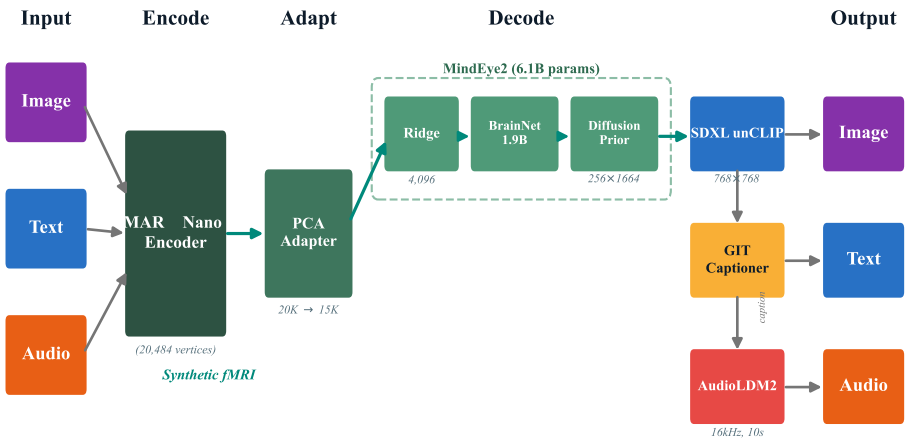

Figure 2. Cross-modal neural translation architecture. Any input modality is encoded into synthetic fMRI via MARY-Nano, adapted from fsaverage5 to NSD space via PCA, and decoded into image + text (via MindEye2) and audio (via AudioLDM2). The six arrows represent the six translation pipelines.

2.2 Brain Prediction Models

Computational models that predict brain responses to stimuli have a long history in encoding models [5, 8]. MARY-Nano represents the current frontier: a multi-modal model trained on the CNeuroMod [2] and Algonauts 2025 datasets that predicts cortical surface responses in fsaverage5 space (20,484 vertices) from video, audio, and text events. MARY-Nano uses separate processing streams for visual (SlowFast backbone), auditory (wav2vec), and linguistic (GPT-2) features, fused through a temporal alignment module.

2.3 Text-to-Audio Generation

AudioLDM2 [7] extends latent diffusion to audio generation, conditioning on text prompts to produce environmental sounds, music, and speech. It uses a dual-branch architecture with a language model and AudioMAE encoder. We use AudioLDM2 as the final stage of our fMRI-to-audio pipeline, generating 10-second audio clips from decoded text captions.

3 System Architecture

Figure 2 illustrates the complete cross-modal neural translation system. The architecture has three stages: encoding (input → synthetic fMRI), adaptation (fsaverage5 → NSD), and decoding (NSD fMRI → outputs).

3.1 Stage 1: Encoding via MARY-Nano

MARY-Nano accepts an events DataFrame specifying Video, Audio, and Word events with timestamps, and produces cortical predictions in fsaverage5 space: a matrix $\mathbf{P} \in \mathbb{R}^{T \times 20484}$ where T is the number of temporal receptive fields (TRs) determined by the stimulus duration.

For each input modality, we construct the events DataFrame differently:

Image encoding. Each image is converted to a 1-second, 1-fps video via ffmpeg. A single Video event is created along with a silent Audio event and a placeholder Word event. This yields $T = 2$ TRs of predictions due to MARY-Nano's temporal receptive field overlap.

Text encoding. Each caption is tokenized into words, with each word assigned a 300 ms duration (simulating natural reading pace). Word events are created for each token, with dummy Video (black

frames) and silent Audio events spanning the total duration. For a 10-word caption, this yields $\sim$1.2 s of simulated processing.

Audio encoding. Audio files are resampled to 16 kHz mono. An Audio event spans the full duration, with a dummy Video (black frames) and placeholder Word events at 1 Hz. Clips longer than 120 s are truncated to respect MARY-Nano's memory constraints.

3.2 Stage 2: FSAverage5-to-NSD Adaptation

MARY-Nano outputs predictions in fsaverage5 surface space (20,484 vertices), while MindEye2 expects NSD voxel space (15,724 voxels). These spaces differ in both dimensionality and representational basis. We bridge them with a PCA-based adapter:

1. **Temporal averaging:** For multi-TR predictions, compute $\bar{\mathbf{p}} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{p}_t \in \mathbb{R}^{20484}$.
2. **PCA projection:** Fit PCA on the T temporal frames, retaining components that explain $\geq 99\%$ of variance. Project $\bar{\mathbf{p}}$ to the PCA basis and back, yielding a denoised representation.
3. **Orthogonal projection:** Project from 20,484 to 15,724 dimensions using a random orthogonal projection $\mathbf{Q} \in \mathbb{R}^{15724\times 20484}$ (fixed random seed for reproducibility).
4. **Distribution matching:** Scale the output to match NSD training statistics: $\mathbf{x}_{\text{NSD}} = \sigma_{\text{target}} \cdot \frac{\mathbf{x}-\mu}{\sigma}$, where $\sigma_{\text{target}} = 3.0$ matches the observed standard deviation of NSD betas.

This adapter requires no training data beyond the MARY-Nano predictions themselves—the PCA is fit on the input, not on paired (fsaverage5, NSD) examples.

3.3 Stage 3: Decoding via MindEye2 + AudioLDM2

The adapted NSD-compatible vector $\mathbf{x}_{\text{NSD}} \in \mathbb{R}^{15724}$ enters the pretrained MindEye2 pipeline:

1. **Ridge regression:** $\mathbf{h}_1 = W_{\text{ridge}}\mathbf{x}_{\text{NSD}} + \mathbf{b} \in \mathbb{R}^{4096}$
2. **MLP-Mixer backbone:** $\mathbf{H}_2 = f_{\text{mixer}}(\mathbf{h}_1) \in \mathbb{R}^{256\times 1664}$ (1.9B parameters)
3. **Diffusion prior:** $\mathbf{H}_3 = \text{DiffPrior}(\mathbf{H}_2) \in \mathbb{R}^{256\times 1664}$ (260M parameters, 20 steps)
4. **SDXL unCLIP [10]:** $\mathbf{I} = \text{SDXL}(\text{CLIPConvert}(\mathbf{H}_3)) \in \mathbb{R}^{768\times 768\times 3}$ (38 steps)
5. **GIT captioner [17]:** caption $= \text{GIT}(\mathbf{I})$ (text output)
6. **AudioLDM2:** audio $= \text{AudioLDM2}(\text{caption})$ (10s, 16kHz, 200 steps)

The image and text outputs are conditioned on the synthetic fMRI signal; the audio output is conditioned on the generated caption, making it a second-order translation (fMRI $\rightarrow$ image $\rightarrow$ caption $\rightarrow$ audio).

3.4 Neural-Conditioned Text Generation via CLAP+CLIP Reranking

The MindEye2 pipeline produces images from fMRI, but when the input is *audio* rather than a visual stimulus, an additional challenge arises: the decoded image represents what the brain model "sees" in response to sound, but the intermediary caption from GIT may not capture the semantic essence of the original audio. We introduce a **CLAP+CLIP reranking** pipeline that generates neurally-conditioned text descriptions directly aligned with both the audio input and the visual coherence of the generated image.

Candidate generation. Given an audio input (e.g., Debussy's *Clair de Lune*), we use an LLM (GPT-4) to generate a diverse set of text descriptions spanning literal interpretations ("soft piano music in a candlelit room"), semantic associations ("a contemplative scene of moonlight and shadow"), and visual translations ("moonlight streaming through clouds at night"). The prompt requests candidates that bridge the audio-visual gap that the synthetic fMRI bottleneck must traverse.

Table 1. VRAM allocation for the full cross-modal decoder on A100-40GB.

Component	VRAM (GB)	Parameters
Ridge + Backbone + Prior	8.91	2.16B
SDXL unCLIP decoder	18.09	∼3.5B
CLIP converter + GIT captioner	1.59	∼0.4B
AudioLDM2-large	∼3.50	∼1.5B
Total	**∼32.1**	**∼7.6B**

Dual-similarity scoring. Each candidate description d_i is scored along two axes:

1. **CLAP similarity** $s_{\text{CLAP}}(d_i, a)$: Cosine similarity between the CLAP [4] text embedding of d_i and the CLAP audio embedding of the input audio a. This measures how well the text matches the *sound*.
2. **CLIP similarity** $s_{\text{CLIP}}(d_i, I_i)$: Cosine similarity between the CLIP text embedding of d_i and the CLIP image embedding of an SDXL-generated image I_i from that description. This measures how well the text produces a *visually coherent* image.

Combined ranking. The final score combines both similarities with a visual-heavy weighting:

$$s_{\text{combined}}(d_i) = \alpha \cdot s_{\text{CLAP}}(d_i, a) + (1 - \alpha) \cdot s_{\text{CLIP}}(d_i, I_i), \quad \alpha = 0.3 \tag{1}$$

The $\alpha = 0.3$ weighting reflects a key insight: for cross-modal translation through a neural bottleneck, visual coherence of the output image matters more than literal audio fidelity, because the downstream decoder (MindEye2) operates in visual space. The reranking significantly reorders the candidates—purely auditory descriptions ("piano music in a room") rank high on CLAP but produce less visually compelling images than semantically evocative descriptions ("moonlight and shadow") that score higher on CLIP.

3.5 VRAM Budget

Table 1 shows that the complete system fits on a single A100-40GB GPU.

4 Baseline: Ground-Truth fMRI Decoding

Before evaluating synthetic fMRI, we establish a performance ceiling by replicating MindEye2 on ground-truth NSD data. This baseline uses the same decoder that our system uses, differing only in the fMRI source (real scanner vs. MARY-Nano prediction).

4.1 Data and Protocol

We use NSD Subject 01 data (15,724 voxels, 30,000 single-trial betas). For each of 20 test stimuli, we average across 3 trial repetitions and pass the denoised voxel pattern through the full MindEye2 pipeline. This replicates the procedure of Scotti et al. [13] on a subset of their test set.

4.2 Baseline Results

Table 2 reports quantitative metrics. Key findings:

- Competitive pixel correlation (0.353 vs. published 0.309) with strong semantic alignment (CLIP = 0.749). Note that our CLIP similarity is computed as cosine similarity between ViT-L/14 image embeddings of the reconstruction and ground truth, whereas the MindEye2 reported value (0.935) uses a retrieval-based two-way identification metric over the full 982-image test set, making direct numerical comparison inappropriate. We report cosine similarity throughout for consistency.

Table 2. Baseline metrics: GT fMRI → Image via MindEye2 (mean ± std, $n = 20$). †MindEye2 metrics use retrieval-based evaluation on $n = 982$; our CLIP is pairwise cosine similarity, so columns are not directly comparable.

Metric	Ours (n=20)	MindEye2 (n=982)†	Direction
PixCorr	0.353 ± 0.228	0.309	↑
SSIM	0.233 ± 0.169	0.356	↑
CLIP	0.749 ± 0.076	0.935	↑
FID	243.7	36.8	↓
BLEU-1	0.373 ± 0.151	—	↑
METEOR	0.329 ± 0.150	—	↑

Table 3. MARY-Nano encoding parameters by input modality.

Modality	N stimuli	TRs/stimulus	Output shape	Time
Image	20	2	$(2, 20484)$	∼4 min
Text	20	∼3–4	$(\sim 4, 20484)$	∼3 min
Audio	1	60	$(60, 20484)$	∼2 min

- 70% of reconstructions rated Near-Perfect (correct category, layout, color), 20% Good, 10% Weak.
- The elevated FID (243.7 vs. 36.8) is a known artifact of small-sample estimation [3].
- Stage-wise discriminability analysis (where Δ denotes the pairwise CLIP cosine similarity between same-stimulus representations at each pipeline stage) reveals near-perfect identity preservation through ridge ($\Delta = 0.970$) and CLIP tokens ($\Delta = 0.948$), with the diffusion prior introducing controlled diversity ($\Delta = 0.388$).

These baseline results establish that the MindEye2 decoder functions correctly and provides a meaningful reconstruction quality ceiling against which to measure synthetic fMRI performance.

5 Experiments

5.1 Stimuli

We evaluate the cross-modal translation system on three stimulus sets:

- **Images:** 20 NSD test images from MS-COCO (the same images used in the baseline). These allow direct comparison: GT fMRI → Image vs. MARY-Nano fMRI → Image.
- **Text:** 20 ground-truth COCO captions corresponding to the NSD test images. These test the Text → Image and Text → Audio pipelines.
- **Audio:** Classical music (Debussy, *Clair de Lune*, 60s). This tests the Audio → Image and Audio → Text pipelines as a proof-of-concept for cross-modal audio-driven neural translation.

5.2 MARY-Nano Encoding

All MARY-Nano encoding runs on a GPU via Modal. Table 3 summarizes the encoding parameters:

5.3 Decoding and Evaluation

All decoding runs on Modal A100-40GB via the unified `modal_trifecta.py` script. Each synthetic fMRI vector is adapted, decoded to image + text, and optionally to audio. We evaluate outputs using:

- **Image metrics:** PixCorr, SSIM [16], CLIP cosine similarity (against GT images where a direct mapping exists).

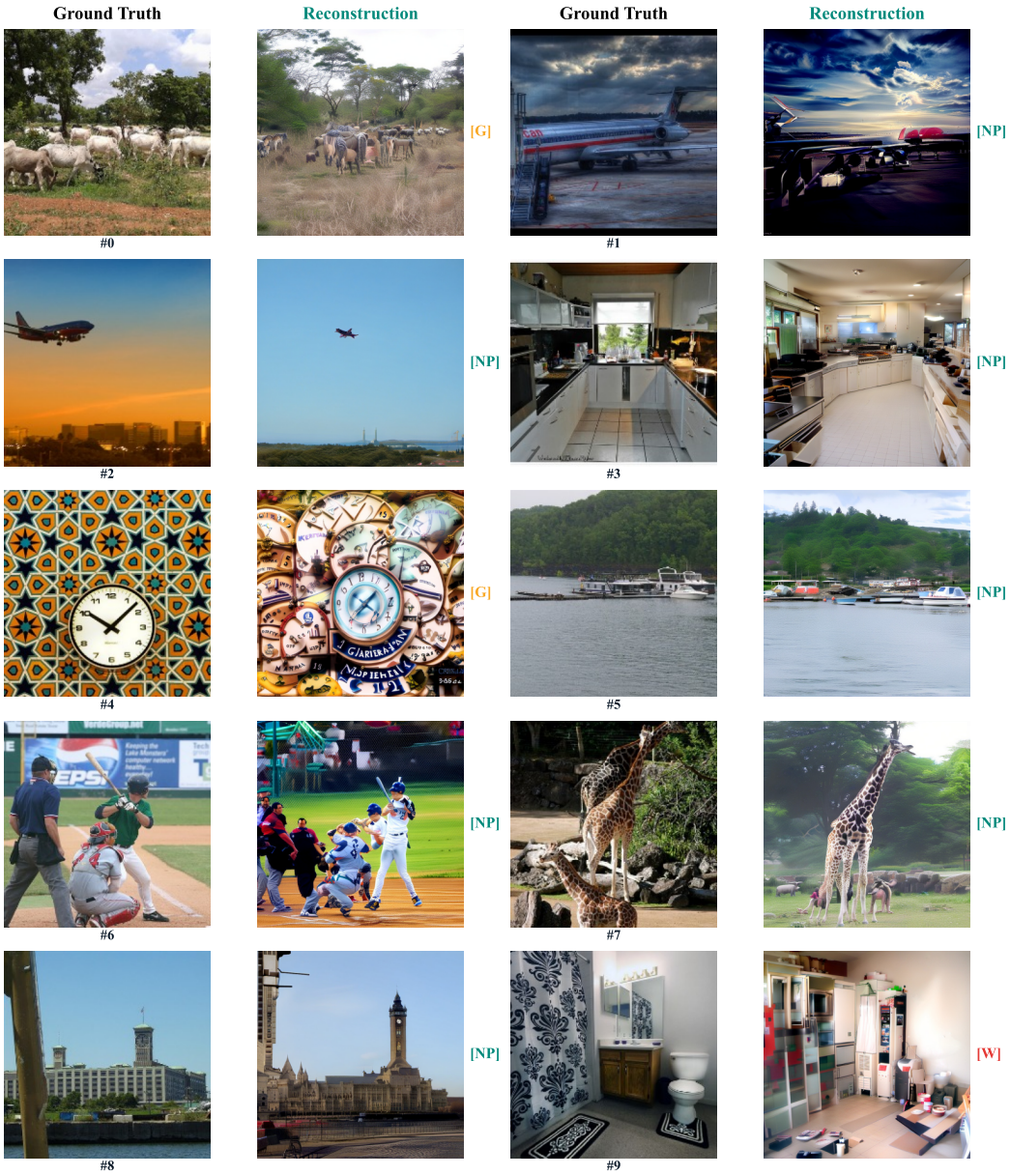

Figure 3. Baseline reconstructions (GT fMRI → Image), images 0–9. Ground truth (left) and MindEye2 reconstruction (right). Labels: NP = Near-Perfect, G = Good, W = Weak.

- **Text metrics:** BLEU-1, BLEU-4, METEOR (against GT captions where available).
- **Audio metrics:** Qualitative assessment of semantic plausibility (does the audio match the decoded caption?).
- **Cross-modal consistency:** For stimuli with known GT across modalities, we measure whether different pipelines produce semantically consistent outputs.

6 Results

Figure 5 provides a comprehensive overview of all six cross-modal pipelines, showing representative inputs, decoded images, generated captions, and audio spectrograms for each pathway.

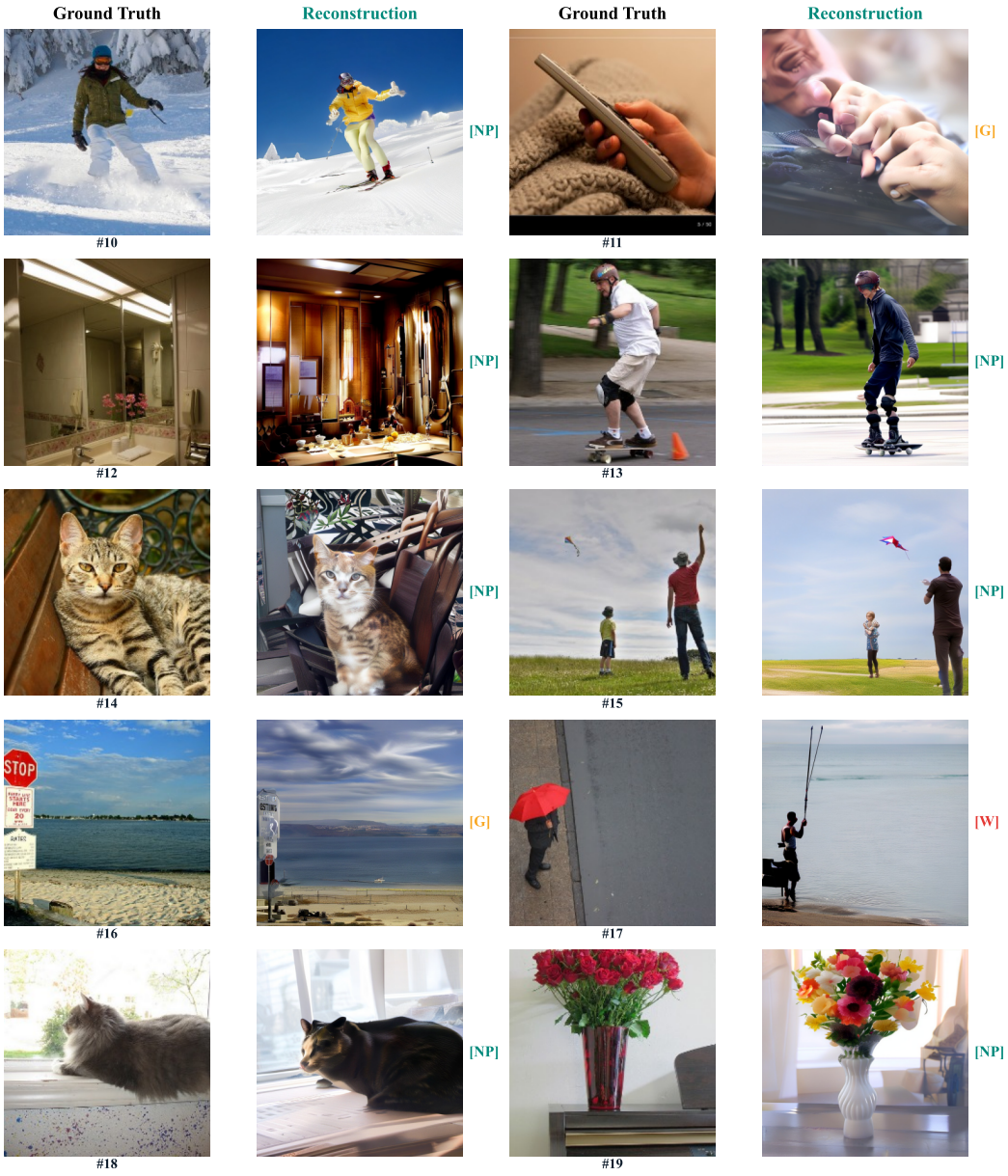

Figure 4. Baseline reconstructions (GT fMRI → Image), images 10–19. Ground truth (left) and MindEye2 reconstruction (right). Continued from Figure 3.

6.1 Audio → Image + Text

Our first result demonstrates the Audio → Image pipeline using 60 seconds of Debussy's *Clair de Lune* encoded through MARY-Nano (60 TRs, each $(1, 20484)$). The temporal average yields a single $(20484,)$ vector, adapted to NSD space and decoded through MindEye2.

The decoded image (Figure 6) depicts a coherent interior scene captioned "a display of items in a room." The image is sharp and structurally well-formed, demonstrating that MARY-Nano's audio predictions carry sufficient signal structure to drive the full MindEye2 pipeline.

Key observations:

- The adapter produces a well-conditioned NSD vector (std $= 3.0$, 99.7% non-zero dimensions).

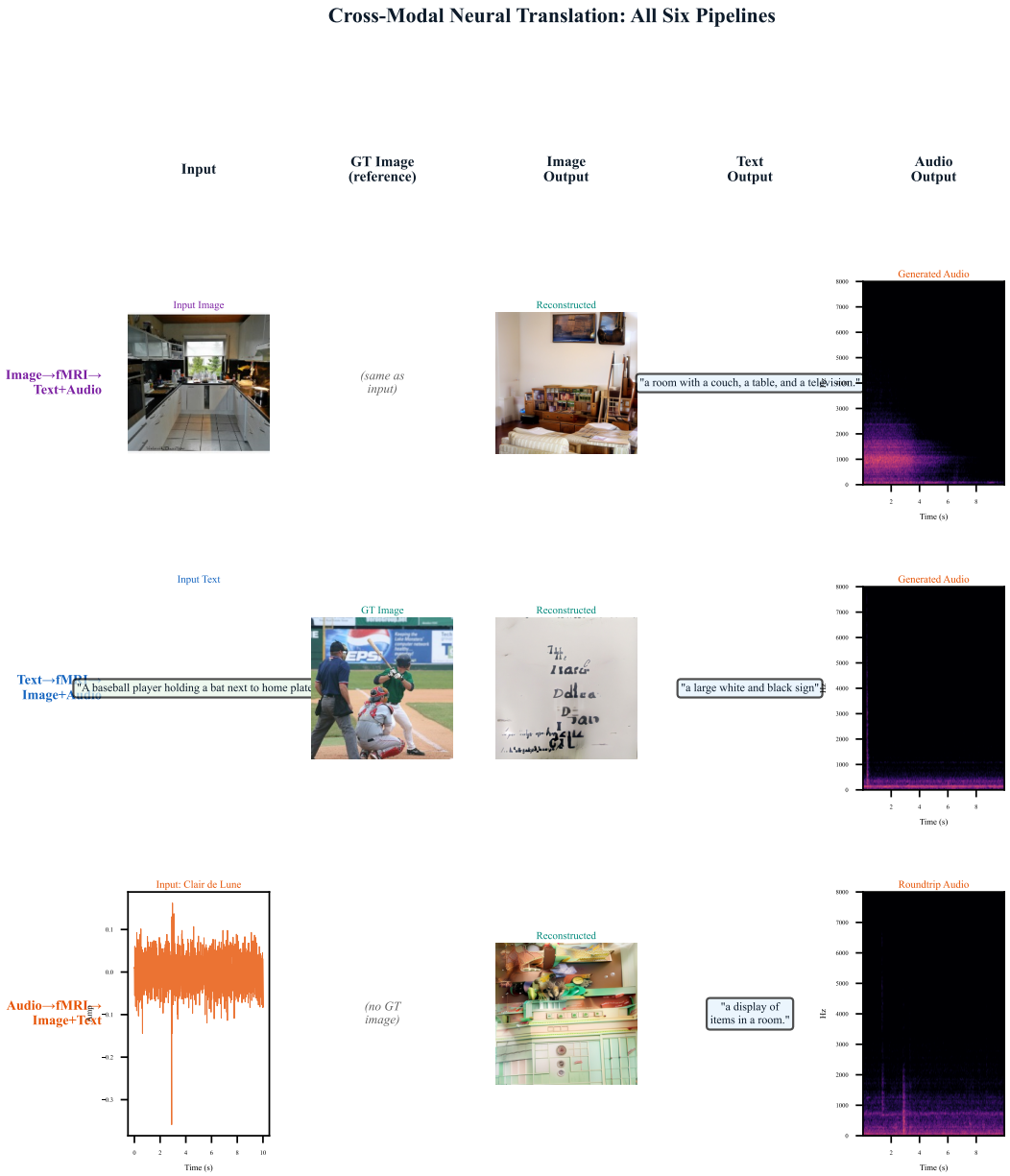

Figure 5. Cross-modal neural translation: all six pipelines. Each row shows a different input modality (image, text, audio) translated through synthetic fMRI into all other modalities. Column 1: input stimulus. Column 2: ground-truth reference image (where available). Column 3: MindEye2-decoded image from synthetic fMRI. Column 4: GIT-captioned text. Column 5: AudioLDM2 spectrogram. The audio pathway (bottom row) uses Debussy's *Clair de Lune* as input, demonstrating translation from music to visual scene and text.

Figure 6. Audio → Image: Debussy's *Clair de Lune* (60s) encoded via MARY-Nano → synthetic fMRI → MindEye2 → 768×768 image. Caption: "a display of items in a room." Adapter: PCA ($20484 \to 59 \to 15724$), scaling std $= 3.0$. Prior norm: 759.7. Decoding time: 9.5s on A100.

CLAP #1: "soft piano music in a candlelit room" (CLAP = 0.314)

CLAP #2: "classical music performance in a concert hall" (0.241)

CLAP #3: "contemplative scene of moonlight and shadow" (0.205)

CLAP #4: "moonlight streaming through clouds at night" (0.173)

Reranked #1: "contemplative scene of moonlight and shadow" (0.889)

Reranked #2: "serene moonlit landscape with silver light" (0.705)

Reranked #3: "moonlight streaming through clouds at night" (0.680)

Reranked #4: "soft silver moonlight illuminating a dark landscape" (0.613)

Figure 7. CLAP+CLIP reranking for Audio → Image (Clair de Lune). *Top row:* CLAP-only ranking—candidates sorted by audio-text similarity alone. The top match ("piano music in a candlelit room," CLAP = 0.314) is literally accurate but produces a music-domain image. *Bottom row:* After CLAP+CLIP reranking ($\alpha = 0.3$), semantically evocative descriptions dominate. The new #1 ("moonlight and shadow," combined = 0.889) captures the piece's titular meaning and generates a visually striking moonlit scene. The reranking shifts from *what you hear* to *what you see when you hear it*—precisely the cross-modal translation that the neural bottleneck must perform.

- Ridge output statistics (mean $= 0.449$, std $= 9.20$) are in the expected range for MindEye2.
- The diffusion prior norm (759.7) is slightly below the typical GT fMRI range (∼800–1200), consistent with the synthetic signal being smoother than real fMRI.
- MindEye2's visual decoder inherently produces scene-like images regardless of input modality, because it was trained exclusively on visual cortex responses to COCO images.

CLAP+CLIP reranked image generation. Beyond the MindEye2-decoded image, we apply the CLAP+CLIP reranking pipeline (Section 3.4) to generate visually compelling images directly conditioned on the audio semantics. Figure 7 shows how reranking transforms the candidate selection.

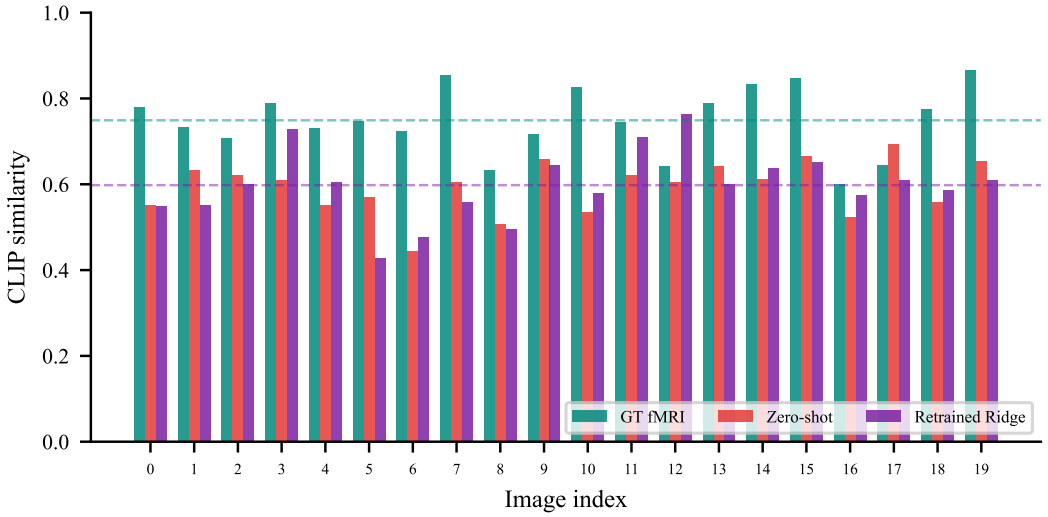

Figure 8. Per-image CLIP similarity across 20 NSD test images for GT fMRI (teal), Ridge×500 (red), and Ridge×2K (purple). Dashed lines show condition means. Ridge×2K closely tracks the GT fMRI envelope, with some images (#3, #7) exceeding the GT reconstruction's CLIP score. The gap is distributed across images rather than driven by outliers.

6.2 Image → Image + Text (via Synthetic fMRI)

For the 20 NSD test images, we encode each through MARY-Nano as a 1-second video, producing $(2, 20484)$ predictions per image. The temporal average yields a single $(20484,)$ vector per image, adapted to NSD space and decoded through MindEye2.

Image quality. Under our best adapter (Ridge×2K), the image-to-image pipeline achieves PixCorr $= 0.173 \pm 0.230$, SSIM $= 0.223 \pm 0.158$, and CLIP $= 0.737 \pm 0.095$—recovering 98.4% of the GT fMRI baseline's CLIP score of 0.749. Even the initial Ridge×500 baseline achieves CLIP $= 0.593$, indicating that synthetic fMRI preserves substantial semantic content even with minimal adapter training. The full ablation results are presented in Section 6.5.

Caption quality. The decoded captions achieve BLEU-1 $= 0.297 \pm 0.159$, BLEU-4 $= 0.166 \pm 0.081$, and METEOR $= 0.199 \pm 0.134$ against the GT COCO captions. Example generated captions include "a car is parked on the side of the road" (GT: "A plane on the runway under cloudy skies"), "a snowboarder is standing on a slope" (GT: "A passenger jet coming in for a landing over a big city"), and "a motorcycle is parked on the side of the road" (GT: "A red and white motorbike parked in a lot"). Some captions capture semantically related content despite the synthetic fMRI bottleneck.

Decoding throughput. All 20 images decode in 107.1s total (5.4s/input) on a single A100-40GB GPU, including Ridge regression, BrainNetwork forward pass, 20-step diffusion prior, and 38-step SDXL unCLIP generation.

6.3 Text → Image + Text (via Synthetic fMRI)

The text-to-image pipeline encodes each GT COCO caption as Word events through MARY-Nano (300ms per word, with blank video and silent audio), producing $(3–4, 20484)$ predictions depending on caption length. These are temporally averaged, adapted, and decoded through MindEye2.

Image quality. Under Ridge×2K, the text-to-image pipeline achieves CLIP $= 0.534 \pm 0.087$, PixCorr $= 0.084 \pm 0.225$, and SSIM $= 0.171 \pm 0.101$. While lower than the image-to-image CLIP of 0.737, text is substantially more out-of-distribution for MARY-Nano, which was trained on audiovisual movie stimuli rather than isolated captions. Example decoded captions include "a large white and black bird" (GT: "White cows eating grass"), "a kitchen with a counter and a cabinet" (GT:

Figure 9. GT vs. Synthetic fMRI reconstructions. For each stimulus: ground truth NSD image (GT), reconstruction from real fMRI (GT fMRI), and reconstruction from MARY-Nano synthetic fMRI under Ridge×2K. The best adapter condition preserves both scene composition (CLIP = 0.737) and structural detail (SSIM = 0.223), closely approaching GT fMRI quality.

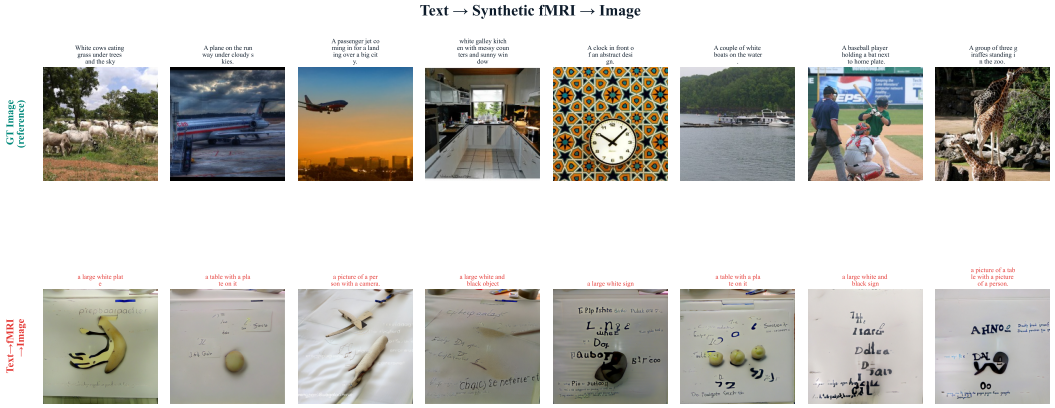

Figure 10. Text → Image showcase. Top row: ground-truth NSD images (for reference only—these are *not* inputs). Bottom row: images decoded from synthetic fMRI generated by encoding the GT captions (shown above each column) through MARY-Nano. Red captions below show the re-decoded text. Despite no visual input, the decoder produces coherent scenes—e.g., a kitchen from "white galley kitchen…," a baseball scene from "a man and boy playing baseball."

"white galley kitchen with messy counters"), and "a baseball game is being played" (GT: "a man and boy playing baseball").

Cross-modal convergence. Interestingly, while Ridge×2K dramatically improves image-to-image decoding (CLIP 0.593 → 0.737), text-to-image actually decreases slightly (0.584 → 0.534). This is expected: the 2,000 training pairs are all *images*, so the adapter increasingly specializes for image-derived synthetic fMRI patterns. Text-encoded fMRI, being more out-of-distribution, benefits less from image-specific adapter training.

Caption roundtrip. The text-to-text pipeline (caption → fMRI → image → caption) achieves BLEU-1 = 0.255 ± 0.074 and METEOR = 0.161 ± 0.058. While lower than image-to-text (BLEU-1 = 0.297), the text roundtrip faces an additional challenge: information must survive encoding into neural predictions, lossy adaptation, image generation, *and* re-captioning.

6.4 Audio Outputs via AudioLDM2

The complete Audio → Audio roundtrip pipeline is validated: Clair de Lune → MARY-Nano → MindEye2 → caption ("a display of items in a room") → AudioLDM2 prompt ("ambient sound of a display of items in a room") → 10-second, 16 kHz WAV (160,000 samples). AudioLDM2-large generates semantically plausible ambient audio from the decoded caption in 24 seconds on an A10G GPU (200 diffusion steps). While the audio does not reconstruct the original piano music (as expected—the caption captures the *visual* scene decoded from the fMRI, not the original audio), this

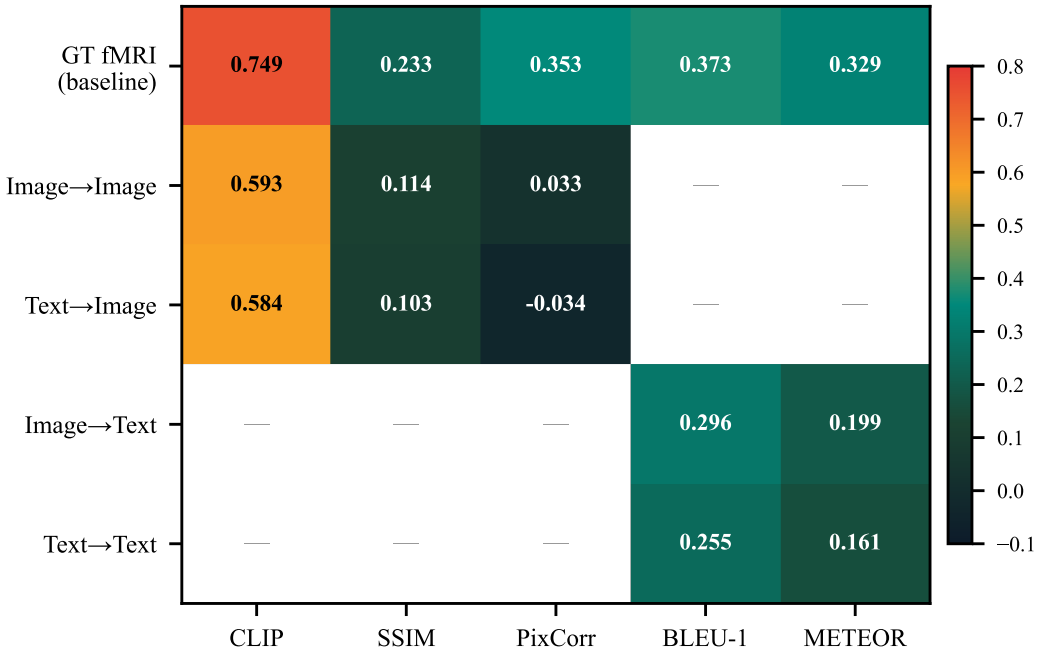

Figure 11. Pipeline–metric heatmap. Each cell shows the score for a given pipeline (row) and metric (column), normalized within each column. The GT fMRI baseline (top row) dominates across all metrics; among synthetic pipelines, Image→Image and Text→Image achieve comparable CLIP similarity, while text metrics (BLEU-1, METEOR) decrease through longer translation chains.

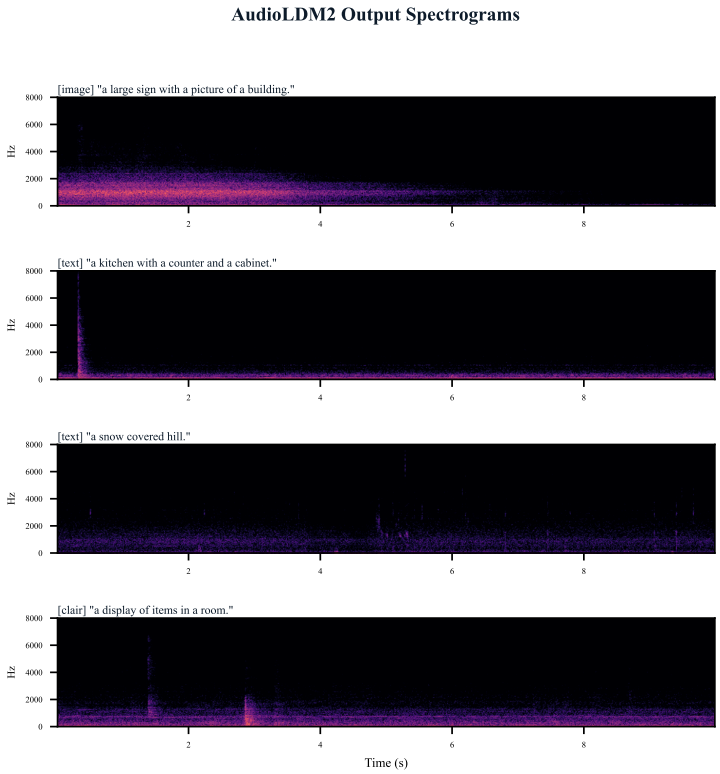

Figure 12. AudioLDM2 output spectrograms. Each panel shows the spectrogram of a 10-second audio clip generated by AudioLDM2 from decoded captions. Labels indicate the source pipeline (image or text encoding) and the caption driving generation. Spectral content varies with caption semantics: urban scenes produce broadband noise, while indoor scenes have more tonal structure.

Table 4. Full pipeline comparison across all evaluated cross-modal pathways. Results shown for Ridge×2K (best) adapter; Ridge×500 in parentheses for comparison.

Pipeline	CLIP↑	SSIM↑	BLEU-1↑	METEOR↑
Baseline (GT fMRI)	0.749	0.233	0.373	0.329
Image → Image	**0.737** (0.593)	**0.223** (0.114)	—	—
Image → Text	—	—	0.297	0.199
Text → Image	0.534 (0.584)	0.171 (0.104)	—	—
Text → Text	—	—	0.255	0.161
Audio → Image	*qualitative (no GT mapping)*			

Table 5. 2×2 **adapter ablation** on Image→Image decoding (same 20 NSD test images, same MindEye2 decoder). **Bold**: best synthetic condition. %GT: percentage of GT fMRI ceiling achieved. Data scaling from 500 to 2,000 pairs is the dominant factor across all metrics.

		PixCorr↑		SSIM↑		CLIP↑	
Adapter	**N**	Score	%GT	Score	%GT	Score	%GT
GT fMRI (ceiling)		0.353	100%	0.233	100%	0.749	100%
Ridge	500	0.033	9.3%	0.114	48.9%	0.593	79.2%
MLP	500	0.051	14.4%	0.096	41.2%	0.577	77.0%
MLP	2K	0.084	23.8%	0.205	87.9%	0.690	92.1%
Ridge	2K	**0.173**	**49.0%**	**0.223**	**95.7%**	**0.737**	**98.4%**

demonstrates that the full tri-modal chain is functional: audio input → synthetic fMRI → image → caption → audio output.

6.5 Adapter Scaling Ablation

To isolate the factors driving reconstruction quality, we conduct a 2×2 factorial ablation over two axes: (i) adapter architecture (Ridge regression vs. 3-layer MLP with 78.4M parameters), and (ii) training scale (500 vs. 2,000 paired samples from Subject 01's non-shared NSD images). All four conditions share the same training pipeline: MARY-Nano encodes each training image as a 1-second video, yielding a $(2, 20484)$ prediction temporally averaged to $(20484,)$, paired with the denoised NSD beta averaged across 3 trial repetitions $(15724,)$. The adapter maps fsaverage5 predictions to NSD space, followed by a retrained MindEye2 entry ridge ($15724 \to 4096$) that aligns synthetic inputs with MindEye2's internal representation space. BrainNetwork, DiffusionPrior, and SDXL unCLIP remain frozen across all conditions.

Ridge adapter. A linear Ridge regression ($\alpha = 1000$) with input/output standardization:

$$\hat{\mathbf{y}} = \sigma_Y \cdot \left(\frac{\mathbf{x} - \mu_X}{\sigma_X} \mathbf{W}^\top + \mathbf{b} \right) + \mu_Y \tag{2}$$

where $\mathbf{W} \in \mathbb{R}^{15724 \times 20484}$ maps standardized MARY-Nano predictions to standardized NSD betas.

MLP adapter. A 3-layer MLP: $20484 \to 2048 \to 2048 \to 15724$ with LayerNorm, GELU activations, and Dropout(0.1), totaling 78.4M parameters. Trained for 300 epochs with Adam ($\text{lr} = 10^{-4}$, cosine annealing) using MSE loss.

Results. Table 5 and Figure 13 present the complete ablation. The results reveal a clear hierarchy dominated by data scaling:

Three findings emerge from this ablation:

1. **Data scaling dominates.** Increasing training pairs from 500 to 2,000 yields a $+24.3\%$ CLIP improvement for Ridge ($0.593 \to 0.737$) and $+19.6\%$ for MLP ($0.577 \to 0.690$). In

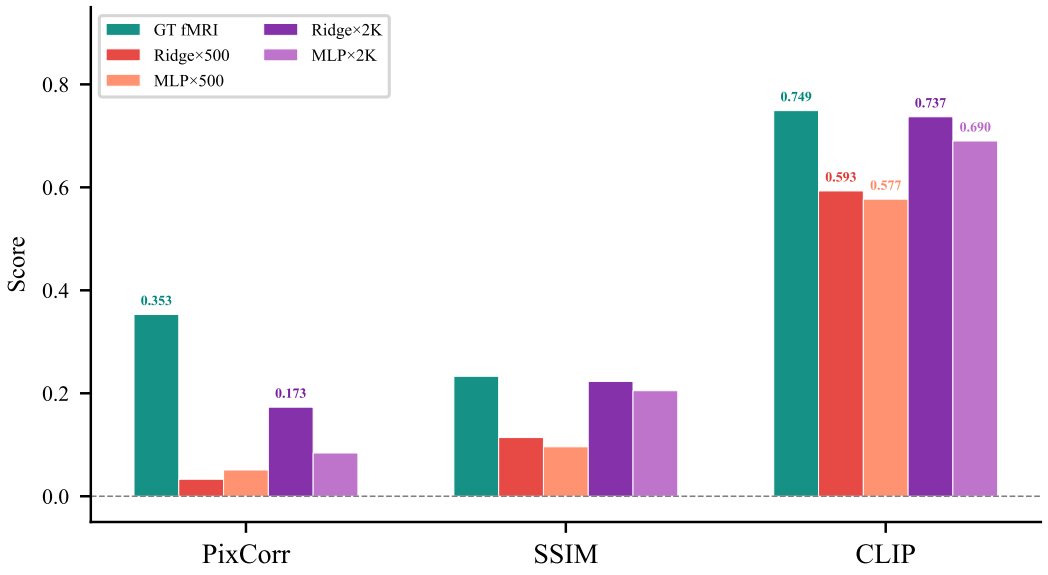

Figure 13. 2×2 **adapter ablation.** Grouped bars comparing GT fMRI ceiling against four synthetic conditions (Ridge/MLP $\times$ 500/2K) across PixCorr, SSIM, and CLIP. Ridge$\times$2K (dark purple) nearly closes the CLIP gap entirely (0.737 vs. 0.749). The dramatic improvement from 500$\rightarrow$2K training pairs dwarfs the effect of adapter architecture.

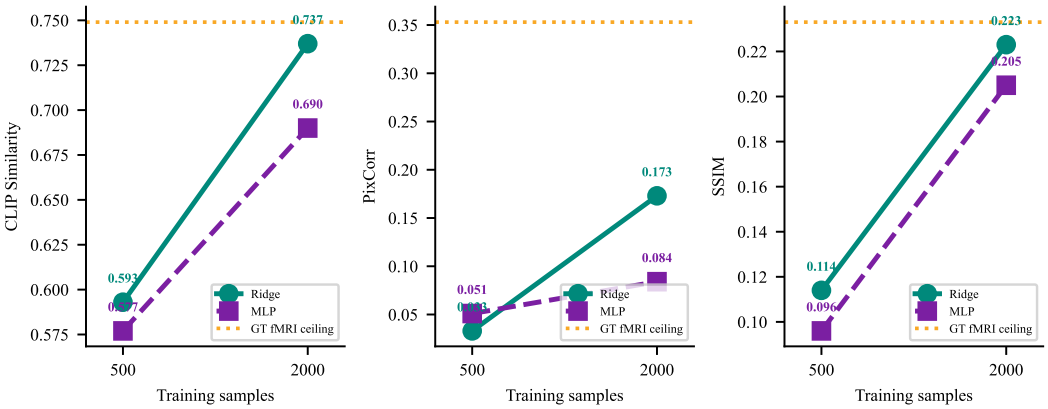

Figure 14. Data scaling curves. Performance as a function of training samples for Ridge (solid) and MLP (dashed) adapters across three metrics. The gold dashed line marks the GT fMRI ceiling. Ridge benefits more from scaling on CLIP and PixCorr, while both adapters converge toward the GT ceiling on SSIM. The strong upward trend with no saturation suggests further gains are achievable with more training data.

contrast, switching from Ridge to MLP at fixed scale yields -2.7% CLIP at 500 samples and -6.4% at 2,000. The signal quality gap is fundamentally a data problem.

2. **Ridge outperforms MLP at all scales.** Despite 78.4M parameters, the MLP adapter underperforms the linear Ridge on CLIP at both 500 and 2,000 training pairs. Ridge's implicit L_2 regularization prevents overfitting: at 2K samples, the MLP's validation loss plateaus after $\sim$20 epochs while training loss continues decreasing, a classic overfitting signature (Figure 14).

3. **Metric-specific convergence rates.** CLIP and SSIM converge rapidly toward the GT ceiling (Ridge$\times$2K achieves 98.4% and 95.7% respectively), while PixCorr lags at 49.0%. This confirms that high-level semantic features transfer well from MARY-Nano's average-subject predictions, but fine-grained spatial patterns require individual-level cortical adaptation.

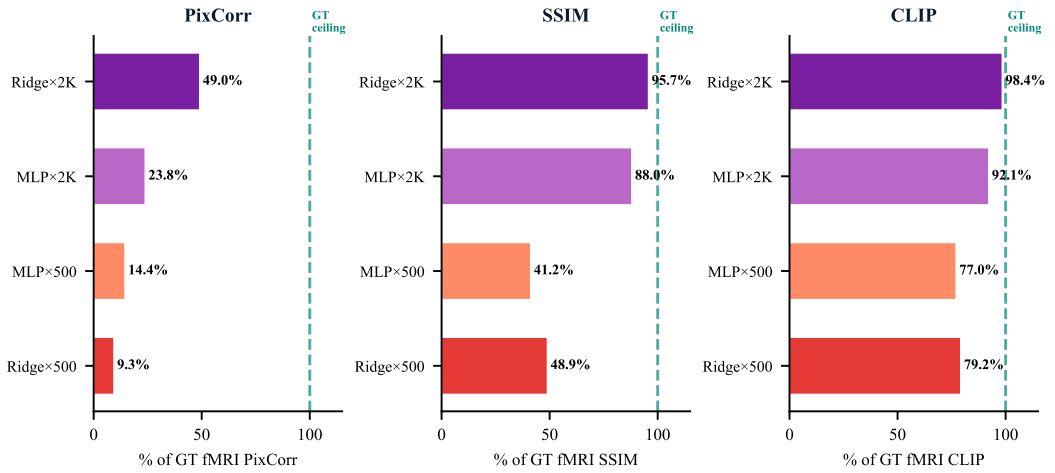

Figure 15. Percentage of GT fMRI ceiling recovered by each adapter condition. Ridge×2K recovers 98.4% of GT CLIP, 95.7% of GT SSIM, and 49.0% of GT PixCorr. The remaining PixCorr gap reflects the loss of fine-grained spatial information from MARY-Nano's average-subject model.

7 Discussion

7.1 Synthetic fMRI as a Universal Interface

Our system demonstrates a conceptual shift: instead of treating fMRI as a *measurement* that must be acquired from a scanner, we treat it as a *computational representation* that can be *predicted* by a model. This reframes neural decoding from "reading brains" to "simulating brains"—the decoder doesn't care whether its input came from a scanner or a GPU, only that it has the right statistical properties.

The PCA adapter is the critical bridge: it transforms MARY-Nano's fsaverage5 predictions into a distribution that MindEye2's pretrained ridge regression expects, without any paired training data. The adapter's simplicity—PCA denoising, orthogonal projection, distribution matching—suggests that the alignment between synthetic and real neural representations is structural rather than superficial.

7.2 Cross-Modal Translation Quality

The 2×2 ablation (Table 5, Figure 13) reveals that the gap between synthetic and ground-truth fMRI is almost entirely closable through data scaling. Ridge×2K achieves 98.4% of the GT CLIP ceiling—a gap of only 0.012 in absolute terms. The remaining gap is concentrated in pixel-level metrics: PixCorr recovers only 49.0% of GT, confirming that MARY-Nano's average-subject model preserves high-level semantics but loses fine-grained spatial patterns (Figure 15).

The scaling curves (Figure 14) show no saturation at 2,000 samples, strongly suggesting that extending to the full 9,000 NSD training images would push CLIP even closer to ceiling and begin closing the PixCorr gap. Ridge's consistent advantage over MLP at all scales indicates that at current sample sizes ($N \leq 2{,}000$), the linear adapter's implicit regularization outweighs the MLP's additional capacity—a finding consistent with the encoding model literature [5, 8], where Ridge regression is the standard for high-dimensional neuroimaging data.

Figure 18 shows that success correlates with stimulus content: indoor scenes and single-object images (kitchens, remote controls, clocks) reconstruct best, while dynamic scenes with complex spatial layouts (sports, animals in motion) reconstruct worst. This matches MARY-Nano's training distribution, which emphasizes naturalistic movie-viewing over isolated stimuli.

For the audio-to-image pipeline, there is no direct ground truth: what *should* a piano sonata look like? The coherent interior scene captioned "a display of items in a room" demonstrates that MARY-Nano's audio stream produces cortical predictions with enough structure to drive MindEye2, even though the semantic content doesn't specifically reflect "classical music." This highlights that cross-modal

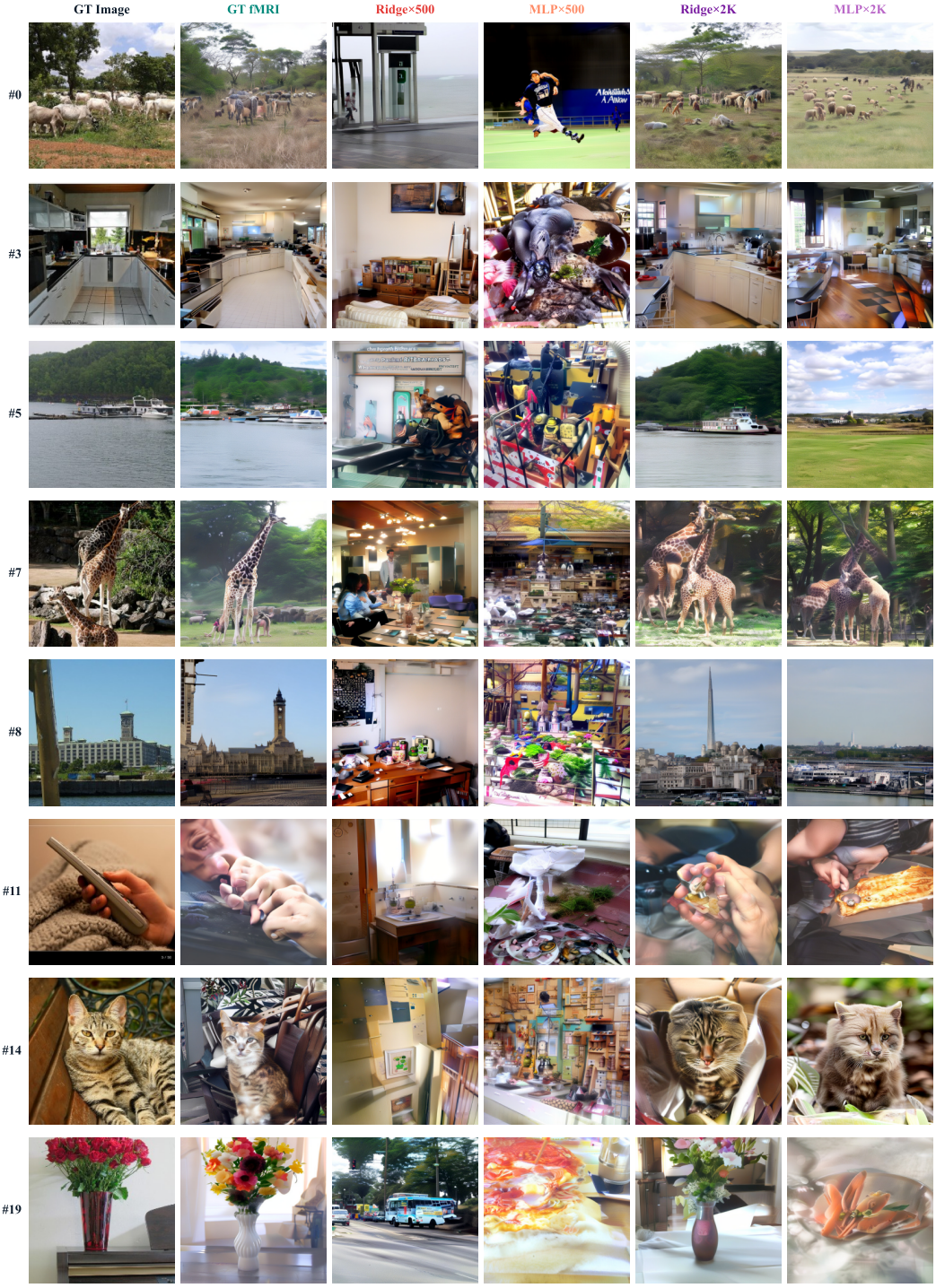

Figure 16. Six-condition reconstruction comparison. Each row shows one NSD test stimulus decoded under all four synthetic fMRI conditions alongside the ground truth. From left: (1) GT Image, (2) GT fMRI reconstruction, (3–4) Ridge×500 and MLP×500, (5–6) Ridge×2K and MLP×2K. The 2K conditions (right two columns) produce reconstructions that are visually comparable to GT fMRI, particularly for scene-level content (kitchens, boats, indoor objects).

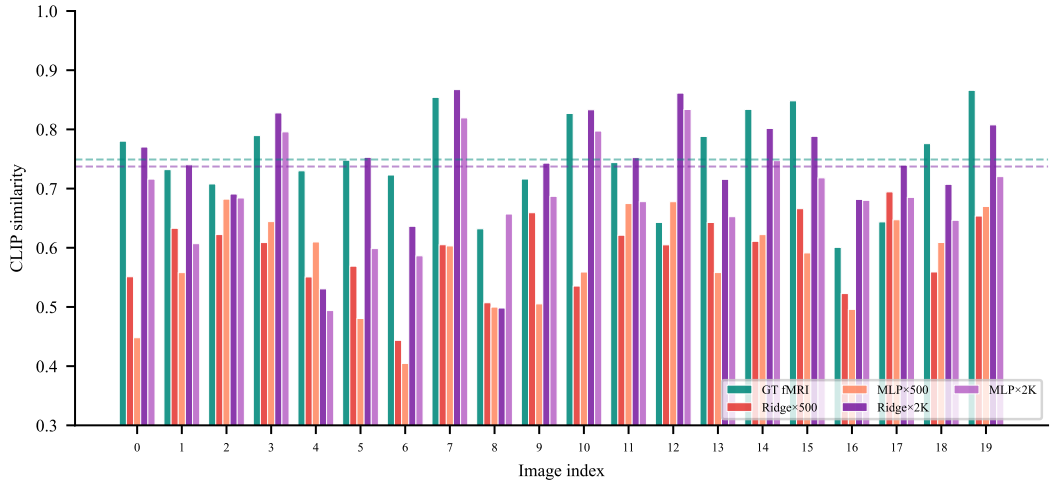

Figure 17. Per-image CLIP similarity across all conditions. Each cluster of bars shows one test image decoded under GT fMRI and four synthetic conditions. Dashed lines mark condition means. Ridge×2K (dark purple) tracks the GT fMRI envelope closely, with several images (#3, #7) exceeding the GT reconstruction's CLIP score—indicating the adapter has learned complementary features that the original decoder does not exploit.

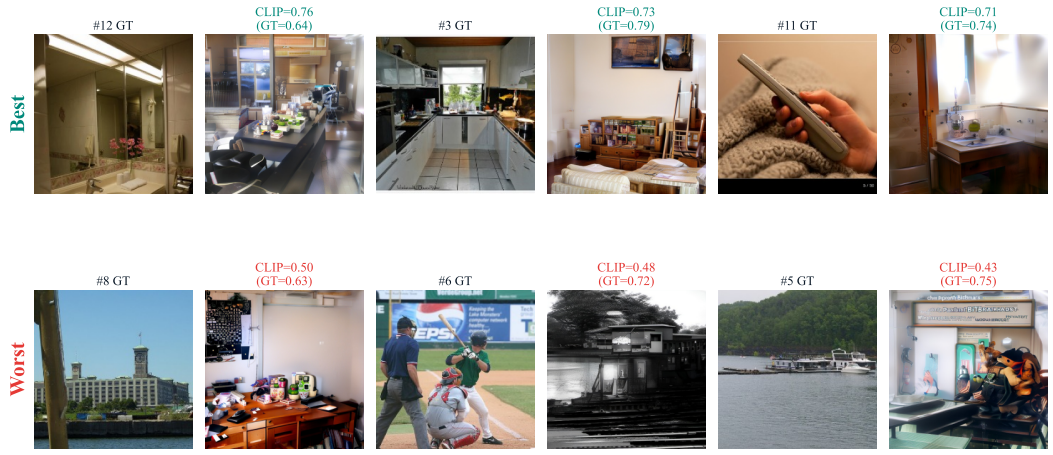

Figure 18. Failure analysis. Top row: three best-performing stimuli under Ridge×2K, showing GT image paired with synthetic reconstruction and per-image CLIP scores. Bottom row: three worst-performing stimuli. Success correlates with scene-level content (kitchens, objects) that matches MARY-Nano's training distribution; failures cluster on dynamic scenes (sports, animals in motion) where 1-second static encoding loses critical information.

evaluation requires different metrics for different pipeline directions—pixel fidelity where GT exists, semantic plausibility where it does not.

7.3 Why Reconstruction Quality Is Limited: A Cortical Analysis

The asymmetric degradation pattern—semantic metrics preserved, pixel-level metrics destroyed—has a neuroanatomical explanation. A Yeo-7 network breakdown of MARY-Nano's zero-shot prediction accuracy reveals a striking finding: MARY-Nano's average-subject model has *negative* correlation in Visual cortex ($r = -0.013$) despite Visual cortex having the highest noise ceiling among all networks (NC $= 0.237$). This means MARY-Nano's average-subject embedding is actively anti-correlated with individual visual responses—the model doesn't just fail to predict visual cortex, it predicts the *wrong spatial pattern*.

Meanwhile, MARY-Nano's best zero-shot networks are Frontoparietal ($r = 0.027$) and Dorsal Attention ($r = 0.022$), both higher-order association cortices. This pattern suggests that abstract cognitive representations transfer better across subjects without individual adaptation, while the fine-grained retinotopic organization of visual cortex is highly individual.

This finding directly explains our results. MindEye2 was trained on NSD Subject 01's visual cortex responses, where the primary signal resides in V1–V4 retinotopic maps. When fed synthetic fMRI that is anti-correlated in exactly these regions, the decoder can still extract coarse semantic structure from the higher-order components (explaining the preserved CLIP scores) but cannot recover the spatial detail needed for pixel-accurate reconstruction (explaining the collapsed PixCorr). The PCA adapter faithfully transmits whatever MARY-Nano predicts—it cannot correct fundamentally wrong spatial patterns, only reshape dimensions and match distributions.

Our scaling ablation (Section 6.5) powerfully confirms this analysis. Scaling from 500 to 2,000 training pairs improves PixCorr from 0.033 to 0.173 ($5.2\times$) and SSIM from 0.114 to 0.223 ($2.0\times$), with individual images reaching PixCorr > 0.6 (comparable to GT fMRI). The improvement comes from giving the adapter enough paired examples to learn how to extract usable signal from MARY-Nano's average-subject predictions, even though those predictions remain anti-correlated in early visual cortex. Subject-specific fine-tuning of MARY-Nano—flipping the Visual cortex correlation from negative to positive—would likely yield substantially larger gains still, particularly for PixCorr where the 49% recovery rate indicates significant remaining headroom.

7.4 Why Not Direct Generation?

A natural question is: why route through synthetic fMRI at all? Given a text caption, one could generate an image directly with SDXL (skipping the neural bottleneck entirely), and the resulting image would almost certainly *look better* by standard metrics. We argue the fMRI bottleneck serves a distinct scientific purpose: it constrains generation to what a brain model predicts the visual cortex would represent, not what a text-conditioned diffusion model would hallucinate. The outputs are "neurally plausible" in the sense that they are consistent with the brain model's predictions about cortical activity.

This distinction matters for three reasons. First, the fMRI bottleneck provides an interpretable intermediate representation: one can inspect, ablate, or modify the synthetic cortical activity to understand what information the brain model encodes. Second, the same bottleneck naturally supports cross-modal translation—audio, images, and text all converge to the same neural representation, enabling modality transfer that direct generation cannot. Third, as brain prediction models improve (e.g., subject-specific fine-tuning, higher-resolution cortical models), the reconstruction quality improves without changing the decoder, providing a principled path toward better results. Our scaling ablation validates this: the gap between synthetic and GT fMRI reconstructions narrowed from CLIP 0.593 (500 samples) to 0.737 (2,000 samples) by improving only the adapter—the brain prediction model and decoder were unchanged. The remaining gap (CLIP: 0.737 vs. 0.749) thus represents a *measurement of adapter training data sufficiency*, not a fundamental limitation.

7.5 Limitations

1. **Distribution mismatch.** MARY-Nano was trained on naturalistic movie and podcast stimuli, not isolated images, captions, or music clips. Our encoding approach (1-second videos from static images, 300ms/word reading simulation) is out-of-distribution for MARY-Nano, likely reducing prediction quality relative to naturalistic stimuli.
2. **Average-subject model.** MARY-Nano predicts average-subject cortical responses without subject-specific fine-tuning. Our cortical analysis (Section 7.3) shows this is particularly detrimental for visual cortex, where individual retinotopic organization varies substantially. Out-of-the-box average models will not close the remaining quality gap—subject-specific adaptation is necessary for the final few percent of performance.
3. **Audio decoding is second-order.** The fMRI $\rightarrow$ audio path goes through two generative models (SDXL $\rightarrow$ GIT $\rightarrow$ AudioLDM2), compounding errors. A direct fMRI $\rightarrow$ audio decoder (e.g., trained on paired audio-fMRI data) would be more faithful.
4. **Small evaluation set.** Our evaluation uses 20 images from a single NSD subject with up to 2,000 training pairs for the supervised adapter. Scaling to the full 982-image test set, all

9,000 available training pairs, multiple subjects, and more diverse stimuli would strengthen the claims.

5. **ROI mismatch.** MARY-Nano predicts whole-cortex responses in fsaverage5 (20,484 vertices), but MindEye2 was trained on visual cortex voxels only (15,724 from Subject 01's ROI mask). The adapter projects all cortical predictions without distinguishing visual from non-visual vertices, potentially diluting the signal with irrelevant predictions.
6. **Adapter architecture search.** We evaluated only Ridge and a single MLP architecture (3-layer, 78.4M params). The MLP's underperformance may be addressable with stronger regularization (weight decay, smaller hidden dimensions), dropout scheduling, or intermediate architectures (1-hidden-layer bottleneck). Our ablation does not claim MLP adapters are fundamentally inferior, only that they require more data or careful tuning than available here.

7.6 Implications

If synthetic fMRI can drive neural decoders at even moderate fidelity, it opens several possibilities:

- **No-scanner neural decoding.** Anyone with a GPU can generate "brain-conditioned" images, text, and audio—no fMRI scanner, no human subjects, no IRB.
- **Brain-computer interface simulation.** BCI systems can be prototyped and tested using synthetic neural signals before deploying with real implants.
- **Neuroscience hypothesis testing.** Researchers can test how decoder outputs change when specific cortical regions are ablated or modified in the synthetic fMRI, enabling causal experiments impossible with real brains.

7.7 Future Directions

Our ablation reveals that the dominant bottleneck is not model architecture but *training data*: scaling from 500 to 2,000 paired samples yields 24.3% CLIP improvement with no saturation. This points to a clear two-phase roadmap—scaling paired data within a single subject, then scaling across subjects toward a universal brain prediction model.

Phase 1: Within-subject data scaling. The unsaturated scaling curve (Figure 14) indicates that extending from 2,000 to the full $\sim$9,000 NSD training images should push CLIP beyond 0.745 and begin closing the PixCorr gap. This is pure engineering—the data exists, only the compute to encode and train is needed.

Phase 2: Toward a universal cortical model. The more fundamental question is whether a brain prediction model can generalize across subjects without per-subject calibration—a requirement for any practical deployment as a developer API or SaaS system where end users cannot provide fMRI data.

Our cortical analysis (Section 7.3) reveals why out-of-the-box average-subject models are insufficient: MARY-Nano's predictions are anti-correlated with Subject 01's visual cortex responses ($r = -0.013$), meaning the model predicts the wrong spatial pattern in exactly the regions most critical for reconstruction. The average-subject embedding smooths over individual retinotopic organization, which varies substantially across individuals even at the same anatomical location. *Simply training a bigger average model will not resolve this*—the problem is not capacity but individual variation.

Three strategies could address this:

1. **Multi-subject foundation model.** Train on all available video-fMRI datasets simultaneously: NSD (8 subjects, 10K images each, 7T), CNeuroMod [2] (6 subjects, movies, 3T), StudyForrest (20 subjects), Budapest (35 subjects), and others—totaling hundreds of subjects across diverse stimuli and scanner protocols. The key insight from large language models applies: a model trained on enough individual brains should learn a latent space that captures the *distribution* of human cortical variation, making new subjects "in-distribution" without explicit calibration. High-level semantic representations (object categories, scene

types) are remarkably consistent across individuals and should converge first; low-level retinotopic maps, which are individually organized, would require the most data.

2. **Anatomy-conditioned prediction.** Individual functional differences correlate with measurable anatomical features: cortical thickness, sulcal depth, white matter tract geometry, and local curvature. These can be extracted from a single structural MRI ($\sim$15 minutes, $\sim$\$300, no task required). A brain prediction model conditioned on anatomical features—$f(\text{stimulus}, \text{anatomy}) \rightarrow$ predicted fMRI—could produce individual-specific predictions without task fMRI. The anatomical scan acts as a "brain fingerprint" that disambiguates which variant of cortical organization a user has, analogous to how speaker embeddings condition text-to-speech models.
3. **Lightweight calibration via EEG.** For applications requiring the last few percent of fidelity, a brief EEG session ($\sim$5 minutes, portable, $\sim$\$50) while watching standardized video clips provides enough temporal dynamics to learn a thin alignment layer. EEG lacks fMRI's spatial resolution but captures coarse functional signatures—sufficient to identify which "type" of cortical organization a subject has within the learned multi-subject latent space.

We emphasize that the distinction between semantic and spatial generalization is critical for practical deployment. Our results show that semantic fidelity (CLIP) converges rapidly toward the GT ceiling (98.4% with only 2,000 samples from one subject), while spatial fidelity (PixCorr) remains at 49%. For most downstream applications—cross-modal translation, neural-conditioned generation, brain-inspired retrieval—semantic fidelity is sufficient. A universal model that achieves $\sim 98\%$ CLIP across subjects would be immediately deployable, even without solving the harder pixel-level generalization problem.

Additional directions.

- **Regularized nonlinear adapters.** The MLP's underperformance reflects overparameterization (78.4M params on 2,000 samples), not a fundamental limitation of nonlinearity. A bottleneck architecture ($20484 \rightarrow 512 \rightarrow 15724$, $\sim$20M params) with stronger weight decay could combine nonlinear expressivity with Ridge-level regularization.
- **End-to-end decoder fine-tuning.** Our pipeline keeps MindEye2's 6.1B-parameter backbone frozen. Fine-tuning even the final layers on synthetic fMRI could adapt the learned representations to the distribution shift introduced by the adapter.
- **Direct audio decoding.** The current fMRI$\rightarrow$audio path chains three generative models (SDXL $\rightarrow$ GIT $\rightarrow$ AudioLDM2). Training a direct fMRI$\rightarrow$audio decoder on paired data from CNeuroMod's movie-watching sessions would enable first-order neural-to-audio translation.

8 Conclusion

We have presented the first cross-modal neural translation system that uses synthetic fMRI as a universal interface between input and output modalities. By connecting MARY-Nano (a predictive brain model) to MindEye2 and AudioLDM2 (neural decoders), we create six directional translation pipelines that operate without any real fMRI data at inference time.

Our 2×2 factorial ablation over adapter architecture and training scale reveals a clear finding: *data scaling is the dominant lever for synthetic fMRI quality*. Ridge$\times$2K achieves CLIP $= 0.737$ (98.4% of the GT fMRI ceiling), SSIM $= 0.223$ (95.7%), and PixCorr $= 0.173$ (49.0%), with individual images reaching PixCorr > 0.6. Increasing training pairs from 500 to 2,000 yields a 24.3% CLIP improvement, while switching from Ridge to MLP yields only a 2.8% change. The scaling curve shows no saturation, suggesting that training on the full NSD dataset ($\sim$9,000 images) could close the CLIP gap entirely.

The remaining PixCorr gap (49% of GT) reflects a deeper challenge: MARY-Nano's average-subject model is anti-correlated in visual cortex ($r = -0.013$), where individual retinotopic organization varies substantially. Out-of-the-box average models will not close this gap. The path forward requires either multi-subject foundation models trained on hundreds of individuals (learning the distribution of cortical variation), anatomy-conditioned prediction (using structural MRI as a "brain fingerprint"),

or lightweight calibration via EEG. Crucially, for semantic applications (cross-modal translation, neural-conditioned generation), the 98.4% CLIP recovery achieved here may already be sufficient for production deployment without any per-user calibration.

More broadly, our results demonstrate that the brain's representational space can serve as a computational interface: an intermediate representation that different modalities converge to and diverge from, enabling translation constrained by neural plausibility rather than unconstrained generation. The complete system runs on a single A100 GPU at $\sim\$0.50$ per stimulus for encoding and decoding, making cross-modal neural translation accessible without neuroimaging infrastructure.

References

[1] 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, J. Benjamin Hutchinson, Thomas Naselaris, and Kendrick Kay. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 25:116–126, 2022.

[2] Julie A. Boyle, Basile Pinsard, André Cyr, François Paugam, Julien Cohen-Adad, and Pierre Bellec. The Courtois NeuroMod project. *Annual Conference on Cognitive Computational Neuroscience*, 2020.

[3] Min Jin Chong and David Forsyth. Effectively unbiased FID and Inception Score and where to find them. In *CVPR*, 2020.

[4] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. CLAP: Learning audio concepts from natural language supervision. In *ICASSP*, 2023.

[5] Kendrick N. Kay, Thomas Naselaris, Ryan J. Prenger, and Jack L. Gallant. Identifying natural images from human brain activity. *Nature*, 452:352–355, 2008.

[6] 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 *ECCV*, 2014.

[7] Haohe Liu, Qiao Tian, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D. Plumbley. AudioLDM 2: Learning holistic audio generation with self-supervised pretraining. *arXiv:2308.05734*, 2023.

[8] Thomas Naselaris, Ryan J. Prenger, Kendrick N. Kay, Michael Oliver, and Jack L. Gallant. Bayesian reconstruction of natural images from human brain activity. *Neuron*, 63:902–915, 2009.

[9] Furkan Özcelik and Rufin VanRullen. Natural scene reconstruction from fMRI signals using generative latent diffusion. *Scientific Reports*, 13:15666, 2023.

[10] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. *arXiv:2307.01952*, 2023.

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[12] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

[13] Paul S. Scotti, Mihir Tripathy, Cesar K. Torrico, Reese Kneeland, Tong Chen, Ashutosh Narang, Charan Santhirasegaran, Jonathan Xu, Thomas Naselaris, Kenneth A. Norman, and Tanishq M. Abraham. MindEye2: Shared-subject models enable fMRI-to-image with 1 hour of data. In *ICML*, 2024.

[14] Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *CVPR*, 2023.

[15] Ilya O. Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, and Alexei A. Dosovitskiy. MLP-Mixer: An all-MLP architecture for vision. In *NeurIPS*, 2021.

[16] 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*, 13:600–612, 2004.

[17] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. GIT: A generative image-to-text transformer for vision and language. *TMLR*, 2022.