Scaling Parameter-Efficient Fine-Tuning for Brain Encoding

BrainLoRA: a paper discussion

Zhuofan Josh Ying, Guanglong Sun, Nolan Ho, Kieran Liu, Rui Yang,
Andrew Chen, Nikolaus Kriegeskorte, Pony Ma, and Hossein Adeli

Takeaway

Small, stimulus-aligned adapters improve whole-brain fMRI prediction on large multimodal backbones; a causal single pass and batched adapters make the approach computationally practical.

arXiv:2610.26802v1 (19 Oct 2026) Publication not found as of 28 Oct 2026 Official code

The task: predict responses to natural movies

Brain encoding

$$\hat{Y} = g_\phi(f_\theta(X))$$

X: synchronized video, audio, transcript.
f_θ: pretrained multimodal backbone.
g_ϕ: temporal alignment and parcel readout.

Paper, Secs. 3–4, Eq. (1), PDF pp. 4–5.

- Target: fMRI responses in 1,000 Schaefer parcels.
- Four subjects, naturalistic audiovisual stimuli.
- Metric: mean parcel-wise temporal Pearson r across subjects; higher is better.

Neural data should shape the representation

Dominant frozen-backbone pipeline

Pretrained features → train a readout on neural data.

- Retains a strong pretrained prior.
- Cannot specialize the backbone to brain prediction.
- MIRAGE uses a frozen Qwen3-Omni and a 1.1B-parameter readout.

Paper, Secs. 1–3, PDF pp. 1–4; Fig. 1; Table 1.

BrainLoRA

Pretrained features → small backbone update → lightweight readout.

- Neural supervision reaches the representation.
- Most pretrained weights remain fixed.
- Main single model trains 61.4M parameters.

PEFT constrains the trainable parameter space

$$\Theta_{\text{frozen}} = \{\theta_0\} \subset \Theta_{\text{PEFT}} = \{\theta_0 + \Delta(\psi) : \psi \in \mathbb{R}^k\} \subset \Theta_{\text{full}} = \mathbb{R}^d, \quad k \ll d,\ \Delta(0) = 0.$$

Scale up

Use a stronger pretrained backbone θ_0.

Scale down

Learn a small adapter ψ from limited neural data.

Scale out

Train many specialized adapters on shared weights.

Paper, Sec. 3, Eq. (2), PDF p. 4; Fig. 1, p. 2.

Validation, ID, and OOD test different questions

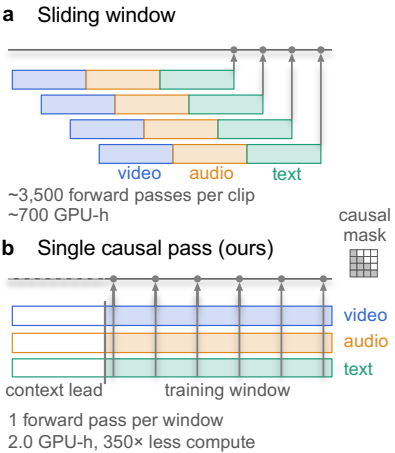
Split	Stimulus	Role
Train	Friends seasons 1–5 and movie10	Fit adapters and readouts
Validation	Friends season 6	Select designs and ensemble weights
ID test	Held-out Friends season 7	Challenge server score
OOD test	Held-out movies	Challenge server score

Controlled single-model comparison

MIRAGE and BrainLoRA use the same Qwen3-Omni backbone. Published baseline scores in Table 1 are cited from their authors; BrainLoRA model choices use season 6 only.

Paper, Sec. 4, PDF p. 4; Table 1, p. 6; App. D, pp. 18–19.

A causal pass replaces repeated sliding windows



- Baseline: re-encode overlapping context at every prediction time.
- BrainLoRA: one pass per 149 s window produces all 100 TR features.
- Add a 60 s context lead to restore the start of each window.
- Decoder causality helps; audio and vision front ends are locally bidirectional.

Paper, Sec. 4, Fig. 2, PDF p. 5; Apps. B and F, pp. 16–17, 21.

The adapter updates early attention layers

Qwen3-Omni configuration

- Read decoder layer 32 of 48; truncate there.
- Rank-16 LoRA on $q/k/v/o$ projections through layer 32.
- 8.91M backbone-side trainable parameters.
- Audio and vision towers remain frozen.

Paper, Sec. 4, PDF p. 5; App. S, p. 36.

$$W' = W_0 + \frac{\alpha}{r} BA$$

$$r = 16, \quad \alpha = 32.$$

Standard LoRA notation illustrates the paper's general $\Delta(\psi)$: train a low-rank update while keeping W_0 fixed.

The readout aligns stimulus features with fMRI

Linear readout

1. Pool each modality on a 2 Hz grid.
2. Project 2048 $\rightarrow$ 512 dimensions.
3. Average eight temporal lag blocks.
4. Map $8 \times 3 \times 512 = 12{,}288$ features to 1,000 parcels per subject.

52.3M trainable parameters.

Cross-attention readout

Per-TR queries attend to about 3.2k audio, video, and text tokens, replacing fixed lag averaging. Adds about 0.2M parameters.

The extra gain mainly comes from richer tokens: cross-attention over the same pooled features barely improves r.

Paper, Secs. 4–5, PDF pp. 5–7; Apps. K and S, pp. 27, 37.

Closed-form re-fit improves the final readout

1. Initialize from ridge regression on frozen features.
2. Jointly train LoRA and readout for one epoch: 1,429 windows, one window and four subjects per step.
3. Freeze adapted features; accumulate training moments and re-solve each parcel's final linear map by ridge regression.

Measured contribution

LoRA + linear improves from $r = 0.3202$ before re-fit to $r = 0.3263$ after re-fit: $+0.0061$ on Friends season 6.

Ridge penalties use a holdout within training windows; season 6 is used for design selection and scoring.

Paper, Secs. 4–5, Table 2, PDF pp. 5–7; App. D, pp. 18–19.

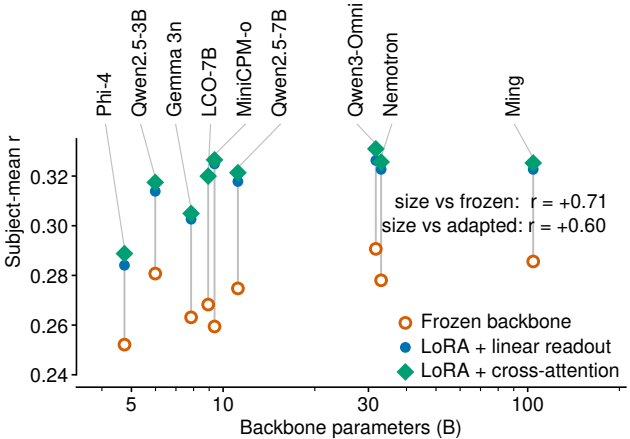
Extraction falls 352-fold; total cost falls 83-fold

One A800-80GB; 331 clips	Sliding + large head	Causal + LoRA
Feature extraction (GPU-h)	701	1.99
Training (GPU-h)	1.88	4.84
Closed-form re-fit (GPU-h)	—	1.67
End-to-end (GPU-h)	**702**	**8.5**
Trainable parameters	1,123M	61.2M

Extraction: $701/1.99 \approx 352\times$; full pipeline: $702/8.5 \approx 83\times$. These measured protocols use different context and training procedures.

Paper, App. C, Table 4, PDF p. 18.

LoRA improves all nine tested backbones



Paper, Sec. 5.1, Fig. 3, PDF p. 5; App. I, Table 12, p. 25.

- Backbones span 4.7B–104.3B total parameters.
- Every frozen-to-LoRA gain is $+0.032$ to $+0.066$ in validation r.
- Model size alone weakly orders accuracy: the 104B model is not best.

A 61M-parameter model beats the published baseline

Qwen3-Omni configuration	Trainable	Val	ID	OOD
MIRAGE single, cited	1.1B	0.319	0.310	0.217
Frozen + linear	52.3M	0.291	0.287	0.187
LoRA + linear	61.2M	0.326	0.322	0.222
Frozen + cross-attn	52.5M	0.312	0.307	0.212
LoRA + cross-attn	**61.4M**	**0.331**	**0.327**	**0.228**
Full tune + cross-attn	20.3B	0.330	0.326	0.225

ID: Friends season 7; OOD: held-out movies. The bold single model exceeds the published 15-model MIRAGE ensemble (ID 0.323, OOD 0.227) on ID and is similar on OOD.

Paper, Sec. 5.2, Table 1, PDF p. 6. Higher r is better.

The adapter supplies the largest component gain

Qwen3-Omni, Friends s06	Mean Pearson r
Frozen backbone + linear	0.2907
LoRA + linear, before re-fit	0.3202
LoRA + linear, after re-fit	**0.3263**
Rank-32 LoRA + linear	0.3268
LoRA + cross-attention	0.3310

- LoRA over matched frozen control: +0.0356.
- Re-fit: +0.0061; rank 16 to 32: only +0.0005.
- Differences below about 0.00077 are not resolved by reported seed spread.

Paper, Sec. 5.3, Table 2, PDF p. 7; App. D, p. 19.

The benefit requires aligned stimulus–brain pairs

Backbone	Adapter training target	Validation r
Frozen	Aligned fMRI	0.2907
LoRA	Aligned fMRI	**0.3263**
Frozen	Shifted fMRI	0.2617
LoRA	Shifted fMRI	0.1874

Shifted targets preserve subject-specific fMRI statistics but break stimulus pairing. All final readouts are re-fit on aligned training data. Shifted LoRA loses on 996/1,000 parcels against its matched shifted frozen control.

Paper, App. L, Table 13, PDF p. 28.

Adapted features transfer within the film domain

Leave one subject out

Train LoRA on three people; fit only the held-out person's readout on their training data. Gain over frozen: $+0.0301$ to $+0.0355$, about 97% of the four-subject gain.

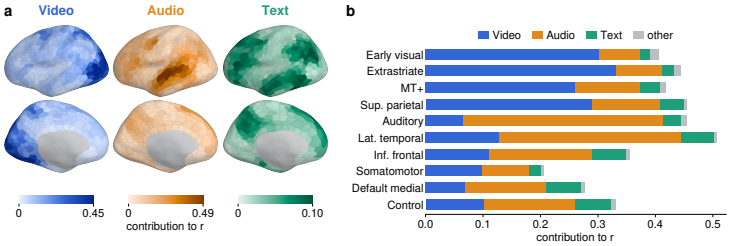
NNDb: new site and scanner

Transfer the Algonauts adapter *and readout* without retraining. Across 20 viewers of a film, r rises from 0.0991 to 0.1223, or $+11.1$ percentage points of NNDb's corrected noise ceiling.

Both evaluations use naturalistic movies; NNDb also changes scanner, participants, preprocessing, and repetition time.

Paper, Apps. M–N, Tables 14–15, PDF pp. 29–30.

Attribution recovers sensory topography

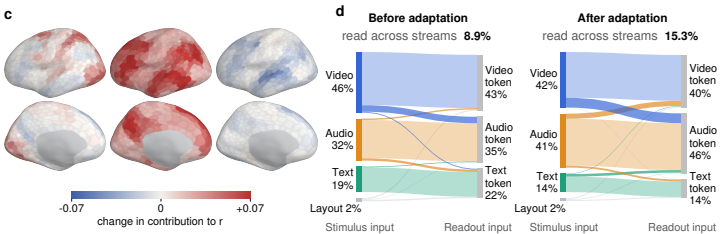


Enumerate all 2^3 video/audio/text coalitions. Replace missing content by mean embeddings while keeping token timing.

Early visual: 72.5% video. Auditory: 72.9% audio. Lateral temporal: 60.3% audio.

Paper, Sec. 5.4, Fig. 4a–b, PDF p. 8; App. Q, pp. 33–34. Percentages are prediction shares; maps show contribution to r.

Adaptation increases audio use and cross-stream mixing



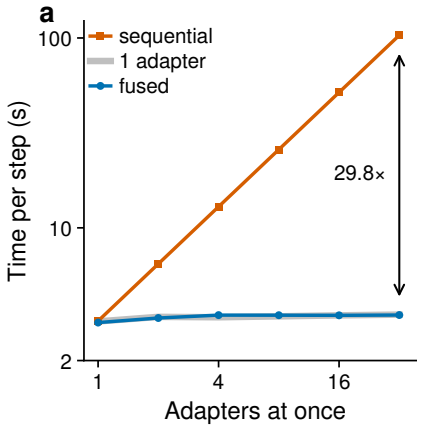
Mean audio contribution:
+39% relative.
Video: +5%; text: −17%.
Cross-stream readout share:

$$8.9\% \longrightarrow 15.3\%$$

+6.4 percentage points.

Paper, Sec. 5.4, Fig. 4c–d, PDF p. 8; App. Q, pp. 33–35. Attribution uses the paper's content-removal baseline.

A fused kernel makes many adapters affordable

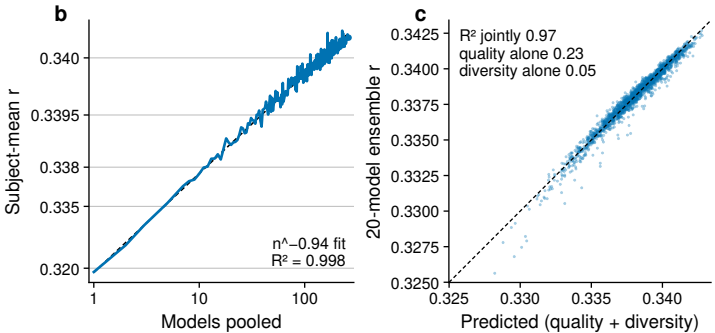


- Route samples to LoRA matrices in one batched gather–matmul.
- At 32 adapters, a fused step costs 1.08× one adapter step.
- Relative to 32 sequential steps: 29.8× faster.

Measured at sequence length 128 on one A800. The gain comes from GPU utilization, not fewer FLOPs; full training gains are smaller.

Paper, Sec. 5.5, Fig. 5a, PDF p. 9; App. C, pp. 18–20.

Ensemble size helps until correlated errors dominate



Paper, Sec. 5.5, Fig. 5b–c, PDF p. 9. The curve approaches a fitted ceiling.

Random subsets fit

$$m(n) = m_\infty - An^{-b}$$

$b \simeq 0.94$. At $n = 20$, member quality and residual diversity jointly explain 97% of score variance in this pool.

The 261-model pool reaches the reported top rank

Ensemble	Val s06	ID s07	OOD movies
MIRAGE, 15 models (cited)	0.335	0.323	0.227
BrainLoRA, selected 15	0.347	0.341	0.247
BrainLoRA, 261	**0.348**	**0.343**	**0.248**

- A parcel-wise minimum-variance combiner is fitted out of fold by episode.
- Different backbones and readouts supply more diversity than seeds.
- First place refers to the public post-challenge leaderboard at paper-writing time.

Paper, Sec. 5.5, Table 1, PDF pp. 6, 9–10; App. D, pp. 18–19.

The model pool still misses shared stimulus signal

Cross-subject residual analysis

The 261-model uniform mean leaves 32.5% [30.0, 35.3] of cross-subject stimulus-locked covariance unexplained.

- A single LoRA + linear model leaves 36.2%.
- Averaging the pool recovers about 3.7 percentage points.
- The residual is slow and distributed; nine low-level stimulus statistics explain it at only $r = 0.027$.

Paper, App. R, Table 18, PDF pp. 35–36. Denominator: inter-subject covariance, not all brain-response variance.

Take-home messages

1. Small, aligned backbone updates improve movie-to-fMRI prediction across all nine tested multimodal backbones.
2. A causal pass cuts extraction cost about 352×; re-fit and richer readout tokens add measurable accuracy.
3. Attribution suggests stronger audio use and cross-stream routing; ensembles improve the score, with correlated errors limiting gains.

Evidence boundary

Main training uses four subjects and one naturalistic movie benchmark. Transfer changes people and scanner but remains in the film domain; a cross-task brain foundation model remains a proposal.

Paper, Sec. 6, PDF p. 10; Apps. M–N and R, pp. 29–30, 35–36.

Paper and source links

- Zhuofan Josh Ying et al., *Scaling Parameter-Efficient Fine-Tuning for Brain Encoding*, arXiv:2610.26802v1, 2026.
- Official BrainLoRA code (linked from the paper and first-author homepage).
- First-author research page.

Publication status checked 28 Oct 2026: no formal venue record found; the author lists the paper as arXiv.

Backup: 60 seconds of context recovers the window start

Context lead	LoRA + linear before re-fit	After re-fit
0 s	0.3148	0.3206
30 s	0.3187	0.3249
60 s	0.3202	**0.3263**
90 s	0.3196	0.3260
120 s	0.3203	**0.3263**

Performance saturates near 60 s while step time rises. These rows use window-local single-pass features.

Paper, App. B, Table 3, PDF p. 16.

Backup: the gain is not unique to LoRA

Adapter family	Adapter parameters	Val s06 r
Frozen	0	0.2907
IA^3 attention + FFN	0.057M	0.3127
IA^3 + MoE experts	3.18M	0.3177
Bottleneck, FFN only	4.26M	0.3164
Bottleneck, attention + FFN	8.52M	0.3190
LoRA rank 16	8.91M	**0.3263**

The shared 52.3M readout is excluded from adapter parameters. Each family uses its screened learning rate.

Paper, App. H, Table 11, PDF p. 24.