CALIBRA-fMRI: TEXT FOR SIGNALS WITHOUT CAPTIONS, READ BY A FROZEN VIDEO–LANGUAGE MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal language models answer many kinds of question in one output space, and extending them to scientific signals requires text that rarely exists and must be generated. Generated text can often be answered without the non-text input, a shortcut usually found only after a model has been trained on it. Task fMRI comes with a controlled experiment and rich metadata, from which answers can be computed and the shortcut controlled at construction. We build CALIBRA-fMRI, whose items span the experimental design, the signal, behavior and the participant in a design space defined by a rule-based item generator. Each item carries a signal-free accuracy, fixed before training by a rule that answers it without the recording, and a model trained on the text alone stays at this level (.252 against .259). Without paired fMRI–text data at pretraining scale, we first verify that the image encoder of a frozen video–language model, never trained on fMRI, is competitive with fMRI foundation models under a linear probe. Building on it, we develop an fMRI aggregator that pools each four-dimensional volume through a cortical atlas, compressing a 15-minute recording from millions of patch tokens into thousands. Coupled with the frozen language model, it forms RIVET, which takes fMRI inputs of any length while training only 21.2M parameters. Two frozen fMRI foundation models with trained heads each answer one side, relations within a segment or participant identity, whereas RIVET answers both when trained on each and, trained on all levels at once, exceeds the signal-free accuracy on 9 of 14 items. It orders states in time without per-step labels and, from three segments on, learns to compare segments given separately, a comparison it carries to unseen paradigms. Trained on sex instead, it retains much less of the moment-to-moment state, consistent with how the two foundation models split as released, so what a model can answer depends on what its input keeps and what its items ask. The construction needs only a record that fixes the answers, offering multimodal language models a way to build corpora that require their non-text input by design.

1 INTRODUCTION

A foundation model for scientific data is asked many different questions about the same measurement, whether they name a category, compare quantities, concern their order in time, or relate several of these at once. Language models, rich in scientific knowledge, answer all of them in one output space that already draws in images, video and time series (Taylor et al., 2022; Singhal et al., 2023; Alayrac et al., 2022; Jin et al., 2024), so scientific foundation models are increasingly joined to them, from molecules and DNA sequences to satellite images (Liu et al., 2023b; de Almeida et al., 2025; Kuckreja et al., 2024). Such a join requires text paired with the signal, and a recording, unlike a photograph, admits no caption written by looking at it, so the text has to be generated from what is known about the recording. Whether a generated corpus can be learned is then established only by training a model on it, while a model for the signal cannot be trained without such a corpus. For fMRI, where both are scarce, the construction of a corpus and that of a model depend on each other.

Generated text must meet two requirements that collected text does not. It must not be answerable without the signal, since a question and its options may carry enough information on their own, and a model trained on such items learns the prior rather than the recording (Geirhos et al., 2020). Such

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.

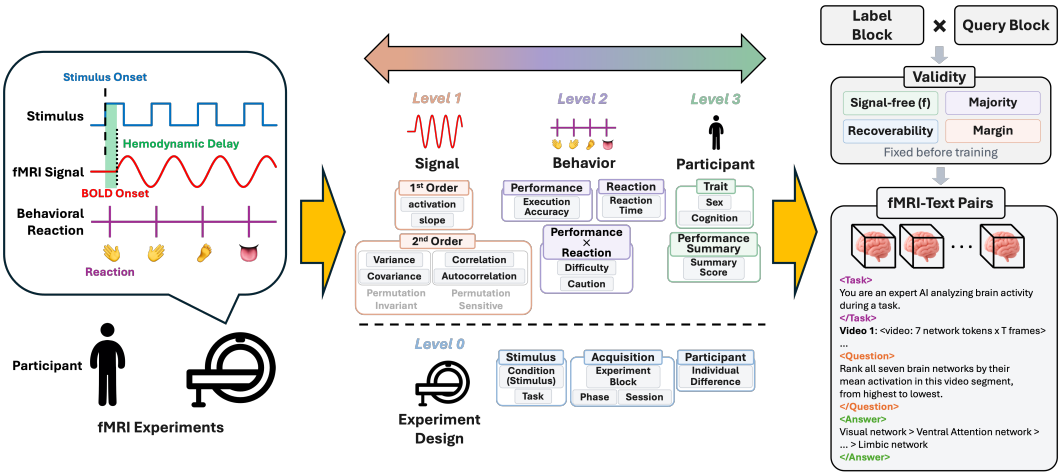

Figure 1: How an fMRI–text pair is made. **Left:** a task fMRI experiment supplies the stimulus record, the signal and the behavioral record on one clock. **Center:** quantities are defined at four levels, from the design of the experiment to measures over the participant. **Right:** a label block and a query block compose into an item, released with its signal-free accuracy f and three further controls, all fixed before training.

shortcuts are usually found only after a benchmark is finished, by withholding its input, as when frontier systems remain at the top of a medical benchmark without the image (Asadi et al., 2026). It must also cover the range of questions that can be asked of the signal. For fMRI, this range runs from the design of an experiment through the signal and the concurrent behavior to the participant. The few fMRI–language models caption a recording through descriptors and predict individual traits well from them (Wei et al., 2026b), but their captions are limited to conventionally derived quantities that summarize a whole recording, such as functional connectivity, gradients and graph metrics, and describe neither the temporal course of the signal nor the behavior it accompanies. Beyond this limited coverage, no fMRI corpus has, to our knowledge, been built to control at construction how far its items can be answered without the signal.

Evaluating a corpus design in turn requires a model that is inexpensive to train for each design and answers every item through a single output space. A model built for fMRI is first pretrained on the signal (Ortega Caro et al., 2024; Wang et al., 2026c) and then requires an additional component for each answer form, from a classification head for a choice to a decoder for a sequence.

For the construction of text, task fMRI provides two sources that most signals lack. The first is recorded alongside the signal, namely the experimental design, the behavior of the participant and participant-level measures, and the second is computed from the signal relative to that design, such as which networks are more active in a given condition. Combining the two, each item poses a question about the recording at one of these four levels (Figure 1). Because every answer is computed rather than annotated, the extent to which an item can be answered without the recording is measured before training and released as its signal-free accuracy f, against which results are read as a normalized gain (§3.1), a procedure that does not depend on the signal being fMRI. A rule-based item generator defines the design space of such items, from which CALIBRA-fMRI is built and which is released in full.

For the model, we build on video–language models, in which an image encoder embeds each frame and a language model integrates the frame embeddings over time (Maaz et al., 2024; Wang et al., 2024). Such encoders transfer well beyond photographs (Zhang et al., 2023; Chen et al., 2025; Daswani et al., 2024), and before adopting one we verified that its vision tower, never trained on four-dimensional fMRI, is competitive under a linear probe with two fMRI foundation models and exceeds both on several targets (§4.1). The four-dimensional nature of fMRI, however, imposes strong constraints on spatio-temporal modeling. Two-dimensional brain parcellation time series reduce each volume to a vector and can span long durations at the cost of spatial detail, whereas volumetric inputs preserve spatial detail but are too large to model over long spans, and foundation

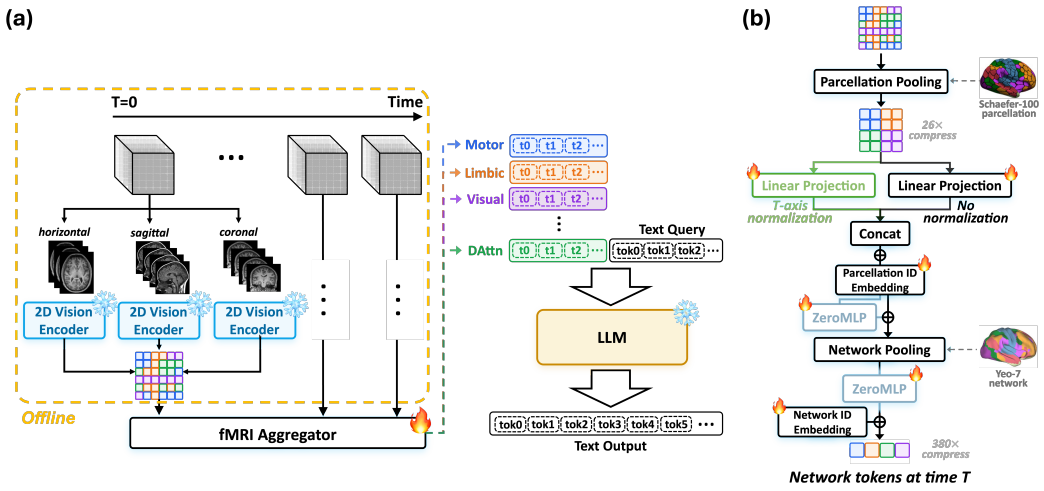

Figure 2: The interface, with flames marking what is trained. **(a)** Each volume is sliced along the three principal axes and read by a frozen vision encoder, whose output is extracted once offline, and the fMRI aggregator pools it into one token stream per functional network, running in time alongside the text query. **(b)** The fMRI aggregator pools patch grids into 100 parcels, projects a raw and a time-standardized channel, and pools them into 7 networks, 370-fold.

models of either kind are typically pretrained on fixed windows of 434 time points (Ortega Caro et al., 2024) or 40 volumes (Wang et al., 2026c). We retain the volumes but encode each one independently and offline with the frozen vision tower, pooling its patch tokens through a fixed cortical atlas, the standard unit of analysis in neuroimaging, into seven network tokens, so that a 15-minute run of more than 900 volumes requires 6,398 tokens instead of some 2.4M and can be trained on a single 40 GB A100 GPU. The autoregressive language model thus models time over sequences of any length, from a few frames to a whole run, which makes the training length a choice of corpus design. With the vision–language model frozen, only the fMRI aggregator that performs this pooling is trained, a dedicated connector of 21.2M parameters, and we refer to the resulting interface as RIVET, for *region-indexed video encoding of task fMRI* (Figure 2).

The corpus and RIVET are built together: the corpus supplies the fMRI–text pairs a video–language model needs, and RIVET makes each corpus design cheap to train and answers all of it in one output space, so together they train a language model on task fMRI, which neither makes possible alone. The items test four capabilities that a model reading fMRI in language should have, namely covering the four levels in one model, reading relations between networks in any answer form, locating states in time, which requires following the hemodynamic response, and integrating segments that arrive separately. As the vision tower never relates one frame to another, the language model carries the last two, which also test whether its video modeling carries over to fMRI. Each capability is evaluated on items built to isolate it and chosen before training, and what a model does not learn is reported as a property of the design. The corpus is thereby a benchmark in its own right, separating three classes of model, zero-shot vision–language models that read only salient contrasts, two fMRI foundation models with trained heads that each address the side their released input keeps, relations and order within a segment or participant identity, and RIVET, which addresses both when trained for each. Trained on all levels at once, RIVET stays above the signal-free accuracy on activation, slope, the order of states, behavioral accuracy and cognition, and reads participant identity only when trained on that item alone, on the social paradigm (§5.4).

Contributions.

- **CALIBRA-fMRI, a benchmark whose items measure their own shortcuts.** Its items join the fMRI signal to the experimental design, the behavior and the participant, each carries its signal-free accuracy, fixed before training, so text-only shortcuts are measured at construction, and all come from a design space the released item generator defines in full.

- **RIVET, an extension of video–language models beyond video, to fMRI.** It reads each volume with a frozen natural-image tower, competitive with fMRI foundation models under a linear probe, and pools it through a cortical atlas into seven tokens, so it models inputs of any length, up to whole runs, in thousands rather than millions of tokens while training only 21.2M parameters. It locates states in time without per-step labels and learns to compare segments given separately, within one output space.
- **What the benchmark reveals: input normalization shapes what a model can answer.** Built and tested with a frozen video–language model alone, the benchmark shows that whether a model keeps what varies within a run or what is invariant within it follows how its input is normalized, as removing the raw channel of RIVET costs sex but not the condition, consistent with how two fMRI foundation models divide as released, and that what the items ask narrows it again.

2 RELATED WORK

Vision–language models as interfaces. Most vision–language models join a pretrained vision encoder to a language model through a small connector trained on paired image and text (Radford et al., 2021; Alayrac et al., 2022; Li et al., 2023; Liu et al., 2023a; Qwen Team, 2025; Wang et al., 2025), and video models leave the temporal modeling to the language model (Maaz et al., 2024; Li et al., 2024; Wang et al., 2024), as a frozen language model does for a reprogrammed time series (Jin et al., 2024). Their image encoders serve a dozen modalities while frozen (Zhang et al., 2023), read signals rendered as images (Chen et al., 2025; Daswani et al., 2024) and tokenize fMRI volumes (Kim et al., 2026), and one reads a recording well enough to caption the image a participant viewed, a text that describes the stimulus rather than the activity (Han et al., 2024). None, to our knowledge, joins fMRI and language as video–language models join video, the image encoder reading each volume and the language model its temporal structure.

Text for signals that have none. Text for scientific measurements must be generated, from a simulator that fixes the ground truth (Johnson et al., 2017), from metadata and prior measurements (Abbasi et al., 2026) or from quantities fitted to templates (Wei et al., 2026b). Generated text must describe its signal and also require it, since an item answerable without its input says nothing about it. The second is measured by withholding the input (Poliak et al., 2018; Gururangan et al., 2018; Asadi et al., 2026) and controlled either by filtering items against a model (Le Bras et al., 2020; Chen et al., 2024; Zhang et al., 2024) or by shaping the answer distribution without one (Johnson et al., 2017). Closest to us, Abbasi et al. (2026) apply three such controls to brain MRI, a generation-time distributional filter, a text-only baseline and a per-template majority as chance, whose reporting conventions we adopt. Their labels come with the data, whereas ours are computed from the signal, so the computation itself is gated before training, by recoverability from the frames given and a margin over the nearest distractor, and rejected cells are released with their f (Appendix A.3, A.4). None of these, to our knowledge, releases with each item the accuracy its text alone affords, fixed before training.

fMRI and language. fMRI has no captions describing the activity pattern itself either, although its stimuli can be captioned and its metadata written out as text. Some describe the stimulus decoded from the recording (Beliy et al., 2026; Qiu et al., 2025; Han et al., 2024). Others describe the participant, through the definition of a diagnostic class (Wei et al., 2026a), clinical records (Gu et al., 2026) or, in concurrent work, traits over static connectivity (Hu et al., 2026). One captions whole recordings through conventionally derived quantities that bring out individual differences, such as functional connectivity and graph metrics (Wei et al., 2026b). None describes the activity pattern of a segment relative to the design that produced it. Nor, to our knowledge, does any corpus join such a description to the stimulus, behavior and participant.

Foundation models for fMRI. Without paired text, fMRI foundation models pretrain on the signal at scale and adapt to each downstream task with a task-specific head (Kim et al., 2023; Ortega Caro et al., 2024; Dong et al., 2024; Yang et al., 2024; Dong et al., 2025; Wang et al., 2026a;c; Lane et al., 2026; Wei et al., 2026c). Task recordings, whose design grounds the signal as rest cannot, are a minority of that pretraining (Wang et al., 2026a) and can be down-weighted by selection criteria tuned

for rest (Wang et al., 2026b), although one model now pretrains across resting, task and naturalistic recordings (Xia et al., 2026). Region-level tokens are established (Gijsen et al., 2026; Dong et al., 2025), and Gijsen et al. (2026) compress parcels onto fixed functional networks as we do.

3 CONSTRUCTION AND INSTANTIATION

3.1 ITEMS THAT MEASURE THEIR OWN SHORTCUTS

Each item is released with its *signal-free accuracy* f. It is the test accuracy of a rule, fitted on the training split, that maps everything the item states apart from the recording to its most frequent answer, floored at chance, and it is computed before training and independently of any evaluated model. It stands in for text-only evaluation with the recording withheld, which is too costly to run on every candidate item, and answers are balanced within each question context so that f is driven toward chance wherever the labels allow. Trained with the recording withheld, a model exceeds f significantly in no seed, on the eight answer forms of §5.2 (.252 against .259, Table 3) or on the fourteen items of Table 2. Results are read through the *normalized gain*

$$g = \frac{a - f}{1 - f}, \tag{1}$$

where a is the score on a set of items in their metric, so that g is 0 at the signal-free accuracy and 1 at a perfect score. Each item also carries further controls, such as the majority answer of its type and the margin of its answer (Appendix A).

3.2 INSTANTIATION ON TASK FMRI

Each recording is a sequence of volumes acquired every 0.72 s under a schedule of condition blocks read from the recorded event files, and labels are stated over the 7 networks (Yeo et al., 2011) formed by the 100 parcels of a cortical parcellation (Schaefer et al., 2018) (Figure 1). Recordings come from task paradigms, such as motor or working memory, and an item refers to one or more segments of consecutive volumes. Labels of the signal, such as a network's mean activation, its linear slope over a segment and its coupling, the correlation of its time course with another network's, are computed from the recording itself, by averaging within each network and standardizing along time within a run, so no choice of input representation can alter an answer. The item generator combines a label block, which fixes what is measured, with a query block, which fixes how it is asked, and labels fall into four levels by the source of the answer, the experimental design (Level 0), statistics of the signal (Level 1), the signal joined to task behavior (Level 2) and measures of the participant (Level 3). A query is combined with a label only where its operation has at least two elements to act on, so the eleven labels of the joint mixture admit 309 of 1,056 combinations, or cells, as the design space, of which the mixture builds 16 in 159K items and the released corpus holds 2.2M (Appendix A.1). Built cells are chosen without reference to any model's accuracy, each isolating a capability of §3.3 or asking a question with an established reading in fMRI research (Table 5), and cells failing the checks of Appendix A.3 are reported, not removed. Splits are by participant, and unparseable generations count as wrong.

3.3 FOUR CAPABILITIES

Table 1 lists what each of the four capabilities requires of a model with trained heads. *Level coverage* asks one model to answer items from all four levels, as one interface to a recording must. *Relational reading* asks for relations between networks within one segment, in closed forms that offer the answers, open-vocabulary forms that name networks or order a few of them, and an open-ended full ranking of the seven networks. *Temporal localization* asks which states a segment contains and in what order, so the number of states is part of the answer and following the hemodynamic response is required. *Cross-segment integration* asks which of several segments, each given separately, comes from a different participant, and since the input already separates participants (§5.4), it tests whether a model learns to compare them rather than whether the information is present.

4 METHOD

4.1 ARCHITECTURE

We use Qwen3-VL-2B-Instruct (Qwen Team, 2025) unmodified and train only an fMRI aggregator of 21.2M parameters, which takes the place of the modality projector for fMRI input only, so images still pass through the model's own projector and the model is otherwise unchanged (Figure 2). For frame t of a run, the frozen vision tower reads the volume V_t as three stacked cross-sections, a fixed matrix W pools its 2,592 patch tokens into the 100 parcels of a cortical atlas (Schaefer et al., 2018), and a second copy is standardized along time within the run,

$$X_t = W^\top \mathrm{ViT}(V_t) \in \mathbb{R}^{100\times d}, \qquad \tilde{X}_t = (X_t - \mu) \oslash \sigma, \tag{2}$$

where μ and σ are the mean and standard deviation of X_t over the run and $d = 2{,}048$. Because the tower and W are fixed, X_t is extracted once offline for every frame and cached, so training reads the cache and passes only through the fMRI aggregator and the frozen language model. For each frame of a segment, the fMRI aggregator produces seven network tokens,

$$\begin{aligned} P_t &= \left[\, X_t A \,\|\, \tilde{X}_t \tilde{A} \,\right] + E_{\mathrm{parcel}}, \\ H_t &= P_t + \mathrm{MLP}_1\big(\mathrm{LN}(P_t)\big), \\ O_t &= \mathrm{MLP}_2\big(\mathrm{LN}(M^\top H_t)\big) + E_{\mathrm{net}} \in \mathbb{R}^{7\times d}, \end{aligned} \tag{3}$$

where A and $\tilde{A}$ project each channel to half width, M is a fixed size-weighted map from parcels to networks (Yeo et al., 2011), E_{parcel} and E_{net} are learned embeddings, and each MLP ends in a zero-initialized layer (Table 7).

We keep the natural-image vision tower because under the probe of the introduction it matches or exceeds both foundation models on every resting-state and task target, trailing only on naturalistic viewing (Appendix C). The 370-fold reduction to seven tokens lets RIVET take whole runs of more than 900 frames on a single card (Figure 3b, Appendix F). The vision tower encodes each frame on its own, so nothing in X_t separates what varies over a run from what is invariant within it. We tested, before any training, whether labels that vary within a run and labels invariant within it call for different inputs, and they do: time-varying labels such as the condition are read better from $\tilde{X}_t$, which removes the invariant part, and time-invariant labels such as sex from X_t, which keeps it (Appendix D). The corpus asks both kinds of question, so the input keeps both channels, which stays within .008 of the better one on either. Ablations attribute nearly all of the result to the fMRI aggregator, whose output carries no residual (one costs .210 and .403 at equal parameter count), and the same aggregator works on InternVL3.5-2B (Wang et al., 2025) (Appendix E).

4.2 TRAINING

Every answer is generated as text, and every answer form is trained under one objective, the cross-entropy of the answer tokens given the visual tokens and the question q,

$$\mathcal{L} = -\sum_{i\in\mathcal{A}} \log p\big(y_i \mid y_{<i},\, O_{1:T},\, q\big), \tag{4}$$

where $\mathcal{A}$ indexes the answer span and only the fMRI aggregator is updated. Each volume enters the sequence as one video frame of seven tokens through the model's own video path (Appendix B).

5 RESULTS

Every experiment below compares three configurations, which differ in what they reuse and in what each task requires to be built (Table 1). All models are trained and evaluated on participant-disjoint splits with checkpoints selected on the validation split, by loss and, for RIVET, also by generated answers, and results are the mean and sample standard deviation over three seeds. The zero-shot models read a heatmap of the recording in one greedy pass and are reported as the best of four models over their renderings, scored by candidate log-probability where marked, as a floor for what a frozen model reads from a rendering rather than as a competitor. The two foundation models,

Table 1: Configurations compared. The last four columns list what a configuration needs beyond its frozen components for each capability of §3.3. fMRI FM and VLM stand for fMRI foundation model and vision–language model, and rules turn predictions into answers.

				Task-specific downstream heads			
Configuration	Encoder	Reads across frames	Text generation	Level coverage §5.1	Relational reading §5.2	Temporal localization §5.3	Cross-segment integration §5.4
fMRI FM + head	fMRI FM	fMRI FM within its window, trained head across windows	✗	6 heads, 8 rules, decoder	1 head, 8 rules	1 GRU, 1 rule	1 head
VLM zero-shot	vision encoder	frozen language model	✓	✗	✗	✗	✗
RIVET	vision encoder + fMRI aggregator	frozen language model	✓	✗	✗	✗	✗

BrainLM (Ortega Caro et al., 2024) on parcel time series and Omni-fMRI (Wang et al., 2026c) on volumes, are frozen and run as released, with each parcel's mean over the run removed for BrainLM and fixed 40-frame windows for Omni-fMRI, and with task-specific components limited in total to the 21.2M trained parameters of our fMRI aggregator. Every result is read against f through the normalized gain of Equation 1, and the full protocol is in Appendix G.

5.1 LEVEL COVERAGE

Level coverage asks whether items from all four levels, in their different answer forms, can be trained together in one model without losing what each teaches alone. Trained on all items at once, RIVET stays above f on activation, slope, the order of states, behavioral accuracy and cognition, and keeps its single-level gains ($+.651$, $+.616$ and $+.171$ against $+.609$, $+.665$ and $+.148$, Table 2). Read through their released inputs, the foundation models each keep one side, the parcel-based model relations between networks ($+.687$ on the full ranking alone) and the volume-based model participant identity ($+.9995$), and with a shared trunk, one joint design among several, the parcel-based model keeps less of its reading of relations ($+.137$). RIVET does not learn participant identity jointly ($+.010$), which it learns alone on the social paradigm (§5.4), and the two coupling items stay at or below f for every model. Level 1 labels are computed from the recording, so a ridge readout recovers them exactly from network time series and with gains of .77 to .89 from standardized parcel time series (Appendix G), an upper reference of which RIVET recovers half to three quarters while answering every form through one output space. The following subsections take up the other three capabilities, each trained alone.

5.2 RELATIONAL READING: ONE MODEL, EIGHT ANSWER FORMS

Relational reading asks whether relations between networks in one segment can be stated in every answer form through one output space. One model trained on the motor paradigm answers all eight forms above their signal-free accuracy, from two-way judgments to a full ranking of the seven networks (Table 3). Ordering three networks transfers to triples never seen in training, at a gain of $+.255$ against $+.329$ for seen triples, so the model composes relations it was trained on into questions it was not. Zero-shot vision–language models stay below it on six of the seven forms they can be scored on and on the unseen triples. The volume-based model reaches a mean gain of $+.028$ against $+.332$ and falls below zero on unseen triples that depart from the typical order ($-.057$ against $+.199$), recovering only the typical order across participants. The parcel-based model, whose parcel time series lie close to the computation of the labels, reaches $+.403$, and $+.322$ on those triples, with a head and rules built for these forms (Table 15).

5.3 TEMPORAL LOCALIZATION AND THE HEMODYNAMIC RESPONSE

A model that reads fMRI in time must at least follow the hemodynamic response, through which a state enters the signal only after several seconds, so we ask for the two to five states of a segment in temporal order. With adjacent repeats merged, where the most frequent training sequence reaches .446, RIVET reaches .806 $\pm$ .023 without any per-step label, within .024 of a GRU head of 21.2M parameters on the parcel-based model trained with the condition of every 14-frame bin (.830 $\pm$ .002, Table 23). Scored exactly, RIVET answers .667 $\pm$ .035 of items ($+.582 \pm .041$ against f, $+.507 \pm .051$ against a frame-count rule), which that head, merging adjacent labels, gets right only when no state recurs in a row (.276), and zero-shot models at most .002. A per-step head falls below the constant answer on the volume-based model, whose 40-frame window exceeds most blocks, and

Table 2: Headline items on the working-memory paradigm as normalized gain g against the signal-free accuracy f, mean over three seeds, for one model trained on all items and, below, for one item per level trained alone. Text-only is RIVET trained with the recording withheld, which Appendix G details along with seed standard deviations, an item counts as above f when g exceeds .1 in every seed, and the highest gain above .1 in a row is in bold. VLM zero-shot entries are the best over models, renderings and scoring by free generation or, marked †, by candidate log-probability, ‡ marks an f above .9, and Level 3 is split by composite in Table 17.

				Text-only	VLM zero-shot	fMRI FM + head		RIVET
				Qwen3-VL-2B	best of 4	BrainLM	Omni-fMRI	Qwen3-VL-2B
Level	Item	n	f	*none*	*heatmap*	*parcel*	*volume*	*volume*
Trained jointly on all items, one model								
L0	Order of conditions in a window	1,624	.062–.063	−.007	.000†	+.171	−.054	**+.616**
	Segment from a different participant	3,232	.250–.258	−.002	+.021	+.146	**+.9995**	+.010
L1	Activation, pairwise	3,248	.500	−.006	+.241†	+.136	+.021	**+.640**
	Activation, highest excluding one	3,246	.167	.000	+.233†	+.051	+.003	**+.503**
	Activation, full ranking	3,248	.305–.316	−.069	−.069	+.137	−.019	**+.651**
	Slope, pairwise	3,248	.500	−.003	+.002	+.031	+.045	**+.441**
	Slope, highest excluding one	3,246	.167	.000	−.004†	+.006	+.010	**+.421**
	Slope, full ranking	3,248	.165–.186	−.129	−.325	+.024	−.022	**+.593**
	Coupling to a reference network‡	1,094	.936–.950	.000	−4.929†	−.025	−.047	−.316
	Coupling pattern of one participant	3,248	.125	−.007	−.143	+.019	+.020	+.003
L2	Difficulty, upper or lower third	129	.500–.530	−.042	+.054	+.061	−.024	**+.155**
	Reaction time, upper or lower third	133	.500–.526	−.049	−.068	−.002	−.059	+.087
	Accuracy, upper or lower third	132	.500	−.005	+.025	+.250	+.199	**+.399**
L3	Ranking three on a cognition composite	1,296	.000	.000	+.039	+.135	**+.196**	+.171
Trained on one level, one representative item per level								
L0	Order of conditions in a window					+.147	−.049	**+.665**
L1	Activation, full ranking					**+.687**	−.002	+.609
L2	Difficulty, upper or lower third					**+.256**	+.045	+.175
L3	Ranking three on a cognition composite					**+.184**	+.174	+.148

Table 3: Answer forms from one model trained on the motor paradigm, on the eight forms in its training, grouped by answer space, and on network triples not seen in training, mean ± sd over three seeds, $n = 868$ per trained form (840 for the top-three set) and 4,812 unseen triples. VLM zero-shot entries score candidate answers by log-probability, best over models and renderings, a dash marks a form with too many answers to enumerate, and ‡ marks the Spearman correlation with the true ranking in place of accuracy, with f that of a fixed ranking from training. Text-only is RIVET trained with the recording withheld, and the highest gain above .1 in a row is in bold.

					Text-only	VLM zero-shot	fMRI FM + head		RIVET	
	Input				Qwen3-VL-2B	best of 4	BrainLM	Omni-fMRI	Qwen3-VL-2B	
					none	*network heatmap*	*parcel*	*volume*	*volume*	
		Question	Answers	f	g	g	g	g	acc.	g
In training	Closed	Is X the highest?	2	.500	−.023	+.071	**+.439**	+.056	.687 ± .053	+.373 ± .106
		Is X higher than Y?	2	.500	+.022	+.194	**+.514**	+.099	.677 ± .023	+.354 ± .046
		Is X among the top three?	2	.533	−.073	+.117	**+.432**	−.065	.672 ± .024	+.298 ± .052
	Open-vocabulary	Which network is at rank r?	7	.143	+.005	+.095	**+.220**	+.024	.182 ± .009	+.046 ± .010
		Which three are highest?	35	.029	−.001	+.064	**+.237**	+.001	.204 ± .014	+.181 ± .015
		Order these three	6	.167–.177	−.028	+.124	**+.354**	+.033	.394 ± .023	+.269 ± .023
		Excluding X, which is highest?	6	.171–.190	+.008	+.356	+.412	+.063	.610 ± .008	**+.524 ± .006**
	Open-ended	Rank all seven	5,040	.210–.225‡	−.273	—	**+.617**	+.009	.694 ± .009‡	+.609 ± .016
Unseen triples	Open-vocabulary	Order these three	6	.167–.169	−.013	+.120	**+.370**	+.021	.380 ± .043	+.255 ± .052

reaches .449 ± .004 on the input of RIVET with an fMRI aggregator trained alongside, as a label that flips within a block adds two states. Per frame, a linear probe decodes the current condition from the hidden states of this model at or above the level of its frozen input (balanced accuracy .924 ± .008 against .895), with a time course that closely follows the hemodynamic pattern of the raw signal (Figure 3a, Appendix I). Trained on sex, the model falls to .756 ± .042 on this probe, so the items decide whether it keeps the time-varying state or the time-invariant trait, both in its input.

5.4 CROSS-SEGMENT INTEGRATION: COMPARING SEGMENTS GIVEN SEPARATELY

Comparison and pooling, the operations of the item generator that take several inputs, reach a language model only through items with several segments, and because the input already carries what separates participants, this item tests whether such items alone teach it to compare and pool. Each

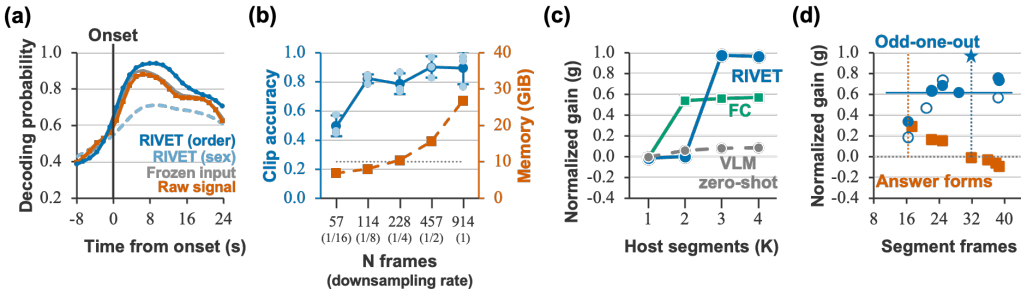

Figure 3: Reading in time, over whole runs, across segments and across paradigms. **(a)** Probe probability of the condition beginning at onset, for RIVET trained on state order or sex, its frozen input and the raw signal (Appendix I). **(b)** Four-way clip identification over whole runs and peak training memory against the frames taken at a fixed stride (Appendix F). **(c)** Gain on the odd-one-out item of the social paradigm against the number of host segments, with the best zero-shot model at each K (Appendix J.2). **(d)** Gain without further training against the mean segment length of each paradigm, for the checkpoints of (c) (filled for three hosts, open for four, star for social, line for language) and the answer-form model of Table 3, training lengths dotted (Appendix J.3).

item presents $K + 1$ segments as separate token sequences, K from one participant, the hosts, and one from another, and asks which comes from a different person (Appendix J.1). Functional connectivity (Finn et al., 2015), the reference, is at chance with one host in all seven task paradigms and reaches $+.42$ to $+.57$ from two hosts on. A parameter-free cosine rule on the raw channel of the input of RIVET, which keeps run-invariant anatomy and intensity, separates participants from two hosts on ($+.997$), as does the volume-based model's embedding, whereas the parcel-based model, whose input removes each parcel's mean, reaches .15 and .12 with a trained head (Table 27). On the social paradigm RIVET is at chance with one host, a control no comparison can solve, and learns the item from three hosts on, one host later than the reference (.538 with two), with gains of .976 and .964 with three and four hosts where zero-shot vision–language models reach at most .085 (Figure 3c, Appendix J.2). Without further training, the same checkpoints exceed f by more than two participant-clustered standard errors in 29 of 33 paradigm-host combinations, the four exceptions losing the format (Figure 3d, Appendix J.3).

6 Discussion and Limitations

Construction. For a signal without captions, whether generated text requires the signal is measurable before training if a record such as a task design determines the answers. With the cells fixed in advance, what a model fails to learn, such as the odd-one-out item with two hosts, is reported, and the three items that isolate a capability form a reference set for later corpus designs. Transfer declines with distance from the training length of 16 to 17 frames, making segment length the next design variable (Figure 3d, Appendix H.2).

Interface. Zero-shot vision–language models miss the order of states, cognition and identity (Table 2), yet their natural-image encoder carries condition, sex and diagnosis under a linear probe, split between raw and time-standardized features (Appendix C, D). We propose tracking the hemodynamic pattern of the raw signal as a minimal check for any encoder of fMRI time series.

Limitations. We claim parity with, not superiority over, domain-pretrained foundation models on these items, and no gain for downstream models. Each capability is trained on one paradigm of one dataset, one family of vision–language models is tested, and matched-budget low-rank adaptation is untested. Splits are not by family, which may inflate participant-trait items but only makes the odd-one-out item harder.

AI USE STATEMENT

We used generative AI tools to draft and revise the conceptual framework and the manuscript, for literature search, and for editing code and figure captions. They were not used to generate data or to formalize mathematical claims: every item is produced by rule from quantities computed from the recordings, and no generative model writes a question, answer or label. Every reported number is traced to a result document in the project repository, every claim about prior work was checked against the cited paper, and the authors read and approved each sentence. We take responsibility for the final content of this work.

ETHICS STATEMENT

We acquired no new data: the recordings of the benchmark come from the Human Connectome Project Young Adult release, collected under the Washington University in St. Louis institutional review board with written informed consent, and the encoder comparison of Appendix C also reads the ABIDE and ADHD-200 releases of the International Neuroimaging Data-sharing Initiative under their data use terms. Our release inherits that release's open-access data use terms. It distributes anonymized participant identifiers with no mapping to the originals or to restricted variables, and participant-level labels only as tertiles. Items about a participant measure whether a representation carries a signal, and accuracy on them is not a basis for inferences about any individual.

REPRODUCIBILITY STATEMENT

All data derive from the Human Connectome Project Young Adult release, whose open-access terms permit redistributing derived data under the same terms, so the benchmark is distributed self-contained (Appendix L). The parameters a regenerated label must match are pinned in Appendix L, and the item generator and the training and evaluation code are released upon acceptance.

REFERENCES

Mohammad H. Abbasi, Favour Nerrise, Shaurnav Ghosh, Ridvan Yesiloglu, Yuncong Mao, Bailey Trang, Mohammad Asadi, Merryn Daniel, Gustavo Chau Loo Kung, Ken Chang, Pavan Pinkesh Shah, Adam Turnbull, Kyan Younes, Seena Dehkharghani, and Ehsan Adeli. NeuroQA: A large-scale image-grounded benchmark for 3D brain MRI understanding. *arXiv preprint arXiv:2605.20525*, 2026.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, et al. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, 2022.

Mohammad Asadi, Jack W. O'Sullivan, Fang Cao, Tahoura Nedaee, Kamyar Fardi, Li Fei-Fei, Ehsan Adeli, and Euan Ashley. Mirage: The illusion of visual understanding. *arXiv preprint arXiv:2603.21687*, 2026.

Roman Beliy, Matias Cosarinsky, Oliver Heinimann, Navve Wasserman, and Michal Irani. Brain-IT-VQA: From brain signals to answers. *arXiv preprint arXiv:2605.29588*, 2026.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, 2024. arXiv:2403.20330.

Mouxiang Chen et al. VisionTS: Visual masked autoencoders are free-lunch zero-shot time series forecasters. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2025.

Mayank Daswani et al. Plots unlock time-series understanding in multimodal models. *arXiv preprint arXiv:2410.02637*, 2024.

Bernardo P. de Almeida, Guillaume Richard, Hugo Dalla-Torre, et al. A multimodal conversational agent for DNA, RNA and protein tasks. *Nature Machine Intelligence*, 7:928–941, 2025. doi: 10.1038/s42256-025-01047-1.

Zijian Dong, Ruilin Li, Yilei Wu, Thuan Tinh Nguyen, Joanna Su Xian Chong, Fang Ji, Nathanael Ren Jie Tong, Christopher Li Hsian Chen, and Juan Helen Zhou. Brain-JEPA: Brain dynamics foundation model with gradient positioning and spatiotemporal masking. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://openreview.net/forum?id=gtU2eLSAmO. Spotlight.

Zijian Dong, Ruilin Li, Joanna Su Xian Chong, Niousha Dehestani, Yinghui Teng, Yi Lin, Zhizhou Li, Yichi Zhang, Yapei Xie, Leon Qi Rong Ooi, B. T. Thomas Yeo, and Juan Helen Zhou. Brain harmony: A multimodal foundation model unifying morphology and function into 1D tokens. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL https://openreview.net/forum?id=tPJg65EB7D.

Emily S. Finn, Xilin Shen, Dustin Scheinost, Monica D. Rosenberg, Jessica Huang, Marvin M. Chun, Xenophon Papademetris, and R. Todd Constable. Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. *Nature Neuroscience*, 18(11):1664–1671, 2015.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665–673, 2020. doi: 10.1038/s42256-020-00257-z.

Sam Gijsen, Marc-Andre Schulz, and Kerstin Ritter. Brain-Semantoks: Learning semantic tokens of brain dynamics with a self-distilled foundation model. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://arxiv.org/abs/2512.11582.

Wentao Gu, Yuquan Li, Xinyang Jiang, Zilong Wang, Dongsheng Li, Zehui Li, Zijian Dong, and Cairong Zhao. Joint adaptation of uni-modal foundation models for multi-modal Alzheimer's disease diagnosis. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=gPTjQxC74G.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 107–112, 2018. doi: 10.18653/v1/N18-2017. URL https://aclanthology.org/N18-2017/.

Jiaming Han et al. OneLLM: One framework to align all modalities with language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Robert K. Heaton, Natacha Akshoomoff, David Tulsky, Dan Mungas, Sandra Weintraub, Sureyya Dikmen, Jennifer Beaumont, Kaitlin B. Casaletto, Kevin Conway, Jerry Slotkin, and Richard Gershon. Reliability and validity of composite scores from the NIH Toolbox Cognition Battery in adults. *Journal of the International Neuropsychological Society*, 20(6):588–598, 2014. doi: 10.1017/S1355617714000241.

Xingcan Hu, Wei Wang, and Li Xiao. FCN-LLM: Empower LLM for brain functional connectivity network understanding via graph-level multi-task instruction tuning. *arXiv preprint arXiv:2603.01135*, 2026. URL https://arxiv.org/abs/2603.01135.

Ming Jin et al. Time-LLM: Time series forecasting by reprogramming large language models. In *International Conference on Learning Representations*, 2024.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1988–1997, 2017. doi: 10.1109/CVPR.2017.215.

Peter Yongho Kim, Junbeom Kwon, Sunghwan Joo, Sangyoon Bae, Donggyu Lee, Yoonho Jung, Shinjae Yoo, Jiook Cha, and Taesup Moon. SwiFT: Swin 4D fMRI transformer. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://arxiv.org/abs/2307.05916.

Peter Yongho Kim, Juhyeon Park, Jungwoo Park, Jubin Choi, Jungwoo Seo, Jiook Cha, and Taesup Moon. Can natural image autoencoders compactly tokenize fMRI volumes for long-range dynamics modeling? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. URL https://arxiv.org/abs/2604.03619.

Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad Shahbaz Khan. GeoChat: Grounded large vision-language model for remote sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27831–27840, 2024.

Connor Lane, Mihir Tripathy, Leema Krishna Murali, Ratna Sagari Grandhi, Shamus Sim Zi Yang, Sam Gijsen, Debojyoti Das, Manish Ram, Utkarsh Kumar Singh, Cesar Kadir Torrico Villanueva, Yuxiang Wei, Will Beddow, Gianfranco Cortés, Suin Cho, Daniel Z. Kaplan, Benjamin Warner, Tanishq Mathew Abraham, and Paul S. Scotti. Scaling vision transformers for functional MRI with flat maps. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2510.13768.

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew E. Peters, Ashish Sabharwal, and Yejin Choi. Adversarial filters of dataset biases. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, 2020. URL http://proceedings.mlr.press/v119/bras20a/bras20a.pdf.

Bo Li, Yuanhan Zhang, Dong Guo, et al. LLaVA-OneVision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, 2023.

Heinrich R. Liesefeld and Markus Janczyk. Combining speed and accuracy to control for speed-accuracy trade-offs(?). *Behavior Research Methods*, 51(1):40–60, 2019. doi: 10.3758/s13428-018-1076-x.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, 2023a.

Zhiyuan Liu, Sihang Li, Yanchen Luo, Hao Fei, Yixin Cao, Kenji Kawaguchi, Xiang Wang, and Tat-Seng Chua. MolCA: Molecular graph-language modeling with cross-modal projector and uni-modal adapter. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 15623–15638, 2023b. doi: 10.18653/v1/2023.emnlp-main.966.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-ChatGPT: Towards detailed video understanding via large vision and language models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

Josue Ortega Caro, Antonio H. de O. Fonseca, Syed A. Rizvi, et al. BrainLM: A foundation model for brain activity recordings. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=RwI7ZEfR27.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics (*SEM)*, 2018. URL https://aclanthology.org/S18-2023/.

Weikang Qiu, Zheng Huang, Haoyu Hu, Aosong Feng, Yujun Yan, and Rex Ying. MindLLM: A subject-agnostic and versatile model for fMRI-to-text decoding. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025. URL https://arxiv.org/abs/2502.15786.

Qwen Team. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

Alexander Schaefer, Ru Kong, Evan M. Gordon, Timothy O. Laumann, Xi-Nian Zuo, Avram J. Holmes, Simon B. Eickhoff, and B. T. Thomas Yeo. Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. *Cerebral Cortex*, 28(9):3095–3114, 2018. doi: 10.1093/cercor/bhx179.

Roy Schwartz and Gabriel Stanovsky. On the limitations of dataset balancing: The lost battle against spurious correlations. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pp. 2182–2194, 2022. URL https://aclanthology.org/2022.findings-naacl.168/.

Karan Singhal, Shekoofeh Azizi, Tao Tu, et al. Large language models encode clinical knowledge. *Nature*, 620:172–180, 2023. doi: 10.1038/s41586-023-06291-2.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*, 2022.

Cheng Wang, Yu Jiang, Zhihao Peng, Chenxin Li, Chang-Bae Bang, Lin Zhao, Wanyi Fu, Jinglei Lv, Jorge Sepulcre, Carl Yang, Lifang He, Tianming Liu, Xue-Jun Kong, Quanzheng Li, Daniel S. Barron, Anqi Qiu, Randy Hirschtick, Byung-Hoon Kim, Hongbin Han, Xiang Li, and Yixuan Yuan. Towards a general-purpose foundation model for functional MRI analysis. *Nature Biomedical Engineering*, 2026a. doi: 10.1038/s41551-026-01666-y.

Mo Wang, Junfeng Xia, Wenhao Ye, Enyu Liu, Kaining Peng, Jianfeng Feng, Quanying Liu, and Hongkai Wen. SLIM-Brain: A data- and training-efficient foundation model for fMRI data analysis. *arXiv preprint arXiv:2512.21881*, 2026b. URL https://arxiv.org/abs/2512.21881.

Mo Wang, Wenhao Ye, Junfeng Xia, Junxiang Zhang, Xuanye Pan, Minghao Xu, Haotian Deng, Hongkai Wen, and Quanying Liu. Omni-fMRI: A universal atlas-free fMRI foundation model. *arXiv preprint arXiv:2601.23090*, 2026c.

Peng Wang, Shuai Bai, et al. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Weiyun Wang et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Xinxu Wei, Kanhao Zhao, Yong Jiao, Lifang He, and Yu Zhang. A brain graph foundation model: Pre-training and prompt-tuning across broad atlases and disorders. In *International Conference on Learning Representations (ICLR)*, 2026a. URL https://openreview.net/forum?id=PeGHkAaRxs.

Yuxiang Wei, Yanteng Zhang, Xi Xiao, Chengxuan Qian, Tianyang Wang, and Vince D. Calhoun. fMRI-LM: Towards a universal foundation model for language-aligned fMRI understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026b. URL https://arxiv.org/abs/2511.21760.

Ziquan Wei, Tingting Dan, and Guorong Wu. Large connectome model: An fMRI foundation model of brain connectomes empowered by brain-environment interaction in multitask learning landscape. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 2155–2163, 2026c.

Junfeng Xia, Wenhao Ye, Xuanye Pan, Xinke Shen, Mo Wang, and Quanying Liu. Brain-DiT: A universal multi-state fMRI foundation model with metadata-conditioned pretraining. *arXiv preprint arXiv:2604.12683*, 2026. URL https://arxiv.org/abs/2604.12683.

Yanwu Yang, Chenfei Ye, Guinan Su, Ziyao Zhang, Zhikai Chang, Hairui Chen, Piu Chan, Yue Yu, and Ting Ma. BrainMass: Advancing brain network analysis for diagnosis with large-scale self-supervised learning. *IEEE Transactions on Medical Imaging*, 2024. doi: 10.1109/TMI.2024.3414476.

B. T. Thomas Yeo, Fenna M. Krienen, Jorge Sepulcre, Mert R. Sabuncu, Danial Lashkari, Marisa Hollinshead, Joshua L. Roffman, Jordan W. Smoller, Lilla Zöllei, Jonathan R. Polimeni, Bruce Fischl, Hesheng Liu, and Randy L. Buckner. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of Neurophysiology*, 106(3):1125–1165, 2011. doi: 10.1152/jn.00338.2011.

Jieyu Zhang, Weikai Huang, Zixian Ma, Oscar Michel, Dong He, Tanmay Gupta, Wei-Chiu Ma, Ali Farhadi, Aniruddha Kembhavi, and Ranjay Krishna. Task me anything. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024) Datasets and Benchmarks Track*, 2024. arXiv:2406.11775.

Yiyuan Zhang, Kaixiong Gong, Kaipeng Zhang, Hongsheng Li, Yu Qiao, Wanli Ouyang, and Xiangyu Yue. Meta-transformer: A unified framework for multimodal learning. *arXiv preprint arXiv:2307.10802*, 2023.

A Admissibility and signal-free accuracy in corpus construction

The item generator builds items by rule from labels at four levels, and every item carries, before any model is trained, the accuracy of a rule that answers it without the recording. Table 4 lists every label used in this paper, its source in the Human Connectome Project Young Adult release, and whether it is recorded there or computed by us. No level has an upper reference, since how far the signal reflects the design (Level 0) is not under our control and a label computed from the recording (Level 1) or aggregated from lower levels (Level 3) cannot carry more than the recording.

A.1 Query axes and admissibility

A label block fixes what is measured and a query block how it is asked, along five axes (Figure 4), and the admissible combinations form the design space of §3.2. The question form (A) selects among inputs by their values, from the highest element through a given rank to a full order, relates them, from a property of one element to agreement between two labels, aggregates over a set, from a count to the most frequent element, or assigns the input to a category defined outside it. The operation (B) reads one input, pools several or compares several, and the remaining axes set whether the question is conditioned on one element (C), covers one segment or several (D), and refers to one label or several (E). A format is admitted only where the axis its operation runs along has at least two elements, which is the length of the label when one input is read or several are pooled and the number of inputs when they are compared. By this rule, selection and aggregation need at least two elements, classification applies only to one, pooling and comparison need several inputs, pooling is not admitted for a label ordered across segments, conditioning needs two elements left once one is held fixed, and a question over several labels cannot be a classification. The rule encodes only what is impossible by construction, so cells that fail when measured stay in the grid and are reported as excluded.

Over the eleven labels of the joint mixture (§5.1) the rule admits 309 of the 1,056 combinations of label and query block, and 288 of 960 without participant identity. The mixture builds 16 of the 309, the fourteen item types of Table 2 with the cognition ranking counted once per composite, in 158,579 items from 1,067 participants of the working-memory paradigm, split by participant. Of the other 293, 2 are built in the datasets of other experiments, 77 are closed by measurement, 71 are deferred as multi-segment forms of cells that already have a single-segment form, 6 are excluded because a behavioral label cannot be carried by a single block, and 137 have not yet been attempted. The released corpus holds 2,234,078 distinct items from 1,097 participants, with no item shared between experiments. Which admissible cells are built was decided without reference to any model's accuracy, by what each asks for (Table 5).

Table 4: Labels of the text–fMRI items used in this paper, by level, with the paradigms they are asked on and their source in the Human Connectome Project Young Adult release. WM is working memory.

Label	Level	Paradigm	Source	Obtained	Used in
Task condition	L0	motor, WM	task event timing files	recorded	App. E, D
Condition sequence	L0	language, WM	task event timing files	recorded	§5.3, 5.1
Movie clip	L0	movie	7T movie runs	recorded	App. F
Participant identity	L0	social, WM	participant index, anonymized	recorded	§5.4, 5.1
Activation, slope, coupling	L1	motor, WM	preprocessed task signal pooled to 7 networks	computed	§5.2, 5.1
Individual coupling	L1	WM	the same, against a population reference from the training split	computed	§5.1
Reaction time, accuracy	L2	WM	E-Prime trial records of the 2-back blocks, per block the median of `Stim.RT` and the mean of `Stim.ACC` over answered trials	recorded	§5.1
Difficulty	L2	WM	reaction time and accuracy standardized within condition and combined as $(z_{\mathrm{RT}} - z_{\mathrm{ACC}})/\sqrt{2}$ (Liesefeld & Janczyk, 2019)	computed	§5.1
Cognition, three composites	L3	WM	age-adjusted NIH Toolbox composites `CogFluidComp`, `CogCrystalComp`, `CogTotalComp`	recorded	§5.1
Sex	L3	motor, WM	open-access demographics	recorded	App. E, D

Table 5: Cells built for the joint mixture and what each is chosen to ask for.

Label	Level	Form	Chosen to ask for
Condition sequence	L0	full order over blocks	when each state began, which the hemodynamic delay separates from the schedule
Activation, slope	L1	pairwise relation, top one with one excluded, full ranking	relations between networks, from a judgment to a ranking
Coupling	L1	top one among six, given a network	which network moves with a given one
Individual coupling	L1	three relations pooled over eight blocks	a participant's coupling against the population
Reaction time, accuracy, difficulty	L2	upper or lower third of participants, with thirds set on the training split, pooled over eight blocks of one	behavior, with difficulty integrating speed and accuracy (Liesefeld & Janczyk, 2019)
Cognition, three composites	L3	order of three participants by score	the composite scores of Heaton et al. (2014)
Participant identity	L0	odd one out over four segments	integration across segments received separately

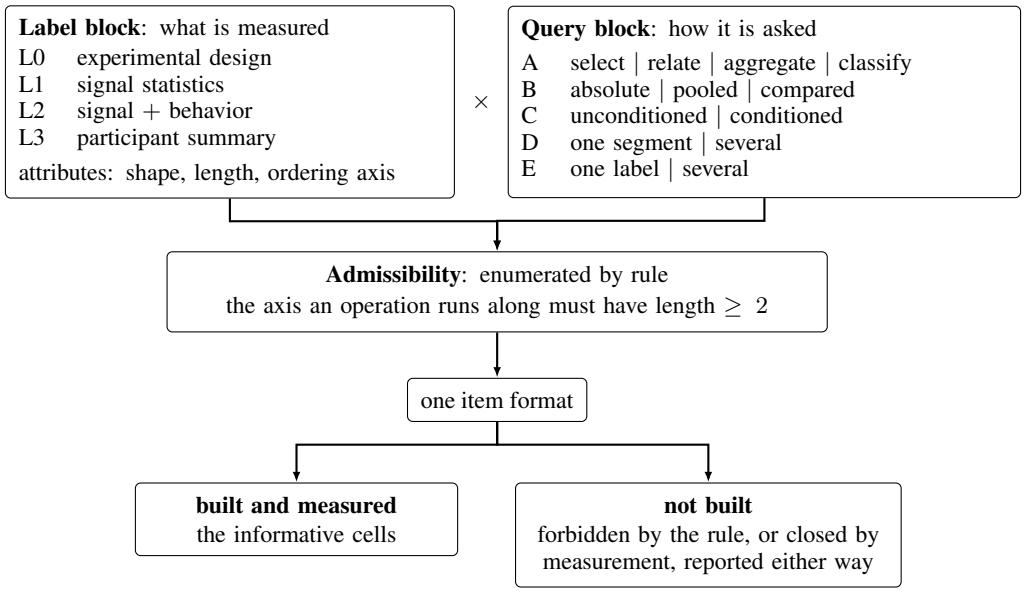

Figure 4: The item generator. A label block and a query block compose into an item format when the admissibility condition permits it.

A.2 Metrics, signal-free accuracy and item controls

Each item is scored with the metric of its form, accuracy for choices and judgments, exact match for sequences, sets and the order of three named networks, and the Spearman correlation for full rankings of networks or participants, with an unparseable answer counted wrong or as zero. The signal-free accuracy f is computed in the same metric by a rule fitted on the training split, which maps everything an item states apart from the recording, such as its form, the networks it names, the rank it asks for and its options, to the most frequent answer, the consensus ranking or the most frequent sequence. A context seen fewer than 20 times in training backs off to a coarser one, ending at the item type. The rule is applied to the test split and floored at chance, since balancing can make the most frequent training answer rare at test, whereas a text-only model can always reach chance by guessing. Results are read as the normalized gain of Equation 1, and a gain over several forms is the mean of their gains, each against its own f.

Three further quantities are fixed with f. The conditional majority, the accuracy of always giving the most frequent answer of an item type, is a looser reference than f, which conditions on the full question. The margin between the correct answer and its nearest distractor removes near ties below 0.10 label standard deviations, whose labels do not reproduce across repeated measurements, and each retained item carries its margin quartile so that difficulty can be graded within a cell. Recoverability, the share of items whose answer from the frames a model receives equals the answer from the whole block, is computed from the signal alone and stands in for human performance, which does not exist for these recordings. Each context contributes at most 20 items to a cell, and f is released for every candidate cell, accepted or rejected, with the reason for each rejection. Balancing can discard signal when shortcut and signal share an input (Schwartz & Stanovsky, 2022), but an item answerable without the recording carries no such signal.

A.3 Admission and exclusion

The gates that admit cells use labels alone, with no model trained and no linear probe read, and their thresholds were fixed before they were computed. A Level 1 predicate is admitted at a recoverability of at least .90 with at least 2,000 items after balancing, which all 294 distinct predicates per statistic pass (recoverability .96 to .999), as do 210 of 210 orderings of three networks on the motor paradigm and 201 of 210 on working memory. For Level 2 and Level 3 labels, a reliability across runs of at least .30 in intraclass correlation separates cells reported in the main text from those in the appendix, and cells below it are still used in training. Reaction time and accuracy pass at .64 and .58, their

agreement in direction fails at $-.06$, and the Level 3 composites are instrument scores measured once per participant, with test–retest reliability reported by Heaton et al. (2014). Most exclusions by construction come from sample size. Correlations among seven networks have 21 free pairs, which a block of median length 38 frames, and as few as 6, cannot estimate, so coupling items pool the blocks of a condition, blocks too short for a cell are dropped from that cell only, and at parcel resolution the correlation matrix of a block is rank-deficient. Labels defined by relations between blocks, such as change, variability or consistency, are dominated by noise, with an intraclass correlation of 0.16 to 0.31 across repeated blocks and most such cells within 0.02 of the conditional majority, whereas labels constant within a participant gain from averaging over blocks.

A.4 Relation to other item-selection criteria

Table 6 places the design-time controls of §2 by what they select on, whether a model is needed to apply them, and what space they search. The corpus selects on the distribution of labels, needs no model, and searches the derivation of labels as well as the templates.

Table 6: How item-selection criteria differ, and what space each searches.

	Selection criterion	Model needed	Space searched
Zhang et al. (2024)	model performance	yes	question templates
Le Bras et al. (2020)	model predictability	yes	instances
Chen et al. (2024)	model responses	yes	instances
Asadi et al. (2026)	model responses	yes	instances
Johnson et al. (2017)	answer distribution	no	question templates
Abbasi et al. (2026)	answer dominance	no	question templates
Ours	label distribution	no	templates + derivation

B Model specification and training

Architecture. The stack of Table 7 is common to every experiment in the paper, and only the fMRI aggregator is trained. The raw channel keeps what is invariant within a run, and the standardized channel removes it to leave variation over time. The per-parcel embedding is added before the parcel block and the pooling, the only place where it can mark parcels, and the per-network embedding after the pooling at the language model's width, where it marks the token slot, and no textual tag names a region.

Training. Each volume enters the sequence as one video frame of seven tokens wrapped in the model's own `<|video_pad|>` markers, so the path that carries video into the language tower carries an fMRI recording unchanged. The text is the tagged question and answer of Figure 1 in the chat template of the backbone without a system turn, the loss is computed on the answer alone, and at test time the model generates greedily for at most 64 tokens. Where item types are mixed, the training files of all item types are concatenated, so each item is seen once per epoch without reweighting, and batches group items of similar length. Every run of Rivet uses AdamW with linear decay after 2 to 20 warmup steps, a learning rate between 5×10^{-6} and 5×10^{-5}, gradient clipping at 1.0 to 2.0, an effective batch of 32 and bf16 on A100 hardware, for five to thirty epochs with the language model frozen. Reported weights are those of the epoch selected on the validation split by loss and by generated answers, and scores are free generation on the test split.

C Natural-image encoder and fMRI foundation models

Read by one linear probe, a vision tower that has never seen an fMRI volume is competitive with two fMRI foundation models on resting-state, task and naturalistic recordings, so pretraining on the domain is not required to reach this regime. The encoders are ours, BrainLM (Ortega Caro et al., 2024) on parcel time series and Omni-fMRI (Wang et al., 2026c) on volumes, chosen in advance so that no task fMRI entered their pretraining.

Table 7: The stack, stage by stage, with the shape of each output. Encoder shapes are per frame. T_{run} is the number of frames in a run and T the number in the segment an item refers to. MLP denotes two linear layers of width 2,048, the second initialized at zero.

Stage	Operation	Output shape	Trained	Parameters
Frozen encoder, per frame				
Resampling	96^3, 96 slices along each of three axes	$(288, 96, 96)$	no	—
Vision tower	patch 16	$(10{,}368, 1{,}024)$	no	—
Merger	2×2 merge, $4{,}096 \to 4{,}096 \to 2{,}048$	$(2{,}592, 2{,}048)$	no	—
Parcel pool	fixed W, $2{,}592 \times 100$	$(100, 2{,}048)$	no	—
Two channels, per run				
Temporal z-scaling	raw, standardized along time in the run	$(T_{\mathrm{run}}, 100, 2{,}048)$	no	—
Concatenation	$[\,\mathrm{raw} \,\|\| \, z\,]$	$(T_{\mathrm{run}}, 100, 4{,}096)$	no	—
Per item				
Segment	frames of the segment	$(T, 100, 4{,}096)$	no	—
fMRI aggregator, the only trained component				
(1) Projection	$2{,}048 \to 1{,}024$ per channel, concatenated	$(T, 100, 2{,}048)$	yes	4,196,352
(2) Parcel block	$p \leftarrow p + E_{\mathrm{parcel}}$		yes	204,800
	$p \leftarrow p + \mathrm{MLP}(\mathrm{LN}(p))$	$(T, 100, 2{,}048)$	yes	8,396,800
(3) Network pool	fixed M, 100×7, size-weighted	$(T, 7, 2{,}048)$	no	—
(4) Output	$q \leftarrow \mathrm{MLP}(\mathrm{LN}(q))$, no residual		yes	8,396,800
	$q \leftarrow q + E_{\mathrm{network}}$	$(T, 7, 2{,}048)$	yes	14,336
Total trained				**21,209,088**
Language model	frozen, no low-rank adaptation	$7T$ tokens in	no	—

Linear probe. The linear probe used throughout the appendix is a multinomial logistic regression with an ℓ_2 penalty chosen on validation, over participant-disjoint 5-folds at three seeds shared by all encoders, on features standardized with the statistics of the training folds, reported as the mean and sample standard deviation over the 15 folds. Condition is scored per block, and sex, constant within a participant, per participant by majority vote over the blocks. The vision tower is read at 100 parcels with the raw and time-standardized channels of §4.1 concatenated, which for a mean over a whole run, as in the resting-state and movie rows, reduces to the raw channel since the standardized channel then averages to zero, and the ordered resting-state entry uses the raw channel alone. Each baseline is read with the input normalization of its release, at the spatial unit, output and layer its authors recommend, fixed before any probe was run, and is given the better of its two representations and, for Omni-fMRI, of two window spacings. BrainLM is run as released on its 434-by-434 input, with a task block's frames placed from the first time point and the rest set to zero, its class token as the block-mean entry and its token means over 14-frame bins as the ordered entry, and resting and movie runs are read in windows of 434 frames. Omni-fMRI is read in 40-frame windows, so on working-memory blocks a median of only 0.480 of a window comes from the block's own condition.

Results. On the block-mean rows of Table 8, where the encoders are comparable, the vision tower is within fold variation of or above both baselines on every resting-state and task target. A domain-pretrained encoder is ahead only on naturalistic viewing, where BrainLM reaches .9364 against .8429, a lead that the phase-encoding direction shared by pairs of clips does not explain, since BrainLM reaches .9366 and .9365 within each pair. The linear probe places Omni-fMRI below and BrainLM above the accuracies published for resting-state sex with each encoder fine-tuned (.9138 against .9554 and .7587 against .6271, Wang et al., 2026c), so it does not systematically understate either. Order adds information only where the label is defined in time, by .094 for ours and .106 for Omni-fMRI over permuted units on the working-memory condition, whereas ordered and permuted entries agree within .003 for all three encoders on resting-state sex. Dimension, which differs by two orders of magnitude between the encoders, does not raise accuracy on its own, as the permuted rows show. The condition rows partly compare temporal resolution and the participant rows rest on a single attribute, so the table supports that domain pretraining is not required rather than a ranking of encoders.

Table 8: One linear probe, three frozen encoders, three acquisitions, with chance or majority in brackets after each target. BrainLM's block-mean entry is its class token, and movie rows come from a single 5-fold split and keep each run's first units common to all four clips.

	Target *Pre-training data*	Time axis	BrainLM *fMRI, parcel*	Omni-fMRI *fMRI, volume*	Vision tower *natural images*
Resting state	HCP-YA, sex [.542]	block mean	$.7587 \pm .0287$	$.9138 \pm .0133$	$.9551 \pm .0157$
		ordered	$.7159 \pm .0372$	$.9099 \pm .0128$	$.9556 \pm .0135$
		permuted	$.7181 \pm .0292$	$.9126 \pm .0101$	$.9549 \pm .0124$
	ABIDE, ASD [.512]	block mean	$.5320 \pm .0330$	$.5491 \pm .0260$	$.6081 \pm .0270$
	ADHD-200, ADHD [.568]	block mean	$.5904 \pm .0325$	$.6666 \pm .0370$	$.6403 \pm .0333$
Task blocks	Motor, condition [.200]	block mean	$.5853 \pm .0064$	$.2835 \pm .0064$	$.9862 \pm .0038$
		ordered	$.9112 \pm .0087$	$.3236 \pm .0062$	$.9860 \pm .0038$
		permuted	$.8702 \pm .0113$	$.2838 \pm .0055$	$.9847 \pm .0040$
	Motor, sex [.541]	block mean	$.5702 \pm .0266$	$.9275 \pm .0190$	$.9485 \pm .0123$
		ordered	$.7228 \pm .0324$	$.9248 \pm .0180$	$.9331 \pm .0187$
		permuted	$.6707 \pm .0434$	$.9266 \pm .0176$	$.9337 \pm .0167$
	WM, condition [.125]	block mean	$.6511 \pm .0112$	$.1694 \pm .0059$	$.8512 \pm .0109$
		ordered	$.8561 \pm .0098$	$.2754 \pm .0058$	$.9039 \pm .0090$
		permuted	$.7852 \pm .0114$	$.1698 \pm .0046$	$.8100 \pm .0109$
	WM, sex [.541]	block mean	$.5835 \pm .0331$	$.9248 \pm .0138$	$.9546 \pm .0115$
		ordered	$.7646 \pm .0289$	$.9291 \pm .0113$	$.9346 \pm .0159$
		permuted	$.6667 \pm .0224$	$.9260 \pm .0120$	$.9319 \pm .0118$
Naturalistic movie	Four clips [.254]	block mean	$.9364 \pm .0322$	$.8329 \pm .0119$	$.8429 \pm .0292$
		ordered	$.9709 \pm .0211$	$.9737 \pm .0180$	$.9888 \pm .0081$
		permuted	$.8648 \pm .0261$	$.5647 \pm .0570$	$.4058 \pm .0342$

D INPUT REPRESENTATION: CHANNELS AND SPATIAL RESOLUTION

Both choices of the input, two channels and seven networks, were made with the linear probe of Appendix C before any fMRI aggregator was trained.

Two channels. Standardizing the frame embeddings along time within a run helps labels that vary over the run and hurts labels that do not, so the input keeps both channels. The vision tower encodes each frame on its own, so its embedding mixes what varies over a run with what is invariant within it, and standardizing along time removes the invariant part. The channels were compared with the linear probe of Appendix C on block means (21,620 motor blocks of 16 frames and 17,342 working-memory blocks of 38) through a shared random projection to 2,048 dimensions per network.

Standardizing raises the condition, which varies within a run (.316 to .971 on motor and .173 to .815 on working memory), and lowers sex, which does not (.935 to .620 and .939 to .673), while carrying both channels stays within .008 of the better one on every target (Figure 5). The gain is not the doubled width, since projecting the pair to one channel's width costs at most .015, and keeping the frames in order is no substitute, reaching only .353 and .305 on the condition.

Trained end to end, the channels split the same way, since removing the raw channel leaves the condition unchanged ($+.006$ and $-.003$, with per-seed differences of both signs) and costs sex .115 and .102 (Wilcoxon $z \leq -8.9$ in every seed, Table 9).

Table 9: The two channels trained end to end, free-generation accuracy as mean $\pm$ standard deviation over three seeds, with chance in brackets.

Channels	Trained	Motor condition [.200]	Motor sex [.500]	WM condition [.125]	WM sex [.500]
raw and standardized	21.21M	$.9618 \pm .0014$	$.9232 \pm .0133$	$.8535 \pm .0059$	$.9006 \pm .0185$
standardized only	17.01M	$.9557 \pm .0053$	$.8080 \pm .0426$	$.8568 \pm .0102$	$.7982 \pm .0121$

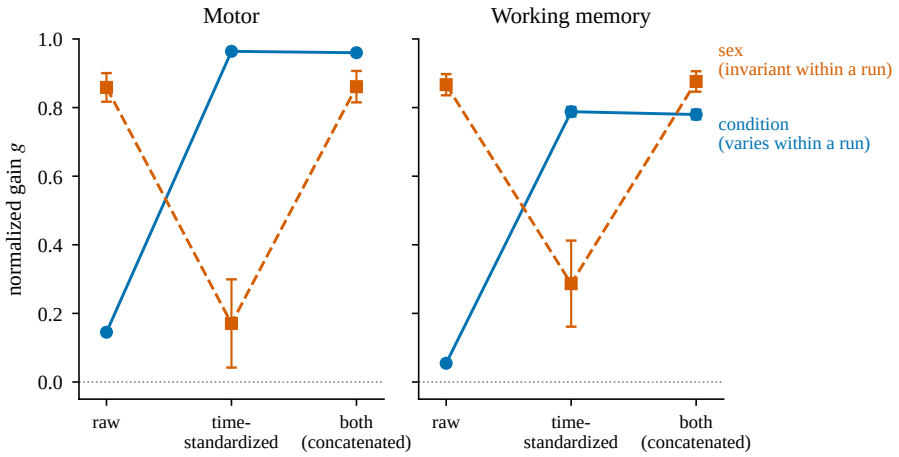

Figure 5: Linear-probe decoding from each channel and from both, as normalized gain against chance for the condition and against the majority class for sex, with bars for standard deviations over 15 folds.

Spatial resolution. Pooling to 100 parcels loses nothing measurable against 400 at a quarter of the storage, and pooling on to 7 networks costs little on the conditions. Four spatial units, 400 and 100 parcels and 17 and 7 networks, are compared with the linear probe of Appendix C on the input of §4.1, so only the spatial unit varies (Table 10). The unpooled patch grid would occupy 14.4 TiB for these two tasks against 1.12 TiB for the recordings, so it is not run and 400 parcels stand in for it.

Table 10: Spatial units, what they cost to store, and what a linear probe recovers from them. Chance is in brackets under each accuracy column head. Starred storage figures are computed rather than measured.

	Cost			Linear probe accuracy			
Spatial unit	Tokens per frame	Storage	Against the source	Motor condition [.200]	Motor sex [.500]	WM condition [.125]	WM sex [.500]
source recordings	—	1.12 TiB	1.00×	—	—	—	—
patch (not run)	2,592	14.4 TiB*	12.8×	—	—	—	—
400 parcels	400	2.22 TiB*	1.98×	.9879 ± .0032	.9442 ± .0138	.8606 ± .0092	.9432 ± .0077
100 parcels	100	569 GiB	0.50×	.9862 ± .0038	.9300 ± .0123	.8512 ± .0109	.9280 ± .0085
17 networks	17	96.7 GiB	0.084×	.9746 ± .0048	.9029 ± .0141	.8085 ± .0116	.8982 ± .0087
7 networks	7	39.8 GiB	0.035×	.9598 ± .0066	.8882 ± .0142	.7623 ± .0100	.8756 ± .0105

Seven networks cost .028 and .098 on the conditions for a 57-fold reduction from 400 parcels, and sex stays between .876 and .944 throughout. Finer units may gain partly from smoothing, since a patch touches 1.83 units at 7 networks and 6.99 at 400 parcels, and a linear probe bounds what a representation contains rather than what a trained model recovers.

E fMRI AGGREGATOR ABLATION

The trained fMRI aggregator carries nearly all of the result, and within it the choices that matter are where the residual and the channel join are placed. Motor blocks (five conditions) and working-memory blocks (eight, two loads by four stimulus categories) are cut from the recorded event files with a fixed hemodynamic shift, participants are split 70/10/20, and three seeds permute participants. The same segments are asked for their condition, scored per block, and for sex, scored per participant by majority vote over the blocks (217 and 218 participants). Accuracy is free generation with an unparseable answer counted wrong, which puts the untrained configuration below chance, since about a quarter of its motor answers and four fifths of its working-memory condition answers do not parse.

Table 11: Design axes of the fMRI aggregator, free-generation accuracy as mean ± standard deviation over three seeds, with chance in brackets.

Configuration	Trained	Motor condition [.200]	Motor sex [.500]	WM condition [.125]	WM sex [.500]
fMRI aggregator	21.21M	$.9618 \pm .0014$	$.9232 \pm .0133$	$.8535 \pm .0059$	$.9006 \pm .0185$
Pooling only, nothing trained	0	$.1592 \pm .0022$	$.4562 \pm .0410$	$.0228 \pm .0024$	$.4144 \pm .0338$
No parcel–network stage	12.61M	$.9379 \pm .0064$	$.9094 \pm .0213$	$.8333 \pm .0072$	$.8670 \pm .0138$
Sample		4,340 blocks, 217 participants		≈3,480 blocks, 218	

Removing the trained fMRI aggregator lowers every cell ($z \leq -11.6$ in a per-participant Wilcoxon test), and removing the parcel–network stage lowers every cell in every seed but one, where $p = .22$ (Table 11). Among implementation choices, placing the residual at the output costs the most (.210 and .403) at the same parameter count, joining the channels at the output is lower on both tasks with 2.6 times the parameters, and removing the parcel or network identity term stays within seed variation (Table 12). No generation of these configurations fails to parse, so what they lose is the content of an answer rather than its form.

Table 12: Implementation choices, condition target, free-generation accuracy as mean ± standard deviation over three seeds, with chance in brackets.

Configuration	Trained	Motor [.200]	WM [.125]
fMRI aggregator	21.21M	$.9618 \pm .0014$	$.8535 \pm .0059$
residual placed at the output	21.21M	$.7522 \pm .0239$	$.4508 \pm .0257$
channels joined at the output	54.80M	$.8507 \pm .0106$	$.5439 \pm .0185$
no parcel position term	21.00M	$.9604 \pm .0019$	$.8578 \pm .0051$
no network identity term	21.19M	$.9561 \pm .0022$	$.8501 \pm .0060$

Backbone. The fMRI aggregator works on another backbone, lower by .028 on the working-memory condition and .089 on sex with InternVL3.5-2B in place of Qwen3-VL-2B, in every seed (Wilcoxon $p \leq .0103$, Table 13). The vision tower, language model, chat template and position encoding change together, with slices resized to 84 pixels so that the tower still yields 2,592 tokens per frame, and under the linear probe of Appendix C the tower alone shows the gap in the same direction (condition .8492 against .8512, sex .9171 against .9546).

Table 13: Backbone swapped, working-memory task, free-generation accuracy as mean ± standard deviation over three seeds, with chance in brackets.

Backbone	Trained	Condition [.125]	Sex [.500]
Qwen3-VL-2B (ours)	21.21M	$.8535 \pm .0059$	$.9006 \pm .0185$
InternVL3.5-2B	21.21M	$.8259 \pm .0027$	$.8119 \pm .0721$

F FRAME SAMPLING AND MEMORY OVER WHOLE RUNS

RIVET trains on whole 7T movie runs of more than 900 frames, about 15 minutes at a repetition time of 1 s, on a single 40 GB A100 GPU at batch one, and its accuracy rises with the frames it sees up to every eighth frame and then levels off. Four naturalistic clips, each viewed once by every participant, give a four-way identification task at a chance level of 0.250, with 76 test items from 19 participants. Each input takes every k-th frame of the whole run for $k \in \{1, 2, 4, 8, 16\}$, a median of 914 down to 57 frames, three seeds set the starting phase of the stride, and every cell trains for 20 epochs and is scored by free generation. Accuracy rises from .4974 with every sixteenth frame to .8226 with every eighth and stays between .7862 and .9022 up to every frame (Figure 3b), above f in every cell and seed (one-sided binomial, $p \leq 1.5 \times 10^{-4}$), though every fourth frame is below every eighth. Two properties of the cohort could stand in for reading the run, and neither does. One

clip's run is shorter (901 against 914 frames), and a rule answering from the frame count reaches .40 to .52, below the model in 14 of 15 cells, while scoring only the three clips of equal length keeps the whole-run accuracy at .8619 ($p \leq 1.9 \times 10^{-7}$ in every seed). Clips share a phase-encoding direction in pairs, and the model resolves the choice within each pair as well as the pair.

Peak training memory grows linearly with the visual tokens, about 5.4 GiB plus 3.4 MiB per token, from 6.97 GiB at every sixteenth frame to 26.85 GiB for the whole run. Measured on an 80 GB card, it is set by the product of sequence length and batch rather than by either, varying by at most 2.0 GiB within a row of Table 14, so the limit on that card is about 12,800 visual tokens per step, which on a 40 GB card is batch one for a whole run. Pooling is what makes this possible, since the tower emits 2,592 tokens per frame and a naturalistic run of 921 frames would be 2,387,232 patch tokens against 6,447 pooled, which rules patches out by memory well before the context limit of 262,144 positions, whereas the foundation models of Appendix C read fixed windows of 40 or 434 frames.

Table 14: Peak training memory (GiB) on an 80 GB card, by visual tokens per step ($T \times 7 \times B$) and sequence length T. The batch B is in parentheses. A dash is a combination that was not run or does not exist.

Tokens per step	T=114	T=228	T=457	T=914
≈3,200	16.3 (4)	15.8 (2)	15.8 (1)	—
≈6,400	27.1 (8)	26.2 (4)	26.0 (2)	26.9 (1)
≈12,800	48.7 (16)	46.9 (8)	46.7 (4)	48.7 (2)
≈25,600	—	*fails* (16)	*fails* (8)	*fails* (4)

G BASELINES AND ZERO-SHOT MODELS

Every baseline answers the same test items as RIVET and is scored with the metrics of Appendix A.2. The coupling pattern of one participant is scored by exact match over its three pairs, and element-wise accuracy against an f of .500 orders the jointly trained models in the same way ($+.013$ for RIVET, $+.034$ for BrainLM and $+.051$ for Omni-fMRI).

Foundation-model heads. The two foundation models of Appendix C are frozen and read as there, the parcel-based one on the frames of an item's segment alone, a shape not seen in its pre-training. Their task-specific components, limited to 21.2M trained parameters in total and selected by validation loss, are the following.

Odd one out. Each segment is embedded on its own, the $K + 1$ embeddings are concatenated, and an MLP with one hidden layer of 512 units classifies the position of the intruder. A rule with no parameters, which names the segment least similar in cosine to the others, is reported beside it as the geometry the embedding already provides.

Answer forms. An MLP head regresses the seven network values from the segment's embedding, pooled by a learned attention where a segment spans several windows, and eight deterministic rules derive every form from them, the full ranking by sorting. A ridge regression from the same embedding, with its penalty chosen on validation participants, is a second readout, and the main text reports the one with the lower validation error. Given the true network values the rules answer every item correctly, so what limits these baselines is how much of a segment's network activity their embeddings carry.

State sequence. A unidirectional GRU over the model's own time units labels each step with the condition of the block it falls in or as preceding the segment, and a rule drops the preceding steps and merges adjacent labels into the sequence.

Joint mixture. Seven groups of items, activation, slope, coupling, cognition, behavior, identity and state sequence, share one trunk, a linear layer with GELU and dropout, followed by pooled regression heads with the rules above, a masked classifier over the three behavior terciles, a permutation-equivariant scorer for identity and a GRU for the state sequence. The loss sums the task losses, each divided by that of a constant prediction, and the configuration is chosen on this normalized validation loss.

In the single configuration each group has its own head, and the population reference for the individual coupling pattern comes from the training split only. The two readouts of each foundation model give nearly the same gains form by form (Table 15).

Table 15: Both readouts on both foundation models for the answer forms of Table 3, normalized gain against f as the mean over three seeds, with the full ranking scored by Spearman correlation and the mean taken over the eight forms.

	BrainLM		Omni-fMRI	
Form	MLP head	Ridge	MLP head	Ridge
Is X the highest?	+.439	+.430	+.055	+.056
Is X higher than Y?	+.514	+.510	+.050	+.099
Is X among the top three?	+.432	+.439	−.075	−.065
Which network is at rank r?	+.220	+.208	−.002	+.024
Which three are highest?	+.237	+.224	+.004	+.001
Order these three	+.354	+.348	−.003	+.033
Excluding X, which is highest?	+.412	+.408	+.040	+.063
Rank all seven, Spearman	+.617	+.613	−.054	+.009
Mean over forms	+.403	+.398	+.002	+.028
Validation error (MSE)	.083	.090–.091	.392–.422	.357

Zero-shot models. Qwen3-VL-2B-Instruct, Qwen3-VL-8B-Instruct, InternVL3.5-2B and InternVL3.5-8B are given a heatmap of the recording at the network or the parcel resolution, together with the question as posed to RIVET, and answer once by greedy decoding. Where candidate answers exist their log-probabilities are also scored, so that a model is not judged on answer format alone, and the main text takes the best per item over models, renderings and the two ways of scoring (Table 16). Most zero-shot scores at or below f come from generations that never close the answer tag within the length limit or that paraphrase the answer, and the coupling item, with f above .9, turns any departure from the most frequent answer into a large negative gain.

Table 16: Zero-shot vision–language models (VLMs) on the headline items, normalized gain against f under free generation with the share of parseable generations in parentheses, and the best gain over the six cells when candidate answers are scored by log-probability, where candidates exist. The identity row averages three data seeds and reads the video number without requiring a closed answer tag.

		Qwen3-VL				InternVL3.5		
Level	Item	2B, network	8B, network	2B, parcel	8B, parcel	2B, network	8B, network	Best by log-probability
L0	Order of conditions in a window	−.067 (0.00)	−.067 (0.00)	−.067 (0.00)	−.067 (0.00)	−.067 (0.01)	−.067 (0.00)	+.000
	Segment from a different participant	+.014 (1.00)	+.021 (1.00)	−.195 (0.40)	+.020 (1.00)	−.001 (1.00)	+.008 (1.00)	—
L1	Activation, pairwise	+.046 (1.00)	+.097 (1.00)	+.014 (1.00)	+.054 (1.00)	−.749 (0.19)	+.239 (1.00)	+.241
	Activation, highest excluding one	−.200 (0.00)	−.200 (0.00)	−.200 (0.00)	−.200 (0.00)	−.200 (1.00)	−.200 (0.99)	+.233
	Activation, full ranking	−.440 (0.00)	−.440 (0.00)	−.440 (0.00)	−.440 (0.00)	−.069 (0.96)	−.440 (0.00)	—
	Slope, pairwise	−.014 (1.00)	+.000 (1.00)	−.009 (1.00)	+.002 (1.00)	−.844 (0.17)	−.009 (1.00)	+.001
	Slope, highest excluding one	−.200 (0.00)	−.200 (0.00)	−.200 (0.00)	−.200 (0.00)	−.200 (0.66)	−.200 (0.65)	−.004
	Slope, full ranking	−.228 (0.00)	−.228 (0.00)	−.228 (0.00)	−.228 (0.00)	−.325 (0.97)	−.228 (0.00)	—
	Coupling to a reference network	−10.77 (1.00)	−14.63 (0.00)	−13.97 (1.00)	−14.63 (0.00)	−14.47 (1.00)	−9.99 (1.00)	−4.93
	Coupling pattern of one participant	−.143 (1.00)	−.143 (0.32)	−.143 (1.00)	−.143 (0.21)	−.143 (0.42)	−.143 (0.87)	—
L2	Difficulty, upper or lower third	−.039 (1.00)	−.039 (1.00)	−.039 (1.00)	−.039 (1.00)	−.039 (1.00)	+.054 (1.00)	—
	Reaction time, upper or lower third	−.083 (1.00)	−.068 (1.00)	−.068 (1.00)	−.068 (1.00)	−.068 (1.00)	−.068 (1.00)	—
	Accuracy, upper or lower third	+.000 (1.00)	+.000 (1.00)	+.000 (1.00)	+.000 (1.00)	+.000 (1.00)	+.025 (1.00)	—
L3	Ranking three on a cognition composite	+.000 (1.00)	+.039 (1.00)	+.000 (1.00)	+.005 (1.00)	−.006 (1.00)	+.032 (1.00)	—

Cognition by composite. The cognition item of Table 2 asks about one of three composite scores, a third of the items each, and the training split gives every permutation equally often, so f is zero for each composite and the gain equals the Spearman correlation (Table 17). Differences between composites, .05 to .13, are well inside the participant-resampled interval of one seed and composite, .33 to .43 wide, so the item is reported pooled.

Table 17: The cognition ranking item by composite score, normalized gain (the Spearman correlation, since f is zero), mean $\pm$ sd over three seeds, with the best of the six zero-shot cells on data seed 0 and 432 test items per composite.

		VLM zero-shot	BrainLM		Omni-fMRI		RIVET
Composite	n	best of 6	trained alone	joint	trained alone	joint	joint
Total	432	+.028	+.227 ± .084	+.161 ± .079	+.222 ± .061	+.215 ± .051	+.196 ± .064
Crystallized	432	+.100	+.154 ± .061	+.135 ± .028	+.206 ± .071	+.234 ± .028	+.170 ± .045
Fluid	432	+.056	+.171 ± .052	+.109 ± .041	+.093 ± .080	+.139 ± .056	+.149 ± .047
All three	1,296	+.039	+.184 ± .050	+.135 ± .045	+.174 ± .070	+.196 ± .035	+.171 ± .033

Text-only control. The text-only column of Table 2 trains RIVET on the same items with the same recipe, with every frame set to zero and the frame count fixed at 16, and evaluates the checkpoint of lowest validation loss. Pooled over the eleven items scored by accuracy, it reaches .302 against an f of .305 ($p > .2$ in every seed). No item exceeds f significantly in any seed ($p < .05$, or twice the standard error for rank correlations), and the two full rankings fall below it in every seed, so f does not inflate the gain of any item. The same recipe applied to the answer-form model of Table 3 reaches .252 against an f of .259 over its eight forms, with no form significantly above f in any seed.

Seed variation. Table 18 gives the seed standard deviation of each gain of RIVET trained jointly in Table 2 and the number of seeds whose gain exceeds .1. Difficulty and reaction time exceed it in two of three seeds and are not counted as above f, and a participant-resampled 95% interval of balanced accuracy lies above f in every seed for behavioral accuracy, in one for difficulty and in none for reaction time.

Table 18: RIVET trained jointly on all items, normalized gain as mean $\pm$ sd over three seeds and the number of seeds with a gain above .1.

Level	Item	Gain	Seeds above .1
L0	Order of conditions in a window	+.616 ± .027	3
	Segment from a different participant	+.010 ± .018	0
L1	Activation, pairwise	+.640 ± .012	3
	Activation, highest excluding one	+.503 ± .019	3
	Activation, full ranking	+.651 ± .005	3
	Slope, pairwise	+.441 ± .046	3
	Slope, highest excluding one	+.421 ± .047	3
	Slope, full ranking	+.593 ± .047	3
	Coupling to a reference network	−.316 ± .026	0
	Coupling pattern of one participant	+.003 ± .003	0
L2	Difficulty, upper or lower third	+.155 ± .061	2
	Reaction time, upper or lower third	+.087 ± .063	2
	Accuracy, upper or lower third	+.399 ± .012	3
L3	Ranking three on a cognition composite	+.171 ± .033	3

Signal-space readout. Level 1 labels are computed from the network time series of the recording, so a closed-form ridge regression from those time series, with the rules of the foundation-model heads, recovers the activation items exactly and the slope items nearly so, and one from the standardized parcel time series, the input closest to that of RIVET, reaches gains of .77 to .89 (Table 19). These readouts see only the frames of the item and are an upper reference for the signal content of these items, whereas the benchmark asks whether one text interface states the same quantities in every answer form.

Table 19: Level 1 items of Table 2 read by ridge regression from the signal, normalized gain against the mean f over three seeds, beside RIVET trained jointly.

Item	Ridge, network time series	Ridge, standardized parcel time series	RIVET, joint
Activation, pairwise	+1.000	+.895	+.640
Activation, highest excluding one	+1.000	+.772	+.503
Activation, full ranking	+1.000	+.868	+.651
Slope, pairwise	+.992	+.890	+.441
Slope, highest excluding one	+.971	+.809	+.421
Slope, full ranking	+.972	+.881	+.593

H RELATIONAL READING: RANKING METRICS AND TRANSFER ACROSS PARADIGMS

H.1 RANKING METRICS

The full ranking is scored by the Spearman correlation in the main text, and exact match, Kendall's τ, the share of network pairs placed in the right order and the overlap of the top three separate the models in the same way (Table 20).

Table 20: The full ranking of seven networks on the motor paradigm under five metrics, mean $\pm$ sd over three seeds with 868 test items each, the signal-free reference being a fixed ranking from training and pairs in order being the share of the 21 network pairs ordered as in the truth. Every metric counts all items, an answer that does not parse as a ranking scoring zero, which applies to at most 1.2% of the items of RIVET and to none of the foundation models.

	Exact match	Spearman ρ	Kendall τ	Pairs in order	Top-three overlap
Fixed ranking from training	.001–.007	.217 $\pm$ .008	.172 $\pm$ .005	.586 $\pm$ .003	.531 $\pm$ .007
RIVET	.032 $\pm$ .007	.694 $\pm$ .009	.581 $\pm$ .009	.786 $\pm$ .003	.748 $\pm$ .002
BrainLM, MLP head	.026 $\pm$ .004	.700 $\pm$ .005	.582 $\pm$ .007	.791 $\pm$ .003	.758
BrainLM, ridge	.022 $\pm$ .001	.697 $\pm$ .006	.578 $\pm$ .007	.789 $\pm$ .004	.753
Omni-fMRI, MLP head	.002 $\pm$ .001	.175 $\pm$ .024	.137 $\pm$ .018	.568 $\pm$ .009	.501 $\pm$ .018
Omni-fMRI, ridge	.000 $\pm$ .001	.224 $\pm$ .007	.179 $\pm$ .006	.589 $\pm$ .003	.520 $\pm$ .003

H.2 TRANSFER ACROSS PARADIGMS

Trained on motor segments of 16 to 17 frames, RIVET transfers without further training to paradigms whose segments are near that length and not to those of 32 frames or more. It is evaluated on four forms, rank r, pairwise, top three and highest excluding one, asked about six other paradigms, with the language paradigm split into its math and story blocks (Table 21). It transfers to language math blocks (+.292) and to relational blocks of 22 to 23 frames (+.164), weakly to emotion blocks of 25 frames (+.154, above f in two of three seeds), and not to the sets of 32 frames or more. Unparseable generations, scored wrong, concentrate in one seed (.65 to .98 of the items of 25 frames or more, against at most .112 for the other two), but the other two seeds also decline with length and stay at or below f on the working-memory and story sets when rescored on well-formed or extracted answers, so the decline is not a failure of format alone. Length coincides with paradigm across the sets and with condition within the language paradigm, so the limit is consistent with segment length without being isolated from the paradigm. BrainLM, whose fixed window and per-task input normalization absorb segment length, transfers to the longer sets as well and Omni-fMRI to none, so the limit lies in the training length of RIVET, not in the signal.

Table 21: Transfer of answer forms from the motor paradigm without further training, normalized gain against f over the four forms asked, ordered by segment length; above f counts seeds whose gain exceeds two participant-clustered standard errors, unparseable gives the share of generations with no tagged answer in seeds 0, 1 and 2, and a dash marks a set not evaluated.

			RIVET			fMRI FM + head	
Target set	Frames	f	Gain	Above f	Unparseable	BrainLM	Omni
Language, math	11–21	.334–.358	$+.292 \pm .018$	3	.002 / .000 / .000	$+.347$	$+.022$
Relational	22–23	.354–.361	$+.164 \pm .074$	3	.109 / .001 / .000	—	—
Emotion	25	.336–.345	$+.154 \pm .142$	2	.646 / .004 / .000	—	—
Social	32	.335–.346	$-.009 \pm .063$	1	.945 / .041 / .002	—	—
Language, story	29–42	.333–.361	$-.029 \pm .183$	1	.966 / .072 / .034	$+.305$	$+.036$
Working memory	38–39	.333–.353	$-.064 \pm .047$	0	.967 / .112 / .038	$+.129$	$+.010$
Gambling	38–39	.344–.348	$-.094 \pm .054$	0	.975 / .074 / .050	—	—

I READOUT CONTRAST AND LINEAR PROBE ON HIDDEN STATES FOR TEMPORAL LOCALIZATION

The model lists the states of a language segment in order well above every signal-free reference, and its hidden states follow the hemodynamic response of the raw signal. An item gives a window of 40 to 70 s of a language-task run, shorter where the run ends (28 to 98 frames in all), and asks for the two to five states it contains in temporal order, a block counting when its onset falls inside the window, with 1,680 participant-disjoint test items at each of three dataset seeds that redraw windows and participants. Table 22 gives exact match, the count of states, the per-position accuracy among items whose count is right, and the order with adjacent repeats merged, each against its signal-free reference and against a rule that reads only the number of frames.

Table 22: Free-generation scores on the state-sequence item, n=1,680, with each measure's signal-free reference in the last row; n/a marks a measure without one.

Training set	Exact	Count	Position given count	Collapsed match
three dataset seeds	$.6669 \pm .0349$	$.7534 \pm .0224$	$.9408 \pm .0129$	$.8062 \pm .0225$
signal-free reference	$.2032 \pm .0058$	$.4679 \pm .0027$	n/a	$.4464 \pm .0072$
frame-count rule	$.3250 \pm .0021$	$.6430 \pm .0184$	n/a	$.4381 \pm .0084$

Readout contrast. Each head of Table 23 is a unidirectional GRU of 21.2M parameters with three classes per step, the two conditions and the steps before the first block, trained on the same stimulus-time labels. Omni-fMRI is read in its 40-frame windows inside the segment, BrainLM in 14-frame bins of the segment's frames, and the frame-level head reads an fMRI aggregator of the RIVET architecture trained from scratch together with it. The score is the order of states with adjacent repeats merged, whose ceiling is that of a head labeling every step correctly, Omni-fMRI drops eight items that contain no whole window, and differences between RIVET and the frame-level head are tested per participant, since a participant's eight items overlap. RIVET is above the frame-level head for at least 177 of 210 participants in every seed.

Table 23: Readout contrast on the state-sequence item of the language paradigm (1,680 held-out items). RIVET, the frame-level head and the parcel-based model's head vary over three dataset seeds, the volume-based model's head over three head seeds on the first.

Input	Time unit	Readout	Trained	Ceiling	Collapsed match
Constant answer (most frequent training sequence)				—	.446
Omni-fMRI	40-frame window	GRU	21.2M	.406	$.300 \pm .026$
BrainLM	14-frame bin	GRU	21.2M	.921	$.830 \pm .002$
Vision tower + fMRI aggregator	frame	GRU	42.4M	.976	$.449 \pm .004$
RIVET	frame	frozen LM	21.2M	—	$.806 \pm .023$

Linear probe on hidden states. The linear probe reads the last-layer hidden states at the vision-token positions, averaged over the seven network tokens, from one teacher-forced pass with causal attention, so the state of a frame sees only the prompt and the frames up to it. Each frame is labeled with the condition of the block containing its stimulus time, its acquisition time minus 4.0 s, rest (6% of frames) is left out, and an ℓ_2-regularized logistic regression is fitted on every other frame over five participant folds, as it is for the raw network signal. Every probed model is read at its checkpoint of lowest validation loss. Balanced accuracy is .924 $\pm$ .008 over three seeds against a majority of .5049, and .883 $\pm$ .004 for the raw signal of the same windows. Around each onset the model's curve has the shape of the raw curve ($r = .987$ between seed-averaged curves) and peaks later, at $+7.84$ to $+9.28$ s against $+6.40$ s in every seed, consistent with the hemodynamic response or with context accumulating under causal attention, and probabilities before time zero rise when a block of the same condition precedes the onset.

Training target. The same probe is applied to RIVET trained on three targets that depend on time to different degrees, the order of states, the full ranking of activation and sex, and to its untrained input, the raw and time-standardized channels pooled to seven networks (Table 24). A target is read as changing what the model keeps when its balanced accuracy differs from that of the order model by more than twice the seed standard deviation, a rule fixed before the models were trained. The order of states stays above the level of the standardized input and the ranking of activation just below it (.881 against .895), whereas sex falls below it in every seed, with a lower peak and a higher pre-onset level, while the model still answers sex correctly for .885 $\pm$ .022 of participants by majority over their windows.

Table 24: Linear probe of the current condition on the language paradigm by training target, balanced accuracy as mean $\pm$ sd over three seeds and the peak and pre-onset level of the onset-aligned curve as the mean over seeds.

Input or target	Varies within a run	Balanced accuracy	Peak (s, value)	At -8 s
Raw signal, seven networks	—	.883 $\pm$ .004	+6.40 (.882)	.396
Frozen input, raw channel	—	.591 $\pm$ .005	+4.96 to +6.40 (.599)	.476
Frozen input, standardized channel	—	.895 $\pm$ .001	+6.40 (.903)	.396
RIVET, order of states	yes	.924 $\pm$ .008	+7.84 to +9.28 (.942)	.388
RIVET, ranking of activation	yes	.881 $\pm$ .007	+7.84 to +9.28 (.882)	.400
RIVET, sex	no	.756 $\pm$ .042	+7.84 to +10.72 (.713)	.429

J INDIVIDUAL DIFFERENCES FROM MULTI-SEGMENT COMPARISON

The information that separates participants is in the recording from two host segments on, and RIVET learns to use it from three hosts on and carries it to six paradigms it was not trained on.

J.1 RAW-SIGNAL REFERENCE

An item gives $K + 1$ task blocks as separate token sequences in arbitrary order, K hosts from one participant and one intruder from another, and asks which comes from a different person (K from one to four, chance $1/(K + 1)$). Run composition is balanced before the intruder is drawn, conditions are not matched, and with one host the two segments are interchangeable, which makes that configuration a control no reading of the comparison can solve. Gains are against f, the most frequent answer position given K, for models that read the question, and against chance for the reference, which reads none. The reference is functional connectivity (Finn et al., 2015), each segment represented by the Fisher-transformed correlations between parcel time series and the intruder taken as the segment of lowest mean cosine to the others, with no learning. It is at chance with one host in all seven paradigms and reaches $+.42$ to $+.57$ from two hosts on (Table 25), and the separability of same- against different-participant pairs rises with the number of segments in every paradigm (area under the curve up by .099 to .139 from one to four segments, reaching .774 to .856).

Table 25: Odd-one-out from the raw signal, seven task paradigms. Each entry is the normalized gain of Equation 1 over chance, mean ± standard deviation over twelve repetitions. Chance is in brackets under each column head.

Paradigm	K=1 [.500]	K=2 [.333]	K=3 [.250]	K=4 [.200]
Motor	$+.007 \pm .022$	$+.417 \pm .030$	$+.421 \pm .014$	$+.436 \pm .015$
WM	$-.009 \pm .037$	$+.494 \pm .019$	$+.511 \pm .022$	$+.528 \pm .015$
Gambling	$+.005 \pm .024$	$+.541 \pm .020$	$+.549 \pm .020$	$+.548 \pm .015$
Language	$+.008 \pm .032$	$+.472 \pm .017$	$+.518 \pm .019$	$+.517 \pm .015$
Social	$-.013 \pm .024$	$+.538 \pm .022$	$+.562 \pm .010$	$+.572 \pm .018$
Relational	$-.001 \pm .026$	$+.468 \pm .021$	$+.488 \pm .020$	$+.493 \pm .018$
Emotion	$-.012 \pm .024$	$+.488 \pm .017$	$+.525 \pm .010$	$+.530 \pm .019$

J.2 Through the language model

RIVET is trained and tested on the social paradigm, which has the largest reference gain at three and four hosts, with 5,010 to 5,027 test items per dataset seed (Table 26). At three and four hosts it is right on at least .95 of items in every seed while f stays near chance. At two hosts, where the raw signal already reaches $+.538$, it does not learn, its validation loss never falling below its chance value as it does by the third epoch at three and four hosts.

Table 26: Odd-one-out through the language readout on the social paradigm, mean ± standard deviation over three dataset seeds, with f given as its range over seeds.

Hosts K	Chance	Accuracy	f	Normalized gain
1	.5000	$.4988 \pm .0116$	.5000–.5134	$-.0114 \pm .0086$
2	.3333	$.3387 \pm .0057$	.3333–.3455	$+.0019 \pm .0160$
3	.2500	$.9823 \pm .0017$	.2500–.2534	$+.9763 \pm .0023$
4	.2000	$.9714 \pm .0135$	.2000–.2077	$+.9640 \pm .0167$

Table 27 sets RIVET beside the raw-signal reference, the foundation-model heads of Appendix G, the parameter-free cosine rule on their embeddings, and the best zero-shot cell under free generation. On Omni-fMRI the rule already solves the item from two hosts on and the head changes it by at most .004, while on BrainLM the head adds about .05 to a rule near chance. The difference follows what each input keeps, since BrainLM's input has each parcel's mean over the run removed, and the same rule on the two channels of RIVET's input reaches $+.997$, $+.996$ and $+.998$ at two to four hosts on the raw channel and $+.161$, $+.092$ and $+.098$ on the standardized one, the level of BrainLM. The input of RIVET thus separates participants at two hosts, so what fails there is learning through the interface rather than the information.

Table 27: Odd-one-out on the social paradigm, normalized gain by configuration, RIVET and VLM zero-shot against f and the other entries against chance. RIVET varies over three dataset seeds and the heads over three training seeds on the first.

			BrainLM		Omni-fMRI		
Hosts K	Raw signal	RIVET	Rule	Head	Rule	Head	VLM zero-shot
1	$-.013$	$-.011$	$-.001$	$-.009$	$-.001$	$-.005$	$-.002$
2	$+.538$	$+.002$	$+.080$	$+.134$	$+.996$	$+.994$	$+.062$
3	$+.562$	$+.976$	$+.104$	$+.153$	$+.993$	$+.994$	$+.081$
4	$+.572$	$+.964$	$+.077$	$+.123$	$+.996$	$+.993$	$+.085$

J.3 Transfer across paradigms

The three social-paradigm checkpoints are evaluated without further training on the other six paradigms at three and four hosts, on 1,000 items per paradigm and number of hosts drawn once

from the test participants, with the answer position balanced so that f equals chance, and gambling has too few blocks per run for four hosts. A checkpoint passes a cell when its accuracy exceeds f by more than two standard errors clustered by host participant, 205 to 210 per cell, so that the items of one participant may be correlated. Of the checkpoint cells, 17 of 18 pass at three hosts and 12 of 15 at four, the four failures coming from checkpoints that lose the answer format on .79 to .96 of their generations, while accuracy on parseable generations is at least .74 in 29 of 33 cells (Table 28). The same cells pass with items treated as independent, with clustering by the other participant and under a bootstrap over host participants, and the narrowest margin is 2.8 clustered standard errors, at four hosts on the motor paradigm. The best checkpoint reaches .950 to .975 on all six paradigms at three hosts against .980 on the social paradigm, and motor, with the shortest segments, is lowest at both numbers of hosts.

Table 28: Odd-one-out transfer from the social paradigm without further training, normalized gain against f (chance) as mean $\pm$ standard deviation over the three checkpoints, with the mean share of unparseable generations and the mean accuracy on parseable ones, 1,000 items per cell; the four-host gambling cell cannot be built.

		Three hosts			Four hosts		
Paradigm	Frames	Gain	Unparseable	Parsed acc.	Gain	Unparseable	Parsed acc.
Working memory	38–39	$+.758 \pm .176$	.053	.869	$+.572 \pm .619$	.303	.941
Gambling	38–39	$+.733 \pm .203$	.066	.862	—	—	—
Emotion	25	$+.686 \pm .356$	.160	.895	$+.740 \pm .268$	.150	.939
Relational	22–23	$+.636 \pm .404$	.205	.902	$+.635 \pm .335$	.201	.905
Language	11–42	$+.616 \pm .319$	.153	.830	$+.471 \pm .466$	.327	.887
Motor	16–17	$+.338 \pm .535$	.289	.757	$+.190 \pm .471$	.500	.769
Social (trained)	32	$+.976 \pm .002$	.000	.982	$+.964 \pm .017$	.000	.971

K PROMPT TEMPLATES

The corpus is released in one template. Slots are in braces, and the run of visual placeholder tokens that stands for a segment is shown as `<video: 7 network tokens x T frames>`, at which position the model receives seven fMRI aggregator outputs per frame.

template

```
<Task>
[{TAG}] {role sentence}
Target quantity: {quantity asked about}.
Answer format: {answer format}.</Task>

Video 1: <video: 7 network tokens x T frames>
...
Video n: <video: 7 network tokens x T frames>

<Question>
{question}</Question>

<Answer>
{answer}
</Answer>
```

The role sentence reads `You are an expert AI analyzing brain activity during a task.`, with the acquisition named where it differs, such as `during movie watching`. Segments are listed as above when an item concerns one person or asks which segment comes from another, and when it compares named persons, which in the joint mixture is the cognition ranking, the segments of each person are enclosed in `<Person_id="Person A">` ... `</Person_id>`. Tables 29 and 30 give the question and answer of one test item for every item type.

The working-memory condition item of the ablation offers the eight options `0BK_BODY` to `2BK_TOOLS`, two loads crossed with four stimulus categories, and the sex item, tagged

[SEX], asks What is the biological sex of this subject? Answer with one of: MALE, FEMALE.. With three hosts the odd-one-out question begins The following are 4 videos. and its options end at (D) Video 4.

Table 29: Items outside the joint mixture, the question and answer of one item each.

Experiment	Tag	Question	Answer
fMRI aggregator ablation, motor condition	[STATE]	Which cognitive condition was the subject in during this fMRI segment? Answer with one of: LEFT_FOOT, LEFT_HAND, RIGHT_FOOT, RIGHT_HAND, TONGUE.	LEFT_FOOT
Whole runs	[MOVIE]	Which movie was this subject watching? Answer with one of: MOVIE1, MOVIE2, MOVIE3, MOVIE4.	MOVIE4
Answer forms, closed	[STMT·ACT]	Is the following true of this segment? "the Visual network has higher mean activation than the Ventral Attention network" Answer True or False.	False
	[STMT·ACT]	Is the following true of this segment? "the Limbic network has the highest mean activation" Answer True or False.	True
	[STMT·ACT]	Is the following true of this segment? "the Frontoparietal network is among the top 3 networks by mean activation" Answer True or False.	True
Answer forms, open vocabulary	[SELECT·ACT]	Which network is ranked 2 by mean activation (1 = highest)?	the Somatomotor network has rank 2 by mean activation (1 = highest)
	[SELECT3·ACT]	Which 3 networks are among the top 3 by mean activation?	the top 3 networks by mean activation are the Visual network, the Ventral Attention network, the Frontoparietal network
	[ORDER3·ACT]	Order these three networks by mean activation (1 = highest), highest first: the Default Mode network, the Dorsal Attention network, the Ventral Attention network.	mean activation is ordered the Dorsal Attention network > the Default Mode network > the Ventral Attention network
	[SELECT·ACT]	Excluding the Visual network, which network has the highest mean activation?	Excluding the Visual network, the Frontoparietal network has the highest mean activation
Answer forms, open ended	[RANK7·ACT]	Rank all 7 brain networks by their mean activation during this segment, from highest to lowest.	the Frontoparietal network > the Visual network > the Default Mode network > the Somatomotor network > the Dorsal Attention network > the Ventral Attention network > the Limbic network
Order of states	[TSEQ]	This fMRI segment spans several task blocks in time. List, in temporal order, the cognitive states the subject went through — each is STORY (listening to a story) or MATH (solving math). Answer as a comma-separated sequence, e.g. STORY, MATH, MATH.	MATH, MATH, MATH, STORY
Odd-one-out, four hosts	[IDENT]	The following are 5 videos. All but one come from the same person. The order of the videos is arbitrary. [Identity] Which video comes from a DIFFERENT person than the others? Options: (A) Video 1 (B) Video 2 (C) Video 3 (D) Video 4 (E) Video 5. Answer with the option letter.	E

Table 30: Item types of the joint mixture by tag, with the question and answer of the first test item of each. The target-quantity and answer-format lines of the template restate what the question asks and the form of its answer.

Tag	Question	Answer
[TSEQ]	This video segment spans several task blocks in time. List, in temporal order, the cognitive states the subject went through — each is 2BK (a 2-back working-memory block) or 0BK (a 0-back matching block). Answer as a comma-separated sequence.	2BK, 0BK
[IDENT]	Videos 1 to 4 are fMRI recordings. Three of them come from the same person and one comes from a different person. Which video comes from the different person?	Video 3 comes from a different person than Videos 1, 2, and 4
[STMT·ACT]	Is the following true of this video segment? "the Limbic network has higher mean activation than the Frontoparietal network" Answer True or False.	True
[SELECT·ACT]	Excluding the Limbic network, which network has the highest mean activation in this video segment?	Excluding the Limbic network, the Ventral Attention network has the highest mean activation
[RANK7·ACT]	Rank all seven brain networks by their mean activation in this video segment, from highest to lowest.	the Visual network > the Dorsal Attention network > ... > the Somatomotor network
[STMT·TREND]	Is the following true of this video segment? "the Somatomotor network shows a more strongly rising temporal trend than the Ventral Attention network" Answer True or False.	True
[SELECT·TREND]	Excluding the Somatomotor network, which network shows the most strongly rising temporal trend in this video segment?	Excluding the Somatomotor network, the Ventral Attention network shows the most strongly rising temporal trend
[RANK7·TREND]	Rank all seven brain networks by their temporal trend in this video segment, from highest to lowest.	the Default Mode network > the Limbic network > ... > the Dorsal Attention network
[FC·SEGMENT]	In this video segment, which network co-activates most with the Somatomotor network? Answer as one statement wrapped in <Answer>...</Answer>.	the Ventral Attention network co-activates most with the Somatomotor network
[FC·INDIVIDUAL]	Videos 1 to 8 are task blocks from one person. Relative to that person's overall level of co-activation, and compared with most people, is each network pair below more or less strongly coupled in this person? (three pairs)	compared with most people, this person shows stronger coupling between ... and weaker coupling between ...
[TERCILE·DIFFICULTY]	Videos 1 to 8 come from one person during 2-back working-memory blocks. Combining speed and accuracy: did this person find the 2-back task among the hardest third or the easiest third of participants? Answer with "hardest third" or "easiest third".	easiest third
[TERCILE·SPEED]	Videos 1 to 8 come from one person during 2-back working-memory blocks. Considering only reaction time, not accuracy: was this person among the slowest third or the fastest third of participants? Answer with "slowest third" or "fastest third".	slowest third
[TERCILE·ACCURACY]	Videos 1 to 8 come from one person during 2-back working-memory blocks. Considering only how many 2-back trials were answered correctly, not speed: was this person among the most accurate third or the least accurate third of participants? Answer with "most accurate third" or "least accurate third".	most accurate third
[RANK3·COGNITION]	Rank the three people by their crystallized intelligence score, highest first. Answer as a single chain separated by " > ".	Person A > Person C > Person B

L RELEASE AND REPRODUCIBILITY

The release has three parts. The *item specification* gives, for every item, the question, the answer, its position in the design space, the recording segment it refers to and the provenance of its label, namely which quantity was computed over which parcels and frames under which normalization. The *derived signals* are the per-parcel time series the items refer to. The *item generator* reproduces the corpus deterministically from those signals with fixed seeds and a checksum manifest, together with a small fixture of derived values against which a user can check a pipeline before regenerating anything. Evaluation and training items are distributed as separate files with the split index over anonymized participant identifiers, so that a reported number refers to a fixed set. The source release permits redistributing derived data under its open-access terms, which the release inherits, so evaluating a model on the derived signals requires no access to the source, whereas a model that reads volumes, as RIVET does, needs the source recordings. Splits separate participants but not families, so relatives in the source release can fall in different splits. We do not distribute the raw imaging, the cached embeddings, the original participant identifiers or any mapping to restricted variables, and participant-level labels are discretized into tertiles rather than redistributed as scores. A label is reproducible only if every parameter of Table 31 agrees, so a user with access to the source release can rederive the signals and check them against these values before comparing anything downstream.

Table 31: Parameters that fix every label.

Parameter	Value
Source release	Human Connectome Project Young Adult, open access
Preprocessing	minimally preprocessed volumetric release
Runs	both runs of each task, one per phase-encoding direction (LR, RL)
Repetition time	0.72 s
Grid	2 mm MNI, $91 \times 109 \times 91$
Parcellation	Schaefer 100 parcels, 7 networks (Schaefer et al., 2018)
Composition	to the 7 networks of Yeo et al. (2011), parcel-size weighted
Segment boundaries	condition blocks read from the recorded event files
Hemodynamic shift	4.0 s added to event times before conversion to frames, identical for every paradigm, with no extension at block ends
Normalization	network signals standardized along time within a run
Splits	over participants, 70/10/20, disjoint, not by family
Seeds	three, permuting participants and leaving segments unchanged