Many-Two-One: Diverse Representations Across Visual Pathways Emerge from A Single Objective

Yingtian Tang[1], Abdulkadir Gokce[1], Khaled Jedoui Al-Karkari[2], Daniel Yamins[2], and Martin Schrimpf[1]

[1]École Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland

[2]Stanford University, Stanford, CA 94305, USA

How the human brain supports diverse behaviours has been debated for decades. The canonical view divides visual processing into distinct "what" and "where/how" streams – however, their origin and independence remain contested. Here, using deep neural network models that accurately predict hours of brain recordings, we computationally characterise how cortex processes dynamic vision. Despite the diversity of cortical regions and thereby supported tasks, we identify two fundamental computations that explain neural activity across visual cortex: object and appearance-free motion recognition. Strikingly, a *single* objective underlies both: these inherent computations in the brain emerge from optimising for understanding world dynamics, and their arrangement is highly distributed and smooth across cortex rather than strictly aligning with the two visual streams. Our results suggest that the human brain's ability to integrate complex information across seemingly distinct representational pathways may originate from the single goal of modelling the world.

NeuroAI | Neuroscience | Human Behaviour | Dynamic Vision | Visual Cortex | Temporal Integration | Task-Driven Modelling

Correspondence: *martin.schrimpf@epfl.ch*

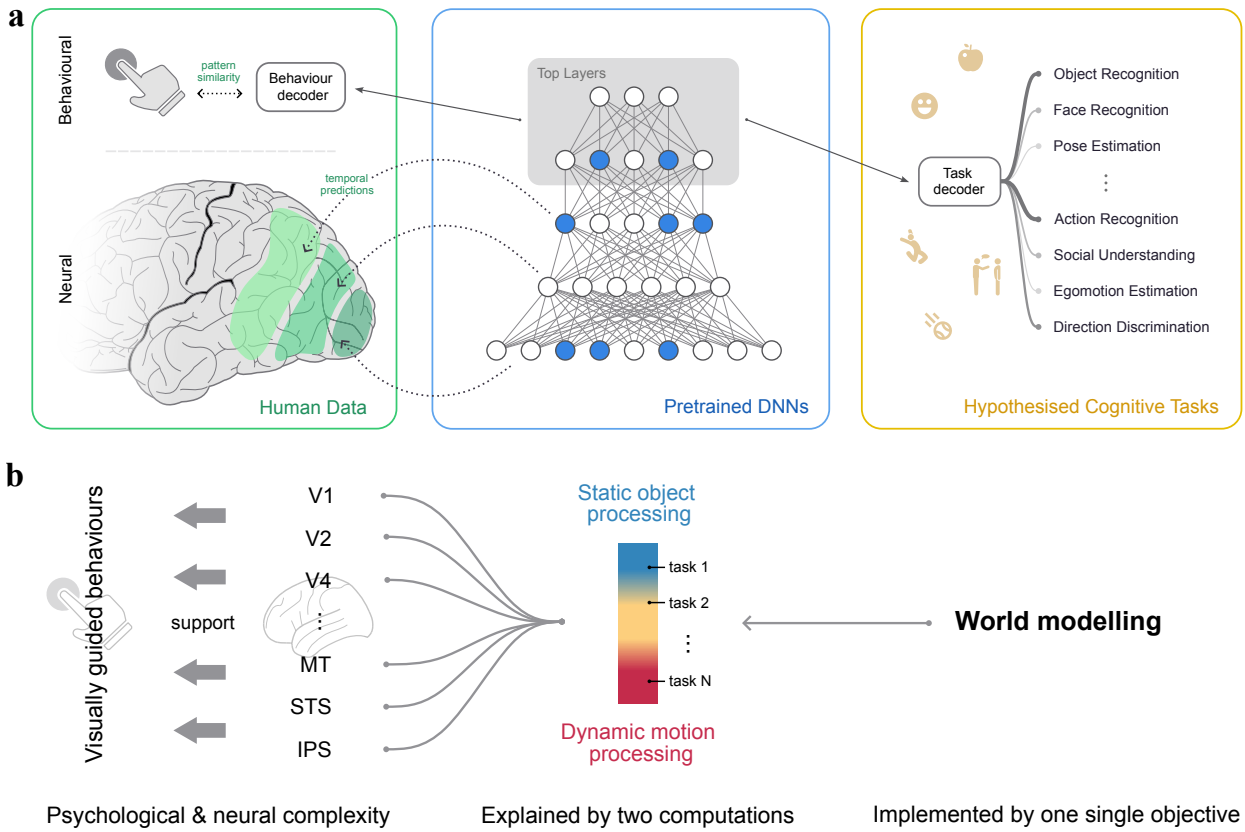


Figure 1. Multi-task computational framework for brain modelling. a. Models bridge neural, behavioural, and machine learning datasets: We test the alignment of representations from stimulus-computable models (*Middle*) to dynamic neural signals (fMRI) and behavioural recordings (*Left*), and connect brain alignment to task performance on diverse cognitive tasks (*Right*), investigating hypotheses of which functional objectives shape representations in the brain. **b. Core claim:** We find that the psychological and neural complexity in diverse brain areas is explained by two core tasks, which can be further reduced to a single overarching objective. Representations across various brain regions supporting many visually guided behaviours arise from two core tasks—object recognition and appearance-free motion recognition—which themselves originate from a single 'world modelling' objective.

Introduction

In order to operate in an ever-changing environment, humans parse dynamic sensory information. From interpreting the hunt of a predator or the escape route of prey, to navigating cars and pedestrians in modern city environments, the brain continuously integrates a wealth of visual sensory input. Visual information is processed by the visual cortex [112, 43, 39], classically described as comprising ventral ("what") and dorsal ("where/how") streams [94]. Within these streams, functionally specialised regions emerge – such as face- [51], scene- [27], or motion-selective [5] areas – supporting the view that the visual system is composed of modular components optimised for specific tasks. However, these functions are not isolated. Increasing evidence [70, 38, 37, 113, 130, 69, 48] suggests that regions across both streams interact dynamically, and their boundaries are graded and smooth rather than discrete. Complex tasks [2, 11] like interpreting dynamic social cues further involve cooperation across streams and regions, indicating that visual processing operates within a broader, distributed network. Despite a coarse localisation of brain regions, a unified understanding of the representations and mechanisms involved in dynamic visual processing remains elusive.

Advances in AI have led to artificial neural networks (ANNs) that achieve human-level performance, with internal machinery abstractly resembling biological neurons. The recent use of task-optimised AI models in systems neuroscience has revolutionised our ability to explain neural representations in distinct regions throughout the brain [139, 67, 119, 65, 101, 98, 68, 120, 118, 121, 18]. These models are optimised for computational tasks without fitting to brain data, and yet exhibit internal representations that are remarkably brain-like [138, 120, 118]. Linking task performance to neural alignment reveals evolutionary and developmental objectives that might have shaped brain representations. This task-driven modelling approach has established object categorisation as a key driver of static vision across the primate visual ventral stream [139, 119], with comparable findings in other systems, such as

auditory [78] and language processing [118, 18]. However, a crucial limitation to date has been the primary focus on static visual stimuli, and the investigation of individual pathways with individual objectives, without a coherent theory of the objective(s) shaping regions across visual cortex.

To address this gap, we here establish a normative computational framework of how dynamic vision is processed across visual cortex in the service of different task demands. We assess the internal representations in 92 pretrained deep neural network models and evaluate their alignment to hours worth of brain recordings across five diverse datasets featuring dynamic stimuli such as human actions, moving objects, and narratives (Fig. 1a). In order to determine which objective(s) best explain the computational nature of brain-like representations, we introduce a novel *multi-task relevance* method that tests which combination of objectives uniquely explains selected brain representations.

Across all model families, we identify dynamic models as state-of-the-art in alignment to visual cortex activity and consequent behavioural choices, surpassing both static and traditional models, and accurately predicting neural activity on a second-by-second basis. These models are optimised via a single computational objective such as prediction across time and space, yet perform well on a multitude of downstreams tasks, and develop brain-like representations for ventral and dorsal streams – which were previously thought to be shaped by distinct task demands (Fig. 1b). Our representational analysis reveals that model features support *both* static object recognition *and* dynamic motion recognition, and that these features distinctively but smoothly map onto the streams in visual cortex. Assessing individual voxels and their behavioural relevance, our computational approach reveals hierarchical processing with increasing feature and motion complexity that unifies previous findings and computationally characterises the pathways supporting action understanding. Overall, our integrative approach unifies previously disparate experimental datasets and findings into a coherent computational framework, advancing our understanding of dynamic integration mechanisms in the visual brain.

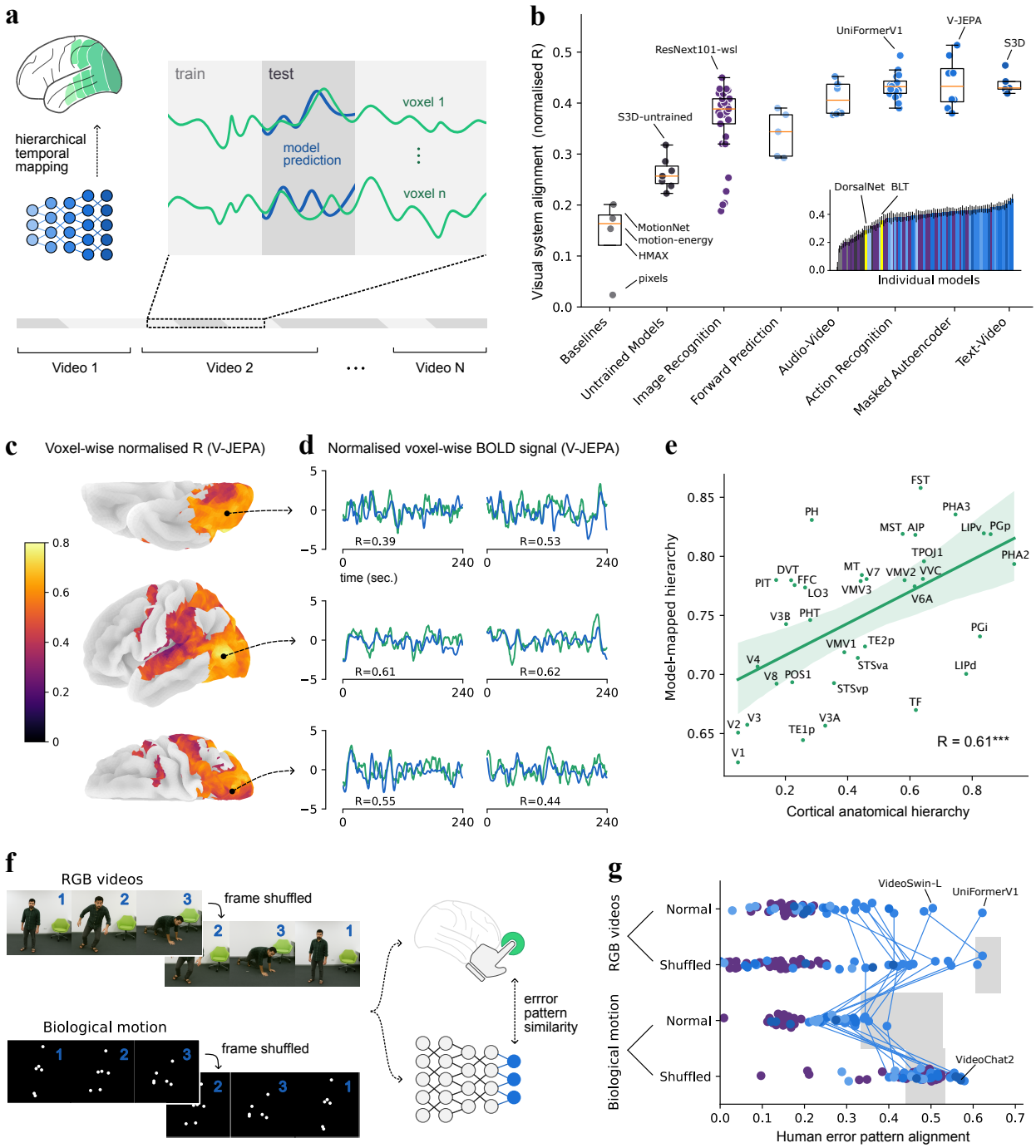


Figure 2. Predictive models of dynamic vision. a. Neural alignment: we map layer-wise representations from stimulus-computable models to dynamic neural signals (fMRI voxel time series) and evaluate similarity across time. **b. Dynamic models are most brain-like:** many dynamic task-optimised models (blue shades) outperform baselines and static models (purple). Inset: individual model scores (details in SI. A.3). **c. Alignment scores of the best model:** normalised fMRI predictivity of V-JEPA on a 3D brain atlas from ventral, symmetrically flipped lateral, and dorsal views. Showing the right hemisphere, similar patterns are observed in the left hemisphere (Fig. S4). **d. Second-by-second predictions:** predicted fMRI responses from V-JEPA for held-out 240-second video clips (blue) compared to ground truth (green) from the three views in panel **c**. **e. Model hierarchy matches cortical hierarchy:** layer assignments averaged across the 20 best-aligned dynamic models strongly correlate with the anatomical hierarchy derived from *Rolls et al.* [114] ($p < 0.001$; regions from *Glasser et al.* [45]; Methods B.3 SI A.6). **f. Behavioural alignment:** we compare action recognition judgements on RGB or biological motion videos (with either shuffled or intact frame order) between humans and models, based on recognition error patterns. **g. Models predict human video error patterns:** model alignment to human errors (dots). Grey shaded areas show inter-subject consistency (95% CI). Lines connect individual top models across conditions.

Results

At the core of our analyses is the evaluation of alignment between artificial neural network models (ANNs) and brain recordings. We use 5 fMRI brain datasets of video viewing, totalling 35.7 subject hours and approximately 2 hours of unique visual input, as well as an additional behavioural dataset of action understanding. The candidate models comprise a repertoire of ANNs pretrained on a diverse set of objectives, including state-of-the-art deep learning approaches spanning image classification to temporal representation learning, such as masked video modelling. We assess model-to-brain alignment using a regression-based metric *across time*, which applies a time-invariant linear mapping from model units to brain voxels (Fig. 2a; Methods B.3).

To investigate which computational objectives shape brain-like representations, we further evaluate the performance of ANNs on 10 ecological tasks (Fig. 1a right). These include object recognition—long a focus of prior work [139, 119]—as well as dynamic tasks like action recognition and social understanding. We evaluate voxel-level task processing and provide evidence that a single functional objective underlies the formation of representations in the visual system.

Task-optimised dynamic neural networks predict cortical dynamics

We first test if models can predict the brain activity elicited by video viewing. Candidate ANNs are sourced from six pretraining classes: image recognition, forward prediction, audio-video correspondence, action recognition, masked autoencoding, and text-video models. All task-optimised model families accurately predict cortical dynamics, up to 0.51 correlation with visually selective voxels across the brain (Fig. 2b, coloured dots).

Task-optimised models outperform baseline and untrained models. Compared to baselines—such as a classic neuroscience model [111, HMAX], spatiotemporal filtering [99, motion-energy], and simple dynamic models [93, 110, MotionNet]), models optimised for image or video tasks achieve significantly higher brain alignment (mean normalised correlation 0.40 vs. best-performing baseline 0.20; $p < 0.001$). Trained models outperform their untrained counterparts ($p = 0.011$; SI A.2), which themselves outperform all baseline models ($p < 0.001$, Fig. 2b). This suggests that a combination of structural priors and task optimisation is critical for brain-like representations. Notably, these models are optimised for a single objective and yet learn representations that predict activity across the visual cortex.

Dynamic models outperform static models. Many dynamic model families show strong brain alignment, significantly outperforming static object recognition models (action recognition: $p < 0.001$; masked autoencoders: $p = 0.015$; text-video models: $p = 0.002$; Fig. 2b).

The most brain-aligned model identified here, V-JEPA [4], learns representations by reconstructing missing features in video segments – achieving a normalised correlation to brain data of 0.51, with reliable predictions across ventral, lateral, and dorsal areas (Fig. 2c,d). Predictivity decreases in regions adjacent to the auditory cortex (SI A.1&A.5), which can be attributed to the unimodal nature of the model.

Model-derived hierarchy matches empirical connectivity. During alignment, we map different layers of the neural network to distinct brain regions based solely on functional correspondence, without using anatomical constraints. Remarkably, the cortical hierarchy emerges from this functional mapping. We quantify this hierarchy by averaging the relative layer positions assigned to each brain region across the 20 best-aligned dynamic models, yielding high split-half consistency ($r = 0.84$). Indeed, this model hierarchy aligns closely ($R = 0.61$, $p < 0.001$, Fig. 2e) with cortical connectivity in the brain [114]. For example, the layer best aligned with V1 appears early in the model, while higher-level regions such as ventral VVC and dorsal MST correspond to later layers (SI A.6).

Dynamic models make human-like decisions in action recognition. Beyond a characterisation of the neural substrate underlying dynamic vision, we link our findings to visual behaviours. Using a large-scale dataset of indoor actions [53], we compare action recognition decisions from models and humans across two presentation formats (RGB and biological motion) and two temporal ordering (normal and shuffled) conditions (Fig. 2f; Methods B.4).

To test behavioural alignment, we examine whether the models exhibit similar error patterns to those of humans (with accuracy results reported in SI A.7). Specifically, we assess the similarity of object-level confusion between humans and models, defined as the probability of misrecognising one action as another [106]. Models closely align with human error patterns, with many models approaching the inter-subject consistency observed among human participants (consistency is undefined for the first condition due to insufficient errors in RGB video trials; Fig. 2g; Methods B.4).

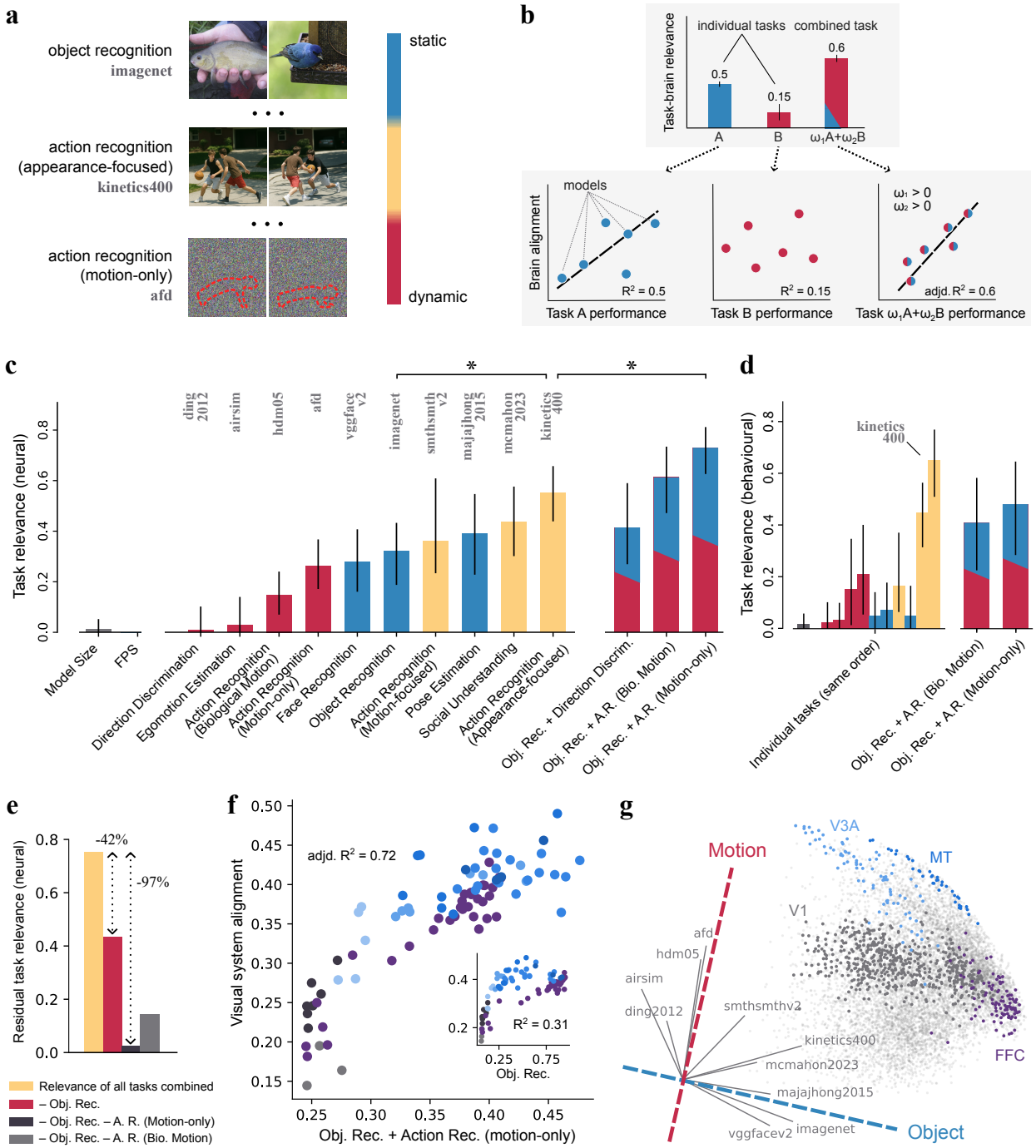


Figure 3. Object and appearance-free motion recognition jointly explain representational alignment. a. Diversity of tasks: We evaluate models' cognitive capacities on a broad range of tasks (complete task suite in SI B.1). Static tasks test for image-based object recognition, while purely dynamic tasks are devoid of appearance cues (motion highlighted with dashed red outlines). **b. Testing cognitive tasks for their relevance to brain function:** To test the influence of multiple tasks on shaping brain-like representations, we assess individual and joint *task relevance* as the correlation between models' task performance and their brain alignment (assessed via R^2; adjusted R^2 for multivariate regressions). **c. Neural task relevance:** Task relevance for brain alignment of controls (grey), individual tasks (single colours), and joint tasks (mixed colours). Action recognition is individually most predictive; combined object and motion recognition explain most variance overall (SI B.2). **d. Behavioural task relevance:** same as in **c**, but evaluated with respect to behavioural alignment. **e. Residual relevance:** neural relevance of all tasks combined after regressing out the contribution of certain tasks. Joint object and motion recognition explain away most of the relevance. **f. Object and motion task performance predicts brain alignment:** correlation between object + motion-only action recognition and brain alignment. The correlation of object recognition alone (inset) is significantly lower ($p < 0.01$). Model groups coloured as in Fig. 2b. Regional results in SI B.4 **g. Low-dimensional task space for brain-like representations:** the 2D eigenspace captures 91% of the variance in task relevance across visually driven voxels in the brain. This space is spanned by two orthogonal basis directions: object and motion recognition.

Two principal computations explain model-brain alignment

The strong alignment of task-optimised neural networks with brain activity prompts the question: *what underlies this alignment?* Evidence from recent studies suggests that brain alignment in individual streams is closely tied to task performance, particularly on ecologically relevant behaviours [139, 119, 65, 118, 97, 88].

Unlike prior work, our study analyses the entire visual system and thus calls for evaluating a broad range of candidate cognitive tasks and their associations with brain regions. Our framework is scalable: via stimulus-computable models, we can directly test arbitrarily many task and stimulus conditions, focusing our investigation on the most striking effects. We examine ten candidate cognitive tasks spanning static, dynamic, and hybrid information processing—including object recognition, action recognition, and social understanding (see Figs. 3a and S10 for sample stimuli). To also test the capacity of processing appearance-free motion (*i.e.,* dynamics without colour or texture cues), we include three tasks: direction discrimination, motion-only action recognition, and biological motion recognition. To assess a model's capacity to support a task independently of its pretraining objective, we evaluate performance using a linear decoding probe (Methods B.5).

Naturalistic hybrid tasks individually yield highest relevance. We identify tasks where models' performance is most correlated with their brain alignment ("*task relevance*"; Fig. 3b). A high correlation suggests that optimising for the task leads to brain-like representations. Appearance-focused action recognition (`kinetics400`) emerges as the single most *relevant* task predicting models' overall brain alignment ($R^2 = 55\%$; Fig. 3c). It significantly surpasses object recognition (`imagenet2012`; $R^2 = 32\%$) which has been closely tied to the primate visual ventral stream [139, 119]. Other tasks that involve processing naturalistic dynamics also show high *task relevance*, including social understanding and motion-focused action recognition. Controls such as model size and frames-per-second (FPS) are not predictive of brain alignment.

Joint task relevance to characterise hybrid processing. Following the intuition that the brain's neural code might be driven by multiple cognitive demands, we sought to explain model-to-brain alignment as a combination of task performances. We introduce a novel *joint task relevance*, which weighs each task to maximise correlation with brain alignment (Fig. 3b; Methods B.6). Applied to all candidate tasks, object and appearance-free motion recognition emerge as the combination that explains the most variance in brain alignment ($adjd.\ R^2 = 73\%$, Fig. 3c right), significantly outperforming any individual task ($p < 0.05$, FDR-corrected; Fig. 3f). Importantly, this combination explains away all other tasks (Fig. 3e), while object recognition alone only accounts for 42% of the contribution (SI B.2). The same combination also shows strong relevance to behavioural alignment ($adjd.\ R^2 = 46\%$, Fig. 3d; SI B.7).

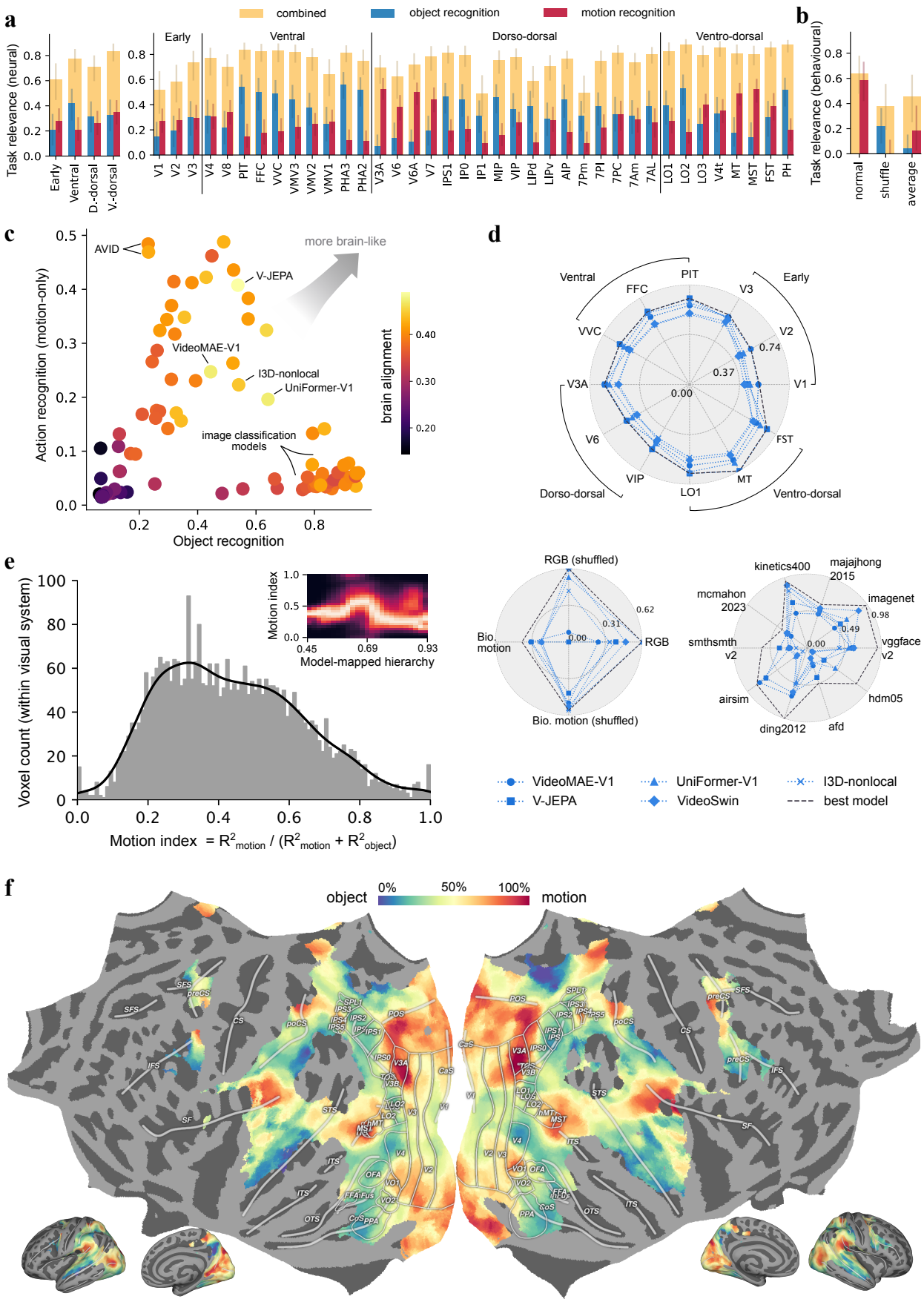
Two tasks explain the complexity of neural and behavioural representations. Since object and appearance-free motion recognition together explain nearly all the variance in neural alignment, we asked: *do these two components span the representational organisation across the visual cortex?* We computed voxel-level task relevance across the 10 cognitive tasks, yielding a 10-dimensional relevance profile for each voxel. Fig. 3g presents the two-dimensional eigenspace projection of these profiles, capturing 91% of the total variance (44% and 47% on the two principal dimensions). The projection can be spanned by two orthogonal axes for motion and object processing, with motion- and object-focused tasks near their respective directions, and naturalistic dynamic tasks centrally, reflecting hybrid processing (grey arrows in Fig. 3g). Patterns also emerge for brain regions: dorsal regions such as V3A and MT show strong motion bias, ventral regions like FFC favour object processing, and early visual areas (*e.g.,* V1) show a balanced profile, integrating both components.

We examine regions across the ventral and dorsal visual streams, including the dorso-dorsal and ventro-dorsal subdivisions. The combined task explains the majority of the variance in brain alignment across nearly all regions (Fig. 4a), with distinct biases toward either object or motion processing in each region. For behavioural relevance, motion recognition underlies decisions for normal videos, whereas object processing dominates with temporally shuffled videos, reflecting two distinct processing mechanisms under these temporal conditions (Fig. 4b).

Brain-like representations emerge from single objectives

Representations across visual cortex can be explained by two tasks: object recognition and appearance-free motion recognition. We next examine how these two computations are distributed across the brain and present evidence that brain-like representations are ultimately implemented by a single unified objective.

Distributed computations across visual brain regions. For all visual brain regions, hybrid processing dominates, with a smooth transition of preferential object and motion processing along visual streams (Fig. 4a right). For example, mid-level regions in both dorsal subdivisions, such as V6A and MT, are biased toward motion processing, but both converge to mixed processing at later stages. Processing of object and motion information is nearly evenly


a
combined
object recognition
motion recognition
Task relevance (neural)
Early
Ventral
Dorso-dorsal
Ventro-dorsal
b
Task relevance (behavioural)
normal
shuffle
average
c
AVID
V-JEPA
VideoMAE-V1
I3D-nonlocal
UniFormer-V1
image classification models
more brain-like
brain alignment
Action recognition (motion-only)
Object recognition
d
PIT
V3
V2
V1
FST
MT
LO1
VIP
V6
V3A
VVC
FFC
Ventral
Early
Dorso-dorsal
Ventro-dorsal
e
Voxel count (within visual system)
Motion index
Model-mapped hierarchy
Motion index = R²motion / (R²motion + R²object)
RGB (shuffled)
Bio. motion
RGB
Bio. motion (shuffled)
kinetics400
majajhong 2015
imagenet
vggface v2
hdm05
afd
ding2012
airsim
smthsmth v2
mcmahon 2023
VideoMAE-V1
V-JEPA
UniFormer-V1
VideoSwin
I3D-nonlocal
best model
f
object
0%
50%
100%
motion

mixed across visual pathways although the ventral stream shows a notable bias toward object recognition (Fig. 4a left). While the classic view of distinct pathways with distinct functions would have suggested strongly biased functional preferences such as object recognition in the ventral pathway and motion recognition in the dorsal pathway, the ubiquitous dominance of hybrid processing and smooth gradient across regions suggest representations might instead be shaped by a single unified objective.

Single-objective models achieve brain-like representational diversity. Models pre-trained on single objectives demonstrate high alignment with the entire visual system (Fig. 2b). The best models achieve high brain alignment not by improving only object or motion recognition, but via an objective that implicitly optimises both (Fig. 4c). Indeed, across visual regions, behaviours, and cognitive tasks, several single-objective models exhibit consistently high alignment and performance (Fig. 4d). Brain-like representations thus can in fact emerge from a unified objective such as action recognition or masked autoencoding, which all implement a form of world modelling.

Absence of multimodality in voxel task relevances suggests a unified functional objective. Beyond regional divisions (Fig. 4a), we investigate the detailed differences in motion and object preferences on a voxel-by-voxel basis. We define a *motion index* to quantify the proportion of motion recognition relevance relative to object recognition relevance. The distribution of motion index values is smooth across voxels (Fig. 4e): rather than exhibiting a bimodal structure indicative of distinct functional clusters, the distribution is approximately unimodal ($p = 0.993$, Hartigan's dip test [84]). This pattern is consistent across the visual hierarchy (inset in Fig.4e; Fig. S25).
Throughout cortex, motion and object preferences across voxels vary smoothly (Fig. 4f), suggesting that these functions are not strictly segregated but rather closely intertwined. Functional organisation in the cortex following a topographic map is a well-established phenomenon [59, 60, 54, 134, 87] and helps explain the bifurcation of specialisation despite a single overarching objective. The distribution of voxel-by-voxel task relevances exhibits spatial clustering which follows a topographic organization – with smooth transitions throughout rather than a strict division into functional streams.

Task-based functional localisation of dynamic vision networks

The motion index map (Fig. 4f) reveals cortical structures corresponding to prior neuroscience findings on object and motion processing in both visual streams. Here, we interpret these structures and link them to behaviour, offering strong validation for our methodology.

Unification of fragmented experimental findings. The division between object and motion processing in the motion index map reveals a functional organisation that aligns with prior neuroscience findings, offering a unifying perspective on fragmented observations across visual streams. Fig. 5a shows that motion processing aligns with visual cortex eccentricity: peripheral regions such as MT and peripheral EVC are more involved in motion processing than foveal EVC, consistent with known retinotopic organisation [12, 135, 56]. Voxels with a bias toward object processing (motion index $< 45\%$) match regions identified in the shape processing network (Fig. 5b), including LO, PHAs, fusiform gyrus, and IPS [37, 30]. This finding is consistent with recent observations that image classification models explain activity not only in the ventral stream but also in dorsal areas [35] (Contrasting the task relevance of face and action recognition, our computational analysis also rediscovers the face processing network, Fig. S17). Voxels biased toward motion processing (motion index $> 55\%$) on the other hand correspond to regions responsive to optic flow and implicated in egomotion encoding (Fig. 5c) [16]. These regions include V1–V3 in early visual cortex; MT and MST in the ventro-dorsal stream; and dorso-dorsal areas such as V3A, V6, and VIP. They also

Figure 4 *(preceding page)*. Diverse visual brain representations originate from a single objective. a. Joint object and motion recognition are most relevant for all visual ROIs: task relevance of object recognition, motion recognition, and their combination to regions of interest (ROIs; SI B.3) along ventral and dorsal pathways. **b. Joint object and motion recognition are most relevant for behaviours across video conditions:** task relevance of action recognition behaviours in normal videos, temporally shuffled videos, and their average. **c. Single objectives jointly capture object and motion recognition for maximum brain alignment**: task performance on object recognition versus motion-only action recognition. Multiple single-objective models successfully implement both capacities, resulting in higher brain alignment. **d. Single objectives yield consistently high neural alignment, behavioural alignment, and task performance:** evaluation of models trained on single objectives across alignment and performance measures. Each connected line is one model. **e. Smooth distribution of motion index across the visual system:** motion index reflects the proportion of motion versus object processing within each voxel. The inset shows the distribution across the full visual hierarchy (filtered from Fig. S25). **f. Cortical map of motion index:** brain-wide motion index map (for voxels with at least 40% inter-subject consistency; Methods B.1). Hybrid processing and smooth transitions are observed in all visual streams.

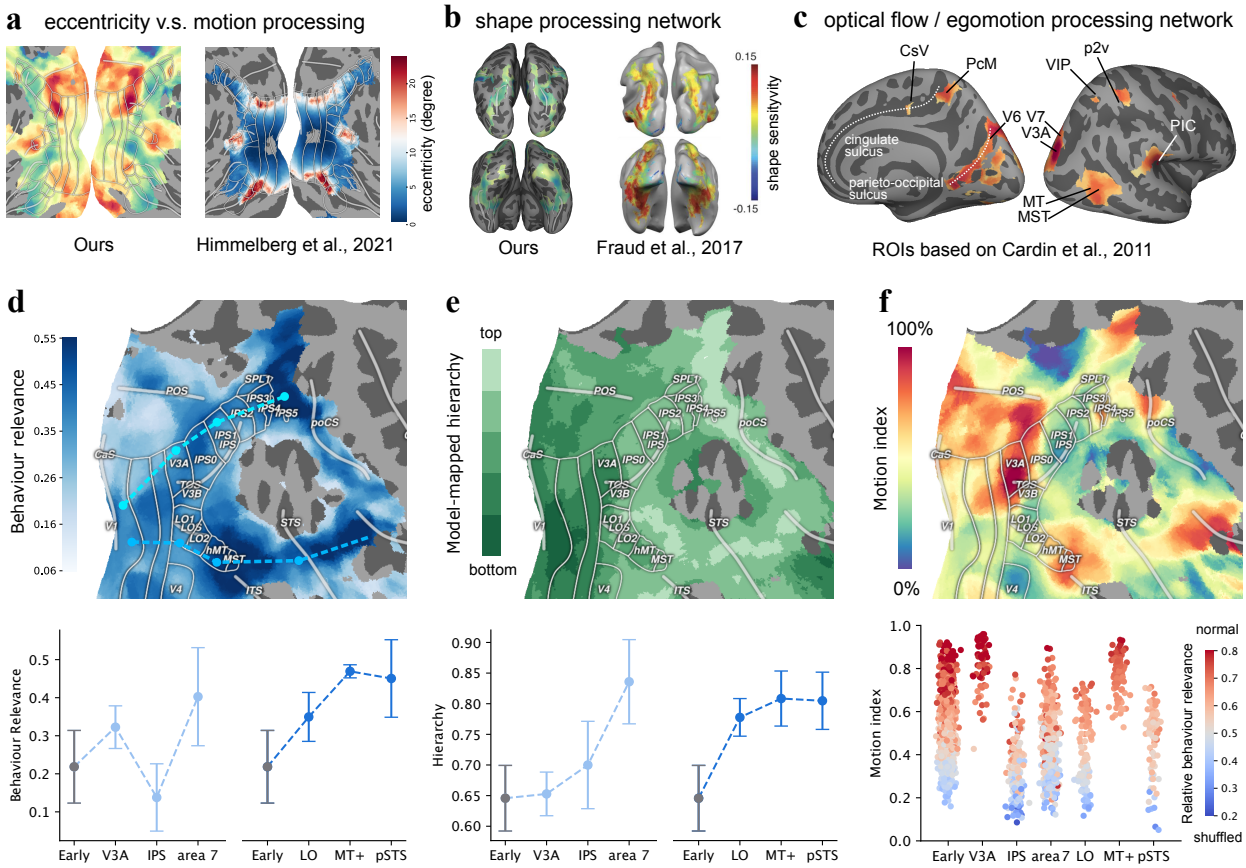


Figure 5. (a-c). Recapitulation of known brain structures: *a.* motion index in early visual cortex (EVC) compared to EVC eccentricity [56]; *b.* object-biased areas of our map ($< 45\%$ motion index) compared to the shape processing network [37]; *c.* motion-biased areas of our map ($> 55\%$ motion index) cover the optical flow/egomotion processing network [16]. **(d-f). Computational characterisation of action understanding pathways:** *d.* behavioural relevance increases along dorso-dorsal and ventro-dorsal streams (full map in Fig. S16). *e.* model-mapped hierarchy increases along dorso-dorsal and ventro-dorsal streams (full map in Fig. S6). *f.* hybrid processing occurs across many regions along the streams; a higher motion index corresponds to greater behavioural relevance under normal temporal video order, and a higher object index for shuffled conditions.

include the parieto-insular cortex (PIC), located between the inferior parietal lobule and the auditory cortex [36]. The motion features encoded across these regions increase in complexity along the visual hierarchy (SI B.5). The spatial functional gradient suggests potential pathways linking these motion-sensitive regions, which we explore next.

Emerging pathways of action understanding. The motion index map suggests processing that starts from initial broad functional processing in early visual cortex and then, through two subdivisions of the dorsal stream, converges to a high-level hybrid nature of representations. We here link this functional analysis to behavioural relevance and computational hierarchy (Fig. 5d,e) for a detailed characterisation of action understanding pathways. Tracing regions along the dorso-dorsal (V3A, IPS, area 7) and ventro-dorsal (LO, MT+, pSTS) streams (blue dash lines in Fig. 5d) reveals increasing behavioural relevance (Fig. 5d, bottom) and computational hierarchy (Fig. 5e). With hybrid object and motion processing across regions, intra-regional subdivisions seem specialised for different conditions of temporal integration: voxels with a higher motion processing index are more behaviourally relevant for normal videos while voxels with an object bias relate more closely to behavioural choices for shuffled videos (Fig. 5f). On the ventro-dorsal stream, the relevant regions seem to extend beyond pSTS and go into the posterior Sylvian fissure, which is adjacent to the posterior insula cortex (PIC). While previous studies have examined this region in the context of egomotion and vestibular processing [36, 104, 16], its potential role within an action understanding pathway as suggested here has been largely overlooked [112, 114, 9, 136, 20].

Discussion

Summary of key results

Our results show that specific artificial neural network models can predict neural and behavioural responses to dynamic visual input with high accuracy: The best models are dynamic in nature, hierarchically organised in a way that resembles the cortical hierarchy, and optimised to understand the visual world. Dynamic visual representations are computationally organised in a space spanned by two principal axes, which are respectively in support of object and appearance-free motion recognition. While these two tasks preferentially map onto classic ventral and dorsal streams, computations throughout visual brain regions are fundamentally distributed and hybrid. Our task-based functional localisation shows that the organisation of representations on tissue follows a smooth interpolation between object and motion processing, exhibiting an overall unimodal distribution. Furthermore, models optimised for a single objective yield representations with high alignment to all brain regions, reproduce key experimental results, and exhibit consistently high behavioural accuracy. Together, these findings point to a unified functional objective shaping the visual system.

Cortical representations are driven by a single objective

Building on decades of research in visual neuroscience and connecting to advances in AI allows us to address the fundamental question of *how* representations are shaped in the brain. Partial accounts of visual cortex representations had pointed to object recognition in areas of the visual ventral stream [139, 119, 14, 72, 68, 98, 62, 120, 19], with few studies on non-ventral areas [93, 34]. We here aim to provide a holistic view of human dynamic vision by identifying models and computational principles for the entire visual cortex. Two core findings emerged, both supporting the idea that the human visual system is optimised for a single 'world understanding' objective.

First, optimising for a single objective yields the most brain-like models. The exact details of the optimisation procedure vary – from predicting masked parts in a video to recognising actions and events – but all attempt to understand the dynamics in the visual world. After training, the emergent representations in these models are highly diverse: they align across areas from early visual cortex through mid-level areas all the way to high-level visual areas, such as FFC, MT, and VIP. Brain-like representations also support various action recognition behaviours and high performance across all tasks we tested.

Second, representations in visual cortex are organised in a smooth and hybrid manner. When we analysed the task relevance of visual regions, we found that a hybrid combination of object and motion recognition is most explanatory and explains away all other tasks; a similar pattern holds for visual behaviours. However, the computational organisation of voxels across visual cortex is smooth: the vast majority of voxels support both object and motion recognition, and the overall distribution is unimodal across the visual hierarchy. If representations were shaped by multiple objectives, we would expect a multimodal distribution that is centred around two or more tasks with many visual regions being entirely dominated by one task; but we only found evidence to the contrary such that voxels' task relevance varies smoothly throughout.

An intriguing possibility is therefore that both the human visual system and successful vision models are unilaterally optimised to understand the visual world. For building more brain-like models, it offers a clear roadmap: modelling the dynamic world is key, and could provide a better scaling trajectory for improving alignment than the saturating object-recognition-only approach [46]. For understanding visual processing, the emergence of distinct computations from a single objective addresses the longstanding 'binding problem' [117]: rather than requiring an explicit integration of distinct features processed in two independent streams, our work suggests that representations throughout visual cortex are inherently hybrid already.

Organisation of representations into pathways

Beyond the origin of visual representations, we also investigated the resulting organisation. The classic view of the visual system is its organisation into ventral and dorsal pathways [49, 92]. Under this perspective, the two streams are typically thought of as disparate systems that perform distinctive roles: the visual ventral ("what") stream extracts object identity and form information for tasks in the space of object recognition, whereas the visual dorsal ("where/how") stream supports spatial perception for motion recognition tasks and to guide actions. The exact interpretation of whether the two streams are distinct modes of input or of output varies [94, 49] but their roles and therefore representations are typically seen as distinct.

To thoroughly test the computational nature of representations, we here used the neural network models to connect brain-like representations and their task relevance by asking: Which tasks are supported by which parcels in the brain? We developed a novel method which jointly considers the relevance of multiple tasks to cortical regions, moving beyond previous analyses on individual tasks [139, 65, 118, 97, 88].

Our multi-task relevance analysis offers a nuanced take on the two visual streams: While representations in ventral and dorsal areas *preferentially* support object over motion recognition and vice versa, computations are consistently hybrid. Our task-based functional localisation offers a new method for mapping the entire processing hierarchy (SI B.8) and further characterises these hybrid computations at voxel-level resolution. The localisation map reproduces a number of previously separate experimental findings on distributed computations across visual streams: Eccentricity [56] and motion processing correspond such that more peripheral regions of early visual cortex are more involved in motion than foveal parts; the shape processing network [37] corresponds with object-biased voxels; and regions implicated in optic flow and egomotion [16] are closely linked to voxels biased toward motion processing. These observations validate our mapping technique. Furthermore, by integrating them into a unified map, we reveal the continuum of object versus motion processing across the entire hierarchy of the visual system, beyond specific regions and experimental contexts.

A potential explanation for this hybrid and continuous organisation is that the unified functional objective of the brain drives feature encoding, while cortical wiring constraints shape how these features are spatially arranged, resulting in a topographically organised functional architecture [87, 26, 10, 66, 108, 23]. This would reconcile with the evidence against strict separation between visual streams [116, 113], and integrate emerging proposals of new pathways [136, 48, 69].

Indeed, our map suggests that the ventro-dorsal stream supporting action understanding extends into regions near the posterior insular cortex (PIC), which are associated with multimodal vestibular-visual (PIC) and auditory-visual (human planum temporale, hPT) processing [21, 52, 109]. This observation aligns with recent studies on anatomical connectivity [52, 109] and underscores the multimodal nature of action understanding and, more broadly, world understanding.

Limitations and future directions

While this and related studies aim to understand the brain in computational terms, adequate brain recordings remain essential for robust findings. High-quality recordings are hard to come by, especially at a scale sufficient for model testing, and most datasets in the field are not collected with model comparisons in mind in the first place. To integrate the 35.7 hours of recordings from 5 different fMRI datasets, we chose to map all the recordings of different subjects onto the *MNI* standard space and focus on group-level analysis only. This likely leads to lower internal consistency for signals in higher-level areas that perform long-term integration and memory [115], such as the prefrontal cortex, medial temporal gyrus, and angular gyrus. The fMRI BOLD signal is also relatively coarse, especially in the time dimension, and prevents us from a finer-grained investigation on millisecond-resolution dynamics. High-resolution datasets at scale with high repetition counts would be ideal, and would likely require larger-scale experimental initiatives. To maximally discern between alternative model candidates, we believe including models in experimental design is a vital step [47, 58].

The discoveries from this work enable many exciting future directions. Due to the solid alignment of the best models, they can now be used to guide future experiments. For instance, past studies have shown that model-selected stimuli can control neural activity [6, 105, 131, 129]. The models identified here potentially enable such model-guided control for temporal stimuli. Further, the best models can be analysed to make sense of their inner workings. Interpreting the function of computational models is a lot more accessible and efficient than biological experiments due to the complete availability of structure, functional connections, and the ability to perform targeted ablations.

The best models identified in this work do not hit the estimated noise ceiling of the data. How can we further improve models of dynamic vision? With the hierarchical motion network culminating in areas associated with multimodal processing, including additional sensory modalities in computational models might be a required next step. It is conceivable that a world modelling objective underlies not just visual cortex but sensory cortices more broadly and the division between "domain-specific" areas might be much more smooth, as we found here for areas in the visual system. In this case, building models that comprehensively understand the world across modalities might not only lead to a more unified model of sensory processing in the brain, but also improve the alignment to individual sensory modalities.

Conclusions

Taken together, our findings suggest that temporal AI vision models serve as viable hypotheses for how the human brain processes visual input in the service of understanding the world. This computational framework together with a novel multi-task relevance method explains how the visual system might be optimised for a single objective — and yet organises its representations along principal axes of object and motion recognition. These core computations connect well to the classic two-stream view of vision, but suggest a smooth and inherently hybrid organisation in the service of integrating dynamic information.

Methods

A. Experimental model and study participant details

We use fMRI recordings from five different studies, three [113, 7, 64] of which use movies lasting 6 to 10 minutes as stimuli, and two [89, 73] use short video clips of duration 3 seconds as stimuli. We also use macaque electrophysiological recordings from four visual areas, collected across two separate datasets with distinct image stimuli comprising objects, textures, and noise patterns [140, 86] (SI A.4). Additionally, we analyse a behavioural dataset involving video viewing and action recognition [53], in which human psychophysics experiments were conducted on Amazon Mechanical Turk. The experimental background can be found in the respective publications.

B. Method details

B.1. Neural and behavioural datasets.

Neural datasets of movie/video watching. We analyse a battery of movie- or video-watching datasets for studying the temporal neural dynamics in the neocortex with pseudo-natural visual stimuli. These datasets include a rich diversity of dynamic visual information and a high number of subject hours. *Berezutskaya et al. 2021* [7] contains fMRI recordings from 30 participants, all of whom performed a passive movie-watching task on a short 6.5-minute audiovisual film. *Keles et al. 2024* [64] includes fMRI and eye-tracking data from 24 participants during an 8-minute excerpt from a movie, after which they were asked to perform a recognition memory test. *Sava-Segal et al. 2023* [115] includes fMRI data from 43 subjects who watched four movies, all of which lasted for about 8 minutes and were followed by a set of tasks to probe the participants' interpretations and reactions. Although we focus only on modelling vision, the above three datasets all used audio-visual movie stimuli. In contrast, the following two use purely visual stimuli: in *Lahner et al. 2024* [73], 10 subjects viewed 1,000 3-second videos with 1-second intertrial intervals and with fixation. They were also presented with null trials 25% of the time during which they were instructed to press a button. Finally, *McMahon et al. (2024)* [89] have four participants who freely viewed 250 3-second videos featuring two actors (with or without interaction between the two actors). The statistics of these datasets are summarised in SI A.1.

MNI space registration and fsaverage extraction. In this study, all fMRI volumes are registered to the *MNI152NLin2009cAsym* template [28] (hereafter abbreviated as *MNI*) either by the original research work or by us. We refer the readers to the individual papers [7, 64, 73, 89, 115] for their detailed preprocessing steps. Then, the cortical signals in the *fsaverage5* mesh space are extracted for the subsequent analyses. For registered volumes, we use *nilearn* [1] to extract signals from the *fsaverage5* mesh space. For non-registered volumes, we directly run *fMRIprep* [28] with default settings and specify the output space to be *fsaverage5*. In the latter process, we also regress out the principal components of estimated nuance factors, retaining the top 95% variance. For all volumes in each dataset used in the study, we average across subjects and repetitions, focusing solely on the group-level analysis (SI A.1). For visualisation purposes (with pycortex [40]; Fig. 4f and Fig. 5), we upsample the data on the *fsaverage5* mesh to the *fsaverage* mesh using a 10-nearest neighbours (KNN) averaging procedure. The locations of the vertices from both meshes in the 3D *MNI* space were taken from *nilearn*.

Normalisation over time. To utilise data from all sources simultaneously, we simply standardise the fMRI signals to have a mean of zero and a variance of one across time, separately for each dataset. This normalisation disregards many differences between the datasets, such as the MRI device, original spatial resolution, thermal noise, *etc.*, making the cross-dataset predictions of fMRI signals particularly challenging. However, this coarse normalisation is sufficient to produce reasonable predictive performance from the models studied in the paper. We plan to investigate finer-grained normalisation in future work.

Internal consistency computation. The neural signals are inherently noisy. To fairly assess the predictions for these signals, we first estimate the internal consistency for each dataset at the group level. Specifically, denote $x_{v,t,s,r}$ as the fMRI BOLD signal at voxel v and time t for subject s at the r-th repetition, the internal consistency c_v for this voxel was computed as:

$$c_v = \mathbb{E}_{S_1,S_2}\left[\text{PearsonR}\left(\sum_{m\in S_1}\sum_{r\in R}\frac{x_{v,t,m,r}}{|S_1||R|}, \sum_{n\in S_2}\sum_{r\in R}\frac{x_{v,t,n,r}}{|S_2||R|}\right)\right] \tag{1}$$

where S_1 and S_2 are even splits of the whole set of subjects, R is the total number of repetitions, and *PearsonR* is the Pearson correlation over time t. The expectation is estimated using 10 random samples of the splits. This internal consistency is then corrected using the Spearman-Brown correction for two splits: $\hat{c}_v = \frac{2c_v}{1+c_v}$. This voxel-wise measure indicates the ceiling of Pearson correlation a predicted signal can have with the averaged group-level

signal. For all voxel-wise analyses, we use this measure to normalise model predictivity. We estimate the normalised ROI-wise (and whole-brain) predictivity by averaging the normalised voxel-wise predictivity within each ROI (or the whole brain). All of these analyses are restricted to the voxels with an internal consistency greater than 0.4. We choose this threshold due to the fact that it covers most of the action observation network [43], which has been implicated in processing dynamic event information while ensuring overall signal reliability.

Behavioural datasets. We use a subset of a recent large-scale behavioural dataset examining participants' action recognition with RGB and biological motion videos [53]. The biological motion presentation focuses exclusively on the key joints of the actors in the videos. This minimalist representation, when contrasted with the detailed RGB videos, provides valuable insights into the motion aspects of human action recognition. Furthermore, the videos are shown in two distinct temporal ordering conditions: normal, where the frames are displayed in their natural chronological sequence; and shuffled, in which the frames are randomly shuffled in time. These conditions allow us to examine the temporal integration over time. The human psychophysics experiments were performed on Amazon Mechanical Turk with 90 recruited participants. Participants were presented with a video (a few seconds long) and a list of ten options, then asked to complete a forced-choice 1-out-of-10 test to identify the action in the video. For the subset we analyse, each of the four conditions includes 600 trials, resulting in a total of 2,400 trials across all conditions. We refer readers to the original paper [53] for details on experimental design, data filtering, and stimuli selection. Beyond the stimuli used in the behavioural experiment, since this dataset itself is a subset of the NTU RGB+D 120 dataset [81], we are able to collect additional data with ground-truth labels across different conditions. These data are used for training the behavioural decoder trained on model outputs, as will be in Methods B.4. Sample images shown in Fig. 2 are drawn from the same dataset. We regenerated them using ChatGPT [102] to ensure anonymity.

B.2. Activation extraction from deep neural networks.

Model groups. We evaluate 92 pretrained deep learning models across many architectures (convolutional neural networks, attention-based transformer models, recurrent models, and their hybrids) that are chosen to represent a diverse range of training objective designs. They include: 1. baselines models based on traditional computational modelling, including "pixels", HMAX [111], motion-energy [99], and MotionNet [110]; 2. the forward-prediction models that were trained to predict the next frame in a video sequence, including PredRNN [107], ConvLSTM [122], SimVP [41], and MIM [133] as implemented by [124]; 3. the audio-video models trained to align audio representation with corresponding video representations, including SeLaVi [3] and AVID [95]; 4. the action recognition models trained to recognise human actions, including convolution-based [128, 137, 17, 32, 31, 8] and transformer-based models [29, 79, 83, 76, 75]; 5. the mask autoencoding models trained to reconstruct missing spatiotemporal patches in videos in the pixel or latent representation space, based on transformer models [127, 132, 4, 33]; 6. the text-video models trained on text-video contrastive matching (S3D-Howto100M [91, 90]) or various video-conditioned text generation tasks such as VideoChat [74] and Video-LLaVA [80]; 7. the image recognition models trained to classify objects, including those identified as among the most brain-like models (ResNext [85], CORnet [72], and VOneNet [22]), AlexNet trained on ImageNet or Stylized ImageNet [24, 71, 42], and 28 custom-trained models spanning various architectures [82, 13, 125, 55]. We also include two dynamic models from prior NeuroAI literature—DorsalNet [93] and BLT [126]—both of which have lower brain alignment relative to most dynamic models in our collection (yellow bars in Fig. 2b). This supports the strength and diversity of our model collection.

Activation extraction from video frames. To emulate humans watching videos, we need to process the video frames using deep learning models in a specific manner. We refer to it as the inferencing process. One ideal inference would be to enforce *causality*, meaning that the activation A_t at time t should not rely on the visual input I in the future, *e.g.,* I_{t+1} at time $t+1$. Therefore, the extraction of A_t incurs the processing of contextual visual inputs $I_{t-c,t}$ from time $t-c$ to time t (c is the duration of the model's temporal context and could be infinity), which is extremely costly for non-recurrent models. On the other hand, one can loosen the constraint on causality and allow the model to produce activations that depend both on the past and the future. Concretely, one specific inferencing method is:

$$A_{t-c,t} = M(I_{t-c,t})$$

where M is the model. In this way, the model processes a whole clip of videos and produces a temporal series of activations $A_{t-c,t}$. Then, $A_{t-c,t}$ is assumed to be evenly distributed across the timestamps to obtain the activation per timestamp. The next inferencing process will run on $I_{t,t+c}$. The process (what we call "block inferencing") is far less computationally costly than the aforementioned process (what we call "causal inferencing"), while the two give strongly correlated brain alignment scores (see Fig. S26). Therefore, we adopt "block" inferencing for models. Furthermore, since it is also computationally infeasible to process the videos of minutes at once, we limit the context duration to $c = 4$ seconds if the model allows flexible input duration. Otherwise, we set the context duration to the model's fixed input duration. Finally, each model processes the video frames with the same FPS when it was

pre-trained, regardless of the FPS of the videos themselves. For static image models, we fixed the processing FPS to be 25. When a video model processes an image, the image is treated as a static video for 1 second.

Recording and downsampling layer activations. We record the model activations from all layers for the subsequent analyses, which consume high disk storage. Therefore, we downsample activations from individual layers to a feature vector of fixed length 5000 (if the layer activation had more than 5000 values) using sparse random projections [77]. This projection has been shown to give highly correlated alignment scores with those with the original activations [46]. This also reduces the variance induced by different input sizes when performing the regression analysis.

B.3. Neural alignment computation.

Temporal mapping between model activations and fMRI recordings. We map the model activations to the fMRI recordings using a fixed linear transformation over time. This procedure embeds the assumption that the mapping between models and the brain should be invariant across time and datasets, thus making the results more generalisable. At the same time, it keeps the computation cost and overfitting effect at the minimal level. For learning this linear transformation, we need to temporally align the model activations and fMRI signals. As mentioned above, the model activations are generated per frame, with the same sampling FPS as when the model was originally pretrained. On the other hand, all of the fMRI datasets we used have 1 volume per second. To align the model activations to the same temporal resolution as the fMRI has, we average the activations with the 1-second time windows. Furthermore, to simulate the fMRI sampling process, we convolute a hemodynamic function implemented by *nilearn* (originally from SPM, with 5 seconds as the peaking time parameter) to the timeseries of model activations. This method produces better fMRI alignment in our preliminary small-scale tests (see Fig. S27).

Cross-validated regression for brain alignment. The brain alignment is defined using correlation over time between the model predictions and the normalised fMRI signals. The predictions are generated from regression, and the whole process is cross-validated for estimating variance. As shown in Fig. 2a, the videos from all datasets are stacked along the time dimension. We segment this long sequence into 15-second video clips and split them into training, testing, and validation sets $(80\%, 10\%, 10\%)$. RidgeCV regression [103] is used to regress the activations from each layer individually to all voxels in the training set. We use 9 different regularisation weights ranging from 10^{-4} to 10^{4} (evenly distributed on a log scale) for the RidgeCV, which picked the best parameter through an approximate leave-one-out error estimation. Then, the regressions were evaluated on the validation set. For each voxel, the best aligned layer (with the least regression error) is determined by the validation set. Finally, the alignment of the model is evaluated as the correlation between model predictions and the ground-truth fMRI signals on the testing set, with the predetermined voxel-layer correspondence from the validation set. The whole process is repeated 10 times for each model with Monte Carlo sampling on the splits to estimate the mean and variance of brain alignment. The choice of using a 15-second clip duration gives a high cross-dataset consistency (0.781) of the motion index map (Fig. 4f). A longer clip duration leads to a substantial drop in consistency (Fig. S22), but still gives qualitatively similar results (SI C). Finally, these brain alignment scores are all normalised by the voxel-level internal consistency (divided by the ceiling) described in Methods B.1.

Layer selection and estimation of model-mapped hierarchy. We hypothesise that the hierarchical layer mapping from the models reflects the true computational hierarchy in the brain. For each model, we only subsample at most 15 layers for brain alignment for the sake of saving storage and runtime. For models with block modules, we try to include all the blocks. Otherwise, we evenly sample layers from the whole model. As mentioned above, there is an assigned layer for each voxel from the regression procedure. We compute the relative position of this layer as k/L, where k is the layer position and L is the total number of layers considered for this specific model. Despite prior works claiming that the layer correspondence is inaccurate [123, 100], we get a rather consistent layer assignment by averaging the 20 most brain-aligned dynamic models (split-half consistency over visually driven voxels: 0.839±0.022). This relative layer assignment is used for analyses in Fig. 2e (Fig. S6), Fig. 4e. To compare the model-mapped hierarchy with an existing reference, we used the region connectivity suggested by *Rolls et al.* [114]. We applied an annealing-based ordering procedure 100 times, recording the position of each region in the hierarchy on each run, and then averaged these positions to obtain a probabilistic hierarchy estimate for each region. We then correlated these reference hierarchy values with the model-mapped hierarchy values in Fig. 2e.

B.4. Behavioural alignment computation.

Behavioural decoders. To compare behavioural judgements on action recognition videos between humans and models, we need to enable models to produce judgements. We achieve this by training behavioural decoders using data distinct from the human behavioural experiments. Specifically, for each presentation condition (RGB and biological motion), we collect 100 samples per action class, resulting in 1,000 samples per condition for fitting the

behavioural decoders. The two decoders are applied to the corresponding groups: one for normal and shuffled RGB videos, and the other for normal and shuffled biological motion videos. For each model, the layer that achieved the highest testing accuracy on the fitting data is selected to generate judgements for the stimuli used in the human behavioural experiments. Both model and human judgements are compared to the ground truth labels, and their accuracies are compared in SI A.7.

Human error pattern alignment. In addition to comparing accuracy between models and humans, we are more interested in whether they exhibited similar error patterns. To analyse this, we apply the one-versus-other object-level performance metric (O_2) [106], adapted to our context as an *action-level* metric. The O_2 matrix is defined as:

$$O_2(i,j) = zscore(hit(i,j)) - zscore(false_alarm(i,j))$$

where i,j represent indices of the *action classes* for the ground truth and model/human choice, and $zscore$ is the inverse of the cumulative Gaussian distribution. The hit rate (hit) is the proportion of trials where videos of action i are correctly labelled when presented against distractor action j, and the false alarm rate ($false_alarm$) is the proportion of trials where videos of action j are incorrectly labelled as action i. Effectively, the O_2 matrix encodes the sensitivity measured by d' for each video-choice combination. The model-human alignment is then computed as the correlation between elements from the model O_2 matrix and the human O_2 matrix, excluding the $i = j$ entries (as we are primarily interested in the error patterns rather than correct classifications).

B.5. Decoding performance on cognitive tasks.

Cognitive task datasets. We consider 10 tasks as the hypothesised functional objectives that drive the visual brain during natural dynamic vision. These tasks and their corresponding datasets are: *object recognition* (`imagenet` [71]): classifying the object categories; *face recognition* (`vggfacev2` [15]): recognising the faces of famous people; *pose estimation* (`majajhong2015` [57]; here we only use the pose data, without the neural recordings): assessing the position and rotation of objects; *social understanding* (`mcmahon-social` [89]): classifying whether two people are interacting/communicating; *appearance-focused action recognition* (`kinetics400` [63]): general-purpose action recognition that includes action classes that are easy to classify based on objects, *e.g.,* playing ping-pong; *motion-focused action recognition* `smthsmthv2` [50]): action recognition that focuses on solely the movement, *e.g.,* putting [something] on [something] (the [something] can be arbitrary objects); *egomotion estimation* (`airsim` [93]): assessing the translational and rotational speeds of objects; *direction discrimination* (`ding2012` [25]): assessing the moving direction of a group of dots with randomness; *motion-only action recognition* (`afd` [61]): recognising the action from optical-flow-embedded white-noise stimuli; *biological-motion-based action recognition* (`hdm05` [96]): recognising the action from moving joints confounded by a background of static random dots. For `hdm05`, we project the 3D joints onto rotations of -45, 0, and 45 degrees and add random dots of the same size to the background. This makes it impossible to recognise the positions of joints from a single frame. The same property also applies to `afd` and `ding2012`, making them purely dynamic. The purely static datasets include `imagenet`, `vggfacev2`, and `majajhong2015`, all including only image stimuli. The rest of the datasets have dynamic stimuli, but studies have been arguing that solving these tasks largely involves processing static information [61, 31]. Fig. S10 shows the samples for every task and how these tasks span the spectrum from purely static to purely dynamic information processing (SI B.1). This spectrum of tasks therefore provides a good opportunity to study the interplay between the two kinds of dichotomous information. All sample images shown in Fig. S10 and Fig. 3 are drawn from the original datasets described above. For any images containing humans, we regenerated the visuals using ChatGPT [102] to ensure anonymity.

Subsampling of the datasets. We have to perform subsampling on the task datasets to limit computation and storage costs. For each regression task, including `airsim` and `majajhong2015`, we randomly sample 5000 instances from the dataset. For other classification tasks, we cap the number of classes to be 100 and the number of samples per class to be 50.

Cross-validated decoding performance of tasks. Even if the models are not originally optimised for these cognitive tasks, we can measure how much their activations *support* them, similarly to neural decoding. This is done by learning regression/classification from the model activations to the various task labels, either continuous or categorical. We apply a cross-validation process that is identical to that of evaluating brain alignment, with the proportions of training, testing, and validation sets to be $80\%, 10\%, 10\%$. The layer selection for decoding is again done with the validation set. However, a major difference is that when selecting the layers, we only consider the top layers from the models. The representation at the top layers is constructed progressively throughout the whole model due to their hierarchical structures. In other words, the task supported by the top layers shapes the representation of all preceding layers. Therefore, we choose to consider the top 20% layers (out of the subsampled layers) and take at least 2 top layers from each model.

B.6. Task relevance computation.

Task relevance. The relevance of a single task to the brain (or brain regions) is simply defined as the explained variance (or R-squared) from correlating the task performance and the whole-brain (or regional) alignment scores across all tested models. A high relevance indicates that by increasing the performance of models on this task, we expect their neural or behavioural alignment to also increase. This further suggests that this task might serve as a functional objective for the computation across the brain or in those specific regions. Similarly, for behavioural alignment, a high relevance implies that the hypothesised cognitive task may drive the model's alignment with human behavioural outputs. It is worth noting that sometimes the task performance can be negatively correlated with the brain alignment. We regarded these cases as artefacts from the limited size of our model pools. Therefore, we constrain all the regressions used for task relevance computation to have non-negative weights.

Joint task relevance. A single area in the brain might serve for multiple different functional goals achieved in the downstream areas. To consider the relevance of multiple tasks in combination, we regress the performance on these tasks to the brain or regional alignment. Then, the joint relevance is defined as the adjusted explained variance (or adjusted R-squared) of this regression:

$$R^2_{adjd} = 1 - \frac{\left(1 - R^2\right)(n-1)}{n-p-1}$$

where R^2 is the explained variance (or R-squared). The adjustment takes into account the number of regressors used. Analogically, the joint relevance reflects how accurately the brain alignment can be predicted from the performance of given tasks, therefore how these tasks drive the neural processing in combination.

Residual task relevance. Since the underlying mechanism of solving two tasks can be similar, *e.g.,* face and object recognition both involve integrating shape and texture information across space, the relevance of the two tasks might overlap with each other. To remove the contribution of a specific task in the joint task relevance, we considered the unique explained variance of the remaining task. Generally, if the whole set of tasks considered is T and the set of tasks we want to remove is $\hat{T}$, the residual task relevance of $T - \hat{T}$ is $R^2_{T-\hat{T}} = R^2_T - R^2_{\hat{T}}$. If the joint relevance of T can be fully explained by the subset $\hat{T}$, the residual task relevance $R^2_{T-\hat{T}}$ should be close to 0. For verifying the statistical significance of the residual task relevance, we conduct permutation tests with 1000 resamples.

Eigenanalysis of Voxel-Level Task Relevance. In Fig. 3g, we visualise the 2D eigenspace of visually driven voxels across the cortex, based on their task relevance profiles. Specifically, each voxel is represented as a point in an M-dimensional space, where $M = 10$ corresponds to the number of cognitive tasks, and the coordinates are given by the relevance for each task, *i.e.,* the vector $(R^2_{t_1}, R^2_{t_2}, \ldots, R^2_{t_M})$. Applying eigendecomposition to the voxel-by-task relevance matrix, we find that the first two principal components together explain 91% of the total variance, with a balanced contribution from each component (47% and 44%, respectively), indicating that the space is evenly spanned in two dimensions. Furthermore, projecting each task vector onto this eigenspace reveals two dominant, approximately orthogonal axes that correspond to object-related and motion-related processing. These axes capture the principal semantic structure underlying the voxel-task relationships.

Motion index of information processing. We consider an intuitive measure of how a brain region processes static versus dynamic information: $R^2_M/(R^2_M + R^2_O)$, where M represents `afd` and O represents `imagenet`. This measure ranges from 0 to 1, indicating from processing only static information to processing only motion information. At the level of 0.5, it indicates an equal processing of both sources of information. To make it more intuitive, we map the range from 0% (purely static) to 50% (equal) and then back to 100% (purely dynamic) in Fig. 4f and Fig. 5f.

C. Quantification and statistical analysis

All analyses are conducted in Python using two-tailed statistical tests. For paired samples (*e.g.,* individual model scores), Student's t-tests are applied, while for non-paired samples (*e.g.,* model group scores and task relevance), Mann-Whitney U tests are used. Multiple comparisons across tasks, voxels, or regions are corrected using the false discovery rate (FDR) method. To estimate the 95% confidence intervals for task relevance, bootstrapping with 1,000 resamples is performed over model scores. On the other hand, the confidence intervals of model scores are directly based on estimated standard deviations from different data splits. To assess whether task relevance or remaining task relevance is significantly above zero, permutation tests with 1,000 resamples are conducted. Statistical significance thresholds are defined as follows: not significant (n.s.), $p > 0.05$; $*p < 0.05$; $**p < 0.01$; $***p < 0.001$, unless stated otherwise.

Acknowledgments

We thank Ben Lönnqvist, Johannes Mehrer, Nancy Kanwisher, and members of the EPFL NeuroAI Lab for helpful feedback. We also thank Yitao Xu and Huzheng Yang for valuable discussions.

References

[1] A. Abraham, F. Pedregosa, M. Eickenberg, P. Gervais, A. Mueller, J. Kossaifi, A. Gramfort, B. Thirion, and G. Varoquaux. Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics*, 8, Feb. 2014. Publisher: Frontiers.

[2] R. Adolphs. The social brain: neural basis of social knowledge. *Annual review of psychology*, 60(1):693–716, 2009.

[3] Y. Asano, M. Patrick, C. Rupprecht, and A. Vedaldi. Labelling unlabelled videos from scratch with multi-modal self-supervision. *Advances in Neural Information Processing Systems*, 33:4660–4671, 2020.

[4] A. Bardes, Q. Garrido, J. Ponce, X. Chen, M. Rabbat, Y. LeCun, M. Assran, and N. Ballas. Revisiting Feature Prediction for Learning Visual Representations from Video, Feb. 2024. arXiv:2404.08471 [cs].

[5] A. Bartels, N. K. Logothetis, and K. Moutoussis. fmri and its interpretations: an illustration on directional selectivity in area v5/mt. *Trends in neurosciences*, 31(9):444–453, 2008.

[6] P. Bashivan, K. Kar, and J. J. DiCarlo. Neural population control via deep image synthesis. *Science*, May 2019. Publisher: American Association for the Advancement of Science.

[7] J. Berezutskaya, M. J. Vansteensel, E. J. Aarnoutse, Z. V. Freudenburg, G. Piantoni, M. P. Branco, and N. F. Ramsey. Open multimodal iEEG-fMRI dataset from naturalistic stimulation with a short audiovisual film. *Scientific Data*, 9:91, Mar. 2022.

[8] G. Bertasius, H. Wang, and L. Torresani. Is Space-Time Attention All You Need for Video Understanding? In *Proceedings of the 38th International Conference on Machine Learning*, pages 813–824. PMLR, July 2021. ISSN: 2640-3498.

[9] F. Binkofski and L. J. Buxbaum. Two action systems in the human brain. *Brain and language*, 127(2):222–229, Nov. 2013.

[10] N. M. Blauch, M. Behrmann, and D. C. Plaut. A connectivity-constrained computational account of topographic organization in primate high-level visual cortex. *Proceedings of the National Academy of Sciences*, 119(3):e2112566119, 2022.

[11] K. Borowiak, C. Maguinness, and K. von Kriegstein. Dorsal-movement and ventral-form regions are functionally connected during visual-speech recognition. *Human Brain Mapping*, 41(4):952–972, 2020.

[12] L. E. Brown, B. A. Halpert, and M. A. Goodale. Peripheral vision for perception and action. *Experimental Brain Research*, 165(1):97–106, Aug. 2005.

[13] T. Bunk, D. Varshneya, V. Vlasov, and A. Nichol. DIET: Lightweight Language Understanding for Dialogue Systems, May 2020. arXiv:2004.09936 [cs].

[14] S. A. Cadena, G. H. Denfield, E. Y. Walker, L. A. Gatys, A. S. Tolias, M. Bethge, and A. S. Ecker. Deep convolutional models improve predictions of macaque V1 responses to natural images. *PLOS Computational Biology*, 15(4):e1006897, Apr. 2019. Publisher: Public Library of Science.

[15] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age, 2018.

[16] V. Cardin and A. T. Smith. Sensitivity of human visual cortical area V6 to stereoscopic depth gradients associated with self-motion. *Journal of Neurophysiology*, 106(3):1240–1249, Sept. 2011.

[17] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[18] C. Caucheteux and J.-R. King. Brains and algorithms partially converge in natural language processing. *Communications Biology*, 5(1):1–10, Feb. 2022. Number: 1 Publisher: Nature Publishing Group.

[19] R. M. Cichy, A. Khosla, D. Pantazis, A. Torralba, and A. Oliva. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, 6(1):27755, June 2016. Number: 1 Publisher: Nature Publishing Group.

[20] K. G. Claeys, D. T. Lindsey, E. De Schutter, and G. A. Orban. A higher order motion region in human inferior parietal lobule: evidence from fMRI. *Neuron*, 40(3):631–642, Oct. 2003.

[21] K. E. Cullen. Vestibular processing during natural self-motion: implications for perception and action. *Nature Reviews Neuroscience*, 20(6):346–363, June 2019. Publisher: Nature Publishing Group.

[22] J. Dapello, T. Marques, M. Schrimpf, F. Geiger, D. Cox, and J. J. DiCarlo. Simulating a Primary Visual Cortex at the Front of CNNs Improves Robustness to Image Perturbations. In *Advances in Neural Information Processing Systems*, volume 33, pages 13073–13087. Curran Associates, Inc., 2020.

[23] M. Deb, M. Deb, and N. Murty. Toponets: High performing vision and language models with brain-like topography. *arXiv preprint arXiv:2501.16396*, 2025.

[24] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[25] L. Ding and J. I. Gold. Neural Correlates of Perceptual Decision Making before, during, and after Decision Commitment in Monkey Frontal Eye Field. *Cerebral Cortex*, 22(5):1052–1067, May 2012.

[26] F. R. Doshi and T. Konkle. Cortical topographic motifs emerge in a self-organized map of object space. *Science Advances*, 9(25):eade8187, 2023.

[27] R. Epstein, A. Harris, D. Stanley, and N. Kanwisher. The parahippocampal place area: recognition, navigation, or encoding? *Neuron*, 23(1):115–125, 1999.

[28] O. Esteban, C. J. Markiewicz, R. W. Blair, C. A. Moodie, A. I. Isik, A. Erramuzpe, J. D. Kent, M. Goncalves, E. DuPre, M. Snyder, H. Oya, S. S. Ghosh, J. Wright, J. Durnez, R. A. Poldrack, and K. J. Gorgolewski. fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature Methods*, 16(1):111–116, Jan. 2019. Publisher: Nature Publishing Group.

[29] H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, J. Malik, and C. Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6824–6835, 2021.

[30] F. Fang and S. He. Cortical responses to invisible objects in the human dorsal and ventral pathways. *Nature neuroscience*, 8(10):1380–1385, 2005.

[31] C. Feichtenhofer. X3d: Expanding architectures for efficient video recognition, 2020.

[32] C. Feichtenhofer, H. Fan, J. Malik, and K. He. SlowFast Networks for Video Recognition, Oct. 2019. arXiv:1812.03982.

[33] C. Feichtenhofer, Y. Li, K. He, et al. Masked autoencoders as spatiotemporal learners. *Advances in neural information processing systems*, 35:35946–35958, 2022.

[34] D. Finzi, E. Margalit, K. Kay, D. L. K. Yamins, and K. Grill-Spector. A single computational objective drives specialization of streams in visual cortex. preprint, Neuroscience, Dec. 2023.

[35] D. Finzi, D. Yamins, K. Kay, and K. Grill-Spector. Do deep convolutional neural networks accurately model representations beyond the ventral stream? In *2022 Conference on Cognitive Computational Neuroscience*, San Francisco, 2022. Cognitive Computational Neuroscience.

[36] S. M. Frank and M. W. Greenlee. The parieto-insular vestibular cortex in humans: more than a single area? *Journal of Neurophysiology*, 120(3):1438–1450, Sept. 2018.

[37] E. Freud, J. C. Culham, D. C. Plaut, and M. Behrmann. The large-scale organization of shape processing in the ventral and dorsal pathways. *eLife*, 6:e27576, 2017.

[38] E. Freud, D. C. Plaut, and M. Behrmann. 'What' Is Happening in the Dorsal Visual Pathway. *Trends in Cognitive Sciences*, 20(10):773–784, Oct. 2016.

[39] C. Galletti and P. Fattori. The dorsal visual stream revisited: Stable circuits or dynamic pathways? *Cortex*, 98:203–217, Jan. 2018.

[40] J. S. Gao, A. G. Huth, M. D. Lescroart, and J. L. Gallant. Pycortex: an interactive surface visualizer for fMRI. *Frontiers in Neuroinformatics*, 9, Sept. 2015. Publisher: Frontiers.

[41] Z. Gao, C. Tan, L. Wu, and S. Z. Li. Simvp: Simpler yet better video prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3170–3180, 2022.

[42] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness, Nov. 2022. arXiv:1811.12231 [cs].

[43] M. A. Giese and G. Rizzolatti. Neural and Computational Mechanisms of Action Processing: Interaction between Visual and Motor Representations. *Neuron*, 88(1):167–180, Oct. 2015.

[44] M. F. Glasser, T. S. Coalson, E. C. Robinson, C. D. Hacker, J. Harwell, E. Yacoub, K. Ugurbil, J. Andersson, C. F. Beckmann, M. Jenkinson, S. M. Smith, and D. C. Van Essen. A multi-modal parcellation of human cerebral cortex. *Nature*, 536(7615):171–178, Aug. 2016. Publisher: Nature Publishing Group.

[45] M. F. Glasser and D. C. V. Essen. Mapping Human Cortical Areas In Vivo Based on Myelin Content as Revealed by T1- and T2-Weighted MRI. *Journal of Neuroscience*, 31(32):11597–11616, Aug. 2011. Publisher: Society for Neuroscience Section: Articles.

[46] A. Gokce and M. Schrimpf. Scaling laws for task-optimized models of the primate visual ventral stream. In *Forty-second International Conference on Machine Learning*, 2025.

[47] T. Golan, P. C. Raju, and N. Kriegeskorte. Controversial stimuli: Pitting neural networks against each other as models of human cognition. *Proceedings of the National Academy of Sciences*, 117(47):29330–29337, 2020.

[48] M. A. Goodale, G. Króliczak, and D. A. Westwood. Dual routes to action: contributions of the dorsal and ventral streams to adaptive behavior. *Progress in brain research*, 149:269–283, 2005.

[49] M. A. Goodale and A. D. Milner. Separate visual pathways for perception and action. *Trends in Neurosciences*, 15(1):20–25, Jan. 1992.

[50] R. Goyal, S. E. Kahou, V. Michalski, J. Materzyńska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag, F. Hoppe, C. Thurau, I. Bax, and R. Memisevic. The "something something" video database for learning and evaluating visual common sense, June 2017. arXiv:1706.04261 [cs].

[51] K. Grill-Spector, K. S. Weiner, K. Kay, and J. Gomez. The functional neuroanatomy of human face perception. *Annual review of vision science*, 3(1):167–196, 2017.

[52] A. Gurtubay-Antolin, C. Battal, C. Maffei, M. Rezk, S. Mattioni, J. Jovicich, and O. Collignon. Direct Structural Connections between Auditory and Visual Motion-Selective Regions in Humans. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 41(11):2393–2405, Mar. 2021.

[53] S. Han, Z. Wang, and M. Zhang. Flow Snapshot Neurons in Action: Deep Neural Networks Generalize to Biological Motion Perception, Oct. 2024. arXiv:2405.16493 [cs].

[54] B. M. Harvey, B. P. Klein, N. Petridou, and S. O. Dumoulin. Topographic representation of numerosity in the human parietal cortex. *Science*, 341(6150):1123–1126, 2013.

[55] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition, Dec. 2015. arXiv:1512.03385 [cs].

[56] M. M. Himmelberg, J. W. Kurzawski, N. C. Benson, D. G. Pelli, M. Carrasco, and J. Winawer. Cross-dataset reproducibility of human retinotopic maps. *NeuroImage*, 244:118609, Dec. 2021.

[57] H. Hong, D. L. K. Yamins, N. J. Majaj, and J. J. DiCarlo. Explicit information for category-orthogonal object properties increases along the ventral stream. *Nature Neuroscience*, 19(4):613–622, Apr. 2016. Number: 4 Publisher: Nature Publishing Group.

[58] E. Hosseini, C. Casto, N. Zaslavsky, C. Conwell, M. Richardson, and E. Fedorenko. Universality of representation in biological and artificial neural networks. *bioRxiv*, 2024.

[59] D. H. Hubel and T. N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1):106, 1962.

[60] C. Humphries, E. Liebenthal, and J. R. Binder. Tonotopic organization of human auditory cortex. *Neuroimage*, 50(3):1202–1211, 2010.

[61] F. Ilic, T. Pock, and R. P. Wildes. Is Appearance Free Action Recognition Possible?, July 2022. arXiv:2207.06261.

[62] K. Kar, J. Kubilius, K. Schmidt, E. B. Issa, and J. J. DiCarlo. Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature Neuroscience*, 22(6):974–983, June 2019. Number: 6 Publisher: Nature Publishing Group.

[63] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman. The Kinetics Human Action Video Dataset, May 2017. arXiv:1705.06950 [cs].

[64] U. Keles, J. Dubois, K. J. M. Le, J. M. Tyszka, D. A. Kahn, C. M. Reed, J. M. Chung, A. N. Mamelak, R. Adolphs, and U. Rutishauser. Multimodal single-neuron, intracranial EEG, and fMRI brain responses during movie watching in human patients. *Scientific Data*, 11(1):214, Feb. 2024. Publisher: Nature Publishing Group.

[65] A. J. Kell, D. L. Yamins, E. N. Shook, S. V. Norman-Haignere, and J. H. McDermott. A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. *Neuron*, 98(3):630–644.e16, May 2018.

[66] T. A. Keller, Q. Gao, and M. Welling. Modeling category-selective cortical regions with topographic variational autoencoders. *arXiv preprint arXiv:2110.13911*, 2021.

[67] S.-M. Khaligh-Razavi and N. Kriegeskorte. Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS computational biology*, 10(11):e1003915, Nov. 2014.

[68] T. C. Kietzmann, C. J. Spoerer, L. K. A. Sörensen, R. M. Cichy, O. Hauk, and N. Kriegeskorte. Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences*, 116(43):21854–21863, Oct. 2019. Company: National Academy of Sciences Distributor: National Academy of Sciences Institution: National Academy of Sciences Label: National Academy of Sciences Publisher: Proceedings of the National Academy of Sciences.

[69] C. S. Konen and S. Kastner. Two hierarchically organized neural systems for object information in human visual cortex. *Nature neuroscience*, 11(2):224–231, 2008.

[70] Z. Kourtzi, H. H. Bülthoff, M. Erb, and W. Grodd. Object-selective responses in the human motion area MT/MST. *Nature Neuroscience*, 5(1):17–18, Jan. 2002.

[71] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[72] J. Kubilius, M. Schrimpf, K. Kar, H. Hong, N. J. Majaj, R. Rajalingham, E. B. Issa, P. Bashivan, J. Prescott-Roy, K. Schmidt, A. Nayebi, D. Bear, D. L. K. Yamins, and J. J. DiCarlo. Brain-Like Object Recognition with High-Performing Shallow Recurrent ANNs, Oct. 2019. arXiv:1909.06161 [cs].

[73] B. Lahner, K. Dwivedi, P. Iamshchinina, M. Graumann, A. Lascelles, G. Roig, A. T. Gifford, B. Pan, S. Jin, N. A. Ratan Murty, K. Kay, A. Oliva, and R. Cichy. Modeling short visual events through the BOLD moments video fMRI dataset and metadata. *Nature Communications*, 15(1):6241, July 2024. Publisher: Nature Publishing Group.

[74] K. Li, Y. He, Y. Wang, Y. Li, W. Wang, P. Luo, Y. Wang, L. Wang, and Y. Qiao. VideoChat: Chat-Centric Video Understanding, Jan. 2024. arXiv:2305.06355.

[75] K. Li, Y. Wang, Y. He, Y. Li, Y. Wang, L. Wang, and Y. Qiao. Uniformerv2: Spatiotemporal learning by arming image vits with video uniformer. *ArXiv*, abs/2211.09552, 2022.

[76] K. Li, Y. Wang, G. Peng, G. Song, Y. Liu, H. Li, and Y. Qiao. Uniformer: Unified transformer for efficient spatial-temporal representation learning. In *International Conference on Learning Representations*, 2022.

[77] P. Li, T. J. Hastie, and K. W. Church. Very sparse random projections. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 287–296, Philadelphia PA USA, Aug. 2006. ACM.

[78] Y. Li, G. K. Anumanchipalli, A. Mohamed, P. Chen, L. H. Carney, J. Lu, J. Wu, and E. F. Chang. Dissecting neural computations in the human auditory pathway using deep neural networks for speech. *Nature Neuroscience*, 26(12):2213–2225, Dec. 2023.

[79] Y. Li, C.-Y. Wu, H. Fan, K. Mangalam, B. Xiong, J. Malik, and C. Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4804–4814, 2022.

[80] B. Lin, Y. Ye, B. Zhu, J. Cui, M. Ning, P. Jin, and L. Yuan. Video-LLaVA: Learning United Visual Representation by Alignment Before Projection, Oct. 2024. arXiv:2311.10122 [cs].

[81] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot. NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(10):2684–2701, Oct. 2020. arXiv:1905.04757 [cs].

[82] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie. A ConvNet for the 2020s, Mar. 2022. arXiv:2201.03545 [cs].

[83] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3202–3211, 2022.

[84] M. Maechler. Package 'diptest'. *R Package Version 0.75–5. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing*, 2013.

[85] D. Mahajan, R. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Bharambe, and L. v. d. Maaten. Exploring the Limits of Weakly Supervised Pretraining, May 2018. arXiv:1805.00932 [cs].

[86] N. J. Majaj, H. Hong, E. A. Solomon, and J. J. DiCarlo. Simple Learned Weighted Sums of Inferior Temporal Neuronal Firing Rates Accurately Predict Human Core Object Recognition Performance. *Journal of Neuroscience*, 35(39):13402–13418, Sept. 2015. Publisher: Society for Neuroscience Section: Articles.

[87] E. Margalit, H. Lee, D. Finzi, J. J. DiCarlo, K. Grill-Spector, and D. L. K. Yamins. A Unifying Principle for the Functional Organization of Visual Cortex. preprint, Neuroscience, May 2023.

[88] A. Marin Vargas, A. Bisi, A. S. Chiappa, C. Versteeg, L. E. Miller, and A. Mathis. Task-driven neural network models predict neural dynamics of proprioception. *Cell*, 187(7):1745–1761.e19, Mar. 2024.

[89] E. McMahon, M. F. Bonner, and L. Isik. Hierarchical organization of social action features along the lateral visual pathway. *Current Biology*, 33(23):5035–5047.e8, Dec. 2023.

[90] A. Miech, J.-B. Alayrac, L. Smaira, I. Laptev, J. Sivic, and A. Zisserman. End-to-End Learning of Visual Representations from Uncurated Instructional Videos, Aug. 2020. arXiv:1912.06430 [cs].

[91] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *ICCV*, 2019.

[92] A. D. Milner. How do the two visual streams interact with each other? *Experimental brain research*, 235:1297–1308, 2017.

[93] P. Mineault, S. Bakhtiari, B. Richards, and C. Pack. Your head is there to move you around: Goal-driven models of the primate dorsal pathway. In *Advances in Neural Information Processing Systems*, volume 34, pages 28757–28771. Curran Associates, Inc., 2021.

[94] M. Mishkin, L. G. Ungerleider, and K. A. Macko. Object vision and spatial vision: two cortical pathways. *Trends in Neurosciences*, 6:414–417, Jan. 1983.

[95] P. Morgado, N. Vasconcelos, and I. Misra. Audio-Visual Instance Discrimination with Cross-Modal Agreement, Mar. 2021. arXiv:2004.12943 [cs].

[96] M. Müller, T. Röder, M. Clausen, B. Eberhardt, B. Krüger, and A. Weber. Documentation mocap database hdm05. Technical Report CG-2007-2, Universität Bonn, June 2007.

[97] A. Nayebi, A. Attinger, M. Campbell, K. Hardcastle, I. Low, C. S. Mallory, G. Mel, B. Sorscher, A. H. Williams, S. Ganguli, L. Giocomo, and D. Yamins. Explaining heterogeneity in medial entorhinal cortex with task-driven neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 12167–12179. Curran Associates, Inc., 2021.

[98] A. Nayebi, D. Bear, J. Kubilius, K. Kar, S. Ganguli, D. Sussillo, J. J. DiCarlo, and D. L. Yamins. Task-Driven Convolutional Recurrent Models of the Visual System. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[99] S. Nishimoto and J. L. Gallant. A Three-Dimensional Spatiotemporal Receptive Field Model Explains Responses of Area MT Neurons to Naturalistic Movies. *Journal of Neuroscience*, 31(41):14551–14564, Oct. 2011. Publisher: Society for Neuroscience Section: Articles.

[100] S. Nonaka, K. Majima, S. C. Aoki, and Y. Kamitani. Brain hierarchy score: Which deep neural networks are hierarchically brain-like? *iScience*, 24(9):103013, Sept. 2021.

[101] P. Olszewski-Kubilius, R. F. Subotnik, L. C. Davis, and F. C. Worrell. Benchmarking Psychosocial Skills Important for Talent Development. *New Directions for Child and Adolescent Development*, 2019(168):161–176, 2019. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cad.20318.

[102] OpenAI. Chatgpt. https://chat.openai.com/chat.

[103] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[104] S. Pitzalis, C. Serra, V. Sulpizio, G. Committeri, F. de Pasquale, P. Fattori, C. Galletti, R. Sepe, and G. Galati. Neural bases of self- and object-motion in a naturalistic vision. *Human Brain Mapping*, 41(4):1084–1111, Nov. 2019.

[105] C. R. Ponce, W. Xiao, P. F. Schade, T. S. Hartmann, G. Kreiman, and M. S. Livingstone. Evolving Images for Visual Neurons Using a Deep Generative Network Reveals Coding Principles and Neuronal Preferences. *Cell*, 177(4):999–1009.e10, May 2019.

[106] R. Rajalingham, E. B. Issa, P. Bashivan, K. Kar, K. Schmidt, and J. J. DiCarlo. Large-Scale, High-Resolution Comparison of the Core Visual Object Recognition Behavior of Humans, Monkeys, and State-of-the-Art Deep Artificial Neural Networks. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 38(33):7255–7269, Aug. 2018.

[107] R. P. Rane, E. Szűgyi, V. Saxena, A. Ofner, and S. Stober. Prednet and predictive coding: A critical review. In *Proceedings of the 2020 international conference on multimedia retrieval*, pages 233–241, 2020.

[108] N. Rathi, J. Mehrer, B. AlKhamissi, T. Binhuraib, N. M. Blauch, and M. Schrimpf. Topolm: brain-like spatio-functional organization in a topographic language model. *arXiv preprint arXiv:2410.11516*, 2024.

[109] M. Rezk, S. Cattoir, C. Battal, V. Occelli, S. Mattioni, and O. Collignon. Shared Representation of Visual and Auditory Motion Directions in the Human Middle-Temporal Cortex. *Current Biology*, 30(12):2289–2299.e8, June 2020. Publisher: Elsevier.

[110] R. Rideaux and A. E. Welchman. But Still It Moves: Static Image Statistics Underlie How We See Motion. *Journal of Neuroscience*, 40(12):2538–2552, Mar. 2020. Publisher: Society for Neuroscience Section: Research Articles.

[111] M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11):1019–1025, Nov. 1999. Publisher: Nature Publishing Group.

[112] G. Rizzolatti and M. Matelli. Two different streams form the dorsal visual system: anatomy and functions. *Experimental Brain Research*, 153(2):146–157, Nov. 2003.

[113] S. Robert, L. G. Ungerleider, and M. Vaziri-Pashkam. Disentangling object category representations driven by dynamic and static visual input, Sept. 2022. Pages: 2022.05.03.490462 Section: New Results.

[114] E. T. Rolls. Two what, two where, visual cortical streams in humans. *Neuroscience & Biobehavioral Reviews*, 160:105650, May 2024.

[115] C. Sava-Segal, C. Richards, M. Leung, and E. S. Finn. Individual differences in neural event segmentation of continuous experiences. *Cerebral Cortex*, 33(13):8164–8178, July 2023.

[116] T. Schenk and R. D. McIntosh. Do we have independent visual streams for perception and action? *Cognitive Neuroscience*, 1(1):52–62, 2010.

[117] H. S. Scholte and E. H. de Haan. Beyond binding: from modular to natural vision. *Trends in Cognitive Sciences*, 2025.

[118] M. Schrimpf, I. A. Blank, G. Tuckute, C. Kauf, E. A. Hosseini, N. Kanwisher, J. B. Tenenbaum, and E. Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118, Nov. 2021. Publisher: Proceedings of the National Academy of Sciences.

[119] M. Schrimpf, J. Kubilius, H. Hong, N. J. Majaj, R. Rajalingham, E. B. Issa, K. Kar, P. Bashivan, J. Prescott-Roy, F. Geiger, K. Schmidt, D. L. K. Yamins, and J. J. DiCarlo. Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like? preprint, Neuroscience, Sept. 2018.

[120] M. Schrimpf, J. Kubilius, M. J. Lee, N. A. R. Murty, R. Ajemian, and J. J. DiCarlo. Integrative Benchmarking to Advance Neurally Mechanistic Models of Human Intelligence. *Neuron*, 108(3):413–423, Nov. 2020. Publisher: Elsevier.

[121] M. Schrimpf, P. McGrath, E. Margalit, and J. J. DiCarlo. Do Topographic Deep ANN Models of the Primate Ventral Stream Predict the Perceptual Effects of Direct IT Cortical Interventions? preprint, Neuroscience, Jan. 2024.

[122] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-k. Wong, and W.-c. Woo. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting, Sept. 2015. arXiv:1506.04214 [cs].

[123] G. St-Yves, E. J. Allen, Y. Wu, K. Kay, and T. Naselaris. Brain-optimized deep neural network models of human visual areas learn non-hierarchical representations. *Nature Communications*, 14(1):3329, June 2023. Publisher: Nature Publishing Group.

[124] C. Tan, S. Li, Z. Gao, W. Guan, Z. Wang, Z. Liu, L. Wu, and S. Z. Li. Openstl: A comprehensive benchmark of spatio-temporal predictive learning. In *Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

[125] M. Tan and Q. V. Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks, Sept. 2020. arXiv:1905.11946 [cs].

[126] S. Thorat, A. Doerig, and T. C. Kietzmann. Characterising representation dynamics in recurrent neural networks for object recognition, Oct. 2023. arXiv:2308.12435 [cs].

[127] Z. Tong, Y. Song, J. Wang, and L. Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.

[128] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.

[129] G. Tuckute, A. Sathe, S. Srikant, M. Taliaferro, M. Wang, M. Schrimpf, K. Kay, and E. Fedorenko. Driving and suppressing the human language network using large language models. *Nature Human Behaviour*, 8(3):544–561, Mar. 2024. Publisher: Nature Publishing Group.

[130] L. M. Vaina, J. Solomon, S. Chowdhury, P. Sinha, and J. W. Belliveau. Functional neuroanatomy of biological motion perception in humans. *Proceedings of the National Academy of Sciences*, 98(20):11656–11661, 2001.

[131] E. Y. Walker, F. H. Sinz, E. Cobos, T. Muhammad, E. Froudarakis, P. G. Fahey, A. S. Ecker, J. Reimer, X. Pitkow, and A. S. Tolias. Inception loops discover what excites neurons most using deep predictive models. *Nature Neuroscience*, 22(12):2060–2065, Dec. 2019. Publisher: Nature Publishing Group.

[132] L. Wang, B. Huang, Z. Zhao, Z. Tong, Y. He, Y. Wang, Y. Wang, and Y. Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14549–14560, 2023.

[133] Y. Wang, J. Zhang, H. Zhu, M. Long, J. Wang, and P. S. Yu. Memory In Memory: A Predictive Neural Network for Learning Higher-Order Non-Stationarity from Spatiotemporal Dynamics, Apr. 2019. arXiv:1811.07490 [cs].

[134] Y. Wong, H. Kwan, W. MacKay, and J. Murphy. Spatial organization of precentral cortex in awake primates. i. somatosensory inputs. *Journal of neurophysiology*, 41(5):1107–1119, 1978.

[135] J. Wu, T. Yan, Z. Zhang, F. Jin, and Q. Guo. Retinotopic mapping of the peripheral visual field to human visual cortex by functional magnetic resonance imaging. *Human Brain Mapping*, 33(7):1727–1740, 2012. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/hbm.21324.

[136] M. F. Wurm and A. Caramazza. Two 'what' pathways for action and object recognition. *Trends in Cognitive Sciences*, 26(2):103–116, Feb. 2022.

[137] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 305–321, 2018.

[138] D. L. K. Yamins and J. J. DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3):356–365, Mar. 2016. Publisher: Nature Publishing Group.

[139] D. L. K. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, June 2014.

[140] C. M. Ziemba, J. Freeman, J. A. Movshon, and E. P. Simoncelli. Selectivity and tolerance for visual texture in macaque v2. *Proceedings of the National Academy of Sciences*, 113(22):E3140–E3149, 2016.