THE HUMAN BRAIN AS A DYNAMIC MIXTURE OF EXPERT MODELS IN VIDEO UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

The human brain is the most efficient and versatile system for processing dynamic visual input. By comparing representations from deep video models to brain activity, we can gain insights into mechanistic solutions for effective video processing, important to better understand the brain and to build better models. Current works in model-brain alignment primarily focus on fMRI measurements, leaving open questions about fine-grained dynamic processing. Here, we introduce the first large-scale benchmarking of both static and temporally-integrating deep neural networks on brain alignment to dynamic electroencephalography (EEG) recordings of short natural videos. We analyze 100+ models across the axes of temporal integration, classification task, architecture and pretraining using our proposed *Cross-Temporal Representational Similarity Analysis (CT-RSA)*, which matches the best time-unfolded model features to dynamically evolving brain responses, distilling 10^7 alignment scores. Our findings reveal novel insights on how continuous visual input is integrated in the brain, beyond the standard temporal processing hierarchy from low to high-level representations. Responses in posterior electrodes, after initial alignment to hierarchical static object processing, best align to mid-level representations of temporally-integrative actions and closely match the unfolding video content. In contrast, responses in frontal electrodes best align with high-level static action representations and show no temporal correspondence to the video. Additionally, state space models show superior alignment to intermediate posterior activity through mid-level action features, in which self-supervised pretraining is also beneficial. We draw a metaphor to a dynamic mixture of expert models for the changing neural preference in tasks and temporal integration reflected in the alignment to different model types across time. We posit that a single best-aligned model would need task-independent training to combine these capacities as well as an architecture that supports dynamic switching.

1 INTRODUCTION

Humans are able to perceive and understand a highly dynamic world efficiently, with neural representations changing dynamically in response to incoming continuous visual information. The framework of representational alignment (Sucholutsky et al., 2023) provides a rich resource to investigate how humans achieve this and for guiding model design. Within computational cognitive neuroscience, this framework is used to identify the mechanisms giving rise to cognition through hypothesis testing with task-performing computational models (Kriegeskorte & Douglas, 2018; Doerig et al., 2023), with neural network computer vision models being the best models capturing neural responses in visual cortex (Yamins et al., 2014; Güçlü & Van Gerven, 2015). Conversely, machine learning can draw from cognitive neuroscience to inform more efficient and robust human-like artificial intelligence, e.g. through implementations of brain-like energy constraints or human-like development Lu et al. (2025a;b). Computational cognitive neuroscience has made remarkable progress by employing large-scale benchmarking as a tool for systematic and reproducible comparisons of model-brain representational alignment (Conwell et al., 2024) under natural stimuli, following the introduction of neural benchmark datasets (Schrimpf et al., 2018; Cichy et al., 2019; 2021; Gifford et al., 2023; 2024; Allen et al., 2022; Hebart et al., 2023; Lahner et al., 2024), mostly with regards to neural processing of static images. However, static images lack temporal context, which strongly affects visual processing in the brain (Willems & Peelen, 2021), as illustrated by the phenomenon of temporal adaptation, where neural responses are modulated by stimulus history (Benda, 2021;

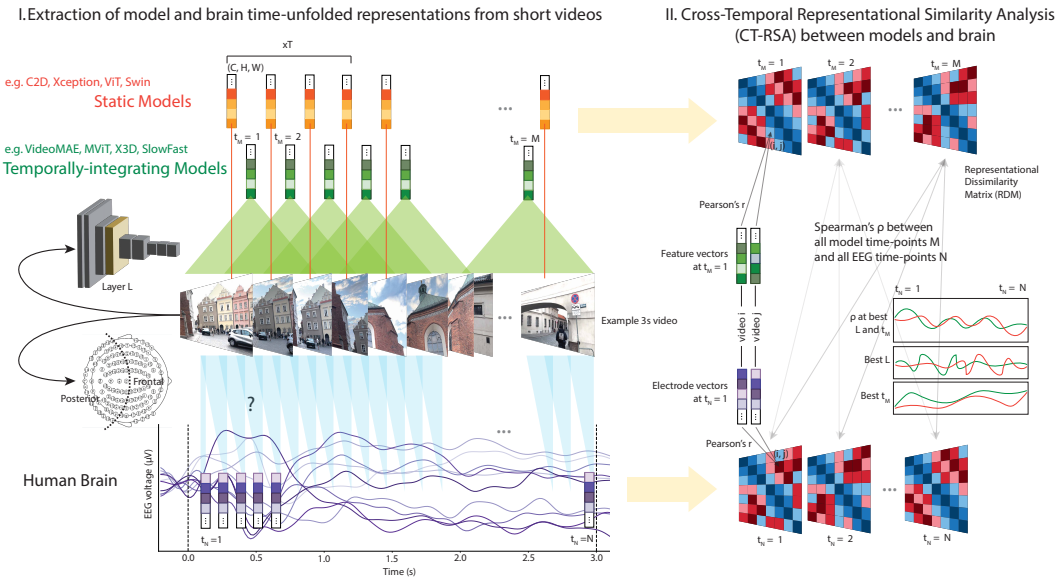

Figure 1: Our method for evaluating alignment of 100+ models with dynamic EEG responses to video. The extraction of time-unfolded representations from both systems (left) is followed by CT-RSA (right), which computes the maximum model-brain alignment across all timepoints and layers of a model. We systematically vary the axes of temporal integration, classification task, architecture, and pretraining - here we highlight differences in model temporal integration.

Brands et al., 2024). Critically, findings from static image perception do not automatically generalize to real-world conditions (Russ et al., 2023; David et al., 2004). This highlights a need for large-scale benchmark initiatives for dynamic video stimuli.

Recent work has mapped video model representations to fMRI (Sartzetaki et al., 2025; Garcia et al., 2025; Tang et al., 2025), showing how key axes of variation in those models (e.g. architecture, task training, degree of temporal integration) affect brain alignment. However, these works did not yet consider two crucial factors: the rich dynamics that characterize neural responses, accessible non-invasively in humans through M/EEG, as well as the dynamics of temporally unfolded model features. Therefore, in this study we employ large-scale benchmarking of 100+ static and temporally-integrating deep neural network models on the dynamic human brain responses to natural videos using EEG, in combination with our proposed *Cross-Temporal Representational Similarity Analysis* (CT-RSA), which matches the best time-unfolded model features across the EEG time-course.

Our contributions are the following:

- We present the first large-scale representational alignment benchmarking on dynamic brain responses (EEG) to natural videos, evaluating 100+ static and temporally-integrating deep neural networks using Cross-Temporal Representational Similarity Analysis (CT-RSA), which matches the best time-unfolded model features across the EEG time-course.
- Our results reveal a changing neural preference for semantic tasks and temporal integration over time, and preferential specialization across the brain hierarchy (posterior-frontal). Posterior activity evolves from alignment with static, hierarchically increasing object processing to mid-level action representations, continuously tracking the video through temporal integration, whereas frontal activity is best captured by early, static semantic action representations.
- Leveraging the time-unfolded model features we show a strong temporal correspondence between model time and EEG time in posterior activity, but not in frontal activity.
- Comparing the additional axes of variation of model architecture and pre-training within models of the same task and temporal integration, we show that state-space models better align to posterior processing, while CNNs better capture frontal activity, and that self-supervised pre-training is beneficial to alignment in earlier posterior processing stages.

Overall, our results suggest that neural representations reflected in fine-grained dynamic brain measurements are not best captured by any single DNN model type but rather resemble a dynamic mixture of expert models that allows switching between semantic tasks and temporal integration, revealing opportunities for brain-inspired representation learning on videos.

2 Related Work

Dynamic brain measurements under static and dynamic natural stimuli. Human brain dynamics are studied at different timescales with distinct methods. For short timescales, usage of M/EEG is essential to make millisecond-level inferences, whereas the sluggish hemodynamic response underlying the fMRI signal prevents such precision. For static stimuli (i.e., images), findings from M/EEG studies of brain dynamics show that discriminable representations emerge as quick as within 50 ms after stimulus onset (Cichy et al., 2014), due to a fast feedforward sweep of visual information processing from early to later stage brain areas, i.e. the cortical hierarchy (Serre et al., 2007). Early neural processing of static images has been found to reflect low-level visual features (Groen et al., 2013), while later neural processing to reflect high-level representations, e.g. categorical processing (Cichy et al., 2014) and task performance (VanRullen & Thorpe, 2001), illustrating a gradual emergence of high-level representations over time: a "temporal hierarchy"(Bankson et al., 2018). This has not only been observed for static objects, but also for static scenes (Greene & Hansen, 2020) and static actions (Zimmermann et al., 2025). From these studies, high-level action information appears to emerge later (250 ms) than high-level object and scene information (150 ms). This may reflect a need for integration of contextual information for action recognition (El-Sourani et al., 2018), with mid-level features like motion patterns object, and scene information being processed in parallel before being integrated at the level of action semantics (Zimmermann et al., 2025).

Human brain dynamics in response to short dynamic stimuli (i.e., videos) are comparatively much less explored. Using fMRI Lahner et al. (2024) found hierarchical correspondence between action recognition model layers and brain regions as in static images, also demonstrating sensitivity to frame-shuffling at different levels and neural temporal correspondence to the video especially in early visual cortex (EVC) - consistent with prior findings on the hierarchy of temporal receptive windows (Hasson et al., 2008; Lerner et al., 2011). Additionally, work by Jung et al. (2025) showed that high-level semantic action features uniquely explain activity over a widespread range of cortex during dynamic natural vision, more so than semantic agent or scene features. A few works have used EEG to investigate brain dynamics to natural videos using domain-specific social interaction videos; Dima et al. (2022) found a temporal hierarchy of visual, action-related and social-affective features, while McMahon et al. (2025) showed that mid and high-level social features are decodable at similar timings, using EEG-fMRI fusion. Using engine-rendered (naturalistic) short videos, Karapetian et al. (2025) also found a temporal processing hierarchy and revealed that motion and action-related features are processed faster during video than static frame presentation. However, large-scale natural video EEG datasets have been missing so far, limiting research on the fine-grained temporal dynamics of real-world visual processing. *In this work,* we are diving in underexplored territory by leveraging a newly collected, large-scale high temporal resolution EEG dataset for domain-general (i.e. not social-specific) and natural (i.e. not rendered) dynamic stimuli.

Benchmarking model representational alignment on short-video fMRI. Very recent works have focused on large-scale benchmarking of model-brain alignment for videos. Garcia et al. (2025) used fMRI of short social interaction videos, comparing 8 video action recognition models with 348 image object recognition models. They found that image models captured representations in EVC and the ventral stream, whereas video models outperformed in the lateral stream, linked to social cognition. Sartzetaki et al. (2025), used fMRI of short domain-general videos, comparing 47 video models, 41 image object recognition models and 11 image action recognition models to disentangle temporal integration from classification task effects. Temporally-integrating models surpassed static models in EVC, through better alignment of middle model layers, while action recognition models surpassed object ones in later stage brain regions via classification layers. Tang et al. (2025), used 5 different video fMRI datasets and compared 92 models, including image object recognition models, video action recognition models, multimodal models, and non-NN baselines. Through correlating brain alignment with model zero-shot performance in different tasks, they found that appearance-free (motion-only) action recognition and object recognition are the two most relevant tasks for brain alignment, with task-agnostic self-supervised models performing best at both tasks and alignment. Current studies are however all based on fMRI, precluding conclusions about fine-grained dynamics. Similarly on the model side, the temporally-evolving internal feature dimension remains unexploited. *In this work,* we present the first large-scale investigation of model alignment with highly dynamic brain representations recorded via EEG during short video viewing. Building on the model set and axes of variation from Sartzetaki et al. (2025), we compare how different model groups and their internal temporal representations align with the brain across processing time.

3 METHODOLOGY

Figure 1 shows an overview of our methodology for measuring alignment of models to dynamic brain responses[1]. We base our methodology on Sartzetaki et al. (2025) but extend it in: (1) applying it to an EEG instead of fMRI dataset, (2) expanding the model architectures sampled, (3) unfolding the temporal dimension of the extracted model features and (4) proposing a temporal extension to the representational alignment metric. We describe these aspects in the next three sections.

3.1 CROSS-TEMPORAL REPRESENTATIONAL SIMILARITY ANALYSIS (CT-RSA)

Neural network models compute representations at subsampled frame intervals, while the mapping of specific frames onto brain responses is uncertain, raising the question of how to align model and brain representations over time. We extend Representational Similarity Analysis (RSA) (Kriegeskorte et al., 2008) to compare all temporally unfolded model features with all EEG timepoints, without imposing assumptions about the relationship between the two. We choose to build on RSA to perform a multivariate analysis that benchmarks emergent brain alignment of model representations[2].

Temporal Representational Dissimilarity Matrices. (A) EEG. At each EEG timepoint t_N we create a super-subject brain Representational Dissimilarity Matrix (RDM) (B_{t_N}), using a particular electrode subset with repetition-average channel vectors v_{s-t_N} for each participant (subject) s. We compute the super-subject RDM by averaging the RDMs of all subjects at that timepoint, as $B_{t_N} = avg_s(B_{s-t_N}),\ B^{ij}_{s-t_N} = 1 - r(v^i_{s-t_N}, v^j_{s-t_N}),\ \forall i, j(i < j),\ 0 \leq j < K, 0 \leq t_N < N$ where K is the total number of videos and r is Pearson correlation. **(B) Models.** We flatten the features of each model layer l and timepoint t_M into a one-dimensional feature vector of length CxHxW, and reduce its dimensionality using Sparse Random Projection followed by Principal Component Analysis to 100 Principal Components to obtain f_{l-t_M}. We compute one RDM per model layer and model time-point as $M^{ij}_{l-t_M} = 1 - r(f^i_{l-t_M}, f^j_{l-t_M}),\ \forall i, j(i < j),\ 0 \leq j < K,\ 0 \leq t_M < M$.

Cross-Temporal correlation of RDMs. To calculate the cross-temporal representational alignment score between the model and EEG timecourses of RDMs, we first compute Spearman's ρ for all combinations of model and EEG timepoints for each layer $R_{l-t_M t_N} = \rho(B_{t_N}, M_{l-t_M})$. We then choose the highest-correlating model layer and model timepoint per EEG timepoint $R_{t_N} = max_{l-t_M}(R_{l-t_M t_N})$. This maximization is essential to retrieve insights from approximately 1k RSA scores that result from all combinations of model layers and timepoints. Over all EEG timepoints, electrode subsets, and models, this allows us to handle over 10^7 RSA scores.

Noise ceiling computation. Because of individual subject variability in brain data, noise ceilings for each electrode subset are computed to compare model RSA scores against the maximum obtainable score given the inter-subject variability, following standard approaches in the field (Nili et al., 2014). For the lower noise ceiling (LNC) we compute a mean RDM across all subjects except one. Then we take the Spearman correlation of the left-out subject RDM and mean RDM, repeat for all the subjects and calculate the average. For the upper noise ceiling (UNC) we take the mean of all RDMs without removing subjects, compute the Spearman correlation of each subject RDM with the mean RDM, and average. The UNC signifies perfect correlation for the amount of noise in the data, often referred to as the maximum amount of variance that can be explained. All further reported alignment scores are scaled by the UNC, and so is the LNC.

Statistical significance. To test if model RDMs correlated significantly with the brain RDMs, we performed permutation tests (Nili et al., 2014), in similar fashion as done in Sartzetaki et al. (2025). For each model RDM at every EEG timepoint we select the model RDM for highest-correlating layer and model timepoint pair and permute the entries of the flattened RDM (excluding the diagonal) 10000 times using the same 10000 random permutations. We then calculate a null distribution per EEG timepoint by computing the Spearman correlations of all permuted RDMs with the brain RDM. For significance of a group of models against zero, we perform a two-tailed sign test between the average null distribution of all models in the group and the average observed Spearman correlation, corrected by subtracting the average observed Spearman correlation before stimulus onset, to

[1]Due to licensing restrictions of the stimuli used in the brain dataset, the video frames shown in the figure are sourced from representative videos captured by the authors themselves and are not subject to copyright.

[2]We utilize the RSA implementation from the Net2Brain python library (Bersch et al., 2025).

account for inflated pre-stimulus correlation scores arising from the maximization inherent to our CT-RSA method. To test for significant differences between two groups of models, we perform a two-tailed sign test between the null distribution created from the across-group differences in the within-group average distributions, and the observed difference in the average of the two model groups' Spearman correlations. When showing group medians instead of means, the above statistical inferences are performed on the medians. All statistical inferences are corrected for multiple comparisons across time points using FDR correction (Benjamini & Hochberg, 1995) with a cluster threshold of two consecutive time points. These tests assess each model group's overall prediction score and its relative predictive ability compared to the other model groups.

3.2 Video Models

We expand upon the original set of 99 video models as used in Sartzetaki et al. (2025) to additionally reflect changes in the current state-of-the-art, such as the introduction of state space models (SSMs) as a scalable linear-complexity alternative to Transformers (Gu & Dao, 2024), for both static vision (Zhu et al., 2024) and video (Li et al., 2024). With the addition of 3 VisionMamba object recognition and 8 VideoMamba action recognition models, our final set of models includes 44 image models trained for object recognition on ImageNet, 10 image models trained for action recognition on Kinetics 400, and 49 video models trained for action recognition on Kinetics 400. Image action models are trained on videos but treat time in a trivial way (separate computations per frame). An additional set of 7 models trained for action recognition on other datasets (Kinetics 710 and Something-Something-v2) was also evaluated, bringing the total to 110 computer vision models tested overall. Comprehensive and extended lists of models can be found in Appendix B.

Temporally unfolded feature extraction. Since the models process a fixed number of frames at a given sampling rate, each video is divided into S sub-clips of length T, differing per specific model. For each layer, we extract features of shape (T, C, H, W) from each sub-clip, and unfold the temporal dimension to obtain TxS=M model timepoint features of shape (C, H, W). We extract features from all higher-level blocks in the models and include the final classification layer.

3.3 EEG Dataset

We utilize the newly collected EEG Moments Dataset EEGMD (to be made public in the near future); the EEG-based extension of the BOLD Moments Dataset (BMD) Lahner et al. (2024), a large scale video fMRI dataset. EEGMD covers the exact same set of 1102 short (3s) naturalistic videos, collected from Moments in Time (Monfort et al., 2019) and Multi-Moments in Time (Monfort et al., 2021). Extensive details on the EEG data acquisition are in Appendix A. The dataset consists of EEG recordings of 6 participants for a "train set" of 1000 videos with 6 repeats and a "test set" of 102 videos with 24 repeats, using a set-up of 128 electrodes. Data was recorded at a sampling rate of 1000 Hz with online filtering (between 0.1 Hz and 100 Hz) and rereferenced (to Fz). In the current analysis we use the test set for the application of RSA, as signal to noise ratio is expected to be higher due the greater number of repetitions, similar to Sartzetaki et al. (2025). The test set videos are representative of the whole dataset, covering a wide range of objects, actions and scenes (for more elaborate description of the video contents see Lahner et al. (2024)).

Preprocessing and electrode selections We performed offline preprocessing using Python and MNE (Gramfort et al., 2013). The continuous EEG data was epoched into trials from -0.2s to 3.5s with respect to stimulus onset, baseline-corrected by subtracting the mean of the pre-stimulus period separately for each trial and channel, and temporally downsampled to 50 Hz. Next, we performed multivariate noise normalisation (MVNN) based on the covariance matrices of each timepoint to reweigh (un)reliable sensors and to (de)emphasise specific spatial frequencies, as recommended for multivariate analyses such as RSA (Guggenmos et al., 2018). We separately analyzed two electrode partitions to make a coarse distinction between parts of the brain: posterior electrodes (35), overlaying visual cortex and commonly used in vision studies, and frontal electrodes (54), covering (pre)frontal cortex, associated with executive functions (see Appendix A for all included electrodes).

4 Results

We assess brain alignment over time for 110 deep neural networks using CT-RSA, considering four main axes of variation: temporal integration, classification task, architecture, and pretraining.

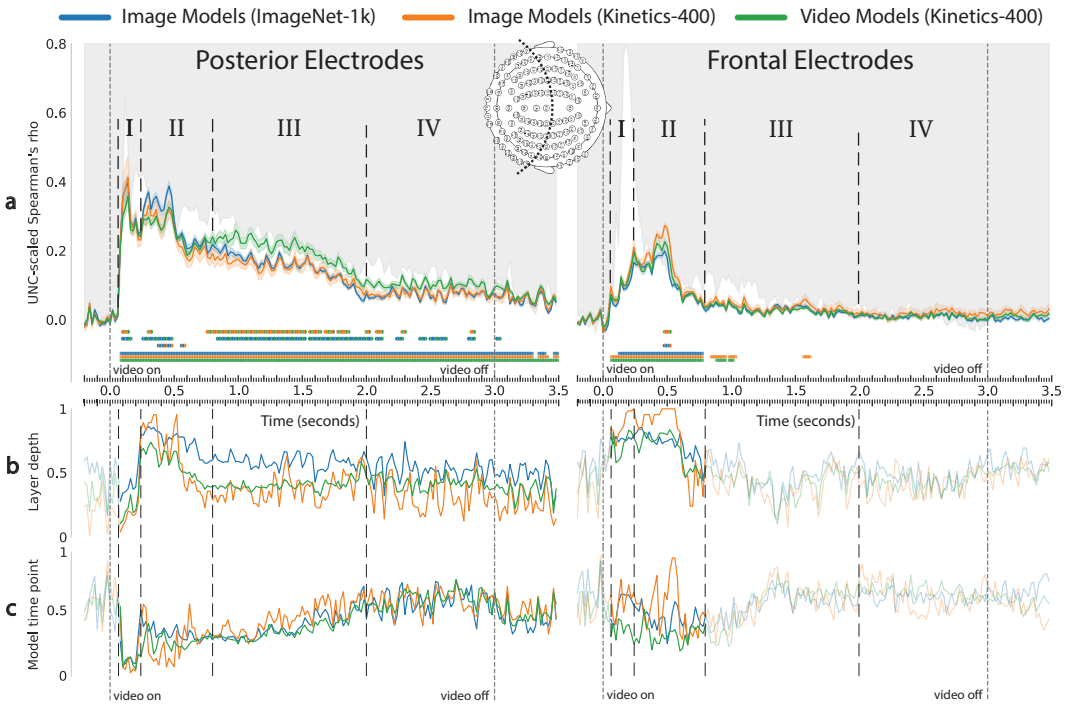

Figure 2: (Left) Task and temporal integration types dynamically interchange in posterior electrodes. (Right) Action task specificity is early and stable in frontal electrodes. Row (a) shows the maximum score over all model timepoints and layers (group average), with significance marked by squares (against zero) and two-colored circles (pairwise), while (b) and (c) show the layer and timepoint that yield the scores. Processing stages I-IV are identified and LNCs are outlined in gray background.

Task and temporal integration types dynamically interchange in posterior electrodes. In Fig. 2 (Left: Posterior) we compare models across the whole EEG time-course by varying (1) the temporal integration, i.e. image (static) v.s. video (temporally-integrating) models, and (2) the classification task, i.e. object (ImageNet-1k) v.s. action recognition (Kinetics-400) models. RSA scores (Fig. 2a) in all model groups are significant against zero from 0.08s, extending to the duration of the whole video and offset. Based on visual inspection of the development of these scores, we distinguish four temporal stages of processing: (I) 0.06s - 0.24s, (II) 0.24s - 0.8s, (III) 0.8s - 2s and (IV) 2s - 3s. To summarize the results of these different stages and explore model variation, we additionally evaluate bin-averaged scores in Fig. 3. We next discuss these stages in turn, simultaneously interpreting the score (Fig. 2a/3a) and the best model layer (Fig. 2b/3b) that gives rise to it. First in **stage I,** scores peak for all model groups at 0.14s and image models significantly outperform video models, with the best model being AlexNet. For all model groups this high correspondence can be accounted for by relatively early model layers (below 0.5), and as the peak in scores decays, by progressively later layers. This suggests that during this stage, posterior processing reflects static low-level information, independent of task. In **stage II,** object recognition models show a clear peak, outperforming the other model groups; the best model here is a DenseNet. In contrast with stage I, the significant scores for all model groups are due to late layers, primarily relating posterior processing at this stage to static high-level object information. During **stage III,** we observe a steady decrease in image model scores, while the score for video models increases and then remains relatively stable, before dropping towards the end of the stage. This leads to video models outranking the other groups (best model being MViT-v2), with scores being driven by mid-level layers. Thus, overall stage III can be summarized as a mid-level temporally-integrative action processing stage. During **stage IV**, scores and layer depth across all model groups remain stable after reaching their lowest point, with video models less significantly outperforming image models than in stage III. Looking across stages in Fig. 2b/3b, action recognition model scores correspond to an earlier layer than object recognition ones in stages I and III, with less clear patterns in the other stages.

The best model timepoint (Fig. 2c/3c) reveals a strong correspondence between model and EEG timepoints in posterior electrodes: early EEG timepoints best correspond to early model timepoints (close to 0), and later EEG timepoints to gradually later model timepoints (yet only up to $\sim$0.6),

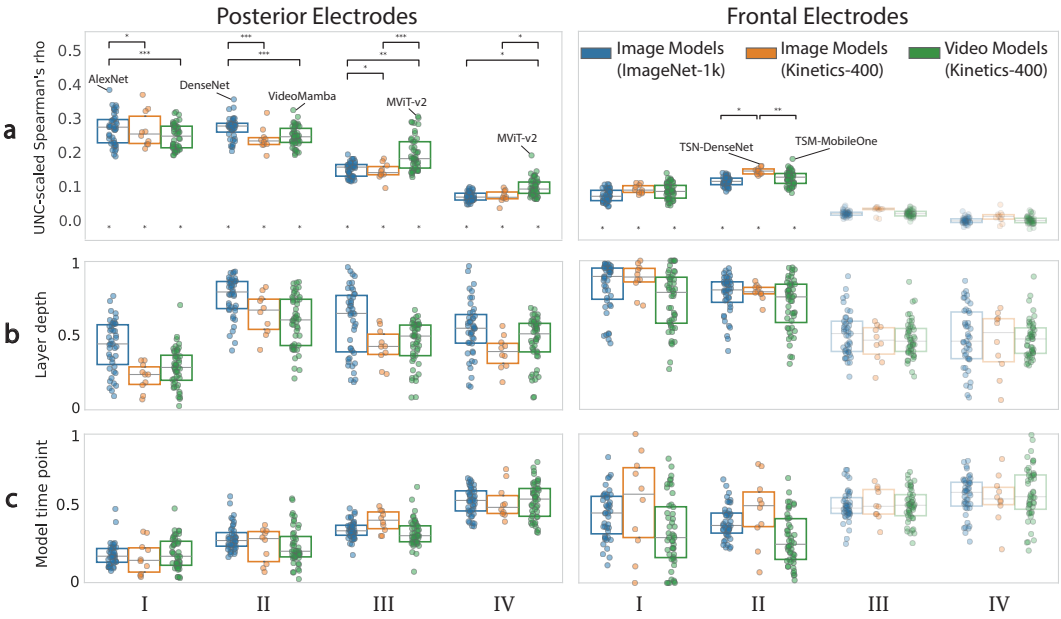

Figure 3: Bin-averaged results showing model variability within each of the temporal processing stages in Fig. 2, outlining best-aligned models in row (a). Significance is marked by a star at $y = 0$ (against zero) and by the respective number of stars for $p < 0.05$, 0.01, and 0.001 on top (pairwise).

highlighting the temporality of posterior neural responses. Note also that where layer depth is high (stage II), the match in model timepoint is less strict.

The analysis demonstrates that posterior processing during video perception is highly temporal, progressing through distinct stages that emphasize different representational demands. An early peak around 0.14s reflects low to mid-level static processing, followed by a rapid shift toward higher-level static object representations. Beyond this point, posterior parts of the brain gradually transition into mid-level temporally-integrative action processing, marking the late stages of video perception.

Action task specificity is early and stable in frontal electrodes. In Fig. 2/3 (Right: Frontal) we show results for the frontal partition. We observe that most neural processing of video information occurs in stages I and II, as model groups show no significant alignment scores afterwards. Scores in this period show two peaks for all model groups, at 0.24s and 0.5s, with static action recognition models significantly outperforming the other groups at the last peak. This results from correspondence with late model layers. Regarding the best model timepoint (Fig. 2c/3c) and in contrast to posterior electrodes, we see that in frontal electrodes, during the period of significant scores, there is no clear temporal correspondence between model and EEG timepoints - with large spread across models (Fig. 3c). These results suggest that during video perception, neural processing reflected in frontal electrodes shows limited temporality and is primarily engaged until 0.8s with high-level, action-related information that is largely independent of within-video dynamics.

State-space video models best capture mid-level intermediate posterior processing. In Fig. 4 we compare brain-alignment of different architectures, i.e. CNNs, Transformers, or SSMs, focusing on the stages I-IV identified in the previous sections. To avoid confounds from task or temporal integration, we focus this analysis on video action recognition models. SSMs are the most brain-aligned to posterior channels in stages I-II, especially in stage II, and through earlier layers (mid-depth). Additionally in stage I, Transformers are significantly better aligned than CNNs. In frontal channels CNNs are most brain-aligned in stages I-II via late layers. In stages III and IV of both channel sets differences are mostly not significant. In Appendix C we show additional comparisons that a) control for model pretraining, and b) compare between object recognition models instead of video action recognition models. The effects of architecture are stable across pretraining types, but differences between object recognition models are more muted. Specifically, static object SSMs do not show the observed advantage for video SSMs in posterior stage II. This suggests that SSMs capture a distinct component of neural processing in posterior channels during phase II, previously linked to static high-level object representation, either through temporally-integrative or mid-level action features.

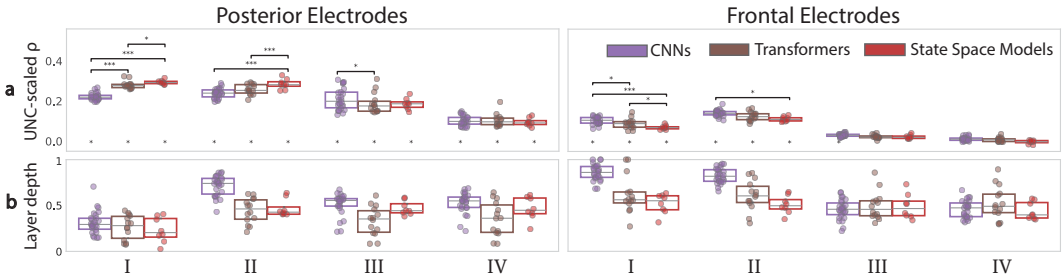

Figure 4: Architecture variation in bin-averaged scores for each stage. (Left) Video SSMs best capture mid-level intermediate posterior processing. (Right) CNNs give a slight edge in alignment to high-level frontal processing.

Self-supervised to no pre-training switch in posterior electrodes. In Fig. 5 we compare the brain alignment of video action recognition models having different types of pretraining, either no pretraining, supervised pretraining on images (object recognition), supervised pretraining on videos (action recognition), or self-supervised pretraining on videos. We observe that in stages I-II of the posterior electrodes self-supervised pretraining achieves superior brain alignment. In stage I it is on par with pretraining on image object recognition, while in stage II it is superior to all other types. In stage III, no pretraining is significantly better than all other types. In stage IV in posterior and in all stages in frontal electrodes, significant differences are less consistent. We again further control for the model architecture in Appendix C, and find that the advantages of self-supervised pretraining in stage II and no pretraining in stage III are robust across architectures. We hypothesize that the advantage of self-supervised pretraining in the primarily "object processing" stage could relate to self-supervision enabling generalization to other tasks, while the benefit of no pre-training in the temporally-integrative stage may reflect avoiding shortcut learning of unrelated patterns Byvshev et al. (2022). Appendix C additionally compares different action recognition fine-tuning datasets (Something-Something-v2, Kinetics710), showing no effects on alignment at any stage or partition.

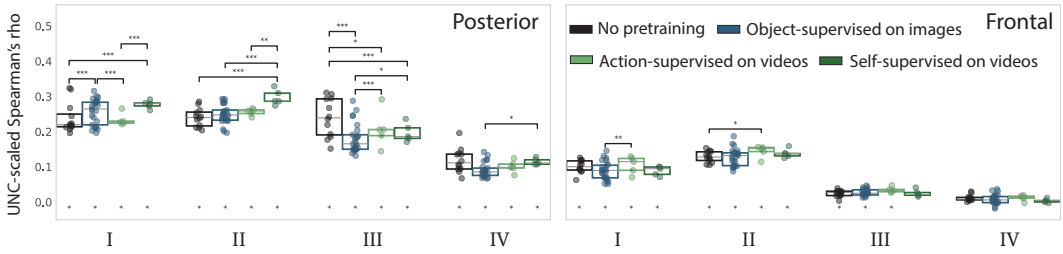

Figure 5: Pretraining variation in bin-averaged scores for each stage. (Left) Switch of benefit from self-supervised (stage II) to no pretraining (stage III) in posterior electrodes. (Right) Slight advantage of supervised video pretraining in frontal electrodes.

5 Discussion and Conclusions

In this work we performed a large-scale model comparison of both static and temporally-integrating deep neural networks to dynamic EEG recordings. Using CT-RSA we make the following main observations: (1) Neural processing during video perception in posterior parts of the brain is highly temporal and unfolds in distinct stages, best captured by different type of representations: from an initial alignment with static low-level features, to mid- and high-level object features, and finally to mid-level temporally integrative action features. (2) In contrast, neural processing during video perception in frontal parts of the brain shows restricted temporality and is best aligned with static high-level action features early in the video. (3) Architecture-wise, temporally-integrating SSMs best capture mid-level representations in posterior electrodes, reflecting a different component of the response in the primarily object-related stage. (4) Self-supervised pretraining helps capture brain representations in the early, primarily object-related stage, while performing no pretraining captures brain representations of later more temporally-integrative stages.

What we learn by moving to dynamic natural stimuli. Moving from static images to more real world dynamic videos, we find similarities and differences in neural processing between these two formats. Congruent with the works of Bankson et al. (2018), Greene & Hansen (2020) and Zimmermann et al. (2025) we find a temporal hierarchy of object and action features in posterior processing,

but only within the duration of phase I and II, i.e. the first 500 ms of processing, consistent with the stimulus durations commonly used in image-based studies. Our findings are also largely consistent with prior work on the neural dynamics of short videos showing a temporal hierarchy of (social) action-related features (Dima et al., 2022; McMahon et al., 2025; Karapetian et al., 2025). However, as we extend beyond the timescales used in these studies (beyond 1s) we discover that temporally-integrative mid-level action features are most brain-aligned until the end of the video, challenging the notion of a strict temporal hierarchy. Notably, this sustained alignment occurs consequently after the peak in alignment with regards to static high-level action features in frontal areas, raising the possibility that feedback from frontal to posterior regions may contribute to shaping later stages of visual processing of dynamic action information. Multiple studies have posed that during object recognition in static context feedback information from prefrontal cortex to posterior regions is critical in shaping behaviorally sufficient object representations, especially under challenging conditions (Goddard et al., 2016; Kar & DiCarlo, 2021; Oyarzo et al., 2024).

What we learn from using dynamic brain measurements for benchmarking. In relation to the results in Sartzetaki et al. (2025), we found high alignment with EVC for mid-layers in temporally-integrating video models. Here, we find a similar advantage for these model features in stages III and IV, i.e. from 0.8s and onwards, in the posterior part of the brain, following initial alignment with high-level object model layers. Combining the insights from both studies, we hypothesize that mid-level dynamic processing could be performed by the EVC after the initial encoding of the video appearance. The high alignment of action recognition models reported in high-level visual cortex by Sartzetaki et al. (2025) was also observed in the current study, but in the frontal electrode partition, likely reflecting an even higher-level cortical stage of processing, during stage II (0.24s–0.8 s). Additionally, leveraging the temporal resolution of EEG, we find that object representations also contribute to video processing in the posterior part of the brain, being the most predictive during early processing but rather briefly, before the processing of dynamic features. A hypothesis for the absence of object contribution in fMRI alignment for video is that, relative to the whole video time-course, object features are processed only briefly, and this transient signal is lost in the aggregation of the fMRI response. Similarly, the enhanced alignment with static processing during stage I (0.06s-0.024s) in posterior electrodes may be too brief to appear in the fMRI signal.

Prospects for model model building. Our findings suggest that the brain performs computations that are not best represented by any single model during the entire EEG time-course, but rather by alternating semantic tasks and temporal integration strategies. We propose (1) that a single model could be best aligned to the whole time-course if it was trained on a sufficiently general objective (e.g. self-supervised masked modeling) so that it can develop experts for object and action recognition, and for temporally-integrating vs. static processing (e.g. with stronger or weak attention weights across frames) (2) that dynamically switching between those experts is a design choice with potential for human-like capabilities (e.g. efficiency). Regarding (1), preliminary results on a single model (VideoMamba; see Appendix C) show that pure self-supervision yields the highest alignment in both stages I and II, but is surpassed by the fine-tuned version at later processing stages. This suggests that although purely self-supervised models may show high alignment due to their general objective (as also suggested in Tang et al. (2025)), they potentially lack expertise from supervised training on actions. Concerning (2), this dynamic switching cannot be found in block-based (i.e. CNN/Transformer) video models, as they process videos in limited fixed temporal blocks; any dynamic switching could then only occur within limited time spans. Architectures better suited to support this property include recurrent neural networks, such as SSMs or RNNs, as they process input sequentially, continuously updating their hidden states. Dynamic switching could then emerge implicitly or be enforced, potentially reducing computational costs via keeping overall network activation low. This exemplifies the potential of brain alignment insights for model building.

Limitations and future work. While our setup involved 100+ models, extending Sartzetaki et al. (2025), specific model types should be investigated in more depth, such as purely self-supervised models and models with different types of recurrent mechanisms (SSMs, LRUs, RNNs). Additionally, fine-grained differences between video models can be further explored by using instead of a coarse top-down grouping, a bottom-up grouping that focuses on variations in temporal feature generalization. Another promising direction is to study alignment with model-based EEG-fMRI fusion (Cichy & Oliva, 2020), combining EEGMD with BMD (Lahner et al., 2024), to leverage the complementary spatial and temporal resolution of these brain measurements.

REFERENCES

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

Brett B Bankson, Martin N Hebart, Iris IA Groen, and Chris I Baker. The temporal evolution of conceptual object representations revealed through models of behavior, semantics and deep neural networks. *NeuroImage*, 178:172–182, 2018.

Jan Benda. Neural adaptation. *Current Biology*, 31(3):R110–R116, 2021.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

Domenic Bersch, Martina G Vilas, Sari Saba-Sadiya, Timothy Schaumlöffel, Kshitij Dwivedi, Christina Sartzetaki, Radoslaw M Cichy, and Gemma Roig. Net2brain: A toolbox to compare artificial vision models with human brain responses. *Frontiers in Neuroinformatics*, 19:1515873, 2025.

Amber Marijn Brands, Sasha Devore, Orrin Devinsky, Werner Doyle, Adeen Flinker, Daniel Friedman, Patricia Dugan, Jonathan Winawer, and Iris Isabelle Anna Groen. Temporal dynamics of short-term neural adaptation across human visual cortex. *PLoS computational biology*, 20(5): e1012161, 2024.

Petr Byvshev, Pascal Mettes, and Yu Xiao. Are 3d convolutional networks inherently biased towards appearance? *Computer Vision and Image Understanding*, 220:103437, 2022.

Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019.

Radoslaw M Cichy and Aude Oliva. Am/eeg-fmri fusion primer: resolving human brain responses in space and time. *Neuron*, 107(5):772–781, 2020.

Radoslaw Martin Cichy, Dimitrios Pantazis, and Aude Oliva. Resolving human object recognition in space and time. *Nature neuroscience*, 17(3):455–462, 2014.

Radoslaw Martin Cichy, Gemma Roig, Alex Andonian, Kshitij Dwivedi, Benjamin Lahner, Alex Lascelles, Yalda Mohsenzadeh, Kandan Ramakrishnan, and Aude Oliva. The algonauts project: A platform for communication between the sciences of biological and artificial intelligence. *arXiv preprint arXiv:1905.05675*, 2019.

Radoslaw Martin Cichy, Kshitij Dwivedi, Benjamin Lahner, Alex Lascelles, Polina Iamshchinina, Monika Graumann, Alex Andonian, NAR Murty, K Kay, Gemma Roig, et al. The algonauts project 2021 challenge: How the human brain makes sense of a world in motion. *arXiv preprint arXiv:2104.13714*, 2021.

Colin Conwell, Jacob S Prince, Kendrick N Kay, George A Alvarez, and Talia Konkle. A large-scale examination of inductive biases shaping high-level visual representation in brains and machines. *Nature communications*, 15(1):9383, 2024.

Stephen V David, William E Vinje, and Jack L Gallant. Natural stimulus statistics alter the receptive field structure of v1 neurons. *Journal of Neuroscience*, 24(31):6991–7006, 2004.

Diana C Dima, Tyler M Tomita, Christopher J Honey, and Leyla Isik. Social-affective features drive human representations of observed actions. *Elife*, 11:e75027, 2022.

Adrien Doerig, Rowan P Sommers, Katja Seeliger, Blake Richards, Jenann Ismael, Grace W Lindsay, Konrad P Kording, Talia Konkle, Marcel AJ Van Gerven, Nikolaus Kriegeskorte, et al. The neuroconnectionist research programme. *Nature Reviews Neuroscience*, 24(7):431–450, 2023.

Nadiya El-Sourani, Moritz F Wurm, Ima Trempler, Gereon R Fink, and Ricarda I Schubotz. Making sense of objects lying around: How contextual objects shape brain activity during action observation. *Neuroimage*, 167:429–437, 2018.

Kathy Garcia, Emalie McMahon, Colin Conwell, Michael Bonner, and Leyla Isik. Modeling dynamic social vision highlights gaps between deep learning and humans. In *The Thirteenth International Conference on Learning Representations*, 2025.

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

Alessandro T Gifford, Domenic Bersch, Marie St-Laurent, Basile Pinsard, Julie Boyle, Lune Bellec, Aude Oliva, Gemma Roig, and Radoslaw M Cichy. The algonauts project 2025 challenge: How the human brain makes sense of multimodal movies. *arXiv preprint arXiv:2501.00504*, 2024.

Erin Goddard, Thomas A Carlson, Nadene Dermody, and Alexandra Woolgar. Representational dynamics of object recognition: Feedforward and feedback information flows. *Neuroimage*, 128: 385–397, 2016.

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pp. 5842–5850, 2017.

Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, et al. Meg and eeg data analysis with mne-python. *Frontiers in Neuroinformatics*, 7:267, 2013.

Michelle R Greene and Bruce C Hansen. Disentangling the independent contributions of visual and conceptual features to the spatiotemporal dynamics of scene categorization. *Journal of Neuroscience*, 40(27):5283–5299, 2020.

Iris IA Groen, Sennay Ghebreab, Hielke Prins, Victor AF Lamme, and H Steven Scholte. From image statistics to scene gist: evoked neural activity reveals transition from low-level natural image structure to scene category. *Journal of Neuroscience*, 33(48):18814–18824, 2013.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=tEYskw1VY2.

Umut Güçlü and Marcel AJ Van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014, 2015.

Matthias Guggenmos, Philipp Sterzer, and Radoslaw Martin Cichy. Multivariate pattern analysis for meg: A comparison of dissimilarity measures. *Neuroimage*, 173:434–447, 2018.

Uri Hasson, Eunice Yang, Ignacio Vallines, David J Heeger, and Nava Rubin. A hierarchy of temporal receptive windows in human cortex. *Journal of neuroscience*, 28(10):2539–2550, 2008.

Martin N Hebart, Oliver Contier, Lina Teichmann, Adam H Rockter, Charles Y Zheng, Alexis Kidder, Anna Corriveau, Maryam Vaziri-Pashkam, and Chris I Baker. Things-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *Elife*, 12:e82580, 2023.

Heejung Jung, Xiaochun Han, Ma Feilong, Jane Han, Deepanshi Shokeen, Cara Van Uden, Isabella Hanson, Andrew C Connolly, James V Haxby, and Samuel A Nastase. Action features dominate cortical representation during natural vision. *bioRxiv*, pp. 2025–01, 2025.

Kohitij Kar and James J DiCarlo. Fast recurrent processing via ventrolateral prefrontal cortex is needed by the primate ventral stream for robust core visual object recognition. *Neuron*, 109(1): 164–176, 2021.

Agnessa Karapetian, Alexander Lenders, Vanshika Bawa, Martin Pflaum, Raphael Leuner, Gemma Roig, Kshitij Dwivedi, and Radoslaw M Cichy. Investigating the temporal dynamics and modelling of mid-level feature representations in humans. *bioRxiv*, pp. 2025–03, 2025.

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

Nikolaus Kriegeskorte and Pamela K Douglas. Cognitive computational neuroscience. *Nature neuroscience*, 21(9):1148–1160, 2018.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:249, 2008.

Benjamin Lahner, Kshitij Dwivedi, Polina Iamshchinina, Monika Graumann, Alex Lascelles, Gemma Roig, Alessandro Thomas Gifford, Bowen Pan, SouYoung Jin, N Apurva Ratan Murty, et al. Modeling short visual events through the bold moments video fmri dataset and metadata. *Nature communications*, 15(1):6241, 2024.

Yulia Lerner, Christopher J Honey, Lauren J Silbert, and Uri Hasson. Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *Journal of neuroscience*, 31(8): 2906–2915, 2011.

Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. Videomamba: State space model for efficient video understanding. In *European conference on computer vision*, pp. 237–255. Springer, 2024.

Zejin Lu, Adrien Doerig, Victoria Bosch, Bas Krahmer, Daniel Kaiser, Radoslaw M Cichy, and Tim C Kietzmann. End-to-end topographic networks as models of cortical map formation and human visual behaviour. *Nature Human Behaviour*, pp. 1–17, 2025a.

Zejin Lu, Sushrut Thorat, Radoslaw M Cichy, and Tim C Kietzmann. Adopting a human developmental visual diet yields robust, shape-based ai vision. *arXiv preprint arXiv:2507.03168*, 2025b.

Emalie McMahon, Elizabeth Jiwon Im, Michael F Bonner, and Leyla Isik. Aspatiotemporal hierarchy for social in-teraction perception in the lateral visual stream. 2025.

Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(2):502–508, 2019.

Mathew Monfort, Bowen Pan, Kandan Ramakrishnan, Alex Andonian, Barry A McNamara, Alex Lascelles, Quanfu Fan, Dan Gutfreund, Rogério Schmidt Feris, and Aude Oliva. Multi-moments in time: Learning and interpreting models for multi-action video understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9434–9445, 2021.

Hamed Nili, Cai Wingfield, Alexander Walther, Li Su, William Marslen-Wilson, and Nikolaus Kriegeskorte. A toolbox for representational similarity analysis. *PLoS computational biology*, 10(4):e1003553, 2014.

Marc R Nuwer, Giancarlo Comi, Ronald Emerson, Anders Fuglsang-Frederiksen, Jean-Michel Guérit, Hermann Hinrichs, Akio Ikeda, Fransisco Jose C Luccas, and Peter Rappelsburger. Ifcn standards for digital recording of clinical eeg. *Electroencephalography and clinical Neurophysiology*, 106(3):259–261, 1998.

Pablo Oyarzo, Johannes JD Singer, Kohitij Kar, and Radoslaw M Cichy. Beyond feedforward: Leveraging discrepancies between humans and convolutional neural networks reveals recurrent dynamics during object recognition. 2024.

Brian E Russ, Kenji W Koyano, Julian Day-Cooney, Neda Perwez, and David A Leopold. Temporal continuity shapes visual responses of macaque face patch neurons. *Neuron*, 111(6):903–914, 2023.

Christina Sartzetaki, Gemma Roig, Cees GM Snoek, and Iris Groen. One hundred neural networks and brains watching videos: Lessons from alignment. In *The Thirteenth International Conference on Learning Representations*, 2025.

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J Majaj, Rishi Rajalingham, Elias B Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, et al. Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, pp. 407007, 2018.

Thomas Serre, Aude Oliva, and Tomaso Poggio. A feedforward architecture accounts for rapid categorization. *Proceedings of the national academy of sciences*, 104(15):6424–6429, 2007.

Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C Love, Erin Grant, Iris Groen, Jascha Achterberg, et al. Getting aligned on representational alignment. *arXiv preprint arXiv:2310.13018*, 2023.

Yingtian Tang, Abdulkadir Gokce, Khaled Jedoui Al-Karkari, Daniel Yamins, and Martin Schrimpf. Many-two-one: Diverse representations across visual pathways emerge from a single objective. *bioRxiv*, pp. 2025–07, 2025.

Rufin VanRullen and Simon J Thorpe. The time course of visual processing: from early perception to decision-making. *Journal of cognitive neuroscience*, 13(4):454–461, 2001.

Roel M Willems and Marius V Peelen. How context changes the neural basis of perception and language. *IScience*, 24(5), 2021.

Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23):8619–8624, 2014.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.

Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=YbHCqn4qF4.

Marius Zimmermann, Angelika Lingnau, Katharina Frey Doehler, Lisa Kaiser, Alice Stephan, and Julian Wieland. The spatiotemporal neural dynamics of action-related features underlying action recognition. 2025.

APPENDIX

A EEG DATASET DETAILS

EEG data during natural video viewing was collected for 6 participants, all with normal or corrected-to-normal vision. Stimuli were the exact same as in Lahner et al. (2024): 1102 3-second second videos, sampled from Monfort et al. (2019; 2021). Stimuli were square-cropped and resized to 268x268 pixels (5°×5° visual angle).

Stimuli were divided in non-overlapping sets, a "train set" of 1000 videos with 6 repeats and a "test set" of 102 videos with 24 repeats. The dataset is accompanied with crowd-sourced metadata, to annotate each clip with five word-level scene, object, and action labels, sampled from Places365 Zhou et al. (2017), THINGS Hebart et al. (2023), and Multi-Moments in Time Monfort et al. (2021) respectively. Additionally, provided were five sentence-level text descriptions, a spoken transcript, a memorability score and a memorability decay rate.

The experiment consisted of a passive viewing paradigm with an orthogonal detection task to have participants stay attentive, where participants had to report after X nr of random trials whether a specific object/scene/action was present. The total experiment consisted of 8 sessions, that each contained 3 repeats of the 250 train videos and 3 repeats of the 102 test videos. Each session consisted of 16 runs of 66 trials, in between which participants could take breaks.

Every trial sequence consisted of 1s of a blank baseline screen, followed by 3s showing a video, another 0.25s baseline and then 2s of blink time. In the case of prompted trials the second baseline was followed by a 3s prompt screen, which was then again followed by the 2s blink time. Stimuli were presented on a grey background with a red fixation cross present during the whole experiment.

EEG data was recorded using a 128-channel actiCAP set up with electrodees arranged according to the standard 10-10 system (Nuwer et al., 1998) and a Brainvision actiCHamp amplifier. Data was recorded at a sampling rate of 1000 Hz with online filtering (between 0.1 Hz and 100 Hz) and rereferenced (to Fz). We performed offline preprocessing using Python and MNE Gramfort et al. (2013). The continuous EEG data was epoched into trials from -0.2s to 3.5s with respect to stimulus onset and subsequently baseline corrected by subtracting the mean of the pre-stimulus interval for each trial and channel separately. The data was then temporally down- sampled to 50 Hz. Next, we performed multivariate noise normalisation (MVNN) based on the covariance matrices of each time-point to reweigh (un)reliable sensors and to (de)emphasise specific spatial frequencies, as recommended for multivariate pattern analysis methods like RSA Guggenmos et al. (2018). We performed analyses on two separate sets of electrodes: a posterior partition consisting of 35 electrodes and a frontal partition consisting of 54 electrodes. The posterior partition contained the following electrodes: 'Pz', 'P3', 'P7', 'O1', 'Oz', 'O2', 'P4', 'P8', 'P1', 'P5', 'PO7', 'PO3', 'POz', 'PO4', 'PO8', 'P6', 'P2', 'P9', 'PPO9h', 'PO9', 'O9', 'OI1h', 'PPO1h'. The frontal partition contained the following electrodes: 'Fp1', 'F3', 'F7', 'FT9', 'FC5', 'FC1', 'FT10', 'FC6', 'FC2', 'F4', 'F8', 'Fp2', 'AF7', 'AF3', 'AFz', 'F1', 'F5', 'FT7', 'FC3', 'FCz', 'FC4', 'FT8', 'F6', 'F2', 'AF4', 'AF8', 'F9', 'AFF1h', 'FFC1h', 'FFC5h', 'FTT7h', 'FCC3h', 'FCC4h', 'FTT8h', 'FFC6h', 'FFC2h', 'AFF2h', 'F10', 'AFp1', 'AFF5h', 'FFT9h', 'FFT7h', 'FFC3h', 'FCC1h', 'FCC5h', 'FTT9h', 'FTT10h', 'FCC6h', 'FCC2h', 'FFC4h', 'FFT8h', 'FFT10h', 'AFF6h'.

B MODEL DETAILS

Table 1: Model families. Action recognition models are trained on Kinetics 400 Kay et al. (2017) ; those also available on other datasets are marked by *a*,*b*.

	Image Object Recognition					Action Recognition					
	CNNs		Transformers		SSMs		CNNs		Transformers		SSMs	
1	AlexNet	2	CAiT	3	VideoMamba	6	CSN	2	MViTv2[b]	8	VideoMamba	
2	DenseNet	2	ConViT			5	I3D	2	TimeSformer			
2	EfficientNet	2	DEiT			1	R2P1D	2	Uniformer			Video
2	RegNet	2	MViTv2			2	SlowFast	2	Uniformerv2[a]			
4	ResNet	3	Swin			4	Slow[a]	1	VideoMAE			
2	ResNeXt	1	Twins			1	TaNet	2	VideoMAEv2			
4	VGG	2	ViT			1	TPN	3	VideoSwin[a]			
2	WideResNet					5	TSM[b]					
2	Inception					2	X3D					
2	RepVGG											
2	SeResNe(X)t					4	C2D	1	TimeSformer			Image
2	Xception					4	TSN[b]	1	TSN			
27		14		3		27+8		14+2		8		

[a] Availability also on Kinetics 710 (Carreira et al., 2019)
[b] Availability also on Something-Something-v2 (Goyal et al., 2017)

Table 2: Exhaustive account of all models.

Image Object Recognition			Action Recognition		
CNNs	Transformers	SSMs	CNNs	Transformers	SSMs
AlexNet	CAiT_S	VideoMamba_T	IR_CSN_R152	MViTv2_S[b]	VideoMamba_T_IN1k_f16
DenseNet161	CAiT_XXS	VideoMamba_S	IR_CSN_R152_BNfrozen_IG65M	MViTv2_B[b]	VideoMamba_T_IN1k_f8
DenseNet201	ConViT_S	VideoMamba_M	IR_CSN_R50_BNfrozen_IG65M	TimeSformer_DivST	VideoMamba_S_IN1k_f16
EfficientNetB3	ConViT_B		IR_CSN_R152_IG65M	TimeSformer_JointST	VideoMamba_S_IN1k_f8
EfficientNetB6	DEiT_S		IP_CSN_R152_IG65M	Uniformer_S	VideoMamba_M_IN1k_f16
RegNetX16gf	DEiT_B		IP_CSN_R152	Uniformer_B	VideoMamba_M_IN1k_f8
RegNetY8gf	MViTv2_S		I3D_R50	Uniformerv2_B[a]	VideoMamba_M_mask_f16
ResNet34	MViTv2_B		I3D_R50_dotprod	Uniformerv2_B_k710pre	VideoMamba_M_mask_f8
ResNet50	Swin_T		I3D_R50_embgauss	VideoMAE_B	
ResNet101	Swin_S		I3D_R50_gauss	VideoMAEv2_S	
ResNet152	Swin_B		I3D_R50_heavy	VideoMAEv2_S	
ResNeXt50	Twins_pcpvt_B		R2P1D_R50	VideoSwin_T	
ResNeXt101	ViT_S		SlowFast_R50	VideoSwin_S[a]	
VGG11	ViT_B		SlowFast_R101	VideoSwin_B	
VGG11BN			Slow_R50		
VGG19			Slow_R101		
VGG19BN			Slow_R50_IN1k[a]		
WideResNet50			Slow_R50_IN1k_embgauss		
WideResNet101			TaNet_R50		
InceptionV3			TPN_R50		
InceptionV4			TSM_R50[b]		
RepVGGa2			TSM_R50_dotprod		
RepVGGb2			TSM_R50_embgauss		
SeResNet50			TSM_R50_gauss		
SeResNeXt50			TSM_MobOne_s4		
Xception41			X3D_S		
Xception71			X3D_M		
			C2D_R50_nopool	TimeSformer_SpaceOnly	
			C2D_R101_nopool	TSN_Swin	
			C2D_R50_pool8		
			C2D_R50_pool16		
			TSN_R50[b]		
			TSN_R101		
			TSN_D161		
			TSN_MobOne_s4		

[a] Availability also on Kinetics 710 (Carreira et al., 2019)
[b] Availability also on Something-Something-v2 (Goyal et al., 2017)

C SUPPLEMENTARY ANALYSES

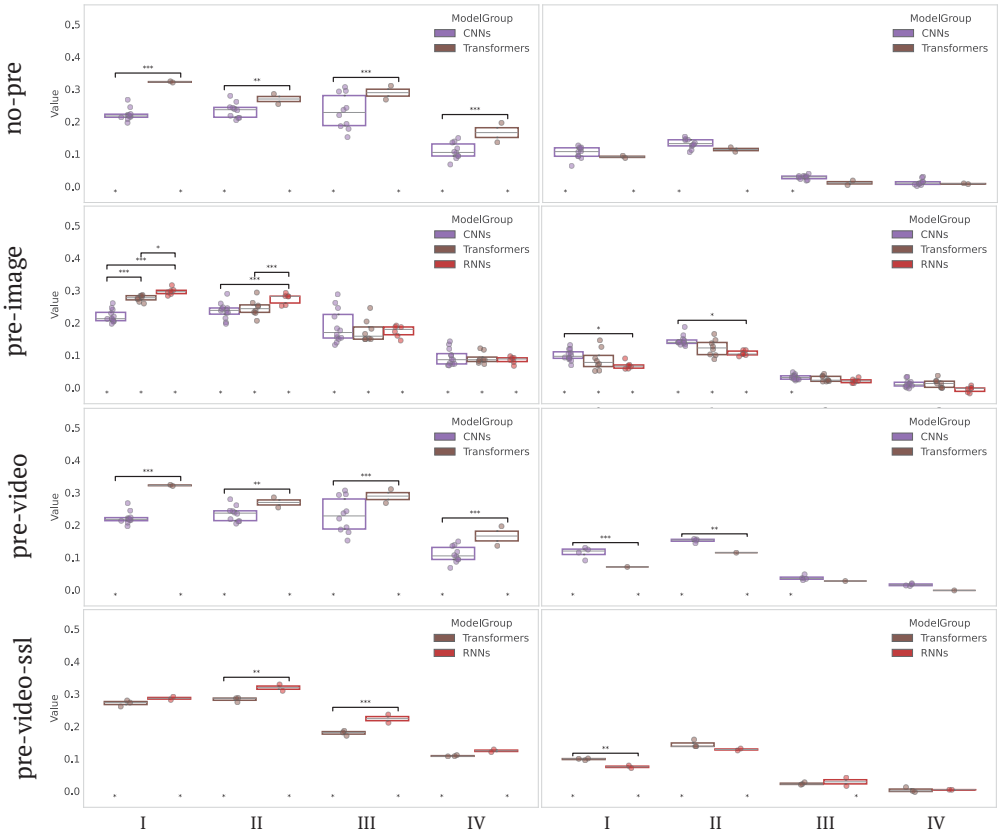

Figure 6: Additional control for pretraining in the architecture comparison of Fig. 4.

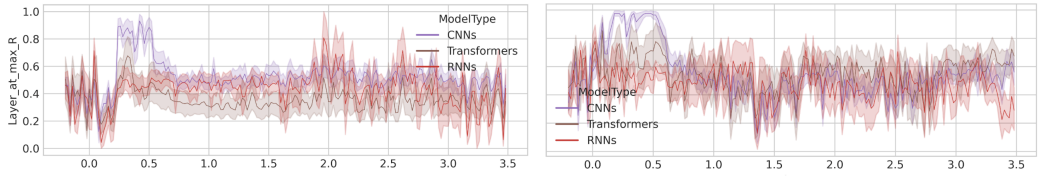

Figure 7: Showing the layer plot of Fig. 4b for the full timecourse.

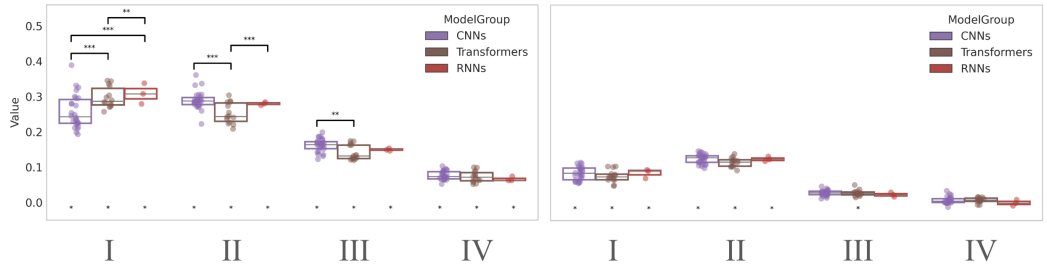

Figure 8: The comparison of Fig. 4 only now within object recognition models.

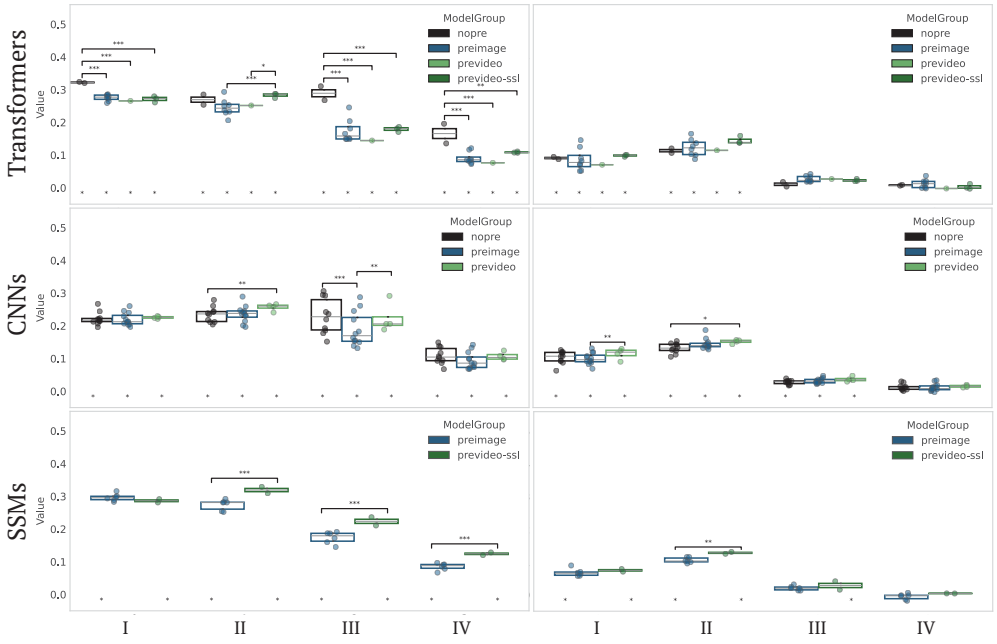

Figure 9: Additional control for architecture in the pretraining comparison of Fig. 5.

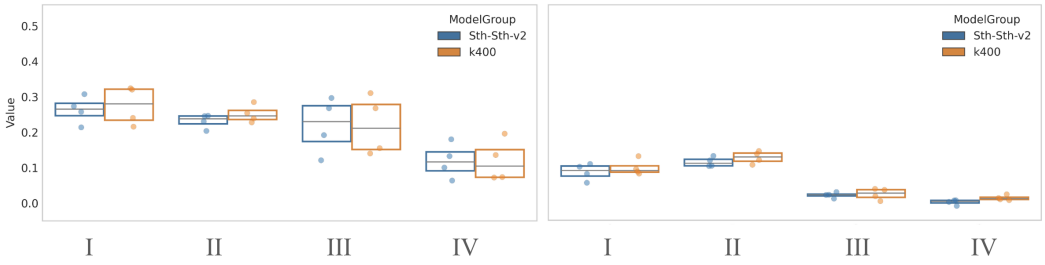

Figure 10: No significant differences between fine-tuning datasets (Sth-Sth-v2 vs. K400).

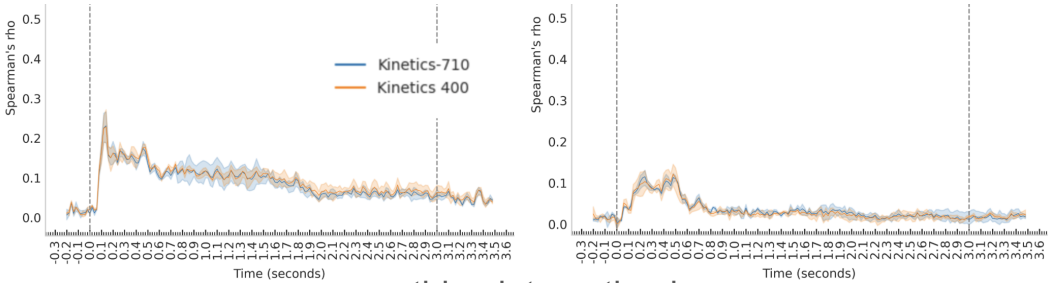

Figure 11: No significant differences between fine-tuning datasets (K710 vs. K400).

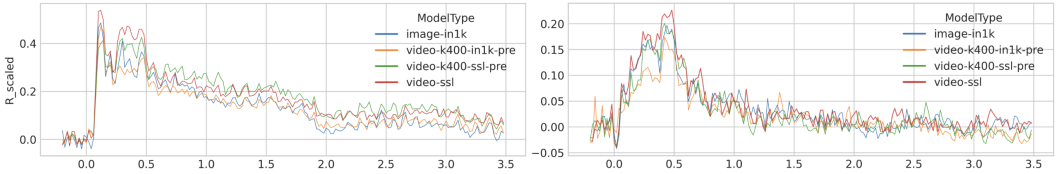

Figure 12: Comparing VideoMamba versions; object (blue), object-pretrained action (orange), self-supervised-pretrained action recognition (green), and pure self-supervised (red).