Action prediction relies on high-level holistic and low-level kinematics priors

Ingmar E.J. de Vries[1,2]*, Floris P. de Lange[2]†, Moritz F. Wurm[1]†

Affiliations

1 Centre for Mind/Brain Sciences, University of Trento, Rovereto, Italy.
2 Donders Centre for Cognitive Neuroimaging, Donders Institute for Brain, Cognition, and Behaviour, Radboud University, Nijmegen, The Netherlands

* Corresponding author: i.e.j.de.vries@gmail.com
† Shared last author.

Abstract

Our brains continuously predict what will happen next, helping us interact smoothly with others and navigate a dynamic world. A key example is biological motion perception, which relies on "priors," or lifelong expectations about how bodies move. Yet it remains unclear how these priors shape the dynamics of neural prediction across hierarchical levels. Using a new dynamic extension to representational similarity analysis (dRSA), we measured how the brain of human participants predicted continuous dance movements recorded with magnetoencephalography. Under normal viewing conditions predictions followed a hierarchy: high-level viewpoint-invariant body motion was predicted earliest, while viewpoint-dependent body motion and low-level pixelwise motion were predicted closer to real-time. Disrupting high-level long-term priors by inverting videos selectively reduced high-level predictions. Instead, disrupting low-level short-term kinematics priors by piecewise scrambling eliminated all motion prediction, while increasing post-stimulus responses. These findings reveal how high- and low-level priors jointly shape predictive perception of naturalistic continuous input.

Teaser

Disrupting high-level priors redirects neural prediction of naturalistic continuous input from high- to mid-levels: an MEG-RSA study

Introduction

Human beings evolved in a complex and ever-changing world. Adaptive behavior such as social interaction requires our brain to continuously update its representation of unfolding events. To safely cross a busy street, we require a real-time representation of the position and motion of other road users. This poses a fundamental challenge: there are inherent delays in processing the steady stream of input along the hierarchy of our sensory systems. Our brain solves this by continuously predicting the sensory input (*1–10*). It is able to do so because it maintains an internal model of the spatio-temporal structure of the world, which contains priors about what future sensory input probably looks like given the current input. While the brain's propensity to form such priors might be innate and automatic (*11–13*), the actual priors are progressively optimized to the statistics of our environment throughout an individual's life (*14*). They determine our perception and prediction of the world (*15–17*). For instance, biological motion perception crucially depends on our priors (*18–22*). Accurate biological motion perception is important as it helps us understand and predict other's actions, moods and intentions (*23*, *24*). Biological motion priors are deeply engraved in our internal model already since early infancy (*25–30*), and include information across hierarchical levels. High-level more stable priors include holistic information regarding configural relations among joints or body parts, global body structure, body form templates, and biomechanical and gravity constraints (*18*, *22*, *31–36*), while low-level priors rely on short-term contextual cues and include information about smooth kinematics including a preference for slower motions (*37–39*). Put simply, we implicitly know how fast bodies and limbs can move, how joints can move, we know about the center of mass and intuitive physics. This knowledge biases our perception, constraints future outcomes and should help us predict.

While the importance of our priors is clear, it remains unclear how these priors exactly determine the representational dynamics of predictions across hierarchical levels, i.e., what input features are predicted and when. Most empirical prediction research only captures a single snapshot or the consequence of prediction rather than the continuous predictions themselves, for instance by comparing the response to expected events and violations of that expectation (*40*, *41*). They evidence that we *can* predict, but not *what* and *when* exactly our brain predicts. Additionally, much empirical evidence for prediction comes from artificial experimental paradigms that use simple isolated (often static) stimuli. Predictability is often induced artificially, for instance by explicit cueing, or by imposing structured sequences of discrete stimuli that are only valid and learned within the context of the experiment (*42–45*). This is in stark contrast to our daily life where we typically make continuous implicit predictions based on our priors that are difficult to capture with these classic paradigms. This leaves critical ambiguities as to how our priors determine what our brain represents and predicts at any given moment in daily life, at which hierarchical level of processing these predictions take place, and how violations of our priors alter our brain's representations.

We recently developed a new dynamic extension to representational similarity analysis (dRSA) that quantifies exactly *what features* our brain represents *when* in rich naturalistic continuous input such as movies (*46*). The influential RSA approach captures the similarity between neural representations and static stimulus models to infer whether the stimulus feature captured by the model is represented in the brain (*47*, *48*). Dynamic RSA extends this to temporally variable models to capture the representational dynamics of unfolding events along a feature dimension of interest. In short, it quantifies strength of the match between continuous representations in the brain and continuous features of naturalistic input, across several hierarchical processing levels at the same time. Besides veridicality of representation (strength of match), it also quantifies at millisecond-precision latency how representations temporally relate to (follow or precede) actual events. Previously we applied this approach to magnetoencephalography (MEG) data and found a temporal hierarchy of predictive neural representations of the future motion of an observed ballet dancer such that high-level view-point invariant body motion was predicted at ~500 msec, view-point dependent body motion at ~200 msec, and low-level pixelwise motion closest to real-time at ~110 msec (*46*). Additionally, neural representation of view-point invariant body posture was represented close to real-time, also indicating an indirect effect of prediction. While these results suggest that our priors allow smooth biological motion to be represented mostly predictively, rather than bottom-up (*46*), it remains unclear to what extend these hierarchical predictive representations depend on our high- and low-level biological motion priors.

Here we apply dRSA to complex continuous action sequences (i.e., ballet dancing) to test how priors across hierarchical levels affect our neural predictions. Specifically, we disrupt high-level holistic priors of body

form and structure, and action structures, as well as gravity priors, by showing the action sequences up-down inverted, which is very different from our typical experience of biological motion, and is an effective way to impede biological motion perception and neural processing, i.e., the so-called inversion effect (*30*, *33*, *49–52*). Indeed, akin to the inversion effect in face (*53*, *54*) and body posture (*31*, *32*) recognition, it has been suggested that inversion in biological motion impairs holistic processing (*33*). At the same time, it leaves low-level visual dynamics intact, making low-level prediction still possible (*55–58*). We therefore hypothesize that disrupting high-level priors by inversion attenuates high-level predictions, and test to what extend this trickles down to lower levels. Additionally, in a separate manipulation we disrupt low-level short-term priors of smooth kinematics by temporal piecewise scrambling (200–500 msec pieces) the videos, which should attenuate prediction across all levels. The piecewise scrambled condition also allows us to test the crucial assumption of predictive processing theories that reducing the predictability of input should result in the representation of more unpredicted information or prediction errors reflecting the actual sensory input (*5*), and which here should be reflected in increased post-stimulus representations. For all our hypotheses we test both whether the veridicality (strength of match) and the latency of representation are affected.

Results

Behavioral task and results

To test the effect of hierarchical priors on predictive neural representations of naturalistic continuous input, we applied dRSA (*46*, *59*) to MEG data of 40 healthy human participants who observed 36 repetitions of 14 unique 5-second ballet dancing sequences under three different viewing conditions, i.e., 12 repetitions per condition (Fig. 1a). That is, participants observed the smooth biological motion sequences either normally (as in (*46*)), up-down inverted, or temporally piecewise scrambled, with pieces uniformly distributed between 200 – 500 msec (Fig. 1b). Participants attended the sequences well, as indicated by high performance on occasional catch trials in which the sequence was unexpectedly occluded for 400 msec, after which the participant had to indicate whether the sequence continued correctly or not (Fig. 1c; top). Specifically, participants responded 77 ± 10, 73 ± 11 and 74 ± 9 percent correct (mean ± std) on the normal, inverted and scrambled viewing conditions in the occlusion task, respectively (Fig. 1d; top, right), with anecdotal evidence for no difference between conditions ($F(2,78) = 2.05$, $p = 0.136$, $\omega^2 = 0.014$, BF = 0.467). Similarly, there was moderate evidence that participants responded equally fast on all three conditions in the occlusion task ($F(2,78) = 1.68$, $p = 0.194$, $\omega^2 = 0.001$, BF = 0.313), with reaction times of 697 ± 141, 716 ± 133, and 706 ± 135 msec on the normal, inverted, and scrambled conditions, respectively (Fig. 1d; bottom, right). While attending the videos participants were asked to keep fixation on a white fixation cross displayed on top of the video. Fixation was ensured by successful detection of an occasional subtle color change of the cross (Fig. 1c; bottom), to which participants responded by button press. Specifically, participants responded 100 ± 12, 94 ± 8, and 89 ± 14 percent correct (median ± IQR) on the normal, inverted and scrambled conditions in the fixation task, respectively (Fig. 1d; top, left), with strong evidence for a difference between conditions ($F(2,78) = 7.79$, $p < 0.001$, $\omega^2 = 0.074$, BF = 44.39), which a post-hoc test indicated was driven by a lower performance on the scrambled compared to both the normal condition ($t(39) = 3.54$, $p = 0.003$, $d = 0.686$, BF = 29.16) and the inverted condition ($t(39) = 3.08$, $p = 0.008$, $d = 0.583$, BF = 9.35). Similarly, there was strong evidence for a difference in reaction time between conditions in the fixation task ($F(2,78) = 7.01$, $p = 0.002$, $\omega^2 = 0.030$, BF = 19.34), with reaction times of 640 ± 145, 736 ± 126, and 757 ± 108 msec on the normal, inverted, and scrambled conditions, respectively (Fig. 1d; bottom, left). This difference was again driven by a slower reaction time on the scrambled compared to both the normal condition ($t(39) = 3.10$, $p = 0.011$, $d = 0.441$, BF = 9.97), and the inverted condition ($t(39) = 2.44$, $p = 0.039$, $d = 0.355$, BF = 2.31). The reduced performance on the fixation task for the scrambled condition is likely due to the color change of the cross being harder to detect amidst the sudden frame changes between scrambled video pieces. Importantly, the lack of a difference in performance on the occlusion task indicates that participants paid attention to the ballet dancer equally well across the three viewing conditions. Last, participants did not systematically choose to either focus only on the fixation cross, or only on the ballet dancer, which would be reflected in a negative correlation between tasks. Instead, there was moderate evidence for no correlation for percentage correct ($r = -0.008$, $p = 0.959$, BF = 0.197), and a positive correlation for reaction time ($r = 0.440$, $p = 0.005$, BF = 9.65), which suggests that participants divided attention across tasks as intended.

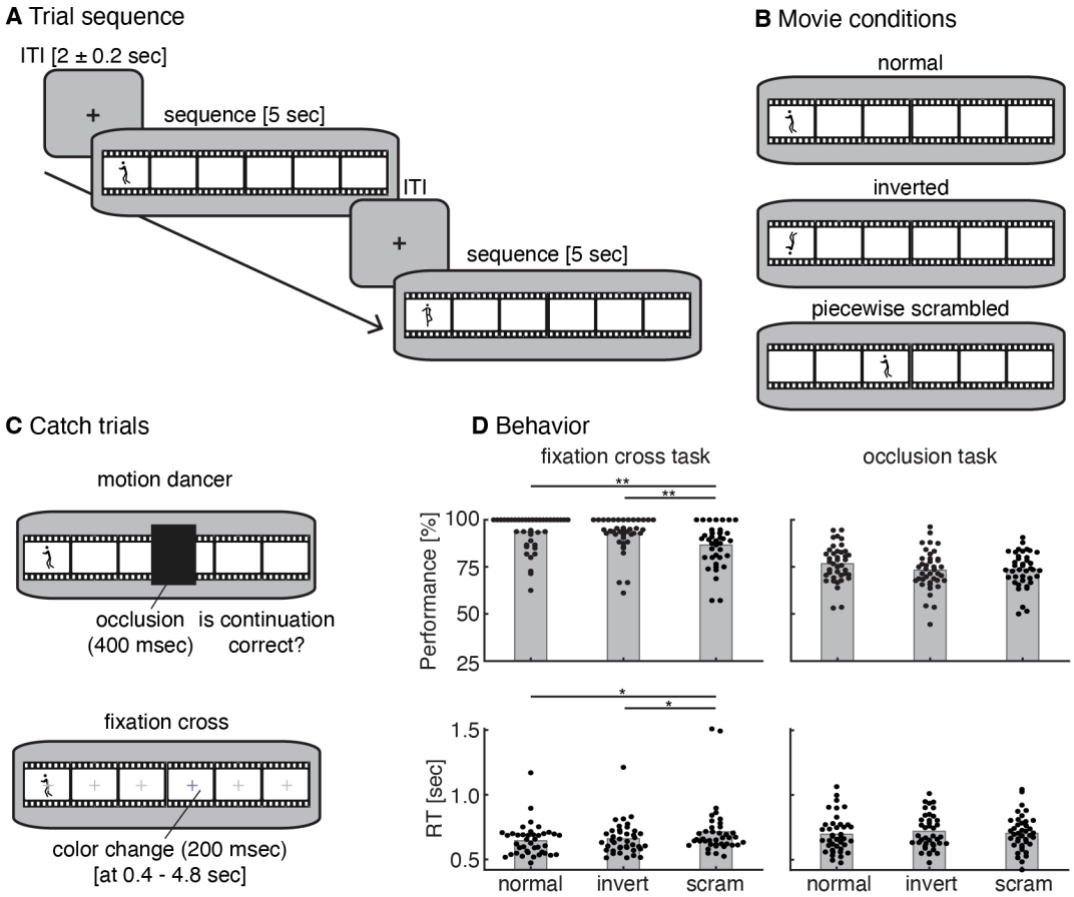

Figure 1. Experimental design and behavior. Please note that real ballet videos were used but for privacy reasons they are replaced in the preprint with illustrations of movie frames. (**A**) A run consisted of 108 trials alternated by a fixation cross (2 ± 0.2 sec, uniformly distributed). A trial consisted of one of 14 unique 5-sec ballet dancing sequences, such that each unique sequence was shown at least 7 times per run. (**B**) Participants observed the ballet sequences in 3 different viewing conditions (33% of trials per condition); normal, up-down inverted, and temporally piecewise scrambled, with pieces uniformly distributed between 200 – 500 msec. (**C**) Top: To ensure attention to the ballet sequence, in 16 trials per run the video was occluded by a black screen for 400 msec, after which the video continued either correctly, or with an incorrect sequence. Participants (dis)confirmed by button press the correct continuation of the video. Bottom: To encourage fixation, on 8 trials per run the fixation cross changed color for 200 msec at a random time between 0.4 and 4.8 seconds after onset, and participants pressed a button upon detection. (**D**) Behavioral results. Percentage correct and trial-averaged RT on catch trials in top and bottom panels, respectively. RT on the fixation task is counted from color change onset, while RT on the occlusion task is counted from occlusion offset. Dots represent single-participant data (n = 40 healthy human participants). Bars represent the group mean. ITI inter-trial interval. * $p < 0.05$, ** $p < 0.01$.

Biological motion is hierarchically predicted across levels of processing

Next, we applied dRSA to the MEG data (for details see Materials and Methods and Fig. S3), to quantify the strength of the match between continuous representations in the brain and continuous feature models of naturalistic input, across several hierarchical processing levels at the same time. The dance videos were modeled at various levels of abstraction across the visual processing hierarchy, from low-level visual (i.e., pixelwise luminance and motion), to mid- and higher-level, perceptually more invariant (i.e., 3-dimensional view-dependent and view-invariant body posture and motion), as to capture a comprehensive characterization of the dance sequences. dRSA quantifies both the veridicality of representation (strength of match), and at millisecond-precision how representations temporally relate to (follow or precede) actual events, as reflected in the peak magnitude and peak latency in dRSA latency plots, respectively (Fig. 2 and 6). In case of pure feedforward processing, one expects a lag between the model and the best-matching neural representation (i.e., due to information transfer from retina to V1), as reflected in a peak to the right of the vertical zero-lag midline. Prediction should reduce this lag, potentially to negative (i.e., left of midline), in which case neural representation predicts the future model state.

First, we replicated our previous findings in the normal viewing condition (*46*) and found that the motion of the ballet dancer was predicted at three distinct times that reflected the order in the visual processing hierarchy, such that high-level view-invariant body motion was predicted furthest in advance at ~500 msec preceding the actual input, while view-dependent body motion and low-level pixelwise motion were predicted at ~170 and ~150 msec, respectively (Fig. 2a and Table S1; see Fig. S1 for all models). We ran a [3x4] ANOVA on

retrieved jackknifed peak latencies with factors model (pixelwise motion, view dependent and view invariant body motion) and ROI, with the first 4 ROIs (i.e., V1, V2, V34, and LOTC) that showed a significant dRSA curve in at least one of the 3 models, but also did an exact replication of our previous study by including only the first 3 ROIs as in (*46*). Most importantly, we replicate the previous findings, i.e., a significant main effect of model on the latency of the three predictive peaks ($F(1.6,64.3) = 63.43$, $p < 0.001$, $\omega^2 = 0.514$, BF = $1.374*10^{14}$ and $F(1.8,69.5) = 57.98$, $p < 0.001$, $\omega^2 = 0.496$, BF = $1.375*10^{13}$, excluding or including LOTC, respectively).

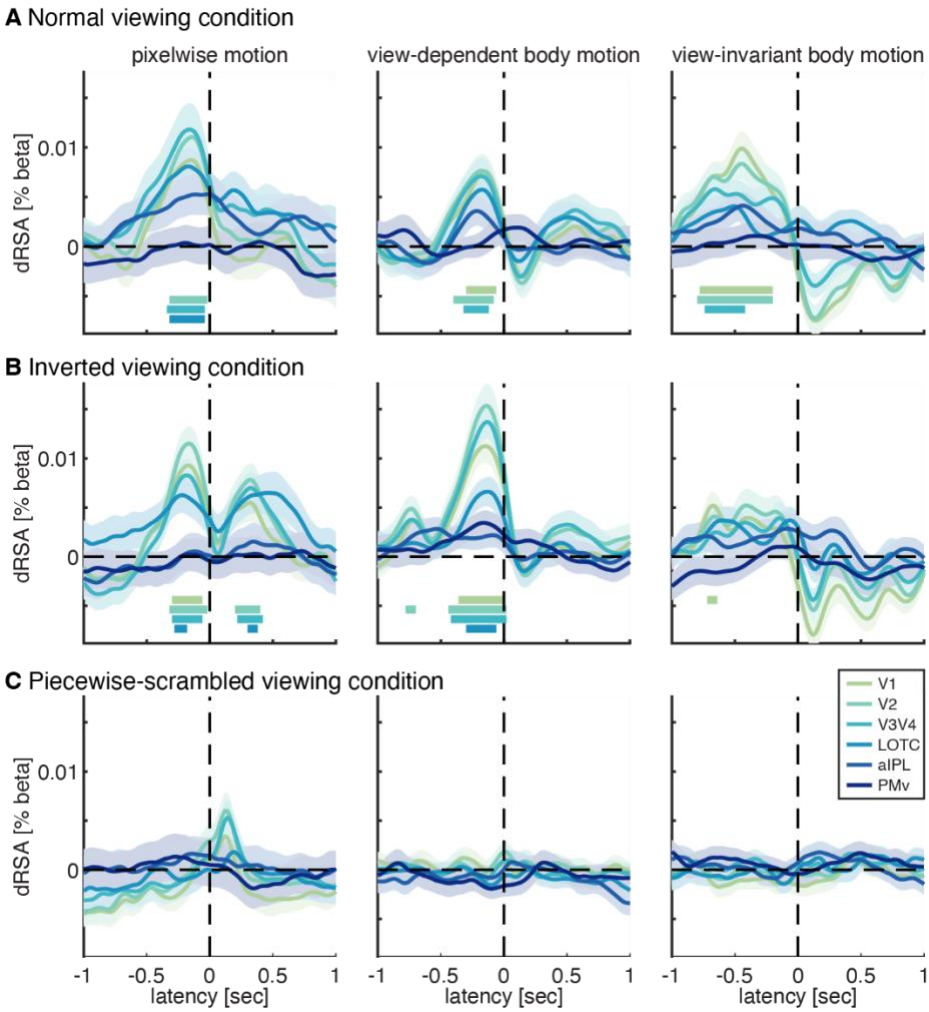

Figure 2. Dynamic RSA results for motion models. Region of interest (ROI)-based analysis, with normalized dRSA regression weights illustrated as latency plots. The three viewing conditions are displayed in rows with normal on top (**A**), inverted in the middle (**B**), and temporally piecewise scrambled at the bottom (**C**). Different stimulus models are displayed in columns with pixelwise motion reflecting optical flow vector direction in each pixel, and body motion reflecting 3D motion of 13 kinematic markers placed on the dancer. Lines and shaded areas indicate participant-average and SEM, respectively, with n = 40 in (**A**) and (**B**), and n = 37 in (**C**). Light-to-dark colors indicate posterior-to-anterior ROIs. Horizontal bars indicate beta weights significantly larger than zero (one-sided t-test with $p < 0.001$ for each time sample), corrected for multiple comparisons across time using cluster-based permutation testing ($p < 0.05$ for cluster-size), with colors matching the respective ROI line plot.

Movie inversion attenuates high-level predictions while enhancing mid-level predictions

To test our hypothesis that disrupting high-level priors attenuates biological motion prediction of specifically high-level (view-invariant) body motion, we compared the motion models between the normal and up-down inverted viewing conditions (Fig. 2a and b, respectively) and indeed observed a significant reduction in the predictive representation of view-invariant body motion. To statistically compare conditions, we averaged dRSA magnitudes in a 100 msec window surrounding a priori selected predictive peaks based on our previous study (i.e., 110 msec, 180 msec, and 510 msec, for pixelwise motion, view-dependent, and view-invariant body motion, respectively (*46*)), and performed a repeated measures ANOVA for each model separately (i.e., time-averages are not comparable between models with different temporal autocorrelation profiles), with [2 condition by 4 ROIs] for pixelwise motion and view dependent body motion, and [2 conditions by 3 ROIs] for view invariant body motion, i.e., taking only ROIs where at least one condition showed a significant dRSA curve (Fig. 2a and b; see Fig. 3 for individual-participant peak magnitudes and condition contrasts). Most

importantly, in line with our hypothesis, prediction at the highest level of view invariant body motion was reduced in the inverted compared to normal viewing condition (Fig. 3c; $F(1,39) = 16.37$, $p < 0.001$, $\omega^2 = 0.095$, BF = 5.366). Surprisingly, prediction of view dependent body motion showed the opposite pattern and was stronger for inverted compared to normal ($F(1,39) = 8.65$, $p = 0.005$, $\omega^2 = 0.054$, BF = 7.955). Last, there was anecdotal evidence for no condition difference for pixelwise motion ($F(1,39) = 0.41$, $p = 0.526$, $\omega^2 < 0.001$, BF = 0.349), indicating that low-level motion prediction is minimally, if at all, affected by high-level priors. These effects together suggest that for naturalistic but up-down inverted biological motion, participants still make predictions of (body) motion but rely on the level at which it is possible for them to make predictions (i.e., view-dependent). In other words, disrupting high-level priors specifically attenuates the highest-level predictive representations, while channeling prediction to lower processing levels. Interestingly, this change in neural strategy seems effective, given that both accuracy and reaction speed on the occlusion task, which arguably requires prediction, were only numerically lower for inverted than normal (Fig. 1d), with anecdotal-to-moderate evidence for no condition difference. A less likely alternative is that people mostly rely on low-level predictions for the task in any case, which are comparable between conditions.

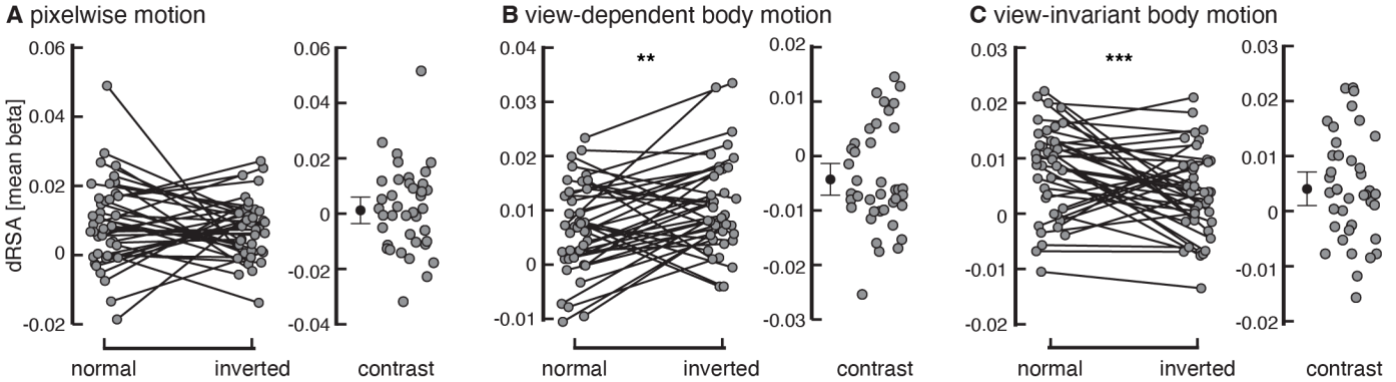

Figure 3. dRSA peak magnitudes for motion models in normal and inverted conditions. Individual participant (n = 40) peak magnitudes reflect the average in the ±50 msec window surrounding a priori selected peak latencies based on previous study. Peaks are also averaged over the first 4 ROIs for the pixelwise motion and view-dependent body motion models (in **A** and **B,** respectively), or the first 3 ROIs in the view-invariant body motion model (in **C**). Dots reflect individual participants in the normal (left) and inverted (right) viewing conditions, connected with thin lines, or in the condition contrast (normal – inverted). The black dot with error bars in the contrast plot reflects the mean ± 95% confidence interval of the condition contrast. ** $p < 0.01$, *** $p < 0.001$.

To test whether up-down inversion also affected the latency of predictive representations, we used bootstrapping with 10.000 bootstrap samples to compare the peak latencies (see Methods) between the normal and inverted conditions, with the hypothesis of earlier prediction (more negative latency) for normal compared to inverted. Specifically, we tested whether the one-tailed 95% confidence interval of the bootstrapped distribution of normal minus inverted peak latencies was lower than zero. Before bootstrapping, we first averaged the dRSA curves over the same ROIs as before that showed significant curves, since peak latency estimation does not make sense for noisy, non-significant and flat curves. In line with the peak magnitudes, we found that only view-invariant body motion was predicted significantly earlier in the normal compared to inverted condition (-448 vs -352 msec, 95% CI: -6 msec), while this was not the case for view-dependent body motion (-165 vs -177 msec, 95% CI: 81 msec), or for pixelwise motion (-217 vs -191 msec, 95% CI: 27 msec). Additionally, we performed the exact same analysis on the centroid latency (i.e., center of mass of the dRSA curve; see Methods). The centroid latency considers latency differences in representational weight that are not captured by the peak (e.g., for non-symmetrical curves) and might be less sensitive to noise than peak latency estimates. Importantly, the centroid analysis showed the exact same pattern where only view-invariant body motion was predicted significantly earlier in the normal compared to inverted condition (-498 vs -418 msec, 95% CI: -16 msec), while this was not the case for view-dependent body motion (-174 vs -172 msec, 95% CI: 38 msec), or for pixelwise motion (-239 vs -227 msec, 95% CI: 16 msec). The latency effects are in line with the magnitude effects, which together show that only view-invariant body motion prediction was affected by inversion.

Another potentially more sensitive way of testing for a latency difference between viewing conditions is by running dRSA between the neural RDMs of the normal and inverted conditions, i.e., computing similarity between the neural RDMs across latencies (Fig. 4a). We again used bootstrapping with 10.000 bootstrap samples to compare the peak latency against zero in the 100 msec window surrounding zero, with the

hypothesis that stimulus representations are similar but delayed in the inverted compared to normal condition, which would result in a negative peak latency. Specifically, we tested whether the one-tailed 95% confidence interval of the bootstrapped distribution of peak latencies was lower than zero. Peak latency was increasingly negative for more anterior areas (Fig. 4a, right panel) with significantly negative latencies at LOTC (-8 msec, 95% CI: -1 msec) and V3/V4 (-2 msec, 95% CI: -2 msec), but not at V1 (0 msec, 95% CI: 12 msec) or V2 (2 msec, 95% CI: 9 msec). Since shared variance between the neural RDMs might be largely driven by low-level visual features that are not predicted, we reran the same analysis but used partial correlation as similarity measure to partial out pixelwise luminance and pixelwise motion magnitude, two low-level visual models that showed strong lagged representations (Fig. S1). This analysis revealed qualitatively equivalent results (Fig. 4b), with significant negative latencies at LOTC (-10 msec, 95% CI: -1 msec) and V3/V4 (-2 msec, 95% CI: -1 msec), but not at V1 (0 msec, 95% CI: 8 msec) or V2 (2 msec, 95% CI: 7 msec). Importantly, this analysis reveals that neural representation of ballet dancing sequences as a whole (i.e., without dissociating between features), is similar between viewing conditions, and in higher visual areas (V3/4 and LOTC) is delayed for the inverted relative to the normal condition. The main dRSA analysis suggests that this is driven by weaker and later prediction of specifically high-level view-invariant body motion.

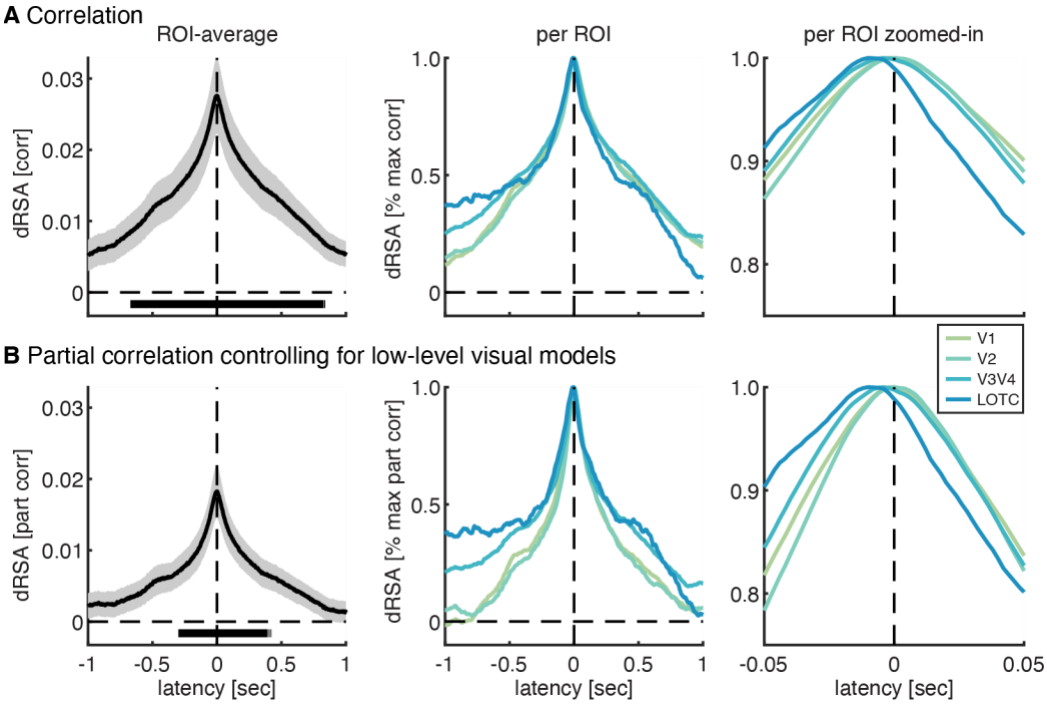

Figure 4. dRSA between neural RDMs of normal and inverted conditions. Similarity between neural RDMs was computed as correlation (**A**), or partial correlation (**B**), where two low-level visual models with lagged representations were partialed out to account for a large part of shared variance between neural RDMs being explained by low-level visual features (i.e., pixelwise luminance and pixelwise motion magnitude; see Methods). The first column reflects the average over the first 4 ROIs. Lines and shaded areas indicate participant-average (n = 40) and SEM, respectively. The horizontal black bars indicate correlation coefficients significantly larger than zero (one-sided t-test with $p < 0.001$ for each time sample), corrected for multiple comparisons across time using cluster-based permutation testing ($p < 0.05$ for cluster-size). The second column reflects the participant-average per ROI, with light-to-dark colors reflecting posterior-to-anterior ROIs, while the third column shows the same but zoomed in. Note that the first column reflects absolute correlation coefficients, while the second and third column reflect proportion of maximum correlation coefficient per ROI. Second and third column are smoothed using a 20 msec sliding window for plotting purposes only.

Piecewise scrambling eliminates predictive representations and causes prediction errors

We hypothesized that if we disrupt low-level short-term kinematics priors by temporal piecewise scrambling, all predictive motion representations would be strongly attenuated, which is indeed what we observed (Fig. 2c; i.e., no significant predictive representations). To compare peak magnitudes, we extracted peak magnitudes at a priori selected peak latencies based on our previous study(*46*), for each participant, condition, ROI and model for statistical comparison using a [2 conditions by 4 ROIs by 3 models] repeated measures ANOVA (see Fig. 5a for individual-participant peak magnitudes and condition contrasts). We used peak magnitudes instead of average magnitudes in a 100 msec window as for the normal vs inverted comparison, since the normal condition models have broader temporal autocorrelations compared to the scrambled condition (Fig. S2), thus biasing averages upwards. Confirming our hypothesis, magnitudes were indeed significantly attenuated for the predictive motion peaks in the scrambled compared to normal condition (i.e.,

main effect of condition; $F(1,36) = 57.27$, $p < 0.001$, $\omega^2 = 0.397$, BF = 1.97×10^6). There was also a minor effect of ROI ($F(2,72.2) = 4.98$, $p = 0.009$, $\omega^2 = 0.028$, BF = 0.633), but moderate evidence for no main effect of model ($F(1.5,53.6) = 1.40$, $p = 0.253$, $\omega^2 = 0.007$, BF = 0.235) and no interactions ($p > 0.101$, BF < 0.256).

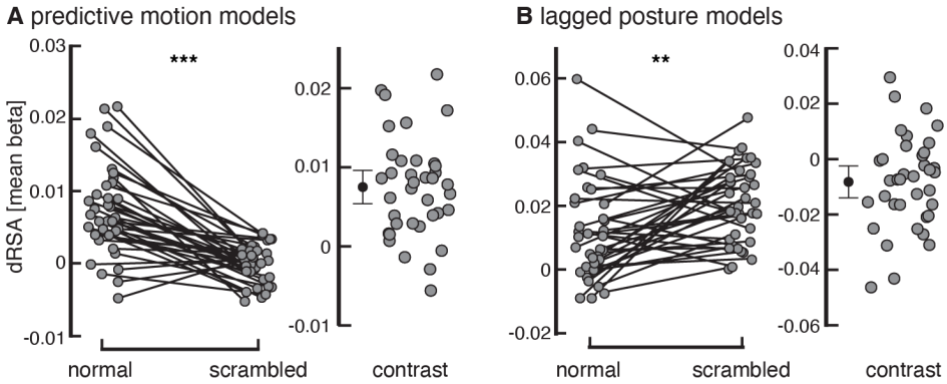

Figure 5. dRSA peak magnitudes in normal and scrambled conditions. Individual participant (n = 37) peak magnitudes. Peaks are averaged over the first 4 ROIs and the 3 predictive motion models (**A**) or 2 the lagged posture models (**B**). Dots reflect individual participants in the normal (left) and scrambled (right) viewing conditions, connected with thin lines, or in the condition contrast (normal – scrambled). The black dot with error bars in the contrast plot reflects the mean ± 95% confidence interval of the condition contrast. ** $p < 0.01$, *** $p < 0.001$.

We further expected that disrupting low-level priors by temporal piecewise scrambling may result in an increase in post-event (lagged) representations of static features such as body posture, which is indeed what we observed (Fig. 6). To compare peak magnitudes between the normal and scrambled condition for the body posture models we did not have very reliable a priori peak latencies since these representations were weak in our previous study(*46*). We therefore estimated peak latencies based on the grand-average dRSA latency plots (averaged over participants, normal and scrambled conditions, and the first four ROIs), and extracted peak magnitudes at that latency for each participant, condition, ROI and model for statistical comparison using a [2 conditions by 4 ROIs by 2 models] repeated measures ANOVA (see Fig. 5b for individual-participant peak magnitudes and condition contrasts). Confirming our expectation, magnitudes were significantly larger for the lagged posture peaks in the scrambled compared to normal condition (i.e., main effect of condition; $F(1,36) = 8.79$, $p = 0.005$, $\omega^2 = 0.070$, BF = 6.58). There was also anecdotal evidence for a small main effect of model ($F(1,36) = 5.67$, $p = 0.023$, $\omega^2 = 0.060$, BF = 2.47), and for an interaction between model and ROI ($F(2.2,80.4) = 3.22$, $p = 0.040$, $\omega^2 = 0.022$, BF = 1.22), but moderate evidence for no main effect of ROI ($F(2.0,71.2) = 2.83$, $p = 0.066$, $\omega^2 = 0.017$, BF = 0.269) and no other interactions ($p > 0.152$, BF < 0.459). These results indicate that if the congruent continuation of a smooth biological motion sequence cannot be predicted, the post-event representation of body posture increases. Such post-event representation of unpredicted information (or prediction errors) is in line with predictive coding theories. Last, we tested whether piecewise scrambling also affected the latency of body posture representations, using bootstrapping as above to test the expectation that piecewise scrambling results in later representations (more positive latency) compared to normal viewing, which would be evidenced by the one-tailed 95% confidence interval of the bootstrapped distribution of normal minus scrambled peak latencies being lower than zero. We found that only the view-invariant body posture was represented delayed in the scrambled compared to normal condition (194 vs 65 msec, 95% CI: -43 msec), while this was not the case for view-dependent body posture (105 vs 133 msec, 95% CI: 101 msec). The centroid latency analysis confirmed this result with view-invariant body posture being delayed in the scrambled compared to normal condition (176 vs 93 msec, 95% CI: -20 msec), while this was not the case for view-dependent body posture (94 vs 187 msec, 95% CI: 149 msec).

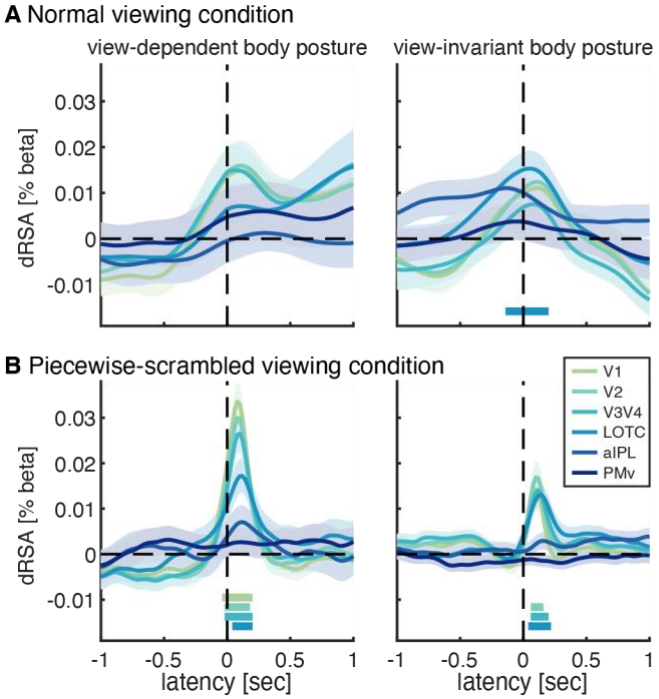

Figure 6. Dynamic RSA results for body posture models. Region of interest (ROI)-based analysis, with normalized dRSA regression weights illustrated as latency plots. Two viewing conditions are displayed in rows with normal on top (**A**) and temporally piecewise scrambled below (**B**). Different stimulus models are displayed in columns with body posture capturing 3D position of 13 kinematic markers placed on the dancer. Lines and shaded areas indicate participant-average and SEM, respectively, with n = 40 and 37 in **A** and **C**, respectively. Light-to-dark colors reflect posterior-to-anterior ROIs. Horizontal bars indicate beta weights significantly larger than zero (one-sided t-test with $p < 0.001$ for each time sample), corrected for multiple comparisons across time using cluster-based permutation testing ($p < 0.05$ for cluster-size), with colors matching the respective ROI line plot.

Discussion

In this study we tested how high- and low-level priors modulate our neural prediction of naturalistic continuous input across hierarchical levels. To this end we applied dynamic RSA to MEG data of healthy human participants observing smooth biological motion sequences (i.e., ballet dancing videos). We disrupted high-level holistic priors of body form and structure, action structures, and gravity, by showing the action sequences up-down inverted, which leaves low-level visual dynamics intact. We disrupted low-level short-term priors of smooth biological kinematics by temporal piecewise scrambling (200–500 msec pieces) the videos. First, we replicated our previous findings (*46*) and found that neural representation of normally viewed smooth biological motion is mostly predictive, with time windows matching hierarchical levels of processing such that high-level view-invariant body motion was predicted earliest, while view-dependent body motion and low-level pixelwise motion were predicted closer to real-time. These findings extend a growing body of evidence for hierarchical prediction (*60–64*), by demonstrating the necessary time windows for such hierarchy. Next, in line with our hypothesis, up-down inversion attenuated neural prediction selectively at the highest processing level of view-invariant body motion, in terms of both strength and prediction latency, while this was not the case for view-dependent body motion and pixelwise motion. In fact, prediction of view-dependent body motion increased, suggesting that predictions were channeled to a level at which prediction was still possible. Interestingly, this change in neural strategy seemed effective, given that performance on the occlusion task, which arguably requires prediction, was only numerically lower for inverted viewing. Low-level pixelwise motion prediction was not affected by inversion at all, indicating that it does not rely on high-level priors. Next, confirming our hypothesis, temporal piecewise scrambling completely eliminated neural prediction at all levels. Instead, lagged post-stimulus neural representations for static body posture increased significantly and became more delayed, which might reflect the unpredicted sensory input or prediction errors.

Accurate biological motion perception and prediction is crucial for our daily social interactions as it helps us understand and predict other's actions, moods and intentions (*23*, *24*). Biological motion priors are deeply engraved in our internal world model already since in early infancy (*27–30*). For instance, biological motion is easily and automatically recognized and distinguished from nonbiological motion in moving dot patterns (i.e., point light displays) in young infants (*27*), and improves further throughout early childhood until reaching adult-level skill at an age of 5 (*28*). It may even be innate as some research found a preference for biological

motion already at birth and across species (*29*, *30*). Biological motion perception crucially depends on both high-level more stable long-term priors that include holistic information about body parts and their configurations, global body structure, and body form, as well as intuitive physics such as gravity that constraints acceleration direction (*22*, *31–34*, *37*, *51*), and low-level short-term priors that include information about smooth biological kinematics (*22*). Evidence for the persistence of high-level priors comes from a series of studies that compared perception of a human-like robot moving mechanically (i.e., incongruency with priors) with an actual robot moving mechanically and a human moving naturally (i.e., congruency), and found that a human-like robot moving mechanically resulted in both behavioral and neural signatures of prediction errors (*18–20*). Additionally, they found that while short-term kinematics priors are used, long-term high-level priors take precedence if available (*22*), evidencing the strength of the high-level priors we have for biological motion (*19*). Priors based on earth's gravity also play an important role in biological motion perception. While the visual system typically has a hard time estimating acceleration of moving objects, acceleration expectations due to gravity are formed in early infancy (*25*), and provide an highly persistent perceptual bias already at an age of 2 (*26*), which persists into adulthood, and is difficult to override (*35*).

While clearly our priors are important for biological motion perception, previous studies did not investigate representational dynamics, leaving it unclear exactly how our priors affect (predictive) representations across hierarchical levels, i.e., what features are predicted when. Here we tested this by disrupting long-term priors by up-down inversion. Inversion should affect our high-level holistic and gravity priors, both of which are built on life-long experience. At the same time, it leaves low-level short-term kinematics priors intact. Indeed, inversion is an effective way to impede biological motion perception and attenuate neural responses (*30*, *49–52*), and newborns have a natural preference for upright versus inverted biological motion (*30*). Such inversion effects have also been observed for static body postures and face recognition. For instance, while possible body postures are recognized slower and less accurately when inverted, this inversion effect is diminished for impossible postures that violate biomechanical constraints (*32*). Similarly, while memory for upright postures is biased by both biomechanical and gravitational priors, memory for inverted postures is not (*31*). These effects are mirrored for face recognition (*54*, *65*, *66*), which is impaired when inverted, but only when recognition relies on holistic processing, not when it relies on featural processing (*53*). While the inversion effect for body postures and faces is typically interpreted as evidence for holistic processing (*32*, *66*), it seems that for biological motion perception also gravitational priors play an important role. That is, when global configuration is destroyed by spatial scrambling, the inversion effect partly remains, which seems mostly due to local biomechanical constraints and acceleration effects, especially of the feet (*37*, *49*, *67*). However, as these studies were restricted to motion direction estimation in simply gait, this unlikely generalizes to predictive processing of more complex three-dimensional movements such as the dance sequences presented here, where high-level holistic priors probably play a more important role. In any case, it is clear that inversion disrupts our long-term experience-based priors. We extend the current literature by showing how inversion affects predictive representational dynamics of biological motion perception. Specifically, inversion does not simply attenuate neural prediction at high processing levels (i.e., view-invariant body motion) but rather redirects prediction to a lower level (i.e., view-dependent body motion). It seems that the view-invariant prediction of biological motion relies on high-level holistic priors, while prediction of view-dependent biological motion may rely less on such long-term stable priors, and more on short-term contextual kinematics priors such as local acceleration effects that are less affected by inversion (*37*, *49*).

We tested the effect of short-term dynamic priors of local kinematics on predictive representations by temporal piecewise scrambling, and found that all motion prediction, including low-level pixelwise motion, was eliminated. Since pieces were between 200 and 500 msec, short-term local kinematics remained intact, and one could expect pixelwise motion prediction that had a latency of only ~150 msec to survive. It could be that indeed mostly higher-level body motion prediction is affected, but that this trickles down to the hierarchically lower pixelwise motion prediction. However, neural prediction of low-level stimuli such as (apparently) moving dots without naturalistic hierarchical embedding is also possible (*55*, *56*). While removing such embedding likely attenuates low-level prediction, it seems unlikely that low-level prediction would disappear altogether. A more likely explanation is that on average our pieces were too short (i.e., 350 msec) to 'start up' any prediction. It should take some time to set up short-term kinematics priors based on local context, and it is an open question how much time is needed after onset of an uninterrupted video segment. Earlier research using temporal piecewise scrambling with ~4 sec pieces in an fMRI movie paradigm found that only higher

area responses were disrupted, while responses in early visual areas were not (*68*). However, ~4 seconds is too long to investigate how low-level visual dynamics are affected. Future research should test how much time is needed to set up our hierarchical priors, for instance by systematically manipulating piece lengths to see at which length high- and low-level prediction disappears. Interestingly here, representational content was dominated by increased and slightly delayed lagged representation for both view-dependent and view-invariant body posture, indicating that if the congruent continuation of a smooth biological motion sequence cannot be predicted, the post-stimulus representation of body posture increases. Stronger post-stimulus representation of unpredicted information is in line with predictive coding, which assumes that sensory information is only processed further post-stimulus if it has not been predicted, or violates a prediction, i.e., a prediction error (*5*, *8*), which is likely the case here. It is also interesting to note that while here we artificially induce sudden segment boundaries, our daily life is filled with event boundaries. Such naturalistic event boundaries are mirrored in hierarchical boundary-induced activation patterns from short events in sensory regions to long events in high-order areas (*69*). Interestingly, such boundary-induced activity patterns in response to repeated movie viewings resemble but precede those of first viewings (*70*, *71*), suggesting neural prediction. Moreover, at naturalistic event boundaries there is worse behavioral prediction (*72*), delayed eye movements to relevant objects (*73*), and a global neural updating signal (*72*), all indicative of an increase in prediction errors demarcating naturalistic boundaries. Similarly, observed action events that are isolated from their naturalistic event sequence and randomly scrambled, lead to less top-down activity, and more bottom-up visual responses (*74*). While these previous studies demonstrate behavioural and neural signatures of reduced prediction and increased prediction errors across (artificial or natural) input boundaries, they do not capture the predictive representational dynamics themselves. We extend these findings by unveiling the exact content and latencies of those reduced predictions and increased prediction errors, i.e., eliminated motion prediction and increased and delayed post-stimulus representation of static body posture, respectively. Future research is needed to test how predictive representational dynamics are modulated by more naturalistic event boundaries.

One caveat is worth mentioning. While we observed lagged representations for body posture in the piecewise scrambled condition that we interpret as unpredicted sensory input, a positive latency precludes inferring that a representation solely reflects bottom-up activation, as top-down expectations might very well have modulated its veridicality (*43*) and processing latency (*75*), i.e., advanced it relative to pure bottom-up processing. It is therefore unclear to what extend these representations are still affected by fast recurrent processing or even prediction for the duration of a video piece. However, a direct comparison with the normal viewing condition does allow us to conclude that post-stimulus representations of body posture are enhanced and delayed, which is indeed in line with predictive processing theories.

To conclude, using a new dynamic RSA approach we demonstrate that high-level view-invariant prediction of biological motion relies on long-term high-level holistic and gravity priors, while prediction of view-dependent biological motion and low-level visual motion relies on short-term kinematics priors based on temporally local context. These findings thus reveal how hierarchical priors shape predictive neural representations of naturalistic continuous input.

Materials and Methods

Participants

Forty-three healthy humans (mean age, 28 ± 6 years) participated in the experiment for monetary compensation. Three participants were excluded from analyses due to low behavioral performance (see below). Participants were relatively diverse and consisted of MSc and PhD students, and postdocs, from all faculties of the University of Trento, and originating from a wide range of countries. All but the first participant was right-handed, all had (corrected-to-)normal vision and were naïve with respect to the purpose of the study. Sixteen participants identified themselves as male. Sex or gender information was not otherwise collected. As we aimed for our findings to be generalizable, we combined all participants in the group-level statistical analyses. All experimental procedures were performed in accordance with the Declaration of Helsinki and were positively reviewed by the Ethical Committee of the University of Trento. Written informed consent was obtained.

Stimuli

We used the same stimuli as in our previous study as they worked well for capturing predictive neural representations across hierarchical levels of processing (*46*). For completeness they are shortly described here. Stimuli consisted of 14 unique 5-sec videos of four smoothly connected ballet dancing figures selected from five unique figures (i.e., passé, pirouette, arabesque, bow and jump; see Table S1 for the figures of each sequence). We selected naturalistic smooth biological motion because this is a rich and evolutionary salient stimulus that can be modeled across levels of processing from low-level visual to higher-level view-invariant actions. All ballet figures were presented from several viewpoints to allow separation of view-dependent and view-invariant representations. Ballet videos were recorded with a Canon PowerShot G9 X Mark II at 50 Hz. Video frames were cropped to 400 × 376 [height × width] pixels. During filming we recorded 3D positions of 13 passive (infrared reflecting) markers placed on the ankles, knees, hips, shoulders, elbows, wrists, and head of the dancer. Marker positions were recorded at 100 Hz using a Qualisys motion capture system (infrared-based marker capture). See https://github.com/Ingmar-de-Vries/DynamicPredictions for all 14 unique videos and temporally aligned 3D kinematic data. Code used for the current analyses will be shared in a new directory.

Stimulus models

All data analyses were performed in MATLAB (version 2023b; MathWorks). We created the exact same stimulus-feature models as in our previous study (*46*), and which are explained below for completeness. We created a total of 9 stimulus-feature models (Fig. S4), to comprehensively characterize the dance sequences at several hierarchical levels of complexity/abstraction from low-level visual information (pixelwise luminance) up to higher-level, perceptually more invariant information (view-invariant body posture and motion). We also included a participant-specific gaze position model to control for eye-movement related variance in the dynamic representational similarity analysis (dRSA). Models based on video data were upsampled from 50 to 100 Hz to match the kinematic data, using shape-preserving piecewise cubic interpolation.

Low-level visual (pixelwise) models (Fig. S4b; left column): RGB values at all 400 × 376 pixels of each video frame were converted to grayscale (i.e., luminance values), smoothed (i.e., spatially low-pass filtered) using a Gaussian kernel with sigma = 5 (i.e., ~11.75 pixels full width at half max(*47*)), and vectorized, thus resulting in a model vector for each time point. For each time point separately, the dissimilarity between the two vectors of two different videos was computed as 1-Pearson correlation, to generate a single entry in the model RDM (representational dissimilarity matrix) at that time point. This was repeated for each stimulus pair to generate a 14 × 14 RDM at a single time point and subsequently repeated over the 500 time points (i.e., frames) to generate 500 model RDMs (see Fig. S3d for an example). Pixelwise motion was estimated with optical flow vectors (Fig. S4b, middle and right column) using MATLAB's estimateFlow.m implementation of the Farnebäck algorithm (*76*), which captures the (apparent) motion between subsequent frames at each pixel. Optical flow vectors were spatially smoothed using the same Gaussian kernel as above (*77*), divided in magnitude and 2D direction, and vectorized, before computing the RDM as above.

Kinematic marker models (Fig. 4c): View-dependent body posture (Fig. 4Sc; first column) was operationalized as the 3D positions of the 13 kinematic markers (i.e., 39 features), relative to a coordinate system with origin placed on the floor in the center of the video frame. View-dependent body motion (Fig. 4Sc; third column) was operationalized as the first derivative of view-dependent body posture, i.e., the difference in the 3D marker positions between two subsequent frames. Body motion vectors, therefore, had a magnitude and a 3D direction. View-invariant body posture and motion (Fig. S4c; second and fourth column) were computed by aligning (i.e., minimize sum of squares) the 3D marker structure of one stimulus with the marker structure of the other stimulus, while keeping its internal structure intact, by using translation, and rotation along the vertical axis only, before computing dissimilarity as above. If in two videos the exact same ballet figure is performed from a different perspective, this would result in high dissimilarity of view-dependent body posture, but high similarity of view-invariant body posture. Since dissimilarity is slightly affected by directionality (i.e., dissimilarity(frame1,frame2aligned) ≠ dissimilarity(frame1aligned,frame2)), dissimilarities were computed in both directions and subsequently averaged. View-dependent and view-invariant body acceleration were operationalized as second derivative of body posture. Since these models only explained a minimal amount of variance in the neural data, they are not illustrated in the main results. However, they did correlate with some of the tested models (see subsection Simulations), and we therefore regressed them out when testing each of the other models (see subsection Dynamic representational similarity analysis for details on regression). Note that to prevent artificially large motion vectors (i.e., for pixelwise and

body motion) across piece-boundaries in the piecewise scrambled condition, we first computed the models, after which model data was piecewise scrambled to match the stimuli observed by the participants.

Since eye movements are a common covariate in multivariate analysis of neuroimaging data(*78*), even in case of a fixation task (i.e., micro-saccades remain), we included participant-specific RDMs based on gaze-position to ensure that the dRSA results for our models of interest were not due to eye movements. Eye tracker data was downsampled to 100 Hz, missing samples were interpolated (e.g., due to blinks; 1.4 ± 3.0% of all samples; mean ± std across participants), and trials with more than 10% missing samples were rejected (2.7 ± 6.5% of all trials; mean ± std across participants). Finally, trials were averaged per condition over all repetitions of the exact same stimulus. Dissimilarity for the RDM was defined as the Euclidean distance between the four eye position values (i.e., x-, and y-position for the left and right eye), as four values are too low for a reliable correlation estimate. Gaze position did not explain much variance in the neural data (Fig. S1), and simulations confirmed that any remaining variance best explained by gaze position was successfully regressed out when testing each of the other models (see Fig. S2 and subsection Simulations).

Experimental design and behavioral task

We used Psychophysics Toolbox (version 3) in MATLAB (version 2012b; MathWorks) to create and run the experiment. Trials consisted of 5-sec ballet videos (Fig. 1a) that were separated by blank screens with a white fixation cross presented for 1.8–2.2 s (uniform distribution). Participants kept fixation on a fixation cross displayed on top of the videos while covertly attending the dancer. Videos were presented in three different viewing conditions (33% of trials per condition); normally as in our previous study (*46*), up-down inverted, or temporally piecewise scrambled with pieces randomly and uniformly distributed between 200–500 msec (Fig. 1b). Piece lengths and scrambled indices were different for each participant and each of the 14 videos. They were determined at start of each recording session and stored for offline analysis to allow reconstruction of the exact observed stimuli per participant. For the first three participants we failed to store this information, which thus led to the exclusion of these 3 participants from any analysis involving the scrambled condition. A run lasted ~14 min and consisted of 108 randomly ordered ballet figures (36 per condition), with the only constraint that the exact same sequence was never directly repeated (i.e., there was always at least one different sequence in between). Attention to the ballet sequence was ensured by means of 16 randomly distributed catch trials per run, in which a black screen occluded the video unexpectedly for 400 msec (Fig. 1c; top). Participants were requested to indicate by button press whether the video continued correctly, or with an incorrect sequence. The occlusion could appear between 2.7 and 4.2 sec after video onset and never crossed the boundary between two ballet figures (which would make the task close to impossible). Additionally, to make the occlusion task possible in the piecewise scrambled condition, the scrambled video was always replaced by the normal unscrambled video from 500 msec before occlusion onset. Fixation was encouraged by means of 8 randomly distributed catch trials per run, in which the fixation cross changed color from white to light purple (RGB=[117 112 179]) for 200 msec at a random time between 0.4 and 4.8 seconds after onset, in response to which participants were requested to press a button. The color was chosen such that the change was only visible when fixating, but not from the periphery. Participants received immediate feedback on each catch trial. To ensure continuous attention to the task and fixation, catch trials were uniformely distributed across each run. Participants performed one practice run outside of the MEG scanner, in which 41% of trials were catch trials to ensure enough practice with the task. During MEG, participants performed the behavioral task well, indicating that they attended the dancer and kept fixation (Fig. 1d). Participants performed either 6 or 7 runs depending on attentional fatigue (6.1 on average), which resulted in an average of 659 trials per participant for further analyses (i.e., ~15.7 trials per unique video per condition). Three participants were excluded from all analyses due to behavioral performance because they responded too fast (< 200 msec) or too slow on the occlusion task in more than 75% of trials (3 participants), or because they had a condition-averaged accuracy on the occlusion task that was lower than the participant-average minus 2 standard deviations (1 participant). This led to a total of 40 remaining participants for the normal and inverted conditions, and 37 for the piecewise scrambled condition.

MEG data collection and preprocessing

MEG was recorded at 1000 Hz using a 306-channel (204 planar gradiometers, 102 magnetometers) VectorView MEG system (Neuromag, Elekta) in a two-layer magnetically shielded room (AK3B, Vacuum Schmelze). A low-pass antialiasing filter at 330 Hz and a DC offset correction was applied online, but no

online or offline high-pass filter was applied, since high-pass filtering with too high a cutoff can have unwanted side-effects including temporal spread of multivariate information (*79*). Before MEG recording, we used the Polhemus FASTRAK electromagnetic localization system to digitize the individual-participant head shape, i.e., the position of three anatomic landmarks (nasion, left and right preauricular points), five head position indicator coils and at least 350 additional points uniformly spread across the participant's head. We digizited landmarks and coils twice to minimize error. To allow offline MEG – MRI co-registration, we acquired the head position at start of each run by passing a small current through the coils. We collected binocular gaze position at 1000 Hz using the SR-Research Eyelink Plus eye tracker. Stimuli were presented using a Vpixx PROPixx projector. All hardware was connected to a DataPixx input-output hub (Vpixx Technologies) to present stimuli, and collect data and button presses with minimal delays, and to store stimulus-onset triggers together with the MEG data. Additionally, a photodiode in the top left corner of the screen (i.e., outside of the video frame and invisible to the participant), registered a square on the screen underneath the photodiode changing from black to white at the onset of each single video frame, for realigning MEG signals to these photodiode signals offline.

As a first preprocessing step, bad sensors (e.g., noisy or with many SQUID jumps) were automatically detected using MNE python's find_bad_channels_maxwell algorithm (*80*), to detect a maximum of 12 noisy sensors for interpolation. Neuromag's MaxFilter implementation (version 2.2) of Signal Source Separation (SSS (*81*)) was used to remove external noise, interpolate bad sensors, and spatially align the head position inside the helmet across runs. All further preprocessing was done using Brainstorm (version 3)(*82*) and Fieldtrip (version 20191113)(*83*) in MATLAB (version 2023b; MathWorks), as well as custom-written MATLAB scripts (shared at https://github.com/Ingmar-de-Vries/DynamicPredictions [will to new directory for current study]). MEG data were loaded into Brainstorm, after which spatial co-registration with participant-specific anatomical MRI scans was refined using the 350+ digitized head points, data were filtered for line noise (50 and 100 Hz, 0.5 Hz filter bandwidth), down-sampled to 500 Hz, and cleaned using independent component analysis (ICA). ICA was applied separately to magneto- and gradiometers, and on a temporary data version that was downsampled to 250 Hz and band-pass filtered at 1–100 Hz to speed up computation and improve ICA, respectively. An average of 2.2 blink and eye-movement, 1.5 cardiac and 0.3 noise related components were detected by visual inspection and removed from the original data. Next, segments were automatically marked as bad using Brainstorm based on low-frequency noise (i.e., 1–7 Hz; e.g., movement or SQUID jumps) or high-frequency noise (i.e., 40–240 Hz; e.g., muscle), with the sensitivity parameter set to 4 or 5 depending on visual inspection of detected segments (participant-average of 4.8 for each frequency band). Next, continuous data were epoched (-1.5 to 6.5 sec locked to video onset), single-trial baseline corrected (-500 to 0 msec), and previously detected bad segment were converted to NaN. This resulted in an average of 614 out of 659 completely clean trials. We then applied source reconstruction on these clean trials only (see below) and only stored the inversion kernels (i.e., 306 x 15.000 matrix) for efficiency. Last, sensor-level data were exported in Fieldtrip format and Fieldtrip functions were used to realign epochs to photodiode onset, temporally smooth data using a 18 msec boxcar kernel, downsample to 50 Hz, and finally average over all ~15.7 repetitions of the same unique video per condition. Note that for the average we ignored the bad segments (i.e., NaNs), as well as the time window after catch-trial onset in the catch-trials.

Source reconstruction and ROI selection

We used minimum-norm estimation (MNE) in Brainstorm with default settings (*84*), to estimate 15,000 source signals distributed on the cortical surface. For 30 participants we used participant-specific anatomical T1 MRI scans, and if unavailable the ICBM152 standard brain for the remaining 10 participants, to construct 3D forward models. Scans were obtained using a Siemens Prisma 3T. Fiducial points were marked automatically in MNI space, after which 3D brains were reconstructed using CAT12 version 2170 (*85*). For the 10 participants without individual scan, the standard brain was warped to the participant's head shape as estimated from the +350 digitized head points. Overlapping spheres were used as forward model. Noise covariance matrices were computed from the baseline-normalized single trials using the -1000 to 0 msec pre-video time window and using both gradio- and magnetometers. Non-normalized current density maps were computed as measure of source activity, with source direction constrained to be normal to the cortex, thus resulting in a single signal at each source. Next, the 15,000 sources were parcellated into 360 parcels according to the Human Connectome Project (HCP) atlas (*86*). Six regions of interest (ROIs) that consisted of one or more parcels were selected a priori consistent with our previous study (Fig. S5) (*46*), i.e., three low-level

visual areas (i.e., V1, V2 and V3+4), and three areas from the action observation network, namely the lateral occipitotemporal cortex (LOTC), the anterior inferior parietal lobe (aIPL), and the ventral premotor cortex (PMv), because this network is involved in processing action stimuli (*87–89*). ROIs were comparable in size with on average 484 sources (min–max = 403–565). Source signals were used as features to compute a neural RDM for each ROI, using 1–Pearson correlation as dissimilarity after centering the data.

Dynamic representational similarity analysis

We applied the same dRSA pipeline with similar parameters as validated in our previous study (*46*), but describe it here shortly for completeness. Time-resolved neural and model RDMs were computed as described above (see also Fig. S3d), resulting in RDMs of 91 features (i.e., 14 × 14 stimuli) by 250 time points (at 50 Hz). RDMs were temporally smoothed using a 30 msec boxcar kernel to improve SNR, and subsequently centered per time point, and standardized across all time points at once to equalize scales between RDMs while keeping temporal structure inherent to a certain model intact. Next, the similarity between neural and model RDMs was computed at each neural-by-model time point (Fig. S3e; bottom panel), and averaged over time for each neural-to-model time lag (i.e., for each diagonal in the 2D dRSA matrix) within the range of -1 to 1 sec. This resulted in a latency-plot (Fig. S3e; top panel), in which a peak to the right of the vertical zero-lag midline indicates that the neural RDM is most similar to a model RDM earlier in time, whereas a peak on the left indicates that the neural RDM is predicting a future model RDM. The different stimulus models inevitably share variance at any given time point, but also across time points due to temporal autocorrelation within models (see subsection Simulations and Fig. S2a and c), which are not problems with dRSA per se, but rather inherent to naturalistic continuous stimuli. We therefore used principal component regression (PCR) to compute similarity, which estimates variance in the neural RDM best explained by each separate model RDM (Fig. S2b and d), while minimizing variance better explained by other models. The neural RDM at a given time point t_y is the response variable Y (91 features by 1 time point), while the regressors consist of the model RDM of interest at t_x (91 x 1), the other 9 model RDMs across latencies from -1 to 1 sec (91 x N_{time} per model), plus the model RDM of interest itself at distant time points (i.e., minimum distance from t_x at which the model shared less than 10% variance with itself at t_x) to minimize effects of model autocorrelation. The regression weight for the model RDM of interest at t_x is the measure of similarity with the neural RDM at t_y. Since this results in >900 regressors (i.e., 91 x 900), we used principal component analysis (PCA) to reduce dimensionality. First, we ran PCA separately per model and selected those components capturing at least 0.1% of total variance. We then combined models and ran a second PCA from which only the first 75 components were selected (to stay well below the theoretical limit of 91 features). These 75 components were used as predictor variables in a linear least-squares regression, after which the PCA loadings were used to project the component regression weights back onto the original regressors, in order to extract the regression weight for the model RDM of interest at t_x. PCR strongly reduces dimensionality of the regressors while maintaining almost all variance. Additionally, it decorrelates regressors which prevents multicollinearity. Crucially, simulations confirmed that PCR was effective at estimating a specific model representation while minimizing variance better explained by other models (see subsection Simulations and Fig. S2).

Temporal subsampling

To increase SNR and generalizability of results we included an additional important step in the dRSA pipeline. At any given time point stimulus-specific feature trajectories determine how far in advance a stimulus can be predicted, which differs for each of the 14 videos. The RDM structure at any time point therefore depends on the exact arbitrary pairwise alignment of the ballet sequences, which in turn causes the pattern across stimulus time (i.e., along the diagonal) in the 2D dRSA matrix to be heterogeneous and idiosyncratic. Since we're not interested in these specific 14 stimuli and their arbitrary alignment per se, we added a temporal subsampling and realignment step to mimic an infinite number of stimuli and thus minimize the idiosyncrasy of our dRSA results (Fig. S6). Specifically, across 1000 iterations a random 3-sec segment was extracted from each video (Fig. S6a; orange boxes). While a different time window was selected per stimulus, the same time window was used for neural and model data within a stimulus, thus keeping their temporal alignment intact (Fig. S6a; dotted vertical orange lines). Next, the new 14 3-sec segments were realigned, and RDMs were computed at each realigned time point tR (Fig. S6b and c), after which similarity between neural and model RDMs was computed for each realigned neural by model time point as described above. This procedure results in a different pairwise alignment of the 14 stimuli on each iteration, thus creating different idiosyncratic dRSA latency curves on each iteration with different shapes and peak latencies (see Fig. 7d in De Vries et al. 2023

(*46*) for an illustration). In the last step dRSA latency curves were averaged across the 1000 iterations, thus reducing dependency of results on arbitrary pairwise alignments of action sequences, make the results more generalizable, and increase SNR. Note that this whole procedure was done separately per participant.

Statistical analysis

First, we tested for which ROI–model–condition combination the dRSA latency curve (i.e., beta weights) were significantly higher than zero. We performed group-level non-parametric permutation testing (25,000 permutations per analysis) with cluster-based correction for multiple comparisons (*90*), with t-values from a one-tailed t-test as test statistic and the sum over t-values within a cluster as measure of cluster size. Clusters were made up of connected time points with p-values below 0.001, and significance of cluster-size was tested at alpha = 0.05. Only significant clusters larger than what might be expected by chance survive this procedure. Significant intervals in the dRSA latency plots are indicated by think horizontal bars with colors matching the respective ROIs (Fig. 2, 6 and S1).

The two main measures of interest we used to compare conditions are the magnitude and latency of the peaks, which capture the strength (veridicality) of the representation, and the timing relative to the stimulus event (i.e., how lagged or predictive) at which the representation is strongest, respectively. Given that peak latencies are difficult if not impossible to capture for weak representations (i.e., a nonsignificant dRSA curve), we restricted the condition comparisons to ROIs where at least one condition showed a significant dRSA curve (Fig. S1). To compare magnitudes of predictive peaks between the normal and inverted conditions, we first averaged dRSA values in a 100 msec window surrounding a priori selected predictive peaks based on our previous results with the same stimuli (*46*) (i.e., ~110 msec, ~180 msec, and ~510 msec, for pixelwise motion, view-dependent, and view-invariant body motion, respectively), and performed a repeated measures ANOVA for each model separately, since time-averages are not comparable between models with different temporal autocorrelation profiles, i.e., averages are biased upwards in case of broader temporal autocorrelation (Fig. S2). This resulted in a [2 condition by 4 ROIs] ANOVA for pixelwise motion and view dependent body motion, and a [2 conditions by 3 ROIs] ANOVA for view invariant body motion. To compare peak magnitudes between the normal and scrambled condition for the body posture models we did not have very reliable a priori peak latencies since these representations were weak in our previous study. We therefore estimated peak latencies based on the grand-average dRSA latency plots (averaged over participants, normal and scrambled conditions, and the first four ROIs), and extracted peak magnitudes at that latency for each participant, condition, ROI and model for statistical comparison using a [2 conditions by 4 ROIs by 3 models] repeated measures ANOVA for predictive peaks and [2 conditions by 4 ROIs by 2 models] ANOVA for lagged body posture peaks. ANOVAs were run in JASP version 0.19.3 with default parameter settings (e.g., Cauchy prior width set to 0.707) (*91*). Frequentist ANOVAs were used for hypothesis testing, while the strength of evidence in favor of or against a given effect was further estimated as Bayes Factor (BF) using Bayesian ANOVAs. Violations of the sphericity assumption were corrected using Huynh-Feldt. For Bayesian ANOVAs we analyzed the effects for matched models, i.e., a direct comparison of the evidence between all models that include a given factor and all equivalent models without that factor.

We used bootstrapping to compare peak latencies between conditions (or against zero for the direct comparison between neural RDMs of normal vs inverted; Fig. 4), a common approach that is sensitive to capture small latency differences (e.g., (*48*)). Specifically, across 10,000 iterations we extracted a random sample (with replacement) of size N from our N participants, estimated the peak latency difference between conditions for each bootstrapped participant, and averaged over participants. These 10,000 bootstrapped averaged peak latency differences were then used to create a distribution, and to test whether the 95% confidence interval of the bootstrapped distribution included zero. Additionally, we performed the exact same bootstrapping analysis on the centroid latency, which is the center of mass of the dRSA curve, or the latency at which the area under the curve to the right and left side of this latency is equal. The centroid latency might be less sensitive to noise than peak latency estimates, as it considers representational weight that is not captured by the peak (e.g., for non-symmetrical curves). Last, for an exact replication of our previous results regarding hierarchical time windows of predictions in the normal condition only (*46*), we computed retrieved jackknifed peak latencies and ran a [3 conditions by 4 ROIs] repeated measures ANOVA.

Simulations

While we used simulations in our previous study to validate the regression-based dRSA pipeline (*46*), we computed them here again for completeness, now including the new conditions. Most importantly, the simulations demonstrate that principal component regression (PCR; see subsection Dynamic representational similarity analysis) is effective at estimating variance in the (simulated) neural RDM best explained by each separate model RDM, while minimizing variance better explained by other models (compare Fig. S2a and c with b and d, respectively). Specifically, each simulated neural RDM consisted of a single model RDM at zero lag with zero added noise. We ran dRSA with PCR beta weights as similarity measure (Fig. S2b and d), as well as correlation for comparison (Fig. S2a and c). Note the nonzero off-midline correlations both across and within models, which demonstrate the complex patterns of shared variance across time, which in turn limits unambiguous inference if correlation is used. Instead, PCR is effective at extracting only the model of interest from the simulations, while largely regressing out other models. Note that the models for the normal and inverted condition are identical, while the piecewise scrambled condition inevitably results in models with narrower temporal autocorrelation. Interestingly, these simulations indicate a different maximum possible PCR beta weight for each model and for the normal/inverted vs scrambled condition (i.e., see zero-lag peak magnitudes in Fig. S2b and d). We used these maximum possible beta weights to normalize the observed beta weights against (i.e., % max beta in Fig. 2, 6 and S1), which allows better comparison between models or conditions with different maximum possible regression weights.

References

1. A. Bubic, D. Yves von Cramon, R. I. Schubotz, Prediction, cognition and the brain. *Frontiers in Human Neuroscience* **4**, 1–15 (2010).

2. F. P. de Lange, M. Heilbron, P. Kok, How Do Expectations Shape Perception? *Trends in Cognitive Sciences* **22**, 764–779 (2018).

3. R. P. N. Rao, D. H. Ballard, Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience* **2**, 79–87 (1999).

4. K. Friston, The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience* **11**, 127–138 (2010).

5. A. M. Bastos, W. M. Usrey, R. A. Adams, G. R. Mangun, P. Fries, K. J. Friston, Canonical Microcircuits for Predictive Coding. *Neuron* **76**, 695–711 (2012).

6. H. Hogendoorn, Perception in real-time: predicting the present, reconstructing the past. *Trends in Cognitive Sciences* **26**, 128–141 (2022).

7. T. S. Lee, D. Mumford, Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America A* **20**, 1434 (2003).

8. L. Aitchison, M. Lengyel, With or without you: predictive coding and Bayesian inference in the brain. *Current Opinion in Neurobiology* **46**, 219–227 (2017).

9. A. Clark, *The Experience Machine: How Our Minds Predict and Shape Reality* (Random House, 2024).

10. G. B. Keller, T. D. Mrsic-Flogel, Predictive Processing: A Canonical Cortical Computation. *Neuron* **100**, 424–435 (2018).

11. A. Dimakou, G. Pezzulo, A. Zangrossi, M. Corbetta, The predictive nature of spontaneous brain activity across scales and species. *Neuron* **113**, 1310–1332 (2025).

12. N. Barascud, M. T. Pearce, T. D. Griffiths, K. J. Friston, M. Chait, Brain responses in humans reveal ideal observer-like sensitivity to complex acoustic patterns. *Proceedings of the National Academy of Sciences* **113**, E616–E625 (2016).

13. P. Series, A. Seitz, Learning what to expect (in visual perception). *Front. Hum. Neurosci.* **7** (2013).

14. P. Berkes, G. Orbán, M. Lengyel, J. Fiser, Spontaneous Cortical Activity Reveals Hallmarks of an Optimal Internal Model of the Environment. *Science* **331**, 83–87 (2011).

15. A. Fiser, D. Mahringer, H. K. Oyibo, A. V. Petersen, M. Leinweber, G. B. Keller, Experience-dependent spatial expectations in mouse visual cortex. *Nat Neurosci* **19**, 1658–1664 (2016).

16. M. Berniker, M. Voss, K. Kording, Learning Priors for Bayesian Computations in the Nervous System. *PLOS ONE* **5**, e12686 (2010).

17. B. Török, D. G. Nagy, M. Kiss, K. Janacsek, D. Németh, G. Orbán, Tracking the contribution of inductive bias to individualised internal models. *PLOS Computational Biology* **18**, e1010182 (2022).

18. B. A. Urgen, M. Kutas, A. P. Saygin, Uncanny valley as a window into predictive processing in the social brain. *Neuropsychologia* **114**, 181–185 (2018).

19. A. P. Saygin, T. Chaminade, H. Ishiguro, J. Driver, C. Frith, The thing that should not be: predictive coding and the uncanny valley in perceiving human and humanoid robot actions. *Soc Cogn Affect Neurosci* **7**, 413–422 (2012).

20. M. B. Tunca, B. A. Urgen, Predictive processing in the cortical network of biological motion perception: An fMRI-DCM study. bioRxiv [Preprint] (2025). https://doi.org/10.1101/2025.03.05.641401.

21. H. O. Elmas, S. Er, A. D. Rezaki, A. Izgi, B. M. Urgen, H. Boyaci, B. A. Urgen, Predictive processing in biological motion perception: Evidence from human behavior. *Perception*, 03010066251355391 (2025).

22. G. Özen, B. A. Urgen, Predictive Processing in Visual Perception of Actions: The Interplay of Long-Term and Short-Term Priors. Social Science Research Network 5289270 [Preprint] (2025). https://doi.org/10.2139/ssrn.5289270.

23. R. Blake, M. Shiffrar, Perception of Human Motion. *Annual Review of Psychology* **58**, 47–73 (2007).

24. J. Schultz, C. D. Frith, Animacy and the prediction of behaviour. *Neuroscience & Biobehavioral Reviews* **140**, 104766 (2022).

25. F. Lacquaniti, G. Bosco, S. Gravano, I. Indovina, B. La Scaleia, V. Maffei, M. Zago, Multisensory Integration and Internal Models for Sensing Gravity Effects in Primates. *BioMed Research International* **2014**, 615854 (2014).

26. B. M. Hood, Gravity does rule for falling events. *Developmental Science* **1**, 59–63 (1998).

27. R. Fox, C. McDaniel, The perception of biological motion by human infants. *Science* **218**, 486–487 (1982).

28. M. Pavlova, I. Krägeloh-Mann, A. Sokolov, N. Birbaumer, Recognition of Point-Light Biological Motion Displays by Young Children. *Perception* **30**, 925–933 (2001).

29. G. Vallortigara, L. Regolin, F. Marconato, Visually Inexperienced Chicks Exhibit Spontaneous Preference for Biological Motion Patterns. *PLOS Biology* **3**, e208 (2005).

30. F. Simion, L. Regolin, H. Bulf, A predisposition for biological motion in the newborn baby. *Proceedings of the National Academy of Sciences* **105**, 809–813 (2008).

31. Q. Han, M. Gandolfo, M. V. Peelen, Prior knowledge biases the visual memory of body postures. *iScience* **27** (2024).

32. C. L. Reed, V. E. Stone, S. Bozova, J. Tanaka, The Body-Inversion Effect. *Psychol Sci* **14**, 302–308 (2003).

33. W. H. Dittrich, Action Categories and the Perception of Biological Motion. *Perception* **22**, 15–22 (1993).

34. G. Marrazzo, F. D. Martino, A. Mukovskiy, M. A. Giese, B. de Gelder, Neural encoding of biomechanically (im)possible human movements in occipitotemporal cortex. bioRxiv [Preprint] (2025). https://doi.org/10.1101/2025.01.07.631720.

35. J. McIntyre, M. Zago, A. Berthoz, F. Lacquaniti, Does the brain model Newton's laws? *Nat Neurosci* **4**, 693–694 (2001).

36. B. Jörges, J. López-Moliner, Gravity as a Strong Prior: Implications for Perception and Action. *Front. Hum. Neurosci.* **11** (2017).

37. D. H. F. Chang, N. F. Troje, Acceleration carries the local inversion effect in biological motion perception. *Journal of Vision* **9**, 19 (2009).

38. Y. Weiss, E. P. Simoncelli, E. H. Adelson, Motion illusions as optimal percepts. *Nat Neurosci* **5**, 598–604 (2002).

39. A. A. Stocker, E. P. Simoncelli, Noise characteristics and prior expectations in human visual speed perception. *Nat Neurosci* **9**, 578–585 (2006).

40. M. I. Garrido, J. M. Kilner, K. E. Stephan, K. J. Friston, The mismatch negativity: A review of underlying mechanisms. *Clinical Neurophysiology* **120**, 453–463 (2009).

41. D. Feuerriegel, R. Vogels, G. Kovács, Evaluating the evidence for expectation suppression in the visual system. *Neuroscience & Biobehavioral Reviews* **126**, 368–381 (2021).

42. F. Aitken, G. Turner, P. Kok, Prior Expectations of Motion Direction Modulate Early Sensory Processing. *The Journal of Neuroscience* **40**, 6389–6397 (2020).

43. P. Kok, J. F. M. Jehee, F. P. de Lange, Less Is More: Expectation Sharpens Representations in the Primary Visual Cortex. *Neuron* **75**, 265–270 (2012).

44. M. Ekman, S. Kusch, F. P. de Lange, Successor-like representation guides the prediction of future events in human visual cortex and hippocampus. *eLife* **12** (2023).

45. D. Richter, T. C. Kietzmann, F. P. de Lange, High-level visual prediction errors in early visual cortex. *PLOS Biology* **22**, e3002829 (2024).

46. I. E. J. de Vries, M. F. Wurm, Predictive neural representations of naturalistic dynamic input. *Nature Communications* **14**, 3858 (2023).

47. N. Kriegeskorte, M. Mur, P. Bandettini, Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience* **2**, 1–28 (2008).

48. R. M. Cichy, D. Pantazis, A. Oliva, Resolving human object recognition in space and time. *Nature Neuroscience* **17**, 455–462 (2014).

49. N. F. Troje, C. Westhoff, The Inversion Effect in Biological Motion Perception: Evidence for a "Life Detector"? *Current Biology* **16**, 821–824 (2006).

50. M. Pavlova, A. Sokolov, Orientation specificity in biological motion perception. *Perception & Psychophysics* **62**, 889–899 (2000).

51. G. Vallortigara, L. Regolin, Gravity bias in the interpretation of biological motion by inexperienced chicks. *Current Biology* **16**, R279–R280 (2006).

52. E. D. Grossman, R. Blake, Brain activity evoked by inverted and imagined biological motion. *Vision Research* **41**, 1475–1482 (2001).

53. A. Freire, K. Lee, L. A. Symons, The Face-Inversion Effect as a Deficit in the Encoding of Configural Information: Direct Evidence. *Perception* **29**, 159–170 (2000).

54. M. J. Farah, J. W. Tanaka, H. M. Drain, What causes the face inversion effect? *Journal of Experimental Psychology: Human Perception and Performance* **21**, 628–634 (1995).

55. T. Blom, D. Feuerriegel, P. Johnson, S. Bode, H. Hogendoorn, Predictions drive neural representations of visual events ahead of incoming sensory information. *Proceedings of the National Academy of Sciences of the United States of America* **117**, 7510–7515 (2020).

56. P. A. Johnson, T. Blom, S. van Gaal, D. Feuerriegel, S. Bode, H. Hogendoorn, Position representations of moving objects align with real-time position in the early visual response. *eLife* **12** (2023).

57. W. Turner, C. Sexton, P. A. Johnson, E. M. Wilson, H. Hogendoorn, Predictable motion is progressively extrapolated across temporally distinct processing stages in the human visual cortex. *PLOS Biology* **23**, e3003189 (2025).

58. W. Turner, C. Sexton, H. Hogendoorn, Neural mechanisms of visual motion extrapolation. *Neuroscience & Biobehavioral Reviews* **156**, 105484 (2024).

59. M. Birk, A. Tesio, I. E. J. de Vries, M. Wurm, Testing predicting coding using MEG-based dynamic RSA. *Journal of Vision* **25**, 2211 (2025).

60. C. Wacongne, E. Labyt, V. van Wassenhove, T. Bekinschtein, L. Naccache, S. Dehaene, Evidence for a hierarchy of predictions and prediction errors in human cortex. *Proceedings of the National Academy of Sciences* **108**, 20754–20759 (2011).

61. M. V. Peelen, E. Berlot, F. P. de Lange, Predictive processing of scenes and objects. *Nat Rev Psychol* **3**, 13–26 (2024).

62. M. Heilbron, K. Armeni, J.-M. Schoffelen, P. Hagoort, F. P. de Lange, A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences* **119**, e2201968119 (2022).

63. S. O. Murray, D. Kersten, B. A. Olshausen, P. Schrater, D. L. Woods, Shape perception reduces activity in human primary visual cortex. *Proceedings of the National Academy of Sciences* **99**, 15164–15169 (2002).

64. T. Nigam, C. M. Schwiedrzik, Predictions enable top-down pattern separation in the macaque face-processing hierarchy. *Nat Commun* **15**, 7196 (2024).

65. G. Yovel, N. Kanwisher, The Neural Basis of the Behavioral Face-Inversion Effect. *Current Biology* **15**, 2256–2262 (2005).

66. J. Taubert, D. Apthorp, D. Aagten-Murphy, D. Alais, The role of holistic processing in face perception: Evidence from the face inversion effect. *Vision Research* **51**, 1273–1278 (2011).

67. M. Hirai, D. H. F. Chang, D. R. Saunders, N. F. Troje, Body Configuration Modulates the Usage of Local Cues to Direction in Biological-Motion Perception. *Psychol Sci* **22**, 1543–1549 (2011).

68. U. Hasson, E. Yang, I. Vallines, D. J. Heeger, N. Rubin, A hierarchy of temporal receptive windows in human cortex. *Journal of Neuroscience* **28**, 2539–2550 (2008).

69. C. Baldassano, J. Chen, A. Zadbood, J. W. Pillow, U. Hasson, K. A. Norman, Discovering Event Structure in Continuous Narrative Perception and Memory. *Neuron* **95**, 709-721.e5 (2017).

70. H. Richardson, R. Saxe, Development of predictive responses in theory of mind brain regions. *Developmental Science* **23** (2020).

71. C. S. Lee, M. Aly, C. Baldassano, Anticipation of temporally structured events in the brain. *eLife* **10** (2021).

72. J. M. Zacks, C. A. Kurby, M. L. Eisenberg, N. Haroutunian, Prediction Error Associated with the Perceptual Segmentation of Naturalistic Events. *J Cogn Neurosci* **23**, 4057–4066 (2011).

73. M. L. Eisenberg, J. M. Zacks, S. Flores, Dynamic prediction during perception of everyday events. *Cogn. Research* **3**, 53 (2018).

74. C. Qin, F. Michon, Y. Onuki, Y. Ishishita, K. Otani, K. Kawai, P. Fries, V. Gazzola, C. Keysers, Predictability alters information flow during action observation in human electrocorticographic activity. *Cell Reports* **42** (2023).

75. Y. Pinto, S. van Gaal, F. P. de Lange, V. A. F. Lamme, A. K. Seth, Expectations accelerate entry of visual stimuli into awareness. *Journal of Vision* **15**, 13 (2015).

76. G. Farnebäck, "Two-Frame Motion Estimation Based on Polynomial Expansion BT - Image Analysis" J. Bigun, T. Gustavsson, Eds. (Springer Berlin Heidelberg, Berlin, Heidelberg, 2003), pp. 363–370.

77. Efros, Berg, Mori, Malik, "Recognizing action at a distance" in *Proceedings Ninth IEEE International Conference on Computer Vision* (IEEE, 2003), pp. 726–733 vol.2.

78. S. C. Quax, N. Dijkstra, M. J. van Staveren, S. E. Bosch, M. A. J. van Gerven, Eye movements explain decodability during perception and cued attention in MEG. *NeuroImage* **195**, 444–453 (2019).

79. J. van Driel, C. N. L. Olivers, J. J. Fahrenfort, High-pass filtering artifacts in multivariate classification of neural time series data. *Journal of Neuroscience Methods* **352**, 109080 (2021).

80. A. Gramfort, M. Luessi, E. Larson, D. A. Engemann, D. Strohmeier, C. Brodbeck, R. Goj, M. Jas, T. Brooks, L. Parkkonen, M. Hämäläinen, MEG and EEG data analysis with MNE-Python. *Front. Neurosci.* **7** (2013).

81. S. Taulu, J. Simola, Spatiotemporal Signal Space Separation Method for Rejecting Nearby Interference in MEG Measurements. *Physics in Medicine & Biology* **51**, 1759 (2006).

82. F. Tadel, S. Baillet, J. C. Mosher, D. Pantazis, R. M. Leahy, Brainstorm: A User-Friendly Application for MEG/EEG Analysis. *Computational Intelligence and Neuroscience* **2011**, 1–13 (2011).

83. R. Oostenveld, P. Fries, E. Maris, J. M. Schoffelen, FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience* **2011**, 156869 (2011).

84. M. S. Hämäläinen, R. J. Ilmoniemi, Interpreting magnetic fields of the brain: minimum norm estimates. *Medical & Biological Engineering & Computing* **32**, 35–42 (1994).

85. C. Gaser, R. Dahnke, P. M. Thompson, F. Kurth, E. Luders, the Alzheimer's Disease Neuroimaging Initiative, CAT: a computational anatomy toolbox for the analysis of structural MRI data. *Gigascience* **13**, giae049 (2024).

86. 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, D. C. Van Essen, A multi-modal parcellation of human cerebral cortex. *Nature* **536**, 171–178 (2016).

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

88. J. M. Kilner, K. J. Friston, C. D. Frith, Predictive coding: An account of the mirror neuron system. *Cognitive Processing* **8**, 159–166 (2007).

89. E. S. Cross, W. Stadler, J. Parkinson, S. Schütz-Bosbach, W. Prinz, The influence of visual training on predicting complex action sequences. *Human Brain Mapping* **34**, 467–486 (2013).

90. E. Maris, R. Oostenveld, Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods* **164**, 177–190 (2007).

91. E.-J. Wagenmakers, J. Love, M. Marsman, T. Jamil, A. Ly, J. Verhagen, R. Selker, Q. F. Gronau, D. Dropmann, B. Boutin, F. Meerhoff, P. Knight, A. Raj, E.-J. van Kesteren, J. van Doorn, M. Šmíra, S. Epskamp, A. Etz, D. Matzke, T. de Jong, D. van den Bergh, A. Sarafoglou, H. Steingroever, K. Derks, J. N. Rouder, R. D. Morey, Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review*, doi: 10.3758/s13423-017-1323-7 (2017).

Acknowledgments

We thank Natasha Bertelsen who volunteered to help with data collection, allowing us to finish just before planned summer closure of the MEG scanner. We thank Carlo Miniussi, former director of CIMeC, where the experiment was conducted, for agreeing to keep the MEG scanner operational 2 weeks longer than planned before summer closure. We thank Danielle E. Parrott for kindly serving as the ballet dancer in our video stimuli.

Funding

European Commission Marie Skłodowska-Curie Actions postdoctoral fellowship 101060807 (IEJV).

Author contributions

Conceptualization: IEJV, FPL, MFW
Methodology: IEJV, FPL, MFW
Investigation: IEJV
Visualization: IEJV
Supervision: FPL, MFW
Writing—original draft: IEJV
Writing—review & editing: IEJV, FPL, MFW

Competing interests

The authors declare that they have no competing interests.

Data and materials availability

The experimental code and stimuli, all MEG, eye-tracking, and behavioral data, and all analysis code needed to run the presented analyses and evaluate the conclusions are freely available. Code and stimuli are available on GitHub: XXX. The data will be shared in BIDS format on the Donders Repository.

Supplementary material

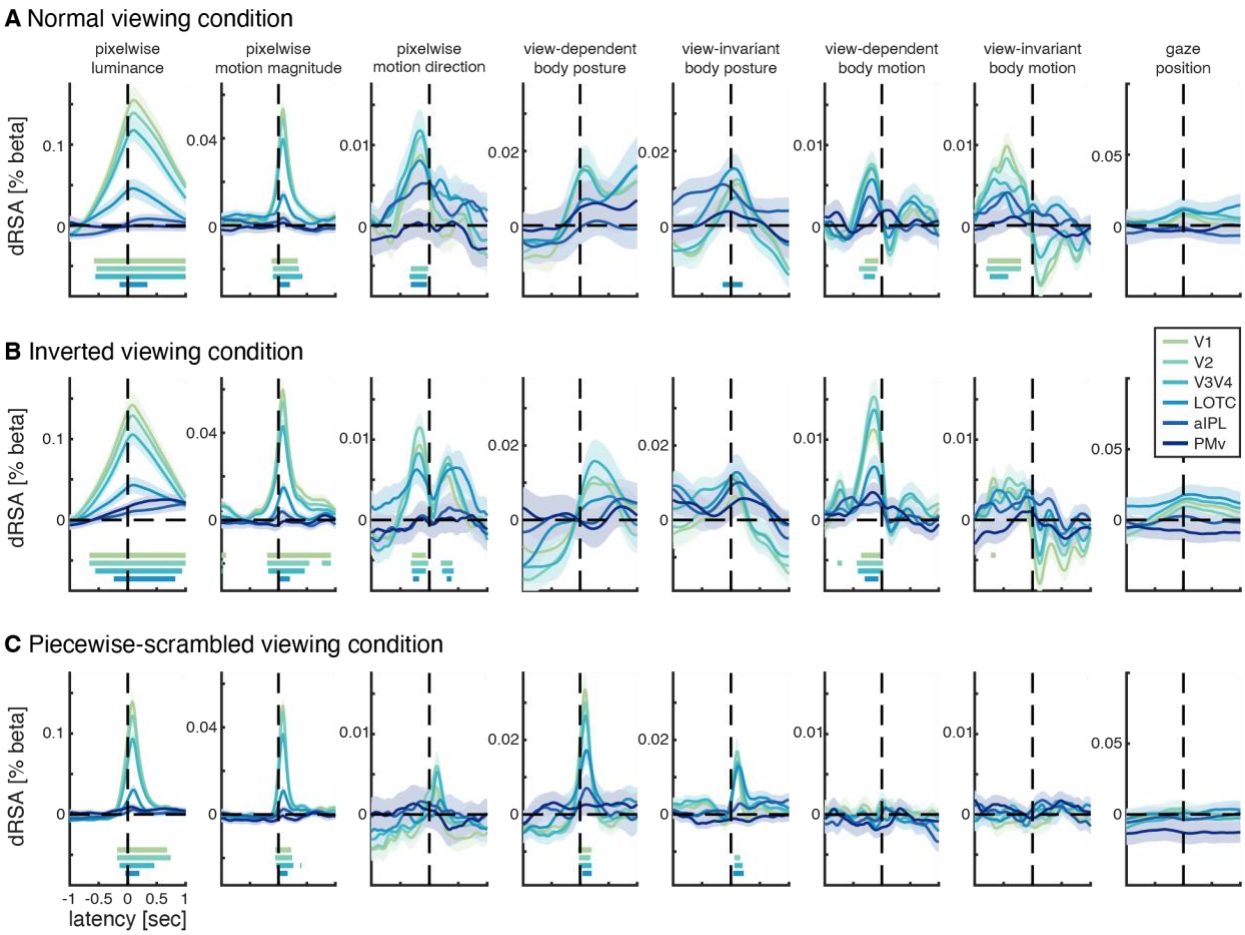

Fig. S1. dRSA results for all models. Region of interest (ROI)-based analysis, with normalized dRSA regression weights illustrated as latency plots for all tested models. The three viewing conditions are displayed in rows with normal on top (**A**), inverted in the middle (**B**), and temporal piecewise scrambled at the bottom (**C**). Different stimulus models are displayed in columns. Lines and shaded areas indicate participant-average and SEM, respectively, with n = 40 in (**A**) and (**B**), and n = 37 in (**C**). Light-to-dark colors indicate posterior-to-anterior ROIs. Horizontal bars indicate beta weights significantly larger than zero (one-sided t-test with $p < 0.001$ for each time sample), corrected for multiple comparisons across time using cluster-based permutation testing ($p < 0.05$ for cluster-size), with colors matching the respective ROI line plot.

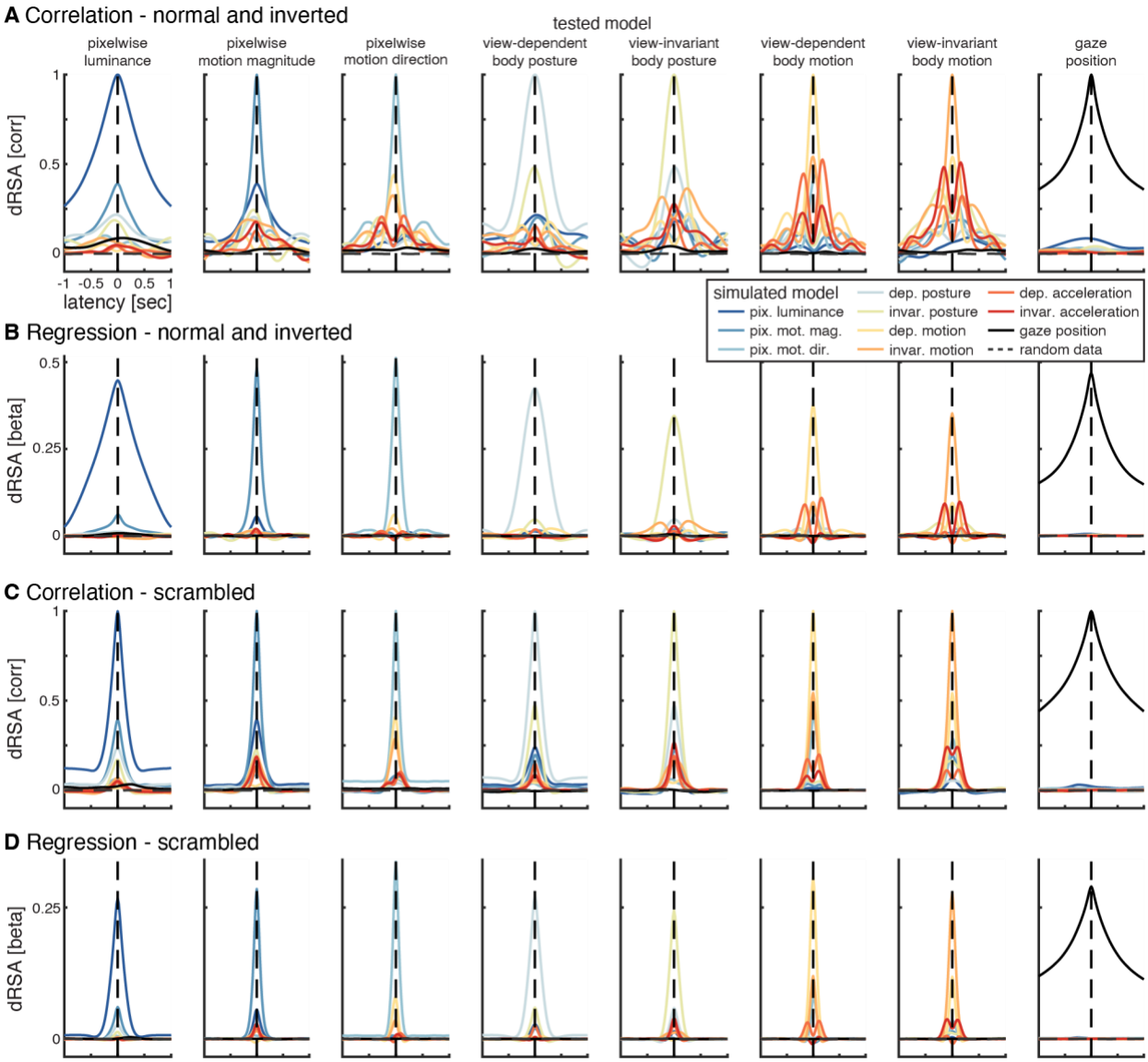

Fig. S2. Simulations. (**A**) dRSA on simulated data using simple correlation as similarity measure, in the normal and inverted viewing conditions. Note that the model RDMs do not differ between the normal and inverted viewing conditions. Colors indicate separate simulations of individual model RDMs. Columns indicate which model was tested. These results illustrate the effect of shared variance between various models and the effect of temporal autocorrelation within a given model. (**B**) Same but with regression weight as similarity measure for dRSA using principal component regression (PCR; see "Materials and Methods" for details). In short, simulated neural RDMs are exactly the same as in (**A**), but at the final step in dRSA (i.e., comparing neural and model RDMs) all models are included in a single regression, and only the beta weight of the tested model (columns) is illustrated. (**C**) and (**D**) Same as (**A**) and (**B**), respectively, but for the piecewise scrambled viewing condition. Importantly, these results indicate that PCR is effective at extracting only the model of interest from the simulated neural RDM, while largely regressing out the other models.

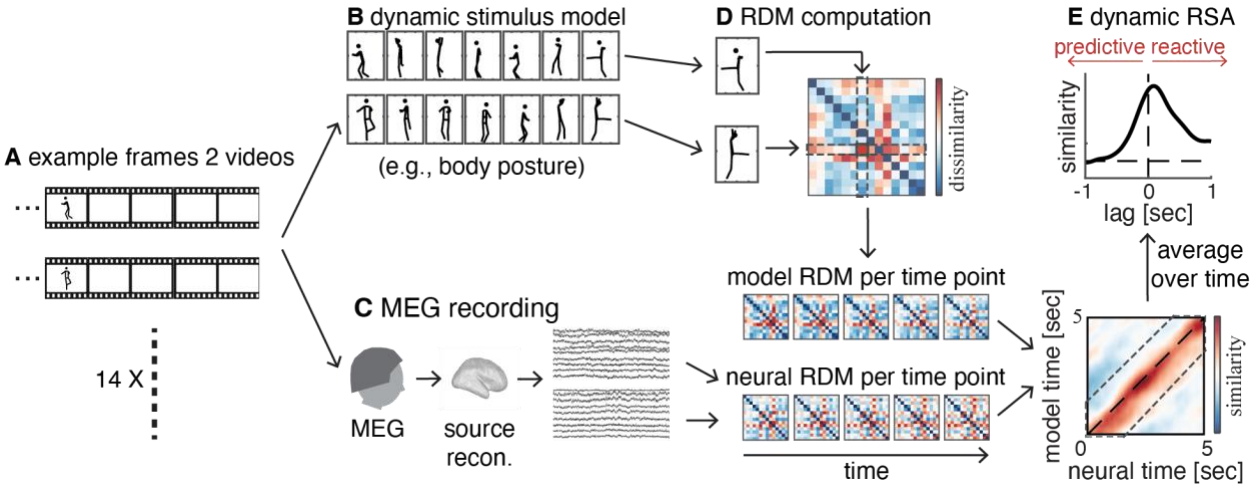

Fig. S3. Dynamic representational similarity analysis (dRSA) approach. Figure adapted from De Vries et al. 2023, Nat Commun, CC BY 4.0 (*1*). Please note that real ballet videos were used but for privacy reasons they are replaced in the preprint with illustrations of movie frames. Please see the original study for the original figure. (**A**) Subjects observed ~46 repetitions of 14 unique 5-sec dancing videos (Table S1) in 3 different viewing conditions (Fig. 1a and b) during MEG (i.e., ~15 repetitions per video per condition). (**B**) Stimuli were characterized at different levels using dynamic stimulus models (e.g., body posture; see Fig. S4 for all models). (**C**) Individual-subject source-reconstructed MEG signals within regions of interest (ROI; see Fig. S5 for ROI definitions) were used as features for subsequent steps. (**D**) Neural and model representational dissimilarity matrices (RDMs) were created at each time point based on pairwise dissimilarity in neural responses to the 14 stimuli and pairwise dissimilarity in stimulus feature models, respectively. Bottom: model and neural RDMs are shown for 5-time points. (**E**) Last, similarity between neural and model RDMs was computed for each neural-by-model time point (lower panel), using regression weights to test a specific model RDM, while regressing out other covarying model RDMs. This approach was validated through simulations (see subsection 'Simulations' and Fig. S2). Last, the 2-dimensional dRSA matrix was averaged along the diagonal to create a latency-plot (i.e., the lag between neural and model RDM; upper panel), in which peaks to the right or left of the vertical zero-lag midline reflect reactive or predictive neural representations, respectively. These dRSA latency plots are computed separately for each subject, ROI, model and condition, and then statistically tested.

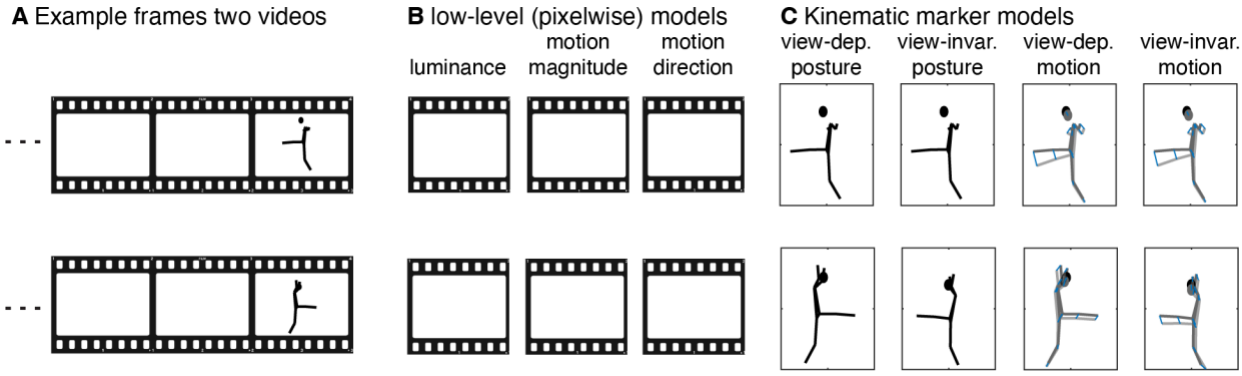

Fig. S4. Illustration of stimulus models. Figure adapted from De Vries et al. (2023) Nat Commun, CC BY 4.0 (*1*). Please note that real ballet videos were used but for privacy reasons they are replaced in the preprint with illustrations of movie frames. Please see the original study for the original figure. (**A**) Example frames of two videos (rows) to be correlated for creating the model RDMs. (**B**) Low-level visual models. Left: pixelwise luminance (spatially smoothed grayscale). Middle: magnitude of pixelwise motion (operationalized as optical flow vectors), with brighter colors indicating higher magnitude. Right: direction of pixelwise motion, with optical flow vectors indicated in blue and scaled 10 times for illustrative purposes. (**C**) Models based on 3D kinematic marker positions, from left to right: view-dependent body posture, view-invariant body posture (i.e., after aligning the kinematic markers between videos without changing their internal structure, through translation and rotation along the vertical axis), view-dependent body motion as indicated by blue lines (i.e., difference in kinematic marker position between two subsequent frames), and view-invariant body motion. Note that the main dRSA regression-based analysis included also view-dependent and view-invariant acceleration models, as well as gaze-position models based on individual participant eye-tracker data.

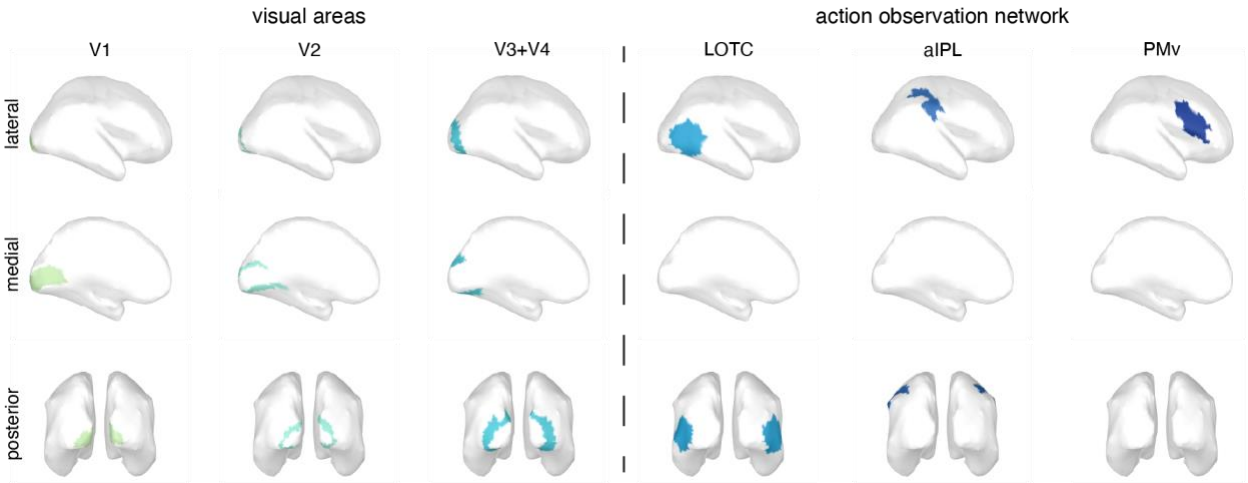

Fig. S5. ROI definitions. Figure reproduced from De Vries et al. 2023, Nat Commun, CC BY 4.0 (*1*). Cortical regions of interest (ROIs) for main analysis based on combinations of parcels as defined according to the Human Connectome Project (HCP) atlas (*2*). ROIs were a priori defined as follows (ROI name = atlas parcels): V1 = V1 [487 vertices]; V2 = V2 [491 vertices]; V3+V4 = V3 and V4 [490 vertices]; LOTC = V4t, FST, MT, MST, LO1, LO2, LO3, PH, PHT, TPOJ2 and TPOJ3 [565 vertices]; aIPL = PF, PFt, AIP and IP2 [403 vertices]; PMv = IFJa, IFJp, 6r, 6v, PEF, IFSp, 44 and 45 [466 vertices]. Note that these vertex amounts are based on the ICBM152 template cortical surface, exact amounts differ slightly between individual subjects. MEG responses at all vertices from both hemispheres were combined into a single vector to compute the neural RDM at a single time point (i.e., pairwise dissimilarity in the MEG response to the 14 action sequences). For visualization, all vertices within a single ROI are given the same color, which matches the color for each ROI in the main results (Fig. 2, 6 and S1). LOTC = lateral occipitotemporal cortex, aIPL = anterior inferior parietal lobe, PMv = ventral premotor cortex.

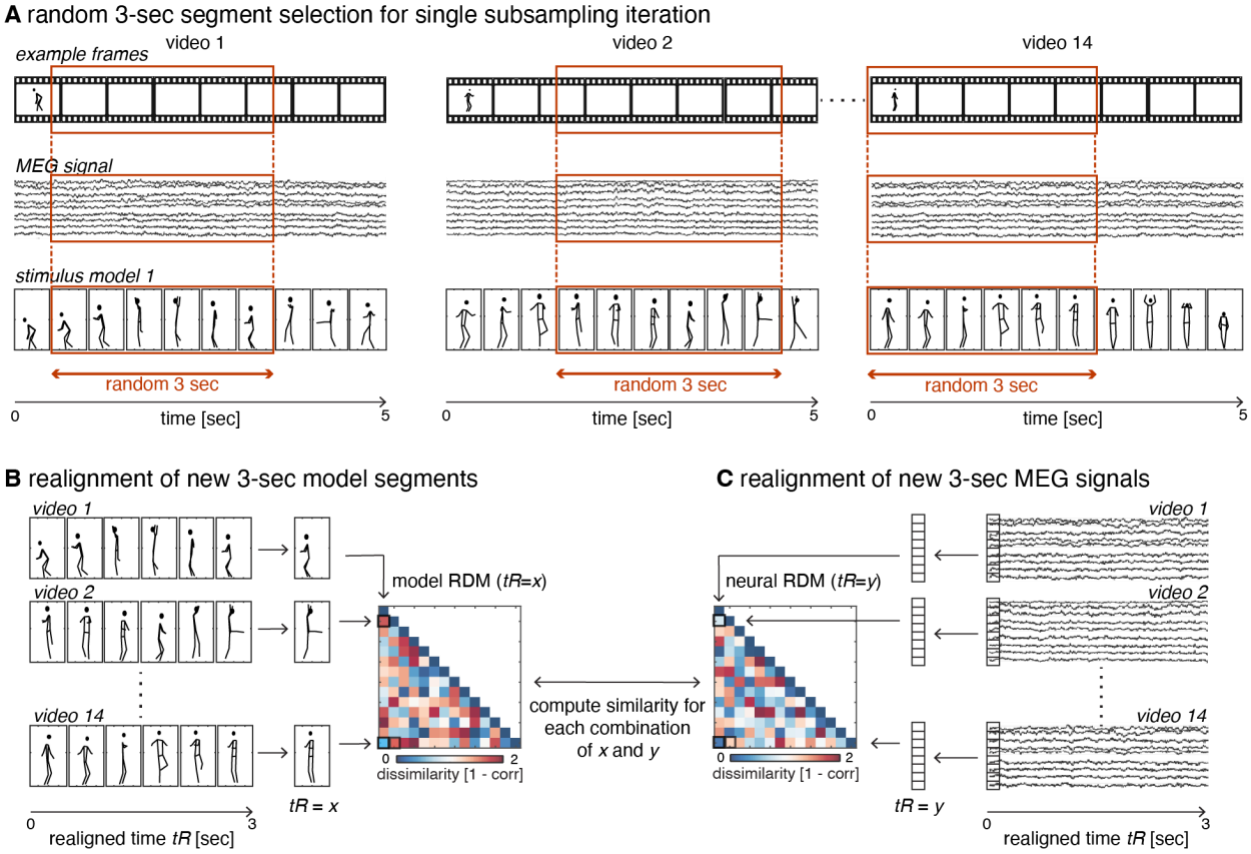

Fig. S6. Temporal subsampling and realignment. Figure reproduced from De Vries et al. 2023, Nat Commun, CC BY 4.0 (*1*). Please note that real ballet videos were used but for privacy reasons they are replaced in the preprint with illustrations of movie frames. Please see the original study for the original figure. Temporal subsampling and realignment were used to attenuate idiosyncratic temporal heterogeneity in dRSA results caused by arbitrary pairwise alignment specific to these 14 stimuli (see subsection "Temporal subsampling" in the "Methods" section). (**A**) On each of 1000 subsampling iterations, a 3-sec segment is randomly selected independently for each of the 14 stimuli (orange boxes). Crucially, while a different random 3-s window was selected for the 14 stimuli, for a given stimulus the same 3-sec window were selected for both neural and model data, thus keeping temporal alignment between those intact (indicated by vertical orange dotted lines). (**B**) and (**C**) Next, the 14 new 3-s segments are realigned, after which model and neural representational dissimilarity matrices (RDMs) are computed at each realigned time point tR. Last, similarity is computed for each combination of realigned model time (x) and neural time (y). Note that the steps in (**A**), (**B**) and (**C**), as well as the last step to compute the dynamic representational similarity analysis (dRSA) curves, are all done within a single subsampling iteration. After 1000 subsampling iterations, dRSA results are averaged over iterations.

Video	First figure	Second figure	Third figure	Fourth figure
1	Bow left	Arabesque left	Pirouette	Passé back
2	Bow left	Passé left	Jump right	Pirouette
3	Bow front	Jump back	Passé front	Arabesque front
4	Pirouette	Bow front	Passé front	Jump right
5	Pirouette	Passé right	Bow front	Jump back
6	Passé right	Bow right	Arabesque right	Pirouette
7	Passé right	Pirouette	Jump left	Arabesque front
8	Passé front	Arabesque front	Bow left	Jump right
9	Jump left	Bow right	Arabesque front	Passé left
10	Jump right	Pirouette	Passé front	Arabesque right
11	Jump back	Arabesque front	Pirouette	Bow front
12	Arabesque left	Pirouette	Jump back	Passé front
13	Arabesque right	Passé right	Pirouette	Bow front
14	Arabesque right	Jump left	Bow right	pirouette

Table S1. Ballet figures per stimulus. Each video stimulus consisted of a unique sequence of 4 smoothly connected ballet figures. The 4 figures were selected from a total of 5 unique figures. Note that all figures were presented from multiple viewpoints, to allow for both view-dependent and view-invariant body posture and motion models. The pirouette was always performed counterclockwise and could start and end at different angles.

		pixel wise lum.	pixel wise motion mag.	predictive pixelwise motion direction	lagged pixelwise motion direction	view dependent body posture	view invariant body posture	view dependent body motion	view invariant body motion
normal	V1	100	80	-	-	-	-	-160	-440
	V2	100	80	-140	-	-	-	-180	-440
	V3/4	100	80	-160	-	-	-	-180	-660
	LOTC	80	80	-160	-	-	40	-	-
	aIPL	-	-	-120	-	-	-	-	-
	vPM	-	-	-	-	-	-		-
inverted	V1	80	80	-160	-	-	-	-140	-680
	V2	80	80	-160	320	-	-	-140	-
	V3/4	80	80	-180	320	-	-	-140	-
	LOTC	100	80	-220	460	-	-	-120	-
	aIPL	-	-	-	-	-	-	-	-
	vPM	-	-	-	-	-	-	-	-
scrambled	V1	80	80	-	-	80	-	-	-
	V2	80	80	-	-	80	100	-	-
	V3/4	80	80	-	-	100	120	-	-
	LOTC	100	80	-	-	120	140	-	-
	aIPL	-	-	-	-	-	-	-	-
	vPM	-	-	-	-	-	-	-	-

Table S2. Peak latencies of the participant-average dRSA curves. Peak latencies are indicated in msec relative to zero lag between neural and model RDMs. Results are shown only for condition-ROI-model combinations with a significant main dRSA result (Fig. 2, 6, and S1), since latency estimation is inaccurate otherwise. We observed both a predictive (negative) and lagged (positive) peak for pixelwise motion direction in the inverted condition (Fig. 2b) and therefore computed peak latency separately for each. LOTC = lateral occipitotemporal cortex, aIPL = anterior inferior parietal lobe, PMv = ventral premotor cortex. Lum. = luminance, mag. = magnitude.

References supplementary material

1. I. E. J. de Vries, M. F. Wurm, Predictive neural representations of naturalistic dynamic input. *Nature Communications* **14**, 3858 (2023).

2. 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, D. C. Van Essen, A multi-modal parcellation of human cerebral cortex. *Nature* **536**, 171–178 (2016).