Selective distractor representations resolve multidimensional interference

Davide Gheza, Michael C. Freund, Thea R. Zalabak, Wouter Kool

Department of Psychological and Brain Sciences
Washington University in St. Louis
St. Louis, MO, USA

Abstract

How can humans manage multiple sources of information competing for attention? To approach this question, we adopted a multi-dimensional task-set interference paradigm that requires individuals to handle distractions from three independent dimensions. Behavioral results suggest that people track prior interference from each dimension to selectively modulate their attentional gain. Testing the mechanism of this adaptation at the neural level requires measuring multi-dimensional task representations. To achieve this, we applied representational similarity analyses and encoding models to human EEG and investigated how the history of interference simultaneously affected the time-resolved representations of target and distractor dimensions. EEG analyses revealed that target and distractor features are initially encoded in parallel, but distractors are rapidly suppressed around 250 ms, with suppression scaling with prior interference. Proactive task-level control enhanced the efficiency of this reactive suppression, without producing sustained preparatory changes. Behavioral and neural effects converged to show that dimension-specific interference and adaptation are associated with the speed of visual integration of each dimension, as captured by a connectionist model with temporal integration. Altogether, these results suggest that proactive control mechanisms operate on the speed and efficiency of a reactive suppression mechanism that constrains selective distractor representations from biasing the response process.

Adaptive behavior requires the flexible integration of a preponderance of sources of information, such as external and internal stimuli, goals, and context. For example, as you read this article, a landscape of constructs—thoughts, images, sensations—compete for your attention. However, to avoid mindless reading, attention must narrow to the text's meaning while filtering distractions. Our ability for attentional control (Miller & Cohen, 2001) enables such orchestration of processing, allowing us to occupy the representational space (Haxby et al., 2014) with attentional sets that emphasize goals and flexibly suppress transient distractors (Botvinick et al., 2001; Gratton et al., 1992). Importantly, contextual factors, such as prior knowledge on the text's topic or time pressure, can help set appropriate control settings in an anticipatory, proactive way (Braver, 2012; Bugg & Crump, 2012). Thus, attentional control does not only require the processing of multiple sources of information, but also a decision about how to weight them. This raises a fundamental question: how does attentional control orchestrate multiple competing representations with our limited representational space?

Although daily life demands flexible switching between sources of information, traditional cognitive control research examines interference from a single distractor. In these tasks (Eriksen & Eriksen, 1974; Liston et al., 2006; Simon & Rudell, 1967; Stroop, 1935), a task-relevant feature signals the correct response, while a second feature primes either a congruent or an incongruent response. For example, in the Stroop task (Stroop, 1935), the automatic tendency to read a color word interferes with the goal of naming the ink color when it is incongruent ("green" printed in red).

Extensive neuroscience research has uncovered how the brain regulates control in such tasks. A key insight from this work is that the magnitude of interference adapts based on recent demands: after experiencing interference, people become less sensitive to distractors (Gratton et al., 1992). Conflict monitoring theory (Botvinick et al., 2001) explains this as the result of a feedback loop that monitors for control demands and uses this to adjust the gain on task-relevant versus task-irrelevant representations. This theory has inspired a generation of computational models describing attention regulation as a conflict control loop (Egner, 2008; Holroyd & McClure, 2015; Jiang et al., 2014; Verguts & Notebaert, 2008; Yeung et al., 2004). Yet these models rely on a single source of interference and thus offer limited predictions for how attentional control operates in complex environments with multiple sources of distraction.

To address this gap, we recently developed the Multidimensional Task-Set Interference task (MULTI)(Gheza & Kool, 2025; Liston et al., 2006). On each trial, participants attend to one of four stimulus dimensions, as indicated by a cue, while the three non-cued dimensions act as simultaneous distractors. Because task cues change every few trials, all distractors remain salient and capable of eliciting interference (Hazeltine et al., 2011; Monsell et al., 2003). The data from this task revealed strong evidence that control adaptation operates in a *dimension-specific* fashion. Conflict from one distractor dimension triggered a reduction in interference from that same dimension but not from others. These findings suggest that attentional control operates through selective modulation of dimension-specific representations, rather than a unitary conflict signal that generalizes across sources of interference.

This dimension-specific adaptation challenges the dominant neuroscientific model of cognitive control at several levels. This account suggests that, in bi-dimensional tasks, neural populations in anterior cingulate cortex (ACC) detect response- or stimulus-level conflict (Ebitz et al., 2020; Fu et al., 2022; Sheth et al., 2012) and signal the need for increased control to distal areas via neural synchronization mediated by theta oscillations (Botvinick et al., 2001; Cavanagh & Frank, 2014; Senoussi et al., 2022; Smith et al., 2015; Verguts, 2017). Dorsolateral prefrontal posterior parietal cortices implement this signal by biasing task-relevant representations (Egner & Hirsch, 2005; Miller & Cohen, 2001) or suppressing distractor processing (Danielmeier et al., 2011; Padmala & Pessoa, 2011; Ritz & Shenhav, 2023), while the subthalamic nucleus transiently suppresses motor output until conflict is resolved (Frank, 2006; Wessel & Anderson, 2024; Zaghloul et al., 2012). Together, these mechanisms support trial-by-trial control adjustments.

However, a single conflict–control loop may be too crude to resolve interference when multiple distractors vary independently. Because a compressed, univariate estimate of conflict cannot identify the source of interference, models relying on this signal are unable to capture dimension-specific control adaptation (Gheza & Kool, 2025). However, a mechanism based on multiple independent conflict loops would require not only a limitless number of specialized neural manifolds within the ACC, but also a multiplexed communication system to relay simultaneous adaptation signals to posterior regions.

An alternative view emerges from recent work on context-dependent neural representations within PFC and ACC prefrontal and cingulate cortices (Bernardi et al., 2020; Ebitz et al., 2020; Flesch et al., 2022; Ritz & Shenhav, 2024). An emerging body of work in humans and non-human primates shows that the representational content of these areas is dynamic and context sensitive (Stokes et al., 2017), allowing task rules to flexibly influence the encoding of stimulus-response contingencies (Kikumoto et al., 2022; Mante et al., 2013). In short, these regions encode task dimensions in partly orthogonal subspaces, with context selectively modulating the response readout axis. Such representational geometry enables flexible selection of relevant features while minimizing interference (Badre et al., 2021; Bernardi et al., 2020; Ritz & Shenhav, 2024; Weber et al., 2023). In the MULTI, dimensionality arises from stimulus complexity rather than rule switching, yet similar mechanisms may permit conflict-driven adjustments to be applied specifically to the distractor dimension that elicited interference. If so, multidimensional control could emerge not from multiple conflict loops, but from dynamic, context-sensitive reconfiguration of representational subspaces.

Here, we apply multivariate methods to human electroencephalography (EEG) data to investigate how attentional control modulates the representation of multiple stimulus dimensions. Although EEG cannot directly measure the activity of individual neural populations, changes in scalp topography necessarily denote changes in the underlying neural generators (Vaughan, 1982). We therefore examined whether the neural representations of cued and non-cued dimensions change trial-by-trial according to distractor-specific conflict history. The high temporal resolution of EEG further allowed us to characterize the within-trial dynamics of these adaptations and whether they reflect reactive or proactive control adjustments (Braver, 2012). To probe sustained, context-based, proactive control, we also manipulated dimension-specific proportion congruency, creating persistent associations between dimensions and control demands. To foreshadow the core findings, we confirm behavioral and model-based predictions about distractor-specific adaptation and identify a reactive-suppression mechanism that operates across temporal scales, underlying both conflict adaptation and proportion-congruency effects.

Results

Multidimensional interference and conflict adaptation

We recorded EEG while twenty-seven participants performed the MULTI (Gheza & Kool, 2025), which requires a context-dependent feature search over two objects that vary in four dimensions: color (blue/orange), shape (oval/rectangle), edge (solid/dashed), and pattern (horizontal/vertical); Fig. 1A). At the start of the experiment, each participant was instructed which feature was the correct choice for each dimension (e.g., blue, oval, dashed vertical), and this mapping stayed fixed throughout the task. On each trial, a brief letter cue indicated which dimension was relevant for selecting between two objects. The other three dimensions acted as distractors which could be congruent (their target feature appeared on the same side as the cued target) or incongruent (on the opposite side; Fig. 1B). Therefore, congruency parametrically varies over a discrete scale from 0 to 3. Cues switched every few trials, rendering each dimension periodically task-relevant and salient (Fig. 1C), and allowing us to track how attention control adapts to changing contexts and distractor congruency histories.

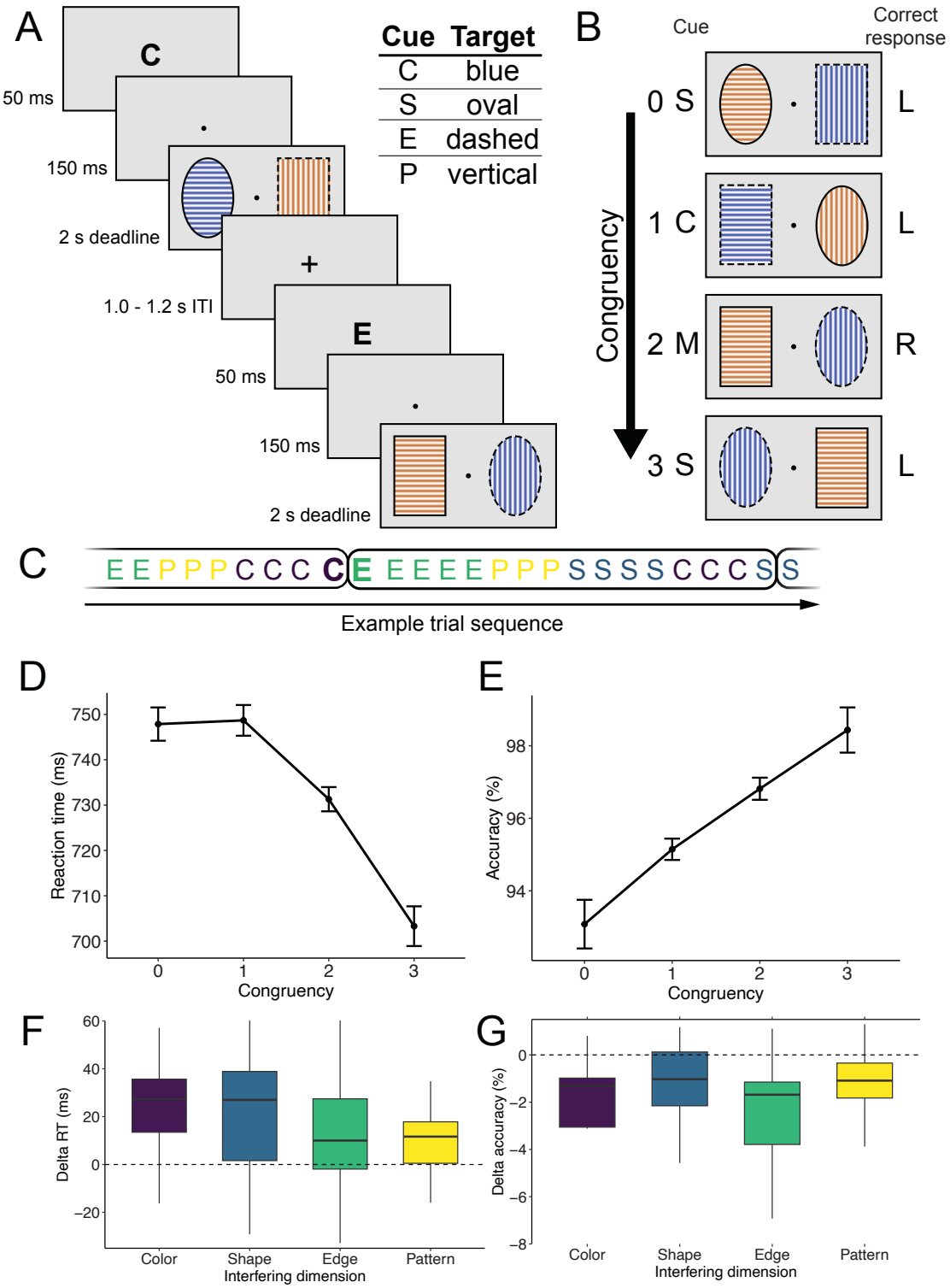

Figure 1. Multidimensional task-set interference task. a, Sequence of two example trials. On each trial, two objects were presented side by side that differ across four dimensions: the colour (orange or blue), the shape (oval or rectangular), the edge (solid or dashed) and the direction of the Gabor (vertical or horizontal). Participants were instructed to pay attention to only one 'task' dimension, previously cued by a letter briefly presented for 50 ms, followed by a fixation dot for 150 ms. For each dimension, participants needed to find a target feature, as instructed at the beginning of the experiment, and select the corresponding object (left (L) or right (R) response). In the example reported here, the target features are the blue colour, the oval shape, the dashed edge and the vertical

pattern. All stimuli remained on screen until a response was given. Response deadline was set to 2 s and the a fixation cross was presented during a variable intertrial interval.

Behavioral results, analyzed using generalized Bayesian linear regressions, confirmed that participants were sensitive to the cumulative interference from all non-cued dimensions (Fig. 1DE). Reaction times decreased and accuracy increased as more distractors supported the correct response, as indicated by strong linear effects of congruency (evidence ratio, ER = ∞). Indeed, all non-cued dimensions produced reliable interference (Fig. 1FG), a difference in task performance between congruent and incongruent trials for a given non-cued dimension for both response time (RT; color: Bayes Factor (BF) = 6.5×10^4, shape: BF = 448, edge: BF = 6, pattern: BF = 24) and accuracy (color: BF = 766, shape: BF = 12, edge: BF = 614, pattern: BF = 70). Although each dimension generated interference, there was substantial heterogeneity in their potency, with the pattern dimension exerting the weakest effect.

Critically, attentional control adapted dynamically to recent conflict (Fig. 2AB). Participants became less susceptible to interference after low-congruency trials, consistent with a congruency sequence effect (CSE; (Gratton et al., 1992)), as confirmed by strong evidence for an interaction between current and previous congruency on both RT and accuracy (ERs = ∞). Following incongruent trials, interference effects were reduced, whereas following fully congruent trials, participants showed slower and less accurate responses on high-conflict trials ($M_{30\text{-}00}$ = 50 ms, -7.2%) and facilitation on low-conflict trials (leading to shorter RTs and higher accuracy; M33-03 = -36 ms, 1.2%).

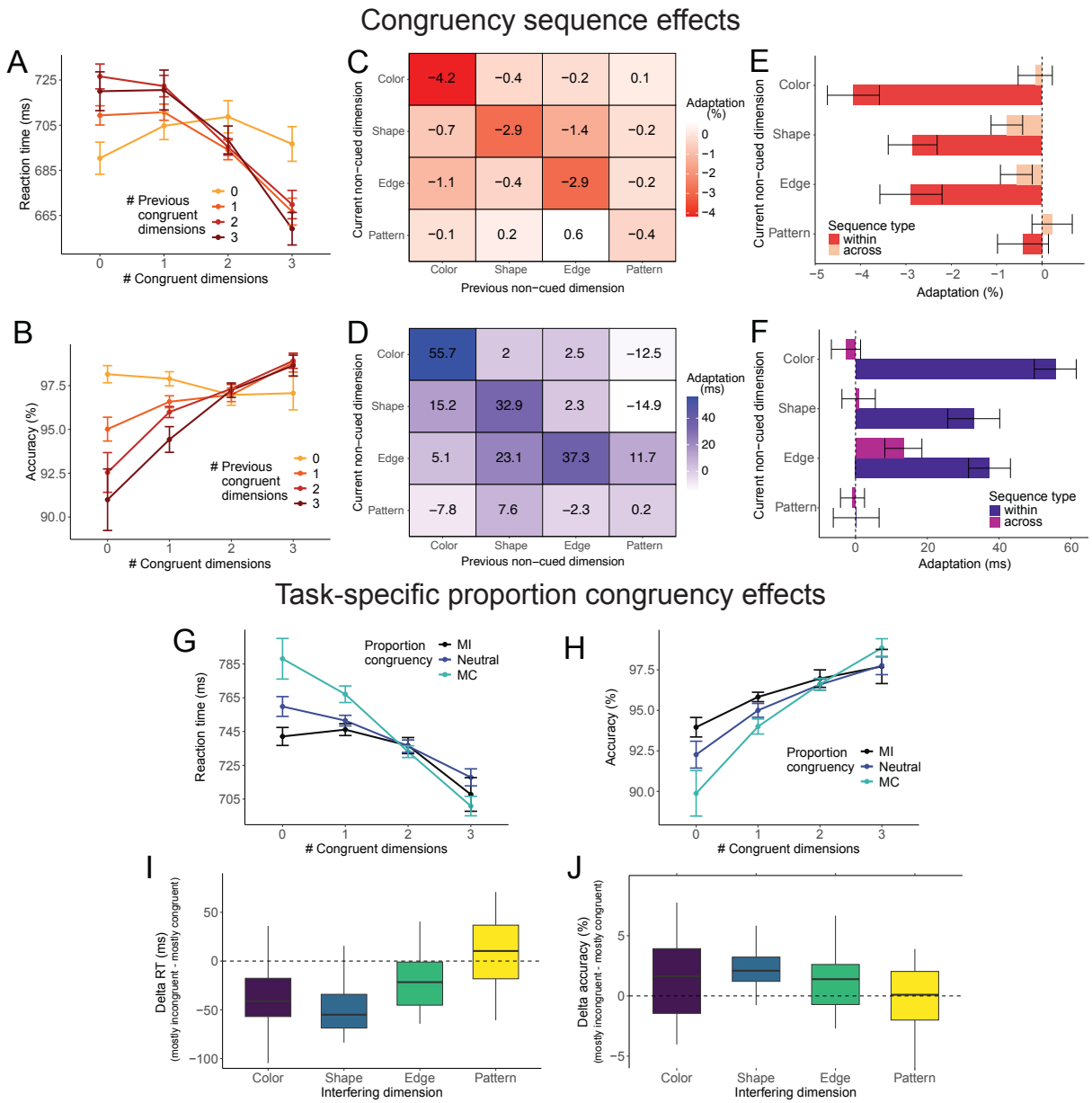

Figure 2

To test whether these adjustments generalize across multidimensional distractors, we compared adaptation within versus across non-cued dimensions. Specifically, we test whether conflict from a given dimension modulated subsequent interference within that same dimension (e.g., previous color congruency on current color sensitivity) and on other dimensions (e.g., previous color congruency on current shape sensitivity). We found strong evidence that adaptation is dimension-specific. Conflict from a given distractor reduced its own subsequent interference but did not transfer to other dimensions (Fig 2CD, diagonal; Fig. 2EF, darker colors). This effect was robust for dimension color (RT: BF = 3249005.3, accuracy: BF = 1540.1), shape (RT: BF = 287.1, accuracy: BF = 10.7), edge (RT: BF = 203.3, accuracy: BF = 20.6), but not for pattern (RT: BF = 0.2, accuracy: BF = 0.7), mirroring the weaker interference of that dimension. Underscoring the specificity, we did not find evidence for across-dimension adaptation (RT: range BFs = 0.2 - 1.9, accuracy: range BFs = 0.2 - 1.3; Fig. 2CD, off-diagonal; Fig. 2EF, lighter colors). These results, which replicate our previous findings (Gheza & Kool, 2025), demonstrate that control in multidimensional environments is achieved through independent adjustment of attentional gain for each distractor dimension, rather than a global conflict signal.

A computational model of dimension-specific adaptation

To mechanistically describe control adaptation in the MULTI, we previously developed a neural network model inspired by conflict-monitoring theory (Botvinick et al., 2001; Gheza & Kool, 2025). This model integrates evidence from four stimulus dimensions using independent pathways, which converge on a response layer (Fig. 3). This model departed from prior models, which assume a single global control loop with conflict measured at the response level. Instead, conflict in our model is measured as entropy (i.e., Hopfield energy, Hopfield, 1982) between dimension-specific representations and the response layer. Implemented as a static neural network, this architecture successfully captured key behavioral signatures such as interference from multiple distractors and dimension-specific CSEs.

However, that model treated conflict as a single scalar per trial, not considering the temporal dynamics of interference within a trial. However, prior work (Koob et al., 2023; Mittelstädt et al., 2023; Ritz & Shenhav, 2023) and our behavior results suggest that interference effects and adaptation may unfold at a sub-trial timescale. To address this, we developed a dynamic leaky competing accumulator (LCA) variant of this model (Fig. 3A). In this model, each pathway integrates evidence according to its own temporal constant (τ), while mutual inhibition between accumulators yields competition between dimensions. This simple extension allows us to capture the heterogeneity in the temporal evolution and strength of distractor interference, such as the weaker effects observed for the pattern dimension.

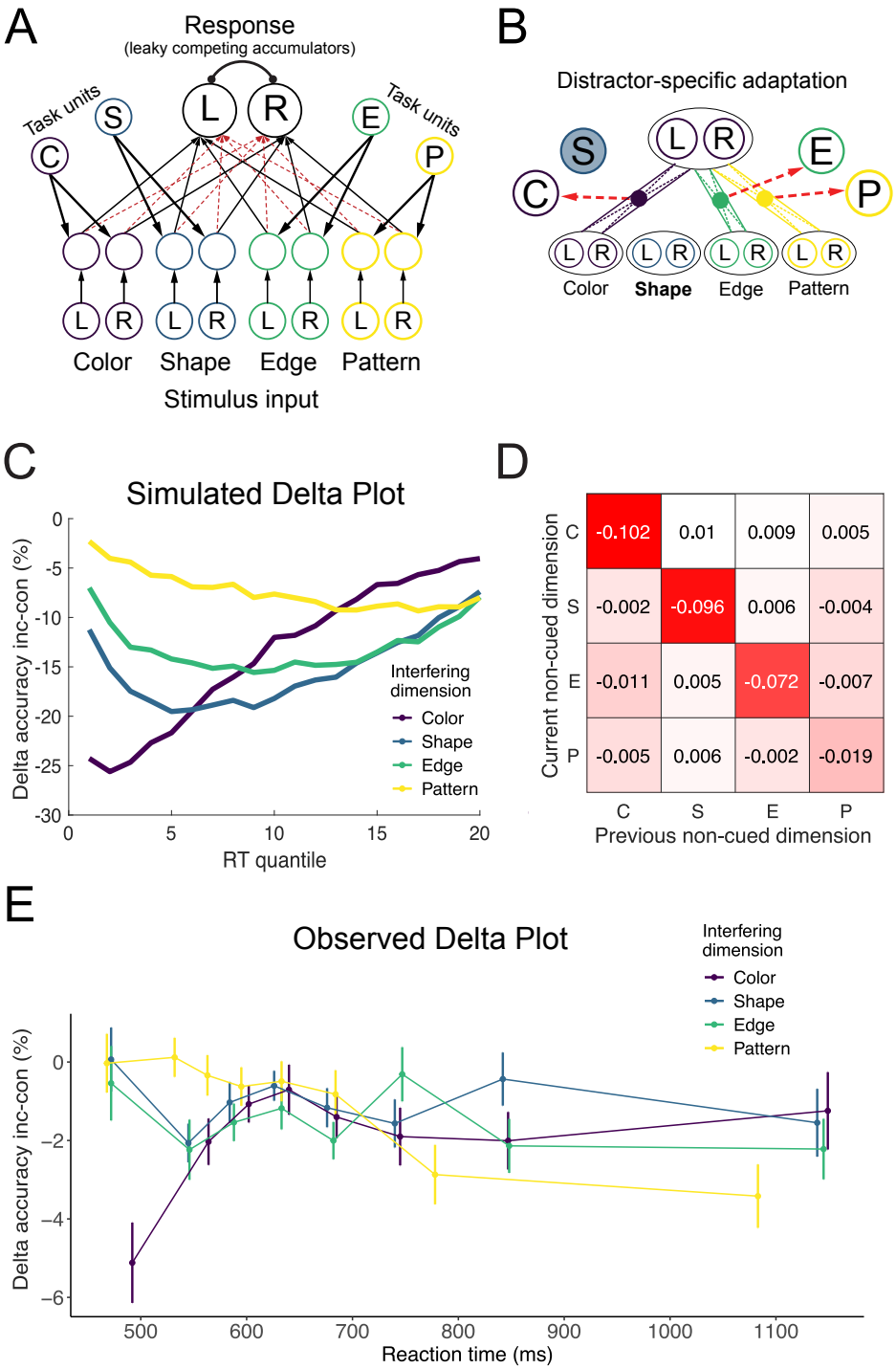

Figure 3. Model architecture generating dimension-specific adaptation and processing dynamics.
Each trial begins with stimulus and task inputs projecting onto an 8-unit hidden layer organized into four dimension pathways. Hidden-layer units integrate their constant net input (stimulus, task cue, and bias) over discrete processing cycles via pathway-specific leaky accumulation, after which activity is passed through a logistic nonlinearity. Hidden activity feeds into two response-layer units, which likewise undergo leaky temporal integration of their weighted hidden inputs and logistic activation. Choice is determined by a pair of opposing evidence accumulators that integrate noisy moment-to-moment evidence proportional to the difference between the two response activations; a response is made when either accumulator reaches threshold, yielding both accuracy and reaction time. During each trial, pathway-specific energy is computed from the weighted interaction between hidden and response activity; the area under its within-trial temporal trajectory provides a trial-level conflict estimate. This

conflict signal adaptively modulates task-cue input on subsequent trials via a slow control process, allowing the model to capture history-dependent adjustments in processing dynamics.

Simulations with this model revealed that adaptation arises from selective suppression of distractor dimensions rather than target enhancement (Fig. 3B). Dimension-specific adaptation occurred even in the face of heterogeneity in the timing and strength of interference (Fig. 3CE). Importantly, weaker dimensions indeed led to reduced adaptation (Fig. 3D), as we observed in our behavioral data (Fig. 2CE). Together, the behavioral results and computational modeling converge on the prediction that neural signatures of adaptation should appear as dimension-specific modulations of distractor representation strength, and that the timing and strength of this adaptation should vary across dimensions in proportion to interference potency.

Frontal midline theta indexes multidimensional control allocation

Frontal midline theta oscillations (FMT; 3–8 Hz) are a hallmark of cognitive control, reliably linked to conflict detection and control engagement in tasks involving stimulus- or response-level interference (Cavanagh & Frank, 2014; M. X. Cohen, 2017; M. X. Cohen & Cavanagh, 2011). However, in the MULTI, interference arises from the interaction of multiple competing stimulus and response representations, rather than a single source. This raises the question of whether the same canonical theta-mediated control mechanisms are engaged under multidimensional control demands. Analyzing theta power therefore provides a converging neural test for the recruitment of cognitive control in the MULTI.

To address this, we examined FMT as a function of prior and current conflict. Theta power increased following stimulus onset and peaked around the response (Fig. 4). Its modulation across types of trial sequences was consistent with a dual role in conflict monitoring and control allocation. Both incongruent-incongruent and congruent-incongruent sequences elicited elevated theta power, with stronger activity for the former sequence type. In contrast, congruent-congruent trial sequences elicited the weakest theta power (Fig. 4AB). FMT was robustly modulated by current congruency at frontocentral electrodes (Fig. 4C, top), consistent with conflict detection. Notably, we also observed marginal effects of previous conflict at occipitoparietal channels (Fig. 4C, bottom), which is suggestive of a role in control allocation prior to the response.

Single-trial multiple regressions (Fig. 4D) further revealed dissociable functional roles across the scalp: frontocentral theta power predicted slower responses, replicating classic signatures of conflict detection (M. X. Cohen & Cavanagh, 2011), whereas parietal and lateral-occipital theta predicted *faster* responses. Together, these observations are suggestive that, in this paradigm, theta may also reflect the implementation of control, specifically the suppression of distractor representations.

Overall, these findings support a canonical conflict-control account of theta oscillations in the MULTI. This observation motivated our subsequent analyses, combining neural decoding and representational similarity analysis (RSA), to characterize how control modulates the neural encoding of targets and distractors.

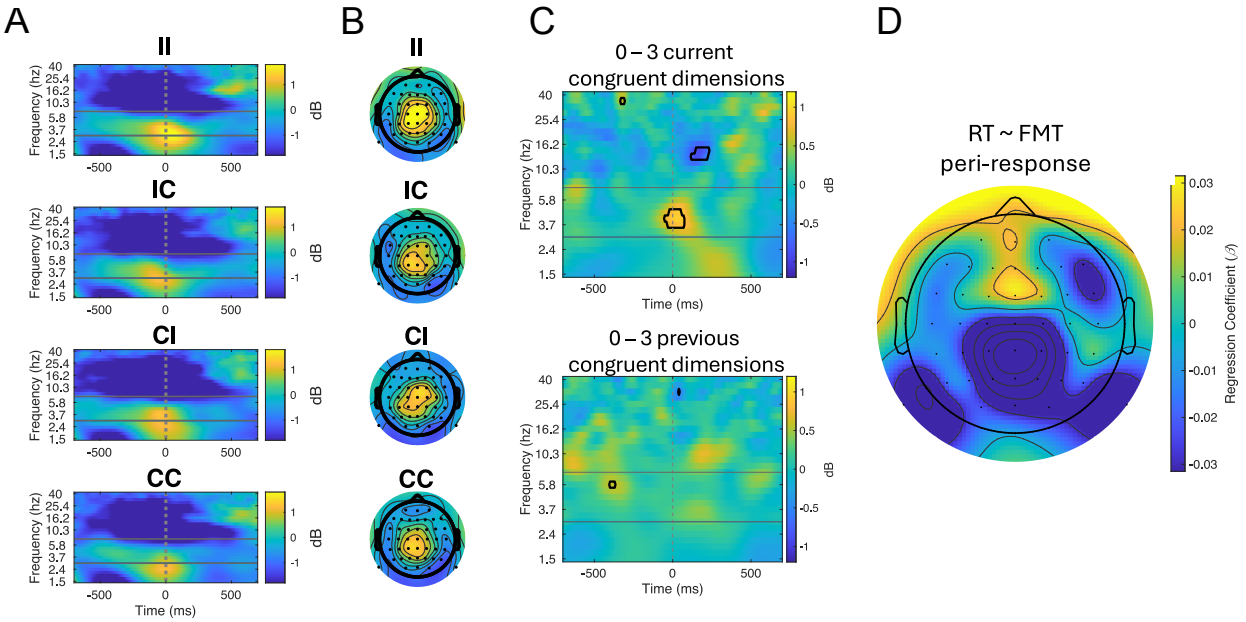

Figure 4. A. Time–frequency plots of response-locked power at frontocentral electrodes for four trial-sequence types (II, IC, CI, CC), where the first letter indicates congruency in the previous trial and the second letter the congruency in the current trial; "I" and "C" correspond to trials with 0 and 3 congruent dimensions, respectively. Data were averaged over frontocentral channels: Cz, FCz, FC1, FC2, C1, C2, CPz, CP1, CP2. Theta (3–8 Hz) power peaks around the response, with II sequences showing the strongest activity and CC the weakest. B. Corresponding scalp topographies of theta power (3–8 Hz, -75 to 75 ms peri-response window), revealing frontocentral modulation as a function of congruency sequence. C. Cluster-corrected time-frequency contrasts (alpha = 0.01, 5000 permutations) showing significant modulation of theta by current congruency (top) and previous congruency (bottom). The top panel shows effects averaged over frontocentral channels (Cz, FCz, FC1, FC2, C1, C2, CPz, CP1, CP2), highlighting robust theta increases for incongruent trials. The bottom panel shows effects over parieto-occipital channels (Pz, POz, CPz, CP1, CP2, P1, P3, P5, PO3, PO7, O1, P9, P2, P4, P6, PO4, PO8, O2, P10), indicating weaker but detectable influences of prior conflict. A-C. Data were response-locked and baseline-normalized using the –800 to –500 ms prestimulus window. Only trials that met all of the following criteria were included: correct responses, following a correct response, not being switch trials, and associated with neutral ITPC. D. Topography of regression coefficients linking peri-response (-200 to 0 ms) theta power to single-trial RTs. Higher theta at frontocentral sites predicts slower responses, whereas parietal and lateral-occipital theta predicts faster responses, consistent with conflict monitoring and control implementation processes, respectively. All correct trials were included in regression models. Single-trial power was not baseline-normalized but z-scored across trials and frequencies. RTs were also z-scored.

Decoding the neural mechanisms of distractor-specific adaptation

To study the temporal dynamics of multidimensional attention control, we analyzed time-resolved neural activity to uncover neural evidence for how and when stimulus representations are expressed and modulated over time. To do this, we adopted the decoding-RSA framework developed by Kikumoto and Mayr (Kikumoto & Mayr, 2020), which combines multivariate encoding with representational similarity analyses to track the strength and temporal evolution of task-related information (Fig. 5A). After successfully training a decoder to detect the cue and response of a given trial from electrodes and spectral power (Fig. 5B), we used confusions in this classification to reveal the neural representations corresponding to the cued task, side of the target feature, and their conjunction.

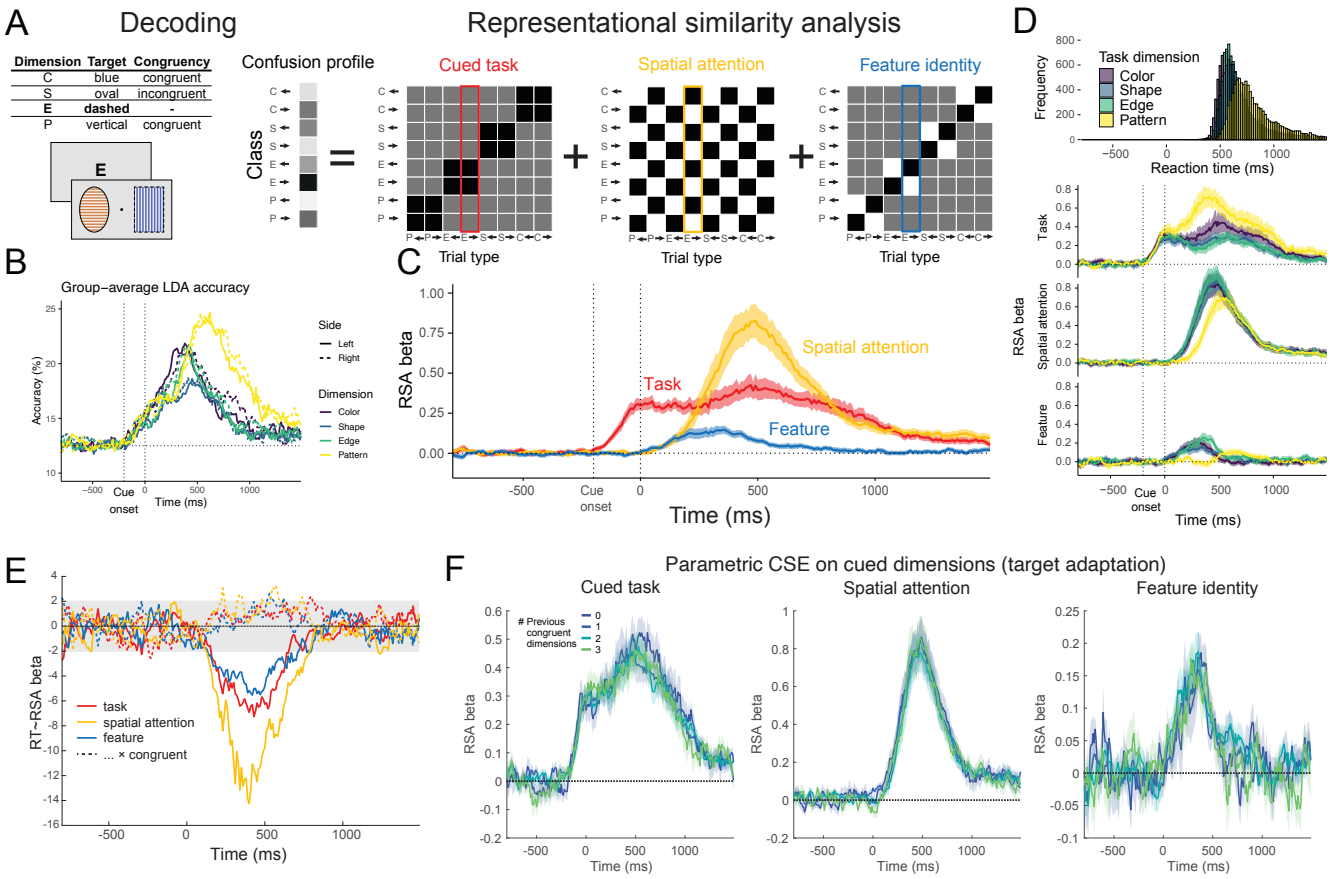

Figure 5

The temporal profile of decoded representations followed the structure of the trial (Fig. 5C). Cue coding emerged rapidly after cue onset (-200 to -150 ms) and showed a bimodal profile: an early plateau around stimulus onset, followed by a second rise around 500 ms. This trajectory, which was shared across cued dimensions, plausibly reflects the initial perceptual encoding of the cue, followed by its later integration with task-rule representations. Choice coding, reflecting activity related to the response or spatial attention, emerged after stimulus onset and peaked just before the median reaction time (~500 ms). Feature coding, the representation of the target identity, shared the initial temporal profile of Choice coding, but peaked and subsided earlier. This suggests that the perceptual encoding of the target feature primarily serves to guide attentional orienting, but that it is not needed to produce a response.

We found similar profiles when we split the results of this encoding-RSA analysis by cued dimension (Fig. 5D). Notably, both Choice and Feature coding were delayed for the pattern dimension. This dimension requires the discrimination of Gabor orientations, which is more challenging when stimuli are lateralized, leading to slower integration of evidence. This is consistent with the observation that responses are slower for this dimension, and that the dimension yields weaker interference when it acts as a distractor. Consistently, the within-trial temporal dynamics of interference (Fig. 3E) revealed that color interference emerged even for the fastest responses, while pattern interference was absent for fast RTs and increased only for slower responses.

Multivariate task representations predict behavior but are stable across conflict history

Next, we confirmed that the neural representations uncovered with the decoding-RSA (Kikumoto et al., 2024; Kikumoto & Mayr, 2020) approach were behaviorally meaningful. For each time point within the RSA profile, we used multi-level models to predict RT from the strength of the Cue, Choice, and Feature coding strengths, and their interaction with parametric congruency. We also added nuisance predictors (task cue, response side, and their interaction) to reduce biases due to systematic RT differences across conditions. All three representations predicted

faster responses, with effects arising at the stimulus onset and peaking around 500 ms (Fig. 5E, solid lines). These results indicate that stronger expression of task-relevant representations facilitates task performance. We also found weak but consistently positive interaction effects with parametric congruency, which suggests that the benefit of these representations was reduced when multiple distractors primed the correct response (Fig. 5E, dashed lines).

We next asked whether the strength of these representations is modulated by recent interference. If trial-to-trial control adaptation operates on the cued task, we should expect stronger (weaker) Cue or Feature coding following trials with low (high) conflict (i.e., target adaptation). To test this, we stratified the single-trial RSA data by the number of congruent dimensions on the previous trial (excluding post-error and post-switch trials). However, the strength of Cue, and Feature representations were unaffected by prior congruency (Fig. 5F). This absence of target modulation mirrors the behavioral results, which suggests that conflict adaptation primarily modulates the strength of distractor representations.

Multivariate distractor representations are shaped by recent conflict

Instead, our behavioral and modeling results predict that attention to non-cued dimensions' features should reflect their history of interference. Specifically, neural representations of previously incongruent distractors should be suppressed, whereas those of previously congruent distractors should be enhanced. To test this hypothesis, we needed to disentangle the representations of all three distractors and the target simultaneously, despite their changing identities and congruencies across trials.

We developed a new decoding-RSA model to jointly estimate the location and identity of each dimension's instructed feature (Fig. 6A). This approach predicts the decoding probabilities from a set of predictors that represented the Cue × Choice conjunction vector of the target feature, as well as the three analogous vectors corresponding to the location of the instructed feature of each non-cued dimension. This design allowed us to detect the representation of distractor features alongside task-relevant information, even though the classifier was trained to discriminate task-relevant information only (e.g., in Fig. 6A, the target feature is the dashed edge). In other words, we tested whether confusion in the classification profile indicates a drift of attention towards distractors.

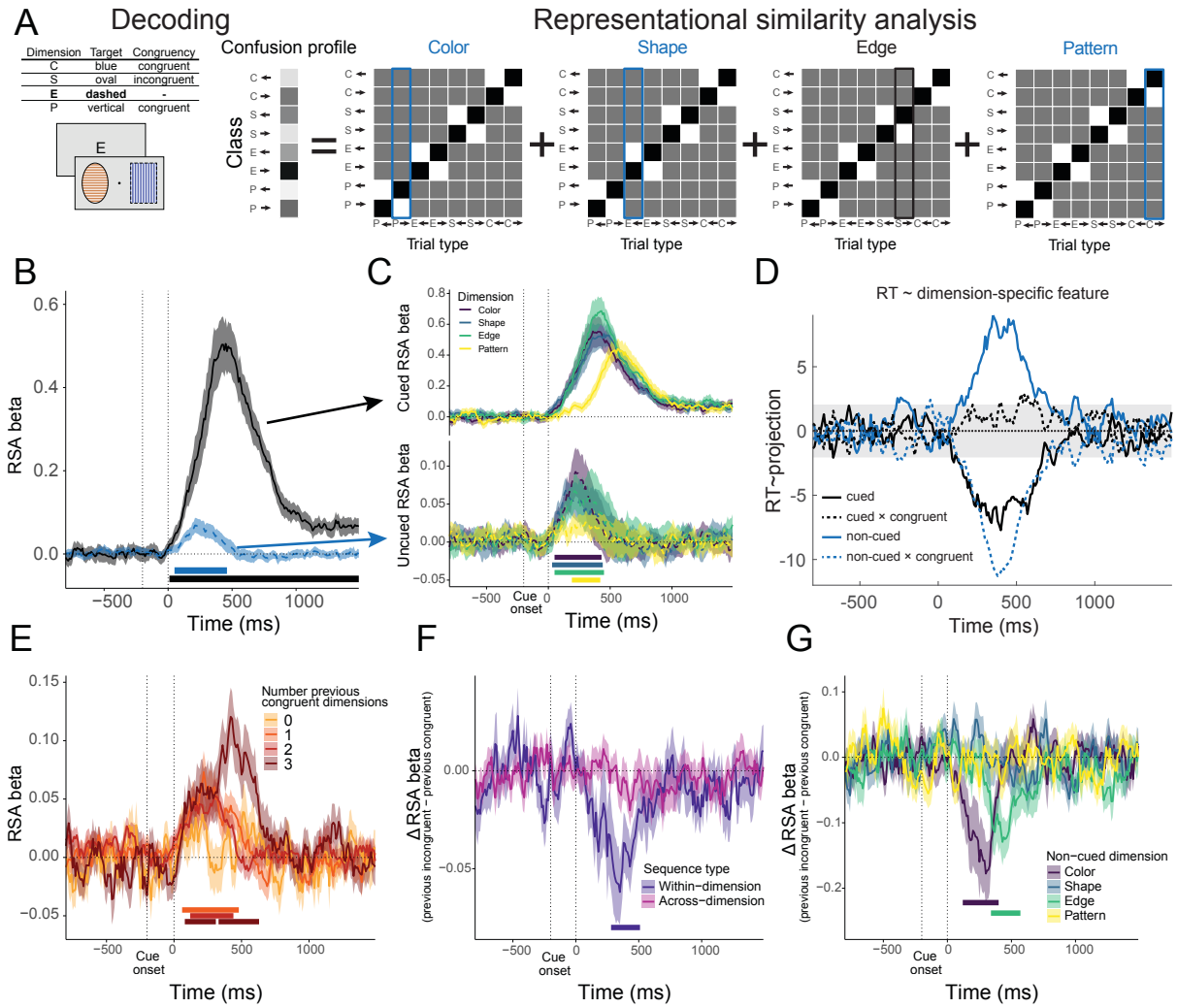

Figure 6

Our results, depicted in Figure 6B, confirmed this hypothesis. Both the cued dimension (black) and the average of the distractor (blue) representations rose after stimulus onset. Importantly, the expression of distractor representations was robust even when controlling for their congruency level across trials. When broken down by non-cued dimension, we were able to reliably detect distinct representations for each distractor (Fig. 6C, bottom) alongside each cued dimension (Fig. 6C, top). The strength of these distractor representations reflected their behavioral interference potency, with the pattern dimension exhibiting notably weaker encoding. For each dimension, distractor coding peaked around 250 ms after stimulus onset, consistent with early attentional capture. These early and short-lived distractor profiles point to a fast-acting inhibitory mechanism that suppresses distractor processing.

Next, to assess the behavioral relevance of these distractor representations, we regressed trial-by-trial variations in reaction times on the strength of each encoded representation using multi-level models. These included the RSA betas for the target feature and the average of the three distractor features, as well as their interactions with parametric congruencies, and the same nuisance predictors mentioned above. We found that target feature representations strongly predicted faster responses (Fig. 6D). More importantly, the representation of distractors either strongly facilitated performance when all non-cued dimensions were congruent, or impaired behavior when they were incongruent. In other words, the impact of distractor representations depended on whether they supported or conflicted with the task goal.

The central hypothesis of our study was that attention control primarily operates on distractor representations. To test this, we stratified the single-trial average distractor coding by the number of congruent dimensions in the preceding trial. If control adaptation specifically targets distractors, their representational strength should vary with recent interference history. This is precisely what we observed. The average distractor representations were strongest following fully congruent trials and weakest and more short-lived following fully incongruent trials, with intermediate levels of congruency falling along this continuum (Fig. 6E). These parametric distractor adaptations mirror the ones observed in behavior (Fig. 2AB).

Finally, we tested whether these adaptations were dimension-specific. Following the logic of the behavioral analyses, we disentangled within- versus across-dimension adaptation effects (Fig. 2CD). For each combination of previous and current distractor, we computed adaptation effects as the difference in representational strength between the temporal profile following incongruent versus congruent trials. Note that, unlike the behavioral analysis, which measured double differences in RTs, these neural adaptation effects capture the simple difference in representation strength for a given distractor following incongruent versus congruent trials (of either the same or different dimension).

The results reveal how adaptation effects unfold over time within a trial. We observed robust within-dimension adaptation effects (on diagonal, blue profile), marked by a negative deflection peaking around 350 ms, but no significant across-dimension effects (Fig. 6F). When examining individual distractor dimensions, color and edge showed the most pronounced adaptation effects, emerging earlier for color (peak: 300 ms) and later for edge (peak: 430 ms). The shape dimensions exhibited a weaker, more diffuse pattern, while no trend was appreciable for pattern (Fig. 6G). Finally, it is worth noting that within-dimension adaptation reflected both the suppression of distractor representations following incongruent trials and their enhancement following congruent ones (Fig. S1).

In sum, these results suggest that distractor features are systematically processed alongside task-relevant information. However, their influence on the response is curtailed by a reactive suppression mechanism. This suppression appears to be triggered after distractor processing begins, but its efficiency increases with prior exposure to similar interference. Thus, prior conflict tunes the speed and strength of this reactive mechanism, effectively reflecting a form of proactive control. Importantly, this form of control operates in a dimension-selective manner, consistent with behavioral adaptation effects and the predictions generated by our model.

Dissociable contributions of adaptation on attentional orienting and feature representations

Our results so far indicate that attention control operates through feature-based attention. Because the side containing each dimension's target feature is selected randomly on every trial, this attention mechanism is inherently location-independent. Indeed, our RSA measure of distractor representations pinpointed a dimension-specific control adaptation mechanism. However, these results do not distinguish between two possible routes by which control regulation could be achieved. One possibility is that control acts on the *strength of feature encoding*, amplifying or suppressing the perceptual representation of the specific stimulus features based on their relevance. Another possibility is that control modulates *the degree of spatial orienting* towards those features, adjusting the deployment of attention across locations to limit the influence of irrelevant information. However, because of the limited degrees of freedom in the classification profile, the RSA approach could not disentangle the feature-based representations of a given distractor (Feature coding) from the spatial attention towards the same features (Choice coding), making it unclear to what degree these mechanisms undergo adaptation.

To resolve this ambiguity, we developed a complementary *encoding* model that separates the feature-based from the spatial components of attention control. This approach adapts state-space regression methods from animal electrophysiology (Mante et al., 2013) for use with EEG. For each trial and time point, we defined four predictors that captured the target location of each cued dimension (i.e., the choices) and projected them onto the EEG data to estimate dimension-specific beta maps. Averaging these beta vectors defined a dimension-general attention axis, which we then residualized from the original vectors to isolate dimension-specific feature coding. This procedure yielded five orthogonal models, four feature-specific and one dimension-general, whose projections onto held-out data provided de-mixed estimates of task representations over time (see Method). The dimension-general attention model captured the strength of spatial attention towards the object containing a given instructed feature. Because

feature locations were randomized across trials, we were still able to estimate dimension-specific effects of spatial attention separately for the cued dimension and or any of the non-cued dimensions.

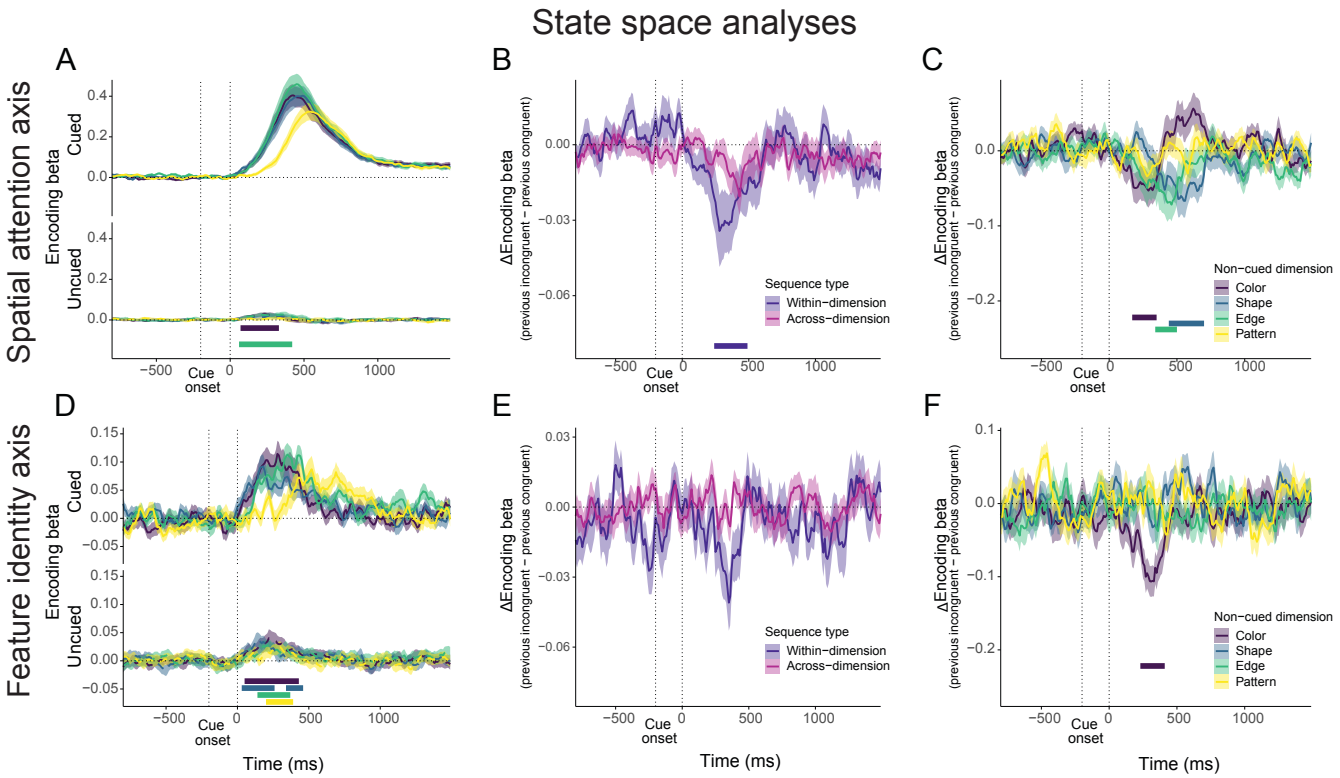

Figure 7

For cued dimensions (Fig 7A, top), the evolution of temporal attention resembled that obtained with decoding-RSA. In contrast, attention to non-cued dimensions (Fig. 7B, bottom) was markedly weaker and short-lived, consistent with rapid attentional capture followed by reactive suppression. Re-aligning trials to response onset further supported this interpretation: spatial attention toward the cued dimension became even stronger and peaked sharply at the time of response, whereas distractor-directed attention disappeared entirely (Supplementary Fig. 2). Thus, the Choice axis for distractors reflects brief orienting responses rather than sustained engagement.

When we examined trial-by-trial adaptations in this signal, we replicated a reliable within-dimension effect, with no appreciable effect across-dimension (Fig. 7BC). Moreover, we saw significant within-dimension adaptation effects for color, shape, and edge (the same dimensions that exhibited behavioral adaptation; Fig. 6G; c.f. Fig. 2C-F), whereas pattern again showed now measurable modulation. The timing of these effects mirrored each dimension's interference dynamics, with pattern having the weakest potency (Fig. 1IJ). In contrast, the timing of adaptation effects for the remaining dimensions aligned with the temporal dynamics of their interference (Fig. 3E).

Finally, the dimension-specific Feature model provided a direct estimate of the visual feature processing for each dimension, regardless of whether it served as the task-relevant target (cued dimension) or as a distractor (non-cued dimension). The analyses using this model confirmed that distractor features are encoded in parallel with target features, consistent with early perceptual processing preceding control (Fig. 7D). However, when we examined within-dimension adaptation, only the color dimension showed reliable modulation (Fig. 7EF). This suggests that most of the conflict adaptation captured by the decoding-RSA approach reflects adjustments along the dimension-general spatial-attention axis. Alternatively, because color interference occurs early and with high potency, its rapid dynamics may have made it uniquely sensitive to feature-level adaptation.

Proactive modulation of distractor suppression explains task-specific conflict learning

So far, our analyses have focused on short-term control adjustments in response to recent conflict. In these analyses, all trials came from *neutral* blocks, where each non-cued dimension was equally likely to appear with congruent or incongruent distractors. We found that attention control quickly adjusts to unpredictable changes in recent conflict. However, cognitive control can also adapt to more stable environmental regularities. For example, when driving, people can learn to use contextual cues (such as heavy traffic) to effectively heighten their attention toward other vehicles and reduce their attention toward irrelevant distractions. To examine this long-term form of control learning, we used a task-specific proportion congruency (TSPC) manipulation (Bugg & Crump, 2012) across blocks. Each task dimension was assigned a likelihood of individual distractors being congruent. At each point in the experiment, one task dimension was mostly incongruent (MI; 25% probability that any distractor was congruent), one was mostly congruent (MC; 75%), and the other two were neutral (50%). These assignments occasionally rotated across blocks. This design established stable, learnable relationships between each task and the expected level of interference, allowing participants to anticipate control demands with sequences of task repetitions, and to adjust attention proactively.

Reaction time and accuracy data revealed that participants adapted control to the TSPC manipulation (Fig. 2GH). Generalized Bayesian linear regression models provided decisive evidence for an interaction between current congruency and TSPC on both RT and accuracy (ERs = ∞). Tasks with MI distractors showed attenuated congruency effects (shallower slopes), reflecting reduced susceptibility to non-cued dimensions. Tasks with MC distractors showed enhanced congruency effects (steeper slopes), reflecting stronger distractor influence. Indeed, compared with MI tasks, trials in MC tasks yielded slower and less accurate response when current congruency was low ($MC_0 – MI_0$ = 41 ms, -4.5%) and faster more accurate response when current congruency was high ($MC_3 – MI_3$ = -26 ms, 0.6%).

The strength of this long-term adaptation mirrored the heterogeneous potency of interference across dimensions (Fig. 2IJ). We observed reliable TSPC adaptation effects on RTs for color (BF = 3.3×10^3), shape (BF = 2.3×10^5), and edge (BF = 57), but not for pattern (BF = 0.1). Accuracy showed a similar pattern, with evidence for adaptation for color (BF = 5.2) and shape (BF = 469), but not edge (BF = 1.2) or pattern (BF = 0.1).

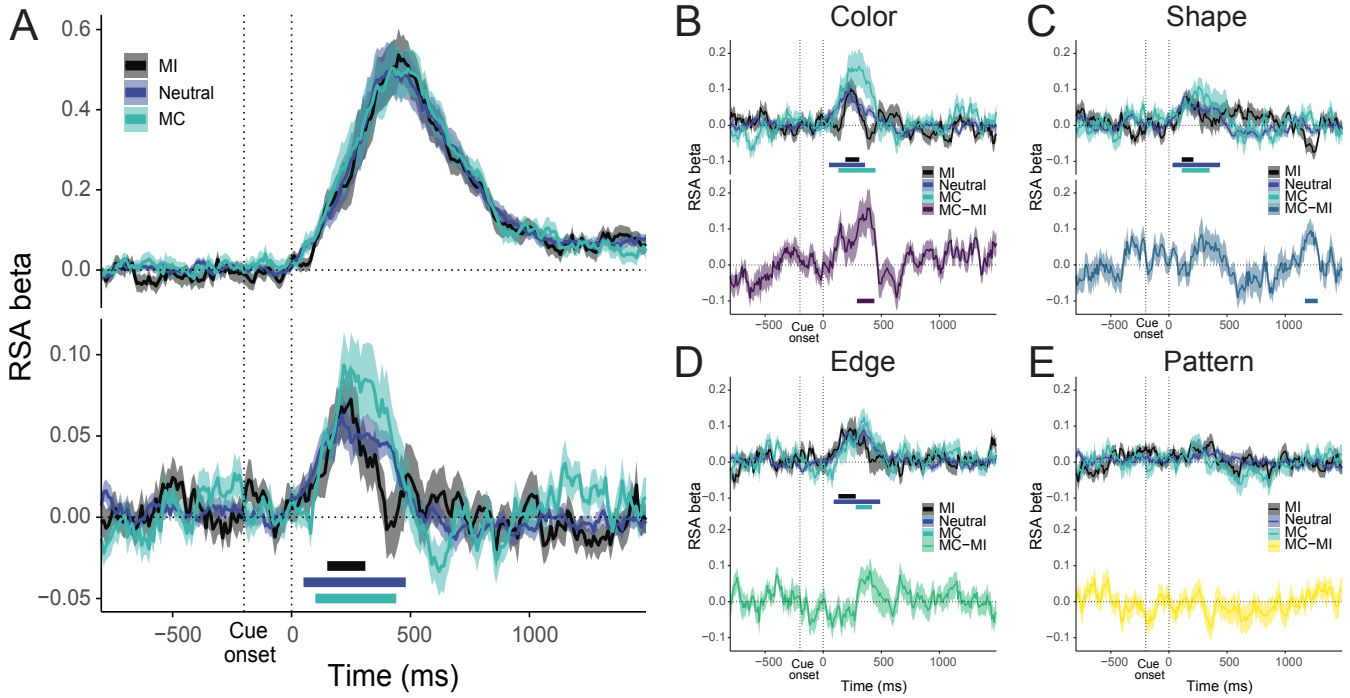

Figure 8

To test whether this form of long-term control adaptation also shaped neural distractor representations, we again used our decoding-RSA model to estimate the strength of cued and distractor features as a function of TSPC (Fig. 8). As described above, this approach provides a combined estimation of the location and identity of each

dimension's instructed feature. However, because the choice component is shared across dimensions, and MI and MC blocks contain different base rates of congruent and incongruent trials, simply averaging across trials would bias estimates. To correct for this, we used a bootstrapping procedure to rebalance the proportion of congruent and incongruent distractors for task dimension and TSPC block. This allowed ue to isolate the genuine effects of long-term control adaptation on feature representations, independent of trial-level congruency bias.

The temporal neural dynamics of control adaptation to TSP closely mirrored those observed for the short-term CSE effects. The TSPC manipulation did not affect representations of the cued dimension (Fig. 8A, top), indicating that proactive adjustments did not enhance target coding. Instead, TSPC modulation emerged in the representations of non-cued features (Fig. 8A, bottom), with the changes in latency and strength being compatible with the reactive suppression of distractors observed for the CSE. This modulation only emerged in the pre-response period. Distractor representations were strongest and most sustained for MC tasks, followed by neutral and MI tasks. Across dimensions, the MC–MI contrast on distractor representations was largest for color, with trending effects for shape and edge, and no adaptation for pattern, paralleling the heterogeneity in behavioral adaptation (Fig. 8B-E).

In sum, these results demonstrate that people can learn stable associations between control settings and task dimensions, allowing them to flexibly balance the benefits and costs of processing distractors across contexts. Once these task-control associations are learned, participants can anticipate upcoming control demands and proactively adjust attention during short task sequences. In MC tasks, for example, relaxing attention to distractors is beneficial because they often support the correct response. In MI tasks, however, heightened control minimizes their disruptive influence. However, this proactive regulation of control seems to be implemented in the form reactive control, making it faster and stronger through learning.

Discussion

The ability for attentional control allows humans to reconfigure information processing so that behavior is guided by current goals rather than irrelevant information. Yet most work in this domain probes this ability using tasks with a single distractor dimension, leaving it unclear how control operates when demands are distributed across multiple dimensions of information. Here, we addressed this question in a setting in which multiple stimulus dimensions independently vary in their relevance and potential to cause interference. Across neural, behavioral, and computational analyses, we found that attentional control adapts to local and global demands at the level of individual distractor dimensions, selectively tuning how the brain represents visual features in multidimensional environments.

Participants performed the MULTI (Gheza & Kool, 2025), a task that requires cognitive control to flexibly orchestrate attention across four visual dimensions to implement context-dependent feature selection. When acting as distractors, each of these dimensions contributed to overall interference. Coupled with a novel analytical approach that combines multivariate decoding of EEG with RSA (Kikumoto et al., 2022, 2024; Kikumoto & Mayr, 2020), this allowed us to trace how neural representations of cued and non-cued dimensions evolve over time. On a fast, trial-to-trial scale, we found that control was adjusted based on the congruency of the previous trial (CSE; Gratton et al., 1992). Despite the constantly changing arrangement of multidimensional distractors, this adjustment was dimension-specific: control selectively modulated the neural representation of a distractor based on whether it had previously caused interference or facilitation, leaving target representations unaffected.

These results have implications for theoretical models of how the brain learns to manage control in structured environments. Classical models of conflict monitoring (Botvinick et al., 2001; Egner, 2008) assume that the brain shapes attention policies using a global control signal calculated from either response- or stimulus-level conflict. Our findings show that the brain does not implement such a global adjustment. Instead, control is adjusted with specificity, learning a control setting that is tailored to individual sources of conflict. This selective pattern of adaptation resembles the behavior of models that learn structured control policies from experience, leveraging the richness of the environment beyond what is strictly necessary to perform the task (Alexander, 2022). Relatedly, reinforcement-learning models of control allocation (Lieder et al., 2018) propose that the system tracks mappings between contexts and control states. Such structured policy learning allows the brain to efficiently latch onto the multidimensional statistical structure of the environment, avoiding the need for costly metacognitive deliberation (Boureau et al., 2015).

A key insight from our results is that attentional control operates on emerging stimulus representations, rather than preventing distractors from being encoded in the first place. Both our decoding-RSA and encoding analyses revealed early parallel processing of target and distractor features, consistent with feature-based models of attention (Gaspelin & Luck, 2018; Liesefeld et al., 2017, 2019). Control emerges shortly thereafter. We found that distractor representations start being suppressed around 250 ms after stimulus onset, and that the strength of this suppression scales with prior interference. By isolating feature-level encoding from a dimension-general axis (Mante et al., 2013), we found that this adaptation primarily involves rapid adjustments to spatial attention, with feature-level adaptation occurring mainly for high-potency dimensions. Indeed, the timing differences emerged selectively in the representation of feature identity and spatial attention, but not task cue. These differences mirrored the heterogeneity in the behavioral interference effects, especially for the delayed pattern dimension. Thus, multidimensional control is achieved through time-sensitive modulation of evolving representations, which is a mechanism that naturally accommodates parallel feature integration and dimension-specific adaptation.

These temporal dynamics inform classical theories of attentional control. In influential biased-competition frameworks (Desimone & Duncan, 1995; Reynolds & Chelazzi, 2004), top-down biases competition between representations toward task-relevant stimuli early in the visual hierarchy. Our work suggests that, even with such early biasing, distractor features are still encoded in parallel that control exerts its strongest effects after perceptual representations have developed. Moreover, it demonstrates that such competition and subsequent modulation can happen between distinct feature dimensions that are spatially separated. Indeed, our work points to new opportunities for electrophysiological work. The MULTI offers a way to study how feature-specific representations of integrated stimuli (Treisman & Gelade, 1980) interact and are modulated by top-down control (Buschman et al., 2014; Miller & Cohen, 2001). One strong prediction from our framework is that neurons tuned to high-potency dimensions would show both stronger representational strength as well as receive larger top-down modulation when their dimension acts as a distractor.

A similar electrophysiological approach may be used to clarify the role of ACC during conflict monitoring and resolution. While extant theories suggest that conflict is unitary, our results suggest that conflict is encoded in a multivariate fashion, allowing the system to regulate attention control with dimension-specific granularity. It is unclear whether the ACC is, by itself, able to encode this signal, or whether it coordinates with posterior regions to estimate the control demands of currently relevant sources of information. After all, the world is replete with a large and heterogeneous set of sources of distraction, and so it is implausible that ACC contains dedicated conflict monitors for each of these. Instead, we suggest that it leverages its inherent mixed selectivity of its neural population to flexibly set up parallel low-dimensional subspaces (Fu et al., 2022), known as manifolds (Flesch et al., 2022), to monitor conflict from multiple sensory regions. From an electrophysiology perspective, this would predict that individual neurons in ACC reliably encode conflict from specific dimensions. Moreover, the proposed ad-hoc configuration of monitors suggests that the involvement of these neurons for the encoding of one type of conflict may be short-lived, context-dependent, dynamically reassigned as task demands shift, and reflect the potency of its current monitoring target.

We found evidence that neural measures of the potency of distractor interference depend on the processing speed of each dimension. Most notably, the pattern dimension was both the weakest distractor, in terms of behavioral interference and control adaptation, and emerged the latest in our neural measures of target feature identity. Our connectionist model accounts for this effect by including a dimension-specific temporal integration parameter, which reproduces within-trial dynamics and variance in dimension-specific adaptation. Future work should adopt recent advances in the model-fitting techniques of RT and choice data (Fengler et al., 2021; Radev et al., 2022) to directly compare this leaky accumulator network to alternative model architectures, such as the diffusion model for conflict (Ulrich et al., 2015). Our model suggests that the evolution of distractor-specific conflict is inherently non-linear, with a central “sweet spot” of maximal returns where integrating additional evidence most strongly influences response selection. This notion accords with previous demonstrations of temporally weighted evidence accumulation in perceptual decision-making (Bronfman et al., 2016; Koob et al., 2023; Mittelstädt et al., 2023; Usher & McClelland, 2001).

Finally, our results suggest that attentional control sets can be tied to abstract task representations, much like they can be anchored to temporal contexts or specific task features (Bugg & Crump, 2012; Dey & Bugg, 2021; Ileri-Tayar et al., 2024). These task-control associations allowed participants to anticipate task demands and proactively adjust

attention based on the task cue. At the neural level, this form of control was expressed in the selective modulation of distractor representations, paralleling short-term conflict adaptation. These results suggest that proactive control can be learned via generalization of repeated exposure to trial-level conflict, and that it relies on mechanisms similar to those supporting reactive, dimension-specific adaptation.

Notably, proactive control, manipulated through our TSPC manipulation, was expressed as greater efficiency in the reactive suppression of distractors of distractors, rather than sustained preparatory changes prior to object onset. We did not observe proactive modulation of task representations that preceded stimulus onset. At face value, these results suggest that proactive control does not involve the tonic maintenance of task-set representations. Specifically, the lack of evidence for sustained changes in scalp-level EEG features does not support trial-long functional changes in the cingulo-opercular executive network (Dosenbach et al., 2007). However, it remains possible that proactive adjustments occur at meso- or micro-levels within smaller neural populations that are difficult to detect with EEG. Finally, it is worth noting that while our TSPC manipulation encouraged task-level proactive control rather than proactive modulation of individual *distractor* dimensions, neural adaptation occurred at the distractor level. Future work should test whether selectively varying the congruency of individual distractor dimensions can elicit more sustained or anticipatory neural modulation by making individual distractors more reliable targets for suppression.

A potential limitation of our EEG and RSA approach is that it depends on the LDA classifier's ability to discriminate classes which are defined only during the cue and stimulus presentation. Thus, the lack of effects outside this window may reflect limited decodability rather than an absence of adaptation per se. More broadly, proactive control does not need to manifest as sustained activity. Instead, it may rely on short, transient bursts of regulation rather than tonic changes across trials (Verguts, 2017). Consistent with this view, a large body of literature shows that theta oscillations reflect control related signaling (Cavanagh & Frank, 2014; M. X. Cohen, 2017; M. X. Cohen & Cavanagh, 2011; Cooper et al., 2019; De Loof et al., 2019; Gheza et al., 2018, 2019; Senoussi et al., 2022) and that it mediates coordination between ACC and PFC (Smith et al., 2019). In our data, theta peaked around the response, reflecting both current and prior conflict, providing evidence that existing traces of control demands can transiently emerge to bias information processing.

Our findings do not rule out other ways in which adaptive control may operate over multidimensional, feature-based attention. As discussed above, one such mechanism is the modulation of strength of stimulus representations: distractors retain their coding geometry, but their activation is selectively attenuated following interference. An alternative is that adaptation could alter the geometry of distractor representations, for example, by rotating their axes away from the task-relevant dimension to reduce future conflict (Ebitz et al., 2020). Though we can't fully disentangle these forms of control adaptation, the reactive suppression in our data is more supportive of the strength-modulation hypothesis.

In light of evidence that PFC can simultaneously represent multiple abstract task features, as observed in context-dependent decisions in non-human primates (Bernardi et al., 2020; Mante et al., 2013; Siegel et al., 2015), one might ask why humans suffer distractor interference at all. However, a comparison between studies on humans and non-human primates (Mante et al., 2013; Siegel et al., 2015) highlights a key difference. Monkeys are trained for weeks or months on the same stimulus-response mappings, showing little interference from non-cued dimensions. This suggests that highly orthogonalized feature representations may be a consequence of extensive training, and not achievable in the short term. Humans, on the other hand, exhibit strong, history-dependent distractor interference, but are rarely exposed to the extended, multi-day training (MacLeod & Dunbar, 1988). Across species, it may be that adaptation initially relies on modulation of coding strength, whereas geometry-based reconfiguration emerges only after long-term consolidation of sensory representations. Future work may implement metrics such as the parallelism score (Ritz & Shenhav, 2024) to evaluate if orthogonality of feature representations emerges in humans, and whether adaptation may then happen via rotations in feature and/or response coding (Mante et al., 2013).

Methods

Participants

We recruited 27 participants (22 females, 5 males, 1 nonbinary; mean age: 27 years, range: 20-39) with no history of neurological disorders and normal or corrected-to-normal vision. Recruitment was restricted to right-handed participants without color blindness. Participants were recruited from the St. Louis community via the Washington University School of Medicine Research Participant Registry and from the Washington University in St. Louis student participant pool. All were compensated with $25. The experimental procedure was approved by the Institutional Review Board of Washington University in St Louis (IRB ID #: 202110190). Written informed consent was obtained before the experiment, and participants were debriefed afterwards. Sex or gender was not analyzed, as it was not pertinent to our hypotheses. We aimed for a sample size of about 25 participants based on the size of behavioral effects observed in our previous work (Gheza & Kool, 2025). However, no prior effect size information was available for the new EEG effects; therefore, no statistical method was employed to define the sample size. One participant was excluded from the analyses due to partial data recording.

Behavioral procedure

To study attentional control over multiple independent stimulus dimensions, we used the MULtidimensional Task-set Interference paradigm (MULTI; Figure 1; Gheza & Kool, 2025), implemented in Psychtoolbox (Kleiner et al., 2007). On each trial of the MULTI, participants attend to a letter cue and choose between two objects, presented side by side, that differ in four dimensions: color, shape, edge, and pattern. For each of these dimensions, there are two features (blue/orange for color, oval/rectangle for shape, solid/dashed for edge, and horizontal/vertical for pattern), which are assigned randomly assigned to the left and right objects on each trial. Each object therefore, consists of one feature from each dimension, and these features are fully crossed across trials. All features are highly discriminable, and the objects are matched for size.

The letter cue indicates which dimension is task-relevant on a given trial. Participants' task was to identify which object contained the target feature for that dimension, and select its side by pressing a corresponding button. At the beginning of the experiment, participants were instructed on which feature was the target for each of the four task dimensions. For instance, one participant might be instructed to choose the object with blue color if the cue is C, the oval shape when it is S, the dashed border when it is B, or the vertical pattern when it is P. The specific target features were randomly assigned for each participant and remained constant throughout the task.

The cued task dimension switched in pseudo-randomized sequences, making each dimension periodically relevant. As in Gheza & Kool (2025), the same dimension was cued for 3 to 5 consecutive trials, before switching to another. All 4 dimensions were cued within 4 switch trials. Therefore, participants constantly cycled through the 4 tasks, organized in a series of 3 to 5 repeat trials. This ensured that all dimensions were cued an equal number of times, and that each of them had been cued relatively recently at each point in the task.

For suitability with EEG data collection, some changes were introduced to the task and trial structure compared to our prior work. Most importantly, the cue and objects display were temporally separated, preventing overlap. Each trial started with a brief presentation of the letter cue (50 ms), followed by a fixation dot (150 ms). Next, the two objects appeared, flanking the fixation dot, and remained on screen until response or until a response deadline of 2 s had passed. Participants indicated their choice by pressing the left or right button on a Cedrus RB-840 response pad (Cedrus Corporation, San Pedro, CA, USA). An incorrect or missing response was immediately followed by negative feedback, presented as a red "x" at the center of the screen for 1 s. The inter-trial interval randomly varied between 1.3 and 1.5 s with fixation cross displayed throughout.

Participants performed 32 blocks of 50 trials each (1600 trials in total), separated by self-paced breaks. Before the main task, participants performed five blocks of practice. These included four blocks in which participants learned to respond to each dimension individually, and a fifth block in which they practiced switching between task dimensions. These practice blocks ensured that participants had mastered the cue–feature mappings before EEG recording began.

In line with the previous design, task-set interference was parametrically manipulated on a discrete congruency scale from 0 to 3. On each trial, any of the three non-cued dimensions could be congruent if they primed the same response side as the cued target feature, or incongruent if they primed the opposite side. This manipulation enabled us to examine trial-by-trial adjustments in attentional control as function of recent multidimensional conflict.

To investigate sustained associations of control settings with task dimensions, we introduced a task-specific proportion congruency (TSPC) manipulation (Bugg & Crump, 2012). This manipulation established stable relationships between cued task dimensions and expected conflict levels. At each point in the experiment, one task was mostly incongruent (MI; i.e., each of the three distractors had a 25% probability of being congruent), one task was mostly congruent (MC; 75% probability), and the remaining two were neutral (50% probability). These assignments rotated three times across the session according to a Latin Square design, with each TSPC mapping lasting eight blocks of 50 trials. This rotation ensured that, across the experiment, each task dimension served once as MI and once as MC, providing a fully counterbalanced design. For each participant, the initial assignment of PC to task dimension was randomized.

Participants were instructed to respond as quickly and as accurately as possible. To minimize eye and head movements, they were requested to always maintain fixation at the center of the screen and to process the left and right objects using peripheral vision. The average duration of the experiment was 2.5 hours, including approximately 1 hour for EEG setup and instructions, 1 hour for task execution, and the remaining time allocated for debriefing and the cleaning of the EEG cap and electrodes.

Behavioral data analysis

Our behavioral analyses followed the procedure reported in our previous behavioral investigation using the MULTI (Gheza & Kool, 2025). Here, we provide a synthesis of that procedure. Switch trials were excluded from the analyses. For congruency sequence effect (CSE) analyses, we also excluded trials following incorrect responses to avoid contamination from post-error slowing (Danielmeier et al., 2011) or other control processes. For proportion congruency analyses, all trials were included. Incorrect trials were further excluded for all RT analyses.

Accuracy and reaction times were modelled using a Bayesian multilevel generalized linear approach implemented in R (R Core Team, 2017) with the 'brms' package (Bürkner, 2017). Models predicting reaction times assumed a shifted lognormal response distribution, with identity and log link functions for the distributional parameters. Models predicting accuracy assumed a Bernoulli distribution with a logit link function. We report medians and 95% credible intervals of the posterior parameter distributions on the log scale for reaction times and the logit scale for accuracy, and estimated marginal means are reported in the response scales (milliseconds for reaction times and percentages for accuracy). All models used with weakly informative priors centered on zero for the parameters of interest and on the raw mean for the intercept. For hypothesis testing, we report evidence ratios (ER).
To assess within- and across dimension adaptation, we calculated CSEs for 16 possible sequences of non-cued dimensions. Each sequence compared the change in interference for a given dimesion after congruent versus congruent trials on another dimension. Some sequences assess within-dimension adaptation (e.g., previous color congruency on current color sensitivity), while others assess across dimension adaptation (e.g., previous shape congruency on current edge sensitivity). For each combination, we computed the adaptation effect by subtracting the interference effect after incongruent trials from the interference after congruent trials.

For the TSPC analyses, we modeled RTs and accuracy as a function of current congruency, TSPC condition, and their interaction. This allowed us to test whether the magnitude of the congruency effect varied as s function of each task's current congruency bias (MI, MC, neutral).

Neural network modeling of conflict adaptation

To capture the temporal dynamics of dimension-specific control adaptation, we implemented a dynamic neural network model of the MULTI (Fig. 3A), inspired by classic architectures of control allocation and conflict monitoring in interference paradigms (Botvinick et al., 2001; J. D. Cohen et al., 1990) and building on the static model introduced by Gheza and Kool (2025). The model simulates how evidence from multiple dimensions accumulates over time and how recent conflict shapes pathway-specific control.

The network consists of four parallel pathways, corresponding to the four dimensions of the task. Each pathway begins with two input units encoding the side of the target feature for that dimension (left vs. right). Each set of input nodes then projects to two units in the hidden layer. These units integrate input from both the stimulus units and a set of task units representing top-down attention towards each dimension. The task units modulate the gain of their corresponding hidden units by offsetting a fixed negative bias, thereby enhancing sensitivity to stimulus input in the attended dimensions. The activation of each task unit depended on whether its corresponding dimension was cued and was further adjusted based on recent dimension-specific conflict (see below). This attention gating mechanism increased the separation between the activations of hidden units for attended dimensions, as the activity of the task units offset the inhibitory bias, allowing hidden units to encode stimulus input with more sensitivity. The units in the hidden layer project to two response nodes (left vs. right) via balanced excitatory and inhibitory connections. Specifically, hidden units representing target features on the left projected positively to the left response node and negatively to the right response node, and vice versa.

Net input into every hidden and response unit was computed as the sum of the activity of projecting units multiplied by their corresponding connection weight, with an added constant bias. This resulting net input was then passed through a logistic activation function, which converted it into an activity value between 0 and 1.

To simulate temporal dynamics, unit activations at the hidden and response layer evolved over time using leaky integration (Usher & McClelland, 2001). At each timestep, the activity of the hidden layer was a weighted combination of its previous activation and its current net input, controlled by a pathway-specific time constant (τ_d). For each timestep t, the net input H into each hidden neuron n in dimension pathway d accumulates a small dimension-specific trickle of evidence:

$$H_{d,n}(t) = (1 - \tau_d)H_{d,n}(t-1) + \tau_d I_{d,n}$$

with $H_{d,n}(0) = 0 \; \forall \; d, n$
where $I_{d,n}$ is the total net input to neuron n in dimension pathway d. This input is constant within a trial and is the sum of stimulus input S, task input T, and bias term b_H:

$$I_{d,n} = S_{d,n} + T_d + \; b_{H,d}$$

This produces dimension-specific temporal integration without lateral interactions at the hidden layer.
To simulate the choice process, the two response units likewise integrated their inputs using a leaky integrator:

$$R_L(t) = (1 - \tau_r)R_L(t-1) + \tau_r\left(\sigma(\boldsymbol{H})^T \boldsymbol{w}_{H,L}\right)$$
$$R_R(t) = (1 - \tau_r)R_R(t-1) + \tau_r\left(\sigma(\boldsymbol{H})^T \boldsymbol{w}_{H,R}\right)$$

with R_L and R_R the activity of the left and right response units, τ_r a response-level time constant, $\boldsymbol{H}$ the activity of the hidden layer at timestep t, $\boldsymbol{w}_{H,L}$ and $\boldsymbol{w}_{H,R}$ the connections between the hidden layer and the right and left response unit, and σ the logistic activation function.

The model's choice mechanism consisted of two opposing evidence accumulators. At each timestep, the instantaneous evidence was proportional to the difference between the two response activations $\Delta(t)$, which drives a pair of opposing noisy accumulators:

$$A_L(t) = A_L(t) + \alpha\Delta(\mathrm{t}) + \; \epsilon_t$$
$$A_R(t) = A_R(t) - \alpha\Delta(\mathrm{t}) + \; \epsilon_t$$

with Gaussian noise $\epsilon_t \sim N(0, \sigma)$. The trial terminated when either accumulator reached a fixed decision threshold θ. The first unit to cross this threshold determined the model's choice, and the corresponding timestep defined the reaction time for that trial. Accuracy was defined by the match between the selected response and the cued target feature's location.

To simulate conflict monitoring, we followed and extended our prior modeling work. As before, we computed distractor-specific conflict E_d as Hopfield energy (Hopfield, 1982), but now for each time point t:

$$E(t) = -\sum_{i,j:i<j} w_{ij} a_i(t) a_j(t)$$

where w_{ij} represents the connection weight between units i and j, a indicates unit activity and both subscripts are indexed over all units in the pathway. Here, we compute this measure of energy between the hidden and response layer separately for each non-cued dimension.

Because of the excitatory and inhibitory configuration of the connections between hidden and response units, the sign and magnitude of the dimension-specific conflict are only a function of the pathway state. If dimension-specific conflict is positive, a non-cued dimension provides evidence inconsistent with the correct response, and negative if it provides consistent evidence

Next, for each dimension, conflict is transformed into a measure of control using the following formula:

$$\mathrm{control}_d(i+1) = \lambda_{control} \cdot \mathrm{control}_d(i+1) + (1 - \lambda_{control})(\alpha \cdot \mathrm{AUC}(E_d))$$

where

$$\mathrm{AUC}(E_d) = \sum_{t=1}^{T} E_d(t)\, \Delta(\mathrm{t})$$

indicates discrete integration of pathway-specific energy across the trial, and with i indexing trials, $\lambda_{control}$ allowing a time-discounted measure of previous conflict, and α a scaling parameter. At the start of each trial, each non-cued dimension's control value was then subtracted from the (zero) input of that dimension. This algorithm implements distractor-specific adaptation, with the model becoming more sensitive to a non-cued dimension after a congruent trial on that dimension, and less sensitive after an incongruent trial.

Model parameters (including connection weights) were hand-tuned within plausible ranges to reproduce the key qualitative behavioral patterns (interference, dimension-specific adaptation, etc.). Similar qualitative results were obtained across a broad range of parameter values, confirming that the reported dynamics do not depend on specific fine-tuned settings.

EEG recordings and preprocessing

EEG data were recorded using a 64-channel Brain Vision actiCHamp Plus system (Brain Vision LLC), placed on a lycra cap (actiCap) following the 10/20 system, and sampled at 500 Hz. Data were collected from 61 scalp sites (Fpz, Fp1, Fz, F3, F7, FC1, FC5, Cz, C3, CP1, CP5, Pz, P3, P9, PO3, PO7, Oz, PO4, PO8, P4, P10, CP2, CP6, C4, FC2, FC6, F4, F8, Fp2, AFz, AF3, AF7, F1, F5, FCz, FC3, FT7, C1, C5, T7, CP3, P1, P5, P7, O1, POz, O2, P2, P6, P8, CPz, CP4, C2, C6, T8, FC4, FT8, F2, F6, AF4, AF8). The Cz electrode served as the online reference site. Three electrodes were placed lateral to the external canthi and below the right eye for measuring horizontal and vertical eye movements (electro-oculogram, EOG), and the ground electrode was placed at channel location AFp1.

The preprocessing pipeline was developed with EEGLab v2021.1 (Delorme & Makeig, 2004). It contained the following steps. The continuous signal was filtered offline with separate high-pass (0.1 Hz lower edge frequency) and low-pass (80 Hz higher edge frequency) Hamming windowed sinc FIR filters. Independent Component Analysis (ICA) was used to identify eye artifacts generated by blinks and saccades, muscle activity, and channel noise. To maximize component separation (Dimigen, 2020), a copy of the data was high-pass filtered at 2.0 Hz and segmented into epochs time-locked to objects onset (-500 to 1500ms) for ICA training; the resulting ICA weights were then transferred to the original dataset. Artefactual components were manually identified by inspecting

activation time courses, topography, power spectrum, and with the aid of an automated classifier (Pion-Tonachini et al., 2019).

Next, stimulus-locked epochs were extracted in a larger interval (-950 to 2050 ms) suited for time/frequency analysis, and the non-artefactual independent components were back-projected to electrode space. Baseline activity (mean voltage during the -500 to -300 ms window) was subtracted from each epoch. An automatic artifact rejection procedure was applied to discard trials with i) voltage exceeding ± 120 μV in the -500 to 1500 ms interval or ii) improbable data based on joint probability of electrode activities.

Clean data were re-referenced to the average channel activity, and the signal at Cz was reconstructed. Trials with incorrect or missing responses were excluded from further EEG analyses. For congruency sequence effects, we further excluded switch trials and trials preceded by incorrect responses.

Time/frequency analysis

Time/frequency analysis of the single-trial EEG data was performed via Morlet wavelet convolution (M. X. Cohen, 2014). We convolved the EEG signal with a set of complex Morlet wavelets, defined as complex sine waves tapered by a Gaussian window ($e^{2i\pi ft}e^{-t^2/2\sigma^2}$), where *i* is the imaginary operator, *t* is time in seconds, *f* is frequency in Hz, and σ is the width of the Gaussian window. The width of the Gaussian was defined as $\sigma = \frac{n}{2\pi f}$, where *n* is the number of sine cycles. Because *n* governs the trade-off between temporal and spectral precision, it was set to increase logarithmically from 2 to 10 with increasing wavelet frequency. The frequencies of the wavelets ranged from 1.5 to 40 Hz in 20 logarithmically spaced steps. It follows that the full-width at half-maximum (FWHM) ranged from 500 to 94 ms with increasing wavelet peak frequency (M. X. Cohen, 2019).

The convolution was performed in the frequency domain and converted to the time domain using an inverse Fourier transform. Instantaneous power estimates were obtained by taking the squared magnitude of the resulting complex-valued signal, and decibel-converted by normalizing with a channel- and frequency-specific baseline. This baseline was defined as the average power across all trials and across 100ms of the inter-trial interval (-800 to -700 ms pre-stimulus, i.e., -600 to -500 ms before the cue onset). The power time series were down-sampled to 100 Hz to reduce the computational load for subsequent analyses.

Decoding-RSA

To link EEG patterns to temporally evolving task representations, we used a representational decoding framework that captures how neural activity encodes information about attended and unattended stimulus dimensions. This approach allowed us to quantify the emergence and temporal dynamics of both target and distractor representations within each trial.

For each trial and time-point, we decoded target- and distractor-relevant information using a two-step procedure consisting of 1) time-resolved multivariate pattern classification ('EEG decoding'), and 2) representation similarity analysis based on the resulting classification profiles ('RSA'). This method ('decoding-RSA') was adapted from Kikumoto and Mayr (2020) for use with the MULTI, incorporating several modifications to accommodate its multidimensional structure and frequency-resolved EEG data.

Decoding

As a first step, we trained linear decoders to classify each trial based on the eight possible conjunctions of task cue (e.g., the letter 'C' for color) and the side of the corresponding instructed feature (e.g., left). These conjunctions define the target feature that participants search for on each trial (e.g., the color blue). The classifier thus aimed to decode which of the possible eight cue × side conjunctions was present in each trial from EEG activity at each time point.

Specifically, we trained independent classifiers for every time point using a Penalized Linear Discriminant Analysis usig MATLAB's 'fitcdiscr' function (*Version: 9.12.0, R2022a)*. The classification feature space was obtained by

concatenating the instantaneous EEG power across all frequency bands and the 61 electrodes. To increase the signal-to-noise ratio, the original 20 frequencies were averaged within the bands Delta (1.5-3.5), Theta (3.6-7.5), Alpha (8-12.5), Beta (13-30), and Gamma (31-40), leading to 305 channel-band features. This approach generates eight-element vectors of probabilities reflecting the likelihood of a given trial to belong to each of the classes at each time point within the epoch.

We adopted a multivariate classification approach, training classifiers to separate the eight class simultaneously, rather than multiple pairwise classifiers. This strategy was essential because it provides probabilistic outputs for all classes. We used these class probability vectors in the subsequent RSA analyses.

We evaluated classifier performance with a repeated *k*-fold cross-validation procedure. For each participant, single-trial EEG data were randomly partitioned into six "folds" with an equal number of observations for every combination of the eight classes and four PC conditions. To ensure balanced training sets, trials exceeding the minimum number of observations across these thirty-two combinations were randomly discarded. At each step of the cross-validation, the classifier was trained on five folds and tested on the remaining fold and discarded trials. This was repeated until all folds served as the test set. This cycle was repeated eight times. Critically, the trial order was shuffled before each cycle, so that each set of folds included a different, random split of the data. This repetition increased the reliability of decoding estimates and maximized trial usage despite the heavy subsampling inherent in the k-fold balancing process (resulting in approximately 50% of trials used in the training per iteration). At each step, the EEG features in train and test trials were submitted to multivariate noise normalization (whitening) by means of a covariance matrix estimated on the full epoch (Guggenmos et al., 2018). During training, the LDA hyperparameter Gamma (shrinkage) was optimized once per time sample within the epoch and then fixed for all following k-fold steps.
For each time sample, a trained classifier was tested on each single trial in the test set. The classification probabilities estimated for each trial were then averaged across all folds and cycles, yielding an eight-element vector for each trial and time sample. Because these classification probability vectors are compositional (i.e., the class probabilities sum to one), a centered log-ratio transformation (Aitchison, 1986) was applied before the subsequent RSA regression analyses.

RSA

In the second step, we used RSA to quantify how the probability profiles produced by the decoders reflect different task representations. For each trial and time point, the classification probability vector was simultaneously regressed onto multiple model vectors, each representing a hypothesized component of task structure. The resulting regression coefficients indexed the strength of neural representations for each component at that moment in time.

We first defined an RSA model (m1) to estimate representations of cued dimensions (Figure 5). This model included three regressors corresponding to 1) the task "cue", 2) the "side" of the instructed feature, and 3) the "cue x side" conjunction (the dot product of cue and side vectors), representing the conjunction that defined the target. These model vectors allowed the estimation of the neural coding of the active task rule, the location of spatial attention, and the identity of the task-relevant feature. Note that, because we only analyzed correct trials, the "side" vector could equally represent spatial attention and response choice. Figure 2A illustrates the RSA procedure for m1 for a given trial, depicting how the model vectors included as RSA regressors depended on the trial type (x-axis).

The trial type is defined by the location of the instructed feature for the cued dimension (i.e., the target feature).
To ensure comparable scaling across regressors, the three model vectors were normalized using their Frobenius norms, and the cue and side matrices were rescaled relative to the conjunction matrix.

A second RSA model (m2) was developed to estimate dimension-specific feature representations of both cued and non-cued dimensions simultaneously. To achieve this, the RSA model included four "Cue x Side" vectors (one per dimension), each representing the location and identity of the instructed feature for a given dimension. For the cued dimension, the encoded side corresponded to the correct response, whereas for each non-cued dimension it depended on the congruency of that dimension. Because congruent and incongruent trials were unevenly distributed across task conditions, subsequent analyses used bootstrapping procedures to ensure unbiased estimates of distractor dimensions (see Bootstrapping procedure below).

Encoding models

The two-step decoding-RSA method is inherently limited in the number of predictors that can be simultaneously included as regressors. Specifically, the low degrees of freedom in the eight-element LDA classification profile required the careful selection of a small number of RSA regressors. For example, in RSA m2, including four Cue x Side vectors allowed us to measure simultaneous representations of cued and non-cued dimensions, but it prevented us from simultaneously accounting for variance solely explained by spatial attention. (This was not an issue for Cue coding, since Cue vectors are orthogonal to Feature and Choice vectors).

To disentangle the spatial attentional capture of distractors (Choice coding) from the representations of their visual properties (Feature coding), we developed a complementary encoding approach. This method was inspired by the state space analyses used in animal electrophysiology (e.g., Mante et al., 2013), but applied to our EEG data (channel × frequency) as the state space.

This encoding approach can be described as a multivariate linear regression. We first defined a matrix of four predictor vectors, each with a length equal to the number of trials. Each vector encoded the location of the target feature for a given cued dimension (the model vectors in Fig. 4A). For each time sample within a trial, we projected this predictor matrix onto the data matrix (305 EEG features × trials), yielding a beta-weights matrix (305 features × 4 dimensions) that quantified how trial-by-trial changes in EEG scalp patterns were modulated by each dimension's feature representations.

To separate shared spatial-attention effects from dimension-specific feature representations, we computed a dimension-general attention axis by averaging across the four dimension-specific beta vectors. This axis captures the shared, dimension-general variance, presumably dominated by lateralized effects of spatial attention. We then residualized this dimension-general axis from each dimension-specific beta vector, yielding five orthogonal encoding axes: four representing dimension-specific visual features and one representing dimension-general attention (i.e., the choice axis).

Projecting the EEG data onto these axes yields de-mixed estimates of task representations, for each trial and time-point within a trial. All encoding models were tested on held-out data following the same k-fold cross-validation procedure implemented for LDA classification.

Bootstrapping procedures

We implemented bootstrapping resampling procedures to prevent biased estimation of choice representations, which could arise because this coding axis was shared across dimensions. Separated procedures were used for the CSE and TSPC analyses.

For the CSE analyses (using either RSA m2 or encoding models), we restricted all analyses to the neutral TSPC blocks to avoid confounding sequential adaptation with task-specific adaptation. In addition, we downsampled the single-trial results to match the number of congruent and incongruent trials for each distractor dimension in both the current and preceding trials. This ensured balanced contributions to trial-by-trial adaptation estimates.

For TSPC analyses, block-wise differences in congruence proportions (MI vs. MC) would distort model estimates if trials were simply averaged. To counter this, we applied a bootstrap procedure to the single-trial RSA results, restoring unbiased proportions of congruent and incongruent distractors within each task dimension and PC condition. Specifically, when computing condition averages, we oversampled trials from the less frequent congruency category until both categories were equally represented.

Significance testing on EEG results

Statistical significance over RSA and encoding timeseries was assessed using cluster-based permutation tests (Maris & Oostenveld, 2007), implemented in MATLAB. We performed 10^4 permutations for each contrast, with

cluster $\alpha = 0.05$. For time-frequency analyses of FMT power we identified significant clusters ($\alpha = 0.01$) using the *EEGLAB* package (Delorme & Makeig, 2004), after averaging over channels of interest.

References

Aitchison, J. (1986). *The statistical analysis of compositional data*. Chapman and Hall.

Alexander, W. H. (2022). *Cognitive control strategies derive from dimension reliability*. PsyArXiv. https://doi.org/10.31234/osf.io/a9bmh

Badre, D., Bhandari, A., Keglovits, H., & Kikumoto, A. (2021). The dimensionality of neural representations for control. *Current Opinion in Behavioral Sciences*, *38*, 20–28. https://doi.org/10.1016/j.cobeha.2020.07.002

Bernardi, S., Benna, M. K., Rigotti, M., Munuera, J., Fusi, S., & Salzman, C. D. (2020). The Geometry of Abstraction in the Hippocampus and Prefrontal Cortex. *Cell*, *183*(4), 954-967.e21. https://doi.org/10.1016/j.cell.2020.09.031

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, *108*(3), 624–652. https://doi.org/10.1037/0033-295X.108.3.624

Boureau, Y. L., Sokol-Hessner, P., & Daw, N. D. (2015). Deciding How To Decide: Self-Control and Meta-Decision Making. *Trends in Cognitive Sciences*, *19*(11), 700–710. https://doi.org/10.1016/j.tics.2015.08.013

Braver, T. S. (2012). The variable nature of cognitive control: A dual mechanisms framework. *Trends in Cognitive Sciences*, *16*(2), 106–113. https://doi.org/10.1016/j.tics.2011.12.010

Bronfman, Z. Z., Brezis, N., & Usher, M. (2016). Non-monotonic Temporal-Weighting Indicates a Dynamically Modulated Evidence-Integration Mechanism. *PLOS Computational Biology*, *12*(2), e1004667. https://doi.org/10.1371/journal.pcbi.1004667

Bugg, J. M., & Crump, M. (2012). In Support of a Distinction between Voluntary and Stimulus-Driven Control: A Review of the Literature on Proportion Congruent Effects. *Frontiers in Psychology*, *3*. https://www.frontiersin.org/articles/10.3389/fpsyg.2012.00367

Bürkner, P.-C. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Buschman, T. J., Miller, E. K., & Miller, E. K. (2014). *Goal-direction and top-down control*.

Cavanagh, J. F., & Frank, M. J. (2014). Frontal theta as a mechanism for cognitive control. *Trends in Cognitive Sciences*, *18*(8), 414–421. https://doi.org/10.1016/j.tics.2014.04.012

Cohen, J. D., McClelland, J. L., & Dunbar, K. (1990). On the Control of Automatic Processes: A Parallel Distributed Processing Account of the Stroop Effect. *Psychological Review*, *97*((3)), 332–361.

Cohen, M. X. (2014). *Analyzing neural time series data: Theory and practice*. MIT Press.

Cohen, M. X. (2017). *Neurophysiological Oscillations and Action Monitoring*.

Cohen, M. X. (2019). A better way to define and describe Morlet wavelets for time-frequency analysis. *NeuroImage*, *199*, 81–86. https://doi.org/10.1016/j.neuroimage.2019.05.048

Cohen, M. X., & Cavanagh, J. F. (2011). Single-trial regression elucidates the role of prefrontal theta oscillations in response conflict. *Frontiers in Psychology*, *2*(FEB), 1–12. https://doi.org/10.3389/fpsyg.2011.00030

Cooper, P. S., Karayanidis, F., McKewen, M., McLellan-Hall, S., Wong, A. S. W., Skippen, P., & Cavanagh, J. F. (2019). Frontal theta predicts specific cognitive control-induced behavioural changes beyond general

reaction time slowing. *NeuroImage*, *189*(January), 130–140. https://doi.org/10.1016/j.neuroimage.2019.01.022

Danielmeier, C., Eichele, T., Forstmann, B. U., Tittgemeyer, M., & Ullsperger, M. (2011). Posterior Medial Frontal Cortex Activity Predicts Post-Error Adaptations in Task-Related Visual and Motor Areas. *The Journal of Neuroscience*, *31*(5), 1780–1789. https://doi.org/10.1523/JNEUROSCI.4299-10.2011

De Loof, E., Vassena, E., Janssens, C., De Taeye, L., Meurs, A., Van Roost, D., Boon, P., Raedt, R., & Verguts, T. (2019). Preparing for hard times: Scalp and intracranial physiological signatures of proactive cognitive control. *Psychophysiology*, *56*(10), 1–14. https://doi.org/10.1111/psyp.13417

Delorme, A., & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*(1), 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009

Desimone, R., & Duncan, J. (1995). Neural Mechanisms of Selective Visual Attention. *Annual Review of Neuroscience*, *18*(Volume 18, 1995), 193–222. https://doi.org/10.1146/annurev.ne.18.030195.001205

Dey, A., & Bugg, J. M. (2021). The Timescale of Control: A Meta-Control Property that Generalizes across Tasks but Varies between Types of Control. *Cognitive, Affective, & Behavioral Neuroscience*, *21*(3), 472–489. https://doi.org/10.3758/s13415-020-00853-x

Dimigen, O. (2020). Optimizing the ICA-based removal of ocular EEG artifacts from free viewing experiments. *NeuroImage*, *207*, 116117. https://doi.org/10.1016/j.neuroimage.2019.116117

Dosenbach, N. U. F., Fair, D. A., Miezin, F. M., Cohen, A. L., Wenger, K. K., Dosenbach, R. A. T., Fox, M. D., Snyder, A. Z., Vincent, J. L., Raichle, M. E., Schlaggar, B. L., & Petersen, S. E. (2007). Distinct brain networks for adaptive and stable task control in humans. *Proceedings of the National Academy of Sciences*, *104*(26), 11073–11078. https://doi.org/10.1073/pnas.0704320104

Ebitz, R. B., Smith, E., Horga, G., Schevon, C., Yates, M., McKhann, G., Botvinick, M., Sheth, S., & Hayden, B. (2020). *Human dorsal anterior cingulate neurons signal conflict by amplifying task-relevant information*. https://doi.org/10.1101/2020.03.14.991745

Egner, T. (2008). Multiple conflict-driven control mechanisms in the human brain. *Trends in Cognitive Sciences*, *12*(10), 374–380. https://doi.org/10.1016/j.tics.2008.07.001

Egner, T., & Hirsch, J. (2005). Cognitive control mechanisms resolve conflict through cortical amplification of task-relevant information. *Nature Neuroscience*, *8*(12), Article 12. https://doi.org/10.1038/nn1594

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics*, *16*(1), 143–149. https://doi.org/10.3758/BF03203267

Fengler, A., Govindarajan, L. N., Chen, T., & Frank, M. J. (2021). Likelihood approximation networks (LANs) for fast inference of simulation models in cognitive neuroscience. *eLife*, *10*, e65074. https://doi.org/10.7554/eLife.65074

Flesch, T., Juechems, K., Dumbalska, T., Saxe, A., & Summerfield, C. (2022). Orthogonal representations for robust context-dependent task performance in brains and neural networks. *Neuron*, *110*(7), 1258-1270.e11. https://doi.org/10.1016/j.neuron.2022.01.005

Frank, M. J. (2006). Hold your horses: A dynamic computational role for the subthalamic nucleus in decision making. *Neural Networks*, *19*(8), 1120–1136. https://doi.org/10.1016/j.neunet.2006.03.006

Fu, Z., Beam, D., Chung, J. M., Reed, C. M., Mamelak, A. N., Adolphs, R., & Rutishauser, U. (2022). The geometry of domain-general performance monitoring in the human medial frontal cortex. *Science*, *376*(6593), eabm9922. https://doi.org/10.1126/science.abm9922

Gaspelin, N., & Luck, S. J. (2018). The Role of Inhibition in Avoiding Distraction by Salient Stimuli. *Trends in Cognitive Sciences*, *22*(1), 79–92. https://doi.org/10.1016/j.tics.2017.11.001

Gheza, D., Bakic, J., Baeken, C., De Raedt, R., & Pourtois, G. (2019). Abnormal approach-related motivation but spared reinforcement learning in MDD: Evidence from fronto-midline Theta oscillations and frontal Alpha asymmetry. *Cognitive, Affective and Behavioral Neuroscience*, *19*(3). https://doi.org/10.3758/s13415-019-00693-4

Gheza, D., De Raedt, R., Baeken, C., & Pourtois, G. (2018). Integration of reward with cost anticipation during performance monitoring revealed by ERPs and EEG spectral perturbations. *NeuroImage*, *173*. https://doi.org/10.1016/j.neuroimage.2018.02.049

Gheza, D., & Kool, W. (2025). Distractor-specific control adaptation in multidimensional environments. *Nature Human Behaviour*, 1–20. https://doi.org/10.1038/s41562-024-02088-z

Gratton, G., Coles, M. G. H., & Donchin, E. (1992). Optimizing the use of information: Strategic control of activation of responses. *Journal of Experimental Psychology: General*, *121*, 480–506. https://doi.org/10.1037/0096-3445.121.4.480

Guggenmos, M., Sterzer, P., & Cichy, R. M. (2018). Multivariate pattern analysis for MEG: A comparison of dissimilarity measures. *NeuroImage*, *173*, 434–447. https://doi.org/10.1016/j.neuroimage.2018.02.044

Haxby, J. V., Connolly, A. C., & Guntupalli, J. S. (2014). Decoding Neural Representational Spaces Using Multivariate Pattern Analysis. *Annual Review of Neuroscience*, *June*, 435–456. https://doi.org/10.1146/annurev-neuro-062012-170325

Hazeltine, E., Lightman, E., Schwarb, H., & Schumacher, E. H. (2011). The boundaries of sequential modulations: Evidence for set-level control. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(6), 1898–1914. https://doi.org/10.1037/a0024662

Holroyd, C. B., & McClure, S. M. (2015). Hierarchical control over effortful behavior by rodent medial frontal cortex: A computational model. *Psychological Review*, *122*(1), 54–83. https://doi.org/10.1037/a0038339

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, *79*(8), 2554–2558. https://doi.org/10.1073/pnas.79.8.2554

Ileri-Tayar, M., Colvett, J. S., & Bugg, J. M. (2024). Between-task transfer of item-specific control is replicable and extends to novel conditions. *Journal of Experimental Psychology: Human Perception and Performance*, No Pagination Specified-No Pagination Specified. https://doi.org/10.1037/xhp0001200

Jiang, J., Heller, K., & Egner, T. (2014). Bayesian modeling of flexible cognitive control. *Neuroscience & Biobehavioral Reviews*, *46*, 30–43. https://doi.org/10.1016/j.neubiorev.2014.06.001

Kikumoto, A., Bhandari, A., Shibata, K., & Badre, D. (2024). A transient high-dimensional geometry affords stable conjunctive subspaces for efficient action selection. *Nature Communications*, *15*(1), 8513. https://doi.org/10.1038/s41467-024-52777-6

Kikumoto, A., & Mayr, U. (2020). Conjunctive representations that integrate stimuli, responses, and rules are critical for action selection. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(19), 10603–10608. https://doi.org/10.1073/pnas.1922166117

Kikumoto, A., Mayr, U., & Badre, D. (2022). The role of conjunctive representations in prioritizing and selecting planned actions. *eLife*, *11*, e80153. https://doi.org/10.7554/eLife.80153

Kleiner, M., Brainard, D., & Pelli, D. (2007). *What's new in Psychtoolbox-3?* https://pure.mpg.de/rest/items/item_1790332/component/file_3136265/content

Koob, V., Mackenzie, I., Ulrich, R., Leuthold, H., & Janczyk, M. (2023). The role of task-relevant and task-irrelevant information in congruency sequence effects: Applying the diffusion model for conflict tasks. *Cognitive Psychology*, *140*, 101528. https://doi.org/10.1016/j.cogpsych.2022.101528

Lieder, F., Shenhav, A., Musslick, S., & Griffiths, T. L. (2018). Rational metareasoning and the plasticity of cognitive control. *PLOS Computational Biology*, *14*(4), e1006043. https://doi.org/10.1371/journal.pcbi.1006043

Liesefeld, H. R., Liesefeld, A. M., Pollmann, S., & Müller, H. J. (2019). Biasing Allocations of Attention via Selective Weighting of Saliency Signals: Behavioral and Neuroimaging Evidence for the Dimension-Weighting Account. In T. Hodgson (Ed.), *Processes of Visuospatial Attention and Working Memory* (pp. 87–113). Springer International Publishing. https://doi.org/10.1007/7854_2018_75

Liesefeld, H. R., Liesefeld, A. M., Töllner, T., & Müller, H. J. (2017). Attentional capture in visual search: Capture and post-capture dynamics revealed by EEG. *NeuroImage*, *156*, 166–173. https://doi.org/10.1016/j.neuroimage.2017.05.016

Liston, C., Matalon, S., Hare, T. A., Davidson, M. C., & Casey, B. J. (2006). Anterior Cingulate and Posterior Parietal Cortices Are Sensitive to Dissociable Forms of Conflict in a Task-Switching Paradigm. *Neuron*, *50*(4), 643–653. https://doi.org/10.1016/j.neuron.2006.04.015

MacLeod, C. M., & Dunbar, K. (1988). Training and Stroop-like interference: Evidence for a continuum of automaticity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(1), 126–135. https://doi.org/10.1037/0278-7393.14.1.126

Mante, V., Sussillo, D., Shenoy, K. V., & Newsome, W. T. (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature*, *503*(7474), 78–84. https://doi.org/10.1038/nature12742

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177–190. https://doi.org/10.1016/j.jneumeth.2007.03.024

MATLAB version: 9.12.0 (R2022a). (2022). [Computer software]. The MathWorks Inc. https://www.mathworks.com

Miller, E. K., & Cohen, J. D. (2001). *An integrative theory of prefrontal cortex function*. 167–202.

Mittelstädt, V., Mackenzie, I. G., Koob, V., & Janczyk, M. (2023). The impact of distractor relevance on the strength and timing of cognitive control: Evidence from delta plots and diffusion model analyses. *Journal of*

Experimental Psychology: Learning, Memory, and Cognition, *49*(10), 1588–1614. https://doi.org/10.1037/xlm0001281

Monsell, S., Sumner, P., & Waters, H. (2003). Task-set reconfiguration with predictable and unpredictable task switches. *Memory & Cognition*, *31*(3), 327–342. https://doi.org/10.3758/BF03194391

Padmala, S., & Pessoa, L. (2011). Reward Reduces Conflict by Enhancing Attentional Control and Biasing Visual Cortical Processing. *Journal of Cognitive Neuroscience*, *23*(11), 3419–3432. https://doi.org/10.1162/jocn_a_00011

Pion-Tonachini, L., Kreutz-Delgado, K., & Makeig, S. (2019). ICLabel: An automated electroencephalographic independent component classifier, dataset, and website. *NeuroImage*, *198*, 181–197. https://doi.org/10.1016/j.neuroimage.2019.05.026

R Core Team. (2017). R. In *R Core Team* [Computer software]. https://doi.org/3-900051-14-3

Radev, S. T., Mertens, U. K., Voss, A., Ardizzone, L., & Köthe, U. (2022). BayesFlow: Learning Complex Stochastic Models With Invertible Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems*, *33*(4), 1452–1466. https://doi.org/10.1109/TNNLS.2020.3042395

Reynolds, J. H., & Chelazzi, L. (2004). ATTENTIONAL MODULATION OF VISUAL PROCESSING. *Annual Review of Neuroscience*, *27*(Volume 27, 2004), 611–647. https://doi.org/10.1146/annurev.neuro.26.041002.131039

Ritz, H., & Shenhav, A. (2023). Humans reconfigure target and distractor processing to address distinct task demands. *Psychological Review*. https://doi.org/10.1037/rev0000442

Ritz, H., & Shenhav, A. (2024). Orthogonal neural encoding of targets and distractors supports multivariate cognitive control. *Nature Human Behaviour*, 1–17. https://doi.org/10.1038/s41562-024-01826-7

Senoussi, M., Verbeke, P., Desender, K., De Loof, E., Talsma, D., & Verguts, T. (2022). Theta oscillations shift towards optimal frequency for cognitive control. *Nature Human Behaviour*, *6*(7), 1000–1013. https://doi.org/10.1038/s41562-022-01335-5

Sheth, S. A., Mian, M. K., Patel, S. R., Asaad, W. F., Williams, Z. M., Dougherty, D. D., Bush, G., & Eskandar, E. N. (2012). Human dorsal anterior cingulate cortex neurons mediate ongoing behavioural adaptation. *Nature*, *488*(7410), 218–221. https://doi.org/10.1038/nature11239

Siegel, M., Buschman, T. J., & Miller, E. K. (2015). Cortical information flow during flexible sensorimotor decisions. *Science*, *348*(6241), 1352–1355. https://doi.org/10.1126/science.aab0551

Simon, J. R., & Rudell, A. P. (1967). Auditory S-R compatibility: The effect of an irrelevant cue on information processing. *Journal of Applied Psychology*, *51*(3), 300–304. https://doi.org/10.1037/h0020586

Smith, E. H., Banks, G. P., Mikell, C. B., Cash, S. S., Patel, S. R., Eskandar, E. N., & Sheth, S. A. (2015). Frequency-Dependent Representation of Reinforcement-Related Information in the Human Medial and Lateral Prefrontal Cortex. *Journal of Neuroscience*, *35*(48), 15827–15836. https://doi.org/10.1523/JNEUROSCI.1864-15.2015

Smith, E. H., Horga, G., Yates, M. J., Mikell, C. B., Banks, G. P., Pathak, Y. J., Schevon, C. A., McKhann, G. M., Hayden, B. Y., Botvinick, M. M., & Sheth, S. A. (2019). Widespread temporal coding of cognitive control in

the human prefrontal cortex. *Nature Neuroscience*, *22*(11), 1883–1891. https://doi.org/10.1038/s41593-019-0494-0

Stokes, M. G., Buschman, T. J., & Miller, E. K. (2017). Dynamic Coding for Flexible Cognitive Control. In T. Egner (Ed.), *The Wiley Handbook of Cognitive Control* (1st ed., pp. 221–241). Wiley. https://doi.org/10.1002/9781118920497.ch13

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*, 643–662. https://doi.org/10.1037/h0054651

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136. https://doi.org/10.1016/0010-0285(80)90005-5

Ulrich, R., Schröter, H., Leuthold, H., & Birngruber, T. (2015). Automatic and controlled stimulus processing in conflict tasks: Superimposed diffusion processes and delta functions. *Cognitive Psychology*, *78*, 148–174. https://doi.org/10.1016/j.cogpsych.2015.02.005

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review*, *108*(3), 550–592. https://doi.org/10.1037/0033-295X.108.3.550

Vaughan, H. G. (1982). The Neural Origins Of Human Event-Related Potentials. *Annals New York Academy of Science*, 125–138.

Verguts, T. (2017). Binding by random bursts: A computational model of cognitive control. *Journal of Cognitive Neuroscience*. https://doi.org/10.1162/jocn_a_01117

Verguts, T., & Notebaert, W. (2008). Hebbian learning of cognitive control: Dealing with specific and nonspecific adaptation. *Psychological Review*, *115*(2), 518–525. https://doi.org/10.1037/0033-295X.115.2.518

Weber, J., Iwama, G., Solbakk, A.-K., Blenkmann, A. O., Larsson, P. G., Ivanovic, J., Knight, R. T., Endestad, T., & Helfrich, R. (2023). Subspace partitioning in the human prefrontal cortex resolves cognitive interference. *Proceedings of the National Academy of Sciences*, *120*(28), e2220523120. https://doi.org/10.1073/pnas.2220523120

Wessel, J. R., & Anderson, M. C. (2024). Neural mechanisms of domain-general inhibitory control. *Trends in Cognitive Sciences*, *28*(2), 124–143. https://doi.org/10.1016/j.tics.2023.09.008

Yeung, N., Botvinick, M. M., & Cohen, J. D. (2004). The Neural Basis of Error Detection: Conflict Monitoring and the Error-Related Negativity. *Psychological Review*, *111*(4), 931–959. https://doi.org/10.1037/0033-295X.111.4.939

Zaghloul, K. A., Weidemann, C. T., Lega, B. C., Jaggi, J. L., Baltuch, G. H., & Kahana, M. J. (2012). Neuronal Activity in the Human Subthalamic Nucleus Encodes Decision Conflict during Action Selection. *The Journal of Neuroscience*, *32*(7), 2453–2460. https://doi.org/10.1523/JNEUROSCI.5815-11.2012