A dynamic spatiotemporal normalization model for continuous vision

Angus F. Chapman[1,2] & Rachel N. Denison[1]

[1]Department of Brain and Psychological Sciences, Boston University, Boston, MA, USA

[2]Lead contact

Corresponding Author: Angus Chapman, angusc@bu.edu

Summary

Perception and neural activity are profoundly shaped by the spatial and temporal context of sensory input, which has been modeled by divisive normalization over space or time. However, theoretical work has largely treated normalization separately within these dimensions and has not explained how future stimuli can suppress past ones. Here we introduce a computational model with a unified spatiotemporal receptive field structure that implements normalization across both space and time and ask whether this model captures the bidirectional effects of temporal context on neural responses and behavior. We found that biphasic temporal receptive fields emerged from this normalization computation, consistent with empirical observations. The model also reproduced several neural response properties, including nonlinear response dynamics, subadditivity, response adaptation, backwards masking, and bidirectional contrast-dependent suppression. Thus, the model captured a wide range of neural and behavioral effects, suggesting that a unified spatiotemporal normalization computation could underlie dynamic stimulus processing and perception.

Keywords: temporal normalization, perception, computational model, continuous vision, visual system

Introduction

Although the world around us is highly dynamic, most theories and models of visual processing consider only a static snapshot of this constantly changing stream of visual stimulation. Less understood is the dynamic neural activity that supports visual perception, which exhibits several non-linear response properties that depend on time-varying stimulus input. Neurons in a range of cortical areas show sensitivity to time-varying stimuli (Mauk & Buonomano, 2004), such as motion-sensitive neurons in V1 and MT, which can be characterized by their *spatiotemporal* receptive fields. More broadly, many neural responses and perception depend not just on current inputs but also on recent stimulus history (Fritsche et al., 2020; Gao et al., 2020; Hasson et al., 2008; Murray et al., 2014; Wolff et al., 2022), and can be modulated by future context as well (Breitmeyer & Öğmen, 2006; Yeshurun et al., 2015). These findings demonstrate that sensory systems are sensitive to temporal structure, which raises questions about what mechanisms exist to process dynamic inputs (Cavanagh et al., 2020; Golesorkhi et al., 2021; Soltani et al., 2021).

A powerful theoretical framework for visual processing is based on the principle of normalization, which has been proposed as a canonical neural computation (Carandini & Heeger, 2012). Normalization is the idea that neurons can have suppressive effects on one another as a function of their tuning preferences, acting to “normalize” overall activity levels within a neural population (Heeger, 1992). Most normalization models are static, with the normalization computation operating across cortical space, which we refer to as spatial normalization. Temporal normalization, in contrast, is the idea that neural activity undergoes normalization across time (Heeger, 1992). Current dynamic normalization models implement temporal normalization by using a temporally filtered and delayed version of the excitatory input drive to compute the normalization signal (Groen et al., 2022; Zhou et al., 2019). Other models combine linear stimulus evoked responses with non-linear compressive—rather than divisive—computations to model how neurons respond to dynamic input (Kim et al., 2024; Kupers et al., 2024). All

these models better predict neural response dynamics than linear-only models and capture several non-linear phenomena (Zhou et al., 2019). However, they also have important limitations. First, the entire excitatory time course must be known in advance, limiting their biological feasibility. Second, normalizing by a delayed copy of the excitatory drive means that only past stimuli can suppress future stimuli, not vice versa. Third, normalization is computed for each neural unit (neuron, voxel, electrode, etc.) in isolation, so to date these models have not considered how temporal normalization may operate across a broader suppressive pool, which has been critical for the success of spatial normalization models.

Here we introduce the dynamic spatiotemporal attention and normalization model (D-STAN), a model of visual processing in which spatial and temporal normalization are integrated within a unified receptive field-based spatiotemporal normalization computation. Model neurons are situated in a recurrent neural network architecture, allowing real-time processing of continuous visual inputs. We analyzed the response properties of this model, focusing on its time-varying behavior and modulation by temporal context. By unifying normalization across space and time, we find that D-STAN captures key non-linear temporal response properties of neurons and predicts human behavior. Critically, unlike previous dynamic normalization models, it can produce bidirectional temporal suppression between stimuli in a sequence, allowing for empirically demonstrated effects of temporal context that operate both forward and backward in time. The current work builds on the normalization model of dynamic attention developed by Denison et al. (2021), which generalized the normalization model of attention (Reynolds & Heeger, 2009) to the time domain. D-STAN advances previous work by implementing excitatory and suppressive drives that depend on recent stimulus history, imbuing the model neurons with receptive fields and normalization computations that are inherently spatiotemporal.

Results

The dynamic spatiotemporal attention and normalization model (D-STAN)

We introduce a model of dynamic visual perception and attention building on (Denison et al., 2021) that produces neural responses and behavioral outputs. The model simulates sensory responses to stimulus input through neurons that are tuned to specific spatial locations and feature values and, critically, integrate inputs over the recent past, giving them spatiotemporal receptive fields (Figure 1A). The "spatial" component of the receptive field can model tuning to any feature represented across cortical space; in the current work we use orientation as the modeled feature. Sensory responses can also be modulated by time-varying attention, though we do not explore attentional modulation here. The sensory responses are read out by a decision layer, which accumulates evidence toward a particular behavioral output. Importantly, we include spatiotemporal normalization at each stage of processing, with excitatory and suppressive drives that contribute to the final neural response (Figure 1B). Responses are continuously updated at each time point with differential equations, allowing us to examine how model parameters affect the dynamics of neural responses as well as the final behavioral output. Thus, the model is an "online" process model that generates neural activity continuously given ongoing stimulus input at each time step (Figure 1C), and whose output can be related to performance on perceptual tasks.

The core of the model is the spatiotemporal normalization computation, which divides the spatiotemporal excitatory drive by the spatiotemporal suppressive drive. To investigate the effects of temporal normalization, we implemented temporal receptive fields in model sensory neurons by allowing the excitatory drive to depend on stimulus input at previous points in time. Specifically, at each time point, the stimulus time course was weighted by an exponential decay function, such that input at more distant points in the past had less effect on the neuron's response. The time constant (τ_E)

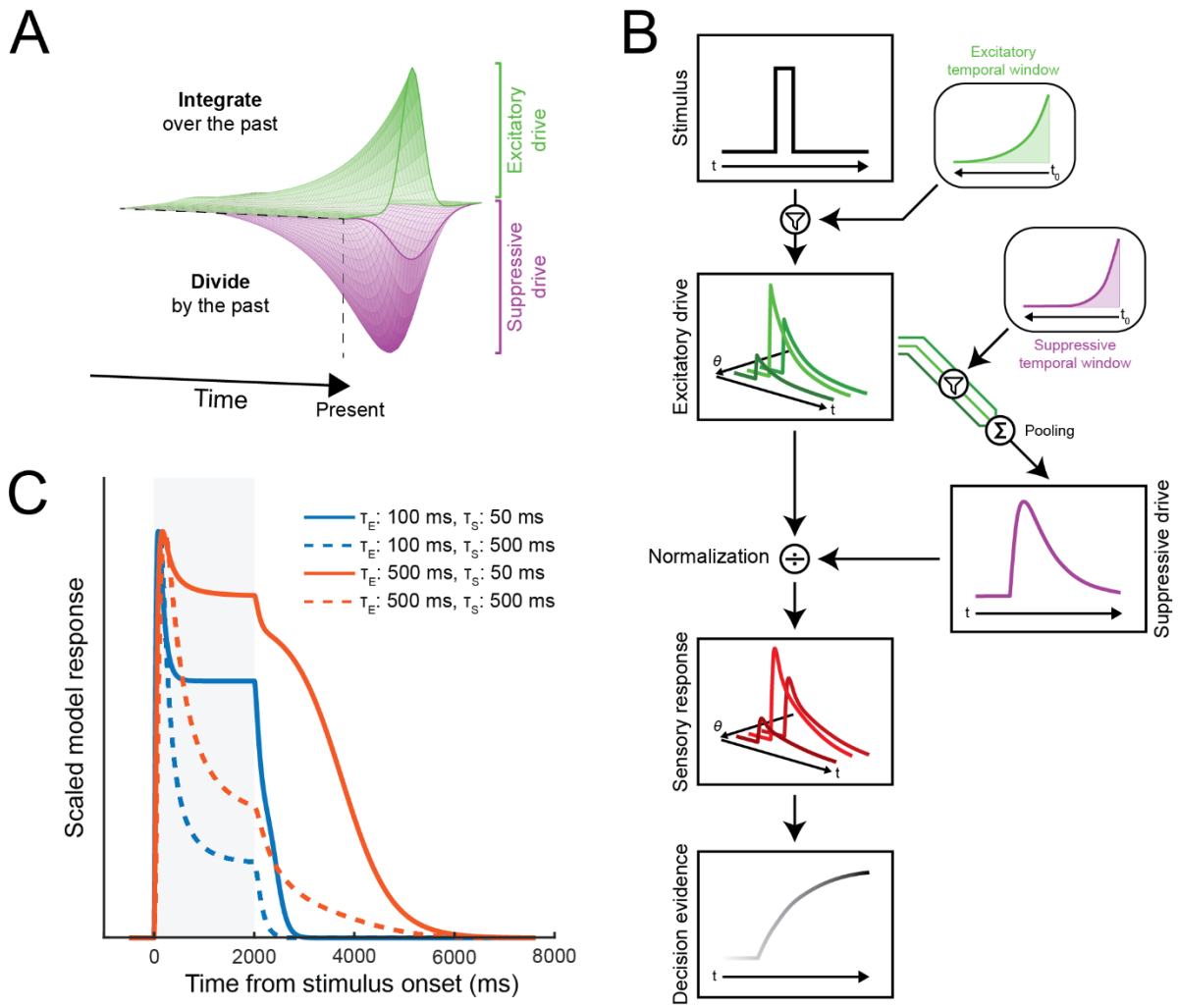

Figure 1. Architecture of the Dynamic Spatiotemporal Attention and Normalization model (D-STAN). A) Spatiotemporal receptive field structure. Model sensory neurons integrate over the recent history of stimulus input, while contributing to a suppressive pool that also extends into the past. Model neurons can be tuned to different spatiotemporal properties of the stimulus input as well as to other stimulus features. B) Schematic of computations in D-STAN. The model receives time-varying stimulus input (here, an oriented stimulus with a specific contrast), which is continuously filtered by the excitatory temporal window. The stimulus input produces excitatory drives in model neurons tuned to different orientations (θ). The excitatory drives of each neuron are continuously filtered and pooled to calculate the suppressive drive, which in turn normalizes the excitatory drives to compute sensory responses. Sensory layer activity is fed into the decision layer, which accumulates evidence about the stimulus orientation. C) Example sensory layer responses to a stimulus presentation (shaded grey region) as a function of different excitatory and suppressive temporal window parameters.

determined the relative weight given to stimuli at each point in time. When τ_E is zero, the model neuron only receives excitatory drive when stimulus input is present, but increasing τ_E results in responses that are driven even after the stimulus ends. Additionally, the suppressive drive of the neural population was pooled across neural units, as in previous work (Denison et al., 2021; Reynolds & Heeger, 2009), but also across time. The suppressive drive was calculated by weighting the excitatory drive across previous time points by an exponential decay function with a separate time constant (τ_S). We performed a series of simulations varying aspects of the stimulus input to sensory layers to examine whether the model could reproduce several different non-linear response properties and to determine the effect of the excitatory and suppressive time constants on these response dynamics.

Stimulus history differentially affects model drives and responses

The dependence of a neuron's current response on recent stimulus inputs defines its temporal receptive field. We used reverse correlation to determine how excitatory and suppressive temporal windows interact to shape the functional temporal receptive fields of model neurons. We first examined how the model responses depended on stimulus history for a specific set of excitatory and suppressive temporal windows (τ_E = 400 ms, τ_S = 100 ms). Random stimulus input was fed into the model sensory layer, driving variable activity across simulations. We calculated how the presence of stimuli up to 1200 ms in the past affected model neuron responses at the current moment by correlating the random stimulus vectors with the observed excitatory drive, suppressive drive, and sensory layer response. As expected, the excitatory drive depended most strongly on input close to the current time, with weights back in time following the exponential excitatory temporal window (Figure 2A). The suppressive drive depended on times in the recent past, and could be approximated by a convolution of the excitatory and suppressive temporal windows (Figure 2B), consistent with how suppressive drives are a temporally-

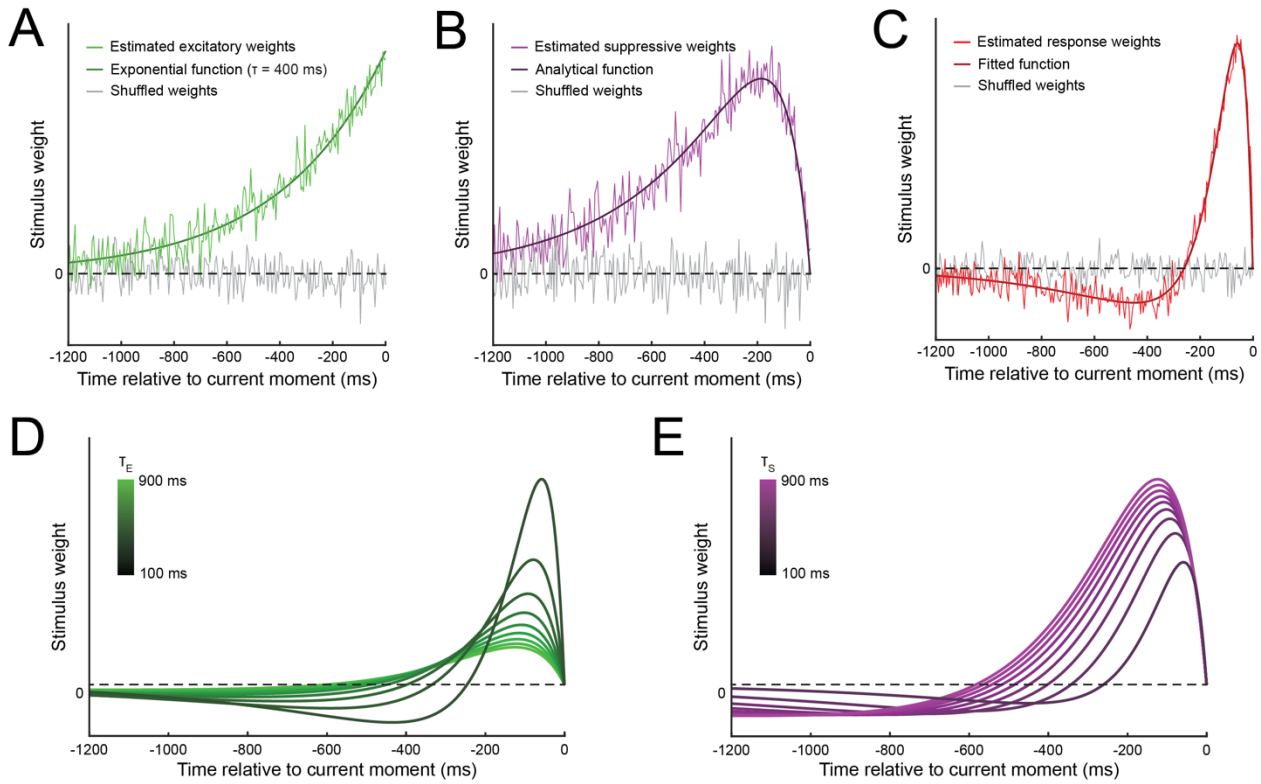

Figure 2. Reverse correlation analysis of model drives and sensory response reveals biphasic temporal receptive field. Stimulus weights were estimated for A) excitatory drives, B) suppressive drives, and C) sensory layer responses. Examples shown reflect simulations with τ_E = 400 ms and τ_S = 100 ms. Each set of estimated weights were fitted using different functional forms, (see main text). We also estimated the sensory layer weight functions (i.e., temporal receptive fields) across a range of parameters for D) excitatory temporal windows, and E) suppressive temporal windows. While varying one parameter, the other was fixed at an intermediate level (400 ms) for visualization purposes; the pattern of results did not depend on the specific fixed value.

weighted sum of all model neurons' excitatory drives, meaning both τ_E and τ_S affect the shape of the suppressive drive.

Reverse correlation revealed that this implementation of normalization generates a biphasic temporal receptive field (Figure 2C), where stimulus input close to the current time drives model neuron responses and input further in the past reduces responses. Notably, the biphasic stimulus weighting

function was not built directly into the model, but emerged through the interaction between the excitatory and suppressive temporal windows. A similar biphasic function in time has been found in empirical recordings from LGN neurons (Mante et al., 2008), and is similar to those used in models of neural temporal dynamics (Kim et al., 2024; Kupers et al., 2024; Zhou et al., 2019). The estimated response weights were well fitted by a difference of Gamma functions, following previous work (Mante et al., 2008; Zhou et al., 2019).

We performed additional simulations to explore the effect of the excitatory and suppressive time constants on the shape of the temporal receptive fields. Increasing the excitatory time constant, while holding the suppressive time constant fixed (τ_S = 400 ms), resulted in a scaling and extension of the temporal receptive field to times further in the past (Figure 2D), as stimuli at these times fell into the longer excitatory temporal windows—a "flattening out" of the response profile. Increasing the suppressive time constant, with the excitatory time constant fixed (τ_E = 400 ms), also resulted in an extension of the temporal receptive field to past times (Figure 2E), although this was a consequence of the suppression being distributed over a longer time span, resulting in less suppression for more recent times. Increasing the suppressive time constant also resulted in longer periods of suppression that extended further into the past, as would be expected given the longer suppressive windows. Thus, the excitatory and suppressive temporal windows generated temporal receptive fields that exhibit a variety of response profiles depending on each time constant.

Reproducing known temporal properties of neuronal responses

We next asked whether D-STAN could reproduce several known non-linear temporal response properties of neurons: 1) transient-sustained dynamics, 2) subadditivity, 3) response adaptation, and 4) backward masking. For each property we asked whether excitatory temporal windows, suppressive

temporal windows, or both were necessary to reproduce the property and how the non-linear temporal effects depended on the excitatory and suppressive time constants in the model.

Transient-sustained dynamics. Neurons typically show an initial transient response followed by sustained activity to a prolonged stimulus (Lisberger & Movshon, 1999; Motter, 2006). We found that the suppressive temporal window was necessary to reproduce these dynamics. When varying the excitatory time constant with no suppressive temporal window ($\tau_S = 0$), responses increased gradually towards a stable activity level following stimulus onset and decreased only after stimulus offset, consistent with linear predictions (Zhou et al., 2019). Higher τ_E resulted in slower rise times (Figure 3A) and consequently longer times to peak (Figure 3B), as well as more prolonged responses after stimulus offset (Figure 3C). The model exhibited slower response dynamics with increasing τ_E, because longer integration windows provided excitatory drive from stimulus input further in the past. In contrast, when varying the suppressive time constant alone ($\tau_E = 0$), model responses showed a more typical transient response, with an initial peak shortly after stimulus onset that decreased to a stable level before stimulus offset (Figure 3D). Higher τ_S resulted in longer times to peak (Figure 3E) and a lower stable activity level relative to peaks (Figure 3F), because suppression integrated more slowly but reached greater levels overall. Varying both time constants generated model responses with a diverse range of temporal profiles (Figure 1C). Thus, suppression alone, or a combination of excitatory and suppressive temporal windows, can capture typical neural dynamics to prolonged stimulus presentations.

Subadditivity. Another hallmark of non-linear response dynamics is that neural responses display temporal subadditivity: doubling the duration of the stimulus leads to a less than doubling of the response amplitude (Groen et al., 2022; Zhou et al., 2018). Increasing the stimulus duration in D-STAN resulted in more prolonged model responses (Figure 4A) that were strongly subbadditive, as quantified by the area under each curve (Figure 4B). We found that increasing either τ_E or τ_S was sufficient to produce subadditive model responses (Figure 4C-D); when both time constants were zero, model

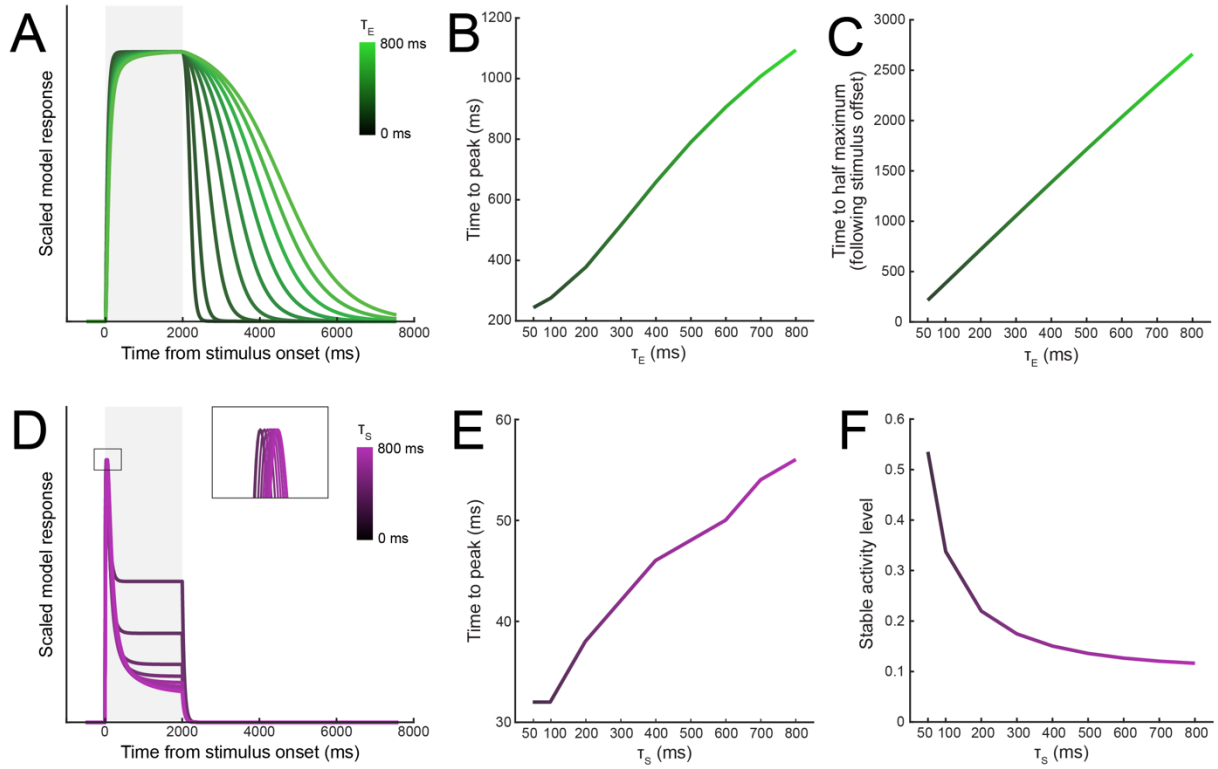

Figure 3. Excitatory and suppressive time constants differentially contribute to transient-sustained response dynamics. A) Effect of the excitatory time constant, ranging from 0-800 ms in 100 ms steps. Shaded region shows the stimulus presentation period. B) Time to peak for sensory responses as a function of τ_E. C) Time for sensory responses to reduce to 50% of maximum as a function of τ_E. D) Effect of the suppressive time constant on model sensory responses. The insert shows the early peak responses. E) Time to peak for the sensory responses as a function of τ_S. F) Stable activity level of sensory responses, relative to the peaks, reached by the end of the stimulus presentation period as a function of τ_S.

responses fell just below the linear prediction, showing only slight subadditivity due to the time constant inherent to the model's recursive computation (τ_R, Equation 6). For values of τ_E and τ_S above zero, each doubling of the stimulus duration increased responses by only ~1.3-1.6×. Combining different parameter values produced similar levels of subadditivity (Figure 4E), where lower values of each time constant resulted in relatively move subadditivity at short stimulus durations, with higher values resulting in more subadditivity at longer stimulus durations.

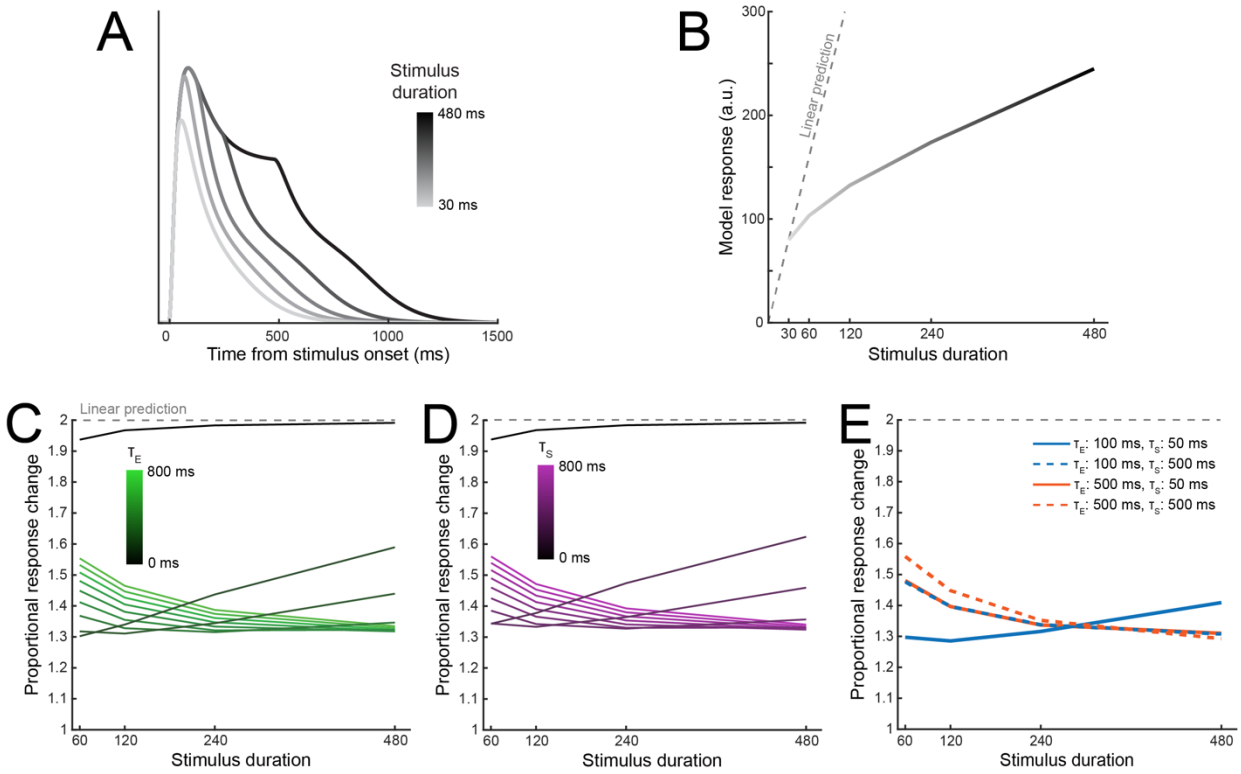

Figure 4. Subadditivity of model responses is robust across excitatory and suppressive time constants. A) Sensory layer responses over time in response to stimuli of varying durations (30, 60, 120, 240, or 480 ms), using τ_E = 100 ms and τ_S = 50 ms. B) Model responses calculated as the area under the curve of the sensory responses across the full simulated trial. Linear predictions are calculated relative to the model response for the 30 ms stimulus duration. C) Effect of τ_E and D) τ_S on subadditivity as a function of stimulus duration. Subadditivity was measured by taking the model response for each stimulus duration and dividing it by the response for a stimulus presented for half of that duration. Values below 2 represent subadditive responses. E) Proportional response changes as a function of different excitatory and suppressive time constants.

Response adaptation. Neurons exhibit response adaptation, such that responses are lower following repeated stimulus presentations (Kohn, 2007; Lisberger & Movshon, 1999; Priebe et al., 2002; Solomon & Kohn, 2014; Vogels, 2016). The magnitude of this adaptation depends on the interstimulus interval (ISI), with shorter ISIs resulting in stronger response adaptation. We observed a similar pattern in model simulations, where responses to the second stimulus (T2) in a sequence of two identical stimuli were lower when the ISI was shorter (e.g., 100 ms; Figure 5A) compared to longer ISIs (e.g., 900 ms;

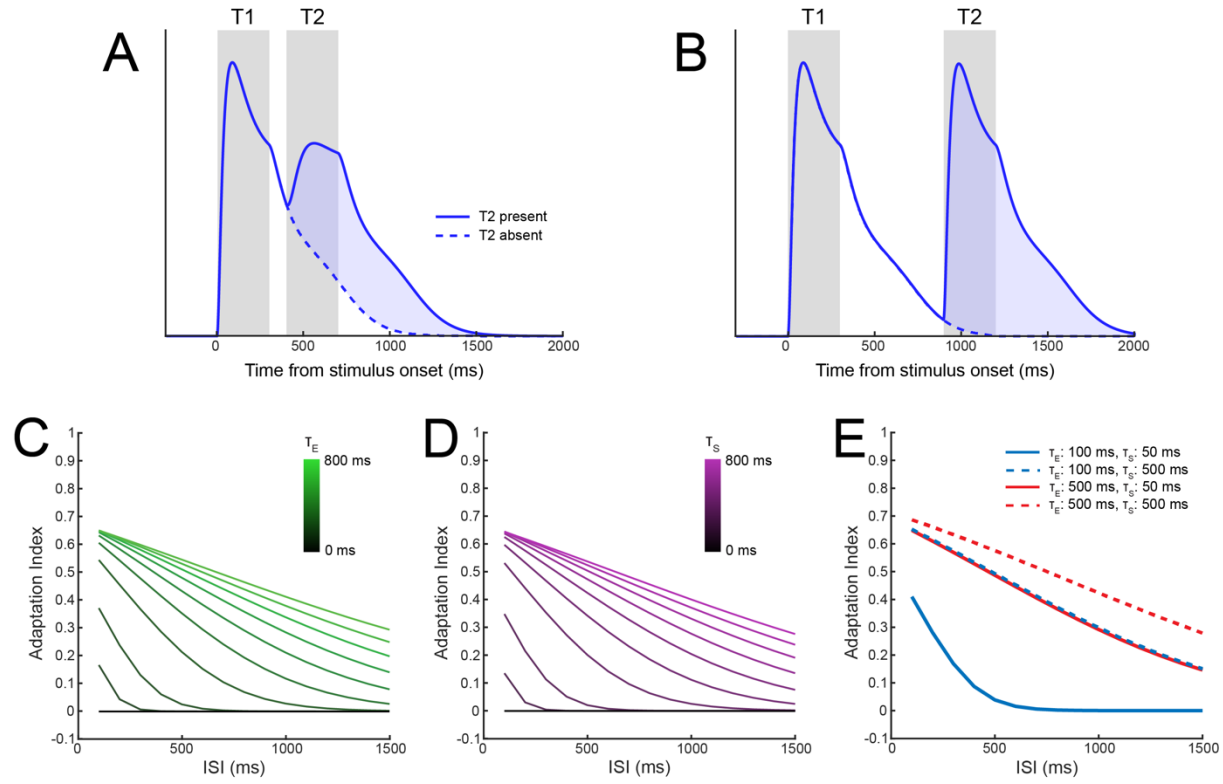

Figure 5. Response adaptation arises from either excitatory or suppressive temporal windows. A) Model sensory responses as a function of the presence of a preceding stimulus with an interstimulus interval of 100 ms, using τ_E = 100 ms and τ_S = 50 ms. Response adaptation is demonstrated by the reduced response to T2 (blue shaded region) relative to T1. B) Response adaptation was nearly eliminated when the ISI was increased to 600 ms. C) Effect of τ_E and D) τ_S on response adaptation for different ISIs. An adaptation index of zero represents no change in responses; positive indices show response adaptation. E) Combining different values of the τ_E and τ_S produces a range of response adaptation profiles.

Figure 5B). We quantified the magnitude of response adaptation by measuring the reduction in the model response to T2 after subtracting out the response to T1 (i.e., the shaded blue region between curves in Figure 5A-B; (Lisberger & Movshon, 1999; Priebe et al., 2002). With increases in either τ_E or τ_S, response adaptation persisted across longer ISIs (Figure 5C-D), as longer time constants allowed the excitatory and/or suppressive drives to extend across longer ISIs, leading to normalization of T2 by T1. For combinations of shorter and longer time constant parameters (Figure 5E), suppression accumulated

as both τ_E and τ_S increased. Thus either excitatory temporal windows, suppressive temporal windows, or both could generate response adaptation, and had quantitatively similar effects on suppression indices.

Empirical work has shown that response adaptation can occur even with non-identical stimuli (Liu et al., 2009; Priebe & Lisberger, 2002), with one study in particular finding that most MT neurons are adapted by a wider range of motion directions than they are responsive to (Priebe & Lisberger, 2002). Such findings are consistent with D-STAN's prediction that excitatory and suppressive tuning widths can differ, but not with models in which each neuron is normalized by a delayed version of its own response (Groen et al., 2022; Zhou et al., 2019). Therefore, we next examined adaptation to sequences of non-identical stimuli in D-STAN. The current simulations pool suppression uniformly across the population of orientation-tuned neurons, so we expected the model to exhibit response adaptation even for a sequence of two orthogonal stimuli. These stimuli drove responses in different neural populations, so we quantified response adaptation in model neurons that were maximally responsive to T2 by measuring the neural response when T1 was present or absent. We observed response adaptation particularly when ISIs were short (Figure 6A vs 6B). Both the excitatory and suppressive time constants contributed to response adaptation, with greater adaptation at higher values of τ_E and τ_S that also persisted across ISIs (Figure 6C-D). Excitatory time constants affected response adaptation less when the stimuli were orthogonal than when they were identical (Figure 5C vs Figure 6C), because the stimuli drove responses in non-overlapping neural subpopulations. However, the effect of the suppressive time constants was equivalent for orthogonal and identical stimuli, because the uniform pooling effectively ignores which neurons are active when computing suppressive drives. Lastly, the time constants interacted to produce varying response adaptation effects across ISIs (Figure 6E). Thus, both the excitatory and suppressive temporal windows contributed to response adaptation even for non-identical stimulus sequences.

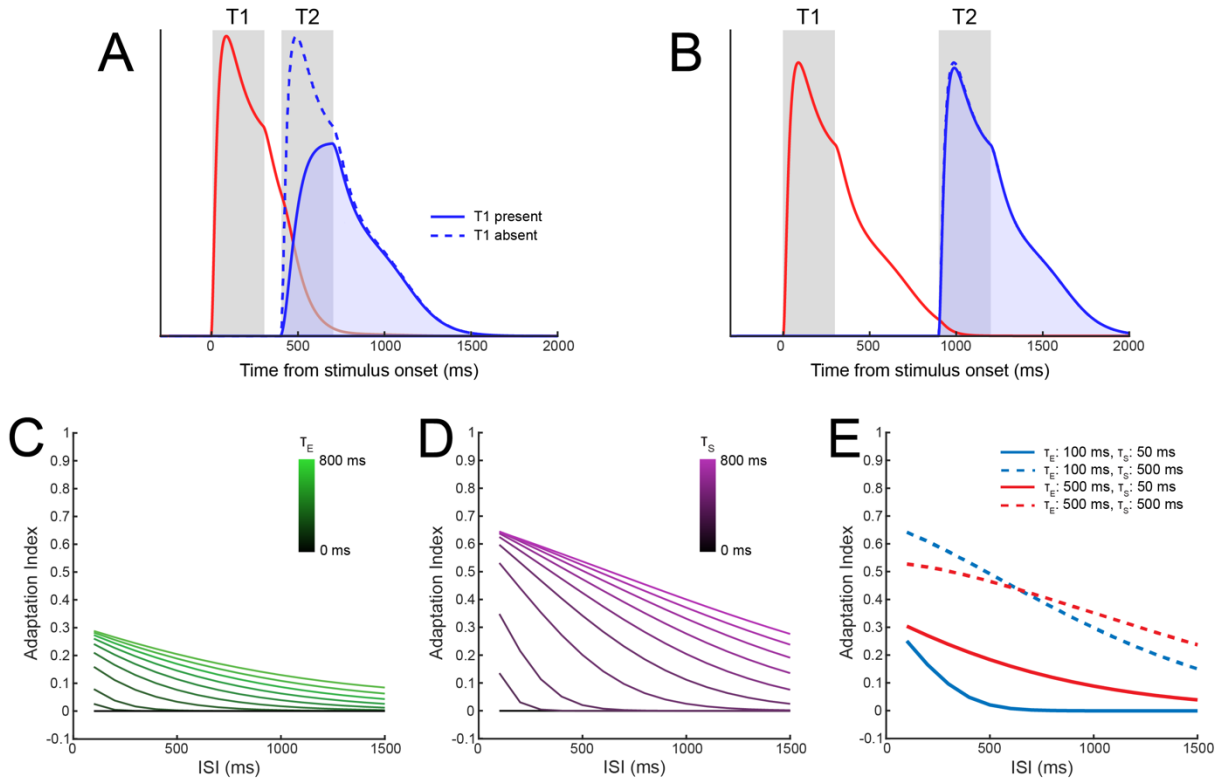

Figure 6. Response adaptation persists for non-identical stimuli. A) Model sensory responses as a function of the presence of a preceding stimulus with an interstimulus interval of 100 ms, using τ_E = 100 ms and τ_S = 50 ms. Response adaptation is demonstrated by the reduction in T2 responses while T1 is present (shaded blue vs. dotted blue region). B) Response adaptation was effectively eliminated when the ISI was increased to 600 ms. C) Effect of τ_E and D) τ_S on response adaptation for different ISIs. An adaptation index of zero represents no change in responses; positive indices show response adaptation. E) Combining different values of the τ_E and τ_S produces a range of response adaptation profiles.

Backward masking. In backward masking, a neuron's ongoing response to an initial stimulus is suppressed by a subsequent stimulus (Breitmeyer & Öğmen, 2006; Enns & Di Lollo, 2000; Kovacs et al., 1995). Similar to response adaptation, the magnitude of backward masking diminishes as the stimulus onset asynchrony (SOA) between stimuli increases. D-STAN exhibited both backward masking and this characteristic SOA dependence, with greater masking at shorter SOAs (e.g., 250 ms; Figure 7A) than longer SOAs (e.g., 500 ms; Figure 7B). We simulated backwards masking using a sequence of two orthogonal stimuli and quantified the degree of masking in model simulations by measuring the

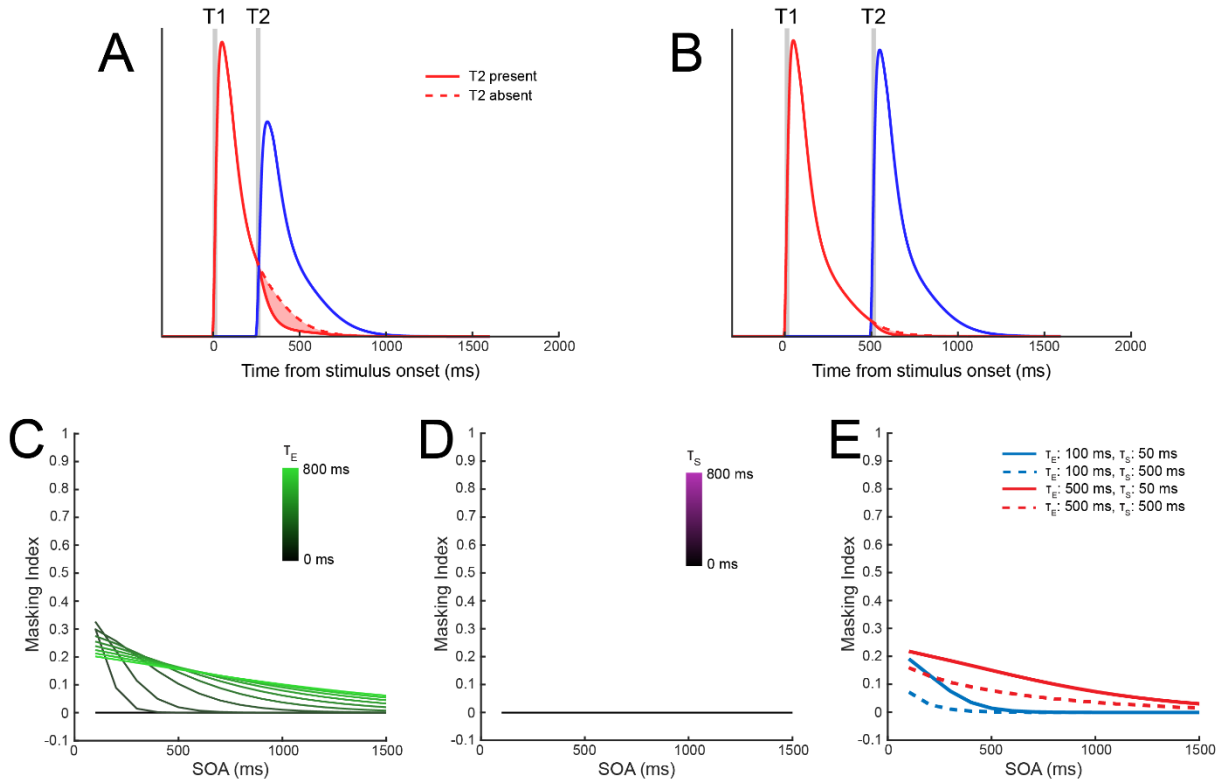

Figure 7. Backward masking requires excitatory temporal windows. A) Model sensory responses for one stimulus (T1, red lines) as a function of the presence of a subsequent stimulus (T2, blue line) with a stimulus onset asynchrony of 250 ms, using τ_E = 100 ms and τ_S = 50 ms. Backward masking is demonstrated by the reduction in T1 responses following the onset of T2, as indicated by the shaded red region. B) Backward masking was eliminated when the SOA was increased to 500 ms. C) Effect of τ_E and D) τ_S on backward masking for different SOAs. E) Combining different values of the τ_E and τ_S affected the pattern of backwards masking.

response to T1 when T2 was present vs absent. Notably, the excitatory temporal window was necessary to produce backward masking, and its strength depended on τ_E (Figure 7C). Longer time constants resulted in backward masking that was generally stronger and persisted across longer intervals. In contrast, masking did not occur with the suppressive window alone, when τ_E was zero (Figure 7D). In this case, there was no temporal overlap in sensory responses to the two stimuli, which meant no normalization of the first stimulus by the second. However, when we explored the interaction between parameters, both τ_E and τ_S affected the magnitude of backward masking (Figure 7E). In general, higher τ_E

values resulted in stronger backward masking, particularly for shorter SOAs (blue vs red lines in Figure 7E), as the extended sensory responses resulted in more overlap between the two stimuli and increased the overall normalization. In contrast, higher τ_S values resulted in reduced backward masking (dashed vs solid lines), as longer suppressive temporal windows tended to reduce the overlap in sensory responses between stimuli (cf. Figure 3D). Thus, the excitatory temporal window was necessary to produce backward masking, but so long as it was present, the strength of masking could be modulated by the suppressive temporal window.

Temporal normalization results in contrast-dependent interactions between stimuli across time

A hallmark of spatial normalization models is that they generate suppressive effects that systematically depend on stimulus contrast (Heeger, 1992; Reynolds & Heeger, 2009). Therefore we next investigated the contrast dependence of temporal context effects produced by D-STAN. Specifically, we examined how model responses are affected by the contrast of a competing stimulus. In D-STAN, contrast increases the magnitude of stimulus input, and thus increases the overall excitatory and suppressive drives in the sensory layers. As such, we expected contrast modulations to have effects similar to our simulations of response adaptation and backward masking: higher contrast (presence vs absence) for one of two stimuli in a sequence would result in reduced model responses to the other.

We first simulated model responses using an SOA of 250 ms, with excitatory and suppressive time constants that resulted in both response adaptation and backward masking (τ_E = 400 ms, τ_S = 100 ms; see Figure 2C). When the contrast of the second stimulus (T2; Figure 8A) was high (64%) vs. low (16%) we observed a reduction in the model response to the first stimulus (T1). Likewise, when the contrast of T1 was high vs. low, the model response to T2 was reduced (Figure 8B). Both effects were due

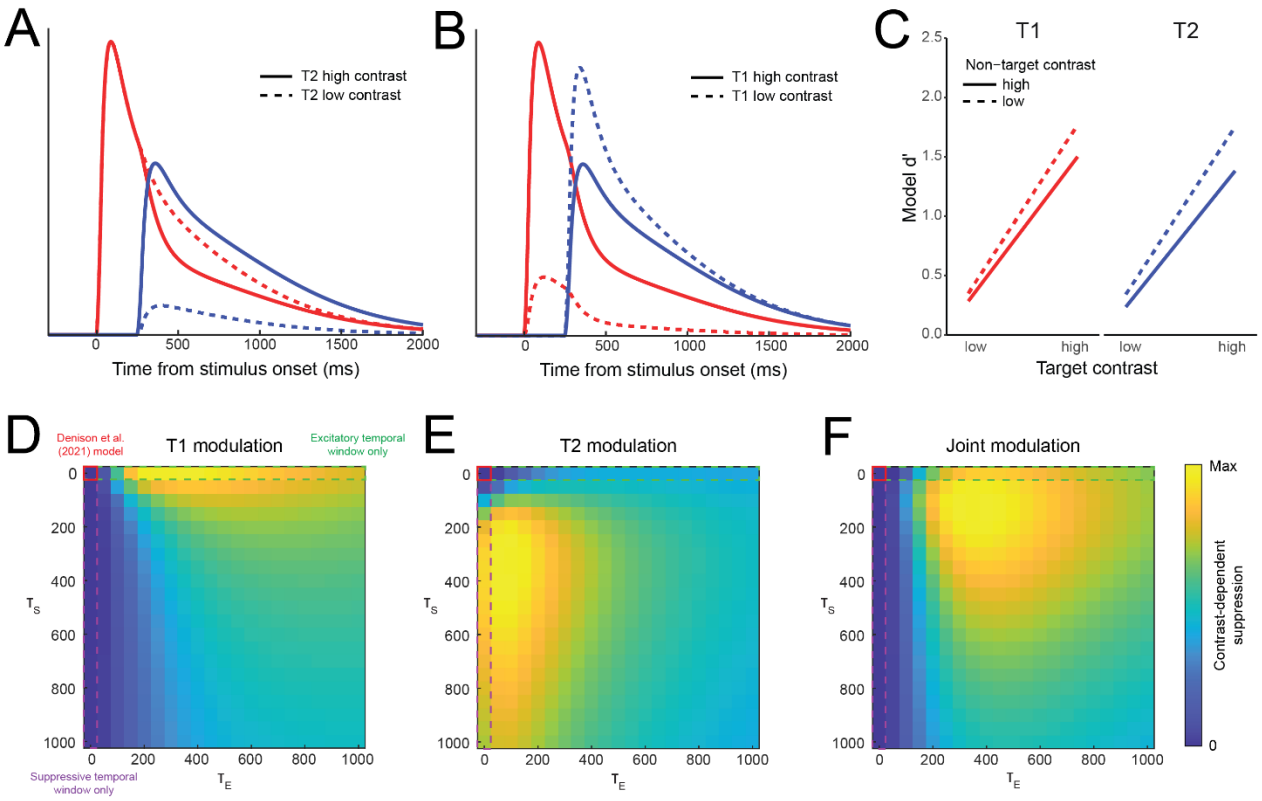

Figure 8. Contrast-dependent suppression forward and backward in time. A) When T1 (red lines) was fixed at high contrast, its response was modulated by the contrast of T2 (red solid line vs dashed line), using τ_E = 400 ms and τ_S = 100 ms. B) Likewise, the response to a high contrast T2 was modulated by the contrast of T1 (blue solid vs dashed lines). C) Modeled behavioral prediction in terms of d' for discriminating clockwise vs counterclockwise stimulus orientations. Performance was lower for each target stimulus when the non-target was presented at higher contrast, consistent with recent empirical observations. D) Effect of τ_E and τ_S on the contrast-dependent modulation of T1, E) T2, and F) the joint modulation of both stimuli, calculated by a pointwise multiplication of individual target heatmaps. A region of the parameter space with moderate τ_E and low τ_S provided the strongest combined modulation of both target stimuli, comparable to human behavior as reported by Epstein and Denison (2023).

to increases in the overall suppressive drive caused by higher contrast stimuli, resulting in stronger normalization of the model response to the other stimulus.

Psychophysical work has demonstrated that orientation discriminability for a target stimulus can be impaired by a non-target stimulus presented either before or after it by 250 ms, with greater impairment for higher vs. lower contrast non-target stimuli (Epstein & Denison, 2023). To determine if

the model could also produce this behavioral pattern, we used the decision layer of D-STAN to calculate discriminability (d') between different target orientations (Figure 8C). Model d' was higher when the non-target was presented at a lower contrast, regardless of the target contrast. Notably, this effect occurred both forward (i.e., the contrast of T1 affected responses to T2) and backward in time (i.e., the contrast of T2 affected responses to T1). In these simulations, excitatory drives were affected by target contrast, with higher target contrast resulting in stronger sensory responses and higher model d' for targets. On the other hand, the effects of non-target contrast were imparted through the suppressive drive, with high contrast non-targets resulting in lower sensory responses and d' for targets.

We also explored how the excitatory and suppressive temporal windows modulated the effects of one stimulus's contrast on the model responses to the other stimulus. We found that temporal windows were necessary for contrast-dependent suppression, as when both time constants were zero—reducing the model to the original Denison et al. (2021) version—there was no modulation for either stimulus (Figure 8D-F, red solid outline). For T1, we found that contrast-dependent suppression was strongest for a wide range of τ_E values (~250-850 ms) combined with a low τ_S (0-100 ms; Figure 8D). In particular, we found that suppression alone (i.e., $\tau_E = 0$) was insufficient to produce the observed effects (Figure 8D, purple dashed outline), because the lack of any extended excitatory drive meant the sensory responses did not overlap in time. This is comparable to the lack of backward masking present in our simulations manipulating only τ_S (cf. Figure 7D). In contrast, modulation of T2 was strongest when τ_S was an intermediate value (~250-650 ms) and τ_E was low but non-zero (~50-150 ms; Figure 8E). To assess what temporal window parameter combinations produced contrast-dependent suppression simultaneously across both T1 and T2 (Figure 8F), we computed a combined modulation index, which showed a range of intermediate τ_E values (300-600 ms) combined with low τ_S (50-200 ms) that resulted in moderate suppressive effects for both T1 and T2 (as in Figure 8C, with $\tau_E = 400$ ms and $\tau_S = 100$ ms). The ability of D-STAN to produce bidirectional contrast-dependent suppression demonstrates how temporal

normalization can account for interactions between stimuli that are separated in time, both affecting ongoing sensory processing as well as behavioral responses.

Discussion

Normalization has proven to be a successful computational principle that predicts neural responses and perception, yet most previous normalization models have been static and not considered how normalization affects responses that occur across time. Here, we demonstrate how spatiotemporal normalization can be implemented in a recurrent neural network model of dynamic visual processing. Our model, D-STAN, is grounded in the framework of normalization models established by Reynolds and Heeger (2009), which was subsequently extended into the temporal domain by Denison et al. (2021). The core of D-STAN is a unified spatiotemporal normalization computation, which provides local contextual modulation across space, time, and features such as orientation. This normalization computation is supported by spatiotemporal receptive fields, which we implemented by allowing the excitatory and suppressive drives for model neurons to depend on recent stimulus history. D-STAN inherits the spatial and featural properties previously described for static normalization models (Reynolds & Heeger, 2009), so here we focused on the temporal dimension and investigated the dynamic response properties of this model.

We report several findings. First, the model sensory neurons exhibit temporal receptive fields with biphasic profiles, similar to neuronal receptive field properties observed empirically (Mante et al., 2008). Notably, this biphasic profile was not incorporated directly into the model but emerged from the interaction between the excitatory and suppressive drives, the exponential temporal windows, and the normalization computation. Second, D-STAN displays several non-linear dynamic response properties that have been observed empirically, including transient-sustained response dynamics (Lisberger &

Movshon, 1999; Motter, 2006), subadditivity with increasing stimulus duration (Groen et al., 2022; Zhou et al., 2018), response adaptation (Kohn, 2007; Priebe et al., 2002; Vogels, 2016), and backward masking (Breitmeyer & Öğmen, 2006; Enns & Di Lollo, 2000; Kovacs et al., 1995). Previous dynamic normalization models have been able to account for several of these properties but have not predicted backward masking (Groen et al., 2022; Zhou et al., 2019). Third, D-STAN predicts bidirectional contrast-dependent suppression between successive stimuli, which has recently been confirmed psychophysically (Epstein & Denison, 2023). D-STAN can thus account for a wide range of neural and behavioral findings in the domain of dynamic vision.

Recent models have attempted to account for several of the dynamic neural response properties we investigated here. Delayed normalization (DN) models, for example, divide a neuron's excitatory drive by a filtered and delayed copy of itself, generating neural response time courses that can be fit to observed fMRI (Zhou et al., 2018) or electrocorticography data (Groen et al., 2022; Zhou et al., 2019). This implementation of normalization produces a typical transient then sustained response, with non-linear response dynamics including subadditivity and response adaptation. Compressive spatiotemporal (CST) models, on the other hand, fit separate sustained and transient channels that are passed through a compressive nonlinearity to produce responses (Kim et al., 2024; Kupers et al., 2024). Although CST models have not been tested for the same set of response dynamics, they are well fit to fMRI BOLD time courses in response to sequences of stimuli with different spatial locations and timings, and reproduce increasing temporal window sizes along the visual hierarchy (Murray et al., 2014; Wolff et al., 2022). In both these models, neural responses in each recorded unit (BOLD in each voxel, local field potentials at each electrode site, single-unit firing rates, etc.) are fit independently, meaning the amount of normalization depends only on the activity in that unit. In contrast, D-STAN implements suppressive pooling, both spatially—via summation across model neurons within layers—and temporally—as a consequence of the temporal windows—such that normalization can be induced by activity generated by

other units and at other times. Pooling suppression across neurons is a property key to previous foundational normalization models (Reynolds & Heeger, 2009) and allows for stimuli outside of the "classical receptive field" (i.e., stimuli that do not directly drive a neuron in isolation) to suppress a neuron's response. Therefore, this pooling allows D-STAN to capture a wider variety of phenomena that are not necessarily localized to a single unit, such as backward masking, contrast-dependent suppression, and response adaptation to non-identical stimuli. Additionally, while DN and CST models both produce continuous neural time courses, these are computed based on *a priori* knowledge of the full stimulus sequence. D-STAN, on the other hand, produces layer responses in a recursive manner, with computations implemented "online" at each timestep. The recurrent nature of the model increases its biological plausibility and, as we have demonstrated, the temporal window parameters allow for flexibility in shaping the response dynamics in a way that can be compared to actual neural recordings.

Our simulations showed that both the excitatory and suppressive temporal window were needed to allow the model to generate the full range of effects reported. When both temporal windows are removed, D-STAN reduces to the Normalization Model of Dynamic Attention (Denison et al., 2021). While this model captures tradeoffs in attention and perception across time, it was unable to reproduce any of the temporal phenomena we examined here. While including just one temporal window (i.e., either τ_E or τ_S non-zero) reproduced some phenomena, such as subadditivity and response adaptation, we found that the excitatory temporal window was necessary for backward masking and to produce contrast-dependent suppression backward in time. The suppressive temporal window was necessary to produce transient-sustained dynamics typical of neural firing (Figure 2D), but together both windows interacted to produce a greater variety of response profiles (Figure 1C) that can allow for more flexibility when comparing to real neural data. Thus, allowing both the excitatory and suppressive drives to depend on activity at previous times contributes to D-STAN's capacity to produce this wide array of phenomena.

Spatial normalization across local neuronal populations has been proposed as a canonical computation in the brain that may provide benefits to coding efficiency (Carandini & Heeger, 2012; Louie & Glimcher, 2012), and evidence for it has been found in a variety of cortical regions (Busse et al., 2009; Carandini et al., 1997; Louie et al., 2011; Rabinowitz et al., 2011; Simoncelli & Heeger, 1998; Zoccolan et al., 2005). Does temporal normalization offer similar computational benefits? Normalization is thought to enhance sensitivity to changes in the environment, and this may be true temporally as well. In line with this interpretation, we found that temporal normalization results in transient-sustained dynamics, subadditive neural responses, and response adaptation, for example, demonstrating that consistent inputs act to reduce overall neural activity (Zhou et al., 2019; Zhou et al., 2018). Temporal normalization may also emphasize differences between stimuli across time, like in the contrast-dependent suppression effects we observed, where model d' was increased for high contrast targets among low contrast non-targets. These effects on coding efficiency also depend on how suppression is pooled across a neural population: if suppression more strongly weighed inputs from similarly tuned neurons, for example, this would increase sensitivity to larger changes in stimulus features over time. Conversely, such tuning offers robustness against small perturbations caused by noise and can promote stability of representations over time. Thus, our results suggest that temporal normalization provides computational benefits consistent with broader conceptualizations of normalization.

D-STAN is a flexible model that can be extended in different ways. First, the simulations we conducted used stimuli that varied along only a single feature dimension (orientation), but the model can handle different spatial and feature dimensions by defining the tuning properties of the sensory neurons. Additionally, the suppressive pooling we implemented summed evenly across the entire neural population, but this too is flexible and can be tuned along different dimensions. Previous delayed normalization models, in which suppression was calculated separately for each neural unit (Groen et al., 2022; Zhou et al., 2019), are essentially a special case of this, where the suppressive pool is tuned highly

narrowly for each neuron. Future work should investigate the consequences of varying the degree of feature tuning for the suppressive field. Second, we chose to define the excitatory and suppressive temporal windows using exponential functions, parameterized by a single time constant; however different parameterizations are possible and might be appropriate depending on the types of neural processing in question. Third, although D-STAN is not a biophysical model of neural activity, it may be interesting to examine how the modeled excitatory and suppressive time constants relate to the dynamics of excitatory and inhibitory neurons within different neural circuits. Notably, recent computational work has demonstrated how normalization can be implemented through recurrent excitation and inhibition (Heeger & Zemlianova, 2020), providing a new class of models with which to examine dynamic processing at the level of cortical circuits. Fourth, we only included a single sensory layer in this version of the model, though it is possible to add multiple hierarchical layers with separate temporal windows, which will be necessary to fit data recorded from multiple cortical regions, such as from fMRI or electrophysiology. Neural dynamics may vary between different regions, given that temporal integration windows have been found to increase along the cortical hierarchy (Chaudhuri et al., 2015; Fritsche et al., 2020; Gao et al., 2020; Murray et al., 2014; Vidaurre et al., 2017; Wolff et al., 2022), suggesting that temporal windows may be successively applied and accumulate over several stages of processing. Finally, D-STAN has the capacity to model the dynamics of voluntary and involuntary attention—which were not included in the current simulations—and so a natural extension is to examine how attention interacts with spatiotemporal normalization to influence ongoing processing.

In summary, we introduced the dynamic spatiotemporal attention and normalization model, D-STAN. The key advance of the model is the integration of standard spatial and feature-based neural tuning functions with excitatory and suppressive temporal windows to generate a unified spatiotemporal normalization computation. We show how this computation results in a spatiotemporal receptive field structure comparable to that seen in physiological recordings of sensory neurons. D-STAN also

reproduces several non-linear response properties, including subadditive responses, response adaptation, and backwards masking, as well as contrast-dependent suppression between stimuli across time, a perceptual phenomenon only recently shown in human observers (Epstein & Denison, 2023). Our model provides advances over other dynamic normalization (Denison et al., 2021; Groen et al., 2022; Zhou et al., 2019) or compressive spatiotemporal (Kim et al., 2024; Kupers et al., 2024) models, such as the recursive neural network architecture which allows for continuous online prediction of layer responses, as well as the flexibility afforded by the temporal window and suppressive pooling structures, which allows both spatial and temporal normalization to be carried out via a single, parsimonious spatiotemporal computation. Overall, D-STAN provides an important step toward the goal of developing real-time process models of dynamic vision.

Resource Availability

Lead contact. The lead contact is Angus Chapman (angusc@bu.edu)

Materials availability. This study did not generate new materials.

Data and code availability. Model code and analysis scripts are available on the project Github repository: https://github.com/denisonlab/dstan. Data produced by several of the analyses are available on OSF: https://osf.io/qy9pa/.

Acknowledgements

This work was supported by Boston University startup funds to R.N.D. We thank the members of the Denison Lab for their feedback on the project, particularly Michael Epstein and Karen Tian for helpful discussions.

Author contributions

Conceptualization: A.F.C., R.N.D. Methodology: A.F.C., R.N.D. Software: A.F.C., R.N.D. Formal Analysis: A.F.C. Visualization: A.F.C. Supervision: R.N.D. Writing – Original Draft: A.F.C. Writing – Review & Editing; A.F.C., R.N.D.

Declaration of interests

We declare no competing interests.

STAR Methods

Method details

Model specification. D-STAN is a hierarchical, recurrent neural network, which models the dynamics of feature-tuned neural populations given time-varying sensory input. D-STAN consists of interconnected sensory and decision layers, that each produce time-varying responses in model neurons, allowing the model to generate predictions about neural activity from continuous input in an online fashion (i.e., time step by time step) as well as to generate predictions about behavioral performance in simple perceptual tasks (Figure 1). D-STAN is built on a modeling foundation established by Denison et al. (2021) and Reynolds and Heeger (2009). The introduction of excitatory and suppressive temporal windows together with a spatiotemporal normalization computation allows D-STAN to generate a rich repertoire of dynamic behavior and to exhibit effects of temporal context in line with empirical observations, which could not be generated by these previous models. D-STAN also contains voluntary and involuntary attention layers, following Denison et al. (2021), though these were removed for all analyses in the current study.

Sensory layer. The sensory layer represents the visual processing stage of the model. As in the normalization model of dynamic attention (Denison et al., 2021), this layer receives stimulus input at each time point, which feeds into N = 12 neurons that are tuned to different feature values. We use orientation as an example feature throughout the reported analyses. In simulations, the time course of stimulus input was represented in the matrix $\boldsymbol{X}$ (with size of M orientations x K time points), with each vector $\boldsymbol{X_t}$ = (0, 0, …, *c*, …, 0) indicating the currently presented orientation at contrast level *c*. When no stimulus was presented, all elements of the vector were zero. The stimulus drive to each neuron at each time point depended on the match between the stimulus orientation and the neuron's orientation tuning function, as determined using a raised cosine function:

$$d_{it} = |\cos(\theta_t - \varphi_i)^m| \quad (1)$$

where θ_t is the orientation of the stimulus shown at time *t*, and φ_i is the preferred orientation of the *i*th neuron. Orientation tuning functions were evenly spaced across the feature space, $\varphi_i = \pi(i-1)/\mathrm{N}$. The exponent *m* controls the width of the tuning curve and was set to 23 (*m* = 2N – 1).

Unlike in previous models, the excitatory drive for each neuron is calculated at each time point using a weighted exponential of recent stimulus drive:

$$e_{it} = \sum_{T=0}^{t} d_{iT} \cdot w_T^E \quad (2)$$

$$w_T^E = \frac{1}{\tau_E} e^{\frac{T-t}{\tau_E}} \quad (3)$$

where *t* is the current time point and τ_E is the time constant determining the amount of weight given to previous time points. In effect, this excitatory temporal window imbues the model neurons with a temporal receptive field, where not only are units responsive to their preferred stimulus at any given time point, but also respond based on the stimulus history as well.

The suppressive drive was also calculated at each time point, and weighted by an exponential:

$$s_t = \sum_i \sum_{T=0}^{t} e_{iT} \cdot w_T^S \quad (4)$$

$$w_T^S = \frac{1}{\tau_S} e^{\frac{T-t}{\tau_S}} \quad (5)$$

Thus, a neuron's current response is suppressed by the activity history of itself and its neighbors, which together comprise the suppressive pool. In the reported simulations, only a single spatial location was modeled, and neurons tuned to all orientations were included with equal weight in the suppressive pool. But in principle, the suppressive pool could also be tuned to locations and features.

Finally, the response of each neuron was updated at every time step using the following differential equation, which functions as a dynamic update of the standard normalization equation (Reynolds & Heeger, 2009), as in the normalization model of dynamic attention (Denison et al., 2021):

$$\tau_R \frac{d}{dt} r_i = -r_i + \frac{e_i}{s + \sigma^n} \tag{6}$$

where r_i is the response of the *i*th neuron within the sensory population, τ_R is a time constant that affects the rate at which the neuron's response increases during stimulus presentation and decreases after stimulus offset, e_i is the excitatory drive of that neuron, s is the (pooled) suppressive drive, σ is a semi-saturation constant that affects the contrast gain of neurons and keeps the denominator non-zero, and n is an exponent that affects the shape of neurons' contrast response functions. Some parameter values (τ_R = 52, n = 1.5) were fixed as fitted to empirical data in previous reports (Denison et al., 2021), however we decreased the semi-saturation constant (σ = 0.1) to account for the fact that we only used a single sensory layer in this version of D-STAN.

Decision layer. The decision layer receives input from the sensory layer, which it uses to compute behavioral output regarding the orientation of the stimuli, as in the normalization model of dynamic attention (Denison et al., 2021). Decisions are generated by two neurons, each encoding for one of the two stimuli (T1 and T2). The excitatory drive for each decision neuron is:

$$e_{jt}^D = r_t^S \cdot w_j^D \tag{7}$$

where each vector w_j^D is designed to compute an optimal linear readout from the sensory responses to the *j*th stimulus, to determine whether the stimulus was rotated clockwise vs. counterclockwise from horizontal or vertical. The vectors project the sensory layer response onto the difference between two templates encoding the population responses to the CW and CCW stimuli along a given orientation axis; the axis of each stimulus (vertical or horizontal) is assumed known. Evidence accumulation is positive for

CW decisions, and negative for CCW decisions, such that the sign of the evidence indicates the decoded orientation within the decision layer, while the magnitude of the evidence indicates the strength of the model decision. The suppressive drive is pooled over the decision neurons, and responses are updated at each time point according to the same differential equation as in sensory layers (τ_D = 10^5, σ = 0.7; (Denison et al., 2021). The long time constant of this layer allows for sustained evidence accumulation.

In contrast to previous model iterations (Denison et al., 2021), each neuron in the decision layer accumulates evidence throughout the duration of the simulated trials, rather than during discrete windows following each stimulus presentation. Because the effects of temporal normalization occur when the excitatory and suppressive drives elicited by the two stimuli overlap, the discrete decision windows were unable to capture any behavioral effects of normalization for T1 at short SOAs. At the end of each simulation, the accumulated evidence for each stimulus is converted into *d'*, a measure of perceptual sensitivity, through multiplicative scaling (s_{T1} = s_{T2} = 1×10^5).

Simulation procedures. All simulations were performed in MATLAB (2022b, MathWorks, Natick, MA). We used a time step Δt of 2 ms throughout simulations. The duration of the simulated trial was varied as necessary to capture model dynamics. When a stimulus was presented, we used a standard contrast level of 64% and presentation duration of 30 ms, unless otherwise specified. We varied the two temporal window time constants τ_E and τ_S across simulations to assess how the temporal windows affected sensory layer dynamics and decision layer outputs.

Estimating temporal receptive fields of model neurons. To assess how stimulus input at different past time points affects model responses at the current time, we performed a simulation and analysis based on reverse correlation, similar to the way a temporal receptive field might be measured in a neurophysiology experiment (Mante, 2008). We modified the stimulus input to the model to be a random binary vector, such that the stimulus drive at each time point was one or zero. We then

performed model simulations for 10,000 different random stimulus vectors with a duration of 1200 ms, resulting in variable excitatory and suppressive drives and sensory layer responses that depended on the stimulus history. To estimate the impact of the stimulus input on each model timeseries, we used reverse correlation to calculate the average change in each measure caused by the presence of a stimulus at each time point by correlating the stimulus input at each time point with the response (excitatory/suppressive drive or layer response) at the final time point (Ringach & Shapley, 2004). Shuffled weights were calculated in the same manner after rearranging the stimulus input vectors so that they were not aligned with the responses from the same simulation.

To characterize the temporal receptive fields and enable comparisons across different parameter settings, we fit the estimated stimulus weights using different functional forms. For the excitatory drive, we compared the estimated stimulus weights with the exponential function that defines the excitatory temporal window. This function (τ_E = 400 ms) had no free parameters, but was adjusted using a scaling parameter to align it to the magnitude of the stimulus weights (the units of which are arbitrary), estimated using simple linear regression. For the suppressive drive, we similarly overlaid a scaled function to the stimulus weights. This function, h_S, was determined by convolving the excitatory and suppressive temporal windows, using their respective time constants. We zero-padded the temporal windows at positive (i.e., future) time points to achieve the resulting functional form. Notably, this function in general can also be calculated by taking the difference between the exponential temporal windows (for t < 0 and $\tau_E \neq \tau_S$):

$$h_S = \left| e^{\frac{t}{\tau_E}} - e^{\frac{t}{\tau_S}} \right| \quad (8)$$

The sensory layer response weights were fitted with a difference of Gamma functions, using the simplified Gamma function form adopted in previous work (Zhou et al., 2019):

$$h_R = te^{\frac{t}{\tau_1}} - kte^{\frac{t}{\tau_2}} \quad (9)$$

with time constants τ_1 and τ_2, and weight k. We fitted the simulated stimulus weights to this function in MATLAB by minimizing the least-squares error using *fminsearch* up to a scaling factor. We performed this optimization 100 times, with random initial values for τ_1 and τ_2 drawn uniformly between 0 and 900, and initial *k* = 0, and selected the best fitting function across all solutions. For the function shown in Figure 2C, the best fitting parameters were: $\widehat{\tau_1} = 305.01, \widehat{\tau_2} = 61.98, \hat{k} = 5.43$.

To assess how the effective temporal receptive fields were affected by the excitatory and suppressive temporal parameters, we performed the model simulations again for each combination of τ_E and τ_S from 100 to 900 ms in 100 ms steps. In the results shown in Figure 2D-E, we took the fitted functions for one parameter fixed at 400 ms, while the other parameter varied.

Effects of temporal receptive fields on neural responses. To assess the effects of the excitatory and suppressive time constants on the model neuron responses, we simulated trials in which a single target stimulus was presented. The stimulus presentation duration was set to 2000 ms to allow time to observe the peak and steady state responses, and the total trial duration was set to 8100 ms, including a 500 ms prestimulus period. We varied τ_E and τ_S separately across 9 levels (50, 100, 200, 300, 400, 500, 600, 700, and 800 ms). To examine the effects of the individual time constants in isolation, for simulations varying τ_E, we fixed τ_S at zero, and vice-versa, effectively removing the excitatory or suppressive temporal receptive fields from the model. Setting $\tau_E = \tau_S = 0$ reduces the model to the previous version of Denison et al. (2021). For these simulations, we extracted response time courses from the sensory layer neuron tuned nearest to the target orientation. We also examined the interaction of the time constants by conducting simulations with different combinations of non-zero values for τ_E and τ_S.

For simulations varying the excitatory time constant, all model neurons reached the same stable activity level, and all neural responses were scaled by dividing the response at all time points by the maximum response within the stimulus presentation window. For each value of τ_E, we found the time point at which the model responses reached 99% of the maximum response ("Time to peak", Figure 3B), and the time point at which responses fell to 50% of the peak value following stimulus offset ("Time to half maximum", Figure 3C). We used 99% of the maximum for the time to peak analysis, because responses approached but did not necessarily reach the numerical maximum until late in the window.

For simulations varying the suppressive time constant, we scaled all responses relative to the first peak (Figure 3D). Shorter suppressive time constants typically resulted in a faster reduction in model responses, reaching a peak sooner but also reducing the overall magnitude. To examine response dynamics relative to the peak response, we therefore normalized the response time course by the peak response during the stimulus presentation window. We calculated the time to peak by measuring the time at which the maximum response was reached ("Time to peak", Figure 3E), as well as the model response 2000 ms after stimulus onset relative to the peak ("Stable activity level", Figure 3F).

Subadditive responses. To assess how neural responses depended on stimulus presentation duration, we again simulated trials with only a single target. For the simulations shown in Figure 4A, we used short time constants (τ_E = 100 ms, τ_S = 50 ms) and varied the stimulus duration by doubling from 30-480 ms, for 5 total duration conditions (30, 60, 120, 240, and 480 ms). We extracted responses from the sensory layer in the model neuron tuned closest to the target orientation, and calculated the model response by summing the sensory response across the entire trial duration, approximating the area under the curves in Figure 4A. To quantify subadditivity in neural responses, we calculated the effect of doubling the stimulus duration on the model response by dividing the response at duration 2x by the response at duration x (e.g., we divided the model response at 60 ms by that at 30 ms; Figure 4C). To assess how subadditivity depends on the excitatory and suppressive time constants, we selected one

shorter and one longer value (τ_E = 100 or 500 ms, τ_S = 50 or 500 ms), and computed the proportional response change for each pair of time constants (Figure 4D).

Response adaptation. Response adaptation refers to the finding that neural responses to stimuli presented shortly after an initial stimulus are typically reduced in magnitude (Kohn, 2007; Motter, 2006; Priebe et al., 2002; Solomon & Kohn, 2014; Vogels, 2016). In our analysis, we therefore aimed to examine how model responses differed to a sequence of two "target" stimuli (referred to as T1 and T2) relative to a single stimulus. We performed separate analyses where T1 and T2 were identical stimuli (i.e., the same feature) or distinct (i.e., orthogonal features). In the first simulation examining the interaction between responses to two identical stimuli, our goal was to assess the magnitude of the response to T2 while subtracting out the activity related to T1. Therefore, we first measured the response to T1 alone (Figure 5A, blue dashed line), and then quantified the activity related to T2 by calculating the difference in responses when T2 was present vs. absent (Figure 5A, blue shaded region). We used a longer stimulus presentation duration of 300 ms, which is typical in the response adaptation literature (Vogels, 2016). For the simulation shown in Figure 5A-B, we again selected short time constants (τ_E = 100 ms, τ_S = 50 ms) and used an interstimulus interval (ISI) of 100 ms. We extracted the sensory response in the maximally selective neuron to the stimulus in two simulations: 1) T2 present, 2) T2 absent. The simulation in Figure 5B was identical, except that the ISI was increased to 600 ms.

To quantify the effect of excitatory and suppressive time constants on the magnitude of response adaptation, we varied each of τ_E and τ_S separately across 10 levels (0, 50, 100, 200, 300, 400, 500, 600, 700, and 800 ms) while keeping the other time constant fixed at zero, and assessed the sensory response across a range of ISIs (100 to 1500 ms, in 100 ms steps). To investigate whether temporal windows were required for response adaptation in this model, we also simulated a condition with τ_E and τ_S both equal to zero. For each value of the time constants, we performed simulations where T1 was present or absent. To isolate the response to T2 when T1 was present, we followed

previous work (Lisberger & Movshon, 1999; Priebe et al., 2002) by first subtracting out the activity elicited by a single stimulus (here, when T2 was absent), as follows:

$$AI = 1 - \frac{\sum_t \left(r_{T2present} - r_{T2absent}\right)}{\sum_t r_{T2absent}} \quad (10)$$

such that a higher adaptation index corresponds to a smaller isolated T2 response and thus stronger response adaptation. To examine how the time constants interact to affect response adaptation, we conducted additional simulations using combinations of shorter and longer time constants (τ_E = 100 or 500 ms, τ_S = 50 or 500 ms), and computed suppression indices across the same range of ISIs (Figure 5E).

To assess the presence of response adaptation for non-identical stimuli (Figure 6), we conducted the same simulations, except that T1 and T2 had orthogonal features, generating activity in distinct model neurons. We computed suppression indices for different parameter combinations similarly to when stimuli were identical, but without the need to subtract out the response when T1 was absent, since the model neurons best tuned to T2 no longer responded to T1:

$$AI = 1 - \frac{\sum_t r_{T1present}}{\sum_t r_{T1absent}} \quad (11)$$

Backward masking. Backward masking refers to the phenomena that perception of a stimulus can be impaired ("masked") by a second stimulus presented shortly after the first (Breitmeyer & Öğmen, 2006; Enns & Di Lollo, 2000; Kovacs et al., 1995). We assessed backward masking in the model using a similar method as for response adaptation, except with simulations comparing the sensory response to T1 as a function of whether T2 was present vs absent (i.e., replacing T1 with T2 present/absent in Eq. 11). Again, we first conducted simulations using orthogonal stimulus orientations, with short time constants (τ_E = 100 ms, τ_S = 50 ms) and compared the effects of backward masking for stimulus onset asynchronies (SOA) of 250 ms (Figure 7A) and 500 ms (Figure 7B). We then quantified backward masking by calculating the response summed over time to T1 when T2 was present relative to when it was

absent. We calculated this masking index across the full range of excitatory and suppressive time constants and SOAs (Figures 7C-D). Finally, we conducted additional simulations assessing the interaction of the two time constants on backward masking using combinations of one shorter and one longer time constant (τ_E = 100 or 500 ms, τ_S = 50 or 500 ms), and computed masking indices across the same range of SOAs (Figure 7E).

Contrast-dependent stimulus interactions. We assessed how the contrast of one stimulus affects the response to the other stimulus through temporal normalization. For initial simulations, we used fixed time constants (τ_E = 400 ms, τ_S = 100 ms), with the SOA (250 ms) and stimulus contrasts (64% vs. 16%) chosen to match the previous behavioral findings. We performed simulations independently manipulating the contrast of both T1 and T2 (64% or 16% contrast level), and extracted model performance for each target. We first examined how the model responses to each target at high contrast was affected by the contrast of the non-target (Figures 8A-B). We then computed the model's behavioral discrimination performance (measured as d') from the output of the decision layer, based on recent empirical findings showing that a high- vs. low-contrast non-target stimulus presented before or after a target stimulus can result in reduced perceptual discriminability of the target (Epstein & Denison, 2023). Model responses were scaled to produce d' values closer to behavioral estimates ($s_{T1} = s_{T2} = 1\times10^4$). Model d' was calculated for each target stimulus based on the target and non-target contrast (Figure 8C).

To assess how contrast-dependent suppression for T1 and T2 was affected by the excitatory and suppressive temporal windows, we performed further simulations in which we varied τ_E and τ_S from 0 to 1000 ms in 50 ms steps. For each simulation, we measured the model d' for each target stimulus (T1 and T2) as a function of the contrast of the non-target (NT), as above. We then calculated a contrast-dependent suppression index for each stimulus as:

$$SI = \frac{d_{NTlow} - d_{NThigh}}{d_{NTlow} + d_{NThigh}} \quad (12)$$

To identify values of τ_E and τ_S that produced suppression in both T1 and T2, we calculated a joint suppression index across the two targets by multiplying the suppression indices of T1 and T2 for each parameter combination. This joint index is maximized when a specific combination of τ_E and τ_S results in contrast-dependent suppression in both stimuli, but not when one or both stimuli show little-to-no suppression.

References

Breitmeyer, B. G., & Öğmen, H. (2006). *Visual masking time slices through conscious and unconscious vision* (2nd ed.). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198530671.001.0001

Busse, L., Wade, A. R., & Carandini, M. (2009). Representation of concurrent stimuli by population activity in visual cortex. *Neuron*, *64*(6), 931-942. https://doi.org/10.1016/j.neuron.2009.11.004

Carandini, M., & Heeger, D. J. (2012). Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, *13*, 51-62. https://doi.org/10.1038/nrn3136

Carandini, M., Heeger, D. J., & Movshon, J. A. (1997). Linearity and normalization in simple cells of the macaque primary visual cortex. *Journal of Neuroscience*, *17*(21), 8621-8644. https://doi.org/10.1523/JNEUROSCI.17-21-08621.1997

Cavanagh, S. E., Hunt, L. T., & Kennerley, S. W. (2020). A Diversity of Intrinsic Timescales Underlie Neural Computations. *Frontiers in Neural Circuits*, *14*, 1-18. https://doi.org/10.3389/fncir.2020.615626

Chaudhuri, R., Knoblauch, K., Gariel, M. A., Kennedy, H., & Wang, X. J. (2015). A Large-Scale Circuit Mechanism for Hierarchical Dynamical Processing in the Primate Cortex. *Neuron*, *88*, 419-431. https://doi.org/10.1016/j.neuron.2015.09.008

Denison, R. N., Carrasco, M., & Heeger, D. J. (2021). A dynamic normalization model of temporal attention. *Nature Human Behaviour*, *5*, 1674-1685. https://doi.org/10.1038/s41562-021-01129-1

Enns, J. T., & Di Lollo, V. (2000). What's new in visual masking? *Trends in Cognitive Science*, *4*(9), 345-352. https://doi.org/10.1016/s1364-6613(00)01520-5

Epstein, M. L., & Denison, R. N. (2023). Perceptual sensitivity depends on the contrast of preceding and following stimuli across hundreds of milliseconds. *Journal of Vision*, *23*, 5018. https://doi.org/10.1167/jov.23.9.5018

Fritsche, M., Lawrence, S. J. D., & de Lange, F. P. (2020). Temporal tuning of repetition suppression across the visual cortex. *Journal of Neurophysiology*, *123*, 224-233. https://doi.org/10.1152/jn.00582.2019

Gao, R., Brink, R. L. V. D., Pfeffer, T., & Voytek, B. (2020). Neuronal timescales are functionally dynamic and shaped by cortical microarchitecture. *eLife*, *9*, e61277. https://doi.org/10.7554/eLife.61277

Golesorkhi, M., Gomez-Pilar, J., Zilio, F., Berberian, N., Wolff, A., Yagoub, M. C. E., & Northoff, G. (2021). The brain and its time: intrinsic neural timescales are key for input processing. *Communications Biology*, *4*, 1-16. https://doi.org/10.1038/s42003-021-02483-6

Groen, I. I. A., Piantoni, G., Montenegro, S., Flinker, A., Devore, S., Devinsky, O., Doyle, W., Dugan, P., Friedman, D., Ramsey, N. F., Petridou, N., & Winawer, J. (2022). Temporal Dynamics of Neural Responses in Human Visual Cortex. *Journal of Neuroscience*, *42*, 7562-7580. https://doi.org/10.1523/JNEUROSCI.1812-21.2022

Hasson, U., Yang, E., Vallines, I., Heeger, D. J., & Rubin, N. (2008). A hierarchy of temporal receptive windows in human cortex. *Journal of Neuroscience*, *28*, 2539-2550. https://doi.org/10.1523/JNEUROSCI.5487-07.2008

Heeger, D. J. (1992). Normalization of cell responses in cat striate cortex. *Visual Neuroscience*, *9*, 181-197. https://doi.org/10.1017/S0952523800009640

Heeger, D. J., & Zemlianova, K. O. (2020). A recurrent circuit implements normalization, simulating the dynamics of V1 activity. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(36), 22494-22505. https://doi.org/10.1073/pnas.2005417117

Kim, I., Kupers, E. R., Lerma-Usabiaga, G., & Grill-Spector, K. (2024). Characterizing Spatiotemporal Population Receptive Fields in Human Visual Cortex with fMRI. *Journal of Neuroscience*, *44*(2). https://doi.org/10.1523/JNEUROSCI.0803-23.2023

Kohn, A. (2007). Visual adaptation: Physiology, mechanisms, and functional benefits. *Journal of Neurophysiology*, *97*(5), 3155-3164. https://doi.org/10.1152/jn.00086.2007

Kovacs, G., Vogels, R., & Orban, G. A. (1995). Cortical Correlate of Pattern Backward-Masking. *Proceedings of the National Academy of Sciences of the United States of America*, *92*(12), 5587-5591. https://doi.org/10.1073/pnas.92.12.5587

Kupers, E. R., Kim, I., & Grill-Spector, K. (2024). Rethinking simultaneous suppression in visual cortex via compressive spatiotemporal population receptive fields. *Nature Communications*, *15*(1), 6885. https://doi.org/10.1038/s41467-024-51243-7

Lisberger, S. G., & Movshon, J. A. (1999). Visual motion analysis for pursuit eye movements in area MT of macaque monkeys. *Journal of Neuroscience*, *19*(6), 2224-2246. https://doi.org/10.1523/JNEUROSCI.19-06-02224.1999

Liu, Y., Murray, S. O., & Jagadeesh, B. (2009). Time Course and Stimulus Dependence of Repetition-Induced Response Suppression in Inferotemporal Cortex. *Journal of Neurophysiology*, *101*(1), 418-436. https://doi.org/10.1152/jn.90960.2008

Louie, K., & Glimcher, P. W. (2012). Efficient coding and the neural representation of value. *Year in Cognitive Neuroscience*, *1251*, 13-32. https://doi.org/10.1111/j.1749-6632.2012.06496.x

Louie, K., Grattan, L. E., & Glimcher, P. W. (2011). Reward Value-Based Gain Control: Divisive Normalization in Parietal Cortex. *Journal of Neuroscience*, *31*(29), 10627-10639. https://doi.org/10.1523/Jneurosci.1237-11.2011

Mante, V., Bonin, V., & Carandini, M. (2008). Functional mechanisms shaping lateral geniculate responses to artificial and natural stimuli. *Neuron*, *58*(4), 625-638. https://doi.org/10.1016/j.neuron.2008.03.011

Mauk, M. D., & Buonomano, D. V. (2004). The neural basis of temporal processing. *Annual Review of Neuroscience*, *27*, 307-340. https://doi.org/10.1146/annurev.neuro.27.070203.144247

Motter, B. C. (2006). Modulation of transient and sustained response components of V4 neurons by temporal crowding in flashed stimulus sequences. *Journal of Neuroscience*, *26*(38), 9683-9694. https://doi.org/10.1523/JNEUROSCI.5495-05.2006

Murray, J. D., Bernacchia, A., Freedman, D. J., Romo, R., Wallis, J. D., Cai, X., Padoa-Schioppa, C., Pasternak, T., Seo, H., Lee, D., & Wang, X. J. (2014). A hierarchy of intrinsic timescales across primate cortex. *Nature Neuroscience*, *17*, 1661-1663. https://doi.org/10.1038/nn.3862

Priebe, N. J., Churchland, M. M., & Lisberger, S. G. (2002). Constraints on the source of short-term motion adaptation in macaque area MT. I. The role of input and intrinsic mechanisms. *Journal of Neurophysiology*, *88*(1), 354-369. https://doi.org/10.1152/jn.00852.2001

Priebe, N. J., & Lisberger, S. G. (2002). Constraints on the source of short-term motion adaptation in macaque area MT. II. Tuning of neural circuit mechanisms. *Journal of Neurophysiology*, *88*(1), 370-382. https://doi.org/10.1152/jn.2002.88.1.370

Rabinowitz, N. C., Willmore, B. D., Schnupp, J. W., & King, A. J. (2011). Contrast gain control in auditory cortex. *Neuron*, *70*(6), 1178-1191. https://doi.org/10.1016/j.neuron.2011.04.030

Reynolds, J. H., & Heeger, D. J. (2009). The Normalization Model of Attention. *Neuron*, *61*, 168-185. https://doi.org/10.1016/j.neuron.2009.01.002

Ringach, D., & Shapley, R. (2004). Reverse correlation in neurophysiology. *Cognitive Science*, *28*(2), 147-166. https://doi.org/10.1016/j.cogsci.2003.11.003

Simoncelli, E. P., & Heeger, D. J. (1998). A model of neuronal responses in visual area MT. *Vision Research*, *38*(5), 743-761. https://doi.org/10.1016/s0042-6989(97)00183-1

Solomon, S. G., & Kohn, A. (2014). Moving Sensory Adaptation beyond Suppressive Effects in Single Neurons. *Current Biology*, *24*(20), R1012-R1022. https://doi.org/10.1016/j.cub.2014.09.001

Soltani, A., Murray, J. D., Seo, H., & Lee, D. (2021). Timescales of cognition in the brain. *Current Opinion in Behavioral Sciences*, *41*, 30-37. https://doi.org/10.1016/j.cobeha.2021.03.003

Vidaurre, D., Smith, S. M., & Woolrich, M. W. (2017). Brain network dynamics are hierarchically organized in time. *Proceedings of the National Academy of Sciences of the United States of America*, *114*, 12827-12832. https://doi.org/10.1073/pnas.1705120114

Vogels, R. (2016). Sources of adaptation of inferior temporal cortical responses. *Cortex*, *80*, 185-195. https://doi.org/10.1016/j.cortex.2015.08.024

Wolff, A., Berberian, N., Golesorkhi, M., Gomez-Pilar, J., Zilio, F., & Northoff, G. (2022). Intrinsic neural timescales: temporal integration and segregation. *Trends in Cognitive Sciences*, *26*, 159-173. https://doi.org/10.1016/j.tics.2021.11.007

Yeshurun, Y., Rashal, E., & Tkacz-Domb, S. (2015). Temporal crowding and its interplay with spatial crowding. *Journal of Vision*, *15*(3). https://doi.org/10.1167/15.3.11

Zhou, J., Benson, N. C., Kay, K., & Winawer, J. (2019). Predicting neuronal dynamics with a delayed gain control model. *PLoS Computational Biology*, *15*, 1-27. https://doi.org/10.1371/journal.pcbi.1007484

Zhou, J., Benson, N. C., Kay, K. N., & Winawer, J. (2018). Compressive temporal summation in human visual cortex. *Journal of Neuroscience*, *38*, 691-709. https://doi.org/10.1523/JNEUROSCI.1724-17.2017

Zoccolan, D., Cox, D. D., & DiCarlo, J. J. (2005). Multiple object response normalization in monkey inferotemporal cortex. *Journal of Neuroscience*, *25*(36), 8150-8164. https://doi.org/10.1523/JNEUROSCI.2058-05.2005