Differential Sustained and Transient Temporal Processing Across Visual Streams

Anthony Stigliani[1], Brianna Jeska[1], & Kalanit Grill-Spector[1,2]

[1] Psychology Department, Stanford University, Stanford, CA, 94305, USA

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

ABSTRACT

How does high-level visual cortex process temporal aspects of our visual experience? By modeling neural responses with millisecond precision in separate sustained and transient channels, we predict fMRI responses for stimuli ranging from 33 ms to 20 s. Using this approach, we discovered that lateral temporal regions process visual transients associated with stimulus onsets and offsets but not sustained visual information. That is, they compute moment-to-moment visual transitions. In contrast, ventral temporal regions process both sustained and transient components, with the former exhibiting adaptation and the latter showing larger contributions from stimulus offsets than onsets. Transient responses in ventral regions may reflect rapid processing of new items at stimulus onset and a memory trace when they are no longer visible at stimulus offset. Together, these findings reveal a fundamental temporal mechanism that distinguishes visual streams and underscore the promise of modeling brain responses with millisecond precision to understand neural computations.

How do high-level visual areas encode the temporal characteristics of our visual experience? The temporal sensitivity of early visual areas has been studied with electrophysiology in non-human primates[1-4] and recently using fMRI in humans[5,6]. However, the nature of temporal processing in high-level visual regions remains a mystery for two main reasons. First, the temporal resolution of noninvasive fMRI measurements is in seconds[7], an order of magnitude longer than the timescale of neural processing, which is in the order of tens of milliseconds. Second, while fMRI responses are roughly linear for stimuli lasting 3-10 s[8], responses in visual cortex exhibit nonlinearities both for briefer stimuli, which generate stronger than expected responses[5,6,8-13], as well as for longer stimuli, which get suppressed due to adaptation[14]. Since the general linear model (GLM) for fMRI[8] is inadequate for modeling responses to such stimuli, the temporal processing characteristics of human high-level visual cortex have remained elusive (but see[12,14-17]).

We hypothesized that if nonlinearities are of neural (rather than BOLD) origin, a new approach that predicts fMRI responses by modeling neural nonlinearities can be applied to characterize temporal processing in high-level visual cortex. Different than the GLM, which predicts fMRI signals directly from the stimulus, the encoding approach first models neural responses to the stimulus and from them predicts fMRI responses. Recent studies show that accurately modeling neural responses to brief visual stimuli at millisecond resolution better predicts fMRI responses than the GLM[5,6,18]. In particular, an encoding model with two temporal channels – one sustained and one transient – predicts fMRI responses in early and intermediate visual areas across a wide range of stimuli varying in duration from 33 ms to 30 s[5]. The encoding approach also enables testing a variety of temporal models and quantifying which best predicts brain responses. By building generative models of neural computations, this approach offers a key framework that can provide insights into multiple facets of temporal processing including integration time windows[19-21], channel contributions[5,18,22-25], and the nature of response nonlinearities[5,6,9-12,18].

We considered three hypotheses regarding temporal processing in high-level visual cortex. One possibility is that temporal processing characteristics are similar across high-level visual regions but differ from those of earlier stages of the visual hierarchy. This hypothesis is

based on results from animal electrophysiology showing longer latencies of responses in higher-level regions compared to primary visual cortex, V1[1], as well as research in humans showing longer temporal receptive windows[19,20] and integration times[21] in ventral temporal cortex (VTC) and lateral temporal cortex (LTC) compared to early visual areas. A second possibility is that temporal processing is uniform across high-level regions that process the same stimulus category (e.g. face-selective regions in VTC and LTC) but differs across regions that process different categories (e.g. face- vs. body-selective regions). This prediction is based on data showing a larger sustained response to long-duration (21 s) images in face- than place-selective regions in VTC[14], and that body-selective regions (but not other category-selective regions) maintain strong responses to fast 8 Hz visual stimulation[15]. These studies suggest that face-selective regions may show larger sustained responses than other high-level regions, and body-selective regions may show larger transient responses. A third possibility is that temporal processing differs across ventral and lateral visual streams rather than across domains. A large body of literature has documented that regions in LTC along the superior temporal sulcus (STS) show heighted responses to biological motion compared to stationary stimuli and other types of motion[26-33], unlike regions in VTC that are not critical for perceiving biological motion[30,31,34]. This predicts that lateral regions may show larger transient responses than ventral regions, which instead may show larger sustained responses.

To test these predictions, we measured fMRI responses in high-level visual areas to images of faces, bodies, and words that were either sustained (one continuous image per trial, durations ranging from 3–20 s) (**Fig. 1a-b**, *experiment 1*), transient [30 flashed, 33 ms long images per trial with interstimulus intervals (ISIs) ranging from 67–633 ms] (**Fig. 1a-b**, *experiment 2*), or intermediate in duration (30 semi-continuous images per trial, durations ranging from 67–633 ms per image with 33-ms ISIs) (**Fig. 1a-b**, *experiment 3*). We also collected a separate functional localizer experiment to independently define regions selective to faces and bodies in VTC and LTC (**Fig. 1c**; Methods). We used face- and body-selective regions as a model system as there are multiple clusters of these regions across the temporal lobe, and face and body regions neighbor on the cortical sheet[35]. This organization enabled us to (i) measure how the temporal dynamics of stimuli affect responses in each region and (ii) test if temporal processing characteristics vary

across regions selective to different categories (e.g., faces or bodies) or across regions in different anatomical locations (e.g., ventral or lateral temporal cortex).

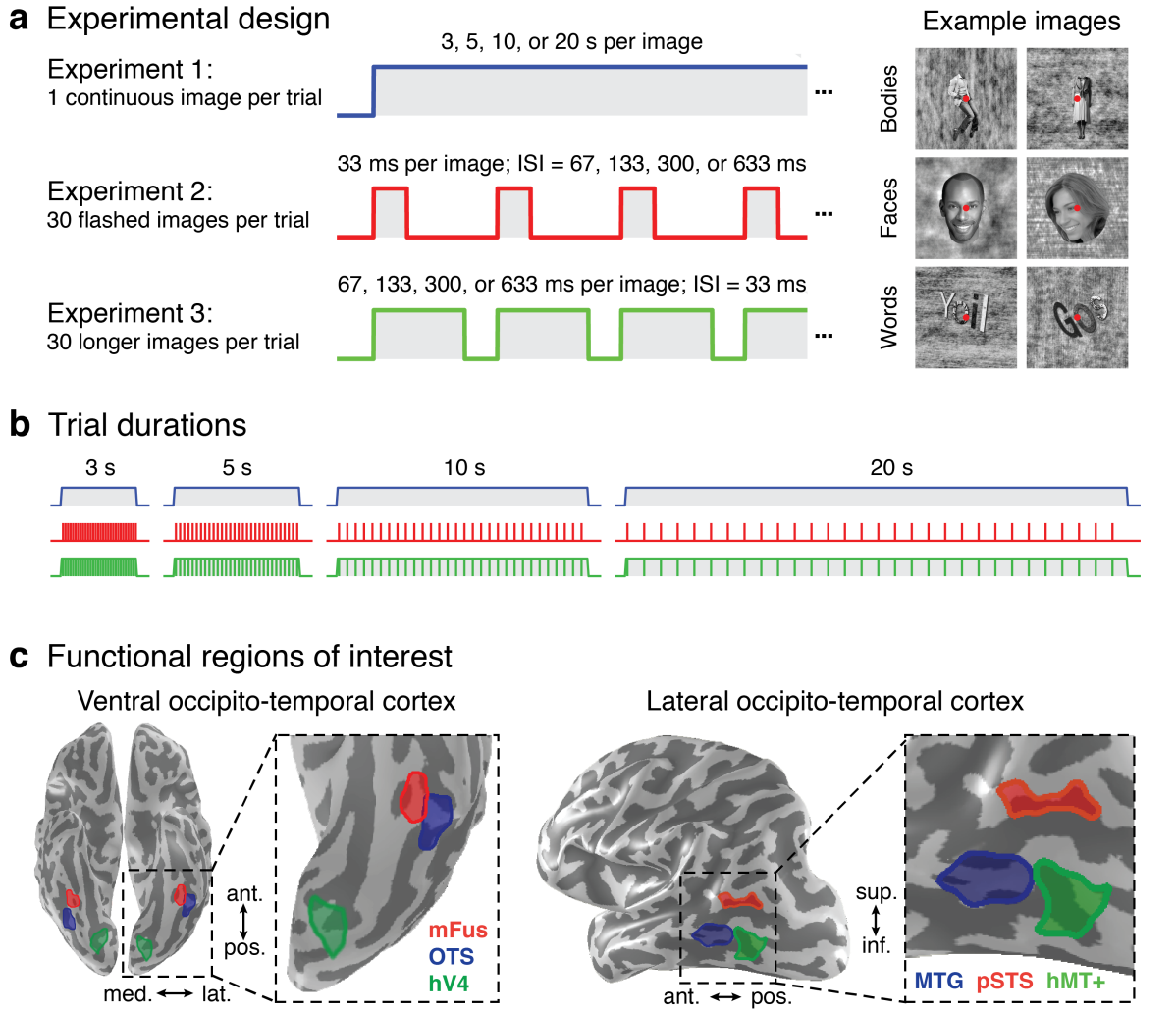

Figure 1. Measuring brain responses to combinations of sustained and transient visual stimuli in high-level visual cortex. (**a**) Participants fixated centrally and viewed images of bodies, faces, and pseudowords (*right*) that were presented in trials of different durations interleaved with 12-s periods of a blank screen (*left*). *Experiment 1:* a single image was shown for the duration of a trial. *Experiment 2:* 30 briefly presented images from the same category (33 ms each), each followed by a blank screen, were presented in each trial. As the trial duration lengthens, the gap between images increases, causing the fraction of the trial containing visual stimulation to decline. *Experiment 3:* 30 semi-continuous images from the same category were presented in each trial with a constant 33-ms blank screen between consecutive images. As the block duration lengthens, the duration of each image progressively increases but the gap does not. (**b**) The same trial durations (3, 5, 10, or 20 s) were utilized across all three experiments, while the rate and duration of visual presentation varied between experiments. Corresponding trials in experiments 1 and 3 have almost the same overall duration of stimulation but different numbers of stimuli, whereas trials in experiments 2 and 3 have the same number of stimuli but different durations of stimulation. The same fixation task was used in the three main experiments. (**c**) Functional regions of interest in ventral temporal cortex (*left*) and lateral temporal cortex (*right*) selective to bodies (OTS and MTG) and faces (IOG and mFus), as well as human V4 (hV4) and human motion-sensitive area (hMT+). Regions in each anatomical section are shown in an example subject.

RESULTS

Responses in High-Level Visual Cortex Exhibit Temporal Nonlinearities

To assess the feasibility of our approach, we first used a standard widely-used GLM[8] to predict fMRI responses in the three main experiments. Then, we compared these predictions to measured fMRI responses from two sample functional regions of interest: a ventral body-selective region and a lateral body-selective region.

In general, the GLM predicts longer responses for longer trials and similar responses in experiments 1 and 3 (**Fig. 2a**, *blue* and *green*). Responses in experiment 1 are predicted to be slightly higher than in experiment 3 because the 33-ms gaps between images in the latter experiment make up 1 s of baseline within each trial. Due to the nature of the hemodynamic response function (HRF), the GLM also predicts that peak response amplitudes in experiments 1 and 3 will increase gradually from 3-s to 10-s trials and subsequently plateau for longer trial durations. In contrast, this model predicts substantially lower responses in experiment 2 compared to the other experiments because the transient 33-ms stimuli in this experiment comprise only a small fraction of each trial duration (**Fig. 2a**, *red*). Therefore, the GLM predicts a progressive decrease in response amplitude from 3-s to 20-s trials in experiment 2, as the fraction of the trial in which stimuli are presented decreases (from 1/3 to 1/20 of the trial).

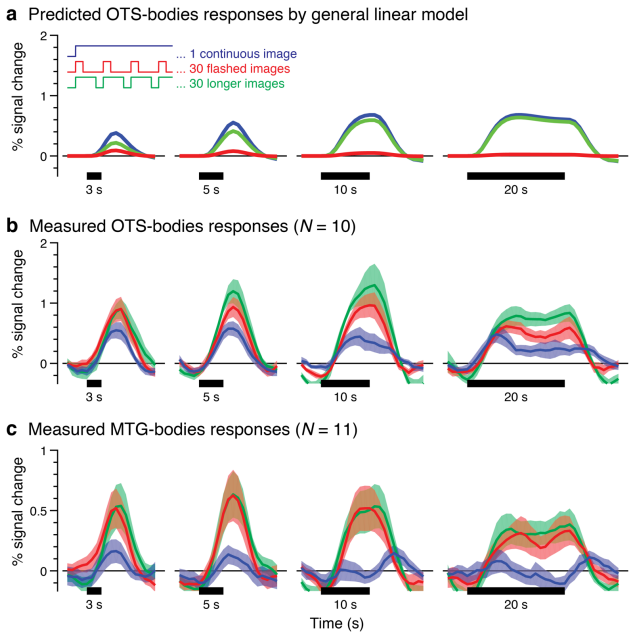

Figure 2. Responses of body-selective regions in ventral and lateral temporal cortex exhibit nonlinearities that are not predicted by a linear model. (**a**) Predicted responses by a GLM for trials containing one continuous image (*blue*), thirty flashed (33 ms) images (*red*), and thirty longer images that span then entire trial duration except for a 33 ms interstimulus interval (ISI) following each image (*green*). Predictors are fit to OTS-bodies responses using data concatenated across all three experiments shown in (b). (**b**) Measured responses in a ventral region on the occipitotemporal sulcus (OTS) selective to bodies (OTS-bodies) during the three experiments. (**c**) Measured responses in a lateral region on the middle temporal gyrus (MTG) selective for bodies (MTG-bodies) during the three experiments. In (b-c), *lines:* mean response time series across participants for trials with body images; *shaded areas:* standard error of the mean (SEM) across participants; *Horizontal black bars:* stimulus duration.

Strikingly, responses to body images in a ventral body-selective region (OTS-bodies; **Fig. 2b**) and a lateral body selective region (MTG-bodies; **Fig. 2c**) both deviate from the predictions of the GLM, but in different ways.

In contrast to the predictions of the GLM, responses in OTS-bodies to trials of 30 flashed images in experiment 2 (**Fig. 2b**, *red*) are substantially higher than in corresponding trial durations in experiment 1, when one stimulus is shown per trial (**Fig. 2b**, *blue*). This occurs despite the fact that stimuli are presented for only a small fraction of each trial duration in experiment 2 compared to experiment 1. Furthermore, peak response amplitudes do not increase with trial duration in experiment 1 as predicted the GLM. Instead, we observe a systematic decrease in response after the first few seconds of stimulation in the 10-s and 20-s trials, which is consistent with prior reports of fMRI adaptation for prolonged stimuli in nearby face- and place-selective regions[14]. Lastly, responses in experiment 3 (**Fig. 2b**, *green*) exceed responses in both experiment 1 (which has only one image per trial but similar overall durations of stimulation) and experiment 2 (which has the same number of images per trial but shorter stimulus durations). This observation suggests that both the number of stimuli in a trial and their duration impact response amplitudes, as in earlier visual areas such as V1 and hV4 (**Fig. S1**).

Unlike OTS-bodies, MTG-bodies illustrates a largely transient response profile with substantially lower responses to the prolonged single images in experiment 1. Notably, for the 10-s and 20-s trials, we observe a transient response following both the onset and the offset of the image but no elevation of response in the middle of the trial (**Fig. 2c**, *blue*). In contrast to the lack of robust responses in experiment 1, MTG-bodies shows surprisingly large responses to briefly flashed stimuli in experiment 2 (**Fig. 2c**, *red*). Additionally, responses in MTG-bodies during experiment 2 (**Fig. 2c**, *red*) and experiment 3 (**Fig. 2c**, *green*), which have the same number of stimuli per trial but different stimulus durations, are similar and both exceed responses in experiment 1. This suggests that, unlike ventral regions, stimulus duration has little impact on MTG-bodies responses, which resemble responses in neighboring motion-sensitive hMT+ (**Fig. S1**).

These data demonstrate that (i) varying the temporal characteristics of visual presentations in the millisecond range has a profound effect on fMRI responses in visual cortex,

(ii) the standard GLM is inadequate for predicting measured fMRI responses to these stimuli in high-level regions, in agreement with prior data in earlier visual areas[5,6,8-13], and (iii) even though OTS-bodies and MTG-bodies prefer the same stimulus category, their temporal response characteristics vastly differ.

An Encoding Model of Temporal Processing in High-Level Visual Cortex

Motivated by the recent success of encoding models that predict fMRI responses in earlier visual areas by modeling neural temporal nonlinearities[5,6,18], we applied a similar approach to predict responses in high-level visual areas. Different than the GLM, the encoding approach first models the neural response in millisecond resolution and then convolves the estimated neural response with a HRF to predict fMRI responses (**Fig. 3**).

Our A+S model consists of two temporal channels[5,18] – a sustained channel and a transient channel – each of which can be modeled using a neural temporal impulse response function (IRF)[2,3,5,18,36] followed by a nonlinearity. The sustained channel is modeled with a monophasic IRF (**Fig. 3b**, *blue channel IRF*), which predicts a sustained neural response for the duration of the stimulus. To capture the gradual decay (adaptation, A) of response observed in ventral regions for sustained images (**Fig. 2b**, *blue*), we apply a nonlinearity to the sustained channel in form of an exponential decay function (Methods). The transient channel is characterized by a biphasic IRF (**Fig. 3b**, *red channel IRF*) that identifies changes to the visual input. That is, it acts like a derivative function, predicting no further increase in the neural response once a stimulus has been presented for longer than the duration of the IRF[5,18]. This channel too has a nonlinearity, as we hypothesize an increase in neural response at both the appearance (onset) and disappearance (offset) of a stimulus. To account for the pronounced transient responses in higher-level visual areas (**Fig. S1**), we apply a flexible compressive nonlinearity on the transient channel using a pair of sigmoid (S) functions, one for the onset and another for the offset (Methods). The predicted fMRI response is generated by convolving the neural response predictors for each channel with the HRF and summing the responses of the two temporal channels (**Fig. 3c**). Since the HRF acts as a low-pass filter, predicted fMRI responses can be downsampled with minimal distortion to match the slower sampling rate of fMRI measurements.

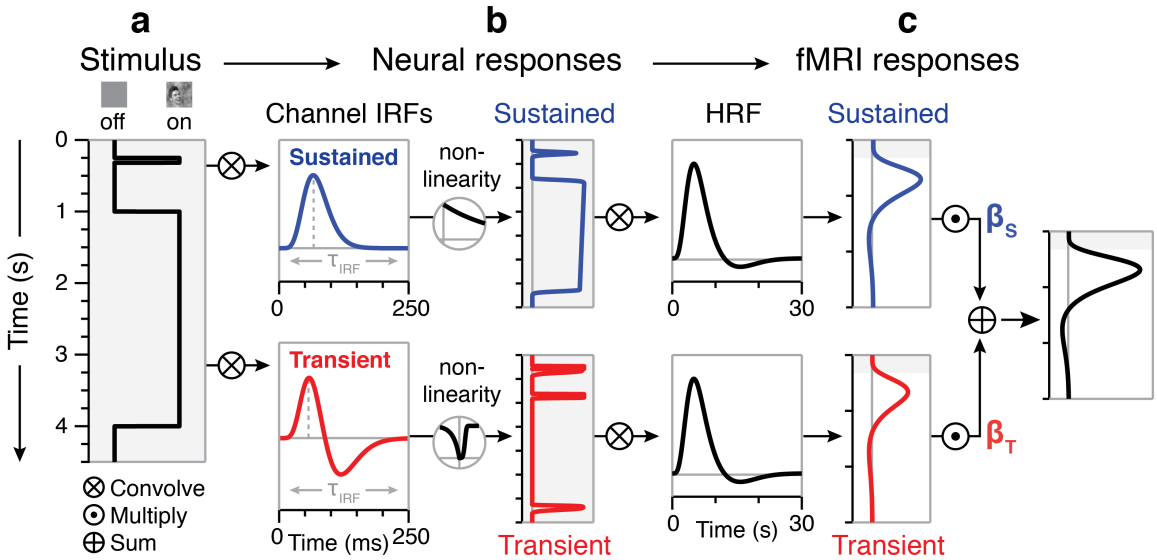

Figure 3. **Optimized two-temporal channel A+S model with adaptation and sigmoid nonlinearities**. (**a**) Transitions between stimulus and baseline screens are coded as a step function representing when a stimulus was on vs. off with millisecond temporal resolution. (**b**) Separate neural responses for the sustained (*blue*) and transient (*red*) channels are modeled by convolving the stimulus vector with an IRF for each channel. An exponential decay function is applied to the sustained channel to model response decrements related to neural adaptation, and a compressive sigmoid nonlinearity is applied to the transient channel to vary the temporal characteristics of "on" and "off" responses (Methods). (**c**) Predictors of sustained and transient fMRI responses are generated by convolving each channel's neural response predictors with the HRF and down-sampling to match the sampling rate of measured fMRI data. The total fMRI response is the sum of the weighted sustained and transient fMRI predictors for each channel. To optimize model parameters and estimate the contributions (β weights) of the sustained (β_S) and transient (β_T) channels, we fit the model to different splits of the data including runs from all three experiments.

We estimated optimized A+S model parameters separately for each participant and region using nonlinear programming and a cross-validation approach. In our procedure, we use half the data from all three experiments to estimate model parameters. Specifically, we estimate a time constant for the neural IRFs (τ), a time constant controlling adaptation of sustained responses (α), and three parameters controlling compression of transient responses (k_{on}, k_{off}, and λ). After optimizing these parameters, we use a GLM to estimate the magnitude of response (β weight) for each channel and stimulus category in our experiments, resulting in three β weights for the sustained channel (one β_S for each category) and three β weights for the transient channel (one β_T for each category). These parameters and weights are then used to predict responses in left-out data and evaluate the model's goodness-of-fit (cross-validated variance explained, x-R^2).

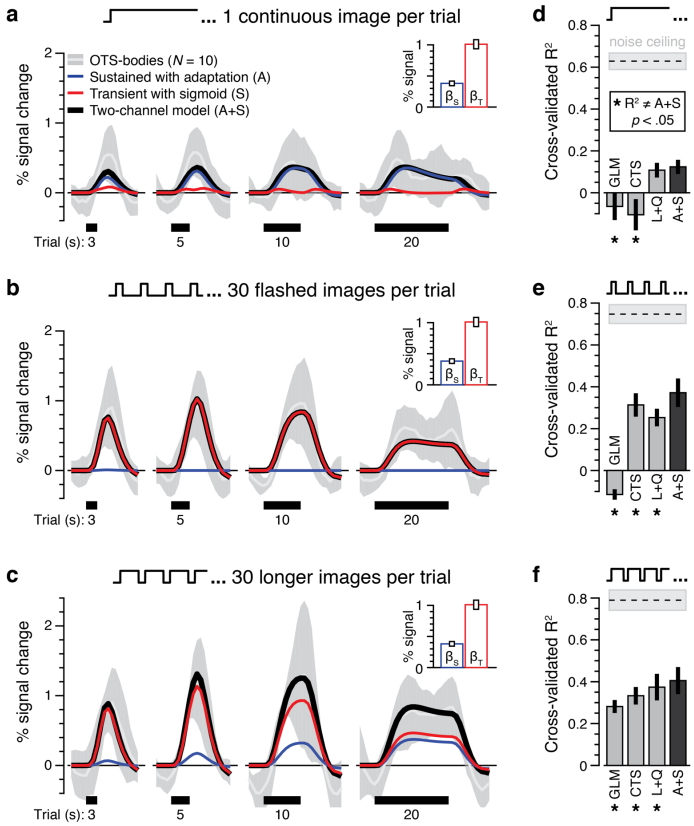

Figure 4. Two-temporal channel model with nonlinearities on both sustained and transient channels predicts responses in ventral temporal cortex. (a-c) Responses and model predictions for body images in OTS-bodies. *White curve:* mean response across 10 participants. *Shaded gray:* standard deviation across participants. *Blue:* predicted response from the sustained channel. *Red:* predicted response from the transient channel: *Black:* sum of responses from both channels. *Inset: mean* contribution (β weight) for each channel ±1 SEM across participants. (**a**) Experiment 1 data, 1 continuous image per trial. (**b**) Experiment 2 data, 30 flashed images per trial. (**c**) Experiment 3 data, 30 longer images per trial. (d-f) Model comparison. Bars show the performance of various models for each experiment presented in (a-c). Models are fit using runs from all three experiments, and cross-validation performance (x-R^2) is calculated in left-out data from each experiment separately. (**d**) Experiment 1. (**e**) Experiment 2. (**f**) Experiment 3. Single-channel models: G*LM,* general linear model[8]; *CTS,* a sustained channel with compressive temporal summation[6]. Dual-channel models: *L+Q,* a linear sustained channel and a transient channel with quadratic nonlinearity[5]; *A+S:* a sustained channel with adaptation and a transient channel with sigmoid nonlinearities. Asterisks denote models with significantly different performance compared to A+S (paired *t*-tests comparing x-R^2 of each model vs. A+S in each experiment).

Comparing the predictions of our optimized A+S model with measured fMRI responses in high-level visual cortex reveals two notable findings. First, our model generates signals that closely track the amplitude of fMRI responses in all three experiments. Second, analysis of x-R^2 shows that our optimized A+S model consistently outperforms other optimized encoding models.

We illustrate these results for one region, OTS-bodies (**Figs. 4**, **S2**); for other regions see **Figs. S3-S5**. Notably, the A+S model closely tracks response amplitudes in all three experiments [**Fig. 4a-c**, compare overall model prediction (*black*) with measured data from OTS-bodies (*gray*)]. Consistent with our predictions, the sustained channel accounts for the bulk of the response in experiment 1 (**Fig. 4a**, *blue*); The transient channel contributes most of the response in experiment 2 (**Fig. 4b**, *red*), and both channels contribute to responses in experiment 3 (**Fig. 4c**).

We compared the performance of our A+S model to the GLM[8], the balloon model[7], four single channel models (L, CTS[6], A, S; **Fig. S3a**), and three alternative two-channel models (L+Q[5,18],

C+Q, A+Q) across all three experiments (Methods; **Figs. S3-S5**). For simplicity, **Fig. 4d-f** compares performance in OTS-bodies for our model vs. three others: the GLM[8], a single-channel model with compressive temporal summation (CTS)[6], and a two-channel model composed of a linear sustained channel and a transient channel with a quadratic nonlinearity (L+Q)[5,18]. Notably, the A+S model predicts OTS-bodies responses in left out data significantly better than the GLM[8], which overestimates responses in experiment 1 and underestimates responses in experiment 2 (GLM vs. A+S: all *t*s > 2.64, *P*s < .05, paired *t*-tests on x-R^2 separately for each experiments) (**Fig. S2a**). The A+S model also outperforms the recently proposed CTS[6] model that enhances early and late portions of the neural response to a stimulus (CTS vs. A+S: all *t*s > 2.60, *P*s < .05). While the CTS model performs considerably better than the GLM in experiment 2, it overestimates responses in experiment 1 with a single continuous images and underestimates responses in experiment 3 with 30 longer images per trial (**Fig. S2b**). In experiments 2 and 3, we also observe a significant advantage of the A+S model compared to the two-temporal channel L+Q model[5,18], which underestimates the large responses to transient stimuli in experiment 2 (**Fig. S2c**) (L+Q vs. A+S: *t*s > 4.06, *P*s < .05; the difference fell short of significance for experiment 1, t_9 = 1.98, *P* = .08).

Thus, an optimized two-temporal channel model with an adaptation nonlinearity in the sustained channel and compressive sigmoid nonlinearities in the transient channel predicts fMRI responses to visual stimuli ranging from milliseconds to seconds in high-level visual cortex with greater accuracy than alternative models.

How do Channel Contributions Differ Across Ventral and Lateral Category-Selective Regions?

Examination of response time series (**Fig. S1**) and channel weights (**Fig. 5**) in body- and face-selective regions in VTC and LTC reveals prominent differences across visual streams.

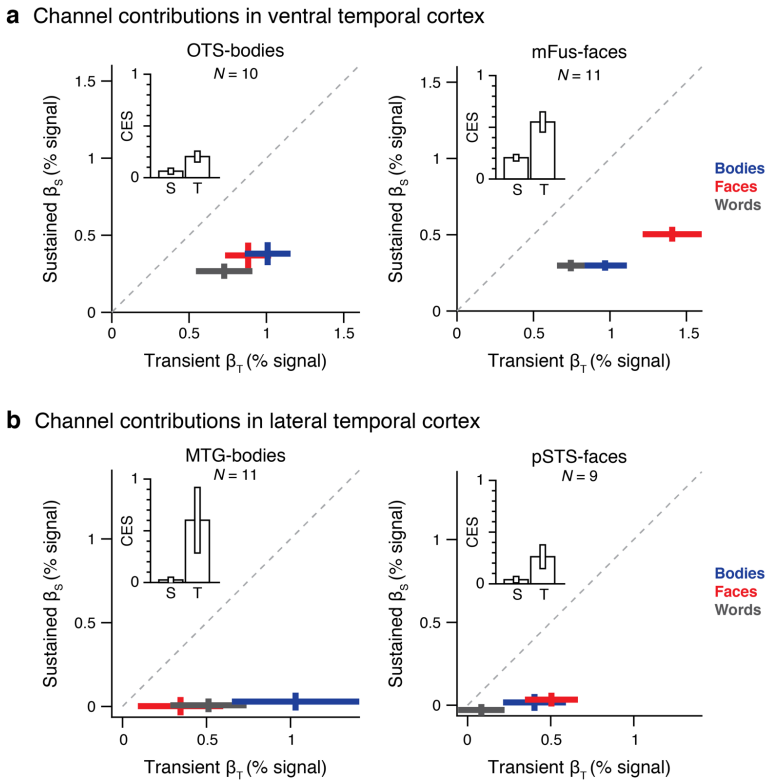

Figure 5. Differential contributions of transient and sustained temporal channels across ventral and lateral regions selective to face and body stimuli. (**a**) Contributions (β weights) of transient (*x* axis) and sustained (*y* axis) channels for each stimulus category estimated by the two-temporal channel A+S model in ventral regions OTS-bodies and mFus-faces. Crosses span ±1 SEM across participants in each axis, and β weights were solved by fitting the model using split halves of the data including runs from all three experiments. Data show average model weights across both splits of the data for each participant. *Red:* response to faces. *Blue:* response to bodies. *Gray:* response to words. *Dashed line*: identity line ($\beta_S = \beta_T$). *Inset*: bars indicate smean contrast effect size (CES) of β weights for the preferred vs. nonpreferred categories in each channel ±1 SEM across participants. (**b**) Same as (a) but for lateral regions MTG-bodies and pSTS-faces.

First, comparing the response time courses of different category-selective regions shows that ventral temporal regions (OTS-bodies and mFus-faces) respond strongly to both the sustained stimuli in experiment 1 and the transient stimuli in experiment 2, whereas lateral temporal regions (MTG-bodies and pSTS-faces) respond strongly to the transient stimuli but minimally to the sustained stimuli (**Fig. S1**). The ratio of sustained and transient channel amplitudes, $\left|\frac{\beta_S}{\beta_T}\right|$, also differs across regions in ventral and lateral aspects of temporal cortex [significant main effect of processing stream, $F_{1,\,107} = 14.27$, $P < .01$, three-way ANOVA with factors of processing stream (ventral/lateral), stimulus category (faces/bodies/words), and preferred category (bodies/faces)]. That is, while both sustained and transient channels contribute to responses in ventral regions (**Fig. 5a**), the transient channel dominates responses in lateral regions (**Fig. 5b**). In fact, zeroing the contribution of the sustained channel slightly improves model performance in lateral regions only (i.e. x-R^2 of the S model is marginally better than the A+S model in hMT+, MTG-bodies, and pSTS-faces; **Fig. S3b**). Moreover, ventral regions

show a characteristic similar to both hV4 (**Fig. S4a**) and occipital category-selective regions (**Fig. S5a**), whereas lateral regions show a characteristic similar to motion-sensitive hMT+ (**Fig. S4a**).

Second, in VTC (OTS-bodies and mFus-faces; **Fig. 5a**), category selectivity – or higher responses to a preferred category vs. other categories – is evident in both sustained and transient channels [all *t*s > 2.26, *P*s < .05, one-tailed *t*-tests comparing the contrast effect size (CES) of β weights for the preferred vs. nonpreferred categories separately for each channel; **Fig. 5a**, *insets*]. That is, responses to faces in mFus are significantly higher than the average responses to words and bodies in both channels (*t*s > 5.25, *P*s < .001, paired *t*-test for each channel; **Fig. 5a**, *right*), and responses to bodies in OTS are higher than average responses to other categories in both channels (*t*s > 2.59, *P*s < .05, paired *t*-test for each channel; **Fig. 5a**, *left*). In contrast, in LTC (MTG-bodies and pSTS-faces; **Fig. 5b**), there is a significant difference in the CES across sustained and transient channels [significant main effect of channel, $F_{1, 8} = 14.88$, $P < .01$, two-way ANOVA with factors of channel (sustained/transient) and preferred category (bodies/faces)]. That is, higher responses to the preferred category are observed only in the transient channel (*t*s > 1.99, *P*s < .05; the effect was not significant in the sustained channel, *t*s < 1.37, *P*s > .10; **Fig. 5b**, *insets*). Thus, these results reveal differential contributions of transient and sustained channels across ventral and lateral category-selective regions.

How do Timing Parameters Vary Across Ventral and Lateral Face and Body Regions?

The parameters in our A+S model were optimized separately for each region within each participant. Thus, for a given ROI, we optimized one time constant for the channel IRFs, one time constant for the adaptation decay function, as well as three sigmoid parameters (Methods). Then, we examined the optimized timing and compression parameters for the sustained and transient channels separately.

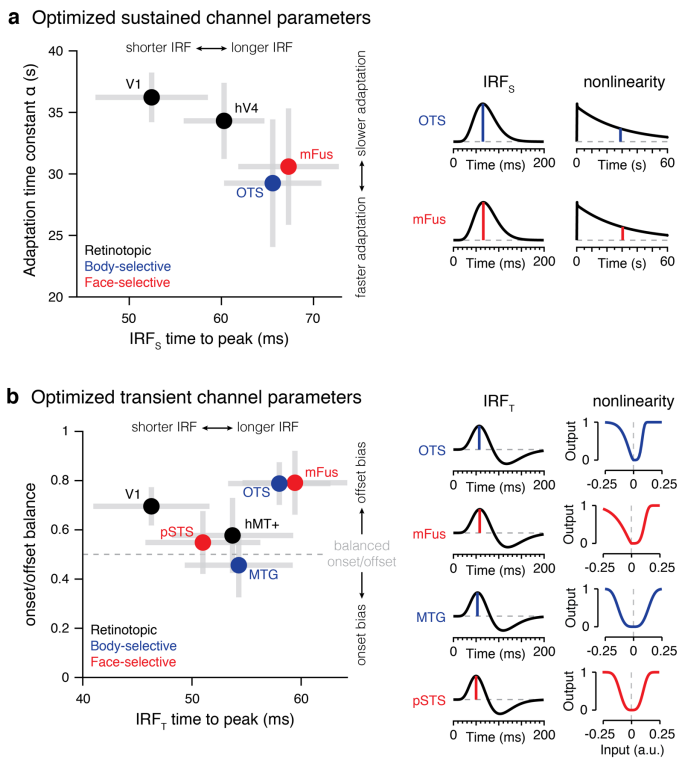

Figure 6. Optimized two-temporal channel model parameters differ across visual cortex. (**a**) Optimized sustained channel parameters. Time to peak of sustained IRF_S (*x* axis) and exponential time constant of the adaptation function (*y* axis) for each region estimated by the two-temporal channel A+S model. Crosses span ±1 SEM across participants in each axis, and parameters were optimized using split halves of the data containing runs from all experiments. Data show model parameters averaged across both splits of the data for each participant. (**b**) Optimized transient channel parameters. Time to peak of transient IRF_T (*x* axis) and onset/offset balance (*y* axis) for each region estimated by the two-temporal channel A+S model (with a zeroed sustained channel in lateral regions). The onset/offset balance metric captures differences in the shapes of the sigmoid nonlienarities used to compress transient "on" and "off" responses, where values larger than 0.5 refect elongation of offset responses compared to onset responses. Crosses span ±1 SEM across participants in each axis, and parameters were optimized using split halves of the data from all experiments. Plots show average model parameters across all splits of the data for each participant. Sample IRFs and nonlinearities shown to the right of (a-b) are generated by averaging optimized model parameters across participants.

For the sustained channel, we assessed how the adaptation decay constant (α) and the time to peak of the neural IRF (IRF_S) vary across ventral regions, omitting lateral regions that did not have significant sustained responses. We discovered a hierarchical progression of longer time to peak and stronger adaptation in the sustained channel ascending from early to later stages of the ventral hierarchy (**Fig. 6a**). That is, the time to peak of the sustained IRF tended to be shorter in V1 than hV4 and shorter in hV4 than in ventral regions OTS-bodies and mFus-faces (**Fig. 6a**, *x axis*). In V1 the mean time to peak of IRF_S was 52±7 ms, which was 14±4 ms faster than in mFus-faces and OTS-bodies. At the same time, the adaptation decay constant gradually decreased going from V1 to ventral regions, indicating more adaptation in mFus-faces and OTS-bodies than in early and intermediate retinotopic regions (**Fig. 6a**, *y axis*). Thus, analyzing timing parameters in the sustained channel revealed two dimensions of hierarchical processing across the ventral stream: time to peak of the neural IRF and decay of adapted responses.

In the transient channel, we examined how the time to peak of the IRF_T varies across regions and if there are asymmetries in the compression of "on" compared to "off" neural responses controlled by the sigmoid shape parameters k_{on} and k_{off}, respectively. Since lower k

values generally elongate transient responses, the relative contribution of the offset component can be indexed by a balance metric, $\frac{k_{on}}{k_{on}+k_{off}}$, where a ratio of 0.5 indicates equal contributions from the onset and the offest of a stimulus to BOLD signals (k_{on}=k_{off}). A ratio < 0.5 indicates a larger contribution of onset than offset responses, and a ratio > 0.5 indicates a larger contribution of offset than onset responses. Like the sustained channel, the transient channel also shows an increase in the time to peak of IRF_T going from V1 to face- and body-selective regions in VTC and LTC (**Fig. 6b**, *x axis*). Interestingly, transients in lateral regions, pSTS-faces and MTG-bodies, show balanced contributions of onset and offset responses (balance metric = 0.50±0.09; **Fig. 6b**, *y axis* and *insets*). In contrast, transients in ventral regions, mFus-faces and OTS-bodies, are dominated by offset responses (balance metric = 0.79±0.08; **Fig. 6b** and *insets*). The surprisingly large offset contribution in VTC indicates that the bulk of the response for the brief stimuli in experiment 2 can be attributed to neural responses that occur after the stimuli are no longer visible, rather than during the initial response to these stimuli.

Thus, comparison of optimized A+S model parameters reveals functional differences between early and later stages of the visual hierarchy, as well as distinct nonlinearities across ventral and lateral regions with the same category preference.

DISCUSSION

Using a temporal encoding approach to explain responses in high-level visual regions, we discovered that an optimized two-temporal channel model consisting of a sustained channel with an adaptation nonlinearity and a transient channel with compressive sigmoid nonlinearities successfully predicts fMRI responses in human high-level visual cortex for stimuli presented for durations ranging from tens of milliseconds to tens of seconds. Critically, the innovative temporal encoding framework we introduce combines in a single computational model several components of temporal processing including time windows of temporal integration[12,16,19-21], channel contributions[5,18,22-25], and nonlinearities in temporal summation[5,6,9-12,18]. Using this approach, we (i) uncover the temporal sensitivity of neural responses in human high-level visual cortex, (ii) find differential temporal characteristics across lateral and ventral category-selective

regions, and (iii) propose a new mechanism – temporal processing – that functionally distinguishes visual processing streams in the human brain.

Differences in Temporal Processing Across Visual Streams

Our results suggest two key differences between temporal processing in the ventral and lateral visual processing streams. First, there are differences in channel contributions. The lateral stream is dominated by responses to visual transients, while the ventral stream responds to both sustained and transient visual information. Transient processing in the lateral stream is consistent the view that lateral regions (including face and body-selective regions in the STS and MTG, respectively) are involved in processing dynamic visual information[26-33]. However, different than prior theories that have implicated these regions in processing biological motion[27-29,37], our data suggest that there is a more fundamental difference between high-level regions in lateral and ventral streams that is driven by differential channel contributions. Second, there are also differences in the dynamics of transient processing across visual streams. Lateral regions show equal increases in neural responses due to the onset and offset of a visual stimulus, suggesting they carry information about moment-to-moment changes in the visual input. However, ventral regions exhibit surprisingly asymmetric contributions from the onset and offset of the stimulus. That is, the accumulation of fMRI responses due to the termination of a stimulus is more pronounced than responses associated with its onset. This difference suggests the intriguing possibility that transient responses in the lateral stream code progressive changes to the visual input, while transient offset responses in the ventral stream may reflect memory traces that are maintained in high-level regions after a stimulus is no longer visible. This prediction is consistent with results from ECoG studies showing that high frequency broadband responses (>60 Hz) in VTC continue for 100-200 ms after the stimulus is off[38-41] and carry stimulus-specific information that may be modulated by attention[40,41].

Observing a strong transient response in lateral regions, MTG-bodies and pSTS-faces, is interesting in the context of classic theories that propose differential contributions of magnocellular (M) and parvocellular (P) inputs to parallel visual streams in the primate visual system[22-24]. The M pathway is thought to code transient visual information and projects to MT,

while the P pathway is thought to code sustained information and projects to V4 and the ventral stream. Our results reveal that the transient channel dominates responses not just in hMT+[5], as suggested by classic theories[22-24], but also in surrounding face- and body-selective regions in LTC.

Different from the predictions of classic theories of a predominant P input to the ventral stream, we find significant contributions from both transient and sustained channels in ventral regions as well as category selectivity in both channels. Surprisingly, transient responses in VTC appear to be larger than sustained responses, which is consistent with later studies reporting that both channels contribute to responses in ventral regions like V4[5,25]. We note that interpreting the relative amplitude of multiple channels is complex, as it depends on the specific implementation of the model and the experimental design. However, we are confident that there are both sustained and transient responses in VTC for two reasons. First, independent of our model, ventral regions respond strongly both to sustained single images (experiment 1) and trains of briefly flashed images (experiment 2). Second, responses in experiment 3, which had combinations of sustained and transient stimulation, exceed those of either experiment 1 or 2.

Finding substantial transient responses in VTC suggests a rethinking of the role of transient processing in the ventral stream and argues against the prevailing theoretical view that the ventral stream just processes static visual information. That is, our data suggests that transient responses in the ventral steam may serve two purposes. First, processing onset transients may foster detection of novel stimuli and rapid extraction of the gist of the visual input. Second, the substantial "off" transient may ignite a memory trace of the stimulus after it is no longer visible.

What Are the Implications for Modeling fMRI Responses Beyond Visual Cortex?

Our data has critical implications for computational models of the brain. We developed a parsimonious yet powerful encoding model that can be applied to estimate nonlinear neural responses and temporal integration windows across cortex with millisecond resolution. While our two-temporal channel model provides a significant improvement in predicting fMRI signals compared to other models, we acknowledge that it does not explain the entire variance of the data. Future research may build upon the present results and improve model predictions by

adding channels and nonlinearities. For example, combining the temporal encoding approach with a spatial encoding approach[42-44] may enable accurate prediction of brain responses to dynamic real-world scenes.

Given the pervasive use of the standard linear model in fMRI research, our results have broad implications for fMRI studies of any part of the brain. We find that varying the timing of stimuli in the millisecond range has a substantial impact on the magnitude of fMRI responses. However, by estimating neural responses in millisecond resolution, we can accurately predict fMRI responses in second resolution for both brief and long visual stimuli. Thus, the temporal encoding approach we pioneered marks a transformative advancement in using fMRI to elucidate temporal processing in the brain because it links fMRI responses to the timescale of neural computations. This innovative approach offers a quantitative framework to identify functional and computational differences across cortex[45,46] in many domains such as audition[47] and working memory[48]. Importantly, the encoding approach can also be applied to study impairments in high-level abilities like reading[49] and mathematical processing[50] that require integrating visual information over space and time.

In sum, our results provide the first comprehensive computational model of temporal processing in high-level visual cortex. The encoding approach we introduce not only reveals functional differences in temporal processing across visual streams but also underscores the importance of modeling brain responses with millisecond precision to better understand the underlying neural computations.

REFERENCES

1 Schmolesky, M. T. *et al.* Signal timing across the macaque visual system. *J Neurophysiol* **79**, 3272-3278 (1998).

2 De Valois, R. L. & Cottaris, N. P. Inputs to directionally selective simple cells in macaque striate cortex. *Proc Natl Acad Sci U S A* **95**, 14488-14493 (1998).

3 Conway, B. R. & Livingstone, M. S. Space-time maps and two-bar interactions of different classes of direction-selective cells in macaque V-1. *J Neurophysiol* **89**, 2726-2742, doi:10.1152/jn.00550.2002 (2003).

4 Nandy, A. S., Mitchell, J. F., Jadi, M. P. & Reynolds, J. H. Neurons in Macaque Area V4 Are Tuned for Complex Spatio-Temporal Patterns. *Neuron* **91**, 920-930, doi:10.1016/j.neuron.2016.07.026 (2016).

5 Stigliani, A., Jeska, B. & Grill-Spector, K. Encoding model of temporal processing in human visual cortex. *Proc Natl Acad Sci U S A* **114**, E11047-E11056, doi:10.1073/pnas.1704877114 (2017).

6 Zhou, J., Benson, N. C., Kay, K. N. & Winawer, J. Compressive Temporal Summation in Human Visual Cortex. *J Neurosci* **38**, 691-709, doi:10.1523/JNEUROSCI.1724-17.2017 (2018).

7 Buxton, R. B., Wong, E. C. & Frank, L. R. Dynamics of blood flow and oxygenation changes during brain activation: the balloon model. *Magn Reson Med* **39**, 855-864 (1998).

8 Boynton, G. M., Engel, S. A., Glover, G. H. & Heeger, D. J. Linear systems analysis of functional magnetic resonance imaging in human V1. *J Neurosci* **16**, 4207-4221 (1996).

9 Huettel, S. A. & McCarthy, G. Evidence for a refractory period in the hemodynamic response to visual stimuli as measured by MRI. *Neuroimage* **11**, 547-553, doi:10.1006/nimg.2000.0553 (2000).

10 Ogawa, S. *et al.* An approach to probe some neural systems interaction by functional MRI at neural time scale down to milliseconds. *Proc Natl Acad Sci U S A* **97**, 11026-11031 (2000).

11 Birn, R. M., Saad, Z. S. & Bandettini, P. A. Spatial heterogeneity of the nonlinear dynamics in the FMRI BOLD response. *Neuroimage* **14**, 817-826, doi:10.1006/nimg.2001.0873 (2001).

12 Mukamel, R., Harel, M., Hendler, T. & Malach, R. Enhanced temporal non-linearities in human object-related occipito-temporal cortex. *Cereb Cortex* **14**, 575-585, doi:10.1093/cercor/bhh019 (2004).

13 Wager, T. D., Vazquez, A., Hernandez, L. & Noll, D. C. Accounting for nonlinear BOLD effects in fMRI: parameter estimates and a model for prediction in rapid event-related studies. *Neuroimage* **25**, 206-218, doi:10.1016/j.neuroimage.2004.11.008 (2005).

14 Gilaie-Dotan, S., Nir, Y. & Malach, R. Regionally-specific adaptation dynamics in human object areas. *Neuroimage* **39**, 1926-1937, doi:10.1016/j.neuroimage.2007.10.010 (2008).

15 Stigliani, A., Weiner, K. S. & Grill-Spector, K. Temporal Processing Capacity in High-Level Visual Cortex Is Domain Specific. *J Neurosci* **35**, 12412-12424, doi:10.1523/JNEUROSCI.4822-14.2015 (2015).

16 McKeeff, T. J., Remus, D. A. & Tong, F. Temporal limitations in object processing across the human ventral visual pathway. *J Neurophysiol* **98**, 382-393, doi:10.1152/jn.00568.2006 (2007).

17 Gentile, F. & Rossion, B. Temporal frequency tuning of cortical face-sensitive areas for individual face perception. *Neuroimage* **90**, 256-265, doi:10.1016/j.neuroimage.2013.11.053 (2014).

18 Horiguchi, H., Nakadomari, S., Misaki, M. & Wandell, B. A. Two temporal channels in human V1 identified using fMRI. *Neuroimage* **47**, 273-280, doi:10.1016/j.neuroimage.2009.03.078 (2009).

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

20 Honey, C. J. *et al.* Slow cortical dynamics and the accumulation of information over long timescales. *Neuron* **76**, 423-434, doi:10.1016/j.neuron.2012.08.011 (2012).

21 Mattar, M. G., Kahn, D. A., Thompson-Schill, S. L. & Aguirre, G. K. Varying Timescales of Stimulus Integration Unite Neural Adaptation and Prototype Formation. *Curr Biol* **26**, 1669-1676, doi:10.1016/j.cub.2016.04.065 (2016).

22 Merigan, W. H. & Maunsell, J. H. How parallel are the primate visual pathways? *Annu Rev Neurosci* **16**, 369-402, doi:10.1146/annurev.ne.16.030193.002101 (1993).

23 Maunsell, J. H., Nealey, T. A. & DePriest, D. D. Magnocellular and parvocellular contributions to responses in the middle temporal visual area (MT) of the macaque monkey. *J Neurosci* **10**, 3323-3334 (1990).

24 Van Essen, D. C. & Gallant, J. L. Neural mechanisms of form and motion processing in the primate visual system. *Neuron* **13**, 1-10 (1994).

25 Ferrera, V. P., Nealey, T. A. & Maunsell, J. H. Responses in macaque visual area V4 following inactivation of the parvocellular and magnocellular LGN pathways. *J Neurosci* **14**, 2080-2088 (1994).

26 Bonda, E., Petrides, M., Ostry, D. & Evans, A. Specific involvement of human parietal systems and the amygdala in the perception of biological motion. *J Neurosci* **16**, 3737-3744 (1996).

27 Puce, A., Allison, T., Asgari, M., Gore, J. C. & McCarthy, G. Differential sensitivity of human visual cortex to faces, letterstrings, and textures: a functional magnetic resonance imaging study. *J Neurosci* **16**, 5205-5215 (1996).

28 Beauchamp, M. S., Lee, K. E., Haxby, J. V. & Martin, A. Parallel visual motion processing streams for manipulable objects and human movements. *Neuron* **34**, 149-159 (2002).

29 Grossman, E. D. & Blake, R. Brain Areas Active during Visual Perception of Biological Motion. *Neuron* **35**, 1167-1175 (2002).
30 Fox, C. J., Moon, S. Y., Iaria, G. & Barton, J. J. The correlates of subjective perception of identity and expression in the face network: an fMRI adaptation study. *Neuroimage* **44**, 569-580, doi:10.1016/j.neuroimage.2008.09.011 (2009).
31 Pitcher, D., Dilks, D. D., Saxe, R. R., Triantafyllou, C. & Kanwisher, N. Differential selectivity for dynamic versus static information in face-selective cortical regions. *Neuroimage* **56**, 2356-2363, doi:10.1016/j.neuroimage.2011.03.067 (2011).
32 Said, C. P., Moore, C. D., Engell, A. D., Todorov, A. & Haxby, J. V. Distributed representations of dynamic facial expressions in the superior temporal sulcus. *J Vis* **10**, 11 (2010).
33 Freiwald, W., Duchaine, B. & Yovel, G. Face Processing Systems: From Neurons to Real-World Social Perception. *Annu Rev Neurosci* **39**, 325-346, doi:10.1146/annurev-neuro-070815-013934 (2016).
34 Gilaie-Dotan, S., Saygin, A. P., Lorenzi, L. J., Rees, G. & Behrmann, M. Ventral aspect of the visual form pathway is not critical for the perception of biological motion. *Proc Natl Acad Sci U S A* **112**, E361-370, doi:10.1073/pnas.1414974112 (2015).
35 Weiner, K. S. & Grill-Spector, K. Sparsely-distributed organization of face and limb activations in human ventral temporal cortex. *Neuroimage* **52**, 1559-1573, doi:10.1016/j.neuroimage.2010.04.262 (2010).
36 Watson, A. B. in *Handbook of Perception and Human Performance* (eds K Boff, L Kaufman, & J Thomas) (Wiley, 1986).
37 Haxby, J. V., Hoffman, E. A. & Gobbini, M. I. The distributed human neural system for face perception. *Trends Cogn Sci* **4**, 223-233 (2000).
38 Fisch, L. *et al.* Neural "ignition": enhanced activation linked to perceptual awareness in human ventral stream visual cortex. *Neuron* **64**, 562-574, doi:10.1016/j.neuron.2009.11.001 (2009).
39 Jacques, C. *et al.* Corresponding ECoG and fMRI category-selective signals in human ventral temporal cortex. *Neuropsychologia* **83**, 14-28, doi:10.1016/j.neuropsychologia.2015.07.024 (2016).
40 Engell, A. D. & McCarthy, G. Selective attention modulates face-specific induced gamma oscillations recorded from ventral occipitotemporal cortex. *J Neurosci* **30**, 8780-8786, doi:10.1523/JNEUROSCI.1575-10.2010 (2010).
41 Davidesco, I. *et al.* Spatial and object-based attention modulates broadband high-frequency responses across the human visual cortical hierarchy. *J Neurosci* **33**, 1228-1240, doi:10.1523/JNEUROSCI.3181-12.2013 (2013).
42 Dumoulin, S. O. & Wandell, B. A. Population receptive field estimates in human visual cortex. *Neuroimage* **39**, 647-660, doi:10.1016/j.neuroimage.2007.09.034 (2008).
43 Kay, K. N., Naselaris, T., Prenger, R. J. & Gallant, J. L. Identifying natural images from human brain activity. *Nature* **452**, 352-355, doi:10.1038/nature06713 (2008).
44 Naselaris, T., Prenger, R. J., Kay, K. N., Oliver, M. & Gallant, J. L. Bayesian reconstruction of natural images from human brain activity. *Neuron* **63**, 902-915, doi:10.1016/j.neuron.2009.09.006 (2009).
45 Yamins, D. L. *et al.* Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc Natl Acad Sci U S A* **111**, 8619-8624, doi:10.1073/pnas.1403112111 (2014).
46 Kriegeskorte, N. Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annu Rev Vis Sci* **1**, 417-446, doi:10.1146/annurev-vision-082114-035447 (2015).
47 Werner, S. & Noppeney, U. The contributions of transient and sustained response codes to audiovisual integration. *Cereb Cortex* **21**, 920-931, doi:10.1093/cercor/bhq161 (2011).
48 Druzgal, T. J. & D'Esposito, M. Dissecting contributions of prefrontal cortex and fusiform face area to face working memory. *J Cogn Neurosci* **15**, 771-784, doi:10.1162/089892903322370708 (2003).
49 Demb, J. B., Boynton, G. M. & Heeger, D. J. Brain activity in visual cortex predicts individual differences in reading performance. *Proc Natl Acad Sci U S A* **94**, 13363-13366 (1997).
50 Harvey, B. M., Klein, B. P., Petridou, N. & Dumoulin, S. O. Topographic representation of numerosity in the human parietal cortex. *Science* **341**, 1123-1126, doi:10.1126/science.1239052 (2013).

ONLINE METHODS

Participants

Twelve participants (6 males, 6 females) with normal or corrected-to-normal vision participated in the main experiments (experiments 1–3). Each individual provided written informed consent and participated in two fMRI sessions: one session for experiments 1 and 2 and another session for experiment 3 and a functional localizer experiment[15]. Seven participants from the main experiments (3 males, 4 females) also underwent population receptive field (pRF) mapping[42] to define retinotopic cortical regions and another experiment to define human motion-sensitive area (hMT+)[51]. The Stanford Internal Review Board on Human Subjects Research approved all protocols.

Temporal Channels Experiments

Visual stimuli

Stimuli consisted of well-controlled grayscale images of faces, bodies, and pseudowords (**Fig. 1a**, *right*) used in our previous publications[15]. Stimuli were presented using an Eiki LC-WUL100L projector (resolution: 1920 x 1200; refresh rate: 60 Hz) that was controlled by an Apple MacBook Pro using MATLAB (http://www.mathworks.com/) and functions from Psychophysics Toolbox[52] (http://psychtoolbox.org). Participants viewed images through an auxiliary mirror mounted on the RF coil with stimuli spanning ~20° of visual angle in each dimension.

Experimental design

To develop a temporal encoding model for high-level visual cortex, we adapted a fMRI paradigm previously used to model contributions of sustained and transient temporal channels in early visual cortex[5]. The three main experiments in this study all used the same stimuli, trial durations, and task but varied the temporal presentation of the images. Critically, a 12-s baseline period (blank gray screen) always came before and after each trial. In all three experiments, participants were instructed to fixate on a small, central dot and respond by button press when it changed color (occurring randomly once every 2–14 s, 8 s on average).

Experiment 1 – one continuous image per trial: Stimuli were shown in trials of varying durations (3, 5, 10, or 20 s per trial) in which a single image was shown for the entire trial. Across

trial durations the number of stimuli and transients (at the onset and offset of each stimulus) are matched but the duration of stimulation varies (**Fig. 1a-b**, *blue*). This experiment was designed enable measurement of fMRI-adaptation for prolonged images[14].

Experiment 2 – 30 flashed images per trial: used the same trial durations as experiment 1, but in each trial we presented 30 different images from the same category. Each image was shown for 33 ms and followed by a blank interstimulus interval (ISI). Across trial durations the number of stimuli, number of transients, and total duration of visual stimulation are matched, but the ISI between consecutive images varied. Each ISI was 67 ms in the 3-s trials, 133 ms in the 5-s trials, 300 ms in the 10-s trials, and 633 ms in the 20-s trials (**Fig. 1a-b**, *red*).

Experiment 3 – 30 longer images per trial: used the same design as experiment 2, except that in each trial we presented 30 images from the same category for longer durations with a constant ISI of 33 ms between images. Image durations varied across trials and were each shown for 67 ms in the 3-s trials, 133 ms in the 5-s trials, 300 ms in the 10-s trials, and 633 ms in the 20-s trials (**Fig. 1a-b**, *green*).

Data acquisition

Functional data were acquired using a simultaneous multi-slice EPI sequence with a multiplexing factor of 3 to obtain near whole-brain coverage with a TR of 1 s. Participants viewed four 270-s runs of each experiment. Each run of each experiment contained one instance of every permutation of stimulus category (face/body/word) and trial duration (3, 5, 10, or 20 s) presented in random order.

Category localizer experiment

To functionally define cortical regions that respond preferentially to specific stimulus categories, we collected three 300-s runs of a standard fMRI category localizer experiment used in our previous publications[15]. Participants were instructed to fixate on a central dot and respond by button press when an image repeated randomly within a block. Code for the experiment is available at https://github.com/VPNL/fLoc.

pRF mapping and hMT+ localizer

To delineate retinotopic boundaries, we acquired four 200-s runs of pRF mapping[42] in a subset of participants from the main experiments. In this experiment, a bar swept across a

circular aperture (40° × 40° of visual angle) in eight directions as participants performed a fixation task. To functionally define hMT+ in the same subset of participants, we collected one 300-s run of a fMRI motion localizer experiment as detailed in our prior publications[5,51].

Magnetic resonance imaging (MRI)

MRI data were collected using a 3T GE Signa MR750 scanner at the Center for Cognitive and Neurobiological Imaging (CNI) at Stanford University.

fMRI: We used a Nova phase-array 32-channel head coil for the main experiments and functional localizer to obtain near whole-brain coverage (48 slices; resolution: 2.4 × 2.4 × 2.4 mm; one-shot T2*-sensitive gradient echo acquisition sequence: FOV = 192 mm, TE = 30 ms, TR = 1000 ms, and flip angle = 76°). We also collected T1-weighted inplane images to align each participant's functional data to their high-resolution whole brain anatomy.

For pRF mapping and the hMT+ localizer, we used a 16-channel visual array coil (28 slices; resolution: 2.4 × 2.4 × 2.4 mm; one-shot T2*-sensitive gradient echo acquisition sequence: FOV = 192 mm, TE = 30 ms, TR = 2000 ms, and flip angle = 77°) and collected T1-weighted inplane images in the same prescription.

Anatomical MRI: We acquired a whole-brain, anatomical volume in each participant using a Nova 32-channel head coil (resolution: 1 × 1 × 1 mm; T1-weighted BRAVO pulse sequence: FOV = 240 mm, TI = 450 ms, and flip angle = 12°).

Data Analysis

Data were analyzed with MATLAB using code from vistasoft (http://github.com/vistalab) and FreeSurfer (http://freesurfer.net). Code used for predicting fMRI responses using a temporal channels approach is available at https://github.com/VPNL/TemporalChannels.

Region of interest (ROI) definition

Category-selective regions were defined in each participant's native anatomical space at a common threshold (t > 3, voxel level, uncorrected) using functional and anatomical criteria detailed in prior publications[15] (**Fig. 1c**). Face-selective ROIs (faces > others) were defined bilaterally in the inferior occipital gyrus (IOG-faces, N = 10), posterior STS (pSTS-faces, N = 9), posterior fusiform gyrus (pFus-faces, N = 11), and mid fusiform gyrus (mFus-faces; N = 11). Body-

selective ROIs (bodies > others) were found bilaterally in the lateral occipital sulcus (LOS-bodies, N = 10), inferior temporal gyrus (ITG-bodies, N = 10), middle temporal gyrus (MTG-bodies, N = 11), and occipitotemporal sulcus (OTS, N = 10).

Visual areas V1 and hV4 were defined in each hemisphere in a subset of participant (N = 7) using data from the pRF mapping experiment. To match the visual field coverage of the stimuli in the main experiments, we restricted ROIs to only included voxels with pRF centers within the central 10°. We also defined bilateral hMT+ in the same subset of participants using data from the motion localizer experiment as in previous publications[5,51].

Optimized two-temporal channel A+S model

To predict responses across all three experiments with a single model, we adapted an encoding approach introduced by prior studies[5,18] that models fMRI responses as the weighted sum of activity across separate sustained and transient temporal channels.

In the procedure illustrated in **Fig. 3**, we first predict neural activity in each channel by convolving the stimulus time course in millisecond resolution (**Fig. 3a**) separately with the neural IRF for the sustained channel (**Fig. 3b**, *blue channel IRF*) and the transient channel (**Fig. 3b**, *red channel IRF*). The sustained channel is characterized by a monophasic IRF_S that generates a response for the entire duration of a stimulus followed by an adaptation nonlinearity — convolved neural responses are multiplied by an exponential decay function beginning at the onset of each stimulus and extending until the onset of the following stimulus. In contrast, the transient channel is characterized by a biphasic IRF_T that generates a brief response at the onset and offset of an image[2,3,36]. Here, convolved responses are passed through sigmoid nonlinearities that allow different levels of compression to be applied to the "on" and "off" responses. Then, the estimated neural responses for each channel are convolved with a hemodynamic response function (HRF) to generate a prediction of the fMRI response (**Fig. 3c**). As such, there are neural nonlinearities in each channel of this model, but a linear relationship is assumed between the neural activity and BOLD responses. Finally, we use a GLM to solve for the contributions (β weights) of the sustained and transient channels, which reflect how much the predicted response from each channel is scaled before the responses of both channels are summed. Thus, the BOLD response to a stimulus can be expressed as

$$\beta_S\,[(\text{stimulus} \otimes IRF_S) \cdot e^{-t/\alpha} \otimes HRF] + \beta_T\,[(\text{stimulus} \otimes IRF_T)^{\sigma} \otimes HRF],$$

where β_S and β_T are the fitted response amplitudes for the sustained and transient channels, respectively; IRF_S and IRF_T are the impulse response functions for the sustained and transient channels, respectively; α determines the exponential decay at time *t* after stimulus onset; σ is a pointwise sigmoid nonlinearity, and *HRF* is the canonical hemodynamic response function.

Modeling nonlinearities in the neural response

We model the IRFs for each channel (**Fig. 3b**) using formulas detailed in our prior publications[5]. Here, we optimize the IRF time constant τ for each region, and the other parameters (taken from Watson[36]) are held constant: κ=1.33, n_1=9, and n_2=10.

Adaptation: To capture fMRI-adaptation[14] effects in the sustained channel, we use an exponential decay function, $e^{-t/\alpha}$, where *t* represents time after stimulus onset, and α indicates when the function declines to a proportion of 1/*e* (~37%) of the initial response.

Sigmoid nonlinearities: To allow different levels of compression to be applied to "on" and "off" responses in the transient channel, we optimize separate sigmoid nonlinearities for the onset and offset responses using cumulative Weibull distribution functions,

$$\sigma_{on}(x) = 1 - e^{-(x/\lambda)^{k_{on}}}$$
$$\sigma_{off}(-x) = 1 - e^{-(x/\lambda)^{k_{off}}},$$

where λ is a sigmoid scale parameter used in both onset and offset nonlinearities; k_{on} is a sigmoid shape parameter that controls the curvature of the onset compression function, and k_{off} is a shape parameter controlling curvature of the offset compression function. Smaller *k* values produce more compressive nonlinearities that elongate transient "on" and "off" responses compared to larger *k* values.

Fitting and optimizing the two-temporal channel model

Since the HRF acts like a temporal low-pass filter, this allows resampling fMRI response predictors to the lower temporal resolution of the measured fMRI data (TR = 1 s) with minimal distortion. These resampled predictors are then compared with measured fMRI responses to estimate the contributions (β weights) of each channel for each category. To normalize the amplitude of predicted fMRI responses for the sustained and transient channels, we match the maximal height of predictors in the design matrix across the two channels. Finally, we used a

GLM to estimate β weights of the sustained and transient channels for each stimulus category by comparing the predicted responses with the mean response time series of each ROI in each participant. To optimize the A+S model time constants (τ and α) and sigmoid parameters (λ, k_{on}, k_{off}) for each region, we used the constrained nonlinear optimization algorithm *fmincon* in MATLAB (*Optimization procedures*).

Validating the optimized two-temporal channel model

We assessed the predictive power of the optimized two-temporal A+S channel model by testing how well it predicts responses from separate runs of data from all three experiments. We first generated predicted neural response time courses by coding the visual stimulation in the left-out runs and convolving it separately with the IRFs of the sustained and transient channels (optimized using a separate split of the data). We then applied the adaptation and sigmoid nonlinearities, which were also optimized with independent data (**Fig. 6**). These transformed neural predictors were next convolved with the HRF and down-sampled to 1 s temporal resolution to match our fMRI acquisition. Finally, we multiplied each channel's fMRI predictors with their respective β weights (estimated for each category in an independent split of the data) before summing the channel responses to predict fMRI responses. We then quantified how well the predicted responses matched the measured response across all data in the validation split.

Model performance was operationalized as cross-validated R^2 (x-R^2), which indexes the proportion of variance explained by β weights and model parameters that were estimated from independent data. While similar to a typical R^2 statistic, x-R^2 can be negative when the residual variance of a poor model prediction exceeds the measured variance in the response. Quantification of x-R^2 within each experiment is presented in **Fig. 4d-f** for OTS-bodies. Performance averaged across all three experiments is shown in **Figs. S3-S5** for all regions.

Testing alternative model architectures

To compare with the performance of our optimized two-temporal channel A+S model with alternatives, we tested five single-channel and three dual-channel models (**Figs. 2**, **4**, **S3-S5**).

General linear model (GLM): To first benchmark our model against a common GLM approach[8], we tested a linear model that predicts fMRI responses with a single convolution of the stimulus with the canonical HRF.

Balloon model (B model): To examine if responses in high-level visual cortex can be explained by a nonlinear hemodynamic model, we implemented the balloon model proposed by Buxton an colleagues[7] using standard parameters detailed in prior publications[5].

Linear sustained channel (L model): Similar to the GLM approach[8] but with two stages of convolution, we tested a single-channel model with a linear sustained channel,

$$\beta[(\text{stimulus} \otimes IRF_S) \otimes HRF],$$

where β is a fitted response amplitude; IRF_S is the impulse response function for the sustained channel, and *HRF* is the canonical HRF.

Sustained channel with compressive temporal summation (CTS, **Figs. 4**, **S2-S5***):* We also implemented a model proposed by Zhou et al.[6] composed of sustained channel with a compressive static power law,

$$\beta[(\text{stimulus} \otimes IRF_S)^{\varepsilon} \otimes HRF],$$

where ε is an optimized compression parameter ranging from 0–1.

Sustained channel with adaptation (A model): Identical to the sustained channel shown in **Fig. 3** (blue), we tested a model composed of a single sustained channel with adaptation,

$$\beta[(\text{stimulus} \otimes IRF_S) \cdot e^{-t/\alpha} \otimes HRF],$$

where α determines the exponential decay at *t* seconds after the onset of a stimulus.

Transient channel with sigmoid nonlinearity (S model): Identical to the transient channel shown in **Fig. 3** (red), we also tested a single-channel model composed of a transient channel with the same sigmoid nonlinearities described above,

$$\beta[(\text{stimulus} \otimes IRF_T)^{\sigma} \otimes HRF],$$

where IRF_T is the impulse response function for the transient channel and σ is a pointwise nonlinearity composed of separate sigmoid functions for onset and offset responses.

Alternative dual-channel models (L+Q, C+Q, and A+Q models): To compare the optimized two-temporal channel model shown in **Fig. 3** (A+S model) to alternative dual-channel models, we tested three variants of our model that all use a transient channel with a quadratic (Q) nonlinearity (squaring) but apply different nonlinearities in the sustained channel (**Fig. S3a**). Combining different combinations of the sustained and transient channels described above, we compared two-channel models composed of a transient channel and either a linear sustained

channel (L+Q model), a sustained channel with CTS (C+Q model), or a sustained channel with adaptation (A+Q model).

Optimization procedures

For all models with a neural IRF, we optimized a single time constant, τ, using formulas described in our prior publications[5]. For models with adaption in the sustained channel (A and A+S), we also optimized an exponential time constant (α). For models with compressive temporal summation (CTS and C+Q models), we instead optimized an exponential compression parameter (ε). For models with a sigmoid nonlinearity in the transient channel (S and A+S models), we optimized three sigmoid parameters (λ, k_{on}, k_{off}). To optimize model parameters, we used the nonlinear optimization algorithm *fmincon* in MATLAB with the following constraints: τ = 4–20 ms, α = 10–40 s, ε = 0.01–1, λ = 0.01–0.5, k_{on} = 0.1–6, and k_{off} = 0.1–6 . The initial values passed to the optimizer for each parameter were τ = 4.93 ms, α = 20 s, ε = 0.1, λ = 0.1, k_{on} = 3, and k_{off} = 3. The cross-validation performance of each model averaged across all three experiments is shown in **Fig. S3b** for category-selective regions in VTC and LTC and in **Figs. S4-S5** for other regions.

Statistical analyses

To test for differences in model performance across regions in VTC and LTC, we used a two-way repeated measures analysis of variance (ANOVA) with factors of model and region (comparing models and regions shown in **Fig. S3b**). We then used paired two-tailed *t*-tests to compare the x-R^2 of our model with others. **Fig. 4d-f** contrasts the performance of our model (A+S) in OTS-bodies against three other models (GLM, CTS, L+Q) for each experiment individually. **Figs. S3-S5** contrast the performance of our model averaged across all three experiments vs. every other model for each region. To assess the level of noise in measurements from different brain regions, we also calculated a noise ceiling for each ROI using the inter-trial variability of responses for each condition as described in our prior publications[5]. The noise ceiling estimate for OTS-bodies in each experiment is plotted in **Fig. 4d-f**, and the average noise ceiling across all three experiments is plotted in **Figs. S3-S5** for each region.

After establishing the validity of our model, we used paired two-tailed *t*-tests to compare β weights estimated by the A+S model for each region's preferred category vs. average contributions for nonpreferred categories, separately for the sustained and transient channels.

To test whether selectivity in the two channels differs across regions preferring bodies and faces in either VTC or LTC, we also used two-way ANOVAs with factors of channel (sustained/transient) and preferred category (bodies/faces) on the difference in channel weights for preferred vs. nonpreferred categories (contrast effect size, CES; **Fig. 5**). To examine whether the proportion of response attributed to sustained vs. transient channels differs across processing streams, stimulus categories, or regions preferring different categories, we then used a three-way ANOVA on channel contribution ratios, $\left|\frac{\beta_S}{\beta_T}\right|$, for each category with factors of stream (ventral/lateral), stimulus (faces/bodies/words), and preferred category (bodies/faces).

Methods References

51 Weiner, K. S. & Grill-Spector, K. Not one extrastriate body area: using anatomical landmarks, hMT+, and visual field maps to parcellate limb-selective activations in human lateral occipitotemporal cortex. *Neuroimage* **56**, 2183-2199, doi:10.1016/j.neuroimage.2011.03.041 (2011).

52 Brainard, D. H. The Psychophysics Toolbox. *Spat Vis* **10**, 433-436 (1997).

SUPPLEMENTARY INFORMATION

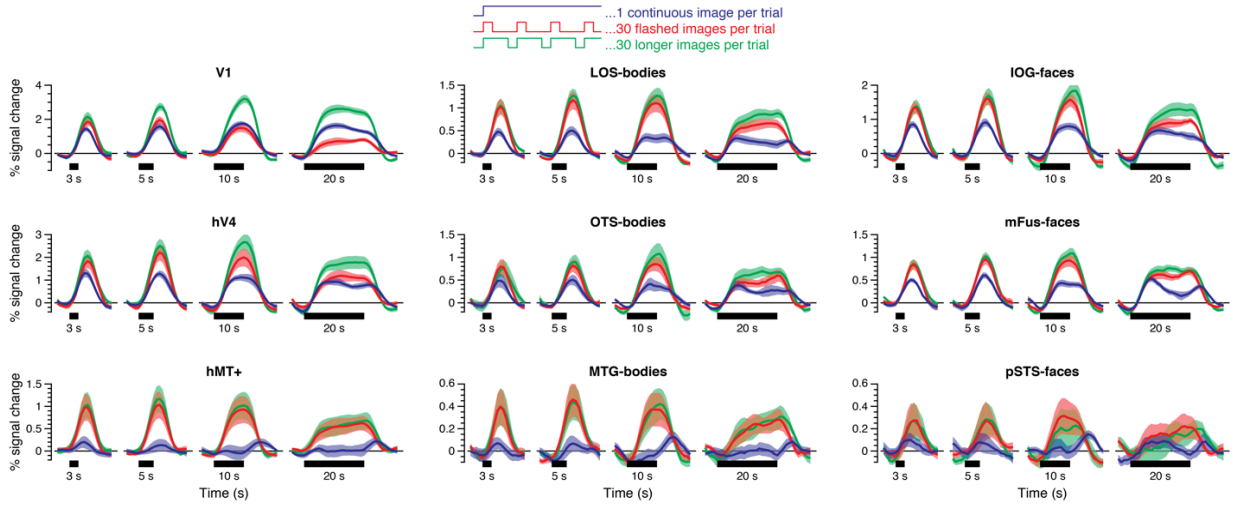

Figure S1. Responses to time-varying stimuli in occipital, ventral, and lateral regions of interest. Measured responses in occipital (V1, LOS-bodies, IOG-faces), ventral (hV4, OTS-bodies, mFus-faces), and lateral (hMT+, MTG-bodies, pSTS-faces) regions of interest in experiment 1 (*blue*), experiment 2 (*red*), and experiment 3 (*green*) averaged across all three stimulus categories. *Lines:* mean response time series across participants; *shaded areas:* standard error of the mean (SEM) across participants; *Horizontal black bars:* trial duration.

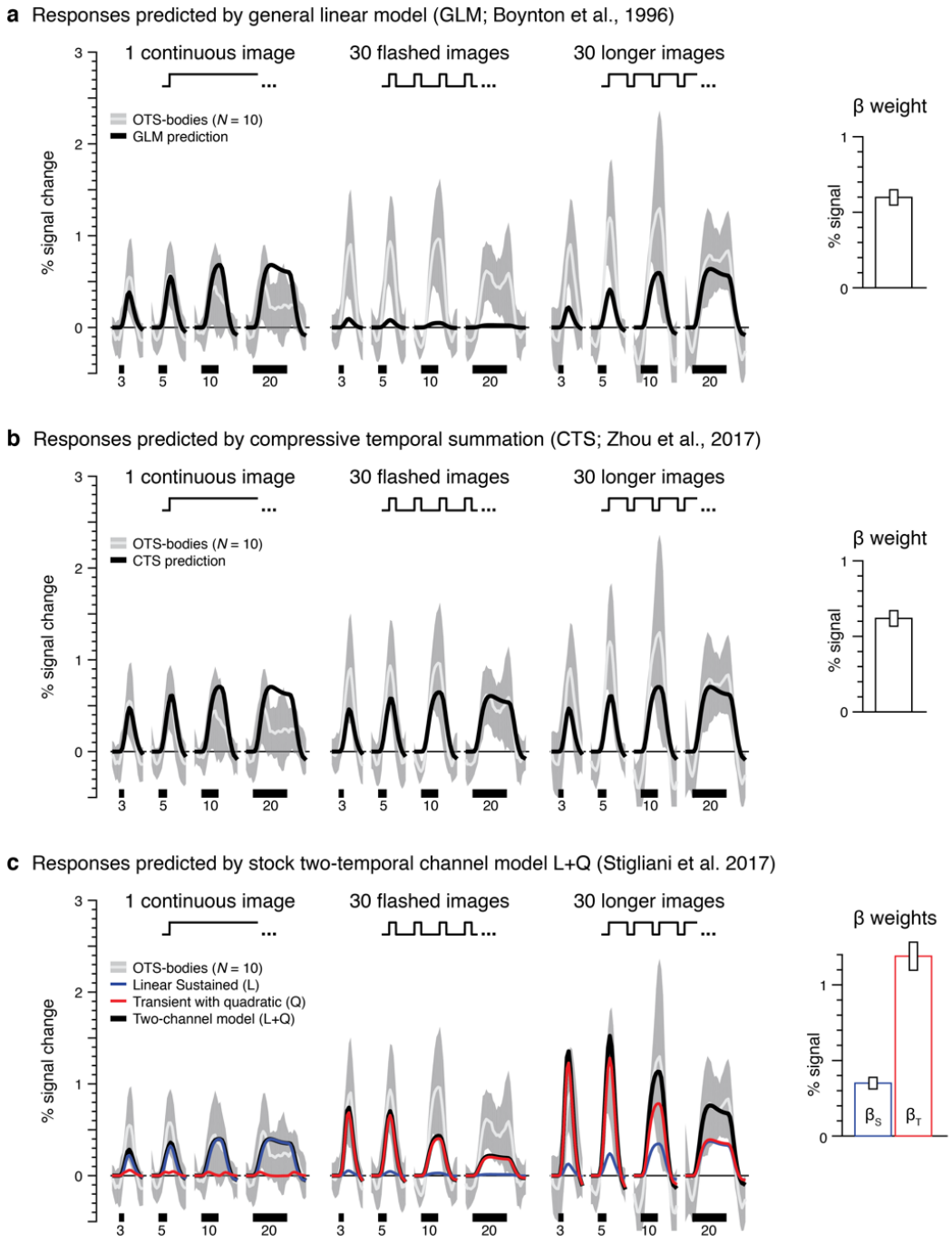

Figure S2. **Comparison of temporal encoding models in OTS-bodies**. (a-c) Responses and model predictions for body images in OTS-bodies for each experiment (*left*) with estimated β weights for each model (*right*). *White curve:* mean response across 10 participants. *Shaded gray:* standard deviation across participants. *Black curve:* overall model prediction. *Horizontal black bar:* trial duration. (**a**) Predictions of a general linear model (GLM)[8]. (**b**) Predictions of a model with compressive temporal summation (CTS)[6]. (**c**) Predictions of the two-temporal channel L+Q model with linear sustained channel and quadratic transient channel. *Blue curve:* predicted response from the sustained channel. *Red curve:* predicted response from the transient channel: *Black curve:* sum of responses from both channels.

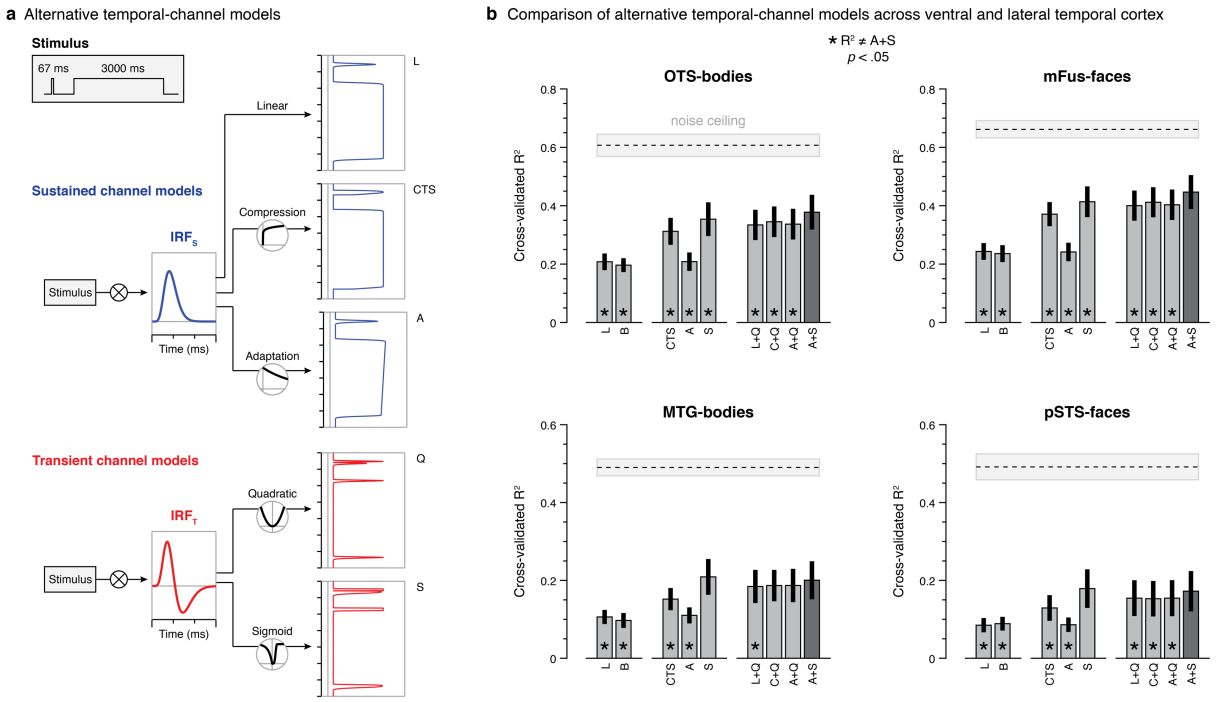

Figure S3. **Comparison of temporal encoding models across visual cortex**. (**a**) Alternative models of sustained (*blue*) and transient (*red*) channels. Schematic depicts neural response predictions generated by different implementations of each channel for both a brief (67 ms) and long (3 s) stimulus. Sustained channel models: *L,* a linear sustained channel; *CTS,* a sustained channel with compressive temporal summation[6]; *A,* a sustained channel with adaptation[8]. Transient channel models: *Q,* a transient channel with a quadratic (squaring) nonlinearity; *S*, a transient channel with a sigmoid nonlinearity. (**b**) Comparison of model performance (cross-validated R^2) in each region averaged across all three experiments. Hemodynamic models: *L*, same as in (a); *B*, balloon model[7]. Single-channel neural models: *CTS*, *A*, and *S*, same as in (a). Two-channel neural models: *L+Q,* a linear sustained channel and a transient channel with a quadratic nonlinearity[5]; *C+Q,* a sustained channel with compressive temporal summation and a transient channel with a quadratic nonlinearity; *A+Q,* a sustained channel with adaptation and a transient channel with a quadratic nonlinearity; *A+S*, a sustained channel with adaptation and a transient channel with a sigmoid nonlinearity. Asterisks denote models with significantly different performance vs. the A+S model.

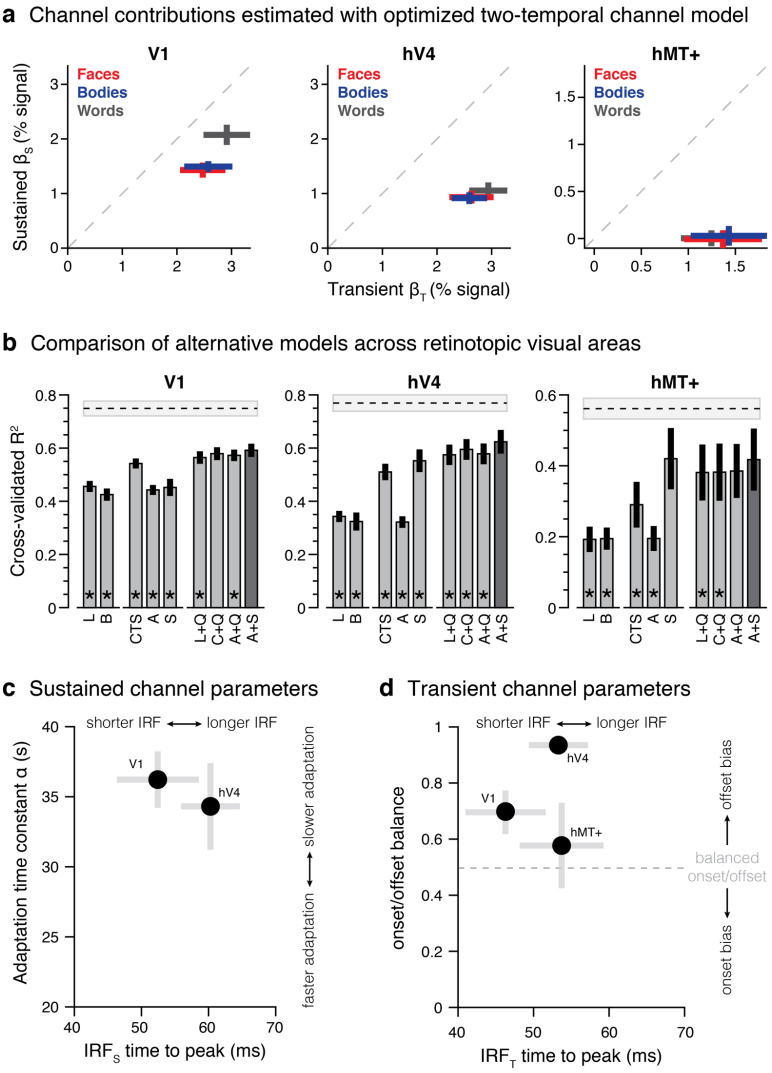

Figure S4. **Contributions of transient and sustained temporal channels across early visual areas.** (**a**) Contributions (β weights) of transient (*x* axis) and sustained (*y* axis) channels for each stimulus category estimated by the two-temporal channel A+S model in V1, hV4, and hMT+. Crosses span ±1 SEM across participants in each axis, and β were solved by fitting the model using data concatenated across all experiments. Data show average model weights across all splits of the data for each participant. *Red:* response to faces. *Blue:* response to bodies. *Gray:* response to words. *Dashed gray*: identity line ($\beta_S = \beta_T$). (**b**) Comparison of model performance (cross-validated R^2) in each region averaged across all three experiments. Hemodynamic models: *L* and *B*. Single-channel neural models: *CTS*, *A*, and *S*. Two-channel neural models: *L+Q*[5], *C+Q*, *A+Q, and A+S*. Asterisks denote models with significantly different performance vs. the A+S model. (**c**) Optimized sustained channel parameters. Time to peak of sustained IRF_S (*x* axis) and exponential time constant of the adaptation function (*y* axis) for each region estimated by the two-temporal channel A+S model. Crosses span ±1 SEM across participants in each axis, and parameters were optimized using split halves of the data containing runs from all experiments. Data show model parameters averaged across both splits of the data for each participant. (**d**) Optimized transient channel parameters. Time to peak of transient IRF_T (*x* axis) and onset/offset balance (*y* axis) for each region estimated by the two-temporal channel A+S model (with a zeroed sustained channel in hMT+). The onset/offset balance metric captures differences in the shapes of the sigmoid nonlienarities used to compress transient "on" and "off" responses, where values larger than 0.5 refect elongation of offset responses compared to onset responses.

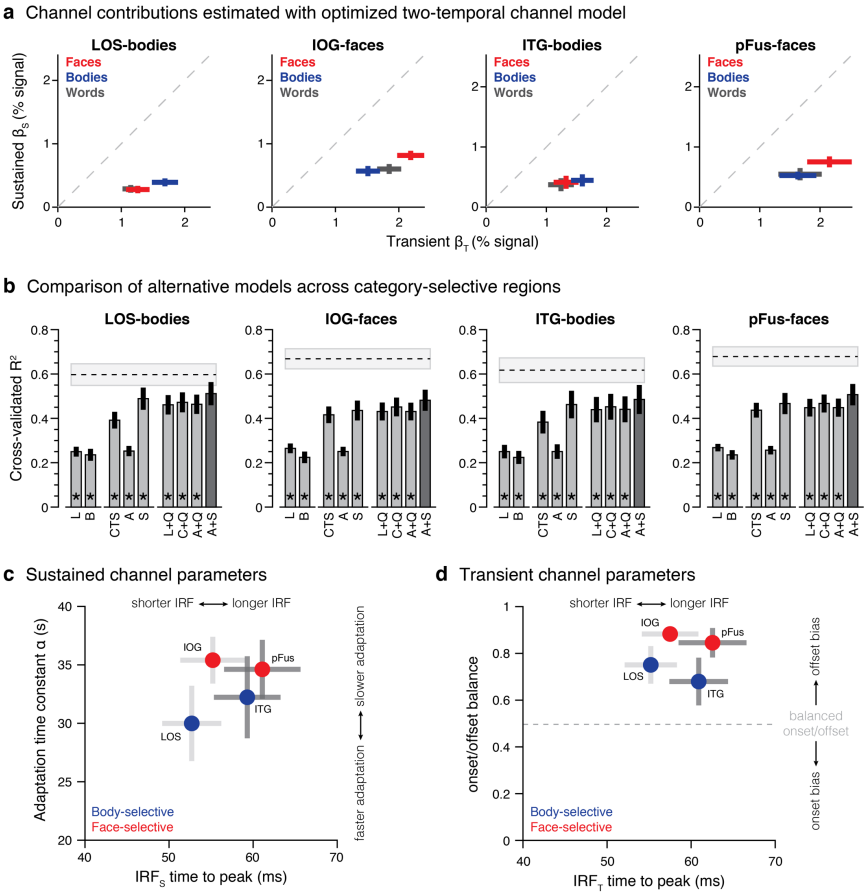

Figure S5. **Contributions of transient and sustained temporal channels across additional category-selective regions.** (**a**) Contributions (β weights) of transient (*x* axis) and sustained (*y* axis) channels for each stimulus category estimated by the two-temporal channel A+S model in lateral occipital cortex (LOS-bodies and IOG-faces) and posterior ventral temporal cortex (ITG-bodies and pFus-faces). Crosses span ±1 SEM across participants in each axis, and β were solved by fitting the model using data concatenated across all experiments. Data show average model weights across all splits of the data for each participant. *Red:* response to faces. *Blue:* response to bodies. *Gray:* response to words. *Dashed gray*: identity line ($\beta_S = \beta_T$). (**b**) Comparison of model performance (cross-validated R^2) in each region averaged across all three experiments. Hemodynamic models: *L* and *B*. Single-channel neural models: *CTS, A,* and *S*. Two-channel neural models: *L+Q*[5], *C+Q, A+Q, and A+S*. Asterisks denote models with significantly different performance vs. the A+S model. (**c**) Optimized sustained channel parameters in lateral occipital cortex (*light gray crosses*) and posterior ventral temporal cortex (*dark gray crosses*). Time to peak of sustained IRF_S (*x* axis) and exponential time constant of the adaptation function (*y* axis) for each region estimated by the two-temporal channel A+S model. Crosses span ±1 SEM across participants in each axis, and parameters were optimized using split halves of the data containing runs from all experiments. Data show model parameters averaged across both splits of the data for each participant. (**d**) Optimized transient channel parameters in lateral occipital cortex (*light gray crosses*) and posterior ventral temporal cortex (*dark gray crosses*). Time to peak of transient IRF_T (*x* axis) and onset/offset balance (*y* axis) for each region estimated by the two-temporal channel A+S model. The onset/offset balance metric captures differences in the shapes of the sigmoid nonlienarities used to compress transient "on" and "off" responses, where values larger than 0.5 refect elongation of offset responses compared to onset responses.