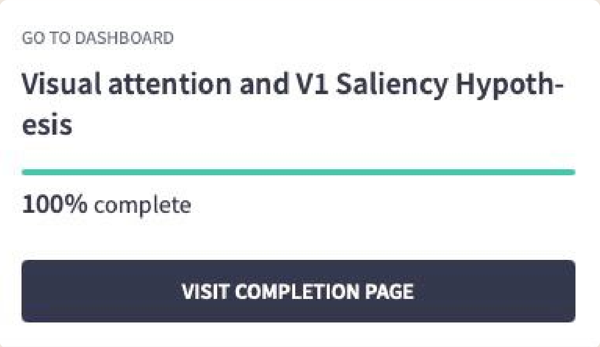
# Visual Attention and V1 Saliency Hypothesis


GO TO DASHBOARD

**Visual attention and V1 Saliency Hypothesis**

**100%** complete

VISIT COMPLETION PAGE

## 1 Visual attention and visual saliency

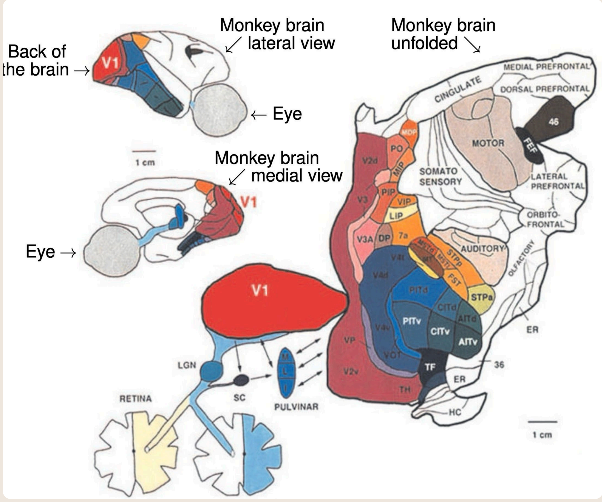
$10^9$ bits/s for inputs, while the attentional bottleneck $\approx 40$bits per second (Szikai 1956)

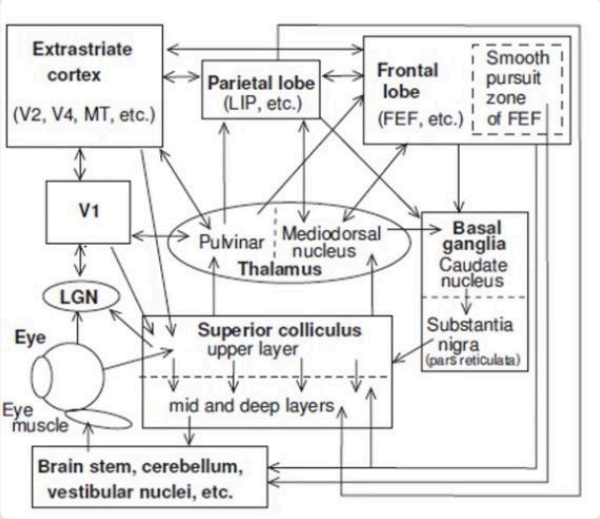
### How to direct attention?

- Behaviorally
  - Overtly (by shifting gaze) vs. Covert (without gaze shift)
  - Top-down (goal-driven) vs. bottom-up (cueing)
    - bottom-up: faster acting, harder to ignore, *so we are easily distracted by things*
    - cue during the sampling -> top-down, cue flash for a while -> bottom-up
  - Space/object/feature-based

    > feature/object-based requires perception and thus is mostly top-down.

- Neurally
  - **FEF** (frontal eye field)
  - **SC** (superior colliculus). It is involved in **eye movement**, thus considered to guide attention.
  - **V1** is massive in its cortical size. Stimulating neurons can evoke eye movement towards the RF location of these neurons.





Can retina alone drive attention, as it also sends signals to SC? #flashcards/neuralmodeling

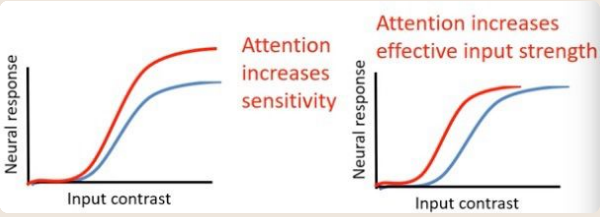
No. After lesioning V1 in monkeys, they are unable to make visually guided **saccades** for weeks. However, for *lower vertebrates* such as fish, the retina plays a significant role in attention.

#flashcards/neuralmodeling

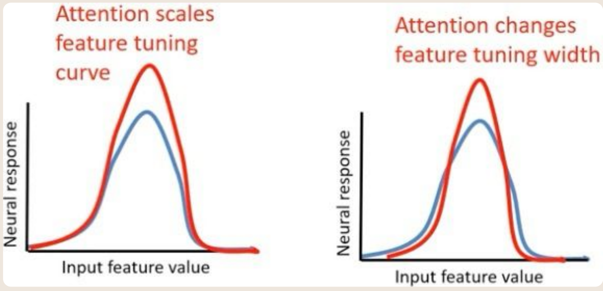
> Which of the following brain regions feed input to the superior colliculus, which gives commands to the brain stem to drive eye movements? -- frontal eye field, V1, parietal cortex, retina, extrastriate cortex(e.g. V2) :: All of them.
> What observation suggests that in monkeys, retinal inputs to the superior colliculus are not sufficient to drive eye movements in normal circumstances? :: Lesioning V1 makes a monkey unable to make visually guided saccades *for weeks*. The research also suggests that the brain can **re-adapt**

### The consequence of attention

- By definition, it makes visual **decoding** better. Faster and more accurate responses
  - Exogenous cueing (hard to ignore) overwrites endogenous cueing at short SOA (time between cue onset and test onset),
- Effects: weak in V1, mostly on neuron responses in V2, V4, etc.
  - increases *sensitivity* or *input strength*



  - *scales* up or *sharpen* the feature tuning curve



- **Biased competition**: Attention selectively *boosts* the neural representation of the attended stimulus, so that it outcompetes other inputs for processing.

#flashcards/neuralmodeling

> In the receptive field of a neuron in V4 or IT, if there are two objects and the monkey directs attention to one of the two objects, what happens to the neural response?:: Due to **biased competition**, it responds as if the other object were not present.
> The effects of top-down attention on neural responses in V1 are typically weaker than those on neural responses in visual cortical areas beyond V1.

## 2 V1 Saliency Hypothesis

**Saliency**: the degree of a spatial location to attract attention in bottom-up manner. The same values could be caused by different features.

**Iso-feature suppression**: nearby neurons tuned to similar features suppress each other, due to intra-cortical interactions in V1. also **Collinear facilitations**

We can measure saliency by **reaction time (RT)**.

Visual search in which the RT is almost *independent of the set size* is called **efficient**, and otherwise **inefficient**, which are considered to depend on underlying neural processes that are *parallel and serial*, respectively.
A feature that supports efficient visual search is defined as a **basic feature dimension**.

> RT reflects more than saliency, as there are more processes than bottom-up attention that happen. Then, how to measure saliency by RT? #flashcards/neuralmodeling
>
> Fixed the top-down factor (and other factors), then the variations in behavior (e.g. RT) reflect the variations in saliency.

## Attention capture by ocular singleton -- a hallmark of V1

The eye-of-origin signal is mainly available in V1 among all the visual cortical areas. Higher visual areas have mostly binocular neurons.

Saliency is measured by the distraction to the signal in the other image.

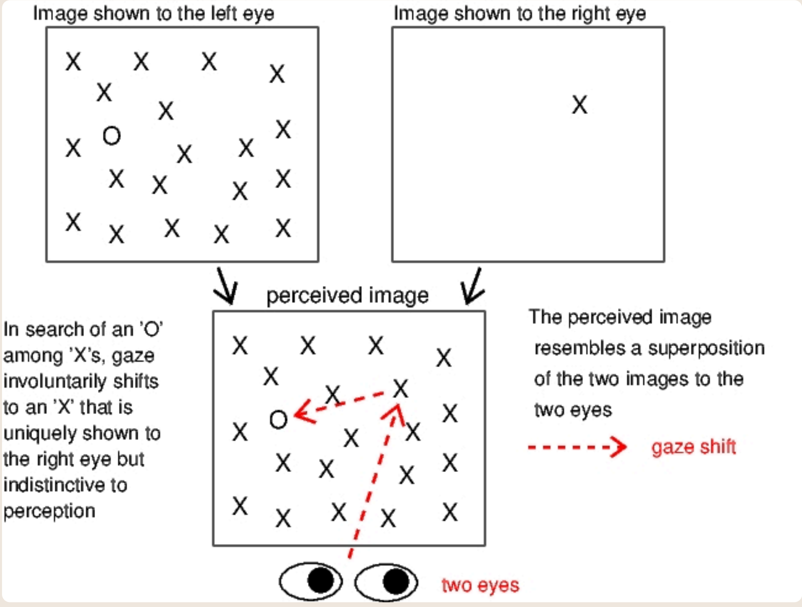


Image shown to the left eye
Image shown to the right eye
perceived image
In search of an 'O' among 'X's, gaze involuntarily shifts to an 'X' that is uniquely shown to the right eye but indistinctive to perception
The perceived image resembles a superposition of the two images to the two eyes
gaze shift
two eyes

#flashcards/neuralmodeling

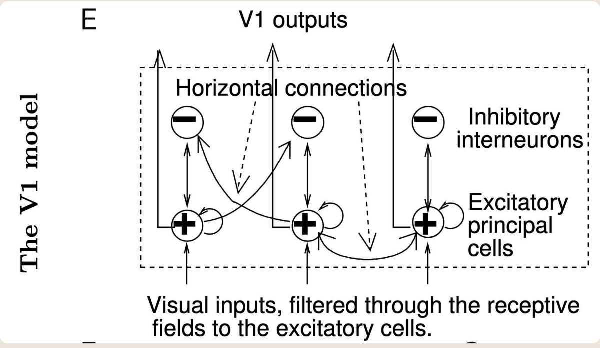
> Saliency in V1 saliency hypothesis refers to the degree of a spatial location to attract attention in bottom-up manner.
> According to the V1 saliency hypothesis, it is the highest response among V1 neural responses to this location that signals the saliency of that location, regardless of the preferred features of these V1 neurons.

# 3 Testing the V1 Saliency Hypothesis

## V1 model: notations and calibrations

Let $i$ and $\theta$ to be the neuron position and the bar angle, and:

- $x_{i\theta}$: state of the principal neuron
- $y_{i\theta}$: state of the interneuron
- $I_{i\theta}$: direct visual input to the E-I pairs by an input bar $\hat{I}_{i\theta}$
- $g_x(x_{i\theta})$: firing rate, sigmoid-like function of $x_{i\theta}$
- $g_y(y_{i\theta})$: firing rate, sigmoid-like function of $y_{i\theta}$


E
V1 outputs
The V1 model
Horizontal connections
Inhibitory interneurons
Excitatory principal cells
Visual inputs, filtered through the receptive fields to the excitatory cells.

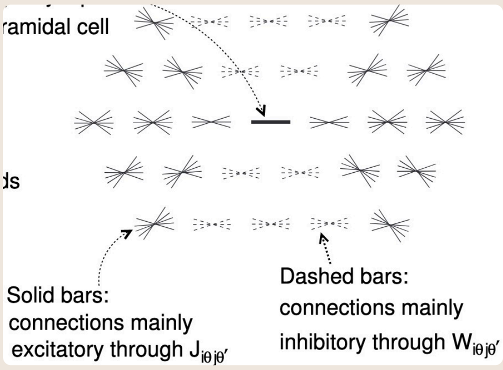
The V1 model

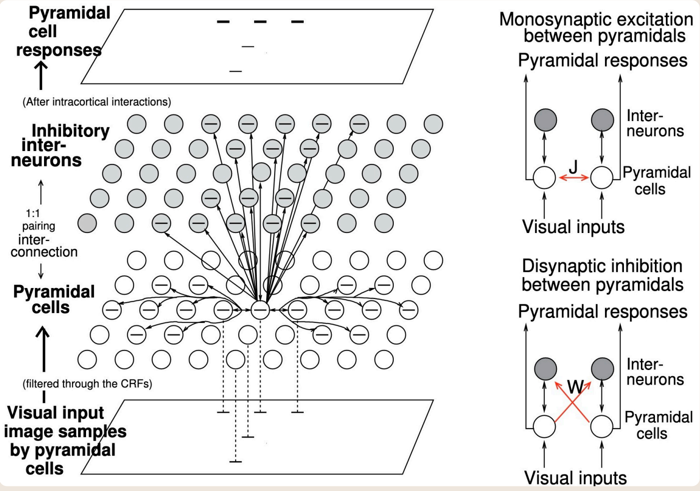
Then the V1 model can be modeled as:

$$\dot{x}_{i\theta} = \underbrace{-\alpha_x x_{i\theta} - g_y(y_{i,\theta}) + I_{i\theta} + J_o g_x(x_{i\theta})}_{\text{Interaction within a single E-I pair}} + \underbrace{\sum_{j\neq i,\theta'} J_{i\theta,j\theta'} g_x(x_{j\theta'})}_{\text{Interaction between hypercolumns}}$$

$$\underbrace{-\sum_{\Delta\theta\neq 0}\psi(\Delta\theta) g_y(y_{i,\theta+\Delta\theta})}_{\text{Di-synaptic suppression between E-I pairs within a hypercolumn}} + \underbrace{I_o}_{\text{Activity normalization term}} + I_{\text{noise}}$$

$$\dot{y}_{i\theta} = \underbrace{-\alpha_y y_{i\theta} + g_x(x_{i\theta})}_{\text{Interaction within a single E-I pair}} + \underbrace{\sum_{j\neq i,\theta'} W_{i\theta,j\theta'} g_x(x_{j\theta'})}_{\text{Interaction between hypercolumns}} + I_{\text{noise}}$$


ramidal cell
ds
Solid bars: connections mainly excitatory through $J_{i\theta j\theta'}$
Dashed bars: connections mainly inhibitory through $W_{i\theta j\theta'}$


Pyramidal cell responses
(After intracortical interactions)
Inhibitory inter-neurons
1:1 pairing inter-connection
Pyramidal cells
(filtered through the CRFs)
Visual input image samples by pyramidal cells
Monosynaptic excitation between pyramidals
Pyramidal responses
Inter-neurons
J
Pyramidal cells
Visual inputs
Disynaptic inhibition between pyramidals
Pyramidal responses
Inter-neurons
W
Pyramidal cells
Visual inputs

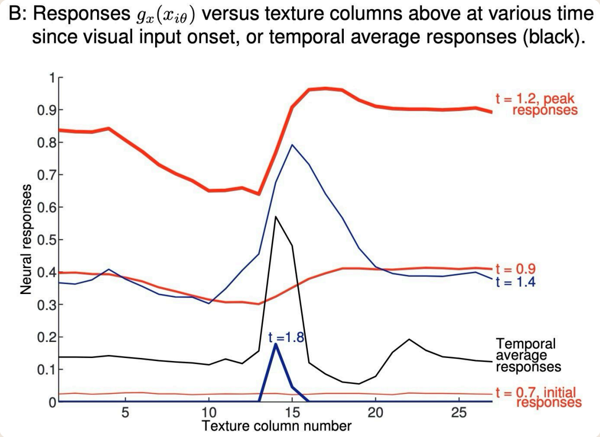
> To prevent hallucinations, it turns out that the E-I dynamic is mathematically essential.

## Model behavior on visual inputs

Temporal dynamics is roughly described as:

1. Visual inputs drive the principal neurons (E), and take a while to activate them fully.
2. Principal neurons activate each other, including the interneurons (I).
3. After interneurons are fully activated, depression is exerted.

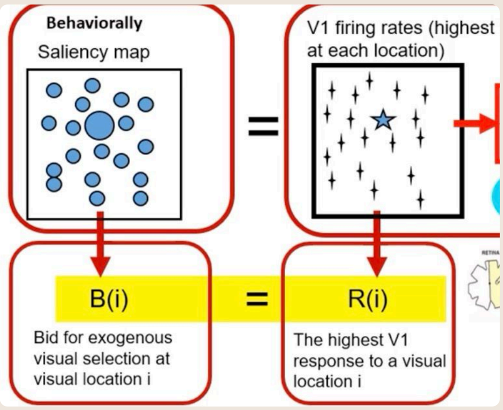
Once the principal neurons are depressed, the interneurons will also lose their drives, leading to another cycle of principal neurons rebounding.



When t=0.7, principal neurons are not fully activated. When t =1.2, responses reach the peak, where depression has not appeared

According to the V1 saliency hypothesis, we use the bid $B_i$ for exogenous visual selection at the visual location $i$, is equal to the highest V1 response $R_i$ to that location.
Namely: $R_i = B_i \equiv \max_\theta[g_x(x_{i\theta})]$. Say $\overline{B} = \langle B_i \rangle$ is the average of $B_i$ over location $i$, and $\sigma_B$ is the standard deviation of $B_i$ over i, then we use the **z-score** to denote the saliency:

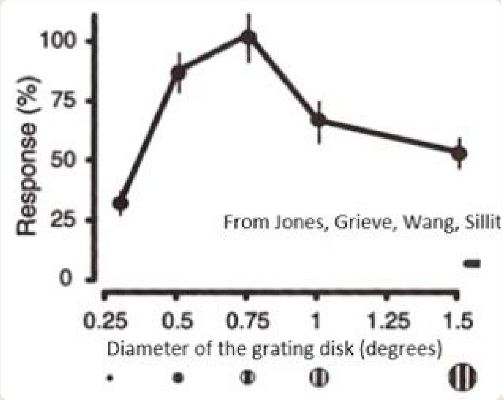
$$z_i \equiv \frac{B_i - \overline{B}}{\sigma_B}$$


Behaviorally
Saliency map
=
V1 firing rates (highest at each location)
B(i)
=
R(i)
Bid for exogenous visual selection at visual location i
The highest V1 response to a visual location i

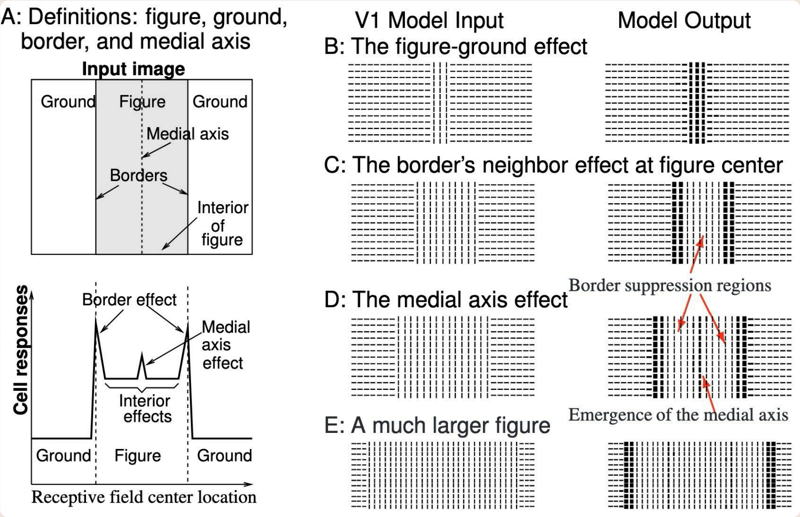
file-20251204133952145.jpg#right

## figure-ground, medial axis, and size effects

The neural response depends on the **size** of the figure, shown by the curve for response versus diameter of a grating disk. --> **figure-ground effect** and **medial axis effect**.

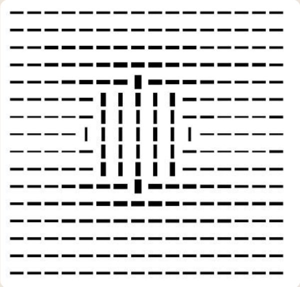


From Jones, Grieve, Wang, Sillit


A: Definitions: figure, ground, border, and medial axis
Input image
Ground
Figure
Ground
Medial axis
Borders
Interior of figure
Cell responses
Border effect
Medial axis effect
Interior effects
Ground
Figure
Ground
Receptive field center location
V1 Model Input
Model Output
B: The figure-ground effect
C: The border's neighbor effect at figure center
Border suppression regions
D: The medial axis effect
Emergence of the medial axis
E: A much larger figure

**Cross-orientation facilitation**: An orientation unspecific surrounding grating can suppress the responses to the border of the central grating, releasing the central response from the border suppression.

The neural response also depends on **input strength (contrast)**, due to collinear facilitation.



Disinhibiting the center response by a surround grating

## Reflections from the model

The V1 model achieves *simultaneously*:

- reproducing the **contextual influences** observed physiologically in real V1 neurons
- being able to **amplify selective deviations** from homogeneity in the input, without hallucinating. Then, the model's behavior by V1SH is consistent with visual saliency behavior.

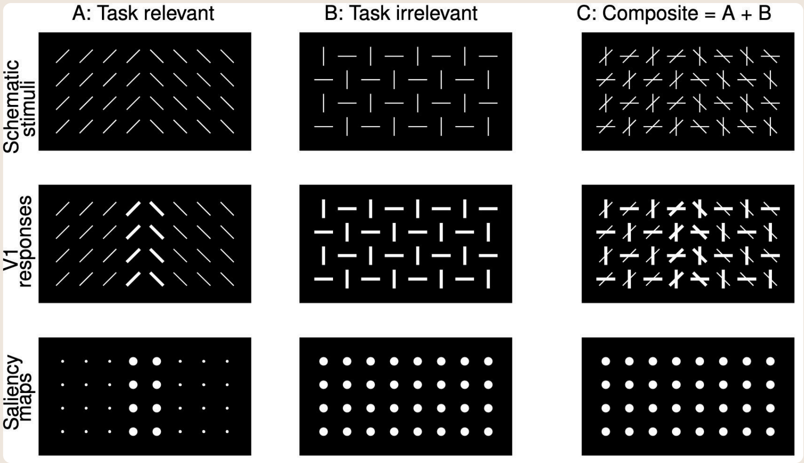
Also, the V1 model

- confirms the intuition that **iso-orientation suppression** is a leading mechanism underlying various saliency effects.
- can signal saliency at locations of *complex shapes* (e.g., circles, crosses)
- illustrates that saliency mechanisms in V1 can have *side effects*, e.g., medial axis effects.

We have done some simulations with a V1 model so far, and more experimental data should be considered to test this theory further.

# 4 Additional psychophysical tests of the V1SH

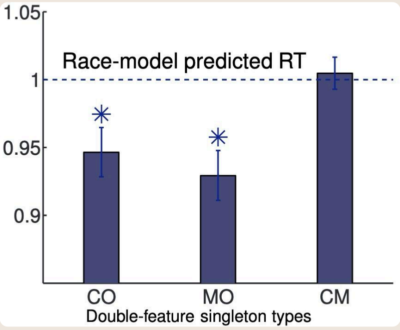
According to the V1SH, the saliency is given by the '*maximum*' rather than the 'summation' of responses, over features at each location.



Psychophysical confirmation of the maximum rule

Considering a visual search task, with a simplistic V1 that has neurons tuned to *color* only and *orientation* only, we have $RT_{CO} = \min(RT_C, RT_O)$. If the V1 has neurons tuned to the *conjunction* of color and orientation, then the prediction should be revised as $RT_{CO} \leq \min(RT_C, RT_O)$. It is called the **race model** prediction.

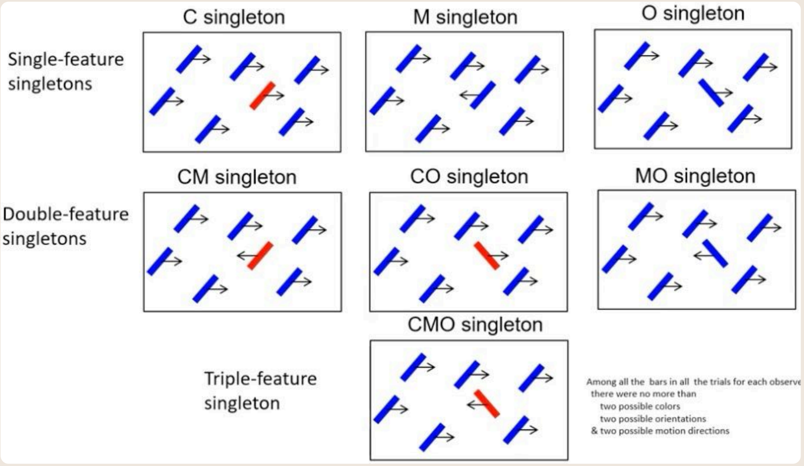
For *color-motion* singleton, the prediction should roughly be $RT_{CM} = \min(RT_C, RT_M)$, as V1 has few cells tuned conjunctively to C and M.
However, we cannot be fully sure whether it is the case, so we turn to triple-feature singletons that more impossible to have neurons tuned to it.



Average normalized RTs

Consider a race model involving C&M&O, we have

$$\min(RT_C, RT_O, RT_M, RT_{CMO}) = \min(RT_{CM}, RT_{CO}, RT_{MO})$$



Among all the bars in all the trials for each observe
there were no more than
two possible colors
two possible orientations
& two possible motion directions

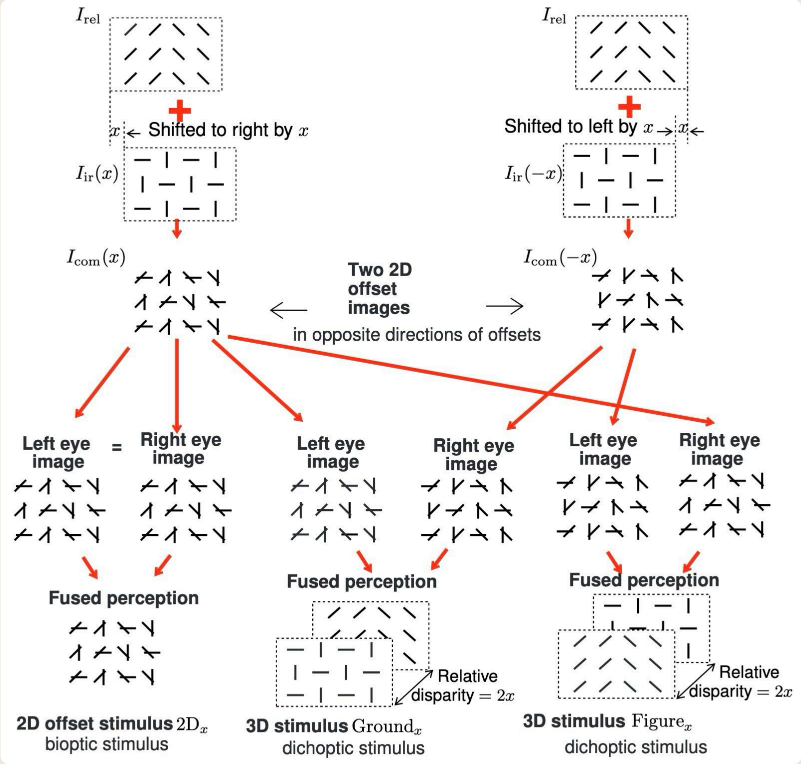
different singletons

> V2 and higher areas have neurons tuned to M&O or the triple features, which means these areas don't contribute to the saliency behavior in such singleton search tasks.

# 5 The roles of V1 and other cortical areas in visual selection

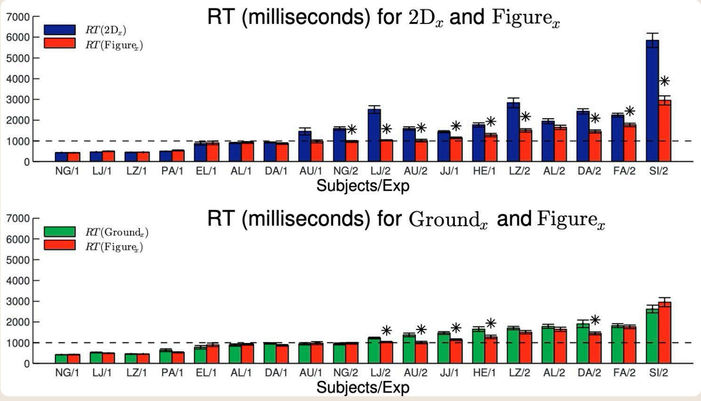
## Using visual depth feature to probe contributions of extrastriate cortex to attentional control

To probe the contribution of the extrastriate cortex, we can use **depth** features, shown by **binocular disparities**. The prediction is that, if the feature-relevant image is at the front, the task should be the easiest, while it is at the back is harder but still easier than the 2D task, i.e., $RT(\text{2D } I_{com}(0)) > RT(\text{3D behind}) > RT(\text{3D front})$.



Constructions of 2D and 3D stimuli

The data shows yes, but only when $RT > 1000ms$. When the actions are fast enough, V1's saliency mechanisms alone guide attention.

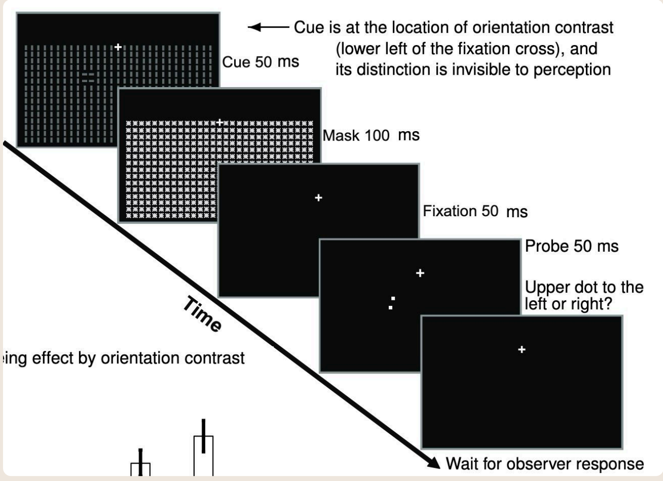


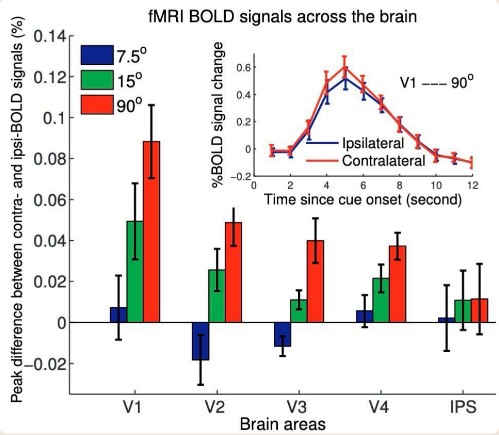
Contributions from 3D visual process to selection

## Salient but indistinguishable inputs activate early visual cortical areas, but not the parietal and frontal areas

Neural activities associated with bottom-up saliency have been observed in LIP (lateral intraparietal area) and FEF (frontal eye field). Does perceptual awareness, more than saliency, contribute to the signal to them?

Stimuli are shown in a very short ($\approx 50ms$) time window, short enough to avoid perceptual awareness. Then test the cuing effect of it to see whether there are some saliency effects. The data said yes.


Cue is at the location of orientation contrast
(lower left of the fixation cross), and
its distinction is invisible to perception
Cue 50 ms
Mask 100 ms
Fixation 50 ms
Probe 50 ms
Upper dot to the
left or right?
Time
ing effect by orientation contrast
Wait for observer response



It is also shown that the saliency difference found in V1 is also found in V2, V3, V4, but not in IPS, FEF, or LGN.

> Pure salient signals without awareness activate early visual cortical areas, but not the parietal and frontal areas, which drive top-down attention.