We appreciate the constructive and detailed review comments suggested by three reviewers. We will revise all typos and proofread the manuscript carefully in a revised version. Followings are our reply to three reviewers' major concerns.

1. Comparison with related works. Earlier works on the computation of opposite neurons (e.g., Zhang et al., eLife 2019; bioRxiv 2018; NIPS 2016) only suggested that opposite neurons compute the likelihood ratio. Here, we derived a mathematically rigorous link using a normative theory, and explained opposite cells' functions in the theoretical framework of causal inference. This we believe is a fundamental, rather than incremental, advance in theoretical neuroscience. In detail, the math derivations of opposite neurons in previous works started directly from a likelihood ratio defined similarly as Eq. (24) in the current paper, which are now grounded on the novel generative model and causal inference as presented in Eqs. (1-23) in the current paper. Secondly, the distribution ratios computed by opposite neurons have distinct physical meanings between current work and earlier works. Now, the integration and segregation models reconstruct their respective best-fit likelihoods over the input directions x (Eq. 16), which are represented by the half of the blue vector (from the origin to the centroid of the parallelogram) and green vectors in Fig. 2B respectively. And opposite neurons compute the ratio between the best-fit likelihoods reconstructed from two models, which are geometrically the two red vectors emanating from the centroid of the parallelogram to either green vector. In contrast, in previous works, opposite neurons compute the posterior ratio representing the disparity of the latent stimulus direction s inferred from each of two cues respectively (Eq. 5 in Zhang 2019), which is geometrically the difference between two green vectors (Fig. 3B in Zhang 2019). Lastly, since the length of the vector representation of the likelihood ratio computed by opposite neurons in current work (Fig. 2B) is half of the one in previous works, in the implementation, opposite neurons average two inputs in our work (Eq. 27), while they sum two inputs together in previous works (e.g., last Eq. on page 19 in Zhang 2019). Taken together, this paper is significant in that we fully developed and established the theoretical link of opposite neurons to causal inference, and illuminate the functional roles of opposite neurons.

As pointed out by reviewer 1, ref. [15] was not able to compute Bayes factor by a linear population code, and we are able to do by including the stimulus strength R in the generative model which leads to two effects. First, the constant Occam factor with respect to inputs simplifies the network implementation (see Sec. 2). Second, the reliability of the best-fit likelihood in the integration model $m_{\rm int}$ ($\hat{\kappa}_{\rm int}/2$ in Eq. 24) is the average of the reliability of two likelihoods ($\hat{\kappa}_{\rm int} \approx \kappa_1 + \kappa_2$ in Eq. 20), because the estimate $\hat{R}_{\rm int}$ is the average of all input spike counts (Eq. 21). In the implementation, this implies the best-fit likelihood of the model $m_{\rm int}$ can be represented by the average of all population inputs, which can be realized by linear operation. On the other hand, without the stimulus strength R as in ref. [15], the reliability of the likelihood ratio (κ_{lp} in Eq. 24) and the Occam factor ratio $\mathrm{OF}(m_{\rm seg})/\mathrm{OF}(m_{\rm int})$ both depend on inputs nonlinearly, hence it is impossible to use linear operations to compute the Bayes factor. Moreover, in a generative model without R, the best-fit likelihood of the model $m_{\rm int}$ is $\mathcal{M}(x_l|\hat{s}_{l,\rm int}, \kappa_l)$, which has the same reliability but different means with the likelihood. This requires a neural circuit to move the location of population inputs (representing the likelihood) in stimulus feature space while keeping the spike count unchanged, which is hard to implement.

2. The implementation of Occam factor. We will discuss the potential implementation of the Occam factor calculation in a biologically plausible neural circuit model in the revised paper. Eq. (22) and line 161-162 states that the Occam factor is a constant invariant to the sufficient statistics of inputs, i.e., x_l and Λ_l ($l = 1, 2$), which indicates the Occam factor ratio can be represented by a fixed parameter in the network, simplifying the network implementation significantly. For example, we can consider a scenario that a downstream neuron whose firing probability represents the logarithm of Bayes factor, i.e., $\ln \mathcal{B}(\mathbf{x}) = \sum_{l=1}^{2} \ln \mathrm{LR}(x_l) + \ln[\mathrm{OF}(m_{\rm seg})/\mathrm{OF}(m_{\rm int})]$. The downstream neuron receives the inputs $\ln \mathrm{LR}(x_l)$ from opposite neurons, and receives a constant background input or has a firing threshold representing the Occam factor ratio $\ln[\mathrm{OF}(m_{\rm seg})/\mathrm{OF}(m_{\rm int})]$.

3. The significance of our work. How neural circuits solve perceptual causal inference is an important and open question in neuroscience (also see reviewer 1's comments). The key contribution of our work is in establishing a theoretical link between the "empirical evidence" of opposite cells found in multi-sensory areas (Fig. 3C-D) and the theoretical framework of causal inference. Our work provides novel insights to the functional roles of opposite cells in causal inference. Thus, the neurophysiological observation of the opposite cells is the "empirical evidence" for the biological plausibility of the Bayes factor computation in neural circuit. We provide the theoretical derivations and simulations to demonstrate how Bayes factor could be computed in opposite neurons, thus linking the normative theory and the underlying neural substrates. More importantly, our theory can provide predictions for verifying the proposed functions of opposite neurons in future experiments. For example, our hypothesis suggests that activating or inactivating opposite neurons' activities would decrease or increase the behavioral choice of integration model (or integration probability) in perception. We will make this prediction more explicit in the Discussion in a revised manuscript.

Related works on causal inference in the field (see references [10, 13, 16]) used generative models with similar structure as our study. Thus, the terminology we used follows the earlier works in the field. We'd like to revise the wording if necessary. We admit that Bayesian model selection (Bayes factor) might not be the only way to solve causal inference, though the existence of the opposite neurons provides strong evidence in supporting it.