We thank the reviewers (R1, R2, R3) for the helpful comments, corrections and suggestions.

The main concern seems to be the **robustness** of our results to various deviations from the idealized scenario considered by our theory (R2, R3). First, while the **hyperparameter** space is too large to explore systematically, simulations suggest the qualitative phenomenology presented in the paper is robust to various model details. The effects of varying the fraction of active/informative neurons are shown in Fig. 2B/3C (we will improve 2B to include all decoders (R2)). We will also document additional parameter variations in the supplementary material (SM). Second, R2 and R3 expressed concern about the precision required for the **optimal modulation weights** in the encoding. Here we show numerically that the results hold qualitatively even when the modulation deviates significantly from the absolute optimal decoding weights ($w = dec_{ML} + \epsilon$ where ϵ is independent gaussian noise; example in panel A) although the overall performance degrades (B). Hence our idea could still apply to a more realistic suboptimal encoding model.

R2 also identified several potential mismatches between our idealized model and real data. First, the modulator could be **multidimensional**. We chose to focus on the unidimensional case because: 1) it is the simplest, 2) mathematically, having a linear combination of gaussian, task-specific modulators does not qualitatively change the problem, though it makes the model harder to parametrize (additionally, if some modulator targeting is not task specific, that would reduce the SNR of all neurons and correspondingly affect all decoders), and 3) while in the Rabinowitz paper there were several modulators, most of variance was accounted for by one (for each hemisphere). Second, the presence of **additive noise**: experimental reports are conflicting (e.g. Goris et al.2014 argue that additive noise is inconsistent with their data); moreover, to date, there is no evidence that this component is functionally targeted. Simulations show that task-invariant additive noise decreases the performance of all decoders, but does not qualitatively change our results (panel C, to be included in SM). These issues will be further detailed in the Discussion.

We chose to focus on classification rather than **estimation** because the experiments showing task-specific modulation use binary discrimination. As R2 rightfully points out, it is important to expand the theory to other tasks. In principle, since estimation also entails learning to appropriately weight informative neurons while ignoring uninformative ones, modulator-labeling should be helpful there too, though the details of the best encoder and decoder will likely change. Including an informative **prior** (R2) should not qualitative change the discussion (it only shifts the threshold), unless the prior directly affects the pattern of modulation. To our knowledge, there is no experimental data supporting this idea.

The model makes several **experimental predictions** (R2), which we are in the process of testing using V1 monkey recordings in a task that dynamically shifts the task-relevant sub-population. The main assumptions of the theory to check: 1) the subpopulation of task-informative neurons is small and hence hard to identify, 2) low-dimensional, shared noise, changing faster than the trial duration, that preferentially targets informative neurons. We further predict that our modulator-guided heuristic decoder should outperform simpler strategies (sign-only or rate-guided). Perhaps counterintuitively, **attention** reduces the variance of the modulator (Rabinowitz et al). Since our theory posits an optimal level of modulation relative to the stimulus-induced variance, we would suggest that attention shifts the level of modulation towards this (empirically estimatable) optimum. The direction of the shift may provide hints about the mechanics of noise generation, something we know little about at the moment (Huang et al. 2019).

There is a small misunderstanding regarding the usage of **MSE** (R1), which we use only for the *modulation weights* estimator, but not for assessing task performance (measured as % correct). Similarly, the procedure used for the **learning** of the MG decoding weights seems unclear (R2). It includes two processes: 1) learning the signs, which happens using end-of-trial feedback, but only needs few examples (see Fig.2A), and 2) learning the absolute optimal decoding weights, which happens within the trial by estimating modulation weights, with variance scaling as $1/T$; the exact learning rate is determined by the strength and time constant of the modulator (see Eq.14). Quantitatively confirming if the modulation is enough to explain the animal's **speed of learning** requires further data analysis (ongoing).

Minor: the explicit reference of the target behavioral 90% level is somewhat misleading and will be removed (R2). We will add the additional references (R2), clarify the equations (R2, R3) and improve Fig.3B visually and for clarification (vertical line=optimum from 3A, middle panel should show precision instead of variance, to reduce confusion) (R1, R2). All the above points will be incorporated in the camera-ready version of the paper, should our submission be accepted.