We would first like to thank the reviewers for their thoughtful commentary and constructive feedback, particularly given the constraints imposed by the COVID-19 pandemic. We are happy to see that the reviewers agree that studying individual differences is an important problem for neuroimaging, and overall appear to lean towards acceptance.

The reviewers do have a number of comments and questions relating to baselines, the importance of characterizing stimulus variation, and the number of available datasets for experiments. Many of these comments are actionable and addressable by camera ready, as we discuss below.

R2: Comparison to MN-SRM, tensor decompositions. We appreciate R2's helpful reference to Shvartsman et al. (2018), which appears to consider spatial variation more explicitly than the original SRM. We were not able to find an open-source implementation of the MN-SRM and as a result were not able to perform a comparison in time for this response. We will attempt to implement this method ourselves and include a comparison in the camera-ready. We will also evaluate whether CP/Tucker tensor decomposition methods could serve as additional baselines, as suggested by R3.

R3+R5: PCA baseline. R3 notes that PCA is not a strong baseline. We agree and it is not intended as such; we will emphasize this more clearly. We appreciate the suggestion from R3 and R5 to perform time-averaging before doing PCA to improve it as a baseline. We did so and found that this did not result in qualitatively different embeddings from the non-averaged analysis in appendix Section A.2. We will update the manuscript and the figures to reflect this change.

R3: How important are the nonlinearities? We have have followed up and trained a version of NTFA in which we replace dense networks with a single linear layer. On our datasets, this results in equivalent reconstruction performance, at the cost of the inferred embeddings looking meaningless (the participant embeddings collapse into the Gaussian prior, and stimulus embeddings lack the interpretable pattern seen in Figure 3).

R5: Cross-validation in MVPA analysis. R5 writes, "feature selection was performed within the cross-validation (as it should be), but NTFA was performed outside." We will clarify that for MVPA, we treat NTFA as an unsupervised feature extractor before any supervised training is performed. This means there is no supervised feature selection. We believe that unsupervised training on all data (which was done as a computational shortcut) is likely not problematic. To verify this, we will re-train NTFA independently for each fold and re-run our analysis.

R2: Notation and dataset tables. We appreciate these suggestions and will include tables that summarize notation, as well as tables with summary statistics for each dataset.

R2: SRM vs TFA. We agree that SRM and TFA as methods have different intended cases (functional alignment and connectivity respectively) and will explain this more clearly in the text. We will also cite Cai et al. (2020).

R3+R5 Why examine stimulus variation? In standard neuroimaging analysis, researchers designate categories of stimuli prior to the experiment and model the BOLD signal with a Gaussian (or similar location-scale) distribution. This approach assumes that variation among stimuli can be treated as noisy deviations from a single mode. However, the assumption that experimenter-designated categories are optimal has been challenged. Cognitive and translational neuroscience suggests that the same category of stimulus-induced state (e.g. fear) or phenotypic trait (e.g. depression), may involve multiple distinct neural pathways. Rather than assume that experimenter-designated categories are correct, NTFA enables researchers to test this assumption. If the stimuli group together into categories naturally, as we found in some of our experiments, then the categorical assumption may be considered well-justified. Where significant intra-category variance appears, NTFA enables researchers to then delve deeper into why stimuli show divergent effects and how that impacts overall conclusions from the findings.

R2+R3+R5: Limitations of NTFA and our present experiments. The three reviewers point out that more interesting evaluation results could be obtained by applying NTFA to further datasets, and that interpretation of participant embeddings has been left to future work. We agree that NTFA is a first step in a longer research program. Participant embeddings in the current model are difficult to interpret because they serve a dual role: they capture both the spatial alignment of latent factors within voxel-space and the per-participant variations in BOLD response to stimuli. The main factor that limits our ability to improve the current model is the availability of datasets that are suitable to the study of individual differences. We are in the process of performing a more comprehensive version of the ThreadVids pilot study, which we hope will allow us to improve upon our current results in future work. We will add discussion to this effect to the manuscript.