We are glad that all reviewers appreciated the soundness of our work, the importance of the hidden stratification (HS) problem we address, and the extensiveness of our evaluations. We thank the reviewers for their thoughtful questions and helpful feedback to improve our paper, and we will incorporate the responses below into our revision.

Computational cost (R4): R4 asks about the important issue of computational cost. As mentioned in App. C.2.4, training with GEORGE takes ≈2-3x as long as training with ERM on the same hardware, as GEORGE first trains an ERM model to obtain a feature representation and then trains a second, robust model. To reduce the cost of GEORGE, one can train the second robust model for fewer epochs, starting from the ERM model instead of from scratch. We also modified our code to use GPU acceleration during the clustering procedure. These changes reduce the runtime of GEORGE to 1.3x that of ERM, while maintaining significant gains in worst-case subclass performance [robust performance] (Table 1). By contrast, simply training an ERM model for longer does not improve robust performance (as also observed in [43]); thus, GEORGE allows one to trade off runtime for (often large) gains in robust performance. With tuning of learning rate schedules and other hyperparameters (HPs), GEORGE's cost could be further reduced.

Causes of HS (R4): R4 asks how GEORGE addresses the different causes of HS. GEORGE primarily focuses on addressing subclass performance gaps that arise from dataset imbalances (unequal fractions of subclasses in the data). As discussed in Sec. 3.2 and App. D.4, we define "inherent hardness" as the *minimum possible* worst-case subclass error that any model can achieve. This may be nonzero due to label noise, insufficient model class expressivity, or insufficient information in the given features to reliably determine the label; by definition, the only way to address subclass performance gaps caused by inherent hardness is to choose a richer model class or improve the data quality.

Origins & robustness of clustering approach (R1): The goal of GEORGE's clustering step is to recover clusters that align with the true subclasses as closely as possible. It should satisfy certain desiderata: 1) auto-select the number of clusters k (as the subclasses may be unknown); 2) be able to identify clusters of very different sizes (as the subclasses may have differing frequencies). 1) motivates our procedure of searching over k and selecting the k yielding the best Silhouette (SIL) score (a metric often used to select k [42]). 2) motivates our "overclustering" procedure described in App. B.3.3: standard methods (e.g. k-means, EM) often fail to identify small clusters, even if they are well-separated, but after overclustering we typically *do* find these. (One could instead simply fix k to a large value; in App. C.2.6 we find that this sometimes works, but has the downside of requiring manual specification, and can spuriously split up larger clusters.) We apply dimensionality reduction (UMAP) as it often improves clustering quality [34] and supports useful visualizations. We thank R1 for asking about this, and will more clearly motivate the design of our clustering procedure in the revision. We hope that building on this method may also be of independent interest. We fixed clustering HPs (e.g. UMAP HPs, overclustering HPs) to be consistent across tasks; tuning per-task would likely further improve performance. Our results are fairly insensitive (no significant performance drop) to reasonable variation in these HPs.

Practical implications of theory (R1): R1 asks about the practical relevance of our theory (e.g. Lemma 1). A key practical takeaway is that if the true data distribution is known, we can estimate the true per-subclass loss R_c by the quantity $\tilde{R}_c$ defined in Lemma 1 (which is computable without requiring subclass labels); Lemma 1 bounds their difference. In practice, the data distribution is typically unknown and must itself be estimated; we use Lemma 2 to deal with this approximation error. Following R1's suggestion, we can empirically validate Lemma 1 for a synthetic mixture-of-Gaussian example (where the data distribution *is* known). From this distribution, we generate varying amounts of datapoints n and then compute $\tilde{R}_c$ and R_c for each subclass; fitting the exponent to our averaged results over several random seeds, we find that $|\tilde{R}_c - R_c|$ converges to 0 at $\approx O(n^{-0.506})$, essentially matching Lemma 1's predicted $O(n^{-0.5})$ rate. We will detail this and additional empirical validations of our theory in the revision.

Additional metrics (R1): We include additional metrics in Table 2 as suggested by R1. In addition to improving robust accuracy, GEORGE improves acc. averaged per-subclass (SCAA); ERM has slightly higher average precision (AP).

		Waterbirds	U-MNIST	CelebA (BiT)
Original results	Robust acc.	82.6	96.1	86.1
	Runtime ratio w.r.t. ERM	2.0	3.2	1.5
Shortened results	Robust acc.	76.4	95.7	86.1
	Runtime ratio w.r.t. ERM	1.3	1.3	1.2
ERM results	Robust acc.	60.4	94.2	41.1

Table 1: Top: Original GEORGE results. (For CelebA, we use BiT embeddings [28], so no ERM model is trained first.) Middle: Modified GEORGE results (fewer epochs in second stage, faster clustering). Bottom: ERM.

Method	Metric	Waterbirds	U-MNIST	CelebA (BiT)
GEORGE	SCAA	89.3	98.4	91.4
	AP	.951	.9986	.883
ERM	SCAA	84.1	98.3	80.8
	AP	.983	.9991	.912

Table 2: Additional classification metrics (ISIC omitted for space). GEORGE improves SCAA, while ERM has higher AP (which is unsurprising as it optimizes for average performance).

Higher-cardinality tasks (R1): R1 asks if GEORGE generalizes to settings with >2 superclasses and/or >2 subclasses per superclass. Our results apply directly to any number of superclasses, and any number of subclasses per superclass. Indeed, the U-MNIST task we evaluate on has 5 subclasses per superclass. Our current results across four datasets provide strong empirical evidence to suggest GEORGE is a promising approach to improve robust performance; we agree that evaluating on even more complex and multiclass datasets is an important area for further work.

Figures (R3, R1): We will include higher-resolution, more readable figures in the revision. We thank R3 for the suggestion to provide a figure to illustrate our overall framework in order to improve clarity; we will also include this.

Typos (R2): We thank R2 for raising the issue of typos; we have since carefully gone through the paper to fix all typos.