We sincerely thank all reviewers for the insightful comments and feedback on our work of learning from failure (LfF).

[R1-1] Multiple bias attributes. One of the strengths of LfF is that it does not require an explicit characterization of bias, enabling the handling of multiple/composite biases by default. Indeed, the CelebA dataset contains a diverse set of attributes that may be spuriously correlated, but LfF performs consistently well, as Table 4 suggests.

[R1-2] Trade-off of debiasing. We *do not* interpret this as a "true" trade-off, as debiasing does not degrade the model's ability to capture the desired correlation; indeed, the performance of LfF is similar on bias-aligned and bias-conflicting sets. Instead, we view the apparent underperformance as a result of "not utilizing a (delusional) spurious correlation."

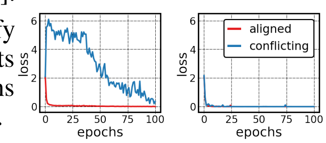
[R1-3] Comparison to SOTA on BAR. Following R1's suggestion, we additionally test ReBias [2] (SOTA among bias-label-free methods) on BAR, using the official code (released after the submission deadline). ReBias achieved 59.73% accuracy, which was lower than 62.98% achieved by LfF; this result will be added to Table 5.

[R2-1] LfF depending on human knowledge. We do not claim LfF to be free of human knowledge and will further clarify this in the final draft. As R2 pointed out, LfF leverages a yes/no type of knowledge: on the given setting, the bias is learned faster than the desired correlation. This is also consistent with our claim that LfF is not "domain-specific" since we followed prevalent use of the word "domain" [2, 22, 29], i.e., groups with the same data types (e.g., natural image) or bias types (e.g., texture). However, this consistency may not hold depending on the definition of "domain."

Hence, we deeply resonate with R2's concern, and we will further clarify the type of knowledge used by LfF and prior work. We will describe (a) LfF as utilizing the aforementioned yes/no type of knowledge and (b) the previous works as utilizing explicit labels or conditioning on the particular bias type, instead of using the term "domain-specific" knowledge. For example, we will modify L2-5 in the abstract by "In this work, we propose a new algorithm utilizing a new yes/no type of knowledge, which does not use explicit labels or presume a particular bias type."

For *empirical evaluation*, we inevitably use additional human knowledge for choosing the attribute pairs, as R2 mentioned. However, we only use the LfF's yes/no type of knowledge for choosing one of the attributes as an undesired bias, e.g., we set `Color` as an undesired bias since it is "easier" than `Digit` for a vanilla model. For *model selection*, we put significant effort into making LfF rely less on human knowledge. Existing hyperparameters, except for GCE parameter q, are obtained from training a vanilla model (using popular architectures) without any prior consideration of the bias. The GCE hyperparameter $q = 0.7$ is simply taken from the original paper [25].

[R2-2] Generalization of the observation. Following R2's suggestion, we further verify our observations' generalizability with the 3D-shapes dataset. We repeat the experiments in Figure 2 using (`ObjectScale`, `WallHue`) attributes (see right). This observation aligns with [27]; CNNs are good at learning local patterns rather than the high-level concepts.



[R2-3] Relation to prior work with combination rules. As an effort to keep the usage of human knowledge minimal, we designed our combination rule without any hyperparameter (in contrast to [2, 28, 29]) though it is not our main contribution. As R2 suggested, we constructed an ablation study on LfF with a combination rule replaced by that of RUBi. Our LfF combination rule achieves 74.01% while that of RUBi achieves 52.41% on the Colored MNIST dataset. We will add more discussions and experiments in the final draft.

[R4-1] Comparison with Group DRO & BAR baseline. While Group DRO assumes and exploits the availability of *worst-case group labels*, LfF achieves a similar or better performance without requiring such additional labels (Tables 2 and 4). This difference makes LfF applicable on datasets without group labels (such as BAR), while Group DRO cannot. Instead of Group DRO, we provide an additional BAR baseline using ReBias [2]; see **[R1-3]**.

[R7-1] Malignancy of bias depending on the algorithm. Our definition of "malignant bias" (see L119) is consistent under the scenarios R7 suggested; we define a spurious correlation as malignant whenever the existence of the correlation leads to a performance degradation *under the given setup*, not as a "global" characteristic.

[R7-2] Alternative word for "easier". We describe the bias as "easier" when it is learned during the early stage, following [1, 31]; still, the expression "learned more thoroughly" which R7 suggested would be appropriate as well.

[R7-3] Alternative schemes for focusing on easy samples. Focal loss is devised to focus on "hard" examples, and thus it cannot replace the GCE loss, which focuses on "easy" examples. Early stopping is not needed for the biased model trained with GCE loss since it remembers the easy examples, even in the later stages (see Figure 4).

[R7-4] Alternative schemes for weighting samples with two networks. We follow the common practice [30] using two networks to distinguish sample groups by leveraging the implicit bias of neural networks. Such practice is known for its high-performance (often better than using one network). We will add the suggested reference in our final draft.

[27] W. Brendel and M. Bethge. Approximating CNNs with bag-of-local-features models ... in ImageNet. ICLR, 2019.
[28] C. Clark, et al. Don't take the easy way out: ... known dataset biases. EMNLP-IJCNLP, 2019.
[29] H. He, et al. Unlearn dataset bias in natural language inference by fitting the residual. EMNLP-IJCNLP, 2019.
[30] J. Li et al. Dividemix: Learning with noisy labels as semi-supervised learning. ICLR, 2020.
[31] S. Sagawa et al. An investigation of why overparameterization exacerbates spurious correlations. ICML, 2020.