We thank all reviewers for their thoughtful feedback. We are encouraged by the overall positive assessment of our work by Reviewers 1–3 (scores 7 7 6), who identified the novelty of our contribution, i.e, bringing to light the important effect of regularization on SGD dynamics, and found that our claims are backed by informative experiments.

The biggest point of criticism, raised by **Reviewer 4**, concerns the novelty of our work. **R4** points to two papers, by two closely related groups of authors:

[1] Poggio, Banburskia, Liao,	PNAS, June 2020.
[2] Liao, Miranda, Hidary, Poggio,	CBMM Memo, July 2018.

The key message of our work is that regularization affects generalization not only by controlling capacity and, thus, conferring stability, but also by its impact on **optimization dynamics:** without it SGD can converge to very poorly generalizing models. **In contrast,** a main claim of [1,2] is that, in the *infinite* time limit and under smoothness assumptions *not* satisfied by ReLU, regularization is unnecessary to SGD, as it converges to good, small norm solutions. Besides the obvious clash, the only connection between [1,2] and our work is that their authors show that one can achieve models with zero classification training loss but increasing amounts of test cross-entropy loss by either using initial weights of increasing variance [1], or by pretraining with an increasing fraction of corrupted labels [2].

In our work we do indeed pretrain with corrupted labels, as in [2], and we will be happy to cite the, unknown to us, CBMM memo in our work. That said, we need to state the following caveats:

1) Reference [2] pretrains with a *mixture* of corrupted and clean labels and shows this leads SGD to minima with worse test **cross-entropy** (not classification) loss. The effect of the drop in test cross-entropy is moderate, i.e., < 0.1 drop; it is difficult to interpret it with regards to classification loss. In contrast, our reported drop is with regards to actual test **classification** loss and is dramatic, i.e., up to 40% degradation. In order to "bury" SGD so deeply:

- We use **completely**, not partially, corrupted labels.
- We employ **data augmentation** on the corrupted examples, i.e., we make several slightly different copies of each example and give it different random labels.
- We train to **full accuracy** on the corrupt data, not just for a fixed number of epochs as [1,2].

Each one of these differences is important in achieving the dramatic drop in generalization that we demonstrate.

2) **More importantly,** the majority of our paper is devoted into analyzing when/how/why SGD can "dig itself out" from bad global minima. [1,2] **say nothing** on the main point of our paper, i.e., the role of regularization in SGD dynamics.

R4 finally alludes that there may be other explanations for our results, e.g., hyperparameter tuning: *"it is unclear if there is something [...] besides the regularization [...] to avoid "bad global minima". How are the hyperparameters [...] adjusted?"* We challenge this: as reported, we do not perform any hyperparameter optimization to help SGD avoid, or get stuck at bad initializers; we solely tune for fast convergence, precisely what is done in general. We find it peculiar that **R4** both missed this and surmises "alternative explanations" in the face of our experimental data.

R1: *"[...] does not offer an explanation as to how different regularization methods help SGD avoid bad global minima."*
Although we don't include this in the paper (but are happy to do so), in the 2D case, all methods seem to equally help SGD escape bad initializers. In the real-data case, none individually leads to SOTA accuracy, but combinations do. We are very interested in understanding the individual effect of these methods, and are conducting relevant experiments.

R2: *"There are a few aspects [...] I would like to see explored more thoroughly. (1) [...] how crucial is the zeroing out of the random subset of pixels? (2) Experiments to more carefully control the learning rate [...] would be very useful.*
(1) Zeroing-out is not crucial and a different technique could be used, e.g., additive noise on pixels, for which similar effects can be observed. (2) Decreasing the LR won't affect the observations, but will lead to slower convergence, while a larger LR won't give SOTA results. We are happy to include these additional experiments for different ranges of LR. As suggested, we will discuss connections with coherent gradients. Thank you!

R3: *"My main concern about this work is the relative lack of discussion and situation in the literature."*
We will significantly expand the discussion on related work along the three most relevant threads: 1) "bad minima exist" and ways to obtain them; 2) the effects of implicit bias; 3) the role of regularization. Due to lack of space here we can't expand on all three above and related literature, but we will do so in the final version of this paper.

R3: *"The paper does not have theoretical results."*
Establishing theory to substantiate our findings would be thrilling but, frankly, these appear far beyond reach, for any meaningful setting. Indeed, per [1], under sufficiently strong assumptions (infinite time, smooth losses), SGD "converges" (in the sense of stationarity) to "wide" minima.

R4: *" For [...] "bad global minima" found in previous works, would regularization help? "*
We ran the experiments and indeed regularization helps massively. The adversarial initialization effect is completely undone.