We thank the reviewers for their informative feedback, indicating improved results (All), that hypotheses are "intuitive" (R4), "convincing and thoroughly justified" (R1), the investigation is "very appropriate"(R4) and "crucial" (R5) for the field, with a "scientifically rigorous", "fair" and "extensive" evaluation (R1,4,5) and 3 of 4 Rs advocating acceptance.

Novelty (R1,R2): Our study indeed builds on previous work including [10,11], however, our work is not merely incremental: As R1 summarized, our manuscript contains multiple independently valuable parts. First, on the conceptual level, we develop and validate a core hypothesis about the anomaly detection failure for generative models on image datasets, explaining it as an effect of model bias and domain prior. We note post-NeurIPS-submission work [1] investigates the same question and comes to consistent conclusions, highlighting that our conceptual contributions address an important research question. Furthermore, to overcome the effects of model bias and domain prior, we use two novel viewpoints (hierarchies of distributions/features) and also derive two novel methods from them. Regarding the first method, while other log-likelihood ratio (LLR) methods have been introduced in [10,11], our LLR method differs in motivation and implementation, as our motivation results in a clear way how to choose the denominator model for the LLR, which we validate in Table 1. Our work provides not only a quantitatively but also qualitatively improved evaluation, as we describe a failure case of [11] that both [10] and [11] did not evaluate (Tab 1). Further, [11] evaluates on the *training* fold of the in-distribution, making their results incomparable due to potential overfitting (see Section 6). Further, our second, novel last-scale-likelihood method shows a completely different way to overcome model bias+domain prior with surprisingly good results, yielding more insight into the anomaly detection problem.

Dependence on Glow Architecture (R1,R2): We purposely kept model architecture and experimental settings similar to prior work to facilitate comparison and focus on the novel ideas. Our LLR method works without adaptation for the autoregressive PCNN++, slightly improving over Glow (Table 1), highlighting our LLR method is not specific to Glow (note that using two separate networks allows to apply LLR to any model). The last-scale method could be evaluated on any model that allows a hierarchical decomposition of the overall likelihood (e.g., VQ-VAE-2).

Applicability of LLR to Other Domains (R2): We believe our LLR method is widely applicable to many domains: (1) In many natural signal domains a suitable general dataset already exists. In this work, we already show this for typical image datasets, and additionally, without adaptation, for medical MRI images. The results of [3] on text/NLP also indicate that a suitable dataset already exists there (Wikitext-2). Since the audio domain also has domain priors like local smoothness, we think LLR to a general audio distribution will also work there, and are happy to include such experiments for the camera-ready. (2) If a suitable general dataset needs to be created, it is important to note that the general distribution does not require labels and may even profit from noisy/unclean data. Therefore there is no principal obstacle preventing collection of such data, including concatenating existing datasets.

Complementarity of the Two Approaches (R1): Our LLR and our last-scale method can be viewed as two independent instances of domain prior removal methods. We do not expect them to benefit each other, as the last scale mostly contains distr.-specific information, so difference subtraction may not help. Influences on their complementarity are an interesting topic for future work, e.g.: How much domain prior info is in last scale of in-dist model? How similar is the domain prior decomposed into Glow model scales for the in-dist and the general model? Do partially joint models help?

Two Separate Networks (R1): Using two networks keeps the LLR method simple and allows to train the general model once for use with different in-distr. models. Still, it is interesting future work to try a joint network (see Discussion p.8).

Correlation Computation (R2): We compute the correlations of different models' likelihoods from all datasets *together* because we analyze relationships *across* datasets for anomaly detection (e.g., else, Fig 3D corr. positive).

Spearman-Correlation Interpretation (R5): One motivation to use the Spearman correlation is that the rank of the likelihoods determine the resulting anomaly detection AUROC value. We redid the analysis with Pearson correlation and find almost identical results, including that likelihoods of a Glow trained on local 8x8 patches have almost perfect correlations with a Glow trained on the full image, even when trained on different datasets. Additionally, earlier scales encode low-level information and account for most of the variance of the overall likelihood (see Section 4 Fig. 3D). That shows local low-level features, beyond being correlated with the likelihood, dominate it. We also confirmed that post-hoc removing low-level features by blurring reduces likelihood correlations (will include in supplementary). Findings of a recent arXiv submission[1] also further reinforce our hypothesis.

Overclaiming wrt MSP-OE (R5): We agree and would modify wording, e.g., to "slightly underperform".

Other Low-Level Features (R5): For convolutional generative models, with few computational steps (our definition of low-level feature) they will extract only local features. Image mean or other global features would require many layers and thus not be low-level for Glow, we will clarify that in Section 2.

More Extensive Related Work (R1,R4, R5): Thanks, we will cite and compare with the work (like BIVA) in revision.

Fig 1A (R2): We will add a missing y-axis-label, i.e., the probability density (of the differences to the local mean across all patches of all images of the entire dataset). As SVHN is locally smooth, its density peaks around zero. Tiny is less smooth than SVHN, so its density is more flattened.

Seeds (R5): Yes, our results are averaged over 3 seeds (same as other exp in Tab S2), will include upon revision.

Other Minor Points (not mentioned due to space limit): We will address them upon revision.

[1] `https://arxiv.org/abs/2006.08545`