We thank all reviewers for their constructive comments.

To Reviewer 1

Q1. *Other inductive biases likely have an effect*: We agree it is possible and will revise Sec. 4 to better reflect this.

Q2. *References on non-generative approaches to outlier detection*: Thanks, we will include them in a revision.

To Reviewer 2

Q1. *Image data are not exactly temporal sequences*: Even though the data is not truly temporal, the conditional distributions $\{p(x_i|x_{<i})\}$ are still well-defined, since the joint distribution $p(\{x_i\}_{i=1}^d)$ is defined. Thus for a random variable (rv) in $\mathbb{R}^d$ following a certain joint distribution (e.g. p_{inlier}), we can always view it as a sequence of d rvs, where the i-th rv is sampled from $p(x_i|x_{<i})$. Subsequently, we can reason about properties of the random sequence, such as IID, MD or WN. We will revise the text to clarify this.

Q2. *How the typicality test is related to IID, WN and MD*: The typicality test is only effective on factorized inlier distributions. So to apply this test, we need to transform the input rv so that the corresponding distribution becomes factorized.[1] Usually the transformation is designed s.t. the resulting rv is also componentwise IID. Such a rv in $\mathbb{R}^d$ can be viewed as a *sequence* of d IID rvs. Thus the typicality test amounts to testing the IID condition on the transformed sequence $T(x)$, using the statistic $\frac{1}{d}\sum_{i=1}^d T_i^2(x)$. It is not related to MD or WN, which are utilized by our test.

Q3. *For general AR models, why do we expect the sequence $R(x)$ to be WN for in-distribution x and not WN for OOD x*: **(i)** We showed that $R(x)$ is always WN when $x \sim p_{inlier}$ in L101-102. Note that we follow the convention in the time series literature, and define WN sequences as uncorrelated sequences with zero mean and unit variance; this definition does not require the sequence to be IID, and is called "weak WN" in some fields. **(ii)** For OOD x, $T(x)$ not being WN is the alternative hypothesis we are interested in. This is because designing a universally effective test is very difficult, if not impossible, given the high dimensionality compared to the limited number of samples. Thus we have to restrict our attention to certain alternative hypotheses (i.e. certain kind of outliers). In this work, we are interested in a variety of natural image outliers which have previously led to confusions. We believe that our alternative hypothesis suits this purpose, as discussed in L121-123 and verified in Appendix A.

Q4. *Test datasets from other domains*: In this work we are mainly interested in natural image outliers, which have previously caused confusion over the calibration of DGMs in literature (also see L21-25, L31-36 in our submission).

To Reviewer 3

Q1. *Differences between the proposed test and the typicality test, and implications in practice (e.g. what happens if we apply the typicality test on non-IID sequences)*: **(i)** They have different assumptions (WN vs IID), and the test statistics are different. **(ii)** The difference in assumptions means that it can be more difficult to apply the typicality test in a principled way, see L77-80. Even if we consider a heuristic application of the typicality test, i.e. to use the test statistics $\frac{1}{d}\sum_i T_i^2(x)$ where $T(x)$ isn't necessarily IID for inlier x, the difference in test statistics means that our test can still be more effective, since it could identify anomalous autocorrelation structures in the (transformed) *outlier* distribution. For example, suppose $T_i(x)$ is WN for inlier x, while for outlier x, $(T_1(x), T_2(x))$ are uniformly sampled from a centered circle with radius $\sqrt{2}$, and for $i > 2$, $T_i(x) = T_{i-2}(x)$. Then for such outliers, $\frac{1}{d}\sum_i T_i^2(x) = 1$, and the typicality test will not be able to detect them; in contrast, our test will detect such outliers, since they have *auto*correlations; see Remark 2.1. The autocorrelation issue is relevant for natural image outliers, as discussed in L121-123 and Appendix A.

Q2. *Results of the typicality test when applied on residuals, or the transformed latents of DGMs*: **(i)** Nalisnick et al (2019) tested it using a flow model (Fig 4(a) therein) and showed that it was not effective. **(ii)** Some methods in Sec 3.1 are equivalent to the typicality test on certain whitened residuals: the LH-2S test using linear models, and the LN-LH2S method in Appendix (Table 5) which works on VAE residuals. Both methods are clearly outperformed by the corresponding WN tests. **(iii)** We also experimented with the typicality test applied to AR-DGM residuals, using the setup of Sec 3.1. The results are similar, with the typicality test outperformed by ours in 5 out of 6 cases.

To Reviewer 4

Q1. *Sample size and CI in Table 1*: For all methods we use the entire test set, except for AR-DGM for which we sample 5×10^4 images from the larger datasets. This leads to a minimum sample size of 10^4 (CIFAR-10 test set). Using the formula R4 provided, we can show that the maximum possible 95% CI is ± 0.011. We will include them in revision.

Q2. *How much does the result change using different AR models*: While we can't train new models due to time constraints, the results in Appendix B are obtained using a smaller-capacity PixelCNN++, and our test still works well. Also note that our test works with a simple linear AR model. Based on these results, it seems reasonable to expect that the results will not change qualitatively as we switch to different models.

[1] This is the approach described in Sec. 2.1. See also L179-181 and footnote 1 in the submission, which discussed an alternative weak typicality test; that test is known to be ineffective, likely due to the lack of any concentration guarantee.