We thank all reviewers for their thorough assessment of our paper.

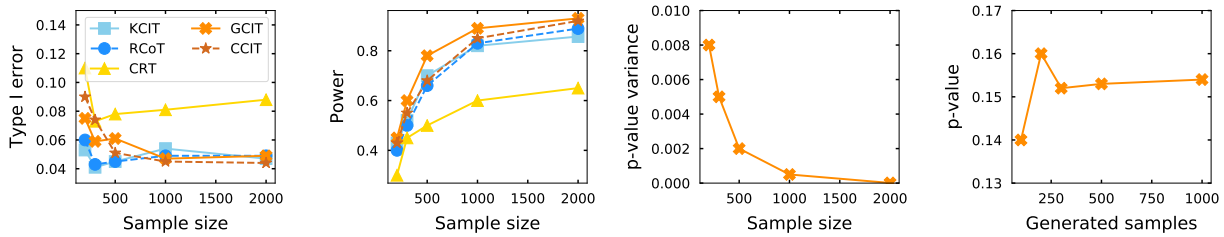

Figure 1: **Leftmost and middle-left panel**: Type I error and power as a function of sample size for data generated under scenario (3) with dimensionality of Z set to 100; **Middle-right panel**: Empirical p-value variance of the GCIT as a function of sample size (computed by generating 100 p-values for each GAN trained on data with the specified size); **Rightmost panel**: Illustration of the convergence of the GCIT's p-values as a function of generated samples.

Response to Reviewer #1.

• *On the robustness of the GCIT for practical applications* - Indeed, our test does depend to some extent on the hyperparameter configurations. However, note that this dependence also exists in alternative tests such as the KCIT and RCoT (e.g. see Figure 1 in [1]), and the CCIT (given that it uses "parametrizable" classifiers). Recall that no ground truth is available to optimize hyperparameters using conventional methods, but we argue that the following procedure can be used to guide hyperparameter selection. We consider artificially inducing conditional independence $(X \perp\!\!\!\perp Y|Z)$ by permuting variables X and Y such as to preserve the marginal dependence in (X, Z) and (Y, Z), as in [2] (further details are also described in our related work). On this data, a well calibrated test is expected to produce uniformly distributed p-values, i.e. the empirical distribution of p-values should be approximately uniform. Our recommendation would be to choose GCIT's hyperparameters with lowest Kolmogorov-Smirnov statistic in comparison to the uniform distribution. This ensures the resulting test produces "well-behaved" p-values and thus prevents to some extent p-value cheating. We will discuss this further in the revised manuscript, thank you for raising this point.

Response to Reviewer #3.

• *On increasing Type I error with λ* - λ determines the influence of $\mathcal{L}_{info}$ in the optimization of the generator (eq. 8). We do discuss the trade-off between power and type I error from a more qualitative, and perhaps intuitive, perspective in Section 3.3. However, insights can also be derived by considering the bound in Theorem 1. Theorem 1 shows that optimal control of the type I error is achieved by optimizing for $\mathcal{L}_G$ in isolation, i.e. $\lambda = 0$. Then, for $\lambda \neq 0$, optimizing for the additional $\mathcal{L}_{info}$ term may converge in practice to a higher $\mathcal{L}_G$, resulting in a higher upper-bound on type I error.

• *On the quality of generated samples and stability of p-values* - We investigate the influence of sample size on the three leftmost panels of Figure 1. The GCIT, as well as most competing tests, have slightly higher type I error in low sample sizes but control type I error successfully with 500 samples or more. In terms of power, our experiments show that we can expect the GCIT to outperform competing tests with 500 samples or more (for dimension of $Z = 100$). Next, we investigate the stability of p-values as a function of sample size; the variance of the empirical p-values quickly drops to 0. This means that for say 500 samples, we can expect the p-values of two independently trained GCITs to be within 0.005 of each other with approximately 95% confidence. The last panel on the right illustrates how quickly the p-value approximation (eq. 3) converges to its population quantity as a function of the number of samples used to compute the approximation i.e. M in eq. 3. The convergence should be at least of order $M^{-1/2}$ by the central limit theorem. As an alternative to a default number of generated samples (previously $M = 1000$), this last experiment led us to modify our implementation to stop sampling from the trained GAN whenever the computed p-value is within $1e^{-3}$ of the mean of the previous 100 computed p-values. This has reduced the computational complexity of the overall procedure while giving better or similar performance, thank you for the suggestion.

Response to Reviewer #4.

• *On the use of GANs as generative models* - From a practical perspective, algorithms based on other generative models can be constructed based on our proposed procedure. However, we chose GANs as they have analytical properties that allow deriving the error bounds in Theorem 1 and enable us to maximize power explicitly with the addition of the information network. In practice, good performance may also be achieved using other flexible generative models. We will mention this as interesting future research.

• *On the quality of generated samples and stability of p-values* - Please kindly refer to the response to Reviewer #3.

[1] Zhang, Kun, et al. "Kernel-based conditional independence test and application in causal discovery." UAI. 2012.
[2] Doran, Gary, et al. "A Permutation-Based Kernel Conditional Independence Test." UAI. 2014.