We are deeply appreciative of reviewers for their feedback amidst these trying circumstances. We are glad that reviewers appreciated the novelty of the approach for a fundamental problem, the rigor of our analysis, simplicity of the approach from an implementation perspective.

[R3] There seems to be a serious misunderstanding, perhaps due to certain terminology used in our paper. In this paper, we are indeed interested in comparing different sampling schemes with respect to the number of samples required. Since we are interested in discrete distributions, most standard sampling schemes, like the two-sample scheme, can be used. But such schemes suffer from an impractical requirement on the number of samples. The tests in hypothesis testing literature operate in a *black-box setting*, wherein we have strong lower bounds [4,36,37] over the required number of samples before a statistically sound conclusion can be drawn. Our setting is not black-box, as pointed out in the Abstract and Introduction. To borrow from folklore in property testing, consider two distributions p and q such that p is the uniform distribution over an N element set S and q is either equal to p or is the uniform distribution over a subset T of S where $|T| = \frac{N}{2}$, that is for every element in T the probability of that element in q is $\frac{2}{N}$. If the set T is unknown, the number of samples one would need to check if q is equal to or far from the uniform distribution p over the set S is at least $\sqrt{N}$. Since N is exponential in the number of bits, the sample complexity has an exponential lower bound. This lower bound is exhibited in [4]. In our setting, since N is exponential in the input size, the sample complexity is also exponential. As mentioned in the Introduction, the usage of conditional sampling allows us to circumvent the hardness, and this work shows how we can handle arbitrary discrete distributions.

We refer the reader to citations [8] and [9] for detailed discussions about property testing literature and the limitations of the standard hypothesis testing setting. While we are aware of the literature on various sampling schemes from hypothesis testing, our paper does not exactly fit in that line of work and instead fits the property testing line of work, as pointed out in the Appendix. Therefore, we did not attempt to cast our results in the notions of null and alternate hypothesis, even though such a characterization is possible but non-standard in property testing literature. Furthermore, we are not sure about the need for such a formulation when Definition 6 and Theorem 1 are self-contained and precise.

[R3] We have access to an Ideal sampler in addition to the weight function (lines 80-85). An explicit description of ground truth distribution in terms of samples is intractable given the large *model count* (see Appendix E2).

[R1, R3] **About the Kernel and the role of randomization.** Kernel, as defined in Definition 7, allows the usage of randomization internally. As discussed on line 229, we use randomization merely to choose the literal on line 6 of Algorithm 2; we will add further discussion to this effect in the final version (We have provided source code with a detailed description). The Kernel in our paper has nothing to do with RKHS. Perhaps we should use a different name to avoid confusion.

[R4] As is the norm, we focused on binary variables, but the techniques work for any arbitrary discrete domain.

[R1, R4]: As a first step to demonstrate the working of the prototype implementation, we focused on log-linear models for which the most efficient technique is to employ the inverse transform. This also allows us to easily obtain a sampler with formal guarantees and two other samplers without guarantees. It is worth noting that STS and variants of UniGen have indeed appeared in AI/ML conferences, so we have taken the first step in addressing the distributions of interest to the AI/ML community. We do agree with [R2] (and also strongly believe) that our algorithmic framework will inspire follow-up work focused on further improvements and a deeper understanding of various complex distributions. Again, our algorithmic analysis is not restricted to log-linear models.

[R1] **Do sampling algorithms satisfy the 'non-adversarial' condition ?** Since conditioning is a fundamental operation in probabilistic reasoning, one would expect that the underlying sampling algorithms support conditioning. Our non-adversariality assumption simply translates to support for conditioning. We will emphasize this further. At the same time, we do agree with the reviewer that one can claim that their sampler supports only one particular distribution, but in such a case, it is hard to see if one can do any better given the strong lower bounds [4,36,37].

[R1] **What does it mean for $\hat{\varphi}$ to have "similar structure" to φ?** $\hat{\varphi}$ is constructed by conjuncting φ with a small formula which allows conditional sampling on the models of φ; thus it can be said that $\hat{\varphi}$ retains the structure of φ.

[R2] **What is the performance like without the constraint obfuscation heuristic, does the naive approach work?** Our preliminary studies showcased the need for constraint obfuscation methodology. We will seek to conduct a longer empirical study and add these results in the final version.

[R4] **The absence of a tight and practical bound.** It is worth emphasizing that Theorem 1 only implies the worst-case behavior of Fulcrum, and lack of tight upper bound does not imply that the performance of Fulcrum would be worse, a claim supported by experimental analysis as well. As a similar example, while there are no bounds better than exponential time for any algorithm to solve SAT, the SAT solvers routinely perform orders of magnitude better than a naive exponential time algorithm. In this view, the design of an efficient algorithm should be viewed as a primary contribution even though its worst-case behavior is hard to analyze.