We thank the reviewers for their thoughtful and constructive feedback, as well as the pointers to related work, which we plan on incorporating in the next revision.

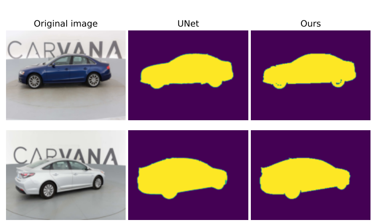
R1 *Comparisons to baselines:* We acknowledge that there is indeed a vast literature of approximate inference methods to compare against. However, as we pointed out in our experiments section, we would like to mention again that Park et al. [24] have provided comparisons and shown that they significantly outperform popular techniques like mean-field approximation, belief propagation, etc. on the binary partition function tasks considered in the paper, which was why we skipped those. Further, the variations in coupling strengths demonstrate that our method works in a variety of temperature settings. For the general k-class MRF, since AIS is a strong sampling-based technique, we focused on running it with different parameter settings to verify the superior performance of our method. However, we would be happy to include other relevant baselines in the general case (e.g. RAISE) in the final version. In our image segmentation experiments, the goal was to demonstrate the capability of our model to actually scale up to real-world MRFs with thousands of nodes, whilst also delivering good results (since the standard k-class relaxation with quadratic constraints would quickly not scale). To put this in context with modern state-of-the-art methods, we provide here a qualitative comparison with a trained UNet (Ronneberger et al. 2015) Figure 1a on the Carvana Image Masking Challenge. As can be observed, the segmentations provided by our method are highly comparable (if not better) to those by UNet.

R2 *Lack of theoretical guarantees and missing references:* We admit the lack of theoretical guarantees for the general k-class case of our algorithm. We remark here that since we are indeed solving a different relaxation with a modified objective and lesser constraints, the analysis of Frieze et al. [9] does not go through as is. But for the special case of $k = 2$, we can provide approximation bounds following $\log \mathbb{E} \exp(x^T Ax) \leq \max x^T Ax$ together with the 0.878 bounds of Goemans-Williamson. We will add the proof for the same in the revision. We remark here that although the classical relaxation (6) does provide theoretical guarantees, solving it becomes infeasible for large MRFs. In contrast, from our empirical results, our relaxation *does* scale up, and we admit that we have traded off a lack of guarantees with practical performance. As for the additional references on discrete integration by hashing pointed out, we will definitely include it (possibly with a comparison on the binary MRF task) in the revised version.

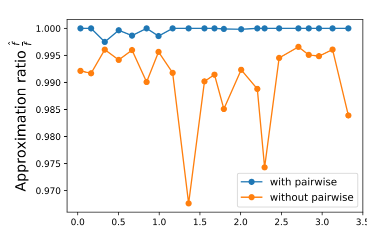
R3 *Tightness of our relaxation:* We acknowledge that our relaxation (11) is indeed looser than the classical k- class relaxation (6). However, due to the quadratic number of pairwise constraints in (6), we remark here again that solving this SDP quickly becomes practically infeasible with increase in the MRF size n. The computational cost to solve this SDP with $\sim n^2$ constraints with a traditional solver would be $O(n^6)$ [1]. Thus, our relaxation, albeit being looser, *does scale up* (as seen in our image segmentation section) to practically large MRFs. Further, we empirically observe that the mode estimates obtained *after randomized rounding* are largely the same on solving the relaxation (6) both with and without the pairwise constraints. This is verified in Figure 1b, where we compute mode estimates on 5-class MRFs by solving (6) both with and without pairwise constraints. Performing randomized rounding on the solutions to both these SDPs deliver mode estimates of practically the same quality across various coupling strengths. Thus, while hugely benefitting in runtime, our relaxation doesn't suffer much at all in quality of solution.

Runtime and complexity-theoretic claims: The naive runtime for segmenting a standard (say 400x400) image by our method (without any GPU parallelization) is roughly ~ 2 minutes. We remark here that performing each segmentation constitutes randomly initializing the v_i vectors and solving (6) via the mixing method as in Algorithm 1, and also performing a few rounds of rounding as in Algorithm 2. However, with parallelization and several optimizations, we believe that there is massive scope for significantly reducing this runtime. We reiterate here that our goal in the segmentation section was to simply demonstrate the capability of our method to scale up. As for the sloppiness in the complexity-theoretic statements in the paper (e.g. naively equating exponential configurations to #P hardness), we will definitely correct them in the final version.

[1] See Sec 6.3.5, Pg. 72 in https://docs.mosek.com/MOSEKModelingCookbook-letter.pdf



(a) Comparison with UNet



(b) Mode estimates w/ and w/o pairwise constraints