We thank the reviewers for the comments, which we will incorporate into the next version. For brevity we denote the reviewers by **[R1][R2][R3][R4]**. We have included additional baselines and ablations in Table 1 (synthetic) and Figure 1 (fuzzing) (described more below). Overall ALOE still performs consistently comparable or better than alternatives.

[R1] Conditional EBM: This extension requires changes only to the parameterizations of energy function, samplers (into $q(x; z)$) without affecting the overall framework. We will elaborate more in our revision.

Table 1: Ablations for ALOE; compared to Table 1 in main paper.

Methods	2sprs	8gauss	cir	moon	pwhl	sroll	ckbd
ALOE	**30.37**	**-0.97**	-0.83	**-0.64**	-0.64	**-0.58**	-1.7
ADE-fac	236.6	65.7	261.7	248.6	187.2	95.3	78.2
ALOE-fac-noEdit	51.24	91.2	5.97	76.8	59.7	15	2.98
ALOE-fac-edit	32.6	3	**-1.5**	1.27	5.02	0.44	**-2.03**
AutoRegressive	32.7	-0.3	-0.8	-0.45	**-1.27**	0.31	-0.2
VAE	35.2	2.09	0.16	1.1	0.85	2.05	-0.77

[R2] ablation on minimizing (7) and local edits: Thanks for the suggestions. We found both were separately helpful through ablations. **a)**To justify the local edits, we use a fully factorized initial q_0, and compare ALOE-fac-noEdit (no further edits) against ALOE-fac-edit (with $\leq$16 edits). ALOE-fac-edit performs much better than the noEdit version. We use a weak q_0 here since we don't need many edits when q_0 is the powerful MLP with no parameter sharing (which is not feasible in realistic tasks). ALOE automatically learns to adapt number of edits, as studied in Fig 3 (left) and Table 2 (right) in main paper. **b)**We also show (7) achieves better results than the REINFORCE objective from ADE [ref 8 in paper], when we compare ADE-fac that uses the same sampler as ALOE-fac-noEdit.

[R2] Table 1 results All methods are evaluated against the same held-out test set.
[R2] Edit-distance bias: We agree with the reviewer. Our experiments show that the bias is not a big issue in practice. If necessary, this bias can be removed: For learning the EBM, we care only about the distribution over end states, and we have the freedom to design q, so we could limit q to generate only shortest paths.
[R2] Use RNN like EBM for fuzzing: As suggested, we include RNN-EBM in Fig 1, which uses RNN as score function and is otherwise the same as our setting. It is indeed better than prefix based sampling, but is still inferior to ALOE in general.
[R2] EBM baselines on other tasks: For program synthesis we mainly evaluate the effect of local edits in our sampler, so the other methods are not applicable; for fuzzing we here include ADE and CD (it is a conditional EBM and PCD's buffer is not directly applicable). From the results in Fig 1 we can see ALOE still outperforms baselines consistently. CD is comparable on `libpng` but for large target like `openjpeg` it performs much worse. ADE performs good initially on some targets but gets worse in the long run. This is due to the lack of diversity, which suggests a potential mode drop problem that is common in REINFORCE based approaches.
[R2] "Clarity: Theorem 1 seems unnecessary": Thanks for your suggestion. Theorem 1 is needed to motivate the "variational form of power method" in Algorithm 2 and in (7). We will make this more clear in our revision.
[R2] Minor "...drawbacks of autoregressive...imprecise": Fair point. We agree that autoregressive models can also be used in a way like EBM during inference, but EBMs can be more general and thus more powerful. We will appropriately weaken the claims. Also thanks for suggestions on typos and notations. We will address.

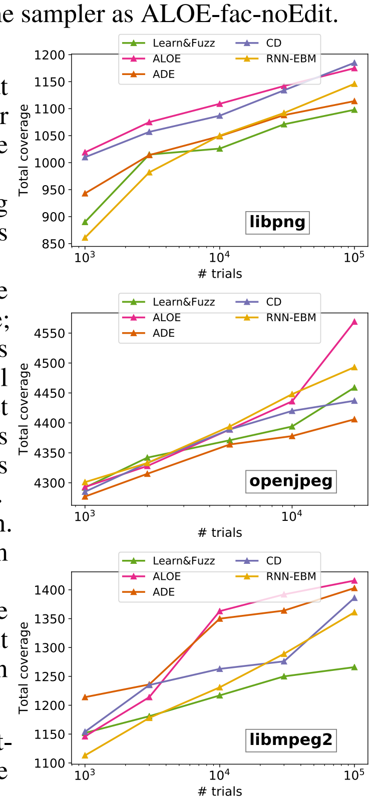
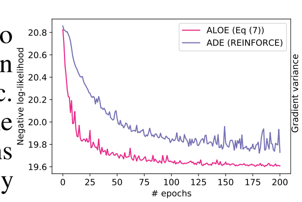


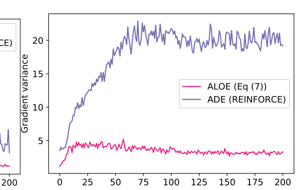
Figure 1: More fuzzing results.

[R3] "...toy-ish domains..." We emphasize that fuzzing is done on real-world softwares with large sample size (see Table A.1 in appendix), where `libfuzzer` baseline is used in commercial. We will explore more application domains in the future.
[R3][R4] other models on toy data: The main purpose of synthetic experiment is to compare different learning methods for the *same* EBM. Nevertheless, we have included autoregressive (with LSTM) and VAE models (with MLP) in Table 1 as suggested. ALOE still performs the best overall. But note that EBMs and the VAE/autoregressive ones use different models and sampling methods.

[R3] "...evaluation...heuristic..." Likelihood is not tracable to compute in EBMs, while using MMD to measure distribution discrepancies is a common protocol rather than a random heuristic.
[R3] "...tricks...domain specific" It is common to serialize the trees (like we used for program synthesis in the paper) and graphs (e.g., SMILES language). Edit-distance can also be defined directly on trees (e.g., `gumtree`) and graphs (`GED`).



Negative Log-likelihood



Gradient Variance

[R4] "... complicated.. variance of REINFORCE" we have included ablations above to justify our design. Regarding the variance, we plot the gradient variance and learning objective during training (estimated via importance sampling) for pinwheel data. We can clearly see ALOE enjoys lower variance than REINFORCE based methods for EBMs.