Table 1: Performance of the relevance and substitution networks of the prover on `iset.mm` validation data.

Human proofs	Synthetic proofs	Generator	Relevance				Substitution		Test proofs found (903 in total)
			Top-1	Top-5	Top-20	MRR	Prob	Accuracy	
7125	0	-	43.27	69.57	89.68	0.5535	0.1723	49.45	378
7125	1M	*MetaGen-IL*	45.10	71.00	89.46	0.5699	0.2554	57.81	398

We thank all reviewers for their thoughtful comments. As suggested by R3, we experimented on another Metamath theorem sets, `iset.mm`, and the results are presented in the table above. Individual questions are addressed below.

R1–There is not that much novelty in the paper. Our novelty is that we propose to generate synthetic theorems by forward reasoning as training data for the theorem prover. While most of prior work uses training data from the proofs (or their variants) of human-written theorems, we demonstrate that the synthetic proofs, which do not contribute to proving human-written theorems, can also be effectively used to advance the prover on human-written theorems. This implies that to train a powerful theorem prover, not only should the prover learn to remember the proofs of given human-written theorems, but also learn to reason about the entire space of potential theorems defined by the existing theorems. We believe this is an important direction that worth more exploration in the AI/TP community.

R1–Give more discussion on the limitations of the proposed approach. We list three conditions to apply our approach to other formal systems in Sec. 4.3, which is one aspect of the limitations of our approach. Another limitation is that we assume a set of human-written theorems as inputs. We expect to explore in the future work the generation of plausible and meaningful synthetic theorems from the basic definitions and axioms only.
R1–The part on Metamath not clearly written. We will clarify the part of Metamath as you suggested.
R1–It's a bit overclaiming by saying that your work is 'orthogonal' to all previous approaches. We believe our main idea is orthogonal to prior work. Our idea is to generate synthetic theorems as training data for the prover, which can also be applied to prior work of neural theorem proving. But we will tone down our text and further clarify.
R1–Any degradation of performance due to overfitting? No. We believe our synthetic theorems do not cause overfitting for two reasons: (1) synthetic theorems are randomly sampled. (2) every new theorem has a unique proof that is different from existing proofs and it contains a novel path of deduction that is not covered by existing data.

To R2: Thank you for pointing us to a list of related work. We discuss how our work differs from these related works below. We will add these discussions to our next revision and revise any inaccurate statements.
R2–Prior work on iterative learning and reasoning. In iterative learning and reasoning, machine-proofs are generated for existing human theorems and are used to train the prover. That is, only new proofs are synthesized and the new proofs are only for existing human theorems, but no new theorems are synthesized. In contrast, our approach synthesize both new theorems and new proofs. Our synthetic theorems and their proofs could cover a much larger space of possible derivations than the proofs of existing human theorems.
R2–Prior work ([3,4,5] from R2) on generating synthetic proof tasks. [3,4,5] extract synthetic proof tasks from the proofs of human-written theorems, such as the intermediate steps or their variants. That is, they extract "sub-proofs" from existing proofs. In contrast, we generate entirely new theorems and new proofs that are not part of any existing proofs.
R2–Prior work on conjecturing. Conjecturing targets on finding meaningful math theorems automatically. Generated conjectures could be either true or false and their proofs are not required. In contrast, each of our synthetic theorem is guaranteed to be correct and its proof is automatically available.
R2–The claim that Holophrasm is "state-of-the-art" should be softened. Thank you for the suggestion. We will soften this language.

R3–Limited experimental evaluations. We have now experimented on another Metamath theorem sets `iset.mm` which formalizes 9371 theorems in intuitionistic logic. As shown in Tab. 1, the prover trained with both human-written data and synthetic data performs better than the Holophrasm baseline, which is consistent with our results on `set.mm`.
R3–Achieving state of the art is not impactful unless controlling computational cost. Our claim of SOTA is against our re-implementation of the baseline in the GPU environment. Thus the computation cost is already controlled for. We will further clarify our claim and release our code such that the future work could compare with us in the GPU environment.
R3–Evaluate the effectiveness of the proposed constraints in Ln. 178 These constraints (that limit what can be invocable theorems) are proposed to simplify our proposed method and to reduce computational cost. Without them, training the prover will be much more costly because there will be more invocable theorems. We are not able to complete this experiment for this rebuttal but will add it to our next revision.
R3–The detail of the substitution network is not clear. We will clearify the substitution network by adding examples.