Theorem-Aware Geometry Problem Solving with Symbolic Reasoning and Theorem Prediction

Pan Lu[1*], **Ran Gong**[1*], **Shibiao Jiang**[2*], **Liang Qiu**[1], **Siyuan Huang**[1],
Xiaodan Liang[3], **Song-Chun Zhu**[1]
[1]UCLA, [2] Zhejiang University, [3]Sun Yat-sen University

Abstract

Geometry problem solving is challenging as it requires abstract problem understanding and symbolic reasoning with axiomatic knowledge. However, current datasets are either small in scale or not publicly available. Thus, we construct a new large-scale benchmark, Geometry3K, consisting of 3,002 geometry problems with dense annotation in formal language. We further propose a novel geometry solving approach with formal language and symbolic reasoning, called *Interpretable Geometry Problem Solver* (Inter-GPS). Inter-GPS first parses the problem text and diagram into formal language automatically via rule-based text parsing and neural object detecting, respectively. Unlike implicit learning in existing methods, Inter-GPS incorporates theorem knowledge as conditional rules and performs symbolic reasoning step by step. Also, a theorem predictor is designed to infer the theorem application sequence fed to the symbolic solver for the more efficient and reasonable searching path. Extensive experiments on the Geometry3K and GEOS datasets demonstrate that Inter-GPS achieves significant improvements over existing methods. The project is available at https://lupantech.github.io/inter-gps.

1 Introduction

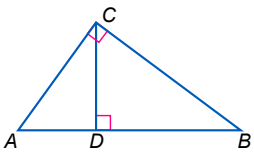
Geometry problem solving is a long-standing challenging task in artificial intelligence and has been gaining more attention in the NLP community recently [28, 14, 25]. Psychologists and educators believe that solving geometric problems requires high-level thinking abilities of symbolic abstraction and logical reasoning [6, 21]. However, if algorithms take the raw problem content, it might encounter challenges to understand the abstract semantics and perform human-like cognitive reasoning for inferring the answer in the geometry domain. Inspired by the ability of a formal language to specify rules and logic in the fields of linguistics and mathematics, we propose to parse the problem inputs into formal language descriptions (see Figure 1) before solving the problems.

Existing methods [29, 26, 27] highly depend on human annotations like symbols in diagrams as the intermediate results, or fail to provide the explicit reasoning processes when predicting the answer. Besides, most current datasets are either small in scale or not publicly available [29, 27], which further hinders the research of geometry problem solving. To overcome these challenges, we first construct a new large-scale benchmark, called Geometry3K, which consists of 3,002 multi-choice problems. In contrast with existing work, we also annotate each problem text and diagram with unified structural descriptions in formal language.

This paper further presents a novel geometry solving approach with formal language and symbolic reasoning, called *Interpretable Geometry Problem Solver* (Inter-GPS). Inter-GPS (Figure 2) develops an automatic parser that translates the problem text via template rules and parses diagrams by a neural object detector into formal language, respectively. In contrast to parameter learning, Inter-GPS

[*]Equal contribution.

The 35th Conference on Neural Information Processing Systems (NeurIPS 2021) Workshop on Math AI for Education (MATHAI4ED).



In triangle *ABC*, *AD* = 3 and *BD* = 14. Find *CD*.
Choices:
A. 6.0 B. 6.5 C. 7.0 D. 8.5
Answer: B

Diagram Formal Language

```
Triangle(A,B,C)
Triangle(A,C,D)
Triangle(B,C,D)
PointLiesOnLine(D,Line(A,B))
Perpendicular(Line(A,C),Line(B,C))
Perpendicular(Line(C,D),Line(A,B))
```

Text Formal Language

```
Triangle(A,B,C)
Equals(LengthOf(Line(A,D)),3)
Equals(LengthOf(Line(B,D)),14)
Find(LengthOf(Line(C,D)))
```

Figure 1: An example in Geometry3K dataset. Each data is annotated with formal language.

formulates the geometry solving task as problem goal searching, and incorporates theorem knowledge as conditional rules to perform symbolic reasoning step by step. Also, we design a theorem predictor to infer the possible theorem application sequence for the efficient searching path. Experiments on Geometry3K and GEOS show Inter-GPS achieves large improvements over existing methods.

2 Geometry3K Dataset

The Geometry3K dataset consists of 3,002 problems, which are collected from two high school textbooks. To the best of our knowledge, currently, it is the largest geometry problem dataset (Table 1). In addition to four elementary shapes (lines, triangles, regular quadrilaterals, and circles) mentioned in that GEOS dataset, Geometry3K contains irregular quadrilaterals and other polygons. Besides, in Geometry3K, there are more unknown variables and operator types that may require equation solving to problem solving. Note that less than 1% of the problems in Geometry3K could be solved when the diagram is not provided. See Appendix C for details of data collection and more data analysis.

Dataset	#qa	#word	#shape	#goal	#var	grade	operator type
GeoShader [2]	102	/	4	1	1	6-10	$\{+, -, \times, \div, \square^2, \sqrt{\square}\}$
GEOS [29]	186	4,343	4	3	1	6-10	$\{+, -, \times, \div, \square^2, \sqrt{\square}\}$
GEOS++ [26]	1,406	/	4	3	1	6-10	$\{+, -, \times, \div, \square^2, \sqrt{\square}\}$
GEOS-OS [27]	2,235	/	4	3	1	6-10	$\{+, -, \times, \div, \square^2, \sqrt{\square}\}$
Geometry3K (ours)	3,002	36,736	6	4	3	6-12	$\{+, -, \times, \div, \square^2, \sqrt{\square}, \sin, \cos, \tan\}$

Table 1: Comparison of our Geometry3K dataset with existing datasets.

3 Geometry Problem Parser

3.1 Text Parser

Given the word sequence of the problem text T, the text parser needs to translate it into a set of literals L_t, a sequence composed of predicates and variables. Inspired by previous works [16, 29, 4] that indicate the rule-based parsing method is able to obtain precise parsing results, we apply this approach with regular expressions to perform text parsing. Semantic parsers [31, 33, 9] using sequence-to-sequence (Seq2Seq) learning methods are not feasible to generate satisfactory literals in Geometry3K for two reasons. Firstly, the limited scale of geometry datasets weakens these highly data-driven methods. Secondly, neural semantic parsers tend to bring noises in generated results while geometry solvers with symbolic reasoning are sensitive to such deviations.

3.2 Diagram Parser

Diagrams provide complementary geometric information that is not mentioned in the problem text. Different from previous works [28, 29] that require manual annotations and fail to deal with special relational symbols such as *parallel*, *perpendicular*, and *isosceles*, we develop an automatic diagram parser to detect varied diagram symbols. The diagram parser first applies Hough Transformation [30] to extract geometry primitives (points, lines, arcs, and circles), following [29]. Then the diagram symbols and text regions are extracted through a strong object detector RetinaNet [18], and the textual content is further recognized by the optical character recognition tool MathPix[1]. After obtaining the primitive set P and symbol set S, we need to ground each symbol with its associated primitives. [29]

[1]https://mathpix.com/

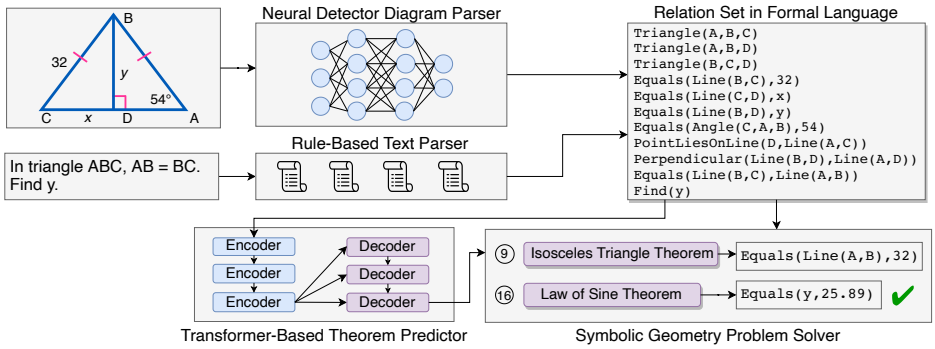


Figure 2: Framework of Inter-GPS. It parses the inputs into a relation set in formal language, and applies the theorem sequence from the theorem predictor to infer the answer via symbolic reasoning.

adapts a greedy approach where each symbol is assigned to the closest primitive without considering its validity. Instead, we formulate the grounding task as an optimization problem with the constraint of geometry relations:

$$\begin{aligned} &\min \sum_{s} dist(s_i, p_j) \times \mathbb{1}\{s_i \text{ assigns to } p_j\} \\ &\text{s.t. } (s_i, p_j) \in \text{Feasibility set } F, \end{aligned} \tag{1}$$

where the $dist$ function measures the Euclidean distance between the symbol s_i and primitive p_j. F defines the geometric constraints for symbol grounding.

4 Geometry Problem Solver

Unlike existing methods [29, 26, 2, 25], Inter-GPS achieves the explicit symbolic reasoning with the theorem knowledge base and the human-readable search process, shown in Figure 2.

4.1 Symbolic Geometry Solver

Inter-GPS takes the relation set $\mathcal{R}$ and the theorem knowledge base set $\mathcal{KB}$ as inputs, and outputs the numeric solution g^* of the problem goal g. The relation set $\mathcal{R}$ defines geometry attributes and relations in the given problem, and is initialized with literals from the text and diagram parsers. $\mathcal{R}$ is further expanded with literals that are derived from definitions of geometry shapes. For example, a triangle is defined as three connected sides. So if there is a *literal* `Triangle(A,B,C)`, six more *literals* (`Ponit(A)`, `Ponit(B)`, `Ponit(C)`, `Line(A,B)`, `Line(B,C)`, `Line(C,A)`) will be appended to $\mathcal{R}$. The theorem set $\mathcal{KB}$ is represented as a set of theorems, where each theorem k_i is written as a conditional rule with a premise p and a conclusion q. For the search step t, if the premise p of k_i matches the current relation set $\mathcal{R}_{t-1}$, the relation set is updated according to the conclusion q:

$$\mathcal{R}_t \leftarrow k_i \wedge \mathcal{R}_{t-1}, k_i \in \mathcal{KB}. \tag{2}$$

After the application of several theorems, equations between the known values and the unknown problem goal g are established, and g could be solved after solving these equations:

$$g^* \leftarrow \textsc{SolveEquation}(\mathcal{R}_t, g). \tag{3}$$

4.2 Theorem Predictor (TP)

As the geometry problems in Geometry3K are collected from high school textbooks, it might need to apply multiple theorems before the problems are solved. Intuitively, one possible search strategy is to use brute force to enumerate candidates in the theorem set randomly. The random search strategy is inefficient and might lead to problems unsolvable as there might be applications of complicated theorems in the early stage. Therefore, an ideal geometry problem solver can solve the problems using reasonable theorem application sequences. Students with good academic performance can solve a problem with prior knowledge learning from a certain amount of problem solving training. Inspired by this phenomenon, we propose a theorem predictor to infer the possible theorem application sequence for inference after multiple attempts on the train data.

Method	All	Angle	Length	Area	Ratio	Line	Triangle	Quad	Circle	Other
Random	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0
Human	56.9	53.7	59.3	57.7	42.9	46.7	53.8	68.7	61.7	58.3
Human Expert	90.9	89.9	92.0	93.9	66.7	95.9	92.2	90.5	89.9	92.3
Q-only	25.3	29.5	21.5	28.3	33.3	21.0	26.0	25.9	25.2	22.2
I-only	27.0	26.2	28.4	24.5	16.7	24.7	26.7	30.1	30.1	25.9
Q+I	26.7	26.2	26.7	28.3	25.0	21.0	28.1	32.2	21.0	25.9
RelNet [5]	29.6	26.2	34.0	20.8	41.7	29.6	33.7	25.2	28.0	25.9
FiLM [22]	31.7	28.7	32.7	**39.6**	33.3	33.3	29.2	33.6	30.8	29.6
FiLM-BERT [9]	32.8	32.9	33.3	30.2	25.0	32.1	32.3	32.2	34.3	33.3
FiLM-BART [17]	33.0	32.1	33.0	35.8	50.0	34.6	32.6	37.1	30.1	37.0
Inter-GPS (ours)	**57.5**	**59.1**	**61.7**	30.2	**50.0**	**59.3**	**66.0**	**52.4**	**45.5**	**48.1**
Inter-GPS (GT)	78.3	83.1	77.9	62.3	75.0	86.4	83.3	77.6	61.5	70.4

Table 2: Evaluation results by our method and compared baselines on the Geometry3K dataset.

As there are no annotated theorem application sequences in Geometry3K, we randomly sample from the theorem set multiple times to generate the application sequences. A generated sequence is regarded as positive if the solver Inter-GPS solves the problem after the application of that sequence. A positive sequence with the minimum length for a problem is seen as pseudo-optimal. Finally, we collect 1,501 samples with the problem and its pseudo-optimal theorem application sequence.

Given the problem formal description $L = \{l_1, ..., l_m\}$, the theorem predictor aims to reconstruct the pseudo-optimal theorem sequence $T = \{t_1, ..., t_n\}$ token by token. We formulate the generation task as a sequence-to-sequence (Seq2Seq) problem and use a transformer-based model [17] to generate theorem sequence tokens. Specifically, the transformer decoder predicts the next theorem order t_i given $T = \{t_1, ..., t_i\}$. The Seq2Seq model is trained to optimize the negative log-likelihood loss: $\mathcal{L}_{\text{TP}} = -\sum_{i=1}^{n} \log p_{\text{TP}}\left(t_i \mid t_1, \ldots, t_{i-1}\right)$, where p_{TP} is the parametrized conditional distribution in the theorem predictor model.

4.3 Low-first Search Strategy

After the application of the theorem sequence predicted by the theorem predictor, it is likely that Inter-GPS still could not find the problem goal. Generally, humans incline to use simple theorems first when solving math problems to reduce complex calculations. If simple theorems are not tangible, they will turn to more complex theorems. On account of that, we apply an efficient search strategy with heuristics driven by subject knowledge. We categorize theorems into two groups: **lower-order** theorem set $\mathcal{KB}_1$ and **higher-order** theorem set $\mathcal{KB}_2$. The lower-order set $\mathcal{KB}_1$ (e.g, *Triangle Angle-Sum Theorem*, *Congruent Triangle Theorem*) only involves two simple operations of addition and subtraction, while $\mathcal{KB}_2$ (e.g, *Law of Sines*) requires complex calculations.

In each following search step after using predicted theorems, we first enumerate theorems in the lower-order set $\mathcal{KB}_1$ to update the relation set $\mathcal{R}$: $\mathcal{R}_t \leftarrow k_i \wedge \mathcal{R}_{t-1}, k_i \in \mathcal{KB}_1$. If lower-order theorems fail to update $\mathcal{R}$ anymore, higher-order theorems are considered to update $\mathcal{R}$: $\mathcal{R}_t \leftarrow k_i \wedge \mathcal{R}_{t-1}, k_i \in \mathcal{KB}_2$. The search process stops once we find the problem goal g or the search steps reach the maximum steps allowed. The whole search algorithm for Inter-GPS is presented in Algorithm 1.

5 Experiments

Table 2 compares the accuracy results of Inter-GPS with baselines on the Geometry3K dataset. Benefiting from symbolic reasoning with theorem knowledge, our Inter-GPS obtains an overall accuracy of 57.5%, significantly superior to all neural baselines. Inter-GPS even attains a better accuracy compared to human beings. Inter-GPS with ground truth formal language gains a further improvement of 20.8%. Inter-GPS also obtains state-of-the-art performance on the GEOS dataset, as shown in Table 11. Please refer to Appendix D for experiment settings and more result analysis.

6 Conclusion

In this paper, we propose a large-scale benchmark, Geometry3K, which consists of 3,002 high-school geometry problems with dense descriptions in formal language. We further propose a novel geometry solving approach, Inter-GPS, which parses the problem as formal language automatically and performs symbolic reasoning over the theorem knowledge base to infer the answer. Experiment results show that Inter-GPS outperforms existing state-of-the-art methods by a large margin.

References

[1] Chris Alvin, Sumit Gulwani, Rupak Majumdar, and Supratik Mukhopadhyay. Synthesis of geometry proof problems. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2014.

[2] Chris Alvin, Sumit Gulwani, Rupak Majumdar, and Supratik Mukhopadhyay. Synthesis of solutions for shaded area geometry problems. In *The Thirtieth International Flairs Conference*, 2017.

[3] Aida Amini, Saadia Gabriel, Peter Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. Mathqa: Towards interpretable math word problem solving with operation-based formalisms. In *Proceedings of Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, 2019.

[4] Mohit Bansal, Kevin Gimpel, and Karen Livescu. Tailoring continuous word representations for dependency parsing. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 809–815, 2014.

[5] Trapit Bansal, Arvind Neelakantan, and Andrew McCallum. Relnet: End-to-end modeling of entities & relations. *arXiv preprint arXiv:1706.07179*, 2017.

[6] Mohan Chinnappan. Schemas and mental models in geometry problem solving. *Educational Studies in Mathematics*, 36(3):201–217, 1998.

[7] Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014.

[8] Shang-Ching Chou, Xiao-Shan Gao, and Jing-Zhong Zhang. Automated generation of readable proofs with geometric invariants. *Journal of Automated Reasoning*, 17(3):325–347, 1996.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT (NAACL)*, 2018.

[10] Wenbin Gan and Xinguo Yu. Automatic understanding and formalization of natural language geometry problems using syntax-semantics models. *International Journal of Innovative Computing, Information and Control*, pages 83–98, 2018.

[11] Wenbin Gan, Xinguo Yu, Ting Zhang, and Mingshu Wang. Automatically proving plane geometry theorems stated by text and diagram. *International Journal of Pattern Recognition and Artificial Intelligence*, 33(07):1940003, 2019.

[12] Feng Han and Song-Chun Zhu. Bottom-up/top-down image parsing by attribute graph grammar. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 2, pages 1778–1785. IEEE, 2005.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[14] Mark Hopkins, Ronan Le Bras, Cristian Petrescu-Prahova, Gabriel Stanovsky, Hannaneh Hajishirzi, and Rik Koncel-Kedziorski. Semeval-2019 task 10: Math question answering. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 893–899, 2019.

[15] Danqing Huang, Shuming Shi, Chin-Yew Lin, and Jian Yin. Learning fine-grained expressions to solve math word problems. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 805–814, 2017.

[16] Terry Koo, Xavier Carreras, and Michael Collins. Simple semi-supervised dependency parsing. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 595–603, 2008.

[17] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7871–7880, 2020.

[18] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017.

[19] Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. In *The 35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks*, 2021.

[20] Takuya Matsuzaki, Takumi Ito, Hidenao Iwane, Hirokazu Anai, and Noriko H Arai. Semantic parsing of pre-university math problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2131–2141, 2017.

[21] Andi Saparuddin Nur and Evy Nurvitasari. Geometry skill analysis in problem solving reviewed from the difference of cognitive style students junior high school. *Journal of Educational Science and Technology (EST)*, 3(3):204–210, 2017.

[22] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 32, 2018.

[23] Jinghui Qin, Lihui Lin, Xiaodan Liang, Rumin Zhang, and Liang Lin. Semantically-aligned universal tree-structured solver for math word problems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3780–3789, 2020.

[24] Subhro Roy and Dan Roth. Mapping to declarative knowledge for word problem solving. *Transactions of the Association for Computational Linguistics (TACL)*, 6:159–172, 2018.

[25] Mrinmaya Sachan, Avinava Dubey, Eduard H Hovy, Tom M Mitchell, Dan Roth, and Eric P Xing. Discourse in multimedia: A case study in extracting geometry knowledge from textbooks. *Computational Linguistics*, 45(4):627–665, 2020.

[26] Mrinmaya Sachan, Kumar Dubey, and Eric Xing. From textbooks to knowledge: A case study in harvesting axiomatic knowledge from textbooks to solve geometry problems. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 773–784, 2017.

[27] Mrinmaya Sachan and Eric Xing. Learning to solve geometry problems from natural language demonstrations in textbooks. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics*, pages 251–261, 2017.

[28] Minjoon Seo, Hannaneh Hajishirzi, Ali Farhadi, and Oren Etzioni. Diagram understanding in geometry questions. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2014.

[29] Minjoon Seo, Hannaneh Hajishirzi, Ali Farhadi, Oren Etzioni, and Clint Malcolm. Solving geometry problems: Combining text and diagram interpretation. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 1466–1476, 2015.

[30] Linda G Shapiro and George C Stockman. *Computer vision*. Prentice Hall, 2001.

[31] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *arXiv preprint arXiv:1409.3215*, 2014.

[32] Kewei Tu, Meng Meng, Mun Wai Lee, Tae Eun Choe, and Song-Chun Zhu. Joint video and text parsing for understanding events and answering queries. *IEEE MultiMedia*, 21(2):42–70, 2014.

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*, 30:5998–6008, 2017.

[34] Lei Wang, Yan Wang, Deng Cai, Dongxiang Zhang, and Xiaojiang Liu. Translating a math word problem to an expression tree. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

[35] Lei Wang, Dongxiang Zhang, Jipeng Zhang, Xing Xu, Lianli Gao, Bing Tian Dai, and Heng Tao Shen. Template-based math word problem solvers with recursive neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 33, pages 7144–7151, 2019.

[36] Wu Wen-Tsun. Basic principles of mechanical theorem proving in elementary geometries. *Journal of Automated Reasoning*, 2(3):221–252, 1986.

[37] Xinguo Yu, Mingshu Wang, Wenbin Gan, Bin He, and Nan Ye. A framework for solving explicit arithmetic word problems and proving plane geometry theorems. *International Journal of Pattern Recognition and Artificial Intelligence*, 33(07):1940005, 2019.

[38] Song-Chun Zhu and David Mumford. A stochastic grammar of images. *Foundations and Trends in Computer Graphics and Vision*, 2(4):259–362, 2006.

Appendix

A Related Work

Datasets for Geometry Problem Solving. Several datasets for geometry problems have been released in recent years. These include GEOS [29], GEOS++ [26], GeoShader [2], GEOS-OS [27], as well as IconQA[19], a newly proposed dataset for icon diagram understanding. However, these datasets are relatively small in scale and contain limited problem types. For example, there are only 102 shaded area problems in GeoShader and 186 problems in GEOS. While GEOS++ and GEOS-OS contain more data of 1,406 and 2,235 problems, respectively, they have not been publicly available yet. Instead, our Geometry3K dataset features 3,002 SAT-style problems collected from two high-school textbooks that cover diverse graph and goal types. Besides, each problem in Geometry3K is annotated with dense descriptions in formal language (defined in Section B), which makes it particularly suited for symbolic reasoning and interpretable problem solving. In order to promote follow-up work in the geometry domain, we release the dataset and evaluation baselines.

Approaches for Geometry Problem Solving. Due to the sparsity of appropriate data, most early works on automated geometry systems focus on geometry theorem proving [36, 8, 37, 11], problem synthesis [1], diagram parsing [28], as well as problem formalization [10]. [29] attempt using computer vision and natural language processing techniques to solve geometry problems with problem understanding. However, the system does not perform explicit reasoning with axiomatic knowledge as it reduces the task to an optimization problem to see which choice can satisfy all constraints. Some recent efforts [26, 25] have been made to incorporate theorem knowledge into problem solving. They feed geometry axioms written as horn clause rules and declarations from the diagram and text parser into logical programs in prolog style to solve the problem. However, these methods fail to provide human-readable solving steps. And parameter learning on horn clause rules and built-in solvers leads to an uncontrollable search process. In contrast, our proposed Inter-GPS implements explicit symbolic reasoning to infer the answer without the help of candidate answers in an interpretable way. Also, a theorem predictor and a low-order first search strategy are employed to provide human-like theorem application steps.

Interpretable Math Problem Solving. Due to the intrinsic requirements of symbolic understanding and logical reasoning, interpretability of solvers plays an essential role in geometry problem solving. While the interpretability of geometry problem solvers is rarely explored, some pioneering work has been proposed in the general math problem solving domain. Broadly there are two main lines of achieving interpretable solving steps for math problems. The first generates intermediate structural results of equation templates [15, 35], operational programs [3] and expression trees [34, 23]. The second line of work with a higher level of interpretability translates the math problems into symbolic language and conducts logical reasoning iteratively to predict the final results [20, 24]. Furthermore, inspired by work on semantic parsing [12, 38, 32], we claim structured diagram parsing and joint semantic representations for text and diagrams is critical in interpretable geometry problem solving.

B Geometry Formal Language

A geometry problem P is defined as a tuple $(t, d, \mathbf{c})$, in which t is the input text, d is the diagram image and $\mathbf{c} = \{c_1, c_2, c_3, c_4\}$ is the multiple-choice candidate set in the format of numerical values. Given the text t and diagram d, an algorithm is required to predict the correct answer $c_i \in \mathbf{c}$. We formally describe the problem in the geometric domain language Ω, a set of literals composed of predicates and arguments. Basic terms used in the geometry problem solver are defined as follows.

Definition 1. A *predicate* is a geometric shape entity, geometric relation, or arithmetic function.

Definition 2. A *literal* is an application of one *predicate* to a set of arguments like variables or constants. A set of *literals* makes up the semantic description from the problem text and diagrams in the formal language space Ω.

Definition 3. A *primitive* is a basic geometric element like a point, a line segment, a circle, or an arc segment extracted from the diagram.

We define 91 *predicates* and their corresponding *literal* templates in the geometry language domain. For development, these *predicates* are categorized into six groups: geometric shapes (Table 3), unary

geometric attributes (Table 4), general geometric attributes (Table 5), binary geometric relations (Table 6), A-IsXOf-B-type geometric relations (Table 7), as well as numerical attributes and relations (Table 8). Moreover, `$` in the literal templates denotes the undetermined shape.

#	Predicates	Literal templates
1	Point	Point(A),Point($)
2	Line	Line(A,B),Line(m),Line($)
3	Angle	Angle(A,B,C),Angle(A),Angle(1),Angle($)
4	Triangle	Triangle(A,B,C),Triangle($),Triangle($1,$2,$3)
5	Quadrilateral	Quadrilateral(A,B,C,D),Quadrilateral(1),Quadrilateral($)
6	Parallelogram	Parallelogram(A,B,C,D),Parallelogram(1),Parallelogram($)
7	Square	Square(A,B,C,D),Square(1),Square($)
8	Rectangle	Rectangle(A,B,C,D),Rectangle(1),Rectangle($)
9	Rhombus	Rhombus(A,B,C,D),Rhombus(1),Rhombus($)
10	Trapezoid	Trapezoid(A,B,C,D),Trapezoid(1),Trapezoid($)
11	Kite	Kite(A,B,C,D),Kite(1),Kite($)
12	Polygon	Polygon($)
13	Pentagon	Pentagon(A,B,C,D,E),Pentagon($)
14	Hexagon	Hexagon(A,B,C,D,E,F),Hexagon($)
15	Heptagon	Heptagon(A,B,C,D,E,F,G),Heptagon($)
16	Octagon	Octagon(A,B,C,D,E,F,G,H),Octagon($)
17	Circle	Circle(A),Circle(1),Circle($)
18	Arc	Arc(A,B),Arc(A,B,C),Arc($)
19	Sector	Sector(O,A,B),Sector($)
20	Shape	Shape($)

Table 3: 20 predicates and corresponding literal templates for geometric shapes.

#	Predicates	Literal templates
1	RightAngle	RightAngle(Angle($))
2	Right	Right(Triangle($))
3	Isosceles	Isosceles(Polygon($))
4	Equilateral	Equilateral(Polygon($))
5	Regular	Regular(Polygon($))
6	Red	Red(Shape($))
7	Blue	Blue(Shape($))
8	Green	Green(Shape($))
9	Shaded	Shaded(Shape($))

Table 4: 9 predicates and corresponding literal templates for unary geometric attributes.

#	Predicates	Literal templates
1	AreaOf	AreaOf(A)
2	PerimeterOf	PerimeterOf(A)
3	RadiusOf	RadiusOf(A)
4	DiameterOf	DiameterOf(A)
5	CircumferenceOf	CircumferenceOf(A)
6	AltitudeOf	AltitudeOf(A)
7	HypotenuseOf	HypotenuseOf(A)
8	SideOf	SideOf(A)
9	WidthOf	WidthOf(A)
10	HeightOf	HeightOf(A)
11	LegOf	LegOf(A)
12	BaseOf	BaseOf(A)
13	MedianOf	MedianOf(A)
14	IntersectionOf	IntersectionOf(A,B)
15	MeasureOf	MeasureOf(A)
16	LengthOf	LengthOf(A)
17	ScaleFactorOf	ScaleFactorOf(A,B)

Table 5: 17 predicates and corresponding literal templates for general geometric attributes .

C Geometry3K Dataset

C.1 Dataset Collection

Most existing datasets for geometry problem solving are relatively small, contain limited problem types, or not publicly available. For instance, the GEOS dataset [29] only contains 186 SAT problems.

#	Predicates	Literal templates
1	PointLiesOnLine	PointLiesOnLine(Point($),Line($1,$2))
2	PointLiesOnCircle	PointLiesOnCircle(Point($),Circle($))
3	Parallel	Parallel(Line($),Line($))
4	Perpendicular	Perpendicular(Line($),Line($))
5	IntersectAt	IntersectAt(Line($),Line($),Line($),Point($))
6	BisectsAngle	BisectsAngle(Line($),Angle($))
7	Congruent	Congruent(Polygon($),Polygon($))
8	Similar	Similar(Polygon($),Polygon($))
9	Tangent	Tangent(Line($),Circle($))
10	Secant	Secant(Line($),Circle($))
11	CircumscribedTo	CircumscribedTo(Shape($),Shape($))
12	InscribedIn	InscribedIn(Shape($),Shape($))

Table 6: 12 predicates and corresponding literal templates for binary geometric relations.

#	Predicates	Literal templates
1	IsMidpointOf	IsMidpointOf(Point($),Line($))
2	IsCentroidOf	IsCentroidOf(Point($),Shape($))
3	IsIncenterOf	IsIncenterOf(Point($),Shape($))
4	IsRadiusOf	IsRadiusOf(Line($),Circle($))
5	IsDiameterOf	IsDiameterOf(Line($),Circle($))
6	IsMidsegmentOf	IsMidsegmentOf(Line($),Triangle($))
7	IsChordOf	IsChordOf(Line($),Circle($))
8	IsSideOf	IsSideOf(Line($),Polygon($))
9	IsHypotenuseOf	IsHypotenuseOf(Line($),Triangle($))
10	IsPerpendicularBisectorOf	IsPerpendicularBisectorOf(Line($),Triangle($))
11	IsAltitudeOf	IsAltitudeOf(Line($),Triangle($))
12	IsMedianOf	IsMedianOf(Line($),Quadrilateral($))
13	IsBaseOf	IsBaseOf(Line($),Quadrilateral($))
14	IsDiagonalOf	IsDiagonalOf(Line($),Quadrilateral($))
15	IsLegOf	IsLegOf(Line($),Trapezoid($))

Table 7: 15 predicates and corresponding literal templates for A-IsXOf-B-type geometric relations.

#	Predicates	Literal templates
1	SinOf	SinOf(Var)
2	CosOf	CosOf(Var)
3	TanOf	TanOf(Var)
4	CotOf	CotOf(Var)
5	HalfOf	HalfOf(Var)
6	SquareOf	SquareOf(Var)
7	SqrtOf	SqrtOf(Var)
8	RatioOf	RatioOf(Var),RatioOf(Var1,Var2)
9	SumOf	SumOf(Var1,Var2,...)
10	AverageOf	AverageOf(Var1,Var2,...)
11	Add	Add(Var1,Var2,...)
12	Mul	Mul(Var1,Var2,...)
13	Sub	Sub(Var1,Var2,...)
14	Div	Div(Var1,Var2,...)
15	Pow	Pow(Var1,Var2)
16	Equals	Equals(Var1,Var2)
17	Find	Find(Var)
18	UseTheorem	UseTheorem(A_B_C)

Table 8: 18 predicates and corresponding literal templates for numerical attributes and relations.

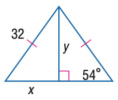
Although there are 1,406 problems in GEOS++ [26], this dataset has not been released to the public yet. Therefore, we build a new large-scale geometry problem benchmark, called Geometry3K. The data is collected from two popular textbooks for high school students across grades 6-12 by two online digital libraries (McGraw-Hill[2], Geometryonline[3]). Groups of well-trained annotators with undergraduate degrees manually collect each problem with its problem text, geometry diagram, four candidate choices, and correct answer. In order to evaluate the fine-grained performance of geometry solvers, we label each problem data with the corresponding problem goal and geometry shapes.

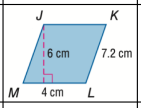
Unlike existing datasets that only collect the problem text and diagrams, we further annotate each data in Geometry3K with dense formal language descriptions that bridge the semantic gap between

[2]https://www.mheducation.com/

[3]www.geometryonline.com

Problem Text	Diagram	Choices	Text Literals	Diagram Literals

Find y. Round to the nearest tenth.		A. 18.8
		B. 23.2
		C. 25.9
		D. 44.0
		Answer: C
			`Find(y)`	`Equals(LengthOf(Line(A,B)),32)`
				`Equals(LengthOf(Line(B,D)),y)`
				`Equals(MeasureOf(Angle(A,C,B)),54)`
				`Equals(LengthOf(Line(A,D)),x)`
				`PointLiesOnLine(D,Line(A,C))`
				`Perpendicular(Line(B,D),Line(C,D))`
				`Equals(LengthOf(Line(A,B)),LengthOf(Line(B,C)))`

Find the perimeter of $\parallelogram$ JKLM.		A. 11.2
		B. 22.4
		C. 24
		D. 44.8
		Answer: B
			`Find(PerimeterOf(Parallelogram(J,K,L,M)))`	`Equals(LengthOf(Line(L,K)),7.2)`
				`Equals(LengthOf(Line(M,L)),4)`
				`Equals(LengthOf(Line(E,J)),6)`
				`PointLiesOnLine(E,Line(M,L))`
				`Perpendicular(Line(J,E),Line(E,L))`

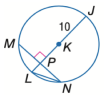
In $\odot$ K, MN = 16 and m $\widehat$ MN = 98. Find the measure of LN. Round to the nearest hundredth.		A. 6.93
		B. 7.50
		C. 8.94
		D. 10.00
		Answer: C
			`Circle(K)`
			`Equals(LengthOf(Line(M,N)),16)`
			`Equals(MeasureOf(Arc(M,N)),98)`
			`Find(LengthOf(Line(L,N)))`
				`Equals(LengthOf(Line(J,K)),10)`
				`Perpendicular(Line(P,K),Line(M,P))`
				`PointLiesOnLine(P,Line(M,N))`
				`PointLiesOnLine(P,Line(L,J))`
				`PointLiesOnLine(P,Line(L,K))`
				`PointLiesOnLine(K,Line(P,J))`
				`PointLiesOnLine(K,Line(L,J))`
				`PointLiesOnCircle(M,Circle(K))`
				`PointLiesOnCircle(J,Circle(K))`
				`PointLiesOnCircle(N,Circle(K))`
				`PointLiesOnCircle(L,Circle(K))`

Figure 3: More data examples in the Geometry3K dataset.

the textual and visual contents as well as benefit the symbolic problem solver. The annotated formal language is used to train and evaluate our proposed problem parsers. Data examples are illustrated in Figure 3.

C.2 Dataset Statistics

The Geometry3K dataset is divided into the train, validation, and test sets with the ratio of 0.7:0.1:0.2, as shown in Table 9. Figure 4 illustrates the question distribution by the number of sentence words. The long tail in the distribution requires the geometry solvers to understand the rich semantics in the textual content.

	Total	Train	Val	Test
Questions	3,002	2,101	300	601
Diagrams	3,002	2,101	300	601
Sentences	4,284	2,993	410	881
Words	30,146	20,882	2,995	6,269
Literals	33,506	24,200	3,001	6,305
Literals (Text)	6,293	4,357	624	1,312
Literals (Diagram)	27,213	19,843	2,377	4,993

Table 9: Basic statistics of Geometry3K.

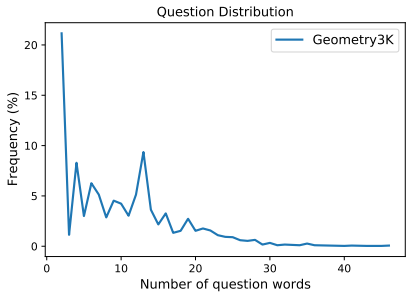


Figure 4: Question length distribution.

There are 6,293 literals for the problem text and 27,213 literals for the diagrams in Geometry3K, respectively. We list the most and least frequent predicates with a frequency greater than 5 in Table 10. It is shown that the predicates for the problem text are more evenly distributed than those for diagrams. This is mainly because the problem text describes diverse geometric shapes, attributes, and relations while diagrams display the basic properties of points, lines, and arcs.

Predicates (Text)	%	Predicates (Diagram)	%
`Find`	19.00	`Line`	30.89
`Line`	14.49	`PointLiesOnLine`	16.66
`Equals`	11.83	`Equals`	15.17
`LengthOf`	9.53	`MeasureOf`	10.46
`MeasureOf`	8.97	`LengthOf`	8.69
.....		.....	
`CircumscribedTo`	0.05	`Triangle`	0.03
`SumOf`	0.04	`Quadrilateral`	0.02
`HeightOf`	0.04	`Kite`	0.01
`BaseOf`	0.04	`HeightOf`	0.01
`IsHypotenuseOf`	0.04	`Square`	0.01

Table 10: Most and least frequent predicates of formal descriptions in Geometry3K (frequency >5).

C.3 Human Performance

As an intellectual task, it is necessary to know the human performance for geometry problems. We push the test-split data of the dataset in the crowdsourcing platform, Amazon Mechanical Turk[4]. Each eligible annotator must have obtained a high school or higher degree and is asked to answer 10 problems in 25 minutes. To ensure annotators solving the problem to the best of their ability, they are further asked to spend at least 7 minutes on the problem set and 10 seconds on each problem. We filter out annotators who do not satisfy the requirement. We also ask dozens of graduates majoring in science or engineering to answer these problems to evaluate human experts' performance. Table 2 shows the human performance. Compared to random guess's accuracy of 25%, humans achieve an overall accuracy of 56.9%, and human experts can achieve a good performance of 90.9%.

D Experiments

D.1 Experimental Settings

Datasets and evaluation metrics. We conduct experiments on the Geometry3K and GEOS [29] datasets. The Geometry3K dataset involves 2,101 training data, 300 validation data, and 601 test data, respectively. The GEOS dataset provides 55 official SAT problems for evaluating geometry solvers. Regarding our proposed Inter-GPS model, if the one closest to the found solution among the four choices is exactly the ground truth, the found solution is considered correct. For a fair comparison, if Inter-GPS fails to output the numeric value of the problem goal within allowed steps, it will randomly choose the one from the four candidates. In terms of compared neural network baselines, the predicted answer has a maximum confidence score among choice candidates.

Baselines. We implement several deep neural network baselines for geometry solvers to compare them with our method. By default, these baselines formalize the geometry problem solving task as a classification problem, fed by the text embedding from a sequence encoder and the diagram representation from a visual encoder. *Q-only* only encodes the problem text in the natural language by a bi-directional Gated Recurrent Unit (Bi-GRU) encoder [7]. *I-only* only encodes the problem diagram by a ResNet-50 encoder [13] as the input. *Q+I* uses Bi-GRU and ResNet-50 to encode the text and diagram, respectively. *RelNet* [5] is implemented for embedding the problem text because it is a strong method for modeling entities and relations. *FiLM* [22] is compared as it achieves effective visual reasoning for answering questions about abstract images. *FiLM-BERT* uses the BERT encoder [9] instead of the GRU encoder, and *FiLM-BART* uses the recently proposed BART encoder [17].

Implementation details. Main hyper-parameters used in the experiments are shown below. For our symbolic solver, a set of 17 geometry theorems is collected to form the knowledge base. For generating positive theorem sequences, each problem is attempted by 100 times with the maximum sequence length of 20. The transformer model used in the theorem predictor has 6 layers, 12 attention heads, and a hidden embedding size of 768. Search steps in Inter-GPS are set up to 100. For the neural solvers, we choose the Adam optimizer and set the learning rate as 0.01, and the maximum epochs are set as 30. Each experiment for Inter-GPS is repeated three times for more precise results.

D.2 Results on the GEOS Dataset

Inter-GPS also obtains state-of-the-art performance on the GEOS dataset over existing geometry solvers, as shown in Table 11.

Method	Acc (%)
GEOS [29]	49
GEOS++ [29]	49
GEOS-OS [27]	52
GEOS++AXIO [26]	55
Inter-GPS (ours)	**67**

Table 11: Evaluation results on the GEOS dataset.

[4] https://www.mturk.com/

D.3 Ablation Study and Discussion.

Search strategies. The overall accuracy and average steps needed for solving problems with different search strategies in Inter-GPS are reported in Table 12. *Predict* refers to the strategy that uses the theorems from the theorem predictor followed by a random theorem sequence. The strategy largely reduces the average steps to 6.5. The final strategy in Inter-GPS applies the predicted theorems first and lower-order theorems in the remain search steps, and gains the best overall accuracy.

Search strategies	Accuracy (%)	# Steps
Random	75.5 ± 0.2	13.2 ± 0.1
Low-first	77.3 ± 0.3	15.1 ± 0.2
Predict	77.5 ± 0.1	**6.5** ± 0.1
Predict+Low-first (final)	**78.3** ± 0.1	7.1 ± 0.1

Table 12: Performance of Inter-GPS with different search strategies.

Problem parsers and literal sources. The rule-based text parser achieves an accuracy of 97% while only 67% for the semantic text parser. Table 13 reports the Inter-GPS performance fed with different sources of literals. With literals generated from our problem solver, Inter-GPS achieves an accuracy of 57.5%. The current text parser performs very well as there is only a slight gap between Inter-GPS with generated text literals and ground truth literals. An improvement of 17.5% for Inter-GPS with annotated diagram literals indicates that there is still much space to improve for the diagram parser.

	Diagram w/o	Diagram	Diagram (GT)
Text w/o	25.0 ± 0.0	46.6 ± 0.7	58.7 ± 0.2
Text	25.4 ± 0.0	57.5 ± 0.2	75.0 ± 0.6
Text (GT)	25.4 ± 0.0	58.0 ± 1.7	**78.3** ± 0.1

Table 13: Performance of Inter-GPS with predicted and ground truth (GT) literals.

Searching step distribution. Figure 5 compares correctly solved problem distribution by the average number of search steps in different strategies. Our final Inter-GPS applies the *Predict+Low-first* strategy, with which 65.97% problems are solved in two steps and 70.06% solved in five steps.

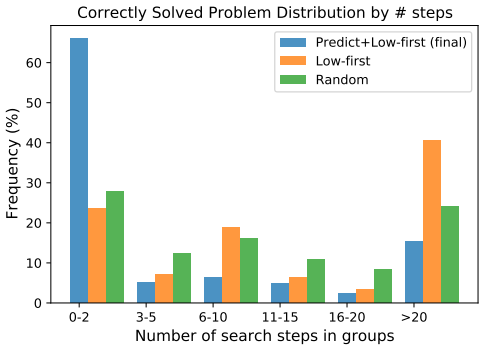


Figure 5: Correctly solved problem distribution by the number of search steps.

Neural geometry solvers. Current neural network baselines for geometry solving fail to achieve satisfactory results in the Geometry3K dataset. It is because there are limited data samples for these neural methods to learn meaningful semantics from the problem inputs. Besides, dense implicit representations might not be suitable for logical reasoning tasks like geometry problem solving. We replace the inputs of problem text and diagram in the *Q+I* baseline with the ground truth textual and visual formal annotations and report the result in Table 14. An improvement of 9.2% indicates the promising potential for neural network models for problem solving if structural representations with rich semantics are learned.

Failure cases. Inter-GPS might not find a solution because of inaccurate parsing results and the incomplete theorem set. Figure 6 illustrates some failure examples for Inter-GPS. For example, diagram parsing tends to fail if there are ambiguous annotations or multiple primitives in the diagram.

	Diagram (visual)	Diagram (formal)
Text (natural)	26.7	35.3
Text (formal)	34.6	35.9

Table 14: Neural solver performance with different representations of the problem text and diagrams.

It is difficult for the text parser to handle nested expressions and uncertain references. And the symbolic solver is still not capable of solving complex problems with combined shapes and shaded areas in the diagrams.

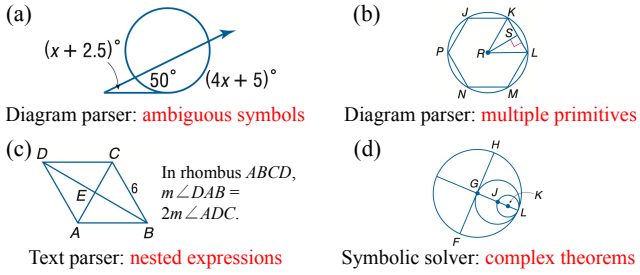


Figure 6: Failure examples for Inter-GPS.

Interpretability in Inter-GPS. Inter-GPS provides an interpretable symbolic solver for geometry problem solving. First, Inter-GPS parses the problem contents into a structural representation of formal language. Second, Inter-GPS performs symbolic reasoning to update the geometric relation set explicitly. Last, Inter-GPS applies reasonable theorems sequentially in the search process.

Algorithm 1 Symbolic Geometry Solver

Input Literals $\mathcal{L}$, goal g, knowledge bases $\mathcal{KB}_1, \mathcal{KB}_2$
Output Numeric goal value g^* and theorem application $\mathcal{S}$

function SEARCH($\mathcal{L}$, g, $\mathcal{KB}_1$, $\mathcal{KB}_2$)
  Initialize relation set $\mathcal{R}_0$ with $\mathcal{L}$, $g^* = \emptyset$, $\mathcal{S} = \emptyset$
  $\mathcal{KB}_p \leftarrow$ THEOPREDICTOR($\mathcal{L}$) ▷ Predicted
  for $k_i \in \mathcal{KB}_p$ **do**
    $\mathcal{R}_t \leftarrow k_i \wedge \mathcal{R}_{t-1}$
    $\mathcal{S}$.APPEND(k_i)
  end for
  $g^* \leftarrow$ SOLVEEQUATION($\mathcal{R}_t, g$)
  if $g^* \neq \emptyset$ **then**
    return g^* and S
  end if
  while $g^* = \emptyset$ and $\mathcal{R}_t$ is updated **do**
    for $k_i \in \mathcal{KB}_1$ **do** ▷ Lower-order
      $\mathcal{R}_t \leftarrow k_i \wedge \mathcal{R}_{t-1}$
      $\mathcal{S}$.APPEND(k_i)
      $g^* \leftarrow$ SOLVEEQUATION($\mathcal{R}_t, g$)
      if $g^* \neq \emptyset$ **then**
        return g^* and S
      end if
    end for
    for $k_i \in \mathcal{KB}_2$ **do** ▷ Higher-order
      $\mathcal{R}_t \leftarrow k_i \wedge \mathcal{R}_{t-1}$
      $\mathcal{S}$.APPEND(k_i)
      $g^* \leftarrow$ SOLVEEQUATION($\mathcal{R}_t, g$)
      if $g^* \neq \emptyset$ **then**
        return g^* and S
      end if
    end for
  end while
end function