UniGeo: Unifying Geometry Logical Reasoning via Reformulating Mathematical Expression

Jiaqi Chen[1,3], **Tong Li**[2], **Jinghui Qin**[4], **Pan Lu**[5],
Liang Lin[1], **Chongyu Chen**[6], **Xiaodan Liang**[1,2*]
[1]Sun Yat-sen University, [2]Shenzhen Campus of Sun Yat-sen University,
[3]The University of Hong Kong, [4]Guangdong University of Technology,
[5]University of California, Los Angeles, [6]DarkMatter AI Research

Abstract

Geometry problem solving is a well-recognized testbed for evaluating the high-level multimodal reasoning capability of deep models. In most existing works, two main geometry problems: calculation and proving, are usually treated as two specific tasks, hindering a deep model to unify its reasoning capability on multiple math tasks. However, in essence, these two tasks have similar problem representations and overlapped math knowledge which can improve the understanding and reasoning ability of a deep model on both two tasks. Therefore, we construct a large-scale **Uni**fied **Geo**metry problem benchmark, **UniGeo**, which contains 4,998 calculation problems and 9,543 proving problems. Each proving problem is annotated with a multi-step proof with reasons and mathematical expressions. The proof can be easily reformulated as a proving sequence that shares the same formats with the annotated program sequence for calculation problems. Naturally, we also present a unified multi-task **Geo**metric Trans**former** framework, **Geoformer**, to tackle calculation and proving problems simultaneously in the form of sequence generation, which finally shows the reasoning ability can be improved on both two tasks by unifying formulation. Furthermore, we propose a Mathematical Expression Pretraining (MEP) method that aims to predict the mathematical expressions in the problem solution, thus improving the Geoformer model. Experiments on the UniGeo demonstrate that our proposed Geoformer obtains state-of-the-art performance by outperforming task-specific model NGS with over 5.6% and 3.2% accuracies on calculation and proving problems, respectively.[1]

1 Introduction

Achieving logical reasoning abilities is still challenging for neural networks, especially in some mathematical reasoning tasks, such as math word problems (MWP) (Zhang et al., 2020a,b; Qin et al., 2020, 2021; Yang et al., 2022a,b; Mishra et al., 2022a,b; Lu et al., 2022), mathematical theorem proving (Li et al., 2020; Welleck et al., 2021), etc. Recently, geometry problem solving (Sachan et al., 2020; Chen et al., 2021; Lu et al., 2021a; Zhang et al., 2022) has also attracted much attention in the NLP community, which requires comprehensive reasoning capabilities in parsing multimodal information and utilizing mathematical knowledge. Specifically, geometry problem solving mainly contains two categories: calculation and proving. For calculation problems, both recent GeoQA (Chen et al., 2021) and Inter-GPS (Lu et al., 2021a) propose multiple-choice geometry problem benchmarks annotated with specific symbolic programs or logic forms, which inspire the neural networks' potential ability to give an in-

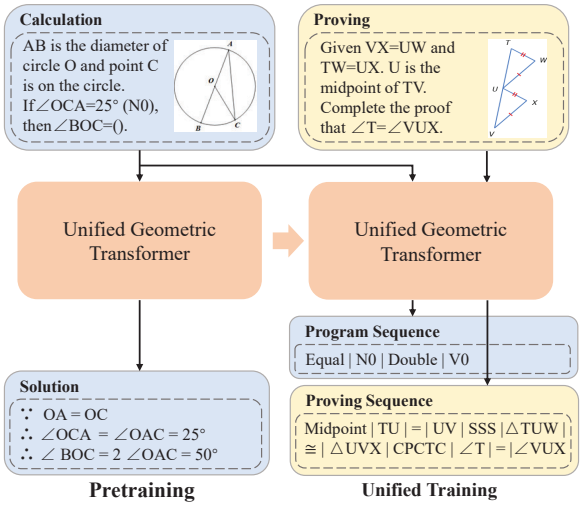


Figure 1: The pipeline of pretraining and unified training of our proposed Geoformer. We pretrain the model by predicting the mathematical expression extracted from the solution of calculation problems. After that, we consider the calculation and proving as downstream tasks, and feed both types of data to Geoformer for unified training.

*Corresponding author.
[1]Data and code: https://github.com/chen-judge/UniGeo

Calculation Problem

AB is the diameter of circle O and point C is on the circle. If ∠OCA = 25° (N0), then ∠BOC=().

A. 30° B. 40° C. 50° D. 60°

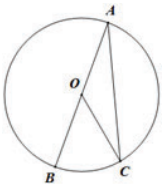
Answer: C.50°

Problem Solution:

∵OA = OC ∴∠OCA=∠OAC=25° ∴∠BOC=2∠OAC=50°

Annotated Program Sequence:

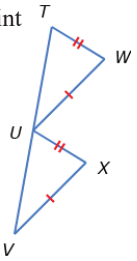
Equal | N0 | Double | V0



Proving Problem

Given VX=UW and TW=UX. U is the midpoint of TV. Complete the proof that ∠T=∠VUX.

Proof	Reasons	Expressions
Step1	Midpoint	TU = UV
Step2	SSS	△TUW ≅ △UVX
Step3	CPCTC	∠T = ∠VUX



Proving Sequence:

Midpoint | TU | = | UV | SSS |△TUW | ≅ | △UVX | CPCTC | ∠T | = |∠VUX

Figure 2: We unify geometry logical reasoning in the proposed UniGeo dataset. Except for the calculation problem provided in the GeoQA benchmark (Chen et al., 2021), we collect some proving problems (right) which contain clear mathematical expressions and corresponding reasons that can be reformulated as proving sequence to unify with the program sequence in the calculation problems.

terpretable prediction. On the subject of geometry proving, the existing work (Chou et al., 1996, 2000; Gan et al., 2019) mainly relies on well-designed proving systems and forward chaining search methods rather than neural-based models. Therefore, there is still a huge gap between the works on these two types of geometry problems, which are usually considered as two areas.

Recently, much work (Raffel et al., 2020; Cho et al., 2021; Lu et al., 2021b; Li et al., 2022; Alayrac et al., 2022) has presented unified models for various vision-language reasoning and generation tasks since the underlying visual/linguistic understanding and reasoning abilities are largely common. Inspired by the mainstream progress, we suppose that a unified model for geometry problem solving is also necessary. To begin, calculation and proving tasks share some fundamental skills and knowledge in geometric reasoning. Therefore, it is desirable to explore the general understanding and reasoning ability of the unified neural network in the math domain. Besides, the unified model doesn't need auxiliary models to determine whether the problem is a calculation or proving problem and develop task-specific models, where cumulative errors can be introduced.

To this end, a framework addressing geometry problems uniformly at both the data level and the model level is valuable and expected. However, the existing proving data is small-scale and annotated in an incompatible format. To achieve our goal, we collect lots of geometry proving data from an online education website and build a single multi-task benchmark, **UniGeo**, in which the provided proof can be reformulated as a causal proving sequence so that the calculation and proving problems are unified in data format, as shown in Figure 2. Our UniGeo contains 4,998 calculation problems and 9,543 proving problems, which can verify the high-level geometry logical reasoning capabilities in neural models.

Taking advantage of the unified formulation of two geometry tasks, we further propose a novel unified **geo**metric trans**former** (**Geoformer**) which is able to handle geometry calculation and proof reasoning simultaneously and outperforms the task-specialized models on both tasks. To learn an efficient Geoformer for unifying geometry logical reasoning, we also propose a mathematical reasoning pre-training method named Mathematical Expression Pretraining (MEP), which is based on the problem solution, since the solution prediction can serve as a universal task for all math problems. Specifically, we extract the mathematical expressions and remove the redundant text description in the solution for MEP. These expressions are rich in implicit math knowledge and can also be formulated as the solution sequence target. We further fine-tune the unified Geoformer to predict program/proving sequences for calculation and proving problems simultaneously. The pipeline of pretraining and unified training is demonstrated in Figure 1. Experiments on the UniGeo benchmark show that our proposed Geoformer achieves state-of-the-art performance, getting 5.6% and 3.2% accuracy improvements on calculation and proving problems, respectively, compared to the task-specific model NGS (Chen et al., 2021).

Our contributions can be summarized as follows:

- We construct a unified geometry reasoning benchmark, named UniGeo, which contains both calculation and proving problems.
- The proving problems in UniGeo are anno-

tated with proof steps, in which the mathematical expressions can be reformulated as proving sequences to match the program sequences in calculation problems.
- We propose a unified geometric transformer framework, which is pretrained by predicting mathematical expressions in the solution and then fine-tuned on calculation and proving problems simultaneously.

2 Related Work

Geometry Problem Solving Several geometry datasets (Seo et al., 2014, 2015; Sachan et al., 2017; Alvin et al., 2017; Sachan and Xing, 2017) have been constructed to facilitate the development of geometry problem solving. However, these previous geometry datasets are either not publicly available or built up with small sizes, which limit the development of relevant research. Besides, the latest datasets (Lu et al., 2021a; Chen et al., 2021; Cao and Xiao, 2022) only focus on the arithmetic calculation skill for geometry problem solving and fail to take into account comprehensive geometry reasoning abilities like logical proving. For instance, GeoQA (Chen et al., 2021) provides 4,998 calculation problems annotated with a symbolic program sequence that corresponds to the problem solution. Instead, we propose a new large-scale geometry dataset, which covers a wide range of sub-tasks and reasoning skills including calculation and proving. To the best of our knowledge, we are the first work to collect so many geometry proving problems for training the neural network and provide detailed sequence annotations corresponding to the proofs which can be unified with calculation problems and facilitate model learning.

Geometry Theorem Proving Theorem proving in the geometry domain (Gelernter et al., 1960; Chou et al., 1996, 2000; Ye et al., 2011; Yu et al., 2019a; Gan et al., 2019) is a long-standing artificial intelligence task. For example, (Chou et al., 1996) developed an initial automated geometry theorem proving system by designing a set of full-angle-based rules. Similarly, the expert system JGEX (Ye et al., 2011) is proposed to prove full-angle geometry problems with a well-defined deductive database. More recently, some pioneering efforts (Li et al., 2020; Tafjord et al., 2020; Welleck et al., 2021) have been attempted to learn automatic proofing systems from large-scale natural language corpus or mathematical propositions. However, how to achieve an automatic neural-based prover in the geometry domain is still less studied. Therefore, we propose a unified Geoformer that can generate proof given a geometry diagram and statements from scratch.

3 Unifying Geometry Reasoning

In this work, we aim to unify geometry logical reasoning for both calculation and proving problems. To this end, we first construct a geometry proving dataset that requires multiple reasoning abilities while solving the problems. Furthermore, we reformulate the proof as sequence form which is consistent with the program sequence in the calculation problems of the current GeoQA (Chen et al., 2021).

3.1 UniGeo Benchmark

3.1.1 Data Collection

We discover an online education website, IXL[2], which contains various types of geometry problems from high school textbooks. We utilize some crawler scripts in Python to crawl a large amount of proving data from this educational website automatically. After selecting the proving problem carefully, we ask some well-trained workers to check the quality of collected data, such as ensuring that each problem has complete diagram and clear proof. All the calculation problems are inherited from the GeoQA dataset, containing 4,998 calculation problems with program sequence annotation which corresponds to problem solution and can be predicted by generative models. We also organize five well-trained college students to translate the problems in GeoQA dataset from Chinese to English so that the language of the two types of geometry data is consistent. Finally, we unify these newly collected proving data with the GeoQA dataset and construct our UniGeo benchmark to be a testbed for unified geometry logical reasoning.

3.1.2 Data Analysis

In this section, we mainly analyze the newly collected proving problems in the UniGeo benchmark. We collect a total of 9,543 proving problems where each data contains a colored geometry diagram, a description text, and the proof with reasons and expressions. There are totally 37 categories of reasons, which are explanations for each step of the

[2]https://www.ixl.com/math/geometry

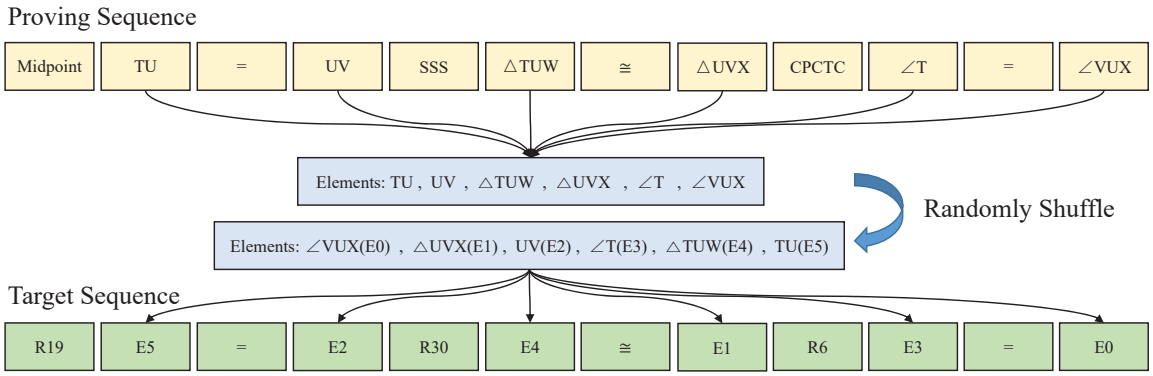


Figure 3: Illustration of converting proving sequence to the target sequence which is considered as training target for the proving problem.

	All	Train	Val	Test
All	9,543	6,675	1,421	1,447
Parallel	443	311	61	71
Triangle	3,035	2,134	452	449
Quadrangle	1,704	1,170	260	274
Congruent	2,808	1,974	414	420
Similarity	1,553	1,086	234	233

Table 1: Statistics for the proving problems in UniGeo. There are five reasoning sub-tasks for geometry proving.

proof, including the reasoning skills or the geometry theorems applied. And the expression is a concrete mathematical proof of each step, consisting of operator and geometry element. For example, in Figure 2, the reason *Midpoint* represents using the definition of midpoint to get an expression *TU* = *UV*. *SSS* stands for "side, side, side" and means that we have two congruent triangles with all three sides equal. *CPCTC* stands for "corresponding parts of congruent triangles are congruent", therefore, we get the final expression $\angle T = \angle VUX$.

As shown in Table 1, the proving data is divided into train, validation, and test splits with a ratio of 7:1.5:1.5. The dataset consists of five sub-tasks which also represent five different topics of proving problems: parallel, triangle, quadrangle, congruent, and similarity. The distribution of these types of proving problems can be viewed in Table 1. In the experiments, we also provide the detailed performance of models on these sub-tasks.

3.2 Reformulate Expressions in the Proof

Based on the collected proving data, we aim to reformulate the mathematical expressions as target sequences to unify with the program sequence in calculation problems, thus achieving a reasonable unified geometry reasoning task. The reasons and expressions in the collected proof are still textual, thus we first translate them into a sequence format. As shown in Figure 3, we organize the proof as the proving sequence which contains three types of tokens: reasons R (e.g., *Midpoint, SSS, CPCTC*), operators OP (e.g., $=, \cong$), and geometry elements E (e.g., *TU,* $\triangle TUW$, etc.). The reasons are inserted in front of the proof expressions (including operators and elements) to form the proving sequence.

Moreover, we reformulate the proving sequence as the final target sequence which can be predicted by generative models. As mentioned in Section 3.1.2, we have summarized all the reasons into a set, thus, each reason can be considered as a token R_i, where i is the index in the predefined reasons set. For the operators, we just reserve their original representation as the tokens. As for geometry elements, however, we first fetch all the geometry elements in the proving sequence, and construct the list of geometry elements. To increase the diversity of proving problems, we randomly shuffle these elements to form a new elements list and convert each element in the proving sequence to a token E_i, where i corresponds to its position in the shuffled geometry elements list. Benefiting from this, we produce diverse target sequences. Even if similar topics may exist in the training and testing sets, the target sequence tends to be completely different, avoiding that the model simply learns some typical proof patterns. Note that the shuffled elements list will also be added as text to the end of the problem text and fed into the model while training.

In summary, by reformulating the expression-based proof as the target sequence, we define a multimodal high-level reasoning task. This scheme is adopted for the following reasons. First, it sim-

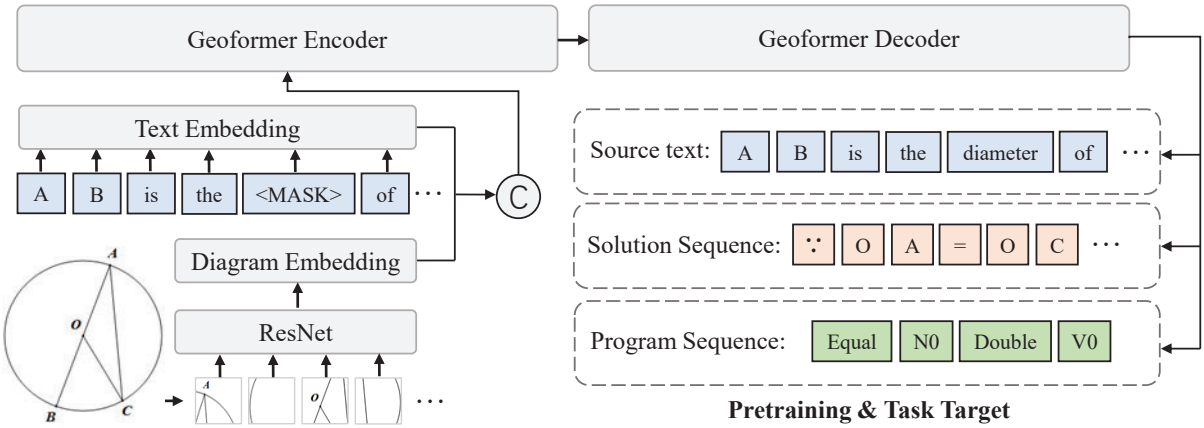


Figure 4: An illustration of our proposed geometric transformer. We concatenate the embeddings of text and diagram, which are fed into the transformer encoder-decoder to generate target sequence. For pretraining, the targets are source text and the mathematical expressions extracted from the solution. And during the fine-tuning stage, the training objective is program sequence or proving sequence. Note that we achieve unified fine-tuning for calculation and proving problems simultaneously. For brevity, this illustration shows only one example of a calculation problem.

plifies the task representation with a clear sequence prediction. Second, although the target token is reduced to a smaller space, it is still challenging for models to understand the correspondence between inputs (including problem diagram, text, and the elements to be selected) and target token. Third, by applying the expression reformulation, we unify the proving problem with the calculation problem to construct the UniGeo benchmark, which requires multiple reasoning capabilities.

4 Unified Geometric Transformer

4.1 Overview

Although the NGS model (Chen et al., 2021) is designed for geometry calculation problems, its performance degrades about 5% after unified training on the UniGeo (Table 2). Therefore, based on the VL-T5 (Cho et al., 2021) model which is capable of handling multiple multimodal tasks uniformly, we propose a geometric transformer (Geoformer) that can conduct comprehensive reasoning on both calculation and proving problems. Figure 4 demonstrates the structure of the model, which consists of a bidirectional multimodal encoder and an autoregressive text decoder. In order to promote the performance of Geoformer, we first pretrain it using the solution provided in calculation problems, as well as applying masked LM task to enhance the representation of the text encoder. At the fine-tuning stage, we train the model end-to-end with calculation and proving problems simultaneously to acquire stronger comprehensive reasoning ability on geometry problem solving, rather than optimizing the model on two tasks separately.

4.2 Unified Pretraining

4.2.1 Mathematical Expression Pretraining

Different from the popular pretraining paradigm that fine-tuning models with large-scale natural corpus, geometry problems are mainly described by some mathematical languages and also solved by mathematical knowledge, which is far from natural language. Therefore, we propose Mathematical Expression Pretraining (MEP) to pretrain our unified Geoformer with the mathematical corpus.

Formulating Solution Sequence The GeoQA dataset provides a problem solution that explains the idea and process of solving the problem in the form of text description. Similar to formulating the proof expression into the sequence, mathematical expression in the solution can also be reformulated into solution sequence for prediction. We remove the redundant text description in the solution and only utilize the mathematical expressions for pretraining. Specifically, we only reserve the geometry elements entities, operation symbols, and numbers, all of which involve abundant geometry mathematical knowledge and can be organized as solution sequence. In addition, the number in solution is replaced with a token NS_i where i represents the order that the number appears in the solution. Different from taking word-level tokenization for nat-

ural text description, we adopt char-level tokenization for geometry elements since some geometry elements share common characters with specific geometric meanings. For example, both line OC and $\angle OCA$ contain points O and C, but this relation will disappear if OC and $\angle OCA$ are considered as basic tokens. In summary, the formulated solution sequence has rich mathematical knowledge and can be learned by models to enhance the understanding of mathematical reasoning process.

4.2.2 Masked Language Modeling

We also explore applying the Masked Language Modeling (MLM) task for solving geometry problems. Following (Cho et al., 2021), we mask 30% of input text tokens with `<mask>` tokens. Then the model is trained to recover the masked text in a unified text generation manner.

4.3 Fine-tuning Unified Geoformer

We combine the above two pretraining tasks to pretrain the unified geometric transformer. After that, fine-tuning the unified Geoformer is straightforward since we have unified the outputs of all downstream tasks into a sequence format. We load the weights from the pretrained model and keep the weights of the diagram encoder fixed, following the NGS model (Chen et al., 2021). Then, we optimize the rest parts of the model end-to-end using a mixture of calculation and proving data.

4.4 Unified Training Objective

All of the pre-training and fine-tuning tasks in this work are unified in the form of text generation, thus sharing the same training objective. The generation loss L_g is the negative log-likelihood (NLL) of the target sequence:

$$\mathcal{L}_g(\boldsymbol{\theta}) = \frac{1}{L}\sum_{t=1}^{L} \log P_t(y_t|\boldsymbol{x}, y_1, ..., y_{t-1}; \boldsymbol{\theta}),$$

where $\boldsymbol{\theta}$ are the parameters of the entire Geoformer architecture except for the diagram encoder, $\boldsymbol{x}$ is the input of both problem text and the extracted diagram feature, $\boldsymbol{y}_t$ are the target tokens, $\boldsymbol{P}_t$ is the distribution of the next token, $\boldsymbol{L}$ is the length of sequence.

5 Experiments

5.1 Experimental Settings

Datasets We conduct experiments on the UniGeo, containing GeoQA (Chen et al., 2021) dataset and our newly collected proving problems. The GeoQA dataset involves 4,998 calculation problems with corresponding annotated program sequence, which illustrates the calculating process of the given problems and is considered as training and testing target. Besides, the GeoQA also provides the problem solution which is not utilized by previous works but is used for pretraining in this work. We also construct a proving dataset with 9,543 problems, which are split to train, validate, and test subsets in a ratio of 7.0: 1.5: 1.5. We further define five sub-tasks: Parallel, Triangle, Quadrangle, Congruent, and Similarity, to provide the detailed performance of models. To unify geometry reasoning, we also translate the Chinese calculation problems into English, so that the language of calculation and proving problems are consistent. We also have considered the Inter-GPS dataset (Lu et al., 2021a). However, it mainly adopts the rule-based parser to translate the problem text into formal language and doesn't have the sequence annotation which can be unified with the proving sequence in our work. Therefore, the Inter-GPS dataset is not compatible with unified training on both calculation and proving problems, and we mainly conduct experiments on GeoQA and newly collected proving data.

Evaluation Metrics For the calculation problems, we follow the evaluation metrics in GeoQA, i,e, the accuracy of solving all the problems and two main subsets: angle and length problems. Following the IsarStep (Li et al., 2020), we adopt top-1 accuracy and top-10 accuracy for evaluating the proofs. Top-K accuracy computes the percentage where the ground-truth proof is among the top K generated proving sequence. Since the models possibly generate alternative valid proving sequences that are not completely consistent with the provided proof, we mainly use more reasonable top-10 accuracy for evaluating the proving problems.

Implementation Details We fill the diagram with a white background to make it equal in length and width, and resize it to 224×224, which is further split into 49 patches with a size of 32×32 each. Then we apply ResNet (He et al., 2016) to extract patch features which are further mapped into flattened 1D sequences to construct final diagram embeddings. Our Geoformer is implemented by PyTorch (Paszke et al., 2017). We use the Adam (Loshchilov and Hutter, 2017) optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning

Methods	Data	Calculation (%)			Proving (%)					
		All	Angle	Length	All	Par.	Tri.	Qua.	Con.	Sim.
FiLM (Perez et al., 2017)	Calculation	31.7	34.0	29.7	-	-	-	-	-	-
RN (Santoro et al., 2017)	Calculation	38.0	42.8	32.5	-	-	-	-	-	-
MCAN (Yu et al., 2019b)	Calculation	39.7	45.0	34.6	-	-	-	-	-	-
BERT (Devlin et al., 2018)	Calculation	54.7	65.8	42.1	-	-	-	-	-	-
NGS (Chen et al., 2021)	Calculation	56.9	69.8	39.2	-	-	-	-	-	-
Geoformer (Ours)	Calculation	60.3	71.5	**49.1**	-	-	-	-	-	-
BERT	Proving	-	-	-	48.0	15.5	48.1	28.5	49.5	77.6
NGS	Proving	-	-	-	53.2	13.2	56.6	29.8	57.1	**79.4**
Geoformer (Ours)	Proving	-	-	-	55.7	19.4	68.3	20.4	60.6	72.5
BERT	UniGeo	52.0	63.1	39.2	48.1	15.4	48.0	**31.7**	49.5	75.1
NGS	UniGeo	51.9	63.6	38.8	47.4	11.2	46.9	31.3	48.3	77.6
Geoformer (Ours)	UniGeo	60.9	72.2	48.8	55.8	18.1	68.8	20.4	60.3	73.3
Geoformer + Pretraining (Ours)	UniGeo	**62.5**	**75.5**	48.8	**56.4**	**19.4**	**69.4**	20.4	**60.3**	75.0

Table 2: The accuracy comparison of various methods and baseline models under different data settings. The newly collected proving problems provide five sub-tasks (as defined in Table 1) for evaluation.

rate is $2e^{-4}$, the batch size is set to 10, and models are trained within 100 epochs. We train our unified Geoformer on randomly shuffled calculation problems and proving problems simultaneously. For pretraining, we maintain the settings as mentioned above, but replace the training label with the solution sequence and set the learning rate to $5e^{-4}$.

Methods	Data	Top-1	Top-10
NGS	UniGeo	17.4	47.4
NGS + Pretraining	UniGeo	19.2	49.6
Geoformer	UniGeo	50.2	55.8
Geoformer + Pretraining	UniGeo	**51.3**	**56.4**

Table 3: Performance comparison on proving problems with different evaluation metrics.

5.2 Experimental Results

Table 2 demonstrates the results of our methods and baselines on the calculation and proving problems. We divided the experiments into three parts according to the data used by the model, i.e., the calculation problems from the GeoQA dataset, our newly collected proving problems, and the unified benchmark of both calculation and proving problems. A detailed analysis is shown below.

Baselines FiLM (Perez et al., 2017), RN (Santoro et al., 2017), MCAN (Yu et al., 2019b) are three multimodal models with strong cross-modal reasoning abilities that well address the compositional language and elementary visual reasoning benchmark, CLEVR (Johnson et al., 2017). They can predict the possibly correct option in calculation problems by using visual question answering. However, this approach does not work well in geometry problem solving since the MCAN achieves the answer accuracy of only 39.7%. The “BERT" model here refers to "BERT2Prog + Diagram" in GeoQA that BERT and ResNet are utilized to encode text and diagram data separately. Finally, the features of these two modalities are fused to guide the generation of target sequence. The NGS model is specially designed for solving the calculation problems in the GeoQA dataset. We also re-run the experiment on the English version of the GeoQA dataset using the NGS model and obtain a performance of 56.9%.

The performance comparison on proving problems We conduct some experiments on the collected proving problems. In Table 2, Par, Tri, Qua, Con, Sim represent five sub-tasks respectively. When using proving data only, the NGS model achieves a total performance of 53.2%. The proposed Geoformer obtains a top-10 accuracy of 55.7% on proving problems. There is a huge difference in the performance of sub-tasks due to the difficulty of various geometric reasoning skills varies greatly. The accuracy rate of proving parallel related problems is only 19.4%, while proving similarity is relatively simple, which can reach an accuracy of 72.5%. Table 3 also provides the results of top-1 accuracy metric. When applying pretraining

Methods	Calculation	Proving
Geoformer	60.9	55.8
Geoformer + MLM	61.3	56.2
Geoformer + MEP	61.8	56.1
Geoformer + MLM + MEP	**62.5**	**56.4**

Table 4: Ablation study for different pretraining methods. MLM and MEP represent masked language modeling and mathematical expression pretraining.

on the unified NGS and Geoformer models, we can get a 19.2% and 51.3% top-1 accuracy respectively.

The performance of unified training Our motivation is to unify the geometry logical reasoning and we have already unified the data format. Thus, apart from training on calculation and proving problems separately, we design the unified Geoformer, which is trained with the mixture of both types of problems. It can be observed that the NGS model suffers a severe performance decline when trained on both tasks simultaneously, in which the accuracy of calculation and proving problems decrease 5.0% and 5.8% respectively. However, our proposed Geoformer avoids this phenomenon and obtains an impressive performance on two tasks simultaneously. Specifically, the unified Geoformer achieves 60.9% and 55.8% accuracy on calculation and proving problems, outperforming two task-specific Geoformer models on two geometry tasks. The reasoning ability can be enhanced on both two tasks with the unified formulation.

The effectiveness of pretraining To further promote the performance of unified Geoformer, We extract a large number of mathematical expressions from the solution of calculation problem as the pre-training target. These expressions are rich in implicit math knowledge and can also be formulated as solution sequence. Applying the pretraining method, the Geoformer+Pretraining model is further improved to 62.5% and 56.4% accuracy on calculation and proving problems respectively, obtaining 5.6% and 3.2% accuracy improvement compared to task-specialized NGS models, which achieves state-of-the-art performance on UniGeo benchmark.

5.3 Ablation Study

We explore the effectiveness of different pretraining settings for the ablation study. In Table 4, we experiment unified Geoformer with two pretraining ways: masked language modeling (MLM) and mathematical expression pretraining (MEP). Using only MLM, the performance of the Geoformer model does not change significantly. When MEP is used alone, the performance of the model on calculation problems is improved obviously. When using both pre-training methods, the model makes an improvement significantly on both types of problems, obtaining the highest 62.5% and 56.4% on calculation and proving problems, respectively. Thus, we apply this setting to the training of Geoformer.

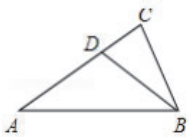
5.4 Case Study

As shown in Figure 5, we conduct a case study for discussing the ability and limitation of our proposed unified Geoformer. For the left case, the unified Geoformer works well on the calculation problem, benefiting from the multi-task learning framework. As we can see in the problem solution, it first utilizes the knowledge of similar triangles to get $\triangle BCD \sim \triangle ACB$, and then lists a proportional relation to get $CD = 3/6 * 3 = 1.5$. Compared to task-specialized Geoformer that predicts a wrong program sequence, the prediction made by our unified Geoformer is completely consistent with ground truth. This is probably because the unified model acquires a stronger understanding of similar triangle knowledge after simultaneously training on proving problems (containing many problems proving similar triangles). Therefore, multi-task learning is beneficial in geometry reasoning. We also select a typical failure case. The unified Geoformer chooses two wrong geometry elements for the proof steps and also fails to give the last two critical proof steps. The geometry problems are still challenging for current neural-based approaches.

6 Conclusion

Recently, geometry problem solving has attracted much attention in AI research while previous works mainly focus on geometry calculation problems. It is significant to explore the unified reasoning abilities of neural models on multiple math tasks. Therefore, we integrate geometry calculation and proving problems, and construct a unified geometry benchmark, UniGeo, containing 9,543 proving problems with proof reasons and mathematical expressions that can be reformulated as proving sequence to unify with the program sequence of calculation problems. We also propose a unified

In △ABC, D is a point on AC, if ∠DBC=∠A, BC=3 (N0), AC=6 (N1), then the length of CD is ().



Solution:
∵∠DBC=∠A,∠C=∠C,∴△BCD∽△ACB,
∴CD/BC=BC/AC ∴CD/3=3/6 ∴CD=1.5

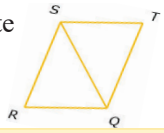
Unified Geoformer & Ground Truth:

Proportion | N0 | N1 | N0

Specialized Geoformer:

Add | N0 | N1 | Proportion | N0 | V0 | N1

QR // ST and QT // RS . Complete the proof that RS≅QT



Ground Truth:

Alternate Interior Angles Theorem | ∠RQS | ≅ | ∠QST |
Alternate Interior Angles Theorem | ∠QSR | ≅ | ∠SQT |
Reflexive Property of Congruence | QS | ≅ | QS |
ASA | △QRS | ≅ | △STQ | CPCTC | RS | ≅ | QT

Unified Geoformer:

Alternate Interior Angles Theorem | ∠RQS | ≅ | ∠SQT |
Alternate Interior Angles Theorem | ∠QSR | ≅ | RS |
Reflexive Property of Congruence | QS | ≅ | QS |

Figure 5: The left calculation case shows a situation where a unified Geoformer works better than a task-specialized Geoformer since the related similar triangle knowledge also exists in proving problems. Through multi-task learning, the model is enhanced on the understanding of similar triangle problems. In the failure proving case on the right, the Geoformer model outputs some incorrect proof (red) and misses part of the proof (the **bold** in ground truth).

Geoformer that can address calculation and proving problems simultaneously. Besides, a mathematical expression pretraining way is proposed to promote the performance of the unified Geoformer. Experiments show that our Geoformer can well address two challenging geometry tasks with a single set of model weights, outperforming task-specialized models and obtaining state-of-the-art performance.

Limitations

To explore the logical reasoning ability of neural network models in the geometry domain, we propose a unified method for two major and similar tasks (calculation and proving) in geometry problems. Although we have achieved state-of-the-art performance on these two tasks simultaneously, the unified Geoformer still has some limitations. First, the answer accuracy of the neural-network-based approaches is still far from the real-world application when addressing such complex tasks which require high-level reasoning ability. Second, the data construction of such mathematical logical reasoning tasks requires a heavy manual collection and annotation process, which also limits the type and difficulty of geometry problems, thereby leading to the failure of neural network models to learn and process more sophisticated cases.

Acknowledgements

This work was supported in part by National Key R&D Program of China under Grant No.2020AAA0109700, National Natural Science Foundation of China (NSFC) under Grant No.U19A2073, Grant No.61976233 and Grant No.62206314, Guangdong Province Basic and Applied Basic Research (Regional Joint Fund-Key) Grant No.2019B1515120039, Guangdong Outstanding Youth Fund (Grant No.2021B1515020061), GuangDong Basic and Applied Basic Research Foundation under Grant No.2022A1515011835, China Postdoctoral Science Foundation under Grant No.2021M703687, Shenzhen Fundamental Research Program (Project No.RCYX20200714114642083) and CAAI-Huawei MindSpore Open Fund. And the Open Project of Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, Anhui University, No.MMC202107. We thank MindSpore for the partial support of this work, which is a new deep learning computing framwork[3].

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*.

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

Jie Cao and Jing Xiao. 2022. An augmented benchmark dataset for geometric question answering through dual parallel text encoding. In *Proceedings of the 29th International Conference on Computational Linguistics (COLING)*, pages 1511–1520.

Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric P Xing, and Liang Lin. 2021. Geoqa: A geometric question answering benchmark

[3]https://www.mindspore.cn/

towards multimodal numerical reasoning. In *Findings of the Association for Computational Linguistics (ACL-IJCNLP)*.

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, pages 1931–1942. PMLR.

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

Shang-Ching Chou, Xiao-Shan Gao, and Jing-Zhong Zhang. 2000. A deductive database approach to automated geometry theorem proving and discovering. *Journal of Automated Reasoning*, 25(3):219–246.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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

Herbert Gelernter, James R Hansen, and Donald W Loveland. 1960. Empirical explorations of the geometry theorem machine. In *Western Joint IRE-AIEE-ACM Computer Conference*, pages 143–149.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2901–2910.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*.

Wenda Li, Lei Yu, Yuhuai Wu, and Lawrence C Paulson. 2020. Isarstep: a benchmark for high-level mathematical reasoning. In *The International Conference on Learning Representations (ICLR)*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. 2021a. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. In *The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*.

Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. 2022. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. *arXiv preprint arXiv:2209.14610*.

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

Swaroop Mishra, Matthew Finlayson, Pan Lu, Leonard Tang, Sean Welleck, Chitta Baral, Tanmay Rajpurohit, Oyvind Tafjord, Ashish Sabharwal, Peter Clark, and Ashwin Kalyan. 2022a. Lila: A unified benchmark for mathematical reasoning. In *The 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Swaroop Mishra, Arindam Mitra, Neeraj Varshney, Bhavdeep Sachdeva, Peter Clark, Chitta Baral, and Ashwin Kalyan. 2022b. Numglue: A suite of fundamental yet challenging mathematical reasoning tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3505–3523.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch. .

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. 2017. Film: Visual reasoning with a general conditioning layer. *arXiv preprint arXiv:1709.07871*.

Jinghui Qin, Xiaodan Liang, Yining Hong, Jianheng Tang, and Liang Lin. 2021. Neural-symbolic solver for math word problems with auxiliary tasks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*.

Jinghui Qin, Lihui Lin, Xiaodan Liang, Rumin Zhang, and Liang Lin. 2020. 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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

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

Mrinmaya Sachan, Kumar Dubey, and Eric Xing. 2017. 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.

Mrinmaya Sachan and Eric Xing. 2017. 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.

Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. 2017. A simple neural network module for relational reasoning. In *Advances in neural information processing systems (NeurIPS)*, pages 4967–4976.

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

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

Oyvind Tafjord, Bhavana Dalvi Mishra, and Peter Clark. 2020. Proofwriter: Generating implications, proofs, and abductive statements over natural language. In *Findings of the Association for Computational Linguistics (ACL-IJCNLP)*.

Sean Welleck, Jiacheng Liu, Ronan Le Bras, Hannaneh Hajishirzi, Yejin Choi, and Kyunghyun Cho. 2021. Naturalproofs: Mathematical theorem proving in natural language. In *The 35th Conference on Neural Information Processing Systems (NeurIPS)*.

ZhiCheng Yang, Jinghui Qin, Jiaqi Chen, and Xiaodan Liang. 2022a. Unbiased math word problems benchmark for mitigating solving bias. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1401–1408. Association for Computational Linguistics.

Zhicheng Yang, Jinghui Qin, Jiaqi Chen, Liang Lin, and Xiaodan Liang. 2022b. Logicsolver: Towards interpretable math word problem solving with logical prompt-enhanced learning. *arXiv preprint arXiv:2205.08232*.

Zheng Ye, Shang-Ching Chou, and Xiao-Shan Gao. 2011. An introduction to java geometry expert. In *International Workshop on Automated Deduction in Geometry*, pages 189–195. Springer.

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

Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. 2019b. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 6281–6290.

Jipeng Zhang, Ka Wei LEE, Ee-Peng Lim, Wei Qin, Lei Wang, Jie Shao, Qianru Sun, et al. 2020a. Teacher-student networks with multiple decoders for solving math word problem. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI)*.

Jipeng Zhang, Lei Wang, Roy Ka-Wei Lee, Yi Bin, Yan Wang, Jie Shao, and Ee-Peng Lim. 2020b. Graph-to-tree learning for solving math word problems. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3928–3937.

Ming-Liang Zhang, Fei Yin, Yi-Han Hao, and Cheng-Lin Liu. 2022. Plane geometry diagram parsing. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1636–1643. International Joint Conferences on Artificial Intelligence Organization. Main Track.