A Survey of Deep Learning for Geometry Problem Solving

Anonymous ACL submission

Abstract

Geometry problem solving, a crucial aspect of mathematical reasoning, is vital across various domains, including education, the assessment of AI's mathematical abilities, and multimodal capability evaluation. The recent surge in deep learning technologies, particularly the emergence of multimodal large language models, has significantly accelerated research in this area. This paper presents a survey of the applications of deep learning in geometry problem solving, including (i) a comprehensive summary of the relevant tasks in geometry problem solving; (ii) a thorough review of related deep learning methods; (iii) a detailed analysis of evaluation metrics and methods; and (iv) a critical discussion of the current challenges and future directions that can be explored. Our objective is to offer a comprehensive and practical reference of deep learning for geometry problem solving, thereby fostering further advancements in this field. We create a continuously updated list of papers: `https://anonymous.4open.science/r/papers-4Km8Pz2Q`.

1 Introduction

As a core aspect of mathematical reasoning, **Geometry Problem Solving** (**GPS**) has long been closely tied to education and the assessment of mathematical proficiency in Artificial Intelligence (AI) systems (Narboux et al., 2018). Given the inherent connection between geometry problems and diagrams, GPS has naturally emerged as a representative multimodal mathematical task. Solving geometry problems in the format of educational exams requires AI systems not only to interpret geometric diagrams but also to perform robust logical reasoning and numerical computation, making it an ideal benchmark for assessing perception and reasoning in deep learning models. In recent years, the rise of Multimodal Large Language Models (MLLMs) has further advanced this field, showcasing the great potential of deep learning in complex visual understanding and reasoning tasks. The number of papers on deep learning for GPS has grown rapidly, from just 1 in 2018 to 110 in 2024, and continues to increase in 2025 (see Figure 5).

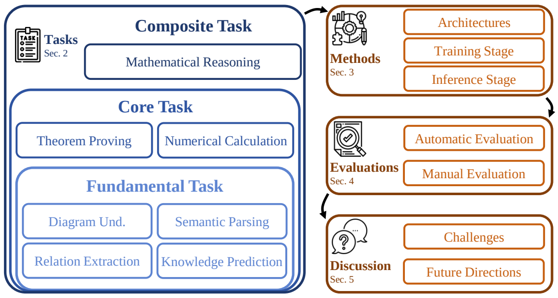

Figure 1: Overview of the survey's structure

Although many surveys have reviewed deep learning methods and Large Language Models (LLMs) in the broader field of mathematical reasoning (Lu et al., 2023; Ahn et al., 2024; Saraf et al., 2024; Yan et al., 2024), the subfield of GPS remains underexplored compared to other mathematical areas (Zhang, 2022; Li et al., 2024e). Recent surveys on GPS are relatively limited in scope—either concentrating solely on multimodal plane geometry problems (Cho et al., 2025b), or lacking a comprehensive summary of relevant datasets and deep learning methods (Zhao et al., 2025b).

In this study, we began with several classic papers in this field, conducted a single round of forward and backward snowballing, searched Google Scholar with the keyword "geometry", and manually screened to ensure the relevance of the papers. As a result, we collected more than **310** academic papers that involved deep learning for GPS, and conducted a comprehensive and in-depth survey.

In the following sections, we will first summarize the tasks related to GPS in depth (§2). Then, we will comprehensively review the various methods used in the field of GPS (§3). After that, we perform a systematic analysis of the evaluation metrics and methods (§4). Finally, we will discuss the current challenges facing this field and look forward to future development directions (§5).

2 Geometry Problem Solving Tasks

In this section, we outline the tasks related to GPS, which are categorized into fundamental, core, and composite tasks. Fundamental tasks cover the basic abilities required for solving geometry problems, core tasks are directly tied to GPS, and composite tasks treat GPS as part of broader complex tasks. The taxonomy of tasks and datasets is shown in Figure 2, and a detailed summary of the datasets can be found in Table 1 and Table 2.

2.1 Fundamental Tasks

In order to solve geometry problems, a deep learning system must first have a variety of fundamental capabilities, including understanding geometric diagrams, semantic parsing of geometry problem texts, extraction of geometric relationships, and prediction of geometric knowledge.

Geometric Diagram Understanding. Geometric diagram understanding is committed to fully understanding the information in geometric diagrams. It consists of multiple subtasks at different levels. First, detect and identify basic geometric elements (such as points, lines, angles, and polygons) and their attributes (such as quantity and size) (Lu et al., 2015; Song et al., 2017, 2020). This task is called *Geometric Element Recognition*. Second, based on the recognition of geometric elements, further identify and construct the structure and spatial relationship between elements (Xia and Yu, 2021; Huang et al., 2023), namely *Geometric Structure Recognition*. These two tasks are often jointly considered as *Geometric Perception* tasks (Kamoi et al., 2024; Xing et al., 2024). Third, based on geometric perception capabilities, generate formal language for geometric diagrams (Hao et al., 2022; Wei et al., 2024). This task is also known as *Geometric Diagram Parsing*. Finally, some studies use natural language to provide an accurate description of geometric diagrams. These descriptions are either generated based on diagram parsing or directly generated from geometric diagrams (Zhang and Moshfeghi, 2024; Huang et al., 2025f), which is referred to as *Geometric Diagram Captioning*.

Semantic Parsing for geometry problem texts. Semantic parsing is essential for converting problem text into machine-readable formal statements (Matsuzaki et al., 2017), and was a core component of early deep learning frameworks for GPS (Joshi et al., 2018; Sun et al., 2019). Geometry problem texts often contain multiple sentences and complex geometric information, making cross-sentence references and domain-specific content challenging (Hopkins et al., 2017). Some studies also integrate diagram parsing with semantic parsing, aiming to achieve the joint parsing of text and diagrams (Boob et al., 2023; Zhou et al., 2024c).

Geometric Relation Extraction. Geometric relation extraction is a well-defined task that involves extracting geometric relationships either from the question text (Huang et al., 2022), or jointly from both text and diagrams (Gan et al., 2017), and representing them in structured formats such as triples (Zhou et al., 2022) or knowledge graphs (Wang et al., 2025h). The model achieves a deep understanding of the problem by extracting geometric relationships in geometry problems rather than using natural language (Gan et al., 2019b,a).

Geometric Knowledge Prediction. Geometric knowledge prediction aims to evaluate the model's understanding of geometry by predicting the geometric principles (Xu et al., 2025b) and theorems (Lu et al., 2021) (i.e., geometric knowledge) required to solve geometry problems (Ning et al., 2025). The model needs to predict the relevant geometric knowledge required to solve the problem based on the input question and apply it in the reasoning process (Wu et al., 2024a).

2.2 Core Tasks

GPS can be categorized into geometry theorem proving and geometric numerical calculation (Chen et al., 2022). On the premise of having the capabilities covered by the fundamental tasks, the model needs to solve geometry problems in the format of educational exams. See Figure 3 for an example.

Geometry Theorem Proving. Geometry theorem proving is a long-standing task in the field of AI (Gelernter et al., 1960; Kapur, 1986). The input is a geometry theorem that requires proof, and the goal is to output a detailed derivation process of the proof, usually focusing on plane geometry.

Geometric Numerical Calculation. Geometric numerical calculation has gradually emerged with the introduction of new datasets in recent years (Seo et al., 2015; Sachan et al., 2017). The input is a geometry problem involving the calculation of a certain geometric value (such as length or angle), and the desired output is a concise answer to the problem, without necessarily providing a

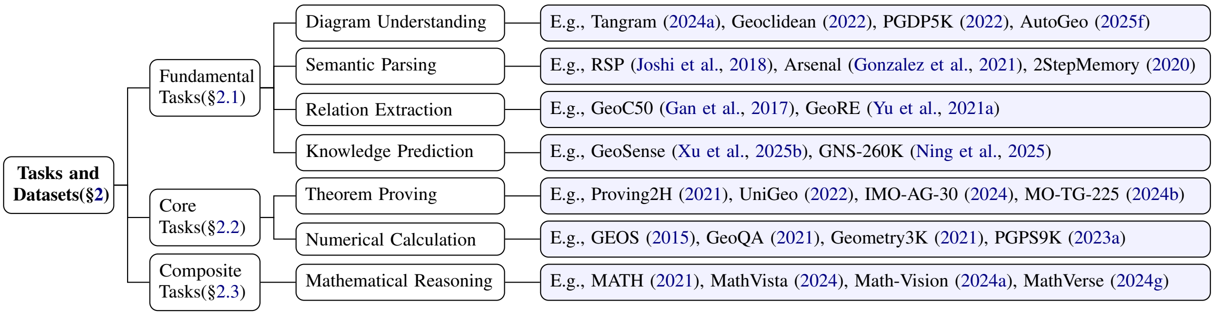

Figure 2: Taxonomy of Tasks and Datasets for Geometry Problem Solving.

complete reasoning process. Its question types can usually be divided into several categories, including plane geometry, solid geometry, and analytic geometry.

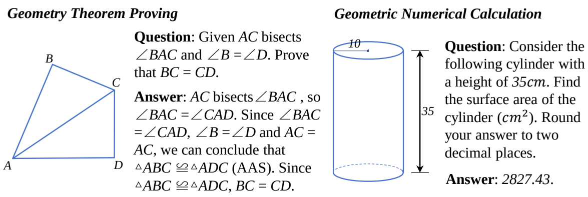

Figure 3: An example of geometry theorem proving and geometric numerical calculation problem.

2.3 Composite Tasks

Recently, GPS has also often appeared as a subtask of composite tasks, mainly used to explore the model's ability in mathematical reasoning.

Mathematical Reasoning. Geometry is an important part of mathematics, and geometric diagrams are also a typical type of mathematical synthetic image. Therefore, geometry problems are often included in single-modal or multi-modal mathematical benchmarks (Hendrycks et al., 2021; Lu et al., 2024) to evaluate the performance of models in mathematical reasoning tasks.

3 Methods for Geometry Problem Solving

This section comprehensively reviews deep learning methods for GPS. We first introduce the relevant architectures, then classify and summarize other methods according to the training and inference stages. The taxonomy of these methods is shown in Figure 4.

3.1 Architectures for Geometry Problem Solving

In GPS, the classic deep learning architecture is the Encoder-Decoder architecture (Sutskever et al., 2014), which also encompasses the widely used MLLMs in recent years. Other architectures have also been explored, including Generative Adversarial Networks (GANs)(Goodfellow et al., 2014), Graph Neural Networks (GNNs) (Scarselli et al., 2008), and Decoder-Only architectures. These architectures are outlined in more detail in Table 4.

3.1.1 Encoder-Decoder Architecture

The encoder-decoder architecture can be divided into the following five key parts: text encoder, diagram encoder, multimodal fusion module, decoder, and optional knowledge module.

Text Encoder. Text encoder can convert the text content of the geometry problem into formalized statements or encode it into vectors, enabling deep learning systems to process the text information. Early studies usually use Long Short-Term Memory network (LSTM) (Hochreiter and Schmidhuber, 1997), Gated Recurrent Unit (GRU) (Cho et al., 2014) and their bidirectional variants as text encoders, while more recent work employs Transformers (Vaswani et al., 2017) or pre-trained language models.

Diagram Encoder. Parsing geometric diagrams into formal statements or encoding them into vector information is of great significance for solving multimodal geometry problems. Early studies mainly used various Convolutional Neural Networks (CNNs) (LeCun et al., 1998) to encode geometric diagrams, while recent studies have widely used pre-trained diagram encoders (Dosovitskiy et al.; Radford et al., 2021). In addition, there are also studies that use LSTM, GNN, and other structures for diagram parsing.

Multimodal Fusion Module. For multimodal geometry problems, the multimodal fusion module fuses and aligns text and diagram information extracted from the original problem or encoders, then passes it to the decoder. Some works use a co-attention module (Yu et al., 2019) for multimodal fusion, and in MLLMs, structures such as MLP (Liu et al., 2024a) are widely used. Addition-

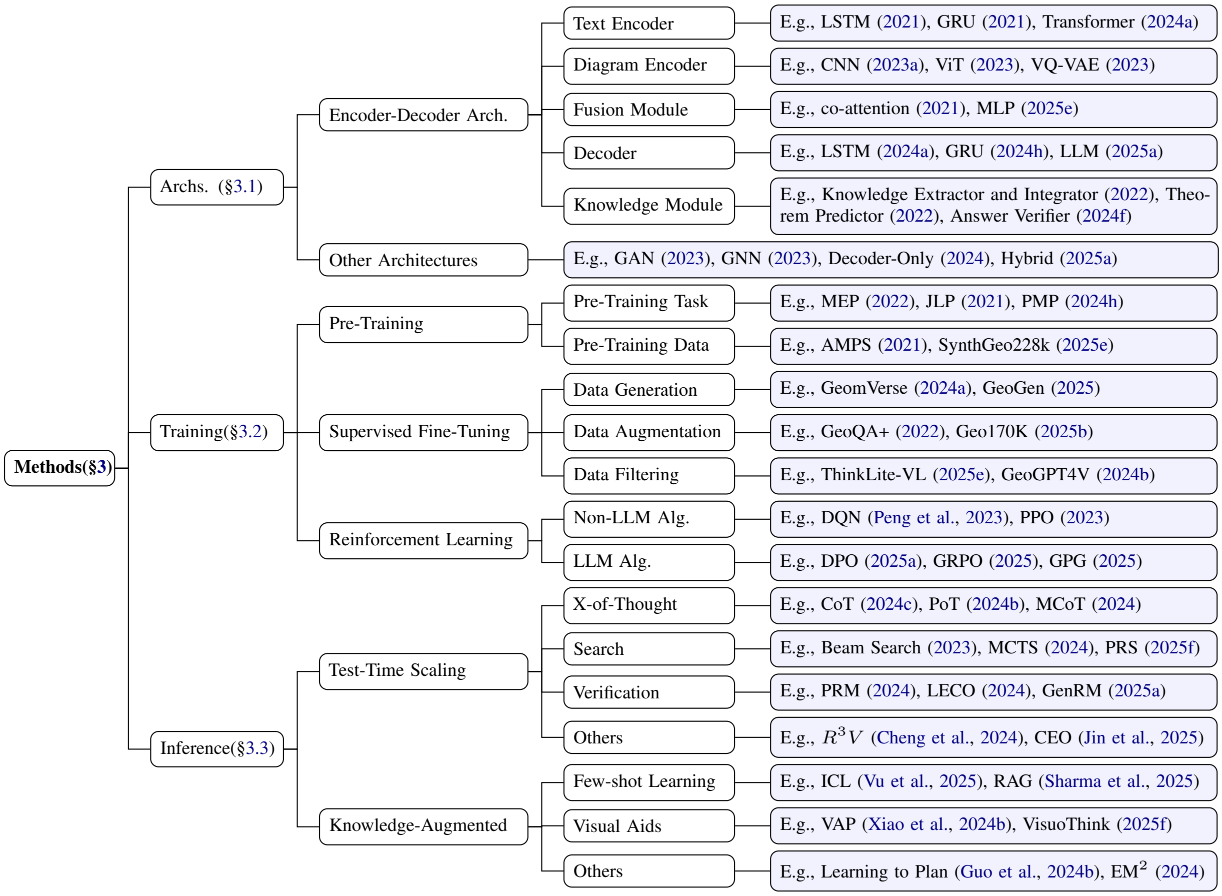

Figure 4: Taxonomy of Deep Learning Methods for Geometry Problem Solving.

ally, some studies treat this module together with the decoder as a unified encoder-decoder architecture.

Decoder. This module decodes the geometric knowledge and information to output the final answer to the question. Many studies use LSTM or GRU as the decoder of deep learning systems. In addition, there are also a lot of studies using pre-trained LLMs.

Knowledge Module. Some GPS systems integrate knowledge modules based on deep neural networks to more efficiently retrieve and apply knowledge and theorems in the field of geometry and verify the correctness of solutions. The knowledge modules can be mainly divided into three categories: the first is *Knowledge Extractor and Integrator*, which is used to extract and integrate geometric knowledge (Xiao et al., 2024a); the second is *Theorem Predictor*, which is used to predict the geometric theorems required for the current solution step (Guo and Jian, 2022); and the third is *Answer Verifier*, which is used to ensure the correctness of the solution (Pan et al., 2025).

More details about the encoder-decoder architecture can be found in Appendix C.

3.1.2 Other Architectures

In addition to the encoder-decoder architecture, some deep learning systems for solving geometry problems have adopted other architectures. Song et al. (2023) adopts **GAN** architecture, while Peng et al. (2023); Huang et al. (2024) use **GNN** to solve geometry problems. Many studies adopt **Decoder-Only Architecture**, for example, the AlphaGeometry series (Trinh et al., 2024; Sinha et al.; Chervonyi et al., 2025) uses a trained Transformer to solve IMO geometry problems, and some work directly uses LLMs to solve unimodal geometry problems (Tong et al., 2024; Tang et al., 2024b). Other studies have combined LLMs with other deep learning architectures (Zhao et al., 2025a; Cheng et al., 2025a), or multiple LLMs (Gao et al., 2024; Lei et al., 2024; Liu et al., 2025b), to build **Hybrid Architectures** for GPS.

3.2 Training Stage for Geometry Problem Solving

3.2.1 Pre-Training

Pre-Training Task. Beyond directly applying pre-trained models to geometry problems, many works design targeted pre-training tasks to enhance performance. Some focus on the *textual modality*: Chen et al. (2022) proposes Mathematical Expression Pretraining (MEP) to capture mathematical knowledge, while Zhang et al. (2023a, 2024f); Ma et al. (2024a) adopt Masked Language Modeling (MLM) to improve understanding and generation of textual descriptions. Others target *diagram encoders*: Chen et al. (2021) introduces Jigsaw Location Prediction (JLP) and Geometry Elements Prediction (GEP), while Ning et al. (2023) applies Masked Image Modeling (MIM) and Multi-Label Classification (MLC) to optimize the diagram encoder. There are also tasks focusing on *matching multimodal relationships*, such as LANS (Li et al., 2024h) with Structural-Semantic Pretraining (SSP) and Point-Match Pretraining (PMP), and SANS (Lin et al., 2025) with Dual-Branch Visual-Textual Points Matching (DB-VTPM).

Pre-Training Data. To address the scarcity of geometric pre-training data, AMPS (Hendrycks et al., 2021) and InfiMM-WebMath-40B (Han et al., 2024) offer large-scale mathematical and multimodal datasets, boosting model performance on geometry tasks. Given the gap between real-world images and geometric diagrams, some works construct dedicated datasets for diagram encoder pre-training. Geo-ViT (Xia et al., 2025) compiles 120K+ diagrams for ViT training; CLIP-Math (Zhang et al., 2025c), GeoCLIP (Cho et al., 2025a), GeoGLIP (Zhang et al., 2025d), and DFE-GPS (Zhang et al., 2025e) use synthetic data for geometry-focused visual pretraining.

3.2.2 Supervised Fine-tuning

In GPS, deep learning models typically require Supervised Fine-Tuning (SFT), where training data plays a key role. In addition to collecting data from textbooks, exams, and the Internet, many studies focus on data generation, augmentation, and filtering of training data.

Data Generation. *Rule-based* approaches synthesize geometry problems using predefined generators (Kim and Chun, 2022; Trinh et al., 2024; Kamoi et al., 2024; Huang et al., 2025b) or program templates that build complex diagrams from basic elements (Kazemi et al., 2024a; Zhang et al., 2025c; Sun et al., 2025c). Recent studies further produce high-quality question-answer pairs with reasoning steps by multi-component pipelines (Pan et al., 2025; Fu et al., 2025). *LLM-based* methods generate questions based on math concepts (Tang et al., 2024b; Huang et al., 2025e), with frameworks like GeoUni (Cheng et al., 2025a) and hybrid strategies combining rule-based image generation with LLM-based QA synthesis (Deng et al., 2024). *Agent-based* approaches are also emerging (Lee et al., 2025; Wen et al., 2025), including competition-grade problems from Tonggeometry (Zhang et al., 2024b).

Data Augmentation. To improve robustness, many works apply rule-based augmentation to diversify text and diagrams (Cao and Xiao, 2022; Zhang et al., 2023a, 2024d; Xiao and Zhang, 2023; Lin et al., 2025; Zhuang et al., 2025), use geometry theorems to create new problems (Zhang et al., 2023c; Wu et al., 2024a), or adopt LLMs to generate diverse QA pairs (Tong et al., 2024; Shi et al., 2024; Anand et al., 2024a; Jaiswal et al., 2024; Cheng et al., 2024). In addition, reasoning ability is enhanced by adding annotated *reasoning traces*, including CoT (Gao et al., 2025b; Chen et al., 2024c; Sun et al., 2025b; Luo et al., 2025; Huang et al., 2025d; Ning et al., 2025), PoT (Li et al., 2024d; Sharma et al., 2025), and long CoT (Xu et al., 2024a; Xiang et al., 2024; Xu et al., 2025a; Du et al., 2025). Other works improve geometric understanding by generating aligned *diagrams* for unimodal geometry problems (Zhao et al., 2024; Cai et al., 2024b) or incorporating *diagram descriptions* such as literals and captions (Tey; Zhang et al., 2025e; Xia et al., 2025; Huang et al., 2025f).

Data Filtering. Sun et al. (2025b); Fu et al. (2025); Wang et al. (2025e) use search algorithms to screen data quality and difficulty, while Cai et al. (2024b); Han et al. (2024); Luo et al. (2025); Jia et al. (2025); Huang et al. (2025e) use LLMs to score samples to screen out high-quality data.

3.2.3 Reinforcement Learning

Reinforcement Learning (RL) can significantly improve the geometric reasoning capabilities of deep learning models.

Non-LLM Algorithms. Some studies have used Deep Reinforcement Learning (DRL) methods without LLM to solve geometry problems (Zou et al., 2024), such as the Deep Q-Network (DQN) (Mnih et al., 2013) algorithm (Peng et al., 2023; Huang et al., 2024) and the Proximal Policy Optimization (PPO) (Schulman et al., 2017) algorithm (Xiao and Zhang, 2023).

LLM Algorithms. In LLM-based approaches, RL is typically introduced after SFT. Common algorithms include PPO (Peng et al., 2024, 2025), Direct Preference Optimization (DPO) (Rafailov et al., 2023; Zhang et al., 2025c; Xu et al., 2025a; Huang et al., 2025b), Group Relative Policy Optimization (GRPO) (Guo et al., 2025; Huang et al., 2025d; Deng et al., 2025b; Tan et al., 2025; Deng et al., 2025a; Huang et al., 2025c; Chen et al., 2025b; Liu et al., 2025a; Wang et al., 2025i), and Group Policy Gradient (GPG) (Chu et al., 2025).

3.3 Inference Stage for Geometry Problem Solving

3.3.1 Test-Time Scaling

Test-Time Scaling (TTS) has recently gained attention for significantly enhancing model reasoning during inference.

X-of-Thought. X-of-Thought methods encourage LLMs to produce longer, more diverse outputs, which consume more computational resources than generating only short samples (Zhang et al., 2025b). Many works adopt different CoT (Wei et al., 2022) for GPS (Xu et al., 2024c; Fu et al., 2024; Taveekitworachai et al., 2024), some of which involve multiple rounds of interaction with the model (Zhang et al., 2023d; Zheng et al.). To boost arithmetic accuracy, PoT (Chen et al., 2023a) is used to generate complete programs (DAS et al., 2024; Chen et al., 2024b) or distributed subprograms (Singh et al., 2025). Some studies combine CoT and PoT (Duan et al., 2024; Liu et al., 2023), or integrate CoT with external tools (Qian et al., 2023; Gou et al., 2024). In addition, multimodal CoT approaches generate formal (Zhou et al., 2024c) or natural language (Jia et al., 2024; Tey; Singh et al., 2024) diagram descriptions before reasoning.

Search Methods. Many deep learning systems for GPS integrate tree-based search algorithms to enhance robustness, including Beam Search (Trinh et al., 2024; Chervonyi et al., 2025; Peng et al., 2023; Zhang et al., 2024d; Xu et al., 2024a), Monte Carlo Tree Search (MCTS) (Coulom, 2006; Zou et al., 2024; Rabby et al., 2024; Yao et al., 2024; Dong et al., 2024; Wu et al., 2025), and Predictive Rollout Search (PRS) (Wang et al., 2025f). Graph search is also explored (Xiong et al., 2024).

Verification Methods. A reliable verification method is crucial in TTS. Process Reward Models (PRMs) assess reasoning quality and often guide search paths (Xiang et al., 2024; Luo et al., 2025; Wang et al., 2025c; Tu et al., 2025; Dong et al., 2024; Hu et al., 2025). Other methods include using logits-based confidence (Yuxuan et al., 2024) or training an outcome verifier (Zhang et al., 2025a).

Others. Cheng et al. (2024) uses an LLM to select correct answers from multiple generated candidate solutions, while Jin et al. (2025) proposes an agent framework to manage multiple agents and their reasoning strategies dynamically.

3.3.2 Knowledge-Augmented Inference

Knowledge-augmented inference enhances reasoning by incorporating external knowledge sources.

Few-shot Learning. Few-shot learning (Brown et al., 2020) guides models in solving similar geometry problems. Several studies provide examples through In-Context Learning (ICL) (Agrawal et al., 2024; Cheng et al., 2025b), some of which provide examples based on basic skills (Chen et al., 2024a), some incorporate curriculum learning methods (Vu et al., 2025), and some place text in images (Wang et al., 2024b). Others follow the Retrieval-Augmented Generation (RAG) paradigm to retrieve similar examples as hints (Xu et al., 2024b; Jaiswal et al., 2024; Sharma et al., 2025).

Visual Aids. For GPS, some studies process the corresponding geometric diagrams during the inference stage to help solve the problem. Xiao et al. (2024b) uses drawing tools to convert text problems into multimodal input for reasoning, while Hu et al.; Chen et al. (2025c); Qi et al. (2025); Wang et al. (2025f) facilitate GPS by drawing auxiliary lines or highlighting key features on diagrams.

Others. Guo et al. (2024b) employs learned task plans to guide reasoning, and Yin et al. (2024) leverages explicit memory updates to utilize contextual knowledge captured during training.

4 Evaluations for Geometry Problem Solving

In this section, we summarize the evaluation methods for GPS, including automatic and manual approaches.

4.1 Automatic Evaluation

Automatic metrics include performance-based metrics (outcome-based metrics and process-based metrics) and efficiency-based metrics.

4.1.1 Performance-Based Metrics

Outcome-Based Metrics. Outcome-based metrics focus on measuring the accuracy of final answers without considering reasoning details. *Top-k Accuracy* (*Top-k Acc*) and *Pass@n* (*P@n*) are two main metrics for answer accuracy, measuring the proportion of cases where a correct answer appears in the top k predictions and the proportion of problems solved correctly at least once within n attempts, respectively. Other works also employ outcome-based metrics such as *choice* (proportion of selecting the correct answer from multiple-choice options, or randomly if undetermined) (Zhang et al., 2023a), *F1 score* (considering both precision and recall) (Mishra et al., 2022b; Cheng et al., 2025b), *maj@k* (proportion of obtaining the correct answer via majority vote among k samples) (Yue et al., 2024a), *number of correct and wrong answers* (Dou et al., 2024), and *competition scores* such as SAT (Seo et al., 2015) or IMO scores (Trinh et al., 2024). Most metrics are evaluated using rule-based methods, with some adopting the "LLM-as-a-Judge" paradigm (Li et al., 2024a).

Process-Based Metrics. Recently, increasing attention has been paid to the reasoning process of deep learning systems, beyond just the final results, to further improve model performance. To assess the executability of the reasoning process, *Completion* (Zhang et al., 2023a) measures the accuracy of selecting the first executable solution, while *No Result* (Chen et al., 2021) indicates the ratio of cases where the reasoning program fails to produce output. To evaluate the correctness of reasoning on benchmarks with standard CoT answers (Jaiswal et al., 2024; Qiao et al., 2024), some studies use metrics such as *N-gram Similarity* (Ma et al., 2024b), *Step Accuracy Rate* (Wang et al., 2025b), and *CoT-E score* (Chen et al., 2025a), and extract step answers via rule-based methods or LLMs. For other process-based metrics that are hard to quantify, such as step accuracy without reference CoT (Zhang et al., 2024g; Liu et al., 2024c; Zhou et al., 2024b) or logical coherence of CoT (Zhang et al., 2025e), scoring is typically done with the help of LLMs.

4.1.2 Efficiency-Based Metrics

Efficiency-based metrics measure the model's resource consumption and efficiency performance during reasoning, including the time required to solve the problem (Alvin et al., 2017), the failure rate within a time limit (*timeout*) (Zhang et al., 2023c), the number of inference steps (Wu et al., 2024a; Fang et al., 2024), and the cost of running the model (Balunović et al., 2025).

4.2 Manual Evaluation

Manual evaluation, which is rarely used in GPS, involves experts or annotators directly checking the model's output or reasoning process. Core uses include: (1) evaluating the correctness of complex answers (e.g., judging whether $\frac{1}{\sqrt{2}}$ equals $\frac{\sqrt{2}}{2}$) (Wu et al., 2023); (2) assessing the interpretability of the reasoning process (Sachan et al., 2017; Trinh et al., 2024). Additionally, many studies manually check the reasons for wrong and correct answers, which is also called a case study (Lu et al., 2021; He et al., 2024a).

5 Discussion

5.1 Challenges

Data. First, *current GPS data have significant limitations*. In terms of task type, geometry theorem proving is seriously underrepresented compared to numerical calculation. In terms of geometry type, solid and analytic geometry are lacking relative to plane geometry. In terms of language type, the data is mostly in English and Chinese, with little in other languages. Second, *a large gap remains between synthetic data and real exam questions*. Although recent methods can generate large-scale synthetic data for training, their performance improvement is still limited (Pan et al., 2025; Fu et al., 2025), which highlights the need for methods to synthesize more realistic and effective data. Additionally, *most datasets lack annotations for intermediate steps and reasoning processes* (Shi et al., 2024), which future work should address. More discussion is in Appendix A.

Evaluation. First, *question types are monotonous*. Existing benchmarks mainly use multiple-choice questions for evaluation (see Table 1), allowing models to guess and compromising evaluation accuracy. Some works mitigate this by permuting options (Liu et al., 2024b) or by not providing candidate options (Fu et al., 2025), but these methods have not yet become widespread. Second, *there is no standard method for evaluating the reasoning process*. As the demand for model improvement grows, reasoning evaluation for GPS has gained attention. However, existing methods lack unified standards, and more precise criteria are needed to better identify and address model deficiencies (Park et al., 2024). Additionally, *current benchmarks may lack robustness*, as model performance often varies under slight perturbations (Wang et al., 2025d; Zhou et al., 2024d). Finally, *some datasets may appear in training data*, compromising fair evaluation (Park et al., 2024), underscoring the need for more authoritative evaluation methods.

Capability. Current deep learning systems still show notable deficiencies in solving geometry problems. Given the multimodal nature of most problems, the model's geometric *visual perception* ability is crucial. However, studies show that adding diagrams often lowers accuracy compared to using text alone (Zhang et al., 2025e; Onuoha et al., 2025). In multimodal settings, spatial perception of diagrams remains a major bottleneck limiting overall performance (Sun et al., 2024; Xing et al., 2024; Kamoi et al., 2024; Zhang et al., 2024c). Studies show that deep learning models struggle to detect (Okada et al., 2023; Cho et al., 2025a) and perceive (Wang et al., 2025d; Weng et al., 2025) geometric angles, and often fail to accurately recognize line lengths (Wei et al., 2024; Huang et al., 2025b). These weaknesses may stem from the one-dimensional nature of model architectures (Sun et al., 2025d), the limited resolution of visual encoders (Zhang et al., 2024a; Zhu et al., 2025), and their training on natural images (Hsu et al., 2022; Sharma et al., 2025), all of which hinder performance on geometric figures. Additionally, many models continue to struggle with *arithmetic accuracy*. Some adopt symbolic or formal reasoning (Ning et al., 2025; Chervonyi et al., 2025), while others use external computation modules to mitigate this limitation (Duan et al., 2024; Zhang et al., 2024f; Pan et al., 2024). LLMs may also develop a *mindset*, such as defaulting to coordinate system construction (Sun et al., 2025d), which can fail when such strategies are inapplicable.

5.2 Future Directions

Combining Perception and Reasoning. Studies show that visual perception and reasoning errors are the primary causes of model failures (Park et al., 2024; Wang et al., 2025b). While early efforts targeted reasoning improvements, recent research has shifted toward perception; however, effectively integrating both remains a key challenge. These two aspects are not mutually exclusive but rather complementary. For example, *better modality alignment tasks* can be designed for specialized visual encoders or modules to enhance reasoning; *more efficient multimodal CoT methods* can be explored to achieve deeper integration of perception and reasoning; and *more effective RL strategies*, including training set design and reward mechanisms can be developed. Notably, training sets designed for SFT may not be suitable for RL (Chen et al., 2025b), which calls for careful consideration from the perspectives of diversity and generalization.

Using Cognitive Pattern. Cognitive pattern is a comprehensive approach that simulates human cognitive processes in understanding and solving complex problems (Kurbatov et al., 2021, 2022). Originating from early problem-solving research, many GPS strategies mimicking human problem-solving have proven effective (Zhou and Yu, 2021; Rao et al., 2022), such as *highlighting key information* in diagrams and texts; *referencing diagram annotations*; *adding auxiliary lines*, *coordinate axes*, and other diagram elements to clarify geometric structures; *applying relevant theorems and knowledge*; and using *curriculum learning* to progressively enhance problem-solving ability. However, these methods remain underutilized in current deep learning systems and warrant further investigation.

Educational System. Before the rise of deep learning, many systems and tools had already been developed for geometry education, such as automatic scoring (Mendis et al., 2017), theorem discovery (Kovács and Yu, 2021), and problem-solving systems (Kang et al., 2016; Kurbatov et al., 2020; Kurbatov, 2021; Kurbatov and Fominykh, 2022; Li et al., 2024c), aimed at supporting teaching and learning. However, in the deep learning era, intelligent systems for geometry education remain relatively scarce. Automated GPS is seen as a key direction for future intelligent education (Yang et al., 2023). While recent AI tools have shown progress in solving geometry problems, they still face challenges in becoming effective educational tools—such as limited multi-language support and insufficient visual interaction. Their real-world capabilities remain constrained, and dedicated educational agents are still rare, highlighting the urgent need for further research to tackle the complex demands of this field.

6 Conclusion

In this paper, we present a comprehensive and systematic survey of GPS. We summarize the relevant tasks, deep learning methods, and evaluation approaches, and provide an in-depth analysis of the limitations of current data, evaluation, and model capabilities. Finally, we look forward to possible future research directions and highlight the broad scope for exploration in this field. This article aims to provide readers who are interested in this field with a comprehensive and practical resource to meet their research needs.

Limitations

Our survey focuses on the intersection of deep learning and GPS tasks in the past decade, and may not fully present the development process of the entire field. In addition, given the rapid development of this field, our survey may not timely reflect the latest developments and progress before and after the survey. Furthermore, our survey is mainly dedicated to summarizing existing research work, and there are limitations in experimental analysis. Despite these limitations, this survey still provides a valuable overview of the current status and main trends in the field of deep learning for GPS, which is expected to provide a useful reference for researchers and practitioners in this field.

References

Vansh Agrawal, Pratham Singla, Amitoj Singh Miglani, Shivank Garg, and Ayush Mangal. 2024. Give me a hint: Can llms take a hint to solve math problems? In *The 4th Workshop on Mathematical Reasoning and AI at NeurIPS'24*.

Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. 2024. Large language models for mathematical reasoning: Progresses and challenges. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 225–237.

Chris Alvin, Sumit Gulwani, Rupak Majumdar, and Supratik Mukhopadhyay. 2017. Synthesis of solutions for shaded area geometry problems. In *FLAIRS*, pages 14–19.

Avinash Anand, Raj Jaiswal, Abhishek Dharmadhikari, Atharva Marathe, Harsh Popat, Harshil Mital, Ashwin R Nair, Kritarth Prasad, Sidharth Kumar, Astha Verma, and 1 others. 2024a. Geovqa: A comprehensive multimodal geometry dataset for secondary education. In *2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 102–108. IEEE.

Avinash Anand, Raj Jaiswal, Abhishek Dharmadhikari, Atharva Marathe, Harsh Parimal Popat, Harshil Mital, Kritarth Prasad, Rajiv Ratn Shah, and Roger Zimmermann. 2024b. Improving multimodal llms ability in geometry problem solving, reasoning, and multistep scoring. *arXiv preprint arXiv:2412.00846*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Mislav Balunović, Jasper Dekoninck, Nikola Jovanović, Ivo Petrov, and Martin Vechev. 2025. Mathconstruct: Challenging llm reasoning with constructive proofs. *arXiv preprint arXiv:2502.10197*.

Yves Bertot, Frédérique Guilhot, and Loic Pottier. 2004. Visualizing geometrical statements with geoview. *Electronic Notes in Theoretical Computer Science*, 103:49–65.

Archana Boob and Mansi Radke. 2024a. Leveraging two-level deep learning classifiers for 2d shape recognition to automatically solve geometry math word problems. *Pattern Analysis and Applications*, 27(3):102.

Archana Boob and Mansi Radke. 2025. Elementarycqt: A new dataset and its deep learning analysis for 2d geometric shape recognition. *SN Computer Science*, 6(1):1–14.

Archana Boob and Mansi A Radke. 2024b. 2d shape detection for solving geometry word problems. *IETE Journal of Research*, 70(6):5617–5632.

Archana Boob, Shiva Reddy, Deep Walke, Harshini Pillarisetti, Shreeya Shukla, and Mansi Radke. 2024. Automatic extraction of structured information from elementary level geometry questions into logic forms. *Multimedia Tools and Applications*, pages 1–25.

Archana Narayandas Boob, Prajakta Dnyaneshwar Bodakhe, Mansi Anup Radke, and Umesh Ashok Deshpande. 2023. Extracting structured information from the textual description of geometry word problems. In *Proceedings of the 2023 7th International Conference on Natural Language Processing and Information Retrieval*, pages 31–37.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Huanqia Cai, Yijun Yang, and Zhifeng Li. 2024a. System-2 mathematical reasoning via enriched instruction tuning. *arXiv preprint arXiv:2412.16964*.

Shihao Cai, Keqin Bao, Hangyu Guo, Jizhi Zhang, Jun Song, and Bo Zheng. 2024b. Geogpt4v: Towards geometric multi-modal large language models with geometric image generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 750–766.

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*, pages 1511–1520.

Hyunsik Chae, Seungwoo Yoon, Chloe Yewon Chun, Gyehun Go, Yongin Cho, Gyeongmin Lee, and Ernest K Ryu. 2024. Decomposing complex visual comprehension into atomic visual skills for vision language models. In *The 4th Workshop on Mathematical Reasoning and AI at NeurIPS'24*.

Felix Chen, Hangjie Yuan, Yunqiu Xu, Tao Feng, Jun Cen, Pengwei Liu, Zeying Huang, and Yi Yang. 2025a. Mathflow: Enhancing the perceptual flow of mllms for visual mathematical problems. *arXiv preprint arXiv:2503.16549*.

Hardy Chen, Haoqin Tu, Fali Wang, Hui Liu, Xianfeng Tang, Xinya Du, Yuyin Zhou, and Cihang Xie. 2025b. Sft or rl? an early investigation into training r1-like reasoning large vision-language models. *arXiv preprint arXiv:2504.11468*.

Jiaao Chen, Xiaoman Pan, Dian Yu, Kaiqiang Song, Xiaoyang Wang, Dong Yu, and Jianshu Chen. 2024a. Skills-in-context: Unlocking compositionality in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13838–13890.

Jiaqi Chen, Tong Li, Jinghui Qin, Pan Lu, Liang Lin, Chongyu Chen, and Xiaodan Liang. 2022. Unigeo: Unifying geometry logical reasoning via reformulating mathematical expression. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3313–3323.

Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric Xing, and Liang Lin. 2021. Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 513–523.

Jingchang Chen, Hongxuan Tang, Zheng Chu, Qianglong Chen, Zekun Wang, Ming Liu, and Bing Qin. 2024b. Divide-and-conquer meets consensus: Unleashing the power of functions in code generation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. 2024c. M3cot: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8199–8221.

Steven-Shine Chen, Jimin Lee, and Paul Pu Liang. 2025c. Interactive sketchpad: A multimodal tutoring system for collaborative, visual problem-solving. *arXiv preprint arXiv:2503.16434*.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W Cohen. 2023a. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *Transactions on Machine Learning Research*.

Wenhu Chen, Ming Yin, Max Ku, Pan Lu, Yixin Wan, Xueguang Ma, Jianyu Xu, Xinyi Wang, and Tony Xia. 2023b. Theoremqa: A theorem-driven question answering dataset. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7889–7901.

Yan Chen, Xiaoqing Lu, Jingwei Qu, and Zhi Tang. 2016. Analysis of stroke intersection for overlapping pgf elements. In *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, pages 245–250. IEEE.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024d. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198.

Jo-Ku Cheng, Zeren Zhang, Ran Chen, Jingyang Deng, Ziran Qin, and Jinwen Ma. 2025a. Geouni: A unified model for generating geometry diagrams, problems and problem solutions. *arXiv preprint arXiv:2504.10146*.

Kanzhi Cheng, Yantao Li, Fangzhi Xu, Jianbing Zhang, Hao Zhou, and Yang Liu. 2024. Vision-language models can self-improve reasoning via reflection. *arXiv preprint arXiv:2411.00855*.

Zihui Cheng, Qiguang Chen, Jin Zhang, Hao Fei, Xiaocheng Feng, Wanxiang Che, Min Li, and Libo Qin. 2025b. Comt: A novel benchmark for chain of multi-modal thought on large vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23678–23686.

Yuri Chervonyi, Trieu H Trinh, Miroslav Olšák, Xiaomeng Yang, Hoang Nguyen, Marcelo Menegali, Junehyuk Jung, Vikas Verma, Quoc V Le, and Thang Luong. 2025. Gold-medalist performance in solving olympiad geometry with alphageometry2. *arXiv preprint arXiv:2502.03544*.

Kyunghyun Cho, Bart van Merriënboer, Çağlar G"ulçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734.

Seunghyuk Cho, Zhenyue Qin, Yang Liu, Youngbin Choi, Seungbeom Lee, and Dongwoo Kim. 2025a. Geodano: Geometric vlm with domain agnostic vision encoder. *arXiv preprint arXiv:2502.11360*.

Seunghyuk Cho, Zhenyue Qin, Yang Liu, Youngbin Choi, Seungbeom Lee, and Dongwoo Kim. 2025b. Plane geometry problem solving with multi-modal reasoning: A survey. *arXiv preprint arXiv:2505.14340*.

Xiangxiang Chu, Hailang Huang, Xiao Zhang, Fei Wei, and Yong Wang. 2025. Gpg: A simple and strong reinforcement learning baseline for model reasoning. *arXiv preprint arXiv:2504.02546*.

Rémi Coulom. 2006. Efficient selectivity and backup operators in monte-carlo tree search. In *International conference on computers and games*, pages 72–83. Springer.

DEBRUP DAS, Debopriyo Banerjee, Somak Aditya, and Ashish Kulkarni. 2024. Mathsensei: Mathematical reasoning with a tool-augmented large language model. In *ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models*.

Arash Gholami Davoodi, Seyed Pouyan Mousavi Davoudi, and Pouya Pezeshkpour. 2025. LLMs are not intelligent thinkers: Introducing mathematical topic tree benchmark for comprehensive evaluation of LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3127–3140, Albuquerque, New Mexico. Association for Computational Linguistics.

Huilin Deng, Ding Zou, Rui Ma, Hongchen Luo, Yang Cao, and Yu Kang. 2025a. Boosting the generalization and reasoning of vision language models with curriculum reinforcement learning. *arXiv preprint arXiv:2503.07065*.

Linger Deng, Yuliang Liu, Bohan Li, Dongliang Luo, Liang Wu, Chengquan Zhang, Pengyuan Lyu, Ziyang Zhang, Gang Zhang, Errui Ding, and 1 others. 2024. R-cot: Reverse chain-of-thought problem generation for geometric reasoning in large multimodal models. *arXiv preprint arXiv:2410.17885*.

Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. 2025b. Openvlthinker: An early exploration to complex vision-language reasoning via iterative self-improvement. *arXiv preprint arXiv:2503.17352*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Guanting Dong, Chenghao Zhang, Mengjie Deng, Yutao Zhu, Zhicheng Dou, and Ji-Rong Wen. 2024. Progressive multimodal reasoning via active retrieval. *arXiv preprint arXiv:2412.14835*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, and 1 others. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Jinghao Dou, Xiaopan Lyu, Xinguo Yu, and Hao Wu. 2024. An enhanced relation-flow algorithm for solving number line problems. In *2024 International Conference on Intelligent Education and Intelligent Research (IEIR)*, pages 1–6. IEEE.

Yifan Du, Zikang Liu, Yifan Li, Wayne Xin Zhao, Yuqi Huo, Bingning Wang, Weipeng Chen, Zheng Liu, Zhongyuan Wang, and Ji-Rong Wen. 2025. Virgo: A preliminary exploration on reproducing o1-like mllm. *arXiv preprint arXiv:2501.01904*.

Xiuliang Duan, Dating Tan, Liangda Fang, Yuyu Zhou, Chaobo He, Ziliang Chen, Lusheng Wu, Guanliang Chen, Zhiguo Gong, Weiqi Luo, and 1 others. 2024. Reason-and-execute prompting: Enhancing multimodal large language models for solving geometry questions. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6959–6968.

Anas El Korchi and Youssef Ghanou. 2020. 2d geometric shapes dataset–for machine learning and pattern recognition. *Data in Brief*, 32:106090.

Jeffrey L Elman. 1990. Finding structure in time. *Cognitive science*, 14(2):179–211.

Meng Fang, Xiangpeng Wan, Fei Lu, Fei Xing, and Kai Zou. 2024. Mathodyssey: Benchmarking mathematical problem-solving skills in large language models using odyssey math data. *arXiv preprint arXiv:2406.18321*.

Daocheng Fu, Zijun Chen, Renqiu Xia, Qi Liu, Yuan Feng, Hongbin Zhou, Renrui Zhang, Shiyang Feng, Peng Gao, Junchi Yan, and 1 others. 2025. Trustgeogen: Scalable and formal-verified data engine for trustworthy multi-modal geometric problem solving. *arXiv preprint arXiv:2504.15780*.

Hongguang Fu, Jingzhong Zhang, Xiuqin Zhong, Mingkai Zha, and Li Liu. 2019. Robot for mathematics college entrance examination. In *Electronic Proceedings of the 24th Asian Technology Conference in Mathematics, Mathematics and Technology, LLC*.

Jinlan Fu, Shenzhen Huangfu, Hang Yan, See-Kiong Ng, and Xipeng Qiu. 2024. Hint-before-solving prompting: Guiding llms to effectively utilize encoded knowledge. *arXiv preprint arXiv:2402.14310*.

Wenbin Gan, Xinguo Yu, Sichao Lai, and Lei Xiang. 2016. Plane geometry diagram retrieval by using hierarchical searching strategy. In *Proceedings of the International Conference on Internet Multimedia Computing and Service*, pages 201–206.

Wenbin Gan, Xinguo Yu, Chao Sun, Bin He, and Mingshu Wang. 2017. Understanding plane geometry problems by integrating relations extracted from text and diagram. In *Image and Video Technology: 8th Pacific-Rim Symposium, PSIVT 2017, Wuhan, China, November 20-24, 2017, Revised Selected Papers 8*, pages 366–381. Springer.

Wenbin Gan, Xinguo Yu, and Mingshu Wang. 2019a. Automatic understanding and formalization of plane geometry proving problems in natural language: A supervised approach. *International Journal on Artificial Intelligence Tools*, 28(04):1940003.

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

Bofei Gao, Feifan Song, Zhe Yang, Zefan Cai, Yibo Miao, Chenghao Ma, Shanghaoran Quan, Liang Chen, Qingxiu Dong, Runxin Xu, and 1 others. 2025a. Omni-math: A universal olympiad level mathematic benchmark for large language models. In *The Thirteenth International Conference on Learning Representations*.

Chang Gao, Haiyun Jiang, Deng Cai, Shuming Shi, and Wai Lam. 2024. Strategyllm: Large language models as strategy generators, executors, optimizers, and evaluators for problem solving. *Advances in Neural Information Processing Systems*, 37:96797–96846.

Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing HONG, Jianhua Han, Hang Xu, Zhenguo Li, and Lingpeng Kong. 2025b. G-LLaVA: Solving geometric problem with multi-modal large language model. In *The Thirteenth International Conference on Learning Representations*.

Xiao-Shan Gao and Qiang Lin. 2004. Mmp/geometer–a software package for automated geometric reasoning. In *Automated Deduction in Geometry: 4th International Workshop, ADG 2002, Hagenberg Castle, Austria, September 4-6, 2002. Revised Papers 4*, pages 44–66. Springer.

Herbert Gelernter, James R Hansen, and Donald W Loveland. 1960. Empirical explorations of the geometry theorem machine. In *Papers presented at the May 3-5, 1960, western joint IRE-AIEE-ACM computer conference*, pages 143–149.

Ross Girshick. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, and 1 others. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Salwa Tabet Gonzalez, Stéphane Graham-Lengrand, Julien Narboux, and Natarajan Shankar. 2021. Semantic parsing of geometry statements using supervised machine learning on synthetic data. In *NatFoM 2021-CICM Workshop*.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yujiu Yang, Minlie Huang, Nan Duan, Weizhu Chen, and 1 others. 2024. Tora: A tool-integrated reasoning agent for mathematical problem solving. In *The Twelfth International Conference on Learning Representations*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Fucheng Guo and Pengpeng Jian. 2022. A graph convolutional network feature learning framework for interpretable geometry problem solving. In *2022 International Conference on Intelligent Education and Intelligent Research (IEIR)*, pages 59–64. IEEE.

Jarvis Guo, Tuney Zheng, Yuelin Bai, Bo Li, Yubo Wang, King Zhu, Yizhi Li, Graham Neubig, Wenhu Chen, and Xiang Yue. 2024a. Mammoth-vl: Eliciting multimodal reasoning with instruction tuning at scale. *arXiv preprint arXiv:2412.05237*.

Yiduo Guo, Yaobo Liang, Chenfei Wu, Wenshan Wu, Dongyan Zhao, and Nan Duan. 2024b. Learning to plan by updating natural language. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10062–10098.

Xiaotian Han, Yiren Jian, Xuefeng Hu, Haogeng Liu, Yiqi Wang, Qihang Fan, Yuang Ai, Huaibo Huang, Ran He, Zhenheng Yang, and 1 others. 2024. Infimm-webmath-40b: Advancing multimodal pre-training for enhanced mathematical reasoning. In *The 4th Workshop on Mathematical Reasoning and AI at NeurIPS'24*.

Yihan Hao, Mingliang Zhang, Fei Yin, and Lin-Lin Huang. 2022. Pgdp5k: A diagram parsing dataset for plane geometry problems. In *2022 26th international conference on pattern recognition (ICPR)*, pages 1763–1769. IEEE.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, and 1 others. 2024a. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3828–3850.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969.

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.

Yiming He, Jia Zou, Xiaokai Zhang, Na Zhu, and Tuo Leng. 2024b. Fgeo-tp: A language model-enhanced solver for euclidean geometry problems. *Symmetry*, 16(4):421.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Mark Hopkins, Cristian Petrescu-Prahova, Roie Levin, Ronan Le Bras, Alvaro Herrasti, and Vidur Joshi. 2017. Beyond sentential semantic parsing: Tackling the math sat with a cascade of tree transducers. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 795–804.

Joy Hsu, Jiajun Wu, and Noah Goodman. 2022. Geoclidean: Few-shot generalization in euclidean geometry. *Advances in Neural Information Processing Systems*, 35:39007–39019.

Pengfei Hu, Zhenrong Zhang, Qikai Chang, Shuhang Liu, Jiefeng Ma, Jun Du, Jianshu Zhang, Quan Liu, Jianqing Gao, Feng Ma, and 1 others. 2025. Prm-bas: Enhancing multimodal reasoning through prm-guided beam annealing search. *arXiv preprint arXiv:2504.10222*.

Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Ranjay Krishna. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. In *NeurIPS 2024 Workshop on Behavioral Machine Learning*.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708.

Kaixuan Huang, Jiacheng Guo, Zihao Li, Xiang Ji, Jiawei Ge, Wenzhe Li, Yingqing Guo, Tianle Cai, Hui Yuan, Runzhe Wang, Yue Wu, Ming Yin, Shange Tang, Yangsibo Huang, Chi Jin, Xinyun Chen, Chiyuan Zhang, and Mengdi Wang. 2025a. MATH-perturb: Benchmarking LLMs' math reasoning abilities against hard perturbations. In *Workshop on Reasoning and Planning for Large Language Models*.

Kung-Hsiang Huang, Can Qin, Haoyi Qiu, Philippe Laban, Shafiq Joty, Caiming Xiong, and Chien-Sheng Wu. 2025b. Why vision language models struggle with visual arithmetic? towards enhanced chart and geometry understanding. In *Findings of the Association for Computational Linguistics: ACL 2025*.

Litian Huang, Xinguo Yu, and Bin He. 2022. A novel geometry problem understanding method based on uniform vectorized syntax-semantics model. In *2022 International Conference on Intelligent Education and Intelligent Research (IEIR)*, pages 78–85. IEEE.

Litian Huang, Xinguo Yu, Lei Niu, and Zihan Feng. 2023. Solving algebraic problems with geometry diagrams using syntax-semantics diagram understanding. *Computers, Materials & Continua*, 77(1).

Litian Huang, Xinguo Yu, Feng Xiong, Bin He, Shengbing Tang, and Jiawen Fu. 2024. Hologram reasoning for solving algebra problems with geometry diagrams. *arXiv preprint arXiv:2408.10592*.

Qihan Huang, Long Chan, Jinlong Liu, Wanggui He, Hao Jiang, Mingli Song, Jingyuan Chen, Chang Yao, and Jie Song. 2025c. Boosting mllm reasoning with text-debiased hint-grpo. *arXiv preprint arXiv:2503.23905*.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. 2025d. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*.

Yiming Huang, Xiao Liu, Yeyun Gong, Zhibin Gou, Yelong Shen, Nan Duan, and Weizhu Chen. 2025e. Key-point-driven data synthesis with its enhancement on mathematical reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24176–24184.

Zihan Huang, Tao Wu, Wang Lin, Shengyu Zhang, Jingyuan Chen, and Fei Wu. 2025f. Autogeo: Automating geometric image dataset creation for enhanced geometry understanding. *IEEE Transactions on Multimedia*.

Anca-Elena Iordan. 2022. Usage of stacked long short-term memory for recognition of 3d analytic geometry elements. In *ICAART (3)*, pages 745–752.

Shachar Itzhaky, Sumit Gulwani, Neil Immerman, and Mooly Sagiv. 2013. Solving geometry problems using a combination of symbolic and numerical reasoning. In *Logic for Programming, Artificial Intelligence, and Reasoning: 19th International Conference, LPAR-19, Stellenbosch, South Africa, December 14-19, 2013. Proceedings 19*, pages 457–472. Springer.

Raj Jaiswal, Avinash Anand, and Rajiv Ratn Shah. 2024. Advancing multimodal llms: A focus on geometry problem solving reasoning and sequential scoring. In *Proceedings of the 6th ACM International Conference on Multimedia in Asia*, pages 1–7.

Predrag Janičić and Julien Narboux. 2021. Automated generation of illustrations for synthetic geometry proofs. In *Automated Deduction in Geometry*, volume 352, pages 91–102. Open Publishing Association.

Ishadi Jayasinghe and Surangika Ranathunga. 2020. Two-step memory networks for deep semantic parsing of geometry word problems. In *International Conference on Current Trends in Theory and Practice of Informatics*, pages 676–685. Springer.

Mengzhao Jia, Zhihan Zhang, Wenhao Yu, Fangkai Jiao, and Meng Jiang. 2024. Describe-then-reason: Improving multimodal mathematical reasoning through visual comprehension training. *arXiv preprint arXiv:2404.14604*.

Yiming Jia, Jiachen Li, Xiang Yue, Bo Li, Ping Nie, Kai Zou, and Wenhu Chen. 2025. Visualwebinstruct: Scaling up multimodal instruction data through web search. *arXiv preprint arXiv:2503.10582*.

Pengpeng Jian, Fucheng Guo, Cong Pan, Yanli Wang, Yangrui Yang, and Yang Li. 2023a. Interpretable geometry problem solving using improved retinanet and graph convolutional network. *Electronics*, 12(22):4578.

Pengpeng Jian, Fucheng Guo, Yanli Wang, and Yang Li. 2023b. Solving geometry problems via feature learning and contrastive learning of multimodal data. *CMES-Computer Modeling in Engineering & Sciences*, 136(2).

Zhihuan Jiang, Zhen Yang, Jinhao Chen, Zhengxiao Du, Weihan Wang, Bin Xu, and Jie Tang. 2024. Visscience: An extensive benchmark for evaluating k12 educational multi-modal scientific reasoning. *arXiv preprint arXiv:2409.13730*.

Can Jin, Hongwu Peng, Qixin Zhang, Yujin Tang, Dimitris N Metaxas, and Tong Che. 2025. Two heads are better than one: Test-time scaling of multi-agent collaborative reasoning. *arXiv preprint arXiv:2504.09772*.

Vidur Joshi, Matthew Peters, and Mark Hopkins. 2018. Extending a parser to distant domains using a few dozen partially annotated examples. *arXiv preprint arXiv:1805.06556*.

Ryo Kamoi, Yusen Zhang, Sarkar Snigdha Sarathi Das, Ranran Haoran Zhang, and Rui Zhang. 2024. Visonlyqa: Large vision language models still struggle with visual perception of geometric information. *arXiv preprint arXiv:2412.00947*.

Bo Kang, Arun Kulshreshth, and Joseph J LaViola Jr. 2016. Analyticalink: An interactive learning environment for math word problem solving. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*, pages 419–430.

Deepak Kapur. 1986. Using gröbner bases to reason about geometry problems. *Journal of Symbolic Computation*, 2(4):399–408.

Mehran Kazemi, Hamidreza Alvari, Ankit Anand, Jialin Wu, Xi Chen, and Radu Soricut. 2024a. Geomverse: A systematic evaluation of large models for geometric reasoning. In *AI for Math Workshop@ ICML 2024*.

Mehran Kazemi, Nishanth Dikkala, Ankit Anand, Petar Devic, Ishita Dasgupta, Fangyu Liu, Bahare Fatemi, Pranjal Awasthi, Sreenivas Gollapudi, Dee Guo, and 1 others. 2024b. Remi: A dataset for reasoning with multiple images. *Advances in Neural Information Processing Systems*, 37:60088–60109.

Jiwoo Kim, Youngbin Kim, Ilwoong Baek, JinYeong Bak, and Jongwuk Lee. 2023. It ain't over: A multi-aspect diverse math word problem dataset. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14984–15011.

Kangmin Kim and Chanjun Chun. 2022. Synthetic data generator for solving korean arithmetic word problem. *Mathematics*, 10(19):3525.

Thomas N Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*.

Zoltán Kovács and Jonathan H. Yu. 2021. Automated discovery of geometrical theorems in geogebra. In *Proceedings 10th International Workshop on Theorem Proving Components for Educational Software, ThEdu@CADE 2021, (Remote) Carnegie Mellon University, Pittsburgh, PA, United States, 11 July 2021*, volume 354 of *EPTCS*, pages 1–12.

Ryan Krueger, Jesse Michael Han, and Daniel Selsam. 2021. Automatically building diagrams for olympiad geometry problems. In *CADE*, pages 577–588.

Sergey Kurbatov, Igor Fominykh, and Aleksandr Vorobyev. 2021. Cognitive patterns for semantic presentation of natural-language descriptions of well-formalizable problems. In *Russian Conference on Artificial Intelligence*, pages 317–330. Springer.

Sergey S Kurbatov and Igor B Fominykh. 2022. Complex modeling of inductive and deductive reasoning by the example of a planimetric problem solver. In *International Conference on Intelligent Information Technologies for Industry*, pages 454–462. Springer.

Sergey S Kurbatov, Igor B Fominykh, and Aleksandr B Vorobyev. 2020. Ontology-controlled geometric solver. In *Artificial Intelligence: 18th Russian Conference, RCAI 2020, Moscow, Russia, October 10–16, 2020, Proceedings 18*, pages 262–273. Springer.

Sergey S Kurbatov, Xenia A Naidenova, Vyacheslav P Ganapolsky, and Tatyana A Martirova. 2022. Natural language processing and functioning ontological solver with visualization in an integrated system. In *Proceedings of SAI Intelligent Systems Conference*, pages 669–682. Springer.

Sergeyi S Kurbatov. 2021. Linguistic processor integration for solving planimetric problems. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 15(4):1–14.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Jeongwoo Lee, Kwangsuk Park, and Jihyeon Park. 2025. Vista: Visual integrated system for tailored automation in math problem generation using llm. In *Large Foundation Models for Educational Assessment*, pages 136–156. PMLR.

Bin Lei, Yi Zhang, Shan Zuo, Ali Payani, and Caiwen Ding. 2024. Macm: Utilizing a multi-agent system for condition mining in solving complex mathematical problems. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, and 1 others. 2024a. From generation to judgment: Opportunities and challenges of llm-as-a-judge. *arXiv preprint arXiv:2411.16594*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. 2024b. Multi-modal arxiv: A dataset for improving scientific comprehension of large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14369–14387.

Lei Li, Zongkai Yang, Mao Chen, Xicheng Peng, Jianwen Sun, Zhonghua Yan, and Sannyuya Liu. 2024c. Automated generation of geometry proof problems based on point geometry identity. *Journal of Automated Reasoning*, 68(2):11.

Nianqi Li, Zujie Liang, Siyu Yuan, Jiaqing Liang, Feng Wei, and Yanghua Xiao. 2024d. Multilingpot: Enhancing mathematical reasoning with multilingual program fine-tuning. *arXiv preprint arXiv:2412.12609*.

Zhaoyu Li, Jialiang Sun, Logan Murphy, Qidong Su, Zenan Li, Xian Zhang, Kaiyu Yang, and Xujie Si. 2024e. A survey on deep learning for theorem proving. In *First Conference on Language Modeling*.

Zhaoyu Li, Jialiang Sun, Logan Murphy, Qidong Su, Zenan Li, Xian Zhang, Kaiyu Yang, and Xujie Si. 2024f. A survey on deep learning for theorem proving. In *First Conference on Language Modeling*.

Zhihao Li, Yao Du, Yang Liu, Yan Zhang, Yufang Liu, Mengdi Zhang, and Xunliang Cai. 2024g. Eagle: Elevating geometric reasoning through llm-empowered visual instruction tuning. *arXiv preprint arXiv:2408.11397*.

Zhong-Zhi Li, Ming-Liang Zhang, Fei Yin, and Cheng-Lin Liu. 2024h. Lans: A layout-aware neural solver for plane geometry problem. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 2596–2608.

Zhongzhi Li, Ming-Liang Zhang, Pei-Jie Wang, Jian Xu, Rui-Song Zhang, Yin Fei, Zhi-Long Ji, Jin-Feng Bai, Zhen-Ru Pan, Jiaxin Zhang, and 1 others. 2025. Cmmath: A chinese multi-modal math skill evaluation benchmark for foundation models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2690–2726.

Zhenwen Liang, Tianyu Yang, Jipeng Zhang, and Xiangliang Zhang. 2023. Unimath: A foundational and multimodal mathematical reasoner. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7126–7133.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Zi-Hao Lin, Shun-Xin Xiao, Zi-Rong Chen, Jian-Min Li, Da-Han Wang, and Xu-Yao Zhang. 2025. Sans: Spatial-aware neural solver for plane geometry problem. In *International Conference on Pattern Recognition*, pages 183–196. Springer.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Hongwei Liu, Zilong Zheng, Yuxuan Qiao, Haodong Duan, Zhiwei Fei, Fengzhe Zhou, Wenwei Zhang, Songyang Zhang, Dahua Lin, and Kai Chen. 2024b. Mathbench: Evaluating the theory and application proficiency of llms with a hierarchical mathematics benchmark. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6884–6915.

Lu Liu, Xiaoqing Lu, Songping Fu, Jingwei Qu, Liangcai Gao, and Zhi Tang. 2014a. Plane geometry figure retrieval based on bilayer geometric attributed graph matching. In *2014 22nd International Conference on Pattern Recognition*, pages 309–314. IEEE.

Lu Liu, Xiaoqing Lu, Keqiang Li, Jingwei Qu, Liangcai Gao, and Zhi Tang. 2014b. Plane geometry figure retrieval with bag of shapes. In *2014 11th IAPR International Workshop on Document Analysis Systems*, pages 1–5. IEEE.

Lu Liu, Xiaoqing Lu, Yuan Liao, Yongtao Wang, and Zhi Tang. 2016. Improving retrieval of plane geometry figure with learning to rank. *Pattern Recognition Letters*, 83:423–429.

Qing Tang Liu, Huan Huang, and Lin Jing Wu. 2012. Using restricted natural language for geometric construction. *Applied Mechanics and Materials*, 145:465–469.

Tengxiao Liu, Qipeng Guo, Yuqing Yang, Xiangkun Hu, Yue Zhang, Xipeng Qiu, and Zheng Zhang. 2023. Plan, verify and switch: Integrated reasoning with diverse x-of-thoughts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2807–2822.

Wentao Liu, Qianjun Pan, Yi Zhang, Zhuo Liu, Ji Wu, Jie Zhou, Aimin Zhou, Qin Chen, Bo Jiang, and Liang He. 2024c. Cmm-math: A chinese multimodal math dataset to evaluate and enhance the mathematics reasoning of large multimodal models. *arXiv preprint arXiv:2409.02834*.

Xiangyan Liu, Jinjie Ni, Zijian Wu, Chao Du, Longxu Dou, Haonan Wang, Tianyu Pang, and Michael Qizhe Shieh. 2025a. Noisyrollout: Reinforcing visual reasoning with data augmentation. *arXiv preprint arXiv:2504.13055*.

Yexiang Liu, Jie Cao, Zekun Li, Ran He, and Tieniu Tan. 2025b. Breaking mental set to improve reasoning through diverse multi-agent debate. In *The Thirteenth International Conference on Learning Representations*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations*.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. 2021. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. *arXiv preprint arXiv:2105.04165*.

Pan Lu, Liang Qiu, Wenhao Yu, Sean Welleck, and Kai-Wei Chang. 2023. A survey of deep learning for mathematical reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14605–14631.

Xiaoqing Lu, Lu Liu, Zhi Tang, and Haibin Ling. 2015. Overlapped-triangle analysis with hierarchical ranking of dominance. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 791–795. IEEE.

Ruilin Luo, Zhuofan Zheng, Yifan Wang, Yiyao Yu, Xinzhe Ni, Zicheng Lin, Jin Zeng, and Yujiu Yang. 2025. Ursa: Understanding and verifying chain-of-thought reasoning in multimodal mathematics. *arXiv preprint arXiv:2501.04686*.

Bin Ma, Pengpeng Jian, Cong Pan, Yanli Wang, and Wei Ma. 2024a. A geometric neural solving method based on a diagram text information fusion analysis. *Scientific Reports*, 14(1):31906.

Jingkun Ma, Runzhe Zhan, Derek F Wong, Yang Li, Di Sun, Hou Pong Chan, and Lidia S Chao. 2024b. Visaidmath: Benchmarking visual-aided mathematical reasoning. *arXiv preprint arXiv:2410.22995*.

Jaroslav Macke, Jiri Sedlar, Miroslav Olsak, Josef Urban, and Josef Sivic. 2021. Learning to solve geometric construction problems from images. In *International Conference on Intelligent Computer Mathematics*, pages 167–184. Springer.

Abeer Mahgoub, Ghada Khoriba, and ElHassan Anas ElSabry. 2024. Mathematical problem solving in arabic: Assessing large language models. *Procedia Computer Science*, 244:86–95.

Takuya Matsuzaki, Takumi Ito, Hidenao Iwane, Hirokazu Anai, and Noriko H Arai. 2017. Semantic parsing of pre-university math problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2131–2141.

Maxwell-Jia. 2024. Aime 2024. https://huggingface.co/datasets/Maxwell-Jia/AIME_2024. Accessed: 2025-06-06.

Chamupathi Mendis, Dhanushka Lahiru, Naduni Pamudika, Supun Madushanka, Surangika Ranathunga, and Gihan Dias. 2017. Automatic assessment of student answers for geometric theorem proving questions. In *2017 Moratuwa Engineering Research Conference (MERCon)*, pages 413–418. IEEE.

Swaroop Mishra, Matthew Finlayson, Pan Lu, Leonard Tang, Sean Welleck, Chitta Baral, Tanmay Rajpurohit, Oyvind Tafjord, Ashish Sabharwal, Peter Clark, and 1 others. 2022a. Lila: A unified benchmark for mathematical reasoning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5807–5832.

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 (Volume 1: Long Papers)*, pages 3505–3523.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. 2013. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*.

Spyridon Mouselinos, Henryk Michalewski, and Mateusz Malinowski. 2024. Beyond lines and circles: Unveiling the geometric reasoning gap in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6192–6222.

Logan Murphy, Kaiyu Yang, Jialiang Sun, Zhaoyu Li, Anima Anandkumar, and Xujie Si. 2024. Autoformalizing euclidean geometry. In *International Conference on Machine Learning*, pages 36847–36893. PMLR.

Julien Narboux, Predrag Janicic, and Jacques Fleuriot. 2018. Computer-assisted theorem proving in synthetic geometry. *Handbook of Geometric Constraint Systems Principles*, pages 25–73.

Maizhen Ning, Qiu-Feng Wang, Kaizhu Huang, and Xiaowei Huang. 2023. A symbolic characters aware model for solving geometry problems. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7767–7775.

Maizhen Ning, Zihao Zhou, Qiufeng Wang, Xiaowei Huang, and Kaizhu Huang. 2025. Gns: Solving plane geometry problems by neural-symbolic reasoning with multi-modal llms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24957–24965.

Kazuaki Okada, Masashi Kudo, and Hayato Yamana. 2023. Estimating answer strategies using online handwritten data: A study using geometry problems. In *Proceedings of the 15th International Conference on Education Technology and Computers*, pages 308–314.

Chibuike Onuoha, Yusuke Haga, Dilshad Ferdousi, and Truong Cong Thang. 2025. Multimodal large language models for high school mathematical reasoning: Impact of input modality and artifacts. *Authorea Preprints*.

Cong Pan, Pengpeng Jian, Bin Ma, Ling Feng, and Xuemei Liu. 2023. The geometric neural solution combined with text diagram parsing. In *2023 International Conference on Intelligent Education and Intelligent Research (IEIR)*, pages 1–7. IEEE.

Yicheng Pan, Zhenrong Zhang, Pengfei Hu, Jiefeng Ma, Jun Du, Jianshu Zhang, Quan Liu, Jianqing Gao, and Feng Ma. 2025. Enhancing the geometric problem-solving ability of multimodal llms via symbolic-neural integration. *arXiv preprint arXiv:2504.12773*.

Yicheng Pan, Zhenrong Zhang, Jiefeng Ma, Pengfei Hu, Jun Du, Qing Wang, Jianshu Zhang, Dan Liu, and Si Wei. 2024. Maths: Multimodal transformer-based human-readable solver. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE.

Jieun Park, Sungeun Park, Hyungbae Jeon, and Joon-Ho Lim. 2024. What is the true performance of large multimodal models in visual context-based mathematical reasoning? an analysis of multiple datasets and future research directions. In *2024 15th International Conference on Information and Communication Technology Convergence (ICTC)*, pages 854–859. IEEE.

Shuai Peng, Di Fu, Liangcai Gao, Xiuqin Zhong, Hongguang Fu, and Zhi Tang. 2024. Multimath: Bridging visual and mathematical reasoning for large language models. *arXiv preprint arXiv:2409.00147*.

Shuai Peng, Di Fu, Yijun Liang, Liangcai Gao, and Zhi Tang. 2023. Geodrl: A self-learning framework for geometry problem solving using reinforcement learning in deductive reasoning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13468–13480.

Yingzhe Peng, Gongrui Zhang, Miaosen Zhang, Zhiyuan You, Jie Liu, Qipeng Zhu, Kai Yang, Xingzhong Xu, Xin Geng, and Xu Yang. 2025. Lmm-r1: Empowering 3b lmms with strong reasoning abilities through two-stage rule-based rl. *arXiv preprint arXiv:2503.07536*.

Ji Qi, Ming Ding, Weihan Wang, Yushi Bai, Qingsong Lv, Wenyi Hong, Bin Xu, Lei Hou, Juanzi Li, Yuxiao Dong, and Jie Tang. 2025. Cogcom: A visual language model with chain-of-manipulations reasoning. In *The Thirteenth International Conference on Learning Representations*.

Cheng Qian, Chi Han, Yi Fung, Yujia Qin, Zhiyuan Liu, and Heng Ji. 2023. Creator: Tool creation for disentangling abstract and concrete reasoning of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6922–6939.

Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma GongQue, Shanglin Lei, Zhe Wei, Miaoxuan Zhang, and 1 others. 2024. We-math: Does your large multimodal model achieve human-like mathematical reasoning? *arXiv preprint arXiv:2407.01284*.

Jingwei Qu, Xiaoqing Lu, Songping Fu, and Zhi Tang. 2016. Improving pgf retrieval effectiveness with active learning. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 1125–1130. IEEE.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Gollam Rabby, Farhana Keya, Parvez Zamil, and Sören Auer. 2024. Mc-nest–enhancing mathematical reasoning in large language models with a monte carlo nash equilibrium self-refine tree. *arXiv preprint arXiv:2411.15645*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741.

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

Andrea Raffo, Andrea Ranieri, Chiara Romanengo, Bianca Falcidieno, and Silvia Biasotti. 2024. Curveml: a benchmark for evaluating and training learning-based methods of classification, recognition, and fitting of plane curves. *The Visual Computer*, pages 1–21.

Yongsheng Rao, Lanxing Xie, Hao Guan, Jing Li, and Qixin Zhou. 2022. A method for expanding predicates and rules in automated geometry reasoning system. *Mathematics*, 10(7):1177.

Mrinmaya Sachan, Avinava Dubey, Eduard H Hovy, Tom M Mitchell, Dan Roth, and Eric P Xing. 2019. 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 the 2017 conference on empirical methods in natural language processing*, 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 (* SEM 2017)*, pages 251–261.

V Sanh. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *Proceedings of Thirty-third Conference on Neural Information Processing Systems (NIPS2019)*.

Amrutesh Saraf, Pooja Kamat, Shilpa Gite, Satish Kumar, and Ketan Kotecha. 2024. Towards robust automated math problem solving: a survey of statistical and deep learning approaches. *Evolutionary Intelligence*, 17(5):3113–3150.

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2008. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80.

Daniel Scher. 1999. Lifting the curtain: The evolution of the geometer's sketchpad. *The Mathematics Educator*, 10(2).

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Minjoon Seo, Hannaneh Hajishirzi, Ali Farhadi, Oren Etzioni, and Clint Malcolm. 2015. Solving geometry problems: Combining text and diagram interpretation. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1466–1476.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Aditya Sharma, Aman Dalmia, Mehran Kazemi, Amal Zouaq, and Christopher Pal. 2025. GeoCoder: Solving geometry problems by generating modular code through vision-language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7340–7356, Albuquerque, New Mexico. Association for Computational Linguistics.

Yan Shengyuan and Zhong Xiuqin. 2024. Geo-qwen: A geometry problem-solving method based on generative large language models and heuristic reasoning. In *2024 21st International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, pages 1–9. IEEE.

Wenhao Shi, Zhiqiang Hu, Yi Bin, Junhua Liu, Yang Yang, See Kiong Ng, Lidong Bing, and Roy Lee. 2024. Math-llava: Bootstrapping mathematical reasoning for multimodal large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4663–4680.

Ayush Singh, Mansi Gupta, Shivank Garg, Abhinav Kumar, and Vansh Agrawal. 2024. Beyond captioning: Task-specific prompting for improved vlm performance in mathematical reasoning. *arXiv preprint arXiv:2410.05928*.

Kunal Singh, Ankan Biswas, Sayandeep Bhowmick, Pradeep Moturi, and Siva Kishore Gollapalli. 2025.

SBSC: step-by-step coding for improving mathematical olympiad performance. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Shiven Sinha, Ameya Prabhu, Ponnurangam Kumaraguru, Siddharth Bhat, and Matthias Bethge. Wu's method boosts symbolic ai to rival silver medalists and alphageometry to outperform gold medalists at imo geometry. In *The 4th Workshop on Mathematical Reasoning and AI at NeurIPS'24*.

Bing Song, Gang Xiong, Zhen Shen, Fenghua Zhu, Yisheng Lv, and Peijun Ye. 2023. Geometry problem solving based on counter-factual evolutionary reasoning. In *2023 IEEE 19th International Conference on Automation Science and Engineering (CASE)*, pages 1–6. IEEE.

Dan Song, Dongming Wang, and Xiaoyu Chen. 2017. Retrieving geometric information from images: the case of hand-drawn diagrams. *Data Mining and Knowledge Discovery*, 31(4):934–971.

Dan Song, Qiong Yao, and Junjie Lu. 2020. A novel geometric information retrieval tool for images of geometric diagrams. In *2020 International Conference on Information Science and Education (ICISE-IE)*, pages 403–411. IEEE.

Wei Song, Yadong Li, Jianhua Xu, Guowei Wu, Lingfeng Ming, Kexin Yi, Weihua Luo, Houyi Li, Yi Du, Fangda Guo, and 1 others. 2024. M3gia: A cognition inspired multilingual and multimodal general intelligence ability benchmark. *arXiv preprint arXiv:2406.05343*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, and 1 others. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *TRANSACTIONS ON MACHINE LEARNING RESEARCH*.

Saurabh Srivastava, Anto PV, Shashank Menon, Ajay Sukumar, Alan Philipose, Stevin Prince, Sooraj Thomas, and 1 others. 2024. Functional benchmarks for robust evaluation of reasoning performance, and the reasoning gap. *arXiv preprint arXiv:2402.19450*.

Haoxiang Sun, Yingqian Min, Zhipeng Chen, Wayne Xin Zhao, Zheng Liu, Zhongyuan Wang, Lei Fang, and Ji-Rong Wen. 2025a. Challenging the boundaries of reasoning: An olympiad-level math benchmark for large language models. *arXiv preprint arXiv:2503.21380*.

Kai Sun, Yushi Bai, Ji Qi, Lei Hou, and Juanzi Li. 2024. Mm-math: Advancing multimodal math evaluation with process evaluation and fine-grained classification. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1358–1375.

Linzhuang Sun, Hao Liang, Jingxuan Wei, Bihui Yu, Tianpeng Li, Fan Yang, Zenan Zhou, and Wentao Zhang. 2025b. Mm-verify: Enhancing multimodal reasoning with chain-of-thought verification. *arXiv preprint arXiv:2502.13383*.

Ruiyong Sun, Yijia Zhao, Qi Zhang, Keyu Ding, Shijin Wang, and Cui Wei. 2019. A neural semantic parser for math problems incorporating multi-sentence information. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 18(4):1–16.

Yanpeng Sun, Shan Zhang, Wei Tang, Aotian Chen, Piotr Koniusz, Kai Zou, Yuan Xue, and Anton van den Hengel. 2025c. Mathglance: Multimodal large language models do not know where to look in mathematical diagrams. *arXiv preprint arXiv:2503.20745*.

Yiyou Sun, Georgia Zhou, Hao Wang, Dacheng Li, Nouha Dziri, and Dawn Song. 2025d. Climbing the ladder of reasoning: What llms can-and still can't-solve after sft? *arXiv preprint arXiv:2504.11741*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and 1 others. 2023. Challenging big-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051.

Huajie Tan, Yuheng Ji, Xiaoshuai Hao, Minglan Lin, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. 2025. Reason-rft: Reinforcement fine-tuning for visual reasoning. *arXiv preprint arXiv:2503.20752*.

Jiamin Tang, Chao Zhang, Xudong Zhu, and Mengchi Liu. 2024a. Tangram: A challenging benchmark for geometric element recognizing. *arXiv preprint arXiv:2408.13854*.

Zhengyang Tang, Xingxing Zhang, Benyou Wang, and Furu Wei. 2024b. Mathscale: Scaling instruction tuning for mathematical reasoning. In *International Conference on Machine Learning*, pages 47885–47900. PMLR.

Pittawat Taveekitworachai, Febri Abdullah, and Ruck Thawonmas. 2024. Null-shot prompting: rethinking prompting large language models with hallucination. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13321–13361.

Joseph Tey. Understanding how vision-language models reason when solving visual math problems.

Yuxuan Tong, Xiwen Zhang, Rui Wang, Ruidong Wu, and Junxian He. 2024. Dart-math: Difficulty-aware rejection tuning for mathematical problem-solving. *Advances in Neural Information Processing Systems*, 37:7821–7846.

Trieu H Trinh, Yuhuai Wu, Quoc V Le, He He, and Thang Luong. 2024. Solving olympiad geometry without human demonstrations. *Nature*, 625(7995):476–482.

Shih-Hung Tsai, Chao-Chun Liang, Hsin-Min Wang, and Keh-Yih Su. 2021. Sequence to general tree: Knowledge-guided geometry word problem solving. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 964–972.

George Tsoukalas, Jasper Lee, John Jennings, Jimmy Xin, Michelle Ding, Michael Jennings, Amitayush Thakur, and Swarat Chaudhuri. 2024. Putnambench: Evaluating neural theorem-provers on the putnam mathematical competition. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Haoqin Tu, Weitao Feng, Hardy Chen, Hui Liu, Xianfeng Tang, and Cihang Xie. 2025. Vilbench: A suite for vision-language process reward modeling. *arXiv preprint arXiv:2503.20271*.

Aaron Van Den Oord, Oriol Vinyals, and 1 others. 2017. Neural discrete representation learning. *Advances in neural information processing systems*, 30.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Duc Anh Vu, Cong-Duy Nguyen, Xiaobao Wu, Nhat Hoang, Mingzhe Du, Thong Nguyen, and Anh Tuan Luu. 2025. Curriculum demonstration selection for in-context learning. In *Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing*, pages 1004–1006.

Junxiao Wang, Ting Zhang, Heng Yu, Jingdong Wang, and Hua Huang. 2025a. Magicgeo: Training-free text-guided geometric diagram generation. *arXiv preprint arXiv:2502.13855*.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. 2024a. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169.

Lei Wang, Wanyu Xu, Zhiqiang Hu, Yihuai Lan, Shan Dong, Hao Wang, Roy Ka-Wei Lee, and Ee-Peng Lim. 2024b. All in an aggregated image for in-image learning. *arXiv preprint arXiv:2402.17971*.

Long Wang and Jianhui Ma. 2024. Mutli-step chain-of-thought in geometry problem solving. In *2024 4th International Conference on Electronic Information Engineering and Computer Science (EIECS)*, pages 1113–1117. IEEE.

Peijie Wang, Zhong-Zhi Li, Fei Yin, Xin Yang, Dekang Ran, and Cheng-Lin Liu. 2025b. Mv-math: Evaluating multimodal math reasoning in multi-visual contexts. *arXiv preprint arXiv:2502.20808*.

Weiyun Wang, Zhangwei Gao, Lianjie Chen, Zhe Chen, Jinguo Zhu, Xiangyu Zhao, Yangzhou Liu, Yue Cao, Shenglong Ye, Xizhou Zhu, and 1 others. 2025c. Visualprm: An effective process reward model for multimodal reasoning. *arXiv preprint arXiv:2503.10291*.

Xiaofeng Wang, Yiming Wang, Wenhong Zhu, and Rui Wang. 2025d. Do large language models truly understand geometric structures? In *The Thirteenth International Conference on Learning Representations*.

Xiyao Wang, Zhengyuan Yang, Chao Feng, Hongjin Lu, Linjie Li, Chung-Ching Lin, Kevin Lin, Furong Huang, and Lijuan Wang. 2025e. Sota with less: Mcts-guided sample selection for data-efficient visual reasoning self-improvement. *arXiv preprint arXiv:2504.07934*.

Yikun Wang, Siyin Wang, Qinyuan Cheng, Zhaoye Fei, Liang Ding, Qipeng Guo, Dacheng Tao, and Xipeng Qiu. 2025f. Visuothink: Empowering lvlm reasoning with multimodal tree search. *arXiv preprint arXiv:2504.09130*.

Yiming Wang, Pei Zhang, Jialong Tang, Haoran Wei, Baosong Yang, Rui Wang, Chenshu Sun, Feitong Sun, Jiran Zhang, Junxuan Wu, and 1 others. 2025g. Polymath: Evaluating mathematical reasoning in multilingual contexts. *arXiv preprint arXiv:2504.18428*.

Ying Wang, Wei Zhou, Yongsheng Rao, and Hao Guan. 2025h. A knowledge and semantic fusion method for automatic geometry problem understanding. *Applied Sciences*, 15(7):3857.

Yiping Wang, Qing Yang, Zhiyuan Zeng, Liliang Ren, Lucas Liu, Baolin Peng, Hao Cheng, Xuehai He, Kuan Wang, Jianfeng Gao, and 1 others. 2025i. Reinforcement learning for reasoning in large language models with one training example. *arXiv preprint arXiv:2504.20571*.

Zhikai Wang, Jiashuo Sun, Wenqi Zhang, Zhiqiang Hu, Xin Li, Fan Wang, and Deli Zhao. 2025j. Benchmarking multimodal mathematical reasoning with explicit visual dependency. *arXiv preprint arXiv:2504.18589*.

Haoran Wei, Youyang Yin, Yumeng Li, Jia Wang, Liang Zhao, Jianjian Sun, Zheng Ge, Xiangyu Zhang, and Daxin Jiang. 2024. Slow perception: Let's perceive geometric figures step-by-step. *arXiv preprint arXiv:2412.20631*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Zixin Wen, Yifu Cai, Kyle Lee, Sam Estep, Joshua Sunshine, Aarti Singh, Yuejie Chi, and Wode Ni. 2025. Feynman: Knowledge-infused diagramming agent for scaling visual reasoning data.

Tengjin Weng, Jingyi Wang, Wenhao Jiang, and Zhong Ming. 2025. Visnumbench: Evaluating number sense of multimodal large language models. *arXiv preprint arXiv:2503.14939*.

Man Fai Wong, Xintong Qi, and Chee Wei Tan. 2023. Euclidnet: Deep visual reasoning for constructible problems in geometry. *Advances in Artificial Intelligence and Machine Learning*, 3(1):839–853.

Haoyi Wu, Wenyang Hui, Yezeng Chen, Weiqi Wu, Kewei Tu, and Yi Zhou. 2023. Conic10k: A challenging math problem understanding and reasoning dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6444–6458.

Jinyang Wu, Mingkuan Feng, Shuai Zhang, Ruihan Jin, Feihu Che, Zengqi Wen, and Jianhua Tao. 2025. Boosting multimodal reasoning with mcts-automated structured thinking. *arXiv preprint arXiv:2502.02339*.

Wenjun Wu, Lingling Zhang, Jun Liu, Xi Tang, Yaxian Wang, Shaowei Wang, and Qianying Wang. 2024a. E-gps: Explainable geometry problem solving via top-down solver and bottom-up generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13828–13837.

Yanan Wu, Jie Liu, Xingyuan Bu, Jiaheng Liu, Zhanhui Zhou, Yuanxing Zhang, Chenchen Zhang, ZhiqiBai ZhiqiBai, Haibin Chen, Tiezheng Ge, and 1 others. 2024b. Conceptmath: A bilingual concept-wise benchmark for measuring mathematical reasoning of large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6815–6839.

Jing Xia and Xinguo Yu. 2021. A paradigm of diagram understanding in problem solving. In *2021 IEEE International Conference on Engineering, Technology & Education (TALE)*, pages 531–535. IEEE.

Renqiu Xia, Mingsheng Li, Hancheng Ye, Wenjie Wu, Hongbin Zhou, Jiakang Yuan, Tianshuo Peng, Xinyu Cai, Xiangchao Yan, Bin Wang, Conghui He, Botian Shi, Tao Chen, Junchi Yan, and Bo Zhang. 2025. Geox: Geometric problem solving through unified formalized vision-language pre-training. In *The Thirteenth International Conference on Learning Representations*.

Kun Xiang, Zhili Liu, Zihao Jiang, Yunshuang Nie, Kaixin Cai, Yiyang Yin, Runhui Huang, Haoxiang Fan, Hanhui Li, Weiran Huang, and 1 others. 2025. Can atomic step decomposition enhance the self-structured reasoning of multimodal large models? *arXiv preprint arXiv:2503.06252*.

Kun Xiang, Zhili Liu, Zihao Jiang, Yunshuang Nie, Runhui Huang, Haoxiang Fan, Hanhui Li, Weiran Huang, Yihan Zeng, Jianhua Han, and 1 others. 2024. Atomthink: A slow thinking framework for multimodal mathematical reasoning. *arXiv preprint arXiv:2411.11930*.

Tong Xiao, Jiayu Liu, Zhenya Huang, Jinze Wu, Jing Sha, Shijin Wang, and Enhong Chen. 2024a. Learning to solve geometry problems via simulating human dual-reasoning process. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 6559–6568.

Ziyang Xiao and Dongxiang Zhang. 2023. A deep reinforcement learning agent for geometry online tutoring. *Knowledge and Information Systems*, 65(4):1611–1625.

Ziyang Xiao, Dongxiang Zhang, Xiongwei Han, Xiaojin Fu, Wing Yin Yu, Tao Zhong, Sai Wu, Yuan Wang, Jianwei Yin, and Gang Chen. 2024b. Enhancing llm reasoning via vision-augmented prompting. *Advances in Neural Information Processing Systems*, 37:28772–28797.

Shangyu Xing, Changhao Xiang, Yuteng Han, Yifan Yue, Zhen Wu, Xinyu Liu, Zhangtai Wu, Fei Zhao, and Xinyu Dai. 2024. Gepbench: Evaluating fundamental geometric perception for multimodal large language models. *arXiv preprint arXiv:2412.21036*.

Siheng Xiong, Ali Payani, Yuan Yang, and Faramarz Fekri. 2024. Deliberate reasoning for llms as structure-aware planning with accurate world model. *arXiv preprint arXiv:2410.03136*.

Guowei Xu, Peng Jin, Li Hao, Yibing Song, Lichao Sun, and Li Yuan. 2024a. Llava-o1: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*.

Haotian Xu, Xing Wu, Weinong Wang, Zhongzhi Li, Da Zheng, Boyuan Chen, Yi Hu, Shijia Kang, Jiaming Ji, Yingying Zhang, and 1 others. 2025a. Redstar: Does scaling long-cot data unlock better slow-reasoning systems? *arXiv preprint arXiv:2501.11284*.

Liangyu Xu, Yingxiu Zhao, Jingyun Wang, Yingyao Wang, Bu Pi, Chen Wang, Mingliang Zhang, Jihao Gu, Xiang Li, Xiaoyong Zhu, and 1 others. 2025b. Geosense: Evaluating identification and application of geometric principles in multimodal reasoning. *arXiv preprint arXiv:2504.12597*.

Shihao Xu, Yiyang Luo, and Wei Shi. 2024b. Geo-llava: A large multi-modal model for solving geometry math problems with meta in-context learning. In *Proceedings of the 2nd Workshop on Large Generative Models Meet Multimodal Applications*, pages 11–15.

Weijia Xu, Andrzej Banburski, and Nebojsa Jojic. 2024c. Reprompting: Automated chain-of-thought prompt inference through gibbs sampling. In *International Conference on Machine Learning*, pages 54852–54865. PMLR.

Tianfan Xue, Jianzhuang Liu, and Xiaoou Tang. 2010. Object cut: Complex 3d object reconstruction through line drawing separation. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1149–1156. IEEE.

Tianfan Xue, Jianzhuang Liu, and Xiaoou Tang. 2012. Example-based 3d object reconstruction from line drawings. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 302–309. IEEE.

Kai Yan, Yufei Xu, Zhengyin Du, Xuesong Yao, Zheyu Wang, Xiaowen Guo, and Jiecao Chen. 2025. Recitation over reasoning: How cutting-edge language models can fail on elementary school-level reasoning problems? *arXiv preprint arXiv:2504.00509*.

Yibo Yan, Jiamin Su, Jianxiang He, Fangteng Fu, Xu Zheng, Yuanhuiyi Lyu, Kun Wang, Shen Wang, Qingsong Wen, and Xuming Hu. 2024. A survey of mathematical reasoning in the era of multimodal large language model: Benchmark, method & challenges. *arXiv preprint arXiv:2412.11936*.

Bo Yang, Qingping Yang, Yingwei Ma, and Runtao Liu. 2024a. Utmath: Math evaluation with unit test via reasoning-to-coding thoughts. *arXiv preprint arXiv:2411.07240*.

Linjie Yang, Jianzhuang Liu, and Xiaoou Tang. 2013. Complex 3d general object reconstruction from line drawings. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1433–1440.

Yingxuan Yang, Bo Huang, Siyuan Qi, Chao Feng, Haoyi Hu, Yuxuan Zhu, Jinbo Hu, Haoran Zhao, Ziyi He, Xiao Liu, and 1 others. 2025. Who's the mvp? a game-theoretic evaluation benchmark for modular attribution in llm agents. *arXiv preprint arXiv:2502.00510*.

Yulong Yang, Hongguang Fu, Xiuqin Zhong, and Tianming Zhang. 2023. Suffi-gpsc: Sufficient geometry problem solution checking with symbolic computation and logical reasoning. In *2023 20th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, pages 1–9. IEEE.

Zhen Yang, Jinhao Chen, Zhengxiao Du, Wenmeng Yu, Weihan Wang, Wenyi Hong, Zhihuan Jiang, Bin Xu, and Jie Tang. 2024b. Mathglm-vision: Solving mathematical problems with multi-modal large language model. *arXiv preprint arXiv:2409.13729*.

Huanjin Yao, Jiaxing Huang, Wenhao Wu, Jingyi Zhang, Yibo Wang, Shunyu Liu, Yingjie Wang, Yuxin Song, Haocheng Feng, Li Shen, and 1 others. 2024. Mulberry: Empowering mllm with o1-like reasoning and reflection via collective monte carlo tree search. *arXiv preprint arXiv:2412.18319*.

Katherine Ye, Wode Ni, Max Krieger, Dor Ma'ayan, Jenna Wise, Jonathan Aldrich, Joshua Sunshine, and Keenan Crane. 2020. Penrose: from mathematical notation to beautiful diagrams. *ACM Transactions on Graphics (TOG)*, 39(4):144–1.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Zhiyuan Zeng, Qinyuan Cheng, Xipeng Qiu, and Xuan-Jing Huang. 2024. Explicit memory learning with expectation maximization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16618–16635.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, and 1 others. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Wei Yu, Mengzhu Wang, Xiaodong Wang, Xun Zhou, Yongfu Zha, Yongjian Zhang, Shuyu Miao, and Jingdong Liu. 2021a. Geore: A relation extraction dataset for chinese geometry problems. In *35th Conference on Neural Information Processing Systems (NeurIPS 2021) Workshop on Math AI for Education (MATHAI4ED)*.

Xinguo Yu, Wenbin Gan, Danfeng Yang, and Sichao Lai. 2015. Automatic reconstruction of plane geometry figures in documents. In *2015 International Conference of Educational Innovation through Technology (EITT)*, pages 46–50. IEEE.

Xinguo Yu, Yixing Geng, and Zihan Feng. 2021b. Solving solid geometric calculation problems in text. In *2021 IEEE International Conference on Engineering, Technology & Education (TALE)*, pages 525–530. IEEE.

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

Albert S Yue, Lovish Madaan, Ted Moskovitz, DJ Strouse, and Aaditya K Singh. 2024a. Harp: A challenging human-annotated math reasoning benchmark. *arXiv preprint arXiv:2412.08819*.

Xiang Yue, Tianyu Zheng, Ge Zhang, and Wenhu Chen. 2024b. Mammoth2: Scaling instructions from the web. *Advances in Neural Information Processing Systems*, 37:90629–90660.

YAO Yuxuan, Han Wu, Zhijiang Guo, Zhou Biyan, Jiahui Gao, Sichun Luo, Hanxu Hou, Xiaojin Fu, and Linqi Song. 2024. Learning from correctness without prompting makes llm efficient reasoner. In *First Conference on Language Modeling*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986.

Chao Zhang, Jing Xiao, and Xiaoyang Fu. 2024a. Enhancing geometry problem solving with attention mechanism and super-resolution. In *2024 5th International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE)*, pages 112–115. IEEE.

Chi Zhang, Jiajun Song, Siyu Li, Yitao Liang, Yuxi Ma, Wei Wang, Yixin Zhu, and Song-Chun Zhu. 2024b. Proposing and solving olympiad geometry with guided tree search. *arXiv preprint arXiv:2412.10673*.

Dongxiang Zhang. 2022. Deep learning in automatic math word problem solvers. In *AI in Learning: Designing the Future*, pages 233–246. Springer International Publishing Cham.

Jiarui Zhang, Ollie Liu, Tianyu Yu, Jinyi Hu, and Willie Neiswanger. 2024c. Euclid: Supercharging multimodal llms with synthetic high-fidelity visual descriptions. *arXiv preprint arXiv:2412.08737*.

Jiaxin Zhang, Yinghui Jiang, and Yashar Moshfeghi. 2024d. Gaps: geometry-aware problem solver. *arXiv preprint arXiv:2401.16287*.

Jiaxin Zhang, Zhong-Zhi Li, Ming-Liang Zhang, Fei Yin, Cheng-Lin Liu, and Yashar Moshfeghi. 2024e. Geoeval: Benchmark for evaluating llms and multimodal models on geometry problem-solving. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 1258–1276.

Jiaxin Zhang and Yashar Moshfeghi. 2024. Gold: Geometry problem solver with natural language description. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 263–278.

Lunjun Zhang, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. 2025a. Generative verifiers: Reward modeling as next-token prediction. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Ming-Liang Zhang, Zhong-Zhi Li, Fei Yin, Liang Lin, and Cheng-Lin Liu. 2024f. Fuse, reason and verify: Geometry problem solving with parsed clauses from diagram. *arXiv preprint arXiv:2407.07327*.

Ming-Liang Zhang, Fei Yin, Yi-Han Hao, and Cheng-Lin Liu. 2022a. Learning to understand plane geometry diagram [c]. In *36th Conference on Neural Information Processing Systems (NeurIPS) Workshop on MATH-AI*.

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

Ming-Liang Zhang, Fei Yin, and Cheng-Lin Liu. 2023a. A multi-modal neural geometric solver with textual clauses parsed from diagram. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 3374–3382.

Qiyuan Zhang, Fuyuan Lyu, Zexu Sun, Lei Wang, Weixu Zhang, Wenyue Hua, Haolun Wu, Zhihan Guo, Yufei Wang, Niklas Muennighoff, and 1 others. 2025b. A survey on test-time scaling in large language models: What, how, where, and how well? *arXiv preprint arXiv:2503.24235*.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, and 1 others. 2024g. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pages 169–186. Springer.

Renrui Zhang, Xinyu Wei, Dongzhi Jiang, Ziyu Guo, Yichi Zhang, Chengzhuo Tong, Jiaming Liu, Aojun Zhou, Shanghang Zhang, Peng Gao, and Hongsheng Li. 2025c. MAVIS: Mathematical visual instruction tuning with an automatic data engine. In *The Thirteenth International Conference on Learning Representations*.

Shan Zhang, Aotian Chen, Yanpeng Sun, Jindong Gu, Yi-Yu Zheng, Piotr Koniusz, Kai Zou, Anton van den Hengel, and Yuan Xue. 2025d. Open eyes, then reason: Fine-grained visual mathematical understanding in mllms. *arXiv preprint arXiv:2501.06430*.

Wenxuan Zhang, Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023b. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. *Advances in Neural Information Processing Systems*, 36:5484–5505.

Xiaokai Zhang, Na Zhu, Yiming He, Jia Zou, Qike Huang, Xiaoxiao Jin, Yanjun Guo, Chenyang Mao, Yang Li, Zhe Zhu, and 1 others. 2023c. Formalgeo: An extensible formalized framework for olympiad geometric problem solving. *arXiv preprint arXiv:2310.18021*.

Xiaokai Zhang, Na Zhu, Cheng Qin, Yang Li, Zhenbing Zeng, and Tuo Leng. 2024h. Fgeo-hypergnet: Geometric problem solving integrating formal symbolic system and hypergraph neural network. *arXiv preprint arXiv:2402.11461*.

Xiaokai Zhang, Na Zhu, Cheng Qin, LI Yang, Zhenbing Zeng, and Tuo Leng. 2024i. Formal representation and solution of plane geometric problems. In *The 4th Workshop on Mathematical Reasoning and AI at NeurIPS'24*.

Yifan Zhang, Jingqin Yang, Yang Yuan, and Andrew Chi-Chih Yao. 2023d. Cumulative reasoning with large language models. *arXiv preprint arXiv:2308.04371*.

Zeren Zhang, Jo-Ku Cheng, Jingyang Deng, Lu Tian, Jinwen Ma, Ziran Qin, Xiaokai Zhang, Na Zhu, and Tuo Leng. 2025e. Diagram formalization enhanced multi-modal geometry problem solver. In *ICASSP*

2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1–5. IEEE.

Junbo Zhao, Ting Zhang, Jiayu Sun, Mi Tian, and Hua Huang. 2025a. Pi-gps: Enhancing geometry problem solving by unleashing the power of diagrammatic information. *arXiv preprint arXiv:2503.05543*.

Xueliang Zhao, Xinting Huang, Tingchen Fu, Qintong Li, Shansan Gong, Lemao Liu, Wei Bi, and Lingpeng Kong. 2024. Bba: Bi-modal behavioral alignment for reasoning with large vision-language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 7255–7279.

Yurui Zhao, Xiang Wang, Jiahong Liu, Irwin King, and Zhitao Huang. 2025b. Towards geometry problem solving in the large model era: A survey. *arXiv preprint arXiv:2506.02690*.

Chuanyang Zheng, Zhengying Liu, Enze Xie, Zhenguo Li, and Yu Li. Progressive-hint prompting improves reasoning in large language models. In *AI for Math Workshop@ ICML 2024*.

Jinxin Zheng, Yongtao Wang, and Zhi Tang. 2015. Solid geometric object reconstruction from single line drawing image. In *International Conference on Computer Graphics Theory and Applications*, volume 2, pages 391–400. SCITEPRESS.

Jinxin Zheng, Yongtao Wang, and Zhi Tang. 2016a. Context-aware geometric object reconstruction for mobile education. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 367–371.

Jinxin Zheng, Yongtao Wang, and Zhi Tang. 2016b. Recovering solid geometric object from single line drawing image. *Multimedia Tools and Applications*, 75:10153–10174.

Hu Zhengyu and Zhong Xiuqin. 2023. A precise text-to-diagram generation method for elementary geometry. In *2023 20th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, pages 1–7. IEEE.

Junjie Zhou, Zheng Liu, Shitao Xiao, Bo Zhao, and Yongping Xiong. 2024a. Vista: Visualized text embedding for universal multi-modal retrieval. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3185–3200.

Mingrui Zhou and Xinguo Yu. 2021. Proving geometric problem by adding auxiliary lines-based on hypothetical test. In *International Conference on Artificial Intelligence in Education Technology*, pages 151–161. Springer.

Minxuan Zhou, Hao Liang, Tianpeng Li, Zhiyu Wu, Mingan Lin, Linzhuang Sun, Yaqi Zhou, Yan Zhang, Xiaoqin Huang, Yicong Chen, and 1 others. 2024b. Mathscape: Evaluating mllms in multimodal math scenarios through a hierarchical benchmark. *arXiv preprint arXiv:2408.07543*.

Motian Zhou, Chen Wu, and Chao Sun. 2024c. Evaluating automated geometric problem solving with formal language generation on large multimodal models. In *2024 International Conference on Intelligent Education and Intelligent Research (IEIR)*, pages 1–7. IEEE.

Wei Zhou, Ruixi Xu, Hao Guan, Jietong Zhao, and Yongsheng Rao. 2022. Research on geometry problem text understanding based on bidirectional lstm-crf. In *2022 9th International Conference on Digital Home (ICDH)*, pages 121–127. IEEE.

Zihao Zhou, Shudong Liu, Maizhen Ning, Wei Liu, Jindong Wang, Derek F Wong, Xiaowei Huang, Qiufeng Wang, and Kaizhu Huang. 2024d. Is your model really a good math reasoner? evaluating mathematical reasoning with checklist. *arXiv preprint arXiv:2407.08733*.

Na Zhu, Xiaokai Zhang, Qike Huang, Fangzhen Zhu, Zhenbing Zeng, and Tuo Leng. 2025. Fgeo-parser: Autoformalization and solution of plane geometric problems. *Symmetry*, 17(1):8.

Wenwen Zhuang, Xin Huang, Xiantao Zhang, and Jin Zeng. 2025. Math-puma: Progressive upward multimodal alignment to enhance mathematical reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 26183–26191.

Changqing Zou, Tianfan Xue, Xiaojiang Peng, Honghua Li, Baochang Zhang, Ping Tan, and Jianzhuang Liu. 2016. An example-based approach to 3d man-made object reconstruction from line drawings. *Pattern Recognition*, 60:543–553.

Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. 2025. Dynamath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models. In *The Thirteenth International Conference on Learning Representations*.

Jia Zou, Xiaokai Zhang, Yiming He, Na Zhu, and Tuo Leng. 2024. Fgeo-drl: Deductive reasoning for geometric problems through deep reinforcement learning. *Symmetry*, 16(4):437.

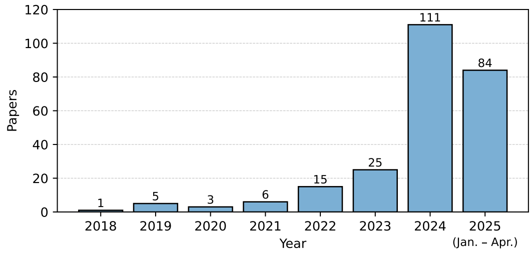

Figure 5: Papers on deep learning for geometry problem solving over the years (data for 2025 is up to April).

A Geometry Problem Solving Datasets

In this section, we further analyze various datasets for GPS. Table 1 and Table 2 provide a comprehensive summary of these datasets related to GPS tasks from multiple perspectives, including dataset name, task type, geometry type, grade level, problem source, presence of images, language, question format, rationale availability, sizes of training, validation, and test sets, as well as open-source status; a check mark indicates open-source datasets with links to the corresponding resources.

The current data for geometry theorem proving remains insufficient. Existing academic research predominantly centers on geometric numerical calculations, whereas studies on geometry theorem proving are relatively limited, and relevant data resources are still lacking. Despite sharing many similarities in problem formulation and underlying mathematical concepts (Chen et al., 2022), proof problems and calculation problems have distinct characteristics and challenges. Therefore, both types of geometry problems deserve equal attention.

The current data for solid geometry and analytic geometry remains insufficient. Most datasets used in GPS tasks are concentrated in plane geometry, while data for other geometry types—such as solid geometry (Yu et al., 2021b) and analytic geometry (Wu et al., 2023)—remain limited. One study notes that existing solid geometry problems are often overly simple and regular (Xu et al., 2025b), with diagrams containing only basic visual elements and rarely involving complex geometric combinations, thereby restricting progress in this area. Even within plane geometry, high-quality evaluation datasets are still scarce.

The current data sources remain limited. While existing datasets are generally authentic and reliable, they are often small in scale. Recently, due to the shortage of real-world data and concerns over copyright, many large-scale datasets have been constructed via data augmentation or programmatic synthesis (Gao et al., 2025b; Pan et al., 2025). However, the synthetic data often falls short in terms of realism, diversity, and quality, making it difficult to serve as a full substitute for real data.

The current data coverage of language and question types remains limited. In terms of language, existing datasets primarily cover English and Chinese, while authentic data involving other native languages (Zhang et al., 2023b; Song et al., 2024) remains notably limited. This limits evaluation in the context of various national exams and reduces fairness. In terms of question types, most are multiple-choice, which allows models to guess answers and impairs accurate assessment of model reasoning ability.

The current datasets remain lacking in rationale annotations. Most datasets do not provide detailed annotations of intermediate reasoning steps (Shi et al., 2024). Even when rationales are included, they often lack standardized formatting and sufficient granularity, falling short of the needs for evaluating step-by-step reasoning. Moreover, the rationale annotations are typically presented in natural language, which may not meet the needs of deep learning systems that operate in formal languages.

B Other Geometry Tasks

In addition to GPS, some other geometry-related tasks, which have similar fundamental tasks, have not been systematically summarized. More details of the corresponding datasets can be found in Table 3.

B.1 Geometric Diagram Generation

This task is dedicated to generating high-quality geometric diagrams. It aims to facilitate a deeper understanding of geometry problems and related applications such as image editing, thereby providing strong support for the field of education.

Geometric Diagram Reconstruction. This task is one of the earlier works in the field of geometry. It aims to use existing simple sketches or preliminarily drawn images to reconstruct a clearer and more standardized complete image, thereby helping users to understand and visualize the image content more intuitively (Yu et al., 2015). One of the key challenges is to reconstruct 3D geometry

Datasets	Task	Type	Grade	Source	Image	Language	Question	Rationale	Trainval Size	Test Size	Opensource
					Fundamental Tasks						
GeoC50 (2017)	RE	**P**	-	exist (dataset)	✓	zh	FR	-	-	50	✗
2Dgeometricshapes (2020)	ER	**P**	-	program	✓	-	CQ	-	36000	54000	✓
GeoRE (2021a)	RE	**P**	6-12	Internet	✗	zh	FR	-	10000	2901	✓
PGDP5K (2022; 2022b; 2022a)	DP	**P**	6-12	exist, textbook	✓	en	FR	-	4000	1000	✓
Geoclidean (2022)	SR	**P**	-	program	✓	en	YN	-	185	555	✓
BBH-geometricshapes (2023)	ER	**P**	-	program	✗	en	MC	-	-	250	✓
Tangram (2024a)	ER	**P**, **S**	1-12	exam, textbook	✓	en	NR	-	-	4320	✓
GeoCQT (2024a)	ER	**P**	-	exist, textbook	✓	-	CQ	-	∼11200	∼2800	✗
SP-1 (2024)	DP	**P**	-	program	✓	en	FR	-	200000	480	✓
ElementaryGeometryQA (2024)	DP, SP	**P**	1-5	textbook	✓	en	FR	-	-	∼500	✗
Geoperception (2024c)	SR	**P**	6-12	exist	✓	en	SA	-	-	11657	✓
GePBench (2024)	ER, SR	**P**	-	program	✓	en	MC	-	∼300000	285000	✗
CurveML (2024)	ER	**P**	-	program	✓	-	CQ	-	468000	52000	✓
AVSBench* (2024)	ER, SR	**P**	-	program	✓	en	MC, FR	-	-	5073	✓
VisOnlyQA* (2024)	ER, SR	**P**	-	exist, program	✓	en	MC, YN	-	70000	1600	✓
AutoGeo-100k (2025f)	DC	**P**	-	program	✓	en	FR	-	100000	-	✓
Geo170K-alignment (2025b)	DC	**P**	6-12	exist	✓	en	FR	-	60252	-	✓
GeomRel (2025d)	SR	**P**	-	program	✗	en	MC	-	-	2629	✓
ElementaryCQT (2025)	ER	**P**	-	program	✓	-	CQ	-	342000	38000	✓
SynthGeo228K (2025e)	DP, DC	**P**	-	program	✓	en	FR	-	205491	22833	✓
formalgeo-structure774k (2025e)	DP, DC	**P**	6-12	exist	✓	en	FR	-	∼774000	-	✗
VGPR (2025a)	ER, SR	**P**	-	program	✓	en	MC	-	300000	50000	✗
GeoX-alignment (2025)	DC	**P**	-	Internet	✓	en	FR	-	6232	-	✓
VisNumBench* (2025)	ER, SR	**P**	-	exist, prog, web	✓	en	MC	-	-	1913	✓
GeoPeP* (2025c)	ER, SR	**P**, **S**	-	program	✓	en	FR	nl	200000	-	✓
MathGlance* (2025c)	ER, SR	**P**, **S**	-	exist, program	✓	en	MC, YN, FR	-	-	1609	✓
CogAlign-Probing* (2025b)	SR	**P**	-	program	✓	en	YN	-	44000	4000	✓
CogAlign-train* (2025b)	SR	**P**, **S**	-	program	✓	en	FR	-	64000	-	✓
					Core Tasks						
GEOS (2015)	NC	**P**	6-10	exam	✓	en	MC	-	67	119	✓
GeoShader (2017)	NC	**P**	6-10	textbook, exam	✓	en	NR	-	-	102	✗
GEOS++ (2017; 2019)	NC	**P**	6-10	textbook	✓	en	MC	-	500	906	✗
GEOS-OS (2017)	NC	**P**	6-10	textbook	✓	en	MC	demonstration	2235	-	✗
Geometry3K (2021)	NC	**P**	6-12	online library	✓	en	MC	-	2401	601	✓
GeoQA (2021)	NC	**P**	6-12	exam	✓	zh	MC	program	4244	754	✓
Geometry3Dcalculation (2021b)	NC	**S**	-	website	✗	en, zh	NR	-	-	140	✗
Proving2H (2021)	TP	**P**	6-9	textbook, Internet	✗	zh	FR	-	-	110	✗
GeometryQA (2021)	NC	**P**	1-6	exist	✗	zh	NR	equations	1118	280	✓
GeoQA+ (2022)	NC	**P**	6-12	website	✓	zh	MC	program	12054	-	✓
UniGeo (2022)	TP, NC	**P**	9-12	website, exist	✓	en	MC, FR	program	12340	2201	✓
BIG-bench-IG (2022)	NC	**P**	-	program	✗	en	NR	-	-	250000	✓
PGPS9K (2023a)	NC	**P**	6-12	exist, textbook	✓	en	NR	program	8022	1000	✓
formalgeo7k (2024i)	NC	**P**	6-12	exist	✓	en, zh	NR	formal	∼5934	∼1047	✓
formalgeo-imo (2023c)	TP	**P**	-	online	✓	en, zh	FR	formal	-	18	✓
Conic10K (2023)	NC	**A**	10-12	website	✗	zh	FR	nl	8793	2068	✓
GeomVerse (2024a)	NC	**P**	-	program	✓	en	NR	nl	11190	29000	✓
IMO-AG-30 (2024)	TP	**P**	-	exam	✗	en	FR	-	-	30	✓
aug-Geo3K (2024a)	NC	**P**	6-12	exist	✓	en	MC	nl	13783	3824	✗
GeoEval (2024e)	NC	**P**, **S**, **A**	1-12	exist, online	✓	en	MC	-	-	5050	✓
GeoGPT4V-GPS (2024b)	TP, NC	**P**	6-12	exist	✓	en, zh	MC, FR	nl	16557	-	✓
GeoVQA (2024a)	TP, NC	**P**, **S**	6-12	textbook	✓	en	NR, FR	nl	4440	150	✗
GeoMath (2024b)	TP, NC	**S**	10-12	website	✓	en	NR, FB, FR	nl	9155	906	✗
GeoMM (2024)	NC	**P**	-	program	✓	en	NR	nl	87000	-	✓
GPSM4K (2024b; 2024)	TP, NC	**P**	7-12	textbook	✓	en	NR, FR	nl	4272	1068	✗
NBLP (2024)	NC	**P**	7-9	textbook, exam	✓	en	NR, YN	-	-	100	✗
G-MATH (2024)	NC	**P**, **S**	9-12	exist	✓	en	FR	-	-	187	✗
MathCheck-GEO (2024d)	MR	**P**	6-12	exist	✓	en	NR, YN, FR	nl	-	1440	✓
MO-TG-225 (2024b)	TP	**P**	-	exam	✗	en	FR	-	-	225	✗
Geo170K-qa (2025b)	NC	**P**	6-12	exist	✓	en	MC	nl	117205	-	✓
FormalGeo7K-v2 (2025)	NC	**P**	6-12	exist	✓	en, zh	NR	formal	5950	1050	✓
VerMulti-Geo (2025)	NC	**P**	6-12	exist	✓	en	MC	-	15000	-	✗
GeoMath-8K (2025)	NC	**P**	6-12	exist	✓	en	MC	-	4500	820	✗
GNS-260K (2025)	KP, NC	**P**	6-12	exist	✓	en	MC, NR, SA	program, nl	260017	-	✗
GeoExpand (2025)	TP, NC	**P**	6-12	exist	✓	en	MC, FR	nl	45526	-	✓
GeoSynth (2025)	TP, NC	**P**	-	program	✓	en	MC, FR	nl	62868	-	✓
IMO-AG-50 (2025)	TP	**P**	-	exam	✗	en	FR	-	-	50	✗
GeoTrust (2025)	NC	**P**	-	program	✓	en	NR	nl	∼200000	240	✗
GeoSense (2025b)	KP, NC	**P**, **S**	6-12	exist, website	✓	en, zh	MC, FR	-	-	1789	✗
formalgeo-reasoning238k (2025e)	NC	**P**	6-12	exist	✓	en	NR	nl	∼238000	-	✗

Table 1: A summarization of geometry problem solving datasets for fundamental tasks and core tasks. Task: ER: geometric element recognition, SR: geometric structure recognition, DP: geometric diagram parsing, DC: geometric diagram captioning, SP: semantic parsing for geometry problem texts, RE: geometric relation extraction, KP: geometric knowledge prediction, TP: geometry theorem proving, NC: geometric numerical calculation. Type: **P**: plane geometry, **S**: solid geometry, **A**: analytic geometry. Question: MC: multiple-choice, NR: numerical response, FR: free-response, FB: fill-in-the-blank, YN: yes-or-no, SA: short-answer, CQ: classification question. Rationale: nl: natural language. * indicates that the dataset contains more than just geometry-related content.

Datasets	Task	Type	Grade	Source	Image	Language	Question	Rationale	Trainval Size	Test Size	Opensource
						Composite Tasks					
MATH (2021)	MR	**P, S**	9-12	exam	✗	en	NR	nl	7500	5000	✓
AMPS (2021)	MR	**P, S**	1-12	website, program	✗	en	NR, FR	nl	∼5100k	-	✓
NumGLUE (2022b)	MR	**P**	6-10	exist	✗	en	CQ	-	81466	10583	✓
Lila (2022a)	MR	**P, S**	-	exist	✗	en	MC, FB, FR	program	107052	26763	✓
DMath (2023)	MR	**P**	1-6	handcraft	✗	en, kr	NR	program	7943	2079	✓
TheoremQA (2023b)	MR	**P**	-	Internet, expert	✓	en	MC, YN, FR	-	-	∼350	✓
M3Exam (2023b)	MR	**P, S**	1-12	exam	✓	9 lang.	MC	-	-	12317	✓
OlympiadBench (2024a)	MR	**P, S**	-	exam	✗	en, zh	NR, FR	nl	-	8476	✓
MathVista (2024)	MR	**P**	6-12	exist	✓	en, zh	MC, FR	-	-	6141	✓
MathVerse (2024g)	MR	**P, S**	-	exist, website	✓	en	MC, FR	nl	-	15672	✓
Math-Vision (2024a)	MR	**P, S, A**	1-12	exam	✓	en	MC, FR	-	-	3040	✓
MM-Math (2024)	MR	**P**	7-9	website	✓	en	FR	nl	-	5929	✓
We-Math (2024)	MR	**P, S**	3-6	website	✓	en	MC	-	-	6524	✓
VisAidMath (2024b)	MR	**P, S, A**	7-12	exam	✓	en	MC, FR, YN	-	-	1200	✗
CMM-Math (2024c)	MR	**P, S, A**	1-12	exam	✓	zh	MC, FB, YN, FR	nl	22248	5821	✓
MathScape (2024b)	MR	**P, S**	1-12	homework, exam	✓	en	NR, FB, FR	nl	-	1325	✓
VisScience (2024)	MR	**P, S**	1-12	-	✓	en, zh	MC, FR	-	-	3000	✗
ArXivQA (2024b)	MR	**P**	-	paper	✓	en	MC	nl	100000	-	✓
ReMI (2024b)	MR	**P**	-	-	✓	en	MC, NR, FR	-	-	2600	✓
MathV360K (2024)	MR	**P**	9-16	exist	✓	en	MC, NR, FR	-	360000	-	✓
MultiMath-300K (2024)	MR	**P, S**	1-12	textbook, exam	✓	en, zh	FB, NR, FR	nl	290227	8443	✓
InfiMM-WebMath-40B (2024)	MR	-	-	website	✓	en, zh	-	-	∼24000k	-	✓
MathVL (2024b)	MR	**P, S, A**	1-12	exist, private	✓	en	MC, FB, FR	nl	484914	2000	✗
ArMATH (2024)	MR	**P**	1-6	school	✗	ar	FR	-	-	200	✗
M3CoT (2024c)	MR	**P**	-	exist	✓	en	MC	nl	9100	2359	✓
MathOdyssey (2024)	MR	**P, S**	10-16	expert	✗	en	MC, YN, FR	nl	-	387	✓
PutnamBench (2024)	MR	-	-	exam	✗	en	FR	formal	-	1692	✓
ConcepetMath (2024b)	MR	**P, S**	1-9	website, textbook	✗	en, zh	NR	-	-	4011	✓
MATH() (2024)	MR	**P, S**	9-12	exist	✗	en	NR	-	-	2060	✗
MathBench (2024b)	MR	**P, S**	1-16	exist, exam	✗	en, zh	MC, FR	-	-	3709	✓
HARP (2024a)	MR	**P**	-	website	✗	en	MC, SA, FR	nl	-	5409	✓
M3GIA (2024)	MR	**P**	6-12	exam	✓	6 lang.	MC	-	-	1800	✓
DART-Math (2024)	MR	**P, S**	9-12	exist	✗	en	NR	nl	∼1180k	-	✓
MathScaleQA (2024b)	MR	**P, S**	1-16	exist, exam	✗	en	FR	nl	2000000	-	✓
UTMath (2024a)	MR	**P, S**	-	OEIS	✗	en	NR	-	-	1053	✓
MultiLingPoT (2024d)	MR	**P, S**	9-12	exist	✗	program	NR	program	41134	-	✓
EITMath (2024a)	MR	**P, S**	9-12	exist	✗	en	NR	nl	15000	-	✗
AIME2024 (2024)	MR	**P, S**	-	exam	✗	en	NR	nl	-	30	✓
AMATH-SFT (2024; 2025)	MR	**P**	-	exist	✓	en	MC, FR	nl	∼124000	-	✓
MMathCoT-1M (2025)	MR	**P**	-	exist	✓	en	MC, NR, FR	nl	∼1020k	-	✓
DynaMath (2025)	MR	**P, S, A**	1-16	exist, website	✓	en	MC, FR	nl	-	5010	✓
CoMT (2025b)	MR	**P**	-	exist	✓	en	MC	nl	-	3853	✓
Diagramma (2025)	MR	**P**	-	program	✓	en	MC	-	-	1058	✗
MV-MATH (2025b)	MR	**P, S, A**	1-12	textbook, exam	✓	en	MC, FR	nl	-	2009	✓
CMMaTH (2025)	MR	**P, S, A**	1-12	website	✓	zh	MC, FR	nl	-	23856	✗
Math-PUMA-1M (2025)	MR	**P, S**	-	exist, online, prog	✓	en	FR	nl	996000	-	✓
VisualWebInstruct (2025)	MR	**P**	1-16	exist, Internet	✓	en	-	nl	906160	-	✓
MAVIS-Instruct (2025c)	MR	**P, S, A**	-	exist, program	✓	en	MC, FR	nl	834000	-	✓
FlowVerse (2025a)	MR	**P, S**	9-12	website	✓	en, zh	MC, FR	nl	-	2000	✓
Omni-Math (2025a)	MR	**P, S, D**	-	exam	✗	en	NR, FR	nl	-	4428	✓
MathConstruct (2025)	MR	**p**	10-16	exam	✗	en	FR	-	-	126	✓
VCBench (2025j)	MR	**P, S**	1-6	textbook	✓	en	MC	-	-	1720	✓
OlymMATH (2025a)	MR	**P, S, A**	-	textbook, exam	✗	en, zh	NR	-	-	200	✓
RoR-Bench (2025)	MR	**P, S**	1-6	Internet	✓	zh	FR	nl	-	215	✓
PolyMath (2025g)	MR	**P, S, A**	1-16	exist, Internet	✗	18 lang.	NR	-	-	9000	✓
MaTT (2025)	MR	**P, S, A**	-	reference book	✗	en	MC	-	-	1958	✓
CapaBench (2025)	MR	**P, S**	9-12	exist	✗	en	NR	nl	-	1545	✓
MATH-Perturb (2025a)	MR	**P**	9-12	exist	✗	en	NR	-	-	558	✓
M500 (2025)	MR	**P**	-	exam, exist	✗	en	NR, FR	nl	500	-	✓
KPMATH-M (2025e)	MR	**P, S**	9-12	exist	✗	en	NR	nl	252000	-	✗

Table 2: A summarization of geometry problem solving datasets for composite tasks. Task: MR: mathematical reasoning. Type: **P**: plane geometry, **S**: solid geometry, **A**: analytic geometry, **D**: differential geometry. Question: MC: multiple-choice, NR: numerical response, FR: free-response, FB: fill-in-the-blank, YN: yes-or-no, SA: short-answer, CQ: classification question. Rationale: nl: natural language.

Datasets	Task	Type	Grade	Source	Image	Language	Question	Rationale	Trainval Size	Test Size	Opensource
Other Geometry Tasks											
GMBL (2021)	TD	**P**	-	exam	✗	en	GD	-	-	39	✓
LeanEuclid (2024)	AF	**P**	-	exist, textbook	✓	en	FR	-	140	33	✓
Euclidea (2024)	CP	**P**	-	website	✗	en	FR	nl	-	98	✗
PyEuclidea (2024)	CP	**P**	-	website	✗	program	FR	-	-	98	✓
MagicGeoBench (2025a)	TD	**P**	6-12	exam	✗	en	GD	-	-	220	✗
GeoX-pretrain (2025)	DG	**P**	-	web, textbook	✓	-	GD	-	127912	-	✓

Table 3: A summarization of datasets for other geometry tasks. Task: TD: geometric text-to-diagram; CP: geometric construction problem; DG: geometric diagram generation; AF: geometric autoformalization. Type: **P**: plane geometry. Question: FR: free-response, GD: geometric diagram. Rationale: nl: natural language.

from 2D single line drawing images (Xue et al., 2010, 2012; Yang et al., 2013), even if the input image is incomplete or inaccurate (Zheng et al., 2015, 2016b,a; Zou et al., 2016).

Geometric Text-to-Diagram. This task requires the system to be able to generate corresponding geometric diagrams from the natural language description of the geometry problem. This ability will significantly enhance the solution system's understanding, enabling it to more accurately interpret geometric propositions presented in flexible and diverse forms (Liu et al., 2012). In addition to traditional rule-based methods (Janičić and Narboux, 2021; Krueger et al., 2021; Trinh et al., 2024), some recent studies have begun to use deep learning technology to build related systems (Zhengyu and Xiuqin, 2023; Wang et al., 2025a; Cheng et al., 2025a). MagicGeoBench (Wang et al., 2025a) provides a dataset of 220 plane geometry problems from middle school mathematics exams, designed to evaluate the performance of text-to-diagram geometry generation models.

In addition to the above approaches, various other techniques have been developed for generating geometric diagrams. Some tools, such as GeoGebra[1] and Geometer's Sketchpad (Scher, 1999), support interactive constructions using virtual ruler and compass operations to generate geometric diagrams. Additionally, non-interactive methods have also been proposed to automatically derive such constructions (Bertot et al., 2004; Itzhaky et al., 2013). To support more forms of geometric diagram generation, some studies have explored a wider range of methods to construct geometric diagrams. These methods include techniques like algebraic numerical optimization (Gao and Lin, 2004) and constrained numerical optimization (Ye et al., 2020).

This task is also related to GPS. GeoX (Xia et al., 2025) builds a pre-trained dataset containing more than 120,000 plane geometry images and tunes the visual encoder-decoder architecture using the mask auto-encoding scheme to obtain a visual encoder that fully understands geometric diagrams. Additionally, some GPS work uses related methods to perform data enhancement on unimodal geometry problems and generate corresponding diagrams to obtain multimodal data (Cai et al., 2024b; Zhao et al., 2024; Xiao et al., 2024b).

B.2 Geometric Construction Problem

Geometric construction problems, similar to problems in GPS, are also part of educational exams. Such tasks aim to use traditional ruler and compass construction methods to find an effective way to construct the desired figure.

In recent years, some studies have tried to use deep learning systems to solve geometric construction problems. In the online geometric construction game Euclidea[2], Macke et al. (2021); Wong et al. (2023) uses Mask R-CNN (He et al., 2017) to solve difficult geometric construction problems using a purely image-based method. Additionally, Mouselinos et al. (2024) converts the Euclidea problem into a Python format and solves it using a multi-agent framework based on LLMs. This provides us with new ideas and inspires us to further explore the application potential of deep learning systems in cognitive fields such as planning and auxiliary line addition.

B.3 Geometric Figure Retrieval

Before the widespread application of deep learning methods, the search problem for plane geometry figures had always been an important topic in the field of scientific research (Liu et al., 2014a,b; Gan et al., 2016; Chen et al., 2016; Qu et al., 2016; Liu et al., 2016). With the advancement of computer

[1] https://www.geogebra.org

[2] https://www.euclidea.xyz

technology, plane geometry retrieval may no longer be challenging in the era of deep learning. However, retrieving more complex solid geometry and irregular geometric figures may still be a direction worth studying.

B.4 Geometric Autoformalization

Autoformalization is a subtask of theorem proving (Li et al., 2024f). Autoformalization is a subtask of theorem proving (Li et al., 2024f). A few studies focus on automatically converting informal geometry problems and proofs into formal theorems and proofs verifiable by machines. LeanEuclid (Murphy et al., 2024) is a 173-problem geometric autoformalization dataset designed to test whether AI can understand mathematical problems and solutions written by humans and convert them into formal theorems and proofs.

C Encoder-Decoder Architecture for Geometry Problem Solving

In this section, we further elaborate on the deep learning components of the encoder-decoder architecture used for GPS. Table 4 provides a detailed summary of these components.

C.1 Text Encoder

Besides rule-based methods (Lu et al., 2021), early research works typically use Recurrent Neural Networks (RNNs) (Elman, 1990) to parse (Joshi et al., 2018; Gonzalez et al., 2021) or encode (Tsai et al., 2021; Chen et al., 2021) geometry problem texts. Common models include LSTM, GRU, and their bidirectional variants, BiLSTM and BiGRU. Some works employ Transformer (Vaswani et al., 2017) to encode text (Zhang et al., 2023a; Ma et al., 2024a; Zhang et al., 2024f). Additionally, some research works use pre-trained language models for text encoding (Jian et al., 2023b; Huang et al., 2022; Zhu et al., 2025), such as BERT (Devlin et al., 2019) and T5 (Raffel et al., 2020). Moreover, the dual encoder structure of RoBERTa (Liu et al., 2019) plus BiLSTM also shows good results (Cao and Xiao, 2022; Ning et al., 2023; Xiao et al., 2024a; Zhang et al., 2024a).

C.2 Diagram Encoder

Early studies primarily used CNNs to encode geometric diagrams (Zhang et al., 2023a, 2024f; Zhang and Moshfeghi, 2024), with network architectures including RetinaNet (Lin et al., 2017) and its DenseNet (Huang et al., 2017) variants (Lu et al., 2021; Guo and Jian, 2022; Jian et al., 2023a; Huang et al., 2022; Ma et al., 2024a), ResNet (He et al., 2016) and its ConvNeXt (Liu et al., 2022) variants (Chen et al., 2021; Cao and Xiao, 2022; Zhang et al., 2024a,d), and Fast R-CNN (Girshick, 2015; Jian et al., 2023b). Recently, studies have widely adopted pre-trained diagram encoders, such as ViT (Dosovitskiy et al.), ViTMAE (He et al., 2022), CLIP-ViT (Radford et al., 2021), SigLIP (Zhai et al., 2023), and Swin-Transformer (Liu et al., 2021), primarily for building MLLMs. Furthermore, Iordan (2022), Zhang et al. (2022b), and Zhu et al. (2025) use LSTM, GNN, and BLIP (Li et al., 2022) to parse geometric diagrams, respectively, while UniMath (Liang et al., 2023) encodes diagrams through VQVAE (Van Den Oord et al., 2017).

Some other studies use a CNN-Transformer hybrid architecture to integrate the functions of a text encoder and a diagram encoder into a multimodal encoder (Li et al., 2024h; Lin et al., 2025).

C.3 Multimodal Fusion Module

Drawing inspiration from Yu et al. (2019), many studies introduce a co-attention module to comprehensively fuse and align text and image representations (Chen et al., 2021; Cao and Xiao, 2022; Ning et al., 2023; Pan et al., 2023; Ma et al., 2024a). Many MLLMs also incorporate multimodal fusion modules to enhance their multimodal understanding capabilities. For example, LLaVA-v1.5 (Liu et al., 2024a) and MAmmoTH-VL (Guo et al., 2024a) both use a two-layer MLP visual-language connector (Shi et al., 2024; Li et al., 2024g; Xu et al., 2024b; Sharma et al., 2025; Ning et al., 2025; Jia et al., 2025); GLM-4V (GLM et al., 2024) and Qwen2.5-VL (Qwen et al., 2025) use MLP to map image representations to text space (Yang et al., 2024b; Pan et al., 2025; Peng et al., 2025); and InternVL2 (Chen et al., 2024d) uses the QLLaMA architecture (Deng et al., 2024; Xu et al., 2025a). Additionally, some studies consider this module and the subsequent decoder as an overall encoder-decoder structure (Jian et al., 2023b; Zhang et al., 2023a; Li et al., 2024h; Lin et al., 2025; Liang et al., 2023; Zhang and Moshfeghi, 2024), employing self-attention units, BiGRU, and T5-Encoder.

C.4 Decoder

Many studies utilize LSTM (Chen et al., 2021; Cao and Xiao, 2022; Ning et al., 2023; Pan et al., 2023; Xiao et al., 2024a; Zhang et al., 2024a;

Ma et al., 2024a) or GRU (Tsai et al., 2021; Jian et al., 2023b; Zhang et al., 2023a; Li et al., 2024h; Zhang et al., 2024d,f) as decoders in deep learning systems, which may also integrate attention mechanisms. Other studies employ pre-trained language models as decoders. For example, Liang et al. (2023) and (Zhang and Moshfeghi, 2024) use the T5-Decoder, (Pan et al., 2024) chooses BERT, Peng et al. (2024) uses DeepSeekMath-RL (Shao et al., 2024), and Zhuang et al. (2025); Shengyuan and Xiuqin (2024); Zhang et al. (2024c) use the Qwenseries model (Bai et al., 2023) as the decoder. In addition, Zhang et al. (2025c) uses MAmmoTH2 (Yue et al., 2024b), Zhang et al. (2025e) chooses Yi-1.5 (Young et al., 2024), and Cho et al. (2025a) uses Llama 3 (Grattafiori et al., 2024).

C.5 Knowledge Module

Knowledge Extractor and Integrator. Some studies construct geometric knowledge frameworks using knowledge graphs. Fu et al. (2019) and Zhou et al. (2022) use BiLSTM to extract geometric relationships, while Tsai et al. (2021) embeds knowledge graphs into vector space using Graph Convolutional Network (GCN) (Kipf and Welling, 2017). Xu et al. (2024b) and Sharma et al. (2025) use CLIP and VISTA (Zhou et al., 2024a) models to encode geometric problems for retrieving similar problems. Additionally, Xiao et al. (2024a) builds a complete knowledge system through LSTM.

Theorem Predictor. The theorem predictor is used to predict the geometric theorems needed for the current solution step to derive a formal solution path. Guo and Jian (2022); Jian et al. (2023a) encodes the structural information of the formal language through GCN and uses a BiLSTM-GRU based Sequence-to-Sequence (Seq2Seq) architecture (Sutskever et al., 2014) for theorem prediction. In addition, many studies use a Transformer-based Seq2Seq architecture for prediction (Lu et al., 2021; Wu et al., 2024a; Zhang et al., 2024h), and some introduce the T5 model (Yang et al., 2023; He et al., 2024b; Shengyuan and Xiuqin, 2024). Furthermore, Zou et al. (2024) leverages DistilBERT (Sanh, 2019) to guide the training of theorem predictors.

Answer Verifier. Ensuring the correctness of the solution logic is one of the key steps in solving geometry problems. In addition to the traditional rule-based verification method (Zhang et al., 2024f), Pan et al. (2025) introduces a pre-trained LLM (Qwen et al., 2025) to verify the solution steps.

Paper	Task	Network	Text Encoder	Diagram Encoder	Fusion Module	Decoder	Knowledge Module
Fundamental Tasks							
RSP (2018)	SP	BiLSTM	BiLSTM	-	-	-	-
2StepMemory (2020)	SP	attention	attention	-	-	-	-
GIRTOOLS (2020)	ER	VGG16	-	-	-	-	-
Arsenal (2021)	SP	Seq2Seq	RNN	-	-	-†	-
PGDPNet (2022b)	DP	FPN+GNN†	-	FPN+GNN†	-	-	-
UV-S2 (2022)	RE	-	BERT	RetinaNet	-	-	-
BiLSTM-CRF (2022)	RE	BiLSTM	-	-	-	-	-
Stacked LSTM (2022)	DP	LSTM	-	LSTM	-	-	-
2DGeoShapeNet (2024b)	ER	CNN	-	-	-	-	-
Euclid (2024c)	SR	MLLM	-	ConvNeXt	MLP	Qwen-2.5	-
FGeo-Parser (2025)	DP, SP	-	T5	BLIP	-	-	-
Core Tasks - Encoder-Decoder Architecture							
Inter-GPS (2021)	NC	-	-	RetinaNet	-	-	Transformer
NGS (2021)	NC	-	LSTM	ResNet101	co-attention	LSTM†	-
S2G (2021)	NC	-	BiGRU	-	-	GRU†	GCN
GCN-FL (2022)	NC	-	-	DenseNet121+FPN	-	-	GCN+BiLSTM-GRU
DPE-NGS (2022)	NC	-	Bi-LSTM+RoBERTa	ResNet101	co-attention	LSTM†	-
Geoformer (2022)	TP, NC	MLLM	-		VL-T5		-
MCL (2023b)	NC	-	BERT	Faster R-CNN	attention	attention+GRU	-
PGPSNet (2023a)	NC	-	Transformer	CNN	BiGRU	GRU	-
UniMath (2023)	TP, NC	MLLM	-	VQ-VAE	T5		-
RetinaNet+GCN (2023a)	NC	-	-	DenseNet121+FPN	-	-	GCN+BiLSTM-GRU
SCA-GPS (2023)	NC	-	Bi-LSTM+RoBERTa	ViT	co-attention	LSTM†	-
TD-Parsing (2023)	NC	-	-	DenseNet121	co-attention	LSTM†	-
SUFFI-GPSC (2023)	NC	-	-	-	-	-	T5
LANS (2024h)	NC	-	ResNet10+Transformer		BiGRU†	GRU	-
GAPS (2024d)	TP, NC	-	-	ResNet	VL-T5	GRU	-
E-GPS (2024a)	NC	-	-	PGDPNet	-	-	Transformer
FGeo-TP (2024b)	NC	-	-	-	-	-	Transformer
FGeo-DRL (2024)	NC	-	-	-	-	-	DistilBERT
FGeo-HyperGNet (2024h)	NC	-	-	-	-	-	Transformer
GOLD (2024)	TP, NC	MLLM	-	FPN+MobileNetV2+CNN	T5		-
DualGeoSolver (2024a)	NC	-	Bi-LSTM+RoBERTa	ViTMAE	co-attention	LSTM†	LSTM
Math-LLaVA (2024)	NC	MLLM	-		LLaVA-1.5		-
PGPSNet-v2 (2024f)	NC	-	Transformer+BiGRU	CNN	-	GRU	-
EAGLE (2024g)	NC	MLLM	-		LLaVA-1.5		-
MultiMath (2024)	NC	MLLM	-	CLIP-ViT	MLP	DeepSeekMath-RL	-
MathGLM-Vision (2024b)	NC	MLLM	-		GLM-4V		-
ATB-NGS (2024a)	NC		RoBERTa+BiLSTM	Real-ESRGAN+ResNet101	co-attention	LSTM†	-
Geo-Qwen (2024)	NC	MLLM	-	PGDPNet	-	Qwen2.5	T5
Geo-LLaVA (2024b)	NC	MLLM	-		LLaVA-1.5		CLIP
GNS-DTIF (2024a)	NC	-	Transformer	DenseNet121+GCN	GRU+co-attention	LSTM	-
MATHS (2024)	TP, NC	-	-	Swin-Transformer	-	BERT	-
R-CoT (2024)	NC	MLLM	-		InternVL2		-
SANS (2025)	NC	-	CNN+Transformer†		BiGRU†	GRU	-
G-LLaVA (2025b)	NC	MLLM	-		LLaVA-1.5		-
MAVIS (2025c)	NC	MLLM	-	CLIP-Math	MLP	MAmmoTH2	-
GeoX (2025)	TP, NC	MLLM	-	Geo-ViT	GS-Former	Geo-LLM	-
DFE-GPS (2025e)	NC	MLLM	-	SigLIP	MLP	Yi-1.5	-
GeoDANO (2025a)	NC	MLLM	-	GeoCLIP	MLP	LLama-3	-
Math-PUMA (2025)	NC	MLLM	-	SigLIP	MLP	Qwen2	-
GeoCoder (2025)	NC	MLLM	-		LLaVA-1.5		VISTA
MAmmoTH-VL2 (2025)	NC	MLLM	-		MAmmoTH-VL		-
GNS (2025)	NC	MLLM	-		LLaVA-1.5		-
GeoGen (2025)	NC	MLLM	-		Qwen2.5-VL		Qwen2.5
RedStar-Geo (2025a)	NC	MLLM	-		InternVL2		-
SVE-Math (2025d)	NC	MLLM	-	GeoGLIP	MLP	Qwen2.5Math	-
MGT-Geo (2025)	NC	MLLM	-		Qwen2.5-VL		-
Core Tasks - Other Architecture							
GAN+CfER (2023)	NC	cGAN	-	-	-	-	-
GeoDRL (2023)	NC	GNN	-	-	-	-	-
HGR (2024)	NC	GNN	-	-	-	-	-

Table 4: A summarization of deep learning architectures for geometry problem solving system. Task: DP: geometric diagram parsing, SP: semantic parsing for geometry problem texts, ER: geometric element recognition, DP: geometric diagram parsing, SR: geometric structure recognition, RE: geometric relation extraction, TP: geometry theorem proving, NC: geometric numerical calculation. † indicates the presence of the attention mechanism.