

# Towards Geometry Problem Solving in the Large Model Era: A Survey

Yurui Zhao<sup>1</sup> Xiang Wang<sup>1</sup> Jiahong Liu<sup>2</sup> Irwin King<sup>2</sup> Zhitao Huang<sup>1</sup>

## Abstract

Geometry problem solving (GPS) represents a critical frontier in artificial intelligence, with profound applications in education, computer-aided design, and computational graphics. Despite its significance, automating GPS remains challenging due to the dual demands of spatial understanding and rigorous logical reasoning. Recent advances in large models have enabled notable breakthroughs, particularly for SAT-level problems, yet the field remains fragmented across methodologies, benchmarks, and evaluation frameworks. This survey systematically synthesizes GPS advancements through three core dimensions: (1) benchmark construction, (2) textual and diagrammatic parsing, and (3) reasoning paradigms. We further propose a unified analytical paradigm, assess current limitations, and identify emerging opportunities to guide future research toward human-level geometric reasoning, including automated benchmark generation and interpretable neuro-symbolic integration.

## 1. Introduction

Geometry problem solving (GPS) has long posed a persistent challenge in mathematical reasoning and artificial intelligence research (Bobrow et al., 1964; Chou et al., 1996). Successfully automating GPS demands three core capabilities: (1) parsing geometric information from diagrams and textual descriptions to extract spatial relationships (Seo et al., 2015; Zhang et al., 2022), (2) constructing logical reasoning chains to deduce stepwise solutions (Itzhaky et al., 2013; Chen et al., 2021; 2022), and (3) performing numerical calculations to derive precise answers (Lu et al., 2021; Chen et al., 2022). It is worth noting that merging logical reason-

ing and numerical calculation in GPS is common because geometric concepts inherently link to numbers, enabling iterative verification, unified symbolic-numerical representation, and supporting automated geometric modeling.

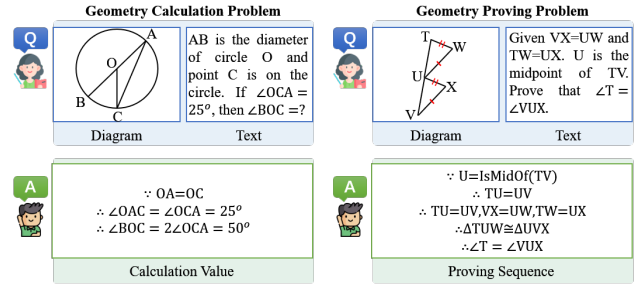


Figure 1. Schematic plot of GPS, which contains geometry calculation problem and geometry proving problem.

Early research in GPS primarily focused on symbolic solvers (Seo et al., 2015; Sachan & Xing, 2017), which relied on computationally intensive search algorithms and basic deductive reasoning. Though these systems achieved progress in constrained settings, their dependence on rigid rule-based frameworks limited applicability to elementary problems (Grade 6–10 level). Recent continuous breakthroughs in artificial intelligence technologies, especially with the coming of the large model era (Liu et al., 2023; Bai et al., 2023; Achiam et al., 2023; Grattafiori et al., 2024; Liu et al., 2025), propel GPS into a new wave of advancement. Advances in GPS methodologies are now bridging these historical limitations, positioning the technology as a foundational tool for emerging interdisciplinary applications, including automated theorem proving (Loveland, 2016), computer-aided design (CAD) (Bi & Wang, 2020), and AI-driven educational systems (Lin et al., 2023).

However, the inherent complexity of GPS introduces critical challenges that impede system performance across three key areas: benchmark construction, information parsing, and logical reasoning. First, creating high-quality, fine-grained, and large-scale benchmarks demands extensive expert annotation (Pan et al., 2025; Deng et al., 2024), limiting dataset scalability. Second, parsing geometric primitives requires accurately interpreting complex layouts and relationships between visual elements (Zhang et al., 2022; 2025). Third,

<sup>1</sup>College of Electronic Science and Technology, National University of Defense Technology, Changsha, China <sup>2</sup>Department of Computer Science and Engineering, Chinese University of Hong Kong, Hong Kong, China. Correspondence to: Xiang Wang <xwang@nuds.edu.cn>.

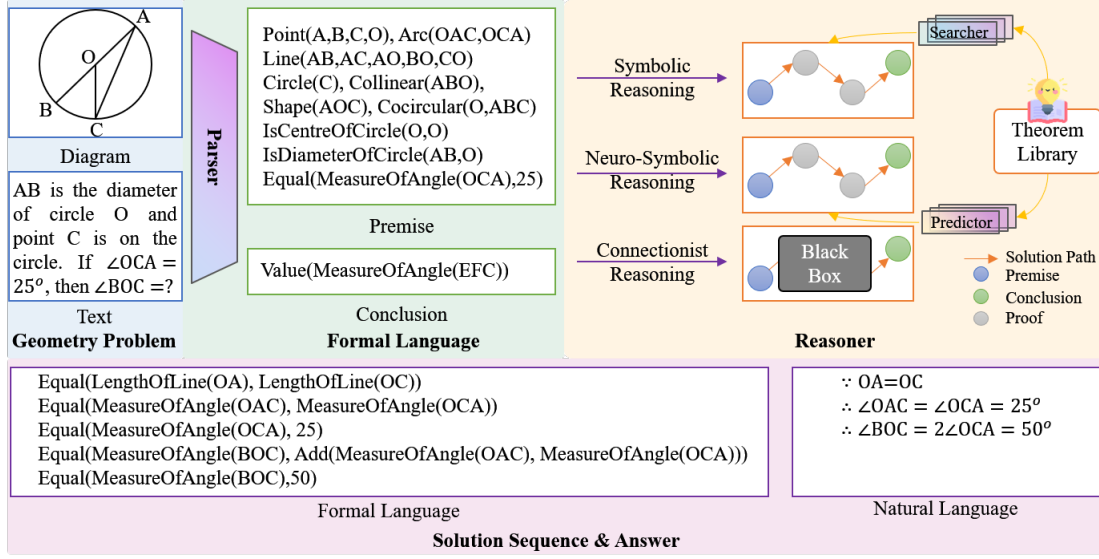


Figure 2. Paradigm of geometry problem solving.

logical reasoning frameworks must balance interpretability with reasoning efficiency (Lu et al., 2021; Ning et al., 2025), especially in zero-shot cases.

The current GPS research landscape, marked by diverse methodologies, fragmented technical approaches, and heterogeneous evaluation benchmarks, has driven innovation but also created a disconnected knowledge base. To address this, we systematically collect, analyze, and synthesize the existing literatures, offering the GPS community a unified perspective and actionable roadmap. The main contributions of this survey are summarized as follows: **(1) Comprehensive Survey:** To the best of our knowledge, this represents the first systematic survey dedicated specifically to GPS, providing the research community with essential foundational knowledge and analytical insights. **(2) Structured Taxonomy:** We propose a three-layer taxonomy for current literature addressing core technical challenges, i.e., benchmark construction, parsing, and reasoning. **(3) Unified Reasoning Paradigm:** We develop a systematic paradigm integrating symbolic reasoning, connectionist reasoning, and neuro-symbolic reasoning into a unified paradigm for GPS. **(4) Summary of Benchmarks:** We analyze benchmark design methodologies and highlight automated construction as a pivotal direction for scalable, high-quality dataset synthesis.

## 2. Problem Description and Paradigm

### 2.1. Problem Statement

GPS encompasses the computational resolution of geometric problems, which fundamentally divide into two categories (as illustrated in Fig. 1): calculation problems de-

termine quantitative measurements like lengths, areas, and angles (Seo et al., 2015; Chen et al., 2021), while proof problems establish geometric truths through deductive arguments (Trinh et al., 2024; Chervonyi et al., 2025). The left panel of Fig. 1 exemplifies a calculation problem: applying the central angle theorem (where an inscribed angle subtending an arc is half its central angle), we derive  $\angle BOC = 2\angle OCA$  to compute  $\angle BOC = 50^\circ$ . The right panel demonstrates a proof problem: establishing triangle congruence via the Side-Side-Side (SSS) criterion to deduce equality of corresponding angles.

Though distinct in objectives, these problem types exhibit deep interconnections: calculation techniques often rely on proof-derived theorems, while computational results provide intuition guiding formal proofs (Chen et al., 2022; Trinh et al., 2024). Given this inherent interdependence, modern computational frameworks increasingly adopt integrated approaches that unify both problem types within cohesive GPS systems. Formally, we represent the GPS task as

$$(\mathcal{S}, \mathcal{A}) = GPS(\mathcal{X} \oplus \mathcal{Y}, \mathcal{Q}, \mathcal{T}), \quad (1)$$

where  $\mathcal{X}$  and  $\mathcal{Y}$  denote problem statements in diagrams and text, respectively.  $\mathcal{T}$  represents the external knowledge library.  $\mathcal{S}$  and  $\mathcal{A}$  are respectively the solution sequence and the answer for the question  $\mathcal{Q}$ .

### 2.2. GPS Paradigm

We propose a systematic paradigm for GPS that encompasses two fundamental stages: parsing and reasoning, as illustrated in Fig. 2. This paradigm can be formally repre-

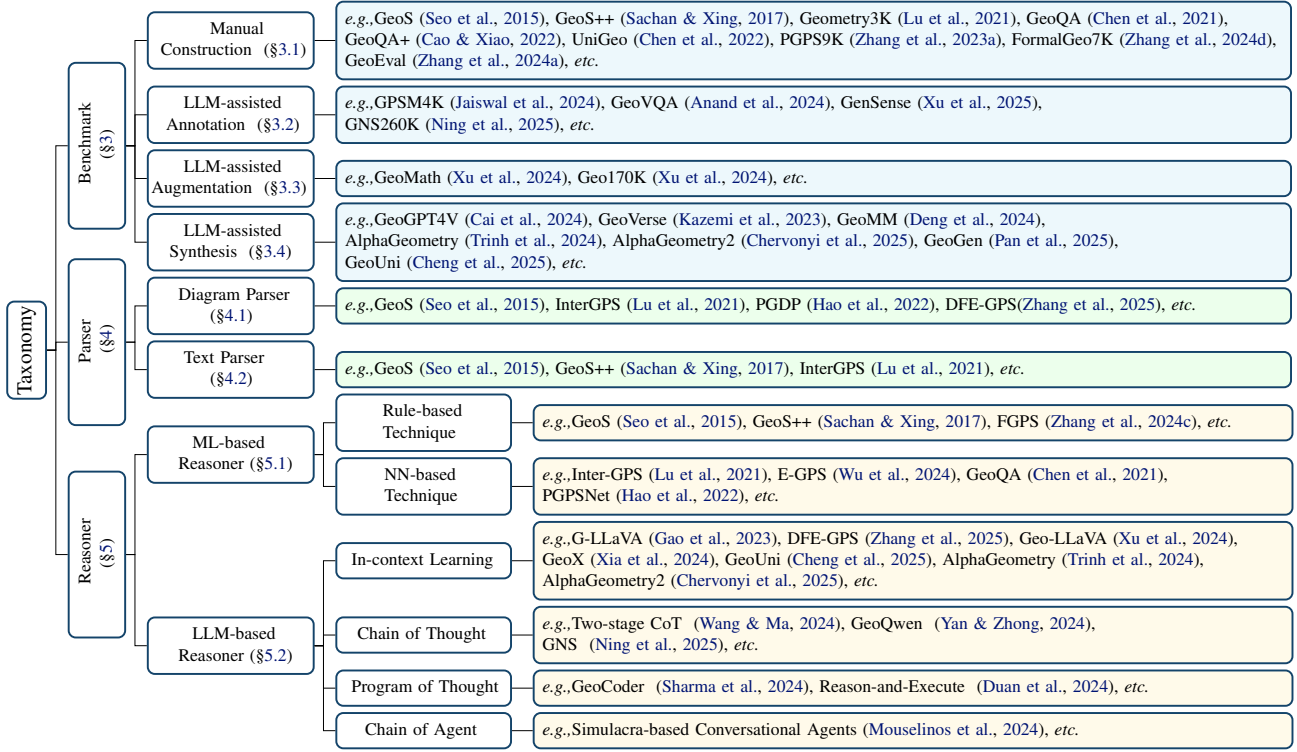


Figure 3. Taxonomy of Geometry Problem Solving.

sented as:

$$(\mathcal{S}, \mathcal{A}) = g(f(\mathcal{X} \oplus \mathcal{Y}, \mathcal{Q}), \mathcal{T}), \quad (2)$$

where  $f$  and  $g$  denote the parsing and reasoning function respectively.

**Parsing Stage:** Given a geometry problem  $(\mathcal{X} \oplus \mathcal{Y}, \mathcal{Q})$ , the parsing component is to design a function  $f$  that transforms the multimodal input into formal languages:  $(\tilde{\mathcal{X}} \oplus \tilde{\mathcal{Y}}, \tilde{\mathcal{Q}}) = f(\mathcal{X} \oplus \mathcal{Y}, \mathcal{Q})$ , where  $\tilde{\mathcal{X}}$ ,  $\tilde{\mathcal{Y}}$ , and  $\tilde{\mathcal{Q}}$  are in the formal language space  $\Omega$ .

**Reasoning Stage:** The reasoning component  $g$  constructs a valid logical chain  $\mathcal{S}$  and calculates accurate values  $\mathcal{A}$  by leveraging a comprehensive theorem database  $\mathcal{T}$ . Our framework supports three distinct reasoning paradigms: (1) Symbolic reasoning employs search-based policies to identify and apply the most appropriate theorems from the knowledge base through systematic exploration. (2) Neuro-symbolic reasoning integrates neural networks as theorem predictors while maintaining the logical rigor of formal deductive reasoning. (3) Connectionist reasoning constructs end-to-end neural architectures that approximate the entire reasoning process as a black-box mapping function. The first two approaches maintain strict adherence to formal logical deduction using theorems from the knowledge repository, while the connectionist approach relies on learned

representations to approximate the reasoning process.

### 2.3. Proposed Taxonomy

A rigorous approach to automated GPS must address three core challenges: **(1) Benchmark Construction:** *What constitutes a well-defined geometric problem?* **(2) Parser:** *How can problems be formally interpreted?* **(3) Reasoner:** *What systematic methods yield correct solutions?* Our paradigm decomposes these into modular components (as illustrated in Fig. 3), with subsequent sections detailing their implementation: Section 3 introduces the Benchmark, which answers (i) by formalizing problem scope and diverse datasets. Section 4 presents the parser, addressing (ii) via examining parsing methodologies including formal language space, diagram parser and text parser. Section 5 discusses the reasoner, resolving (iii) through contrasting two dominant reasoning paradigms: ML frameworks and large language model (LLM)-based systems.

## 3. Benchmark Construction

This section reviews the evolution of GPS benchmarks, as depicted in Fig. 4. Since 2015, benchmark development has focused on enhancing precision, expanding problem variety, and increasing overall scale. The advent of LLMs significantly accelerated this trend, primarily due

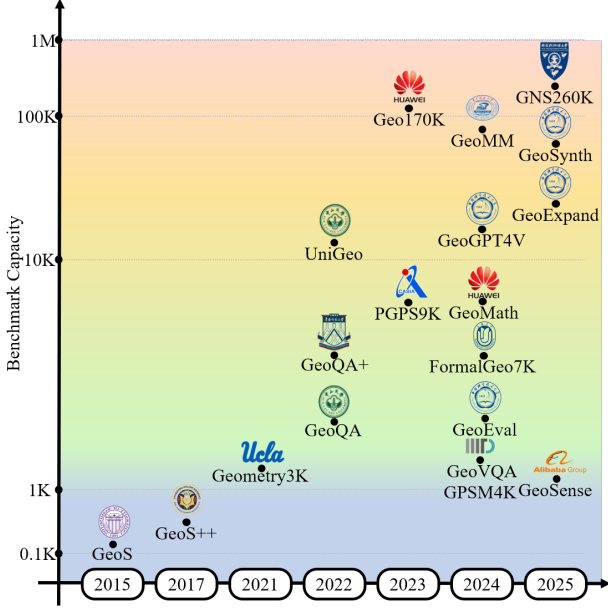


Figure 4. Research trend of benchmarks.

to two factors: LLMs underscored the importance of data scale for superior performance, and they provided powerful tools for automated problem generation, transforming benchmark construction from a manual to a scalable process. Existing benchmarks are categorized by their construction technique: manual construction, LLM-assisted annotation, LLM-assisted augmentation, and LLM-assisted synthesis. Further details on these benchmarks are provided in Appendix 8.1.

### 3.1. Manual Construction

GeoS built the first dataset for GPS (Seo et al., 2015), covering 186 SAT plane geometry questions where every question has a textual description in English accompanied by a diagram and multiple choices. To further enhance the dataset scale, GeoS++ collected a total of 1,406 SAT style questions with more complex concepts (Sachan & Xing, 2017) across grades 6-10, while Geometry3K collected 3,002 problems from textbooks across grades 6-12 (Lu et al., 2021). Considering solution processes are also important in the GPS, PGPS9K added solution sequences as labels (Zhang et al., 2023a) and constructed a dataset of 9,022. To construct benchmarks with Chinese middle school exams, GeoQA (Chen et al., 2021) and advanced GeoQA+ (Cao & Xiao, 2022) were proposed. The annotated operations required to solve the problem in GeoQA are limited to a maximum of 4 steps, but that in GeoQA+ is with up to 8. To unify the proving problem with the calculation problem to construct the UniGeo benchmark (Chen et al., 2022), UniGeo collects 9,543 proving problems and heritage 4,998 calculation problems from GeoQA. Considering imperfect annotation

limits the GPS to be SAT-level, Zhang *et al.* proposed a formal geometry theory and extended the annotation to the IMO-level (Zhang et al., 2024d; 2023b). Thus, two datasets are constructed, i.e., FormalGeo7k and FormalGeo-IMO. The former contains 6,981 geometry problems (SAT-level to IMO-level), while the latter includes 18 IMO-level challenging geometry problems. GeoEval integrates seven public datasets and newly collected geometry problems, covering areas such as plane geometry, solid geometry, and analytic geometry, to provide a benchmark for the GPS field (Zhang et al., 2024a). It consists of four subsets: GeoEval-2000, which includes 2,000 problems; GeoEval-backward, containing 750 problems designed for backward reasoning; GeoEval-aug, featuring 2,000 augmented problems with varied rephrasing; and GeoEval-hard, comprising 300 problems that focus on solid and analytic geometry.

### 3.2. LLM-assisted Annotation

LLMs exhibit dual capabilities in understanding multimodal geometric content and generating structured solution sequences, enabling their application in benchmark annotation. Researchers leverage LLMs to generate step-by-step solution labels, considering that detailed reasoning descriptions significantly enhance human comprehension. GPSM4K (Jaiswal et al., 2024) and GeoVQA (Anand et al., 2024) constitute complementary benchmarks sourced from Indian mathematics textbooks (Grades 6–12), collectively containing approximately 4,400 calculation and proof problems. Both leverage Gemini Pro Vision to decompose textbook solutions into fundamental reasoning components, enhancing computational interpretability. GPSM4K employs Gemini Vision Pro for automated diagram captioning, enriching visual accessibility, while GeoVQA utilizes Gemini Vision Pro + Gemini Pro to dissect problems systematically, covering identifying key concepts, performing computations, and structuring solutions for multimodal question answering. GeoSense (Xu et al., 2025) employs LLMs to generate principle-level annotations for 1,789 geometric problems sourced from existing benchmarks and educational platforms. These annotations codify essential geometric knowledge into three classes, i.e., definitions, theorems, and formulas. To ensure reliability, Xu *et al.* developed a semi-automated pipeline where GPT-4o and human experts collaboratively identify necessary principles, contextualize their application within geometric diagrams, and verify mathematical expressions for computational accuracy. GNS260K addresses a critical gap in plane geometry datasets: the absence of natural language solution descriptions (Ning et al., 2025). Leveraging GPT-4’s advanced reasoning capabilities, it generates human-interpretable solving traces for 9,426 unique diagrams, scaled through augmentation to 18,852 knowledge prediction samples, 86,732 symbolic parsing samples, and 154,433 problem reasoning samples. The an-



notation methodology employs a two-stage pipeline: (1) parsed problem clauses and symbolic solutions serve as structured input. (2) GPT-4 converts formal solutions into pedagogical narratives using description-focused prompts. This comprehensive annotation enables chain-of-thought reasoning in GPS systems, significantly enhancing both interpretability and educational applicability.

### 3.3. LLM-assisted Augmentation

LLMs demonstrate significant generative capabilities for enhancing dataset quality. By automatically generating diverse variants and refining questions from existing data, LLMs can substantially improve a dataset’s diversity, complexity, and scale. This approach has become a widely adopted strategy for data augmentation. GeoMath exemplifies this technique by utilizing GPT-3.5 for data augmentation through text rewriting and image caption generation. Specifically, it rephrases original problem statements in five distinct ways, expanding the sample size sixfold while increasing linguistic diversity. This method generates richer training examples, strengthening model generalization (Xu et al., 2024). Building on this paradigm, the Geo170K dataset employs ChatGPT to augment existing geometric datasets (GeoQA+ and Geometry3K). Key augmentation strategies include (1) equation solving and value scaling, (2) re-formulating conditions as unknowns, and (3) systematic sentence paraphrasing. Geo170K provides both alignment data (linking problems to geometric diagrams) and instruction data (problem-solving steps) derived from GeoQA+ and Geometry3K training sets (Xu et al., 2024). The final dataset comprises approximately 60,000 geometric image-caption pairs and approximately 110,000 question-answer pairs, representing a significant scale advancement in geometric reasoning resources.

### 3.4. LLM-assisted Synthesis

Researchers have increasingly leveraged LLMs for generating geometry problems due to their powerful capabilities. The primary goal is to utilize generative models to create accurate, non-contradictory problems, ensuring that the required theorems are clearly defined and the reasoning steps for problem-solving are controllable. Image-generation oriented methods, exemplified by GeoGPT4V (Cai et al., 2024), employ multi-modal fusion through a three-stage framework encompassing problem simplification, Wolfram code generation, and image scoring, utilizing 4.9K generated samples alongside 19K open-source data to achieve text-image correspondence for intuitive model training. Reverse reasoning methodologies, represented by GeoMVerse (Kazemi et al., 2023) and GeoMM (Deng et al., 2024), adopt backward generation strategies that construct multi-hop reasoning problems through predefined geometric constructs, with GeoMM specifically implementing a Re-

verse Chain-of-Thought framework processing 20 geometric shapes across 87K samples to generate multi-step reasoning question-answer pairs. Symbolic reasoning approaches, notably AlphaGeometry (Trinh et al., 2024; Chervonyi et al., 2025) and GeoGen (Pan et al., 2025), leverage symbolic engines and traceback algorithms to derive mathematical facts from random geometric diagrams, with AlphaGeometry2 scaling to 0.3 billion training samples while GeoGen integrates symbolic reasoning with large language model capabilities, producing 45,526 question-answer pairs through GeoExpand and 129,230 initial diagrams via GeoSynth’s predicate sampling methodology. Knowledge-directed generation, as demonstrated by GeoUni (Cheng et al., 2025), represents the methodological frontier by precisely generating geometric diagrams aligned with specific knowledge points, ensuring tight correspondence between textual problems and visual representations for customized educational applications. These methodological developments reveal a systematic evolution from elementary text-image pairing to sophisticated knowledge-directed generation, characterized by fundamental distinctions in generation strategies (forward versus reverse mechanisms), control mechanisms (difficulty adjustment versus knowledge targeting), data scalability (thousand-level samples to billion-scale datasets), and application domains (general-purpose generation versus educational customization), thereby establishing a comprehensive theoretical framework that advances geometric problem generation through diversified technical paradigms and scalable methodological innovations.

## 4. Parser

In our proposed paradigm, parsers  $f$  are designed to map a geometry problem  $(\mathcal{X} \oplus \mathcal{Y}, \mathcal{Q})$  into a formal language representation  $(\tilde{\mathcal{X}} \oplus \tilde{\mathcal{Y}}, \tilde{\mathcal{Q}})$ , where the output belongs to a formal language space  $\Omega$ . A formal language is a rigorously defined system of symbols and rules, commonly used in computer science and mathematics to describe and analyze structured information. In the context of GPS, formal languages enable the translation of geometric elements into symbolic expressions that support precise logical manipulation. Compared to latent features, formal representations offer several advantages: they provide greater transparency and verifiability, facilitate robust symbolic reasoning, and allow for seamless integration of expert geometric knowledge (Seo et al., 2015; Chen et al., 2021; Trinh et al., 2024; Chervonyi et al., 2025; Zhang et al., 2023b). In contrast, latent features often lack interpretability, struggle with rigorous logical inference, generalize poorly to novel structures, and face difficulties in incorporating domain expertise (Cao & Xiao, 2022; Ning et al., 2025; Zhang et al., 2023a; Xia et al., 2024). Hence, this paper focuses on surveying parsers that utilize formal languages, owing to their inherent advantages.

#### 4.1. Diagram Parser

The development of diagram parsing techniques has shown significant advancement over time. GeoS pioneered this area using a publicly available diagram parser that provided confidence scores for literals and identified visual elements by maximizing pixel coverage, text-visual agreement, and element coherence (Seo et al., 2014). This approach could identify various shapes including lines, circles, and polygons, while serving dual purposes: computing diagram scores and extracting visual literals unavailable from text. Inter-GPS subsequently improved upon this foundation by developing a fully automatic parser without manual intervention (Lu et al., 2021). This system applied Hough Transformation to extract geometric primitives, then employed RetinaNet for detecting diagram symbols and text regions, with Math-Pix handling optical character recognition. Unlike previous methods, Inter-GPS successfully managed special relational symbols such as parallel, perpendicular, and isosceles markings. PGDP further advanced the field through deep learning and graph reasoning (Hao et al., 2022), introducing an end-to-end model (PGDPNet) that utilized modified instance segmentation for primitive extraction and graph neural networks for relation parsing and classification, incorporating both geometric features and prior knowledge. The most recent innovation, DFE-GPS-SigLIP, represents a multimodal approach integrating three key components (Zhang et al., 2025): a Diagram Formalizer, a Projection module, and a large language model. This system processes diagram features through SigLIP (Vision Encoder), aligns diverse inputs within the LLM’s semantic space.

#### 4.2. Text Parser

The development of text parsing techniques for geometry problems has undergone multiple evolutionary stages. GeoS implemented a three-phase parsing pipeline: initially mapping textual terms to corresponding geometric concepts, subsequently identifying relationships between these concepts, and finally processing implications and coordinating conjunctions (Seo et al., 2014). GeoS employed a hypergraph structure to represent literals, where nodes corresponded to geometric concepts (constants, variables, functions, or predicates) and edges captured the relationships between concepts. The system first identified concepts within the text, then learned affinity scores for hypergraph edges, and ultimately completed relationships to satisfy type matching requirements in the formal language. GeoS++ enhanced this approach by adopting a part-based log-linear model that integrated multiple steps into a unified framework while maintaining a similar three-phase process (Sachan & Xing, 2017). However, GeoS++ decomposed its model into representing concepts and representing relationships and then utilized a rule-based method similar to GeoS for relation completion. Inter-GPS diverged from these approaches

by employing template rules to transform problem text into formal language (Lu et al., 2021). This system deliberately avoided sequence-to-sequence learning methods due to the limited scale of geometry datasets and the sensitivity of symbolic reasoning to noise, instead implementing rule-based parsing techniques with regular expressions to achieve more precise parsing results.

### 5. Reasoner

In this section, we introduce the reasoner, the central component in GPS systems. While Section 2.2 details the three reasoning paradigms (symbolic, neuro-symbolic, and connectionist), this section classifies reasoners by their implementation tools: Machine Learning-based (ML-based) and LLM-based approaches. Details are presented in Appendix 8.2

#### 5.1. ML-based Techniques

##### 5.1.1. RULE-BASED TECHNIQUE

Rule-based techniques in GPS, often rooted in data mining, leverage the deductive nature of geometric reasoning. These approaches model geometry problems as logical systems: given values act as initial facts, geometric principles serve as rules, and the derivation of unknown values functions as logical deduction (Seo et al., 2015; Sachan & Xing, 2017; Zhang et al., 2024c;a).

GeoS pioneered this paradigm by translating geometry problems into logical expressions. Its solver utilized submodular optimization with greedy selection and basin-hopping combined with sequential least squares programming to maximize global constraint satisfaction (Seo et al., 2015). Extending this foundation, GeoS++ parsed axiom information from mathematics textbooks and employed log-linear models to score axiom application sequences, effectively learning optimal theorem application strategies (Sachan & Xing, 2017). FGPS further advanced the field by incorporating symmetric problem-solving algorithms: forward search derives new conditions from initial ones until the goal is reached, while backward search decomposes goals into sub-goals until they match known conditions (Zhang et al., 2024c;d). This approach organizes the problem-solving process into a hypertree structure, where conditions are hypernodes and theorems are hyperedges. Experiments show that various search strategies (e.g., breadth-first, depth-first, random, and beam) offer differing performance benefits based on problem characteristics.

##### 5.1.2. NN-BASED TECHNIQUE

Recent research in computational geometry has integrated neural networks (NNs) into problem-solving frameworks through two distinct approaches: as theorem predictors

within neuro-symbolic systems and as end-to-end reasoners in connectionist paradigms.

In the **neuro-symbolic paradigm**, NNs function as theorem predictors to guide symbolic reasoning. Inter-GPS exemplifies this approach by formulating GPS as goal-directed search, incorporating theorem knowledge as conditional rules for step-by-step symbolic reasoning (Lu et al., 2021). The integration of a theorem predictor significantly enhances search efficiency by inferring likely theorem application sequences, addressing the limitations of brute-force enumeration strategies. Wu et al. advanced this paradigm with their Explainable Top-Down Problem Solver (TD-PS), which mimics human expert reasoning by starting from the target and working backward (Wu et al., 2024). This approach employs target decomposition and condition processing mechanisms to ensure explainability while using neural-guided theorem prediction to constrain the search space effectively.

Alternatively, **connectionist reasoning approaches** employ NNs as direct reasoners. GeoQA utilizes an LSTM decoder with attention mechanisms over multimodal information to generate sequential programs guided by embeddings (Chen et al., 2021). PGPSNet further refines this approach by encoding diagram images with CNN and problem text with pre-trained language models (Hao et al., 2022), fusing these modalities through bidirectional GRU encoders. A notable innovation in PGPSNet is its self-limited GRU decoder, which significantly reduces representation and search spaces to accelerate training and inference. This architecture employs simplified clauses describing basic relationships rather than complex multi-order logic forms, achieving performance comparable to more complex tree decoders but with substantially improved computational efficiency.

## 5.2. LLM-based Techniques

### 5.2.1. IN-CONTEXT LEARNING

In-Context Learning (ICL) has emerged as a significant capability of LLMs, enabling them to perform new tasks by conditioning on demonstration examples provided in the input prompt without explicit parameter updates. Recent research has leveraged this capability to address the complex domain of GPS through several innovative approaches.

G-LLaVA pioneered a method that concatenates mapped image features with text embeddings as input to LLMs, demonstrating superior performance over GPT4-V on geometry problems (Gao et al., 2023). Building on this foundation, DFE-GPS introduced a more structured approach by processing three input types, i.e., diagram features extracted by a Diagram Encoder, formal diagram language representations, and natural language inputs (Zhang et al., 2025). Further advancing the field, Geo-LLaVA enhanced

performance through dataset augmentation and Retrieval Augmentation Generation (RAG) (Xu et al., 2024), while GeoX employed a Generator-and-Sampler Transformer to create geometry content-aware queries and eliminate uninformative representations (Xia et al., 2024). GeoX’s formalized pre-training strategy demonstrated particular effectiveness in downstream geometry tasks through its generation of Minimal Sufficient Representations. GeoUni introduced a significant methodological advancement with Geo-Reasoning-Adapter trained via a comprehensive reward function encompassing format, formalization, and accuracy components (Cheng et al., 2025). This approach substantially improved performance on multiple-choice and open-ended geometry questions. Departing from purely connectionist reasoning approaches, AlphaGeometry and AlphaGeometry2 positioned LLMs as auxiliary line predictors within a neuro-symbolic paradigm, using models like Gemini to predict auxiliary line construction before verification through symbolic reasoning (Trinh et al., 2024; Chervonyi et al., 2025).

### 5.2.2. CHAIN OF THOUGHT

Chain-of-Thought (CoT) prompting has emerged as a significant technique for enhancing language models’ reasoning capabilities by generating intermediate steps before producing final answers. This approach is particularly valuable for geometric problem solving, which requires multi-hop mathematical reasoning across both textual and visual information.

Research from GeomVerse demonstrates that while Vision-Language Models (VLMs) perform adequately on simpler problems, their effectiveness significantly diminishes with increased reasoning complexity (Kazemi et al., 2023). However, training these models with CoT prompting substantially improves their geometric reasoning capabilities, especially when finetuned to generate both solutions and reasoning processes.

Several innovative approaches have extended the basic CoT paradigm. Wang et al. proposed a two-step zero-shot methodology that first generates a diagram graph to enhance the model’s comprehension of geometric information, followed by problem-solving using this intermediate representation (Wang & Ma, 2024). Ning et al. introduced GNS, a neural-symbolic framework that parses problems into symbolic clauses for explicit geometric comprehension and precise computation (Ning et al., 2025). Further advancing this field, Yan et al. developed GeoQwen, conceptualizing the problem-solving process as a reasoning tree where each node represents a solution step (Yan & Zhong, 2024). This approach implements backtracking when reaching impasses and employs a theorem predictor to identify relevant theorems, thereby reducing the search space and



improving reasoning efficiency.

### 5.2.3. PROGRAM OF THOUGHT

Program-of-Thought (PoT) prompting is an advanced technique where large language models are guided to generate executable code, representing the reasoning steps required to solve a problem. A persistent challenge with CoT finetuned Visual Language Models (VLMs) is their susceptibility to calculation errors and incorrect formula application, particularly in geometry problems requiring extended reasoning chains.

To address these limitations, Sharma et al. introduced Geo-Coder, proposing modular code-finetuning as an alternative approach (Sharma et al., 2024). This method generates Python code utilizing a predefined geometry library, offering three key advantages: deterministic calculations through code execution, reduced formula errors through predefined mathematical functions, and enhanced interpretability via templated print statements. AlphaEvolve transforms geometric problems into algorithmic challenges aimed at finding the optimal geometric configurations under specific conditions (Novikov et al., 2025). It employs a multi-stage heuristic search approach to explore a vast solution space in search of the best algorithm to solve geometric problems.

Research by Duan et al. suggests that reasoning ability surpasses computational power in importance for geometric problem-solving. Recognizing the complementary strengths of CoT excelling in reasoning while PoT provides computational precision has led to hybrid methodologies (Duan et al.). The Reason-and-Execute (RaE) framework exemplifies this integration, embedding PoT within CoT as an execution mechanism. RaE employs reverse thinking to identify necessary geometric knowledge and logical steps while generating executable code blocks for precise arithmetic operations (Duan et al., 2024).

### 5.2.4. CHAIN OF AGENT

Mouselinos et al. adopted simulacra-based conversational agents to construct Chain of Agent for GPS (Mouselinos et al., 2024). In mathematical reasoning, a Chain of Agents could be employed by assigning one agent to parse and understand the word problem, another to identify the relevant mathematical operations or theorems, a third to perform the symbolic or numerical calculations, and a final agent to synthesize the steps and present the solution. Studies suggest that the resulting cross-domain collaborative dialogue enables more sophisticated geometric problem-solving than traditional single-agent approaches by leveraging complementary cognitive strengths and distributing complex reasoning processes across specialized components.

## 6. Challenges and Future Work

**Enhanced Multimodal Understanding and Logical Reasoning Capabilities:** Future advancements will focus on two key LLM capabilities: (1) enhanced multimodal understanding of dynamic geometry (Goldenberg & Cuoco, 2012), including motion analysis and property evolution tracking, and (2) robust logical reasoning for complex analytic geometry (Fischer, 2006).

**Deep Fusion of Algebraic and Geometric Methods:** Future LLM-powered GPS systems will integrate algebraic and geometric reasoning through three key capabilities: (1) automatic conversion of geometric problems into algebraic formulations, (2) application of symbolic computation tools (e.g., Wu’s method (Wen-Tsun, 1984), Gröbner bases (Becker et al., 1993)), and (3) geometric interpretation of algebraic results. This cross-paradigm approach is expected to facilitate breakthroughs in automated theorem proving, particularly for complex geometric inequalities that current methods struggle with.

**Creative Inductive Capabilities:** Future LLMs will revolutionize geometric research by autonomously generating novel conjectures (Chen et al., 2025; Li et al., 2025) and discovering hidden patterns in complex structures. By combining learned geometric knowledge with symbolic reasoning, these systems will serve as AI collaborators, proposing testable hypotheses, identifying new symmetries, and accelerating mathematical exploration beyond human intuition alone, while maintaining formal rigor.

**Self-Verification Ability:** Causal reasoning modules can act as an internal monitor in the GPS task, flagging steps that violate fundamental spatial or relational expectations before errors propagate (Kiciman et al., 2023; Wang, 2024). Integrating intuitive, causality-based plausibility checks can greatly enhance the robustness and reliability of automated GPS by enabling real-time detection and correction of reasoning errors, similar to human problem-solving.

## 7. Conclusion

This survey offers a comprehensive and systematic overview of GPS, tracing its developmental trajectory, especially within the era of large AI models. We establish a systematic paradigm and propose a structured taxonomy, classifying existing methodologies into three key technical perspectives: benchmark construction, parser, and reasoner, with each component further subdivided. Furthermore, we discuss current limitations and challenges within the field, identifying several promising directions for future research. This work provides valuable insights and a foundational framework to catalyze the advancement of GPS.



## References

- Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- Anand, A., Jaiswal, R., Dharmadhikari, A., Marathe, A., Popat, H., Mital, H., Nair, A. R., Prasad, K., Kumar, S., Verma, A., et al. Geovqa: A comprehensive multimodal geometry dataset for secondary education. In *2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, pp. 102–108. IEEE, 2024.
- Awais, M., Ahmed, T., Aslam, M., Rehman, A., Alamri, F. S., Bahaj, S. A., and Saba, T. Mathvision: An accessible intelligent agent for visually impaired people to understand mathematical equations. *IEEE Access*, 2024.
- Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.
- Becker, T., Weispfenning, V., Becker, T., and Weispfenning, V. *Gröbner bases*. Springer, 1993.
- Bi, Z. and Wang, X. *Computer aided design and manufacturing*. John Wiley & Sons, 2020.
- Bobrow, D. et al. Natural language input for a computer problem solving system. 1964.
- Cai, S., Bao, K., Guo, H., Zhang, J., Song, J., and Zheng, B. Geogpt4v: Towards geometric multi-modal large language models with geometric image generation. *arXiv preprint arXiv:2406.11503*, 2024.
- Cao, J. and Xiao, J. An augmented benchmark dataset for geometric question answering through dual parallel text encoding. In *Proceedings of the 29th international conference on computational linguistics*, pp. 1511–1520, 2022.
- Chen, J., Tang, J., Qin, J., Liang, X., Liu, L., Xing, E. P., and Lin, L. Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning. *arXiv preprint arXiv:2105.14517*, 2021.
- Chen, J., Li, T., Qin, J., Lu, P., Lin, L., Chen, C., and Liang, X. Unigeo: Unifying geometry logical reasoning via reformulating mathematical expression. *arXiv preprint arXiv:2212.02746*, 2022.
- Chen, Q., Qin, L., Liu, J., Peng, D., Guan, J., Wang, P., Hu, M., Zhou, Y., Gao, T., and Che, W. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *arXiv preprint arXiv:2503.09567*, 2025.
- Cheng, J.-K., Zhang, Z., Chen, R., Deng, J., Qin, Z., and Ma, J. Geouni: A unified model for generating geometry diagrams, problems and problem solutions. *arXiv preprint arXiv:2504.10146*, 2025.
- Chervonyi, Y., Trinh, T. H., Olšák, M., Yang, X., Nguyen, H., Menegali, M., Jung, J., Verma, V., Le, Q. V., and Luong, T. Gold-medalist performance in solving olympiad geometry with alphageometry2. *arXiv preprint arXiv:2502.03544*, 2025.
- Chou, S.-C., Gao, X.-S., and Zhang, J.-Z. Automated generation of readable proofs with geometric invariants: Ii. theorem proving with full-angles. *Journal of Automated Reasoning*, 17(3):349–370, 1996.
- Deng, L., Liu, Y., Li, B., Luo, D., Wu, L., Zhang, C., Lyu, P., Zhang, Z., Zhang, G., Ding, E., et al. R-cot: Reverse chain-of-thought problem generation for geometric reasoning in large multimodal models. *arXiv preprint arXiv:2410.17885*, 2024.
- Duan, X., Tan, D., Fang, L., Guan, Q., Zhou, Y., Huang, X., Gong, Z., and Huang, C. Towards geometry problems solving employing gpt-4 vision with few-shot prompting: An empirical study of what matters.
- Duan, X., Tan, D., Fang, L., Zhou, Y., He, C., Chen, Z., Wu, L., Chen, G., Gong, Z., Luo, W., et al. Reason-and-execute prompting: Enhancing multi-modal large language models for solving geometry questions. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 6959–6968, 2024.
- Fischer, G. *Complex analytic geometry*, volume 538. Springer, 2006.
- Gao, J., Pi, R., Zhang, J., Ye, J., Zhong, W., Wang, Y., Hong, L., Han, J., Xu, H., Li, Z., et al. G-llava: Solving geometric problem with multi-modal large language model. *arXiv preprint arXiv:2312.11370*, 2023.
- Goldenberg, E. P. and Cuoco, A. A. What is dynamic geometry? In *Designing learning environments for developing understanding of geometry and space*, pp. 351–367. Routledge, 2012.
- Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- Hao, Y., Zhang, M., Yin, F., and Huang, L.-L. Pgdp5k: A diagram parsing dataset for plane geometry problems. In *2022 26th international conference on pattern recognition (ICPR)*, pp. 1763–1769. IEEE, 2022.

- Itzhaky, S., Gulwani, S., Immerman, N., and Sagiv, M. 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*, pp. 457–472. Springer, 2013.
- Jaiswal, R., Anand, A., and Shah, R. R. Advancing multi-modal llms: A focus on geometry problem solving reasoning and sequential scoring. In *Proceedings of the 6th ACM International Conference on Multimedia in Asia*, pp. 1–7, 2024.
- Kazemi, M., Alvari, H., Anand, A., Wu, J., Chen, X., and Soricut, R. Geomverse: A systematic evaluation of large models for geometric reasoning. *arXiv preprint arXiv:2312.12241*, 2023.
- Kiciman, E., Ness, R., Sharma, A., and Tan, C. Causal reasoning and large language models: Opening a new frontier for causality. *Transactions on Machine Learning Research*, 2023.
- Li, Z.-Z., Zhang, D., Zhang, M.-L., Zhang, J., Liu, Z., Yao, Y., Xu, H., Zheng, J., Wang, P.-J., Chen, X., et al. From system 1 to system 2: A survey of reasoning large language models. *arXiv preprint arXiv:2502.17419*, 2025.
- Lin, C.-C., Huang, A. Y., and Lu, O. H. Artificial intelligence in intelligent tutoring systems toward sustainable education: a systematic review. *Smart Learning Environments*, 10(1):41, 2023.
- Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.
- Liu, J., Qiu, Z., Li, Z., Dai, Q., Zhu, J., Hu, M., Yang, M., and King, I. A survey of personalized large language models: Progress and future directions. *arXiv preprint arXiv:2502.11528*, 2025.
- Loveland, D. W. *Automated theorem proving: A logical basis*. Elsevier, 2016.
- Lu, P., Gong, R., Jiang, S., Qiu, L., Huang, S., Liang, X., and Zhu, S.-C. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. *arXiv preprint arXiv:2105.04165*, 2021.
- Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., Cheng, H., Chang, K.-W., Galley, M., and Gao, J. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.
- Mouselinos, S., Michalewski, H., and Malinowski, M. Beyond lines and circles: Unveiling the geometric reasoning gap in large language models. *arXiv preprint arXiv:2402.03877*, 2024.
- Ning, M., Zhou, Z., Wang, Q., Huang, X., and Huang, K. Gns: Solving plane geometry problems by neural-symbolic reasoning with multi-modal llms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 24957–24965, 2025.
- Novikov, A., Vu, N., Eisenberger, M., Dupont, E., Huang, P.-S., Wagner, A. Z., Shirobokov, S., Kozlovskii, B., Ruiz, F. J. R., Mehrabian, A., Kumar, M. P., See, A., Chaudhuri, S., Holland, G., Davies, A., Nowozin, S., Kohli, P., and Balog, M. Alphaevolve: A coding agent for scientific and algorithmic discovery. *Google DeepMind*, 2025. URL [https://colab.research.google.com/github/google-deepmind/alphaevolve\\_results/blob/master/mathematical\\_results.ipynb](https://colab.research.google.com/github/google-deepmind/alphaevolve_results/blob/master/mathematical_results.ipynb). Accessed: 2025-05-25.
- Pan, Y., Zhang, Z., Hu, P., Ma, J., Du, J., Zhang, J., Liu, Q., Gao, J., and Ma, F. Enhancing the geometric problem-solving ability of multimodal llms via symbolic-neural integration. *arXiv preprint arXiv:2504.12773*, 2025.
- Sachan, M. and Xing, E. 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)*, pp. 251–261, 2017.
- Seo, M., Hajishirzi, H., Farhadi, A., Etzioni, O., and Malcolm, C. Solving geometry problems: Combining text and diagram interpretation. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pp. 1466–1476, 2015.
- Seo, M. J., Hajishirzi, H., Farhadi, A., and Etzioni, O. Diagram understanding in geometry questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28, 2014.
- Sharma, A., Dalmia, A., Kazemi, M., Zouaq, A., and Pal, C. J. Geocoder: Solving geometry problems by generating modular code through vision-language models. *arXiv preprint arXiv:2410.13510*, 2024.
- Tang, J., Zhang, C., Zhu, X., and Liu, M. Tangram: A challenging benchmark for geometric element recognizing. *arXiv preprint arXiv:2408.13854*, 2024.
- Trinh, T. H., Wu, Y., Le, Q. V., He, H., and Luong, T. Solving olympiad geometry without human demonstrations. *Nature*, 625(7995):476–482, 2024.

- Wang, L. and Ma, J. Mutli-step chain-of-thought in geometry problem solving. In *2024 4th International Conference on Electronic Information Engineering and Computer Science (EIECS)*, pp. 1113–1117. IEEE, 2024.
- Wang, Z. Causalbench: A comprehensive benchmark for evaluating causal reasoning capabilities of large language models. In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)*, pp. 143–151, 2024.
- Wen-Tsün, W. On the decision problem and the mechanization of theorem-proving in elementary geometry<sup>1</sup>. *Automated Theorem Proving: After 25 Years: After 25 Years*, 89:213, 1984.
- Wu, W., Zhang, L., Liu, J., Tang, X., Wang, Y., Wang, S., and Wang, Q. 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*, pp. 13828–13837, 2024.
- Xia, R., Li, M., Ye, H., Wu, W., Zhou, H., Yuan, J., Peng, T., Cai, X., Yan, X., Wang, B., et al. Geox: Geometric problem solving through unified formalized vision-language pre-training. *arXiv preprint arXiv:2412.11863*, 2024.
- Xu, L., Zhao, Y., Wang, J., Wang, Y., Pi, B., Wang, C., Zhang, M., Gu, J., Li, X., Zhu, X., et al. Geosense: Evaluating identification and application of geometric principles in multimodal reasoning. *arXiv preprint arXiv:2504.12597*, 2025.
- Xu, S., Luo, Y., and Shi, W. 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*, pp. 11–15, 2024.
- Yan, S. and Zhong, X. 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)*, pp. 1–9. IEEE, 2024.
- Zhang, J., Li, Z., Zhang, M., Yin, F., Liu, C., and Moshfeghi, Y. Geoeval: benchmark for evaluating llms and multi-modal models on geometry problem-solving. *arXiv preprint arXiv:2402.10104*, 2024a.
- Zhang, M.-L., Yin, F., Hao, Y.-H., and Liu, C.-L. Plane geometry diagram parsing. *arXiv preprint arXiv:2205.09363*, 2022.
- Zhang, M.-L., Yin, F., and Liu, C.-L. A multi-modal neural geometric solver with textual clauses parsed from diagram. *arXiv preprint arXiv:2302.11097*, 2023a.
- Zhang, R., Jiang, D., Zhang, Y., Lin, H., Guo, Z., Qiu, P., Zhou, A., Lu, P., Chang, K.-W., Qiao, Y., et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pp. 169–186. Springer, 2024b.
- Zhang, X., Zhu, N., He, Y., Zou, J., Huang, Q., Jin, X., Guo, Y., Mao, C., Li, Y., Zhu, Z., et al. Formalgeo: An extensible formalized framework for olympiad geometric problem solving. *arXiv preprint arXiv:2310.18021*, 2023b.
- Zhang, X., Zhu, N., He, Y., Zou, J., Qin, C., Li, Y., and Leng, T. Fgeo-sss: A search-based symbolic solver for human-like automated geometric reasoning. *Symmetry*, 16(4):404, 2024c.
- Zhang, X., Zhu, N., Qin, C., Yang, L., Zeng, Z., and Leng, T. Formal representation and solution of plane geometric problems. In *The 4th Workshop on Mathematical Reasoning and AI at NeurIPS’24*, 2024d.
- Zhang, Z., Cheng, J.-K., Deng, J., Tian, L., Ma, J., Qin, Z., Zhang, X., Zhu, N., and Leng, T. Diagram formalization enhanced multi-modal geometry problem solver. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

## 8. Appendix

### 8.1. Details for Benchmarks

Detailed information about benchmarks is presented in Table. 1.

### 8.2. Details for Reasoners

Detailed information about benchmarks is presented in Table. 2.

### 8.3. Related Benchmarks

**Benchmarks for Diagram Parsing:** Since diagrams play an important role in GPS, many researchers concentrate on constructing benchmarks for geometry diagrams with fine-grained annotation. The PGDP5K dataset contains 5,000 diagram samples (Hao et al., 2022), consisting of 1,813 non-duplicated images from the Geometry3K dataset and other 3,187 images collected from three popular textbooks across grades 6-12 on mathematics website. In contrast to previous datasets, diagrams in PGDP5K have more complex layouts such as multiple classes of primitives and complicated primitive relations. Tangram, includes 1,080 diverse geometric diagrams sourced from primary and secondary school exams, competitions, and textbooks, covering from

simple basic geometric shapes to complex combinations, aiming to evaluate the performance of LMMs on geometric element recognition (Tang et al., 2024). Each diagram is associated with four questions about the diagram. SynthGeo228K is a collect of over 228,000 geometric diagrams, generated from a single template by modifying point positions and orientations, with rotation being a common data augmentation method (Zhang et al., 2025). SynthGeo228K, comprising 462 templates, provides a comprehensive exploration of basic geometry through the relationships between points, lines, and triangles, polygons, and circles.

**Benchmarks for Mathematical Visual Reasoning:** Recent advances in multimodal learning have produced several specialized datasets for mathematical visual reasoning. MathVista (Lu et al., 2023), MathVerse (Zhang et al., 2024b), and MathVision (Awais et al., 2024) represent significant contributions to this field, all designed to evaluate multimodal models’ performance through rich metadata annotations. These datasets differ primarily in their data collection approaches: MathVista integrates existing resources, MathVerse adapts publicly available materials, and MathVision collects entirely new samples from mathematical competitions. MathVista comprises 6,141 samples with detailed metadata annotations including question types, answer types, and task categories. MathVerse contains 2,612 high-quality mathematical problems with diagrams, each transformed into six variants with varying information distribution across modalities: text-dominant, text-lite, text-only, vision-intensive, vision-dominant, and vision-only. MathVision features 3,040 mathematics-visual question-answer pairs, meticulously curated through a four-stage filtering process by ten university students. Besides, math reasoning with visual content is also an interesting topic.



Table 1. Benchmarks in geometry problem solving.

| Benchmark            | Scale   | Level                      | Type                     | Format                                    | Source                                                             |
|----------------------|---------|----------------------------|--------------------------|-------------------------------------------|--------------------------------------------------------------------|
| GeoS                 | 186     | SAT-level<br>(Grades 6-10) | Calculation              | Choice-Answer                             | SAT                                                                |
| GeoS++               | 1,406   | SAT-level<br>(Grades 6-10) | Calculation              | Choice-Answer                             | SAT & Textbook                                                     |
| Geometry3K           | 3,002   | SAT-level<br>(Grades 6-12) | Calculation              | Choice-Answer                             | Textbook                                                           |
| PGPS9K               | 9,022   | SAT-level<br>(Grades 6-12) | Calculation              | Solution Sequence                         | Textbook                                                           |
| GeoQA                | 4,998   | SAT-level<br>(Grades 6-12) | Calculation              | Key point & Solution<br>Sequence          | Exams (Chinese)                                                    |
| GeoQA+               | 7,528   | SAT-level<br>(Grades 6-12) | Calculation              | Key point & Solution<br>Sequence          | Exams (Chinese)                                                    |
| UniGeo               | 14,541  | SAT-level<br>(Grades 6-12) | Calculation &<br>Proving | Solution Sequence                         | Online Education Website                                           |
| FormalGeo7K          | 6,987   | SAT-level to<br>IMO-level  | Calculation &<br>Proving | Solution Sequence                         | Online Website                                                     |
| GeoEval-<br>2000     | 2,000   | SAT-level                  | Calculation &<br>Proving | Solution Sequence                         | Geometry3K, PGPS9K,<br>UniGeo, MATH, GeoQA+,<br>GeometryQA, MathQA |
| GeoEval-<br>backward | 750     | SAT-level                  | Calculation &<br>Proving | Solution Sequence<br>(Backward Reasoning) | Geometry3K, PGPS9K,<br>UniGeo, MATH, GeoQA+,<br>GeometryQA, MathQA |
| GeoEval-aug          | 2,000   | SAT-level                  | Calculation &<br>Proving | Solution Sequence                         | Geometry3K, PGPS9K,<br>UniGeo, MATH, GeoQA+,<br>GeometryQA, MathQA |
| GeoEval-hard         | 300     | SAT-level                  | Calculation &<br>Proving | Solution Sequence                         | Geometry3K, PGPS9K,<br>UniGeo, MATH, GeoQA+,<br>GeometryQA, MathQA |
| GPMS4K &<br>GeoVQA   | 4,440   | SAT-level<br>(Grades 6-12) | Calculation &<br>Proving | Solution Sequence                         | Textbook (Indian)                                                  |
| GeoSense             | 1,789   | SAT-level                  | Calculation              | Solution Sequence                         | Online Website<br>(English & Chinese))                             |
| GNS260K              | 154,433 | SAT-level<br>(Grades 6-12) | Calculation              | Solution Sequence                         | Synthesis<br>(PGPS9K, GeoQA+, Geo170K)                             |
| GeoMath              | 9,155   | SAT-level                  | Calculation &<br>Proving | Solution Sequence                         | Online Education Website<br>(Chinese)                              |
| Geo170K              | 110,000 | SAT-level<br>(Grades 6-12) | Calculation              | Solution Sequence                         | Synthesis<br>(Geometry3K, GeoQA+)                                  |
| GeoGPT4V             | 23,955  | SAT-level<br>(Grades 6-12) | Calculation              | Solution Sequence                         | Synthesis<br>(UniGeo-Calculation,<br>Geometry3K, GeoQA+)           |
| GeoMM                | 87,000  | SAT-level                  | Calculation              | Solution Sequence                         | Synthesis                                                          |
| GeoExpand            | 45,526  | SAT-level                  | Calculation              | Solution Sequence                         | Synthesis<br>(Geometry3K and PGPS9K)                               |
| GeoSynth             | 62,868  | SAT-level                  | Calculation              | Solution Sequence                         | Synthesis                                                          |

Table 2. Reasoners in geometry problem solving.

| Reasoner                                   | Technique  | Paradigm       | Reasoner                                                         | Technique  | Paradigm       |
|--------------------------------------------|------------|----------------|------------------------------------------------------------------|------------|----------------|
| <b>ML-based Techniques</b>                 |            |                |                                                                  |            |                |
| GeoS<br>(Seo et al., 2015)                 | Rule-based | Symbolic       | GeoS+<br>(Sachan & Xing, 2017)                                   | Rule-based | Symbolic       |
| FGPS<br>(Zhang et al., 2024c)              | Rule-based | Symbolic       | InterGPS<br>(Lu et al., 2021)                                    | NN-based   | Neuro-Symbolic |
| E-GPS<br>(Wu et al., 2024)                 | NN-based   | Neuro-Symbolic | GeoQA<br>(Chen et al., 2021)                                     | NN-based   | Connectionist  |
| PGPSNet<br>(Hao et al., 2022)              | NN-based   | Connectionist  |                                                                  |            |                |
| <b>LLM-based Techniques</b>                |            |                |                                                                  |            |                |
| G-LLaVA<br>(Gao et al., 2023)              | ICL        | Connectionist  | DFE-GPS<br>(Zhang et al., 2025)                                  | ICL        | Connectionist  |
| Geo-LLaVA<br>(Xu et al., 2024)             | ICL        | Connectionist  | GeoX<br>(Xia et al., 2024)                                       | ICL        | Connectionist  |
| GeoUni<br>(Cheng et al., 2025)             | ICL        | Connectionist  | AlphaGeometry<br>(Trinh et al., 2024)                            | ICL        | Neuro-Symbolic |
| AlphaGeometry2<br>(Chervonyi et al., 2025) | ICL        | Neuro-Symbolic | GeomVerse<br>(Kazemi et al., 2023)                               | CoT        | Connectionist  |
| Two-Stage CoT<br>(Wang & Ma, 2024)         | CoT        | Connectionist  | GNS<br>(Ning et al., 2025)                                       | CoT        | Connectionist  |
| GeoQwen<br>(Yan & Zhong, 2024)             | CoT        | Connectionist  | GeoCoder<br>(Sharma et al., 2024)                                | PoT        | Connectionist  |
| RaE<br>(Duan et al., 2024)                 | PoT        | Connectionist  | Simulacra-<br>Conversational Agents<br>(Mouselinos et al., 2024) | CoA        | Connectionist  |