Encode Geometric Diagram as Geo-Graph in Geometry Problem Solving

Wenjun Wu[1,2], **Lingling Zhang**[1,2*], **Bo Zhao**[1,2], **Bo Li**[1,3], **Xinyu Zhang**[1,2], **Yaqiang Wu**[4]

[1]School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, 710049, China
[2]Ministry of Education Key Laboratory of Intelligent Networks and Network Security, Xi'an Jiaotong University, Xi'an, 710049, China
[3]Shaanxi Province Key Laboratory of Big Data Knowledge Engineering, Xi'an Jiaotong University, Xi'an, 710049, China
[4]Lenovo Research
zhanglling@xjtu.edu.cn, nickjunwork@163.com

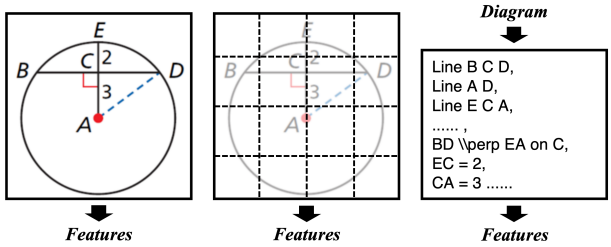
Abstract

Geometry Problem Solving has become a hot topic these years due to its complexity of enabling the machine with geometric abstraction, multi-modal reasoning and mathematical capabilities. Majority of research works place their attention on the fusion of multi-modal data or the synergistic combination of neural and symbolic systems for performance improvement. However, their neglect of the unique characteristics of geometric diagrams, which distinguish them from natural images, impedes the further exploring of critical information in geometric diagrams. In this work, we introduce the novel concept of geo-graph and propose the Geo-Graph Geometry Problem Solving model which encodes the geometric diagram from a new perspective. The geo-graph is designed to include semantic, structural and spatial information in the diagram, which is crucial to subsequent problem reasoning stage. To facilitate the model's comprehension of the actual layout of geometric diagram, spatial and connecting attentions are devised to serve as intrinsic knowledge guidance for feature propagation. An extra cross-modal attention is used as external guidance to instruct the encoding of geo-graph to be related to specific problem target. Fused multi-modal features are then sent into a commonly used encoder-decoder framework for final solution generation. The model is first trained with three carefully designed pre-training tasks to establish its fundamental knowledge of geo-graph, leveraging numerous varied samples generated through a geo-graph-based augmentation method. Experiments on popular geometry problem solving datasets demonstrate the effectiveness and superiority of our model for geometric diagram encoding.

Code — https://github.com/nicktech-git/Geo-Graph

1 Introduction

Geometry Problem Solving (GPS) has drawn growing attention from the automated reasoning community (Liu et al. 2023; Trinh et al. 2024; Zhao et al. 2025; Zhang et al. 2025a) due to its collaborative use of symbolic and neural systems for high-order logical reasoning. In practical educational scenarios, the task aims at obtaining both the numerical answer and corresponding solutions with given geometric diagram and problem text. The text usually contains some basic descriptions of the geometry problem and final goal, while the diagram is a concrete visualization of the geometric conditions. Therefore, it requires to first understand the multi-modal geometric information and then reason for the answer, where geometric abstraction, multi-modal reasoning and mathematical capabilities of the machine are indispensable.



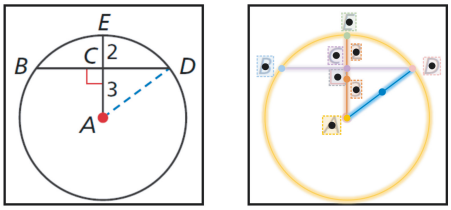
(a) the whole diagram (b) patches (c) formal languages

Figure 1: Different strategies of encoding geometric diagram in existing works. (a) extracts generic features of the whole diagram; (b) divides the diagram into patches to learn local features with spatial information; (c) transforms the diagram into structural languages to retain structural information.

Recent works mainly form two mainlines. Neural-based approaches (Zhang, Yin, and Liu 2023; Ning et al. 2023; Gao et al. 2025) transform the geometry problem into multi-modal features and feed them into neural networks to predict solution sequences. Symbolic-based approaches (Wu et al. 2024; Peng et al. 2023; Zhao et al. 2025) parse the problem into predefined formal languages and continuously deduct for the answer with symbolic engines (Lu et al. 2021). While there has been increasing studies (Xia et al. 2025; Wang et al. 2025; Zhang et al. 2025c) utilizing the power of large language models (LLMs), their underlying principles can still be traced back to the two aforementioned approaches. However, regardless of the varied data processing mechanisms employed, existing methods generally exhibit limited consideration for the unique characteristics of geometric diagrams in visual information understanding, which are distinct from those of natural images.

We present three typical strategies to encode geomet-

*Corresponding author.



(a) Original Diagram **(b) Geo-Graph**

Figure 2: Illustration of geo-graph. (a) is the original geometric diagram from PGPS9K data set; (b) is a simplified visualization of virtual geo-graph that corresponds to the actual layout of diagram, where centers of geometric and non-geometric primitives are represented by colored and black dots respectively. For easier correspondence in vision, the primitives are painted or grounded the same color as their paired primitive centers. For instance, the circle in (b) is painted yellow as same as its centre point and the text `E` is grounded in a green box as same as the point below it.

ric diagrams in fig. 1. Some early works (Zhang, Yin, and Liu 2023) treat geometric diagrams equivalently to natural images, extracting the generic features of a whole diagram by using CNN. This is primarily beneficial for learning only certain overall content-related information, often failing to capture fine-grained geometric properties. Other works (Chen et al. 2021; Ning et al. 2023) recognize the significance of layout clues contained in diagram, thus dividing the diagram into patches and aligning local visual features with characters in text. This is also widely adopted by GPS researches (Li et al. 2024a; Cho et al. 2025) that involve LLMs with pretrained vision encoders (Radford et al. 2021). While such approach effectively captures the spacial information of individual points, it inevitably breaks the geometric structures in the diagram, where the visual connecting of primitives is highly sensitive. For example, `Line BD` in fig. 1(b) is divided into four patches and could be easily recognized as multiple lines if the network is not adequately trained. To preserve the structural information, some studies (Zhang, Yin, and Liu 2023; Li et al. 2024b; Zhang et al. 2025c) transform the diagram into structural clauses as in fig. 1(c), which are then processed by language models to explore the geometric relations. However, this strategy lacks consideration of spatial feature, which is another critical determinant in understanding and solving geometry problems. Additionally, a high risk of information loss exists in such vision-to-language transformation due to the suboptimal performances of existing parsers.

In contrast, we contend that the underlying motivation behind the formalization and drafting of geometric diagrams suggests their representation and modeling as graphs. We follow the classification principles of previous works (Zhang et al. 2022) to divide the fundamental elements in the diagram into geometric primitives (*i.e.* points, lines and circles) and non-geometric primitives (*i.e.* symbols and texts). Therefore, geometric diagram serves as the concrete visualization of symbolized primitives and their abstract relations, illustrating specific geometric conditions. As shown in fig. 2(a), `Line BD` $\perp$ `Line AE` is visualized as `Line BD` and `Line AE` being perpendicularly intersecting at `Point C` with a vertical bar connecting them three. To this end, graph provides inherent advantages in modeling the representations of multiple elements and their intrinsic structural relations. Moreover, spatial clues are also instrumental for accurate diagram understanding. For example in fig. 2(a), text `2` and `3` are paired and allocated to `Line AE` according to their relative locations. Hence, the graph is required to include spatial information of the primitives for more fine-grained encoding of the diagram. We name it as geo-graph, depicted in fig. 2(b). It is actually the reconstruction of visual geometric conditions within a virtual environment, projected from the real geometric diagram.

To achieve this, we propose the Geo-Graph Geometry Problem Solving (G^3PS) model which provides a new perspective to encode geometric diagram. It first parses the basic geometric and non-geometric primitives, which are taken as nodes of the geo-graph. In details, the semantic and spatial embeddings of the primitives are fused to compose the initial features of the nodes. Next, a graph transformer is utilized to encode geo-graph based on its node representations, guided by a multi-attention mechanism devised from three aspects: spatial attention, connecting attention and cross-modal target attention. On one hand, the spatial and connecting relations are integrated into the attention as intrinsic knowledge to help the model understand the actual layout of geometric diagram. On the other hand, the relevance between primitives and problem goal is transformed as cross-modal attention for target-guided encoding of geo-graph. For textual data, the original problem text is extended by structural and semantic clauses and processed by language model. The multi-modal features are then fed into an encoder-decoder framework for problem reasoning to finally predict the executable solution sequence and generate the answer. Additionally, three pre-training tasks are carefully designed to help the graph transformer establish fundamental knowledge of geo-graph. Specifically, the model is trained to predict the classes of nodes, the locations of nodes and indicative relations within node pairs. To provide diverse samples of geometric diagram content, a novel geo-graph-based data augmentation method is also implemented and used in both pre-training and training stages. Our contributions can be summarized as four-folds:

- We introduce the concept of geo-graph and the Geo-Graph Geometry Problem Solving model, encoding the geometric diagram from a novel perspective. To the best of our knowledge, it is the first work to transform the diagram into geo-graph, considering all semantic, structural and spatial information.
- We devise three attentions to guide the model's comprehension and encoding of geo-graph with both intrinsic knowledge and external cross-modal information.
- We design three pre-training tasks and a novel geo-graph-based data augmentation method to enhance model learning in scenarios with limited training samples.
- Experiments over popular GPS datasets demonstrate the effectiveness of G^3PS in geometric diagram encoding.

2 Methodology

Given a geometry problem (D, T), where D is a geometric diagram and T is problem description in natural language, the task aims at obtaining the final answer of problem goal and corresponding solutions. The overall framework of our G^{3}PS is illustrated in fig. 3.

2.1 Geometric Diagram Encoding

While geometric diagrams contain rich geometric conditions, their unique characteristics that differentiate them from natural images have been overlooked by prior research. Therefore, the geo-graph is introduced for enhanced exploration of spatial and structural information in diagrams.

Geo-Graph Construction For a geometric diagram D, the commonly used parser (Zhang et al. 2022) is used to first detect K^{ng} non-geometric primitives and segment K^g geometric primitives with their locations. The corresponding visual or semantic d^{sem}-dimension embedding is taken as the initial semantic feature of nodes in geo-graph, denoted as $V = \{v_1, v_2, \cdots, v_K\}$ where $K = K^g + K^{ng}$. Different from nodes in normal graphs, the nodes in geo-graph are location-sensitive, which means even if the nodes have similar semantics, they represent different primitives when they have different locations in the diagram. Therefore, a mapping rule is used to generate the initial d^{loc}-dimension location features $P = \{p_1, p_2, \cdots, p_K\}$ of the nodes from their bounding boxes or segmentation locations. Then, the semantic features and location features are concatenated after two multi-layer perceptrons to compose the initial node representations $N = \{n_1, n_2, \cdots, n_K\}$:

$$n_i = \left[MLP_1\left(v_i\right); MLP_2\left(p_i\right)\right], \tag{1}$$

where $n_i \in \mathbb{R}^{d^{emb}}$. For inclusion of as much potential information as possible, the geo-graph is first considered as a fully connected undirected graph.

Multi-attention for Geo-Graph To better propagate information among nodes, the Graphormer (Ying et al. 2021) is used to encode geo-graph for its strong modeling capabilities. However, necessary guidance is needed to instruct the transformer to learn the actual geometric layout and focus on parts that benefit solving the problem. To this end, three kinds of supervision are proposed to provide the model with intrinsic knowledge and external guidance as attention bias, namely spatial attention, connecting attention and cross-modal target attention.

Spatial Attention. For a pair of nodes (n_i, n_j), the closer they are when corresponded back to the diagram, the larger possibility that they have correlation and constitute one geometric condition. For example, the text symbols "2" and "3" in fig. 2 should be paired with `Line AE` to define its length rather than `Line AD`. Thus, the relative distances between nodes are calculated according to the center coordinates (x_i, y_i) and (x_j, y_j) of their locations to generate a symmetric spatial attention matrix $M^S_{i,j}\,(1 \le i, j \le K)$:

$$M^S_{i,j} = 1 - Norm\left(\sqrt{\left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2}\right), \tag{2}$$

where $Norm\left(\cdot\right)$ function norms the value of each $M^S_{i,j}$ into $[0, 1]$ according to the minimum and maximum values.

Connecting Attention. Despite the spatial relevance, there is another stronger relation to constrain the induction of geometric conditions, *i.e.* connecting relevance. Geometric primitives that are visually connected in the diagram form a concrete geometric structure, such as `Line BD` intersecting with `Circle A` at `Point B` and `D` in fig. 2. In order to facilitate the model to understand such structural clue, a symmetric connecting attention matrix $M^C_{i,j}\,(1 \le i, j \le K)$ is generated from the prediction of parser:

$$M^C_{i,j} = \begin{cases} 1, \text{if } n_i \text{ is connected to } n_j \\ 0, \text{if } n_i \text{ is not connected to } n_j \end{cases}. \tag{3}$$

Notice that such prediction provides valuable reference but is not definite, ergo serving as one of the intrinsic attention guidance instead of a filtering mask.

Cross-modal Target Attention. Intrinsic guidance focuses on the diagram itself, while external information from the geometry problem is also needed to concentrate on key parts that could eventually lead to the answer. Hence, the problem target representation e^{tg}_* from section 2.2 is used to compare the degrees of correlation between it and all nodes, followed by a softmax function:

$$a^T_j = \frac{exp\left(sim\left(n_j, e^{tg}_*\right)\right)}{\sum_{j=1}^{K} exp\left(sim\left(n_j, e^{tg}_*\right)\right)}, \tag{4}$$

where $sim\left(\cdot\right)$ can be any similarity function and $1 \le j \le K$. Since such attention should be directly related to the global representation of geo-graph, a virtual global node n^g is introduced by averaging the features of N. Thus, the target attention matrix $M^T_{i,j}\,(1 \le i, j \le K+1)$ is generated by allocating the correlation degree a^T_j to the nodes with respect to the global node n^g:

$$M^T_{i,j} = \begin{cases} a^T_{j-1}, \text{ if } i = 1, 1 < j \le K+1 \\ 0, \text{ if } 1 < i \le K+1, 1 \le j \le K+1 \\ m, \text{ if } i = 1, j = 1 \end{cases}, \tag{5}$$

where m is set to 0.5 empirically. To merge both intrinsic attention and external target guidance, a final attention matrix $M_{i,j}\,(1 \le i, j \le K+1)$ is derived by aligning each attention matrix with its corresponding position, which is also illustrated in fig. 3 (a):

$$M_{i,j} = \begin{cases} \alpha_1 M^T_{i,j}, \text{ if } i = 1 \text{ or } j = 1 \\ \alpha_2 M^S_{i-1,j-1} + \alpha_3 M^C_{i-1,j-1}, \text{otherwise} \end{cases}, \tag{6}$$

where α_1, α_2 and α_3 are hyper-parameters.

Global Representation Encoding Under internal and external guidance, the graph transformer is instructed to encode the high-level semantics of the geo-graph from local nodes to the global node. For a L-layer graph transformer, the l-th layer $f^{GT}_l\left(\cdot\right)$ takes in the global node n^g_{l-1} and local nodes N_{l-1} from previous layer with the attention matrix M obtained by eq. (6), and outputs the new representations n^g_l and N_l:

$$n^g_l, N_l = f^{GT}_l\left(n^g_{l-1}, N_{l-1}, M\right). \tag{7}$$

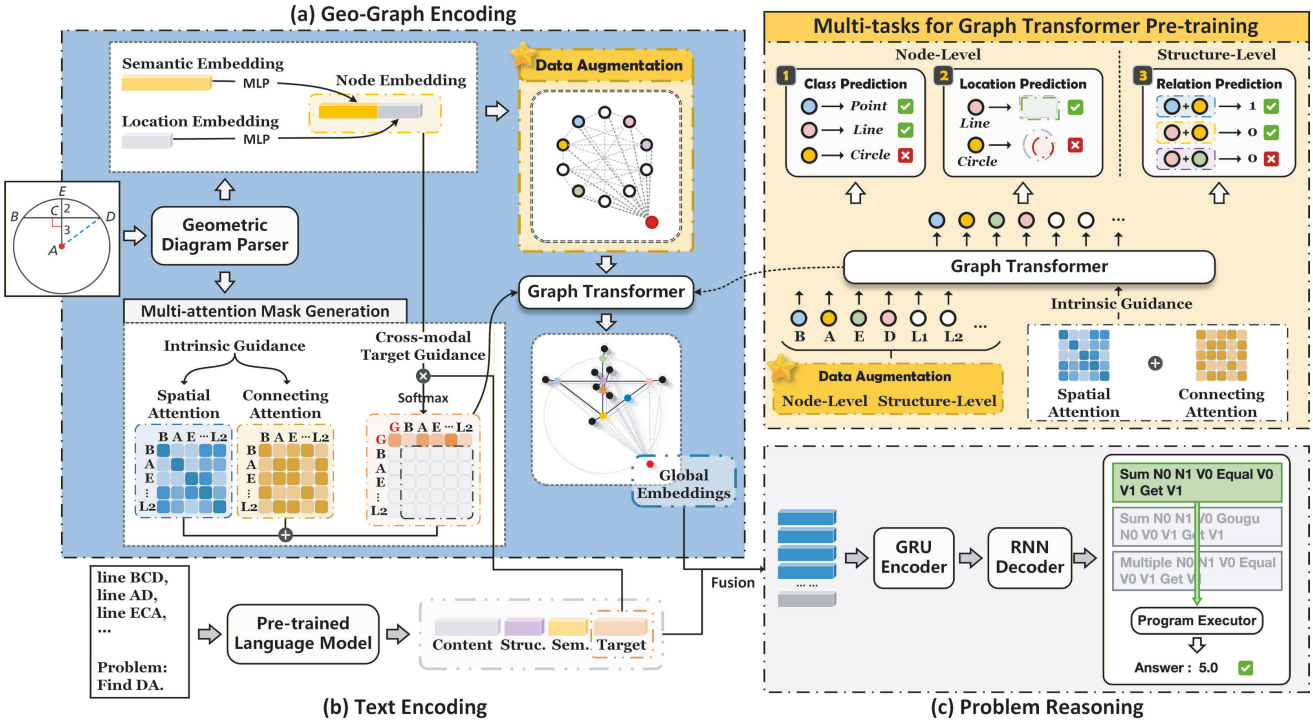


Figure 3: The overall framework of G^{3}PS which includes four parts. In (a), the proposed geo-graph is constructed and then encoded by pre-trained graph transformer under multi-attention guidance. Textual features are generated by a language model in (b). Combined multi-modal features are fed into an encoder-decoder network to obtain final solution and answer in (c). In pre-training stage, three auxiliary tasks and data augmentation method are used, as depicted in the top-right yellow box.

The representations of global node of all L layers are taken as the guided encoding results of the whole geo-graph, denoted as $N^g = \{n_1^g, n_2^g, \cdots, n_L^g\}$.

2.2 Text Encoding

Problem text includes not only the problem target, but also some crucial geometric conditions that are complementary to diagram. Following previous works (Zhang, Yin, and Liu 2023; Li et al. 2024b), the structural and semantic clauses parsed from the diagram are combined with the original problem text. It effectively strengthens the understanding of geometric conditions from the textual perspective. In details, the extended text $T = \{T^{cont}, T^{tg}\}$ is sent into a pre-trained language model to obtain its generic features $E = \{E^{cont}, E^{tg}\}$ from the semantic, class and section aspects of each word token:

$$E = f_{sem}^{LM}(T) + f_{cls}^{LM}(T) + f_{sec}^{LM}(T), \tag{8}$$

where $f_*^{LM}(\cdot)$ are corresponding embedding layers, *cont* and tg refer to extended content and problem target. For cross-modal interaction, the representation of problem target e_*^{tg} is generated by averaging the word features in E^{tg} after a linear projection:

$$e_*^{tg} = avg\left(Linear\left(E^{tg}\right)\right), \tag{9}$$

which is considered as external target guidance in eq. (4).

2.3 Problem Reasoning

As the encoding results of geo-graph and problem text (*i.e.* N^g and E) have been acquired after problem understanding stage, the features from both modalities are then concatenated for mutual complement of information. We adopt the commonly used encoder-decoder structure (Chen et al. 2021, 2022; Zhang, Yin, and Liu 2023) in reasoning stage for effectiveness and simplicity. Specifically, a GRU-based encoder and a self-limited GRU-based decoder (Zhang, Yin, and Liu 2023) are used to explore high-level geometric clues and predict the solution $S = \{s_1, s_2, \cdots, s_H\}$. An offline symbolic executor is employed to calculate the final answer a^* of the problem according to the predicted executable program S. In practice, the model is optimized to generate correct solution sequence by Cross Entropy Loss:

$$L = -\sum_{i=1}^{H}\sum_{j=1}^{C} y_{i,j}^s \cdot \log q\left(s_{i,j}\right), \tag{10}$$

where H is the length of S, C is the size of self-limited sequence vocabulary space and $y_{i,j}^s$ is the ground-truth label.

2.4 Multi-task Pre-training

Although the graph transformer excels at modeling complicated graph structures, it requires sufficient training samples to support the optimization of its parameters. However, fine-grained GPS data sets have always been hard to acquire and

annotate, making it a natural challenge for learning of deep neural networks in GPS. To alleviate this, three pre-training tasks are designed to help the model establish fundamental knowledge of geo-graph, namely class prediction, location prediction and relation prediction.

Class Prediction. Each node representation is supposed to maintain key semantic features for its class classification (*e.g.*, point or text) after multi-layer information propagation among the neighbors. We randomly mask and replace the class tags of partial nodes with mask token or other word, and urge the model to predict their ground-truth class labels:

$$q_i^{cls} = f_{cls}\left(f^{GT}\left(n_i\right)\right), \tag{11}$$

where $f_{cls}\left(\cdot\right)$ is a one-layer linear projection and q_i^{cls} is predicted class tag of node n_i.

Location Prediction. As one major difference of geo-graph from ordinary graphs is that each node contains the location of represented primitive in the diagram, the model should learn to merge and retain such spatial information in the nodes. Therefore, it is expected to predict the precise locations q^{loc} of nodes using a one-layer linear function $f_{reg}\left(\cdot\right)$:

$$q_i^{loc} = f_{reg}\left(f^{GT}\left(n_i\right)\right). \tag{12}$$

Relation Prediction. The indicative relation between nodes is crucial for understanding the structure of geo-graph. A well-trained model should be able to differentiate strong and weak correlations between nodes, which forms the foundation for subsequent high-level tasks such as graph encoding and problem reasoning. To achieve this, the relation prediction task is applied to discriminate whether a strong bond relation exists between two nodes:

$$q_{i,j}^{rel} = f_{rel}\left(f^{GT}\left(n_i\right) - f^{GT}\left(n_j\right)\right), \tag{13}$$

where $q_{i,j}^{rel} \in \{0, 1\}$ and $f_{rel}\left(\cdot\right)$ is a multi-layer perceptron. To optimize the model, the commonly used Cross Entropy Loss is applied for class prediction and relation prediction (*i.e.*, L_{cls} and L_{rel}), while the Smooth L1 loss is used for location prediction (*i.e.*, L_{reg}). Hence, the final loss function in pre-training stage is defined as $L_{pre} = L_{cls} + \beta_1 L_{reg} + \beta_2 L_{rel}$ to teach the model at both node-level and structure-level, where β_1 and β_2 are manually set hyper-parameters.

2.5 Data Augmentation

While pre-training can efficiently help the model build fundamental knowledge of geo-graph, there is still a risk of over-fitting on small scale datasets. Previous works adopted some basic approaches to augment diagram and text, such as diagram flipping, token replacement and representation transposition. These methods indeed generate various representations of each sample, but they are constrained by the original geometric conditions which ultimately leads to the same reasoning path of model. Instead, we seek for a way that could actually provide samples of diverse geometric conditions, which is also one of the reasons for introducing the concept of geo-graph. As shown in fig. 4, we assume that any geometric diagram, regardless of how complicated

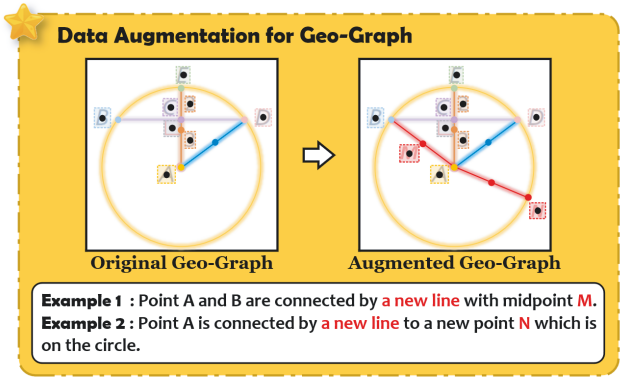


Figure 4: Data augmentation method for geo-graph. Taking an optimized geo-graph as an example which is transformed from the geometric diagram, the content in red on the right are added by our augmentation method.

it could be, can be transformed into the form of geo-graph. Therefore, it is feasible for us to make editions on primitives and structures as follows:

- *Primitive Addition*: To randomly add new virtual primitive node in the geo-graph such as points, lines and circles. Each virtual node is assigned the same attribute categories as real nodes but with various labels.
- *Structure Addition*: To randomly add new indicative relations between real nodes and virtual nodes. Each relation added indicates the join of a new geometric condition.

These two strategies are not independent but interrelated and they are compatible with aforementioned basic augmentation methods in both pre-training and training stage. Because they actually enrich the geometric conditions with confusing noise, the model is further encouraged to explore the decisive content for feasible solutions. We decide not to add any numerical attributes (*e.g.*, length of line or measure of angle) because it may lead to conflicts with original geometric conditions and affect the solution discovery.

3 Experiments

3.1 Datasets

We mainly compare the performances on the following two popular and appropriately annotated GPS datasets.

PGPS9K (Zhang, Yin, and Liu 2023). A fine-grained GPS dataset with diagram annotation and interpretable solution programs. It consists of 9,022 geometry problems paired with non-duplicate 4,000 geometric diagrams, part of which are selected from Geometry3K (Lu et al. 2021). For comprehensive examination of model performance, the dataset is split into 8,022 training and 1,000 testing samples according to all problem types with a ratio of 8:1.

Geometry3K* (Zhang, Yin, and Liu 2023). A derivation of the original Geometry3K, which is enriched by PGPS9K but keeps the test set of 589 disjoint samples. It is a typical benchmark to measure the performances of both symbolic-based and neural-based models. We follow previous works

Method	Geometry3K*			PGPS9K			Parameters
	Completion	Choice	Top-3	Completion	Choice	Top-3	
Human Expert	-	90.9	-	-	-	-	-
InterGPS-*pred* (Lu et al. 2021)	44.6	56.9	-	-	-	-	-
InterGPS-*diagram* (Lu et al. 2021)	64.2	71.7	-	59.8	68.0	-	-
InterGPS-*GT* (Lu et al. 2021)	69.0	75.9	-	-	-	-	-
GeoDRL-*pred* (Peng et al. 2023)	-	68.4	-	-	-	-	-
E-GPS-*pred* (Wu et al. 2024)	-	67.9	-	-	-	-	-
Pi-GPS (Zhao et al. 2025)	<u>70.6</u>	77.8	-	61.4	69.8	-	-
LLaVA-v1.5 (Liu et al. 2024)	7.6	11.2	-	6.3	9.1	-	7B
Qwen-VL (Bai et al. 2023)	22.1	26.7	-	20.1	23.2	-	7B
GPT-4V (Achiam et al. 2023)	38.6	42.3	-	31.8	40.3	-	-
GeoGen-SFT-7B (Pan et al. 2025)	46.3	58.4	-	43.9	54.3	-	7B
Neural Solver (Lu et al. 2021)	-	35.9	-	-	-	-	-
NGS (Chen et al. 2021)	35.3	58.8	62.0	34.1	46.1	60.9	80M
Geoformer (Chen et al. 2022)	36.8	59.3	62.5	35.6	47.3	62.3	267M
SCA-GPS (Ning et al. 2023)	-	76.7	-	-	-	-	>310M
LANS (Li et al. 2024b)	**71.3**	<u>82.3</u>	<u>82.0</u>	<u>66.1</u>	<u>73.8</u>	<u>81.7</u>	26M
– *baseline* (PGPSNet)	65.0	77.9	80.7	62.7	70.4	79.5	<u>23M</u>
G^3PS (ours)	<u>70.6</u>	**85.1**	**83.4**	**66.6**	**77.5**	**81.8**	**16.6M**

Table 1: Comparison of GPS on Geometry3K* and PGPS9K. Best/second best results are in bold/underlined.

and choose this data set for its consideration for neural-based methods that need more training data.

3.2 Baselines and Evaluation Metrics

We choose the state-of-the-art and some typical methods as competitors for detailed comparison.

Symbolic-based methods. Inter-GPS (Lu et al. 2021) made the first attempt to continuously deduct for new geometric conditions with a predefined rule base. Following previous work (Li et al. 2024b), we list out its performances under different situations, where *-pred*, *-diagram* and *-GT* refer to using predicted parsing results of diagram and text, annotations for diagram, and annotations for both modalities, respectively. GeoDRL (Peng et al. 2023) constructs a complicated logic graph from formal language conditions and predicts each theorem application with an improved rule base. E-GPS (Wu et al. 2024) proposes to first reason for the solution in a top-down decomposition manner and then solve the problem by obtained solution. Pi-GPS (Zhao et al. 2025) utilizes MLLMs to disambiguate text based on diagram. AlphaGeometry (Trinh et al. 2024) is not selected, which is designed for geometry problem proving, due to its rule base being inapplicable for educational use and calculation, as well as its lack of diagram processing capabilities.

Neural-based methods. Earlier typical works such as NGS (Chen et al. 2021) and Geoformer (Chen et al. 2022) relies more on textual modality without using diagram parsers. For consistency, we report their performances from (Li et al. 2024b). SCA-GPS (Ning et al. 2023) enhances the model's understanding of geometric conditions by aligning the symbolic characters in diagram and language, which introduces visual information. PGPSNet (Zhang, Yin, and Liu 2023), treated as our baseline model, extends the problem text with structural and semantic clauses parsed from diagram and extracts features of diagram by CNN. LANS (Li et al. 2024b) further improves the layout awareness of model by aligning the points in patches of diagram with the characters. We also recognize the attention from the community on performances of large language models, *e.g.*, LLaVA (Liu et al. 2024) and GPT-4V (Achiam et al. 2023), to solve geometry problems and borrowed results from (Li et al. 2024b).

Evaluation Metrics. Following the settings of previous works (Li et al. 2024b), three evaluation metrics are mainly used to assess the problem solving performances, namely *Completion*, *Choice* and *Top-k*. In *Completion*, the problem is solved correctly if the first executable solution from model output obtains the correct numerical answer. It changes into *Top-k* for looser tolerance where any one within top k solution programs can lead to the ground-truth answer. *Choice* is prepared for more practical scenarios, where model is forced to choose one closest answer out of candidates or by random if it fails to, and is considered correct if the choice is made right. For more detailed analysis, an extra evaluation metric *Solution Accuracy* is added to further measure the reasoning abilities, where the solution is seen as correct if it is shorter than or equal as the annotated programs under *Completion*.

3.3 Implementation Details

We simply set the graph transformer with 6-layers, 8-head, 256-dimension input size and 512-dimension hidden size. The d^{loc}, d^{sem} and d^{emb} are set as 9, 256 and 256, respectively. In consistency with baseline PGPSNet, the parameter settings of encoder and decoder remain the same. Empirically, we assign $\{1, 0.5, 0.5\}$ and $\{10, 1\}$ to hyperparameters $\{\alpha_1, \alpha_2, \alpha_3\}$ and $\{\beta_1, \beta_2\}$. The overall model is optimized by Adam optimizer with initial learning rates 0.001, 0.0001 and 0.0005 for normal layers, pre-trained language model and pre-trained graph transformer.

Method	Solution Accuracy	
	Geometry3K*	PGPS9K
PGPSNet	64.4	61.0
G^{3}PS	**69.8**	**64.5**

Table 2: Comparison of quality of solutions. Best results are in bold.

3.4 Performance Comparison

The overall performance results are recorded in table 1, from which we mainly make the following observations:
(1) G^{3}PS achieves the state-of-the-art performances over almost all evaluation metrics on both datasets. Specifically, it obtains an average improvement of 5.2% and 4.4% on Geometry3K* and PGPS9K with a decrease of 6.4M parameters, compared with the baseline model PGPSNet. It demonstrates the effectiveness of our G^{3}PS to encode information in geometric diagram, and indicates that understanding geometric diagrams by graph representations is promising.
(2) Among the neural-based methods, our G^{3}PS achieves the best overall performances with the least parameter scale. In details, G^{3}PS outperforms the previous SOTA model LANS over almost all metrics with a sum of 7.8% improvement gain, while reducing the parameter count from 26M to 16.6M. Note that LANS integrates an additional LA-FA module based on the encode-decoder framework of baseline which we use. The results further demonstrates the efficiency of G^{3}PS to explore crucial information in diagrams.
(3) Symbolic-based methods can be improved with ease by augmentation of rules and exhaustive search, but they are sensitive to correctness of parsing results. Take InterGPS-*diagram* and InterGPS-*pred* as examples, the large performance gap between with and without diagram annotation indicates the importance of understanding geometric diagram.
(4) Large vision language models present excellent performances over general tasks, but are still unsatisfactory on GPS task which requires geometric abstraction, multi-modal reasoning and mature mathematical capabilities.

We further evaluate the solution quality of G^{3}PS in comparison with the baseline model PGPSNet in table 2. It can be seen that G^{3}PS is more capable of generating shorter solutions that lead to correct answers. It mainly benefits from the enhanced understanding of geometric diagrams, thereby reducing the unnecessary steps of solution.

3.5 Ablation Studies and Discussion

We have carried out extensive experiments for ablation studies to verify the effectiveness of different parts. As shown in table 3, the model performance is constrained in all situations when any part of the proposed method is removed, which means they all contribute to the final performance improvement. In details, connecting and spatial attentions are both necessary for establishing basic understanding of geo-graph, while external target guidance seems even more important because it instructs the model to concentrate on the problem goal-related primitives. Without target guidance, the encoding of geo-graph will be general and rough. Removing pre-training tasks does not prevent the model from learning key information with attention provided, but it is helpful to build geo-graph knowledge for better optimization. Augmentation for geo-graph and problem is important to provide various training samples in case of over-fitting. With all parts disabled, the encoding of geo-graph would rather become noise and results in performance decrease.

Pre.	Aug.	Attentions			Geometry3K*	
		Spa.	Con.	Tgt.	Completion	Solution
					64.2	63.3
✓	✓				67.7	67.4
✓	✓		✓	✓	68.6	67.6
✓	✓	✓		✓	68.3	68.1
✓	✓	✓	✓		68.1	67.4
✓		✓	✓	✓	67.4	66.7
	✓	✓	✓	✓	68.8	68.1
✓	✓	✓	✓	✓	**70.6**	**69.8**

Table 3: Ablation study results. Pre., Aug., Spa., Con., Tgt. are the abbreviations for pre-training, data augmentation, spatial attention, connecting attention and target guidance, respectively. Solution metric refers to *Solution Accuracy*. ✓indicates enabling the module. Best results are in bold.

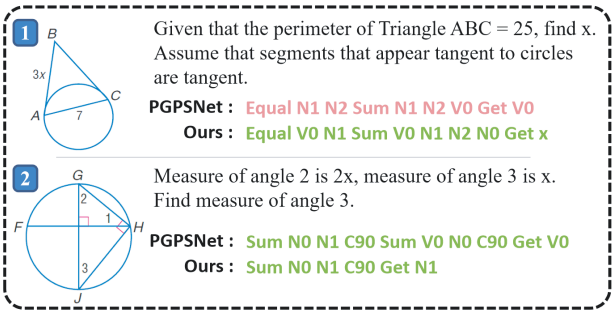


Figure 5: Examples of case study. The colored lines are solution programs generated by the models, where red means it leads to wrong answer and green refers to correct solutions.

Case Study. Two typical cases are shown in fig. 5 for further discussion. In case 1, PGPSNet predicts the same solving logic but with wrong operands, while G^{3}PS successfully selects related sides of the triangle and builds correct equations. In case 2, both models obtain the ground-truth answer. However, the mastery of the layout of diagram enables G^{3}PS to reason within one triangle and generate shorter solutions.

4 Conclusion

In this paper, we introduce a novel way of encoding geometric diagram by geo-graph to provide valuable spatial and structural information for subsequent problem reasoning. The method is developed based on characteristics of geometric diagrams, rather than adopting same approach for natural images. We will take a further step to integrate our method with symbolic-based models for improved theorem prediction and LLM-based models for feature extraction, as well as other educational reasoning task (Zhang et al. 2025b).

Acknowledgments

This work was supported by National Natural Science Foundation of China (No.62137002, 62293550, 62293553, 62293554, 62450005, 62437002, 62477036), "LENOVO-XJTU" Intelligent Industry Joint Laboratory Project, the Natural Science Basic Research Program of Shaanxi (2023-JC-YB-593), the Youth AI Talents Fund of China Association of Automation (Grant No._HBRC-JKYZD-2024-311), and the China Scholarship Council Program.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Chen, J.; Li, T.; Qin, J.; Lu, P.; Lin, L.; Chen, C.; and Liang, X. 2022. UniGeo: Unifying Geometry Logical Reasoning via Reformulating Mathematical Expression. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 3313–3323.

Chen, J.; Tang, J.; Qin, J.; Liang, X.; Liu, L.; Xing, E.; and Lin, L. 2021. GeoQA: A Geometric Question Answering Benchmark Towards Multimodal Numerical Reasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 513–523.

Cho, S.; Qin, Z.; Liu, Y.; Choi, Y.; Lee, S.; and Kim, D. 2025. GeoDANO: Geometric VLM with Domain Agnostic Vision Encoder. *arXiv preprint arXiv:2502.11360*.

Gao, J.; Pi, R.; Zhang, J.; Ye, J.; Zhong, W.; Wang, Y.; HONG, L.; Han, J.; Xu, H.; Li, Z.; et al. 2025. G-LLaVA: Solving Geometric Problem with Multi-Modal Large Language Model. In *The Thirteenth International Conference on Learning Representations*.

Li, Z.; Du, Y.; Liu, Y.; Zhang, Y.; Liu, Y.; Zhang, M.; and Cai, X. 2024a. Eagle: Elevating geometric reasoning through llm-empowered visual instruction tuning. *arXiv preprint arXiv:2408.11397*.

Li, Z.-Z.; Zhang, M.-L.; Yin, F.; and Liu, C.-L. 2024b. LANS: A Layout-Aware Neural Solver for Plane Geometry Problem. In *Findings of the Association for Computational Linguistics ACL 2024*, 2596–2608.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Liu, W.; Hu, H.; Zhou, J.; Ding, Y.; Li, J.; Zeng, J.; He, M.; Chen, Q.; Jiang, B.; Zhou, A.; et al. 2023. Mathematical Language Models: A Survey. *CoRR*.

Lu, P.; Gong, R.; Jiang, S.; Qiu, L.; Huang, S.; Liang, X.; and Zhu, S.-c. 2021. Inter-GPS: Interpretable Geometry Problem Solving with Formal Language and Symbolic Reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 6774–6786.

Ning, M.; Wang, Q.-F.; Huang, K.; and Huang, X. 2023. A Symbolic Characters Aware Model for Solving Geometry Problems. In *Proceedings of the 31st ACM International Conference on Multimedia*, 7767–7775.

Pan, Y.; Zhang, Z.; Hu, P.; Ma, J.; Du, J.; Zhang, J.; Liu, Q.; Gao, J.; and Ma, F. 2025. Enhancing the Geometric Problem-Solving Ability of Multimodal LLMs via Symbolic-Neural Integration. *arXiv preprint arXiv:2504.12773*.

Peng, S.; Fu, D.; Liang, Y.; Gao, L.; and Tang, Z. 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*, 13468–13480.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Trinh, T. H.; Wu, Y.; Le, Q. V.; He, H.; and Luong, T. 2024. Solving olympiad geometry without human demonstrations. *Nature*, 625(7995): 476–482.

Wang, Y.; Wang, Y.; Wang, D.; Peng, Z.; Guo, Q.; Tao, D.; and Wang, J. 2025. Geometryzero: Improving geometry solving for llm with group contrastive policy optimization. *arXiv preprint arXiv:2506.07160*.

Wu, W.; Zhang, L.; Liu, J.; Tang, X.; Wang, Y.; Wang, S.; and Wang, Q. 2024. 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*, 13828–13837.

Xia, R.; Li, M.; Ye, H.; Wu, W.; Zhou, H.; Yuan, J.; Peng, T.; Cai, X.; Yan, X.; Wang, B.; et al. 2025. GeoX: Geometric Problem Solving Through Unified Formalized Vision-Language Pre-training. In *The Thirteenth International Conference on Learning Representations*.

Ying, C.; Cai, T.; Luo, S.; Zheng, S.; Ke, G.; He, D.; Shen, Y.; and Liu, T.-Y. 2021. Do transformers really perform badly for graph representation? *Advances in neural information processing systems*, 34: 28877–28888.

Zhang, J.; Wang, Z.; Wang, Z.; Zhang, X.; Xu, F.; Lin, Q.; Mao, R.; Cambria, E.; and Liu, J. 2025a. Maps: A multi-agent framework based on big seven personality and socratic guidance for multimodal scientific problem solving. *arXiv preprint arXiv:2503.16905*.

Zhang, M.-L.; Yin, F.; Hao, Y.-H.; and Liu, C.-L. 2022. Plane Geometry Diagram Parsing. In Raedt, L. D., ed., *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, 1636–1643. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Zhang, M.-L.; Yin, F.; and Liu, C.-L. 2023. A multi-modal neural geometric solver with textual clauses parsed from diagram. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 3374–3382.

Zhang, X.; Dong, Y.; Wu, Y.; Huang, J.; Jia, C.; Fernando, B.; Shou, M. Z.; Zhang, L.; and Liu, J. 2025b. PhysReason: A Comprehensive Benchmark towards Physics-Based Reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 16593–16615. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-251-0.

Zhang, Z.; Cheng, J.-K.; Deng, J.; Tian, L.; Ma, J.; Qin, Z.; Zhang, X.; Zhu, N.; and Leng, T. 2025c. Diagram formalization enhanced multi-modal geometry problem solver. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Zhao, J.; Zhang, T.; Sun, J.; Tian, M.; and Huang, H. 2025. Pi-GPS: Enhancing Geometry Problem Solving by Unleashing the Power of Diagrammatic Information. *arXiv preprint arXiv:2503.05543*.