Scilight

Journal of Educational Technology and Innovation
https://jeti.thewsu.org/index.php/cieti

Article

Enhancing Geometric Diagram Parsing with Deep Textual Semantics

Xuhui Zhang [1], Yanli Wang [2], Pengpeng Jian [1,*] and Lei Wu [1]

[1] Information Engineering Institute, North China University of Water Resources and Electric Power, Zhengzhou 450046, China
[2] Institute of Marxism, Henan University of Economics and Law, Zhengzhou 450046, China
* Correspondence: jianpengpeng@ncwu.edu.cn; Tel.: +86-13140168760

How To Cite: Zhang, X., Wang, Y., Jian, P., & Wu, L. (2026). Enhancing Geometric Diagram Parsing with Deep Textual Semantics. *Journal of Educational Technology and Innovation*, *8*(2), 1–9. https://doi.org/10.61414/y5wfx754

Received: 7 November 2025
Revised: 18 March 2026
Accepted: 25 April 2026
Published: 30 June 2026

Abstract: A complete analysis of a geometric diagram hinges on interpreting both its fundamental primitives and the accompanying natural language text, yet existing models struggle to process the rich semantics within these descriptions, often leading to ambiguity and restricted reasoning. To address this, our work introduces a method that deeply integrates a Transformer-based text encoder within a sophisticated visual parsing architecture. Central to our approach is a novel Semantic-Guided Cross-Attention mechanism, which uses a global sentence representation as a semantic query to dynamically guide the model's focus toward the most relevant visual primitives based on the textual context. This end-to-end process generates context-aware visual features that are then processed by a Graph Neural Network (GNN) to perform robust cross-modal reasoning. Validated on the large-scale PGDP5K and IMP-Geometry3K datasets, our method demonstrates substantial accuracy improvements in relationship parsing and geometric proposition generation, especially in challenging cases involving text-diagram ambiguity, and significantly surpasses current state-of-the-art baselines by offering a more effective framework for fusing deep textual semantics with visual information.

Keywords: geometric problem solving; multimodal learning; geometric diagram parsing; cross-attention; DistilBERT

1. Introduction

Automatic Geometric Problem Solving (AGP) is a long-standing challenge in artificial intelligence with significant potential applications in intelligent education platforms, adaptive learning systems, and teaching-assistant tools. A typical geometry problem is composed of a geometric diagram and a problem text, which together define the problem. The first critical step towards solving such problems is Geometric Diagram Parsing, which aims to interpret the diagram by identifying its constituent geometric primitives (e.g., points, lines, circles) and the relationships between them.

Initial approaches to Automated Geometric Proofing (AGP) were built upon symbolic reasoning and expert systems. These methods functioned by using hand-engineered rules and logical axioms to represent geometric knowledge. Their core limitation, however, was a lack of flexibility. While these systems excelled at formal logical deduction, they were not well-suited to manage the ambiguity and noise present in actual diagrams and text. This inherent brittleness made them difficult to scale and constrained their use in practical scenarios.

The advent of computer vision and machine learning brought a paradigm shift. Subsequent methods began to integrate traditional image processing techniques, like the Hough transform, to detect basic primitives from diagram images. For instance, systems like GeoS (Seo et al., 2015) combined text parsing with diagram interpretation, but their reliance on rule-based or template-based parsing limited their ability to handle diverse and

complex natural language expressions. These methods often struggled to bridge the "semantic gap" between raw visual features and the high-level concepts described in the text.

Deep learning models, particularly Graph Neural Networks (GNNs), now represent the leading approach in this area. Models such as PGDPNet (Zhang et al., 2022) for instance, excel at parsing clean diagrams to extract primitives and their relationships. Their core weakness, however, is a reliance on purely visual information, which means they often fail to incorporate the full context provided in natural language descriptions. This limitation becomes critical when a diagram is ambiguous. For example, crucial information needed for a correct interpretation—such as a line being an angle bisector—may only be stated in the text. Models that ignore this textual information are therefore susceptible to making incorrect assumptions, which in turn compromises the validity of any subsequent reasoning. This disconnect between visual and textual understanding, often termed the semantic gap, remains a primary obstacle for advancing current AGP systems.

To address these limitations, this paper argues that a truly effective AGP system must combine robust visual understanding capabilities with equally powerful deep text comprehension. To do this, we propose a novel model that deeply integrates a Pre-trained Language Model (PLM) into the geometric diagram parsing pipeline. Our main contributions are as follows:

We design and implement a novel text encoding module that leverages a pre-trained language model (DistilBERT) for deep contextual semantic modeling of the problem text.

We propose a semantic-guided cross-attention mechanism. This mechanism transforms textual semantics into attention weights in the visual space, allowing the model to dynamically focus on key visual regions according to the text description, thereby achieving effective, end-to-end text-diagram alignment.

We demonstrate the effectiveness of our method through comprehensive experiments. Our model significantly outperforms strong baselines on two large-scale datasets, particularly in parsing text-dependent relationships and improving downstream task performance.

2. Methods

Our work is situated at the intersection of geometric diagram parsing, natural language understanding, and multimodal reasoning. We review relevant literature from these three perspectives.

2.1. Geometric Diagram Parsing

The interpretation of geometric diagrams is a foundational task in AGP. Early methods often relied on traditional computer vision techniques like the Hough transform to detect simple primitives. However, these methods were sensitive to image quality and struggled with complex, cluttered diagrams. The paradigm shifted with the advent of deep learning. Modern approaches reframe primitive detection as a standard object detection task, employing powerful convolutional neural network backbones such as Feature Pyramid Networks (FPN) (Lin et al., 2017). A state-of-the-art example in this area is PGDPNet (Zhang et al., 2022), which forms the visual basis of our work. PGDPNet pioneers a two-stage framework: first, it uses an object detection module (based on FCOS (Tian et al., 2019)) to identify geometric primitives (points, lines, circles). Second, it constructs a graph from these primitives and employs a Graph Neural Network (GNN) (Xu et al., 2017) to infer the relationships between them (e.g., incidence, parallelism). PGDPNet has demonstrated remarkable performance in parsing diagrams from a purely visual standpoint. However, its core limitation is that it does not incorporate the rich semantic information from the accompanying problem text. This makes it vulnerable to errors when diagrams are ambiguous or underspecified, a gap our work aims to fill.

2.2. Text Understanding in Geometry Problems

Parallel to diagram parsing, understanding the problem text is equally crucial. Early systems like GeoS (Seo et al., 2015) and InterGPS (Lu et al., 2021) utilized rule-based parsers, keyword matching, or formal language grammars to extract information from text. While effective for structured or simplified language, these methods lack the flexibility to handle the full diversity of natural language expressions.

The development of Pre-trained Language Models (PLMs) such as BERT (Devlin et al., 2019) and its variants like DistilBERT (Sanh et al., 2019), RoBERTa (Liu et al., 2019), and DeBERTa (He et al., 2020) has revolutionized natural language understanding. These models, pre-trained on vast text corpora, can capture deep contextual dependencies and nuances in meaning, moving far beyond surface-level pattern matching. Their powerful semantic representation capabilities offer a promising path to overcome the limitations of older text-parsing techniques in the geometry domain. Our work is among the first to deeply integrate a powerful PLM into a GNN-based parsing framework for this task.

2.3. Multimodal Fusion

To solve geometry problems effectively, information from both the diagram (vision) and the text (language) must be intelligently fused. This aligns with a broader trend in AI research on multimodal learning. In fields like Visual Question Answering (VQA) (Antol et al., 2015) and image-text retrieval, sophisticated fusion models like LXMERT (Tan & Bansal, 2019), UNITER (Chen et al., 2020), VL-BERT (Su et al., 2019), and ViLBERT (Lu et al., 2019) have achieved great success.

A key mechanism in these models is cross-modal attention (Vaswani et al., 2017), which allows elements from one modality (e.g., words in a question) to attend to and highlight relevant regions in another modality (e.g., objects in an image). This enables a fine-grained, dynamic alignment between text and vision. Inspired by these advances, our work designs a novel Semantic-Guided Cross-Attention mechanism tailored for the geometry domain. Unlike generic fusion, our method uses the high-level semantic "intent" extracted from the entire problem text to guide the visual parser's focus, enabling a more targeted and effective text-diagram integration.

3. Results

Our model addresses the semantic gap between visual diagrams and natural language by deeply integrating a text encoding module with a visual parsing framework. As illustrated in Figure 1, the architecture first processes the visual and textual inputs independently. These parallel streams then converge at a novel fusion stage, where the semantics from the text are used to guide the model's interpretation of the diagram. This process produces context-aware visual features, which are ultimately used by a graph-based reasoning network to determine the final geometric relationships.

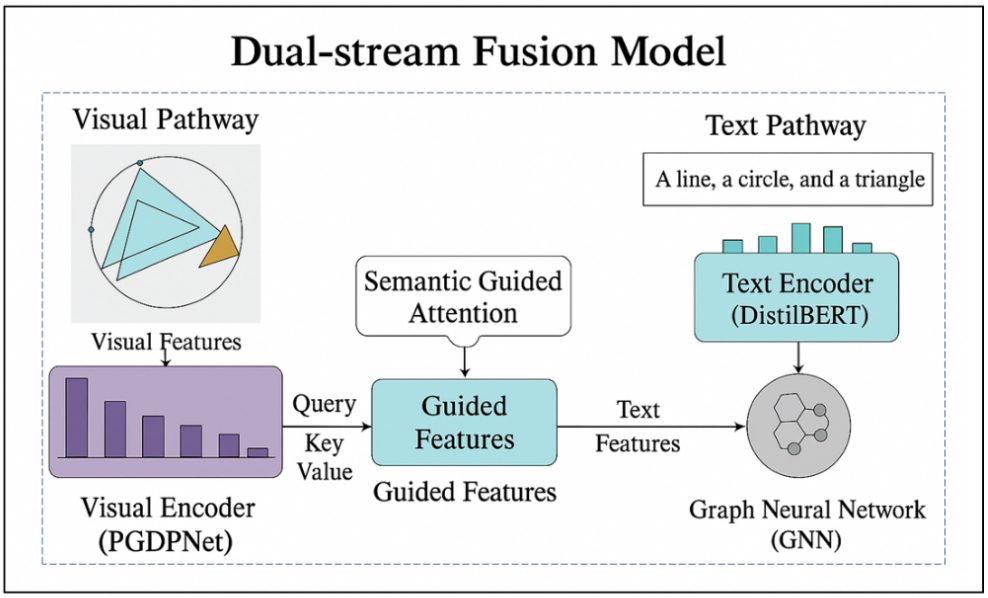

Figure 1. The overall architecture of our proposed dual-stream fusion model.

3.1. Overall Architecture

Our model operates on a diagram-text pair as input. Its architecture is composed of two primary streams: a Visual Pathway and a Textual Pathway. The Visual Pathway is responsible for identifying geometric primitives within the diagram and extracting their low-level visual features, leveraging the robust front-end of the PGDPNet (Zhang et al., 2022) model. Concurrently, the Textual Pathway processes the problem's natural language description using a pre-trained Transformer, DistilBERT (Sanh et al., 2019), to produce a single, high-level semantic vector that captures the global context and intent of the text.

The innovation of our model lies in how these two streams are merged. We introduce a Semantic-Guided Cross-Attention mechanism, which serves as the core fusion module. This module uses the textual semantic vector as a query to dynamically re-weight the visual features, effectively allowing the text to direct the model's attention to the most relevant primitives in the diagram. The resulting semantically-enriched visual features are then organized into a graph structure. Finally, a Graph Neural Network (GNN) reasons over this graph to model the complex interactions between primitives and predict their relationships, producing the final structured interpretation of the problem.

3.2. Modality-Specific Feature Encoding

Visual Pathway: Primitive Feature Extraction: The foundation of our visual understanding is built upon the powerful visual parsing capabilities of PGDPNet (Zhang et al., 2022). This pathway takes the diagram image as input and processes it through a convolutional backbone network fine-tuned for geometric object detection. This network identifies the locations and classes of various geometric primitives, such as points, lines, and circles. For each of the N primitives detected, a corresponding feature vector vi is extracted. This vector is a rich, low-level representation encoding essential visual attributes, including the primitive's class identity (e.g., 'point' vs. 'line'), precise coordinates, and other geometric properties like radius for circles or endpoints for line segments. The collective output is a set of N feature vectors, $V = \{v_1, v_2, \ldots, v_N\}$ is crucial to note that at this stage, these features are "context-agnostic"—they describe the visual appearance of the primitives but are unaware of the relationships and constraints specified in the accompanying text.

Textual Pathway: Semantic Context Encoding: To unlock the information contained in the problem text, our textual pathway aims to produce a representation that is not only comprehensive but also structured for guiding visual interpretation. We employ DistilBERT (Sanh et al., 2019), a distilled and efficient version of the BERT model. The input text is first tokenized into sub-word units using the WordPiece tokenizer and then passed through the multi-layer Transformer architecture of DistilBERT. The model's bidirectional self-attention mechanism allows it to build a deep contextual understanding, capturing complex syntactic structures and semantic nuances (e.g., distinguishing "perpendicular to" from "parallel to"). To distill this rich, token-level understanding into a single, actionable vector, we take the final hidden state corresponding to the special [CLS] token. This vector, denoted Ctext, serves as a global sentence embedding. It represents a holistic summary of the text's meaning, effectively capturing the core "intent" or "constraints" of the problem statement, which is ideal for guiding the subsequent fusion process.

3.3. Semantic-Guided Multimodal Fusion

The central challenge in multimodal geometric parsing is to effectively align textual semantics with visual elements. We address this with a Semantic-Guided Cross-Attention mechanism, which enables a dynamic, context-aware fusion of the two modalities. Unlike naive approaches like simple feature concatenation, which treat all visual elements equally, our mechanism allows the model to learn which primitives are most relevant to the given text.

The mechanism operates by casting the fusion problem into the standard attention framework (Vaswani et al., 2017). The global text semantic vector Ctext is designated as the Query, representing the high-level question or constraint being imposed on the diagram. The set of N visual primitive features V is used to derive both the Key and Value sets through learnable linear projections. This formulation elegantly captures the intuition of "semantic guidance": the text's intent (Query) "inspects" the visual content (Keys) to compute attention scores, which then determine how much "focus" to place on each primitive when constructing the new representation (by weighting the Values). The computation follows the scaled dot-product attention formula:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

where Q, K, and V are the linearly projected Query, Key, and Value matrices, and dk is the dimension of the key vectors, used for scaling. The output of this operation is a new set of visual features where each primitive's representation has been modulated by its relevance to the text. To retain the original visual information, this attention-weighted representation is fused with the initial visual features V (e.g., through element-wise addition or concatenation) to produce the final, context-aware features Vfused. This process empowers the model to first "read" the text and then intelligently "focus" on the specific parts of the diagram needed for accurate reasoning.

3.4. Relational Reasoning with a Graph Neural Network

Once the primitive features have been enriched with textual context to form Vfused, the final task is to infer the complex web of relationships between them. We model this problem using a graph-based approach. A fully connected graph is constructed where each of the N primitives corresponds to a node. Each node is initialized with its respective context-aware feature vector from Vfused. The goal is then to predict the class of each edge in this graph, which represents the relationship between the two connected nodes.

To perform this sophisticated relational reasoning, we employ a Graph Neural Network (GNN). Specifically, we adopt the Edge-Gated Graph Attention Network (EGAT) (Veličković et al., 2018) from the PGDPNet baseline, as it is well-suited for capturing nuanced relationships. The GNN operates through an iterative message-passing

process. In each iteration, every node aggregates feature information from its neighbors, and the EGAT mechanism uses an attention-like gating to control the flow of information along each edge. This iterative refinement allows the model to reason about higher-order dependencies (e.g., how the relationship between A and B might be influenced by C). After a fixed number of iterations, the final node embeddings, which now encode information about their local graph neighborhood, are used to make the final classification for each pairwise relationship. By feeding the GNN with semantically-rich features from the start, our model is far more capable of correctly identifying relationships that depend heavily or entirely on the textual description, thus leading to a more robust and accurate final parsing.

4. Experiments and Analysis

We evaluate the performance of our proposed method through a series of experiments. The experimental setup is based on two widely used datasets: PGDP5K (Seo et al., 2015) and a re-annotated version of IMP-Geometry3K (Seo et al., 2015). PGDP5K contains 5,000 samples with rich annotations of primitives and relationships. IMP-Geometry3K contains more complex textual descriptions, making it suitable for testing the model's generalization ability. Related geometric QA benchmarks include GeoQA (Chen et al., 2021) and MathQA (Amini et al., 2019). Our method is compared against two strong baselines: PGDPNet (Zhang et al., 2022) and InterGPS (Lu et al., 2021). We strictly follow the evaluation metrics defined in the PGDPNet paper, including F1 score and Full Relation Accuracy for relationship parsing, and Totally Same and Almost Same for geometric proposition generation. All models are implemented in PyTorch. The text encoding module uses the distilbert-base-uncased model from the Hugging Face library, and other hyperparameter settings are consistent with PGDPNet (Zhang et al., 2022).

As shown in Table 1, our method on the PGDP5K dataset significantly outperforms PGDPNet on all relation parsing metrics ($p < 0.05$). The performance improvement is particularly pronounced for relationships like Text2Geo and Sym2Geo, which are highly dependent on non-visual information. The Full Relation Accuracy increased from 81.5% to 83.0%, demonstrating an enhancement in the model's overall reasoning capability.

Table 1. Comparison of relationship parsing results on PGDP5K.

Model	Geo2Geo F1	Text2Geo F1	Sym2Geo F1	Full Rel. Acc.
PGDPNet (Baseline)	98.68%	97.60%	96.53%	81.5%
Our Method	98.86%	98.04%	97.13%	83.0%

As shown in Table 2, the advantages of our method become even more pronounced in downstream tasks. On PGDP5K, the Totally Same accuracy for geometric proposition generation improved by 6.5%. When the parsing results were fed into the solver of InterGPS (Lu et al., 2021), the final problem-solving accuracy also increased by 4.3%.

Table 2. Downstream task performance comparison (PGDP5K).

Task	PGDPNet	Our Method
Proposition Generation (Totally Same)	78.2%	84.7%
Problem Solving (Accuracy)	70.0%	74.3%

Additional analysis of our model focused on two areas: computational efficiency and the specific contribution of each architectural component. Regarding efficiency (Table 3), our choice of DistilBERT over the larger BERT-base model struck an effective balance, achieving a substantial performance boost while introducing only minimal overhead in parameter count and inference time. A series of ablation studies (Table 4) then confirmed the necessity of our design. These tests showed that removing the text module caused performance to drop to the baseline, while a simplified fusion method—injecting the global CLS vector without our attention mechanism—provided little to no advantage. This comparison clearly demonstrates that the proposed semantic-guided cross-attention is the crucial mechanism enabling the effective use of textual information.

Table 3. Model efficiency comparison.

Model Configuration	Parameters (M)	Inference Time (ms/Diagram)	Full Rel. Acc.
PGDPNet (Baseline)	12.4	35.2	81.5%
Our Method (w/DistilBERT)	78.6	42.8	83.0%
Our Method (w/BERT-base)	122.1	58.1	83.2%

Table 4. Ablation study results (PGDP5K).

Ablation Configuration	Text2Geo F1	Full Rel. Acc.
Full Model (Ours)	98.04%	83.0%
without DistilBERT (i.e., PGDPNet)	97.60%	81.5%
with CLS Injection only (no attention)	97.82%	82.1%
Replace with BERT-base	98.10%	83.2%

The training dynamics, visualized in Figure 2, reveal a key advantage of our method. Compared to the PGDPNet baseline, our model converges faster and reaches a lower final loss value. We attribute this superior learning behavior to the integration of textual semantics, which appears to provide a stronger and clearer supervisory signal early in the training process. This ultimately enables the model to learn more efficiently and achieve a more accurate fit to the data.

Figure 2. Comparison of training loss convergence curves.

A case study in Figure 3 illustrates our model's ability to resolve ambiguities that cause the baseline to fail. Visualizing the semantic-guided attention mechanism provides an intuitive explanation for this capability. For instance, when the text describes a "perpendicular" relationship, the model's attention correctly converges on the corresponding visual elements in the diagram. This not only validates the model's reasoning process but also makes its decisions more interpretable.

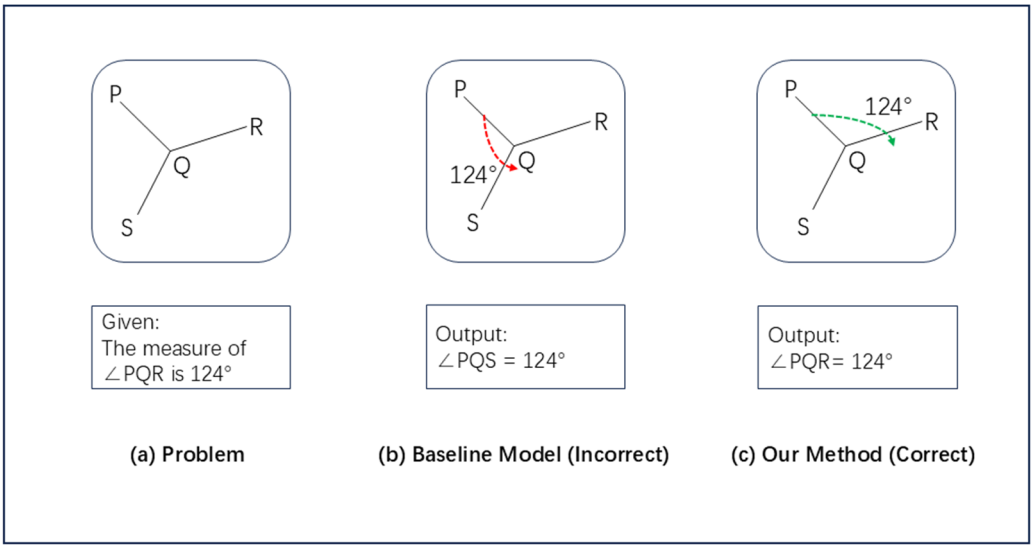

Figure 3. A case study comparing our method with the baseline.

5. Conclusions and Outlook

This paper addresses the deficiency of existing geometric diagram parsing models in deep text semantic understanding by proposing a novel method. By deeply integrating a Transformer-based text encoding module into an advanced visual parsing framework and designing a novel semantic-guided cross-attention mechanism for text-diagram fusion, our model significantly enhances the ability to parse primitive relationships in complex and ambiguous scenarios. Experimental results on two large-scale datasets strongly demonstrate the effectiveness of our method, which significantly surpasses strong baseline models across multiple levels, including relationship parsing, geometric proposition generation, and final problem-solving.

The core insight of this research is that future intelligent geometry solvers must be "dual-specialists" in both visual and language understanding; relying solely on either modality will inevitably hit a performance ceiling. Our work provides an effective and feasible framework for building such a multi-modal collaborative reasoning system. Its principles also offer valuable insights for other text-diagram reasoning tasks that require structured understanding, such as scientific literature chart analysis. Furthermore, the synergy between multimodal parsing and structured output aligns with the AI-Integrated Writing (AWAI) framework (Zhao, 2025), which provides a theoretical basis for rethinking authorship and knowledge synthesis in the era of generative AI.

Despite its notable success, our method has some limitations. First, the DistilBERT model we use is a general-purpose language model with limited adaptability to domain-specific terminology. We found that on a subset of data containing specialized terms like "foot of the perpendicular" and "tangent-chord angle", the model's relationship parsing F1 score dropped by 5.7%, highlighting the need for domain adaptation. Second, the current text-diagram alignment mechanism, while effective, is relatively simple. Future work could explore more sophisticated fusion strategies, such as introducing span-based matching mechanisms or leveraging prior knowledge from knowledge graphs. Finally, the complexity of text in existing datasets still has room for improvement. Based on this, future research directions include: developing domain-specific language models; exploring more sophisticated text-diagram alignment and interaction mechanisms; and constructing more open-ended and challenging multi-modal geometry problem datasets.

Author Contributions

Conceptualization: X.Z. and Y.W.; Methodology: X.Z. and P.J.; Software: X.Z. and L.W.; Validation: Y.W., P.J. and L.W.; Formal analysis: X.Z.; Investigation: P.J. and L.W.; Resources: P.J.; Data Curation: X.Z. and L.W.; Writing—Original Draft Preparation: X.Z.; Writing—Review & Editing: Y.W. and P.J.; Visualization: X.Z. and L.W.; Supervision: P.J. and Y.W.; Project Administration: P.J.; Funding Acquisition: P.J. All authors have read and agreed to the published version of the manuscript.

Funding

This research was supported by the General Project of Natural Science Foundation of Henan Province (262300421801) and Soft Science Project of Henan Province (No. 262400410529).

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

The data presented in this study are openly available. The PGDP5K and IMP-Geometry3K datasets used for validation can be accessed through their respective original publication sources.

Conflicts of Interest

The authors declare no conflict of interest.

Use of AI and AI-Assisted Technologies

No AI tools were utilized for this paper.

References

Amini, A., Gabriel, S., Lin, S., Koncel-Kedziorski, R., Choi, Y., & Hajishirzi, H. (2019, June 2–7). *MathQA: Towards interpretable math word problem solving with operation-based formalisms* [Conference session]. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT) (pp. 2957–2967), Minneapolis, MN, USA.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., & Parikh, D. (2015, December 7–13). *VQA: Visual question answering* [Conference session]. 2015 IEEE International Conference on Computer Vision (ICCV) (pp. 2425–2433), Santiago, Chile. https://doi.org/10.1109/ICCV.2015.279.

Chen, J., Tang, J., Qin, J., Xia, T., Leng, Z., Lin, X., Hao, Y., Wei, S., Ji, P., & Liang, X. (2021, August 1–6). *GeoQA: A geometric question answering benchmark towards multimodal numerical reasoning* [Conference session]. Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021 (pp. 4887–4898), Online.

Chen, Y.-C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., & Liu, J. (2020, August 23–28). *UNITER: Universal image-text representation learning* [Conference session]. 16th European Conference on Computer Vision (ECCV) (pp. 104–120), Glasgow, UK.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, June 2–7). BERT: Pre-training of deep bidirectional transformers for language understanding [Conference session]. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT) (pp. 4171–4186), Minneapolis, MN, USA. https://doi.org/10.18653/v1/N19-1423.

He, P., Liu, X., Gao, J., & Chen, W. (2020). *DeBERTa: Decoding-enhanced BERT with disentangled attention*. arXiv. https://doi.org/10.48550/arXiv.2006.03654

Lin, T.-Y., Dollár, P., Girshick, R., & He, K. (2017, July 21–26). *Feature pyramid networks for object detection* [Conference session]. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 2117–2125), Honolulu, HI, USA.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A robustly optimized BERT pretraining approach*. arXiv. https://doi.org/10.48550/arXiv.1907.11692

Lu, J., Batra, D., Parikh, D., & Lee, S. (2019, December 8–14). *ViLBERT: Pretraining for ground-and-vision language tasks* [Conference session]. Advances in Neural Information Processing Systems (NeurIPS 2019) (Vol. 32, pp. 1–12), Vancouver, BC, Canada.

Lu, P., Qiu, L., Chen, J., Xia, T., Zhao, Y., Zhang, W., Zhou, Y., & Wu, Y. N. (2021, August 1–6). Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning [Conference session]. 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP) (pp. 5946–5958), Virtual.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). *DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter*. arXiv. https://doi.org/10.48550/arXiv.1910.01108.

Seo, M. J., Hajishirzi, H., Farhadi, A., Etzioni, O., & Malcolm, C. (2015, September 17–21). *Solving geometry problems: Combining text and diagram interpretation* [Conference session]. 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1456–1466), Lisbon, Portugal.

Su, W., Lin, X., Wang, W., & Li, J. (2019). *VL-BERT: Pre-training of generic visual-linguistic representations*. arXiv. https://doi.org/10.48550/arXiv.1908.08530.

Tan, H., & Bansal, M. (2019, November 3–7). *LXMERT: Learning cross-modality encoder representations from transformers* [Conference session]. 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 5100–5111), Hong Kong, China.

Tian, Z., Shen, C., Chen, H., & He, T. (2019, October 27–November 2). *FCOS: Fully convolutional one-stage object detection* [Conference session]. 2019 IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 9627–9636), Seoul, Republic of Korea.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017, December 4–9). *Attention is all you need* [Conference session]. Advances in Neural Information Processing Systems (NIPS 2017) (Vol. 30, pp. 5998–6008), Long Beach, CA, USA.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018, April 30–May 3). *Graph attention networks* [Conference session]. 2018 Conference on Learning Representations (ICLR), Vancouver, BC, Canada.

Xu, D., Zhu, Y., Choy, C. B., & Li, F.-F. (2017, July 21–26). *Scene graph generation by iterative message passing* [Conference session]. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 5410–5419), Honolulu, HI, USA.

Zhang, M. L., Yin, F., Hao, Y. H., & Liu, C. L. (2022, July 23–29). *Plane geometry diagram parsing* [Conference session]. 31st International Joint Conference on Artificial Intelligence (IJCAI) (pp. 1636–1643), Vienna, Austria.

Zhao, C. (2025). Rethinking authorship in the age of AI: Reflections on the AI-integrated writing framework (AWAI). *Journal of Educational Technology and Innovation*, *7*(2), 25–38. https://doi.org/10.61414/h2a5bt21.