Pi-GPS: Enhancing Geometry Problem Solving by Unleashing the Power of Diagrammatic Information

Junbo Zhao, Ting Zhang, Jiayu Sun, Mi Tian, Hua Huang

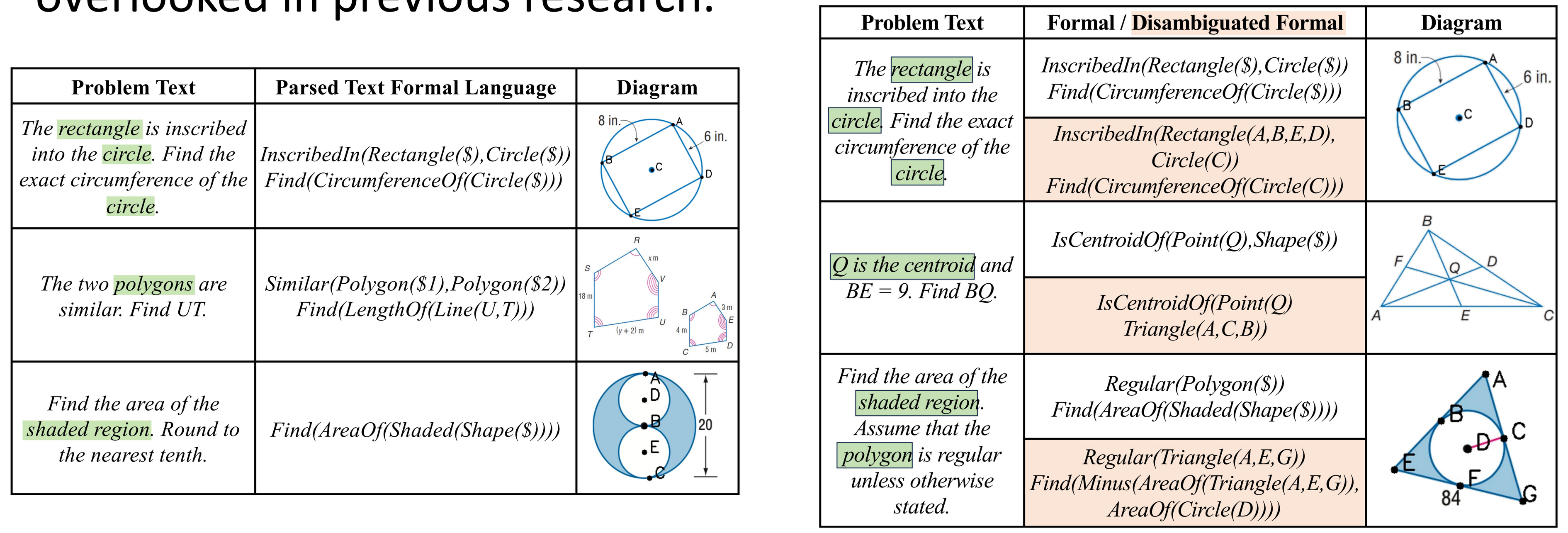
BEIJING NORMAL UNIVERSITY · TAL 好未来

ICCV OCT 19-23, 2025 HONOLULU HAWAII

Motivation & Background

Geometry Problem Solving (GPS) is critical for intelligent education, requiring the integration of visual diagrams and textual descriptions.

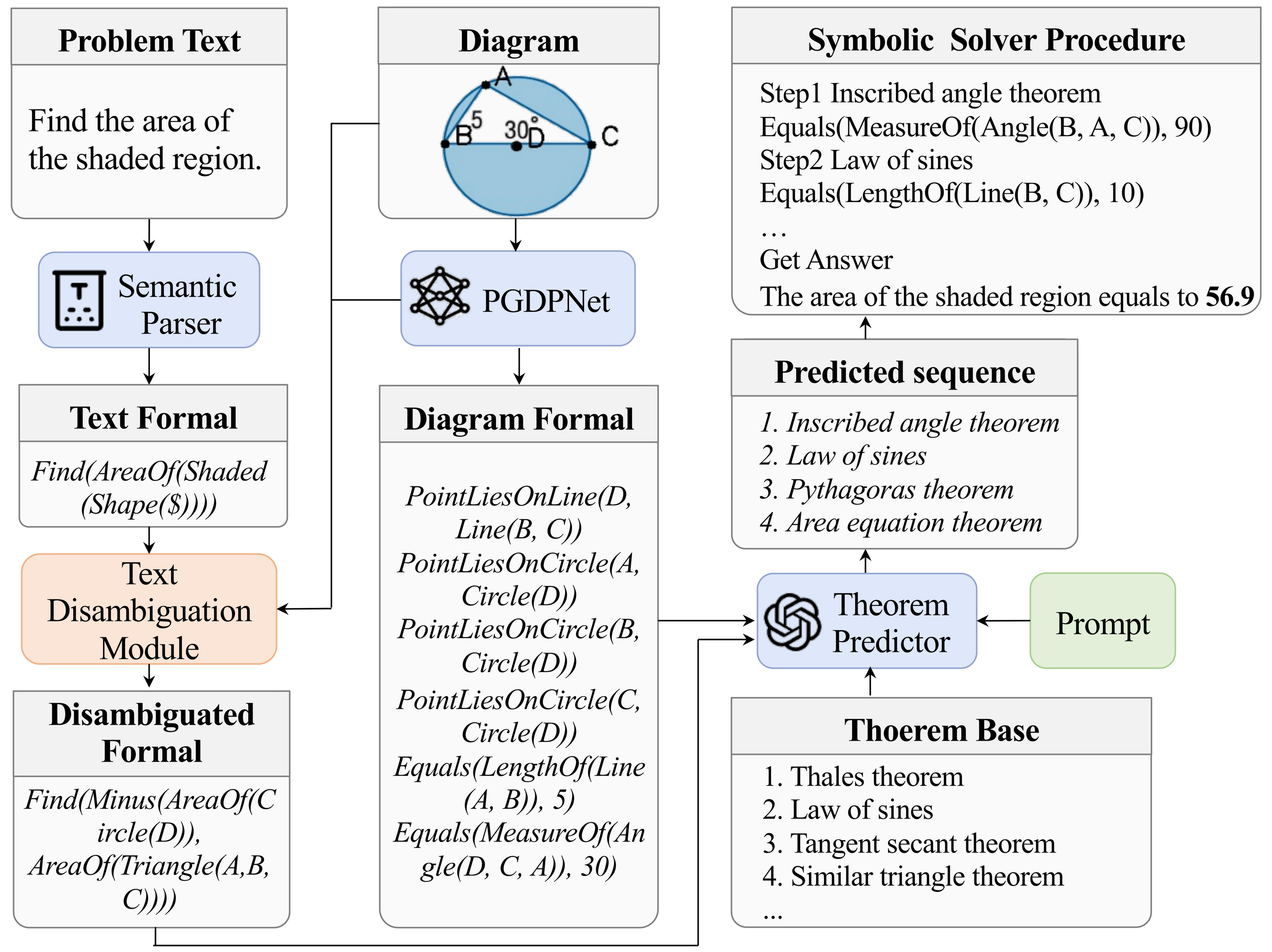
- **Key Challenge**: Textual ambiguities often hinder problem comprehension, while diagrams can effectively resolve these ambiguities.
- **Existing Approaches**: Symbolic and neural methods struggle with precise alignment between text and diagrams, leading to errors in reasoning.
- **Our Perspective**: Text ambiguity is a major bottleneck in GPS, yet largely overlooked in previous research.



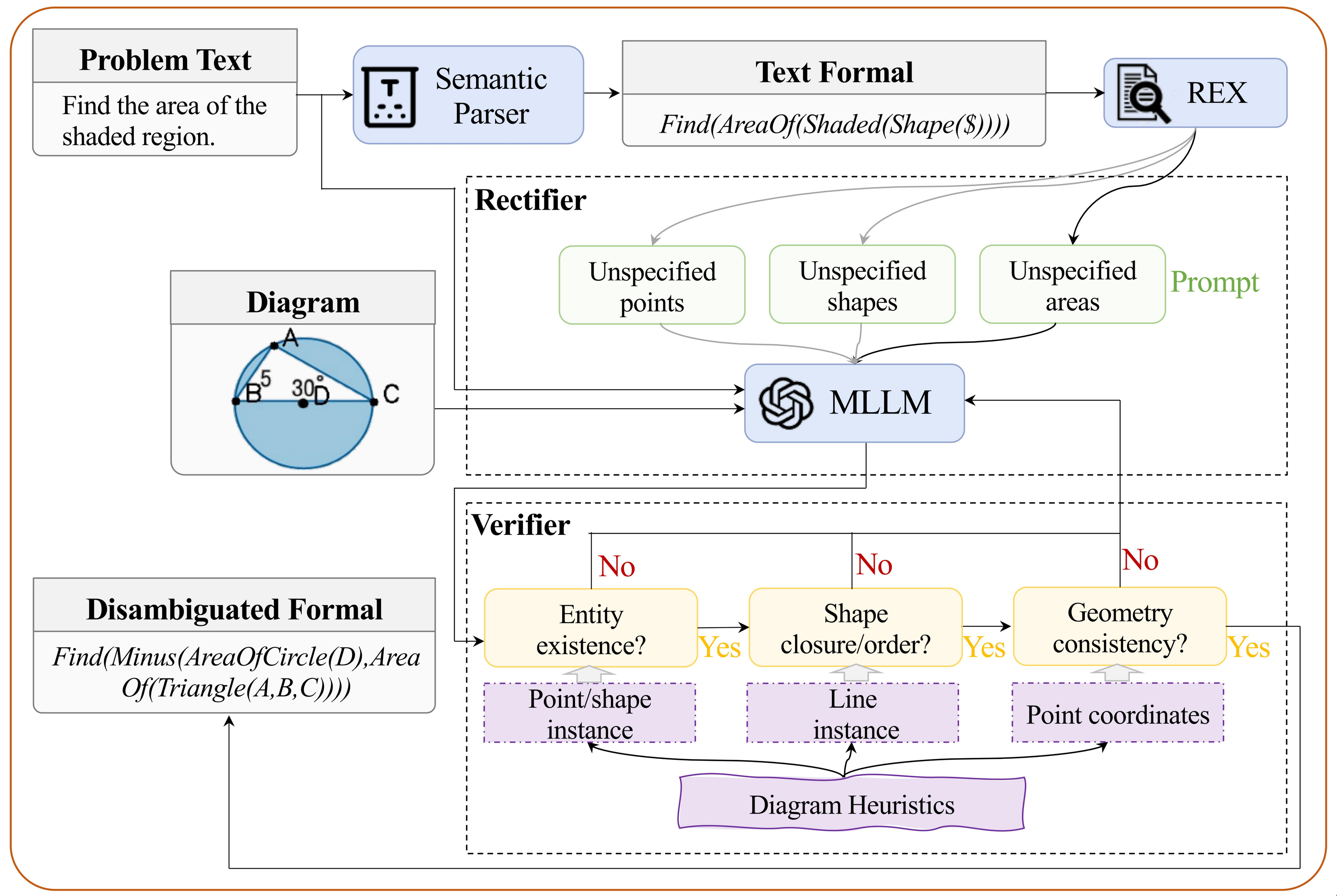
Method: The Pi-GPS Framework

- **Parser:**
  text (rule-based); diagram (PGDPNet).
- **Disambiguator:**
  rectifier (MLLM; three ambiguities); verifier (diagram heuristics match entities/structure; curb hallucinations).
- **Reasoner:**
  theorem predictor (LLM/o3-mini); symbolic solver (executes sequence; interpretable).

Pi-GPS framework



Text disambiguation module



Experiments & Results

Key Findings:

- Pi-GPS achieves a nearly 10% improvement over previous state-of-the-art methods.
- The text disambiguation module is critical, delivering consistent performance gains.
- The verifier ensures reliable alignment and prevents MLLM hallucination.

Method	Completion	Choice
Ours w/o Text disam.	63.2	72.3
+ Rectifier (general prompt)	62.4	71.9
+ Rectifier (specific prompt)	64.2	73.3
+ Verifier	**70.6**	**77.8**

Table 4. Illustrating the roles of the rectifier and verifier in the text disambiguation module on Geometry3K.

Task	Models	Completion	Choice
Direct solv. (MLLM)	GPT-4o	34.8	58.6
	Gemini 2	38.9	60.7
Direct solv. (LLM)	GPT-4o	36.5	59.7
	DeepSeek-R1	63.9	72.2
	o3-mini	66.4	75.5
	o3-mini w/o Text disam.	61.4	70.4
Theorem pred. (LLM)	o3-mini (ours)	**70.6**	**77.8**

Category	Method	Geometry3K Completion	Geometry3K Choice	PGPS9K Completion	PGPS9K Choice
MLLMs	Qwen-VL [7]	22.1	26.7	20.1	23.2
	GPT-4o [1]	34.8	58.6	33.3	51.0
	Claude 3.5 Sonnet [6]	32.0	56.4	27.6	45.9
	Gemini 2 [15]	38.9	60.7	38.2	56.8
Neural Methods	NGS [8]	35.3	58.8	34.1	46.1
	Geoformer [10]	36.8	59.3	35.6	47.3
	SCA-GPS [23]	-	76.7	-	-
	GOLD* [40]	-	62.7	-	60.6
	PGPSNet-v2-S* [43]	65.2	76.4	60.3	69.2
	LANS (Diagram GT)* [18]	72.1	82.3	66.7	74.0
Neural-symbolic Methods	Inter-GPS [21]	43.4	57.5	-	-
	GeoDRL [25]	57.9	68.4	55.6	66.7
	E-GPS [36]	-	67.9	-	-
	Pi-GPS (ours)	**70.6**	**77.8**	**61.4**	**69.8**

Table 2. Comparison on Geometry3K and PGPS9K. Our method achieves the best performance (highlighted in bold) compared to the neural-symbolic methods. Note that all baselines except LANS use parsed results, while LANS uses textual clauses and point positions from diagram annotations. * indicates that GOLD, PGPSNet and LANS are trained on the larger dataset, PGPS9K.

Conclusion & Limitation

- **Contribution**: Pi-GPS adds a rectifier–verifier that leverages diagram context to disambiguate text, boosting geometric problem-solving.
- **Impact**: Underscores the importance of ambiguity resolution in multimodal mathematical reasoning.
- **Limitation**: Still constrained by brittle text-to-formal parsing, weak detection of subtle diagram relations, and an incomplete theorem base.