GEOX: Geometric Problem Solving Through Unified Formalized Vision-Language Pretraining

ICLR 2025 (8,8,6,6)

Presenter: Haotian Liu

December 9, 2024

Outline

Outline

Background

Task

- Use Multi-modal Large Language Models to deal with automatic Geometry Problem Solving(GPS).

Limitations

- Differences between geometric diagram-symbol and natural image-text.
- Lack of automated verification in the problem-solving process.
- Current methods are limited by their task-specific designs, making them less effective for broader geometric problems.
- Using natural language to describe geometric diagrams introduces a significant amount of redundant information.

Related Work

Multi-modal Large Language Models

- G-LLaVA and MAVIS train LLM on the constructed geometry datasets with descriptions in natural language form. However, unable to provide the answer as required, and incorrect solving steps still result in correct answers.

Geometry Problem Solving

- Rule-based Methods: rely on external tools like OCR to parse diagrams into texts, which are then used for logical reasoning based on path search and condition matching.
- Neural approaches: use networks to predict solving steps via program sequences, which are then executed by the solver.

Contributions

Contributions

- Propose GeoX, aiming to build geometric generalist models by modeling geometric tasks into a unified formulation.
- Propose GS-Former, which effectively bridges the modality gap between geometric diagrams and formalized language.
- Compared with previous generalist and specialized models, our GeoX achieves competitive performance on different datasets

Outline

Method Overview

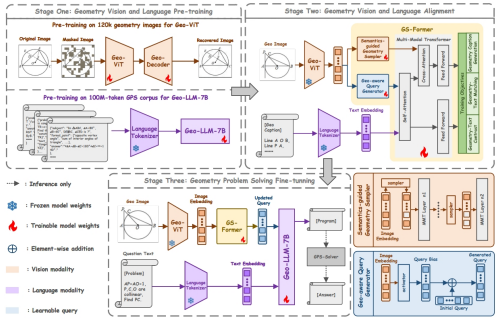
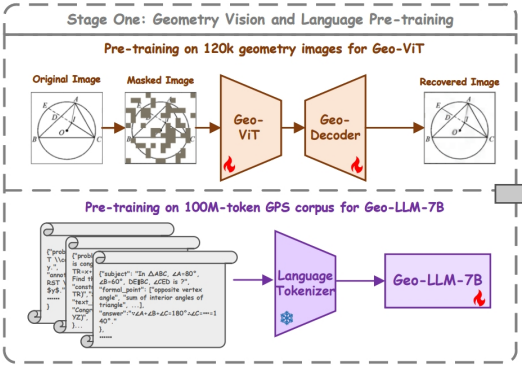


Figure 2: Overview of GeoX for training. We present a versatile method for automatic geometric problem solving through unified formalized vision-language pre-training, which comprises three progressive stages.

- **Unimodal Pre-training**: aiming to enhance the GeoX's ability to understand geometric diagrams and symbols.
- **Geometry-Language Alignment**: facilitate the pre-trained unimodal models for performing cross-modal alignment.
- **End-to-end Instruction Tuning**: boosting GeoX's capacity to comprehend geometric problems and generate formal solution programs.

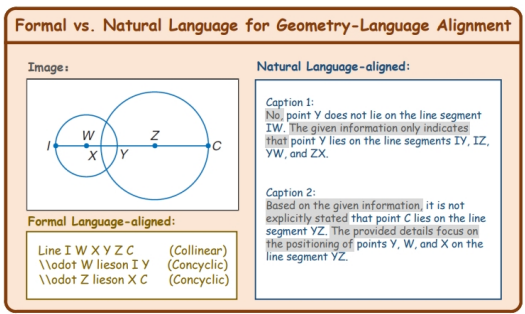
Unimodal Pre-training



Geometry Encoder: collect more than 120K diagrams, using the masked auto-encoding scheme to fine-tune on ViT and obtain Geo-ViT.

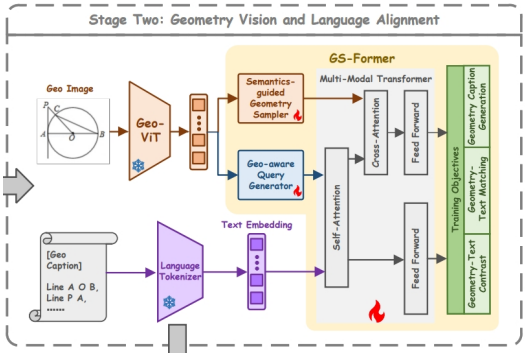
Symbol Decoder: build a 100M-token geometric corpus, use auto-regressive language modeling objective to fine-tune on the LLEMMA-7B and obtain Geo-LLM-7B.

Data Engine



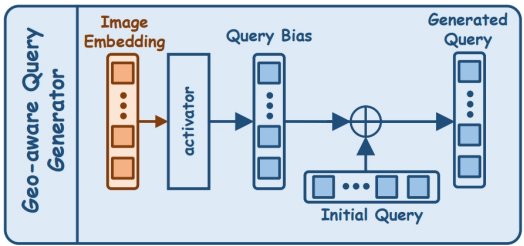
- Build a formalized diagram-caption dataset that delves into the spatial relationships at a granular level.
- Identify and describe the relative positions and connections between points. Two relationships: Collinear and Concyclic.
- This dataset contains 6232 geometric images annotated by experts.

Generator-and-Sampler Transformer



With the formalized geometry-language dataset, GeoX learns a unified representation space for geometry and formalized language through the Generator-and-Sampler Transformer.

Geo-aware Query Generator.



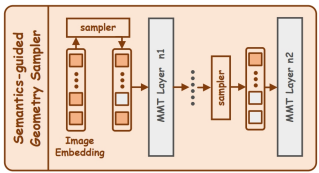
- Previous methods extract visual features using a set of static query tokens.
- GQG leverages visual features from the encoder, aggregates context using attention and pooling, and projects them onto learnable queries.

Semantics-guided Geometry Sampler.

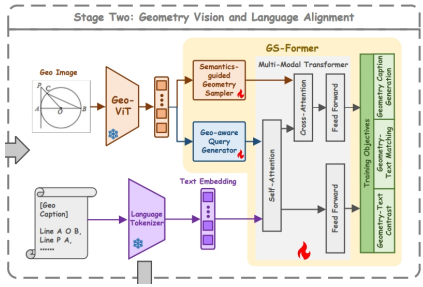
- SGS is tasked with predicting a binary mask $M = \{m_i^j | i \in K, j \in N\}$, with each $m_j^i \in \{0, 1\}$ determining whether to retain or discard representations.
- This module receives the previous mask M^{i-1} and visual features as inputs, using a linear layer to obtain retention probabilities p^i.
- To enable differentiable sampling from probabilities, use the parameterization with Gumbel-Softmax:

$$M^i = M^{i-1} \odot \text{Gumbel-Softmax}(P^i) \tag{1}$$

$\odot$ is the Hadamard product, i and $i-1$ represents the previous stage and current stage.



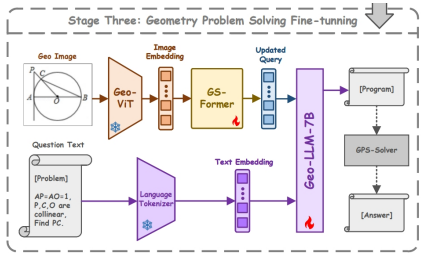
Multi-Modal Transformer.



- Following BLIP-2, introduce a multimodal alignment loss L_{align} to optimize GS-Former, incorporating three training objective. Further impose a sparsification term L_{spr} to prevent trivial solutions where all mask values m_j^i are set to 1:

$$\mathcal{L}_p = \mathcal{L}_{\text{align}} + \lambda \mathcal{L}_{\text{spr}}\,, \text{ where } \mathcal{L}_{spr} = \frac{1}{KN} \sum_{i \in K, j \in N} \left\| m_j^i \right\|_1 . \tag{2}$$

End-to-End Visual Instruction Tuning.



- Diagrams are processed by Geo-ViT and GS-Former to extract visual tokens T_g, which are projected into the language space. Geo-LLM combines T_g and instruction tokens T_p to generate solutions in an auto-regressive manner, optimized using cross-entropy loss.

$$\mathcal{L}_t = -\sum_{l} \log P\left(s_l \mid s_{i,i\in[1:l-1]}; T_g; T_p\right). \tag{3}$$

Outline

Datasets

- **GeoQA:** 4998 real-world geometry problems with detailed solutions.
- **Geometry3K:** 3002 detailed geometry problems, less than 1% can be solved without diagrams.
- **GeoQA+:** enhances GeoQA by adding 2,518 newly annotated geometric problems.
- **UniGeo:** including 9543 proving problems and 4998 calculation problems from GeoQA.
- **PGDP5K:** 5000 images with detailed annotations including geometric primitives, symbols, etc.
- **PGPS9K:** 9022 problems with detailed annotations for diagrams and solution programs.
- **G-LLaVA:** over 110k question-answer pairs created by the GPT-API.

Metrics

- **GeoQA and UniGeo:** use top-1 and top-10 accuracies.
- **Geometry3K and PGPS9K:** use the metrics of Completion, Choice, and Top-3.
- To evaluate MLLMs in solving complex geometry problems, utilizing Completion (which requires models to provide answers directly) and Choice (which involves selecting from given options)

Overall Performance

Table 1: Comparison of various methods on the GeoQA benchmark with different accuracy metrics.

Methods	Metric	Total	Angle	Length
Generalists				
mPLUG-Owl2 (Ye et al., 2023)		16.0	16.5	15.9
LLaVA-v1.5 (Liu et al., 2024)		20.7	20.9	19.8
Qwen-VL (Bai et al., 2023)		24.4	23.7	24.4
GPT-4V (OpenAI, 2023)		43.4	39.3	49.8
Specialists	Top-1			
LLaVA-v1.5 (Liu et al., 2024)+**Solver**		9.4	14.9	3.2
NGS(Chen et al., 2021)		46.3	-	-
UniMath-T5(Liang et al., 2023)		49.6	-	-
UniMath-Flan-T5(Liang et al., 2023)		50.0	-	-
GeoX (Ours)		**54.9**	**62.8**	**45.2**

Methods	Metric	Total	Angle	Length
Specialists				
LLaVA-v1.5 (Liu et al., 2024)+**Solver**		29.2	40.5	15.9
FiLM(Perez et al., 2018)		31.7	34.0	29.7
RN(Santoro et al., 2017)		38.0	42.8	32.5
MCAN(Yu et al., 2019)		39.7	45.0	34.6
BERT (Kenton & Toutanova, 2019)	Top-10	54.7	65.8	42.1
NGS(Chen et al., 2021)		56.9	69.8	39.2
Geoformer(Chen et al., 2022)		60.3	71.5	49.1
DPE-NGS(Cao & Xiao, 2022)		62.7	74.9	47.7
SCA-GPS(Ning et al., 2023)		64.1	74.9	50.1
GeoX (Ours)		**69.0**	**78.2**	**58.0**

Table 2: Comparison of model performance on UniGeo for geometry calculation and proof problems.

Methods	Metric	Calculation(%)			Proving (%)					
		All ↑	*Angle* ↑	*Length* ↑	*All* ↑	*Par.* ↑	*Tri.* ↑	*Qua.* ↑	*Con.* ↑	*Sim.* ↑
Generalists										
mPLUG-Owl2 (Ye et al., 2023)		18.7	18.7	19.1	-	-	-	-	-	-
LLaVA-v1.5 (Liu et al., 2024)		24.0	26.4	21.6	-	-	-	-	-	-
Qwen-VL (Bai et al., 2023)		24.4	24.2	25.4	-	-	-	-	-	-
GPT-4V (OpenAI, 2023)		47.9	45.8	51.6	-	-	-	-	-	-
Specialists	Top-1									
LLaVA-v1.5 (Liu et al., 2024)+**Solver**		16.1	19.2	13.1	1.0	0.0	1.1	0.4	0.2	3.0
Geoformer (Chen et al., 2022)		46.8	57.8	35.0	51.3	13.9	63.8	20.4	56.1	64.0
UniMath-T5-base (Liang et al., 2023)		-	-	-	82.9	-	-	-	-	-
UniMath-Flan-T5-base (Liang et al., 2023)		-	-	-	83.0	-	-	-	-	-
GeoX (Ours)		**54.4**	**63.1**	**43.1**	**97.8**	**77.8**	**100.0**	**95.4**	**99.5**	**99.2**
Specialists										
LLaVA-v1.5 (Liu et al., 2024)+**Solver**		43.0	51.3	35.3	11.3	0.0	16.2	5.0	2.9	27.5
BERT (Kenton & Toutanova, 2019)		52.0	63.1	39.2	48.1	15.4	48.0	31.7	49.5	75.1
NGS (Chen et al., 2021)	Top-10	51.9	63.6	38.8	47.4	11.2	46.9	31.3	48.3	77.6
Geoformer (Chen et al., 2022)		62.5	75.5	48.8	56.4	19.4	69.4	20.4	60.3	75.0
GeoX (Ours)		**68.6**	**76.7**	**58.3**	**99.5**	**97.2**	**100.0**	**97.7**	**100.0**	**100.0**

GeoX outperforms both Generalist Models and Specialist Models across different datasets.

Insightful Analyses

Effectiveness of Uni-modal Pre-training

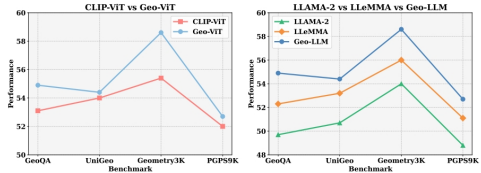


Figure 3: Effectiveness of Uni-modal Pre-training. We compare the widely used CLIP-ViT-B and our Geo-ViT-B, along with three LLM models: LLAMA-2-7B, LLEMMA-7B, and our Geo-LLM-7B.

Compared to general-purpose models or mathematical models, the pre-trained model demonstrates superior results across various geometry benchmarks.

Insightful Analyses

Effectiveness of Geometry-Language Alignment

Table 5: Effectiveness of geometry-language alignment.

Module	Alignment	Language	Geometry3K			PGPS9K		
			Completion ↑	*Choice* ↑	*Top* − 3 ↑	*Completion* ↑	*Choice* ↑	*Top* − 3 ↑
-	×	-	33.1	54.0	48.2	31.5	43.6	50.1
GS-Former	×	-	48.6	65.7	63.2	42.7	54.3	56.8
	✓	Natural	55.7	71.5	67.2	52.2	62.2	**67.1**
	✓	Formal	**58.6**	**72.5**	**69.4**	**52.7**	**63.3**	65.4

The introduction of GS-Former significantly boosts performance. Formal language is more effective for GPS than natural language.

Ablation of Modules in GS-Former

Table 6: Ablation study of modules in GS-Former, assessing the contribution of GQG and SGS modules when GS-Former is utilised for geometry-formal language alignment.

Geo-aware Query Generator	Semantics-guided Geometry Sampler	Geometry3K			PGPS9K		
		Completion ↑	*Choice* ↑	*Top* − 3 ↑	*Completion* ↑	*Choice* ↑	*Top* − 3 ↑
×	×	55.0	70.3	68.3	49.8	59.9	64.6
✓	×	57.4	71.7	68.1	50.8	62.0	64.3
✓	✓	**58.6**	**72.5**	**69.4**	**52.7**	**63.3**	**65.4**

Case Study

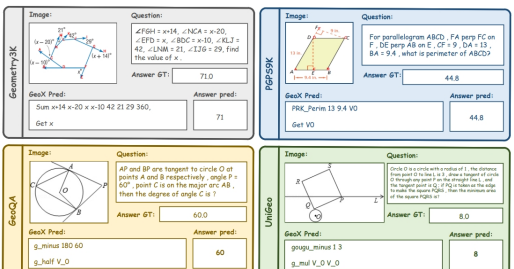


Figure 4: Visualization results on four datasets by our GeoX.

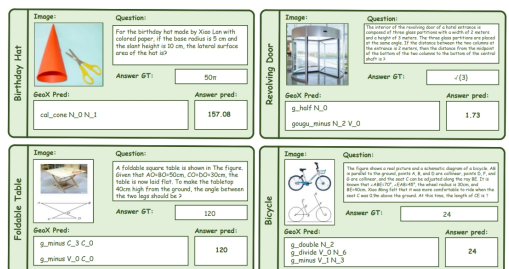


Figure 5: Four visualized examples of geometric problem in natural images solved by our GeoX.

The model can also deal with the natural images.

Outline

Conclusion

- This paper introduces GeoX, a novel multi-modal large model specifically designed for automatic Geometry Problem Solving (GPS) tasks.
- **Advantage:**
  1. It verifies that formalized vision-language learning is beneficial for learning informative representations for automatic GPS tasks.
  2. It can produce formalized process descriptions, which enhance the interpretability of GPS and the correctness of the solution process.
- **Disadvantage:** The entire procedure appears to be quite resource-intensive.

Thank you!