LATENTEUCLID: DATASET PROPOSAL AND ANALYSIS

Anshul Sawant * **Canchen Li*** **Joanna Smolska-Moreno*** **Shilin Ma*** **Steven Liu***
{anshulsa, canchenl, jsmolska, shilinm, stevenl5}@andrew.cmu.edu

1 [4 POINTS] PROBLEM DEFINITION AND DATASET CHOICE

The problem we study is the brittleness of language-centric reasoning in multimodal geometric tasks, where correct solutions depend on maintaining precise spatial structure throughout multi-step inference. Empirical analysis on the GeoThoughts benchmark Shi et al. (2025) shows that even state-of-the-art multimodal large language models (MLLMs), when provided with both diagrams and text, frequently form internally inconsistent geometric representations. These inconsistencies lead to systematic errors despite the models producing fluent and logically coherent language-based reasoning.

We hypothesize that this limitation is not inherent to language-based reasoning itself, but arises from the dominance of token-level Chain-of-Thought reasoning during inference. In such settings, visual information may be weakly grounded and fail to sufficiently constrain intermediate reasoning states, allowing linguistically coherent but geometrically incorrect interpretations to persist. Our objective is therefore to investigate whether strengthening the role of latent visual or multimodal representations in the reasoning process—either by reasoning directly in latent space or by more tightly integrating latent representations into Chain-of-Thought reasoning—can better preserve spatial structure and mitigate these failure modes.

To study this problem, we adopt the **GeoThoughts** dataset Shi et al. (2025). GeoThoughts pairs geometric diagrams with natural language questions and reference reasoning annotations grounded in valid geometric constraints. The dataset is particularly suitable because it highlights errors caused by incorrect or unstable spatial representations rather than missing visual input or symbolic knowledge, making it an appropriate testbed for evaluating latent-space reasoning approaches in multimodal geometric problem solving.

1.1 [0.5 POINTS] WHAT PHENOMENA OR TASK DOES THIS DATASET HELP ADDRESS?

GeoThoughts addresses the task of multimodal geometric reasoning, where correct problem solving requires inferring and manipulating spatial relationships encoded in diagrams under textual constraints. Unlike standard visual question answering or mathematical word problems, success in this setting depends on maintaining a coherent internal representation of geometric structure—such as relative orientation, parallelism, and adjacency—across multiple reasoning steps.

Crucially, the dataset enables the study of reasoning *behavior* rather than only final outcomes. The presence of reference reasoning annotations allows model outputs to be evaluated with respect to the validity of underlying geometric relationships, making it possible to analyze where reasoning diverges from structurally correct interpretations. This supports investigation into how errors emerge during reasoning, particularly in cases where logically consistent language-based reasoning is built on an incorrect spatial representation.

1.2 [0.5 POINTS] WHAT ABOUT THIS TASK IS FUNDAMENTALLY MULTIMODAL?

This task is fundamentally multimodal because the information required for correct reasoning is distributed across visual and textual modalities in a complementary, non-redundant manner. Textual descriptions specify symbolic constraints (e.g., angle measures or parallel relations), while diagrams encode spatial structure such as relative position, orientation, and topological relationships that are not fully recoverable from language alone.

* Everyone Contributed Equally – Alphabetical order by First Name

Beyond input modality, the challenge is inherently representational and temporal. Solving geometric problems requires maintaining and updating a grounded spatial representation throughout a multi-step reasoning process. Empirical results on GeoThoughts show that even when visual inputs are available, models often fail to preserve correct spatial structure over extended chains of reasoning, suggesting that the difficulty lies not in perception, but in cross-modal representation maintenance.

As a result, geometric reasoning serves as a strong testbed for evaluating whether multimodal models perform genuine cross-modal reasoning, as opposed to relying on language-dominant heuristics with weak visual grounding. This aligns closely with the course's emphasis on representation alignment and the distinction between accessing multimodal information and reasoning over it in the latent space.

1.3 HYPOTHESES

[1 points] Hypothesis For a subset of geometric problems in GeoThoughts, textual descriptions alone are sufficient to reconstruct the underlying geometry with low ambiguity. We hypothesize that removing those cases from the data set would lead to even bigger improvement on related problems when finetuning with GeoThought [Anshul: IIRC, Monet already does this for the datasets it considers].

[1 points] Hypothesis Encoding geometric diagrams as structured graphs with node and edge attributes will lead to more consistent geometric reasoning than pixel-based image encodings, particularly for problems requiring preservation of spatial relations across multiple reasoning steps. We plan to test this hypothesis by either manually converting the images into a set of vertices and edges or through prompt engineering that encourages the model to look at the diagrams first. [Anshul: Another possibility is to use a frontier model to generate these graphs alongside the images. Yet another possibility is some affine transformations of the geometric figures (reflection, rotation etc) that maintain the semantics.]

[1 points] Hypothesis Geometric reasoning can be carried out directly in latent visual or multimodal representation spaces without explicit token-level Chain-of-Thought. Inspired by latent-native reasoning paradigms such as VL-JEPA Chen et al. (2025), as well as approaches that integrate latent reasoning within LLM-based architectures, such as Monet Wang et al. (2025b), we hypothesize that latent-space inference can better preserve geometric consistency than language-mediated reasoning, particularly for problems requiring stable spatial representations [Anshul: won't this imply that we need no language reasoning to solve the problem. Is that true for most of the problems?].

2 [6 POINTS] DATASET ANALYSIS

2.1 [1 POINTS] DATASET PROPERTIES

We will use the GeoThoughts benchmark as our dataset, which is composed of $17,077$ problems that evaluate the mathematical reasoning capabilities of vision-language models in the subject of geometry. Each entry in our dataset contains a problem string that outlines the geometry question in text, a corresponding image to the question posed, and a solution string that gives a sample reasoning chain that derives the final answer.

The total size of the GeoThoughts dataset is 120 MB.

2.2 [0.5 POINTS] COMPUTE REQUIREMENTS

1. Files (can fit in RAM?)
   The GeoThoughts dataset is publicly available on Hugging Face (GeoThoughts). The dataset has an on-disk size of approximately 120 MB. While the in-memory footprint may increase due to image decoding and text processing, the dataset is typically accessed in a batched or streaming manner. Under these assumptions, RAM is not expected to be a limiting factor.
2. Models (can fit on GCP/AWS GPUs?)
   We plan to experiment with representative multimodal large language models at a similar scale to those evaluated in the GeoThoughts benchmark. In particular, we consider models in the 7B parameter range that support joint vision–language reasoning, such as Qwen2.5-VL-7B-Instruct Bai et al. (2025), Qwen3-VL-8B and Monet-7B Wang et al. (2025b), or comparable alternatives. Models of this scale are expected to be feasible to host on standard AWS GPU instances.

2.3 [2 POINTS] MODALITY ANALYSIS

1. **Question Type** To better characterize the scope and diversity of the dataset, we analyzed the distribution of question types. The findings indicate a non-uniform distribution, with a predominant focus on problems involving angle and length calculations.

Question Type Distribution in Geo-Thought-10K Dataset						
Type	**Angle**	**Length**	**Area**	**Similarity**	**Coordinate**	**Other**
Percentage (%)	63.5	18.3	9.5	5.0	3.4	0.2

Table 13. Question Type Distribution in Geo-Thought-10K Dataset

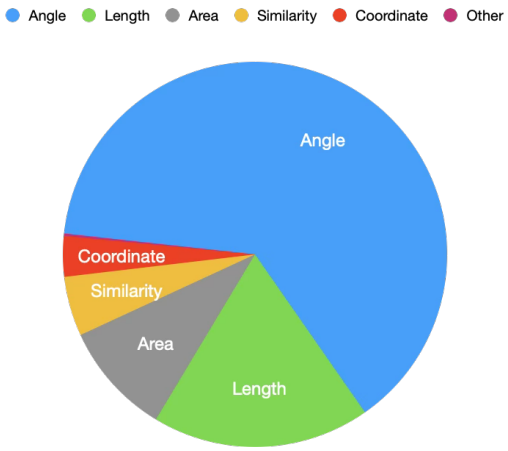

Figure 1: Data from Shi et al. (2025)

2. **Lexical diversity and Sentence Length** We have quantified the lexical diversity and word count distributions across the dataset's question, reasoning (thinking), and answer components, as detailed in fig. 2. Analysis of the Type/Token Ratio (TTR) reveals low to moderate lexical diversity across samples. Word count distributions indicate that questions are typically concise (under 100 words) and answers follow a normal distribution centered around 300 words. Notably, the reasoning segments exhibit significantly greater length; this discrepancy underscores a potential opportunity for latent space reasoning to optimize the traditionally verbose Chain-of-Thought process. Finally, an analysis of the top 50 non-stop words confirms a high density of geometric terminology, with a primary concentration on angle-related concepts.

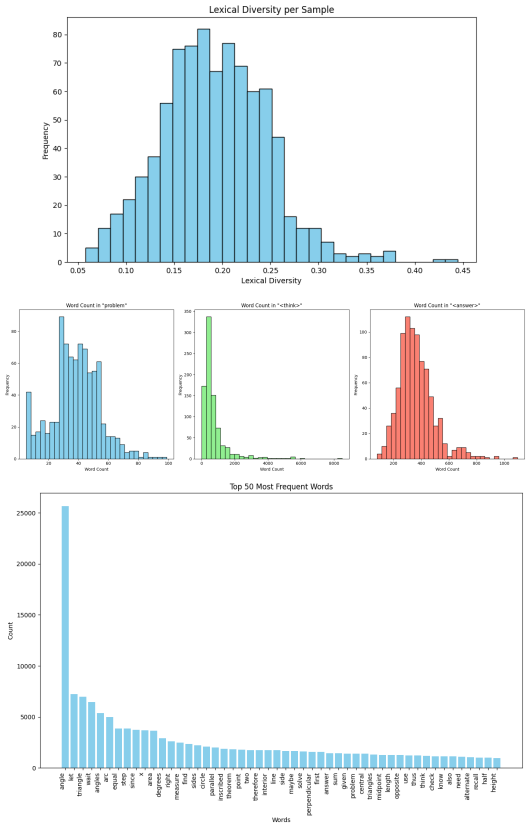

Figure 2: Based on 5% sample of the GeoThought dataset

3. **Visualization Related CoT** To quantify the extent of diagrammatic integration within the reasoning trajectories, we identified a set of keywords indicative of visual grounding: *image, diagram, figure, plot, chart, visualize, graph, sketch, draw, drawing, illustration, and picture*. Their occurrence frequencies, detailed in fig. 3, suggest that visual information may be underutilized in the current Chain-of-Thought process. These results establish a preliminary metric for data filtering, allowing us to prioritize samples with more explicit cross-modal engagement.

4. **Image Repetition** To ensure dataset integrity, we retained only those generated questions that yielded consistent answers. We analyzed the frequency of diagram repetition to investigate the impact of this filtering process on the visual distribution, with results detailed in fig. 4. The distribution shows a modal frequency of two, with a maximum of 20 occurrences per image. Notably, repetition detection was conducted via image hashing; this approach identifies identical pixel-level representations while treating diagrams with differing vertex labels as distinct samples.

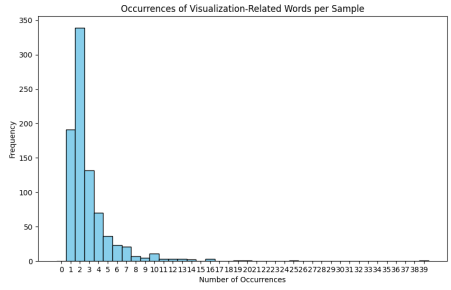

Figure 3: Based on 5% sample of the GeoThought dataset

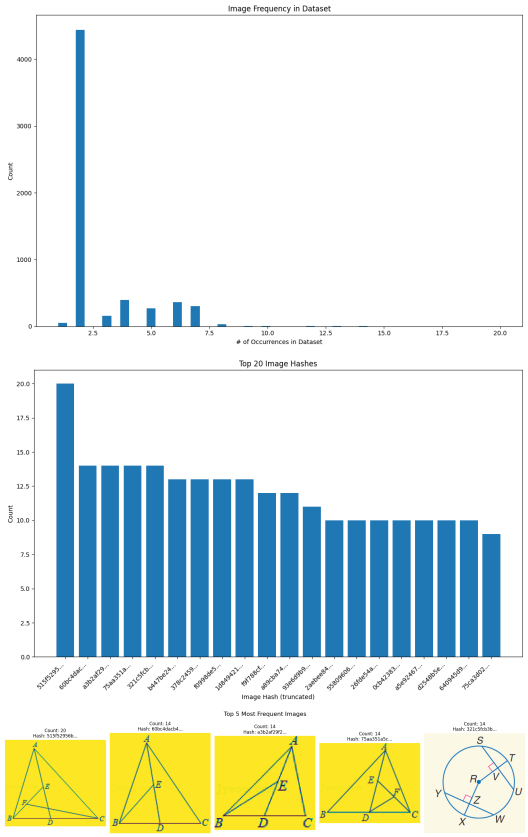

Figure 4: Based on all images in the GeoThought10K dataset

2.4 [0.5 POINTS] METRICS USED

- **Accuracy**: To evaluate the accuracy of model predictions on the dataset, our primary evaluation metric will be accuracy. Due to the open-ended nature of our GeoThoughts dataset, it may be difficult to meaningfully capture the degree of correctness for any given final generated output. Thus, we will evaluate whether our model successfully predicts the exact final answer in the solution.

- **ReasonEval** Xia et al. (2025): Due to the complexity of outputs when generating reasoning steps for that geometric tasks, we will use an LLM-judge to evaluate the quality of our outputs. The ReasonEval evaluation framework produces scores for the validity and redundancy of our chain-of-thought reasoning.

- **Self-Consistency**: We would expect that a model that is trained to properly reason across tasks in a specialized domain will consistently arrive at the same conclusion. Thus, we would like to examine how often our model will correctly respond to each geometric task across multiple generations.

2.5 [2 POINTS] BASELINES

Given that the GeoThoughts dataset was released in October 2025, there are currently no external works utilizing it. However, the authors evaluated their model against two primary benchmarks, which we also adopt for our analysis.

- **Geometry3K** Lu et al. (2021): Geometry3K is a large-scale benchmark containing 3,002 SAT-style multi-choice geometry problems collected from high-school textbooks for grades 6 through 12. The dataset covers a diverse range of geometric shapes and is densely annotated with unified structural descriptions in a formal language to support symbolic reasoning. It is particularly challenging for multimodal models because less than 1% of the problems are solvable without the associated diagram, emphasizing the necessity of joint textual and visual comprehension.
  Papers using this baseline: Ping et al. (2025) Yan & Zhong (2024) Zhang et al. (2023) Zhang et al. (2024)
- **GeoQA** Chen et al. (2021): GeoQA is a dataset featuring 4,998 multiple-choice geometric problems sourced from real Chinese middle school mathematics examinations. Each problem consists of textual descriptions and associated diagrams categorized primarily into angle calculations, length measurements, and area computations. To facilitate explainable numerical reasoning, the dataset is uniquely annotated with domain-specific programs that represent the sequential mathematical operations required to solve each task.
  Papers using this baseline: Wang et al. (2025a) Liu et al. (2025) Zhang et al. (2025)

3 TEAM

3.1 EXPERTISE

We have the following expertise in the underlying modalities required by this task:

1. Anshul Sawant: Works on training LLMs and designing agentic systems for brand understanding as an Applied Scientist at Amazon. Worked on TPU scheduling optimization at Google. Has experience with Graph Transformer for predicting h/w failures and teaching LLMs to schedule, fine tuning LLMs for automated root cause analysis among other things.
2. Canchen Li: Works on OS software engineering and on-device acoustic model evaluation. Previously has experience in computer vision, reinforcement learning, and variational autoencoders. Holds a Master's degree in Mobile and Internet of Things (IoT) Engineering and dual Bachelor's degrees in Computer Science and Statistics.
3. Joanna Smolska-Moreno: Works as a software engineer with applied machine learning. Currently mostly focused on training small language models (around 1B parameters) and knowledge distillation tasks. Holds a Master's degree in Computer Science.
4. Shilin Ma: Works on data-driven investment strategies. Holds a Master's degree in Computer Science and a Master's degree in Mathematical Sciences. Has experience with computer graphics, modal logic and interactive theorem proving.
5. Steven Liu: Holds a Bachelor's degree in Mathematics-Computer Science. Has an interest in LLM reasoning and uncertainty quantification.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Delong Chen, Mustafa Shukor, Theo Moutakanni, Willy Chung, Jade Yu, Tejaswi Kasarla, Allen Bolourchi, Yann LeCun, and Pascale Fung. Vl-jepa: Joint embedding predictive architecture for vision-language. *arXiv preprint arXiv:2512.10942*, 2025.

Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric Xing, and Liang Lin. GeoQA: A geometric question answering benchmark towards multimodal numerical reasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 513–523, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.46. URL https://aclanthology.org/2021.findings-acl.46/.

Bing Liu, Wenqiang Yv, Xuzheng Yang, Shichang Wang, Junzhuo Liu, Peng Wang, Guoqing Wang, Yang Yang, and Heng Tao Shen. Georef: Referring expressions in geometry via task formulation, synthetic supervision, and reinforced mllm-based solutions. *ArXiv*, abs/2509.21050, 2025. URL https://api.semanticscholar.org/CorpusID:281526109.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Huang Siyuan, Xiaodan Liang, and Song Zhu. Intergps: 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)*, pp. 6774–6786, 01 2021. doi: 10.18653/v1/2021.acl-long.528.

Bowen Ping, Minnan Luo, Zhuohang Dang, Chenxi Wang, and Chengyou Jia. Autogps: Automated geometry problem solving via multimodal formalization and deductive reasoning, 2025. URL https://arxiv.org/abs/2505.23381.

Nannan Shi, Chuanyu Qin, Shipeng Song, and Man Luo. Geothought: A dataset for enhancing mathematical geometry reasoning in vision-language models, 2025. URL https://arxiv.org/abs/2510.21881.

Lihong Wang, Liangqi Li, Weiwei Feng, Jiamin Wu, Changtao Miao, Tieru Wu, Rui Ma, Bo Zhang, and Zhe Li. Virc: Enhancing visual interleaved mathematical cot with reason chunking. *ArXiv*, abs/2512.14654, 2025a. URL https://api.semanticscholar.org/CorpusID:283909230.

Qixun Wang, Yang Shi, Yifei Wang, Yuanxing Zhang, Pengfei Wan, Kun Gai, Xianghua Ying, and Yisen Wang. Monet: Reasoning in latent visual space beyond images and language, 2025b. URL https://arxiv.org/abs/2511.21395.

Shijie Xia, Xuefeng Li, Yixin Liu, Tongshuang Wu, and Pengfei Liu. Evaluating mathematical reasoning beyond accuracy, 2025. URL https://arxiv.org/abs/2404.05692.

Shengyuan Yan and Xiuqin Zhong. Geo-qwen: A geometry problem-solving method based on generative large language models and heuristic reasoning. *2024 21st International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, pp. 1–9, 2024. URL https://api.semanticscholar.org/CorpusID:276452016.

Ming-Liang Zhang, Fei Yin, and Cheng-Lin Liu. A multi-modal neural geometric solver with textual clauses parsed from diagram. In *International Joint Conference on Artificial Intelligence*, 2023. URL https://api.semanticscholar.org/CorpusID:257078982.

Ming-Liang Zhang, Zhongzhi Li, Fei Yin, Liang Lin, and Chenglin Liu. Fuse, reason and verify: Geometry problem solving with parsed clauses from diagram. *ArXiv*, abs/2407.07327, 2024. URL https://api.semanticscholar.org/CorpusID:271088754.

Yuhao Zhang, Dingxin Hu, Ting-Ting Yu, Hao Liu, and Yiting Liu. Geofm: Enhancing geometric reasoning of mllms via synthetic data generation through formal language. *ArXiv*, abs/2510.27448, 2025. URL https://api.semanticscholar.org/CorpusID:282719039.